Article- Discoveries

Title: Using all gene families vastly expands data available for phylogenomic inference in primates

Authors: Megan L. Smith[1], Dan Vanderpool[1], and Matthew W. Hahn[1]

[1]Department of Biology and Department of Computer Science, Indiana University, Bloomington, Indiana, USA

Corresponding Author: mls16@indiana.edu

Abstract

Traditionally, single-copy orthologs have been the gold standard in phylogenomics. Most phylogenomic studies identify putative single-copy orthologs by using clustering approaches and retaining families with a single sequence from each species. However, this approach can severely limit the amount of data available by excluding larger families. Recent methodological advances have suggested several ways to include data from larger families. For instance, tree-based decomposition methods facilitate the extraction of orthologs from large families. Additionally, several popular methods for species tree inference appear to be robust to the inclusion of paralogs, and hence could use all of the data from larger families. Here, we explore the effects of using all families for phylogenetic inference using genomes from 26 primate species. We compare single-copy families, orthologs extracted using tree-based decomposition approaches, and all families with all data (i.e., including orthologs and paralogs). We explore several species tree inference methods, finding that across all nodes of the tree except one, identical trees are returned across nearly all datasets and methods. As in previous studies, the relationships among Platyrrhini remain contentious; however, the tree inference methods matter more than the dataset used. We also assess the effects of each dataset on branch length estimates, measures of phylogenetic uncertainty and concordance, and in detecting introgression. Our results demonstrate that using data from larger gene families drastically increases the number of genes available for phylogenetic inference and leads to consistent estimates of branch lengths, nodal certainty and concordance, and inferences of introgression.

Keywords: phylogenetics, orthologs, paralogs, concatenation, coalescence

Introduction

Advances in sequencing technology have led to the availability of more genomic data than ever before, and the promise of phylogenomics is the application of this data to infer species relationships (Scornavacca et al. 2020). Essential to the application of genomic data to phylogenetic inference is the identification of homologous genes, or genes that share a common ancestor. Homologous genes may share a common ancestor due to speciation (orthologs) or duplication (paralogs). Since the terms ortholog and paralog were coined (Fitch 1970), orthologs have been considered the appropriate genes for phylogenetic inference because they are related only through speciation events, and therefore are thought to best reflect species relationships. Thus, identifying orthologs is a central part of most phylogenomic pipelines.

Nearly all pipelines for extracting putative orthologs from genomic data begin with a clustering step (Figure 1). Clustering approaches aim to identify sets of homologous genes. While the details vary, these approaches generally begin with pairwise comparisons of all sequences across genomes, identify putative pairwise homologs, and then use clustering approaches to attempt to group many sets of these genes together (reviewed in Altenhoff et al. 2019). The end-products of graph-based clustering approaches are clusters of orthologs and paralogs—i.e., gene families. Since most phylogenetic methods were designed for use with orthologs (and a single sequence per taxon), these groups must be further processed for downstream phylogenetic inference.

Three primary approaches have been used to process families for downstream inference (Figure 1; Step 1). The first and most common is to extract clusters with only a single copy in each species—these represent putative single-copy orthologs. Using single-copy families is generally seen as a conservative approach in phylogenomics, as these genes are likely to be orthologs; this choice also limits the amount of further downstream processing needed. However, the number of genes that are single-copy in all sampled species decreases sharply as additional species are included in the analyses (Emms and Kelly 2018), limiting the usefulness of this approach in many phylogenetic contexts.

In lieu of relying only on single-copy clusters, tree-based decomposition approaches (Willson et al. 2021) for orthology detection can be applied to extract orthologous genes from clusters that may have more than one copy in one or more species (Figure 1; Step 2). Tree-based decomposition approaches attempt to infer whether nodes in gene trees represent duplication or

speciation events, followed by the extraction of orthologs based on these node labels (reviewed in Altenhoff et al. 2019). Early tree-based approaches relied on gene tree reconciliation to a known species tree (e.g., Goodman et al. 1979), limiting their utility in cases where the species tree is unknown or uncertain. However, recent approaches have relaxed these requirements. For example, the method LOFT relies on a species overlap approach to identify duplication nodes in gene trees (van der Heijden et al. 2007). Similarly, the software package Agalma (Dunn et al. 2013), the methods of Yang and Smith (2014), and DISCO (Willson et al. 2021) extract subtrees without duplicates to generate sets of orthologs. While the exact implementations vary, in general tree-based decomposition approaches aim to extract orthologous genes from families of any size. Tree-based approaches allow researchers to vastly increase the number of genes retained compared to using only the single-copy clusters. However, these approaches require that users construct gene trees and perform ortholog extraction for each gene family, and they are therefore more computationally intensive than relying on single-copy clusters alone (Figure 1).

Finally, families containing both orthologs and paralogs could be used for phylogenetic inference. Although orthologs have traditionally been considered the appropriate genes for phylogenetics, methods for estimating phylogenies from data including paralogs were introduced more than forty years ago (Goodman et al. 1979; reviewed in Smith and Hahn 2021a). Recently, several popular methods for species tree estimation have been shown to be robust to the presence of paralogs (Hill et al. 2020; Legried et al. 2020; Markin and Eulenstein 2020; Yan et al. 2021). Of particular interest, quartet-based methods, such as ASTRAL (Zhang et al. 2018), should be robust to the inclusion of paralogs because the most common quartet is still expected to match the species tree even in the presence of gene duplication and loss. Given that all ortholog extraction methods may erroneously lead to the inclusion of paralogs, using methods that are robust to their inclusion is likely a good strategy no matter the method employed to process the output of clustering methods.

Though there have been several empirical comparisons between ortholog-detection methods (e.g., Fernández et al. 2018; Kallal et al. 2018; Altenhoff et al. 2019), along with several simulation-based (e.g., Legried et al. 2020; Zhang et al. 2020; Yan et al. 2021) and empirical (e.g., Yan et al. 2021) studies evaluating the effects of paralog inclusion on phylogenetic inference, several questions remain. First, a comparison of inference on single-copy clusters to tree-based decomposition methods and methods that use all of the data (i.e., use orthologs and

98 paralogs for phylogenetic inference) would shed light on the advantages of the three approaches.
99 In addition, the joint effects of dataset, missing data requirements, and gene and species tree
100 inference method on species tree topology will provide information on the importance of each.
101 Finally, questions remain about the effects of the dataset used on branch length estimates,
102 measures of nodal support, and tests for introgression.
103 To address these questions, we reanalyze a recently published phylogenomic dataset that
104 includes 26 species of primates and 3 outgroups (Vanderpool et al. 2020). The data consist of
105 whole genomes from all 29 species. In the original study, Vanderpool et al. restricted inference
106 to 1,730 single-copy clusters present in 27 of the 29 studied species, a relatively small proportion
107 of the >20,000 genes available from each species; the species tree was inferred using
108 concatenated maximum likelihood, concatenated maximum parsimony, and quartet-based
109 approaches applied to gene trees inferred using both maximum likelihood and maximum
110 parsimony. The authors found robust relationships among all species except the Platyrrhini
111 ("New World Monkeys"), for which inferences differed across species-tree and gene-tree
112 inference methods. In this paper, we compare inferences from three major types of dataset:
113 single-copy families, orthologs extracted from larger families using tree-based decomposition
114 approaches, and all families including all data (orthologs + paralogs). These datasets are then
115 compared in three different phylogenetic applications. First, we compare the species trees
116 inferred from these datasets using several methods, including concatenation-based and gene-tree-
117 based approaches. Second, we compare several measures of nodal support and nodal
118 consistency, as well as branch length estimates across datasets. Finally, we perform tests of
119 introgression and compare results across different datasets. Our results suggest minimal effects
120 of the dataset used on downstream phylogenetic inference, while highlighting the fact that both
121 tree-based decomposition approaches and approaches using both orthologs and paralogs greatly
122 expand the amount of data available.
123
124 **Results**
125 ***Using all gene families vastly expands the data available for phylogenetics***
126 We compared three types of datasets produced by clustering approaches: single-copy
127 clusters, orthologs extracted from all clusters using tree-based decomposition approaches, and all
128 clusters (orthologs + paralogs) (Figure 1). For all datasets we considered both a stringent

129 missing-data threshold (only those genes present in at least 27 of the 29 sampled species;
130 MIN27) and a relaxed missing-data threshold (only those genes present in at least 4 of the 29
131 sampled species; MIN4). First, we considered only clusters that included a single gene from each
132 species (single-copy clusters; SCCs). While these genes are not guaranteed to be orthologs—due
133 to the potential inclusion of pseudoorthologs (Doolittle and Brown 1994; Koonin 2005)—this is
134 considered a safe approach and is often employed in phylogenomics. As expected, this dataset
135 included the fewest genes (Table 1).

136 Tree-based decomposition approaches aim to extract orthologous genes from any
137 cluster/family. We constructed gene trees for all clusters, and then used several tree-based
138 approaches to extract orthologous genes. First, we considered those clusters in which all
139 duplications were specific to a single lineage and kept a single gene-copy from this lineage.
140 When duplications are restricted to a single lineage, choosing one of the copies as the ortholog
141 cannot mislead phylogenetic inference regardless of which sequence is retained (see Figure 1d
142 from Smith and Hahn 2021a; Figure S1a). This dataset ("lineage-specific duplicates"; LSDs)
143 included more than 4X as many genes as the SCC dataset (Table 1). Next, we further expanded
144 our criteria to include those clusters with duplications specific to a pair of lineages ("two-species
145 duplicates"; TSDs; Figure S1b). Such duplications also cannot mislead topological inference,
146 though picking a non-orthologous pair could lead to longer branches. It is straightforward to pick
147 the most closely related pair of genes from the two species, which should not mislead either
148 topological or branch length inferences; including these genes further expanded the dataset
149 compared to the LSD dataset (Table 1).

150 We considered two tree-based decomposition approaches from Yang and Smith (Yang
151 and Smith 2014): minimum inclusion (MI) and monophyletic outgroups (MO). The MI approach
152 takes a gene tree and iteratively extracts subtrees with the highest number of taxa without taxon
153 duplication, until it cannot extract anymore subtrees with the minimum number of taxa. The MO
154 approach considers only those gene trees with a monophyletic outgroup, roots the tree, and infers
155 gene duplications from the root to the tips, pruning at nodes with duplications. These two
156 approaches were each applied to three datasets: the original gene trees, the original gene trees
157 trimmed to remove lineage-specific duplicates, and the original gene trees trimmed to remove
158 both lineage-specific and two-lineage duplicates. We explored the effects of additional filtering
159 and alternative parameters for the MI approach; as these changes had minimal effects, the results

160 are presented in the Supplementary Material (Appendix A). We also considered a new tree-based
161 decomposition approach: subtree extraction (SE). In this approach, we midpoint-root gene trees,
162 trimming away lineage-specific and two-lineage duplicates. We then extract subtrees that include
163 a single representative from each taxon (i.e., subtrees with no duplicates) and keep those trees
164 that meet minimum taxon-sampling thresholds (Figure S1c, d).

165 All tree-based approaches further expanded the amount of data available (Table 1). Since
166 the SE and MI approaches are highly similar (neither requires an outgroup, and both aim to
167 extract subtrees with no duplication events), we further examined the genes extracted using the
168 two approaches. We compared the MI dataset with TSDs trimmed and a minimum of 27 taxa to
169 the SE dataset with a minimum of 27 taxa sampled (this method trims TSDs internally). The
170 number of trees extracted using the two approaches was very similar (12,046 vs 12,198 genes in
171 the MI and SE datasets, respectively). For the 12,046 trees in the MI dataset, there was no analog
172 in the SE dataset for 2.4%, there was an identical tree in the SE dataset for 92.7%, and there was
173 a similar tree in the SE dataset for 4.8% (median Robinson-Foulds distance of these trees=2.0).
174 Thus, the MI and SE approaches extract very similar subsets of trees from the original clusters.

175 Finally, we considered two approaches that made no attempt to remove paralogs from the
176 dataset. We considered one dataset in which all orthologs and paralogs were included ("All
177 Paralogs"). This dataset was the most complete, as, even though it had fewer gene trees than
178 some tree-based approaches, the gene trees from these tree-based approaches are subtrees
179 extracted from this full dataset. This dataset cannot be analyzed using concatenation methods
180 because these approaches require an alignment that includes a single sequence for each species.
181 To address this, and to evaluate the effects of stochastic sampling of paralogs, we also included a
182 dataset in which a single gene (without regard to whether it was an ortholog or paralog) was
183 sampled at random from each species ("One Paralogs").

184 In total, we considered 20 datasets each with MIN4 and MIN27 taxon sampling. The
185 number of gene families ranged from 1,820 to 27,900 (Table 1). Clearly, considering only SCCs
186 drastically restricts the amount of data available, in terms of the number of gene trees (Table 1),
187 the number of decisive sites for each branch of the species tree (Figure 2A), and the number of
188 gene trees informative about each branch of the species tree (Figure 2B). All other datasets are
189 subsets of the All Paralogs dataset, and thus this dataset is necessarily the most informative.
190 Apart from the All Paralogs dataset, including a randomly sampled paralog (One Paralogs) leads

191 to the most decisive sites (Figure 2A), though they are not necessarily the most accurate sites
192 (see below and Figure 3). MI and SE lead to the most informative gene trees (Figure 2B).
193
194 ***Species tree inference is largely consistent across datasets***
195 We inferred species trees using six approaches: ASTRAL (Sayyari and Mirarab 2016;
196 Zhang et al. 2018; Rabiee et al. 2019) on maximum likelihood (ML) gene trees, ASTRAL on
197 maximum parsimony (MP) gene trees, ASTRID (Vachaspati and Warnow 2015) on ML gene
198 trees, ASTRID on MP gene trees, concatenated ML inference in IQ-Tree (Nguyen et al. 2015),
199 concatenated MP inference in PAUP* (Swofford 2001), SVDQuartets (Chifman and Kubatko
200 2014), and ASTRAL-Pro (Zhang et al. 2020) on MP and ML gene trees. ML gene trees were
201 inferred in IQ-Tree, while MP gene trees were inferred in PAUP*. ASTRAL, ASTRID,
202 concatenated ML, and concatenated MP were all developed with orthologs in mind, but
203 ASTRAL has subsequently been demonstrated to be statistically consistent under models of gene
204 duplication and loss (Hill et al. 2020; Legried et al. 2020; Markin and Eulenstein 2020).
205 ASTRAL-Pro, on the other hand, was designed with paralogs in mind, and was only applied to
206 the All Paralogs datasets.
207 Across all nodes of the species tree, except for the relationships among the Platyrrhini
208 (discussed below), an identical phylogeny was recovered across all datasets and species tree
209 inference methods (Figure 3) with two exceptions. When concatenated MP or SVDQuartets was
210 used to infer a species tree from the One Paralogs dataset (MIN27), *Macaca fasciularis* was
211 recovered as sister to *M. nemestrina* rather than *M. mulatta*, as in all other datasets and previous
212 studies (e.g., Vanderpool et al. 2020), but bootstrap support for this relationship was low (55%)
213 in the SVDQuartets analysis. Additionally, when SVDQuartets was used to infer a species tree
214 from the One Paralogs (MIN4) dataset, *Mandrillus leucophaeus* was recovered as sister to a
215 clade containing *Cerocebus atys, Papio anubis,* and *Theropithecus gelada*, rather than sister to
216 *Cerocebus atys* as in other analyses and previous studies, but bootstrap support for this
217 relationship was also low (< 50%).
218 Branch support values were also highly similar across filtering methods. Local posterior
219 probabilities were 1.0 in ASTRAL for all datasets and for all nodes except the contentious node
220 in the Platyrrhini. All bootstrap support values in the concatenated ML analyses were 100, and
221 all bootstrap support values were 100 in the concatenated MP analyses except for in the One

222 Paralogs (MIN27) dataset, which also had topological issues among macaques as mentioned
223 above. Similarly, in all SVDQuartets analyses bootstrap values were 99 or 100 except in the
224 Platyrrhini and One Paralogs datasets.
225 In addition to branch support values, we calculated measures of genealogical
226 discordance: gene and site concordance factors (Minh et al. 2020). These analyses were carried
227 out for all datasets except All Paralogs, because it is not possible to calculate these statistics for
228 this dataset. For all datasets except the One Paralogs dataset, site and gene concordance factors
229 were highly similar across datasets (Figure 3A, B, C). Concordance in the One Paralogs dataset
230 was consistently lower, as would be expected from the random sampling of homologs. In some
231 cases, gene concordance factors were slightly lower for the SCC and MO dataset than for the
232 other datasets (Figure 3B); this seems to be due to more genes that fall into the 'paraphyly'
233 category (i.e., genes for which at least one of the reference clades for a particular branch is not
234 monophyletic), rather than for more genes supporting either of the two minor topologies. gCFs
235 and sCFs for the MIN4 datasets are shown in Supplemental Figure S2.
236
237 ***Resolution of the Platyrrhini radiation varies across species tree and gene tree inference***
238 ***methods***
239 As in Vanderpool et al. (2020), we found uncertainty around relationships among the Platyrrhini.
240 Concatenated ML analyses and gene-tree based analyses that relied on gene trees inferred using
241 ML preferred a symmetric tree, with *Saimairi boliviensis* and *Cebus capucinus imitator* as sister
242 species and *Callithrix jacchus and Aotus nancymaae* as sister species (topology 1 in Figure 4A).
243 However, concatenated MP and gene-tree based analyses that relied on gene trees inferred using
244 MP preferred an asymmetric topology, with *Saimiri boliviensis* and *Cebus capucinus imitator*
245 sister and *Aotus nancymaae* sister to these two (topology 2 in Figure 4A). Finally, SVDQuartets
246 preferred a third topology that placed *Callithrix* sister to *Saimairi boliviensis* and *Cebus*
247 *capucinus* (topology 3 in Figure 4A).
248 Gene and site concordance factors clarify these results. A slight majority of ML gene
249 trees prefer topology 1 (Figure 4B), a majority of MP gene trees prefer topology 2 (Figure 4B),
250 while slightly more sites support topology 2 than support topology 1 (Figure 4C). While the
251 results from SVDQuartets may seem counterintuitive at first, SVDQuartets relies on symmetry
252 between the two minor topologies to infer the third topology as the correct topology. Since there

253 are relatively equal numbers of sites supporting topologies 1 and 2, it is therefore expected that
254 SVDQuartets would prefer topology 3, even though fewer sites support this topology. Results for
255 the MIN4 dataset are similar and are shown in Figure S3.
256 To further investigate the causes of disagreement among these taxa, we focused on the
257 SCC dataset with MIN27 filtering to compare ML and MP gene trees. For each gene, we
258 recorded the ML and MP gene tree topology and the site concordance factor with respect to the
259 focal node, as well as various summary statistics about each locus (number of site patterns,
260 number of parsimony informative sites, tree length, etc.). The percentage of sites supporting the
261 best topology was highest when ML and MP gene trees agreed (Supplemental Figure S6a,c).
262 Additionally, there was more variance in sCFs within a gene (i.e., the number of sites supporting
263 each topology differed more) when ML and MP gene trees agreed (Supplemental Figure S6a,b).
264 This suggests that for genes with similar numbers of sites supporting multiple topologies, ML
265 and MP were more likely to infer conflicting gene trees. Notably, 17.6 percent of gene trees
266 supported Tree 1 under both ML and MP inference, while 18.8 percent of the gene trees
267 supported Tree 2 under both ML and MP inference.
268
269 ***Branch length estimates are largely consistent across datasets***
270 We inferred branch lengths using two approaches. In general, our results suggest that all
271 methods that extract orthologs perform similarly and should lead to reliable estimates of branch
272 lengths. First, we estimated branch lengths in units of substitutions per site using concatenated
273 ML (i.e., site-based branch lengths). We expect that the inclusion of paralogs will lead to
274 overestimates of site-based branch lengths, since the divergence times of paralogs should pre-
275 date divergence times of orthologs. As expected, estimated site-based branch lengths for the One
276 Paralogs dataset are longer than those estimated for the SCC dataset (Figure 5a, b). For all other
277 MIN27 datasets, estimated site-based branch lengths were highly similar to those from the SCC
278 dataset (Figure 5 c, d). However, there are some inconsistencies with site-based branch lengths
279 for terminal branches (Figure 5d), and all site-based branch lengths are more variable for the
280 MIN4 datasets (Supplemental Figure S4).
281 We also inferred discordance-based branch lengths in coalescent units using ASTRAL
282 for the ML gene tree datasets. We expect that the inclusion of paralogs will lead to
283 underestimated discordance-based branch lengths, because datasets with paralogs should have

284 higher levels of discordance. As expected, estimated discordance-based branch lengths from the
285 All and One Paralogs dataset using ASTRAL are shorter than those estimated from the All
286 Paralogs dataset using ASTRAL-Pro, a method that accounts for the extra discordance caused by
287 the inclusion of paralogs (Figure 5 e, f, g). In general, across all datasets except the two including
288 paralogs (All and One), discordance-based branch lengths were highly similar to those estimated
289 in ASTRAL-Pro (Figure 5g). However, there were some surprising results. Specifically, the SCC
290 and MO datasets led to slightly shorter discordance-based branch length estimates than both
291 ASTRAL-Pro and the datasets from other tree-based decomposition methods (Figure 5g). In
292 addition, all discordance-based branch length estimates are relatively short, which could be
293 explained by difficulties estimating the lengths of longer branches with very little gene tree
294 discordance (i.e., for which all (or most) genes support a single topology) in ASTRAL.
295
296 ***Tests for introgression are consistent across datasets***
297 To test for introgression, we looked for a deviation from the expected number of alternate
298 gene tree topologies using the statistic Δ (Huson et al. 2005; Vanderpool et al. 2020). We used
299 only the ML gene trees from each dataset for this analysis. There was evidence of introgression
300 across several branches of the primate phylogeny (Figure 6A), and values of Δ were similar
301 across datasets (Figure 6B). Notably, there was evidence of introgression in a majority of tests at
302 the contentious node in the Platyrrhini, which may explain difficulties inferring the species tree
303 topology at this node. There was also evidence of introgression in the macaques, as found by
304 Vanderpool et al. (2020). Deeper in the tree, results were more suspect, with tests on some
305 datasets suggesting introgression while others did not (Figure 6B). Results of introgression tests
306 were similar with less stringent missing-data filters (Supplemental Figure S5).
307
308 **Discussion**
309 Our results demonstrate that, regardless of the dataset used, the inferred species tree
310 topology is largely stable across datasets. Regardless of whether all families, families with only a
311 single copy per species, or large families from which orthologs are extracted were used, the only
312 disagreements between trees were with respect to relationships among the Platyrrhini; in this
313 case the species tree inference method was a larger determinant of results than the particular
314 dataset (Figure 4). Despite the overall similarity among results, when a single gene was

randomly sampled per species results were unstable in two cases, suggesting—unsurprisingly—that such a sampling strategy is not ideal. Based on the results presented here, when whole genome sequence data are available, using all of families output by clustering methods followed by the application of gene-tree decomposition methods can greatly expand the data available without sacrificing the accuracy of inference. These results align with several recent simulation and theoretical studies focusing on gene-tree based methods and demonstrating their robustness to paralog inclusion (Legried et al. 2020; Yan et al. 2021).

Despite the stability of inference across most of the tree, there remains disagreement about relationships amongst the Platyrrhini, a notably contentious node (Perelman et al. 2011; Springer et al. 2012; Perez et al. 2013; Jameson Kiesling et al. 2015; Schrago and Seuánez 2019; Wang et al. 2019; Vanderpool et al. 2020). As in Vanderpool et al. (2020), we find that both concatenated ML and ASTRAL based on ML gene trees favor a symmetrical topology (tree 1 in Figure 4A). A bias towards the symmetrical 4-taxon tree is expected when using ML in the presence of recombination and when the time between speciation events is short (Kubatko and Degnan 2007; Roch and Steel 2015). Although the bias on ML under these conditions is often linked to concatenation methods, if the gene trees themselves are inaccurate due to the concatenation of multiple unique histories (e.g., among exons; Mendes et al. 2019), then the same bias in inferred trees can occur. Bias in the gene trees can then lead to bias in the methods that they are used as input to (e.g., ASTRAL). Note that this bias does not affect inferences under maximum parsimony (Mendes and Hahn 2018). Furthermore, there are nearly equal numbers of trees supporting the two best-supported topologies in the primate data (Figure 4B), which suggests two things: first, choosing the best topology will be difficult no matter what method is used, as the evidence in favor of one topology over the other is minimal. Second, there is likely some introgression, since we would otherwise expect equal numbers of the two minor topologies. We do not see equal numbers of the two minor topologies, as confirmed by significant tests for introgression in this clade (Figure 6). Finally, a detailed comparison of SCC gene trees inferred by both ML and MP suggests that genes whose topologies disagreed across the two approaches did not support either topology as strongly as genes for which ML and MP agreed (Figure S6). Of the gene trees that agreed across ML and MP inference, more supported Tree 2 than supported Tree 1 (Figure 4A). Thus, of the genes for which methods agree, more support the asymmetric topology than the symmetric topology (as in Vanderpool et al. 2020).

346 We also compared branch length estimates and tests for introgression across datasets.
347 Branch length estimates are largely consistent across datasets, with the exception of datasets that
348 explicitly include paralogs, which led to biases in expected directions for both discordance-based
349 and site-based branch lengths. Site-based branch lengths are very consistent across all datasets
350 except the One Paralogs dataset when stringent filters for missing data are applied. When
351 paralogs are included, site-based branch lengths are overestimated, as expected (e.g., Siu-Ting et
352 al. 2019). Discordance-based branch lengths (i.e., those estimated in ASTRAL) are
353 underestimated for datasets including paralogs, because these datasets have higher levels of
354 discordance. These methods accommodate increased discordance by positing a shorter time
355 between speciation events. Otherwise, discordance-based branch lengths are largely similar
356 across datasets, though the SCC and MO datasets appear to have slightly shorter estimated
357 branch lengths than all other methods (Figure 5e). Given the consistency of results across tree-
358 based decomposition methods, as well as ASTRAL-Pro, and the vastly larger number of gene
359 trees used in these cases, we suggest that discordance-based branch lengths may actually be
360 underestimated for the SCC and MO datasets. This result is consistent with lower gCFs in these
361 datasets (Figure 3b), and suggests that branch lengths estimated from these datasets may be
362 inaccurate because they include pseudoorthologs.
363 To our knowledge, this is the first evaluation of the effects of including more than just
364 single-copy families on tests for introgression based on the asymmetry in minor topology
365 frequencies. We expected that the inclusion of paralogs would not bias such tests, because under
366 models that include duplication and loss the two minor topologies should occur in equal
367 frequencies (Smith and Hahn 2021a; Smith and Hahn 2021b). Our results largely confirm these
368 expectations: although there is variation in whether or not tests are significant across datasets,
369 estimates of Δ are very similar (Figure 6B). At some nodes there is consistent evidence for
370 introgression across datasets, suggesting a strong signal of asymmetry: for example in the
371 macaques and among the Platyrrhini. Deeper in the tree, there may be more gene tree error (e.g.,
372 due to long-branch attraction), since introgression is detected for some datasets and not for others
373 (Figure 6B).
374 Phylogenetics based on whole-genome sequences almost always begins by identifying
375 homologous genes via clustering. The clustering process operationally defines gene families,
376 using clustering methods that range from very simple to very complex. While the single-copy

377 clusters output by any one of these methods have most often been used in phylogenetics, there is
378 nothing inherently more suitable about these clusters. First, SCCs may not be orthologs, due to
379 the presence of pseudoorthologs—paralogs that are mistaken as orthologs due to differential
380 patterns of gene duplication and loss (Doolittle and Brown 1994; Koonin 2005). In other words,
381 having only a single representative sequence in each species does not guarantee that all the
382 sampled genes are orthologs. Second, and more importantly, the size of clusters identified by
383 clustering approaches is determined by parameters set by the user. For example, in OrthoMCL
384 (Li 2003) the inflation parameter determines the size of output clusters: by changing this
385 parameter, users can identify larger or smaller clusters. Because genes are related to all other
386 genes via a long history of duplication and divergence (with a few exceptions; Knowles and
387 McLysaght 2009; Zhao et al. 2014), there is no single level of similarity that uniquely identifies
388 gene families (Demuth and Hahn 2009). However, users can choose the value of the inflation
389 parameter that identifies more, smaller clusters in order to find more single-copy clusters; this
390 does not mean these genes do not have paralogs, only that more distant paralogs were not
391 included at this clustering threshold. Many clustering methods aim to form groups of genes that
392 descend from a single common ancestor in the studied taxa (e.g., Emms and Kelly 2015), though
393 this does not ensure a lack of duplication events since the common ancestor. While tree-based
394 decomposition approaches still rely on the clustering step to initially identify the homologs from
395 which gene trees are built, their output is directly related to the definitions of orthologs and
396 paralogs, and is more easily interpreted in a phylogenetic context. By applying these
397 decomposition approaches to larger clusters, researchers can avoid arbitrary determinants of
398 which clusters are single-copy and can instead attempt to extract as many sets of orthologs as
399 possible. Not only does this approach increase the amount of data available, but it also uses
400 criteria more directly linked to the evolutionary history of gene families.
401 In conclusion, our results suggest that methods for species tree inference are accurate
402 across datasets whether single-copy clusters are used, tree-based decomposition methods are
403 used to extract orthologs from larger clusters, or all clusters (i.e., orthologs and paralogs) are
404 included. Our results highlight the benefits of using all gene families by showing that the amount
405 of data used can be increased by an order of magnitude (Table 1; Figure 2). While even the
406 smallest dataset was sufficient for accurate species tree inference in this case study, that is not
407 always the case (e.g., Emms and Kelly 2018; Thomas et al. 2020). Finally, more data facilitates

408 inferences beyond species tree topology, including branch length estimates and the detection of
409 introgression.
410
411 **Materials and Methods**
412 ***Dataset and alignment***
413 The full sets of protein-encoding genes for 26 primates and 3 non-primates were obtained
414 as in (Vanderpool et al. 2020), and clusters were obtained as in that study. Briefly, an all-by-all
415 BLASTP search (Altschul et al. 1990; Camacho et al. 2009) was executed, and the longest
416 isoform of each protein-coding gene from each species was used. Then, the mcl algorithm (Van
417 Dongen 2000) as implemented in FastOrtho (Wattam et al. 2014) with an inflation parameter of
418 5 was used to cluster the BLASTP output. CDSs for each cluster that included samples from at
419 least four species were aligned, cleaned and trimmed as in (Vanderpool et al. 2020). Sequences
420 were aligned by codon using GUIDANCE2 (Sela et al. 2015) with MAFFT v7.407 (Katoh and
421 Standley 2013) with 60 bootstrap replicates. Sequence residues with GUIDANCE scores < 0.93
422 were converted to gaps and sites with >50% gaps were removed using Trimalv1.4rev22
423 (Capella-Gutiérrez et al. 2009). Alignments shorter than 200 bp and alignments that were
424 invariant or contained no parsimony informative characters were removed from further analyses.
425 A subset of alignments that could not be aligned by codon were aligned by nucleotide, and
426 subsequent steps were as with the codon-aligned dataset. In total 18,484 alignments were used in
427 downstream analyses.
428
429 ***Gene tree inference***
430 We inferred gene trees from all alignments with at least four species (18,484 alignments)
431 in IQ-TREE v2.0.6 (Nguyen et al. 2015) with nucleotide substitution models selected using
432 ModelFinder (Kalyaanamoorthy et al. 2017) as implemented in IQ-TREE. The full IQ-TREE
433 command used on each alignment was 'iqtree2 -s *alignment name* -m MFP -c 1 -pre *alignment*
434 *name*'. We also inferred gene trees from all 18,484 alignments using the maximum parsimony
435 criterion in PAUP* v 4.0a (Swofford 2001). We treated gaps as missing data, obtained a starting
436 tree via random stepwise addition, held a single tree at each step, and used the TBR branch-
437 swapping algorithm with a reconnection limit of 8. We kept a maximum of 1000 trees and did
438 not collapse zero-length branches.

439

440 ***Filtering***

441 We considered three major groups of filtering methods:

442 1) Single-copy clusters (SCCs): We considered a dataset that consisted only of those clusters
443 that included a single gene copy from each species.

444 2) Tree-based decomposition approaches: We considered several methods that involved
445 trimming branches of gene trees to extract orthologs. All custom branch-cutting operations
446 were written in python3 and used the python package ete3 (Huerta-Cepas et al. 2016) to read,
447 traverse, trim, and output gene trees and modified sequence alignments. We used postorder
448 node traversal when traversing trees, and prior to custom trimming operations, we midpoint-
449 rooted gene trees.

450 i. Lineage-specific duplicates: In this dataset we identified gene duplications that were
451 specific to a single species. For such lineage-specific duplicates, we selected the sequence
452 copy that was closest in length to the median length of sequences in the alignment, kept
453 that copy, and trimmed the other copy or copies from both the alignment and the gene tree.

454 ii. Two-lineage duplicates: To expand our data beyond lineage-specific duplicates, in addition
455 to trimming lineage-specific duplicates, we identified gene duplications specific to a pair of
456 species. For such duplicates, we selected the two sequence copies with the minimum
457 branch distance separating them and trimmed the remaining copies from the tree and the
458 alignment.

459 iii. Minimum Inclusion: We applied the minimum inclusion (MI) approach described in (Yang
460 and Smith 2014) to trim gene trees. We used the python script provided by Yang and Smith
461 (2014; prune_paralogs_MI.py) and used as input one of three sets of gene trees: the original
462 18,484 gene trees, the original 18,484 gene trees with lineage-specific duplicates trimmed,
463 and the original 18,484 gene trees with lineage-specific and two-lineage duplicates
464 trimmed. For the MI approach, branches longer than a specified threshold are trimmed to
465 remove potential pseudoorthologs; we used the following branch-length cutoffs: 0.4
466 substitutions per site for the ML gene trees and 500 changes for MP trees. We explored
467 additional cutoffs in the Supplementary Material (Appendix A).

468 iv. Monophyletic Outgroups: We also applied the monophyletic outgroups (MO) approach
469 described in (Yang and Smith 2014) to trim gene trees. We used the python script provided

470 by Yang and Smith (2014; prune_paralogs_MO.py) and used as input one of three sets of
471 gene trees: the original 18,484 gene trees, the original 18,484 gene trees with lineage-
472 specific duplicates trimmed, and the original 18,484 gene trees with lineage-specific and
473 two-lineage specific duplicates trimmed.
474 v. Subtree Extraction: Finally, we evaluated a new tree-based decomposition approach
475 introduced here (subtree extraction; SE). In this approach we start by midpoint-rooting
476 gene trees, followed by trimming lineage-specific and two-lineage duplicates. We then
477 extract subtrees with a single representative from each taxon (i.e., subtrees with no
478 duplicates) and keep those subtrees that meet minimum taxon-sampling thresholds.
479 3) Paralog methods: We considered two approaches that included paralogs in addition to
480 orthologs. First, we included all genes (All Paralogs). Additionally, we randomly sampled a
481 single gene (without regard to orthology) per species (One Paralogs).
482 For all datasets, we considered a stringent (minimum of 27 of 29 taxa) and relaxed (minimum of
483 4 of 29 taxa) missing data threshold.
484
485 ***Species tree inference***
486 We inferred species trees using six methods. Three methods inferred species trees from
487 concatenated datasets: maximum parsimony, maximum likelihood, and SVDQuartets. To infer
488 an MP tree from the concatenated datasets we used PAUP* v4.0a (build 168) (Swofford 2001).
489 We set the criterion to parsimony, and used 500 bootstrap replicates to assess nodal support. For
490 all other options we used PAUP* defaults. To infer an ML tree from the concatenated dataset we
491 used IQ-TREE v2.0.6 (Nguyen et al. 2015) with nucleotide substitution models selected using
492 ModelFinder (Kalyaanamoorthy et al. 2017) as implemented in IQ-TREE. We used an edge-
493 linked, proportional partition model (Chernomor et al. 2016) and 1000 ultrafast bootstrap
494 replicates (Hoang et al. 2018). The full IQ-TREE command used on each alignment was 'iqtree2
495 -s *alignment name* -p *partition file name* -c 1 -pre *alignment name -B 1000*'. For three
496 alignments, IQTree v2.0.6 failed to run, and, based on a suggestion from the developers, we
497 reverted to IQTree v.1.6.12 to infer the species trees for these alignments. For these three
498 alignments, the full IQ-TREE command used was 'iqtree -s *alignment name* -spp *partition file*
499 *name* -pre *alignment name* -bb 1000 -nt 4'. Finally, to infer a species tree from the concatenated
500 alignments using SVDQuartets, we used PAUP* v4.0a (build 168) (Swofford 2001). We

evaluated all quartets and treated ambiguous sites as missing to infer the species tree topology. To assess nodal support, we evaluated 10,000 random quartets for each of 100 bootstrap replicates. We used the multilocus bootstrapping option and again treated ambiguous sites as missing.

In addition to the three concatenation-based methods, we inferred species trees using three gene-tree based methods. Prior to inferring species trees or estimating discordance (see below) from filtered gene trees, we collapsed all zero-length branches. For each gene tree, we did the following: First, we midpoint-rooted the gene tree. Then, we calculated site concordance factors using IQ-Tree v2.0.6 (Minh et al. 2020) for the alignment with the rooted gene tree as the reference tree. We used 100 randomly sampled quartets to compute the sCF, collapsing any nodes where sN == 0; in other words, any nodes for which no sites were informative.

We inferred a species tree using ASTRAL-III v5.7.3 (Sayyari and Mirarab 2016; Zhang et al. 2018; Rabiee et al. 2019). ASTRAL infers a species tree from a set of gene trees by extracting quartets and finding the species tree that maximizes the number of shared quartet trees. It has been demonstrated to be consistent under the multispecies coalescent (MSC) model (Mirarab et al. 2014) and under models of gene duplication and loss (Legried et al. 2020). Gene trees obtained using ML and MP, from all datasets described above, and with zero-length branches collapsed, were used as input to ASTRAL; local posterior probabilities were used to assess nodal support. For the All Paralogs dataset, we used the mapping file and treated each gene copy as a separate individual. Additionally, we inferred species trees using ASTRID v2.2.1 (Vachaspati and Warnow 2015), again using the filtered and zero-length collapsed ML and MP gene trees as input. ASTRID is a distance-based approach that estimates species trees using internode distances and is statistically consistent under the MSC model (Vachaspati and Warnow 2015). As in ASTRAL, for the All Paralogs dataset we treated gene copies from the same species as individuals using the mapping file. Finally, we inferred species trees from the All Paralogs datasets using ASTRAL-Pro (Zhang et al. 2020). ASTRAL-Pro uses an internal rooting-and-tagging algorithm to label nodes as duplication or speciation nodes, and then infers quartets using only speciation nodes before finding the species tree that maximizes the number of shared quartet trees. ASTRAL-Pro has been shown to be statistically consistent under a model of gene duplication and loss (Zhang et al. 2020).

532 ***Assessing discordance***

533 To assess levels of discordance across datasets we calculated site and gene concordance
534 factors in IQ-Tree v2.0.6 (Minh et al. 2020). We used as the reference tree the tree shown in
535 Figure 3, and to estimate sCFs, we used 1000 randomly sampled quartets. gCFs were estimated
536 for filtered ML and MP gene trees after zero-length branches were collapsed. sCFs were
537 estimated for the alignments that resulted from filtering the ML gene trees.
538

539 ***Testing for introgression***

540 We used the approach used in (Vanderpool et al. 2020) to test for introgression. Briefly,
541 the introgression test assesses whether there is a deviation from the expected equal numbers of
542 alternative tree topologies (under the MSC model without gene flow) using the statistic Δ (Huson
543 et al. 2005), where
544

545 $$\Delta = \frac{Number\ of\ DF1\ trees - Number\ of\ DF2\ trees}{Number\ of\ DF1\ trees + Number\ of\ DF2\ trees}$$

546 *DF1* represents the most common minor topology, and *DF2* represents the least common minor
547 topology. In the absence of introgression, Δ is expected to be equal to zero. To test whether
548 deviations from zero were significant, we followed the procedure of (Vanderpool et al. 2020) and
549 used 2,000 datasets generated by resampling gene trees with replacement, considering only those
550 nodes where more than five percent of trees were discordant. This distribution was used to
551 calculate *Z*-scores and *p*-values for the observed Δ statistic, and for each filtered dataset, we
552 corrected for multiple comparisons using the Dunn-Sidák correction (Dunn 1959; Šidák 1967).
553
554

Acknowledgements

This work was supported by a National Science Foundation postdoctoral fellowship to MLS (DBI-2009989) and an NSF grant to MWH (DEB-1936187).

Data Availability

Scripts used for filtering gene trees are available on GitHub (github.com/meganlsmith/Primate_Paralogs). Alignments, gene trees, and species trees are available from FigShare (doi: 10.6084/m9.figshare.16653025).

564 **Tables**

565

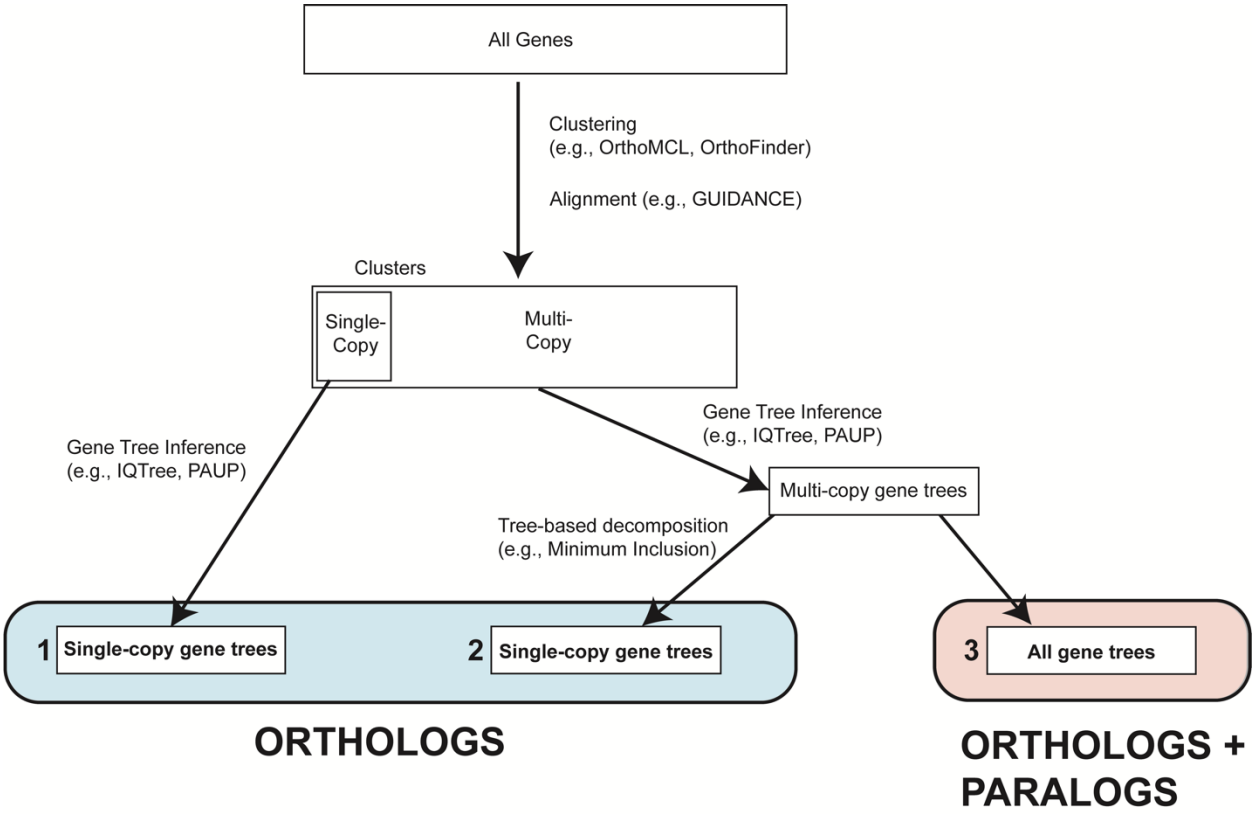
Filter	MIN4	MIN27
Single-copy clusters	5771	1820
Lineage-specific duplicates (LSD)	13627	7693
Two-species duplicates (TSD)	14931	8719
Minimum Inclusion	27880	4849
Minimum Inclusion (LSD)	22360	11479
Minimum Inclusion (TSD)	21793	12046
Monophyletic Outgroups	9724	4805
Monophyletic Outgroups (LSD)	16962	10222
Monophyletic Outgroups (TSD)	17104	10254
Subtree Extraction	20562	12198
All Paralogs	18484	11555
One Paralogs	18484	11555

566

567 **Table 1: Number of genes trees included with different filtering approaches.** LSD and TSD
568 indicate when lineage-specific and both lineage-specific and two-lineage specific duplicates were
569 trimmed; the subtree extraction method trims these automatically. The MIN4 dataset required a
570 minimum of 4 taxa (out of 29 total), while the MIN27 dataset required a minimum of 27 taxa.

571

572 **Figures**



573

574

575 **Figure 1.** Conceptual overview of methods for inferring species trees from genomic data. We
576 begin with All Genes, clustering them into gene families. We can then use single-copy ortholog
577 clusters for inference (Dataset 1), use tree-based decomposition approaches to extract orthologs
578 from all clusters (Dataset 2), or infer species trees from all clusters (i.e., from datasets including
579 orthologs and paralogs; Dataset 3).

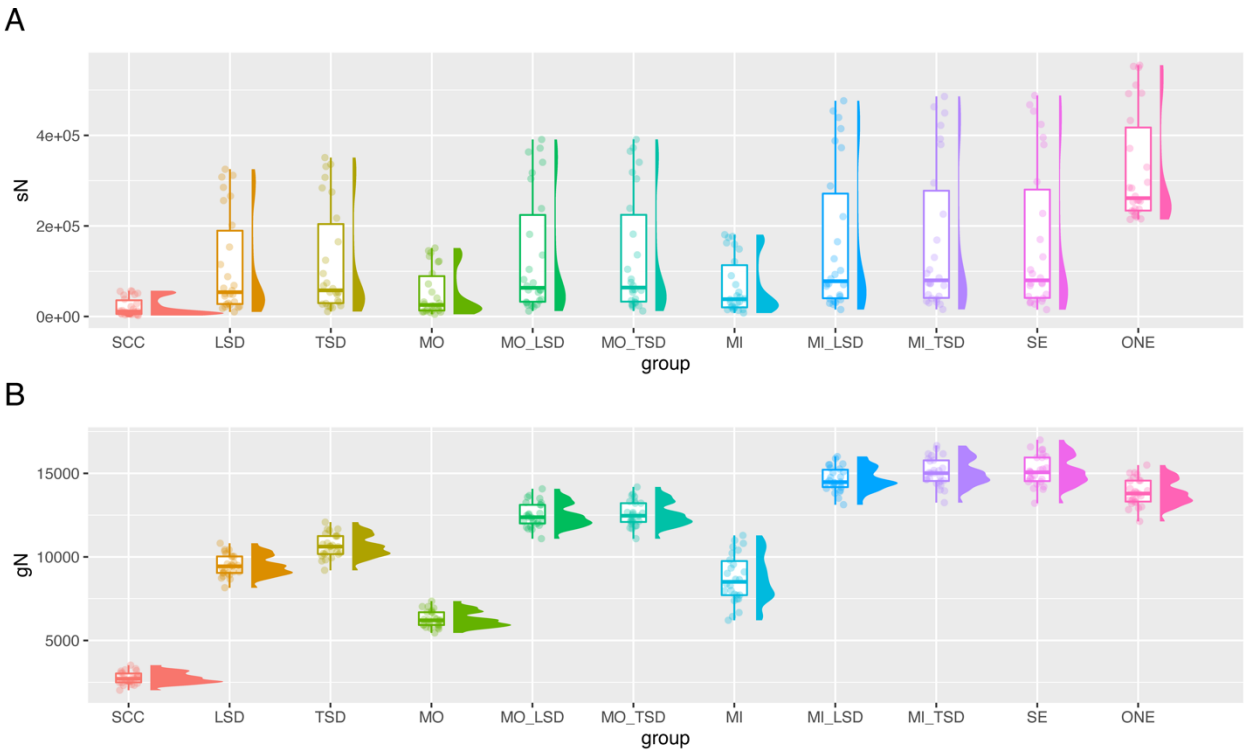


Figure 2. Numbers of informative genes and sites across datasets using the MIN27 datasets. A) The distribution of the number of decisive sites (across branches) as calculated in IQ-Tree. Decisive sites are defined in Minh et al. (2020). B) The distribution of the number of decisive gene trees (across branches) as calculated in IQ-Tree. Decisive gene trees are defined in Minh et al. (2020). SCC=single-copy clusters; LSD=lineage-specific duplicates; TSD=two-species duplicates; MO=monophyletic outgroup; MI=minimum inclusion; SE=subtree extraction; ONE=one paralogs.

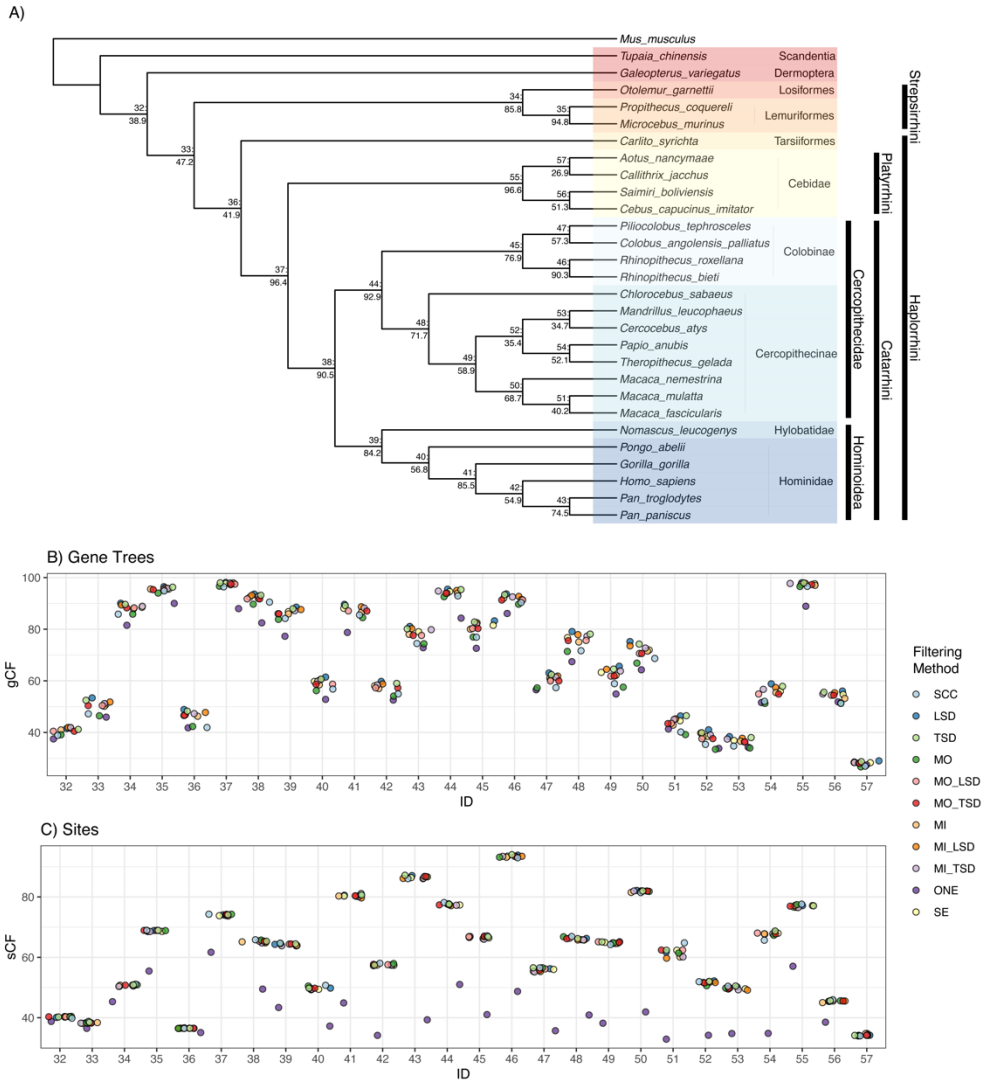


Figure 3. Gene (gCF) and site (sCF) concordance factors among primate datasets using ML gene trees (MIN27). A) Primate phylogeny from ASTRAL using the ML gene trees (all input datasets give the same topology). Nodes show Node ID: gCF values from the SCC dataset. B) Distribution of gCF values across datasets. C) Distribution of sCF values across datasets. Node IDs correspond to the numbers displayed on the tree in panel A. SCC=single-copy clusters; LSD=lineage-specific duplicates; TSD=two-species duplicates; MO=monophyletic outgroup; MI=minimum inclusion; SE=subtree extraction; ONE=one paralogs.

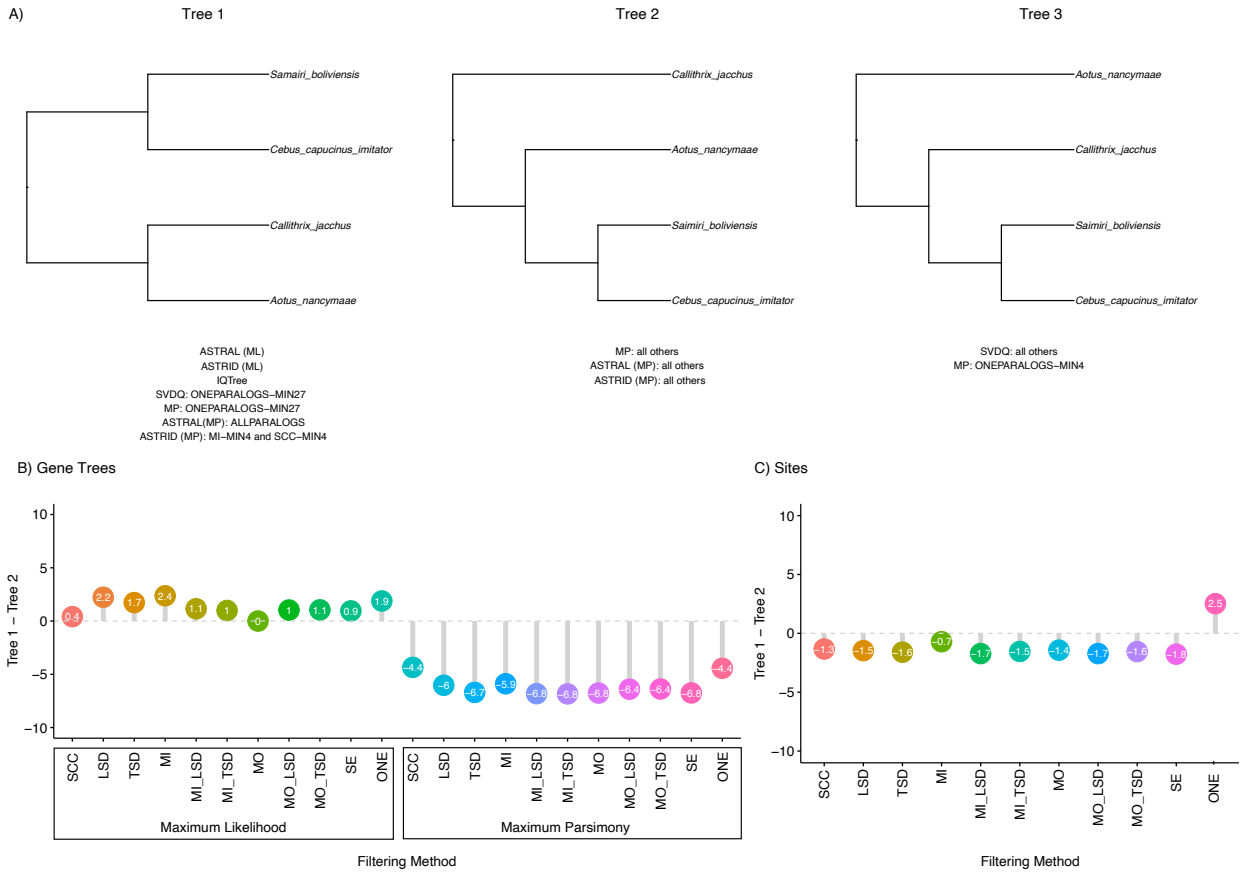


Figure 4. Alternative resolutions of Platyrrhini relationships. A) The three most common tree topologies. Below each resolution, inference methods and filtering approaches that supported the topology are listed. B) The percentage of gene trees supporting Tree 1 minus the percentage of gene trees supporting Tree 2 for ML and MP gene trees across datasets. C) The percentage of sites supporting Tree 1 minus the percentage of sites supporting Tree 2 across datasets. SCC=single-copy clusters; LSD=lineage-specific duplicates; TSD=two-species duplicates; MO=monophyletic outgroup; MI=minimum inclusion; SE=subtree extraction; ONE=one paralogs. Results in B and C from MIN27 datasets.

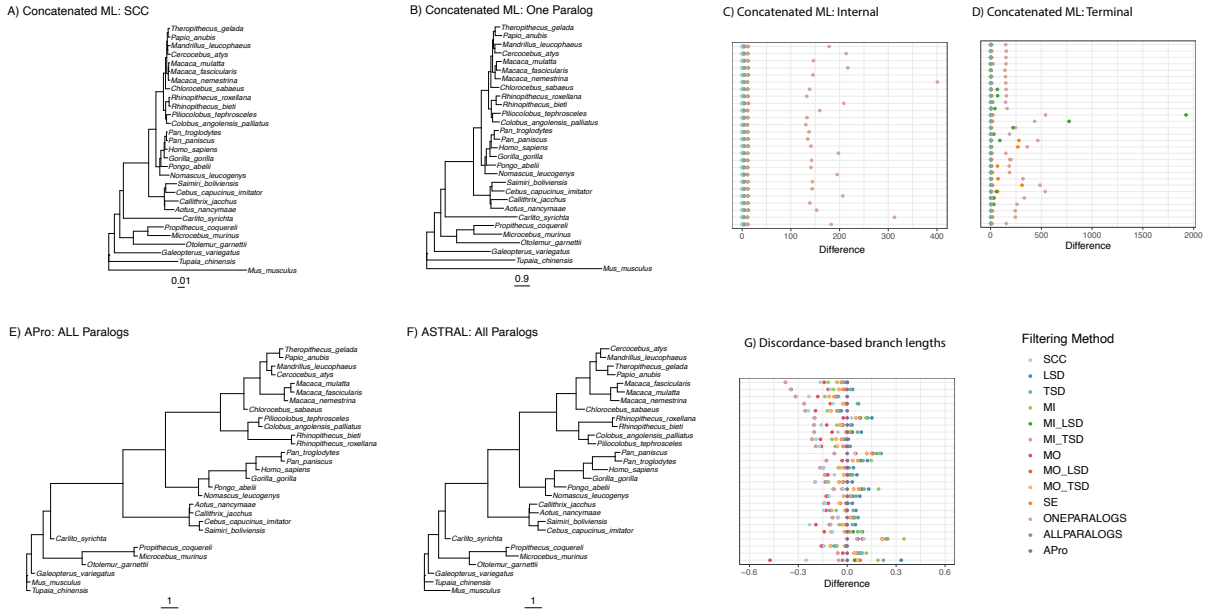


Figure 5. Branch lengths across datasets and species tree inference methods. Site-based branch lengths estimated using concatenated ML when A) SCCs and B) one randomly selected paralog per species are used for inference. Note the different scales in panels A and B. C) Difference between site-based branch lengths for internal branches from the SCC dataset and all other datasets, normalized by SCC branch length. D) Same as in panel C, but for terminal branches. Discordance-based branch lengths calculated on the All Paralogs dataset when E) ASTRAL-Pro and F) ASTRAL are used for inference. Note that terminal branch lengths are arbitrary in these panels. G) Difference between discordance-based branch lengths estimated with ASTRAL-Pro (APro) and all other methods, normalized by APro branch length. Colors represent different filtering methods, and each row is a different branch. SCC=single-copy clusters; LSD=lineage-specific duplicates; TSD=two-species duplicates; MO=monophyletic outgroup; MI=minimum inclusion; SE=subtree extraction; ONE=one paralogs. Results from MIN27 datasets.

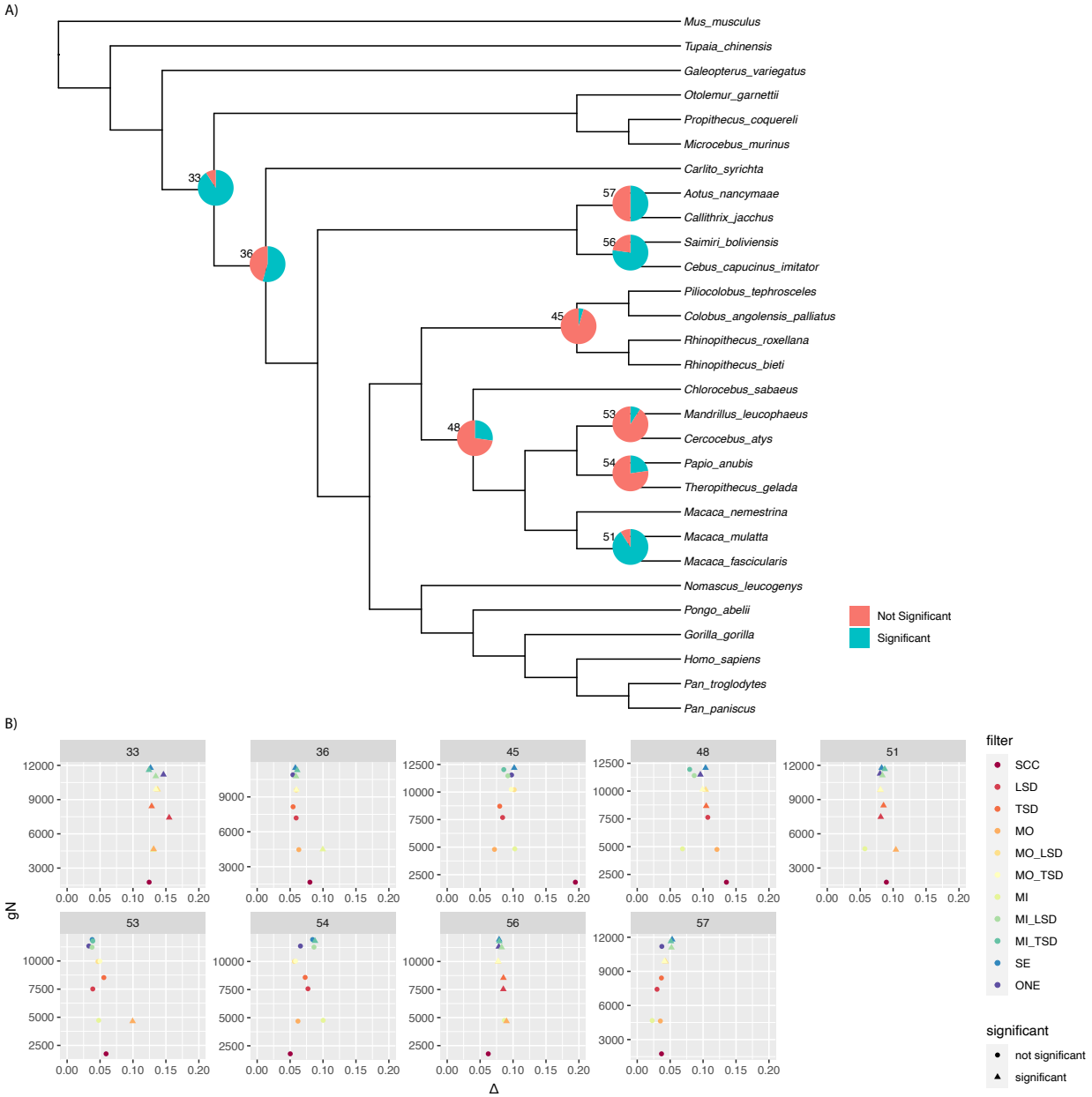


Figure 6. Results of introgression tests on MIN27 ML gene trees. A) Pie charts are shown for branches with any significant introgression tests. Numbers are node numbers. B) For all branches with some significant tests, we show the number of informative genes versus Δ. Observations are colored by filtering method, and shapes indicate whether a particular test was significant. SCC=single-copy clusters; LSD=lineage-specific duplicates; TSD=two-species duplicates; MO=monophyletic outgroup; MI=minimum inclusion; SE=subtree extraction; ONE=one paralogs.

633 **Reference**

634 Altenhoff AM, Glover NM, Dessimoz C. 2019. Inferring orthology and paralogy. In: Anisimova
635 M, editor. Evolutionary genomics: Statistical and computational methods. New York,
636 NY: Springer. p. 149–175. Available from: https://doi.org/10.1007/978-1-4939-9074-0_5

637 Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool.
638 *J Mol Biol* 215:403–410.

639 Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009.
640 BLAST+: architecture and applications. *BMC Bioinformatics* 10:421.

641 Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. 2009. trimAl: a tool for automated
642 alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* 25:1972–1973.

643 Chernomor O, von Haeseler A, Minh BQ. 2016. Terrace aware data structure for phylogenomic
644 inference from supermatrices. *Syst Biol* 65:997–1008.

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

647 Demuth JP, Hahn MW. 2009. The life and death of gene families. *BioEssays* 31:29–39.

648 Doolittle WF, Brown JR. 1994. Tempo, mode, the progenote, and the universal root. *PNAS*
649 91:6721–6728.

650 Dunn CW, Howison M, Zapata F. 2013. Agalma: an automated phylogenomics workflow. *BMC*
651 *Bioinformatics* 14:330.

652 Dunn OJ. 1959. Confidence intervals for the means of dependent, normally distributed variables.
653 *J Am Stat Assoc* 54:613–621.

654 Emms DM, Kelly S. 2015. OrthoFinder: solving fundamental biases in whole genome
655 comparisons dramatically improves orthogroup inference accuracy. *Genome Biol* 16:157.

656 Emms DM, Kelly S. 2018. STAG: Species Tree Inference from All Genes.
657 bioRxiv:10.1101/267914.

658 Fernández R, Kallal RJ, Dimitrov D, Ballesteros JA, Arnedo MA, Giribet G, Hormiga G. 2018.
659 Phylogenomics, diversification dynamics, and comparative transcriptomics across the
660 spider tree of life. *Curr Biol* 28:1489–1497.

661 Fitch WM. 1970. Distinguishing homologous from analogous proteins. *Syst Zool* 19:99–113.

662 Goodman M, Czelusniak J, Moore GW, Romero-Herrera AE, Matsuda G. 1979. Fitting the gene
663 lineage into its species lineage, a parsimony strategy illustrated by cladograms
664 constructed from globin sequences. *Syst Biol* 28:132–163.

665 Hill M, Legried B, Roch S. 2020. Species tree estimation under joint modeling of coalescence
666 and duplication: sample complexity of quartet methods. arXiv:2007.06697.

667 Hoang DT, Chernomor O, von Haeseler A, Minh BQ, Vinh LS. 2018. UFBoot2: Improving the
668 ultrafast bootstrap approximation. *Mol Biol Evol* 35:518–522.

669 Huerta-Cepas J, Serra F, Bork P. 2016. ETE 3: reconstruction, analysis, and visualization of
670 phylogenomic data. *Mol Biol Evol* 33:1635–1638.

671 Huson DH, Klöpper T, Lockhart PJ, Steel MA. 2005. Reconstruction of reticulate networks from
672 gene trees. In: Miyano S, Mesirov J, Kasif S, Istrail S, Pevzner PA, Waterman M, editors.
673 Research in Computational Molecular Biology. Vol. 3500. Lecture Notes in Computer
674 Science. Berlin, Heidelberg: Springer Berlin Heidelberg. p. 233–249. Available from:
675 http://link.springer.com/10.1007/11415770_18

676 Jameson Kiesling NM, Yi SV, Xu K, Gianluca Sperone F, Wildman DE. 2015. The tempo and
677 mode of New World monkey evolution and biogeography in the context of phylogenomic
678 analysis. *Mol Phylogenet Evol* 82:386–399.

679 Kallal RJ, Fernández R, Giribet G, Hormiga G. 2018. A phylotranscriptomic backbone of the
680 orb-weaving spider family Araneidae (Arachnida, Araneae) supported by multiple
681 methodological approaches. *Mol Phylogenet Evol* 126:129–140.

682 Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermiin LS. 2017. ModelFinder:
683 fast model selection for accurate phylogenetic estimates. *Nature Methods* 14:587–589.

684 Katoh K, Standley DM. 2013. MAFFT Multiple Sequence Alignment Software Version 7:
685 Improvements in Performance and Usability. *Mol Biol Evol* 30:772–780.

686 Knowles DG, McLysaght A. 2009. Recent de novo origin of human protein-coding genes.
687 *Genome Res* 19:1752–1759.

688 Koonin EV. 2005. Orthologs, paralogs, and evolutionary genomics. *Annu Rev Genet* 39:309–
689 338.

690 Kubatko LS, Degnan JH. 2007. Inconsistency of phylogenetic estimates from concatenated data
691 under coalescence. *Syst Biol* 56:17–24.

692 Legried B, Molloy EK, Warnow T, Roch S. 2020. Polynomial-time statistical estimation of
693 species trees under gene duplication and loss. *J Comput Biol* 28:452-468.

694 Li L. 2003. OrthoMCL: Identification of ortholog groups for eukaryotic genomes. *Genome Res*
695 13:2178–2189.

696 Markin A, Eulenstein O. 2020. Quartet-Based inference methods are statistically consistent
697 under the unified duplication-loss-coalescence model. *arXiv*:2004.04299.

698 Mendes FK, Livera AP, Hahn MW. 2019. The perils of intralocus recombination for inferences
699 of molecular convergence. *Philos Trans R Soc London [Biol]* 374:20180244.

700 Minh BQ, Hahn MW, Lanfear R. 2020. New methods to calculate concordance factors for
701 phylogenomic datasets. *Mol Biol Evol* 37:2727–2733.

702 Mirarab S, Reaz R, Bayzid MdS, Zimmermann T, Swenson MS, Warnow T. 2014. ASTRAL:
703 genome-scale coalescent-based species tree estimation. *Bioinformatics* 30:i541–i548.

704 Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ. 2015. IQ-TREE: A Fast and Effective
705 Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. *Mol Biol Evol*
706 32:268–274.

Perelman P, Johnson WE, Roos C, Seuánez HN, Horvath JE, Moreira MAM, Kessing B, Pontius J, Roelke M, Rumpler Y, et al. 2011. A Molecular Phylogeny of Living Primates. *PLOS Genetics* 7:e1001342.

Perez SI, Tejedor MF, Novo NM, Aristide L. 2013. Divergence Times and the Evolutionary Radiation of New World Monkeys (Platyrrhini, Primates): An Analysis of Fossil and Molecular Data. *PLOS ONE* 8:e68029.

Rabiee M, Sayyari E, Mirarab S. 2019. Multi-allele species reconstruction using ASTRAL. *Mol Phylogenet Evol* 130:286–296.

Roch S, Steel M. 2015. Likelihood-based tree reconstruction on a concatenation of aligned sequence data sets can be statistically inconsistent. *Theor Popul Biol* 100:56–62.

Sayyari E, Mirarab S. 2016. Fast coalescent-based computation of local branch support from quartet frequencies. *Mol Biol Evol* 33:1654–1668.

Schrago CG, Seuánez HN. 2019. Large ancestral effective population size explains the difficult phylogenetic placement of owl monkeys. *Am J Primatol* 81:e22955.

Scornavacca C, Delsuc F, Galtier N. 2020. Phylogenetics in the genomic era. Open access book available from https://hal.inria.fr/PGE/.

Sela I, Ashkenazy H, Katoh K, Pupko T. 2015. GUIDANCE2: accurate detection of unreliable alignment regions accounting for the uncertainty of multiple parameters. *Nucleic Acids Res* 43:W7–W14.

Šidák Z. 1967. Rectangular confidence regions for the means of multivariate normal distributions. *J Am Stat Assoc* 62:626–633.

Siu-Ting K, Torres-Sánchez M, San Mauro D, Wilcockson D, Wilkinson M, Pisani D, O'Connell MJ, Creevey CJ. 2019. Inadvertent paralog inclusion drives artifactual topologies and timetree estimates in phylogenomics. *Mol Biol Evol* 36:1344–1356.

731 Smith ML, Hahn MW. 2021a. New approaches for inferring phylogenies in the presence of
732 paralogs. *TIG* 37:174–187.

733 Smith ML, Hahn MW. 2021b. The frequency and topology of pseudoorthologs.
734 bioRxiv:2021.02.17.431499.

735 Springer MS, Meredith RW, Gatesy J, Emerling CA, Park J, Rabosky DL, Stadler T, Steiner C,
736 Ryder OA, Janečka JE, et al. 2012. Macroevolutionary dynamics and historical
737 biogeography of primate diversification inferred from a species supermatrix. *PLOS ONE*
738 7:e49521.

739 Swofford DL. 2001. Paup*: Phylogenetic analysis using parsimony (and other methods) 4.0. B5.

740 Thomas GWC, Dohmen E, Hughes DST, Murali SC, Poelchau M, Glastad K, Anstead CA,
741 Ayoub NA, Batterham P, Bellair M, et al. 2020. Gene content evolution in the
742 arthropods. *Genome Biol* 21:15.

743 Vachaspati P, Warnow T. 2015. ASTRID: Accurate species trees from internode distances. *BMC*
744 *Genomics* 16:S3.

745 van der Heijden RT, Snel B, van Noort V, Huynen MA. 2007. Orthology prediction at scalable
746 resolution by phylogenetic tree analysis. *BMC Bioinformatics* 8:83.

747 Van Dongen SM. 2000. Graph clustering by flow simulation.

748 Vanderpool D, Minh BQ, Lanfear R, Hughes D, Murali S, Harris RA, Raveendran M, Muzny
749 DM, Hibbins MS, Williamson RJ, et al. 2020. Primate phylogenomics uncovers multiple
750 rapid radiations and ancient interspecific introgression. *PLOS Biology* 18:e3000954.

751 Wang X, Lim BK, Ting N, Hu J, Liang Y, Roos C, Yu L. 2019. Reconstructing the phylogeny of
752 new world monkeys (platyrrhini): evidence from multiple non-coding loci. *Curr Zool*
753 65:579–588.

754 Wattam AR, Abraham D, Dalay O, Disz TL, Driscoll T, Gabbard JL, Gillespie JJ, Gough R, Hix
755 D, Kenyon R, et al. 2014. PATRIC, the bacterial bioinformatics database and analysis
756 resource. *Nucleic Acids Res* 42:D581–D591.

757 Willson J, Roddur MS, Liu B, Zaharias P, Warnow T. 2021. DISCO: Species tree inference
758 using multi-copy gene family tree decomposition. *Syst Biol* syab070.

759 Yan Z, Smith ML, Du P, Hahn MW, Nakhleh L. 2021. Species tree inference on data with
760 paralogs is accurate using methods intended to deal with incomplete lineage sorting. *Syst*
761 *Biol* syab056.

762 Yang Y, Smith SA. 2014. Orthology inference in nonmodel organisms using transcriptomes and
763 low-coverage genomes: Improving accuracy and matrix occupancy for phylogenomics.
764 *Mol Biol Evol* 31:3081–3092.

765 Zhang C, Rabiee M, Sayyari E, Mirarab S. 2018. ASTRAL-III: polynomial time species tree
766 reconstruction from partially resolved gene trees. *BMC Bioinformatics* 19:153.

767 Zhang C, Scornavacca C, Molloy EK, Mirarab S. 2020. ASTRAL-Pro: quartet-based species-tree
768 inference despite paralogy. *Mol Biol Evol* 37:3292–3307.

769 Zhao L, Saelao P, Jones CD, Begun DJ. 2014. Origin and spread of de novo genes in *Drosophila*
770 *melanogaster* populations. *Science* 343:769–772.