

A peer-reviewed version of this preprint was published in PeerJ on 14 February 2019.

[View the peer-reviewed version](https://peerj.com/articles/6399) (peerj.com/articles/6399), which is the preferred citable publication unless you specifically need to cite this preprint.

Bravo GA, Antonelli A, Bacon CD, Bartoszek K, Blom MPK, Huynh S, Jones G, Knowles LL, Lamichhaney S, Marcussen T, Morlon H, Nakhleh LK, Oxelman B, Pfeil B, Schliep A, Wahlberg N, Werneck FP, Wiedenhoeft J, Willows-Munro S, Edwards SV. 2019. Embracing heterogeneity: coalescing the Tree of Life and the future of phylogenomics. PeerJ 7:e6399
<https://doi.org/10.7717/peerj.6399>

1 Article submission to *PeerJ*
2 Manuscript category: Literature Review Articles
3 Collection: “*Endless forms: Advances in evolutionary analyses of biodiversity*”
4
5

6 Article title: **Embracing heterogeneity: building the Tree of Life and the future of phylogenomics**
7
8

9 Gustavo A. Bravo*¹, Alexandre Antonelli^{1,2,3,4}, Christine D. Bacon^{2,3}, Krzysztof Bartoszek⁵, Mozes P. K.
10 Blom⁶, Stella Huynh⁷, Graham Jones³, L. Lacey Knowles⁸, Sangeet Lamichhane¹, Thomas Marcussen⁹,
11 H el ene Morlon¹⁰, Luay K. Nakhleh¹¹, Bengt Oxelman^{2,3}, Bernard Pfeil³, Alexander Schliep¹², Niklas
12 Wahlberg¹³, Fernanda P. Werneck¹⁴, John Wiedenhoeft^{12,15}, Sandi Willows-Munro¹⁶, Scott V. Edwards^{1,17}
13

14 ¹ Harvard University, Department of Organismic and Evolutionary Biology, Museum of Comparative
15 Zoology, 26 Oxford St., Cambridge, MA 02138, USA.

16 ² Gothenburg Global Biodiversity Centre, Box 461, SE-405 30 G teborg, Sweden.

17 ³ University of Gothenburg, Department of Biological and Environmental Sciences, Box 461, 405 30
18 G teborg, Sweden.

19 ⁴ Gothenburg Botanical Garden, Carl Skottsbergs gata 22A, SE-41319 G teborg, Sweden.

20 ⁵ Link ping University, Department of Computer and Information Science, SE-581 83 Link ping,
21 Sweden.

22 ⁶ Swedish Museum of Natural History, Department of Bioinformatics and Genetics, Box 50007, 104 05
23 Stockholm, Sweden.

24 ⁷ Universit  de Neuch tel, Institut de Biologie, Rue Emile-Argand 11, Bureau D322, Neuch tel,
25 Switzerland.

26 ⁸ University of Michigan, Department of Ecology and Evolutionary Biology, Ann Arbor, MI 48109,
27 USA.

28 ⁹ University of Oslo, Centre for Ecological and Evolutionary Synthesis, P.O. Box 1066 Blindem, NO-
29 0316, Oslo, Norway.

30 ¹⁰ Institut de Biologie de l’Ecole Normale Sup rieure (IBENS), CNRS UMR 8197, INSERM U1024,
31 Ecole Normale Sup rieure, Paris Sciences et Lettres (PSL) Research University, F-75005 Paris, France.

32 ¹¹ Rice University, Department of Computer Science, 6100 Main St. MS 132, Houston, TX 77005, USA.

33 ¹² Chalmers University of Technology and University of Gothenburg, Department of Computer Science
34 and Engineering, Computing Science Division, SE-412 96, G teborg, Sweden.

35 ¹³ Lund University, Department of Biology, S lvegatan 37, SE-223 62 Lund, Sweden.

36 ¹⁴ Instituto Nacional de Pesquisa da Amaz nia, Coordena o de Biodiversidade, Programa de Cole o es
37 Cient ficas Biol gicas, 69060-000 Manaus, AM, Brazil.

38 ¹⁵ Rutgers University, Department of Computer Science, 110 Frelinghuysen Rd. Piscataway, NJ 08854,
39 USA.

40 ¹⁶ University of Kwazulu-Natal, School of Life Sciences, Pietermaritzburg, South Africa.

41 ¹⁷ Chalmers University of Technology and University of Gothenburg, Gothenburg Centre for Advanced
42 Studies in Science and Technology, SE-412 96, G teborg, Sweden.

43
44 *Corresponding author: Gustavo A. Bravo, E-mail: gustavo_bravo@fas.harvard.edu
45
46

47 **Abstract**

48 Building the Tree of Life (ToL) is a major challenge of modern biology, requiring major advances in
49 cyberinfrastructure, data collection, theory, and more. Here, we argue that phylogenomics stands to
50 benefit by embracing the many heterogeneous genomic signals emerging from the first decade of large-
51 scale phylogenetic analysis spawned by High-throughput sequencing (HTS). Such signals include those
52 most commonly encountered in phylogenomic datasets, such as incomplete lineage sorting, but also those
53 reticulate processes emerging with greater frequency, such as recombination and introgression. We
54 suggest that methods of data acquisition and the types of markers used in phylogenomics will remain
55 restricted until *a posteriori* methods of marker choice are made possible with routine whole-genome
56 sequencing of taxa of interest. We discuss limitations and potential extensions of a major model
57 supporting innovation in phylogenomics today, the multispecies coalescent model. Macroevolutionary
58 models that use phylogenies, such as character mapping, often ignore the heterogeneity on which building
59 phylogenies increasingly rely, and suggest that assimilating such heterogeneity is an important goal
60 moving forward. Finally, we argue that an integrative cyberinfrastructure linking all steps of the process
61 of building the ToL, from specimen acquisition in the field to publication and tracking of phylogenomic
62 data, as well as a culture that values contributors to each step, are essential for progress.

63
64 **KEYWORDS:** gene flow, genome, multispecies coalescent model, retroelement, speciation,
65 transcriptome.

66 **I. Introduction**

67 Charles Dickens famously wrote in *A Tale of Two Cities* “It was the best of times it was the worst of
68 times.” The same could be said about phylogenomics today. Phylogenomics has been invigorated with the
69 introduction of high-throughput sequencing (HTS) and increased breadth of phylogenomic sampling,
70 which have allowed researchers interested in the Tree of Life to scale up in several dimensions, placing
71 both fields squarely in the era of ‘big data’. Additionally, conceptual advances and improvements of
72 statistical models used to analyze these data are helping bridge what some have perceived as a gap
73 between phylogenetics and phylogeography (e.g., Felsenstein, 1988; Huson, 2006; Edwards et al., 2016a).
74 However, as datasets become larger, researchers are inevitably faced with a plethora of heterogeneous
75 signals that often appear to depart from a dichotomously-branching phylogeny (Kunin, Goldovsky &
76 Darzentas, 2005; Jeffroy et al., 2006; Mallet, Besansky & Hahn, 2015). These signals cover an
77 increasingly large array of biological processes at the level of genes and genomes, as well as individual
78 organisms and populations, including processes such as recombination, hybridization, gene flow, and
79 polyploidization. These heterogeneous signals can be thought of as conflicting, but in truth, they are
80 simply a record of the singular history that we commonly refer to as the Tree of Life. One of the grand
81 challenges of evolutionary biology is deciphering this history, whether at the level of genes, populations,
82 species, or genomes. In this perspective piece, we argue that unabashedly embracing this heterogeneity
83 conceptually and analytically will lead to increased insight into the Tree of Life and its underlying
84 processes.

85 A key concept introduced by the scaling up from phylogeography to phylogenomics is the
86 continuum of processes and analytical methods, – the so-called phylogeography-phylogenetics continuum
87 (Edwards et al., 2016a). We argue here that bridging this continuum is critical for advancing
88 phylogenetics. This can be done by either developing phylogenomic approaches that acknowledge and
89 explicitly account for phylogeographic processes, or by determining the regions of parameter space (e.g.,
90 branch lengths in tree, level of gene flow) if any, where such within-species processes are not relevant.

91 For example, the choice of markers in a given phylogenomics project is currently guided more by
92 convenience and cost than by evaluating the biological properties and phylogenomic signals in those data;
93 but comparisons of signals across various types of markers (e.g., transcriptomes, noncoding regions)
94 reveal that marker choice is a critical step toward shedding light on the history of populations and
95 unraveling potential processes underlying such history (Rokas et al., 2003; Cutter, 2013; Jarvis et al.,
96 2014; Reddy et al., 2017). On the analysis side, we are in desperate need of methods that can handle the
97 increasingly large data sets being produced by empiricists, but at the same time there is a desire to include
98 increasingly diverse sources of signal in estimates of divergence times, biogeographic history, and models
99 of diversification (Delsuc, Brinkmann & Philippe, 2005; Jeffroy et al., 2006; Kumar et al., 2012). Finding
100 the balance between breadth, depth, and computational feasibility in project design and statistical analysis
101 is crucial for the field today.

102 A general issue in thinking about the future of phylogenomics is – what do researchers want in
103 the realm of phylogenomics? What does society want? Researchers in phylogenomics are motivated by
104 many factors. Some are excited about building the Tree of Life. Others are less interested in the tree itself,
105 but instead are focused on studying conservation, ecological, and evolutionary processes within a
106 phylogenetic framework. Society will likely benefit from results and archiving practices for data and
107 genetic resources that ensure longevity, reproducibility, and relevance to societal problems. Are the
108 priorities that society places on the many disciplines feeding into scientific efforts toward the Tree of Life
109 – fieldwork, museum collections, databases – appropriate for this grand mission? Although we cannot
110 possibly answer all of these questions within the scope of this perspective, we hope to at least spur
111 discussion on the wide range of field, laboratory, conceptual, and societal issues that allow
112 phylogenomics to move forward.

113

114 **II. Data generation and data types in phylogenomics**

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

133

134 ***Data quality***

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

140 examples of errors in phylogenomic data sets have been reported in the literature, including the use of
141 unintended paralogous sequences in alignments (e.g., Struck, 2013); mistaking the genome sequence of
142 one species for another (Philippe et al., 2011); and inclusion of genome sequence from parasites into the
143 genome of the host (Kumar et al., 2013). However, the incidence of smaller errors in alignments that are
144 not easily discerned from natural allelic variation, such as base miscalls or misplaced indels, are probably
145 much more widespread than has been reported in the literature. Combined with the sensitivity of some
146 phylogenomic datasets to individual loci or Single Nucleotide Polymorphisms (SNP) within loci (Shen,
147 Hittinger & Rokas, 2017), such errors could have damaging consequences for phylogenomic studies, both
148 for topologies and even more so for branch lengths of phylogenetic trees (Marcussen et al., 2014;
149 Bleidorn, 2017).

150 Sequencing high-quality samples from well-archived voucher specimens is a good first step to
151 increase reproducibility and alleviate issues related to sample identity (Peterson et al., 2007; Pleijel et al.,
152 2008; Chakrabarty, 2010; Turney et al., 2015). For individual phylogenomic studies, wholesale manual
153 inspection of every locus is unsustainable (Irisarri et al. 2017), but spot checks of a subset of the data
154 (e.g., 5-10% of the alignments) is a recommended best practice (Phillipe et al., 2011) that is beginning to
155 be encouraged in peer review and in published papers (Montague et al., 2014; Liu et al., 2017). Such
156 checking is important not only for new data generated by a given study, but also for data downloaded
157 from public repositories such as NCBI and Orthomam, which are well known to contain errors (Wesche,
158 Gaffney & Keightley, 2004; Ranwez et al., 2007; Douzery et al., 2014). Because several databases do not
159 include the raw sequence data it is often impossible to evaluate whether oddities may derive from poor
160 sequencing. Robust pipelines for flagging poorly aligned sites or non-homologous sequences, based on
161 existing tools or novel scripts such as Gblocks (Castresana, 2000; Talavera & Castresana, 2007) or
162 TrimAl (Capella-Gutiérrez, Silla-Martínez & Gabaldon, 2009) are gradually being put into practice
163 (Marcussen et al., 2014; He et al., 2016; Irisarri et al., 2017).

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

172

173 ***Data generation and marker development***

174 *Genome reduction methods:* A growing number of genome reduction methods are now providing
175 empiricists with the means to generate genomic subsets suitable for phylogenetic and phylogeographic
176 inference (reviewed by McCormack et al., 2013; Leaché & Oaks 2017; Lemmon & Lemmon, 2013). For
177 phylogenomics, most prominently featured are sequence-capture, focusing on highly conserved regions
178 (e.g., Faircloth et al., 2012; Lemmon, Emme & Lemmon, 2012; reviewed by Jones & Good, 2016) and
179 transcriptomes (e.g., Misof et al., 2014; Cohen et al., 2016; Fernández et al., 2014; Park et al., 2015;
180 Simion et al., 2017; Irisarri et al., 2017), but phylogenomic trees have also been constructed based on
181 restriction-digest methods that focus on single nucleotide polymorphisms or SNPs (Leaché, Chavez &
182 Jones, 2015; Harvey et al., 2016) and analysis of transposable elements (e.g., Suh, Smeds & Ellegren,
183 2015). This diversity of marker types for phylogenetics should be celebrated, but each marker type brings
184 with it a list of pros and cons. For example, many questions in the higher level phylogenetics of animals
185 and plants have so far relied almost exclusively on transcriptome data, for the simple reason that non-
186 coding portions of the genome are difficult if not impossible to align and analyze. However, the uncritical
187 use of transcriptomes in phylogenetics is not without caveats. At high taxonomic levels, coding regions
188 can exhibit extreme levels of among-taxon base composition, sometimes resulting in strong violations of

189 phylogenetic models (Romiguier et al., 2016; Romiguier & Roux, 2017). Coding regions can exhibit
190 reduced levels of incomplete lineage sorting (ILS) compared to noncoding regions (Scally et al., 2012).
191 Such reduced ILS could in fact be helpful in building complex phylogenies with rapid radiations
192 (Edwards, 2009b), but it will certainly distort estimated branch lengths when coalescent methods, which
193 assume neutrality, are used. SNPs have been advocated by some authors (Leaché & Oaks, 2017), but the
194 available methods for analyzing such data are still extremely limited. For example, concatenation and two
195 coalescent methods (SNAPP and SVD quartets: Bryant et al., 2012; Chifman & Kubatko, 2014) have
196 recently been highlighted as the main methods available for phylogenomic analysis of SNPs (Leaché &
197 Oaks, 2017). But each of these methods has its shortcomings. It is likely that concatenation of SNPs will
198 be misleading for many of the same reasons that concatenation of sequence-based markers are misleading
199 (Kubatko and Degnan 2007). SNAPP, a coalescent method suitable for analysis of SNPs (Bryant et al.,
200 2012), works well only on relatively small data sets, and it is unclear how well SVD quartets performs on
201 some data sets (Shi & Yang, 2018). Although SNPs do provide a helpful route around the oft-violated
202 assumption in coalescent models of no recombination within loci (Bryant et al., 2012), and are widely
203 seen as excellent markers for phylogeography and population genetics, it remains to be seen how
204 powerful they are at high phylogenetic levels.

205 Despite the diversity of marker types for phylogenomics, it remains unclear whether features
206 specific to each marker type can ultimately result in phylogenomic datasets that can strongly mislead. For
207 example, incongruence in the phylogeny of modern birds developed by Jarvis et al. (2014; 48 whole
208 genomes) and Prum et al. (2015; 259 anchored phylogenomics loci, 198 species) has recently been
209 attributed to differences in marker type rather than number of taxa (Reddy et al., 2017). Whereas Jarvis et
210 al. (2014) used primarily noncoding loci because they observed gross incongruence when using coding
211 regions, the loci used by Prum et al. (2015), although nominally focused on broadly “anchored”
212 conserved regions, in fact was dominated by coding regions. Thus, at least one or the other marker type or
213 their analysis are likely inappropriate when applied across modern birds. These data type effects can stem

214 from multiple sources. Selection on exons might lead to localized differences in effective population size
215 across the genome and previous studies have highlighted base composition heterogeneity within exons
216 across taxa (Figuet et al., 2015; Scally et al., 2012). On the other hand, alignment quality of introns and
217 ultraconserved elements (UCEs) can sometimes be less than desired (Edwards, Cloutier & Baker, 2017).
218 Clearly marker effects can potentially have substantial consequences on species tree estimates and need to
219 be further evaluated and compared side-by-side by the phylogenetic community (Shen, Hittinger &
220 Rokas, 2017).

221
222 *A priori versus a posteriori selection of loci for phylogenomics:* In an ideal world, phylogeneticists
223 would have whole and fully annotated genomes of all taxa available, allowing them to select loci for
224 phylogenomics based on the relative merits of different loci. This *a posteriori* selection of loci for
225 phylogenomics is clearly a long-term goal that will yield greater choice and justification for specific loci
226 over the *a priori* selection protocol that now dominates the field. Today, the loci for phylogenomics are
227 based primarily on ease of collection and alignment but many potentially useful regions of the genome are
228 ignored by each specific method, whether it is transcriptomes or hybrid-capture. Thus, an attractive aspect
229 of whole-genome sequencing (WGS) for phylogenomics is to have the opportunity to select markers *a*
230 *posteriori* once genomes are in hand (e.g., Edwards, Cloutier & Baker 2017; Fig. 1). WGS also allows for
231 further expansion into different research fields and questions based on the same initial data. In contrast, *a*
232 *priori* marker selection often limits the kinds of questions and methods that researchers can apply and
233 represents a real constraint for phylogenomics and other disciplines.

234 An important constraint for using WGS for downstream phylogenomic analyses is genome
235 quality. Obtaining high-coverage well-assembled and thoroughly annotated genomes is still very
236 expensive and time-consuming, and even low coverage genomes are still outside reach for large portions
237 of the community. However, even low coverage genomes can yield a modest number of markers for
238 phylogenomics, and in the short term might yield data sets allowing a broader diversity of markers for

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

241

242 *More taxa versus more loci*: The question of whether to add more genes or more taxa was a dominant
243 theme in phylogenetics in the 1990s and early 2000s (e.g., Hillis, 1996; Kim, 1996), and remains an
244 important theme guiding phylogenomics today. After much debate in the literature (e.g., Hillis, 1996;
245 Graybeal, 1998; Hillis, 1998; Poe, 1998; Mitchell, Mitter & Regier, 2000), the initial consensus view
246 from the Sanger sequencing era of phylogenetics, is that adding more taxa generally improves
247 phylogenetic analysis more so than more markers (e.g., Hillis, 1996; Graybeal, 1998; Poe, 1998).

248 However, phylogenomics is adding a new twist to this consensus, both from the standpoint of data
249 acquisition and from theory (e.g., Rokas, 2005; Nabhan & Sarkar, 2012; Xi et al., 2012; Patel, Kimball &
250 Braun, 2013). Amassing large data sets, both in terms of more taxa and more loci, is still a guiding
251 principle of phylogenomics. But with the ability now to bring together many different types of markers in
252 a single analysis, and to analyze them in ways that were not previously available, the “more taxa vs. more
253 genes” debate is becoming more nuanced (Nabhan & Sarkar, 2012). For example, recent work shows that
254 this debate can be highly context-specific and model-dependent (e.g., Baurain, Brinkmann & Philippe,
255 2006; Dell Ampio et al., 2013; Edwards et al., 2016b). Also, some phylogenetic methods, such as
256 coalescent methods, appear to be more robust to limited taxon sampling than traditional methods like
257 concatenation (Song et al., 2012; Liu, Xi & Davis, 2015). Some researchers favor “horizontal” data
258 matrices, wherein the number of loci far exceeds the number of taxa, whereas other researchers favor
259 “vertical” matrices, where many taxa are analyzed at just a few (1-5) loci. Whereas the PCR era of
260 phylogenetics was often dominated by vertical matrices, HTS is allowing data matrices to become more
261 horizontal (Fig. 2). Scaling up in both dimensions will be crucial for improved phylogenies, and the
262 number of loci required to resolve a given phylogenetic problem, at least in a coalescent framework, is

263 often a function of the coalescent branch lengths in the phylogenetic tree being resolved, with longer
264 branches requiring fewer loci (Edwards et al., 2007; Huang et al., 2010).

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

281

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

288 that can contribute to gene tree discord. At the same time, the variance in gene tree topologies could also
289 have been caused by errors in gene tree estimation. Such observations have been used to argue that the
290 accuracy of gene tree inference should be maximized or at least evaluated, but it is not clear what criteria
291 should be used to filter sets of gene trees. For example, filters can be based on rates of molecular
292 evolution (Klopfstein, Massingham & Goldman, 2017), levels of phylogenetic informativeness (Fong et
293 al., 2012), or on the cause of gene-tree discord itself, if known (Huang et al., 2010). Chen, Liang & Zhang
294 (2015) found that selecting genes whose trees contained a well-known uncontested long branch in a given
295 species phylogeny was a better way to improve phylogenomic signal than selecting genes based on
296 characteristics of sequence evolution. However, the effects of such culling on the distribution of gene
297 trees, and whether it could distort the distribution so that it no longer conforms to models like the
298 multispecies coalescent, are unknown, and potentially of concern (but see Huang et al., 2017). We need
299 further studies on the effects of different types of phylogenomic filters on the properties of large-scale
300 phylogenomic datasets.

301

302 *Heterozygosity and Intra-Individual Site Polymorphisms*: One of the prevalent occurrences in organisms
303 with multiple ploidies are intra-individual polymorphisms and heterozygosity is, of course, common in
304 diploid organisms. However, confident identification of such polymorphism has always been challenging
305 (Garrick, Sunnucks & Dyer, 2010; Lischer, Excoffier & Heckel, 2014; Schrepff et al., 2016) and many
306 data sets do not permit statistical approaches, such as PHASE (Stephens, Smith & Donnelly, 2001) to
307 robustly determine haplotypes of different alleles (Garrick, Sunnucks & Dyer, 2010). Consequently, in
308 phylogenetics, heterozygosity and intra-specific polymorphic sites are either accommodated using UIPAC
309 ambiguity codes or ignored entirely or by randomly selecting alleles (Iqbal et al., 2012). In fact, most
310 “one allele per individual/species” phylogenomic data sets consists of haplotypes that in fact do not occur
311 in nature, in so far as many methods yield single haplotypes consisting of consensus or other haplotype
312 summaries from diploid organisms. The fact that HTS produces several reads of the same region allows

313 the identification of heterozygosity and intra-specific polymorphic sites represents an untapped
314 opportunity to incorporate intra-individual variation in our phylogenetic estimates (Lischer, Excoffier &
315 Heckel, 2014; Schrepff et al., 2016; Andermann et al., 2018). Recent models have been proposed to
316 improve calling and sorting such polymorphisms (De Maio, Schlötterer & Kosiol, 2013; Lischer,
317 Excoffier & Heckel, 2014; Potts et al., 2014; Schrepff et al., 2016) and, although results of different
318 studies vary (Kubatko, Gibbs & Bloomquist, 2011; Lischer et al. 2014), estimation of individual,
319 naturally occurring haplotypes has been shown to improve phylogenomic reconstructions based on
320 genome-scale data (Andermann et al., 2018).

321
322 *Rare genomic changes:* As noted above, molecular phylogenetics has primarily used alignments of
323 sequence-level data for phylogenetic inference. This bias is perhaps driven by the notion that genome
324 evolution occurs by aggregating small changes, such as point substitutions, over time, but more likely it is
325 due to the challenges of characterizing rare genomic changes in genomes, such as indels, transpositions,
326 inversions, and other large-scale genomic events (Rokas & Holland, 2000; Boore 2006; Bleidorn 2017).
327 This emphasis on sequence data has produced a vast ecosystem of algorithms tailored to analyze such
328 data, but most phylogeneticists would agree that rare genomic changes would be a welcome addition to
329 the toolkit of phylogenomics, since they are generally regarded as highly informative markers, providing
330 strong evidence of homology and monophyly (Boore 2006; Rogozin et al., 2008). With the increased
331 availability and affordability of WGS, our view of genome plasticity has changed drastically in recent
332 years and we are now capable of exploring other genomic features beyond the signals encapsulated in
333 DNA or amino acid sequences. The question then arises of how to identify and utilize these rare genomic
334 markers. Genome-level characters will likely have different evolutionary properties than sequence-based
335 markers, suggesting that one of the biggest challenges we face for incorporating genomic changes into
336 phylogenetic analyses is to find informative evolutionary models and tools suited for these kinds of data.

337

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

347

348 *Indels and transpositions*: Indels and transpositions are two types of molecular characters that are
349 underutilized in phylogenomics, the former perhaps because standard methods of analysis often treat
350 indels as missing data and the latter because they are technically challenging to collect without whole
351 genome data. Indels have been used sporadically in phylogenomics and several have argued for their
352 utility and informativeness, given appropriate analytical tools (Jarvis et al., 2014; Ashkenazy et al., 2014;
353 Roncal et al., 2016). Murphy et al. (2007) used indels in protein-coding regions to bolster estimates of
354 mammalian phylogeny and found that the Atlantogenata hypothesis was supported after scrutinizing
355 proteome-wide indels for spurious alignments and orthology. The Avian Phylogenomics Project found
356 that indels had less homoplasy than SNPs and, despite showing high levels of ILS, was largely congruent
357 with other markers across the avian tree. Transposable elements arguably are even more highly favored
358 by phylogenomics researchers, but are much more difficult to isolate and analyze and have been used
359 principally across various studies in mammals and birds (Kaiser et al., 2007; Churakov et al., 2010;
360 Kriegs et al., 2010; Suh et al., 2011; Baker et al., 2014). Whereas they are generally considered to have a
361 low rate of homoplasy, most researchers agree that they can in some circumstances exhibit insertional
362 homoplasy. Moreover, no marker is immune to the challenges of ILS, and transposable elements and

363 indels are no exception (Matzke et al., 2012; Suh, Smeds & Ellegren, 2015). Still, the exceptional
364 resolution afforded by some studies employing transposable elements is exciting, and we expect this
365 marker type to increase in use as whole genomes are collected with higher frequency.

366

367 *Copy Number Variations (CNV)*: The 1000 Genomes Project estimates that in humans about 20 million
368 base pairs are affected by structural variants, including copy number variations (CNV) and large deletions
369 (1000 Genomes Project Consortium, 2015), suggesting that these types of mutations encompass a higher
370 fraction of the genome than do SNPs in humans. A CNV is a DNA segment of at least one kilobase (kb)
371 that varies in copy number compared with a reference genome (Redon et al., 2006). CNVs appear as
372 deletions, insertions, duplications, and complex multi-site variants (Fredman et al., 2004). Such a
373 profusion of CNVs across human genomes has proven useful in tracking population structure (Sjödín &
374 Jakobsson, 2012), but still remains underappreciated in phylogenetics.

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

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

389

390 **III. Concepts and models in phylogenomics**

391 For decades, phylogenetics has struggled with how best to translate evolutionary changes in DNA
392 sequences and other characters into phylogenies, and genomic data are no exception to this trend.

393 Phylogenomics is still in a developing stage of formulating models that effectively represent the
394 underlying mechanisms for genome-scale variation while remaining efficient and within reasonable
395 analytical and bioinformatic capacities. The current focus on models and evolutionary forces generating
396 the patterns that we recover as branching and reticulation events in our phylogenetic reconstructions is a
397 healthy one, and can be extended to other important topics in phylogenomics, such as species
398 delimitation, character mapping, and trait evolution (e.g., Yang and Rannala 2014). All of these areas are
399 developing rapidly and are in need of updated models and bioinformatics applications to cope with the
400 heterogeneity brought by genome-scale data.

401

402 *The multispecies coalescent (MSC) model*

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

412 histories across loci scattered throughout the genome, the MSC lays at the core of the conceptual
413 framework to deal with genome-scale data (e.g., Rannala & Yang, 2008; Liu et al., 2015). In the few
414 instances in which model comparison and fit has been evaluated (Liu and Pearl 2007; Edwards et al.
415 2007), the MSC vastly outperforms concatenation. This of course does not mean that the MSC is the
416 correct, or even an adequate, model for phylogenomic data, and we need more tests of model adequacy
417 and fit, using Bayesian methods for example (Reid et al., 2014). Despite concerns regarding some of its
418 implementations when dealing with genomic data (e.g., Springer & Gatesy, 2016), there is consensus
419 among systematists that the MSC is a powerful theoretical model for phylogenomics and that there is
420 room for refinement and improvement for its applications (e.g., Edwards et al., 2016b, Xu & Yang, 2016).

421

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

423 Given the huge computational difficulties involved in modelling all the complexities of evolutionary
424 processes in a statistical framework, there is interest in methods that will accommodate genome-scale data
425 for large numbers of species. The utility of such methods cannot be overstated: the rapid rise of large-
426 scale genomic data sets has clearly outstripped theoretical and computational methods required to analyze
427 them. For example, although progress is being made regarding scalability of full Bayesian methods of
428 species phylogeny inference (e.g., Ogilvie, Bouckaert & Drummond, 2017), they are still unable to
429 accommodate large phylogenomic datasets, which often consist of hundreds of species for thousands of
430 loci (Supplementary Table S1). A common approach to speeding up species phylogeny inference consists
431 of ‘two-step’ methods, wherein gene trees are estimated first and separately from the species phylogeny;
432 then, using various summaries of the coalescent process for collections of gene trees, a species phylogeny
433 is estimated. Many useful methods for estimating species phylogenies in this way have been proposed
434 (see Marcussen et al., 2014; Liu, Wu & Yu, 2015; Mirarab & Warnow, 2015; Mirarab, Bayzid &
435 Warnow, 2016), taking advantage of various summaries of the coalescent process, such as the average
436 ranks of pairs of species in the collection of gene trees (e.g., STAR: Liu et al., 2009; ASTRAL-II:

437 Mirarab et al., 2015) or the distribution of gene trees containing triplets of species (e.g., MP-EST; Liu, Yu
438 & Edwards, 2010). Some of these two-step methods, while approximate, nonetheless allow for statistical
439 testing in a likelihood framework. For example, MP-EST can evaluate the (pseudo)likelihood of two
440 proposed species phylogenies given a collection of gene trees and the difference in likelihood can be used
441 to evaluate two proposed species phylogenies against each other. However, such statistical approaches
442 have rarely been used thus far, and bootstrapping or approximate posterior probabilities on branches are
443 by far the most common statistics applied to species phylogenies (Sayyari & Mirarab, 2016). Speeding up
444 the estimation process using two-step methods can be effective, but it can also accumulate errors or
445 misallocate sources of variance which cannot be corrected at later stages (Xu & Yang, 2016). If gene trees
446 are biased or uninformative, then downstream analyses for species phylogeny estimation or species
447 delimitation may similarly be compromised (e.g., Olave et al., 2014). For example, MP-EST can
448 sometimes perform poorly when Phym1 is used to build low-information gene trees because Phym1 may
449 produce biased gene trees when the alignments contain very similar sequences (Xi, Liu & Davis, 2015).
450 This may account for the lower performance of MP-EST compared to ASTRAL in some simulation
451 conditions, because ASTRAL resolves input polytomies and zero-length branches in gene trees more
452 appropriately. This difference between MP-EST and ASTRAL is eliminated when RaxML is used to
453 build gene trees (Xi, Liu & Davis, 2015).

454

455 ***Beyond the multispecies coalescent model***

456 *Reticulation at multiple levels challenges the standard multispecies coalescent model*

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

462 higher level will cause correlated branching patterns at lower levels. At the same time, reticulations at
463 lower levels, such as recombination acting within genes, will cause inference problems at higher levels,
464 such as estimating population histories. Crucially, however, it is only recombination that will break one
465 key element driving many recent models of phylogenetics and population histories, namely dichotomous
466 gene trees. Reticulations at levels of organization higher than the genome, such as the fusing of
467 populations, as well as gene duplication, will still yield collections of dichotomous gene trees, even if the
468 higher-level history is reticulated. Ultimately, the additive effects of these reticulate processes result in
469 our observed phylogenetic reconstructions, and we expect all of these scenarios to produce bifurcating,
470 dichotomous gene trees. From a modelling point of view, another key distinction is whether at the species
471 level, we still have a phylogeny that is tree-like, or whether a network is needed. The process whereby
472 two populations jointly produce a third requires a network to model properly. Allopolyploidy is another
473 situation requiring a network. There are several statistical methods for inferring homoploid networks (Yu
474 et al., 2014; Solis-Lemus & Ané, 2016; Wen et al., 2016; Wen & Nakhleh, 2018), species histories under
475 allopolyploidy (Jones, Sagitov & Oxelman, 2013), and some two-step methods such as PADRE (Huber et
476 al., 2006; Lott et al., 2009). In general, dealing with multiple simultaneous violations of the MSC, such as
477 introgression and allopolyploidy, remains challenging. It is likely that the history of many radiations
478 involves parts of the genome with a dichotomous history and parts that exhibit reticulation, demanding
479 methods that accommodate both scenarios. Alternatively, rather than trying to accommodate multiple
480 processes in our methods for phylogenetic inference, we might instead focus our attention on subsets of
481 loci that would not violate the MSC (e.g., Knowles et al., 2018). In cases where processes other than
482 incomplete lineage sorting are contributing to gene tree discord (i.e., the distribution of trees is
483 statistically inconsistent with expectations under the MSC; see Smith et al., 2015), loci consistent with the
484 MSC model might be identified (e.g., separated from loci with horizontal gene transfer), using the newly
485 developed program CLASSIPHY (Huang et al., 2017).

486 Models accommodating a dichotomous divergence with gene flow are somewhat limited. For
487 example, in IMA2 (Hey & Nielsen, 2004; Hey & Nielsen, 2007; Hey, 2010) the species phylogeny must
488 be known and fairly small; in the method of Dalquen et al. (2017), both the species phylogeny and gene
489 trees are restricted to three tips. Looking forward, it may be useful to deal with two sub-problems: The
490 first sub-problem is estimating the species phylogeny despite some migration, for example by identifying
491 which loci are interfering with the species phylogeny inference or causing reticulations in the form of
492 gene flow. The second sub-problem is to incorporate a gradual speciation process (Fig. 6), where gene
493 flow after speciation slowly declines, perhaps according to some simple function like an exponential.
494 Such a model would capture what is thought to be a more common speciation process than the
495 instantaneous process modelled by the MSC (Jones 2017).

496 In some cases, it is possible to model the same situation with either a species network or a tree
497 with gene flow. Long (1991) discussed two models of admixture: Intermixture and gene flow, illustrated
498 in Figure 7. The phylogenetics community has mainly focused on methods for inference under the
499 intermixture model (e.g., the multispecies network coalescent; Yu et al., 2014), whereas the population
500 genetics community has focused more on models including gene flow (e.g., IM, admixture graphs, G-
501 PhoCS, Phrapl). While some initial work to test inference based on one of these models on data generated
502 by the other has recently appeared (Wen & Nakhleh, 2018; Solís-Lemus et al., 2017; Zhang et al., 2018),
503 much more work is needed to bring together these two lines of work. Simulations and comparisons of
504 observed and expected summary statistics, such as the site-frequency spectrum (Excoffier et al., 2013),
505 have proven especially useful in distinguishing such scenarios (Fig. 7).

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

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

518

519 *Polyploidy and the challenges of analyzing gene duplication and loss*

520 The MSC model describes well allelic lineages and the mutations they accumulate (Fig. 3; Degnan &
521 Rosenberg, 2009; Liu, Xi & Davis, 2015). The simple MSC model is challenging to apply to evolutionary
522 events in which the evolving entities (genes or paralogs) duplicate and occasionally go extinct during the
523 evolutionary history of the populations/species and thus cannot be sampled in contemporary population or
524 species. Estimating the existence and number of these “ghost” lineages remains challenging. For example,
525 how can we detect duplication events if one of the duplicated loci is lost in descendant lineages? In the
526 case of polyploidy, two (or more) genomes having separate evolutionary histories end up together in a
527 single individual. What consequences for evolutionary history do genomic conflicts and dosage variation
528 in gene expression impose? Polyploidy also raises technical issues, such as whether or not homoeologous
529 sequences are recovered in standard genomic surveys.

530 The complication that gene duplication and loss (GDL) brings to the inference of species
531 phylogenies has long been recognized (Fitch, 1970). It is therefore surprising that practical solutions to
532 the problem of GDL are almost non-existent, with empirical examples usually based on *ad hoc* methods
533 and deductions. Ancient duplications where most additional copies are retained in descendent species can
534 be fairly easy to diagnose based on phylogeny (Oxelman et al., 2004; Pfeil et al., 2004). However,

535 resolving duplications becomes more difficult when copy number changes quickly (Ashfield et al., 2012),
536 or when duplications are recent and copy loss is complete or nearly so, thus returning the locus to a
537 single-copy state (Ramadugu et al., 2013). In the latter case, the phylogenetic pattern can mimic that of
538 ILS and become indistinguishable from it (Sousa et al., 2017), generally leaving no trace at all of the loss.

539 Why is GDL so challenging to implement in theory? The topological and coalescent-time
540 similarities between ILS and GDL complicates extending the MSC to include both processes, unless copy
541 number exceeds one in at least some samples (Fig. 4). Assuming that allelic and homoeologous variation
542 is not confused with the copy number of independently duplicated genes, at the very least, duplicated
543 genes could be handled as independent loci with missing data for some samples, and ordinary MSC
544 inference undertaken. When copy loss is complete, or when the duplication is so recent so as to conflate
545 allelic versus copy variation, these GDL loci have little effect on species phylogeny inference and
546 divergence times, especially if the algorithms used employ averages over coalescence times or other
547 parameters across many gene trees (Liu et al., 2009; Sousa et al., 2017). At high proportions, though, they
548 may cause serious issues for phylogenetic reconstruction, because the unexpected positions of gene
549 duplications in a species phylogeny, coupled with random copy loss, means that no specific pattern is
550 expected among the affected gene trees (Fig. 4). This scenario contrasts with the retention of ancestral
551 polymorphisms, where we know that branches in short species phylogenies (in coalescent units) are the
552 cause (Rosenberg & Nordborg, 2002). Thus, we expect deeply coalescing lineages to occur in specific
553 parts of a species phylogeny with a limited number of topological outcomes and branch lengths limited by
554 effective population size, which is not the case for duplicated genes. A recent approach to identifying
555 genes that are single copy, but have nonetheless been affected by GDL, was made using the genomic
556 location of the loci (Sousa et al., 2017), and could prove useful for distinguishing GDL and ILS.

557

558 *Recombination*

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

568 Researchers have started to examine the first question and found a detectable effect of
569 recombination only under extreme levels of ILS and gene tree heterogeneity (e.g., Lanier & Knowles,
570 2012). However, more analyses and studies are still needed to explore a wider range of factors and
571 parameters that could affect species phylogeny inference when the assumption of recombination-free loci
572 is violated. For answering the second question, one approach involves combining the multispecies
573 coalescent with hidden Markov models (e.g., Hobolth et al., 2007). These methods suffer from the “state
574 explosion problem”, where individual states are needed for the different coalescent histories, and they
575 increase rapidly with the number of taxa in the dataset, making them infeasible except for very small (~4
576 taxa) datasets. New methods that scale to larger datasets are needed if such approaches are to be useful in
577 practice. A different direction is to devise novel methods for inferring species phylogeny while assuming
578 that the genealogies of the individual loci could take the form of an ancestral recombination graph (ARG:
579 Siepel, 2009).

580 Extending these approaches to address recombination would require the development of new
581 models that significantly extend the multispecies coalescent to account for ARGs within the branches of a
582 species phylogeny. For two-step species tree methods, this entails developing new methods that infer
583 ARGs for the individual loci and methods that infer species phylogenies from collections of ARGs. For

584 single-step (Bayesian) methods, novel developments are needed to sample species phylogenies, locus-
585 specific ARGs, and their related parameters. It will also be important to better understand the conditions
586 under which ignoring recombination will still yield reasonable estimates of phylogeny. Extending the
587 theory to accommodate ARGs may be of intrinsic interest, but if the parameter space in which
588 recombination is relevant is very small, then practitioners may be able to ignore recombination.

589

590 *Species concepts and delimitation*

591 Coalescent methods have played an important role in the development and critical evaluation of species
592 delimitation methods because they provide hypotheses for species boundaries based on genetic and can
593 now be integrated with phenotypic data (e.g., Solis-Lemus et al., 2015). Irrespective of traditional species
594 concepts, it is essential that the entities at the tips of the species tree do not violate the assumptions of the
595 MSC, wherein the definition of species is mathematically clear-cut (e.g., Rannala & Yang, 2003, Degnan
596 & Rosenberg, 2009): the branches of the species tree constitute species or populations that do not
597 exchange genes. However, the MSC model also carries strict assumptions about the divergence process if
598 the delimited units are to be interpreted as species. Specifically, it is important to emphasize that in the
599 “standard” MSC model, these species represent populations that, immediately after divergence, no longer
600 experience gene flow. Therefore, the species of the MSC model do not necessarily correspond with
601 species as a taxonomic rank, defined by traditional species concepts (Heled & Drummond, 2010): “MSC”
602 species could simply be populations by other criteria, so long as they have ceased to exchange genes,
603 even for a short period of time. In other words, a species tree built under the MSC might then be
604 interpreted as a depiction of the history of the barriers to gene flow among diverging structured
605 populations (Sukumaran & Knowles 2017). Therefore, in those species-phylogeny methods requiring *a*
606 *priori* assignments of individuals to species, such assignments may strongly influence the inferred species
607 phylogeny, in the same way that hybridization will have serious consequences on an estimated species
608 phylogeny (Leaché et al., 2014).

609 Recently, several MSC-based methods that have the ability to simultaneously perform species
610 delimitation and estimate the species phylogenies have been developed and implemented (e.g., Yang &
611 Rannala, 2014; Jones, Aydin, & Oxelman, 2015; Jones, 2016). These methods seem to consistently
612 recover the correct number of “MSC species” given the assumptions of the model. However, it is
613 probable that the assumption of no gene flow between the descendant populations is often violated and
614 that most reproductive isolation processes are gradual or episodic rather than sudden and permanent (e.g.,
615 Rosindell et al., 2010). There is thus need for methods that perform simultaneous species phylogeny
616 estimation and assignment of individuals to species while taking into account the limitations of the MSC
617 (Jones, Aydin, & Oxelman, 2015).

618 If one prefers a species concept that affirms that most recently diverged populations are
619 necessarily reproductively isolated, current methods will overestimate the number of species as defined
620 by traditional species concepts, and will likely reveal instead intraspecific population structure
621 (Sukumaran & Knowles, 2017). Toprak et al. (2016) used DISSECT (Jones, Aydin & Oxelman, 2015)
622 but also employed checks as to the integrity of various hypotheses of species boundaries suggested by the
623 data. From a computational point of view, any species delimitation method will need an operational
624 definition of species. Therefore, a possible development of MSC-based species delimitation methods
625 could be allowing migration and assuming that speciation is complete when a certain proportion of the
626 migrations is reached or when the migration rate is sufficiently low. However, this solution will not be
627 suited for the protracted speciation model because other kinds of information besides the movement of
628 genes will still be needed to identify when a clade becomes reproductively isolated. Possibly the best way
629 to avoid confusion is to restrict the word “species” to taxonomy and base it on multiple sources of
630 information which are synthesized in an integrative fashion (Dayrat, 2005; Will, Mishler & Wheeler,
631 2005; Bacon et al., 2012; Solis-Lemus, et al., 2015), and refer to the reproductively isolated units of MSC
632 analysis as “MSC units” or “MSC taxa”.

633

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

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

642 Macroevolutionary models focus on long-term processes, in terms of both species richness and
643 phenotypic diversity. They rely on two types of models: birth-death models of diversification aimed at
644 understanding how and why speciation and extinction rates vary through time and across lineages (Hey
645 1992; Nee, Mooers & Harvey, 1992; see Stadler 2013 and Morlon, 2014 for review) and models of trait
646 evolution aimed at understanding the mode and tempo of phenotypic evolution (Felsenstein, 1973; see
647 Pennell & Harmon, 2013 and Manceau, Lambert & Morlon, 2017 for reviews). These models are
648 typically constructed at the level of species, ignoring the populations or individuals that constitute these
649 species (but see Manceau, Lambert & Morlon, 2015 and Rosindell, Harmon & Etienne, 2015 for
650 exceptions). As a consequence, microevolutionary processes such as coalescence have informed
651 phylogenetic methods for building phylogenies more so than have macroevolutionary methods that use
652 them. For example, the most widely used phylogenetic dating methods generally do not acknowledge the
653 critical distinction between speciation times, which are usually of primary interest, and coalescence times,
654 which are often assumed to represent speciation times but in fact represent events older than the
655 divergence of the species concerned (Edwards & Beerli, 2000; dos Reis, Donoghue & Yang, 2016;
656 Angelis & dos Reis 2015). In addition, macroevolutionary models are fit to species phylogenies
657 (diversification models) or a combination of species phylogenies and phenotypic data (trait evolution
658 models), most often assuming that evolution is best represented by a species tree, not a network (but see

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

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

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

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

687

688 ***Mapping trait evolution on heterogeneous genomic datasets***

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

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

712

713 *Tree-free methods of character evolution*

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

726 Tree-free comparative methods work by integrating over the space of trees (under a given
727 branching process model). For example, under a pure birth model and with enough tip measurements, the
728 optimum value of the OU process can be estimated as the sample average (Bartoszek & Sagitov, 2015a).
729 Similar results have now been derived for other models of tree growth that include extinction (Adamczak
730 and Miłoś, 2014; 2015; Ané, Ho & Roch, 2017). Similarly, the rate of adaptation under the OU process,
731 often modeled as the stationary variance – the ratio of the squared “rate of evolution” (sigma parameter in
732 the OU model) and twice the “rate of adaptation” (the alpha parameter) can be estimated as the sample
733 variance (Sagitov and Bartoszek 2015a). Teasing sigma and alpha apart, however, requires a tree. The key

734 parameter of the BM model, the rate of evolution, is similarly estimable directly from the trait sample
735 (Bartoszek & Sagitov, 2015b; Crawford & Suchard, 2013), whereas the root state cannot be consistently
736 estimated without a tree (Ané, 2008; Sagitov & Bartoszek, 2012). In addition to providing tree-free
737 estimators of some model parameters, the studies mentioned above also derived Central Limit Theorems
738 that allow computing confidence intervals around these point estimates as well as the sample sizes needed
739 to obtain reliable estimates.

740

741 *Extinct and unsampled species*

742 A notable case when phylogenomics and macroevolution do meet is in the treatment of extinct or
743 unsampled species in phylogenetic reconstruction and dating. Despite the avalanche of genomes for an
744 increasing number of species, we still lack sequence data for most species, making it difficult to place
745 them in a phylogeny. Some researchers (e.g., Jetz et al., 2012; Tonini et al., 2015) have opted to impute
746 the phylogenetic relationships of unsampled species. In this case, polytomies are often resolved by using
747 distributions of branching times obtained from macroevolutionary birth-death models (Kuhn, 2011).
748 While such approaches elicit a culture clash between those who laboriously build trees and those that
749 simply use them, there are other approaches stemming from macroevolution that are less offending to
750 phylogeny builders. For example, recent results using conditioned birth-death processes (e.g., Gernhard,
751 2008a; b; Sagitov & Bartoszek, 2012) show that under constant rate processes the size of the clade
752 contributes information on the height of the tree and also on the coalescence times. Such results can be
753 used to improve the calibration and node dating of the phylogeny when some species are not sampled.
754 One would expect that ignoring the non-sequenced species would incur a bias resulting in shorter tree
755 heights, because less time is usually required to generate fewer tips. Conditioned branching process
756 models can help alleviate this bias. Also, macroevolutionary birth-death models are used as branching
757 process priors in Bayesian molecular dating. The availability of likelihood expressions for incompletely
758 sampled phylogenies (Stadler 2009; Stadler & Steel 2012; Morlon et al, 2011) thus allow to date

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

761

762 **V. Building, updating and sustaining the Tree of Life**

763 *Scalability challenges*

764 Inferring the phylogeny of all living organisms represents a different challenge than inferring the
765 relationships of just a few terminals; often the scale at which new methods are developed and tested is on
766 this latter scale. For instance, for eukaryotes alone, recent conservative estimates indicate that there are
767 ~8.7 million species on Earth and only 9-14% of them have been formally described (Mora et al., 2011).
768 Furthermore, out of 2.6 million taxa currently represented in the Open Tree of Life
769 (<https://tree.opentreeoflife.org>; Hinchliff et al., 2015), only ~55,000 were gathered from hard-data
770 phylogenies, whereas phylogenetic affinities of the rest were inferred from current taxonomic
771 classifications (McTavish et al., 2015; 2017). These observations suggest that the vast majority of taxa on
772 Earth still await formal taxonomic description and placement in the Tree of Life (Mora et al., 2011;
773 McTavish et al., 2017). One common challenge that phylogeneticists encounter towards that end is the
774 difficulty in accessing samples from rare, endangered, or extinct taxa, particularly in countries where
775 collecting and exporting is not possible. Recent genomic techniques now allow successful results in
776 obtaining valuable DNA data from museum specimens (e.g., Staats et al., 2013; Hykin, Bi & McGuire,
777 2015; McCormack, Tsai & Faircloth, 2016; McCormack et al., 2017; Ruane & Austin, 2017), and here,
778 we advocate for routine use of these resources to enhance research in phylogenomics and
779 phylogeography.

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

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

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

803

804 ***Community initiatives***

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

809 provides a synthesis of previously published phylogenies merged through supertree and other grafting
810 methods. One issue faced by the initiative is that it relies on authors uploading their phylogenetic trees to
811 open data repositories, such as Dryad Data Repository (<http://datadryad.org/pages/organization>; Vision,
812 2010) or TreeBase (Sanderson et al., 1994; Piel et al., 2009), which at least until recently only occurred in
813 about 17% of cases (Drew, 2013). Substantial curatorial efforts are also critical to facilitate reusability of
814 deposited trees (McTavish et al., 2015). A different approach was taken by Antonelli et al. (2017), who
815 developed a framework for continuously inferring time-calibrated large phylogenies from raw sequence
816 data deposited in GenBank (Clark et al., 2016) in a multi-step method. Similarly, various tools have been
817 developed to make information contained in the Tree of Life available for the general public (e.g.,
818 Rosindell & Harmon, 2012; Harmon et al., 2013).

819

820 *Mapping the Tree of Life*

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

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

835

836 ***Best practices for building the Tree of Life***

837 *Data must be well curated and publicly available*

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

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

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

874

875 *The need of adequate curation of analytical tools*

876 In the same way that data must be adequately stored and curated, analytical tools must be available for
877 future use and should guarantee proper reproducibility (Wilson et al., 2014). One of the reasons behind
878 the dramatic increase in the number of phylogeographic and phylogenetic studies during the last 20 years
879 is the proliferation of software and bioinformatic tools to process and analyze these data. Thanks to these
880 new methods, it is now possible to implement a wide array of theoretical models that sustain the fields of
881 phylogenomics and phylogeography. As stated above, genome-wide data have notoriously increased the
882 necessity to expand our analytical models, ultimately leading to a stronger demand for computing

883 resources. Given their key role in phylogenomic research, it is advisable that software development,
884 documentation, and availability follow the best possible practices (e.g., Leprevost et al., 2014; Wilson et
885 al., 2014; Guang et al., 2016). Having both data and analytical tools adequately stored and accessible to
886 the public not only will ensure high reproducibility of previous studies, but, more importantly, will
887 facilitate continuing the construction of the Tree of Life (McTavish et al., 2017).

888

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

890 Some current publishing practices in the scientific community may unintentionally represent hurdles
891 toward the ultimate end of collecting and disseminating phylogenetic data on which to build a Tree of
892 Life. For instance, the increasing need in many countries and communities for publishing high-impact
893 papers understandably often discourages researchers from releasing their data until their studies are
894 complete and have passed the peer-review process. This is partially explained by the heavy emphasis of
895 top journals on unusually novel and flashy findings as compared to those studies that represent more
896 modest, but just as critical, advances in the understanding of the phylogenetic relationships of the groups.
897 Similarly, this urge to publish high-impact papers often impedes adequate long-term studies that could
898 potentially generate a wider variety of basic data. With the cultural emphasis on impact and numbers and
899 rates of publication, in practice there is often a penalty for long-term studies. Our current climate often
900 values novel results produced in the short term. Consequently, as a community, we must reach an
901 equilibrium between short- and long-term scientific production in a way that values both, encouraging
902 high impact studies bringing radical reorganizations of the Tree of Life, without hurting lower impact
903 research and the ongoing search for innovation.

904 Moreover, because building the Tree of Life is a slow and daunting task, it is important that, as a
905 scientific community, all contributors to the process receive proper recognition for their contributions,
906 thereby keeping motivation high and retaining our best talent. Unfortunately, some contributors, both
907 institutions and roles within them, receive less recognition in this grand task than others. For example,

908 field biologists that obtain basic natural history information and specimens used for building the Tree of
909 Life (Suarez & Tsutsui, 2004), and the natural history museums that house those specimens, are often not
910 recognized sufficiently. As a community, we have been following a trend in which, perhaps inadvertently,
911 we do not value as much the production of basic biological and natural history data. This can certainly be
912 recognized in our national funding practices, which often do not support basic taxonomic or natural
913 history fieldwork at the expense of flashier end-uses of biological specimens. Specimens are the
914 foundation of most phylogenomic and phylogeographic studies, and we should find standard mechanisms
915 not only to acknowledge, but also to encourage the production of these data in an integrative framework.
916 It is time to strengthen those initiatives aimed at recognizing scientific production beyond citations of
917 peer-reviewed literature (e.g., ORCID; <https://orcid.org>) by giving also credit to the production and
918 impact of basic biology datasets and collected specimens. Providing credit for depositing and generating
919 data by tracking, for example, number of access and downloads or number of studies using genetic data
920 associated to specimens, could represent a formal recognition of the importance of producing and sharing
921 basic biological data could help bridge the gap between naturalists, taxonomists, empiricists, and
922 mathematicians invested on the study of life history.

923 It will be exciting to have objective estimates that allow tracking the direct and indirect impact of
924 how these data and samples are being used. We are confident that such initiatives will highlight the
925 importance of continuing field- and museum-based research in various fields of biological research
926 (Buerki & Baker, 2016). Furthermore, such cultural shifts will undoubtedly encourage discerning young
927 minds to embrace basic biological research in their academic endeavors, rather than embracing more
928 lucrative and societally appreciated applied fields.

929

930 **VI. Conclusions**

931 In this perspective, we have attempted to cover ground in the vast arena of issues facing modern
932 phylogenomics today. We have seen how genome-scale phylogenomics, currently on a strong footing as a

933 result of the multispecies coalescent model, is increasingly infiltrated by models that recognize reticulate
934 processes, such as recombination and introgression. By contrast, macroevolutionary models that use
935 phylogenies have yet to embrace the heterogeneity that currently drives many theoretical innovations in
936 phylogenetic reconstruction itself. We have emphasized the need for the phylogenomics community to
937 embrace high standards of data quality, curation and accessibility in its long-term pursuit of the Tree of
938 Life. Such a grand mission requires value and recognition placed not only on the end products of the
939 process, such as publications and trees, but also on the natural history specimens on which phylogenies
940 are based and which are cared for by the community of natural history museums. Building the tree of life
941 will require contributions from all sectors of biological and related sciences – from field biology to theory
942 and everything in between – and robust cyberinfrastructures to integrate these diverse and increasingly
943 massive data streams.

944

945 **ACKNOWLEDGEMENTS**

946 This paper is a product of the ‘Origin of Biodiversity Workshop’ organized by Chalmers University of
947 Technology and the University of Gothenburg, under the auspices of the Gothenburg Centre for
948 Advanced Studies (GoCAS). We are particularly grateful to the GoCAS organizers and facilitators, in
949 particular Karin Hårding, Mattias Marklund, Bernt Wennberg, Sandra Johansson, and Lotta Fernström.
950 We thank Johnathan Clark, Alison Cloutier, Phil Grayson, Kathrin Näpflin, Flavia Termignoni, Jonathan
951 Schmitt, Simon Sin, João Tonini, and Pengcheng Wang for help compiling Supplementary Table 1.
952 Thomas Couvreur and Tobias Andermann provided useful comments that improved the contents of this
953 manuscript.

954

955 **FUNDING STATEMENT**

956 The Gothenburg Center for Advanced Studies (GoCas) workshop ‘Origins of Biodiversity’ was funded by
957 Chalmers University of Technology and the University of Gothenburg. The following researchers are

958 supported by scholarship or research grants from the following agencies: Swedish Research Council
959 (B.O., A.A.), US National Science Foundation, the European Research Council under the European
960 Union’s Seventh Framework Programme (FP/2007-2013, ERC Grant Agreement n. 331024 to A.A.), the
961 Swedish Foundation for Strategic Research and a Wallenberg Academy Fellowship (A.A.). F.P.W. would
962 like to thank Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Partnerships for
963 Enhanced Engagement in Research from the U.S. National Academy of Sciences, the U.S. Agency of
964 International Development (PEER NAS/USAID), and the L’Oreal-Unesco For Women in Science
965 Program.

966

967 **AUTHOR CONTRIBUTIONS**

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

972

973 **CONFLICT OF INTEREST**

974 The authors declare no conflict of interests.

975

976 **REFERENCES**

977 1000 Genomes Project Consortium. 2015. A global reference for human genetic variation. *Nature* 526:
978 68–74.

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

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

- 983 Albalat, R., Cañestro, C. 2016. Evolution by gene loss. *Nature Reviews Genetics* 17: 379–391.
- 984 Ané, C., Ho, L. S. T., Roch, S. 2017. Phase transition on the convergence rate of parameter estimation
985 under an Ornstein-Uhlenbeck diffusion on a tree. *Journal of Mathematical Biology* 74: 355–385.
- 986 Ané, C. 2008. Analysis of comparative data with hierarchical autocorrelation. *Annals of Applied*
987 *Statistics* 2: 1078–1102.
- 988 Andermann, T., Fernandes, A. M., Olsson, U., Topel, M., Pfeil, B., Oxelman, B., Aleixo, A., Faircloth,
989 B. C., Antonelli, A. 2018. Allele Phasing Greatly Improves the Phylogenetic Utility of
990 Ultraconserved Elements. *BioRxiv* doi: <https://doi.org/10.1101/255752>.
- 991 Angelis, K., dos Reis, M. 2015. The impact of ancestral population size and incomplete lineage sorting on
992 Bayesian estimation of species divergence times. *Current Zoology* 61:874–885.
- 993 Antonelli, A., Hettling, H., Condamine, F. L., Vos, K., Nielsson, R. H., Sanderson, J., Sauquet,
994 H., Scharn, R., Silvestro, D., Töpel, M., Bacon, C.D., Oxelman, B., Vos, R. A. 2017. Towards a Self-
995 Updating Platform for Estimating Rates of Speciation and Migration, Ages, and Relationships of
996 Taxa. *Systematic Biology* 66:152–166.
- 997 Ashfield, T., Egan, A. N., Pfeil, B. E., Chen, N. W. G., Podicheti, R., Ratnaparkhe, M. B., Ameline-
998 Torregrosa, C., Denny, R., Cannon, S., Doyle, J. J., Geffroy, V., Roe, B. A., Saghai-Marooof, M. A.,
999 Young, N. D., Innes, R. W. 2012. Evolution of a complex disease resistance gene cluster in diploid
1000 *Phaseolus* and tetraploid *Glycine*. *Plant Physiology* 159: 336–354.
- 1001 Ashkenazy, H., Cohen, O., Pupko, T., Huchon, D. 2014. Indel Reliability in Indel-Based Phylogenetic
1002 Inference. *Genome Biology and Evolution* 6: 3199–3209.
- 1003 Bacon, C. D., McKenna, M. J., Simmons, M. P., Wagner, W. L. 2012. Evaluating multiple criteria for
1004 species delimitation: an empirical example using Hawaiian palms (*Arecaceae: Pritchardia*). *BMC*
1005 *Evolutionary Biology* 2012: 12–23.
- 1006 Baker, A. J., Haddrath, O., McPherson, J. D., Cloutier, A. 2014. Genomic Support for a Moa-Tinamou
1007 Clade and Adaptive Morphological Convergence in Flightless Ratites. *Molecular Biology and*

1008 Evolution 31: 1686–1696.

1009 Baird, N. A., Etter, P. D., Atwood, T. S., Currey, M. C., Shiver, A. L., Lewis, Z. A., Selker, E. U.,
1010 Cresko, W. A., Johnson, E. A. Rapid SNP discovery and genetic mapping using sequenced RAD
1011 markers. PLoS One 3(10):e3376.

1012 Bartoszek, K. 2014. Quantifying the effects of anagenetic and cladogenetic evolution. Mathematical
1013 Biosciences 254: 42–57.

1014 Bartoszek, K. 2016. Phylogenetic effective sample size. Journal of Theoretical Biology 407: 371–386.

1015 Bartoszek, K., Sagitov, S. 2015a. Phylogenetic confidence intervals for the optimal trait value. Journal of
1016 Applied Probability 52: 1115–1132.

1017 Bartoszek, K., Sagitov, S. 2015b. A consistent estimator of the evolutionary rate. Journal of Theoretical
1018 Biology 371: 69–78.

1019 Bastide, P., Solis-Lemus, C., Kriebel, R., Sparks, K. W., Ané, C. 2017. Phylogenetic Comparative
1020 Methods on Phylogenetic Networks with Reticulations. BioRxiv doi: <https://doi.org/10.1101/194050>

1021 Baurain, D., Brinkmann, H., Philippe, H. 2006. Lack of Resolution in the Animal Phylogeny: Closely
1022 Spaced Cladogeneses or Undetected Systematic Errors? Molecular Biology and Evolution 24: 6–9.

1023 Belyaev, D. K. 1969. Domestication of animals. Science Journal 4: 47–52.

1024 Betancur, R., Naylor, G. J. P., Ortí, G. 2014. Conserved genes, sampling error, and phylogenomic
1025 inference. Systematic Biology 63: 257–262.

1026 Bleidorn, C. 2017. Sources of Error and Incongruence in Phylogenomic Analyses. In: Phylogenomics.
1027 Cham: Springer International Publishing, 173–193.

1028 Blom, M. P. K. 2015. EAPhy: A Flexible Tool for High-throughput Quality Filtering of Exon-alignments
1029 and Data Processing for Phylogenetic Methods. PLoS ToL.

1030 Boore, J. L., Collins, T. M., Stanton, D., Daehler, L. L., & Brown, W. M. 1995. Nature. 376: 163–165.

1031 Boore, J. L., Daehler, L. L., Brown, W. M. Complete sequence, gene arrangement, and genetic code of
1032 mitochondrial DNA of the cephalochordate *Branchiostoma floridae* (Amphioxus). Molecular

1033 Biology and Evolution 16: 410–418.

1034 Boore, J. L. 2006. The use of genome-level characters for phylogenetic reconstruction. Trends in Ecology
1035 and Evolution 21:439–446

1036 Bryant, D., Bouckaert, R., Felsenstein, J., Rosenberg, N. A., RoyChoudhury, A. 2012. Inferring species
1037 trees directly from biallelic genetic markers: bypassing gene trees in a full coalescent analysis.
1038 Molecular Biology and Evolution 29: 1917–1932.

1039 Buerki, S., Baker, W. J. 2016. Collections-based research in the genomics era. Biological Journal of the
1040 Linnean Society 117: 5–10.

1041 Burbrink, F. T., Pyron, R. A. 2011. The Impact of Gene-Tree/Species-Tree Discordance on
1042 Diversification-Rate Estimation. Evolution 65: 1851–1861.

1043 Capella-Gutierrez, S., Silla-Martinez, J. M., Gabaldon, T. 2009. trimAl: a tool for automated alignment
1044 trimming in large-scale phylogenetic analyses. Bioinformatics 25: 1972–1973.

1045 Castresana, J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic
1046 analysis. Molecular Biology and Evolution 17: 540–552

1047 Chakrabarty, P. 2010. Genotypes: a concept to help integrate molecular phylogenetics and taxonomy.
1048 Zootaxa 2632: 67–68.

1049 Chen, M. Y., Liang, D., Zhang, P. 2015. Selecting Question-Specific Genes to Reduce Incongruence in
1050 Phylogenomics: A Case Study of Jawed Vertebrate Backbone Phylogeny. Systematic Biology 64:
1051 1104–1120.

1052 Chifman, J., Kubatko, L. 2014. Quartet inference from SNP data under the coalescent model.
1053 Bioinformatics 30: 3317–3324.

1054 Churakov, G., Sadasivuni, M. K, Rosenbloom, K. R., Huchon, D., Brosius, J., Schmitz, J. 2010. Rodent
1055 Evolution: Back to the Root. Molecular Biology and Evolution 27: 1315–1326.

1056 Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., Sayers, E.W. 2016. GenBank. Nucleic Acids
1057 Research 44: D67–D72.

1058 Cohen, O., Doron, S., Wurtzel, O., Dar, D., Edelheit, S., Karunker, I., Mick, E., Sorek, R. 2016.
1059 Comparative transcriptomics across the prokaryotic Tree of Life. *Nucleic Acids Research* 44: W46–
1060 W53.

1061 Costello, M. J., Wieczorek, J. 2014. Best practice for biodiversity data management and publication.
1062 *Biological Conservation* 173: 68–73.

1063 Crawford, F. W., Suchard, M. A., 2013. Diversity, disparity, and evolutionary rate estimation for
1064 unresolved Yule trees. *Systematic Biology* 62: 439–455.

1065 Cutter, A. D. 2013. Integrating phylogenetics, phylogeography and population genetics through genomes
1066 and evolutionary theory. *Molecular Phylogenetics and Evolution* 69: 1172–1185.

1067 Dalquen, D. A., Zhu, T., Yang, Z. 2017. Maximum Likelihood Implementation of an Isolation-with-
1068 Migration Model for Three Species. *Systematic Biology* 66: 379–398.

1069 Dayrat, B. 2005. Towards integrative taxonomy. *Biological Journal of the Linnean Society* 85: 407–415

1070 De Maio, N., Schlötterer, C., Kosiol, C. 2013. Linking Great Apes Genome Evolution across Time Scales
1071 Using Polymorphism-Aware Phylogenetic Models. *Molecular Biology and Evolution* 30: 2249–
1072 2262.

1073 DeBolt, S. 2010. Copy Number Variation Shapes Genome Diversity in Arabidopsis Over Immediate
1074 Family Generational Scales. *Genome Biology and Evolution* 2: 441–453.

1075 Degnan, J. H., Rosenberg, N. A. 2009. Gene tree discordance, phylogenetic inference and the
1076 multispecies coalescent. *Trends in Ecology & Evolution* 24: 332–340.

1077 Dell Ampio, E., Meusemann, K., Szucsich, N. U., Peters, R. S., Meyer, B., Borner, J., Petersen, M.,
1078 Aberer, A. J., Stamatakis, A., Walz, M. G., Minh, B. Q., Haeseler, von A., Ebersberger, I., Pass, G.
1079 N., Misof, B. 2013. Decisive Data Sets in Phylogenomics: Lessons from Studies on the Phylogenetic
1080 Relationships of Primarily Wingless Insects. *Molecular Biology and Evolution* 31: 239–249.

1081 Delsuc, F., Brinkmann, H., Philippe, H. 2005. Phylogenomics and the reconstruction of the Tree of Life.
1082 *Nature Reviews Genetics* 6: 361–375.

1083 dos Reis, M., Donoghue, P. C. J., Yang, Z. 2016. Bayesian molecular clock dating of species divergences
1084 in the genomics era. *Nature Reviews Genetics* 17:71–80.

1085 Douzery, E. J. P., Scornavacca, C., Romiguier, J., Belkhir, K., Galtier, N., Delsuc, F., Ranwez, V. 2014.
1086 OrthoMaM v8: A Database of Orthologous Exons and Coding Sequences for Comparative Genomics
1087 in Mammals. *Molecular Biology and Evolution* 31:1923–1928.

1088 Doyle, J. J. 1992. Gene trees and species trees: molecular systematics as one character taxonomy.
1089 *Systematic Botany* 17: 144-163.

1090 Drew, B. T. 2013. Data deposition: Missing data mean holes in Tree of Life. *Nature* 493: 305.

1091 Dunn, C. W., Howinson, M., Zapata, F. 2013. Agalma: an automated phylogenomics workflow. *BMC*
1092 *Bioinformatics* 14: 330.

1093 Edwards, S. V., Beerli, P. Perspective: gene divergence, population divergence, and the variance in
1094 coalescence time in phylogeographic studies. *Evolution* 54: 1839–1854.

1095 Edwards, S. V., Cloutier, A., Baker, A. J. 2017. Conserved Nonexonic Elements: A Novel Class of
1096 Marker for Phylogenomics. *Systematic Biology* 66: 1028–1044.

1097 Edwards, S. V., Potter, S., Schmitt, C. J., Bragg, J. G., Moritz, C. 2016a. Reticulation, divergence, and the
1098 phylogeography-phylogenetics continuum. *Proceedings of the National Academy of Sciences* 113:
1099 8025–8032.

1100 Edwards, S. V., Xi, Z., Janke, A., Faircloth, B. C., McCormack, J. E., Glenn, T. C., Zhong, B., Wu, S.,
1101 Lemmon, E. M., Lemmon, A. R., Leaché, A. D., Liu, L., Davis, C. C. 2016b. Implementing and
1102 testing the multispecies coalescent model: A valuable paradigm for phylogenomics. *Molecular*
1103 *Phylogenetics and Evolution* 94: 447–462.

1104 Edwards, S. V. 2009a. Is a new and general theory of molecular systematics emerging? *Evolution* 63:1–
1105 19.

1106 Edwards, S. V. 2009b. Natural selection and phylogenetic analysis. *Proceedings of the National Academy*
1107 *of Sciences of the United States of America* 106: 8799–8800.

- 1108 Elmer, K. R., Meyer, A. 2011. Adaptation in the age of ecological genomics: insights from parallelism
1109 and convergence. *Trends in Ecology and Evolution* 26: 298–306.
- 1110 Excoffier, L., Dupanloup, I., Huerta-Sánchez, E., Sousa, V. C., Foll, M. 2013. Robust demographic
1111 inference from genomic and SNP data. *PLoS Genet* 9: e1003905.
- 1112 Faircloth, B. C., McCormack, J. E., Crawford, N. G., Harvey, M. G., Brumfield, R. T., Glenn, T. C. 2012.
1113 Ultraconserved Elements Anchor Thousands of Genetic Markers Spanning Multiple Evolutionary
1114 Timescales. *Systematic Biology* 61: 717–726.
- 1115 Faurby, S., Svenning, J. C. 2015. A species-level phylogeny of all extant and late Quaternary extinct
1116 mammals using a novel heuristic-hierarchical Bayesian approach. *Molecular Phylogenetics and*
1117 *Evolution* 84: 14–26.
- 1118 Felsenstein, J. 1973. Maximum-likelihood estimation of evolutionary trees from continuous characters.
1119 *American Journal of Human Genetics* 25:471–492.
- 1120 Felsenstein, J. 1985. Phylogenies and the comparative method. *American Naturalist* 125: 1–15.
- 1121 Felsenstein, J. 1988. Phylogenies from molecular sequences: inference and reliability. *Annual Review of*
1122 *Genetics* 22: 521–565.
- 1123 Felsenstein, J. 2012. A comparative method for both discrete and continuous characters using the
1124 threshold model. *American Naturalist* 179: 145–156.
- 1125 Fernández, R., Laumer C. E., Vahtera, V., Libro, S., Kaluziak, S., Sharma, P. P., Pérez-Morro, A. R.,
1126 Edgecombe, G. D., Giribert, G. 2014. Evaluating Topological Conflict in Centipede Phylogeny
1127 Using Transcriptomic Data Sets. *Molecular Biology and Evolution* 31: 1500–1513
- 1128 Figuet, E., Ballenghien, M., Romiguier, J., Galtier, N. 2015. Biased Gene Conversion and GC-Content
1129 Evolution in the Coding Sequences of Reptiles and Vertebrates. *Genome Biology and Evolution* 7:
1130 240–250.
- 1131 Fitch, W. M. 1970. Distinguishing homologous from analogous proteins. *Systematic Zoology* 19: 99–113.
- 1132 Fong, J. J., Brown, J. M., Fujita, M. K., Boussau, B. 2012. A Phylogenomic Approach to Vertebrate

1133 Phylogeny Supports a Turtle-Archosaur Affinity and a Possible Paraphyletic Lissamphibia. PLoS
1134 ONE 7: e48990–14.

1135 Fredman, D., White, S. J., Potter, S., Eichler, E. E., Dunnen, J. T. D., Brookes, A. J. 2004. Complex SNP-
1136 related sequence variation in segmental genome duplications. *Nature Genetics* 36:861–866.

1137 Garrick, R. C., Sunnucks, P., Dyer, R. J. 2010. Nuclear gene phylogeography using PHASE: dealing with
1138 unresolved genotypes, lost alleles, and systematic bias in parameter estimation. *BMC Evolutionary*
1139 *Biology* 10: 118.

1140 Garrick, R. C., Bonatelli, I. A., Hyseni, C., Morales, A., Pelletier, T. A., Perez, M. F., Rice, E., Satler, J.
1141 D., Symula, R. E., Thomé, M. T. C. 2015. The evolution of phylogeographic data sets. *Molecular*
1142 *Ecology* 24: 1164–1171.

1143 Gernhard, T., 2008a. The conditioned reconstructed process. *Journal of Theoretical Biology* 253: 769–
1144 778.

1145 Gernhard, T. 2008b. New analytic results for speciation times in neutral models. *Bulletin of Mathematical*
1146 *Biology* 70: 1082–1097.

1147 Ghiurcuta, C. G., Moret, B. M. 2014. Evaluating synteny for improved comparative studies.
1148 *Bioinformatics* 30: i9–i18.

1149 Goolsby, E. W., Bruggeman, J., Ané, C. 2017. Rphylopar: fast multivariate phylogenetic comparative
1150 methods for missing data and within-species variation. *Methods in Ecology and Evolution* 8: 22–27.

1151 Graybeal, A. 1998. Is it better to add taxa or characters to a difficult phylogenetic problem? *Systematic*
1152 *Biology* 47: 9–17.

1153 Guang, A., Zapata, F., Howison, M., Lawrence, C. E., Dunn, C. W. 2016. An Integrated Perspective on
1154 Phylogenetic Workflows. *Trends in Ecology and Evolution* 31: 116–126.

1155 Gusfield, D. 2015. Persistent phylogeny. In: New York, New York, USA: ACM Press, 443–451.

1156 Hahn, M. W., Nakhleh, L. 2016. Irrational exuberance for resolved species trees. *Evolution* 70: 7–17.

1157 Hammer, M. F., Woerner, A. E., Mendez, F. L., Watkins, J. C., Wall, J. D. 2011. Genetic evidence for

1158 archaic admixture in Africa. *Proceedings of the National Academy of Sciences* 108: 15123–15128.

1159 Hansen, T. F. 1997. Stabilizing selection and the comparative analysis of adaptation. *Evolution* 51: 1341–

1160 1351.

1161 Harmon, L. J., Baumes, J., Hughes, C., Soberón, J., Specht, C. D., Tumer, W., Lisle, C., Thacker, R. W.

1162 2013. Arbor: Comparative Analysis Workflows for the Tree of Life. *PLoS Currents* 5:

1163 ecurrents.tol.099161de5eabdee073fd3d21a44518dc.

1164 Harvey, M. G., Smith, B. T., Glenn, T. C., Faircloth, B. C., Brumfield, R. T. 2016. Sequence Capture

1165 versus Restriction Site Associated DNA Sequencing for Shallow Systematics. *Systematic Biology*

1166 65: 910–924.

1167 He, D., R. Sierra, Pawlowski, J., Baldauf, S. L. 2016. Reducing long-branch effects in multi-protein data

1168 uncovers a close relationship between *Alveolata* and *Rhizaria*. *Molecular Phylogenetics and*

1169 *Evolution* 101: 1–7.

1170 Heled, J., Drummond, A. J. 2010. Bayesian inference of species trees from multilocus data. *Molecular*

1171 *Biology and Evolution* 27: 570–580.

1172 Hey, J., Nielsen, R. 2004. Multilocus methods for estimating population sizes, migration rates and

1173 divergence time, with applications to the divergence of *Drosophila pseudoobscura* and *D. persimilis*.

1174 *Genetics* 167: 747–760.

1175 Hey, J., Nielsen, R. 2007. Integration within the Felsenstein equation for improved Markov chain Monte

1176 Carlo methods in population genetics. *Proceedings of the National Academy of Sciences USA* 104:

1177 2785–2790.

1178 Hey, J. 1992. Using phylogenetic trees to study speciation and extinction. *Evolution* 46: 627–640.

1179 Hey, J. 2010. Isolation with Migration Models for More Than Two Populations. *Molecular Biology and*

1180 *Evolution* 27: 905–920.

1181 Hiller, M., Schaar, B. T., Indjeian, V. B., Kingsley, D. M., Hagey, L. R., Bejerano, G. 2012. A “forward

1182 genomics” approach links genotype to phenotype using independent phenotypic losses among related

1183 species. *Cell Reports* 2: 817–23.

1184 Hillis, D. M. 1996. Inferring complex phylogenies. *Nature* 383: 130–131.

1185 Hillis, D. M. 1998. Taxonomic sampling, phylogenetic accuracy, and investigator bias. *Systematic*
1186 *Biology* 47: 3–8.

1187 Hinchliff, C. E., Smith, S. A., Allman, J. F., Burleigh, J. G., Chaudhary, R., Coghill, L. M., Crandall, K.
1188 A., Deng, J., Drew, B. T., Gazis, R., Gude, K., Hibbett, D. S., Katz, L. A., Laughinghouse, H. D. IV,
1189 McTavish, E. J., Midford, P. E., Owen, C. L., Ree, R. H., Rees, J. A., Soltis, D. E., Williams, T.,
1190 Cranston, K. A. 2015. Synthesis of phylogeny and taxonomy into a comprehensive Tree of Life.
1191 *Proceedings of the National Academy of Sciences USA* 112: 12764–12769.

1192 Hoban, S., Kelley, J. L., Lotterhos, K. E., Antolin, M. F., Bradburd, G., Lowry, D. B., Poss, M. L., Reed,
1193 L. K., Storfer, A., Whitlock, M. C. 2016. Finding the Genomic Basis of Local Adaptation: Pitfalls,
1194 Practical Solutions, and Future Directions. *American Naturalist* 188: 379–397.

1195 Hobolth, A., Christensen, O. F., Mailund, T., Schierup, M. H. 2007. Genomic relationships and speciation
1196 times of human, chimpanzee, and gorilla inferred from a coalescent hidden Markov model. *PLoS*
1197 *Genet.* 3, e7.

1198 Ho, L. S. T., Ané, C. 2014. A linear-time algorithm for Gaussian and non-Gaussian trait evolution
1199 models. *Systematic Biology* 63: 397–408.

1200 Huang, H. T., He, Q. I., Kubatko, L. S., Knowles, L. L. 2010. Sources of error inherent in species-tree
1201 estimation: impact of mutational and coalescent effects on accuracy and implications for choosing
1202 among different methods. *Systematic Biology* 59: 573–583.

1203 Huang, H., Sukumaran, J., Smith, S. A., Knowles, L. L. 2017. Cause of gene tree discord? Distinguishing
1204 incomplete lineage sorting and lateral gene transfer in phylogenetics. *PeerJ Preprints*
1205 <https://doi.org/10.7287/peerj.preprints.3489v1>.

1206 Huber, K. T., Oxelman, B., Lott, M., Moulton, V. 2006. Reconstructing the evolutionary history of
1207 polyploids from multilabeled trees. *Molecular Biology and Evolution* 23: 1784–91.

1208 Huson, D. H. 2006. Application of Phylogenetic Networks in Evolutionary Studies. *Molecular Biology*
1209 *and Evolution* 23: 254–267.

1210 Hykin, S. M., Bi, K., McGuire, J. A. 2015. Fixing Formalin: A Method to Recover Genomic-Scale DNA
1211 Sequence Data from Formalin-Fixed Museum Specimens Using High-Throughput Sequencing. *PLoS*
1212 *ONE* 10(10): e0141579.

1213 Iqbal, Z., Caccamo, M., Turner, I., Fliccek, P., McVean, G. 2012. De novo assembly and genotyping of
1214 variants using colored de Bruijn graphs. *Nature Genetics* 44: 226–232.

1215 Irisarri, I., Baurain, D., Brinkmann, H., Delsuc, F., Sire, J., Kupfer, A., Petersen, J., Jarek, M., Meyer, A.,
1216 Vences, M., Philippe, H. 2017. Phylotranscriptomic consolidation of the jawed vertebrate timetree.
1217 *Nature Ecology and Evolution* 1: 1370–1378.

1218 Jarvis, E. D., Mirarab, S., Aberer, A. J., Li, B., Houde, P., Li, C., Ho, S. Y. W., Faircloth, B. C., Nabholz,
1219 B., Howard, J. T., Suh, A., Weber, C. C., da Fonseca, R. R., Li, J., Zhang, F., Li, H., Zhou, L.,
1220 Narula, N., Liu, L., Ganapathy, G., Boussau, B., Bayzid, M. S., Zavidovych, V., Subramanian, S.,
1221 Gabaldon, T., Capella-Gutierrez, S., Huerta-Cepas, J., Rekepalli, B., Munch, K., Schierup, M.,
1222 Lindow, B., Warren, W. C., Ray, D., Green, R. E., Bruford, M. W., Zhan, X., Dixon, A., Li, S., Li,
1223 N., Huang, Y., Derryberry, E. P., Bertelsen, M. F., Sheldon, F. H., Brumfield, R. T., Mello, C. V.,
1224 Lovell, P. V., Wirthlin, M., Schneider, M. P. C., Prosdocimi, F., Samaniego, J. A., Velazquez, A. M.
1225 V., Alfaro-Nunez, A., Campos, P. F., Petersen, B., Sicheritz-Ponten, T., Pas, A., Bailey, T., Scofield,
1226 P., Bunce, M., Lambert, D. M., Zhou, Q., Perelman, P., Driskell, A. C., Shapiro, B., Xiong, Z., Zeng,
1227 Y., Liu, S., Li, Z., Liu, B., Wu, K., Xiao, J., Yinqi, X., Zheng, Q., Zhang, Y., Yang, H., Wang, J.,
1228 Smeds, L., Rheindt, F. E., Braun, M., Fjeldsa, J., Orlando, L., Barker, F. K., Jønsson, K. A., Johnson,
1229 W., Koepfli, K. P., O'Brien, S., Haussler, D., Ryder, O. A., Rahbek, C., Willerslev, E., Graves, G. R.,
1230 Glenn, T. C., McCormack, J., Burt, D., Ellegren, H., Alstrom, P., Edwards, S. V., Stamatakis, A.,
1231 Mindell, D. P., Cracraft, J., Braun, E. L., Warnow, T., Jun. W., Gilbert, M. T. P., Zhang, G. 2014.
1232 Whole-genome analyses resolve early branches in the Tree of Life of modern birds. *Science*

1233 346:1320–1331.

1234 Jeffroy, O., Brinkmann, H., Delsuc, F., Philippe, H. 2006. Phylogenomics: the beginning of
1235 incongruence? *Trends in Genetics* 22: 225–231.

1236 Jetz, W., Thomas, G. H., Joy, J. B., Hartmann, K, Mooers, A. O. 2012. The global diversity of birds in
1237 space and time. *Nature* 491: 444–448.

1238 Jhwueng, D. C., O’Meara, B. 2015. Trait evolution on phylogenetic networks. *BioRxiv*
1239 doi: <https://doi.org/10.1101/023986>

1240 Jones, M. R., Good, J. M. 2016. Targeted capture in evolutionary and ecological genomics. *Molecular*
1241 *Ecology* 25: 185–202.

1242 Jones, G. 2016. Algorithmic improvements to species delimitation and phylogeny estimation under the
1243 multispecies coalescent. *Journal of Mathematical Biology* 74: 447–467.

1244 Jones, G. R. 2017. Divergence estimation in the presence of incomplete lineage sorting and migration.
1245 *bioRxiv*. <https://www.biorxiv.org/content/early/2017/10/16/174342>

1246 Jones, G., Sagitov, S., Oxelman, B. 2013. Statistical Inference of Allopolyploid Species Networks in the
1247 Presence of Incomplete Lineage Sorting. *Systematic Biology* 62: 467–478.

1248 Jones, G., Aydin, Z., Oxelman, B. 2015. DISSECT: an assignment-free Bayesian discovery method for
1249 species delimitation under the multispecies coalescent. *Bioinformatics* 31: 991–998.

1250 Kaiser, V. B., van Tuinen, M., Ellegren, H. 2007. Insertion events of CR1 retrotransposable elements
1251 elucidate the phylogenetic branching order in galliform birds. *Molecular Biology and Evolution* 24:
1252 338–347.

1253 Kidd, D. M. 2010. Geophylogenies and the Map of Life. *Systematic Biology* 59: 741–752.

1254 Kim, J. 1996. General inconsistency conditions for maximum parsimony: effects of branch lengths and
1255 increasing numbers of taxa. *Systematic Biology* 46: 363–374.

1256 Kingman, J. F. C. 1982. On the genealogy of large populations. *Journal of Applied Probability* 19: 27–43.

1257 Klopstein, S., Massingham, T., Goldman, N. 2017. More on the Best Evolutionary Rate for Phylogenetic

1258 Analysis. *Systematic Biology* 66: 769–785.

1259 Knowles, L. L., Smith, S. A. Huang, H., Sukumaran, J. 2018. A matter of phylogenetic scale:
1260 distinguishing incomplete lineage sorting from lateral gene transfer as the cause of gene tree discord
1261 in recent versus deep diversification histories. *American Journal of Botany*, revision in review.

1262 Kocot, K. M., Struck, T. H., Merkel, J., Waits, D. S., Todt, C., Brannock, P. M., Weese, D. A., Cannon, J.
1263 T., Moroz, L. L., Lieb, B., Halanych, K. M. 2017. Phylogenomics of Lophotrochozoa with
1264 Consideration of Systematic Error. *Systematic Biology* 66: 256–282.

1265 Kowada, L. A. B., Doerr, D., Dantas, S., Stoye, J. 2016. New Genome Similarity Measures Based on
1266 Conserved Gene Adjacencies. In: Singh M. (eds) *Research in Computational Molecular Biology*.
1267 RECOMB 2016. *Lecture Notes in Computer Science*, vol 9649. Springer, Cham.

1268 Kriegs, J. O., Zemann, A., Churakov, G., Matzke, A., Ohme, M., Zischler, H., Brosius, J., Kryger, U.,
1269 Schmitz, J. 2010. Retroposon Insertions Provide Insights into Deep Lagomorph Evolution. *Molecular*
1270 *Biology and Evolution* 27: 2678–2681.

1271 Kubatko, L. S., Degnan, J. H. 2007. Inconsistency of phylogenetic estimates from concatenated data
1272 under coalescence. *Systematic Biology* 56: 17–24.

1273 Kubatko, L. S., Gibbs, H. L., Bloomquist, E. W. 2011. Inferring species-level phylogenies and taxonomic
1274 distinctiveness using multilocus data in *Sistrurus* rattlesnakes. *Systematic Biology* 60: 393–409.

1275 Kuhn, T.S., Mooers, A. Ø., Thomas, G. H. 2011. A Simple Polytomy Resolver for Dated Phylogenies.
1276 *Methods in Ecology and Evolution* 2: 427–36.

1277 Kumar, S., Filipski, A. J., Battistuzzi, F. U., Kosakovsky, P., Pond, S. L., Tamura, K. 2012. Statistics and
1278 Truth in Phylogenomics. *Molecular Biology and Evolution* 29:457–472.

1279 Kumar, S., Jones, M., Koutsovoulos, G., Clarke, M. 2013. Blobology: exploring raw genome data for
1280 contaminants, symbionts and parasites using taxon-annotated GC-coverage plots. *Frontiers in*
1281 *Genetics* 4: 237.

1282 Kunin, V., Goldovsky, L., Darzentas, N. 2005. The net of life: reconstructing the microbial phylogenetic

1283 network. *Genome Research* 15: 954–959.

1284 Lamichhaney, S., Berglund, B., Almén, M.S., Maqbool, K., Grabherr, M., Martinez-Barrio, A.,
1285 Promerova, M., Rubin, C.J., Wang, C., Zamani, N., Grant, B.R., Grant, P.R., Webster, M.T.,
1286 Andersson, L. 2015. Evolution of Darwin's finches and their beaks revealed by genome sequencing.
1287 *Nature* 518: 371–375.

1288 Lanier, H. C., Knowles, L. L. 2012. Is Recombination a Problem for Species-Tree Analyses? *Systematic*
1289 *Biology* 61: 691–701.

1290 Leaché, A. D., Oaks, J. R. 2017. The Utility of Single Nucleotide Polymorphism (SNP) Data in
1291 Phylogenetics. *Annual Review of Ecology, Evolution, and Systematics* 48: 69–84.

1292 Leaché, A. D., Chavez, A. S., Jones, L. N. 2015. Phylogenomics of phrynosomatid lizards: conflicting
1293 signals from sequence capture versus restriction site associated DNA sequencing. *Genome Biology*
1294 7: 706–719.

1295 Leaché, A. D., Harris, R. B., Rannala B., Yanz Z. 2014. The influence of gene flow on species tree
1296 estimation: A simulation study. *Systematic Biology* 63: 17–30.

1297 Lemmon, A. R., Emme, S. A., Lemmon, E. M. 2012. Anchored Hybrid Enrichment for Massively High-
1298 Throughput Phylogenomics. *Systematic Biology* 61: 727–744.

1299 Lemmon, E. M., Lemmon, A. R. 2013. High-Throughput Genomic Data in Systematics and
1300 Phylogenetics. *Annual Review of Ecology Evolution and Systematics* 44: 99–121.

1301 Leprevost, F. V., Barbosa, V. C., Francisco, E. L., Perez-Riverol, Y., Carvalho, P. C. 2014. On best
1302 practices in the development of bioinformatics software. *Frontiers in Genetics* 5: 199.

1303 Li, C., Hofreiter, M., Straube, N., Corrigan, S., Naylor, G. J. P. 2013. Capturing protein-coding genes
1304 across highly divergent species. *BioTechniques* 54: 1–5.

1305 Lin, Y., Hu, F., Tang, J. Moret, B. M. 2013. Maximum likelihood phylogenetic reconstruction from high-
1306 resolution whole-genome data and a tree of 68 eukaryotes. *Pacific Symposium on Biocomputing*
1307 285–296.

- 1308 Lischer, H. E., Excoffier, L., Heckel, G. 2014. Ignoring heterozygous sites biases phylogenomic estimates
1309 of divergence times: implications for the evolutionary history of *Microtus voles*. *Molecular Biology*
1310 *and Evolution* 31: 817–831.
- 1311 Liu, L., Pearl, D. K. 2007. Species trees from gene trees: reconstructing Bayesian posterior distributions
1312 of a species phylogeny using estimated gene tree distributions. *Systematic Biology* 56: 504–514.
- 1313 Liu, L., Wu, S., Yu, L. 2015. Coalescent methods for estimating species trees from phylogenomic data.
1314 *Journal of Systematics and Evolution* 53: 380–390.
- 1315 Liu, L., Xi, Z., Wu, S., Davis, C. C., Edwards, S. V. 2015. Estimating phylogenetic trees from genome-
1316 scale data. *Annals of the New York Academy of Sciences* 1360: 36–53.
- 1317 Liu, L., Xi, Z., Davis, C. C. 2015. Coalescent Methods Are Robust to the Simultaneous Effects of Long
1318 Branches and Incomplete Lineage Sorting. *Molecular Biology and Evolution* 32: 791–805.
- 1319 Liu, L., Yu, L., Edwards, S. V. 2010. A maximum pseudo-likelihood approach for estimating species
1320 trees under the coalescent model. *BMC Evolutionary Biology* 10: 302.
- 1321 Liu, L., Yu, L., Pearl, D. K., Edwards, S. V. 2009. Estimating species phylogenies using coalescence
1322 times among sequences. *Systematic Biology* 58: 468–477.
- 1323 Liu, L., Zhang, J., Rheindt, F. E., Lei, F., Qu, Y., Wang, Y., Sullivan, C., Nie, W., Wang, J., Yang, F.,
1324 Chen, J., Edwards, S. V., Meng, J., Wu, S. 2017. Genomic evidence reveals a radiation of placental
1325 mammals uninterrupted by the KPg boundary. *Proceedings of the National Academy of Science of*
1326 *the USA* 114: E7282–7290.
- 1327 Long, J. C. 1991. The genetic structure of admixed populations. *Genetics* 127: 417–418.
- 1328 Lott, M., Spillner, A., Huber, K. T., Moulton, V. 2009. PADRE: A package for analyzing and displaying
1329 reticulate evolution. *Bioinformatics* 25: 1199–2000.
- 1330 Maddison, W. P. 1997. Gene trees in species trees. *Systematic Biology* 46: 523–536.
- 1331 Magadum, S., Banerjee, U., Murugan, P., Gangapur, D., Ravikesavan. R. 2013. Gene duplication as a
1332 major force in evolution. *Journal of Genetics* 92: 155–161.

- 1333 Mallet, J., Besansky, N., Hahn, M. W. 2015. How reticulated are species? *BioEssays* 38: 140–149.
- 1334 Manceau, M., Lambert, A., Morlon, H. 2015. Phylogenies support out-of-equilibrium models of
1335 biodiversity. *Ecology Letters* 18: 347–356.
- 1336 Manceau, M., Lambert, A., Morlon, H. 2017. A unifying comparative phylogenetic framework including
1337 traits coevolving across interacting lineages. *Systematic Biology* 66: 551–568.
- 1338 Mandáková, T., Joly, S., Krzywinski, M., Mummenhoff, K., Lysak, M. A. 2010. Fast Diploidization in
1339 Close Mesopolyploid Relatives of *Arabidopsis*. *The Plant Cell* 22: 2277–2290.
- 1340 Marcovitz, A., Jia, R., Bejerano, G. 2016. “Reverse Genomics” Predicts Function of Human Conserved
1341 Noncoding Elements. *Molecular Biology and Evolution* 33: 1358–1369.
- 1342 Marcussen, T., Sandve, S. R., Heier, L., Spannagl, M., Pfeifer, M., The International Wheat Genome
1343 Sequencing Consortium, Jakobsen, K. S., Wulff, B. B. H., Steuernagel, B., Mayer, K. F. X., Olsen,
1344 O. A. 2014. Ancient hybridizations among the ancestral genomes of bread wheat. *Science* 345:
1345 1250092.
- 1346 Marcussen, T., Heier, L., Brysting, A. K., Oxelman, B., Jakobsen, K. S. 2015. From gene trees to a dated
1347 allopolyploid network: Insights from the angiosperm genus *Viola* (Violaceae). *Systematic Biology*
1348 64: 84–101.
- 1349 Matzke, A., Churakov, G., Berkes, P., Arms, E. M., Kelsey, D., Brosius, J., Kriegs, J. O., Schmitz, J.
1350 2012. Retroposon Insertion Patterns of Neoavian Birds: Strong Evidence for an Extensive Incomplete
1351 Lineage Sorting Era. *Molecular Biology and Evolution* 29: 1497–1501.
- 1352 McCormack, J. E., Hird, S. M., Zellmer, A. J., Carstens, B. C., Brumfield, R. T. 2013. Applications of
1353 next-generation sequencing to phylogeography and phylogenetics. *Molecular Phylogenetics and*
1354 *Evolution* 66: 526–538.
- 1355 McCormack, J. E., Tsai, W. L. E., Faircloth, B. C. 2016. Sequence capture of ultraconserved elements
1356 from bird museum specimens. *Molecular Ecology Resources* 16: 1189–1203.

- 1357 McCormack, J. E., Rodríguez-Gómez, F., Tsai, W. L. E., Faircloth, B. C. 2017. Transforming Museum
1358 Specimens into Genomic Resources. Pp. 143–156 in M. S. Webster (editor), *The Extended*
1359 *Specimen: Emerging Frontiers in Collections-based Ornithological Research*. *Studies in Avian*
1360 *Biology* (no. 50), CRC Press, Boca Raton, FL.
- 1361 McTavish, E. J., Drew, B. T., Redelings, B., Cranston, K. A. 2017. How and why to build a unified Tree
1362 of Life. *BioEssays* 1700114.
- 1363 McTavish, E. J., Hinchliff, C. E., Allman, J. F., Brown, J., Cranston, K. A., Holder, M. T., Rees, J. A.,
1364 Smith, S. A. 2015. Phylesystem: A git-based data store for community-curated phylogenetic
1365 estimates. *Bioinformatics* 31: 2794–2800.
- 1366 Meiklejohn, K. A., Faircloth, B. C., Glenn, T. C., Kimball, R. T., Braun, E. L. 2016. Analysis of a Rapid
1367 Evolutionary Radiation Using Ultraconserved Elements: Evidence for a Bias in Some Multispecies
1368 Coalescent Methods. *Systematic Biology* 65: 612–627.
- 1369 Mendes, F. K., Hahn, M. W. 2016. Gene Tree Discordance Causes Apparent Substitution Rate Variation.
1370 *Systematic Biology* 65: 711–721.
- 1371 Meyer, M., Kircher, M., Gansauge, M. T., Li, H., Racimo, F., Mallick, S., Schraiber, J. G., Jay, F., Prüfer,
1372 K., de Filippo, C., Sudmant, P. H., Alkan, C., Fu, Q., Do, R., Rohland, N., Tandon, A., Siebauer,
1373 M., Green, R. E., Bryc, K., Briggs, A. W., Stenzel, U., Dabney, J., Shendure, J., Kitzman, J.,
1374 Hammer, M. F., Shunkov, M. V., Derevianko, A. P., Patterson, N., Andrés, A. M., Eichler, E. E.,
1375 Slatkin, M., Reich, D., Kelso, J., Pääbo, S. 2012. A high-coverage genome sequence from an archaic
1376 Denisovan individual. *Science* 338: 222–226.
- 1377 Mirarab, S., Warnow, T. 2015. ASTRAL-II: coalescent-based species tree estimation with many hundreds
1378 of taxa and thousands of genes. *Bioinformatics* 31: i44–i52.
- 1379 Mirarab, S., Bayzid, M. S., Warnow, T. 2016. Evaluating summary methods for multilocus species tree
1380 estimation in the presence of incomplete lineage sorting. *Systematic Biology* 65: 366–380.
- 1381 Misof, B., Liu, S., Meusemann, K., Peters, R. S., et al. 2014. Phylogenomics resolves the timing and

1382 pattern of insect evolution. *Science* 346: 763–767.

1383 Mitov, V., Stadler, T. 2017. POUMM: An R-package for Bayesian Inference of Phylogenetic Heritability.
1384 bioRxiv, <https://doi.org/10.1101/115089>.

1385 Mitchell, A., Mitter, C., Regier, J. C. 2000. More taxa or more characters revisited: Combining data from
1386 nuclear protein-encoding genes for phylogenetic analyses of Noctuoidea (Insecta: Lepidoptera).
1387 *Systematic Biology* 49: 202–224.

1388 Montague, M. J., Li, G., Gandolfi, B., Khan, R., Aken, B. L., Searle, S. M. J., Minx, P., Hillier, L. W.,
1389 Koboldt, D. C., Davis, B. W., Driscoll, C. A., Barr, C. S., Blackistone, K., Quilez, J., Lorente-
1390 Galdos, B., Marques-Bonet, T., Alkan, C., Thomas, G. W. C., Hahn, M. W., Menotti-Raymond, M.,
1391 O'Brien, S. J., Wilson, R. K., Lyons, L. A., Murphy, W. J., Warren, W. C. 2014. Comparative
1392 analysis of the domestic cat genome reveals genetic signatures underlying feline biology and
1393 domestication. *Proceedings of the National Academy of Sciences USA* 111: 17230–17235.

1394 Mora, C., Tittensor, D. P., Adl, S., Simpson, A. G. B., Worm, B. 2011. How Many Species Are There on
1395 Earth and in the Ocean? *PLoS Biology* 9(8):e1001127.

1396 Morlon, H. 2014. Phylogenetic approaches for studying diversification. *Ecology Letters* 17: 508–525.

1397 Morlon, H., Parsons, T.L., Plotkin, J. 2011. Reconciling molecular phylogenies with the fossil record
1398 *Proceedings of the National Academy of Sciences* 108: 16327–16332.

1399 Mulder, W. H., Crawford, F. W. 2015. On the distribution of interspecies correlation for Markov models
1400 of character evolution on Yule trees. *Journal of Theoretical Biology* 364: 275–283.

1401 Murphy, W. J., Larkin, D. M., Everts-van der Wind, A, Bourque, G., Tesler, G., Auvil, L., Beever, J. E,
1402 Chowdhary, B. P., Galibert, F., Gatzke, L., Hitte, C., Meyers, S. N., Milan, D., Ostrander, E. A.,
1403 Pape, G., Parker, H. G., Raudsepp, T., Rogatcheva, M. B., Schook, L. B., Skow, L. C., Welge, M.,
1404 Womack, J. E., O'Brien, S. J., Pevzner, P. A., Lewin, H. A. 2005. Dynamics of mammalian
1405 chromosome evolution inferred from multispecies comparative maps. *Science* 309: 613–617.

1406 Murphy, W. J., Pringle, T. H., Crider, T. A., Springer, M. S., Miller, W. 2007. Using genomic data to

1407 unravel the root of the placental mammal phylogeny. *Genome Research* 17: 413–421.

1408 Nabhan, A. R., Sarkar, I. N. 2012. The impact of taxon sampling on phylogenetic inference: a review of
1409 two decades of controversy. *Briefings in Bioinformatics* 13: 122–134.

1410 Nee, S., Mooers, A. O., Harvey, P. H. 1992. Tempo and mode of evolution revealed from molecular
1411 phylogenies. *Proceedings of the National Academy of Sciences USA* 89: 8322–8326.

1412 Ogilvie, H. A., Bouckaert, R. R., Drummond, A. J. 2017. StarBEAST2 Brings Faster Species Tree
1413 Inference and Accurate Estimates of Substitution Rates. *Molecular Biology and Evolution* 34: 2101–
1414 2114.

1415 Olave, M., Sola E., Knowles L. L. 2014. Upstream analyses create problems with DNA-based species
1416 delimitation. *Systematic Biology* 63: 263–271.

1417 Oxelman, B., Yoshikawa, N., McConaughy, B. L., Luo, J., Denton, A. L., Hall, B. D. 2004. RPB2 Gene
1418 Phylogeny in Flowering Plants, with Particular Emphasis on Asterids. *Molecular Phylogenetics and*
1419 *Evolution* 32: 462–79.

1420 Pamilo, P., Nei, M. 1988. Relationships between gene trees and species trees. *Molecular Biology and*
1421 *Evolution* 5: 568–583.

1422 Park, S. D. E., Magee, D. A., McGettigan, P. A., Teasdale, M. D., Edwards, C. J., Lohan, A. J, Murphy,
1423 A., Braud, M., Donoghue, M. T., Liu, Y., Chamberlain, A. T, Rue-Albrecht, K., Schroeder, S.,
1424 Spillane, C., Tai, S., Bradley, D. G., Sonstegard, T. S., Loftus, B. J., McHugh, D. E. 2015. Genome
1425 sequencing of the extinct Eurasian wild aurochs, *Bos primigenius*, illuminates the phylogeography
1426 and evolution of cattle. *Genome Biology* 16: 234.

1427 Patel, S., Kimball., R.T., Braun, E. L. 2013. Error in Phylogenetic Estimation for Bushes in the Tree of
1428 Life. *Journal of Phylogenetics and Evolutionary Biology* 1: 110.

1429 Pease, J. B., Haak, D. C., Hahn, M. W., Moyle, L. C. 2016. Phylogenomics Reveals Three Sources of
1430 Adaptive Variation during a Rapid Radiation. *PloS Biology* 14: e1002379.

1431 Peñalba, J. V., Smith, L. L., Tonione, M. A., Sass, C., Hykin, S. M., Skipwith, P. L., McGuire, J. A.,

- 1432 Bowie, R. C. K., Moritz, C. 2014. Sequence capture using PCR-generated probes: a cost-effective
1433 method of targeted high-throughput sequencing for non-model organisms. *Molecular Ecology*
1434 *Resources* 14: 1000–1010.
- 1435 Pennell, M. W., Harmon, L. J. 2013. An integrative view of phylogenetic comparative methods:
1436 Connections to population genetics, community ecology, and paleobiology. *Annals of the New York*
1437 *Academy of Sciences* 1289: 90–105.
- 1438 Peterson, A. T., Moyle, R. G., Nyári, Á. S., Robbins, M. B., Brumfield, R. T., Remsen, J. V. Jr. 2007. The
1439 need for proper vouchers in phylogenetic studies of birds. *Molecular Phylogenetics and Evolution*
1440 45: 1042–1044.
- 1441 Pfeil, B. E., C. L. Brubaker, L. A. Craven, Crisp, M. D. 2004. Paralogy and orthology in the Malvaceae
1442 rpb2 gene family: Investigation of gene duplication in *Hibiscus*. *Molecular Biology and Evolution*
1443 21: 1428–1437.
- 1444 Philippe, H., Brinkmann, H., Lavrov, D. V., Littlewood, D. T. J., Manuel, M., Wörheide, G., Baurain, D.
1445 2011. Resolving Difficult Phylogenetic Questions: Why More Sequences Are Not Enough. *PLoS*
1446 *Biology* 9: e1000602.
- 1447 Piel, W. H., Chan, L., Dominus, M. J., Ruan, J., Vos, R. A., Tannen, V. 2009. Treebase v. 2: A Database
1448 of Phylogenetic Knowledge. *e-Biosphere*.
- 1449 Pleijel, F., Jondelius, U., Norlinder, E., Nygren, A., Oxelman, B., Schander, C., Sundberg, P., Thollesson,
1450 M. 2008. Phylogenies without roots? A plea for the use of vouchers in molecular phylogenetic
1451 studies. *Molecular Phylogenetics and Evolution* 48: 369–371.
- 1452 Poe, S. 1998. Sensitivity of phylogeny estimation to taxonomic sampling. *Systematic Biology* 47: 18–31.
- 1453 Potts, A. J., Hedderson, T. A., Grimm, G. W. 2014. Constructing Phylogenies in the Presence Of Intra-
1454 Individual Site Polymorphisms (2ISPs) with a Focus on the Nuclear Ribosomal Cistron. *Systematic*
1455 *Biology* 63: 1–16.
- 1456 Prum, R. O., Berv, J. S., Dornburg, A., Field, D. J., Townsend, J. P. 2015. A comprehensive phylogeny of

1457 birds (Aves) using targeted next-generation DNA sequencing. *Nature* 526: 569–573.

1458 Pyron, R. A. 2015. Post-molecular systematics and the future of phylogenetics. *Trends in Ecology &*
1459 *Evolution* 30: 384–389.

1460 Quintero, I., P. Keil., W. Jetz., F. W. Crawford. 2015. Historical Biogeography Using Species
1461 Geographical Ranges. *Systematic Biology* 64: 1059–1073.

1462 Ramadugu, C., Pfeil, B. E., Manjunath, K. L., Lee, R. F., Maureira-Butler, I. J., Roose, M. L. 2013.
1463 Coalescence simulation testing of hybridization versus lineage sorting in *Citrus* (Rutaceae) using six
1464 nuclear genes. *PLoS One* 8: e68410.

1465 Rannala, B., Yang, Z. H. 2003. Bayes estimations of species divergence times and ancestral population
1466 sizes using DNA sequences from multiple loci. *Genetics* 164: 1645–1656.

1467 Rannala, B., Yang Z. H. 2008. Phylogenetic inference using whole genomes. *Annual Review of*
1468 *Genomics and Human Genetics* 9: 217–231.

1469 Ranwez, V., Delsuc, F. D. R., Ranwez, S., Belkhir, K., Tilak, M-K., Douzery, E. J. 2007. OrthoMaM: A
1470 database of orthologous genomic markers for placental mammal phylogenetics. *BMC Evolutionary*
1471 *Biology* 7: 241–12.

1472 Rasmussen, M. D., Kellis, M. 2012. Unified modeling of gene duplication, loss, and coalescence using a
1473 locus tree. *Genome Research* 22: 755–765.

1474 Reddy, S., Kimball, R. T., Pandey, A., Hosner, P. A., Braun, M. J., Hackett, S. J., Han, K-L., Harshman,
1475 J., Huddleston, C. J., Kingston, S., Marks, B. D., Miglia, K. J., Moore, W. S., Sheldon, F. H., Witt, C.
1476 C., Yuri, T., Braun, E. L. 2017. Why Do Phylogenomic Data Sets Yield Conflicting Trees? Data
1477 Type Influences the Avian Tree of Life more than Taxon Sampling. *Systematic Biology* 66: 857–
1478 879.

1479 Redon, R., Ishikawa, S., Fitch, K. R., Feuk, L., Perry, G. H., Andrews, T. D., Fiegler, H., Shaperro, M. H.,
1480 Carson, A. R., Chen, W., Cho, E. K., Dallaire, S., Freeman, J. L., González, J. R., Gratacòs, M. N.,
1481 Huang, J., Kalaitzopoulos, D., Komura, D., MacDonald, J. R., Marshall, C. R., Mei, R.,

1482 Montgomery, L., Nishimura, K., Okamura, K., Shen, F., Somerville, M. J., Tchinda, J., Valsesia, A.,
1483 Woodwark, C., Yang, F., Zhang, J., Zerjal, T., Zhang, J., Armengol, L., Conrad, D. F., Estivill, X.,
1484 Tyler-Smith, C., Carter, N. P., Aburatani, H., Lee, C., Jones, K. W., Scherer, S. W., Hurles, M. E.
1485 2006. Global variation in copy number in the human genome. *Nature* 444: 444–454.

1486 Reid, N. M., Hird, S. M., Brown, J. M., Pelletier, T. A., McVay, J. D., Satler, J. D., Carstens, B. C. 2014.
1487 Poor fit to the multispecies coalescent is widely detectable in empirical data. *Systematic Biology* 63:
1488 322–333.

1489 Rogozin, I. B., Thomson, K., Csürös, M., Carmel, L., Koonin, E. V. 2008. Homoplasy in genome-wide
1490 analysis of rare amino acid replacements: the molecular-evolutionary basis for Vavilov's law of
1491 homologous series. *Biology Direct* 2008 3: 7.

1492 Rogozin, I. B., Wolf, Y. I., Babenko, B. N., Koonin, E. V. 2006. Dollo parsimony and the reconstruction
1493 of genome evolution. In: Albert VA ed. *Parsimony, Phylogeny, and Genomics*. Oxford University
1494 Press, p.p. 190–200.

1495 Rokas, A. 2005. More Genes or More Taxa? The Relative Contribution of Gene Number and Taxon
1496 Number to Phylogenetic Accuracy. *Molecular Biology and Evolution* 22: 1337–1344.

1497 Rokas, A., Holland, P. W. H. 2000. Rare genomic changes as a tool for phylogenetics. *Trends in Ecology*
1498 *and Evolution* 15: 454–459.

1499 Rokas, A., Williams, B. L., King, N., Carroll, S. B. 2003. Genome-scale approaches to resolving
1500 incongruence in molecular phylogenies. *Nature* 425: 798–804.

1501 Romiguier, J., Cameron, S. A., Woodard, S. H., Fischman, B. J., Keller, L., Praz, C. J. 2016.
1502 Phylogenomics Controlling for Base Compositional Bias Reveals a Single Origin of Eusociality in
1503 Corbiculate Bees. *Molecular Biology and Evolution* 33: 670–678.

1504 Romiguier, J., Roux, C. 2017. Analytical Biases Associated with GC-Content in Molecular Evolution.
1505 *Frontiers in Genetics* 8: 16.

1506 Roncal, J., Guyot, R., Hamon, P., Crouzillat, D., Rigoreau, M., Konan, O. N., Rakotomalala, J. J., Nowak,
1507 M. D., Davis, A. P., de Kochko, A. 2016. Active transposable elements recover species boundaries
1508 and geographic structure in Madagascan coffee species. *Molecular Genetics and Genomics* 291: 155–
1509 168.

1510 Rosenberg, N. A., Nordborg, M. 2002. Genealogical trees, coalescent theory and the analysis of genetic
1511 polymorphisms. *Nature Reviews Genetics* 3: 380–390.

1512 Rosindell, J., Cornell, S. J., Hubbell S. P., Etienne, R. S. 2010. Protracted speciation revitalizes the
1513 neutral theory of biodiversity. *Ecology Letters* 13: 716–727.

1514 Rosindell, J., Harmon, L. J. 2012. OneZoom: A Fractal Explorer for the Tree of Life. *PLoS Biology*
1515 10(10): e1001406.

1516 Rosindell, J., Harmon, L. J., Etienne, R. S. 2015. Unifying ecology and macroevolution with individual-
1517 based theory. *Ecology Letters* 18: 472–482.

1518 Ruane, S., Austin, C. C. 2017. Phylogenomics using formalin-fixed and 100+ year-old intractable natural
1519 history specimens. *Molecular Ecology Resources* 17: 1003–1008.

1520 Sagitov, S., Bartoszek, K. 2012. Interspecies correlation for neutrally evolving traits. *Journal of*
1521 *Theoretical Biology* 309: 11–19.

1522 Sanderson, M. J., Donoghue, M. J., Piel, W. H., Eriksson, T. 1994. TreeBASE: a prototype database of
1523 phylogenetic analyses and an interactive tool for browsing the phylogeny of life. *American Journal of*
1524 *Botany* 81: 183.

1525 Sayyari, E., Mirarab, S. 2016. Fast Coalescent-Based Computation of Local Branch Support from Quartet
1526 Frequencies. *Molecular Biology and Evolution* 33: 1654–1668.

1527 Scally, A., Dutheil, J. Y., Hillier, L. W., Jordan, G. E., Goodhead, I., Herrero, J., Hobolth, A.,
1528 Lappalainen, T., Mailund, T., Marques-Bonet, T., McCarthy, S., Montgomery, S. H., Schwalie, P. C.,
1529 Tang, Y. A., Ward, M. C., Xue, Y., Yngvadottir, B., Alkan, C., Andersen, L. N., Ayub, Q., Ball, E.

1530 V., Beal, K., Bradley, B. J., Chen, Y., Clee, C. M., Fitzgerald, S., Graves, T. A., Gu, Y., Heath, P.,
1531 Heger, A., Karakoc, E., Kolb-Kokocinski, A., Laird, G. K., Lunter, G., Meader, S., Mort, M.,
1532 Mullikin, J. C., Munch, K., O'Connor, T. D., Phillips, A. D., Prado-Martinez, J., Rogers, A. S.,
1533 Sajjadian, S., Schmidt, D., Shaw, K., Simpson, J. T., Stenson, P. D., Turner, D. J., Vigilant, L.,
1534 Vilella, A. J., Whitener, W., Zhu, B., Cooper, D. N., de Jong, P., Dermitzakis, E. T., Eichler, E. E.,
1535 Flicek, P., Goldman, N., Mundy, N. I., Ning, Z., Odom, D. T., Ponting, C. P., Quail, M. A., Ryder, O.
1536 A., Searle, S. M., Warren, W. C., Wilson, R. K., Schierup, M. H., Rogers, J., Tyler-Smith, C.,
1537 Durbin, R. 2012. Insights into hominid evolution from the gorilla genome sequence. *Nature* 483:
1538 169–175.

1539 Schrempf, D., Minh, B. Q., De Maio, N., von Haeseler, A., Kosiol, C. 2016. Reversible polymorphism-
1540 aware phylogenetic models and their application to tree inference. *Journal of Theoretical Biology*
1541 407: 362–370.

1542 Schwartz, R., Schäffer, A. A. 2017. The evolution of tumour phylogenetics: Principles and practice.
1543 *Nature Reviews Genetics* 18: 213–229.

1544 Shen, X-X., Hittinger, C. T., Rokas, A. 2017. Contentious relationships in phylogenomic studies can be
1545 driven by a handful of genes. *Nature Ecology and Evolution* 1: 0126.

1546 Shi, C.-M., Yang, Z. 2018. Coalescent-Based Analyses of Genomic Sequence Data Provide a Robust
1547 Resolution of Phylogenetic Relationships among Major Groups of Gibbons. *Molecular Biology and*
1548 *Evolution* 35: 159–179.

1549 Siepel, A. 2009. Phylogenomics of primates and their ancestral populations. *Genome Research* 19: 1929–
1550 1941.

1551 Silvestro, D., Schnitzler, J., Liow, L. H., Antonelli, A., Salamin, N. 2014. Bayesian estimation of
1552 speciation and extinction from incomplete fossil occurrence data. *Systematic Biology* 63: 349–367.

1553 Simion, P., Philippe, H., Baurain, D., Muriel, J., Richter, D. J., Di Franco, A., Roure, B., Satoh, N.,
1554 Quéinnec, E., Ereskovsky, A. 2017. A Large and Consistent Phylogenomic Dataset Supports

1555 Sponges as the Sister Group to All Other Animals. *Current Biology* 27: 958–967.

1556 Sjödin, P., Jakobsson, M. 2012. Population genetic nature of copy number variation. *Population Genetic*
1557 *Nature of Copy Number Variation*. In: Feuk L. (eds) *Genomic Structural Variants. Methods in*
1558 *Molecular Biology (Methods and Protocols)*, vol 838. Springer, New York, NY.

1559 Smith, S. A., Moore, M. J., Brown, J. W., Yang, Y. 2015. Analysis of phylogenomic datasets reveals
1560 conflict, concordance, and gene duplications with examples from animals and plants. *BMC*
1561 *Evolutionary Biology* 15: 150.

1562 Solis-Lemus, C., Ané, C. 2016. Inferring Phylogenetic Networks with Maximum Pseudolikelihood under
1563 Incomplete Lineage Sorting. *PLoS Genet* 12: e1005896.

1564 Solis-Lemus, C., Knowles, L. L., Ané, C. 2015. Bayesian species delimitation combining multiple genes
1565 and traits in a unified framework. *Evolution* 69: 492–507.

1566 Solis-Lemus, C., Bastide, P., Ané, C. 2017. PhyloNetworks: A Package for Phylogenetic Networks.
1567 *Molecular Biology and Evolution* 34: 3292–3298.

1568 Song, S., Liu, L., Edwards, S. V., Wu, S. 2012. Resolving conflict in eutherian mammal phylogeny using
1569 phylogenomics and the multispecies coalescent model. *Proceedings of the National Academy of*
1570 *Sciences USA* 112: E6079–E6079.

1571 Sousa, F., Bertrand, Y. J. K., Doyle, J. J., Oxelman, B., Pfeil, B. E. 2017. Using genomic location and
1572 coalescent simulation to investigate gene tree discordance in *Medicago* L. *Systematic Biology* 66:
1573 934–949.

1574 Springer, M. S., Gatesy, J. 2016. The gene tree delusion. *Molecular Phylogenetics and Evolution* 94: 1–
1575 33.

1576 Staats, M., Erkens, R. H. J., van de Vossenberg, B., Wieringa, J. J., Kraaijeveld, K., Stielow, B., Geml,
1577 J., Richardson, J. E., Bakker, F. T. 2013. Genomic treasure troves: complete genome sequencing of
1578 herbarium and insect museum specimens. *PLoS ONE* 8: e69189.

1579 Stadler, T., 2009. On incomplete sampling under birth-death models and connections to the sampling-

1580 based coalescent. *Journal of Theoretical Biology* 261: 58–68.

1581 Stadler, T. 2013. Recovering speciation and extinction dynamics based on phylogenies. *Journal of*
1582 *Evolutionary Biology* 26: 1203–1219.

1583 Stadler, T., Steel, M., 2012. Distribution of branch lengths and phylogenetic diversity under
1584 homogeneous speciation models. *Journal of Theoretical Biology* 297: 33–40.

1585 Stephens, M., Smith, N.J., Donnelly, P. 2001. A new statistical method for haplotype reconstruction from
1586 population data. *The American Journal of Human Genetics* 68: 978–989.

1587 Struck, T. H. 2013. The Impact of Paralogy on Phylogenomic Studies – A Case Study on Annelid
1588 Relationships. *PLoS ONE* 8: e62892.

1589 Suarez, A. V., Tutsui, N. D. 2004. The value of museum collections for research and society. *BioScience*
1590 54: 66–74.

1591 Suh, A., Paus, M., Kiefmann, M., Churakov, G., Franke, F. A., Brosius, J., Kriegs, J. O., Schmitz, J.
1592 2011. Mesozoic retroposons reveal parrots as the closest living relatives of passerine birds. *Nature*
1593 *Communications* 2: 443.

1594 Suh, A., Smeds, L. A., Ellegren, H. 2015. The Dynamics of Incomplete Lineage Sorting across the
1595 Ancient Adaptive Radiation of Neoavian Birds. *PLoS Biology* 13: e1002224–18.

1596 Sukumaran, J., Knowles, L. L., 2017. Multispecies coalescent delimits structure, not species. *Proceedings*
1597 *of the National Academy of Sciences USA* 114: 1607–1612.

1598 Talavera, G., Castresana, J. 2007. Improvement of phylogenies after removing divergent and ambiguously
1599 aligned blocks from protein sequence alignments. *Systematic Biology* 56: 564–577.

1600 Tang, J., Moret, B. M. E., Cui, L., dePamphilis, C. W. 2004. Phylogenetic reconstruction from arbitrary
1601 gene-order data. in *Proceedings. Fourth IEEE Symposium on Bioinformatics and Bioengineering*
1602 592–599.

1603 Tonini, J. F. R., Beard, K. H., Ferreira, R. B., Jetz, W., Pyron, R. A. 2016. Fully-sampled phylogenies of
1604 squamates reveal evolutionary patterns in threat status. *Biological Conservation* 204: 23–31.

1605 Toprak, Z., Pfeil, B.E., Jones, G., Marcussen, T., Ertekin, A.S., Oxelman, B. 2016. Species Delimitation
1606 Without Prior knowledge: DISSECT Reveals Extensive Cryptic Speciation in the *Silene aegyptiaca*
1607 complex (Caryophyllaceae). *Molecular Phylogenetics and Evolution* 102: 1–8.

1608 Turney, S., Cameron, E. R., Cloutier, C. A, Buddle, C. M. 2015. Non-repeatable science: assessing the
1609 frequency of voucher specimen deposition reveals that most arthropod research cannot be verified.
1610 *PeerJ* 3: e1168–16.

1611 Vision, T. 2010. Open data and the social contract of scientific publishing. *BioScience* 60: 330.

1612 Wen, D., Nakhleh L. 2018. Co-estimating Reticulate Phylogenies and Gene Trees from Multi-locus
1613 Sequence Data. *Systematic Biology* in press.

1614 Wen, D., Yu Y., Hahn M. W., Nakhleh L. 2016. Reticulate evolutionary history and extensive
1615 introgression in mosquito species revealed by phylogenetic network analysis. *Molecular Ecology* 25:
1616 2361–2372.

1617 Wesche, P. L., Gaffney, D. J., Keightley, P. D. 2004. DNA Sequence Error Rates in Genbank Records
1618 Estimated Using the Mouse Genome as a Reference. *DNA Sequence* 15(5–6): 362–64.

1619 Wiedenhoeft, J., Brugel, E., Schliep, A. 2016. Fast Bayesian Inference of Copy Number Variants using
1620 Hidden Markov Models with Wavelet Compression. *PLoS Computational Biology* 12(5): e1004871.

1621 Will, K. P., Mishler, B. D., Wheeler, Q. D. 2005. The perils of DNA Barcoding and the need for
1622 integrative taxonomy. *Systematic Biology* 54: 844–851.

1623 Wilson, G., Aruliah, D. A., Brown, C. T., Chue-Hong N. P., Davis, M., Guy, R. T., Haddock, S. H. D,
1624 Huff, K. D., Mitchell, I. M., Plumbley, M. D, Waugh, B., White, E. P., Wilson, P. 2014. Best
1625 Practices for Scientific Computing. *PLoS Biology* 12(1): e1001745.

1626 Wu, Y. C., Rasmussen, M. D., Bansal, M. S., Kellis, M. 2013. TreeFix: statistically informed gene tree
1627 error correction using species trees. *Systematic Biology* 62: 110–120.

1628 Wu, Y. C., Rasmussen, M. D., Bansal, M. S., Kellis M. 2014. Most parsimonious reconciliation in the
1629 presence of gene duplication, loss, and deep coalescence using labeled coalescent trees. *Genome*

1630 Research 24: 475–486.

1631 Xi, Z., Ruhfel, B. R., Schaefer, H., Amorim, A. M., Sugumaran, M., Wurdack, K. J., Endress, P. K.,
1632 Matthews, M. L., Stevens, P. F., Mathews, S., Davis, C. C. 2012. Phylogenomics and a posteriori
1633 data partitioning resolve the Cretaceous angiosperm radiation Malpighiales. *Proceedings of the*
1634 *National Academy of Sciences USA* 109: 17519–17524.

1635 Xi, Z., Liu, L., Davis, C. C. 2015. Genes with minimal phylogenetic information are problematic for
1636 coalescent analyses when gene tree estimation is biased. *Molecular Phylogenetics and Evolution* 92:
1637 63–71.

1638 Xu, B., Yang, Z. 2016. Challenges in Species Tree Estimation Under the Multispecies Coalescent Model.
1639 *Genetics* 204: 1353–1368.

1640 Yang Z, Rannala B, 2014. Unguided species delimitation using DNA sequence data from multiple loci.
1641 *Molecular Biology and Evolution* 31: 3125–3135.

1642 Yu, Y., Dong J., Liu K. J., Nakhleh, L. 2014. Maximum likelihood inference of reticulate evolutionary
1643 histories. *Proceedings of the National Academy of Sciences USA* 111: 16448–16453.

1644 Zanne, A. E., Tank, D. C., Cornwell, W. K., Eastman, J. M., Smith, S. A, FitzJohn, R. G., McGlenn, D. J.,
1645 O’Meara, B. C., Moles, A. T., Reich, P. B., Royer, D. L., Soltis, D. E., Stevens, P. F., Westoby, M.,
1646 Wright, I. J., Aarssen, L., Bertin, R. I. Calaminus, A., Govaerts, R., Hemmings, F., Leishman, M. R.,
1647 Oleksyn, J., Soltis, P. S., Swenson, N. G., Warman, L., Beaulieu, J. M. 2014. Three keys to the
1648 radiation of angiosperms into freezing environments. *Nature* 506: 89–92.

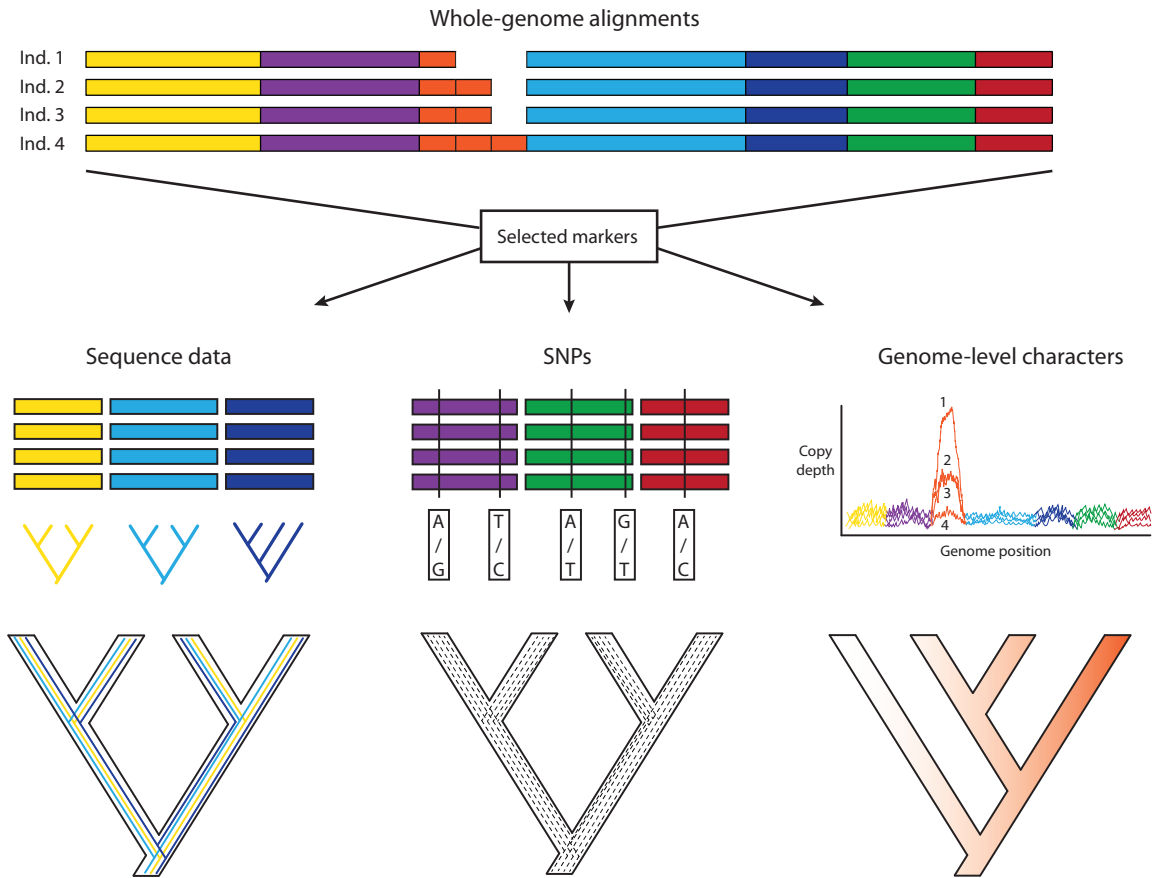
1649 Zhang, C., Ogilvie, H. A., Drummond, A. J., Stadler, T. 2018. Bayesian Inference of Species Networks
1650 from Multilocus Sequence Data. *Molecular Biology and Evolution*, in press.

1651

1652

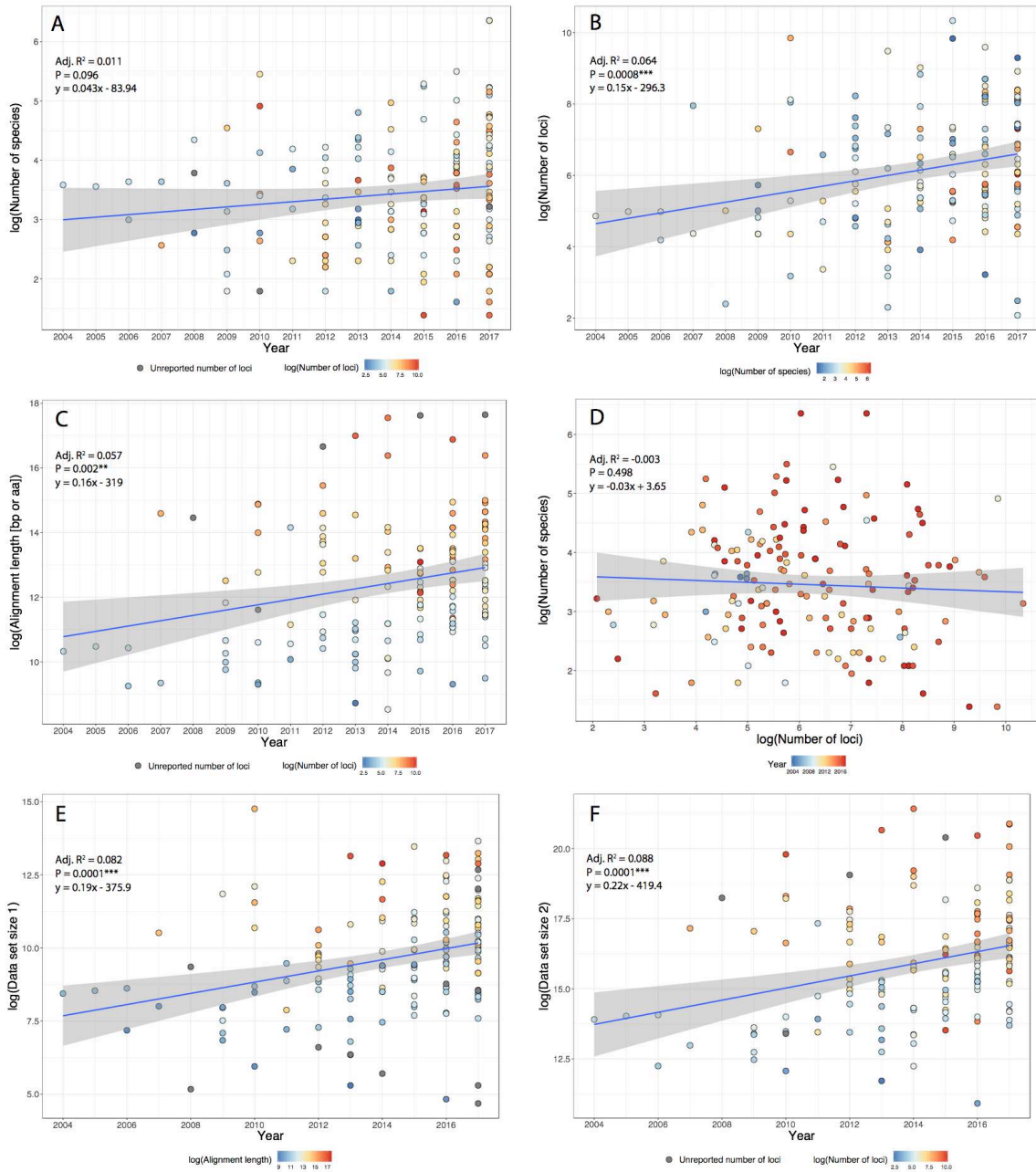
1653 **Figures**

1654 Figure 1. *A posteriori* marker selection from whole genome alignments for phylogenomics and
 1655 phylogeography. Whole genome analysis (top) permits researchers to choose different markers for
 1656 specific purposes. By contrast, subsampling methods such as Rad-seq or hybrid capture, which dominate
 1657 phylogenomics today, usually yield a specific set of markers that the researcher has chosen *a priori*. The
 1658 generation of WGA thus greatly increases the use of genomic data in biological research, beyond the
 1659 initial goals of the researcher producing those data.
 1660



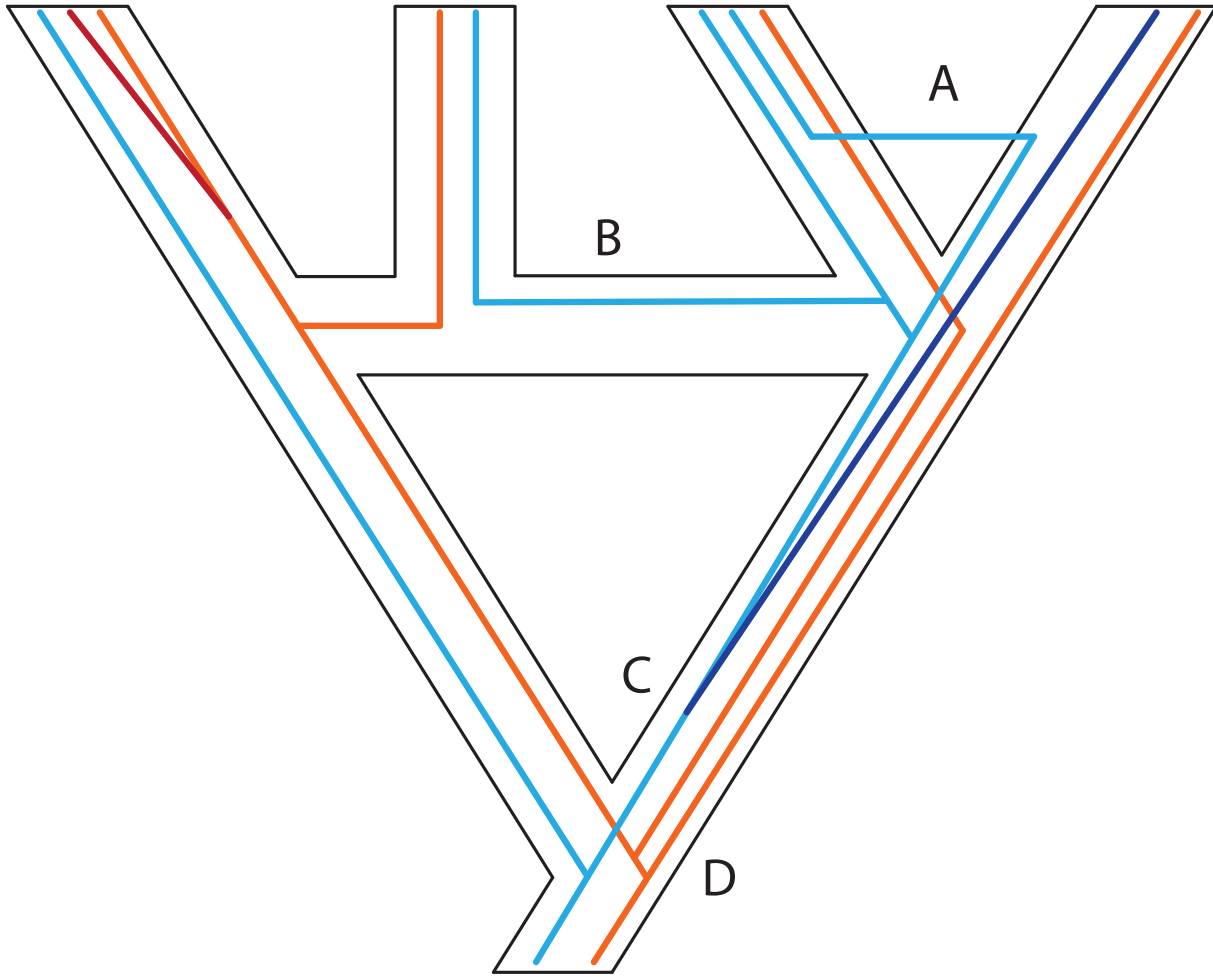
1661
 1662

1663 Figure 2. Trends in phylogenomic data sets since the emergence of HTS. Based on a sample of 164
 1664 phylogenomic papers published since 2004 (see Supplementary Table S1), we observed no increase in the
 1665 number of species per data set over time (A). On the other hand, there is a significant increase in the
 1666 number of loci (B), total alignment length (C), and total data set size, as measured by the product of
 1667 species times locus number (Data set size 1, E) and species times total alignment length (Data set size 2,
 1668 F). Moreover, the advent of HTS does not support the notion of a tradeoff between the number of species
 1669 and the number of loci in phylogenomic studies.
 1670



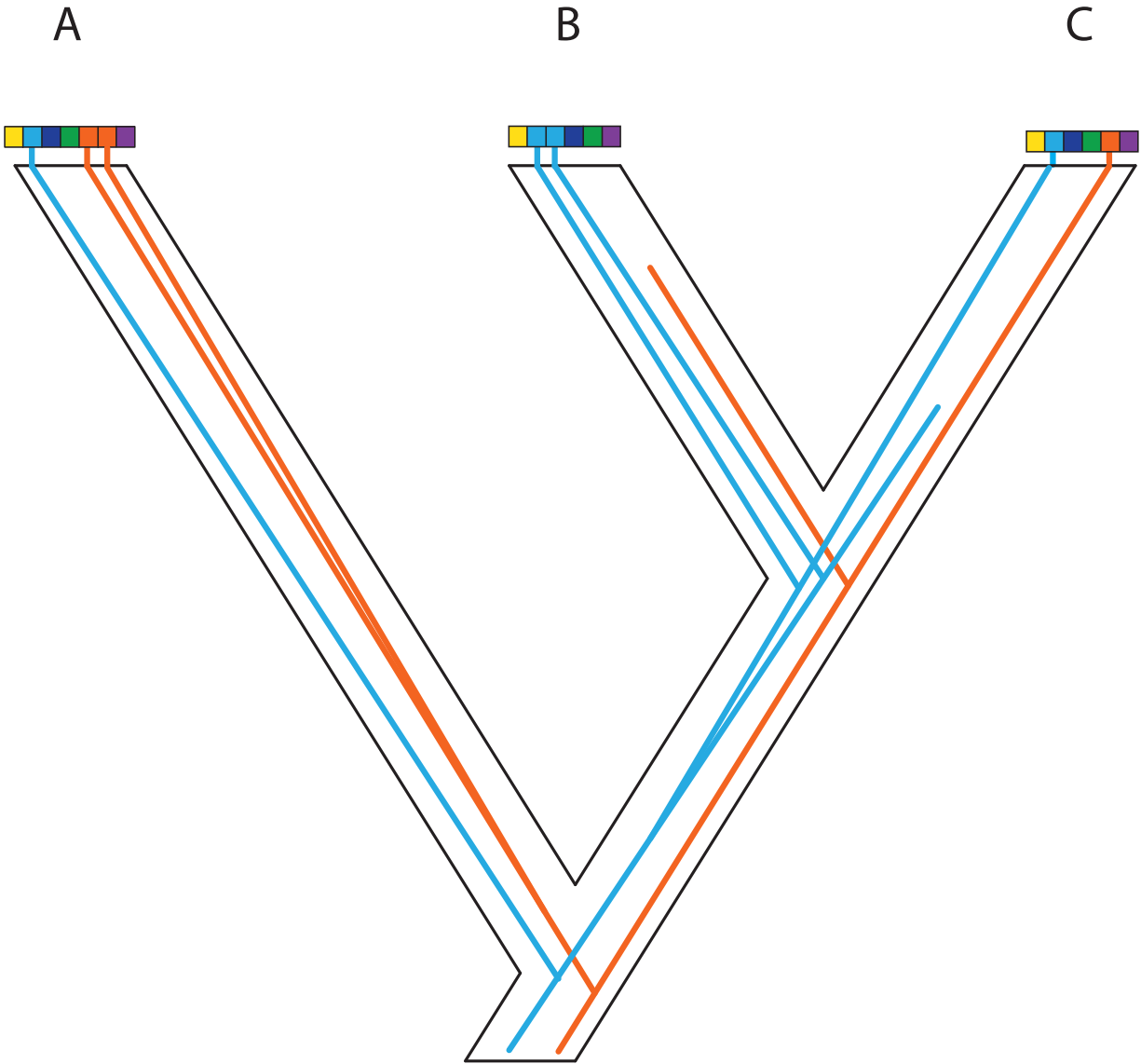
1671

1672 Figure 3. Some examples of violations of the multispecies coalescent. In event A, there is gene flow; in
1673 event B there is homoploid hybridization; in event C, there is a gene duplication; and in event D,
1674 incomplete lineage sorting. All of these processes contribute to gene tree heterogeneity but fall outside the
1675 standard multispecies coalescent model. Importantly, all of these processes also yield strictly dichotomous
1676 gene trees, whereas recombination (not illustrated here) does not. This implies that tree building without
1677 considering the multispecies coalescent could, in this case, lead to erroneous estimation of tree topology
1678 and divergence times.
1679
1680



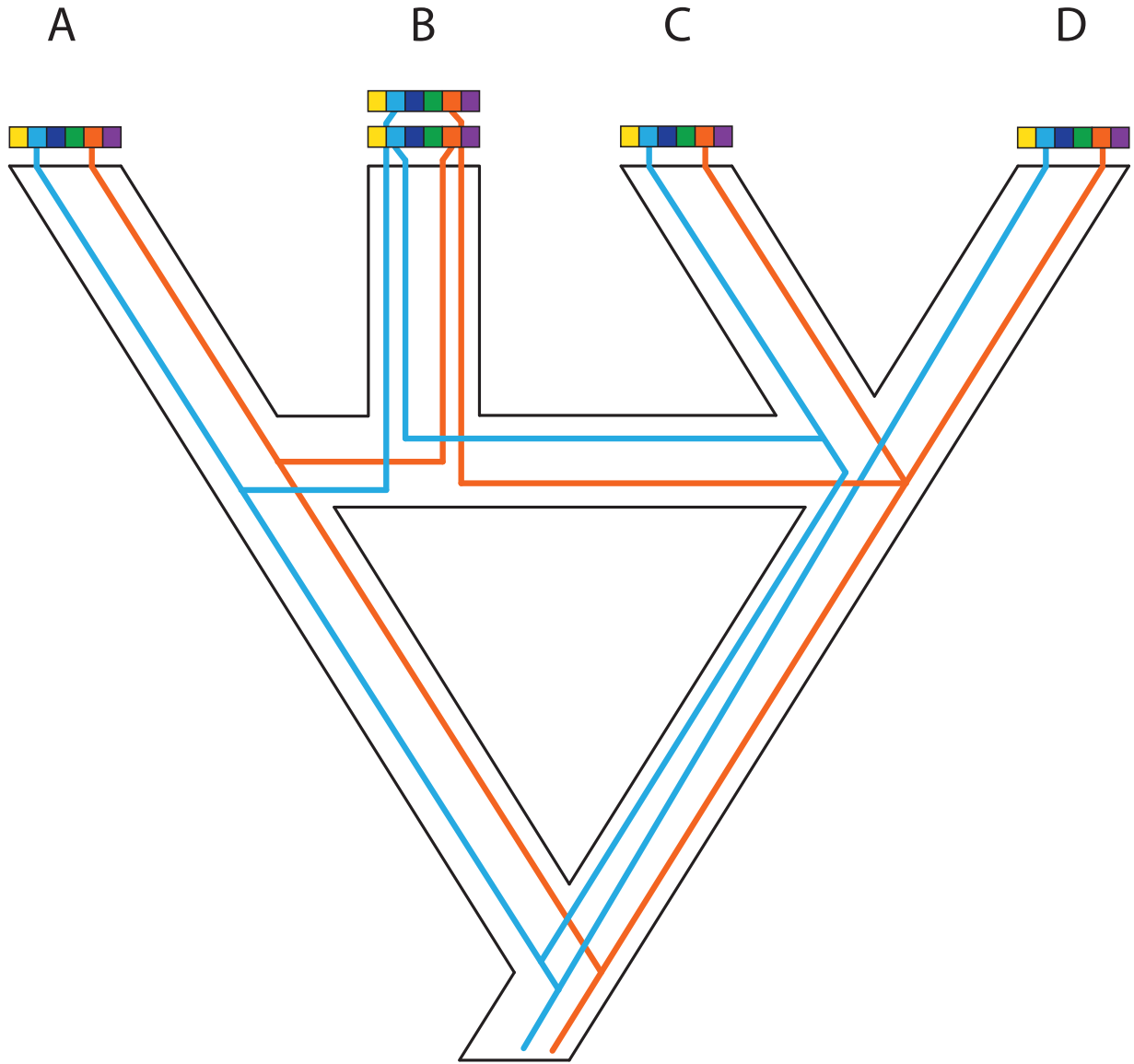
1681

1682 Figure 4. Gene duplication and loss creates patterns that can mimic incomplete lineage sorting and other
1683 processes. Genes and genomes of three species A, B, and C. Multi-colored bars show (parts of) their
1684 genomes with a number of loci indicated in different colors. The orange gene is duplicated in species A
1685 and it was lost in species B. The blue gene was duplicated before the divergence between B and C.
1686 However, both copies are maintained in species B and only one copy persists in species C. The
1687 duplication and loss history of these two genes may cause serious issues for phylogenetic reconstruction
1688 because no specific pattern can be expected between them.
1689



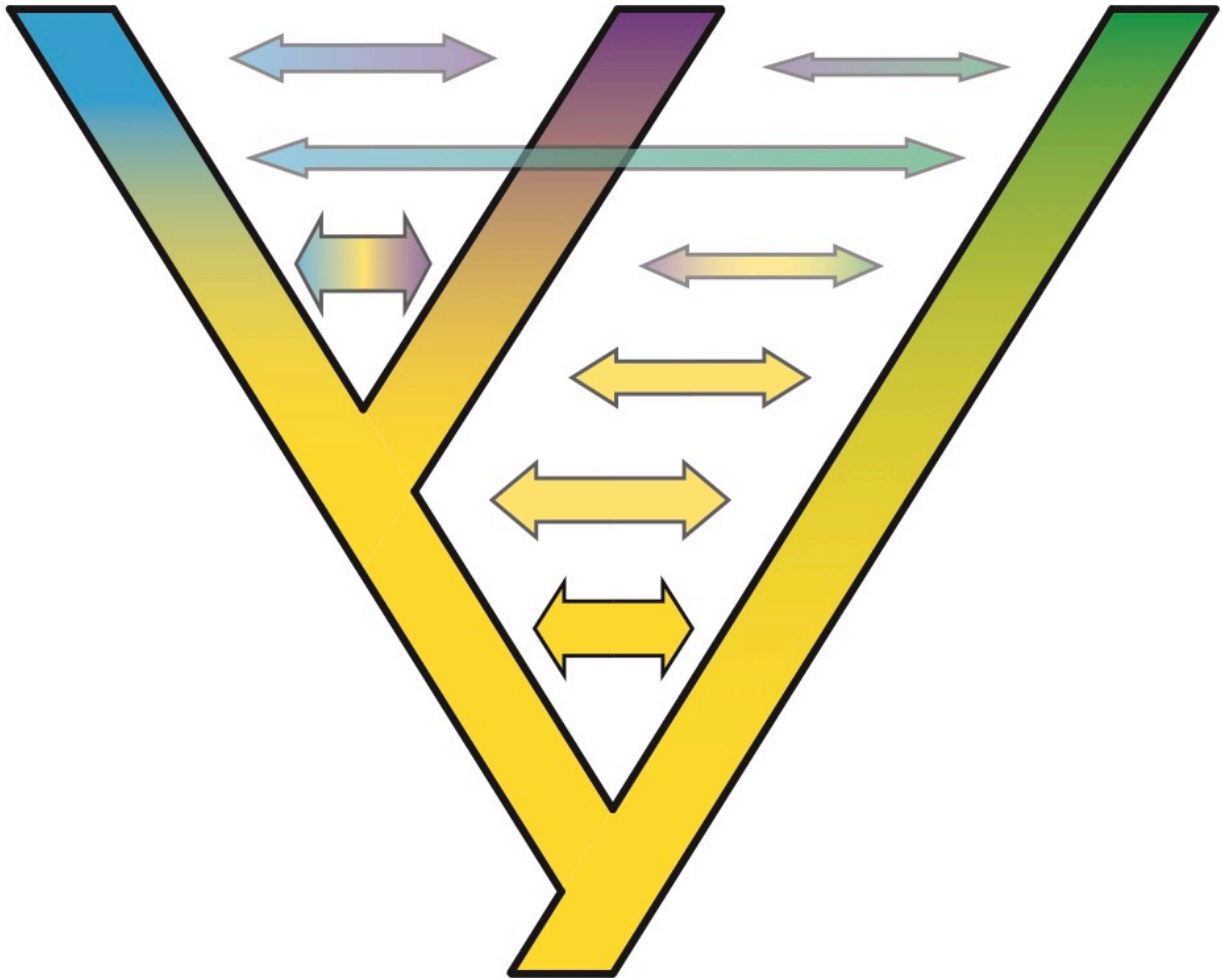
1690

1691 Figure 5. Complex patterns of gene lineages with polyploidization and interspecific gene flow. Genes and
1692 genomes of four species A, B, C and D. Multi-colored bars show (parts of) genomes with a number of
1693 loci indicated in different colors. Two gene trees, one orange and one blue, evolve within the species
1694 network. Species B is an allopolyploid containing two genomes.
1695
1696



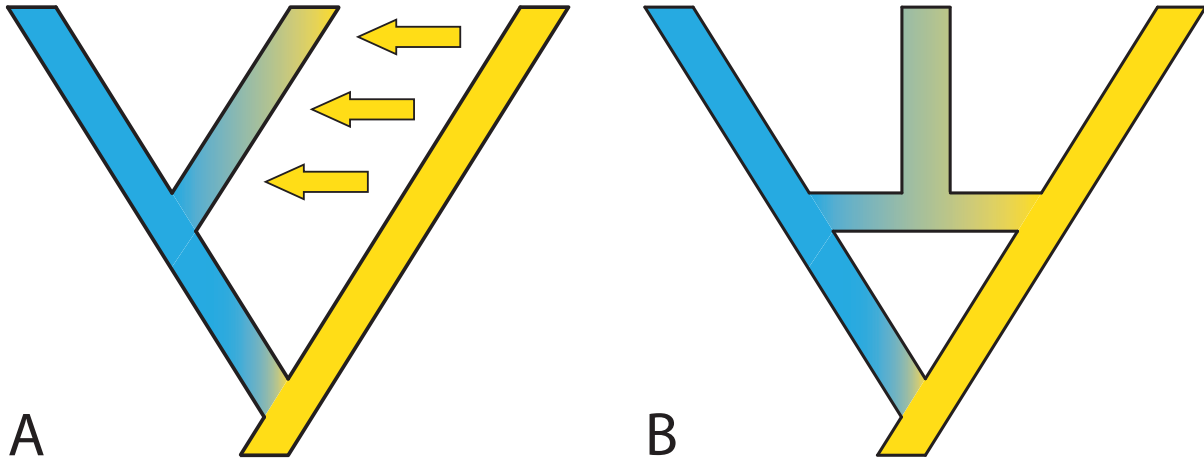
1697

1698 Figure 6. Gradual speciation, or isolation-with migration. After starting to split, gene flow between
1699 species decreases gradually. Such a gradual decrease in the extent of gene flow between species might
1700 present an especially useful extension of the standard multispecies coalescent model.
1701
1702



1703

1704 Figure 7. Two possible species phylogenies producing similar observations at present time. On the left
1705 (A), there is a species tree with gene flow. On the right (B), there is a species network with homoploid
1706 hybridization. Distinguishing two such scenarios usually requires simulations and comparison of
1707 observed and expected summary statistics.
1708
1709



1710