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

Article title: **Embracing heterogeneity: building 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 major 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. 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 major model supporting innovation in phylogenomics today, the multispecies coalescent model. 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 to each step, are essential for progress.

KEYWORDS: gene flow, genome, multispecies coalescent model, retroelement, speciation, transcriptome.

I. Introduction

Charles Dickens famously wrote in *A Tale of Two Cities* "It was the best of times it was the worst of times." The same could be said about phylogenomics today. Phylogenomics has been invigorated with the introduction of high-throughput sequencing (HTS) and increased breadth of phylogenomic sampling, which have allowed researchers interested in the Tree of Life to scale up in several dimensions, 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 that often appear to 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. These heterogeneous signals 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 Tree of Life. One of the grand challenges of evolutionary biology is deciphering this history, whether at the level of genes, populations, species, or genomes. In this perspective piece, we argue that unabashedly embracing this heterogeneity conceptually and analytically will lead to increased insight into the Tree of Life and its underlying processes.

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.

A general issue in thinking about the future of phylogenomics is – what do researchers want in the realm of phylogenomics? What does society want? Researchers in phylogenomics are motivated by many factors. Some are excited about building the Tree of Life. Others are less interested in the tree itself, but instead are focused on studying conservation, ecological, and evolutionary processes within a phylogenetic framework. Society will likely benefit from results and archiving practices for data and genetic resources that ensure longevity, reproducibility, and relevance to societal problems. Are the priorities that society places on the many disciplines feeding into scientific efforts toward the Tree of Life – 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.

II. Data generation and data types in phylogenomics

One of the fundamental challenges in evolutionary biology is to estimate a Tree of Life for all species. The potential impact of such large phylogenies is reflected in their publication in the highest impact journals, but also in their broad contribution, which extends beyond big data, to methodological innovations and downstream understanding of macroevolutionary processes (e.g., coalescent methods of species tree inference; accounting for hybridization and unsampled species or localities in datasets; understanding community or genome evolution through large-scale phylogenetics). Hence, the phylogenomics community now places a high priority on very large-scale trees, whether in terms of number of taxa, number of genes, or both. The current need for large phylogenies and the high priority placed on them by high-impact journals can result in short-cuts, wherein large-scale phylogenetic trees are cobbled together from disparate existing sources, even taxonomy, but often without hard data behind the placement of many species (Jetz et al., 2012; Zanne et al., 2014; Faurby & Svenning, 2015). At the same time, however, hypothesis-testing in areas such as macroevolution, macroecology, biodiversity, and systematics require these large-scale trees, even as they present challenges being built on high quality data. The phylogenetic knowledge on which we lay a foundation for downstream analyses must be robust, and therefore it is essential that the input phylogenetic hypotheses themselves are robust (Pyron, 2015). Indeed, the current bottlenecks in large-scale phylogenomic data do not appear to be the sequencing, but rather the compilation of high quality, well-curated genomic resources that can fuel phylogenomics for the next century (e.g., Global Genome Initiative, www.mnh.si.edu/ggi/).

Data quality

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

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26449v3 | CC BY 4.0 Open Access | rec: 30 Jan 2018, publ: 30 Jan 2018

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 Polyorphisms (SNP) within loci (Shen, Hittinger & Rokas, 2017), such errors could have damaging consequences for phylogenomic studies, both for topologies and even more so for branch lengths of phylogenetic trees (Marcussen et al., 2014; Bleidorn, 2017).

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). 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 & Castresana, 2007) or TrimAl (Capella-Gutiérrez, Silla-Martínez & Gabaldon, 2009) are gradually being put into practice (Marcussen et al., 2014; He et al., 2016; Irisarri et al., 2017).

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26449v3 | CC BY 4.0 Open Access | rec: 30 Jan 2018, publ: 30 Jan 2018

Coding regions, whether derived from transcriptomes or whole-genome data, are particularly amenable to spot checking of alignments and to filtering out of low-quality data with bioinformatic pipelines (e.g., Dunn et al., 2013; Blom, 2015). Coding regions have the advantage of allowing amino acids to guide alignments, which is particularly useful for highly divergent sequences. Stop codons can help flag errors or genuine pseudogenes. Examining gene tree topologies is also widely used to detect paralogs in phylogenomic data (e.g., Betancur, Naylor & Ortí, 2014). Examining gene trees for aberrantly long branch lengths can also reveal misalignments (e.g., He et al., 2016); sensitivity analyses of various methods for indirectly detecting errors in alignments are sorely needed.

Data generation and marker development

Genome reduction methods: A growing number of genome reduction methods are now providing empiricists with the means to generate genomic subsets suitable for phylogenetic and phylogeographic inference (reviewed by McCormack et al., 2013; Leaché & Oaks 2017; Lemmon & Lemmon, 2013). For phylogenomics, most prominently featured are sequence-capture, focusing on highly conserved regions (e.g., Faircloth et al., 2012; Lemmon, Emme & Lemmon, 2012; reviewed by Jones & Good, 2016) and transcriptomes (e.g., Misof et al., 2014; Cohen et al., 2016; Fernández et al., 2014; Park et al., 2015; Simion et al., 2017; Irisarri et al., 2017), but phylogenomic trees have also been constructed based on restriction-digest methods that focus on single nucleotide polymorphisms or SNPs (Leaché, Chavez & Jones, 2015; Harvey et al., 2016) and analysis of transposable elements (e.g., Suh, Smeds & Ellegren, 2015). This diversity of marker types for phylogenetics should be celebrated, but each marker type brings with it a list of pros and cons. For example, many questions in the higher level phylogenetics of animals and plants have so far relied almost exclusively on transcriptome data, for the simple reason that non-coding portions of the genome are difficult if not impossible to align and analyze. However, the uncritical use of transcriptomes in 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

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26449v3 | CC BY 4.0 Open Access | rec: 30 Jan 2018, publ: 30 Jan 2018

phylogenetic models (Romigiuer 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, 2009b), but it will certainly distort estimated branch lengths when coalescent methods, which assume neutrality, are used. SNPs have been advocated by some authors (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 concatenation of sequence-based markers are misleading (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 oft-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 high 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, in fact was dominated by coding regions. Thus, at least one or the other marker type or their analysis are likely inappropriate when applied across modern birds. These data type effects can stem

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26449v3 | CC BY 4.0 Open Access | rec: 30 Jan 2018, publ: 30 Jan 2018

from multiple sources. Selection on exons might lead to localized differences in effective population size across the genome and previous studies have highlighted 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).

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* selection of loci for phylogenomics is clearly a long-term goal that will yield greater choice and justification for specific loci over the *a priori* selection protocol that now dominates the field. Today, the loci for phylogenomics are based primarily on ease of collection and alignment but many potentially useful regions of the genome are ignored by each specific method, whether it is transcriptomes or hybrid-capture. 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 also allows for further expansion into different research fields and questions based on the same initial data. In constrast, *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 community. However, even low coverage genomes can yield a modest number of markers for phylogenomics, and in the short term might yield data sets allowing a broader diversity of markers for

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26449v3 | CC BY 4.0 Open Access | rec: 30 Jan 2018, publ: 30 Jan 2018

analysis. Although we are fully aware of its constraints, we are particularly excited about the potential that we see in routinely using WGS to assemble phylogenomic data sets.

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 an important theme guiding 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). Scaling up in both dimensions will be crucial for improved phylogenies, and the number of loci required to resolve a given phylogenetic problem, at least in a coalescent framework, is

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26449v3 | CC BY 4.0 Open Access | rec: 30 Jan 2018, publ: 30 Jan 2018

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).

To study how researchers have resolved challenges of balancing numbers of taxa versus numbers of loci, we quantified trends in phylogenomic data set size and structure over the past 13 years, drawing data from 164 data sets across diverse taxa (Supplementary Table S1). We found that, whereas the number of species per paper has not increased significantly over time (Fig. 2A), there were significant increases with time in number of loci (Fig. 2B), total length of sequence analyzed (Fig. 2C), as well as total data set size, as measured by the product of species times locus number (Fig. 2E) or species times total alignment length (Fig. 2F). These trends mirror similar trends evaluated for the size of data sets in phylogeography (Garrick et al., 2015). Surprisingly, we found no evidence for a tradeoff between the number of species investigated and the number of loci analyzed (Fig. 2D); perhaps HTS data sets have plateaued somewhat in terms of number of loci, whereas the number of species analyzed is more a function of the questions being asked and the clade being investigated. Regardless, we suspect that, in general, the number of loci and total alignment lengths in phylogenomic data sets are likely a function of resources and sequencing effort. The era of whole genome sequencing in phylogenomics is still dawning, given that most studies thus far have used targeted approaches for sampling loci (Supplementary Table S1). We suspect that once whole genome sequencing on a clade-wide basis become routine, we will witness yet another jump in the sizes of phylogenomic data sets.

Filtering heterogeneous phylogenomic data sets: Recent studies show that the addition of more loci and more taxa can result in higher levels of gene-tree discordance (e.g., Smith et al., 2015; Shen, Hittinger & Rokas, 2017). This is not unexpected - as the number of taxa and loci increase, the greater the likelihood that the dataset will capture the heterogeneous evolutionary history (e.g., incomplete lineage sorting [ILS], lateral gene transfer [LGT], hybridization, gene duplication and loss [GDL]) and patterns of molecular evolution (e.g., noise/lack of signal in the sequences, and nonstationarity in base composition)

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26449v3 | CC BY 4.0 Open Access | rec: 30 Jan 2018, publ: 30 Jan 2018

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 long branch in a given species phylogeny was a better way to improve phylogenomic signal than selecting genes based on characteristics of sequence evolution. However, the effects of such culling on the distribution of gene trees, and whether it could distort the distribution so that it no longer conforms to models like the multispecies coalescent, are unknown, and potentially of concern (but see Huang et al., 2017). We need further studies on the effects of different types of phylogenomic filters on the properties of large-scale phylogenomic datasets.

Heterozygosity and Intra-Individual Site Polymorphisms: One of the prevalent occurrences in organisms with multiple ploidies are intra-individual polymorphisms and heterozygosity is, of course, common in diploid organisms. However, confident identification of such polymorphism has always been 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 either accommodated using UIPAC ambiguity codes or ignored entirely or by randomly selecting alleles (Iqbal et al., 2012). In fact, most "one allele per individual/species" phylogenomic data sets consists of haplotypes that in fact do not occur in nature, in so far as many methods 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 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 ocurring haplotypes has been shown to improve phylogenomic reconstructions based on genome-scale data (Andermann et al., 2018).

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 in genomes, 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 phylogenomics, since they are generally regarded as highly informative markers, providing strong evidence of homology and monophyly (Boore 2006; Rogozin et al., 2008). With the increased availability and affordability of WGS, our view of genome plasticity has changed drastically in recent years and we are now capable of exploring other genomic features beyond the signals encapsulated in DNA or amino acid sequences. The question then arises of how to identify and utilize these rare genomic markers. Genome-level characters will likely have different evolutionary properties than sequence-based markers, suggesting that one of the biggest challenges we face for incorporating genomic changes into phylogenetic analyses is to find informative evolutionary models and tools suited for these kinds of data.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26449v3 | CC BY 4.0 Open Access | rec: 30 Jan 2018, publ: 30 Jan 2018

Gene order and synteny: Computational algorithms to use gene order and rearrangements as markers in phylogenetics (Tang et al., 2004; Ghiurcuta & Moret, 2014; Kowada et al., 2016) were spurred in part by the seminal paper by Boore, Daehler & Brown (1999) using mitochondrial gene rearrangements to understand the phylogeny of arthropods. Initially, algorithms for making use of gene order and synteny were applied primarily to microbial genomes, but recent efforts have extended such methods to the analysis of eukaryotes as well (see Lin et al., 2013). Gene order and synteny appear most promising at high phylogenetic levels, although we still do not know how informative gene order will be at many levels, such as within mammals. Chromosomal rearrangements appear highly dynamic in some groups, such as mammals, and further study of their use in phylogenomics is warranted (Murphy et al., 2005).

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). 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.

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 human genomes has proven useful in tracking population structure (Sjödin & Jakobsson, 2012), but still remains underappreciated in phylogenetics.

Newly available methods allow inference of CNV at high resolution with great accuracy (Wiedenhoeft, Brugel & Schliep, 2016). The frequency with which CNVs occur in animal and plant populations raises the question of how informative they would be at higher phylogenetic levels, and whether they would incur unwanted homoplasy that would obscure homology and phylogenetic relationships. For example, some CNVs evolve so quickly that they can be used with success at the sub-individual level, for example, in tracking clonal evolution of cancer cells (Schwartz & Schäffer, 2017). Such fast evolution may mean that these markers are less useful at higher levels of biological organization. Additionally, the adaptive nature of CNVs may or may not facilitate clear phylogenetic signals. For example, a study in *Arabidopsis thaliana* (DeBolt, 2010) showed that adaptation to novel cognitive environments, or to varying temperatures, is associated with mutations in CNVs. If CNVs are to become a useful tool in phylogenomics or phylogeography, we must understand their microevolutionary properties in greater detail. For example, the pattern of evolution of CNVs, wherein deletions of genetic

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26449v3 | CC BY 4.0 Open Access | rec: 30 Jan 2018, publ: 30 Jan 2018

material may not easily revert, resulting in a type of Dollo evolution, might help clarify the overall structure of the models applied to them (Rogozin et al., 2006; Gusfield, 2015).

III. Concepts and models in phylogenomics

For decades, phylogenetics has struggled with how best to translate evolutionary changes in DNA sequences and other characters into phylogenies, and genomic data are no exception to this trend. Phylogenomics is still in a developing stage of formulating models that effectively represent the underlying mechanisms for genome-scale variation while remaining efficient and within reasonable analytical and bioinformatic capacities. The current focus on models and evolutionary forces generating the patterns that we recover as branching and reticulation events in our phylogenetic reconstructions is a healthy one, and can be extended to other important topics in phylogenomics, such as species delimitation, character mapping, and trait evolution (e.g., Yang and Rannala 2014). All of these areas are developing rapidly and are in need of updated models and bioinformatics applications to cope with the heterogeneity brought by genome-scale data.

The multispecies coalescent (MSC) model

One of the key practical advances in molecular phylogenetics has been the incorporation of gene tree stochasticity into the inference of species phylogenies, via the multispecies coalescent model (MSC: Rannala & Yang, 2003; Liu & Pearl, 2007; Heled & Drummond, 2010). The MSC allows gene trees to be inferred with their own histories, including coalescent-appropriate branching models, but contained within independent but connected lineages within a species phylogeny, with speciation-appropriate branching models (Degnan & Rosenberg, 2009). The main conceptual advance has been to understand and separately manage the variation at different levels of biological organization – an advance that began years ago (Doyle 1992; Maddison, 1997; Pamilo and Nei 1988), but has only recently been widely embraced and put into practice (Edwards, 2009a). Given its ability to accommodate heterogeneous

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26449v3 | CC BY 4.0 Open Access | rec: 30 Jan 2018, publ: 30 Jan 2018

histories across loci scattered throughout the genome, the MSC lays at the core of the conceptual framework to deal with genome-scale data (e.g., Rannala & Yang, 2008; Liu et al., 2015). In the few instances in which model comparison and fit has been evaluated (Liu and Pearl 2007; Edwards et al. 2007), the MSC vastly outperforms concatenation. This of course does not mean that the MSC is the correct, or even an adequate, model for phylogenomic data, and we need more tests of model adequacy and fit, using Bayesian methods for example (Reid et al., 2014). Despite concerns regarding some of its implementations when dealing with genomic data (e.g., Springer & Gatesy, 2016), there is consensus among systematists that the MSC is a powerful theoretical model for phylogenomics and that there is room for refinement and improvement for its applications (e.g., Edwards et al., 2016b, Xu & Yang, 2016).

Bypassing full likelihood models by relying on summaries of the coalescent process

Given the huge computational difficulties involved in modelling all the complexities of evolutionary processes in a statistical framework, there is interest in methods that will accommodate genome-scale data for large numbers of species. The utility of such methods cannot be overstated: the rapid rise of large-scale genomic data sets has clearly outstripped theoretical and computational methods required to analyze them. For example, although progress is being made regarding scalability of full Bayesian methods of species phylogeny inference (e.g., Ogilvie, Bouckaert & 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 various summaries 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 et al., 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 et al., 2014). For example, MP-EST can sometimes perform poorly when Phyml 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 is used to build gene trees (Xi, Liu & Davis, 2015).

Beyond the multispecies coalescent model

Reticulation at multiple levels challenges the standard multispecies coalescent model

The phylogenetic processes of branching and reticulation can operate at several levels of organization, including within genes, within genomes, and within populations or species (Figs. 3 and 5). For example, recombination can cause reticulations within genes, allopolyploidization can cause reticulations at the level of whole genomes, and introgression and hybridization can cause reticulations at the level of populations. These levels are nested so that branching processes (and in part reticulations) acting at a

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26449v3 | CC BY 4.0 Open Access | rec: 30 Jan 2018, publ: 30 Jan 2018

higher level will cause correlated branching patterns at lower levels. At the same time, reticulations at lower levels, such as recombination acting within genes, will cause inference problems at higher levels, such as estimating population histories. Crucially, however, it is only recombination that will break one key element driving many recent models of phylogenetics and population histories, namely dichotomous gene trees. Reticulations at levels of organization higher than the genome, such as the fusing of populations, as well as gene duplication, will still yield collections of dichotomous gene trees, even if the higher-level history is reticulated. Ultimately, the additive effects of these reticulate processes result in our observed phylogenetic reconstructions, and we expect all of these scenarios to produce bifurcating, dichotomous gene trees. From a modelling point of view, another key distinction is whether at the species level, we still have a phylogeny that is tree-like, or whether a network is needed. The process whereby two populations jointly produce a third requires a network to model properly. Allopolyploidy is another situation requiring a network. There are several statistical methods for inferring homoploid networks (Yu et al., 2014; Solis-Lemus & Ané, 2016; Wen et al., 2016; Wen & Nakhleh, 2018), species histories under allopolyploidy (Jones, Sagitov & Oxelman, 2013), and some two-step methods such as PADRE (Huber et al., 2006; Lott et al., 2009). In general, dealing with multiple simultaneous violations of the MSC, such as introgression and allopolyploidy, remains challenging. It is likely that the history of many radiations involves parts of the genome with a dichotomous history and parts that exhibit reticulation, demanding methods that accommodate both scenarios. Alternatively, rather than 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 are contributing 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 might be identified (e.g., separated from loci with horizontal gene transfer), using the newly developed program CLASSIPHY (Huang et al., 2017).

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26449v3 | CC BY 4.0 Open Access | rec: 30 Jan 2018, publ: 30 Jan 2018

Models accommodating a 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 et al. (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 sub-problem is estimating the species phylogeny despite some 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 2017).

In some cases, it is possible to model the same situation with either a species network or a tree with gene flow. Long (1991) discussed two models of admixture: Intermixture and gene flow, illustrated in Figure 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; Zhang et al., 2018), much more work is needed to bring together these two lines of work. Simulations and comparisons of observed and expected summary statistics, such as the site-frequency spectrum (Excoffier et al., 2013), have proven especially useful in distinguishing such scenarios (Fig. 7).

Reticulation in the form of gene flow or introgression is probably the most difficult violation of the MSC to address, in part because the number of potential trees accommodating a reticulating network is even higher than the already high number of trees for a given number of taxa. There is at least one issue where reticulation presents an opportunity as well as a challenge. Any kind of gene flow/hybridization means that there is the possibility of inferring the existence of extinct species, because extinct species

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26449v3 | CC BY 4.0 Open Access | rec: 30 Jan 2018, publ: 30 Jan 2018

contribute novel alleles that exceed the coalescence time of most alleles in the focal species under study (Hammer et al., 2011). Well-known examples are the documented presence of Neanderthal genes in most human genomes due to introgression (e.g., Meyer et al., 2012) and the presence of genomes derived from now-extinct diploids in extant allopolyploids (i.e. meso-allopoloploids; e.g., Mandáková et al., 2010; Marcussen et al., 2015). Some current models can explain the data as containing genetic information from extinct species, but they do not model the full species phylogeny: such a generalized approach seems a promising avenue to explore.

Polyploidy and the challenges of analyzing gene duplication and loss

The MSC model describes well allelic lineages and the mutations they accumulate (Fig. 3; Degnan & Rosenberg, 2009; Liu, Xi & Davis, 2015). The simple MSC model is challenging to apply to evolutionary events in which the evolving entities (genes or paralogs) duplicate and occasionally go extinct during the evolutionary history of the populations/species and thus cannot be sampled in contemporary population or species. Estimating the existence and number of these "ghost" lineages remains challenging. For example, how can we detect duplication events if one of the duplicated loci is lost in descendant lineages? In the 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,

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26449v3 | CC BY 4.0 Open Access | rec: 30 Jan 2018, publ: 30 Jan 2018

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, and ordinary MSC inference undertaken. 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 limited number of topological outcomes and branch lengths limited by effective population size, which is not the case for duplicated genes. A recent approach to identifying genes that are single copy, but have nonetheless been affected by GDL, was made using the genomic location of the loci (Sousa et al., 2017), and could prove useful for distinguishging GDL and ILS.

Recombination

All existing methods for coalescent estimation of species trees and networks make two important assumptions, namely that (1) there is free recombination between loci, and (2) there is no recombination within a locus. These two assumptions address a key concept distinguishing MSC models from concatenation or supermatrix models: it is the conditional independence of loci, mediated by recombination between loci, and not the ability to address ILS or discordance among genes *per se*. Moving forward, three important questions to address are: (1) How robust are methods to the presence of recombination within loci and/or to the violation of independence among loci? (2) How should we model recombination within the species phylogeny inference framework? and (3) How do we detect it and differentiate recombination-free loci?

Researchers have started to examine the first question and found a detectable effect of recombination only under extreme levels of ILS and gene tree heterogeneity (e.g., Lanier & Knowles, 2012). However, more analyses and studies are still needed to explore a wider range of factors and parameters that could affect species phylogeny inference when the assumption of recombination-free loci is violated. For answering the second question, one approach involves combining the multispecies coalescent with hidden Markov models (e.g., Hobolth et al., 2007). These methods suffer from the "state explosion problem", where individual states are needed for the different coalescent histories, and they increase rapidly with the number of taxa in the dataset, making them infeasible except for very small (~4 taxa) datasets. New methods that scale to larger datasets are needed if such approaches are to be useful in practice. A different direction is to devise novel methods for inferring species phylogeny while assuming that the genealogies of the individual loci could take the form of an ancestral recombination graph (ARG: Siepel, 2009).

Extending these approaches to address recombination would require the development of new models that significantly extend the multispecies coalescent to account for ARGs within the branches of a species phylogeny. For two-step species tree methods, this entails developing new methods that infer ARGs for the individual loci and methods that infer species phylogenies from collections of ARGs. For single-step (Bayesian) methods, novel developments are needed to sample species phylogenies, locus-specific ARGs, and their related parameters. It will also be important to better understand the conditions under which ignoring recombination will still yield reasonable estimates of phylogeny. Extending the theory to accommodate ARGs may be of intrinsic interest, but if the parameter space in which recombination is relevant is very small, then practitioners may be able to ignore recombination.

Species concepts and delimitation

Coalescent methods have played an important role in the development and critical evaluation of species delimitation methods because they provide hypotheses for species boundaries based on genetic and can now be integrated with phenotypic data (e.g., Solis-Lemus et al., 2015). Irrespective of traditional species concepts, it is essential that the entities at the tips of the species tree do not violate the assumptions of the MSC, wherein the definition of species is mathematically clear-cut (e.g., Rannala & Yang, 2003, Degnan & Rosenberg, 2009): the branches of the species tree constitute species or populations that do not exchange genes. However, the MSC model also carries strict assumptions about the divergence process if the delimited units are to be interpreted as species. Specifically, it is important to emphasize that in the "standard" MSC model, these species represent populations that, immediately after divergence, no longer experience gene flow. Therefore, the species of the MSC model do not necessarily correspond with species as a taxonomic rank, defined by traditional species concepts (Heled & Drummond, 2010): "MSC" species could simply be populations by other criteria, so long as they have ceased to exchange genes, even for a short period of 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 (Sukumaran & Knowles 2017). Therefore, in those species-phylogeny 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, 2016). 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 taking into account 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 information which are synthesized in an integrative fashion (Dayrat, 2005; Will, Mishler & Wheeler, 2005; Bacon et al., 2012; Solis-Lemus, et al., 2015), and refer to the reproductively isolated units of MSC analysis as "MSC units" or "MSC taxa".

IV: Models at the intersection of phylogenomics, phylogeography, and macroevolution

Phylogenomics and macroevolution represent two ends of a research spectrum, with one end focusing on building phylogenies and the other end on using them. In many important respects, these two sub-disciplines have remained distinct and non-communicative. On the one hand, phylogenomics and phylogeography have not exhaustively aimed to address the type of questions - related to diversification and trait evolution - that macroevolution focuses on. On the other hand, macroevolution ignores many kinds of complexities inherent to the phylogeny building process that phylogenomics has recently begun to address.

Macroevolutionary models focus on long-term processes, in terms of both species richness and phenotypic diversity. They rely on two types of models: birth-death models of diversification aimed at understanding how and why speciation and extinction rates vary through time and across lineages (Hey 1992; Nee, Mooers & Harvey, 1992; see Stadler 2013 and Morlon, 2014 for review) and models of trait evolution aimed at understanding the mode and tempo of phenotypic evolution (Felsenstein, 1973; see Pennell & Harmon, 2013 and Manceau, Lambert & Morlon, 2017 for reviews). These models are typically constructed at the level of species, ignoring the populations or individuals that constitute these species (but see Manceau, Lambert & Morlon, 2015 and Rosindell, Harmon & Etienne, 2015 for exceptions). As a consequence, microevolutionary processes such as coalescence have informed phylogenetic methods for building phylogenies more so than have macroevolutionary methods that use them. For example, the most widely used phylogenetic dating methods generally do not acknowledge the critical distinction between speciation times, which are usually of primary interest, and coalescence times, which are often assumed to represent speciation times but in fact represent events older than the divergence of the species concerned (Edwards & Beerli, 2000; dos Reis, Donoghue & Yang, 2016; Angelis & dos Reis 2015). In addition, macroevolutionary models are fit to species phylogenies (diversification models) or a combination of species phylogenies and phenotypic data (trait evolution models), most often assuming that evolution is best represented by a species tree, not a network (but see

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26449v3 | CC BY 4.0 Open Access | rec: 30 Jan 2018, publ: 30 Jan 2018

Jhwueng & O'Meara, 2015; Bastide et al., 2017; Solis-Lemus et al., 2017 for models of trait evolution on networks), and that the species phylogeny is known. Nearly all models that use phylogenies to study character evolution assume a single underlying species phylogeny on which characters evolve. But it has become evident recently that different characters often might in principle have different phylogenies, for the same reason that genes themselves might have different phylogenies (Hahn & Nahkleh, 2016). Analyzing incongruences between character evolution inferred from the species tree versus from gene trees that are more directly linked to the character under study would provide a refined understanding of character evolution. Recent work on the phylogeny of quantitative characters may be helpful in this endeavor (Felsenstein 2012).

Developing research projects that integrate the heterogeneity currently experienced by phylogenomics and macroevolution will bring important new insights into the evolutionary process. For example, developing diversification and phenotypic evolution models to be fit to networks rather than dichotomous trees will allow estimates of rates of hybrid speciation and phenotypic evolution as well as a better understanding of factors influencing such rates (see Bastide et al., 2017). Embracing genetic heterogeneity and the incongruence between gene trees and species phylogenies when applying macroevolutionary models could help us to better understand how speciation proceeds, and also to analyze the coupling between genetic and phenotypic evolution (e.g., is phenotypic convergence coupled or not with genetic convergence in relevant genes?). Developing macroevolutionary models accounting for within-species heterogeneity linked to biogeography could help us understand how biogeographic structuring influences speciation, extinction, and phenotypic evolution.

More generally, evolutionary biologists have not yet thought much about the type of new questions that we are going to be able to address if we are given genomic data at the tips of all species from a phylogeny. Such data could allow us to gain an integrative understanding of three fundamental aspects of evolution: evolution at the molecular level, at the phenotypic level, and at the clade level, as well as the links among them. Are rates of evolution at these three levels correlated? If so, how? Do

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26449v3 | CC BY 4.0 Open Access | rec: 30 Jan 2018, publ: 30 Jan 2018

features of genomes or of genome evolution, such as quantity of transposable elements, substitution rates, number of gene duplications, influence rates of diversification and phenotypic evolution? Clearly, we are only at the beginning of exploring these new possibilities.

Mapping trait evolution on heterogeneous genomic datasets

Mapping the genomic basis of phenotypic traits is a major trend in evolutionary biology today (Elmer & Meyer, 2011; Hoban et al., 2016). Such mapping can be conducted in the context of populations of a single species or, increasingly, via comparisons of species on a phylogeny (e.g., Hiller et al., 2012; Marcovitz, Jia & Bejerano, 2016). Phylogenetic genome-wide association studies ("PhyloGWAS") methods identify genomic features in coding or non-coding DNA that exhibit unusual patterns of evolution on branches concerned with repeated evolution of phenotypes, thereby drawing connections between the genomic and phenotypic levels (Pease et al., 2016). Such phylogenomic mapping usually assumes a single phylogeny, the species phylogeny, as a framework for analysis, and therefore ignores genomic heterogeneity. To make phyloGWAS mapping most efficient it might be more appropriate to use the local topology in the genome for inference and estimation of ancestral states. Estimating genotype-phenotype associations solely on the species phylogeny might yield misleading results regarding the origin and evolution of phenotypic traits (Hahn & Nakleh, 2016). Heterogeneity across gene histories has been traditionally considered as "biological noise" when using comparative genomics to map traits, but of course such heterogeneity is the focus of gene mapping efforts at lower taxonomic levels. Genome-wide or gene-specific selective sweeps associated with the evolution of a particular phenotypic trait are a major source of genetic heterogeneity among closely related populations or species, and can be captured using outlier statistics, such as Fst or Dxy (Pease et al., 2016). Such selective sweep mapping of genes with large phenotypic effect can now be accomplished with high resolution and precision in genomically poorly studied organisms (Lamichhaney et al., 2015). Apart from providing valuable knowledge on the genetic basis of trait diversification, such data are providing increasing support to the fact that cases of

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26449v3 | CC BY 4.0 Open Access | rec: 30 Jan 2018, publ: 30 Jan 2018

genetic heterogeneity can be profitably used in the effort to understand and resolve evolutionary history, rather than considering it "biological noise." Such thinking needs to be incorporated into comparative genomics more frequently.

Tree-free methods of character evolution

We have seen that incorporating phylogenetic heterogeneity is a challenge for macroevolutionary models of character evolution. At the other end of the spectrum are a class of methods (so called "tree-free methods") that attempt to draw inferences and principles about trait evolution without assuming a particular phylogeny. The common situation when analyzing character or trait data correlated by a phylogeny is to assume a stochastic process for the trait, commonly a variation of the Brownian motion (BM; Felsenstein, 1985) or Ornstein-Uhlenbeck (OU; Hansen, 1997) processes. Then, using the estimated phylogeny and measured trait data for each species, the parameters of various evolutionary processes – trait variation, patterns and rates of change, etc. - are estimated, often using maximum-likelihood or Bayesian approaches (see Pennell & Harmon, 2013 and Manceau, Lambert & Morlon, 2017 for reviews). However, given the various logistical and technical challenges of inferring robust phylogenies, exploring tree-free methods might represent a useful mechanism for guiding the study of character evolution for certain groups.

Tree-free comparative methods work by integrating over the space of trees (under a given branching process model). For example, under a pure birth model and with enough tip measurements, the optimum value of the OU process can be estimated as the sample average (Bartoszek & Sagitov, 2015a). 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 (Sagitov and Bartoszek 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.

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 genomes 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., 2015) have opted 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, 2011). While such approaches elicit a culture clash between those who laboriously build trees and those that simply use them, there are other approaches stemming from macroevolution that are less offending to phylogeny builders. For example, recent results using conditioned birth-death processes (e.g., Gernhard, 2008a; b; Sagitov & Bartoszek, 2012) show that under constant rate processes the size of the clade contributes information on the height of the tree and also on the coalescence times. Such results can be used to improve the calibration and node dating of the phylogeny when some species are not sampled. One would expect that ignoring the non-sequenced species would incur a bias resulting in shorter tree heights, because less time is usually required to generate fewer tips. Conditioned branching process models can help alleviate this bias. Also, macroevolutionary birth-death models are used as branching process priors in Bayesian molecular dating. The availability of likelihood expressions for incompletely sampled phylogenies (Stadler 2009; Stadler & Steel 2012; Morlon et al, 2011) thus allow to date

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26449v3 | CC BY 4.0 Open Access | rec: 30 Jan 2018, publ: 30 Jan 2018

phylogenies while accounting for the fact that we have observed only a certain fraction of unsampled species.

V. Building, updating and sustaining the Tree of Life

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 Tree of Life (Mora et al., 2011; McTavish et al., 2017). One common challenge that phylogeneticists encounter towards that end is the difficulty in accessing samples from rare, endangered, or extinct taxa, particularly in countries where collecting and exporting is not possible. Recent genomic techniques now allow successful results in obtaining valuable 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 enhance research in phylogenomics and phylogeography.

Despite the great increase in the generation of genomic data across organisms, we are often forced to use simpler, less realistic phylogenetic methods and assumptions to deal with large, heterogeneous datasets. For instance, the popular phylogenetic software program *BEAST (Heled & Drummond, 2010) is not capable of dealing with more than a few hundred taxa and some dozen loci at a

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26449v3 | CC BY 4.0 Open Access | rec: 30 Jan 2018, publ: 30 Jan 2018

time for a common analysis, and only recently the release of StarBeast2 allows for the use of thousands of loci for tens of taxa (Ogilvie, Bouckaert & Drummond, 2017). To tackle this problem, we encourage the continuing development of methods that are fully scalable and ideally only increase analytical time linearly rather than exponentially with the number of taxa and loci. Phylogenetic methods should also be fully parallelizable (in order to run natively in computer clusters) and contain checkpoints, i.e., be able to resume the analyses from the latest logged file in case an analysis crashes or the user wishes to evaluate partial results. Another point of possible improvement is in dealing with new sequences to be added to a previously large dataset: should the analysis start from scratch, or could there be substantial time gains by letting those sequences find their placement in the phylogeny 'on the fly'?

Large scale phylogenies should ideally be based on the best (or most comprehensive) available datasets in terms of taxonomic and molecular sampling and be constructed from the data itself. However, even supermatrix inference conducted under a single analysis can add bias on tree heights and coalescence times when performed across unbalanced sampled clades (a very common case for species-rich clades or understudied taxa), and therefore affect downstream analyses that rely on these parameters (e.g., biogeography, trait evolution, diversification rates). Computing optimally populated datasets that combine the largest number of taxa and loci simultaneously is a complex mathematical problem, but recent approaches (e.g., SUPERSMART; Antonelli et al., 2017) attempt to overcome it objectively, such as applying the knapsack problem to phylogenetics by packing the optimal choice of species and suitable alignments into a minimally sparse supermatrix.

Community initiatives

Building the Tree of life is a grand challenge in molecular phylogenetics, and one that cannot be accomplished by a single person or institution's efforts. Several initiatives have been developed in recent years to coordinate efforts and provide the research community with synthetic information. A prominent project is the Open Tree of Life (https://tree.opentreeoflife.org/; Hinchliff et al., 2015). This project

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26449v3 | CC BY 4.0 Open Access | rec: 30 Jan 2018, publ: 30 Jan 2018

provides a synthesis of previously published phylogenies merged through supertree and other grafting methods. One issue faced by the initiative is that it relies on authors uploading their phylogenetic trees to open data repositories, such as Dryad Data Repository (http://datadryad.org/pages/organization; Vision, 2010) or TreeBase (Sanderson et al., 1994; Piel et al., 2009), which at least until recently only occurred in about 17% of cases (Drew, 2013). Substantial curatorial efforts are also critical to facilitate reusability of deposited trees (McTavish et al., 2015). A different approach was taken by Antonelli et al. (2017), who developed a framework for continuously inferring time-calibrated large phylogenies from raw sequence data deposited in GenBank (Clark et al., 2016) in a multi-step method. Similarly, various tools have been developed to make information contained in the Tree of Life available for the general public (e.g., Rosindell & Harmon, 2012; Harmon et al., 2013).

Mapping the Tree of Life

While progress has been made in mapping species distributions at the large scale aiming for improved conservation practices (e.g., the Map of Life collaborative project; https://mol.org/), most initiatives do not map the tips of phylogenetic trees directly onto the geographic space, and therefore are limited by current taxonomic knowledge. As spatial variation in biodiversity results from interactions between evolutionary history and environmental factors, explicit connections between the tips of the Tree of Life and geographic ranges will greatly improve biogeographic inferences (Quintero et al., 2015) and our understanding of biodiversity patterns and future trends. Advances in mapping the Tree of Life through earth history using genomic-based phylogenetic inferences over broad scales and explicit spatial models (e.g., geophylogenies and continuous diffusion models: Kidd, 2010) depend directly on locality data that should be made available in raw and ready-to-use formats. Data sharing policies for associated data, such as geographic coordinates and voucher information, is not well established among journals. We argue that editorial boards should try as best as possible to establish data policies that value and encourage the

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26449v3 | CC BY 4.0 Open Access | rec: 30 Jan 2018, publ: 30 Jan 2018

deposit of geographic data associated to vouchered specimens and other associated information available for future reference.

Best practices for building the Tree of Life

Data must be well curated and publicly available

As we are now entrenched into the era of big data in biological sciences, adequate reproducibility must be a fundamental endeavor of biodiversity research. Therefore, data publication in open-access repositories represents a powerful tool that not only ensures long-term storage and public availability for future research, but also serves as a vehicle for clarifying intellectual rights and scientific merits (Costello & Wieczorek, 2014). The exponential growth in the amount of genomic scale data and the increased dependence on the availability of each other's' data to answer complex biological questions means that there is a need for improved data management, analysis, and accessibility. Biocuration, the activity of organizing, representing, and making biological information accessible to both biologists and bioinformaticians, has now become an important consideration in building, updating, and sustaining the Tree of Life (McTavish et al, 2017). GenBank has been the main open access repository for annotated collections of publicly available molecular data. Although the data stored in this database usually lists information such as organism of origin and publication details, the utility of molecular data in this database to answer multiple biological questions, such as biogeographic patterns of biodiversity, is often hampered by lack of associated information such as collection locality or attachment to a specific voucher specimen. We propose two urgent actions to advance this key field. First, authors should be encouraged to submit molecular data that is linked to voucher specimens deposited in recognized scientific collection and museums. Second, authors, journals, and curators should encourage all molecular data submitted to include information such as collection locality and details of voucher specimens. In this regard, other global initiatives such as the International Barcode of Life Project (iBOL; http://www.ibolproject.org) have had great success linking molecular data with morphological and distributional data. When all the

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26449v3 | CC BY 4.0 Open Access | rec: 30 Jan 2018, publ: 30 Jan 2018

data produced or published are curated to high standards and made accessible as soon as available, biological research will be able to process massive amounts of complex data much more quickly.

Submitting sequence and tree data during publication is now routine. However, making available all analytical methods such as software and code used to process and analyze data is less widely employed by the phylogenetic community. Facilities such as TreeBase, Dryad Digital Repository, and Github (https://github.com/) provide a platform for the curated storage of the data and bioinformatic pipelines underlying the scientific literature (see McTavish et al., 2015; 2017). Authors and journals should require all published research to include links to raw data, processed data, and all analytical methods used to produce the results presented. In general, we advocate for following best practices of data management and publication to ensure the quality and utility of phylogenomic data and their associated biological information (see Costello & Wieczorek, 2014 for a review). In putting together Figure 2, for example, we found that basic information on a given phylogenomic study, such as the number of species or sequences analyzed, or the total number of base pairs in an alignment, were often not reported or difficult to recover; including such information in easy-to-access tables prior to article acceptance would greatly facilitate meta-analyses and syntheses as the number of studies grows (Supplementary Table S1).

The need of adequate curation of analytical tools

In the same way that data must be adequately stored and curated, analytical tools must be available for future use and should guarantee proper reproducibility (Wilson et al., 2014). 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

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26449v3 | CC BY 4.0 Open Access | rec: 30 Jan 2018, publ: 30 Jan 2018

resources. 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). 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 Tree of Life (McTavish et al., 2017).

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 Tree of Life. 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 Tree of Life, without hurting lower impact research and the ongoing search for innovation.

Moreover, because building the Tree of Life 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,

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26449v3 | CC BY 4.0 Open Access | rec: 30 Jan 2018, publ: 30 Jan 2018

field biologists that obtain basic natural history information and specimens used for building the Tree of Life (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.

VI. Conclusions

In this perspective, we have attempted to cover ground in the vast arena of issues facing modern phylogenomics today. We have seen how genome-scale phylogenomics, currently on a strong footing as a result of the multispecies coalescent model, is increasingly infiltrated by models that recognize reticulate processes, such as recombination and introgression. By 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 Tree of Life. 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 tree of life 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 and Tobias Andermann provided useful comments that improved the contents of this manuscript.

FUNDING STATEMENT

The Gothenburg Center for Advanced Studies (GoCas) workshop 'Origins of Biodiversity' was funded by Chalmers University of Technology and the University of Gothenburg. The following researchers are supported by scholarship or research grants from the following agencies: Swedish Research Council (B.O., A.A.), US National Science Foundation, the European Research Council under the European Union's Seventh Framework Programme (FP/2007-2013, ERC Grant Agreement n. 331024 to A.A.), the Swedish Foundation for Strategic Research and a Wallenberg Academy Fellowship (A.A.). F.P.W. would like to thank Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Partnerships for Enhanced Engagement in Research from the U.S. National Academy of Sciences, the U.S. Agency of International Development (PEER NAS/USAID), and the L'Oreal-Unesco For Women in Science Program.

AUTHOR CONTRIBUTIONS

B. O. and S. V. E. conceived and led the project; G. A. B. and S. V. E. compiled and coordinated writing the manuscript; all authors participated in the discussions held during May 15–19, 2017 under the 'Origin of Biodiversity' Workshop in Göteborg, Sweden, read, and approved the final version submitted for publication.

CONFLICT OF INTEREST

The authors declare no conflict of interests.

REFERENCES

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

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

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

Albalat, R., Cañestro, C. 2016. Evolution by gene loss. Nature Reviews Genetics 17: 379–391.

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.

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

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. BioRxiv doi: https://doi.org/10.1101/255752.

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.

Ashfield, T., Egan, A. N., Pfeil, B. E., Chen, N. W. G., Podicheti, R., Ratnaparkhe, M. B., Ameline-Torregrosa, C., Denny, R., Cannon, S., Doyle, J. J., Geffroy, V., Roe, B. A., Saghai-Maroof, M. A., Young, N. D., Innes, R. W. 2012. Evolution of a complex disease resistance gene cluster in diploid Phaseolus and tetraploid Glycine. Plant Physiology 159: 336–354.

Ashkenazy, H., Cohen, O., Pupko, T., Huchon, D. 2014. Indel Reliability in Indel-Based Phylogenetic Inference. Genome Biology and Evolution 6: 3199–3209.

Bacon, C. D., McKenna, M. J., Simmons, M. P., Wagner, W. L. 2012. Evaluating multiple criteria for species delimitation: an empirical example using Hawaiian palms (Arecaceae: *Pritchardia*). BMC Evolutionary Biology 2012: 12–23.

Baker, A. J., Haddrath, O., McPherson, J. D., Cloutier, A. 2014. Genomic Support for a Moa-Tinamou Clade and Adaptive Morphological Convergence in Flightless Ratites. Molecular Biology and

Evolution 31: 1686–1696.

Baird, N. A., Etter, P. D., Atwood, T. S., Currey, M. C., Shiver, A. L., Lewis, Z. A., Selker, E. U., Cresko, W. A., Johnson, E. A. Rapid SNP discovery and genetic mapping using sequenced RAD markers. PLoS One 3(10):e3376.

Bartoszek, K. 2014. Quantifying the effects of anagenetic and cladogenetic evolution. Mathematical Biosciences 254: 42–57.

Bartoszek, K. 2016. Phylogenetic effective sample size. Journal of Theoretical Biology 407: 371–386.

Bartoszek, K., Sagitov, S. 2015a. Phylogenetic confidence intervals for the optimal trait value. Journal of Applied Probability 52: 1115–1132.

Bartoszek, K., Sagitov, S. 2015b. A consistent estimator of the evolutionary rate. Journal of Theoretical Biology 371: 69–78.

Bastide, P., Solis-Lemus, C., Kriebel, R., Sparks, K. W., Ané, C. 2017. Phylogenetic Comparative Methods on Phylogenetic Networks with Reticulations. BioRxiv doi: https://doi.org/10.1101/194050

Baurain, D., Brinkmann, H., Philippe, H. 2006. Lack of Resolution in the Animal Phylogeny: Closely Spaced Cladogeneses or Undetected Systematic Errors? Molecular Biology and Evolution 24: 6–9.

Belyaev, D. K. 1969. Domestication of animals. Science Journal 4: 47–52.

Betancur, R., Naylor, G. J. P., Ortí, G. 2014. Conserved genes, sampling error, and phylogenomic inference. Systematic Biology 63: 257–262.

Bleidorn, C. 2017. Sources of Error and Incongruence in Phylogenomic Analyses. In: Phylogenomics. Cham: Springer International Publishing, 173–193.

Blom, M. P. K. 2015. EAPhy: A Flexible Tool for High-throughput Quality Filtering of Exon-alignments and Data Processing for Phylogenetic Methods. PLoS ToL.

Boore, J. L., Collins, T. M., Stanton, D., Daehler, L. L., & Brown, W. M. 1995. Nature. 376: 163–165.

Boore, J. L., Daehler, L. L., Brown, W. M. Complete sequence, gene arrangement, and genetic code of mitochondrial DNA of the cephalochordate *Branchiostoma floridae* (Amphioxus). Molecular

Biology and Evolution 16: 410–418.

Boore, J. L. 2006. The use of genome-level characters for phylogenetic reconstruction. Trends in Ecology and Evolution 21:439–446

Bryant, D., Bouckaert, R., Felsenstein, J., Rosenberg, N. A., RoyChoudhury, A. 2012. Inferring species trees directly from biallelic genetic markers: bypassing gene trees in a full coalescent analysis. Molecular Biology and Evolution 29: 1917–1932.

Buerki, S., Baker, W. J. 2016. Collections-based research in the genomics era. Biological Journal of the Linnean Society 117: 5–10.

Burbrink, F. T., Pyron, R. A. 2011. The Impact of Gene-Tree/Species-Tree Discordance on Diversification-Rate Estimation. Evolution 65: 1851–1861.

Capella-Gutierrez, S., Silla-Martinez, J. M., Gabaldon, T. 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics 25: 1972–1973.

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

Chakrabarty, P. 2010. Genetypes: a concept to help integrate molecular phylogenetics and taxonomy. Zootaxa 2632: 67–68.

Chen, M. Y., Liang, D., Zhang, P. 2015. Selecting Question-Specific Genes to Reduce Incongruence in Phylogenomics: A Case Study of Jawed Vertebrate Backbone Phylogeny. Systematic Biology 64: 1104–1120.

Chifman, J., Kubatko, L. 2014. Quartet inference from SNP data under the coalescent model. Bioinformatics 30: 3317–3324.

Churakov, G., Sadasivuni, M. K, Rosenbloom, K. R., Huchon, D., Brosius, J., Schmitz, J. 2010. Rodent Evolution: Back to the Root. Molecular Biology and Evolution 27: 1315–1326.

Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., Sayers, E.W. 2016. GenBank. Nucleic Acids Research 44: D67–D72.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26449v3 | CC BY 4.0 Open Access | rec: 30 Jan 2018, publ: 30 Jan 2018

Cohen, O., Doron, S., Wurtzel, O., Dar, D., Edelheit, S., Karunker, I., Mick, E., Sorek, R. 2016. Comparative transcriptomics across the prokaryotic Tree of Life. Nucleic Acids Research 44: W46–W53.

Costello, M. J., Wieczorek, J. 2014. Best practice for biodiversity data management and publication. Biological Conservation 173: 68–73.

Crawford, F. W., Suchard, M. A., 2013. Diversity, disparity, and evolutionary rate estimation for unresolved Yule trees. Systematic Biology 62: 439–455.

Cutter, A. D. 2013. Integrating phylogenetics, phylogeography and population genetics through genomes and evolutionary theory. Molecular Phylogenetics and Evolution 69: 1172–1185.

Dalquen, D. A., Zhu, T., Yang, Z. 2017. Maximum Likelihood Implementation of an Isolation-with-Migration Model for Three Species. Systematic Biology 66: 379–398.

Dayrat, B. 2005. Towards integrative taxonomy. Biological Journal of the Linnean Society 85: 407–415

De Maio, N., Schlötterer, C., Kosiol, C. 2013. Linking Great Apes Genome Evolution across Time Scales Using Polymorphism-Aware Phylogenetic Models. Molecular Biology and Evolution 30: 2249–2262.

DeBolt, S. 2010. Copy Number Variation Shapes Genome Diversity in Arabidopsis Over Immediate Family Generational Scales. Genome Biology and Evolution 2: 441–453.

Degnan, J. H., Rosenberg, N. A. 2009. Gene tree discordance, phylogenetic inference and the multispecies coalescent. Trends in Ecology & Evolution 24: 332–340.

Dell Ampio, E., Meusemann, K., Szucsich, N. U., Peters, R. S., Meyer, B., Borner, J., Petersen, M., Aberer, A. J., Stamatakis, A., Walzl, M. G., Minh, B. Q., Haeseler, von A., Ebersberger, I., Pass, G. N., Misof, B. 2013. Decisive Data Sets in Phylogenomics: Lessons from Studies on the Phylogenetic Relationships of Primarily Wingless Insects. Molecular Biology and Evolution 31: 239–249.

Delsuc, F., Brinkmann, H., Philippe, H. 2005. Phylogenomics and the reconstruction of the Tree of Life. Nature Reviews Genetics 6: 361–375.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26449v3 | CC BY 4.0 Open Access | rec: 30 Jan 2018, publ: 30 Jan 2018

dos Reis, M., Donoghue, P. C. J., Yang, Z. 2016. Bayesian molecular clock dating of species divergences in the genomics era. Nature Reviews Genetics 17:71–80.

Douzery, E. J. P., Scornavacca, C., Romiguier, J., Belkhir, K., Galtier, N., Delsuc, F., Ranwez, V. 2014. OrthoMaM v8: A Database of Orthologous Exons and Coding Sequences for Comparative Genomics in Mammals. Molecular Biology and Evolution 31:1923–1928.

Doyle, J. J. 1992. Gene trees and species trees: molecular systematics as one character taxonomy. Systematic Botany 17: 144-163.

Drew, B. T. 2013. Data deposition: Missing data mean holes in Tree of Life. Nature 493: 305.

Dunn, C. W., Howinson, M., Zapata, F. 2013. Agalma: an automated phylogenomics workflow. BMC Bioinformatics 14: 330.

Edwards, S. V., Beerli, P. Perspective: gene divergence, population divergence, and the variance in coalescence time in phylogeographic studies. Evolution 54: 1839–1854.

Edwards, S. V., Cloutier, A., Baker, A. J. 2017. Conserved Nonexonic Elements: A Novel Class of Marker for Phylogenomics. Systematic Biology 66: 1028–1044.

Edwards, S. V., Potter, S., Schmitt, C. J., Bragg, J. G., Moritz, C. 2016a. Reticulation, divergence, and the phylogeography-phylogenetics continuum. Proceedings of the National Academy of Sciences 113: 8025–8032.

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

Edwards, S. V. 2009a. Is a new and general theory of molecular systematics emerging? Evolution 63:1–19.

Edwards, S. V. 2009b. Natural selection and phylogenetic analysis. Proceedings of the National Academy of Sciences of the United States of America 106: 8799–8800.

Elmer, K. R., Meyer, A. 2011. Adaptation in the age of ecological genomics: insights from parallelism and convergence. Trends in Ecology and Evolution 26: 298–306.

Excoffier, L., Dupanloup, I., Huerta-Sánchez, E., Sousa, V. C., Foll, M. 2013. Robust demographic inference from genomic and SNP data. PLoS Genet 9: e1003905.

Faircloth, B. C., McCormack, J. E., Crawford, N. G., Harvey, M. G., Brumfield, R. T., Glenn, T. C. 2012. Ultraconserved Elements Anchor Thousands of Genetic Markers Spanning Multiple Evolutionary Timescales. Systematic Biology 61: 717–726.

Faurby, S., Svenning, J. C. 2015. A species-level phylogeny of all extant and late Quaternary extinct mammals using a novel heuristic-hierarchical Bayesian approach. Molecular Phylogenetics and Evolution 84: 14–26.

Felsenstein, J. 1973. Maximum-likelihood estimation of evolutionary trees from continuous characters. American Journal of Human Genetics 25:471–492.

Felsenstein, J. 1985. Phylogenies and the comparative method. American Naturalist 125: 1–15.

Felsenstein, J. 1988. Phylogenies from molecular sequences: inference and reliability. Annual Review of Genetics 22: 521–565.

Felsenstein, J. 2012. A comparative method for both discrete and continuous characters using the threshold model. American Naturalist 179: 145–156.

Fernández, R, Laumer C. E., Vahtera, V., Libro, S., Kaluziak, S., Sharma, P. P., Pérez-Morro, A. R., Edgecombe, G. D., Giribert, G. 2014. Evaluating Topological Conflict in Centipede Phylogeny Using Transcriptomic Data Sets. Molecular Biology and Evolution 31: 1500–1513

Figuet, E., Ballenghien, M., Romiguier, J., Galtier, N. 2015. Biased Gene Conversion and GC-Content Evolution in the Coding Sequences of Reptiles and Vertebrates. Genome Biology and Evolution 7: 240–250.

Fitch, W. M. 1970. Distinguishing homologous from analogous proteins. Systematic Zoology 19: 99–113.

Fong, J. J., Brown, J. M., Fujita, M. K., Boussau, B. 2012. A Phylogenomic Approach to Vertebrate

Phylogeny Supports a Turtle-Archosaur Affinity and a Possible Paraphyletic Lissamphibia. PLoS ONE 7: e48990–14.

Fredman, D., White, S. J., Potter, S., Eichler, E. E, Dunnen, J. T. D., Brookes, A. J. 2004. Complex SNP-related sequence variation in segmental genome duplications. Nature Genetics 36:861–866.

Garrick, R. C., Sunnucks, P., Dyer, R. J. 2010. Nuclear gene phylogeography using PHASE: dealing with unresolved genotypes, lost alleles, and systematic bias in parameter estimation. BMC Evolutionary Biology 10: 118.

Garrick, R. C., Bonatelli, I. A., Hyseni, C., Morales, A., Pelletier, T. A., Perez, M. F., Rice, E., Satler, J. D., Symula, R. E., Thomé, M. T. C. 2015. The evolution of phylogeographic data sets. Molecular Ecology 24: 1164–1171.

Gernhard, T., 2008a. The conditioned reconstructed process. Journal of Theoretical Biology 253: 769–778.

Gernhard, T. 2008b. New analytic results for speciation times in neutral models. Bulletin of Mathematical Biology 70: 1082–1097.

Ghiurcuta, C. G., Moret, B. M. 2014. Evaluating synteny for improved comparative studies. Bioinformatics 30: i9–i18.

Goolsby, E. W., Bruggeman, J., Ané, C. 2017. Rphylopars: fast multivariate phylogenetic comparative methods for missing data and within-species variation. Methods in Ecology and Evolution 8: 22–27.

Graybeal, A. 1998. Is it better to add taxa or characters to a difficult phylogenetic problem? Systematic Biology 47: 9–17.

Guang, A., Zapata, F., Howison, M., Lawrence, C. E., Dunn, C. W. 2016. An Integrated Perspective on Phylogenetic Workflows. Trends in Ecology and Evolution 31: 116–126.

Gusfield, D. 2015. Persistent phylogeny. In: New York, New York, USA: ACM Press, 443–451.

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

Hammer, M. F., Woerner, A. E., Mendez, F. L., Watkins, J. C., Wall., J. D. 2011. Genetic evidence for archaic admixture in Africa. Proceedings of the National Academy of Sciences 108: 15123–15128.

Hansen, T. F. 1997. Stabilizing selection and the comparative analysis of adaptation. Evolution 51: 1341–1351.

Harmon, L. J., Baumes, J., Hughes, C., Soberón, J., Specht, C. D., Tumer, W., Lisle, C., Thacker, R. W. 2013. Arbor: Comparative Analysis Workflows for the Tree of Life. PLoS Currents 5: ecurrents.tol.099161de5eabdee073fd3d21a44518dc.

Harvey, M. G., Smith, B. T., Glenn, T. C., Faircloth, B. C., Brumfield, R. T. 2016. Sequence Capture versus Restriction Site Associated DNA Sequencing for Shallow Systematics. Systematic Biology 65: 910–924.

He, D., R. Sierra, Pawlowski, J., Baldauf, S. L. 2016. Reducing long-branch effects in multi-protein data uncovers a close relationship between *Alveolata* and *Rhizaria*. Molecular Phylogenetics and Evolution 101: 1–7.

Heled, J., Drummond, A. J. 2010. Bayesian inference of species trees from multilocus data. Molecular Biology and Evolution 27: 570–580.

Hey, J., Nielsen, R. 2004. Multilocus methods for estimating population sizes, migration rates and divergence time, with applications to the divergence of *Drosophila pseudoobscura* and *D. persimilis*. Genetics 167: 747–760.

Hey, J., Nielsen, R. 2007. Integration within the Felsenstein equation for improved Markov chain Monte Carlo methods in population genetics. Proceedings of the National Academy of Sciences USA 104: 2785–2790.

Hey, J. 1992. Using phylogenetic trees to study speciation and extinction. Evolution 46: 627–640.

Hey, J. 2010. Isolation with Migration Models for More Than Two Populations. Molecular Biology and Evolution 27: 905–920.

Hiller, M., Schaar, B. T., Indjeian, V. B., Kingsley, D. M., Hagey, L. R., Bejerano, G. 2012. A “forward genomics” approach links genotype to phenotype using independent phenotypic losses among related species. Cell Reports 2: 817–23.

Hillis, D. M. 1996. Inferring complex phylogenies. Nature 383: 130–131.

Hillis, D. M. 1998. Taxonomic sampling, phylogenetic accuracy, and investigator bias. Systematic Biology 47: 3–8.

Hinchliff, C. E., Smith, S. A., Allman, J. F., Burleigh, J. G., Chaudhary, R., Coghill, L. M., Crandall, K. A., Deng, J., Drew, B. T., Gazis, R., Gude, K., Hibbett, D. S., Katz, L. A., Laughinghouse, H. D. IV, McTavish, E. J., Midford, P. E., Owen, C. L., Ree, R. H., Rees, J. A., Soltis, D. E., Williams, T., Cranston, K. A. 2015. Synthesis of phylogeny and taxonomy into a comprehensive Tree of Life. Proceedings of the National Academy of Sciences USA 112: 12764–12769.

Hoban, S., Kelley, J. L., Lotterhos, K. E., Antolin, M. F., Bradburd, G., Lowry, D. B., Poss, M. L., Reed, L. K., Storfer, A., Whitlock, M. C. 2016. Finding the Genomic Basis of Local Adaptation: Pitfalls, Practical Solutions, and Future Directions. American Naturalist 188: 379–397.

Hobolth, A., Christensen, O. F., Mailund, T., Schierup, M. H. 2007. Genomic relationships and speciation times of human, chimpanzee, and gorilla inferred from a coalescent hidden Markov model. PLoS Genet. 3, e7.

Ho, L. S. T., Ané, C. 2014. A linear-time algorithm for Gaussian and non-Gaussian trait evolution models. Systematic Biology 63: 397–408.

Huang, H. T., He, Q. I., Kubatko, L. S., Knowles, L. L. 2010. Sources of error inherent in species-tree estimation: impact of mutational and coalescent effects on accuracy and implications for choosing among different methods. Systematic Biology 59: 573–583.

Huang, H., Sukumaran, J., Smith, S. A., Knowles, L. L. 2017. Cause of gene tree discord? Distinguishing incomplete lineage sorting and lateral gene transfer in phylogenetics. PeerJ Preprints https://doi.org/10.7287/peerj.preprints.3489v1.

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.

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

346:1320–1331.

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

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

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

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

Jones, G. 2016. Algorithmic improvements to species delimitation and phylogeny estimation under the multispecies coalescent. Journal of Mathematical Biology 74: 447–467.

Jones, G. R. 2017. Divergence estimation in the presence of incomplete lineage sorting and migration. bioRxiv. https://www.biorxiv.org/content/early/2017/10/16/174342

Jones, G., Sagitov, S., Oxelman, B. 2013. Statistical Inference of Allopolyploid Species Networks in the Presence of Incomplete Lineage Sorting. Systematic Biology 62: 467–478.

Jones, G., Aydin, Z., Oxelman, B. 2015. DISSECT: an assignment-free Bayesian discovery method for species delimitation under the multispecies coalescent. Bioinformatics 31: 991–998.

Kaiser, V. B., van Tuinen, M., Ellegren, H. 2007. Insertion events of CR1 retrotransposable elements elucidate the phylogenetic branching order in galliform birds. Molecular Biology and Evolution 24: 338–347.

Kidd, D. M. 2010. Geophylogenies and the Map of Life. Systematic Biology 59: 741–752.

Kim, J. 1996. General inconsistency conditions for maximum parsimony: effects of branch lengths and increasing numbers of taxa. Systematic Biology 46: 363–374.

Kingman, J. F. C. 1982. On the genealogy of large populations. Journal of Applied Probability 19: 27–43.

Klopfstein, S., Massingham, T., Goldman, N. 2017. More on the Best Evolutionary Rate for Phylogenetic

Analysis. Systematic Biology 66: 769–785.

Knowles, L. L., Smith, S. A. Huang, H., Sukumaran, J. 2018. A matter of phylogenetic scale: distinguishing incomplete lineage sorting from lateral gene transfer as the cause of gene tree discord in recent versus deep diversification histories. American Journal of Botany, revision in review.

Kocot, K. M., Struck, T. H., Merkel, J., Waits, D. S., Todt, C., Brannock, P. M., Weese, D. A., Cannon, J. T., Moroz, L. L., Lieb, B., Halanych, K. M. 2017. Phylogenomics of Lophotrochozoa with Consideration of Systematic Error. Systematic Biology 66: 256–282.

Kowada, L. A. B., Doerr, D., Dantas, S., Stoye, J. 2016. New Genome Similarity Measures Based on Conserved Gene Adjacencies. In: Singh M. (eds) Research in Computational Molecular Biology. RECOMB 2016. Lecture Notes in Computer Science, vol 9649. Springer, Cham.

Kriegs, J. O., Zemann, A., Churakov, G., Matzke, A., Ohme, M., Zischler, H., Brosius, J., Kryger, U., Schmitz, J. 2010. Retroposon Insertions Provide Insights into Deep Lagomorph Evolution. Molecular Biology and Evolution 27: 2678–2681.

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

Kubatko, L. S., Gibbs, H. L., Bloomquist, E. W. 2011. Inferring species-level phylogenies and taxonomic distinctiveness using multilocus data in *Sistrurus* rattlesnakes. Systematic Biology 60: 393–409.

Kuhn, T.S., Mooers, A. Ø., Thomas, G. H. 2011. A Simple Polytomy Resolver for Dated Phylogenies. Methods in Ecology and Evolution 2: 427–36.

Kumar, S., Filipski, A. J., Battistuzzi, F. U., Kosakovsky, Pond, S. L, Tamura, K. 2012. Statistics and Truth in Phylogenomics. Molecular Biology and Evolution 29:457–472.

Kumar, S., Jones, M., Koutsovoulos, G., Clarke, M. 2013. Blobology: exploring raw genome data for contaminants, symbionts and parasites using taxon-annotated GC-coverage plots. Frontiers in Genetics 4: 237.

Kunin, V., Goldovsky, L., Darzentas, N. 2005. The net of life: reconstructing the microbial phylogenetic network. Genome Research 15: 954–959.

Lamichhaney, S., Berglund, B., Almén, M.S., Maqbool, K., Grabherr, M., Martinez-Barrio, A., Promerova, M., Rubin, C.J., Wang, C., Zamani, N., Grant, B.R., Grant, P.R., Webster, M.T., Andersson, L. 2015. Evolution of Darwin's finches and their beaks revealed by genome sequencing. Nature 518: 371–375.

Lanier, H. C., Knowles, L. L. 2012. Is Recombination a Problem for Species-Tree Analyses? Systematic Biology 61: 691–701.

Leaché, A. D., Oaks, J. R. 2017. The Utility of Single Nucleotide Polymorphism (SNP) Data in Phylogenetics. Annual Review of Ecology, Evolution, and Systematics 48: 69–84.

Leaché, A. D., Chavez, A. S., Jones, L. N. 2015. Phylogenomics of phrynosomatid lizards: conflicting signals from sequence capture versus restriction site associated DNA sequencing. Genome Biology 7: 706–719.

Leaché, A. D., Harris, R. B., Rannala B., Yanz Z. 2014. The influence of gene flow on species tree estimation: A simulation study. Systematic Biology 63: 17–30.

Lemmon, A. R., Emme, S. A., Lemmon, E. M. 2012. Anchored Hybrid Enrichment for Massively High-Throughput Phylogenomics. Systematic Biology 61: 727–744.

Lemmon, E. M., Lemmon, A. R. 2013. High-Throughput Genomic Data in Systematics and Phylogenetics. Annual Review of Ecology Evolution and Systematics 44: 99–121.

Leprevost, F. V., Barbosa, V. C., Francisco, E. L., Perez-Riverol, Y., Carvalho, P. C. 2014. On best practices in the development of bioinformatics software. Frontiers in Genetics 5: 199.

Li, C., Hofreiter, M., Straube, N., Corrigan, S., Naylor, G. J. P. 2013. Capturing protein-coding genes across highly divergent species. BioTechniques 54: 1–5.

Lin, Y., Hu, F., Tang, J. Moret, B. M. 2013. Maximum likelihood phylogenetic reconstruction from high-resolution whole-genome data and a tree of 68 eukaryotes. Pacific Symposium on Biocomputing 285–296.

Lischer, H. E., Excoffier, L., Heckel, G. 2014. Ignoring heterozygous sites biases phylogenomic estimates of divergence times: implications for the evolutionary history of Microtus voles. Molecular Biology and Evolution 31: 817–831.

Liu, L., Pearl, D. K. 2007. Species trees from gene trees: reconstructing Bayesian posterior distributions of a species phylogeny using estimated gene tree distributions. Systematic Biology 56: 504–514.

Liu, L., Wu, S., Yu, L. 2015. Coalescent methods for estimating species trees from phylogenomic data. Journal of Systematics and Evolution 53: 380–390.

Liu, L., Xi, Z., Wu, S., Davis, C. C., Edwards, S. V. 2015. Estimating phylogenetic trees from genome-scale data. Annals of the New York Academy of Sciences 1360: 36–53.

Liu, L., Xi, Z., Davis, C. C. 2015. Coalescent Methods Are Robust to the Simultaneous Effects of Long Branches and Incomplete Lineage Sorting. Molecular Biology and Evolution 32: 791–805.

Liu, L., Yu, L., Edwards, S. V. 2010. A maximum pseudo-likelihood approach for estimating species trees under the coalescent model. BMC Evolutionary Biology 10: 302.

Liu, L., Yu, L., Pearl, D. K., Edwards, S. V. 2009. Estimating species phylogenies using coalescence times among sequences. Systematic Biology 58: 468–477.

Liu, L., Zhang, J., Rheindt, F. E., Lei, F., Qu, Y., Wang, Y., Sullivan, C., Nie, W., Wang, J., Yang, F., Chen, J., Edwards, S. V., Meng, J., Wu, S. 2017. Genomic evidence reveals a radiation of placental mammals uninterrupted by the KPg boundary. Proceedings of the National Academy of Science of the USA 114: E7282–7290.

Long, J. C. 1991. The genetic structure of admixed populations. Genetics 127: 417–418.

Lott, M., Spillner, A., Huber, K. T., Moulton, V. 2009. PADRE: A package for analyzing and displaying reticulate evolution. Bioinformatics 25: 1199–2000.

Maddison, W. P. 1997. Gene trees in species trees. Systematic Biology 46: 523–536.

Magadum, S., Banerjee, U., Murugan, P., Gangapur, D., Ravikesavan. R. 2013. Gene duplication as a major force in evolution. Journal of Genetics 92: 155–161.

Mallet, J., Besansky, N., Hahn, M. W. 2015. How reticulated are species? BioEssays 38: 140–149.

Manceau, M., Lambert, A., Morlon, H. 2015. Phylogenies support out-of-equilibrium models of biodiversity. Ecology Letters 18: 347–356.

Manceau, M., Lambert, A., Morlon, H. 2017. A unifying comparative phylogenetic framework including traits coevolving across interacting lineages. Systematic Biology 66: 551–568.

Mandáková, T., Joly, S., Krzywinski, M., Mummenhoff, K., Lysak, M. A. 2010. Fast Diploidization in Close Mesopolyploid Relatives of *Arabidopsis*. The Plant Cell 22: 2277–2290.

Marcovitz, A., Jia, R., Bejerano, G. 2016. "Reverse Genomics" Predicts Function of Human Conserved Noncoding Elements. Molecular Biology and Evolution 33: 1358–1369.

Marcussen, T., Sandve, S. R., Heier, L., Spannagl, M., Pfeifer, M., The International Wheat Genome Sequencing Consortium, Jakobsen, K. S., Wulff, B. B. H., Steuernagel, B., Mayer, K. F. X., Olsen, O. A. 2014. Ancient hybridizations among the ancestral genomes of bread wheat. Science 345: 1250092.

Marcussen, T., Heier, L., Brysting, A. K., Oxelman, B., Jakobsen, K. S. 2015. From gene trees to a dated allopolyploid network: Insights from the angiosperm genus *Viola* (Violaceae). Systematic Biology 64: 84–101.

Matzke, A., Churakov, G., Berkes, P., Arms, E. M., Kelsey, D., Brosius, J., Kriegs, J. O., Schmitz, J. 2012. Retroposon Insertion Patterns of Neoavian Birds: Strong Evidence for an Extensive Incomplete Lineage Sorting Era. Molecular Biology and Evolution 29: 1497–1501.

McCormack, J. E., Hird, S. M., Zellmer, A. J., Carstens, B. C., Brumfield, R. T. 2013. Applications of next-generation sequencing to phylogeography and phylogenetics. Molecular Phylogenetics and Evolution 66: 526–538.

McCormack, J. E., Tsai, W. L. E., Faircloth, B. C. 2016. Sequence capture of ultraconserved elements from bird museum specimens. Molecular Ecology Resources 16: 1189–1203.

McCormack, J. E., Rodríguez-Gómez, F., Tsai, W. L. E., Faircloth, B. C. 2017. Transforming Museum Specimens into Genomic Resources. Pp. 143–156 *in* M. S. Webster (editor), The Extended Specimen: Emerging Frontiers in Collections-based Ornithological Research. Studies in Avian Biology (no. 50), CRC Press, Boca Raton, FL.

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.

Meiklejohn, K. A., Faircloth, B. C., Glenn, T. C., Kimball, R. T., Braun, E. L. 2016. Analysis of a Rapid Evolutionary Radiation Using Ultraconserved Elements: Evidence for a Bias in Some Multispecies Coalescent Methods. Systematic Biology 65: 612–627.

Mendes, F. K., Hahn, M. W. 2016. Gene Tree Discordance Causes Apparent Substitution Rate Variation. Systematic Biology 65: 711–721.

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., et al. 2014. Phylogenomics resolves the timing and pattern of insect evolution. Science 346: 763–767.

Mitov, V., Stadler, T. 2017. POUMM: An R-package for Bayesian Inference of Phylogenetic Heritability. bioRxiv, https://doi.org/10.1101/115089.

Mitchell, A., Mitter, C., Regier, J. C. 2000. More taxa or more characters revisited: Combining data from nuclear protein-encoding genes for phylogenetic analyses of Noctuoidea (Insecta: Lepidoptera). Systematic Biology 49: 202–224.

Montague, M. J., Li, G., Gandolfi, B., Khan, R., Aken, B. L., Searle, S. M. J., Minx, P., Hillier, L. W., Koboldt, D. C., Davis, B. W., Driscoll, C. A., Barr, C. S., Blackistone, K., Quilez, J., Lorente-Galdos, B., Marques-Bonet, T., Alkan, C., Thomas, G. W. C., Hahn, M. W., Menotti-Raymond, M., O’Brien, S. J., Wilson, R. K., Lyons, L. A., Murphy, W. J., Warren, W. C. 2014. Comparative analysis of the domestic cat genome reveals genetic signatures underlying feline biology and domestication. Proceedings of the National Academy of Sciences USA 111: 17230–17235.

Mora, C., Tittensor, D. P., Adl, S., Simpson, A. G. B., Worm, B. 2011. How Many Species Are There on Earth and in the Ocean? PLoS Biology 9(8):e1001127.

Morlon, H. 2014. Phylogenetic approaches for studying diversification. Ecology Letters 17: 508–525.

Morlon, H., Parsons, T.L., Plotkin, J. 2011. Reconciling molecular phylogenies with the fossil record Proceedings of the National Academy of Sciences 108: 16327–16332.

Mulder, W. H., Crawford, F. W. 2015. On the distribution of interspecies correlation for Markov models of character evolution on Yule trees. Journal of Theoretical Biology 364: 275–283.

Murphy, W. J., Larkin, D. M., Everts-van der Wind, A, Bourque, G., Tesler, G., Auvil, L., Beever, J. E, Chowdhary, B. P., Galibert, F., Gatzke, L., Hitte, C., Meyers, S. N., Milan, D., Ostrander, E. A., Pape, G., Parker, H. G., Raudsepp, T., Rogatcheva, M. B., Schook, L. B., Skow, L. C., Welge, M., Womack, J. E., O'Brien, S. J., Pevzner, P. A., Lewin, H. A. 2005. Dynamics of mammalian chromosome evolution inferred from multispecies comparative maps. Science 309: 613–617.

Murphy, W. J., Pringle, T. H., Crider, T. A., Springer, M. S., Miller, W. 2007. Using genomic data to unravel the root of the placental mammal phylogeny. Genome Research 17: 413–421.

Nabhan, A. R., Sarkar, I. N. 2012. The impact of taxon sampling on phylogenetic inference: a review of two decades of controversy. Briefings in Bioinformatics 13: 122–134.

Nee, S., Mooers, A. O., Harvey, P. H. 1992. Tempo and mode of evolution revealed from molecular phylogenies. Proceedings of the National Academy of Sciences USA 89: 8322–8326.

Ogilvie, H. A., Bouckaert, R. R., Drummond, A. J. 2017. StarBEAST2 Brings Faster Species Tree Inference and Accurate Estimates of Substitution Rates. Molecular Biology and Evolution 34: 2101–2114.

Olave, M., Sola E., Knowles L. L. 2014. Upstream analyses create problems with DNA-based species delimitation. Systematic Biology 63: 263–271.

Oxelman, B., Yoshikawa, N., McConaughy, B. L., Luo, J., Denton, A. L., Hall, B. D. 2004. RPB2 Gene Phylogeny in Flowering Plants, with Particular Emphasis on Asterids. Molecular Phylogenetics and Evolution 32: 462–79.

Pamilo, P., Nei, M. 1988. Relationships between gene trees and species trees. Molecular Biology and Evolution 5: 568–583.

Park, S. D. E., Magee, D. A., McGettigan, P. A., Teasdale, M. D., Edwards, C. J., Lohan, A. J, Murphy, A., Braud, M., Donoghue, M. T., Liu, Y., Chamberlain, A. T, Rue-Albrecht, K., Schroeder, S., Spillane, C., Tai, S., Bradley, D. G., Sonstegard, T. S., Loftus, B. J., McHugh, D. E. 2015. Genome 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.

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.

Peñalba, J. V., Smith, L. L., Tonione, M. A., Sass, C., Hykin, S. M., Skipwith, P. L., McGuire, J. A., Bowie, R. C. K., Moritz, C. 2014. Sequence capture using PCR-generated probes: a cost-effective method of targeted high-throughput sequencing for non-model organisms. Molecular Ecology Resources 14: 1000–1010.

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.

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, M. 2008. Phylogenies without roots? A plea for the use of vouchers in molecular phylogenetic studies. Molecular Phylogenetics and Evolution 48: 369–371.

Poe, S. 1998. Sensitivity of phylogeny estimation to taxonomic sampling. Systematic Biology 47: 18–31.

Potts, A. J., Hedderson, T. A., Grimm, G. W. 2014. Constructing Phylogenies in the Presence Of Intra-Individual Site Polymorphisms (2ISPs) with a Focus on the Nuclear Ribosomal Cistron. Systematic Biology 63: 1–16.

Prum, R. O., Berv, J. S., Dornburg, A., Field, D. J., Townsend, J. P. 2015. A comprehensive phylogeny of birds (Aves) using targeted next-generation DNA sequencing. Nature 526: 569–573.

Pyron, R. A. 2015. Post-molecular systematics and the future of phylogenetics. Trends in Ecology & Evolution 30: 384–389.

Quintero, I., P. Keil., W. Jetz., F. W. Crawford. 2015. Historical Biogeography Using Species Geographical Ranges. Systematic Biology 64: 1059–1073.

Ramadugu, C., Pfeil, B. E., Manjunath, K. L., Lee, R. F., Maureira-Butler, I. J., Roose, M. L. 2013. Coalescence simulation testing of hybridization versus lineage sorting in *Citrus* (Rutaceae) using six nuclear genes. PLoS One 8: e68410.

Rannala, B., Yang, Z. H. 2003. Bayes estimations of species divergence times and ancestral population sizes using DNA sequences from multiple loci. Genetics 164: 1645–1656.

Rannala, B., Yang Z. H. 2008. Phylogenetic inference using whole genomes. Annual Review of Genomics and Human Genetics 9: 217–231.

Ranwez, V., Delsuc, F. D. R., Ranwez, S., Belkhir, K., Tilak, M-K., Douzery, E. J. 2007. OrthoMaM: A database of orthologous genomic markers for placental mammal phylogenetics. BMC Evolutionary Biology 7: 241–12.

Rasmussen, M. D., Kellis, M. 2012. Unified modeling of gene duplication, loss, and coalescence using a locus tree. Genome Research 22: 755–765.

Reddy, S., Kimball, R. T., Pandey, A., Hosner, P. A., Braun, M. J., Hackett, S. J., Han, K-L., Harshman, J., Huddleston, C. J., Kingston, S., Marks, B. D., Miglia, K. J., Moore, W. S., Sheldon, F. H., Witt, C. C., Yuri, T., Braun, E. L. 2017. Why Do Phylogenomic Data Sets Yield Conflicting Trees? Data Type Influences the Avian Tree of Life more than Taxon Sampling. Systematic Biology 66: 857–879.

Redon, R., Ishikawa, S., Fitch, K. R., Feuk, L., Perry, G. H., Andrews, T. D., Fiegler, H., Shapero, M. H., Carson, A. R., Chen, W., Cho, E. K., Dallaire, S., Freeman, J. L., González, J. R., Gratacòs, M. N., Huang, J., Kalaitzopoulos, D., Komura, D., MacDonald, J. R., Marshall, C. R., Mei, R., Montgomery, L., Nishimura, K., Okamura, K., Shen, F., Somerville, M. J., Tchinda, J., Valsesia, A., Woodwark, C., Yang, F., Zhang, J., Zerjal, T., Zhang, J., Armengol, L., Conrad, D. F., Estivill, X., Tyler-Smith, C., Carter, N. P., Aburatani, H., Lee, C., Jones, K. W., Scherer, S. W., Hurles, M. E. 2006. Global variation in copy number in the human genome. Nature 444: 444–454.

Reid, N. M., Hird, S. M., Brown, J. M., Pelletier, T. A., McVay, J. D., Satler, J. D., Carstens, B. C. 2014. Poor fit to the multispecies coalescent is widely detectable in empirical data. Systematic Biology 63: 322–333.

Rogozin, I. B., Thomson, K., Csürös, M., Carmel, L., Koonin, E. V. 2008. Homoplasy in genome-wide analysis of rare amino acid replacements: the molecular-evolutionary basis for Vavilov's law of homologous series. Biology Direct 2008 3: 7.

Rogozin, I. B., Wolf, Y. I., Babenko, B. N., Koonin, E. V. 2006. Dollo parsimony and the reconstruction of genome evolution. In: Albert VA ed. Parsimony, Phylogeny, and Genomics. Oxford University Press, p.p. 190–200.

Rokas, A. 2005. More Genes or More Taxa? The Relative Contribution of Gene Number and Taxon Number to Phylogenetic Accuracy. Molecular Biology and Evolution 22: 1337–1344.

Rokas, A., Holland, P. W. H. 2000. Rare genomic changes as a tool for phylogenetics. Trends in Ecology and Evolution 15: 454–459.

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

Romiguier, J., Cameron, S. A., Woodard, S. H., Fischman, B. J., Keller, L., Praz, C. J. 2016. Phylogenomics Controlling for Base Compositional Bias Reveals a Single Origin of Eusociality in Corbiculate Bees. Molecular Biology and Evolution 33: 670–678.

Romiguier, J., Roux, C. 2017. Analytical Biases Associated with GC-Content in Molecular Evolution. Frontiers in Genetics 8: 16.

Roncal, J., Guyot, R., Hamon, P., Crouzillat, D., Rigoreau, M., Konan, O. N., Rakotomalala, J. J., Nowak, M. D., Davis, A. P., de Kochko, A. 2016. Active transposable elements recover species boundaries and geographic structure in Madagascan coffee species. Molecular Genetics and Genomics 291: 155–168.

Rosenberg, N. A., Nordborg, M. 2002. Genealogical trees, coalescent theory and the analysis of genetic polymorphisms. Nature Reviews Genetics 3: 380–390.

Rosindell, J., Cornell, S. J., Hubbell S. P., Etienne, R. S. 2010. Protracted speciation revitalizes the neutral theory of biodiversity. Ecology Letters 13: 716–727.

Rosindell, J., Harmon, L. J. 2012. OneZoom: A Fractal Explorer for the Tree of Life. PLoS Biology 10(10): e1001406.

Rosindell, J., Harmon, L. J., Etienne, R. S. 2015. Unifying ecology and macroevolution with individual-based theory. Ecology Letters 18: 472–482.

Ruane, S., Austin, C. C. 2017. Phylogenomics using formalin-fixed and 100+ year-old intractable natural history specimens. Molecular Ecology Resources 17: 1003–1008.

Sagitov, S., Bartoszek, K. 2012. Interspecies correlation for neutrally evolving traits. Journal of Theoretical Biology 309: 11–19.

Sanderson, M. J., Donoghue, M. J., Piel, W. H., Eriksson, T. 1994. TreeBASE: a prototype database of phylogenetic analyses and an interactive tool for browsing the phylogeny of life. American Journal of Botany 81: 183.

Sayyari, E., Mirarab, S. 2016. Fast Coalescent-Based Computation of Local Branch Support from Quartet Frequencies. Molecular Biology and Evolution 33: 1654–1668.

Scally, A., Dutheil, J. Y., Hillier, L. W., Jordan, G. E., Goodhead, I., Herrero, J., Hobolth, A., Lappalainen, T., Mailund, T., Marques-Bonet, T., McCarthy, S., Montgomery, S. H., Schwalie, P. C., Tang, Y. A., Ward, M. C., Xue, Y., Yngvadottir, B., Alkan, C., Andersen, L. N., Ayub, Q., Ball, E.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26449v3 | CC BY 4.0 Open Access | rec: 30 Jan 2018, publ: 30 Jan 2018

V., Beal, K., Bradley, B. J., Chen, Y., Clee, C. M., Fitzgerald, S., Graves, T. A., Gu, Y., Heath, P., Heger, A., Karakoc, E., Kolb-Kokocinski, A., Laird, G. K., Lunter, G., Meader, S., Mort, M., Mullikin, J. C., Munch, K., O'Connor, T. D., Phillips, A. D., Prado-Martinez, J., Rogers, A. S., Sajjadian, S., Schmidt, D., Shaw, K., Simpson, J. T., Stenson, P. D., Turner, D. J., Vigilant, L., Vilella, A. J., Whitener, W., Zhu, B., Cooper, D. N., de Jong, P., Dermitzakis, E. T., Eichler, E. E., Flicek, P., Goldman, N., Mundy, N. I., Ning, Z., Odom, D. T., Ponting, C. P., Quail, M. A., Ryder, O. A., Searle, S. M., Warren, W. C., Wilson, R. K., Schierup, M. H., Rogers, J., Tyler-Smith, C., Durbin, R. 2012. Insights into hominid evolution from the gorilla genome sequence. Nature 483: 169–175.

Schrempf, D., Minh, B. Q., De Maio, N., von Haeseler, A., Kosiol. C. 2016. Reversible polymorphism-aware phylogenetic models and their application to tree inference. Journal of Theoretical Biology 407: 362–370.

Schwartz, R., Schäffer, A. A. 2017. The evolution of tumour phylogenetics: Principles and practice. Nature Reviews Genetics 18: 213–229.

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

Shi, C.-M., Yang, Z. 2018. Coalescent-Based Analyses of Genomic Sequence Data Provide a Robust Resolution of Phylogenetic Relationships among Major Groups of Gibbons. Molecular Biology and Evolution 35: 159–179.

Siepel, A. 2009. Phylogenomics of primates and their ancestral populations. Genome Research 19: 1929–1941.

Silvestro, D., Schnitzler, J., Liow, L. H., Antonelli, A., Salamin, N. 2014. Bayesian estimation of speciation and extinction from incomplete fossil occurrence data. Systematic Biology 63: 349–367.

Simion, P., Philippe, H., Baurain, D., Muriel, J., Richter, D. J., Di Franco, A., Roure, B., Satoh, N., Quéinnec, E., Ereskovsky, A. 2017. A Large and Consistent Phylogenomic Dataset Supports

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26449v3 | CC BY 4.0 Open Access | rec: 30 Jan 2018, publ: 30 Jan 2018

Sponges as the Sister Group to All Other Animals. Current Biology 27: 958–967.

Sjödin, P., Jakobsson, M. 2012. Population genetic nature of copy number variation. Population Genetic Nature of Copy Number Variation. In: Feuk L. (eds) Genomic Structural Variants. Methods in Molecular Biology (Methods and Protocols), vol 838. Springer, New York, NY.

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

Solis-Lemus, C., Ané, C. 2016. Inferring Phylogenetic Networks with Maximum Pseudolikelihood under Incomplete Lineage Sorting. PLoS Genet 12: e1005896.

Solis-Lemus, C., Knowles, L. L., Ané, C. 2015. Bayesian species delimitation combining multiple genes and traits in a unified framework. Evolution 69: 492–507.

Solis-Lemus, C., Bastide, P., Ané, C. 2017. PhyloNetworks: A Package for Phylogenetic Networks. Molecular Biology and Evolution 34: 3292–3298.

Song, S., Liu, L., Edwards, S. V., Wu, S. 2012. Resolving conflict in eutherian mammal phylogeny using phylogenomics and the multispecies coalescent model. Proceedings of the National Academy of Sciences USA 112: E6079–E6079.

Sousa, F., Bertrand, Y. J. K., Doyle, J. J., Oxelman, B., Pfeil, B. E. 2017. Using genomic location and coalescent simulation to investigate gene tree discordance in *Medicago* L. Systematic Biology 66: 934–949.

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

Staats, M., Erkens, R. H. J., van de Vossenberg, B., Wieringa, J. J., Kraaijeveld, K., Stielow, B., Geml, J., Richardson, J. E., Bakker, F. T. 2013. Genomic treasure troves: complete genome sequencing of herbarium and insect museum specimens. PLoS ONE 8: e69189.

Stadler, T., 2009. On incomplete sampling under birth-death models and connections to the sampling-based coalescent. Journal of Theoretical Biology 261: 58–68.

Stadler, T. 2013. Recovering speciation and extinction dynamics based on phylogenies. Journal of Evolutionary Biology 26: 1203–1219.

Stadler, T., Steel, M., 2012. Distribution of branch lengths and phylogenetic diversity under homogeneous speciation models. Journal of Theoretical Biology 297: 33–40.

Stephens, M., Smith, N.J., Donnelly, P. 2001. A new statistical method for haplotype reconstruction from population data. The American Journal of Human Genetics 68: 978–989.

Struck, T. H. 2013. The Impact of Paralogy on Phylogenomic Studies – A Case Study on Annelid Relationships. PLoS ONE 8: e62892.

Suarez, A. V., Tutsui, N. D. 2004. The value of museum collections for research and society. BioScience 54: 66–74.

Suh, A., Paus, M., Kiefmann, M., Churakov, G., Franke, F. A., Brosius, J., Kriegs, J. O., Schmitz, J. 2011. Mesozoic retroposons reveal parrots as the closest living relatives of passerine birds. Nature Communications 2: 443.

Suh, A., Smeds, L. A., Ellegren, H. 2015. The Dynamics of Incomplete Lineage Sorting across the Ancient Adaptive Radiation of Neoavian Birds. PLoS Biology 13: e1002224–18.

Sukumaran, J., Knowles, L. L., 2017. Multispecies coalescent delimits structure, not species. Proceedings of the National Academy of Sciences USA 114: 1607–1612.

Talavera, G., Castresana, J. 2007. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. Systematic Biology 56: 564–577.

Tang, J., Moret, B. M. E., Cui, L., dePamphilis, C. W. 2004. Phylogenetic reconstruction from arbitrary gene-order data. in Proceedings. Fourth IEEE Symposium on Bioinformatics and Bioengineering 592–599.

Tonini, J. F. R., Beard, K. H., Ferreira, R. B., Jetz, W., Pyron, R. A. 2016. Fully-sampled phylogenies of squamates reveal evolutionary patterns in threat status. Biological Conservation 204: 23–31.

Toprak, Z., Pfeil, B.E., Jones, G., Marcussen, T., Ertekin, A.S., Oxelman, B. 2016. Species Delimitation Without Prior knowledge: DISSECT Reveals Extensive Cryptic Speciation in the *Silene aegyptiaca* complex (Caryophyllaceae). Molecular Phylogenetics and Evolution 102: 1–8.

Turney, S., Cameron, E. R., Cloutier, C. A, Buddle, C. M. 2015. Non-repeatable science: assessing the frequency of voucher specimen deposition reveals that most arthropod research cannot be verified. PeerJ 3: e1168–16.

Vision, T. 2010. Open data and the social contract of scientific publishing. *BioScience* 60: 330.

Wen, D., Nakhleh L. 2018. Co-estimating Reticulate Phylogenies and Gene Trees from Multi-locus Sequence Data. Systematic Biology in press.

Wen, D., Yu Y., Hahn M. W., Nakhleh L. 2016. Reticulate evolutionary history and extensive introgression in mosquito species revealed by phylogenetic network analysis. Molecular Ecology 25: 2361–2372.

Wesche, P. L., Gaffney, D. J., Keightley, P. D. 2004. DNA Sequence Error Rates in Genbank Records Estimated Using the Mouse Genome as a Reference. DNA Sequence 15(5–6): 362–64.

Wiedenhoeft, J., Brugel, E., Schliep, A. 2016. Fast Bayesian Inference of Copy Number Variants using Hidden Markov Models with Wavelet Compression. PLoS Computational Biology 12(5): e1004871.

Will, K. P., Mishler, B. D., Wheeler, Q. D. 2005. The perils of DNA Barcoding and the need for integrative taxonomy. Systematic Biology 54: 844–851.

Wilson, G., Aruliah, D. A., Brown, C. T., Chue-Hong N. P., Davis, M., Guy, R. T., Haddock, S. H. D, Huff, K. D., Mitchell, I. M., Plumbley, M. D, Waugh, B., White, E. P., Wilson, P. 2014. Best Practices for Scientific Computing. PLoS Biology 12(1): e1001745.

Wu, Y. C., Rasmussen, M. D., Bansal, M. S., Kellis, M. 2013. TreeFix: statistically informed gene tree error correction using species trees. Systematic Biology 62: 110–120.

Wu, Y. C., Rasmussen, M. D., Bansal, M. S., Kellis M. 2014. Most parsimonious reconciliation in the presence of gene duplication, loss, and deep coalescence using labeled coalescent trees. Genome

Research 24: 475–486.

Xi, Z., Ruhfel, B. R., Schaefer, H., Amorim, A. M., Sugumaran, M., Wurdack, K. J., Endress, P. K., Matthews, M. L., Stevens, P. F., Mathews, S., Davis, C. C. 2012. Phylogenomics and a posteriori data partitioning resolve the Cretaceous angiosperm radiation Malpighiales. Proceedings of the National Academy of Sciences USA 109: 17519–17524.

Xi, Z., Liu, L., Davis, C. C. 2015. Genes with minimal phylogenetic information are problematic for coalescent analyses when gene tree estimation is biased. Molecular Phylogenetics and Evolution 92: 63–71.

Xu, B., Yang, Z. 2016. Challenges in Species Tree Estimation Under the Multispecies Coalescent Model. Genetics 204: 1353–1368.

Yang Z, Rannala B, 2014. Unguided species delimitation using DNA sequence data from multiple loci. Molecular Biology and Evolution 31: 3125–3135.

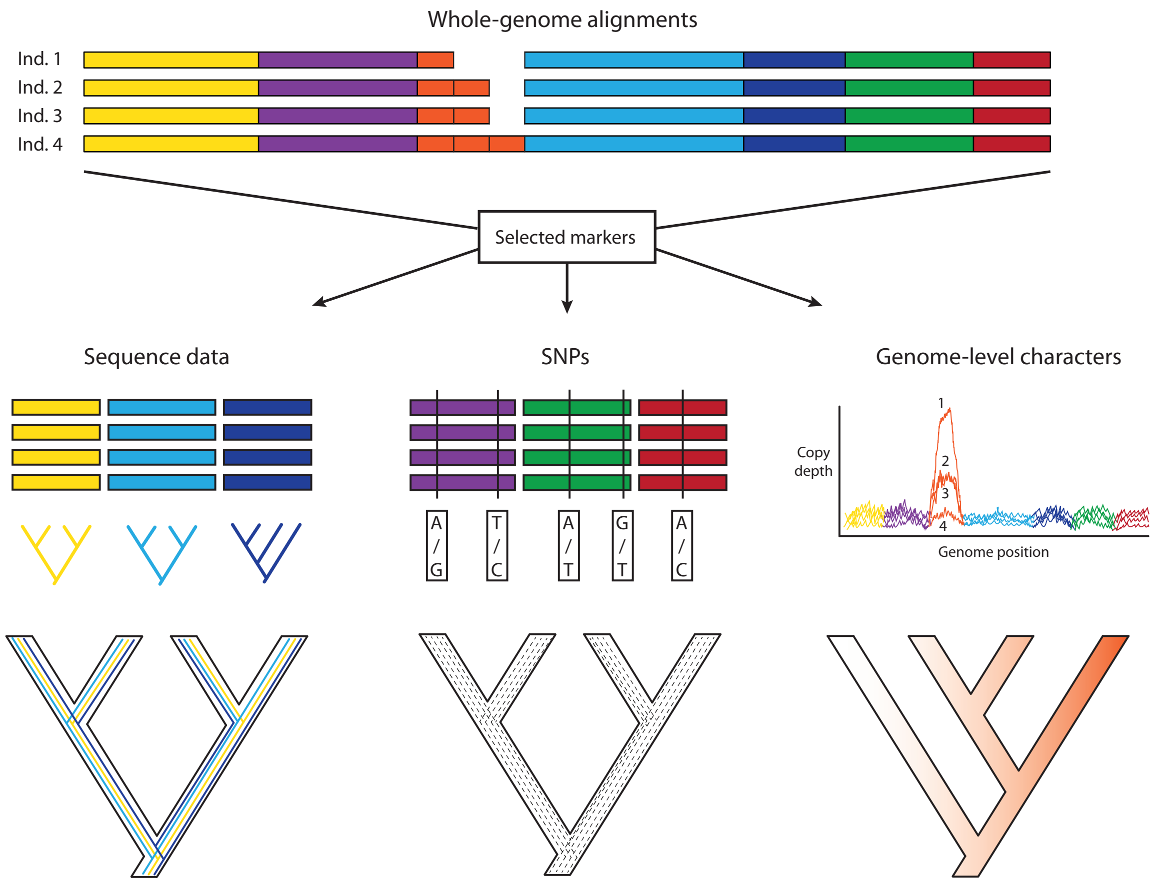
Yu, Y., Dong J., Liu K. J., Nakhleh, L. 2014. Maximum likelihood inference of reticulate evolutionary histories. Proceedings of the National Academy of Sciences USA 111: 16448–16453.

Zanne, A. E., Tank, D. C., Cornwell, W. K., Eastman, J. M., Smith, S. A, FitzJohn, R. G., McGlinn, D. J., O'Meara, B. C., Moles, A. T., Reich, P. B., Royer, D. L., Soltis, D. E., Stevens, P. F., Westoby, M., Wright, I. J., Aarssen, L., Bertin, R. I. Calaminus, A., Govaerts, R., Hemmings, F., Leishman, M. R., Oleksyn, J., Soltis, P. S., Swenson, N. G., Warman, L., Beaulieu, J. M. 2014. Three keys to the radiation of angiosperms into freezing environments. Nature 506: 89–92.

Zhang, C., Ogilvie, H. A., Drummond, A. J., Stadler, T. 2018. Bayesian Inference of Species Networks from Multilocus Sequence Data. Molecular Biology and Evolution, in press.

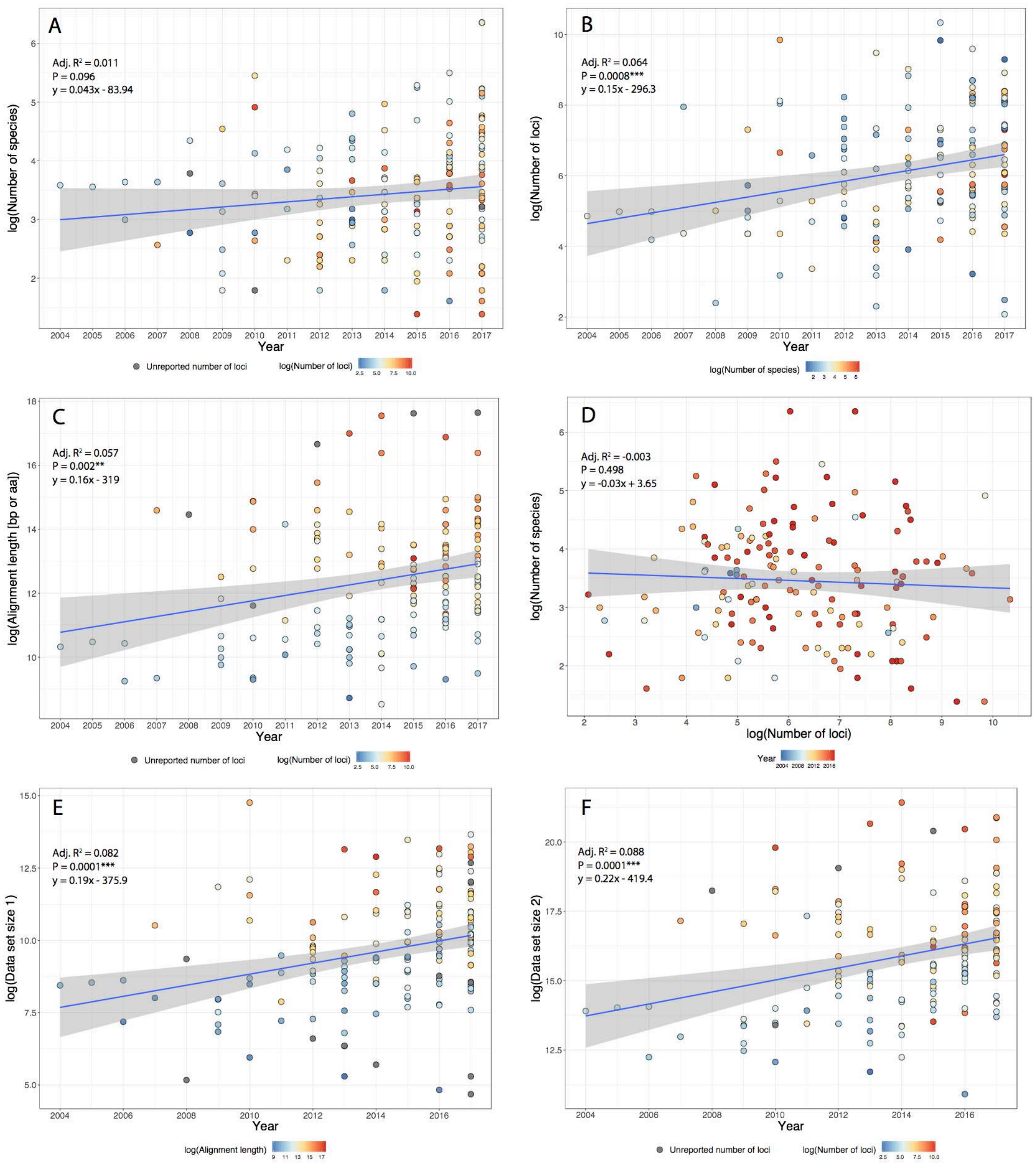
Figures

Figure 1. *A posteriori* marker selection from whole genome alignments for phylogenomics and phylogeography. Whole genome analysis (top) permits researchers to choose different markers for specific purposes. 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.



PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26449v3 | CC BY 4.0 Open Access | rec: 30 Jan 2018, publ: 30 Jan 2018

Figure 2. Trends in phylogenomic data sets since the emergence of HTS. Based on a sample of 164 phylogenemic 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 Preprints | https://doi.org/10.7287/peerj.preprints.26449v3 | CC BY 4.0 Open Access | rec: 30 Jan 2018, publ: 30 Jan 2018

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. This implies that tree building without considering the multispecies coalescent could, in this case, lead to erroneous estimation of tree topology and divergence times.

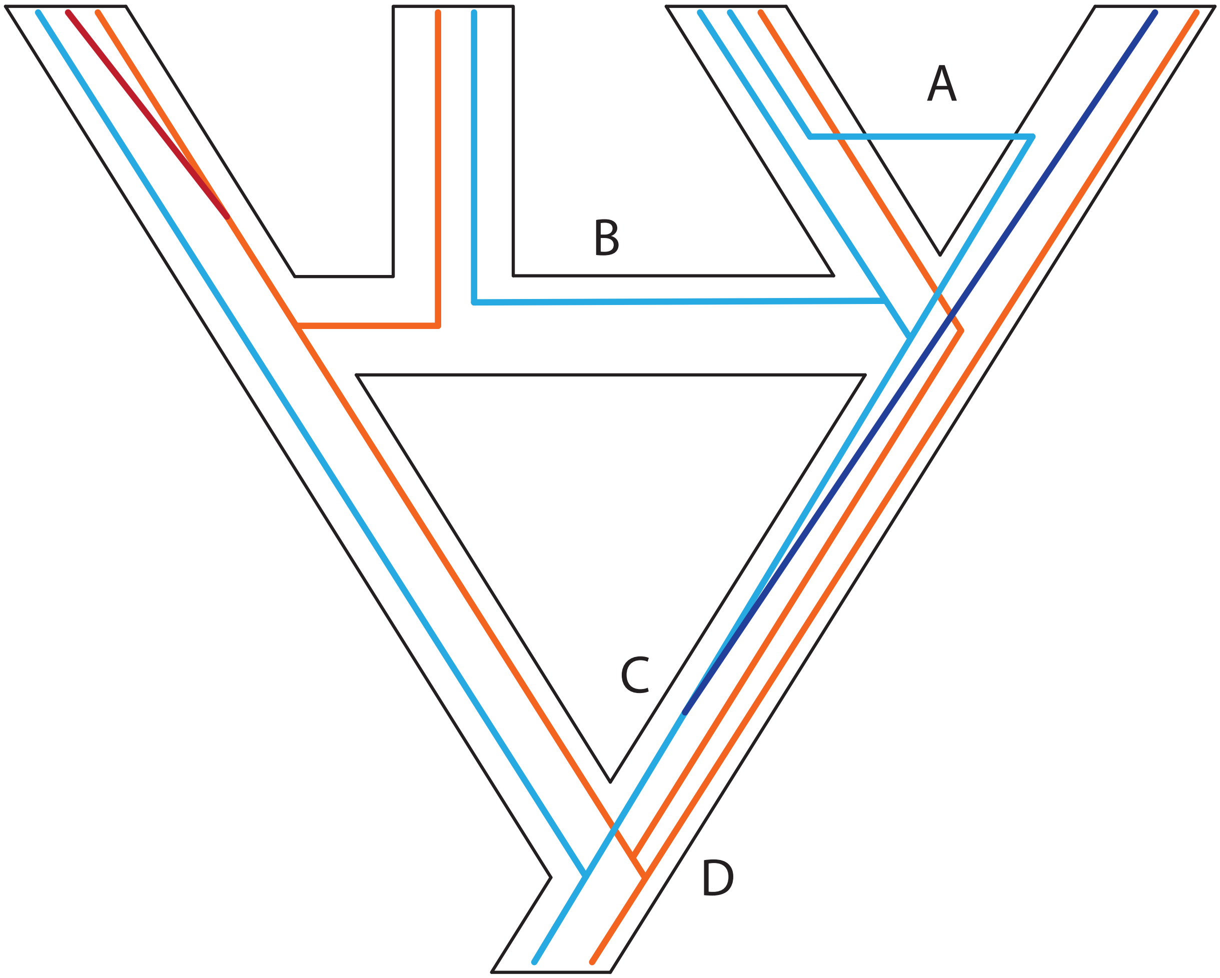


Figure 4. Gene duplication and loss creates patterns that can mimic incomplete lineage sorting and other processes. Genes and genomes of three species A, B, and C. Multi-colored bars show (parts of) their genomes with a number of loci indicated in different colors. The orange gene is duplicated in species 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. 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.

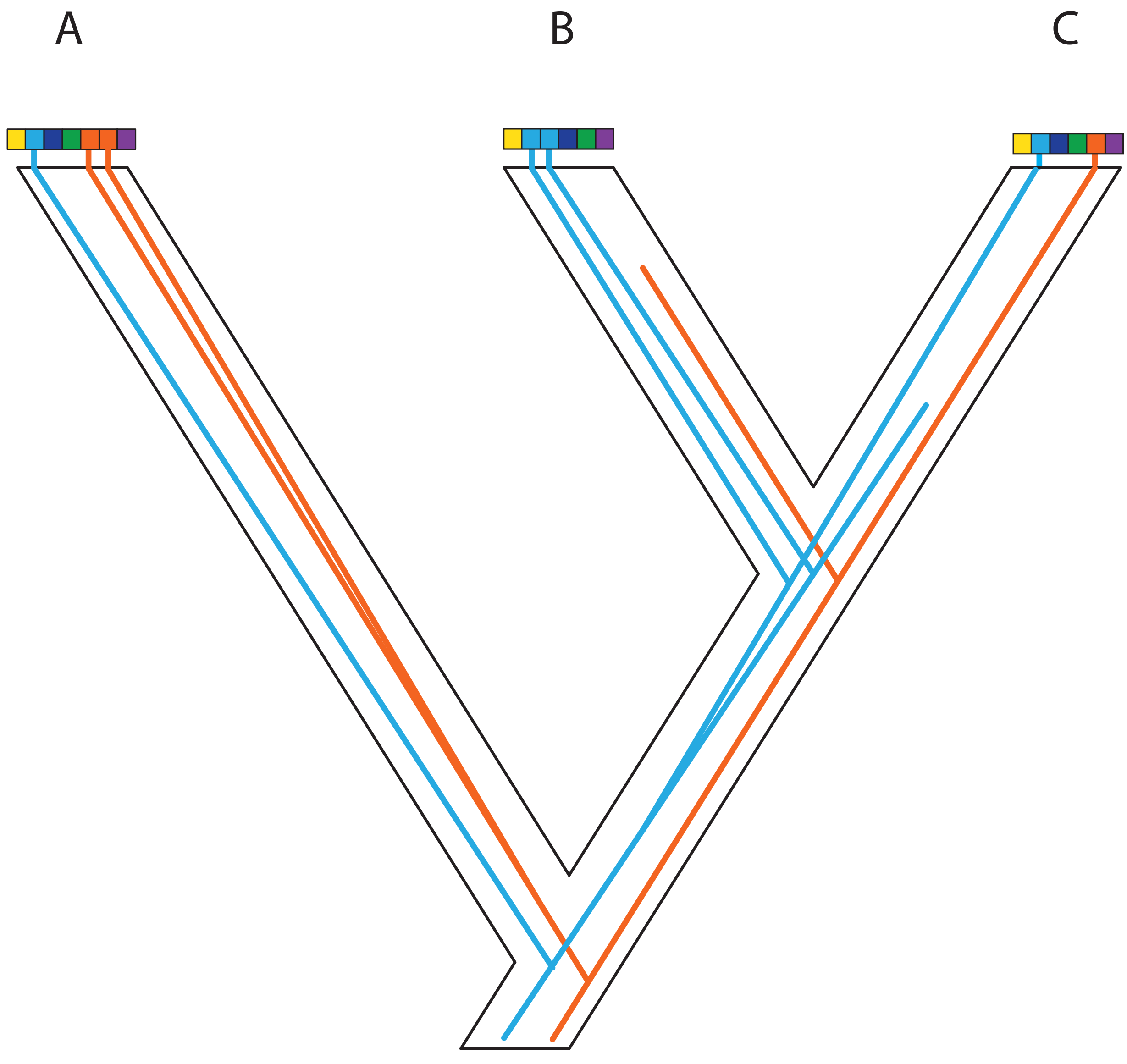
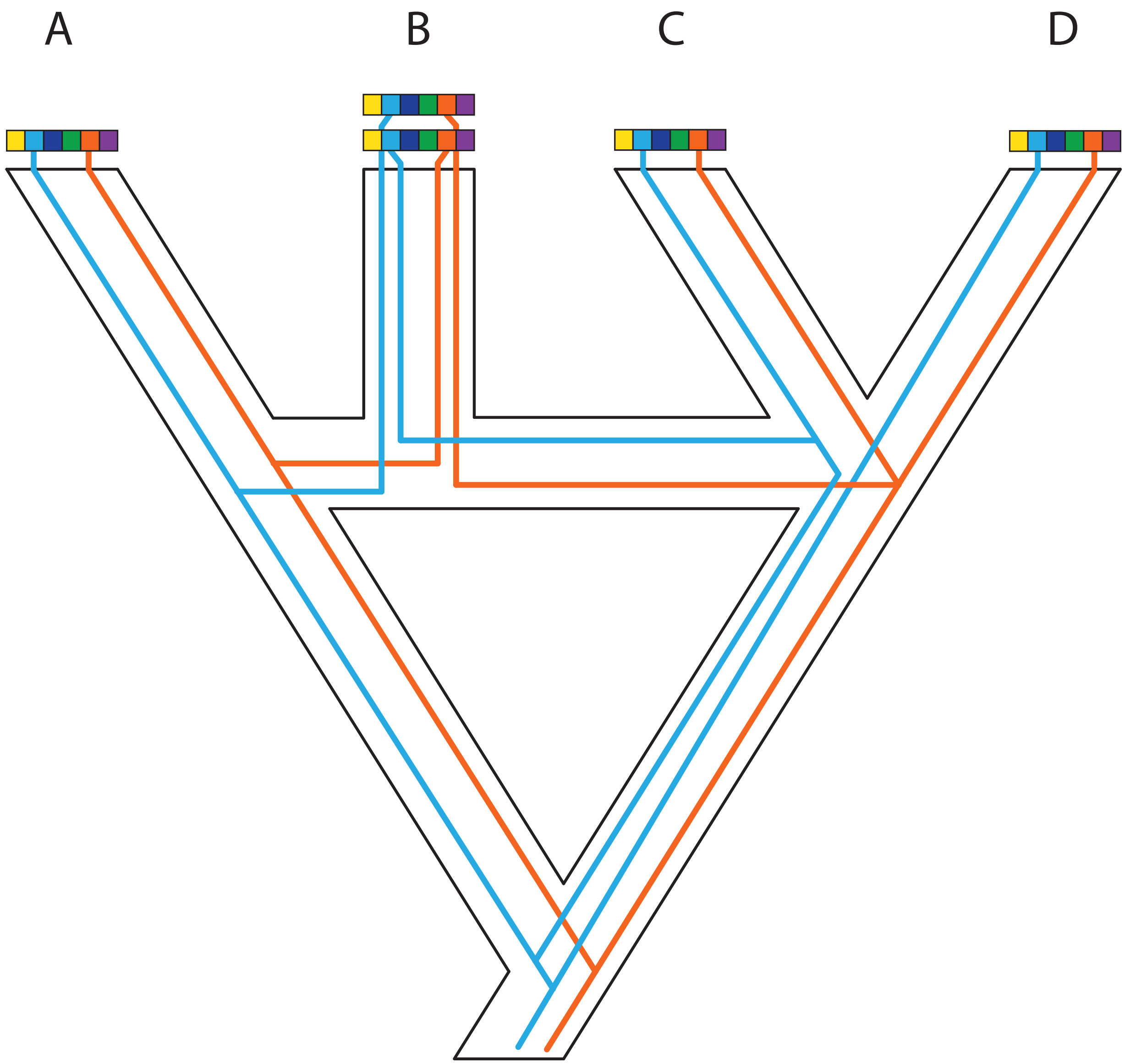


Figure 5. Complex patterns of gene lineages with polyploidization and interspecific gene flow. Genes and genomes of four species A, B, C and D. Multi-colored bars show (parts of) genomes with a number of loci indicated in different colors. Two gene trees, one orange and one blue, evolve within the species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



PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26449v3 | CC BY 4.0 Open Access | rec: 30 Jan 2018, publ: 30 Jan 2018

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.

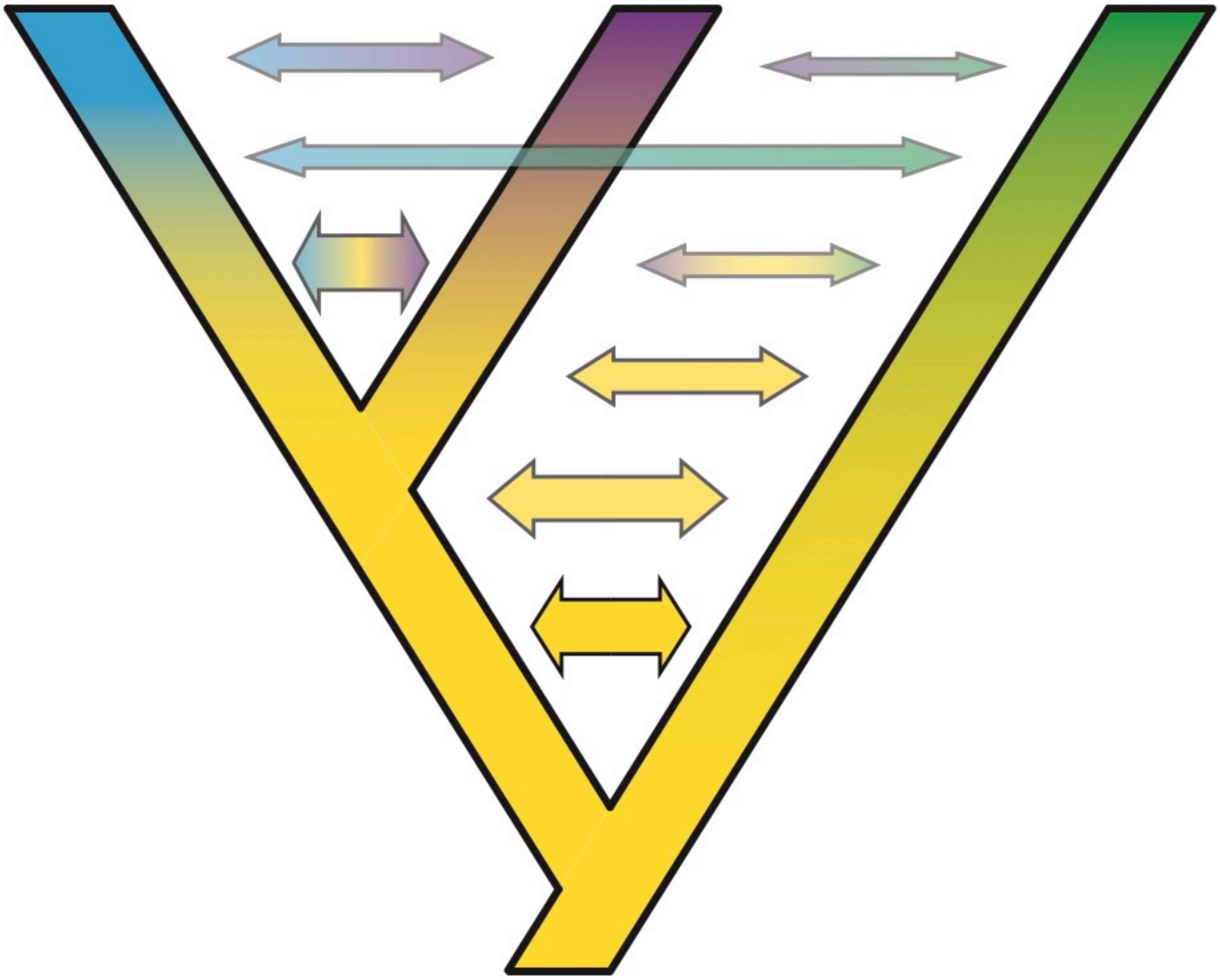
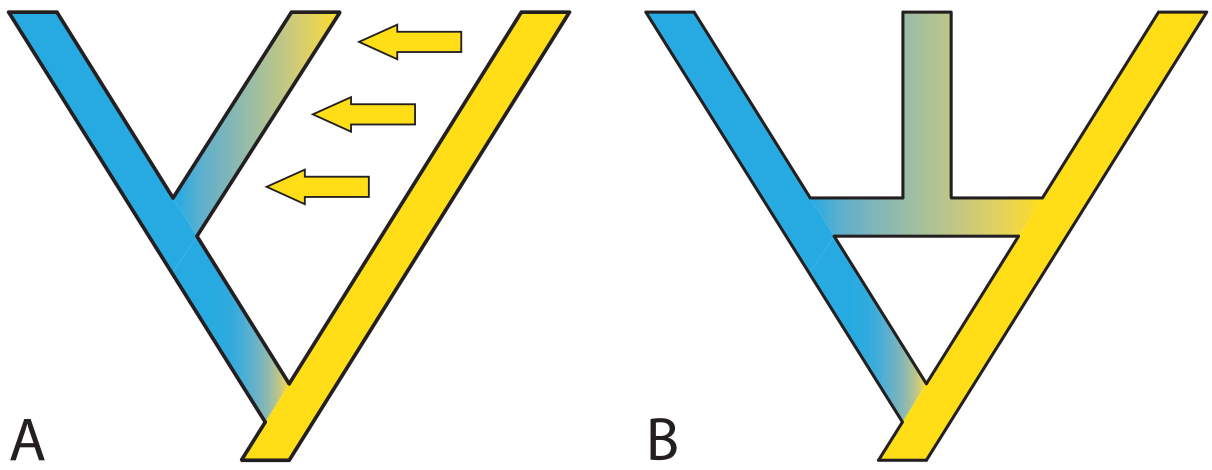


Figure 7. Two possible species phylogenies producing similar observations at present time. On the left (A), there is a species tree with gene flow. On the right (B), there is a species network with homoploid hybridization. Distinguishing two such scenarios usually requires simulations and comparison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



PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.26449v3 | CC BY 4.0 Open Access | rec: 30 Jan 2018, publ: 30 Jan 2018