Scoliid wasp phylogenetics, evolution, and taxonomy and an exploration of the power of phylogenetic posterior predictive checks

By

ZIAD KHOURI
DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Entomology

in the

OFFICE OF GRADUATE STUDIES

of the

UNIVERSITY OF CALIFORNIA

DAVIS

Approved:

Lynn S. Kimsey, Chair

Philip S. Ward

Brian R. Moore

Committee in Charge

2022

Dedication

To my parents, Imad and Natalia Khouri, who instilled in me a love for science; my undergraduate life sciences mentors Pauline Aad, Nada El Ghossein-Maalouf, Tanos Hage, Doris Jaalouk, and Colette Kabrita Bou Serhal, who gave me an excellent education and prepared me to further pursue my research interests; and my instructors in philosophy, literature, astronomy, and physics, Edward Alam, Colette Guldimann, Roger Hajjar, Paul Jahshan, Naji Oueijan, Bassem Sabra, Doumit Salameh, Mary-Angela Willis, and Catherine Zoghbi, who shaped me into a better thinker and equipped me to appreciate their respective subjects in new ways.

Abstract

Scoliid wasps comprise a clade of aculeate insects whose larvae are parasitoids of scarabaeid beetle grubs. While scoliids have been studied and used as biological control agents, research into the group's evolution, as well as the stability of scoliid taxonomy, has been limited by a lack of reliable phylogenies. In Chapter 1, ultraconserved element (UCE) data are used under concatenation and the multispecies coalescent to infer a phylogeny of the Scoliidae. Data filtering experiments using posterior predictive checks and matched-pairs tests of symmetry are performed in order to mitigate potential issues arising from model misspecification. Analyses confirm the position of *Proscolia* as sister to all other extant scoliids. There is also strong support for a sister group relationship between the campsomerine genus *Colpa* and the Scoliini, rendering the Campsomerini non-monophyletic. Campsomerini excluding *Colpa* (hereafter Campsomerini *sensu stricto*) is inferred to be monophyletic, with the Australasian genus *Trisciloa* recovered as sister to the remaining members of the group. Out of nine genera in which more than one species was sampled, *Campsomeriella*, *Dielis*, *Megascolia*, and *Scolia* are inferred to be non-monophyletic. Analyses incorporating fossil data indicate an Early Cretaceous origin of the crown Scoliidae, with the split between Scoliini + *Colpa* and Campsomerini *s.s.* most probably occurring in the Late Cretaceous. Posterior means of Scoliini + *Colpa* and Campsomerini *s.s.* crown ages are estimated to be in the Paleogene, though age 95% HPD intervals extend slightly back past the K-Pg boundary, and analyses including fossils of less certain placement result in more posterior mass on older ages. Estimates of the stem ages of Nearctic scoliid clades are consistent with dispersal across Beringia during the Oligocene or later Eocene. This study provides a foundation for future research into scoliid wasp evolution and biogeography by being the first to leverage genome-scale data and model-based methods. However, the precision of dating analyses performed here is constrained by the paucity of well-

preserved fossils reliably attributable to the scoliid crown group. Despite concluding that the higher-level taxonomy of the Scoliidae is in dire need of revision, the chapter ends with the recommendation that taxonomic changes be predicated on datasets that extend the geographic and taxonomic sampling of the current study.

When used for phylogenetic inference, exonic DNA sequences can be coded in multiple ways, including as nucleotides, amino acids, and codons. In empirical studies, the choice of data type and associated model is often predicated on which model is less expected to be violated in ways that lead to inaccurate inference. Posterior predictive checks are one method for assessing the adequacy of phylogenetic models and potentially providing an indication of inference reliability. In Chapter 2, a simulation-based approach is used to explore how the ability to detect model inadequacy using phylogenetic posterior prediction, as well as the associated inference errors, may vary with data coding. Specifically, data were simulated under multiple models, including codon models featuring process heterogeneity across lineages, selection heterogeneity across sites, and selection for codon usage. Inference and posterior predictive checks were then performed under nucleotide and amino acid models from the GTR family. Some simulation conditions resulted in large differences, between amino acid and nucleotide treatments, in the ability to detect model violation, even when the magnitude of error in an estimate of interest was similar. Moreover, the results of other studies indicating that error in tree length estimation is not always correlated with error in topology reconstruction are corroborated. Although the use of amino acid models generally resulted in more accurate topologies, tree length errors were often greater than for nucleotide models when the data being analyzed were generated using branch-heterogeneous codon models. The results demonstrate that the magnitude and direction of tree length estimation error can depend on both data coding and properties of the data-generating process. Chapter 2 ends with the conclusion that if posterior predictive checks are to be used for

purposes such as data filtering, practical effect size thresholds indicative of low inference reliability must be established separately for amino acid and nucleotide data. Caution and careful selection of models and data coding are recommended when performing analyses where accurate inference of tree length is important.

Existing resources for the identification of Nearctic scoliid wasps have multiple shortcomings including limited geographic and taxonomic coverage, the use of outdated taxon names, and factual errors. Chapter 3 seeks to remedy the situation by providing a new key the Nearctic species. Additionally, molecular phylogenetic analysis and examination of morphological characters are used to demonstrate that specimens identified as *Scolia bicincta* using existing keys and commonly labeled as such in collections belong to two different species. One of these groups is sister to *Scolia dubia*. The other is sister to or conspecific with *Scolia mexicana*. Until the identity of the *Scolia bicincta* type is definitively established, the specimens related to *S. dubia* are treated as *S. bicincta*, and the specimens related to *S. mexicana* are treated as a geographic variant of *S. mexicana*.

Table of Contents

Chapter 1: The evolutionary history of mammoth wasps (Hymenoptera: Scoliidae)

Khouri, Z.[1], Gillung, J.P.[2], Kimsey, L.S.[1]

[1] Bohart Museum of Entomology, University of California, Davis, CA, U.S.A; [2] Lyman Entomological Museum, McGill University, Montreal, Quebec, Canada.

Introduction

Members of the family Scoliidae, sometimes referred to as mammoth wasps, are large fossorial aculeates that comprise one of the most visually striking and easily identifiable hymenopteran clades. The family has a cosmopolitan distribution and includes approximately 560 described species (Osten, 2005). Adult mammoth wasps feed primarily on nectar, with honeydew (Illingworth, 1921) and possibly pollen (Jervis, 1998) also reported as food sources. The larvae develop as ectoparasitoids on the larvae of scarabaeid beetles (Clausen, 1940). Some studies have highlighted interesting aspects of mammoth wasp natural history, such as parasitism of ant inquilines (Burmeister, 1854; Jonkman 1980), pseudocopulation with orchids (Jones & Gray, 1974; Ciotek *et al.*, 2006), fidelity of males to patrolling sites (Tani & Ueno, 2013), and efficient location of subterranean hosts (Inoue & Endo, 2008). Despite this, no study has attempted to reconstruct a phylogeny of the family, which precludes the examination of scoliid biology in an evolutionary context.

The lack of a solid phylogenetic hypothesis has also contributed to a lack of taxonomic clarity and stability. Day *et al.* (1981) referred to the group as "over-burdened nomenclatorially". Subsequently Argaman (1996), while describing the state of scoliid taxonomy as "disastrous", established a new subfamily, 21 new tribes, and 62 new genera without conducting a

phylogenetic analysis. In assembling a checklist of all scoliid species, Osten (2005) ignored the taxonomic changes implemented by Argaman and implicitly synonymized many of the new taxa by placing their type species in other groups (Elliott, 2011; Kimsey & Brothers, 2016). Currently, the need for a thorough taxonomic revision is recognized (Elliott, 2011).

A robust phylogeny is a prerequisite for studies of character evolution, diversification patterns over time, and biogeography, as well as for a natural taxonomy. In turn, the lack of a stable natural taxonomy hampers research by making species determination difficult and by impeding the communication and indexing of scientific information. In the case of mammoth wasps, this is especially apparent in the context of their use as agents for the biological control of scarabaeid pests (Illingworth, 1921; Wilson, 1960; DeBach, 1964). Misidentification of the control agent (for an example, see Elliott (2011) on research by the Queensland Bureau of Sugar Experiment Stations) precludes the repeatability of research and past biological control attempts and means that valuable information discovered in the process cannot easily be traced to the right organism (Rosen, 1986). This is particularly unfortunate, since a large portion of what is currently known about scoliid development, phenology, and host interaction was discovered while evaluating and using mammoth wasps for biological control (e.g. Illingworth, 1921; Miyagi, 1960). In the process of updating the BIOCAT database of introductions of biological control agents, Cock *et al.* (2016) listed Scoliidae among the groups requiring further taxonomic work.

In the present study, we aim to establish a solid foundation for research into mammoth wasp evolution and systematics. We use ultraconserved element (UCE) sequence data (Faircloth *et al.*, 2012; 2015) to infer scoliid phylogenetic trees using concatenation and under the multispecies coalescent (Rannala & Yang, 2003; Degnan & Rosenberg, 2009). Additionally, we leverage existing fossil data to estimate a timeline of scoliid evolution. To better understand potential

biases resulting from model misspecification, we perform data filtering experiments based on matched-pairs tests of symmetry (Bowker, 1948; Ababneh *et al*., 2006; Naser-Khdour *et al*., 2019) and assessments of model adequacy using data-based posterior predictive checks (Bollback, 2002; Huelsenbeck *et al.*, 2001; Doyle *et al.*, 2015).

Methods

Taxon and locus selection

We successfully sequenced 85 specimens of Scoliidae for this study. Taxon selection was aimed at maximizing taxonomic and biogeographic diversity within the limits imposed by the availability of material from which DNA could be extracted. All biogeographic realms are represented, but with weaker sampling in Australasia and the Neotropical and Palearctic regions. We also included previously published data (Johnson *et al*., 2013; Faircloth *et al*., 2015; Branstetter *et al*., 2017a; Peters *et al*., 2018) from six additional scoliid specimens. See Table S1.1 for specimen collection data and resources used for taxonomic determination.

Based on an examination of morphology, we suspected that *Scolia bicincta* may constitute two separate species. We therefore sequenced multiple individuals from each putative species. However, given the focus of the current study on reconstructing the scoliid phylogeny and identifying major clades rather than on species delimitation, we retained only two specimens following a preliminary phylogenetic analysis (see below).

We used the bradynobaenid genus *Apterogyna* as the only outgroup, and mined UCE sequences (see "Sequence quality control, assembly, and UCE identification" section below) from the partial genome published by Johnson *et al*. (2013). No sequences from other bradynobaenid taxa were publicly available, and we were unsuccessful in sequencing the specimens of

Bradynobaenus chubutinus to which we had access. Bradynobaenidae is well-supported as the sister group to Scoliidae (Johnson *et al*., 2013; Branstetter *et al*., 2017a; Peters *et al*., 2018). It is a species-poor clade, making it easier to avoid highly disproportionate taxon sampling, which would be difficult if ants or apoids were used. Adding more distant outgroups also increases the chance that heterogeneity in the evolutionary process across lineages results in more severe violations of homogeneous phylogenetic models.

We used the hymenoptera-v2 ant-specific probe set (Branstetter *et al*., 2017b) targeting 2524 UCEs and 12 nuclear genes ("legacy" markers).

Wet lab methods

We extracted DNA from pinned and ethanol-preserved specimens using QUIAGEN DNeasy Blood & Tissue Kits. Extractions were semi-nondestructive. In the case of pinned specimens, we first removed them from their pins. For most specimens, we made holes in the right side of the thorax using an insect pin, then soaked the specimen in lysis buffer overnight. We used the buffer, now containing DNA, for subsequent extraction steps. We then washed the specimens in 95% ethanol and either dried and remounted them or returned them to ethanol. For especially large specimens (e.g. of *Megascolia*) we only used a sample of thoracic muscle for extraction. For some medium-to-large specimens that are part of longer collection series, we separated the metasoma and the head from the mesosoma, and soaked the mesosoma in lysis buffer overnight. In some cases, quantities of extraction reagents used had to be proportionally adjusted to accommodate specimen size. Finally, we either reassembled the specimen for remounting, or mounted the parts on separate points on the same pin.

We prepared, enriched, and pooled libraries using the hymenoptera-v2 ant-specific probe set following the protocols of Faircloth *et al*. (2015) as modified for use at the Ward Ant Lab (Ward

& Branstetter, 2017). This was done in two separate batches. High-throughput sequencing was performed at the Huntsman Cancer Institute, University of Utah on an Illumina HiSeq 2500 platform (125 cycle paired-end) for the first batch and at the Novogene facility in Sacramento, CA on an Illumina HiSeq 4000 for the second batch.

Sequence quality control, assembly, and UCE identification

After receiving demultiplexed reads, we used three different bioinformatics pipelines for quality control and *de novo* assembly.

Pipeline A:

We performed quality-aware 3' adapter trimming with Scythe (https://github.com/vsbuffalo/scythe) version 0.991. This was followed with 5' adapter trimming with cutadapt (Martin, 2011) version 1.14 using a minimum overlap of 3 and an error tolerance of 0.16. We subsequently trimmed the reads with sickle (Joshi & Fass, 2011) version 1.33 using a quality threshold of 34 and a length threshold of 50. Assembly was done with Trinity (Grabherr *et al*., 2011) version 2.6.6 using a kmer size of 31. We also generated alternative assemblies with Velvet (Zerbino & Birney, 2008) version 1.2.10 and VelvetOptimiser (https://github.com/tseemann/VelvetOptimiser) version 2.2.4. However, the Velvet assemblies yielded significantly fewer UCE-containing contigs (data not shown, available upon request) and were not used for subsequent steps.

Pipeline B:

We used HTStream (https://github.com/s4hts/HTStream) version 1.1.0 for adapter and quality trimming. The HTStream pipeline consisted of the following steps: (1) calculating basic statistics on the raw reads with hts_Stats (2) screening for phiX with hts_SeqScreener, (3) removing

polyA/T sequences with hts_PolyATTrim with minimum size set to 100, (4) screening for adapter contamination with hts_SeqScreener using the i5 and i7 adapter sequences corresponding to each sample, with the kmer size set to 15, and with the percentage-hits argument set to 0.01, (5) a second round of adapter screening with hts_AdapterTrimmer, (6) quality-based 5' and 3' trimming with hts_QWindowTrim, (7) extracting the longest subsequences without "N"s using hts_NTrimmer with the minimum length set to 50, and finally (8) calculating statistics on the processed reads with hts_Stats. In order to speed up read processing, we wrote a python script that can run the pipeline in parallel on more than one sample if the number of available CPU cores is at least twice the number of steps in the pipeline.

We then assembled the reads with Spades (Bankevich *et al*., 2012) using a wrapper script from the phyluce package (Faircloth, 2016), version 1.6.8. Except for increasing allowed memory usage, settings were left at phyluce defaults.

Pipeline C:

We used Illumiprocessor (Faircloth, 2013), a wrapper around Trimmomatic (Bolger *et al*., 2014) and part of the phyluce package, for adapter and quality trimming. Spades was used for *de novo* assembly as in Pipeline B above.

For all pipelines, we used FastQC (Andrews, 2010) to evaluate reads before and after quality-control procedures.

We put reads from the first sequencing batch through Pipeline A and subsequently Pipeline B, while reads from the second sequencing batch were processed with Pipeline B and (with the exception of two samples) Pipeline C. In the case of ingroup taxa with previously published data (*Colpa sexmaculata*, *Colpa alcione*, *Proscolia* sp. EX568, *Scolia hirta*, *Scolia verticalis*, and

Scoliinae sp. EX577), we used the available assemblies and did not redo quality control and assembly. In all cases, we used the phyluce_assembly_match_contigs_to_probes, phyluce_assembly_get_match_counts, phyluce_assembly_get_fastas_from_match_counts, and phyluce_assembly_explode_get_fastas_file scripts to identify UCE-containing contigs and write them to fasta files for downstream analyses.

Pipelines A and B recovered similar numbers of UCEs per sample, although Pipeline B resulted in assemblies with higher N50 as calculated in QUAST (Gurevich *et al*., 2013) version 5.0.2 on both whole assemblies and assemblies filtered to UCE-containing contigs only. Pipelines B and C were close in terms of both number of recovered UCEs and N50. See Tables S1.2-1.3 for details. However, each pipeline recovered some UCEs that the other pipelines did not. Therefore, we combined the assemblies, choosing the longer contig in cases where a contig containing the same UCE was recovered in both assemblies. However, longer contigs may either represent genuine sequence or be the result of assembly errors. We visually inspected alignments prior to most downstream analyses to identify and remove misaligned sequences possibly originating from misassembly.

Due to low UCE yield from some samples in the second sequencing batch (likely due to failed enrichment) and concerns over contamination, we did the following to identify problematic samples: (1) selected loci that were represented by > 75% of taxa, (2) aligned sequences from those loci using MAFFT (Katoh & Standley, 2013), (3) edge-trimmed the alignments using the phyluce_align_get_trimmed_alignments_from_untrimmed script from phyluce, and (4) estimated a phylogeny (Fig. S1.1) using maximum likelihood (ML) with IQTREE (Minh *et al*., 2020; Hoang *et al*., 2018; Chernomor *et al*., 2016; Nguyen *et al*. 2015) version 2.0-rc2 while partitioning by locus and filtering out loci using a matched-pairs test of symmetry (Bowker,

1948; Ababneh *et al*., 2006; Naser-Khdour *et al*., 2019) designed to detect sequences whose evolution violates assumptions of stationarity, reversibility, and homogeneity (Jermiin *et al*., 2017). Thirteen taxa associated with suspected failed enrichments clustered together in two "clades" with very long branches, corroborating the spurious nature of the obtained sequences (Fig. S1.1). These taxa were not used in subsequent phylogenetic analyses and are not included in the counts under the taxon and locus selection section above.

In the case of *Apterogyna*, we mined UCE sequences from the partial genome of Johnson *et al*. (2013). We aligned UCE probes to the contigs using the phyluce_probe_run_multiple_lastzs_sqlite script from the phyluce package. We then extracted matching sequences in fasta format using the phyluce_probe_slice_sequence_from_genomes script, setting the flanking length to 700 bases.

Phylogenetic analysis

Unless otherwise indicated, we performed all multiple-sequence alignments using MAFFT version 7.407 with the E-INS-i algorithm (Altschul, 1998). Preliminary visual inspection of alignments confirmed that they often contain multiple conserved, well-aligned regions separated by ambiguously aligned regions. This better conforms to the assumptions behind the E-INS-i algorithm. L-INS-i (Gotoh, 1993), on the other hand, assumes a single, contiguous alignable region. All edge-trimming was done using the phyluce_align_get_trimmed_alignments_from_untrimmed script. All Bayesian phylogenetic analyses were performed using RevBayes (Höhna *et al*., 2014; 2016) version 1.0.12 unless otherwise indicated. Matched-pairs tests of symmetry in IQTREE refer specifically to MaxSymTest with a 0.05 p-value cutoff.

Analysis 1a:

We performed a preliminary run combining all (non-spurious) data from both sequencing batches, including all *Scolia bicincta* samples. This helped inform which *S. bicincta* samples to retain, as discussed below. We selected loci that had no more that 20% missing data at the site level (after including taxa without data) and estimated a phylogeny using ML with IQTREE while partitioning by locus and filtering out loci using matched-pairs tests of symmetry.

Analysis 1b:

We performed a second ML analysis with the goal of leveraging data from as many loci and taxa as possible while maintaining acceptable total levels of missing data. Given that analysis 1a indicated that samples of *S. bicincta* fall into two distinct clades that are sister to *S. dubia* and *S. mexicana* respectively (Fig. 1.1), we removed all but two *S. bicincta* samples (one from each putative species). In addition to phylogenetic position, the decision on which samples to retain was based on the number of recovered UCEs and on assembly quality statistics calculated using QUAST. We also removed *Scolia hirta* and Scoliinae sp. EX577, both from previously published studies, because they had very high fractions of missing data. After taxon removal, we redid alignment and edge-trimming. We then sorted loci by increasing fraction of missing data at the site level and progressively selected loci until the cumulative fraction of missing data reached 25% (1235 loci were selected at this point). After filtering using matched-pairs tests of symmetry in IQTREE, we retained 727 loci. We concatenated the alignments and selected a substitution and across-site rate variation (ASRV) model (from a pool of substitution models from the GTR (Tavaré, 1986) family and discretized gamma (Yang, 1994) and free-rates ASRV models) for each locus based on Bayesian Information Criterion (BIC) (Schwarz, 1978) scores. We then

estimated a phylogeny and performed 1000 ultrafast bootstrap replicates while leaving other IQTEE settings at default.

Analysis 1c:

In order to account for potential gene-tree-gene-tree conflict due to incomplete lineage sorting, we estimated species trees using the program ASTRAL-MP version 1.15.1 (Yin *et al.*, 2019). Starting with the same set of taxa used in analysis 1b, we redid alignment and edge-trimming, discarding alignments shorter than 600 bases. Given that highly fragmentary sequences can negatively affect accuracy (Sayyari *et al*., 2017), we subsequently removed taxa with more than 50% missing data and discarded alignments that retained fewer than 66 taxa. We then inferred gene trees using IQTREE with model selection settings similar to those in analysis 1b above while also performing matched-pairs tests of symmetry. We based subsequent species tree inference on three sets of gene trees: The first set contained trees corresponding to all loci, the second contained only trees from loci that failed the matched-pairs test of symmetry, and the last contained only trees from loci that passed.

Additionally, we estimated posterior distributions of gene trees in a Bayesian framework under the GTR+G model, followed by posterior predictive simulation (Bollback, 2002; Brown, 2014b; Doyle *et al.*, 2015; Höhna *et al.*, 2018) and calculation of posterior predictive p-values using two test statistics: multinomial likelihood (Goldman, 1993; Bollback, 2002) and chi-squared (Huelsenbeck *et al*., 2001; Foster 2004). Similarly to the ML-based analyses above, we then used different sets of maximum clade credibility (MCC) gene trees and gene tree posterior distribution samples (3000 trees per gene) for species tree inference with ASTRAL-MP. Using an alpha of 0.05 and the Bonferroni correction to account for multiple testing, we treated loci for which the posterior predictive p-value with either test statistic was < 0.025 as loci for which the model was

likely inadequate. This set included 922 of the total 954 loci. For each test statistic, we also split the loci into two sets, each respectively representing loci with the lowest and highest effect sizes for that statistic. Finally, we created another similar pair of sets but based on the Pythagorean sum of the effect sizes for both statistics. When using gene tree posterior distribution samples with ASTRAL, we performed bootstrapping using the -b option and set the number of replicates to 1000.

In datasets used for analyses 1a-c, *Apterogyna* and *Proscolia* have disproportionately high fractions of missing data (49% and 73% respectively) compared to other taxa. However, removing these taxa means the loss of the only outgroup. We therefore took a two-step approach: First, we performed an analysis (2a) only using loci with data available from both *Apterogyna* and *Proscolia* to minimize the potential impact of missing data on the inferred position of the root as well as on the placement of *Proscolia*. However, significantly cutting down the base dataset could result in loss of resolution in some parts of the tree. To address this, we performed another set of analyses (analyses 3a and 3b; see below) excluding *Apterogyna* and *Proscolia* as well as loci used in analysis 2a but conditioning on the position of the root inferred in analysis 2a. This allowed use of the remaining majority of the original data to resolve relationships within Scoliidae.

Analysis 2a:

We started with the same taxon set as for analysis 1b and selected aligned, trimmed fasta files corresponding to the 647 loci that have sequences from both *Apterogyna* and *Proscolia*. We used the biclustering algorithm of Uitert *et al*. (2008) as implemented in the R (Core R Team, 2020) package BicBin (https://github.com/TylerBackman/BicBin) to find large, dense biclusters of taxa and loci. We chose a set of 68 taxa and 484 loci with >99% completeness (presence or absence of

sequence for a given taxon and locus pair treated as a binary value). We then retrieved unaligned, untrimmed fasta files corresponding to the above loci and removed the taxa that are not part of the selected set. The sequences were then aligned and edge-trimmed. Given that the phylogenetic models we planned to use do not directly model indels (gaps are treated as missing data) and that unique indels are unlikely to contribute significant information, we removed all unique indels (i.e. columns where all taxa except one are represented by a gap) from the alignments. Calculating basic alignment statistics using AMAS (Borowiec, 2016) and visually inspecting the alignments in AliView (Larsson, 2014) revealed that *Apterogyna* sequences were (1) sometimes much shorter than those of other taxa for a given locus and (2) sometimes had poorly aligned sections. We therefore only retained alignments containing at least 500 non-ambiguous bases for both *Apterogyna* and *Proscolia*. We then manually trimmed alignment edges that contained no *Apterogyna* sequence and also trimmed any parts with suspected alignment uncertainty while discarding alignments that were poor throughout their length. Any alignments that became shorter than 300 bases were also discarded.

In order to assess model adequacy on the remaining 177 loci, we performed Bayesian phylogenetic analyses under the GTR+G model followed by posterior predictive simulation on each locus individually using the program RevBayes. We calculated the multinomial likelihood and chi-squared (as applied to nucleotide composition across taxa) test statistics and associated posterior predictive p-values and effect sizes on the empirical and simulated data using custom R code. For the purpose of filtering data for which the available model is suspected of being inadequate, one must choose some threshold. In advance of looking at the output, we decided to use an overall alpha of 0.05 and use the Bonferroni correction to account for multiple testing. We therefore discarded loci for which the posterior predictive p-value with either test statistic was < 0.025. We concatenated the remaining 31 alignments and used them for phylogeny estimation.

Each locus was assigned a separate GTR+G substitution model and tree length parameter (i.e. branch length multiplier), while a single vector of branch lengths drawn from a flat Dirichlet prior was shared among partitions. See used Rev scripts for further details. We assessed convergence for numerical parameters through visualization of posterior samples in Tracer (Rambaut *et al.*, 2018) version 1.7. For tree topologies, we made plots comparing posterior probabilities of splits across both runs using the bonsai (May & Moore, 2017) version 0.9 R package and calculated the Average Standard Deviation of Split Frequencies (ASDSF).

Analysis 2b:

Rasnitsyn (1993) identified only one fossil from Shangwang, Shandong, China as unequivocally belonging to the scoliid crown group. This fossil was attributed by Zhang (1989) to the extant species *Scolia prismatica*, currently in the genus *Megacampsomeris*. Yu *et al.* (2021) dated the Shanwang shale to approximately 18.5 Ma, in the early Miocene. Species described in later studies (Rasnitsyn & Martınez-Delclos, 1999; Nel *et al.*, 2013; Zhang *et al.*, 2015) are either connected to the crown Scoliidae by venation characters alone, or are of uncertain placement. This limits the information available to precisely estimate divergence times. Given this limitation and our inability to examine the *M. prismatica* specimen, we chose a conservative approach and estimated a broad timeline of scoliid evolution by calibrating the node representing the most recent common ancestor of Scoliidae and Bradynobaenidae using the age of *Protoscolia normalis*, a putative stem scoliid dated to approximately 125.5 Ma (Haichun *et al.*, 2002). We started with 177 processed alignments from analysis 2a (i.e. the state of the dataset after removal of short alignments but prior to filtering using posterior predictive checks). We then performed analyses on individual loci followed by posterior predictive simulation. We used a birth-death prior on tree topologies and node ages with a scaled beta prior on the root age (125.5 Ma

minimum age, 174.1 Ma maximum age, 132.5 Ma expected age, and a standard deviation of 5.5 Ma) and an uncorrelated lognormal relaxed clock model. See used Rev scripts for further details. After filtering loci in a similar manner to what was done in analysis 2a, we concatenated and analyzed the remaining 63 loci, adding rate multiplier parameters to allow the overall substitution rates to vary among loci.

In addition to the conservative primary analysis, we tested the effect of calibrating additional nodes using fossils of less certain placement. Although it is doubtful that the fossil described by Zhang (1989) belongs to an extant species, for the first additional analysis, we used it to set an 18.5 Ma minimum age (lognormal node age "prior" offset by 18.5, with a mean of 5.0 (mu ≈ 1.44) relative to the offset and a sigma of 0.587405) for the *Megacampsomeris* clade. For the second analysis, we used both the *Megacampsomeris* calibration above as well as a calibration of the scoliid crown group age based on *Araripescolia magnifica* (Nel *et al.*, 2013) (lognormal node age "prior" offset by 112.6 Ma, a mean of 10.0 relative to the offset and a sigma of 0.587405).

Analysis 2c:

In order to account for potential gene-tree-gene-tree conflict due to incomplete lineage sorting, we performed a species tree estimation analysis under the multispecies coalescent (e.g. Rannala & Yang, 2003; Degnan & Rosenberg, 2009) using the BEAST2 (Bouckaert *et al*., 2019) package STACEY (Jones, 2017). We used the same 63 loci from analysis 2b. Collapse weight was drawn from a beta prior with an alpha of 1.0 and a beta of 19.0 (mean 0.05, to reflect the belief that most samples are likely from distinct species). We used a lognormal prior on the popPriorScale parameter with a mean and standard deviation (in real space) of 1.0E-6 and 2.0 respectively. We enabled estimation of the relative death rate, which in this context corresponds to using a birth-death (as opposed to Yule) tree prior, and used a strict clock model. The site model was set to

GTR+G, unlinked among loci. We ran four independent chains and combined and summarized the output using the logcombiner and treeannotator tools packaged with BEAST2.

Analysis 2d:

We additionally performed species tree estimation using ASTRAL-MP. We used the same 177 starting loci from analysis 2a, but reran Bayesian gene tree estimation and posterior predictive simulation after removing taxa which had no data for a given locus. We then assembled sets of loci based on posterior predictive effect sizes in a manner similar to that in analysis 1c.

Analysis 3a:

In order to leverage more data to resolve relationships within Scoliidae, we set up an analysis that conditions on the position of the root inferred in analysis 1b while removing *Proscolia* and *Apterogyna* from the dataset. We followed a locus and taxon selection, alignment, and trimming procedure similar to that in analysis 2a. We chose a set of 72 taxa and 617 loci at 91% completeness from a pool of loci that excludes those used in analysis 1b. After discarding all alignments that, after trimming, were shorter than 300 bases or had more than 25% missing data at the site level, 469 alignments were retained. We did not trim alignments manually at this stage as the number of loci was large and the exclusion of *Apterogyna* and *Proscolia* improved alignment quality (assessed by visual inspection of a subset of alignments). We then ran Bayesian phylogenetic analyses followed by posterior predictive simulation on each individual alignment as in 2a. All alignments which passed filtering, as well as some that did not, were visually evaluated, and in a few cases problematic regions were manually trimmed. One locus was excluded due to very poor alignment. We reran posterior predictive tests on all alignments that have been altered. We then performed a concatenated analysis analogous to that in 2a, which

included all loci that passed filtering and were not subsequently edited and loci which were edited and subsequently passed filtering.

Analysis 3b:

The data processing and phylogenetic analysis procedures were analogous to those of analysis 3a, except we used a birth-death prior on trees and node ages (with no node calibration and with the root age arbitrarily fixed to 100 units) and an uncorrelated lognormal clock model.

Results

Sequence quality control, assembly, and UCE identification

Using pipeline A (Scythe + cutadapt + sickle + Trinity), we recovered 1941.9 UCE-containing contigs on average across specimens from batch 1, which is almost identical to the 1943.7 UCE-containing contigs recovered when using pipeline B (HTStream + Spades). However, the output of pipeline B had higher average N50 (2112.6 versus 1191.4, calculated from on-target contigs only) and a higher average number of UCE-containing contigs longer than 1000 bases (1429.0 versus 968.3).

The differences between outputs from pipelines B and C applied to specimens from batch 2 were in some ways less pronounced. The average number of UCE-containing contigs was 1726.3 and 1785.6 for pipelines B and C respectively, while average values for N50 were 1477.8 and 1482.5 respectively. The average number of on-target contigs longer than 1000 bases was 1036.0 for pipeline B and 1077.6 for pipeline C. When calculating these statistics, we excluded batch 2 samples for which we suspected failed enrichment (see corresponding section under Methods for details and Tables S1.1 and S1.2 for full QUAST statistics).

Overall, we recovered a total of 2495 UCE loci and an average of 1883.6 UCE loci per taxon across 91 taxa (including 6 taxa from previously published studies).

Phylogenetic analysis

Analysis 1a:

A total of 176 loci were retained after all filtering steps and used to estimate a phylogeny by maximum likelihood (Fig. 1.1). We recovered *Proscolia* as the sister group to all remaining Scoliidae, which correspond to the subfamily Scoliinae *sensu* Day *et al*. (1981). The tribe Scoliini is monophyletic. However, in contrast to the assumptions behind the current scoliid taxonomy (Osten, 2005), the genus *Colpa* was recovered as sister to the Scoliini, rendering the Campsomerini paraphyletic.

A clade represented by the scoliine genera *Megascolia*, *Pyrrhoscolia*, and *Carinoscolia* is sister to all other Scoliini, which in turn form three distinct groups. All New World members of the genus *Scolia* form a clade. We recovered *Scolia verticalis*, an Australasian species, as sister to the morphologically unusual Nearctic species *Triscolia ardens*. Given the unexpected nature of this pairing, we conducted an additional analysis (see Supporting Information for details) using (1) the "legacy" markers enriched from *T. ardens* as part of this study and from *S. verticalis* (from Faircloth *et al*. (2015), the source of *S. verticalis* UCE data used in this study), (2) corresponding Sanger data from the same specimen of *S. verticals* (Brady *et al.*, 2006; Ward & Fisher, 2016), and (3) corresponding Sanger data from different specimens of *T. ardens* (Pilgrim *et al*., 2008) and *S. verticalis* (Klopfstein & Ronquist, 2013). Sequences from the specimens used in this study grouped with their corresponding sequences from independent samples (Fig. S1.2), which makes contamination or data curation errors a less likely explanation for the relationship between *T.*

ardens and *S. verticalis* inferred here. All remaining sampled Scoliini form an Old World clade that is sister to the clade consisting of New World *Scolia* + (*T. ardens* + *S. verticalis*).

Samples of *Scolia bicincta* fall into two separate clades: one sister to *Scolia mexicana* and the other sister to *Scolia dubia*. This suggests the two groups belong to different species.

Campsomerini minus *Colpa* (provisionally referred to as Campsomerini *sensu stricto* from here on) is monophyletic. *Trisciloa saussurei* (not to be confused with members of the genus *Triscolia*) is inferred to be the sister taxon to the remaining Campsomerini *sensu stricto*. Within the latter group, all sampled New World taxa form a single clade. The closest relative of this New World clade is the Indomalayan taxon *Colpacampsomeris indica*, followed by a clade including the Afrotropical *Megameris soleata*, the Australiasian *Laevicampsomeris formosa*, and the Indomalayan genus *Megacampsomeris*. *Megacampsomeris* itself is recovered as monophyletic. Taxa occurring in Madagascar, such as *Micromeriella pilosella* and some *Campsomeriella*, have their closest affinities with Afrotropical taxa but do not form a monophyletic group.

Analysis 1b:

We used 727 loci from 76 taxa to reconstruct the tree in Fig. 1.2. The results are largely congruent with those from analysis 1a above, with the exception of the *Triscolia ardens* + *Scolia verticalis* group being recovered as sister to the Old World Scoliini (minus *Megascolia* + *Pyrrhoscolia* + *Carinoscolia*) as opposed to sister to the New World *Scolia*. *Colpa* is still recovered as sister to the Scoliini. The non-monophyly of *Dielis*, due to *Dielis pilipes* being more closely related to *Xanthocampsomeris* than to other *Dielis*, is likewise corroborated.

Analysis 1c:

For all analyses, the topology of the "main" ASTRAL tree (based only on ML or MCC gene trees) was effectively the same as the consensus topology estimated using gene tree posterior distributions and bootstrapping. Differences were limited to quadripartitions with very low support (e.g. 0.46 local posterior probability for most probable resolution, versus 0.35 for the next most probable alternative) or to relationships within species (e.g. *Dielis plumipes*).

The inferred topology based on ML trees from all loci (Fig. 1.3C) agrees with that from analysis 1b above. The topology based only on loci not failing the matched-pairs test of symmetry (Fig. 1.4, Fig. 1.3A) is identical, but with reduced support for the quadripartition involving *Megameris soleata*, *Laevicampsomeris formosa* + *Megacampsomeris*, *Colpacampsomeris indica* + New World Campsomerini, and the remaining Campsomerini. The topology inferred from loci failing the symmetry test (Fig. 1.3B) maintained high support for this quadripartition. On the other hand, the position of *Triscolia ardens* + *Scolia verticalis* became more uncertain, with 0.50 local posterior probability for the same placement as the other analyses above and 0.30 local posterior probability for *Triscolia ardens* + *Scolia verticalis* being sister to the New World *Scolia*.

Results of the analysis using MCC trees (as a way of summarizing tree posterior distributions) from all loci (Fig. 1.5D) agree with the ML-based results above with respect to the Campsomerini *sensu stricto*. However, the placement of *Triscolia ardens* + *Scolia verticalis* is not resolved, with 0.47 and 0.46 local posterior probability for a sister relationship with the sampled Old World *Scolia* and with the New World *Scolia* respectively. The ASTRAL tree based on loci with the lowest combined posterior predictive effect sizes (Fig. 1.5A) is similar to the tree above, with 0.46 local posterior probably in favor of (*Triscolia ardens* + *Scolia verticalis*) + New World *Scolia*, but a slightly lower probability (0.36) in favor *Triscolia ardens* + *Scolia verticalis*

being sister to the Old World *Scolia*. The analysis of loci with highest combined posterior predictive effect sizes (Fig. 1.5B) resulted in stronger (0.82 local posterior probability) support for the (*Triscolia ardens* + *Scolia verticalis*) + New World *Scolia* hypothesis. Unexpectedly, this relationship was likewise supported (0.87 local posterior probability) when using only the 32 loci for which the model was not found to be inadequate (Fig. 1.5C) using posterior predictive checks, but resolution within the Campsomerini was significantly reduced. Crucially, all analyses agree with respect to the placement of *Colpa* as sister to the Scoliini.

Analysis 2a:

The tree in Fig. 1.6 is the Maximum *A Posteriori* (MAP) tree summarized from two independent runs based on 31 loci for which the model was not found to be inadequate. The MCMC exhibited good convergence with respect to topology (see Fig. 1.7A for a comparison of split frequencies between runs). The average standard deviation of split frequencies was approximately 0.001.

This analysis places emphasis on reducing missing data in the outgroup and in *Proscolia*, removing poorly aligned sites, and reducing potential model violation at the expense of dataset size. Despite this, the tree backbone is fully resolved, with only a few shallow nodes having lower posterior probabilities. With respect to the position of the root, the results corroborate those from analyses 1a, 1b, and 1c: *Proscolia* is sister to the Scoliinae, *Colpa* is sister to the Scoliini, and Campsomerini *sensu stricto* is sister to Scoliini + *Colpa*, with Campsomerini in the traditional sense being non-monophyletic. The position of *Triscolia* as sister to an Old World scoliine clade is congruent with that in analysis 1b but not analysis 1a. *Scolia verticalis*, which was recovered as sister to *Tricolia ardens* in previous analyses, was not represented here and in subsequent analyses due to a high proportion of missing data. While *Colpacampsomeris indica* was likewise excluded from this analysis for the same reason, *Megameris soleata* is placed as

sister to the New World Campsomerini instead of being sister to *Laevicampsomeris* + *Megacampsomeris* as in analyses 1a and 1b.

Analysis 2b:

A total of 63 loci were retained post-filtering and used to construct a chronogram (Fig. 1.8). See Fig. 1.7B for a plot of split frequencies from two independent runs. While most clade posterior probabilities are close to 1 and none are lower than 0.94, node age credible intervals are broad due to only one calibration point being available. The crown Scoliini are inferred to have likely originated after the Cretaceous-Paleogene (K-Pg) extinction event. The mean estimated crown ages of Campsomerini *sensu stricto* and of Scoliini + *Colpa* are 49 million years (Ma) and 58 Ma respectively, although the associated 95% highest posterior density (HPD) intervals extend past the K-Pg boundary. The mean estimated age of crown Scoliinae is 84 Ma, with lower and upper bounds of the 95% HPD interval at 56 Ma and 107 Ma respectively. The crown Scoliidae as a whole (and thus the split between Proscoliinae and Scoliinae) has a 95% HPD age interval bounded by 96 Ma and 145 Ma, placing the likely origin of the group in the Early Cretaceous.

Results from the analyses including additional fossil calibrations (Fig. 1.9-1.10) were broadly congruent with the results above, but with greater ages estimated for most nodes after the Scoliinae/Proscoliinae split. When using both additional calibrations, the posterior distributions of ages for Campsomerini *sensu stricto* and of Scoliini + *Colpa* had means of 63 Ma and 69 Ma respectively, with more posterior mass on pre-K-Pg ages compared to the more conservative analysis above.

There are some topological differences between the results of these analyses and the tree from analysis 2a, mostly in the relationships of Old World *Scolia* and the position of *Megameris soleata* as sister to (*Laevicampsomeris* + *Megacampsomeris*) + New World *Campsomerini sensu*

stricto. However, both sets of analyses agree on the placement of *Colpa* as sister to the Scoliini and of *Triscolia ardens* as sister to the Old World *Scolia* clade.

Analysis 2c:

The species or minimal clusters (SMC) tree inferred under the multispecies coalescent using STACEY (Fig. 1.11-1.12) recovered many of the same major clades as the other analyses. However, some relationships, particularly those that had conflicting resolutions among the previous analyses, were poorly resolved. Specifically, while *Colpa* is still sister to a monophyletic Scoliini and the *Megascolia* + *Pyrrhoscolia* + *Carinoscolia* clade is sister to all other Scoliini, the position of *Triscolia ardens* within the latter group is uncertain. *Trisciloa* is still sister to all other members of Campsomerini *sensu stricto*, the New World members of which form a monophyletic group. *Megameris soleata*, *Laevicampsomeris* + *Megacampsomeris*, and the New World Campsomerini *sensu stricto* form a clade, though the relationships among them is uncertain. Likewise, the relationships among this clade, the *Cathimeris* + *Micromeriella* clade, and the *Campsomeriella* + *Tristimeris* clade are not resolved.

Analysis 2d:

The "main" ASTRAL topology, estimated using MCC trees only, was mostly congruent with the consensus topology, estimated using posterior samples and bootstrapping, in the case of the dataset with all loci (Fig. 1.13D) and of the dataset with loci having the highest-third combined posterior predictive effect sizes (Fig. 1.13C), with a few differences in the resolution of shallow nodes with low support. The dataset with loci having the lowest posterior predictive effect sizes showed somewhat bigger differences between the "main" (Fig. 1.13A) and bootstrap consensus (Fig. 1.13B) topologies, the "main" topology notably placing *Proscolia* as sister to the Campsomerini *sensus stricto*, albeit with low support.

The topology inferred using all loci mostly agrees with the results of analysis 2a, 2b, and 2c with the exception of *Megameris soleata* being inferred to be more closely related to *Laevicampsomeris* and *Megacampsomeris* than to the New World Campsomerini clade. Additionally, *Triscolia ardens* is placed as sister to the New World Scoliini, as opposed to being sister to the Old World *Scolia* clade as in analyses 2a and 2b and its position being unresolved as in analysis 2c. Analysis of the subset of loci with the highest combined posterior predictive effect sizes produced results almost identical to those based on all loci. Conversely, as reported above, using loci with the lowest posterior predictive effect sizes resulted in the unexpected placement of *Proscolia* as sister to Campsomerini *sensus stricto*. Relationships were otherwise similar to those inferred using other locus sets, but with lower local posterior probabilities associated with many quadripartitions.

Analyses 3a and 3b:

Analyses 3a and 3b are based on data from 115 and 159 loci respectively. The results (Fig. 1.14-1.15) agree with each other and mostly agree with those from analysis 1b. Differences include *Triscolia ardens* being sister to the Old World scoliine clade and *Megameris soleata* being sister to *Laevicampsomeris formosa*.

Discussion

Phylogenetic results and taxonomic implications

This is the first study to use molecular data to reconstruct the mammoth wasp phylogeny. Our results corroborate some long-standing phylogenetic hypotheses originally based on morphological data while contradicting others. Scoliid taxonomy has historically been unstable and confusing (see Elliott (2011) and Kimsey & Brothers (2016) for commentary). In the

following discussion, we use Osten (2005) as the reference for the current status of taxon names unless otherwise specified. We use Campsomerini *sensu stricto* to refer to Campsomerini excluding *Colpa* and taxa more closely related to *Colpa* than to the Scoliini.

The genus *Proscolia* was originally described by Rasnitsyn (1977), hypothesized to be sister to the remaining extant Scoliidae, and placed in a new subfamily Proscoliinae, with the other extant Scoliidae relegated to the Scoliinae. Day *et al.* (1981) and Osten (2005) maintained this arrangement and treated the former subfamilies Scoliinae and Campsomerinae as the scoliine tribes Scoliini and Campsomerini respectively (Fig. 1.16C). Notable exceptions to this approach include earlier works by Osten (1988, 1993), where he argued against the inclusion of *Proscolia* in the Scoliidae, and Argaman (1996), who radically revised the higher-level scoliid taxonomy without conducting an explicit phylogenetic analysis. Argaman elevated the Campsomerini (minus *Colpa* and its presumed close relatives) back to subfamily rank (Fig. 1.16D) and placed it as sister to the remaining extant Scoliidae (including the Proscoliinae). Pilgrim *et al.* (2008) included three scoliids in their study and placed *Proscolia* as either sister to the other two scoliids or as sister to Bradynobaenidae + other Scoliidae. Two more recent molecular phylogenetic studies of aculeates that included five and three scoliid species respectively (Debevec *et al.*, 2012; Branstetter *et al.*, 2017a) placed *Proscolia* as sister to all other scoliids. All analyses in the present study (Fig. 1.16E) strongly support this placement.

The taxonomic treatment of the species currently comprising the genus *Colpa* has historically varied significantly. To date, none of the taxonomic changes have been supported by phylogenetic analyses. However, the following authors generally presented informal phylogenetic arguments when making taxonomic decisions. Bradley (1950a), using the name *Campsoscolia* for the genus including what is now *Colpa* and *Dasyscolia*, argued for a "basal"

placement of these taxa, presumably meaning they fall outside the clade formed by the remaining Scoliidae (Fig. 1.16A). Betrem (1965) erected the tribe Trielini (emended by Betrem & Bradley (1972) to Trielidini) within the Campsomerinae (Fig. 1.16B) to contain the genera *Trielis* (corresponding to *Campsoscolia* as used by Bradley (1950a) and currently understood (Day *et al.*, 1981) to be a junior synonym of *Colpa*), *Crioscolia* (currently treated as a subgenus of *Colpa*), and *Guigliana*, which was formally described later by Bradley & Betrem (1967). Following the demotion of Campsomerinae to tribe rank by Day *et al.* (1981), *Colpa* and its allies were kept within the Campsomerini (Fig. 1.16C), with the implied relationships being Proscoliinae + (Campsomerini + Scoliini). Argaman (1996) on the other hand, created a new subfamily Colpinae (corresponding to the Trielidini of Betrem and Bradley (1972)) and placed it as sister to the Scoliini (which he elevated to subfamily rank), concluding that the Campsomerini *sensu stricto* (also elevated to subfamily rank) is sister to Proscoliinae + (Colpinae + Scoliinae) (Fig. 1.16D).

Debevec *et al.* (2012) included five scoliid species in their analyses, one of them being *Colpa sexmaculata*, but the main text contains no discussion of *Colpa* and the relationships within the Scoliidae. If we assume the monophyly of Campsomerini *sensu stricto* and of *Colpa* (each only represented by one species), the phylogenies included with the supporting information place Proscoliinae as sister to Campsomerini *sensu stricto* + (*Colpa* + Scoliini). All analyses in the current study agree with the latter hypothesis (Fig. 1.16E) while using a significantly larger dataset and attempting to mitigate the effects of non-randomly-distributed missing data and phylogenetic model violation.

In light of these results, morphological similarities between *Colpa* and the Campsomerini, such as the presence of an articulation between the basal and apical parts of the volsella and the

presence of the second recurrent vein, are likely plesiomorphies. We recommend the exclusion of *Colpa* from Campsomerini when a formal taxonomic revision of Scoliidae is undertaken. However, a phylogenetic analysis establishing the positions of *Guigliana* and *Dasyscolia* (not represented in this study) should be considered a prerequisite of such a revision. Both genera lack the transverse impressed impunctate band on the frons, which serves as the defining feature of *Colpa*, but share with *Colpa* and the Scoliini some mesothoracic characters (Bradley, 1950a; Betrem & Bradley, 1972). If *Guigliana* and *Dasyscolia* form a monophyletic group with *Colpa*, the establishment of a tribe Colpini may be justified. Otherwise, if they are more closely related to or nested within the Scoliini, it may be reasonable to transfer *Colpa*, *Guigliana*, and *Dasyscolia* to that tribe. More complete sampling of this group would also allow the evaluation of its subgeneric classification. The subgenus *Colpa* (*Crioscolia*) has a strongly disjunct distribution in both the New and Old World (Bradley, 1950a). Our results (Fig. 1.2) indicate the paraphyly of *Colpa* (*Colpa*): the Nearctic *Colpa* (*Colpa*) *octomaculata* is more closely related to the Nearctic *Colpa* (*Crioscolia*) *alcione* than it is to the Palearctic *Colpa* (*Colpa*) *sexmaculata*. In addition to allowing a critical evaluation of the phylogenetic validity of *Colpa* subgenera, a molecular phylogeny including more *Colpa* species would contribute significant biogeographic information, as this group appears to have undergone dispersal and/or vicariance events between the Old World and the Americas independently of the Scoliini and the Campsomerini *sensu stricto*.

Campsomerini *sans Colpa* is inferred to be monophyletic in all our analyses, with *Trisciloa* always sister to the remaining members of the group. Likewise, all sampled New World Campsomerini *sensu stricto* form a clade with high support in all analyses. *Colpacampsomeris indica* is consistently inferred to be the closest relative of this New World clade in all analyses in which the former was included. However, we have not sampled any species from South America,

so it remains unknown whether those share a closer relationship with the New World taxa sampled here or with Old World scoliids. *Dielis pilipes* groups with *Xanthocampsomeris* as opposed to with other *Dielis*. This is consistent with *D. pilipes* lacking some prominent morphological characteristics shared by other *Dielis*, such as a medial longitudinal furrow on the clypeus and a deep transverse furrow on the anterior of abdominal sternum II. Bradley (1964) states the opinion that *D. pilipes* should be excluded from *Dielis*, but this change was never formalized.

Megacampsomeris is always monophyletic in these analyses. Other consistently monophyletic groups include (1) *Micromeriella* with *Cathimeris* as the sister taxon and (2) the group consisting of *Campsomeriella*, *Tristimeris*, and some Malagasy species (undescribed or of uncertain taxonomic placement, provisionally labeled *Campsomeriella* sp. in the figures). Both *Tristimeris* and the Malagasy specimens are nested within *Campsomeriella*.

The positions of *Megameris* and *Laevicampsomeris* are uncertain, though they are likely more closely related to the New World Campsomerini, *Megacampsomeris*, and *Colpacampsomeris* than to other Campsomerini. In the current study, they are each represented by only one species. More thorough taxon sampling within these two genera will likely result in less uncertainty regarding their placement.

All analyses conducted here strongly support the monophyly of Scoliini. The first split within the Scoliini gives rise to two clades: one consisting of *Megascolia*, *Pyrrhoscolia*, and *Carinoscolia* and the other consisting of *Scolia* and *Triscolia*. *Megascolia* is consistently non-monophyletic in our analyses. The situation warrants a taxonomic revision, though it should ideally be informed by future phylogenetic studies that are able to sample *Megascolia*, *Pyrrhoscolia*, and

Carinoscolia more completely. Sequencing of multiple *Carinoscolia* species is especially important, given that the genus is suggested to be polyphyletic by Golfetti (2019).

Our sampling of New World species was restricted to the Nearctic, and the affinities of Neotropical scoliines thus remain uncertain. However, all sampled Nearctic *Scolia* form a single clade. The phylogenetic position of *Triscolia ardens* was inconsistent across our analyses. The genus *Triscolia* has a complicated taxonomic history (see Betrem & Bradley, 1964) and currently includes only two Nearctic species, *T. badia* and *T. ardens*. In all phylogenies where *Scolia verticalis* is included, *T. ardens* and *S. verticalis* are sisters. This is somewhat surprising given that *S. verticalis* is an Australasian species. We have mostly ruled out contamination and misidentification (see results section above) as potential explanations. More thorough sampling of scoliines from Australasia, Southeast Asia, and the eastern Palearctic might reveal species related to *S. verticalis* and fill in the gap in distributions, making a relationship with the Nearctic fauna more plausible. It is also possible that the two species of *Triscolia* are the only extant representatives of a previously more widespread lineage. The lack of close relatives of either species in the present study also means they are both subtended by long branches. A combination of the potential for long branch attraction and the disproportionately high fraction of missing data from *S. verticalis* raises the suspicion that the pairing might be artefactual. Regardless of its relationship to *S. verticalis*, *T. ardens* is recovered in our analyses either as closely related to the Nearctic *Scolia* clade or to the Old World *Scolia* clade, making it likely that the genus *Scolia* is paraphyletic irrespective of which placement of *T. ardens* is correct. One potential course of action is to synonymize *Triscolia* with *Scolia*. However, any taxonomic decisions involving *Scolia* should take into account the phylogenetic positions of two other large Scoliine genera, *Liacos* and *Austroscolia*, both of which are not represented in the current study.

Given the proliferation of scoliid generic names attached to groups defined mainly by superficial characters such as color and punctation, it seems likely that there are many examples of distinctive groups within larger genera being given their own generic names, thus rendering the larger genera paraphyletic. Further phylogenetic studies with more complete taxon sampling are needed before a taxonomic revision of scoliid genera is attempted. In the absence of such studies, we recommend proceeding cautiously when describing new species (such as those belonging to the Malagasy scoliid fauna) and avoiding the establishment of new genera or groups of higher rank without first conducting thorough phylogenetic analyses.

Divergence times and biogeography

The precision of node age estimates in the current study is limited by the small number of fossils that can be reliably attributed to the scoliid crown. It might be possible to slightly increase precision by conducting analyses with a broader phylogenetic scope. Including taxa from the Apoidea and Formicoidea could allow fossil data from those clades to inform overall rates of molecular evolution. However, apoids and formicoids being much more diverse than scoliids makes it difficult to sample species evenly across clades, and care must be taken to accommodate for this in any attempted analyses. The increased likelihood of heterogeneity in the evolutionary process becoming problematic as one expands the scope of the analysis should also be considered and addressed. Ultimately, the discovery and description of well-preserved crown fossils is likely to be a necessary prerequisite to achieving scoliid divergence time estimates with better precision and accuracy.

Due to weak sampling from some biogeographic regions, particularly Australasia and the Neotropics, we did not conduct a formal phylogeographic analysis. However, our phylogenetic

results do indicate some biogeographic patterns that could be further investigated in future studies.

We estimated the stem age of the Nearctic Campsomerini *sensu stricto* clade to be between 19 Ma and 46 Ma (95% HPD interval) when calibrating the root age only (Fig. 1.8). Among taxa sampled in this study, the closest relatives of this clade are taxa from Indomalaya, Australasia, and the eastern Palearctic. This suggests a possible exchange of fauna across Beringia during the Oligocene or later Eocene, which is broadly consistent with patterns observed in other animal groups (Jiang *et al.*, 2019). The Nearctic *Scolia* clade has a very similar estimated stem age (19-50 Ma). Analyses using additional (but less reliable, in terms of the phylogenetic placement of the associated fossil) calibrations extend the age 95% credible intervals into the early Eocene. Further refinement of node age estimates, in conjunction with more complete geographic sampling, is needed to evaluate the possibility of late (*c*. 65 Ma) exposures of the Thulean Route (Brikiatis, 2014) contributing to scoliid dispersal.

The phylogenetic position of *Triscolia* is uncertain, and has implications for the number and timing of biotic interchanges between North America and other regions. In addition to resolving the position of *Triscolia*, future phylogenetic studies need to prioritize sampling of the South American and Australasian scoliids. It is currently unclear whether South American Scoliini and Campsomerini *sensu stricto* each represent single lineages or multiple lineages with different biogeographic origins. It is possible that South America harbors relatively young lineages originating from Africa or the Nearctic and dispersing into South America during the Late-Early Eocene or later (Hoffmeister, 2020) and/or more ancient lineages with possible relationships to the Australian and African fauna. Understanding the phylogenetic and biogeographic affinities of South American scoliids, while interesting in itself, is also essential to understanding patterns of

scoliid diversification and answering questions such as why the Campsomerini *sensu stricto* are significantly more diverse than the Scoliini in the New World tropics while the opposite pattern holds in the Afrotropic and Indomalaya (Bradley, 1950b; 1959).

Madagascar is home to members of at least two campsomerine lineages, represented in the current study by one species of *Micromeriella* and several samples (probably from currently undescribed species) falling within the *Campsomeriella* clade. The presence of *M. pilosella* is probably due to a very recent dispersal from mainland Africa, while the Malagasy *Campsomeriella* lineage is older but also most closely related to African species. Given that the Malagasy scoliid fauna has received much less study than that of mainland Africa, it is certainly possible that among the species not sampled in this study there exist representatives of older endemic lineages that are not closely related to either *Micromeriella* or *Campsomeriella*. Our study additionally included two (probably undescribed) species belonging to *Scolia*. The scoliine genera *Liacos* and *Autroscolia* both have representatives on the African mainland, Madagascar, Asia, and Australia (Bradley, 1950b; Osten, 2005; Elliott 2011), while the morphologically distinctive *Mutilloscolia* is confined to Madagascar (Bradley, 1959). None are included in this study and their phylogenetic relationships to other scoliines remain poorly understood. Although it is possible that these genera could be nested within *Scolia* (which is mostly identified by lacking the defining characters of other genera), the apparent lack of morphological characters uniting them specifically with the *Scolia* species sampled here suggests that the current Malagasy scoliine diversity is likely a result of multiple dispersal (and possibly vicariance) events. This is tentatively supported by the morphology-based phylogenies of Golfetti (2019), which place *Austroscolia* and *Liacos* outside the clade formed by all scoliini sampled in this study.

Methodological considerations

Doyle *et al*. (2015) demonstrated the potential utility of filtering data using posterior predictive methods. We made use of a similar approach, albeit limiting it to data-based (Huelsenbeck *et al*., 2001) as opposed to inference-based (Brown, 2014a) tests. Molloy & Warnow (2018) used a simulation-based approach to explore the effect of data filtering using various criteria on species tree inference using ASTRAL (among other methods). They found that excluding loci with high gene tree estimation error can improve the accuracy of species tree inference when levels of incomplete lineage sorting (ILS) were moderate to low. The dependence on ILS levels was explained in terms of the number of gene trees required to accurately reconstruct the species tree increasing with higher levels of ILS. Thus, the negative effect of using fewer genes sometimes outweighed the positive effect of more accurate gene trees (Molloy & Warnow, 2018). In this context, we make the following observations based on our empirical analyses:

Using posterior predictive p-values with "conventional" cutoffs (e.g. 0.05) resulted in the exclusion of the majority of available loci. In some cases (e.g. Fig. 1.13A), this led to an unexpected and implausible species tree topology resulting from ASTRAL analyses (i.e. placement of *Proscolia* as sister to the Campsomerini). This could be a result of too few loci being used. Additionally, one would expect a correlation between the amount of data and the ability to detect model inadequacy, which might lead to the retention of less "informative" loci. This appears to be borne out in analysis 2d, where mean pairwise Robinson-Foulds distances among posterior topology samples were on average higher (73.3 versus 53.8) for the third of loci having the lowest posterior predictive effect sizes compared to the third having the highest. Under these circumstances, a fully-resolved point estimate of the topology might be a worse representation of the gene tree posterior distribution, and variance among gene tree point

estimates might be higher, even if there is no ILS and the underlying posterior distributions are unbiased. This could explain why we observed generally lower quadripartition support values resulting from analyses of loci with lower posterior predictive effect sizes even when the number of loci per analysis was kept constant (Fig. 1.5A, B; Fig. 1.13A, C). In contrast to Fig. 1.13A, a bootstrap-based ASTRAL species tree (Fig. 1.13B) that used samples from the posterior distributions of gene trees (as opposed to point estimates) recovered *Proscolia* in a more plausible position that is also corroborated by our STACEY analysis. Mirarab (2019) observed that using samples from gene tree posteriors does not have the same negative effect on species tree accuracy as does using gene tree bootstrap replicates in a ML framework. We concur with Mirarab (2019) that further investigation is warranted. Potential use cases for this hybrid approach could be datasets with both (1) a limited number of genes available (e.g. from Sanger data) where the accuracy of estimates using ASTRAL with gene-tree point estimates may be lower and (2) with a very large number of terminal taxa where a fully Bayesian approach (model-based coestimation of gene trees and species tree) may be more computationally challenging.

Conflicts of interest

The authors declare that there are no conflicts of interest.

Acknowledgments

We would like to thank M. Hauser (California Department of Food and Agriculture), S.L. Heydon (Bohart Museum of Entomology, University of California, Davis), and K. Williams (California Department of Food and Agriculture) for providing access to scoliid specimens at their respective institutions; B.E. Boudinot, M. Hauser, C. Parker, and T. Zavortink for giving

access to specimens in their personal collections; former members of the Ward Ant Lab, University of California, Davis (UC Davis) M. Borowiec, B.E. Boudinot, and M. Prebus and A. Abrieux from the Chiu Lab (UC Davis) for training the lead author in wet lab techniques; B. Moore (UC Davis) for mentoring the lead author in phylogenetic methods; G. Attardo, J. Bond, J. Chiu, B. Johnson, S. Nadler, and P.S. Ward for the use of their respective labs and equipment at UC Davis; L. Smith (Evolutionary Genetics Lab, Museum of Vertebrate Zoology, University of California, Berkeley) for providing access to a sonicator and training the lead author in its use; B. Boudinot and C. Pagan (Nadler Lab, UC Davis) for assistance with library preparation and enrichment; former FARM HPC cluster sysadmins B. Broadley and T. Thatcher for help with using the cluster; current and former members of the Ward Ant Lab M. Borowiec, B.E. Boudinot, Z. Griebenow, Z. Lieberman, J. Oberski, M. Prebus, and P.S. Ward and of the Moore Lab (UC Davis) E. Espejo, J. Gao, M.R. May, and B. Moore for helpful and insightful discussion; N. Tam for proofreading and providing comments on the manuscript.

References

Ababneh, F., Jermiin, L. S., Ma, C., & Robinson, J. (2006). Matched-pairs tests of homogeneity with applications to homologous nucleotide sequences. *Bioinformatics*, *22*(10), 1225-1231.

Altschul, S. F. (1998). Generalized affine gap costs for protein sequence alignment. *Proteins: Structure, Function, and Bioinformatics*, *32*(1), 88-96.

Andrews S. (2010). FastQC: a quality control tool for high throughput sequence data. URL: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Argaman, Q. (1996). Generic synopsis of Scoliidae (Hymenoptera, Scolioidea). *Annales Historico Naturales-Musei Nationalis Hungarici*, *88*, 171-222.

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., ... & Pevzner, P. A. (2012). SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *Journal of computational biology*, *19*(5), 455-477.

Betrem, J. G. (1965). The African Scoliids and their Affinities. In *XIIth International Congress of Entomology* (p. 120).

Betrem, J. G., & Bradley, J. C. (1964). Annotations on the genera *Triscolia*, *Megascolia* and *Scolia* (Hymenoptera, Scoliidae). *Zoologische Mededelingen*, *39*(43), 433-444.

Betrem J. G. & Bradley J. C. (1972) *The African Campsomerinae (Hymenoptera, Scoliidae)*. Nederlandse Entomologische Vereniging, Amsterdam.

Bollback, J. P. (2002). Bayesian model adequacy and choice in phylogenetics. *Molecular Biology and Evolution*, *19*(7), 1171-1180.

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, *30*(15), 2114-2120.

Borowiec, M. L. (2016). AMAS: a fast tool for alignment manipulation and computing of summary statistics. *PeerJ*, *4*, e1660.

Bouckaert, R., Vaughan, T. G., Barido-Sottani, J., Duchêne, S., Fourment, M., Gavryushkina, A., ... & Drummond, A. J. (2019). BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis. *PLoS computational biology*, *15*(4), e1006650.

Bowker, A. H. (1948). A test for symmetry in contingency tables. *Journal of the american statistical association*, *43*(244), 572-574.

Bradley, J. C. (1950a). The most primitive Scoliidae. *Revista Española de Entomología (tomo extraordinario)*, *1*, 427-438.

Bradley, J. C. (1950b). The Ethiopian Scoliidae. In *Eighth International Congress of Entomology* (p. 1).

Bradley, J. C. (1959). The Scoliidae of Africa. *Annals of the Transvaal Museum*, *23*(4), 331-362.

Bradley, J. C. (1964). Further notes on the American taxa of *Campsomeris* (Hymenoptera: Scoliidae). *Entomological News*, *25*, 101-108.

Bradley, J. C., & Betrem, J. G. (1967). The types of the Scoliidae described by Frederick Smith (Hymenoptera). *Bulletin of The British Museum (Natural History) Entomology*, *20*(7), 289.

Brady, S. G., Schultz, T. R., Fisher, B. L., & Ward, P. S. (2006). Evaluating alternative hypotheses for the early evolution and diversification of ants. *Proceedings of the National Academy of Sciences*, *103*(48), 18172-18177.

Branstetter, M. G., Danforth, B. N., Pitts, J. P., Faircloth, B. C., Ward, P. S., Buffington, M. L., ... & Brady, S. G. (2017a). Phylogenomic insights into the evolution of stinging wasps and the origins of ants and bees. *Current Biology*, *27*(7), 1019-1025.

Branstetter, M. G., Longino, J. T., Ward, P. S., & Faircloth, B. C. (2017b). Enriching the ant tree of life: Enhanced UCE bait set for genome-scale phylogenetics of ants and other Hymenoptera. *Methods in Ecology and Evolution*, *8*(6), 768-776.

Brikiatis, L. (2014). The De Geer, Thulean and Beringia routes: key concepts for understanding early Cenozoic biogeography. *Journal of Biogeography*, *41*(6), 1036-1054.

Brown, J. M. (2014a). Detection of implausible phylogenetic inferences using posterior predictive assessment of model fit. *Systematic biology*, *63*(3), 334-348.

Brown, J. M. (2014b). Predictive approaches to assessing the fit of evolutionary models. *Systematic biology*, *63*(3), 289-292.

Burmeister, H. (1854). Bemerkungen über den allgemeinen Bau und die Geschlechtsunterschiede bei den Arten der Gattung *Scolia* Fabr. *Abhandlungen der Naturforschenden Gesellschaft zu Halle*, *1*(4), 1-46.

Chernomor, O., Von Haeseler, A., & Minh, B. Q. (2016). Terrace aware data structure for phylogenomic inference from supermatrices. *Systematic biology*, *65*(6), 997-1008.

Ciotek, L., Giorgis, P., Benitez-Vieyra, S., & Cocucci, A. A. (2006). First confirmed case of pseudocopulation in terrestrial orchids of South America: pollination of Geoblasta pennicillata (Orchidaceae) by Campsomeris bistrimacula (Hymenoptera, Scoliidae). *Flora-Morphology, Distribution, Functional Ecology of Plants*, *201*(5), 365-369.

Clausen, C. P. (1940). *Entomophagous insects*. McGraw-Hill book Company, Incorporated.

Cock, M. J., Murphy, S. T., Kairo, M. T., Thompson, E., Murphy, R. J., & Francis, A. W. (2016). Trends in the classical biological control of insect pests by insects: an update of the BIOCAT database. *BioControl*, *61*(4), 349-363.

Core R Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL: https://www.R-project.org/

Day, M. C., Else, G. R., & Morgan, D. (1981). The most primitive scoliidae (Hymenoptera). *Journal of natural History*, *15*(4), 671-684.

DeBach, P. (1964). *Biological control of insect pests and weeds*. Reinhold Publishing Corporation, New York.

Debevec, A. H., Cardinal, S., & Danforth, B. N. (2012). Identifying the sister group to the bees: a molecular phylogeny of Aculeata with an emphasis on the superfamily Apoidea. *Zoologica scripta*, *41*(5), 527-535.

Degnan, J. H., & Rosenberg, N. A. (2009). Gene tree discordance, phylogenetic inference and the multispecies coalescent. *Trends in ecology & evolution*, *24*(6), 332-340.

Doyle, V. P., Young, R. E., Naylor, G. J., & Brown, J. M. (2015). Can we identify genes with increased phylogenetic reliability?. *Systematic biology*, *64*(5), 824-837.

Elliott, M. G. (2011). Annotated catalogue of the Australian Scoliidae (Hymenoptera). *Technical Reports of the Australian Museum, Online*, *22*, 1-17.

Faircloth, B. C. (2013) Illumiprocessor: a trimmomatic wrapper for parallel adapter and quality trimming. URL: http://dx.doi.org/10.6079/J9ILL

Faircloth, B. C. (2016). PHYLUCE is a software package for the analysis of conserved genomic loci. *Bioinformatics*, *32*(5), 786-788.

Faircloth, B. C., Branstetter, M. G., White, N. D., & Brady, S. G. (2015). Target enrichment of ultraconserved elements from arthropods provides a genomic perspective on relationships among H ymenoptera. *Molecular ecology resources*, *15*(3), 489-501.

Faircloth, B. C., McCormack, J. E., Crawford, N. G., Harvey, M. G., Brumfield, R. T., & Glenn, T. C. (2012). Ultraconserved elements anchor thousands of genetic markers spanning multiple evolutionary timescales. *Systematic biology*, *61*(5), 717-726.

Foster, P. G. (2004). Modeling compositional heterogeneity. *Systematic biology*, *53*(3), 485-495.

Goldman, N. (1993). Statistical tests of models of DNA substitution. *Journal of molecular evolution*, *36*(2), 182-198.

Golfetti, I. F. (2019). *Análise filogenética de Scolia Fabricius (Hymenoptera, Scoliidae, Scoliinae)*. [Unpublished masters thesis], Instituto de Biociências, Letras e Ciências Exatas. *See http://hdl.handle.net/11449/181962*

Gotoh, O. (1993). Optimal alignment between groups of sequences and its application to multiple sequence alignment. *Bioinformatics*, *9*(3), 361-370.

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., ... & Regev, A. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature biotechnology*, *29*(7), 644.

Gurevich, A., Saveliev, V., Vyahhi, N., & Tesler, G. (2013). QUAST: quality assessment tool for genome assemblies. *Bioinformatics*, *29*(8), 1072-1075.

Haichun, Z., Rasnitsyn, A. P., & Junfeng, Z. (2002). The oldest known scoliid wasps (Insecta, Hymenoptera, Scoliidae) from the Jehol biota of western Liaoning, China. *Cretaceous Research*, *23*(1), 77-86.

Hoang, D. T., Chernomor, O., Von Haeseler, A., Minh, B. Q., & Vinh, L. S. (2018). UFBoot2: improving the ultrafast bootstrap approximation. *Molecular biology and evolution*, *35*(2), 518-522.

Hoffmeister, M. F. C. (2020). From Gondwana to the Great American Biotic Interchange: The Birth of South American Fauna. In *Pilauco: A Late Pleistocene Archaeo-paleontological Site* (pp. 13-32). Springer, Cham.

Höhna, S., Coghill, L. M., Mount, G. G., Thomson, R. C., & Brown, J. M. (2018). P3: Phylogenetic posterior prediction in RevBayes. *Molecular biology and evolution*, *35*(4), 1028-1034.

Höhna, S., Heath, T. A., Boussau, B., Landis, M. J., Ronquist, F., & Huelsenbeck, J. P. (2014). Probabilistic graphical model representation in phylogenetics. *Systematic biology*, *63*(5), 753-771.

Höhna, S., Landis, M. J., Heath, T. A., Boussau, B., Lartillot, N., Moore, B. R., ... & Ronquist, F. (2016). RevBayes: Bayesian phylogenetic inference using graphical models and an interactive model-specification language. *Systematic biology*, *65*(4), 726-736.

Huelsenbeck, J. P., Ronquist, F., Nielsen, R., & Bollback, J. P. (2001). Bayesian inference of phylogeny and its impact on evolutionary biology. *science*, *294*(5550), 2310-2314.

Illingworth, J. F. (1921). Natural enemies of sugar-cane beetles in Queensland. *Queensland Bureau Sugar Experimental Station Division of Entomology Bulletin*, *13*, 1-47.

Inoue, M., & Endo, T. (2008). Below-ground host location by Campsomeriella annulata (Hymenoptera: Scoliidae), a parasitoid of scarabaeid grubs. *Journal of ethology*, *26*(1), 43.

Jermiin, L. S., Jayaswal, V., Ababneh, F. M., & Robinson, J. (2017). Identifying optimal models of evolution. In *Bioinformatics* (pp. 379-420). Humana Press, New York, NY.

Jervis, M. (1998). Functional and evolutionary aspects of mouthpart structure in parasitoid wasps. *Biological Journal of the Linnean Society*, *63*(4), 461-493.

Jiang, D., Klaus, S., Zhang, Y. P., Hillis, D. M., & Li, J. T. (2019). Asymmetric biotic interchange across the Bering land bridge between Eurasia and North America. *National Science Review*, *6*(4), 739-745.

Johnson, B. R., Borowiec, M. L., Chiu, J. C., Lee, E. K., Atallah, J., & Ward, P. S. (2013). Phylogenomics resolves evolutionary relationships among ants, bees, and wasps. *Current Biology*, *23*(20), 2058-2062.

Jones, D. L., & Gray, B. (1974). The pollination of *Calochilus holtzei* F. Muell. *American Orchid Society Bulletin*, *43*, 604-606.

Jones, G. (2017). Algorithmic improvements to species delimitation and phylogeny estimation under the multispecies coalescent. *Journal of mathematical biology*, *74*(1-2), 447-467.

Jonkman, J. C. M. (1980). The external and internal structure and growth of nests of the leaf-cutting ant Atta vollenweideri Forel, 1893 (Hym.: Formicidae) Part I 1. *Zeitschrift für angewandte Entomologie*, *89*(1-5), 158-173.

Joshi, N. A., & Fass, J. N. (2011). Sickle: A sliding-window, adaptive, quality-based trimming tool for FastQ files (Version 1.33). URL: https://github.com/najoshi/sickle

Katoh, K., & Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular biology and evolution*, *30*(4), 772-780.

Kimsey, L. S., & Brothers, D. J. (2016). The life, publications and new taxa of Qabir Argaman (Carol Nagy). *Journal of Hymenoptera Research*, *50*(50), 141-178.

Klopfstein, S., & Ronquist, F. (2013). Convergent intron gains in hymenopteran elongation factor-1α. *Molecular phylogenetics and evolution*, *67*(1), 266-276.

Larsson, A. (2014). AliView: a fast and lightweight alignment viewer and editor for large datasets. *Bioinformatics*, *30*(22), 3276-3278.

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. journal*, *17*(1), 10-12.

May, M. R. & Moore, B. R. (2017). bonsai: Automated phylogenetic MCMC diagnosis. R package version 0.9. URL: https://github.com/mikeryanmay/bonsai

Minh, B. Q., Schmidt, H. A., Chernomor, O., Schrempf, D., Woodhams, M. D., Von Haeseler, A., & Lanfear, R. (2020). IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era. *Molecular biology and evolution*, *37*(5), 1530-1534.

Mirarab, S. (2019). Species tree estimation using ASTRAL: practical considerations. *arXiv preprint arXiv:1904.03826*.

Miyagi, I. (1960). Ecological studies of some Japanese species of Scoliidae. *Transactions of the Shikoku Entomological Society*, *6*, 104-120.

Molloy, E. K., & Warnow, T. (2018). To include or not to include: the impact of gene filtering on species tree estimation methods. *Systematic Biology*, *67*(2), 285-303.

Naser-Khdour, S., Minh, B. Q., Zhang, W., Stone, E. A., & Lanfear, R. (2019). The prevalence and impact of model violations in phylogenetic analysis. *Genome biology and evolution*, *11*(12), 3341-3352.

Nel, A., Escuillie, F., & Garrouste, R. (2013). A new scoliid wasp in the Early Cretaceous Crato Formation in Brazil (Hymenoptera: Scoliidae). *Zootaxa*, *3717*, 395-400.

Nguyen, L. T., Schmidt, H. A., Von Haeseler, A., & Minh, B. Q. (2015). IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular biology and evolution*, *32*(1), 268-274.

Osten, T. (1988). Die Mundwerkzeuge von *Proscolia spectator* Day (Hymenoptera: Aculeata). Ein Beitrag zur Phylogenie der "Scolioidea". *Stuttgarter Beiträge zur Naturkunde*, *415*, 1-30.

Osten, T. (1993). The enigmatic genus Proscolia, its distribution and phylogenetic relationships (Insecta, Hymenoptera). *Biologia Gallo-hellenica*, *20*(1), 177-182.

Osten, T. (2005). Checkliste der Dolchwespen der Welt (Insecta: Hymenoptera, Scoliidae). *Bericht der Naturforschenden Gesellschaft Augsburg*, *62*, 1-62.

Peters, R. S., Niehuis, O., Gunkel, S., Bläser, M., Mayer, C., Podsiadlowski, L., ... & Krogmann, L. (2018). Transcriptome sequence-based phylogeny of chalcidoid wasps (Hymenoptera: Chalcidoidea) reveals a history of rapid radiations, convergence, and evolutionary success. *Molecular Phylogenetics and Evolution*, *120*, 286-296.

Pilgrim, E. M., Von Dohlen, C. D., & Pitts, J. P. (2008). Molecular phylogenetics of Vespoidea indicate paraphyly of the superfamily and novel relationships of its component families and subfamilies. *Zoologica Scripta*, *37*(5), 539-560.

Rambaut, A., Drummond, A. J., Xie, D., Baele, G., & Suchard, M. A. (2018). Posterior summarization in Bayesian phylogenetics using Tracer 1.7. *Systematic biology*, *67*(5), 901.

Rannala, B., & Yang, Z. (2003). Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci. *Genetics*, *164*(4), 1645-1656.

Rasnitsyn AP. (1977). New subfamily of scoliid wasps (Hymenoptera, Scoliidae, Proscoliinae). *Zoologicheskii zhurnal*, *66*, 522-599.

Rasnitsyn, A. P. (1993). Archaeoscoliinae, an extinct subfamily of scoliid wasps (Insecta: Vespida = Hymenoptera: Scoliidae). *Journal of Hymenoptera Research*, *2*(1), 85-96.

Rasnitsyn, A. P., & Martınez-Delclos, X. (1999). New Cretaceous Scoliidae (Vespida = Hymenoptera) from the Lower Cretaceous of Spain and Brazil. *Cretaceous Research*, *20*(6), 767-772.

Rosen, D. (1986). The role of taxonomy in effective biological control programs. *Agriculture, Ecosystems & Environment*, *15*(2-3), 121-129.

Sayyari, E., Whitfield, J. B., & Mirarab, S. (2017). Fragmentary gene sequences negatively impact gene tree and species tree reconstruction. *Molecular biology and evolution*, *34*(12), 3279-3291.

Schwarz, G. (1978). Estimating the dimension of a model. *The annals of statistics*, *6*(2), 461-464.

Tani, S., & Ueno, T. (2013). Site fidelity and long-distance homing by males of solitary parasitic wasps (Hymenoptera: Scoliidae). *The Canadian Entomologist*, *145*(3), 333-337.

Tavaré, S. (1986). Some probabilistic and statistical problems in the analysis of DNA sequences. *Lectures on mathematics in the life sciences*, *17*(2), 57-86.

Uitert, M. V., Meuleman, W., & Wessels, L. (2008). Biclustering sparse binary genomic data. *Journal of Computational Biology*, *15*(10), 1329-1345.

Ward, P. S., & Branstetter, M. G. (2017). The acacia ants revisited: convergent evolution and biogeographic context in an iconic ant/plant mutualism. *Proceedings of the Royal Society B: Biological Sciences*, *284*(1850), 20162569.

Ward, P. S., & Fisher, B. L. (2016). Tales of dracula ants: the evolutionary history of the ant subfamily Amblyoponinae (Hymenoptera: Formicidae). *Systematic Entomology*, *41*(3), 683-693.

Wilson, F. (1960). A review of the biological control of insects and weeds in Australia and Australian New Guinea. *Commonwealth Agricultural Bureaux*.

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

Yin, J., Zhang, C., & Mirarab, S. (2019). ASTRAL-MP: scaling ASTRAL to very large datasets using randomization and parallelization. *Bioinformatics*, *35*(20), 3961-3969.

Yu, J., Pang, X., Fu, W., Hilton, J., Liang, M., Jiang, Z., ... & Zhang, R. (2021). A high-resolution timescale for the Miocene Shanwang diatomaceous shale lagerstätte (China): development of Wavelet Scale Series Analysis for cyclostratigraphy. *Geosciences Journal*, *25*, 561-574.

Zhang, J. (1989). Fossil insects from Shanwang, Shandong, China. Jinan: Shandong Science and Technology Publishing House. [in Chinese].

Zhang, Q., Zhang, H., Rasnitsyn, A. P., & Jarzembowski, E. A. (2015). A new genus of Scoliidae (Insecta: Hymenoptera) from the Lower Cretaceous of northeast China. *Cretaceous Research*, *52*, 579-584.

Zerbino, D. R., & Birney, E. (2008). Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome research*, *18*(5), 821-829.

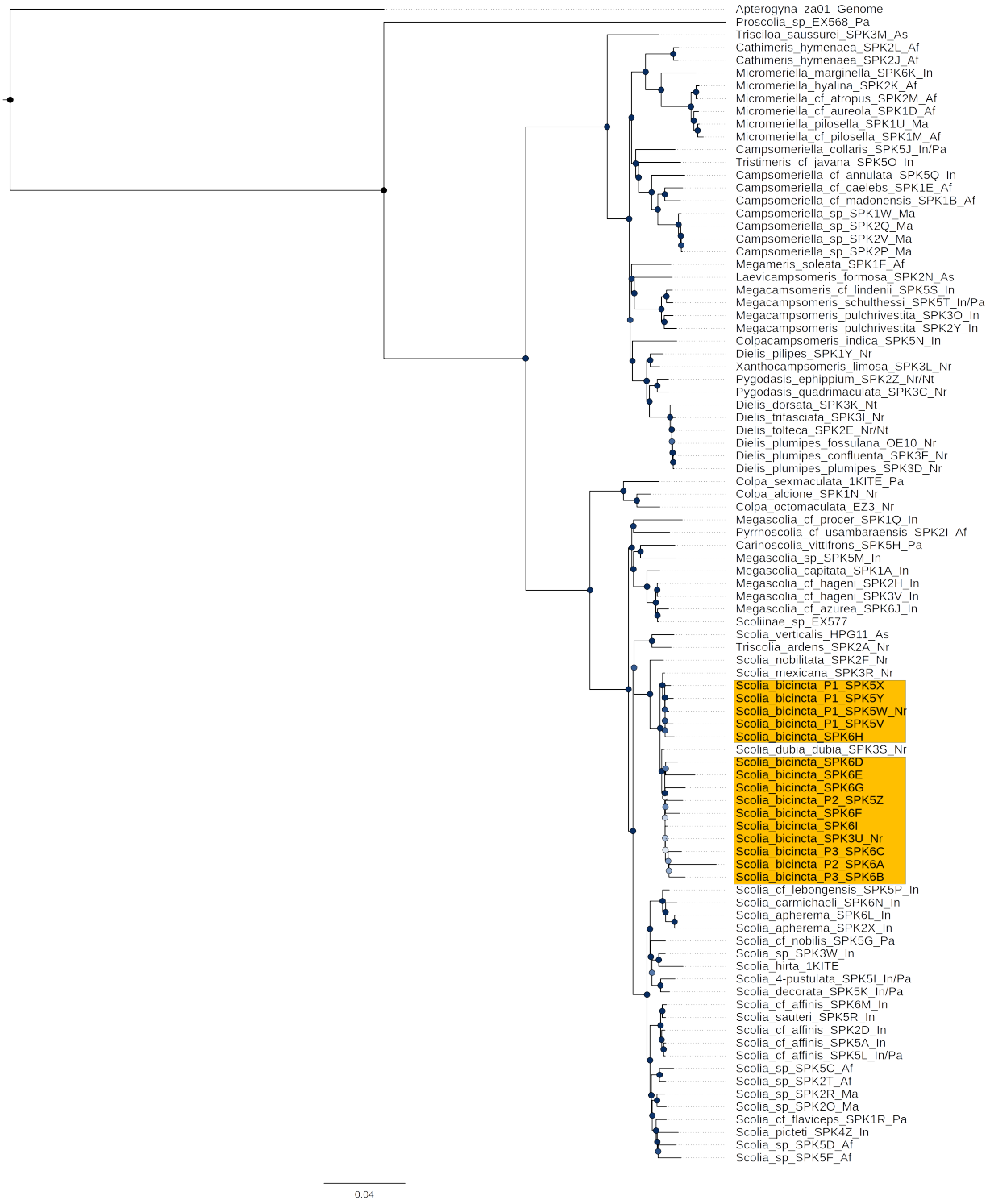

Figure 1.1 Maximum likelihood phylogeny including all samples (analysis 1a). Two distinct clades of *Scolia bicincta* are highlighted in yellow. Node support values are based on Ultrafast Bootstrapping in IQTREE; darker nodes reflect higher support.

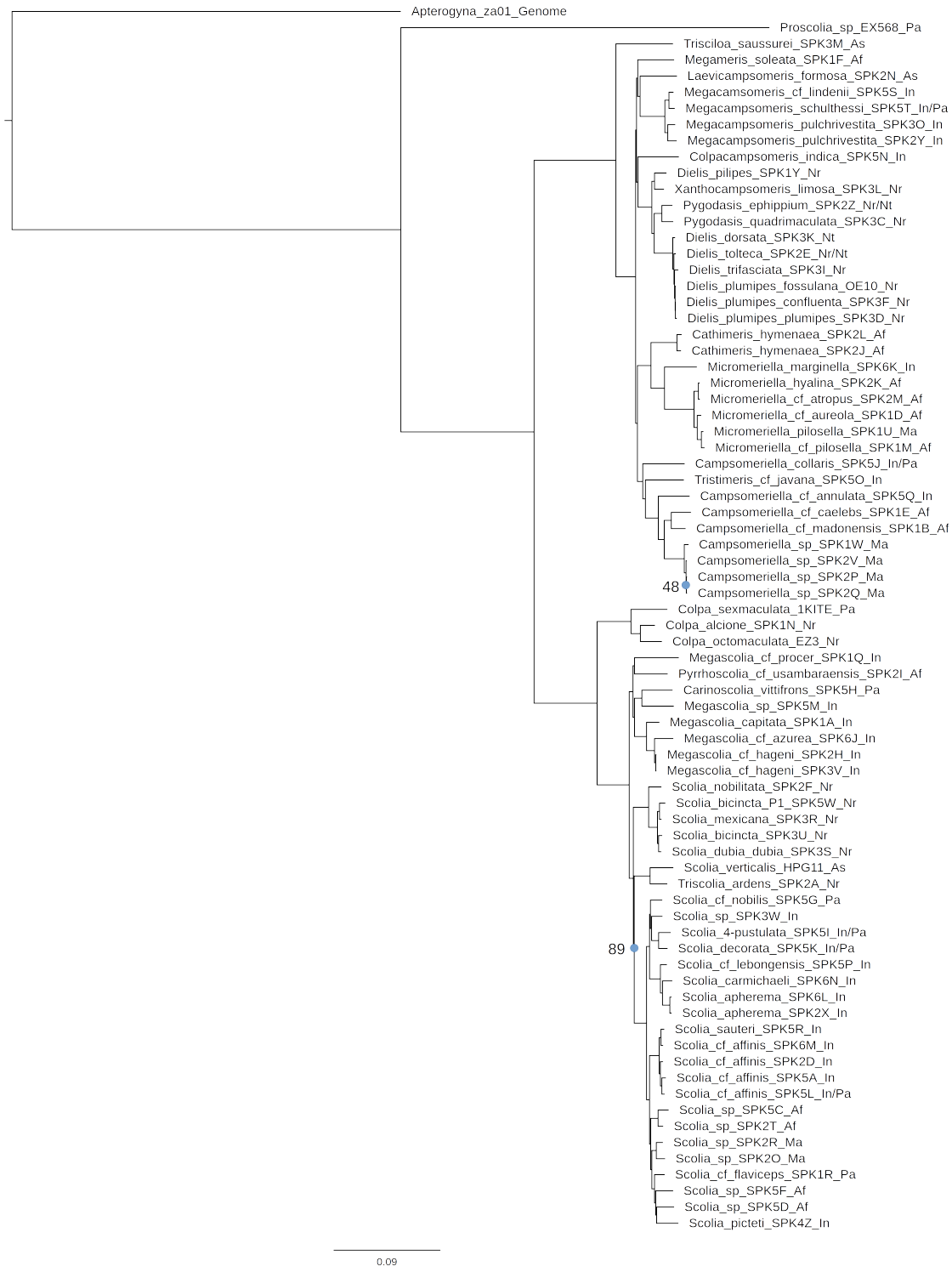

Figure 1.2. Maximum likelihood phylogeny based on 727 UCE loci (analysis 1b). Node support values based on Ultrafast Bootstrapping in IQTREE. All unlabeled internal nodes have 100% bootstrap support.

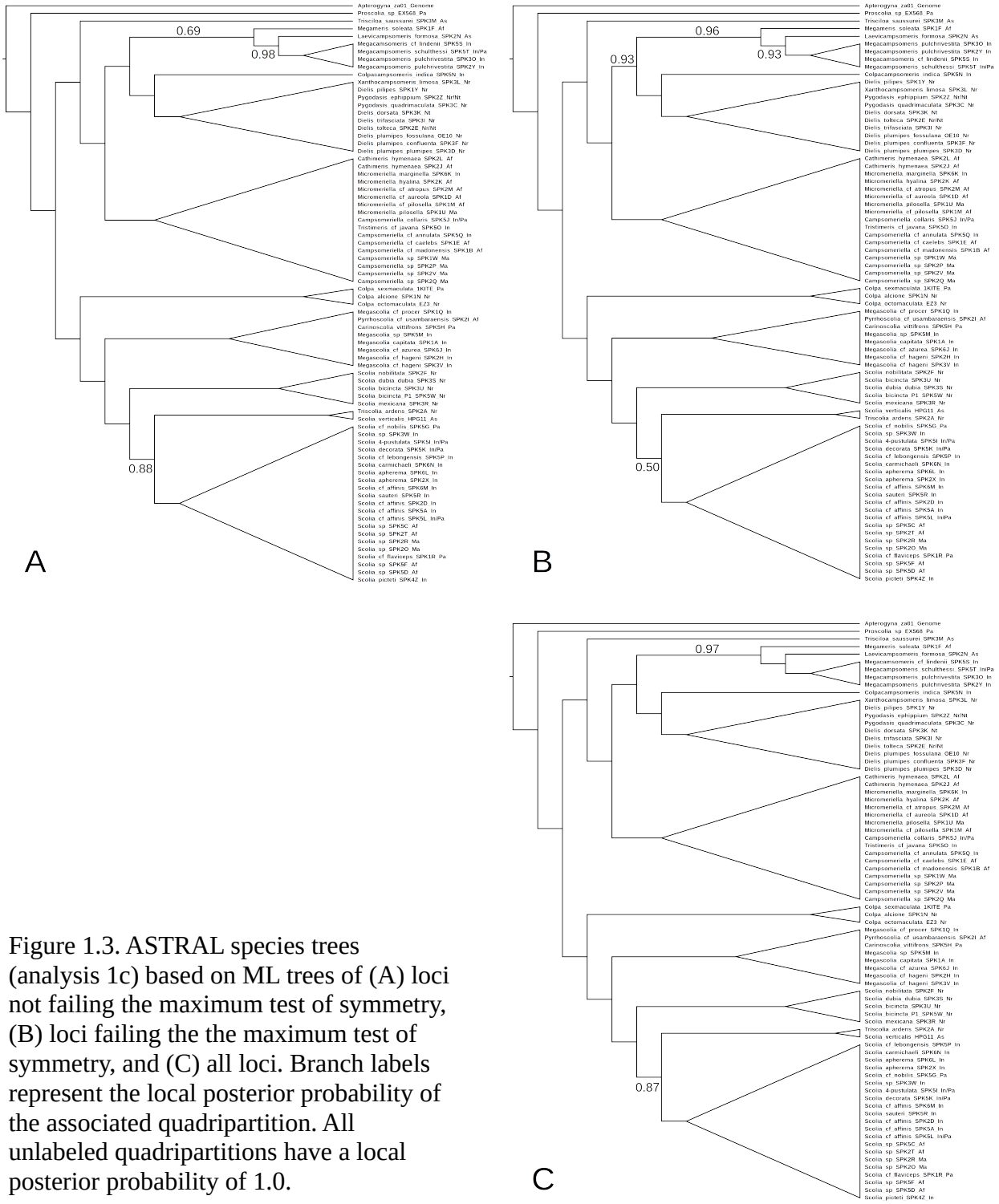

Figure 1.3. ASTRAL species trees (analysis 1c) based on ML trees of (A) loci not failing the maximum test of symmetry, (B) loci failing the the maximum test of symmetry, and (C) all loci. Branch labels represent the local posterior probability of the associated quadripartition. All unlabeled quadripartitions have a local posterior probability of 1.0.

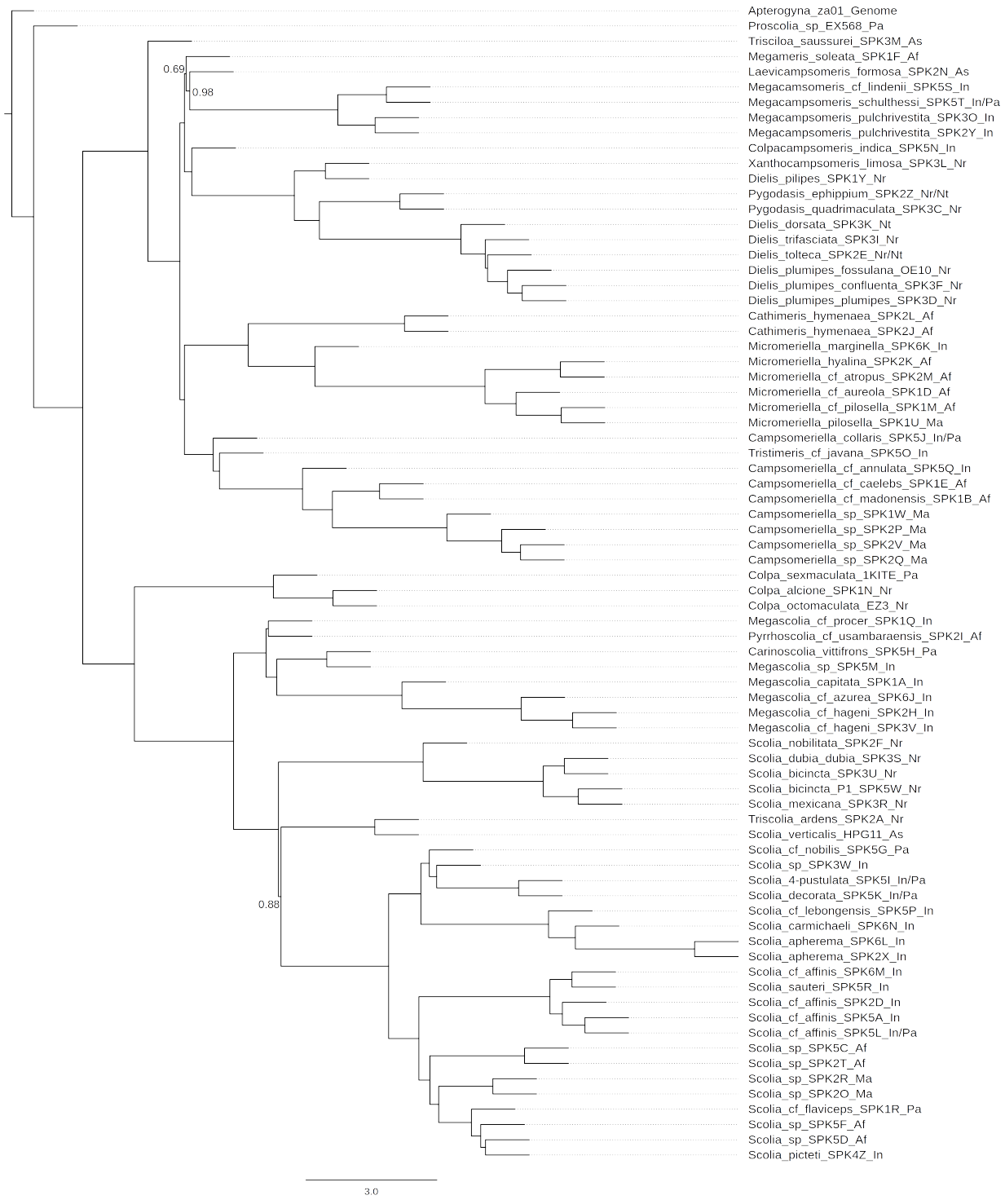

Figure 1.4. ASTRAL species tree (analysis 1c) based on ML trees of loci not failing the maximum test of symmetry. Branch labels represent the local posterior probability of the associated quadripartition. All unlabeled quadripartitions have a local posterior probability of 1.0.

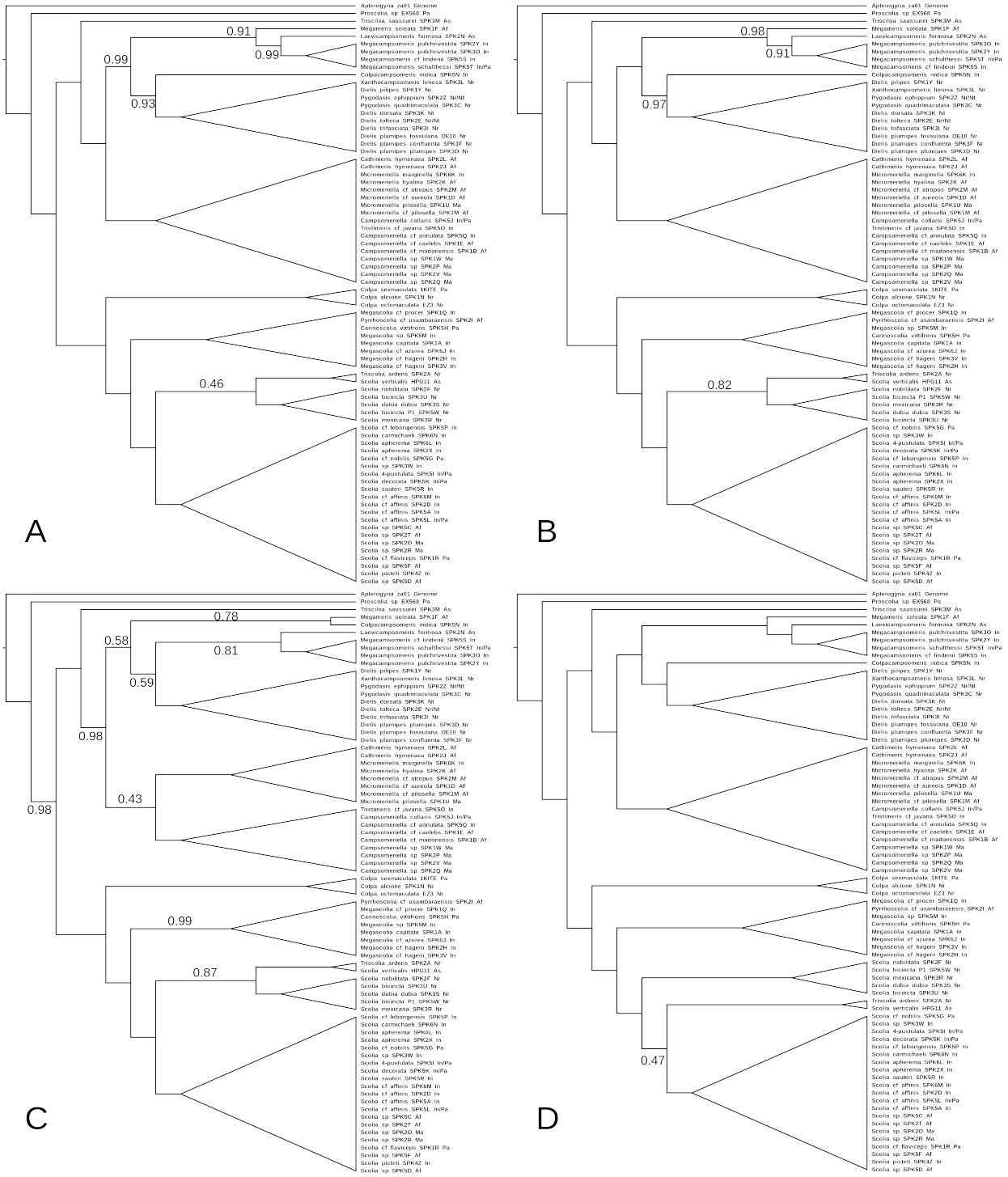

Figure 1.5. ASTRAL species trees (analysis 1c) based on MCC trees of (A) loci having the lowest (1/3) combined posterior predictive effect sizes, (B) loci having the highest (1/3) combined posterior predictive effect sizes, (C) loci for which the model was not found to be inadequate (alpha = 0.05), and (D) all loci. Branch labels represent the local posterior probability of the associated quadripartition. All unlabeled quadripartitions have a local posterior probability of 1.0.

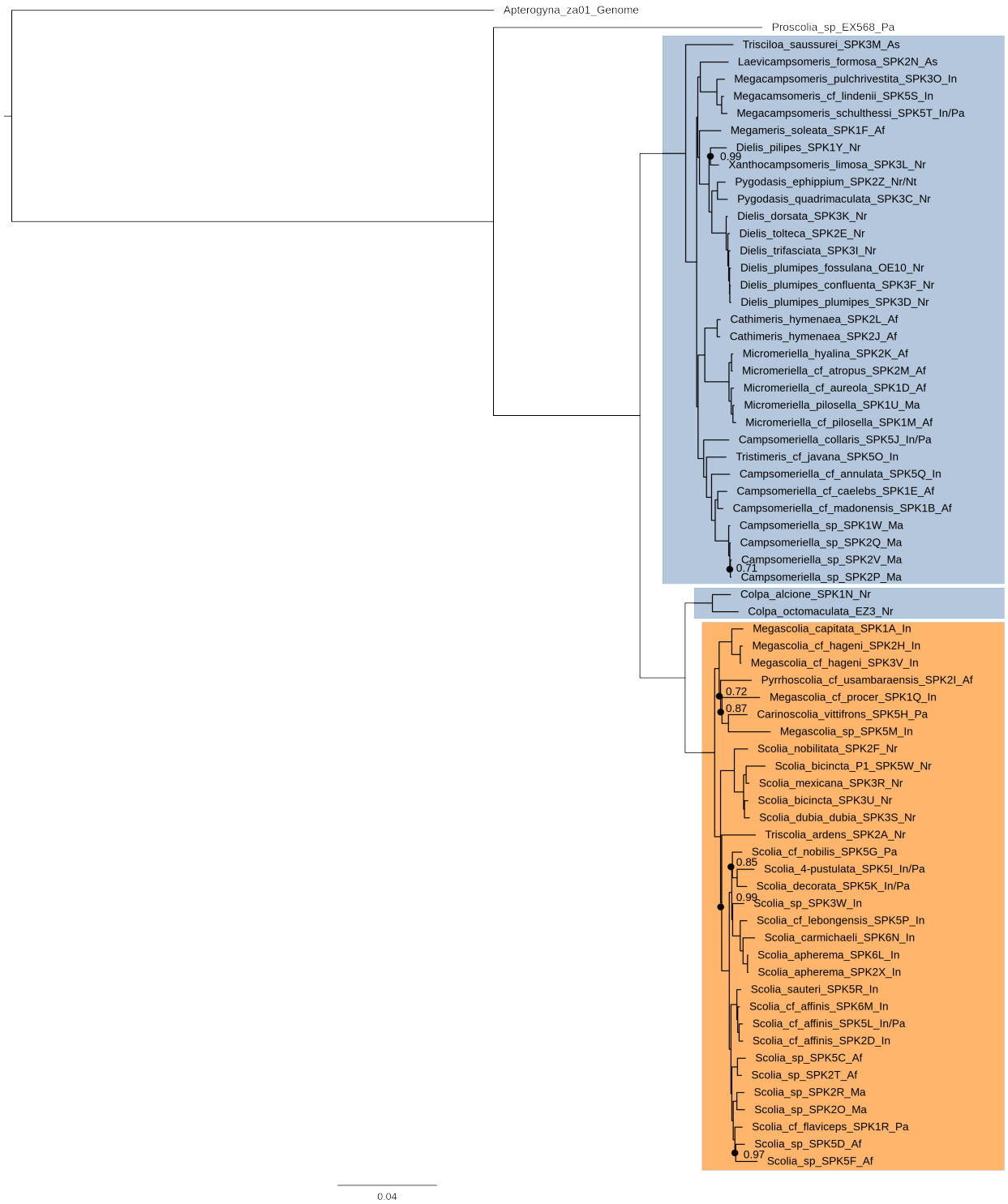

Figure 1.6. Bayesian MAP tree based on 31 UCE loci after data filtering using posterior predictive checks (analysis 2a). All unlabeled internal nodes have posterior probabilities of 1.0. Paraphyletic Campsomerini highlighted in blue; Scoliini highlighted in orange.

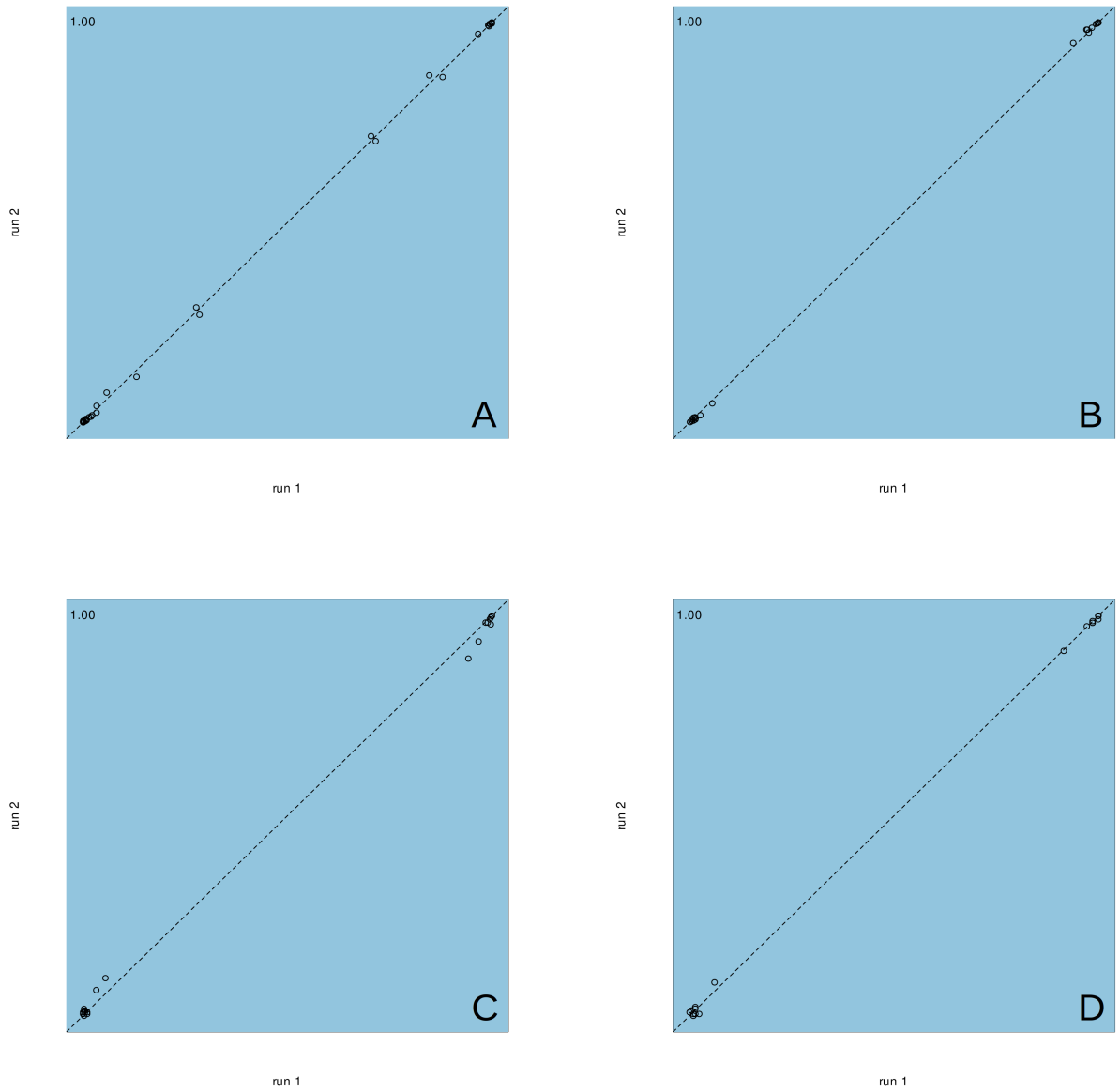

Figure 1.7. Comparison of split posterior probabilities between two independent MCMC runs: (A) analysis 2a; (B) analysis 2b, root calibration only; (C) analysis 2b, root + *Megacampsomeris* calibration; (D) analysis 2b, root + *Megacampsomeris* + Scoliidae calibration. Numbers in the top left corners represent R^2.

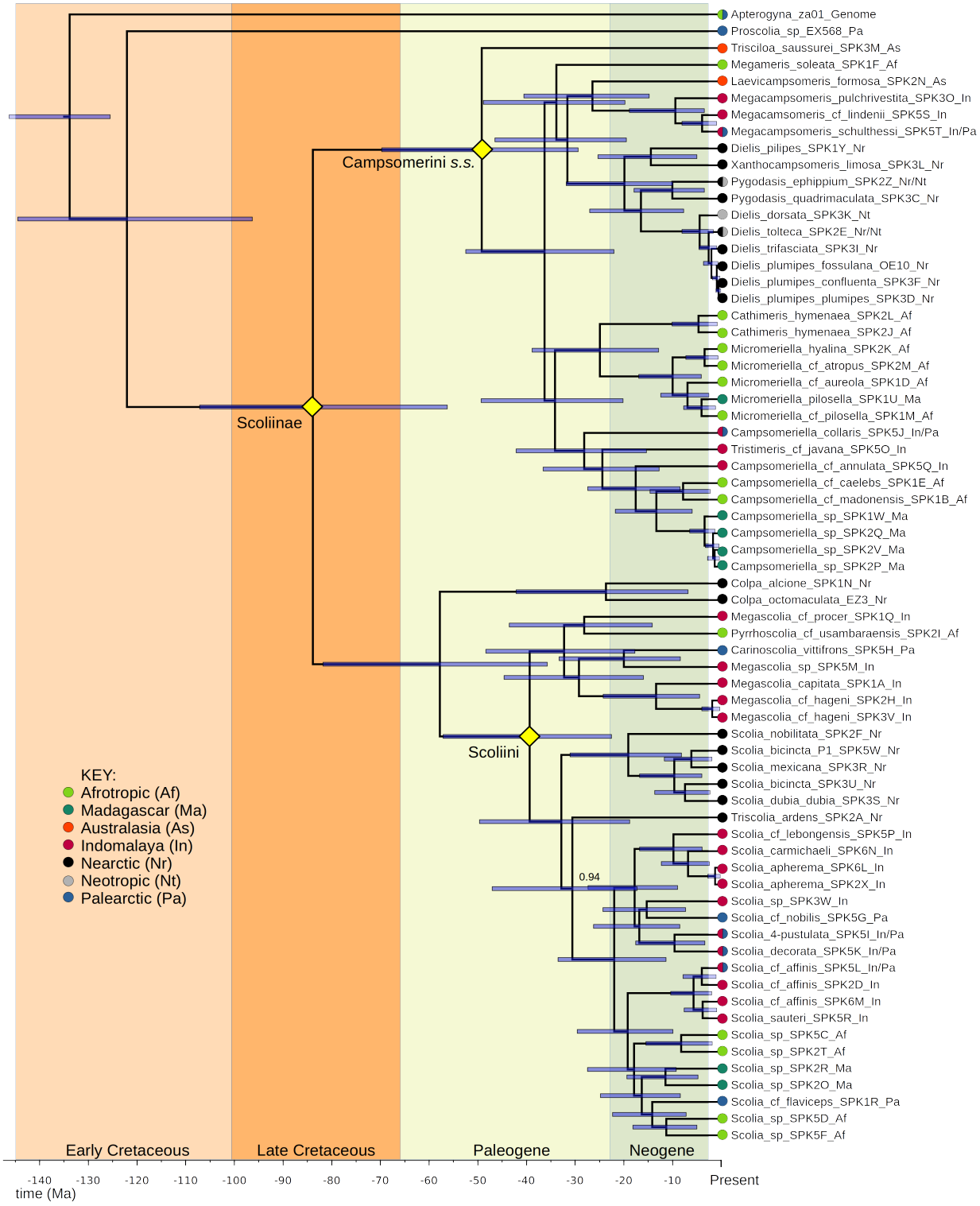

Figure 1.8. Bayesian MAP chronogram based on 63 loci after data filtering using posterior predictive checks (analysis 2b). Node bars represent age 95% HPD intervals. All unlabeled internal nodes have posterior probability of 0.97 or greater. Taxonomic labels indicated with ◇.

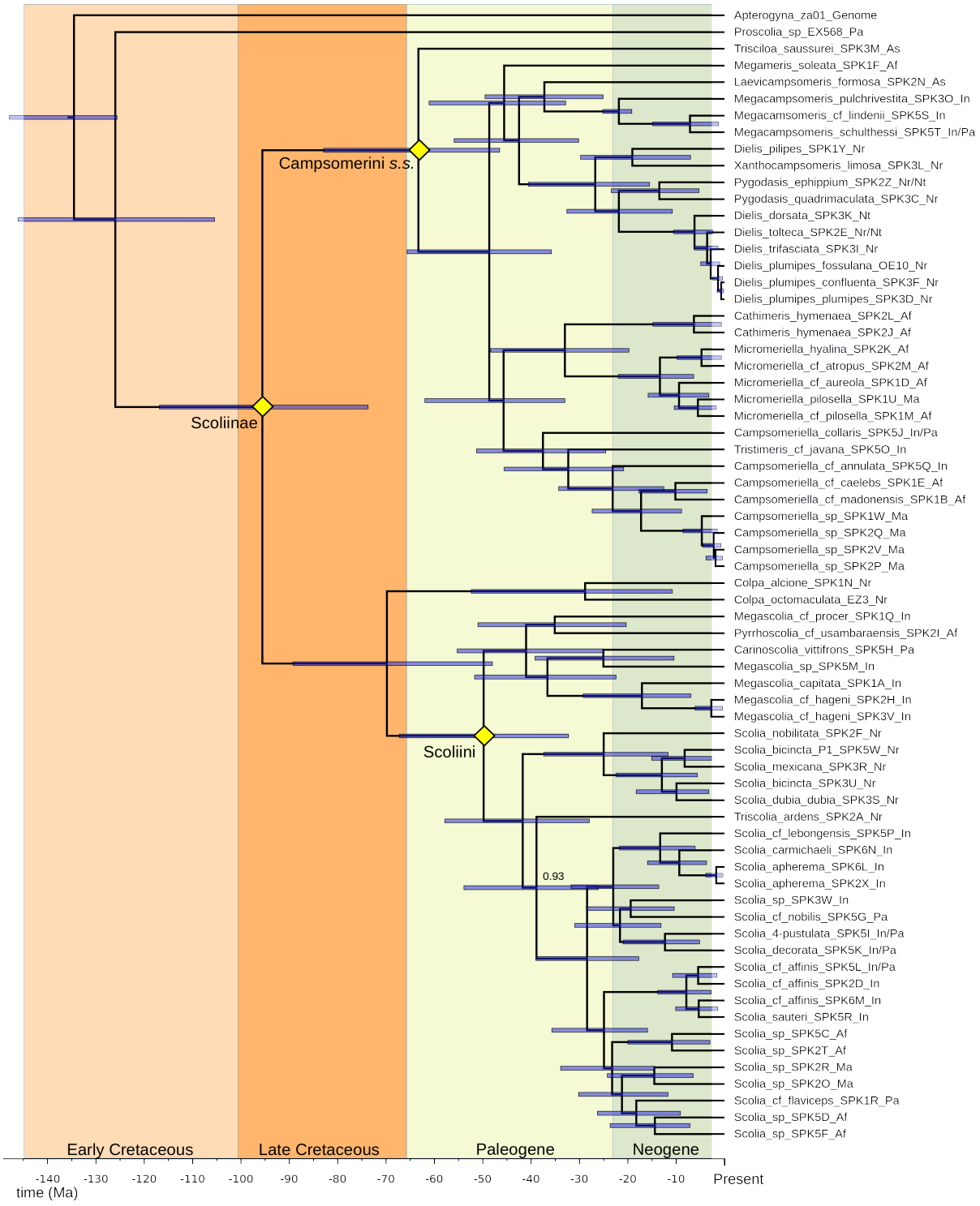

Figure 1.9. Bayesian MAP chronogram using additional calibration on crown *Megacampsomeris* (analysis 2b). All unlabeled internal nodes have posterior probability of 0.96 or greater. Taxonomic labels indicated with ◆.

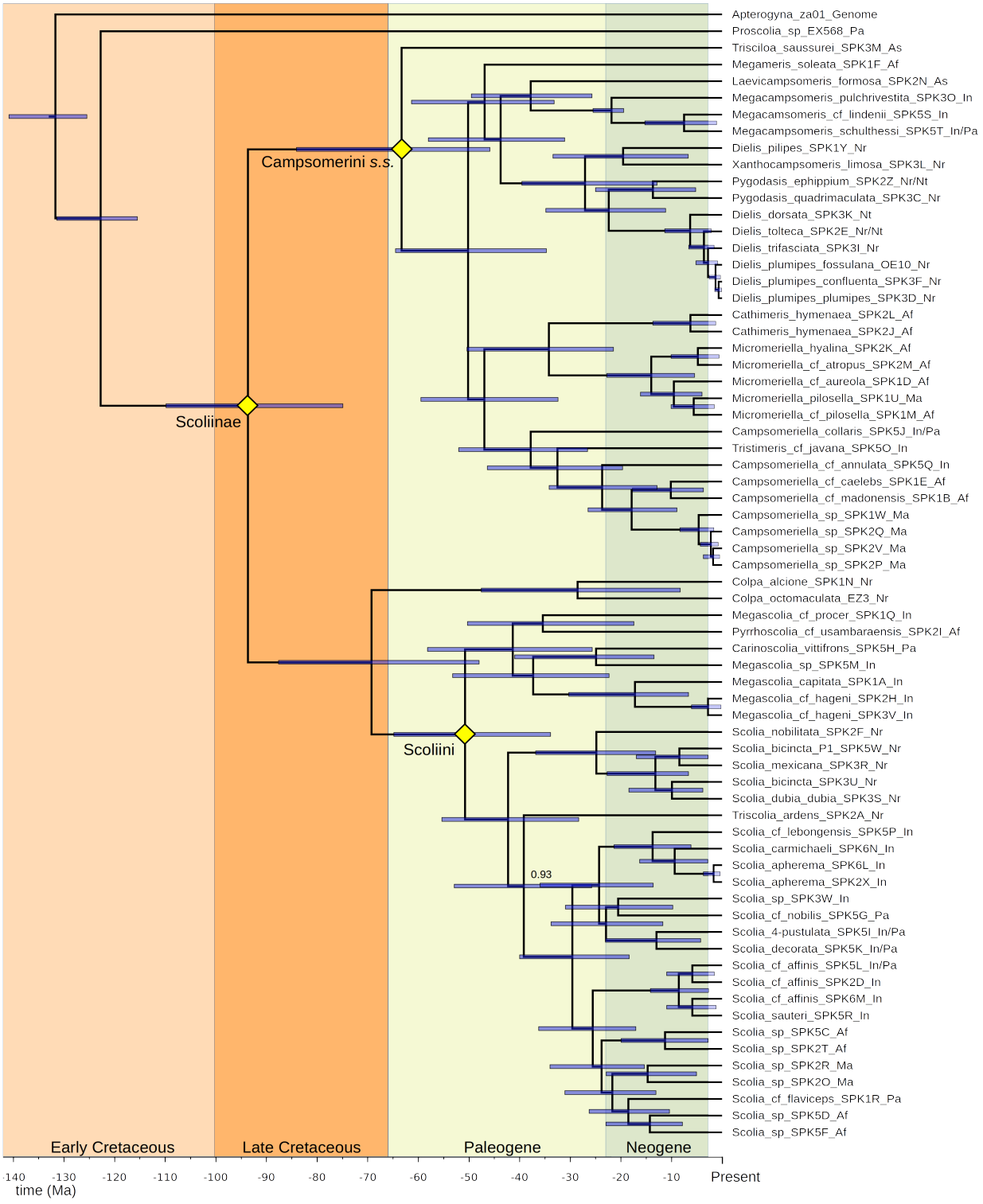

Figure 1.10. Bayesian MAP chronogram using additional calibrations on crown *Megacampsomeris* and crown Scoliidae (analysis 2b). All unlabeled internal nodes have posterior probability of 0.97 or greater. Taxonomic labels indicated with ◆.

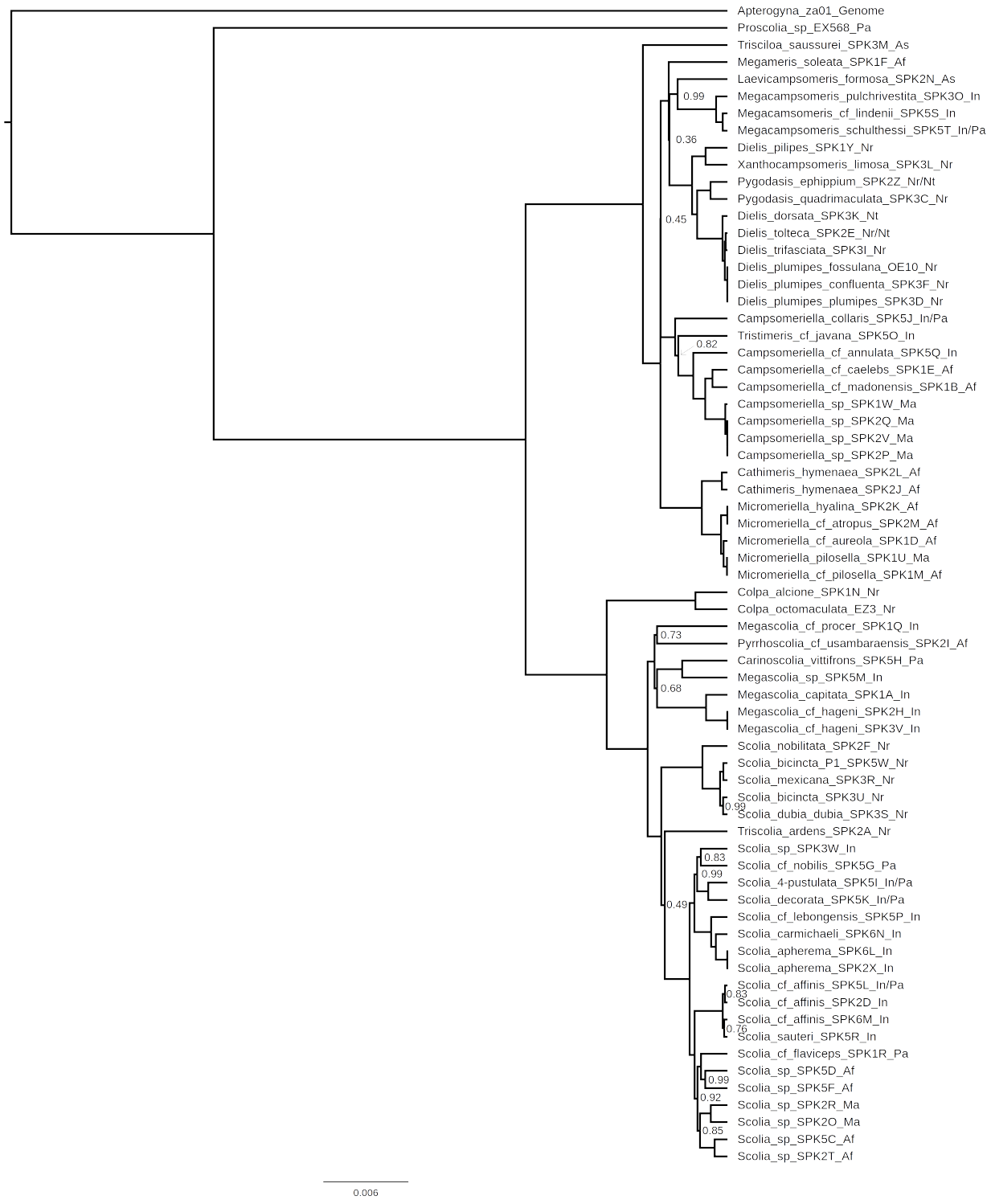

Figure 1.11. MCC species or minimal clusters tree based on 4 independent MCMC chains run using STACEY (analysis 2c). All unlabeled internal nodes have posterior probability of 1.0.

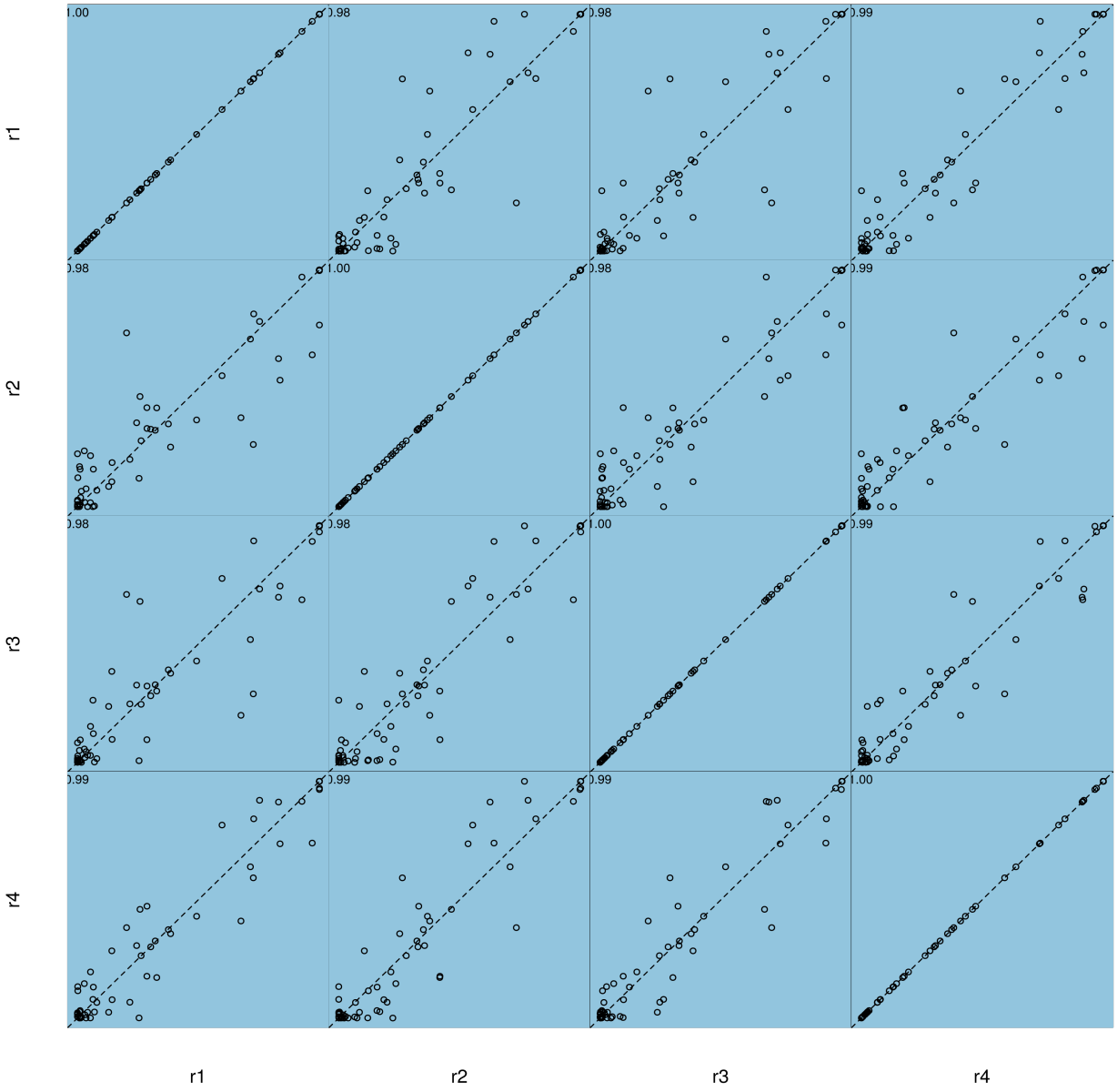

Figure 1.12. Comparison of split posterior probabilities between four independent MCMC runs using STACEY (analysis 2c). Numbers in the top left corners represent R^2.

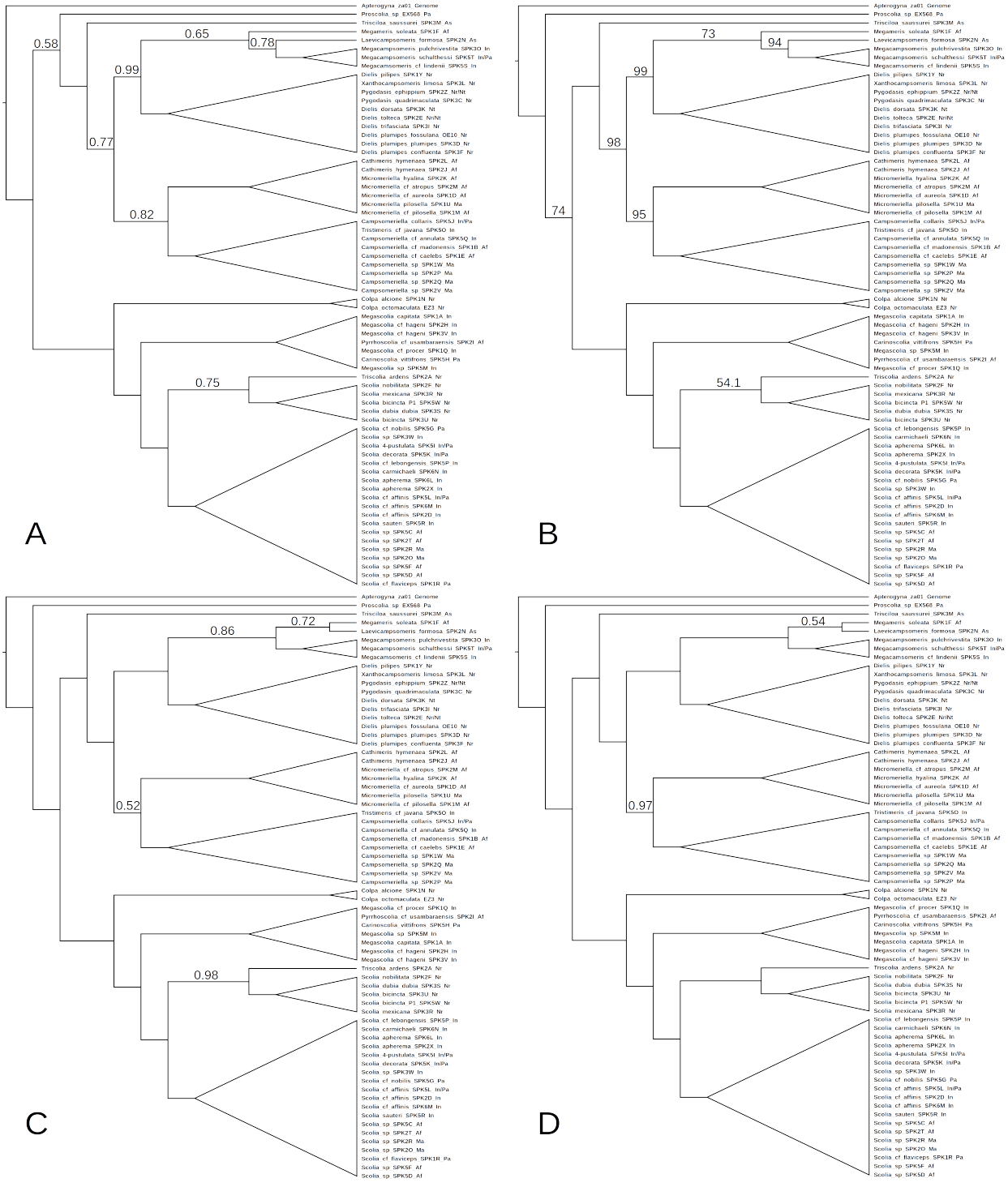

Figure 1.13. ASTRAL species trees (analysis 2d) based on MCC trees of (A) loci having the lowest (1/3) combined posterior predictive effect sizes, (C) loci having the highest (1/3) combined posterior predictive effect sizes, and (D) all loci. (B) is an ASTRAL bootstrap consensus tree using posterior samples of gene trees from loci having the lowest (1/3) combined posterior predictive effect sizes. Branch labels represent the local posterior probability or bootstrap support of/for the associated quadripartition.

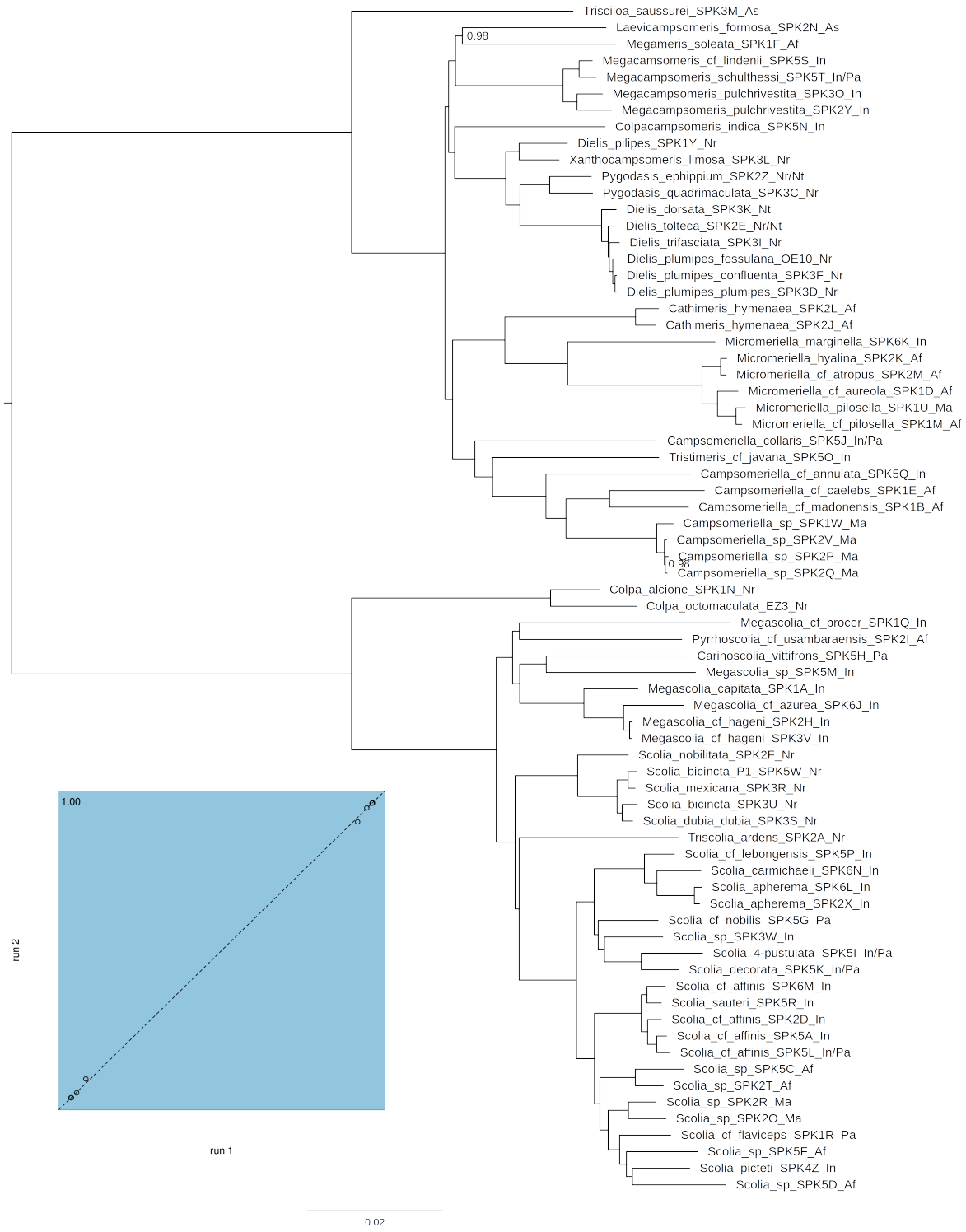

Figure 1.14. Bayesian MAP tree based on 115 UCE loci after data filtering using posterior predictive checks (analysis 3a). All unlabeled internal nodes have posterior probabilities of 1.0. Comparison of split posterior probabilities between two independent MCMC runs on lower left.

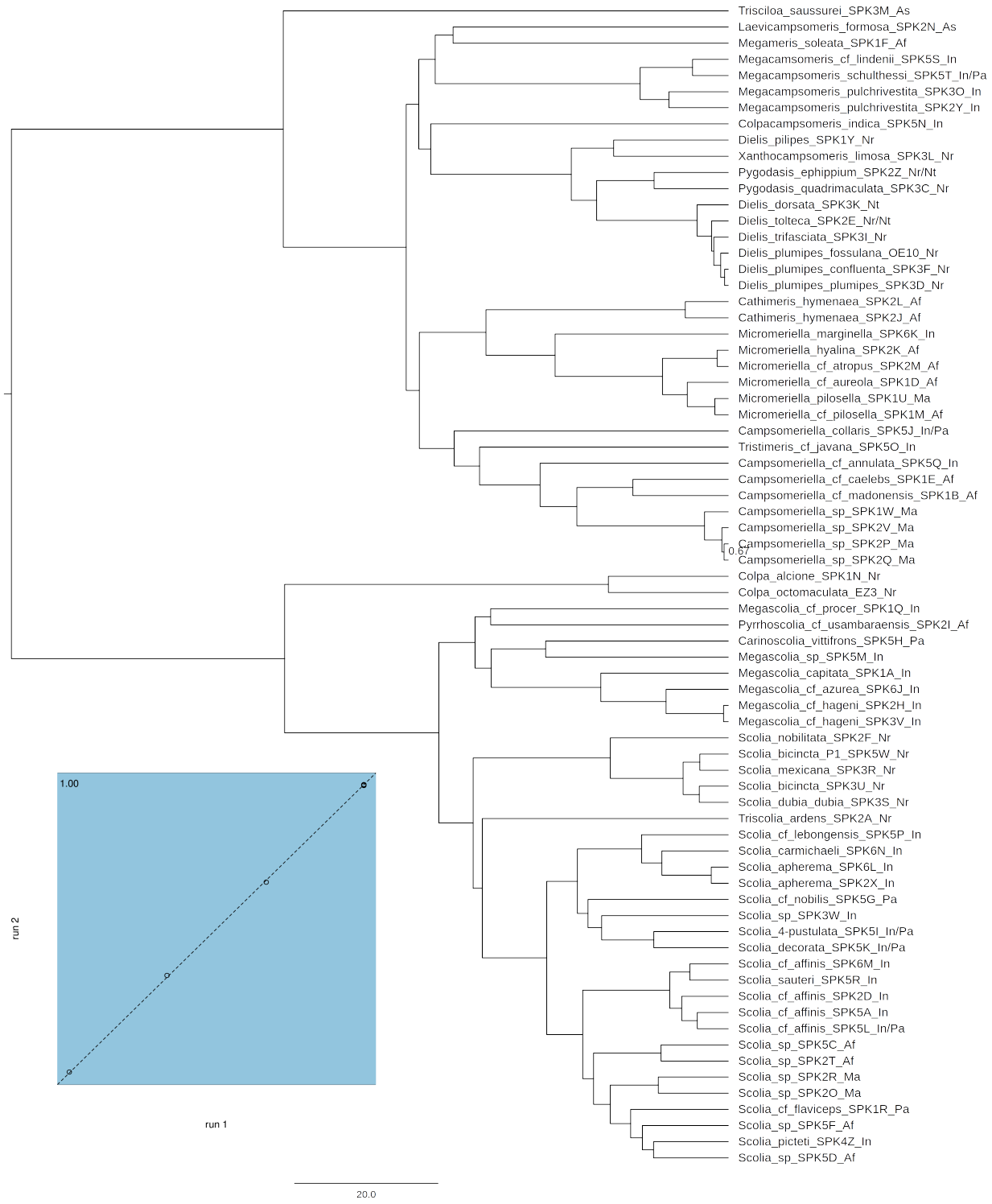

Figure 1.15. Bayesian MAP relative-time chronogram based on 159 UCE loci after data filtering using posterior predictive checks (analysis 3b). All unlabeled internal nodes have posterior probabilities of 1.0. Comparison of split posterior probabilities between two independent MCMC runs on lower left.

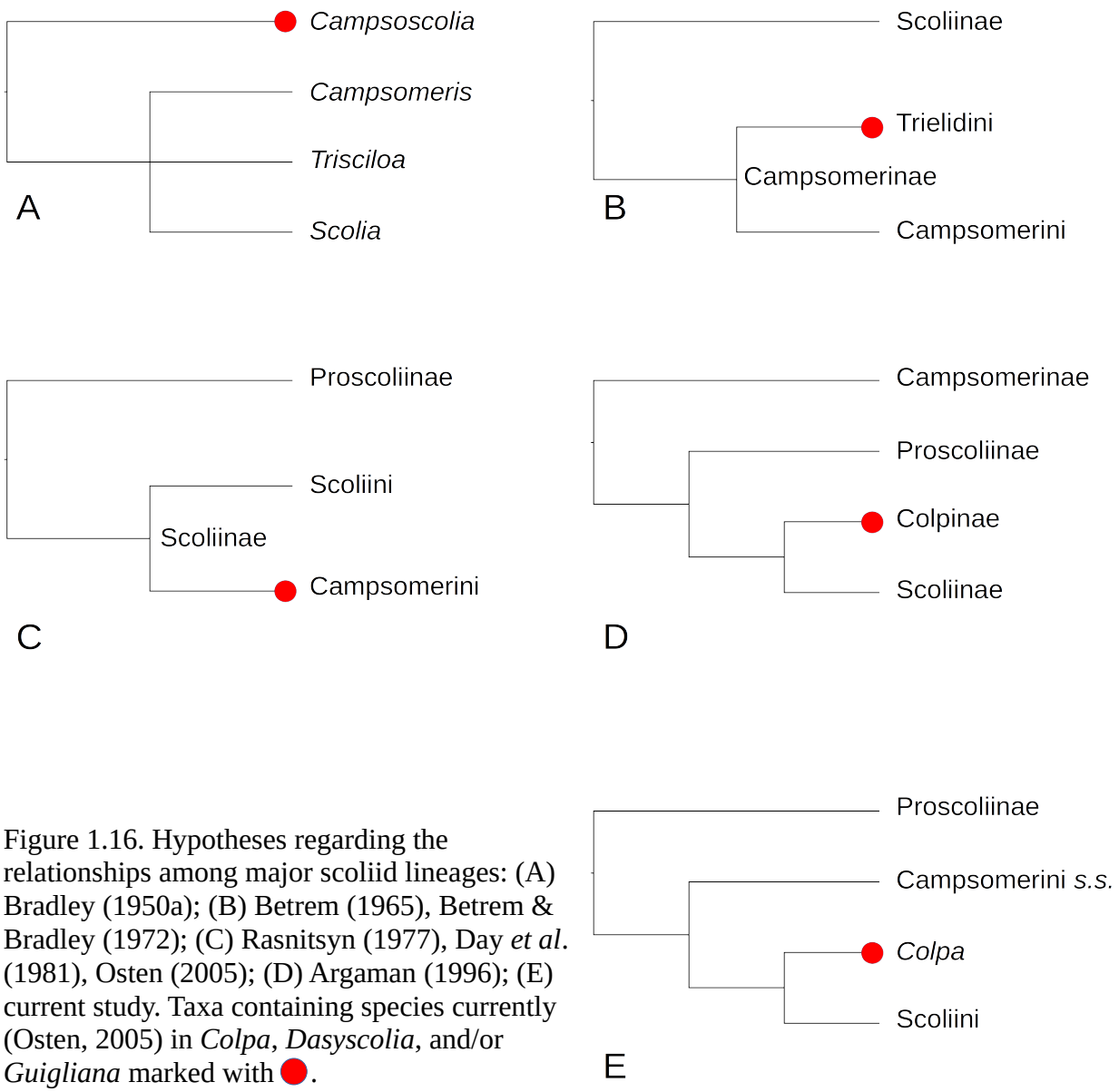

Figure 1.16. Hypotheses regarding the relationships among major scoliid lineages: (A) Bradley (1950a); (B) Betrem (1965), Betrem & Bradley (1972); (C) Rasnitsyn (1977), Day *et al*. (1981), Osten (2005); (D) Argaman (1996); (E) current study. Taxa containing species currently (Osten, 2005) in *Colpa*, *Dasyscolia*, and/or *Guigliana* marked with ●.

Supporting Information

Validation of *Scolia verticalis* and *Triscolia ardens* data

We obtained Sanger data for *T. ardens* from Pilgrim *et al*. (2008) and for *S. verticalis* from Klopfstein & Ronquist (2013), Brady *et al*. (2006), and Ward & Fisher (2016). Sequences from Brady *et al*. (2006) and Ward & Fisher (2016) are from the same specimen used by Faircloth *et al*. (2015) and so not independent from the ones used in this study. See Table S1.4 for loci, accession numbers, and voucher information.

We extracted the corresponding loci from our Scoliini data and from *Colpa* (used as an outgroup) following the same procedure we used to extract UCEs, but using a probe file only containing probes corresponding to exon sequences. We then aligned the sequences with MAFFT (E-INS-i algorithm) and and performed edge-trimming using the phyluce_align_get_trimmed_alignments_from_untrimmed script from the phyluce package. We then visually inspected the alignments, manually removed non-coding regions, and enforced the correct reading frame, validating the amino acid translations by using BLAST to align them against Hymenoptera protein sequences on GenBank.

We initially partitioned the data by locus and codon position, then used IQTREE ModelFinder (Kalyaanamoorthy *et al*., 2017) to select a partition scheme and substitution models using the greedy algorithm for exploring possible partition schemes and BIC. We used the "edge-linked-proportional" model for branch lengths (i.e. a single vector of branch lengths with partition-sepcific rate multipliers) and restricted the substitution model search space using the --mset mrbayes option. We then performed ML tree inference in IQTREE as well as Bayesian inference in RevBayes under the preferred model.

The *S. verticalis* and *T. ardens* sequences grouped with their conspecifics, and *S. verticalis* and *T. ardens* grouped with each other in both analyses (Fig. S1.2). However, support for the clade comprising target-enrichment-derived *S. verticalis* sequences + other *S. verticalis* sequences was weak, and the Sanger sequences from Klopfstein & Ronquist (2013) grouped with the Sanger sequences from Brady *et al*. (2006) and Ward & Fisher (2016) to the exclusion of the target enrichment sequences from Faircloth *et al*. (2015), despite the Faircloth *et al*. (2015), Brady *et al*. (2006), and Ward & Fisher (2016) sequences having been obtained from the same physical specimen. We hypothesize this is due to sequencing and/or data processing errors, likely in the target enrichment pipeline. Nonetheless, these analyses independently corroborate the relatedness of *T. ardens* and *S. verticalis*. This grouping might still be artefactual (e.g. due to model misspecification), but it is unlikely to be due to specimen misidentification or incorrect association of sequences with specimens.

Notes on taxon names used in this study

Scolia affinis Guérin, 1845

Specimens SPK6M (Thailand), SPK2D (Thailand), SKP5A (Sri Lanka), and SPK5L (Bhutan) key out to *Scolia aureipennis* Lepeletier 1845 (tome 3, page 525) using Betrem's (1928) key. SKP5A also keys out to *S. affinis* using Krombein's (1978) key. *S. affinis* was originally described from a specimen labeled as being from Senegal. Krombein (1978), citing "recent papers by Betrem and Bradley" states that *S. aureipennis* is a synonym of *S. affinis* and that the Senegal label on the type is erroneous. See also Bradley (1964, 1974) for further discussion of types and synonymy. Osten (2005) lists *S. aureipennis* as a possible synonym of *Scolia affinis* Guérin, 1845. We provisionally refer to the specimens in this study listed above as *Scolia cf. affinis*. They mostly match descriptions of that species (and its putative synonyms), but show

some differences in morphology and color. This could be attributed to geographic variation within a single widespread species. Specimen SPK6M appears to be genetically more closely related to specimen SPK5R (China) identified as *Scolia superciliaris* than to the other specimens attributed to *S. affinis,* including the other specimen from Thailand SPK2D, which raises the possibility of *S. superciliaris* being a color variant of *S. affinis*. On the other hand, is it is also possible that this group includes multiple species, some currently undescribed. Given the broad geographic range of *S. affinis* as currently understood, settling this matter definitively will likely require examination of material from across the entire Indomalayn region.

Scolia lebongensis Betrem, 1928

Specimen SPK5P (China) keys out to *S. lebongensis* in Betrem 1928, but is also similar to how *S. oculata formosicola* (as *S. formosicola*) is described. However, it does not fully match the description of either species and might represent a currently undescribed closely related species.

Scolia flaviceps Eversmann, 1846

For specimen SPK1R (Lebanon), compare also with *Scolia orientalis*.

Supporting information references

Bartlett, O. C. (1912). The North American Digger Wasps of the Subfamily Scoliinae. *Annals of the Entomological Society of America*, *5*(4), 293-340.

Betrem, J. G. (1928). Monographie der Indo-Australischen Scoliiden Mit zoogeographischen Betrachtungen. *Treubia*, *9* (Suppl.), 1-388.

Bradley, J. C. (1928a). A Revision of the New World Species of *Trielis* a Subgenus of *Campsomeris* (Hymenoptera: Scoliidae). *Transactions of the American Entomological Society (1890-)*, *54*(3), 195-214.

Bradley, J. C. (1928b). The Species of *Campsomeris* (Hymenoptera-Scoliidae) of the Plumipes Group, Inhabiting the United States, the Greater Antilles, and the Bahama Islands. *Proceedings of the Academy of Natural Sciences of Philadelphia*, *80*, 313-337.

Bradley, J. C. (1957). The taxa of Campsomeris (Hymenoptera: Scoliidae) occurring in the New World. *Transactions of the American Entomological Society (1890-)*, *83*(2), 65-77.

Grissell, E. E. (2007). Scoliid Wasps of Florida, Campsomeris, Scolia and Trielis spp. (Insecta: Hymenoptera: Scoliidae). *Institute of Food and Agricultural Sciences Extension Electronic Data Information Source*, 1-9. *See https://edis.ifas.ufl.edu/pdf/IN/IN74500.pdf*

Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K., Von Haeseler, A., & Jermiin, L. S. (2017). ModelFinder: fast model selection for accurate phylogenetic estimates. *Nature methods*, *14*(6), 587-589.

Kim, J. K., & Yoon, I. B. (1993). A taxonomic study of subfamily Scoliinae from Korea (Hymenoptera: Scoliidae). *Entomological Research Bulletin*, *19*, 1-6.

Krombein, K. V. (1978). Biosystematic studies of Ceylonese wasps, II: A monograph of the Scoliidae (Hymenoptera: Scolioidea). *Smithsonian Contributions to Zoology*.

Osten, T. (2000). Die Scoliiden des Mittelmeer-Gebietes und angrenzender Regionen (Hymenoptera). Ein Bestimmungsschlüssel. *Linzer biologische Beiträge*, *32*(2), 537-593.

Rohwer, S. A. (1927). Some scoliid wasps from tropical America. *Journal of the Washington Academy of Sciences*, *17*(6), 150-155.

Chapter 2: Comparing the power of data-based phylogenetic posterior predictive checks in the context of nucleotide and amino acid data

Khouri, Z.[1], Gillung, J.P.[2], May, M.R.[3], Moore, B.[4]

[1] Bohart Museum of Entomology, University of California, Davis, CA, U.S.A.; [2] Lyman Entomological Museum, McGill University, Montreal, Quebec, Canada; [3] Department of Integrative Biology, University of California, Berkeley, CA, U.S.A.; [4] Department of Evolution and Ecology, University of California, Davis, CA, U.S.A.

Introduction

Most modern approaches to phylogenetic reconstruction are model-based and thus potentially prone to error when the chosen model poorly describes the data at hand. The last two decades have seen the development of improved methods to evaluate relative model fit (Lartillot & Philippe, 2006; Fan *et al.*, 2011; Xie *et al.*, 2011; Baele *et al.*, 2012a; 2012b) as well as model adequacy (Huelsenbeck *et al.*, 2001; Bollback, 2002; Brown, 2014a; 2014b; Lewis *et al.*, 2014; Doyle *et al.*, 2015) in a Bayesian context. Despite this, assessment of model adequacy using posterior predictive simulation remains relatively uncommon in empirical phylogenetic studies (but see Williams *et al.* (2020) and May *et al.* (2021) for two recent exceptions).

Statistical power is the probability that a test correctly rejects a hypothesis when it is false. However, in the context of posterior predictive checks, maximizing statistical power may not always be desirable. Models are by necessity simplified representations of the biological processes generating the data and can thus always be rejected given enough data (Gelman, 2013). The degree of correspondence between detectable model violation and error in phylogenetic parameter estimation is an important practical consideration. Inference-based test statistics

(Brown, 2014a) were primarily motivated by the need to develop tests of model adequacy that were more indicative of a model's propensity to lead to inaccurate inferences. On the other hand, assessments of the power of specific posterior predictive tests may be useful if performed using realistic amounts of data and when viewed in relation to inference (in)accuracy.

Such assessments have been performed, typically in studies introducing new tests and using nucleotide data (e.g. Bollback, 2002). Posterior predictive checks (without power analyses) have also been applied to complex amino acid (Blanquart & Lartillot, 2008) and codon models (Rodrigue *et al.*, 2008; 2010) as well as to models of re-coded amino acids (Feuda *et al.*, 2017). Duchêne *et al.* (2016) evaluated the power of various model adequacy tests, including posterior predictive tests of some codon models. However, no study has to our knowledge attempted to compare the power of posterior predictive tests of models describing nucleotide and amino acid coding of the same underlying data.

Coding and subsequently analyzing DNA sequence data in terms of amino acids is often preferred when tree lengths are expected to be high and/or when compositional heterogeneity is pronounced at the nucleotide but not at the amino acid level. However, doing so results in loss of information. For datasets of intermediate tree length, the choice of which data type to use may not be obvious *a priori*, while phylogenetic inferences may differ significantly depending on this choice (e.g. Gillung *et al.*, 2018). In such cases, tests of model adequacy may be helpful, and understanding the statistical power of these tests and how their results may correlate with errors in inference might aid interpretation.

In this study, we employed a Bayesian simulation-based approach to compare the power of data-based posterior predictive tests of nucleotide and amino acid models. We sampled posterior distributions of parameter values under nucleotide and amino acid models applied to empirical

data from acrocerid flies (Gillung *et al.*, 2018), amniotes (Chiari *et al.*, 2012), birds (Jarvis *et al.*, 2015), brittle stars (O'Hara *et al.*, 2017), seed plants (Ran *et al.*, 2018), plastids (Ruhfel *et al.*, 2014), and *Wolbachia* (Comandatore *et al.*, 2013), and codon models applied to data from *Flavivirus* (Moureau *et al.*, 2015). We then used these parameter values to simulate sequence data, including under models featuring process heterogeneity across lineages, selection heterogeneity across sites, and selection for codon usage. We subsequently performed inference followed by posterior predictive simulation under simpler amino acid and nucleotide models. This allowed us to assess the ability of posterior predictive checks using the multinomial likelihood (Goldman, 1993; Bollback, 2002) and chi-squared test statistics (Huelsenbeck *et al.*, 2001; Foster 2004) to detect model violation and to determine the effect of model violation on inference accuracy.

Methods

Simulation parameters

For all non-*Flavivirus* datasets, our general approach was to use IQTREE (Nguyen *et al.*, 2015; Chernomor *et al.*, 2016) version 1.6.12 to get maximum likelihood estimates (MLEs) of the following parameters: stationary frequencies (Fig. 2.1C), exchangeability rates, shape parameter (Fig. 2.1A, 2.2A) of the discretized gamma across-site rate variation (ASRV) model (Yang, 1994), and tree length (Fig. 2.1B, 2.2B). We used a GTR (Tavaré, 1986) substitution matrix for both amino acid and nucleotide data. In all cases, we partitioned the data by locus. For nucleotide data, we additionally partitioned by codon position, estimating all the above parameters separately. Tree topology and relative branch lengths were linked across partitions, but a separate tree length parameter was estimated for each partition (-spp option). For loci without data for some taxa, we rescaled the estimated tree length by dividing it by the number of branches in the

associated gene tree then multiplying the result by the number of branches in the tree containing all taxa (i.e. we preserved the average branch length). We used 6 independent runs for each dataset.

For some simulations (see below), we used parameter MLEs directly. However, in the case of nucleotide stationary frequencies and exchangeability rates, we also wanted the ability to obtain values that have similar distributions to the aggregated MLEs from empirical data. For each set of stationary frequencies and exchangeability rates of interest, we treated their MLEs as data drawn from a Dirichlet distribution with an unknown α parameter (with a uniform prior over the range 0 to 20 for each element of α) and sampled the posterior distribution of α using RevBayes (Höhna *et al.*, 2014; 2016). We then used the means of the posterior distributions of α to specify distributions from which to draw stationary frequencies and exchangeability rates for simulation.

We sub-sampled the empirical datasets as follows (see Table 2.1 for a summary). For amniotes, we used a 300 character minimum alignment length cutoff per locus for amino acid data and an 800 character cutoff for nucleotide data. For both data types, we only retained loci with at least 50% taxon coverage, resulting in subsets of 54 and 85 nucleotide and amino acid alignments respectively. We used more stringent criteria for the larger bird dataset: minimum lengths of 700 and 2100 characters for amino acid and nucleotide alignments respectively, 100% taxon coverage, and a maximum of 7% missing data at the site level. This resulted in reduced amino acid and nucleotide datasets of 104 loci. In the case of brittle stars we used cutoffs of 300 and 900 characters for amino acid and nucleotide data respectively and 100% taxon coverage, resulting in the retention of 76 loci. Plants were treated similarly in terms of minimum taxon coverage, but we used lengths of 500 and 1500 characters as cutoffs, keeping 170 loci. We sub-sampled the plastid dataset down to 284 taxa, and used cutoffs of 200 and 600 characters. 31 loci

were retained. We used the same criteria for *Wolbachia*, but without taxon sub-sampling, and also retained 31 loci.

To obtain empirically grounded parameter values for simulation under codon models, we used a 100 taxon *Flavivirus* dataset from Moureau *et al.* (2015). See Supporting Information for a list of accession numbers. It was chosen as an example of a "difficult" dataset with high rates of evolution as well as because it was previously used by Duchêne *et al.* (2016).

We considered three approaches to estimating a codon tree length for the full dataset: (1) estimating the tree length directly using BAli-Phy (Redelings, 2021) on codon data while taking into account the uncertainty in alignment, tree topology, and model numerical parameters, (2) using BAli-Phy on amino acid data to get maximum *a posteriori* (MAP) estimates of alignment and topology, then using PAML (Yang, 2007) on the corresponding codon data and estimating a maximum likelihood (ML) tree length conditioning on the MAP topology and alignment, and (3) using MAFFT (Katoh & Standley, 2013) to obtain an amino acid alignment, IQTREE with amino acid data to get a ML topology estimate conditioning on that alignment, and PAML to estimate the codon tree length. The first option is likely the most accurate but is also computationally expensive. Given that our goal was getting a ballpark estimate to use for simulation, we opted for the second option as a compromise between accuracy and computational burden.

In order to select a model of amino acid substitution for use in BAli-Phy, we first estimated a preliminary amino acid alignment with MAFFT version 7.471 using the E-INS-i algorithm (Altschul, 1998) and 10,000 cycles of iterative refinement. We then used IQTREE (Minh *et al*., 2020) version 2.0.6 to compare amino acid substitution models using the Bayesian Information Criterion (BIC) (Schwarz, 1978). LG+F+R7 and LG+F+R6 accounted for close to 100% of the

BIC weight, with the closest runners up being other variants of the LG model (Le & Gascuel, 2008).

In order to obtain MAP estimates of alignment and topology, we co-estimated the alignment, phylogeny, and other model parameters in a Bayesian framework using BAli-Phy version 3.5. We used the LG+F+R4 substitution model (reducing the number of rate categories from 7 to 4 to reduce computational cost) and the RS07 indel model (Redelings & Suchard, 2007). We ran 8 independent chains for 1,000 hours. Since some runs exhibited poor mixing and/or appeared to be stuck in regions of lower posterior density, we combined samples from the three best-behaving chains to generate MAP topology and alignment estimates. Mixing and convergence diagnostics were performed using BAli-Phy tools and the bonsai (May & Moore, 2017) version 0.9 R (Core R Team, 2020) package. We subsequently used the alignment (converted to a codon alignment using a custom python script) and topology to estimate a codon tree length under the MutSel+M3 (Yang & Nielsen, 2008; Yang *et al.*, 2000) model using PAML version 4.9j.

We performed kmeans clustering in R and used the fviz_cluster() function from the factoextra package (Kassambara & Mundt, 2020) to visualize differences in codon usage across taxa (Fig. S2.1). Finally, we extracted subsets of the full alignment corresponding to clades of interest: tick-borne flavivirus (TBFV), dengue virus (DENV), West Nile virus (WNV), yellow fever virus (YFV), and Zika virus (ZIKV). This set of clades was selected such that each member of the set has distinct patterns of codon usage. We then sampled posterior distributions of clade parameters under the MutSel+M5+G (Yang & Nielsen, 2008; Yang *et al.*, 2000) model using a custom implementation in RevBayes.

Simulation and inference under a simple amino acid model

Our general approach involved two steps: (1) the creation of a stochastic variable trace with samples from the desired distributions of parameter values followed by (2) using the posterior predictive simulation functionality of RevBayes (Höhna *et al.*, 2018) to simulate sequence data under a specific model and the set of parameter values taken from the trace.

Here, we used unrooted tree topologies with 20 terminals drawn from a uniform distribution and vectors of branch lengths drawn from a symmetrical Dirichlet distribution with a concentration parameter of 1.3. Values of the alpha (shape) parameter associated with the discretized gamma model of across-site rate variation (ASRV) as well as values of the tree length parameter were taken directly from those inferred from empirical data (see section above): each row of the trace, corresponding to one simulated alignment, was constructed to contain parameter estimates from one locus from one of the empirical datasets. We used the WAG (Whelan & Goldman, 2001) substitution matrix and simulated a set of 517 alignments of 400 amino acid sites and another set of 517 alignments of 1800 amino acid sites.

In order to evaluate the ability of posterior predictive checks to detect violation of the equal-rates-across-sites assumption, we performed phylogenetic inference followed by posterior predictive simulation on each alignment first using the true model (WAG+G) then using WAG without a discretized gamma ASRV model.

Simulation and inference under a simple nucleotide model

We ran a similar set of simulations with nucleotide data. This allowed us to compare the ability to detect a fixed level of ASRV model violation when using amino acids to that when using nucleotides. We used the same values for tree topology, branch lengths, tree length, and

discretized gamma ASRV model shape parameter as above. Nucleotide exchangeability rates and stationary frequencies were sampled from Dirichlet distributions with α parameters set to posterior mean estimates associated with second position sites from empirical data (see "Simulation parameters" section above). Similarly to the amino acid simulations, we generated a set of 517 alignments of 400 nucleotide sites and another set of 517 alignments of 1800 nucleotide sites under a GTR+G substitution model. We then performed phylogenetic inference followed by posterior predictive simulation using the GTR+G and GTR models.

Simulation under a branch-heterogenous mutation-selection codon model

We used the HKY-like version of the mutation-selection (MutSel) codon model (Yang & Nielsen, 2008) following the parameterization of Rodrigue *et al.* (2008), i.e. using a vector of 61 codon preference parameters with the constraint that they sum to zero. We implemented this model in RevBayes by writing a python script that generates Rev code defining codon stationary frequencies and exchangeability rates in terms of nucleotide stationary frequencies and exchangeability rates, a transition/transversion ratio, a d_N/d_S ratio, and codon preference parameters. We validated our implementation by simulating data under this model with fixed parameter values and recovering these values by performing maximum likelihood inference under the MutSel model in PAML (Yang, 2007) version 4.9j.

We combined this MutSel matrix with the M5 model (Yang *et al.*, 2000) allowing the d_N/d_S ratio to vary across sites following a discretized gamma distribution. We used the following approach to model shifts in the evolutionary process across the tree we: (1) treated each unrooted tree topology used for simulation as if it were rooted on the branch subtending "taxon01", (2) randomly selected two different internal branches on each of these tree topologies to serve as "breakpoints" (the "root" branch was always treated as the first breakpoint), (3) identified three

sets of branches, each including a breakpoint branch and all its descendants, (4) if a specific branch belonged to two or more sets, we reassigned it exclusively to the set associated with the younger breakpoint branch, and (5) assigned separated values of MutSel+M5 parameters to each of the three resulting sets of branches, with each set of parameters being drawn from the posterior associated with a different subset of the empirical *Flavivirus* dataset. Specifically, tick-borne flavivirus (TBFV) was always used for the root set, and all combinations of dengue virus (DENV), West Nile virus (WNV), yellow fever virus (YFV), and Zika virus (ZIKV) were used for the other breakpoint sets. Additionally, we applied a branch-homogeneous discretized gamma ASRV model acting at the codon level, with the alpha shape parameter drawn from the posterior associated with one of the empirical data subsets above. Root frequencies were set to the stationary frequencies associated with the set of branches including the branch on which the tree was rooted. See Fig. 2.3 for a schematic summary.

In order to generate a pseudo-empirical amino acid model to use for inference, we estimated amino acid stationary frequencies and exchangeability rates under a GTR+G model from simulated data. We ran an analysis using 10 alignments of 1800 amino acids each, simulated under 10 different sets of tree topologies, branch lengths, and breakpoints. We fixed the topology associated with each alignment/partition to the "true" topology used for simulation while estimating a separate tree length and vector of branch lengths for each alignment. Conversely, stationary frequencies, exchangeability rates and the gamma ASRV shape parameter were linked across alignments. The resulting amino acid matrix served as an equivalent for empirical amino acid models such as JTT (Jones *et al.*, 1992), WAG, and LG that is optimized for our simulated data.

We then preformed inference and posterior predictive simulation under this pseudo-empirical amino acid model (EMP+G) and a GTR+G nucleotide model (partitioned by codon position) as outlined in the section above.

Branch-heterogenous codon ASRV and branch-homogeneous d_N/d_S

The simulation scheme above results in heterotachy at the amino acid and nucleotide levels due to d_N/d_S changing differently over the tree for different sets of sites, despite no heterotachy at the codon level since the pattern of codon ASRV is homogeneous. We performed two additional sets of simulations similar to the above, but with the following differences. For the first set, we allowed codon site rates (governed by a discretized gamma ASRV model) to vary across the tree. We used the same breakpoints as those used for parameters of the MutSel+M5 model. Site rate values for each subset of branches were based on posterior samples of the discretized gamma ASRV shape parameter (α) associated with the same empirical data subset as that used for MutSel+M5 parameters (e.g. a subset of branches would have the shape parameter and substitution matrix parameter values all taken from analysis of DENV data). For the second set of simulations, we also used branch-heterogeneous ASRV but kept site d_N/d_S (ω) values constant across the tree, always using the values associated with the "root" branch subset. In both cases, rate (and d_N/d_S, when applicable) categories were partially reordered. If a site belonged to one of the three slowest categories at the root of the tree, it was kept in that category throughout the tree even though the rate itself was allowed to change at breakpoints. Conversely, sites in other categories (e.g. starting out in the second fastest bin) randomly switched categories at breakpoints.

Calculation of test statistics and quantification of topology and tree length estimation accuracy

We calculated the multinomial likelihood test statistic following equation 7 in Bollback (2002). The chi-squared test statistic was calculated on a contingency table of character state counts per taxon, basing the expected frequency of a state on its average frequency across the entire alignment.

All models used for inference in this study are homogeneous across the phylogeny, while some models used for simulating data are not. We expect alignments simulated under models violating the homogeneity assumption to have larger values of chi-squared compared the expectation under a homogeneous model. We therefore calculated one-tailed posterior predictive p-values when using the chi-squared statistic. Conversely, we calculated 2-tailed posterior predictive p-values when using multinomial likelihood by finding a Highest Density Interval (HDI) on the posterior predictive distribution such that one of its boundaries corresponds to the multinomial likelihood of the alignment on which inference was performed, then taking the mass lying outside that HDI. HDI calculations were made using the HDInterval (Meredith & Kruschke, 2020) version 0.2.2 R package. Effect sizes for both statistics were calculated as the absolute value of the difference between the value of the statistic calculated on the alignment used for inference and the median of the posterior predictive distribution, divided by the standard deviation of the posterior predictive distribution.

We measured topology inference accuracy by determining whether the true topology is covered by the 95% highest posterior density set and by estimating the tree distance between the true topology and the MAP topology and the average distance between the true topology and every topology in the posterior sample. We used the following distance metrics: Robinson-Foulds (RF)

distance (Robinson & Foulds, 1981), approximate Subtree Prune-Regraft (SPR) distance (Hein, 1990; de Oliveira Martins, 2008), and Clustering Information Distance (CID) (Smith, 2020a) calculated using the phangorn (Schliep, 2011) version 2.7.1 and TreeDist (Smith, 2020b) version 2.0.3 R packages. Since SPR distances generally followed the same patterns as RF distances, we did not report them in the main text, figures, and tables, though they are included in Supporting Information tables.

To assess tree length inference accuracy, we calculated (1) the percent difference between the posterior mean tree length and the true tree length, (2) the difference between the posterior mean tree length and the true tree length divided by the standard deviation of the posterior distribution, and (3,4) the absolute values of the previous two metrics. For inference on nucleotide data generated by codon models, the true nucleotide tree length was calculated as the codon tree length divided by 3. The amino acid tree length was calculated using the following procedure on the codon instantaneous rate matrix: (1) multiply each instantaneous rate by the stationary frequency of the originating codon to get a "weighted" rate, (2) calculate the ratio of the sum of weighted rates associated with non-synonymous substitutions to the sum of weighted rates associated with synonymous substitutions, and (3) multiply this ratio by the codon tree length.

We used the two-sample Cramer-Von Mises test (Cramér, 1928; Von Mises, 1928; Anderson, 1962) as implemented in the twosample (Dowd, 2020) R package to determine whether sampled test statistics associated with different analyses could be considered to have different distributions. In the case of binary statistics (e.g. whether the true tree is covered), we used Fisher's exact test (Fisher, 1922). Since we performed multiple statistical tests comparing the results of nucleotide and amino acid analyses for each batch of simulations, we used the Benjamini-Yekutieli procedure (Benjamini & Yekutieli, 2001) as implemented in the stats (Core

R Team, 2020) version 4.1.2 R package to control false discovery rate. In order to reduce the number of redundant comparisons, we did not calculate p-values for RF distances and absolute error statistics. All p-values included in figures are uncorrected, while the p-values presented in tables are Benjamini-Yekutieli corrected.

MCMC diagnostics

We visually assessed within-run convergence and mixing of a subset of analyses using Tracer (Rambaut *et al.*, 2018) version 1.7. Since running multiple MCMC chains per inference run was too computationally prohibitive given the total number of analyses performed, we only ran independent replicates of a small subset of analyses and used the bonsai R package to calculated convergence statistics for tree topology and numerical parameters.

Output visualization

We used the following R packages to visualize output and generate figures: RColorBrewer (Neuwirth, 2014) version 1.1.2, geometry (Habel *et al.*, 2019) version 0.4.5, plot3D (Soetaert, 2021) version 1.4, ggplot2 (Wickham, 2016) version 3.3.5, dplyr (Wickham *et al.*, 2021) version 1.0.5, forcats (Wickham, 2021) version 0.5.1, grid (Core R Team, 2021) version 4.1.2.

Results

Simulation and inference under a simple amino acid and nucleotide models

Table 2.2 and Fig. 2.4-2.15 summarize statistics associated with inference accuracy and posterior predictive tests. See Table S2.1 for more statistics and raw values. Error in topology inference, as measured by the mean clustering information distance (CID) between every topology in the posterior sample and the true topology (Fig. 2.6), the CID between the MAP tree and the true tree (Fig. 2.7), and how frequently the true topology is covered by the 95% credible set (Fig.

2.12A, B) indicate overall lower accuracy associated with inference under nucleotide models (versus amino acid models) when across-site rate variation (ASRV) is ignored. When using the true model, the fraction of times the true topology was covered was not significantly below 95% except in the case of the nucleotide GTR+G model (exact binomial test p-value = 0.00785) in combination with an alignment length corresponding to 1200 codons (400 second-position nucleotide sites). Using RF distance instead of CID led to qualitatively similar results (Fig. 2.4-2.5).

Tree length estimation error was not significantly different between nucleotide and amino acid models when the simulation model was used for inference, but both nucleotide and amino acid models consistently underestimated the tree length when ASRV was ignored, with nucleotide models being associated with greater error (Fig. 2.8-2.11). In general, the true tree length was covered much less frequently than the true topology (11% at best in the case of tree length versus 72% at worst in the case of topology) when ASRV was ignored (Fig. 2.12).

For the degree of model violation explored here, posterior predictive tests using the multinomial likelihood statistic had high power: approximately 96% for alignment lengths of 1200 and >99% for alignment lengths of 5400 (Table 2.2, Fig. 2.15A, B). The difference in power when using amino acid versus nucleotide models was not significant. However, multinomial likelihood posterior predictive effect sizes were on average higher when using amino acid models both when using the true model and in the presence of model violation (Fig. 2.13). In contrast to the above, tests using the chi-squared statistic were not sensitive to model violation caused by ignoring ASRV, with power being lower than the type I error (Fig. 2.15C, D). Posterior predictive effect sizes were not significantly different for amino acids and nucleotides when the

true model was used, but effect sizes where larger in the case of amino acids when ASRV was ignored (Fig. 2.14), despite negligible power.

Simulation under a branch-heterogenous mutation-selection codon model

Table 2.3 and Fig. 2.16-2.27 summarize statistics associated with inference accuracy and posterior predictive tests. See Tabble S2.2 for more statistics and raw values. Topology inference under amino acid models was more accurate for all tree lengths when measured in terms of how frequently the true topology is covered by the 95% credible set (Fig. 2.24A). Comparisons of the mean CID between every topology in the posterior sample and the true topology (Fig. 2.18) and the CID between the MAP topology and the true topology (Fig. 2.19) point in the same direction, but are only significant (adjusted p-value < 0.05) for runs with tree length equal to 50 expected substitutions per codon. Assessment using RF distance (Fig. 2.16-2.17) followed the same pattern.

When a tree length of 50 expected substitutions per codon was used for simulation, using amino acid models resulted in underestimating the tree length, while using nucleotide models resulted in tree length overestimation (Fig. 2.20, 2.22; top). When true tree length was increased to 100, estimates using nucleotide models became on average almost unbiased, while estimates using amino acid models underestimated the tree length more strongly (Fig. 2.20, 2.22; middle). At 150 expected substitutions per codon, estimates based on both nucleotide and amino acid models were on average lower than the true tree length (Fig. 2.20, 2.22; bottom). Absolute error (Fig. 2.21, 2.23) was greater in the case of nucleotide models for a true tree length of 50, and greater in the case of amino acid models for true tree lengths of 100 and 150. Nucleotide models were in general associated with higher variance in magnitude of error. The true tree length was covered by the 95% credible interval 38% and 23% of the time when using amino acid and nucleotide

models respectively when the simulation tree length was set to 50 substitutions per codon (Fig. 2.24B; left). For true tree lengths of 100 and 150, the true length was covered 74% and 64% of the time respectively for nucleotide models, and was never covered for amino acid models (Fig. 2.24B; middle and right).

Power of posterior predictive tests using the multinomial likelihood statistic was higher for nucleotide than for amino acid models (95% versus 8%) when the simulation tree length was 50 expected substitutions per codon (Fig. 2.27A; left). At tree lengths of 100 and 150, the model was always rejected (Fig. 2.27A; middle and right). Multinomial likelihood posterior predictive effect sizes (Fig. 2.25) followed a similar pattern to tree length estimation error: at a true tree length of 50, nucleotide models were associated with higher effect sizes, while amino acid models were associated with higher effect sizes at true tree lengths of 100 and 150. Power of tests based on the chi-squared statistic was maximal (i.e. 100%) for nucleotide models at all evaluated tree lengths, while it increased from 26% to 46% with increasing tree length for amino acid models (Fig. 2.27B). Both the average and variance of chi-squared effect sizes was much greater with nucleotide models for all tree lengths (Fig. 2.26).

Branch-heterogenous codon ASRV and branch-homogeneous d_N/d_S

Table 2.4 and Fig. 2.28-2.41 summarize statistics associated with inference accuracy and posterior predictive tests. See Table S2.3 for more statistics and raw values. Topology inference accuracy as measured by the mean CID between every topology in the posterior sample and the true topology (Fig. 2.30), the CID between the MAP topology and the true topology (Fig. 2.31), and the fraction of times the true topology was covered (Fig. 2.36) appeared to indicate slightly more error associated with nucleotide analyses across all tree lengths and simulation models when looking at statistic means only. However, these apparent differences were generally not

significant (adjusted p-values > 0.05 in all cases but one; see Table 2.4). This reduction in power compared to the previous set of analyses is expected given the reduced number of simulations per model (100 versus 600).

In the case of data simulated under branch-heterogeneous ASRV, branch-homogeneous d_N/d_S, and a codon tree length of 50, tree length estimates under amino acid analyses were on average more positively biased (i.e. tree length is overestimated) than the estimates under equivalent nucleotide analyses (Fig. 2.32; top). This is in contrast to the case of homogeneous ASRV and heterogeneous d_N/d_S, where estimates based on amino acid models were negatively biased and nucleotide-based estimates were positively biased. When both ASRV and d_N/d_S were branch-heterogeneous, both nucleotide and amino acid analyses overestimated the tree length. In this case, estimates based on amino acids were slightly less biased than estimates based on nucleotide analyses. When data were simulated under tree lengths of 100 and 150 substitutions per codon (Fig. 2.32; middle, bottom), tree lengths were underestimated across all simulation and inference models, with the relative magnitude of error increasing with tree length.

For all tree lengths, tree length estimation error associated with nucleotide analyses was similar across simulation models. Error was on average positive at a true tree length of 50 but became negative at 100 and more negative at 150. Across all tree lengths, amino acid analyses of data simulated under branch-heterogeneous d_N/d_S and ASRV resulted in percent error that was on average in between that associated with simulation models where only d_N/d_S is heterogeneous and models where only ASRV is heterogeneous. For distributions of absolute percent error, see Fig. 2.33. Overall, while all inference models performed more poorly at a true tree length of 150, the presence of d_N/d_S branch-heterogeneity appears to be correlated with greater error in amino acid analyses.

The mean error in estimated tree length relative to the standard deviation of the posterior distribution (Fig. 2.34) was almost always greater in magnitude in the case of amino acid analyses. The only exception was analyses of data generated with branch-heterogeneous ASRV and homogeneous d_N/d_S at a codon tree length of 100, where the difference was not significant (Table 2.4). However, the mean of the absolute error relative to the standard deviation (Fig. 2.35) was slightly higher for nucleotide analyses of data generated under branch-homogeneous ASRV and heterogeneous d_N/d_S at a codon tree length of 50 (Table S2.3). Following this general trend, the true tree length was covered by the 95% credible interval more often for nucleotide than for amino acid analyses in all cases where the difference was significant (Fig. 2.37). There are thus instances, such as with data generated with branch-heterogeneous ASRV and homogeneous d_N/d_S at a codon tree length of 100, where the posterior mean tree length is on average proportionally closer to the true tree length for amino acid analyses (Fig. 2.32-2.33; middle) but at the same time falls outside the 95% credible interval more often.

Multinomial likelihood posterior predictive effect sizes (Fig. 2.38) for amino acid analyses across all tested tree lengths were larger in the case of branch-heterogeneous ASRV with or without heterogeneous d_N/d_S than in the case of homogeneous ASRV and heterogeneous d_N/d_S. Given that amino acid analyses of data generated with homogeneous ASRV and heterogeneous d_N/d_S resulted in higher error in tree length estimation at true codon tree lengths of 100 and 150, lower multinomial likelihood effect size was not always associated with lower error in these cases. Similarly, nucleotide analyses at the same tree lengths of data with d_N/d_S heterogeneity only, had on average greater tree length estimation error than analyses of data with ASRV heterogeneity only, but had lower multinomial likelihood effect sizes. Finally, pairwise comparisons of nucleotide and amino acid analyses of the same codon data displayed higher effect sizes for amino acid analyses, except in the case of data with d_N/d_S heterogeneity only and

a true tree length of 50. The power of posterior predictive tests using the multinomial likelihood statistic (Fig. 2.40) was maximal across all analyses at tree lengths of 100 and 150. At a tree length of 50, power was high (> 84%) for both amino acid and nucleotide analyses when data were simulated with heterogeneous ASRV. However, power was very low (3% on average) for amino acid analyses of data with d_N/d_S heterogeneity only, while it was high (93% on average) for the corresponding nucleotide analyses.

Chi-squared effect sizes (Fig. 2.39) for amino acid analyses were similar across simulation models, albeit increasing slightly with larger tree lengths. The same pattern held for nucleotide analyses. However, effect sizes were consistently much larger for nucleotide analyses compared to those associated with amino acid analyses of the same data. Power (Fig. 2.41) was maximal for all nucleotide runs. For amino acid runs, power increased slightly with tree length and was similar across simulation models for a given tree length, although analyses of data simulated with heterogeneous ASRV but homogeneous d_N/d_S were associated with slightly higher power.

Discussion

The ability to detect model violation, as well as the error in inferred topologies and tree lengths, varied across data types, true tree lengths, and properties of the simulation model. Our results partially corroborate and extend the findings of previous studies. Duchêne *et al.* (2017) focus on error in topology inference when using nucleotide data but note that error in tree length estimation was not always correlated with topology error. We observe this in our simulations where, for example, topology inference accuracy stayed roughly the same for the nucleotide GTR+G inference model and data simulated under a branch-heterogeneous codon model as the true tree length was increased, but the percent error in the estimated tree length changed dramatically (e.g. from +8.8% to -22.4%). Moreover, we found that although using amino acid

data generally resulted in more accurate reconstruction of tree topology, it was also associated with larger errors in estimated tree length under some simulation conditions. The former result is compatible with the conventional preference for using amino acid data to reconstruct phylogenies of highly divergent taxa. However, the tree length estimation error observed here has the potential to impact analyses focusing on molecular dating, biogeography, and ancestral state reconstruction, since inaccurately estimated branch lengths can bias estimates of divergence times (Phillips, 2009; Schwartz & Mueller, 2010). As such, we recommend using both amino acid and nucleotide coding for dating analyses when possible, and investigating discrepancies in estimates without *a priori* assuming amino acid analyses to be more accurate, especially when using sequences that might be under different selective constraints in different lineages.

Phillips (2009) also demonstrated that the direction of age estimation bias can depend on data coding, with nucleotide coding resulting in overestimates of divergence times and RY coding resulting in underestimates. We demonstrated a similar effect on tree length estimation bias for nucleotide and amino acid coding respectively, although it was limited to a specific combination of true tree length and simulation model. Importantly, we find that the relative magnitude and direction of bias can also be dataset dependent. Our analyses of data simulated under discretized gamma ASRV nucleotide and amino acid models agree with the results of Duchêne *et al.* (2016), who found that amino acid analyses had less biased tree length estimates than nucleotide analyses when ASRV is not accommodated by the inference model. However, the opposite was true when we simulated data under codon models where the strength of selection, and thus amino acid substitution rates, was allowed to vary across sites and branches.

When analyzing empirical data, the primary concern is usually with some aspect of model performance, not the ability to reject a false model *per se*, since all models are "wrong" and can

be rejected given enough data (Gelman, 2013). Posterior predictive p-values also cannot distinguish cases where the test statistic calculated from the empirical dataset falls just outside the distribution predicted by the model and cases where model predictions and empirical reality are very different. For these reasons, posterior predictive effect sizes are sometimes used (e.g. by Doyle *et al.,* 2015) as measures of model (in)adequacy in lieu of p-values. Building on this approach, Duchêne *et al.* (2017) find chi-squared effect size thresholds that allow the identification of instances where topology inference is likely to be biased. Our results indicate that if such an approach is to be extended to comparisons of model adequacy in the context of different coding of the same data, different useful thresholds might need to be established for each case. For example, we found that in the case of data simulated with d_N/d_S branch-heterogeneity, multinomial likelihood effect sizes for amino acid models at a true tree length of 100 substitutions per codon were on average very similar to those for nucleotide models at a tree length of 150 (4.1 versus 4.3 standard deviations respectively). However, the associated respective error in estimated tree lengths was very different (a -18.9% difference between the posterior mean and the true tree length, with the true value never covered by the 95% credible interval for amino acid coding, versus a -13.8% difference and the true tree length covered 64% of the time for nucleotide coding). Additionally, a given test statistic might differ subtly in sensitivity to specific features of the data. Although we generally found multinomial likelihood to be sensitive to cases where the inference model does not sufficiently accommodate branch and/or site heterogeneity in rates, this was not always the case. At larger tree lengths, adding d_N/d_S branch-heterogeneity when codon-level ASRV was already present had a strong effect on tree length estimation error using amino acid data, but had a minimal effect on multinomial likelihood effect size.

Conflicts of interest

The authors declare that there are no conflicts of interest.

Acknowledgments

We would like to thank J. Gao (University of California, Davis), N. Lashinsky (University of California, Davis), and B. Rannala (University of California, Davis) for helpful and insightful discussion; current and former FARM HPC cluster sysadmins B. Broadley, P. Osmani, T. Thatcher, and O. Wild for help with using the cluster; S. Winterton (California Department of Food and Agriculture) for providing computational resources used for preliminary analyses; N. Tam for proofreading and providing comments on the manuscript.

References

Altschul, S. F. (1998). Generalized affine gap costs for protein sequence alignment. *Proteins: Structure, Function, and Bioinformatics*, *32*(1), 88-96.

Anderson, T. W. (1962). On the distribution of the two-sample Cramer-von Mises criterion. *The Annals of Mathematical Statistics*, 1148-1159.

Baele, G., Lemey, P., Bedford, T., Rambaut, A., Suchard, M. A., & Alekseyenko, A. V. (2012a). Improving the accuracy of demographic and molecular clock model comparison while accommodating phylogenetic uncertainty. *Molecular biology and evolution*, *29*(9), 2157-2167.

Baele, G., Li, W. L. S., Drummond, A. J., Suchard, M. A., & Lemey, P. (2012b). Accurate model selection of relaxed molecular clocks in Bayesian phylogenetics. *Molecular biology and evolution*, *30*(2), 239-243.

Benjamini, Y., & Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of statistics*, 1165-1188.

Blanquart, S., & Lartillot, N. (2008). A site-and time-heterogeneous model of amino acid replacement. *Molecular biology and evolution*, *25*(5), 842-858.

Bollback, J. P. (2002). Bayesian model adequacy and choice in phylogenetics. *Molecular Biology and Evolution*, *19*(7), 1171-1180.

Brown, J. M. (2014a). Detection of implausible phylogenetic inferences using posterior predictive assessment of model fit. *Systematic biology*, *63*(3), 334-348.

Brown, J. M. (2014b). Predictive approaches to assessing the fit of evolutionary models. *Systematic biology*, *63*(3), 289-292.

Chiari, Y., Cahais, V., Galtier, N., & Delsuc, F. (2012). Phylogenomic analyses support the position of turtles as the sister group of birds and crocodiles (Archosauria). *Bmc Biology*, *10*(1), 1-15.

Chernomor, O., Von Haeseler, A., & Minh, B. Q. (2016). Terrace aware data structure for phylogenomic inference from supermatrices. *Systematic biology*, *65*(6), 997-1008.

Comandatore, F., Sassera, D., Montagna, M., Kumar, S., Koutsovoulos, G., Thomas, G., ... & Blaxter, M. (2013). Phylogenomics and analysis of shared genes suggest a single transition to mutualism in Wolbachia of nematodes. *Genome biology and evolution*, *5*(9), 1668-1674.

Core R Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Core R Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Cramér, H. (1928). On the composition of elementary errors: First paper: Mathematical deductions. *Scandinavian Actuarial Journal*, *1928*(1), 13-74.

de Oliveira Martins, L., Leal, E., & Kishino, H. (2008). Phylogenetic detection of recombination with a Bayesian prior on the distance between trees. *PLoS One*, *3*(7), e2651.

Dowd, C. (2020). twosamples: Fast Permutation Based Two Sample Tests. R package version 1.1.1. URL: https://CRAN.R-project.org/package=twosamples

Doyle, V. P., Young, R. E., Naylor, G. J., & Brown, J. M. (2015). Can we identify genes with increased phylogenetic reliability?. *Systematic biology*, *64*(5), 824-837.

Duchêne, S., Di Giallonardo, F., & Holmes, E. C. (2016). Substitution model adequacy and assessing the reliability of estimates of virus evolutionary rates and time scales. *Molecular biology and evolution*, *33*(1), 255-267.

Duchêne, D. A., Duchêne, S., & Ho, S. Y. (2017). New statistical criteria detect phylogenetic bias caused by compositional heterogeneity. *Molecular Biology and Evolution*, *34*(6), 1529-1534.

Fan, Y., Wu, R., Chen, M. H., Kuo, L., & Lewis, P. O. (2011). Choosing among partition models in Bayesian phylogenetics. *Molecular biology and evolution*, *28*(1), 523-532.

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

Fisher, R. A. (1922). On the interpretation of χ 2 from contingency tables, and the calculation of P. *Journal of the Royal Statistical Society*, *85*(1), 87-94.

Foster, P. G. (2004). Modeling compositional heterogeneity. *Systematic biology*, *53*(3), 485-495.

Gelman, A. (2013). Two simple examples for understanding posterior p-values whose distributions are far from uniform. *Electronic Journal of Statistics*, *7*, 2595-2602.

Gillung, J. P., Winterton, S. L., Bayless, K. M., Khouri, Z., Borowiec, M. L., Yeates, D., ... & Wiegmann, B. M. (2018). Anchored phylogenomics unravels the evolution of spider flies (Diptera, Acroceridae) and reveals discordance between nucleotides and amino acids. *Molecular phylogenetics and evolution*, *128*, 233-245.

Goldman, N. (1993). Statistical tests of models of DNA substitution. *Journal of molecular evolution*, *36*(2), 182-198.

Habel, K., Grasman, R., Gramacy, R. B., Mozharovskyi, P., & Sterratt, D. C. (2019). geometry: Mesh Generation and Surface Tessellation. R package version 0.4.5. URL: https://CRAN.R-project.org/package=geometry

Hein, J. (1990). Reconstructing evolution of sequences subject to recombination using parsimony. *Mathematical biosciences*, *98*(2), 185-200.

Höhna, S., Coghill, L. M., Mount, G. G., Thomson, R. C., & Brown, J. M. (2018). P3: Phylogenetic posterior prediction in RevBayes. *Molecular biology and evolution*, *35*(4), 1028-1034.

Höhna, S., Heath, T. A., Boussau, B., Landis, M. J., Ronquist, F., & Huelsenbeck, J. P. (2014). Probabilistic graphical model representation in phylogenetics. *Systematic biology*, *63*(5), 753-771.

Höhna, S., Landis, M. J., Heath, T. A., Boussau, B., Lartillot, N., Moore, B. R., ... & Ronquist, F. (2016). RevBayes: Bayesian phylogenetic inference using graphical models and an interactive model-specification language. *Systematic biology*, *65*(4), 726-736.

Huelsenbeck, J. P., Ronquist, F., Nielsen, R., & Bollback, J. P. (2001). Bayesian inference of phylogeny and its impact on evolutionary biology. *science*, *294*(5550), 2310-2314.

Jarvis, E. D., Mirarab, S., Aberer, A. J., Li, B., Houde, P., Li, C., ... & Avian Phylogenomics Consortium. (2015). Phylogenomic analyses data of the avian phylogenomics project. *GigaScience*, *4*(1), s13742-014.

Jones, D. T., Taylor, W. R., & Thornton, J. M. (1992). The rapid generation of mutation data matrices from protein sequences. *Bioinformatics*, *8*(3), 275-282.

Kassambara, A., & Mundt, F. (2020). factoextra: Extract and Visualize the Results of Multivariate Data Analyses. R package version 1.0.7. URL: https://CRAN.R-project.org/package=factoextra

Katoh, K., & Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular biology and evolution*, *30*(4), 772-780.

Lartillot, N., & Philippe, H. (2006). Computing Bayes factors using thermodynamic integration. *Systematic biology*, *55*(2), 195-207.

Le, S. Q., & Gascuel, O. (2008). An improved general amino acid replacement matrix. *Molecular biology and evolution*, *25*(7), 1307-1320.

Lewis, P. O., Xie, W., Chen, M. H., Fan, Y., & Kuo, L. (2014). Posterior predictive Bayesian phylogenetic model selection. *Systematic biology*, *63*(3), 309-321.

May, M. R. & Moore, B. R. (2017). bonsai: Automated phylogenetic MCMC diagnosis. R package version 0.9. URL: https://github.com/mikeryanmay/bonsai

May, M. R., Contreras, D. L., Sundue, M. A., Nagalingum, N. S., Looy, C. V., & Rothfels, C. J. (2021). Inferring the Total-Evidence Timescale of Marattialean Fern Evolution in the Face of Model Sensitivity. Systematic Biology, 70(6), 1232-1255.

Meredith, M., and Kruschke, J. (2020). HDInterval: Highest (Posterior) Density Intervals. R package version 0.2.2. URL: https://CRAN.R-project.org/package=HDInterval

Minh, B. Q., Schmidt, H. A., Chernomor, O., Schrempf, D., Woodhams, M. D., Von Haeseler, A., & Lanfear, R. (2020). IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era. *Molecular biology and evolution*, *37*(5), 1530-1534.

Moureau, G., Cook, S., Lemey, P., Nougairede, A., Forrester, N. L., Khasnatinov, M., ... & De Lamballerie, X. (2015). New insights into flavivirus evolution, taxonomy and biogeographic history, extended by analysis of canonical and alternative coding sequences. *PloS one*, *10*(2), e0117849.

Neuwirth, E., (2014). RColorBrewer: ColorBrewer Palettes. R package version 1.1-2. URL: https://CRAN.R-project.org/package=RColorBrewer

Nguyen, L. T., Schmidt, H. A., Von Haeseler, A., & Minh, B. Q. (2015). IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular biology and evolution*, *32*(1), 268-274.

O'Hara, T. D., Hugall, A. F., Thuy, B., Stöhr, S., & Martynov, A. V. (2017). Restructuring higher taxonomy using broad-scale phylogenomics: the living Ophiuroidea. *Molecular phylogenetics and evolution*, *107*, 415-430.

Phillips, M. J. (2009). Branch-length estimation bias misleads molecular dating for a vertebrate mitochondrial phylogeny. *Gene*, *441*(1-2), 132-140.

Ran, J. H., Shen, T. T., Wang, M. M., & Wang, X. Q. (2018). Phylogenomics resolves the deep phylogeny of seed plants and indicates partial convergent or homoplastic evolution between Gnetales and angiosperms. *Proceedings of the Royal Society B*, *285*(1881), 20181012.

Rambaut, A., Drummond, A. J., Xie, D., Baele, G., & Suchard, M. A. (2018). Posterior summarization in Bayesian phylogenetics using Tracer 1.7. *Systematic biology*, *67*(5), 901.

Redelings, B. D. (2021). BAli-Phy version 3: Model-based co-estimation of alignment and phylogeny. *Bioinformatics*, *37*(18), 3032-3034.

Redelings, B. D., & Suchard, M. A. (2007). Incorporating indel information into phylogeny estimation for rapidly emerging pathogens. *BMC evolutionary biology*, *7*(1), 1-19.

Robinson, D. F., & Foulds, L. R. (1981). Comparison of phylogenetic trees. *Mathematical biosciences*, *53*(1-2), 131-147.

Rodrigue, N., Lartillot, N., & Philippe, H. (2008). Bayesian comparisons of codon substitution models. *Genetics*, *180*(3), 1579-1591.

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

Ruhfel, B. R., Gitzendanner, M. A., Soltis, P. S., Soltis, D. E., & Burleigh, J. G. (2014). From algae to angiosperms–inferring the phylogeny of green plants (Viridiplantae) from 360 plastid genomes. *BMC evolutionary biology*, *14*(1), 1-27.

Schliep, K. P. (2011). phangorn: phylogenetic analysis in R. *Bioinformatics*, *27*(4), 592-593.

Schwartz, R. S., & Mueller, R. L. (2010). Branch length estimation and divergence dating: estimates of error in Bayesian and maximum likelihood frameworks. *BMC Evolutionary Biology*, *10*(1), 1-21.

Schwarz, G. (1978). Estimating the dimension of a model. *The annals of statistics*, 461-464.

Smith, M. R. (2020a). Information theoretic generalized Robinson–Foulds metrics for comparing phylogenetic trees. *Bioinformatics*, *36*(20), 5007-5013.

Smith, M. R. (2020b). TreeDist: distances between phylogenetic trees. *Comprehensive R Archive Network*.

Soetaert, K. (2021). plot3D: Plotting Multi-Dimensional Data. R package version 1.4. URL https://CRAN.R-project.org/package=plot3D

Tavaré, S. (1986). Some probabilistic and statistical problems in the analysis of DNA sequences. *Lectures on mathematics in the life sciences*, *17*(2), 57-86.

Von Mises, R. (1928). Statistik und wahrheit. *Julius Springer*, *20*.

Whelan, S., & Goldman, N. (2001). A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Molecular biology and evolution*, *18*(5), 691-699.

Wickham, H. (2016). *ggplot2: elegant graphics for data analysis*. springer.

Wickham, H. (2021). forcats: Tools for Working with Categorical Variables (Factors). R package version 0.5.1. URL: https://CRAN.R-project.org/package=forcats

Wickham, H., François, R., Henry, L., & Müller, K. (2021). dplyr: A Grammar of Data Manipulation. R package version 1.0.5. URL: https://CRAN.R-project.org/package=dplyr

Williams, T. A., Cox, C. J., Foster, P. G., Szöllősi, G. J., & Embley, T. M. (2020). Phylogenomics provides robust support for a two-domains tree of life. *Nature ecology & evolution*, *4*(1), 138-147.

Xie, W., Lewis, P. O., Fan, Y., Kuo, L., & Chen, M. H. (2011). Improving marginal likelihood estimation for Bayesian phylogenetic model selection. *Systematic biology*, *60*(2), 150-160.

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

Yang, Z., Nielsen, R., Goldman, N., & Pedersen, A. M. K. (2000). Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics*, *155*(1), 431-449.

Yang, Z. (2007). PAML 4: phylogenetic analysis by maximum likelihood. *Molecular biology and evolution*, *24*(8), 1586-1591.

Yang, Z., & Nielsen, R. (2008). Mutation-selection models of codon substitution and their use to estimate selective strengths on codon usage. *Molecular biology and evolution*, *25*(3), 568-579.

Yang, Z., Nielsen, R., Goldman, N., & Pedersen, A. M. K. (2000). Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics*, *155*(1), 431-449.

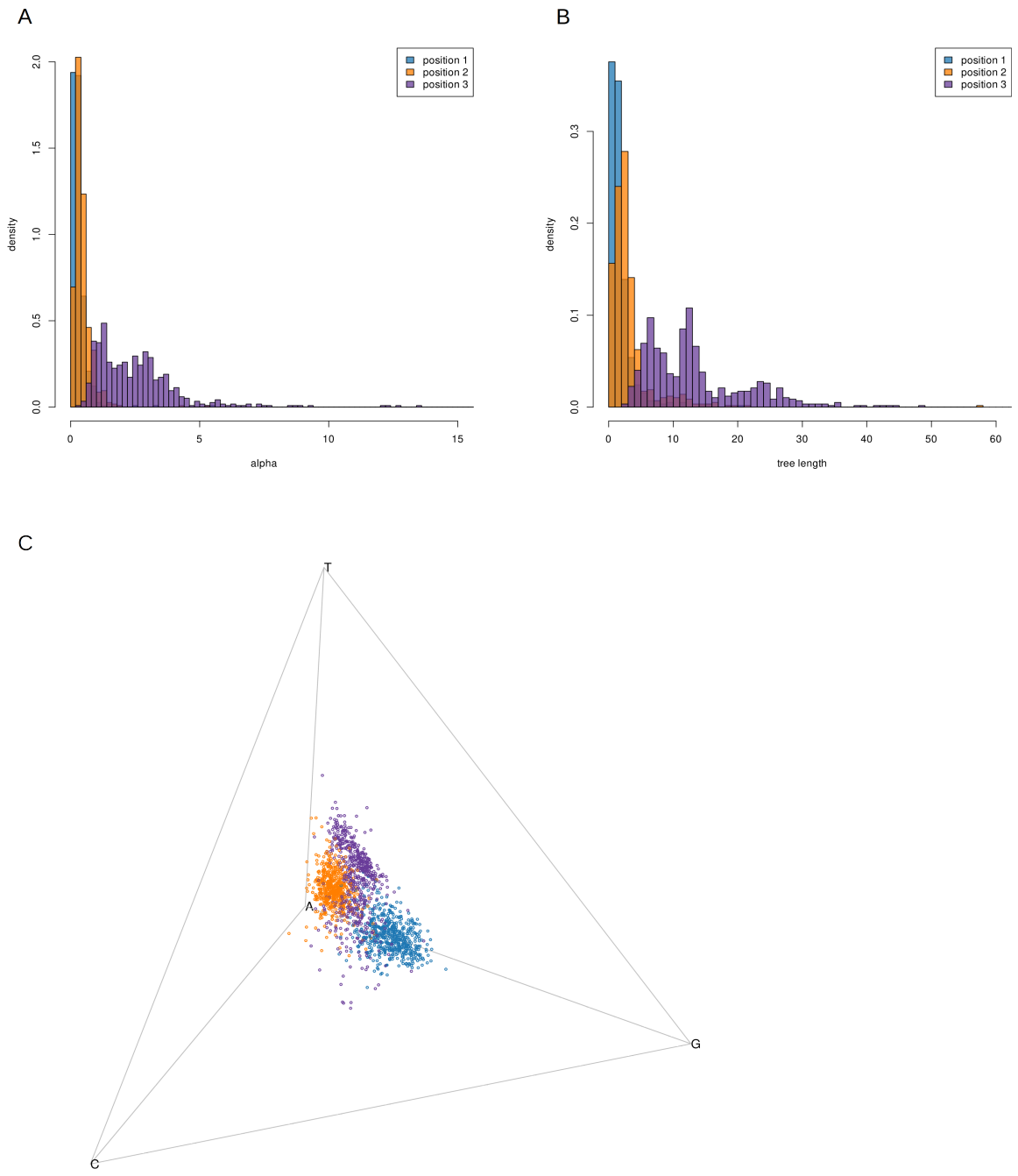

Figure 2.1. Plots of MLEs from analyses of empirical nucleotide data under the GTR+G model. (A) alpha, the shape parameter of the discretized-gamma ASRV model; (B) tree lengths in expected number of nucleotide substitutions per site; (C) nucleotide stationary frequencies.

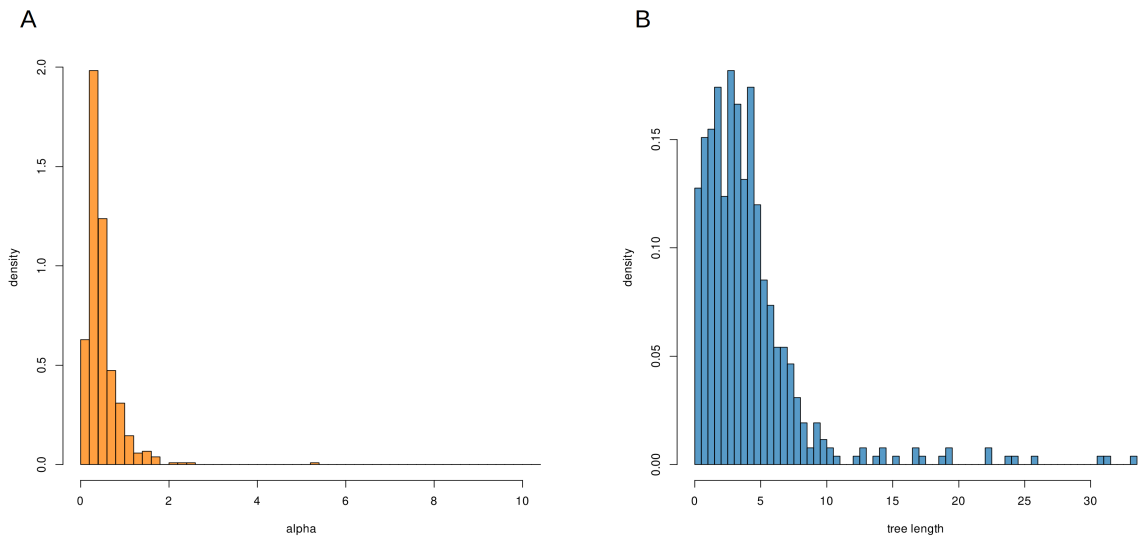

Figure 2.2. Histograms of MLEs from analyses of empirical amino acid data under the GTR+G model. (A) alpha, the shape parameter of the discretized-gamma ASRV model; (B) tree lengths in expected number of amino acid substitutions per site.

Clade and reference	Number of loci in original dataset	Number of taxa in original dataset	Number of loci retained	Number of taxa retained	Amino acid and nucleotide data match
Acrocerids (Gillung *et al.* 2018)	240	50	51	50	Yes
Amniotes (Chiari *et al.*, 2012)	248	16	54 (AA) 85 (Nt)	16	No
Birds (Jarvis *et al.*, 2015)	8295	52	104	52	Yes
Brittle stars (O'Hara *et al.*, 2017)	1484	576	76	224	Yes
Seed plants (Ran *et al.*, 2018)	1308	38	170	38	Yes
Plastids (Ruhfel *et al.*, 2014)	78	360	31	143	No
Wolbachia (Comandatore *et al.*, 2013)	90	16	31	16	No

Table 2.1. Datasets used to estimate parameters for downstream simulations.

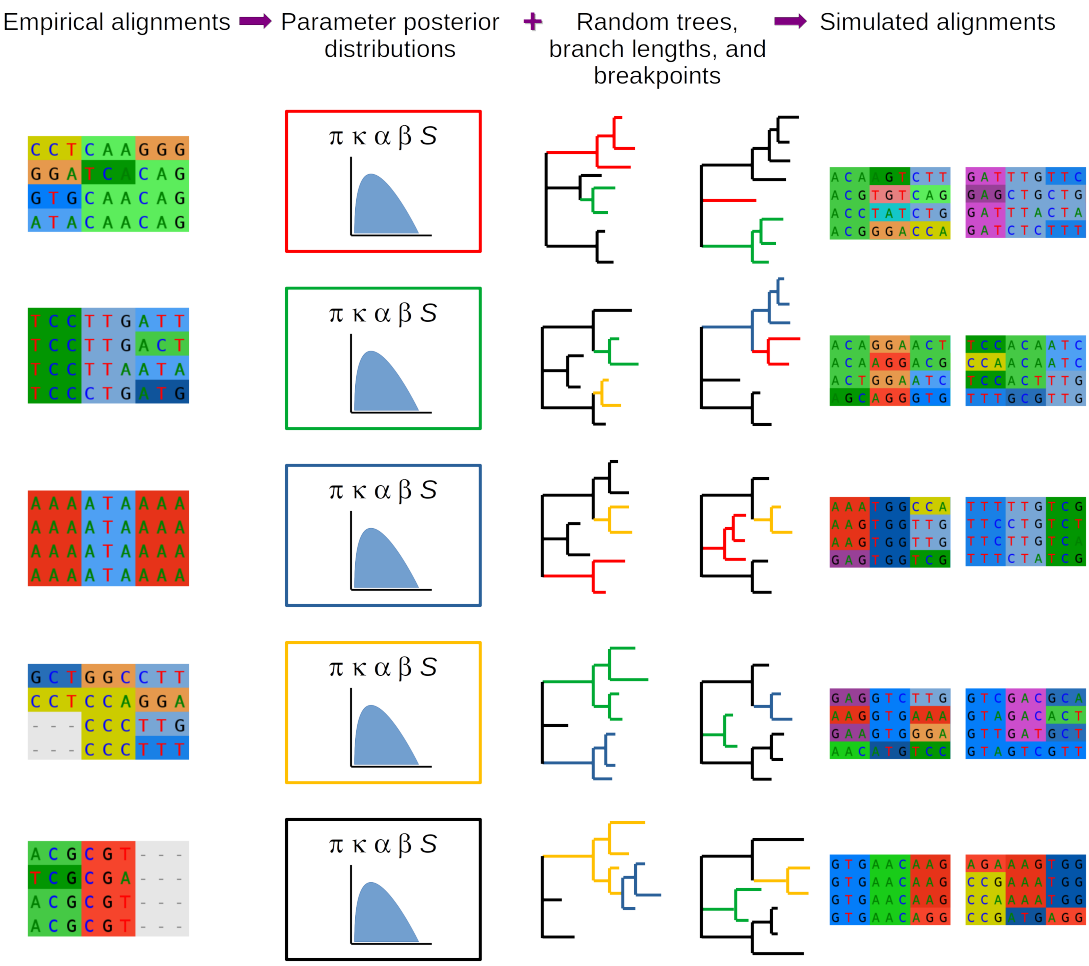

Figure 2.3. Procedure for simulating data under a branch-heterogeneous MutSel model using empirically plausible parameter values. Alignment graphics generated using AliView.

Inference	mean CID to true	CID MAP to true	true topology covered	TL % mean to true	TL mean to true over sd	true TL covered	Multinomial likelihood effect size	Multinomial likelihood reject	Chi-squared effect size	Chi-squared reject
AA WAG+G 1200	1.37 (1.46) *0.000*	1.09 (1.55) *0.000*	0.938 (0.241) 1.000	-0.3% (9.5) 0.788	0.0 (1.1) 0.550	0.921 (0.270) 0.955	0.44 (0.40) *0.000*	0.008 (0.088) 1.000	0.76 (0.60) 0.550	0.039 (0.193) 1.000
Nt GTR+G 1200	2.05 (1.83)	1.53 (1.89)	0.925 (0.264)	-0.3% (8.5)	0.0 (1.2)	0.944 (0.230)	0.34 (0.32)	0.004 (0.062)	0.71 (0.59)	0.039 (0.193)
AA WAG+G 5400	0.40 (0.69) *0.000*	0.32 (0.77) *0.001*	0.990 (0.098) 0.076	-0.3% (4.6) 0.709	0.0 (1.1) 0.605	0.917 (0.276) 1.000	0.43 (0.43) *0.000*	0.019 (0.138) 0.572	0.79 (0.61) 0.056	0.043 (0.202) 1.000
Nt GTR+G 5400	0.69 (0.83)	0.51 (0.93)	0.965 (0.183)	-0.9% (10.4)	-0.1 (1.9)	0.934 (0.248)	0.31 (0.43)	0.006 (0.076)	0.71 (0.56)	0.033 (0.179)
AA WAG 1200	1.43 (1.50) *0.000*	1.28 (1.56) *0.000*	0.793 (0.406) 0.051	22.5% (13.8) *0.000*	10.4 (9.7) *0.000*	0.112 (0.316) 0.122	10.48 (4.60) *0.000*	0.956 (0.206) 1.000	1.20 (0.72) *0.000*	0.002 (0.044) 1.000
Nt GTR 1200	2.17 (1.90)	1.91 (1.99)	0.718 (0.451)	30.8% (14.3)	13.4 (10.7)	0.068 (0.251)	8.12 (4.12)	0.959 (0.198)	0.67 (0.47)	0.008 (0.088)
AA WAG 5400	0.44 (0.74) *0.000*	0.41 (0.83) *0.000*	0.882 (0.323) *0.000*	22.1% (11.7) *0.000*	21.8 (20.2) *0.000*	0.058 (0.234) 0.227	23.28 (8.41) *0.000*	0.988 (0.107) 1.000	1.31 (0.72) *0.000*	0.002 (0.044) 1.000
Nt GTR 5400	0.81 (0.97)	0.75 (1.04)	0.772 (0.420)	30.5% (13.4)	28.6 (22.7)	0.029 (0.168)	17.22 (7.28)	0.992 (0.088)	0.71 (0.48)	0.008 (0.088)

Table 2.2. Topology and tree length inference accuracy and posterior predictive test statistics using amino acid and nucleotide models. Simulation was done using WAG+G for amino acid data and GTR+G for nucleotide data. Displayed values are averages across all performed simulation/inference runs. Standard deviations are in parentheses. Benjamini-Yekutieli adjusted p-values for amino acid versus nucleotide comparisons are listed below the standard deviation of each amino acid entry. Adjusted p-values smaller than 0.05 are italicized and in red.

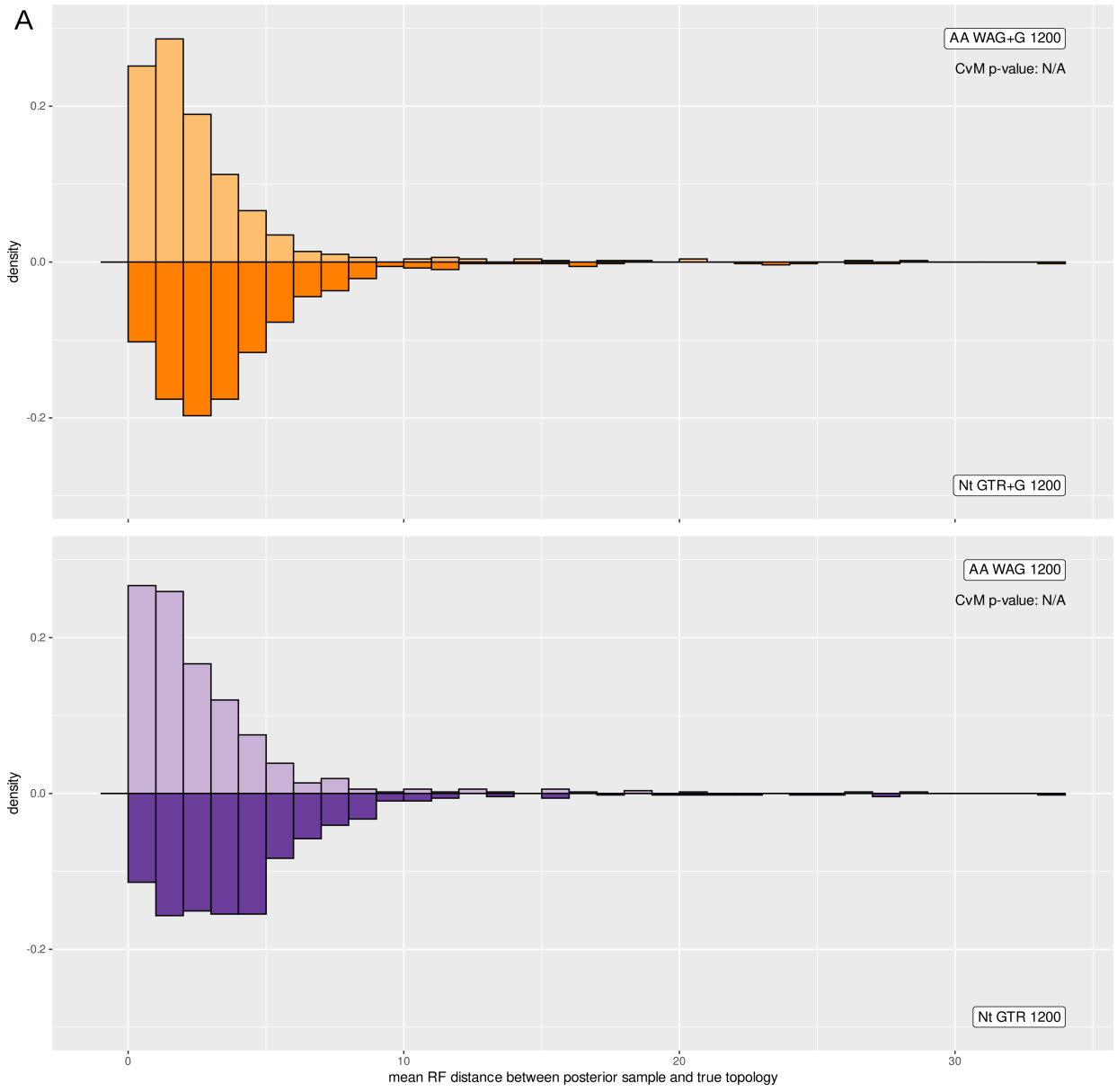

Figure 2.4A. Distributions, across 517 inference runs on different simulated datasets, of the mean RF distance between every topology in the posterior sample and the true topology. Labels in the right corner of each histogram indicate the data type, inference model, and alignment length in equivalent number of codons. P-values comparing inference runs using amino acid versus nucleotide models on the same underlying data were not calculated for this statistic.

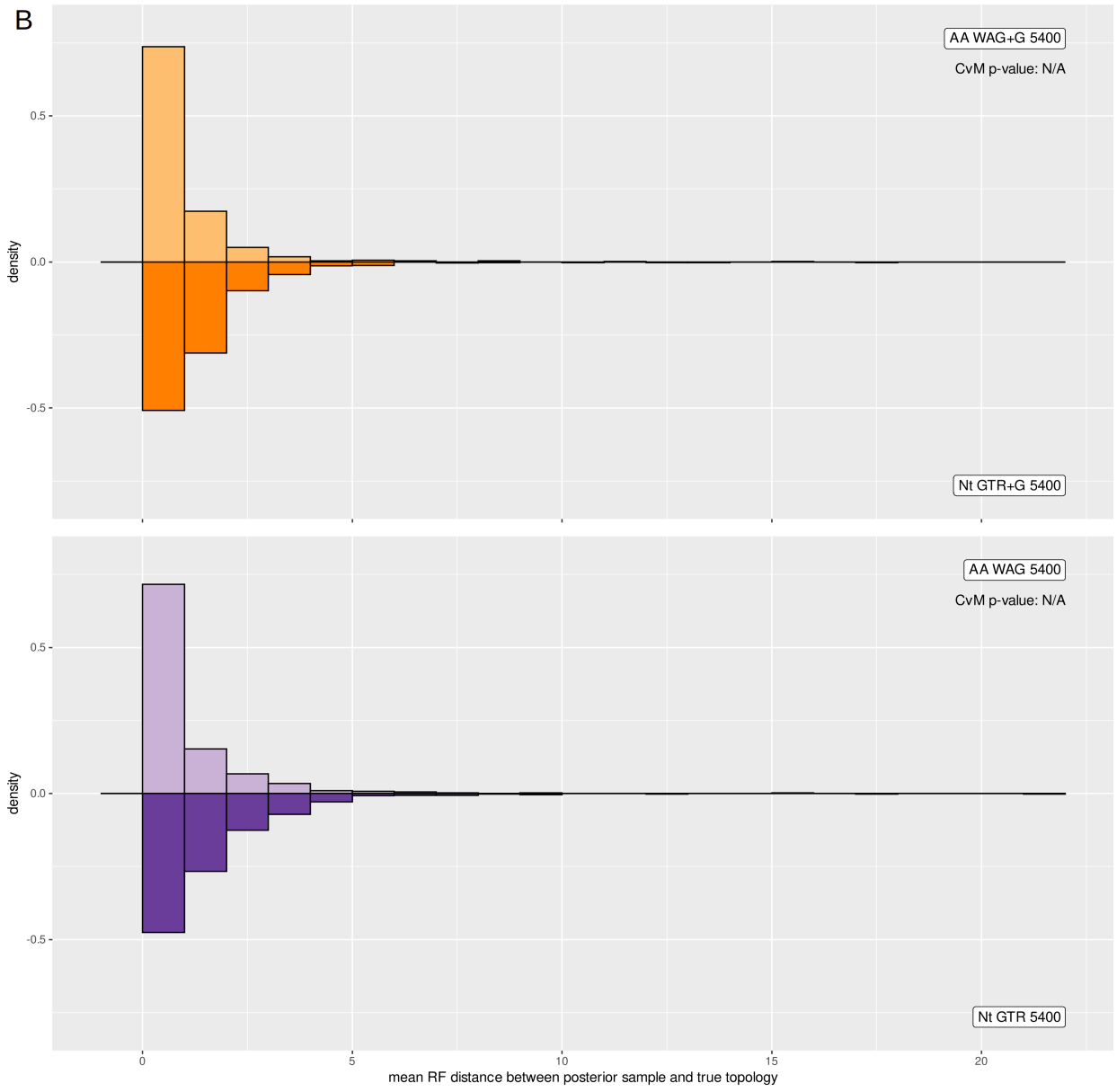

Figure 2.4B. Distributions, across 517 inference runs on different simulated datasets, of the mean RF distance between every topology in the posterior sample and the true topology. Labels in the right corner of each histogram indicate the data type, inference model, and alignment length in equivalent number of codons. P-values comparing inference runs using amino acid versus nucleotide models on the same underlying data were not calculated for this statistic.

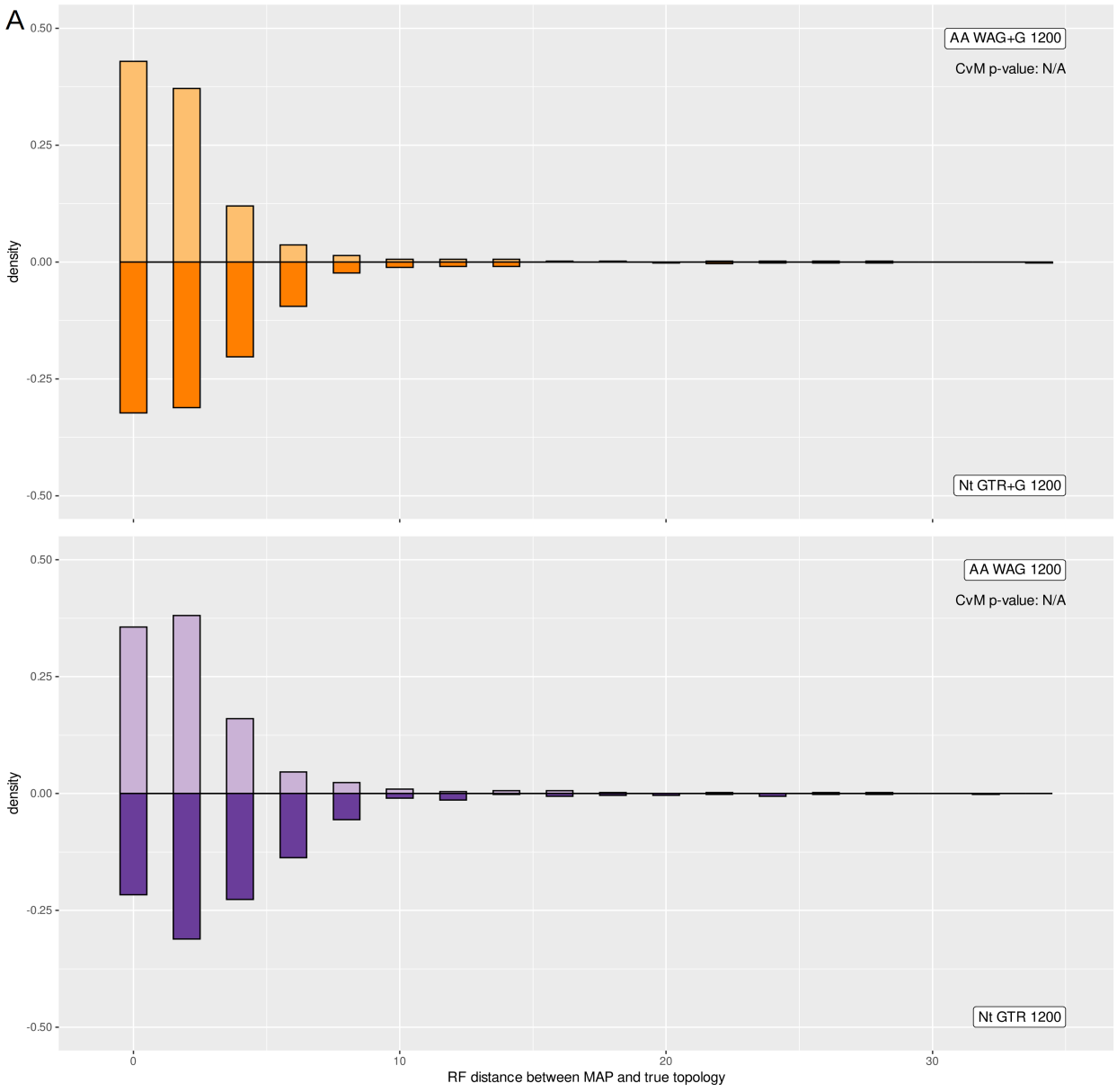

Figure 2.5A. Distributions, across 517 inference runs on different simulated datasets, of the RF distance between the MAP topology and the true topology. Labels in the right corner of each histogram indicate the data type, inference model, and alignment length in equivalent number of codons. P-values comparing inference runs using amino acid versus nucleotide models on the same underlying data were not calculated for this statistic.

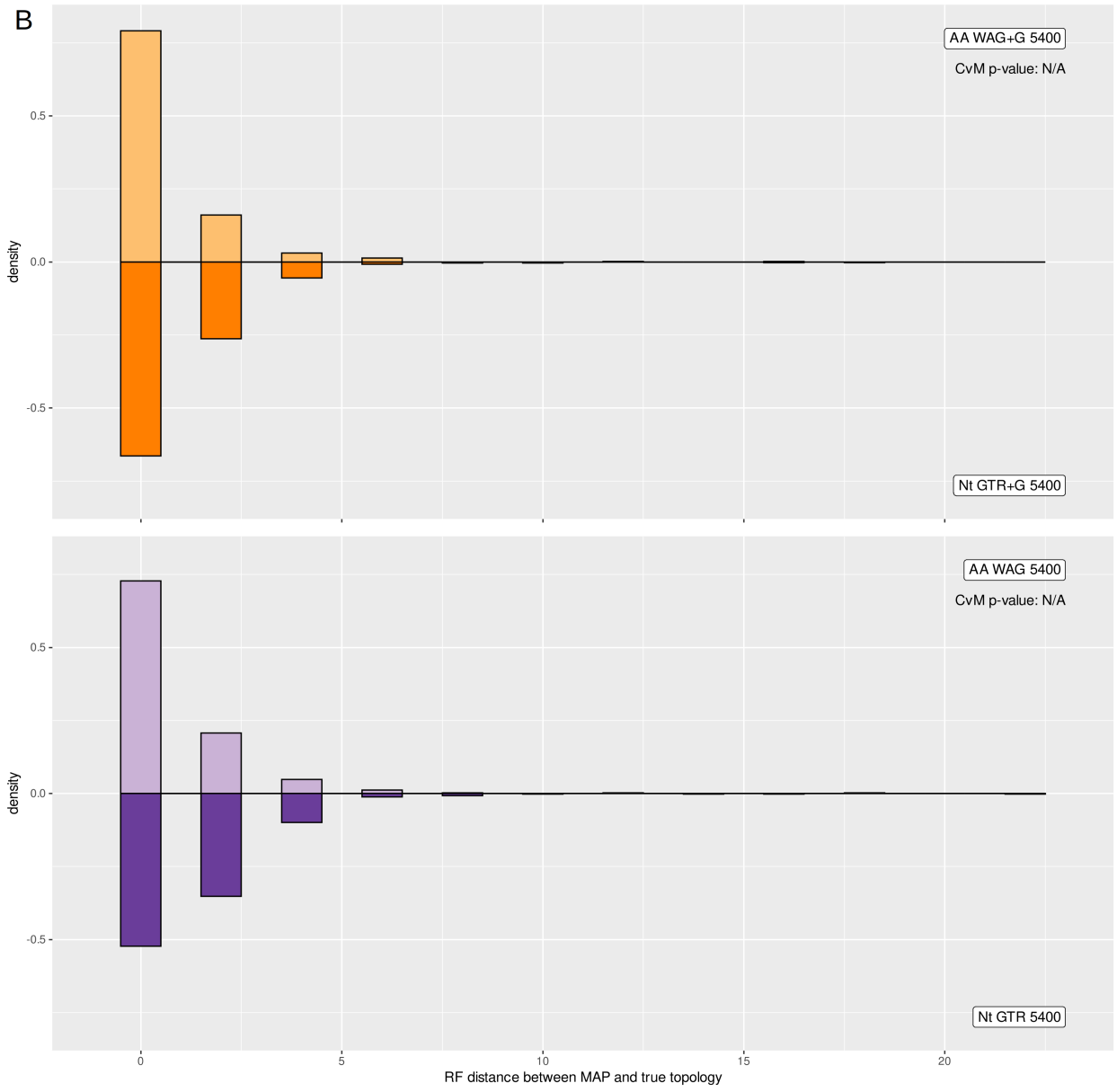

Figure 2.5B. Distributions, across 517 inference runs on different simulated datasets, of the RF distance between the MAP topology and the true topology. Labels in the right corner of each histogram indicate the data type, inference model, and alignment length in equivalent number of codons. P-values comparing inference runs using amino acid versus nucleotide models on the same underlying data were not calculated for this statistic.

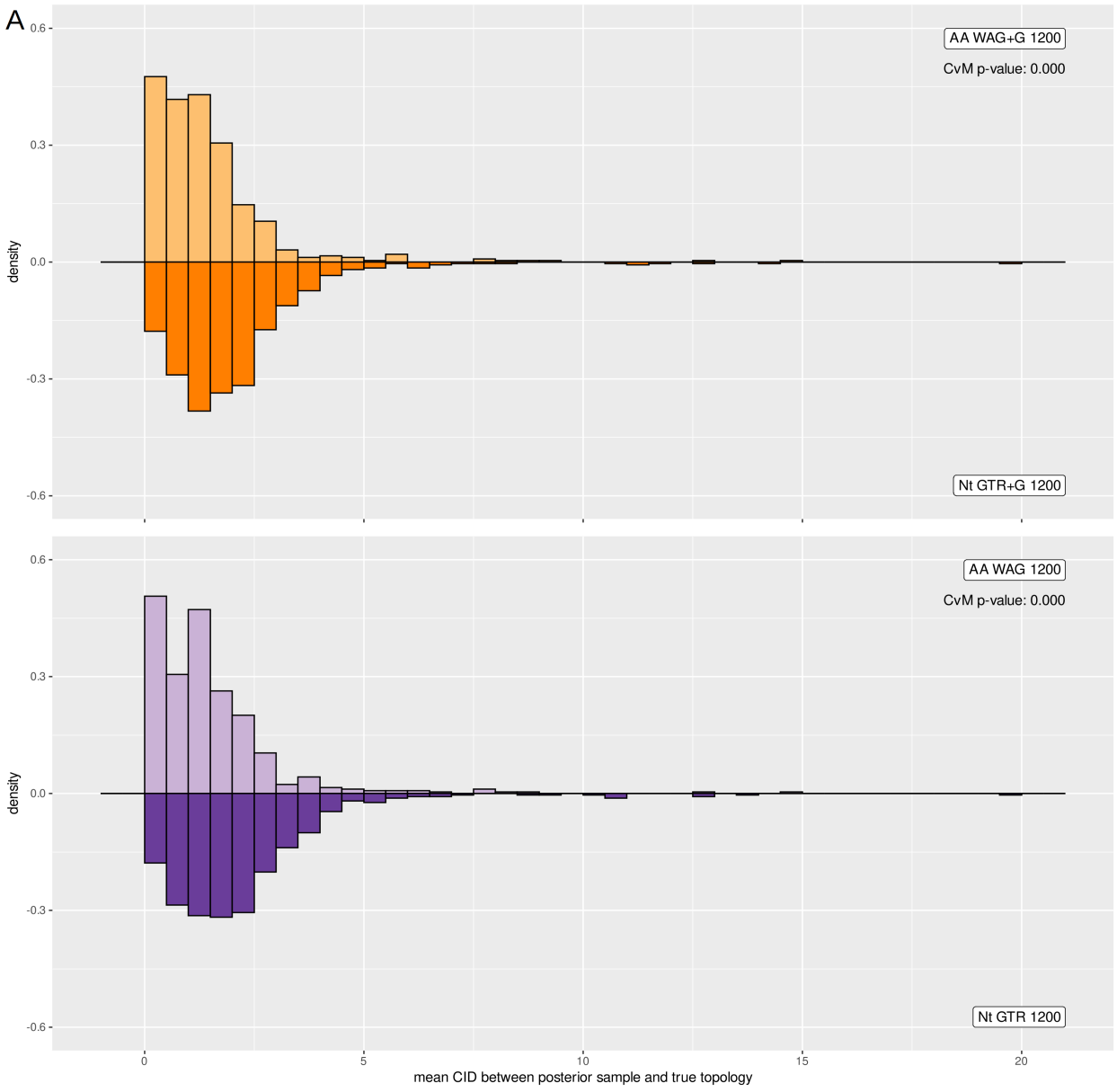

Figure 2.6A. Distributions, across 517 inference runs on different simulated datasets, of the mean CID between every topology in the posterior sample and the true topology. Labels in the right corner of each histogram indicate the data type, inference model, and alignment length in equivalent number of codons. P-values are from two-sample Cramer-Von Mises tests comparing analogous sets of inference runs on amino acid versus nucleotide data.

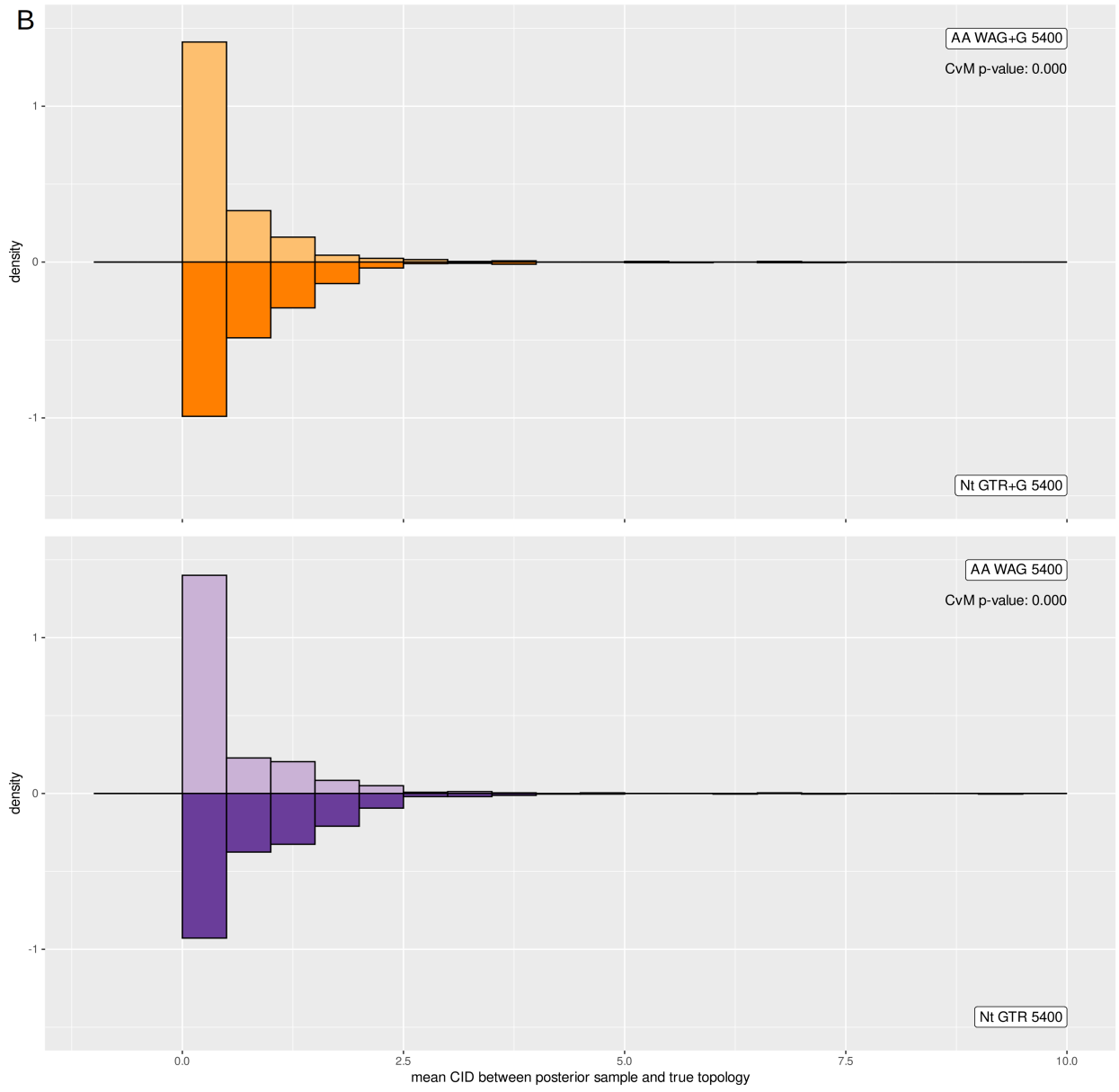

Figure 2.6B. Distributions, across 517 inference runs on different simulated datasets, of the mean CID between every topology in the posterior sample and the true topology. Labels in the right corner of each histogram indicate the data type, inference model, and alignment length in equivalent number of codons. P-values are from two-sample Cramer-Von Mises tests comparing analogous sets of inference runs on amino acid versus nucleotide data.

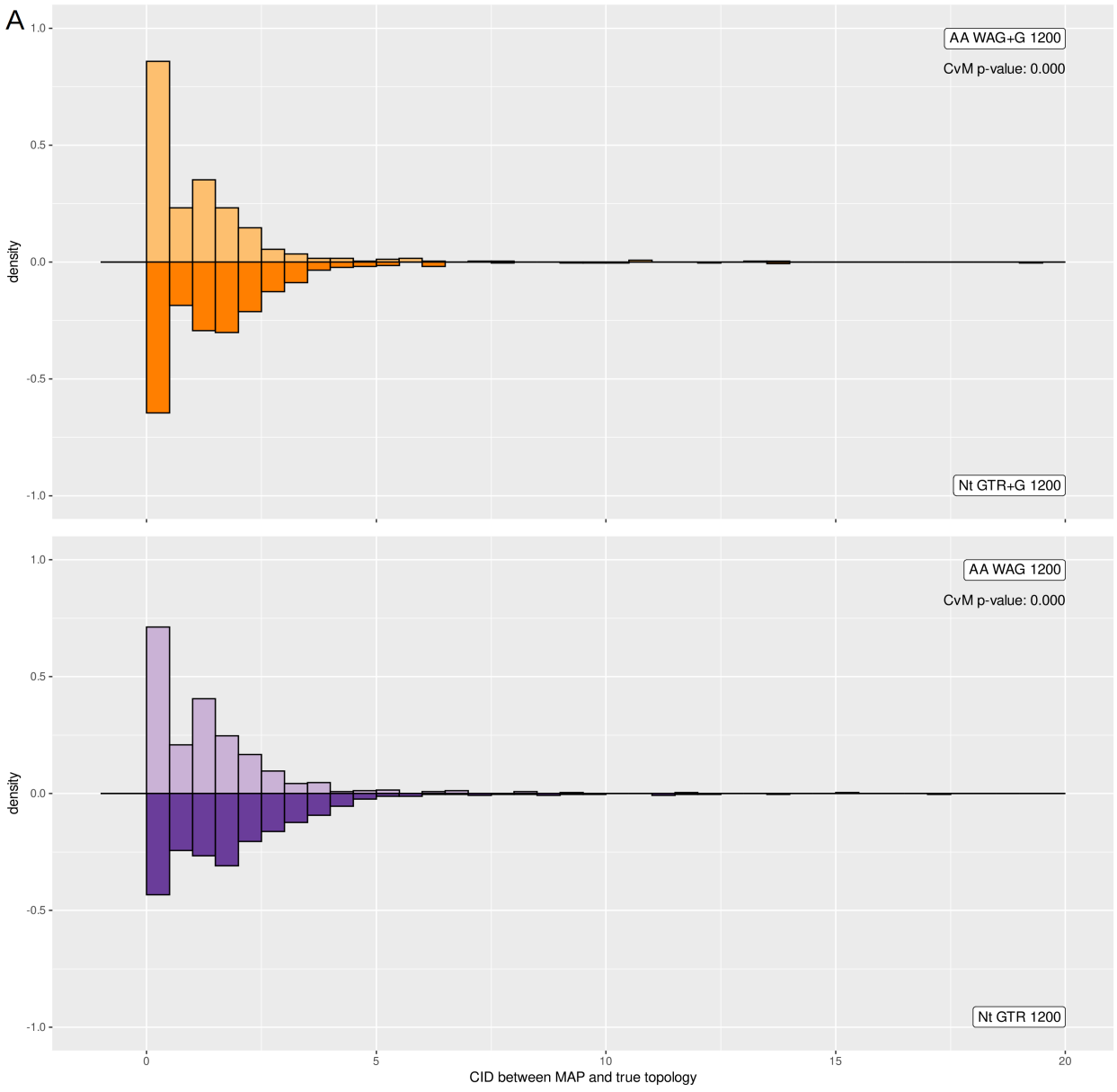

Figure 2.7A. Distributions, across 517 inference runs on different simulated datasets, of the CID between the MAP topology and the true topology. Labels in the right corner of each histogram indicate the data type, inference model, and alignment length in equivalent number of codons. P-values are from two-sample Cramer-Von Mises tests comparing analogous sets of inference runs on amino acid versus nucleotide data.

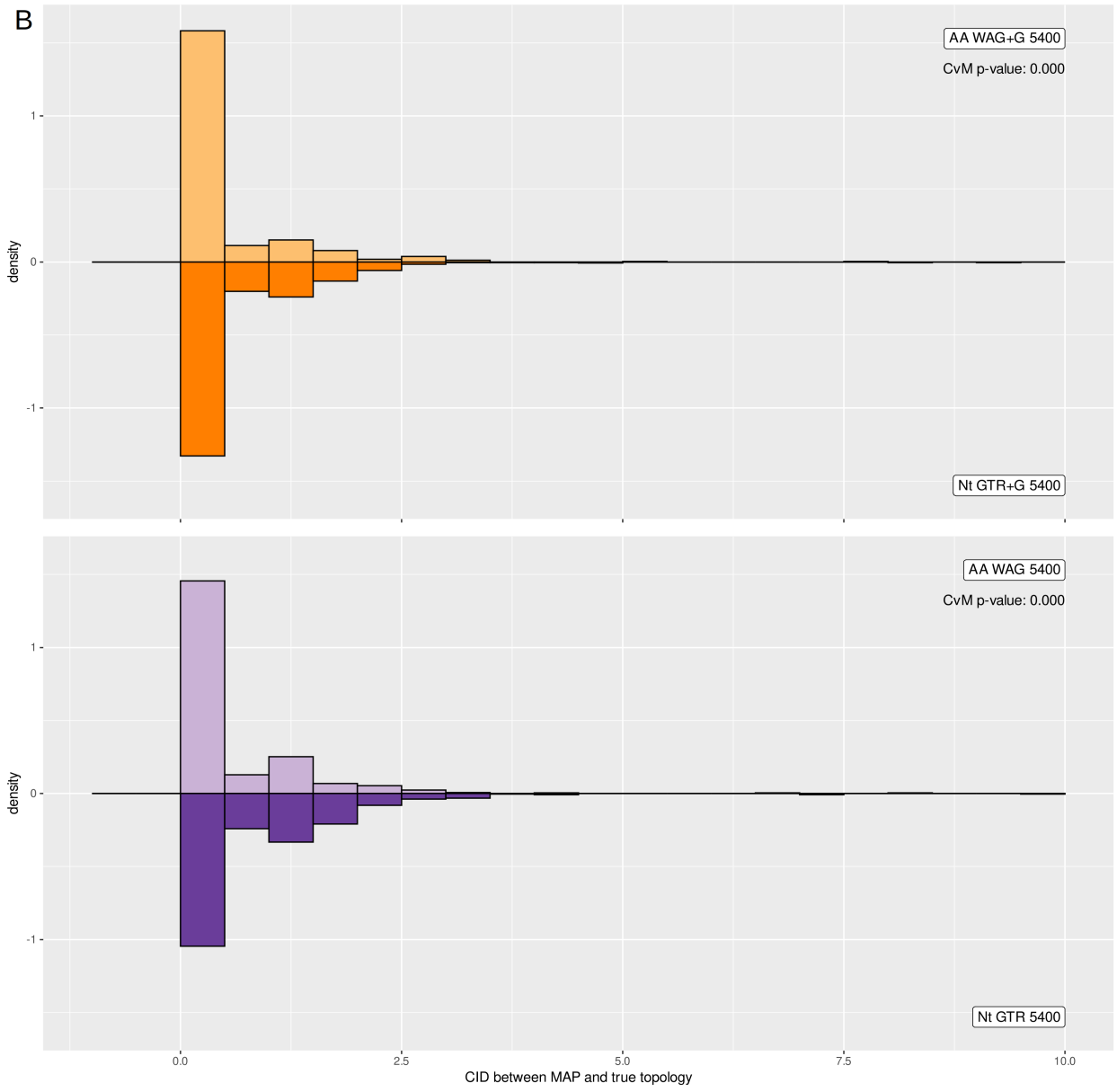

Figure 2.7B. Distributions, across 517 inference runs on different simulated datasets, of the CID between the MAP topology and the true topology. Labels in the right corner of each histogram indicate the data type, inference model, and alignment length in equivalent number of codons. P-values are from two-sample Cramer-Von Mises tests comparing analogous sets of inference runs on amino acid versus nucleotide data.

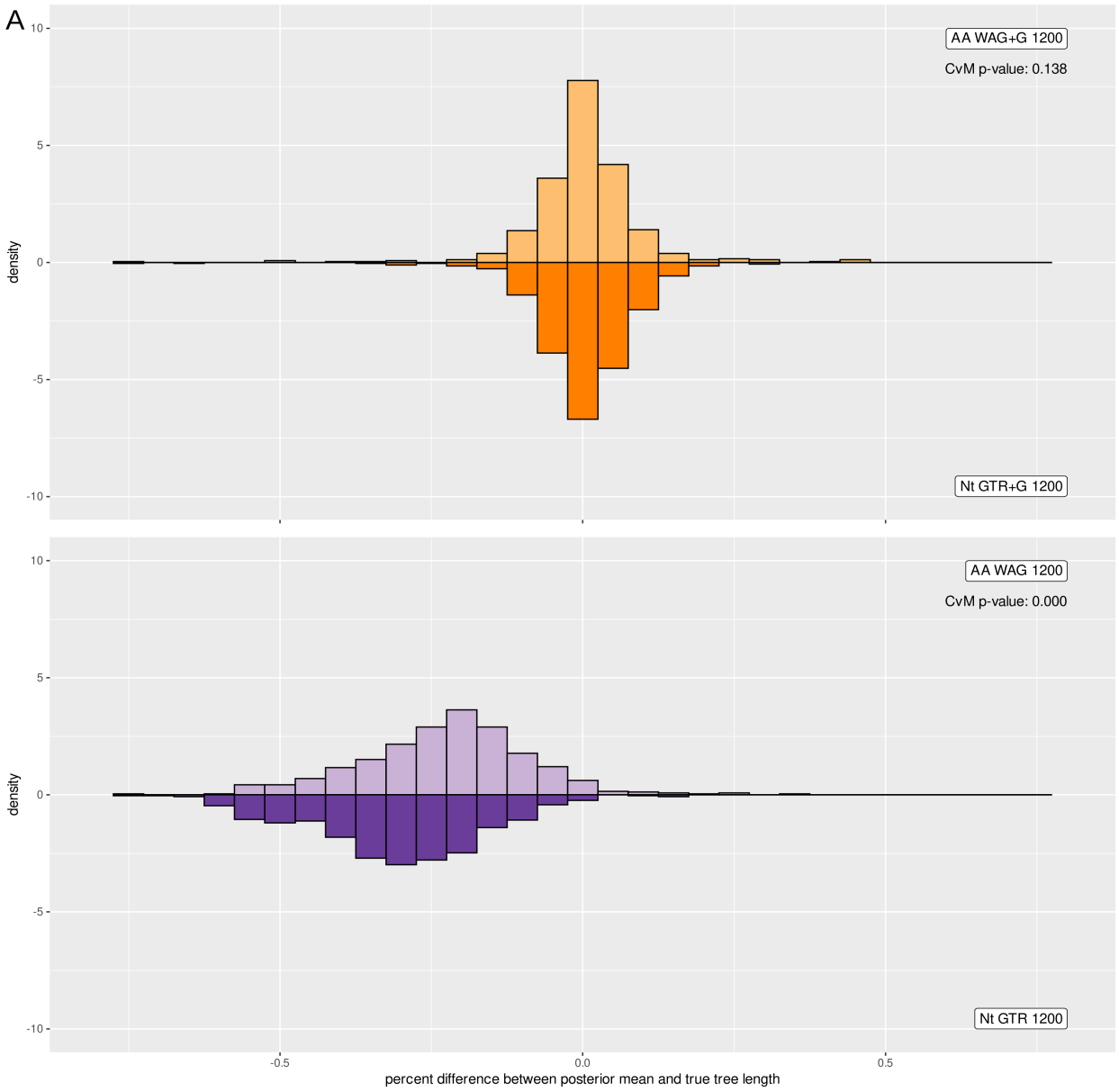

Figure 2.8A. Distributions, across 517 inference runs on different simulated datasets, of the percent difference between estimated and true tree lengths; i.e. (estimated - true) ÷ true. X-axis labels are in decimal (0.5 corresponds to 50%). P-values are from two-sample Cramer-Von Mises tests comparing analogous sets of inference runs on amino acid versus nucleotide data.

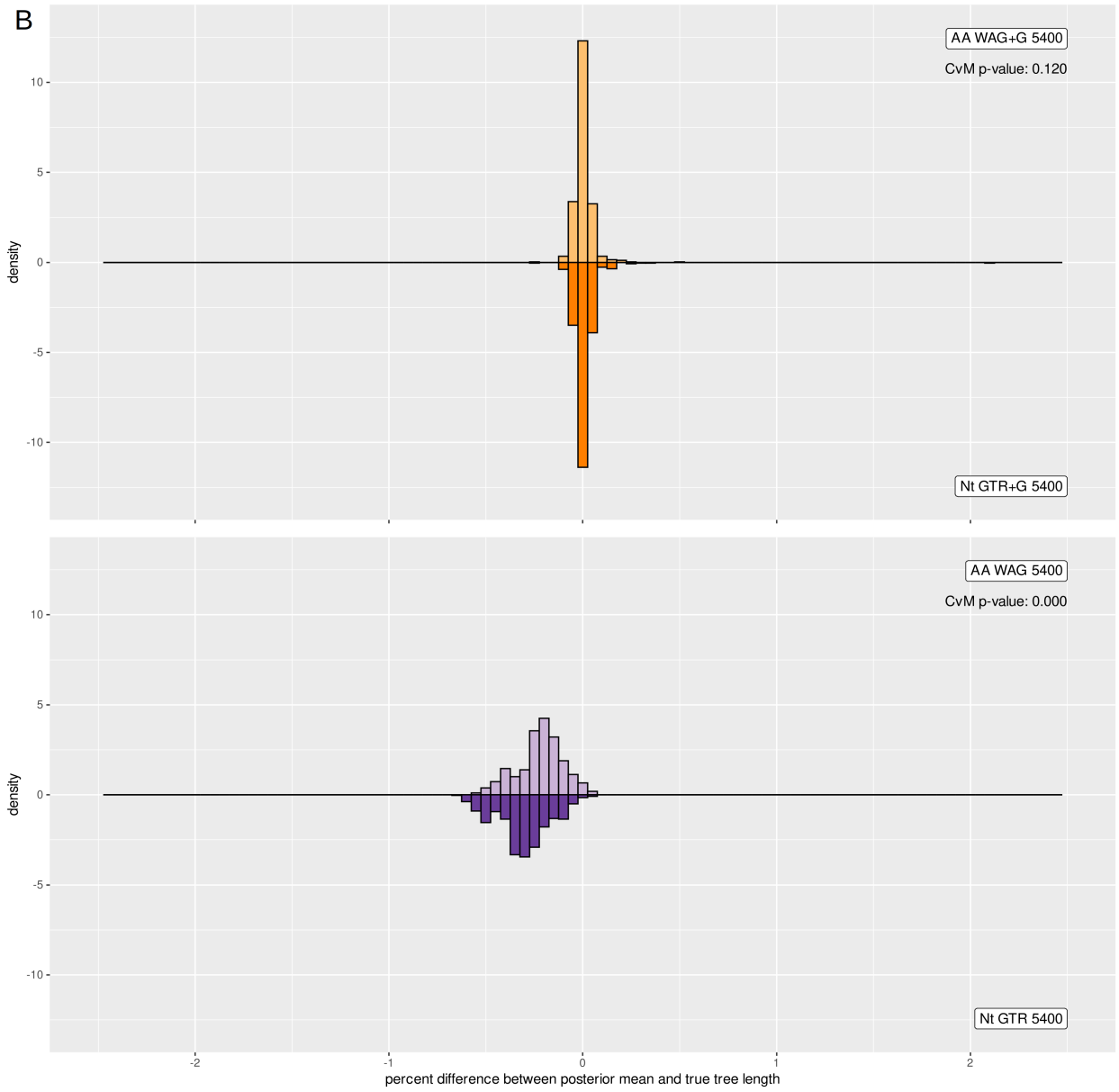

Figure 2.8B. Distributions, across 517 inference runs on different simulated datasets, of the percent difference between estimated and true tree lengths; i.e. (estimated - true) ÷ true. X-axis labels are in decimal (0.5 corresponds to 50%). P-values are from two-sample Cramer-Von Mises tests comparing analogous sets of inference runs on amino acid versus nucleotide data.

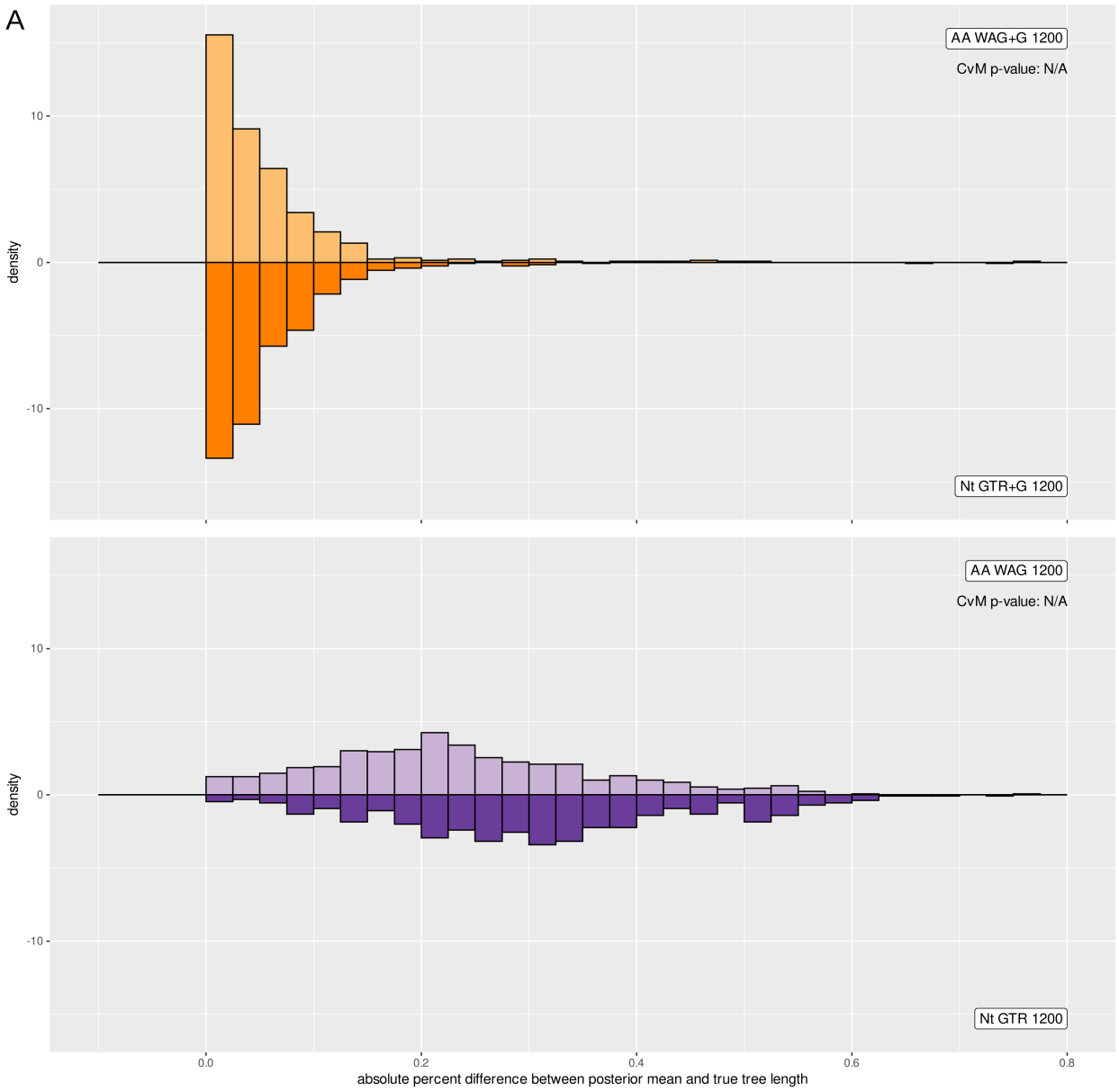

Figure 2.9A. Distributions, across 517 inference runs on different simulated datasets, of the absolute value of the percent difference between estimated and true tree lengths; i.e. |(estimated - true) ÷ true|. X-axis labels are in decimal (0.5 corresponds to 50%). P-values comparing inference runs using amino acid versus nucleotide models on the same underlying data were not calculated for this statistic.

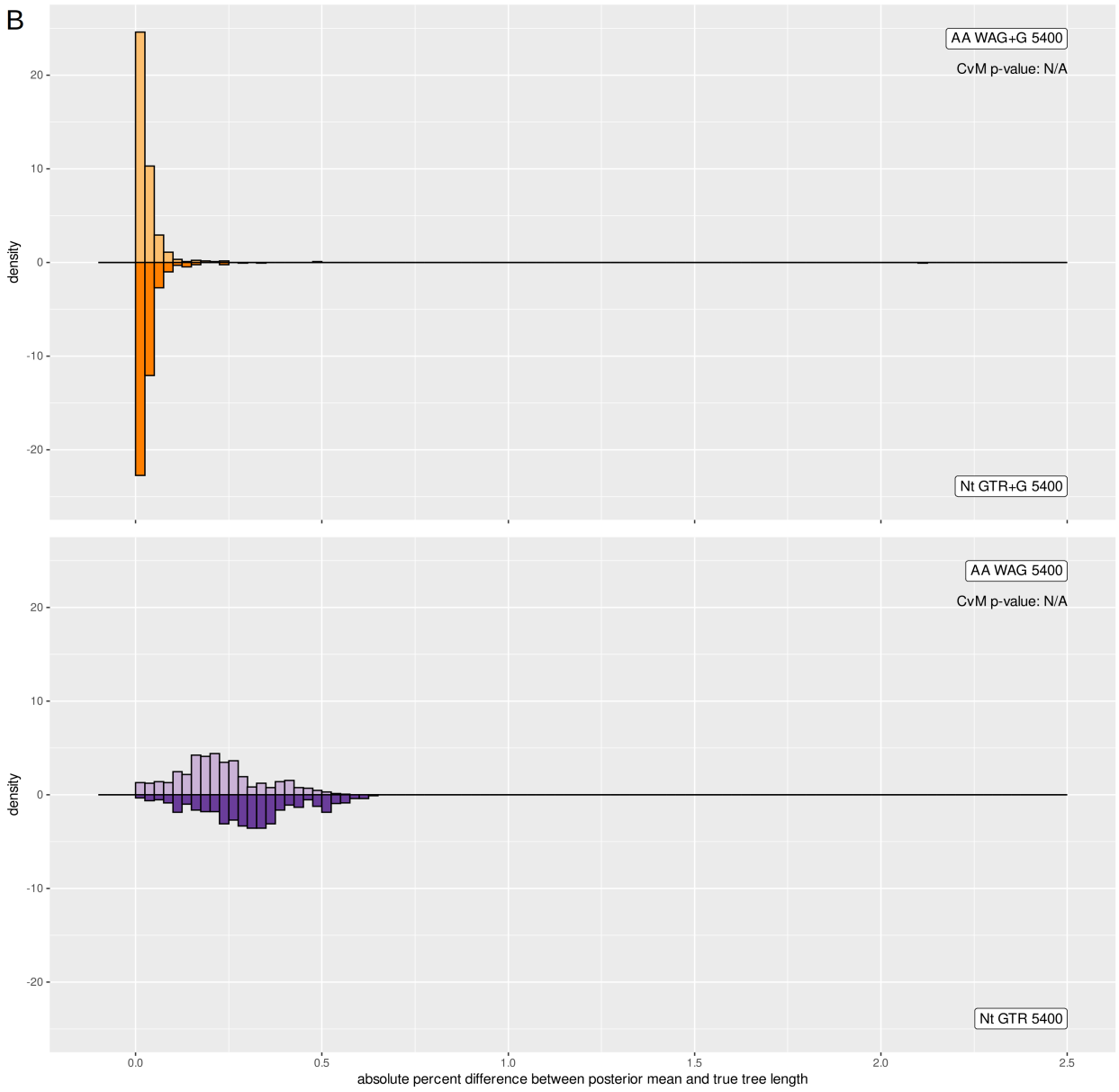

Figure 2.9B. Distributions, across 517 inference runs on different simulated datasets, of the absolute value of the percent difference between estimated and true tree lengths; i.e. |(estimated - true) ÷ true|. X-axis labels are in decimal (0.5 corresponds to 50%). P-values comparing inference runs using amino acid versus nucleotide models on the same underlying data were not calculated for this statistic.

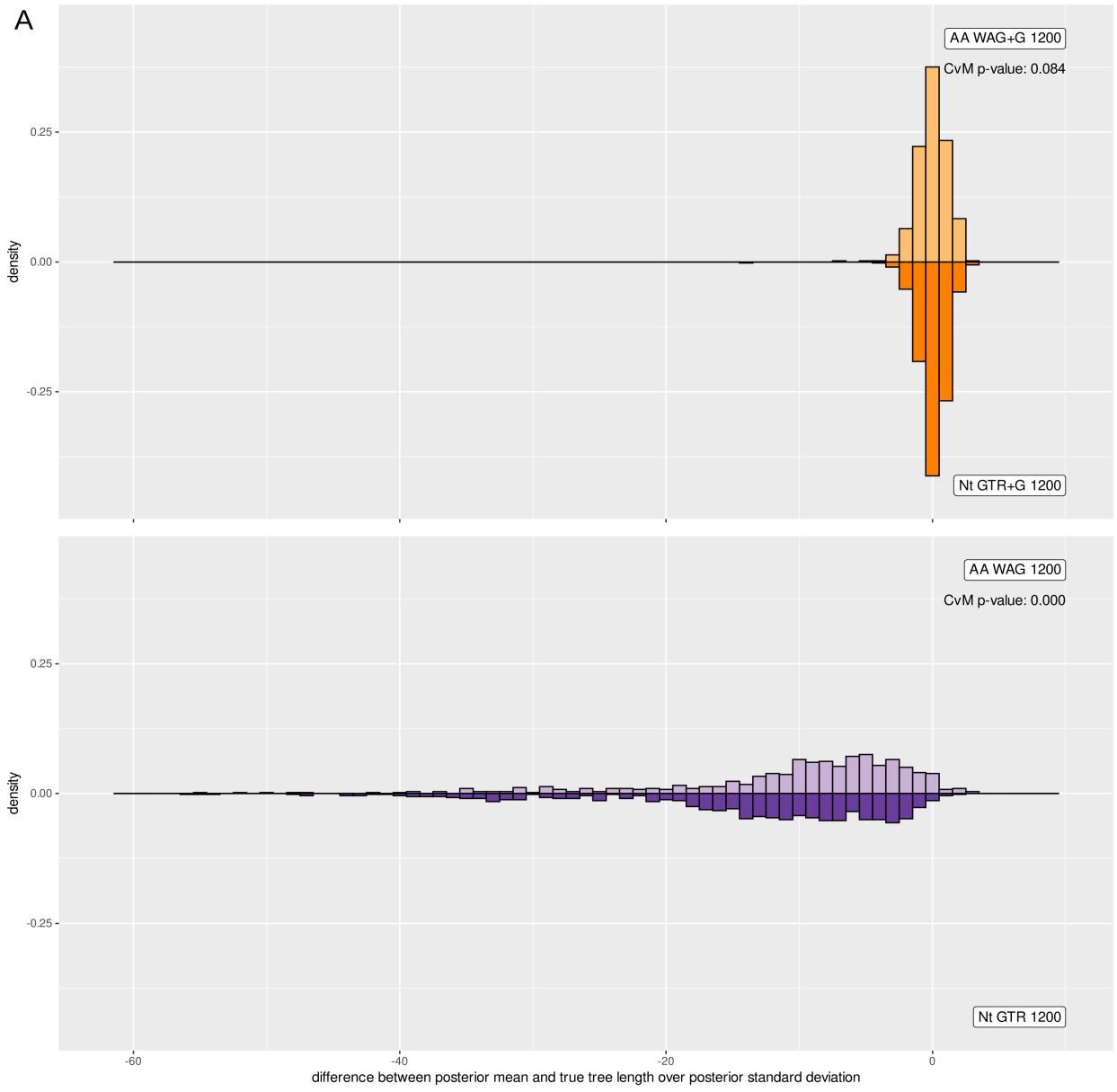

Figure 2.10A. Distributions, across 517 inference runs on different simulated datasets, of the difference between estimated and true tree lengths as a multiple of the standard deviation of the posterior distribution. P-values are from two-sample Cramer-Von Mises tests comparing analogous sets of inference runs on amino acid versus nucleotide data.

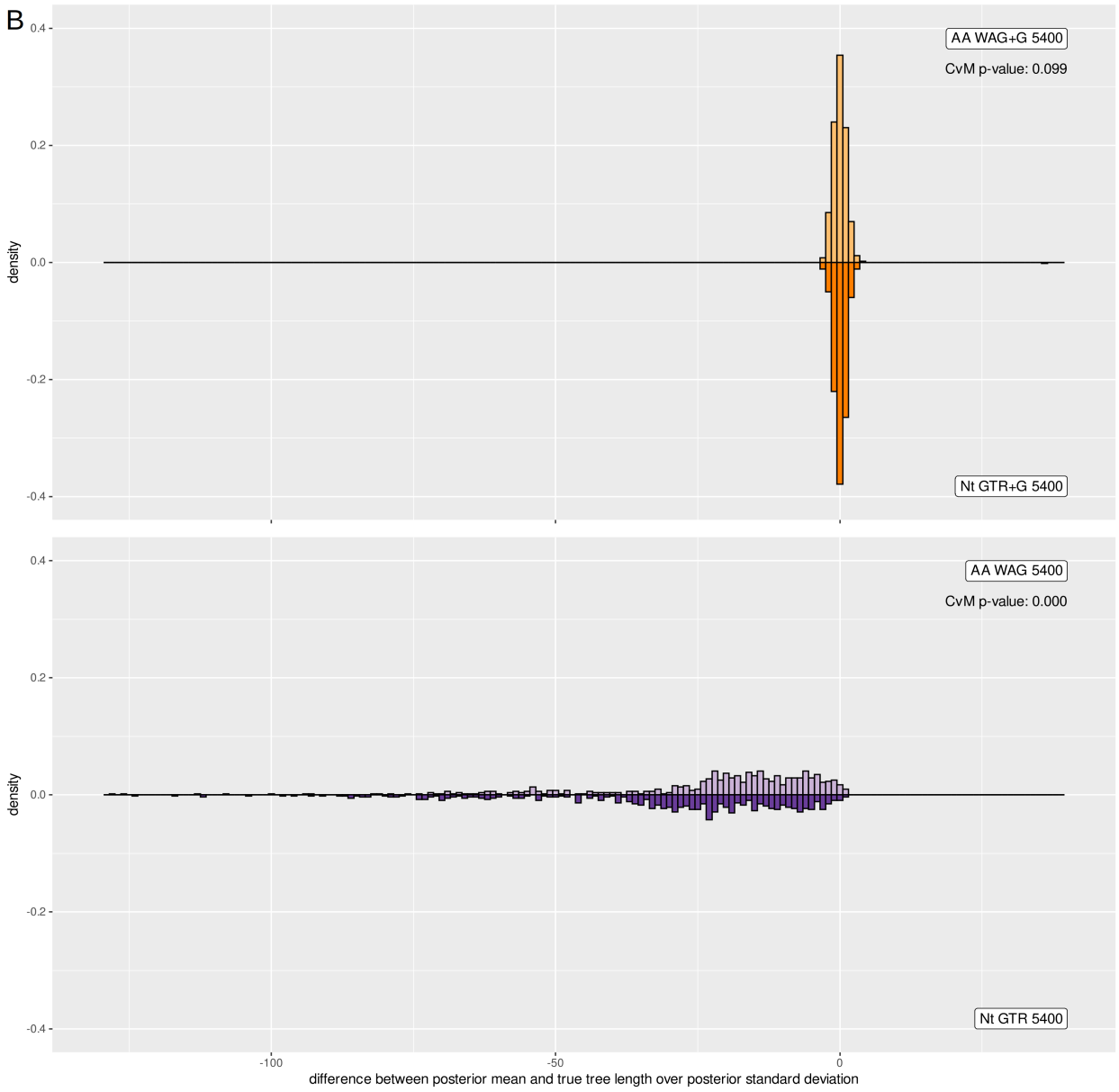

Figure 2.10B. Distributions, across 517 inference runs on different simulated datasets, of the difference between estimated and true tree lengths as a multiple of the standard deviation of the posterior distribution. P-values are from two-sample Cramer-Von Mises tests comparing analogous sets of inference runs on amino acid versus nucleotide data.

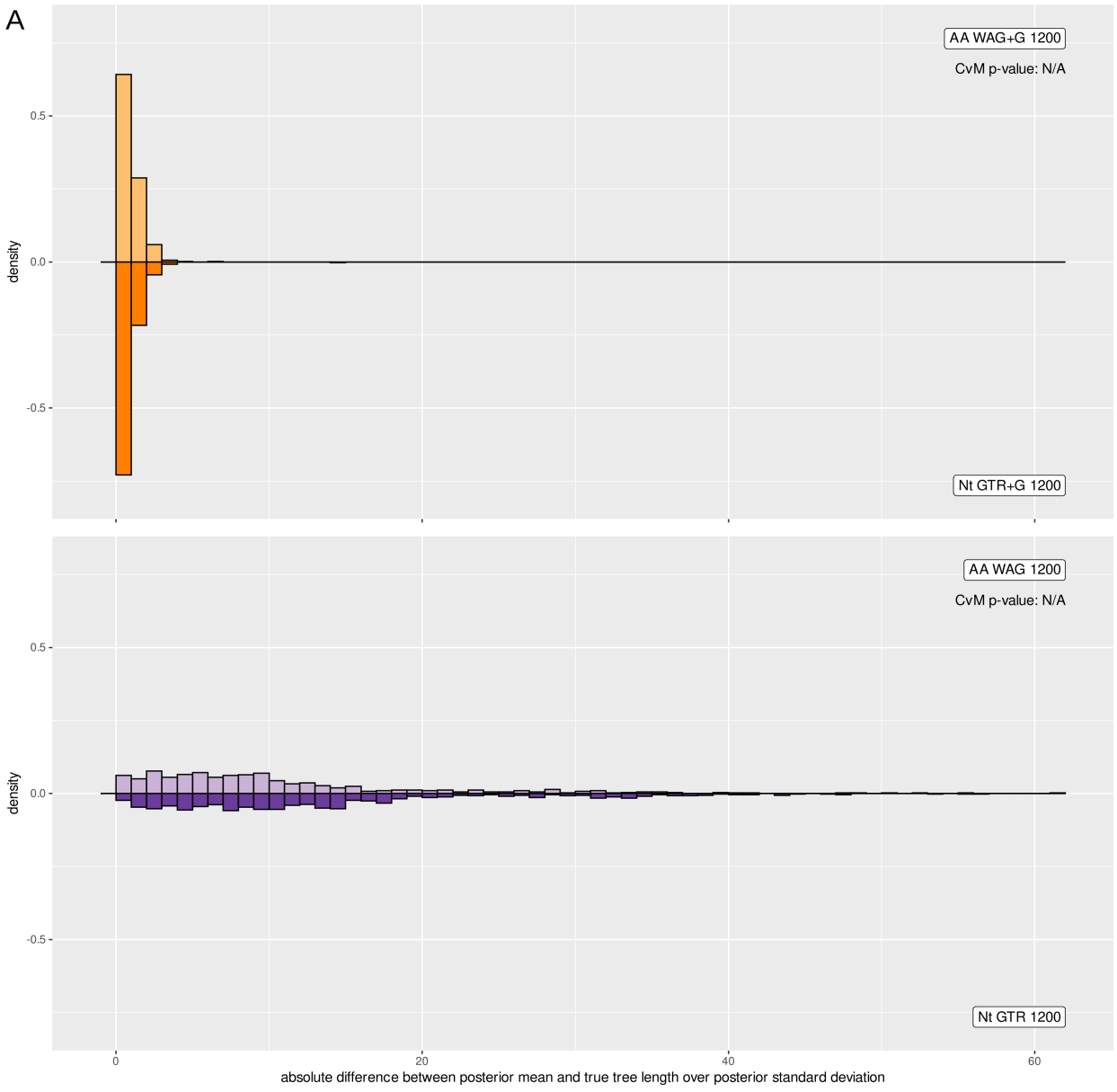

Figure 2.11A. Distributions, across 517 inference runs on different simulated datasets, of the absolute value of the difference between estimated and true tree lengths as a multiple of the standard deviation of the posterior distribution. P-values comparing inference runs using amino acid versus nucleotide models on the same underlying data were not calculated for this statistic.

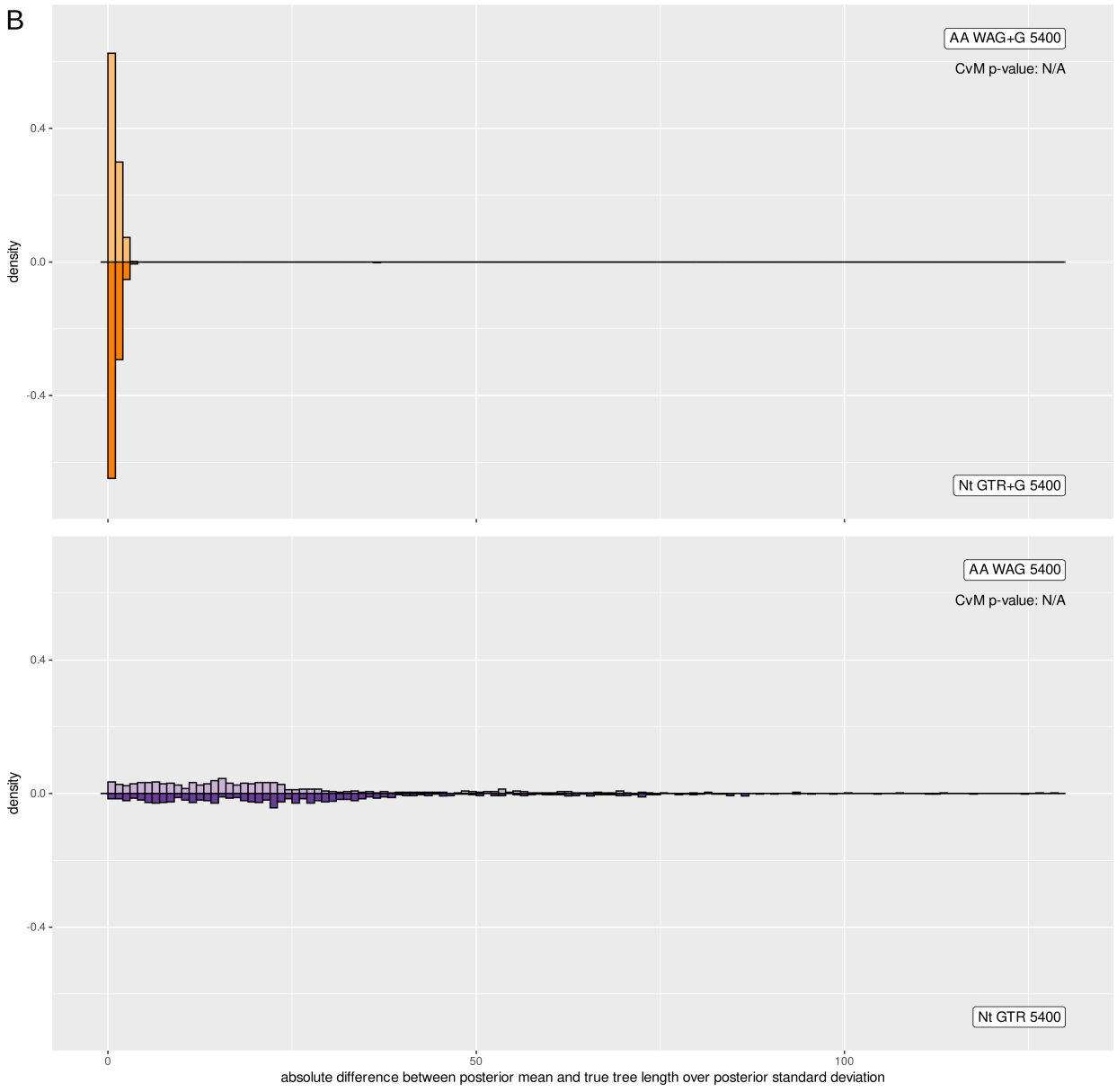

Figure 2.11B. Distributions, across 517 inference runs on different simulated datasets, of the absolute value of the difference between estimated and true tree lengths as a multiple of the standard deviation of the posterior distribution. P-values comparing inference runs using amino acid versus nucleotide models on the same underlying data were not calculated for this statistic.

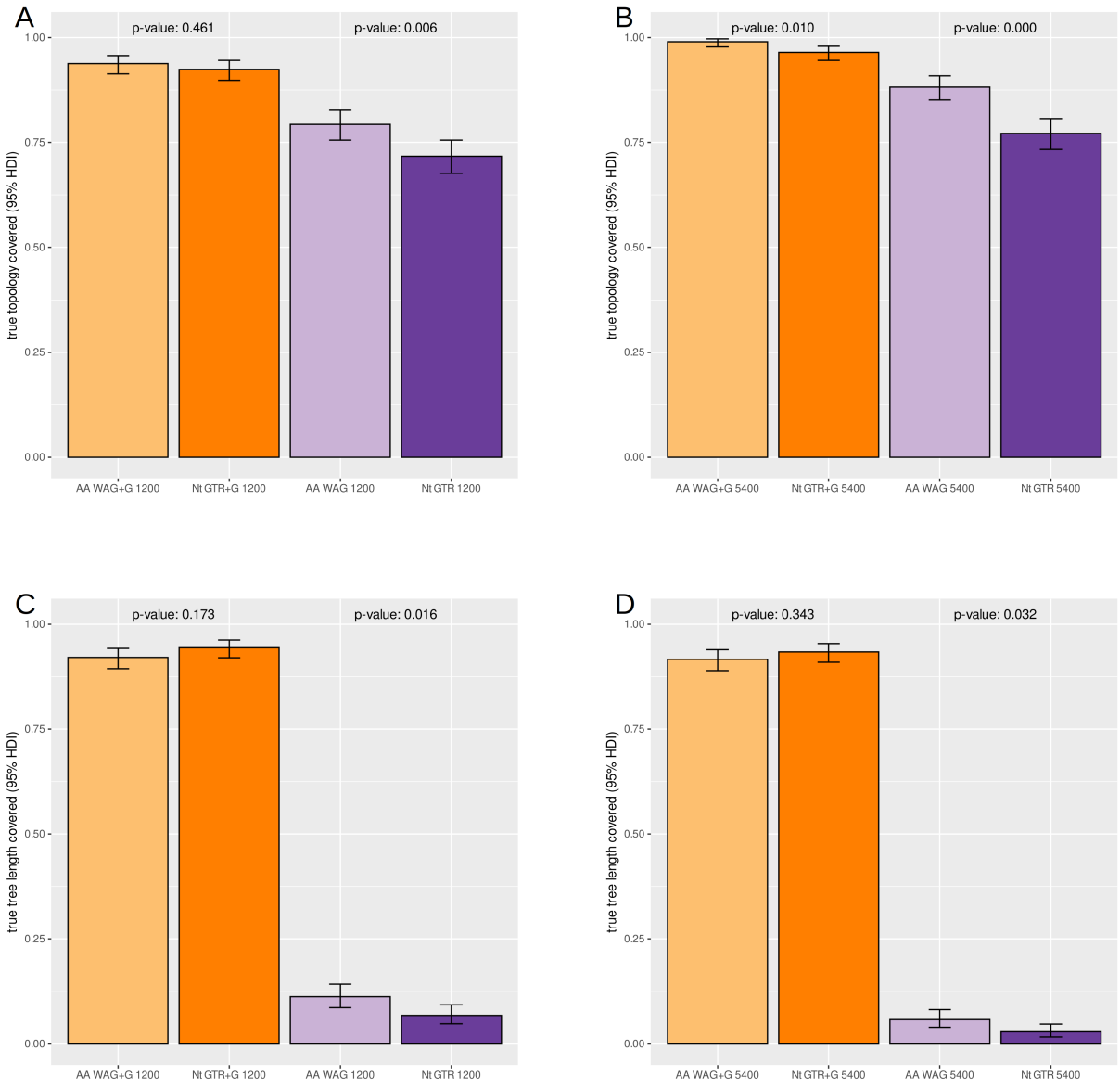

Figure 2.12. Fractions of instances when the true topology (A, B) and true tree length (C, D) are covered by the 95% credible set/interval. Labels beneath each bar indicate the data type, inference model, and alignment length in equivalent number of codons. P-values are from Fisher's exact test comparing analogous sets of inference runs on amino acid versus nucleotide data.

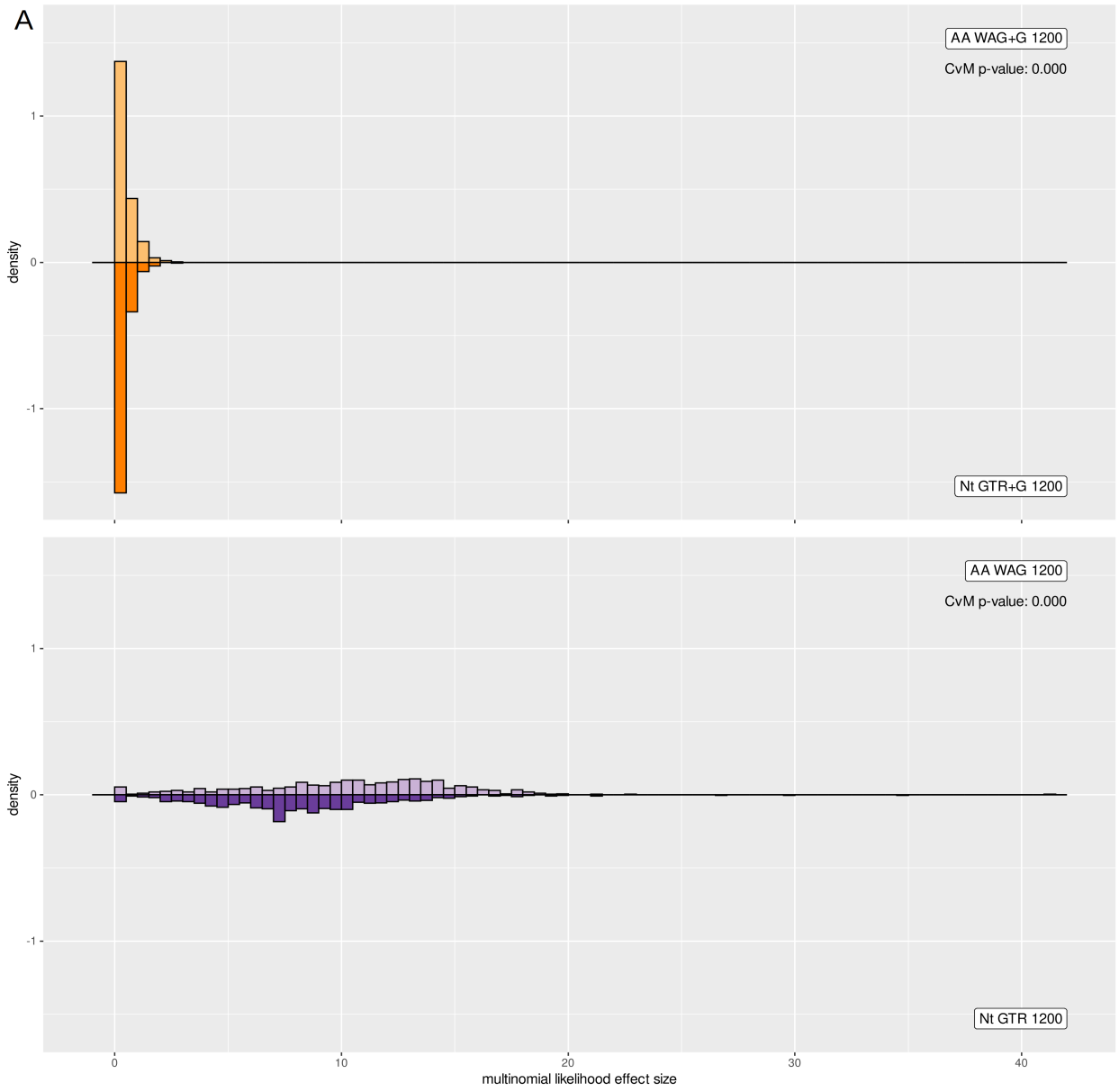

Figure 2.13A. Distributions, across 517 inference runs on different simulated datasets, of the multinomial likelihood posterior predictive effect size. P-values are from two-sample Cramer-Von Mises tests comparing analogous sets of inference runs on amino acid versus nucleotide data.

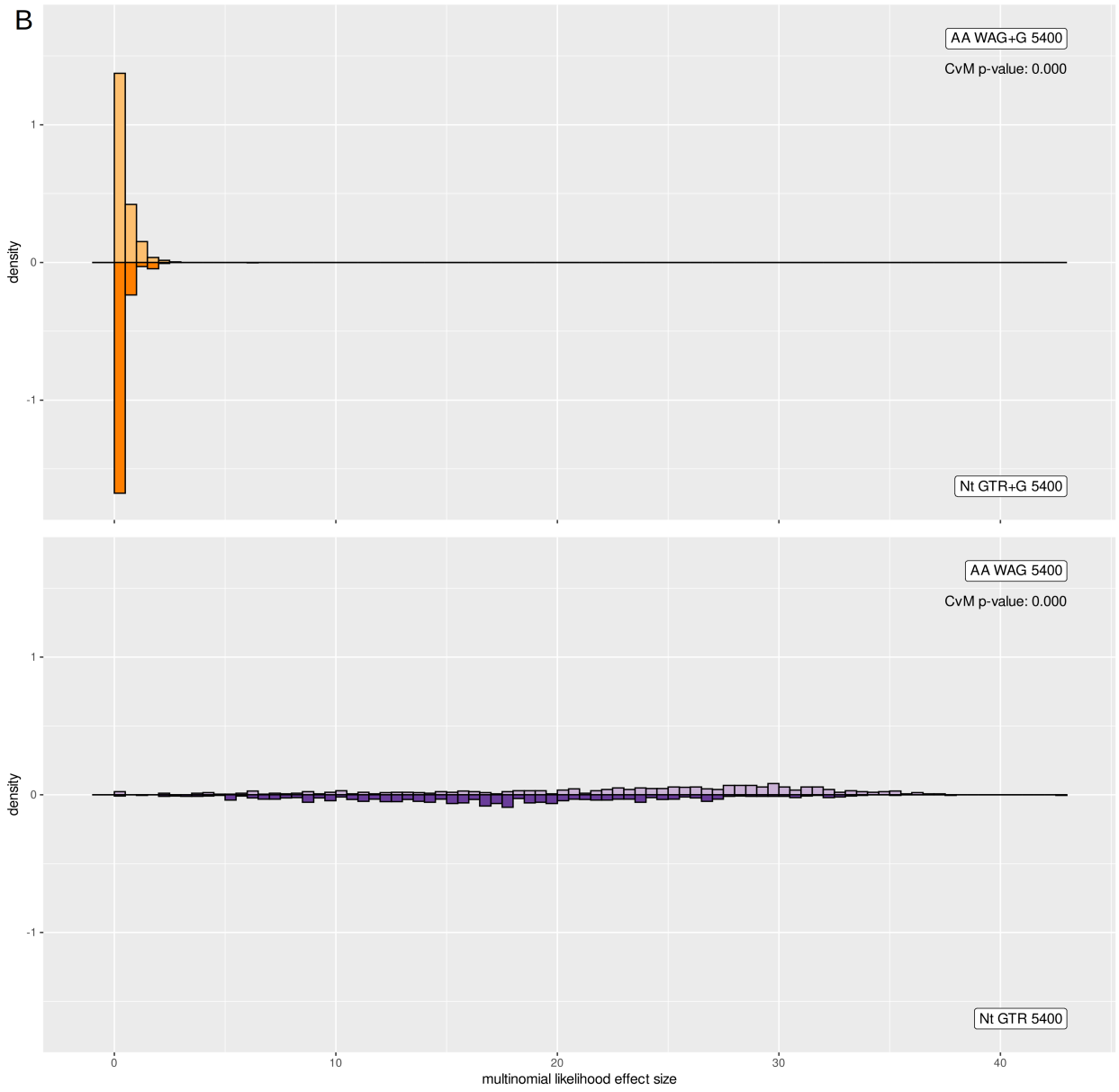

Figure 2.13B. Distributions, across 517 inference runs on different simulated datasets, of the multinomial likelihood posterior predictive effect size. P-values are from two-sample Cramer-Von Mises tests comparing analogous sets of inference runs on amino acid versus nucleotide data.

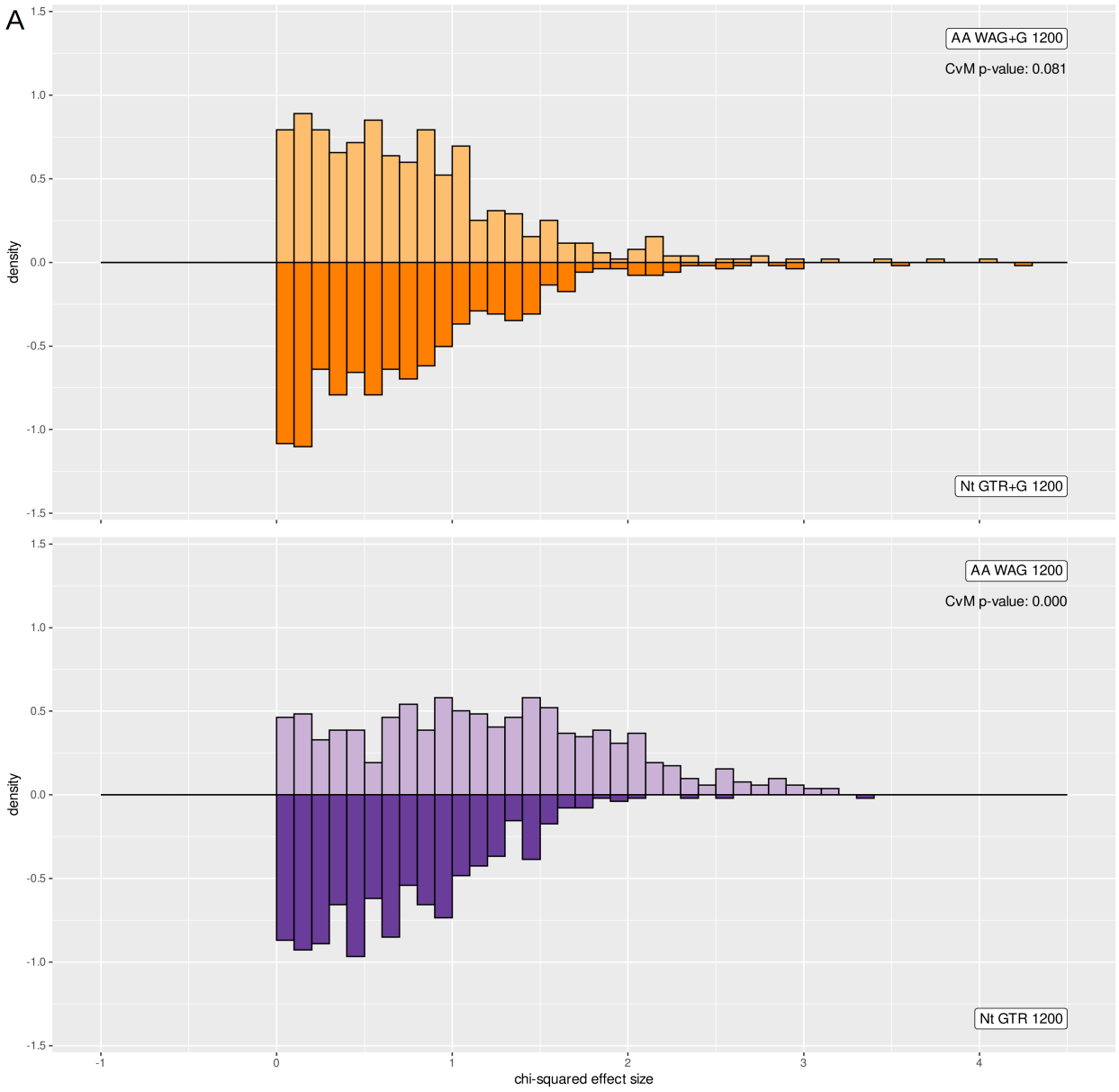

Figure 2.14A. Distributions, across 517 inference runs on different simulated datasets, of the chi-squared posterior predictive effect size. P-values are from two-sample Cramer-Von Mises tests comparing analogous sets of inference runs on amino acid versus nucleotide data.

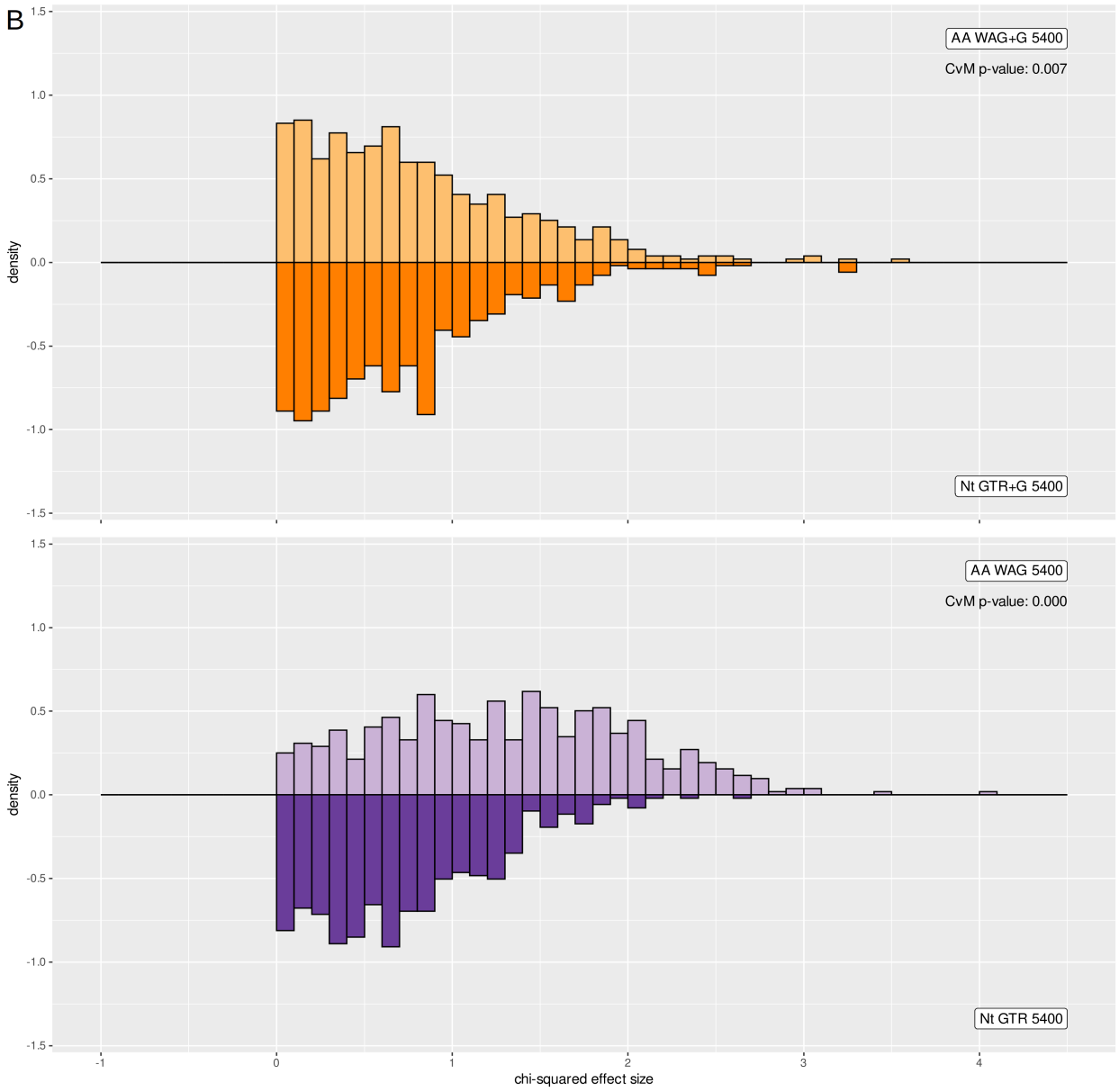

Figure 2.14B. Distributions, across 517 inference runs on different simulated datasets, of the chi-squared posterior predictive effect size. P-values are from two-sample Cramer-Von Mises tests comparing analogous sets of inference runs on amino acid versus nucleotide data.

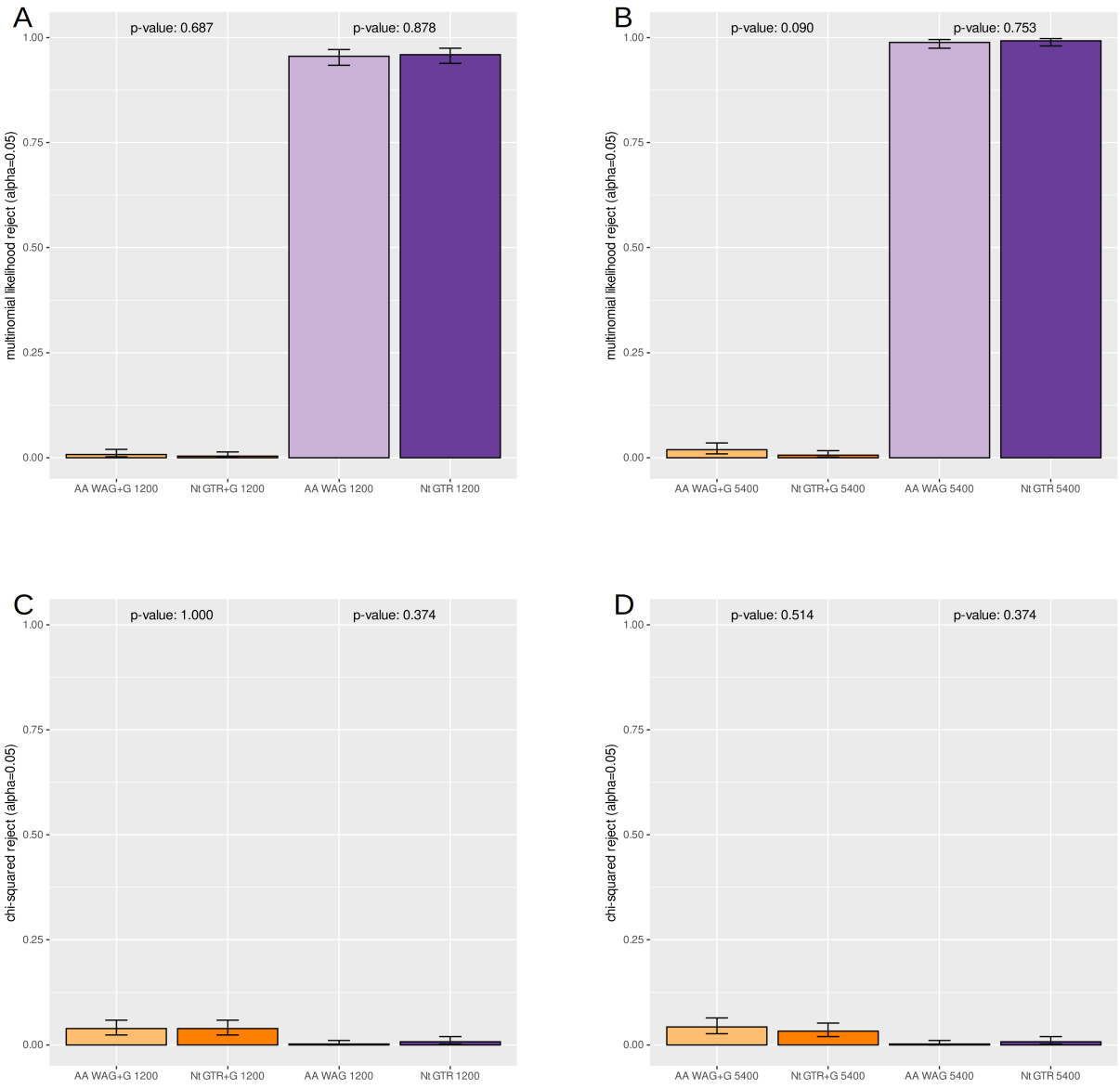

Figure 2.15. Fractions of instances when the inference model is rejected when using the multinomial likelihood (A, B) and chi-squared (C, D) statistics. Labels beneath each bar indicate the data type, inference model, and alignment length in equivalent number of codons. P-values are from Fisher's exact test comparing analogous sets of inference runs on amino acid versus nucleotide data.

Inference	mean CID to true	CID MAP to true	true topology covered	TL % mean to true	TL mean to true over sd	true TL covered	Multinomial likelihood effect size	Multinomial likelihood reject	Chi-squared effect size	Chi-squared reject
AA EMP+G tl50	0.35 (0.49) *0.030*	0.27 (0.55) *0.010*	0.978 (0.146) *0.000*	-7.6% (4.0) *0.000*	-2.5 (1.3) *0.000*	0.378 (0.485) *0.000*	1.46 (0.34) *0.000*	0.082 (0.274) *0.000*	1.30 (1.09) *0.000*	0.263 (0.441) *0.000*
Nt GTR+G tl50	0.43 (0.60)	0.39 (0.66)	0.870 (0.337)	30.0% (26.2)	3.6 (3.3)	0.228 (0.420)	2.54 (0.37)	0.953 (0.211)	18.23 (15.89)	0.980 (0.140)
AA EMP+G tl100	0.37 (0.51) 0.604	0.30 (0.57) 0.092	0.962 (0.192) *0.000*	-18.9% (4.8) *0.000*	-8.4 (2.4) *0.000*	0.000 (0.000) *0.000*	4.12 (0.62) *0.000*	1.000 (0.000) 1.000	1.77 (1.39) *0.000*	0.410 (0.492) *0.000*
Nt GTR+G tl100	0.44 (0.62)	0.39 (0.69)	0.867 (0.340)	6.7% (21.5)	-0.4 (2.1)	0.740 (0.439)	3.37 (0.43)	1.000 (0.000)	19.44 (16.89)	0.988 (0.107)
AA EMP+G tl150	0.40 (0.52) 0.100	0.33 (0.60) 0.171	0.955 (0.207) *0.000*	-26.8% (5.2) *0.000*	-14.2 (3.5) *0.000*	0.000 (0.000) *0.000*	6.76 (1.09) *0.000*	1.000 (0.000) 1.000	2.00 (1.53) *0.000*	0.460 (0.499) *0.000*
Nt GTR+G tl150	0.48 (0.59)	0.42 (0.67)	0.860 (0.347)	-13.8% (16.3)	-2.0 (2.9)	0.638 (0.481)	4.32 (0.43)	1.000 (0.000)	19.84 (16.97)	0.985 (0.122)

Table 2.3. Topology and tree length inference accuracy and posterior predictive test statistics using amino acid and nucleotide models. Simulation was done using a branch-heterogeneous MutSel codon model. All alignments are 1800 codons in length. Displayed values are averages across 600 simulation/inference runs. Standard deviations are in parentheses. Benjamini-Yekutieli adjusted p-values for amino acid versus nucleotide comparisons are listed below the standard deviation of each amino acid entry. Adjusted p-values smaller than 0.05 are italicized and in red.

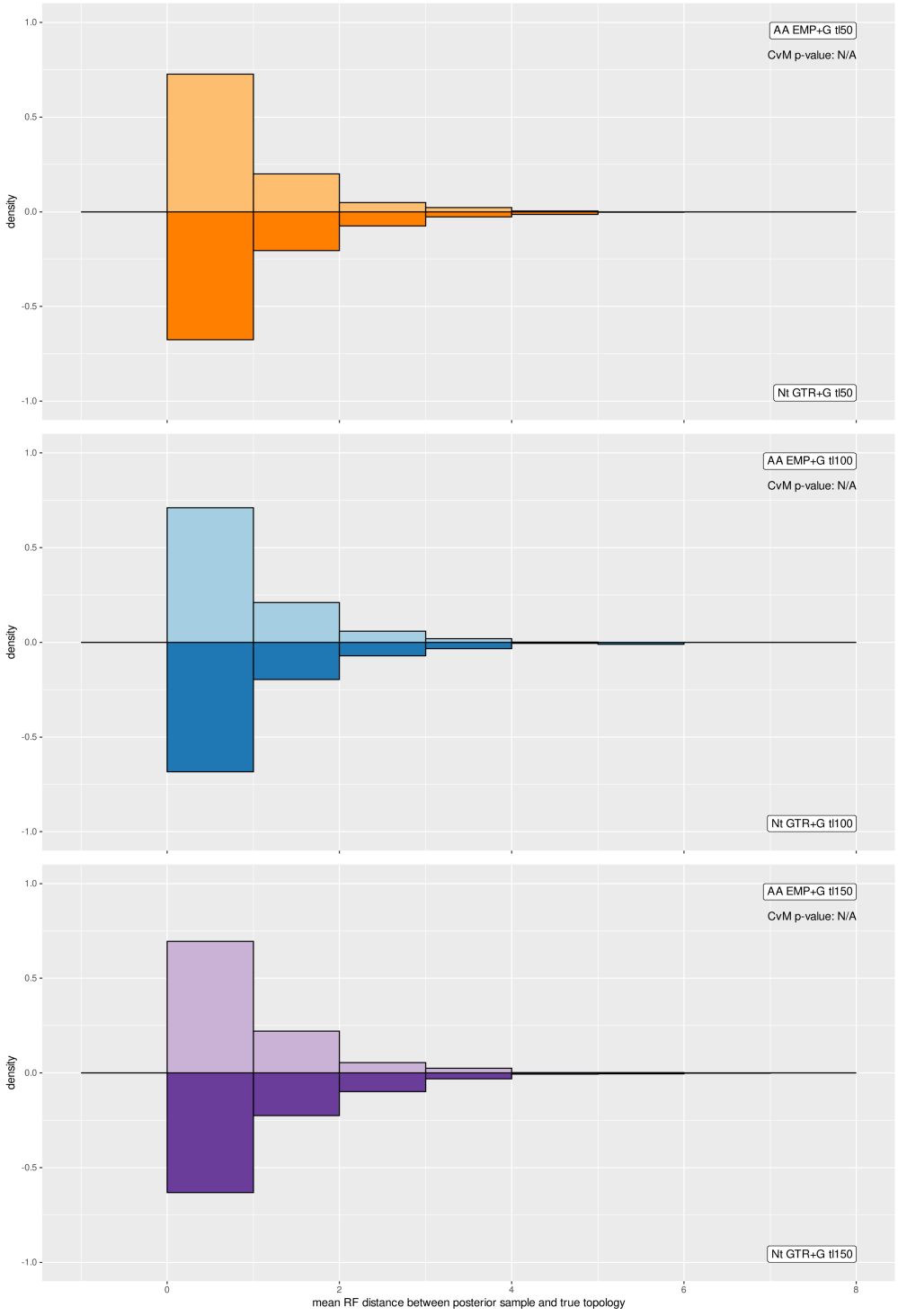

Figure 2.16. Distributions of the mean RF distance between every topology in the posterior sample and the true topology. Labels in the right corner of each histogram indicate the data type, inference model, and tree length (in expected number of substitutions per codon). P-values comparing inference runs using amino acid versus nucleotide models on the same underlying data were not calculated for this statistic.

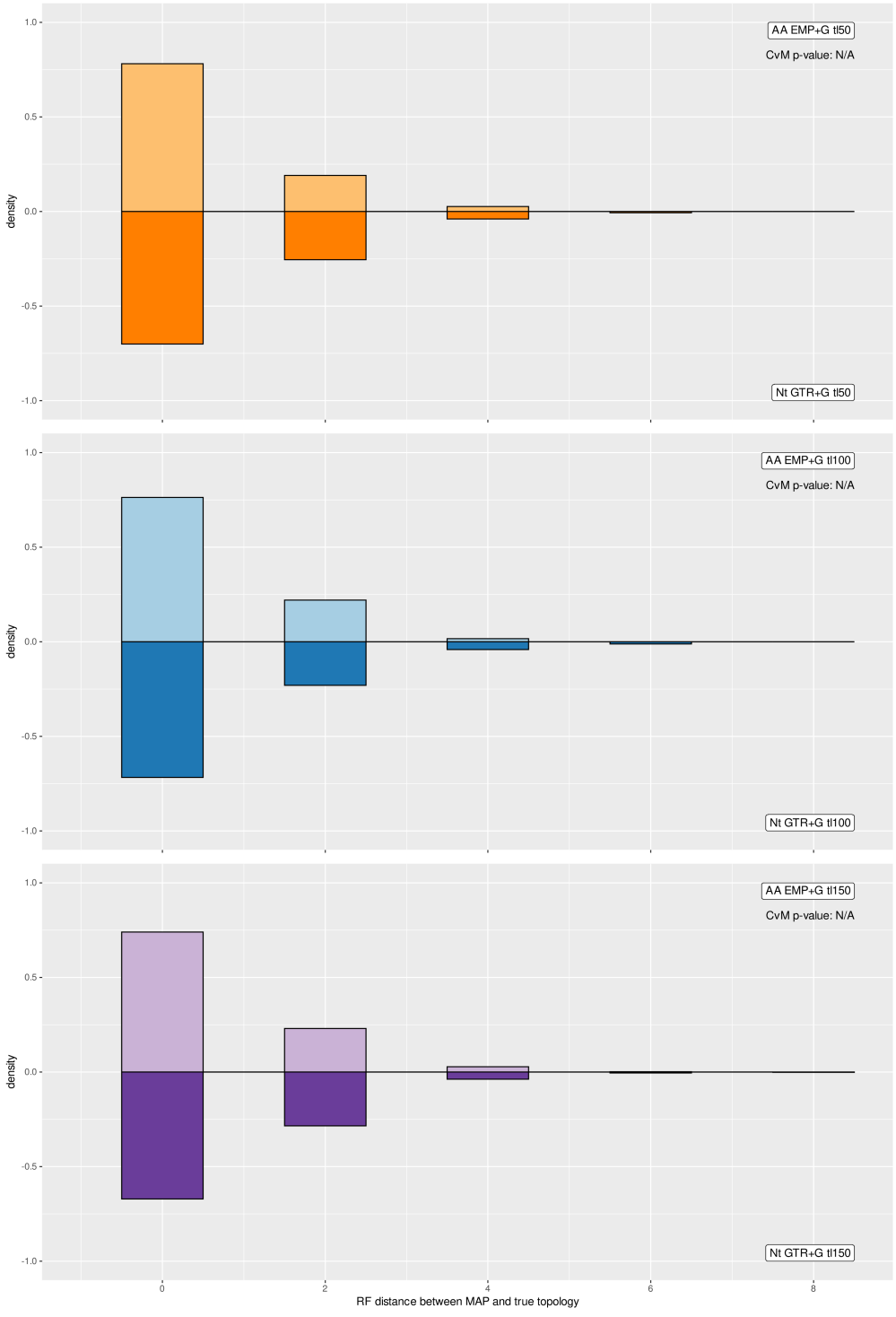

Figure 2.17. Distributions of the RF distance between the MAP topology and the true topology. Labels in the right corner of each histogram indicate the data type, inference model, and tree length (in expected number of substitutions per codon). P-values comparing inference runs using amino acid versus nucleotide models on the same underlying data were not calculated for this statistic.

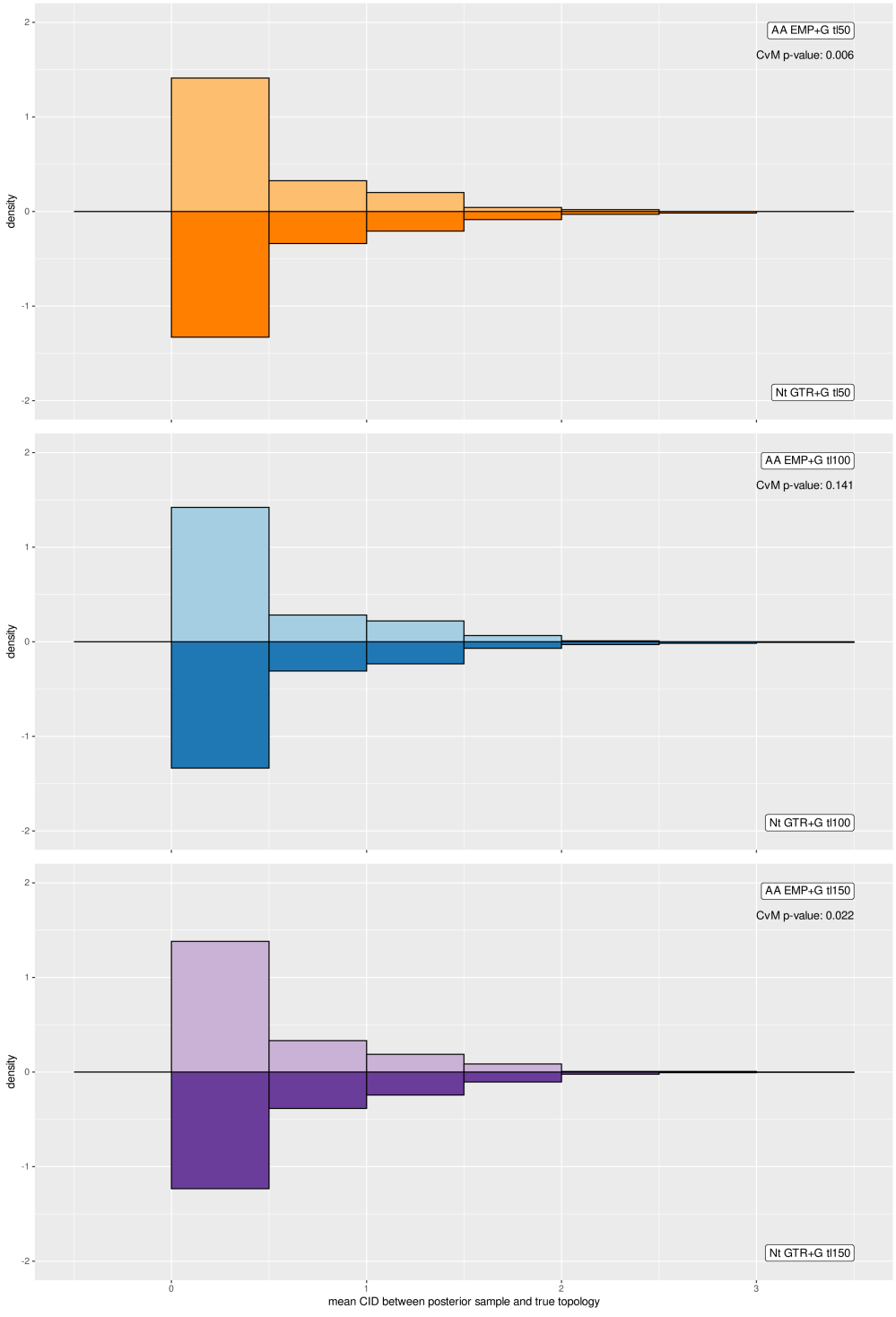

Figure 2.18. Distributions of the mean CID between every topology in the posterior sample and the true topology. Labels in the right corner of each histogram indicate the data type, inference model, and tree length (in expected number of substitutions per codon). P-values were calculated using two-sample Cramer-Von Mises tests comparing inference runs using amino acid versus nucleotide models on the same underlying data.

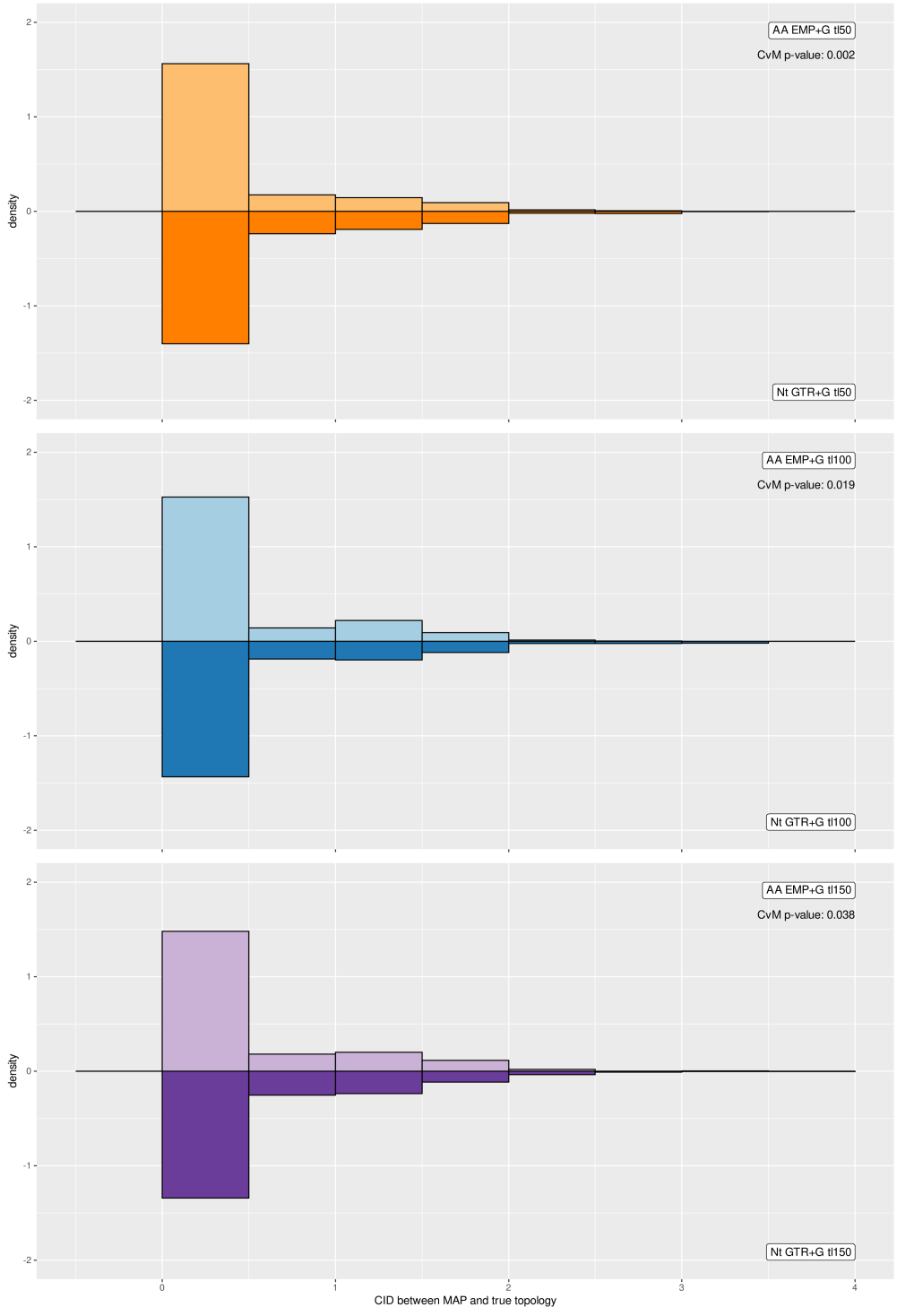

Figure 2.19. Distributions of the CID between the MAP topology and the true topology. Labels in the right corner of each histogram indicate the data type, inference model, and tree length (in expected number of substitutions per codon). P-values were calculated using two-sample Cramer-Von Mises tests comparing inference runs using amino acid versus nucleotide models on the same underlying data.

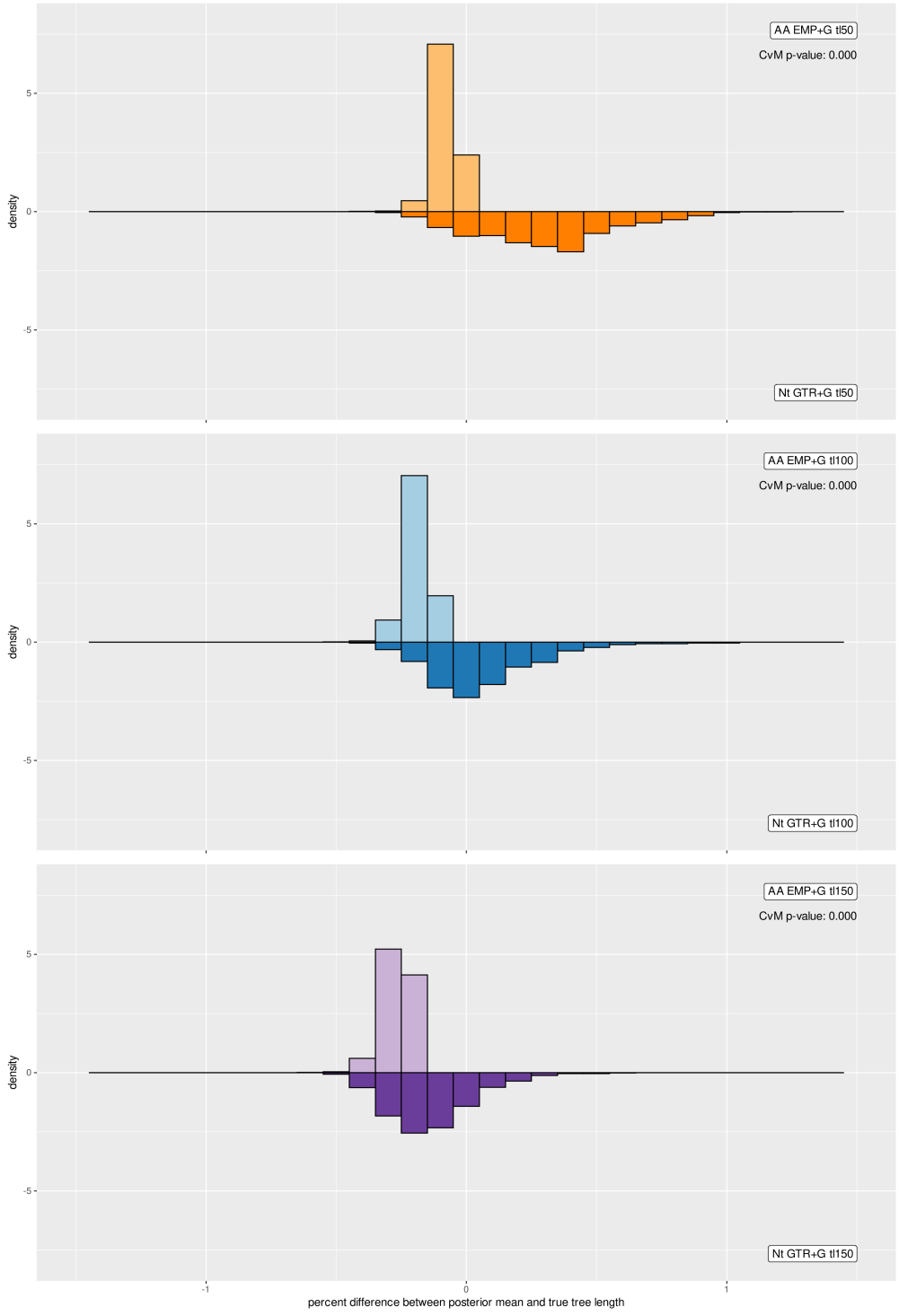

Figure 2.20. Distributions of the percent difference between estimated and true tree lengths; i.e. (estimated - true) ÷ true. X-axis labels are in decimal (1 corresponds to 100%). Labels in the right corner of each histogram indicate the data type, inference model, and tree length (in expected number of substitutions per codon). P-values were calculated using two-sample Cramer-Von Mises tests comparing inference runs using amino acid versus nucleotide models on the same underlying data.

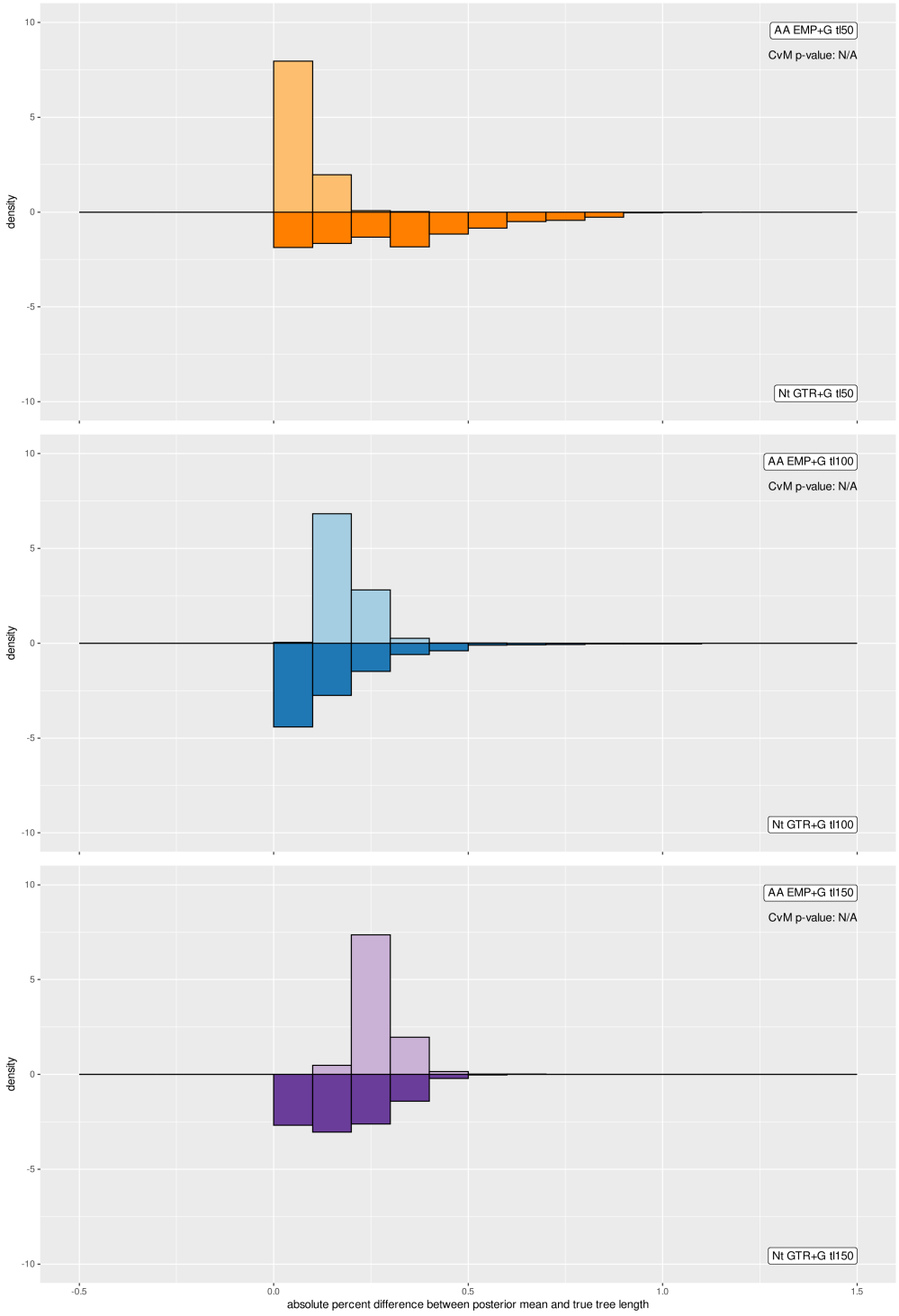

Figure 2.21. Distributions of the absolute value of the percent difference between estimated and true tree lengths; i.e. |(estimated - true) ÷ true|. X-axis labels are in decimal (1 corresponds to 100%). Labels in the right corner of each histogram indicate the data type, inference model, and tree length (in expected number of substitutions per codon). P-values comparing inference runs using amino acid versus nucleotide models on the same underlying data were not calculated for this statistic.

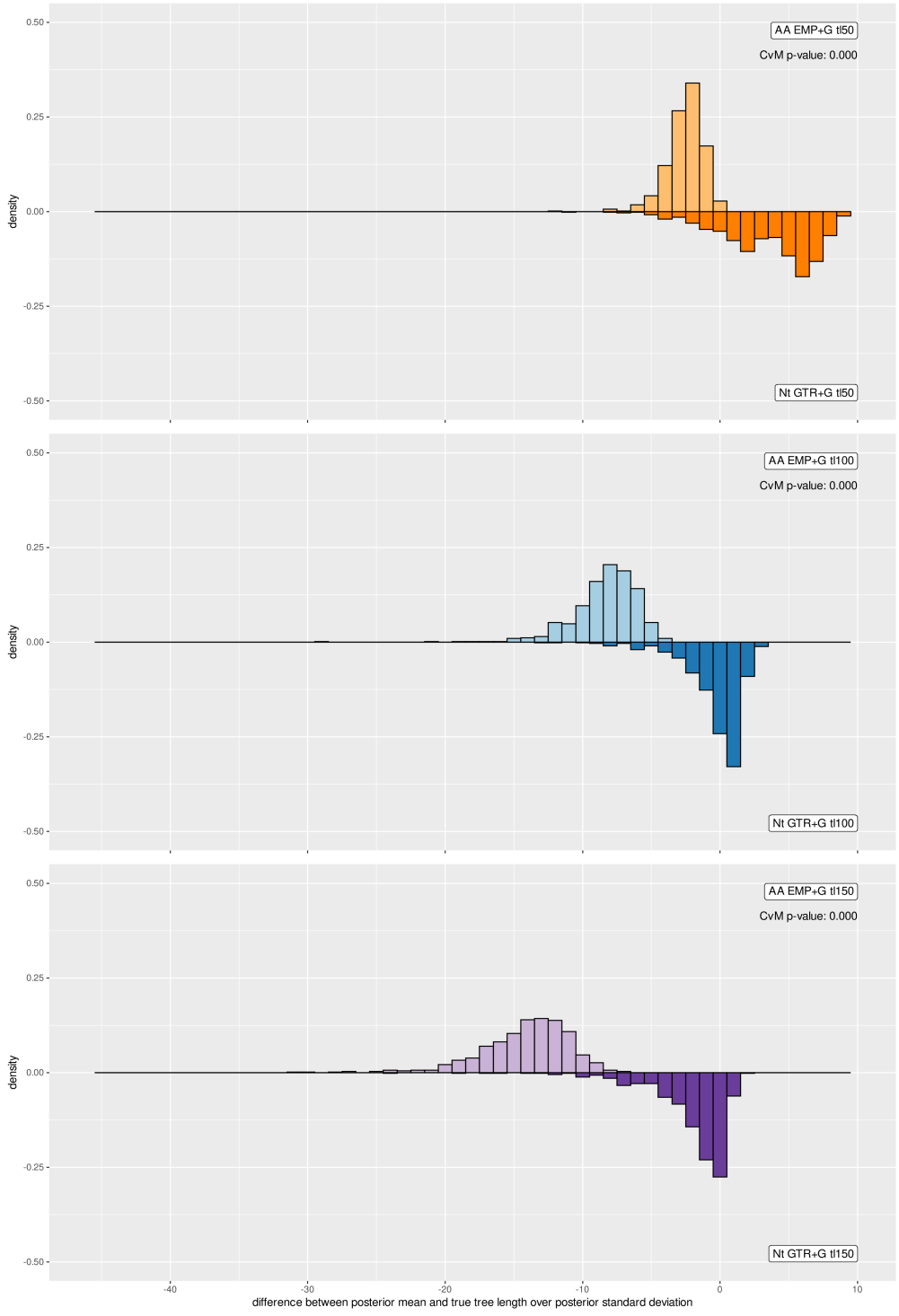

Figure 2.22. Distributions of the difference between estimated and true tree lengths as a multiple of the standard deviation of the posterior distribution. Labels in the right corner of each histogram indicate the data type, inference model, and tree length (in expected number of substitutions per codon). P-values were calculated using two-sample Cramer-Von Mises tests comparing inference runs using amino acid versus nucleotide models on the same underlying data.

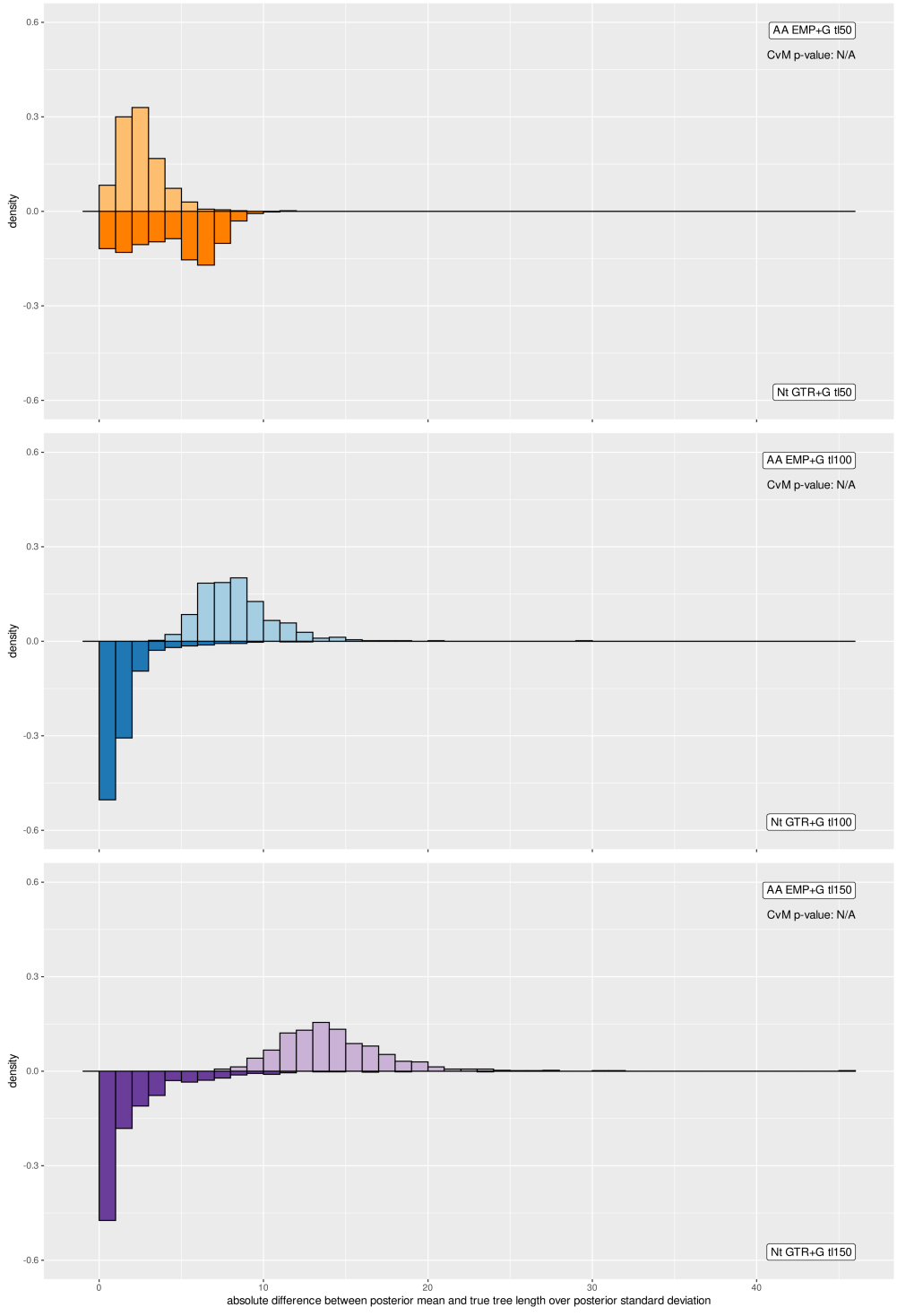

Figure 2.23. Distributions of the absolute value of the difference between estimated and true tree lengths as a multiple of the standard deviation of the posterior distribution. Labels in the right corner of each histogram indicate the data type, inference model, and tree length (in expected number of substitutions per codon). P-values comparing inference runs using amino acid versus nucleotide models on the same underlying data were not calculated for this statistic.

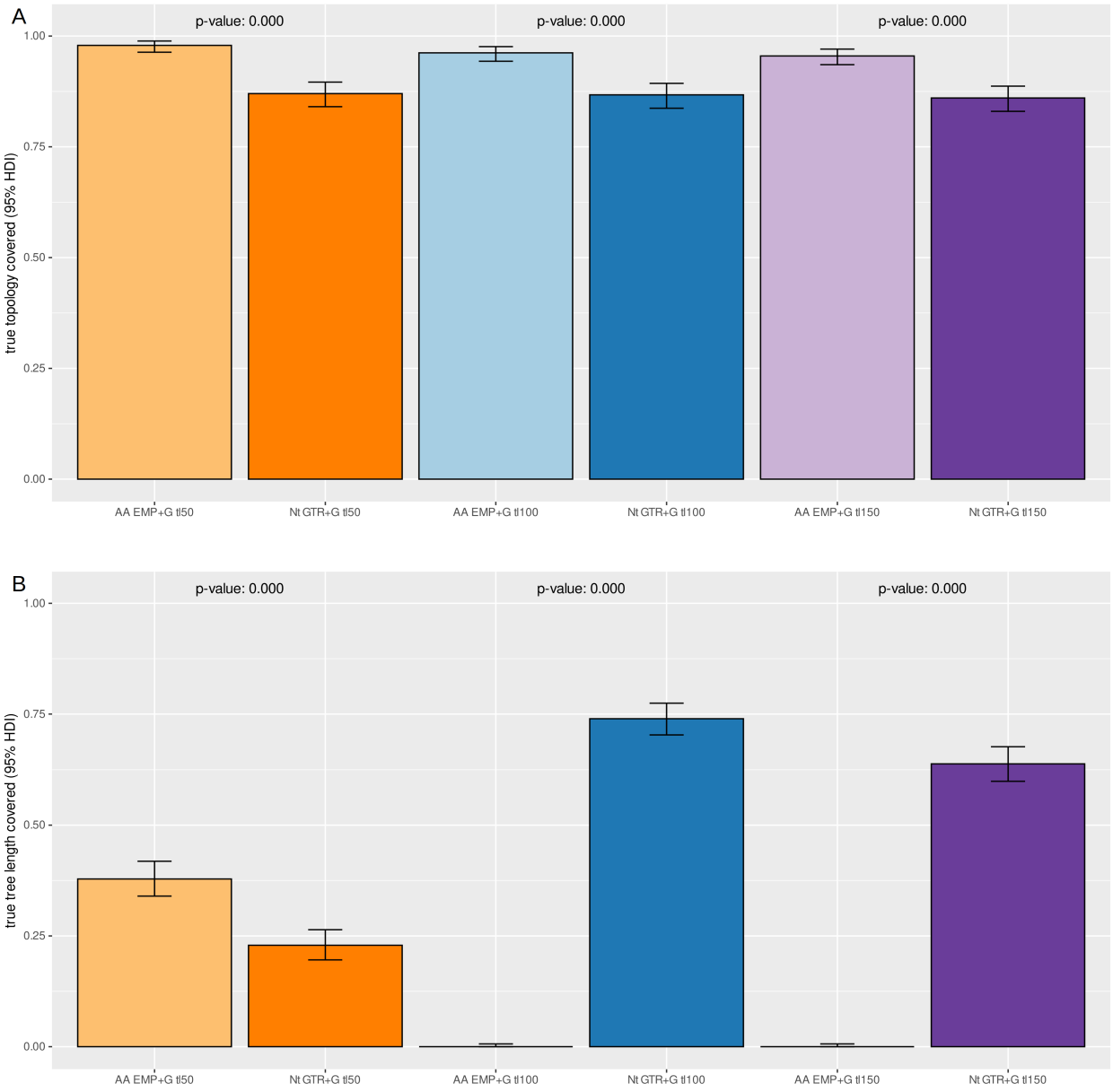

Figure 2.24. Fractions of instances when the true topology (A) and the true tree length (B) are covered by the 95% credible set/interval. Labels beneath each bar indicate the data type, inference model, and tree length (in expected number of substitutions per codon). P-values were calculated using Fisher's exact test comparing inference runs using amino acid versus nucleotide models on the same underlying data.

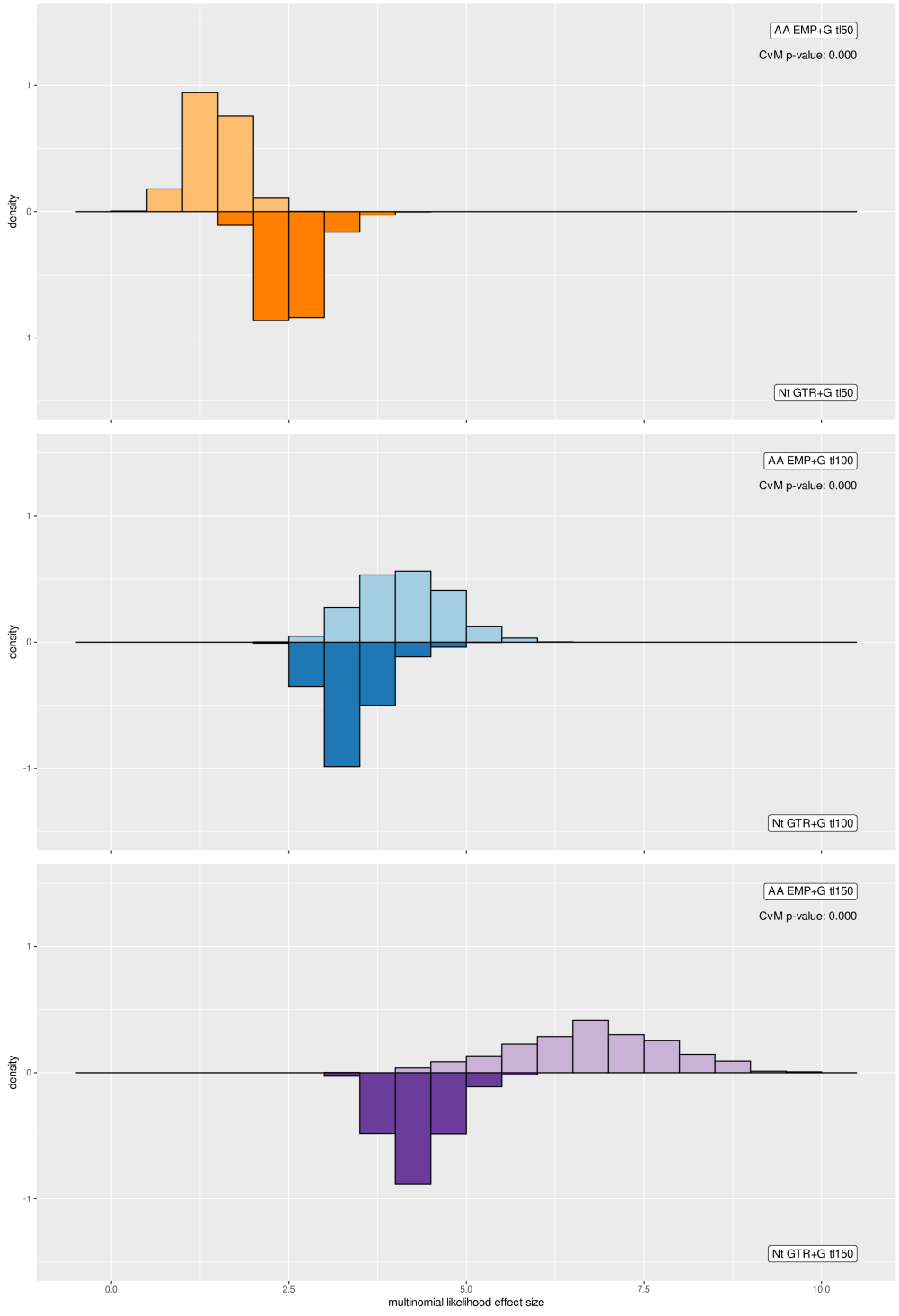

Figure 2.25. Distributions of the multinomial likelihood posterior predictive effect size. Labels in the right corner of each histogram indicate the data type, inference model, and tree length (in expected number of substitutions per codon). P-values were calculated using two-sample Cramer-Von Mises tests comparing inference runs using amino acid versus nucleotide models on the same underlying data.

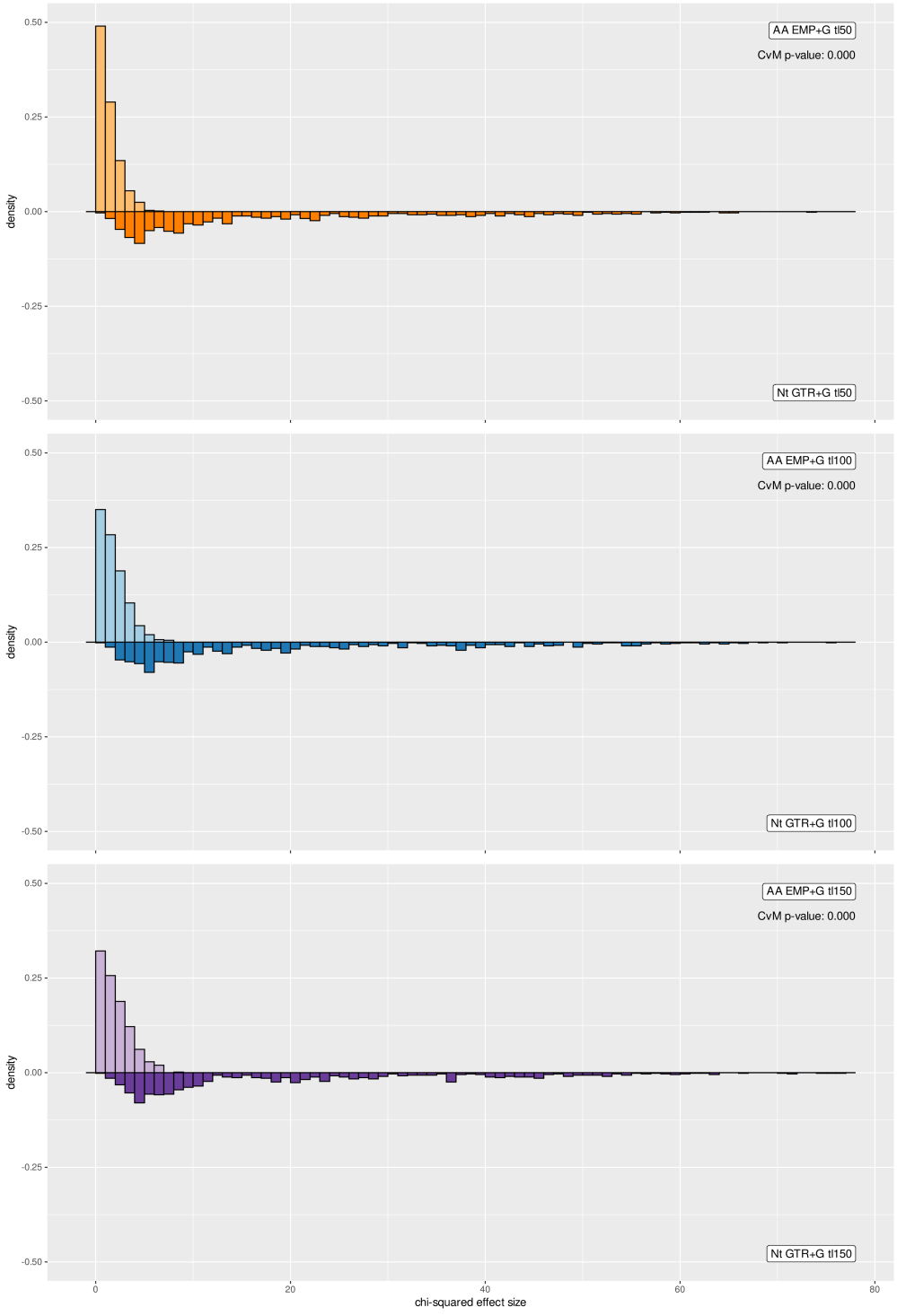

Figure 2.26. Distributions of the chi-squared posterior predictive effect size. Labels in the right corner of each histogram indicate the data type, inference model, and tree length (in expected number of substitutions per codon). P-values were calculated using two-sample Cramer-Von Mises tests comparing inference runs using amino acid versus nucleotide models on the same underlying data.

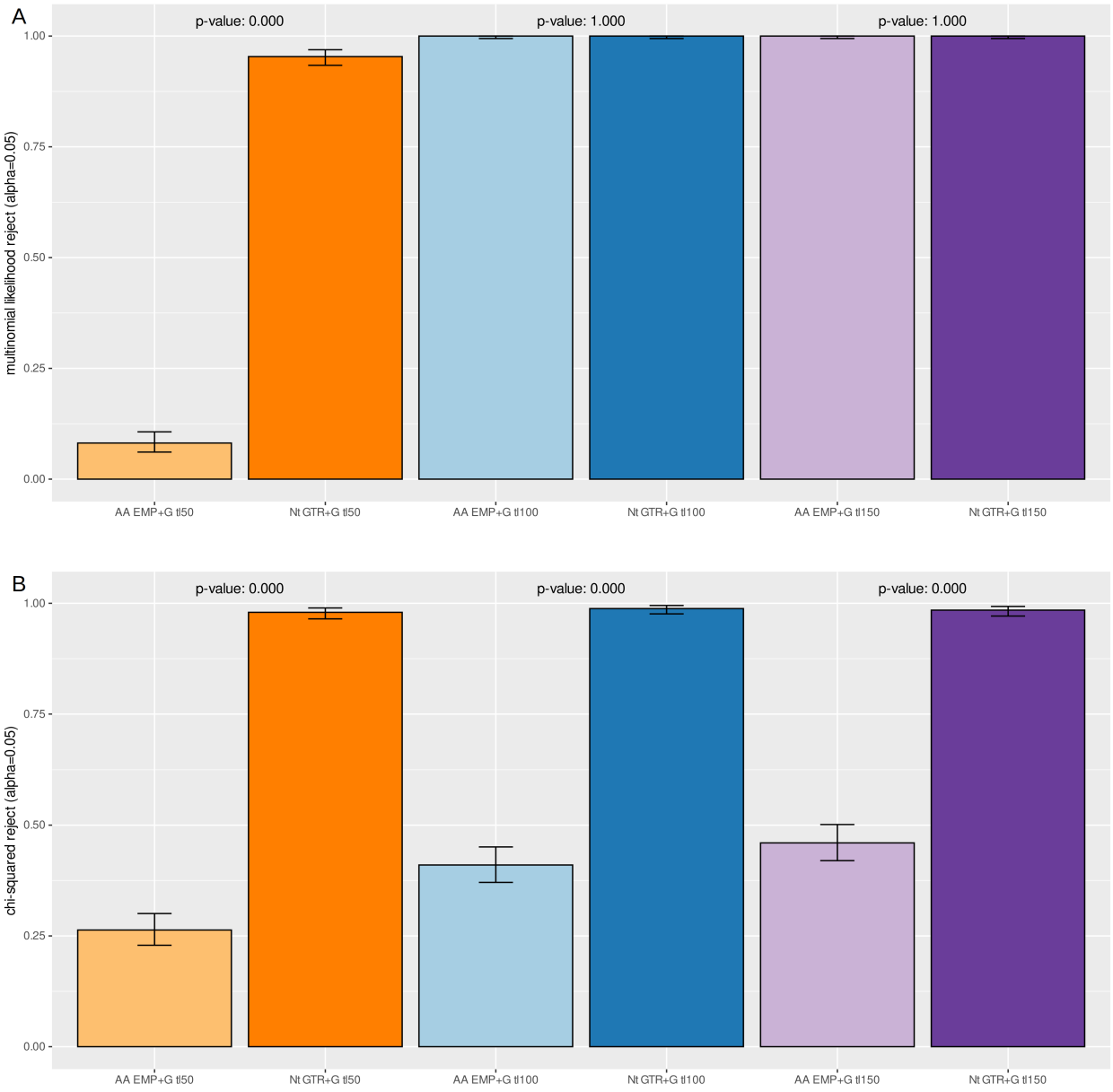

Figure 2.27. Fractions of instances when the inference model is rejected when using the multinomial likelihood statistic (A) and the chi-squared statistic (B). Labels beneath each bar indicate the data type, inference model, and tree length. P-values were calculated using Fisher's exact test comparing inference runs using amino acid versus nucleotide models on the same underlying data.

Inference	mean CID to true	CID MAP to true	true topology covered	TL % mean to true	TL mean to true over sd	true TL covered	Multinomial likelihood effect size	Multinomial likelihood reject	Chi-squared effect size	Chi-squared reject
AA EMP+G tl50 h. ω	0.29 (0.42) 0.899	0.23 (0.51) 1.000	0.99 (0.10) *0.022*	-8.2% (4.6%) *0.000*	-2.7 (1.6) *0.000*	0.33 (0.47) 1.000	1.40 (0.31) *0.000*	0.03 (0.17) *0.000*	1.46 (1.20) *0.000*	0.32 (0.47) *0.000*
Nt GTR+G tl50 h. ω	0.36 (0.59)	0.34 (0.64)	0.88 (0.33)	8.8% (18.6%)	1.3 (3.9)	0.32 (0.47)	2.48 (0.33)	0.93 (0.26)	29.02 (14.92)	1.00 (0.00)
AA EMP+G tl50 h. α	0.26 (0.40) 1.000	0.20 (0.46) 1.000	0.97 (0.17) 0.393	11.9% (6.2%) *0.000*	3.6 (1.7) *0.000*	0.19 (0.39) *0.000*	2.81 (0.52) *0.000*	0.97 (0.17) *0.025*	1.89 (1.54) *0.000*	0.44 (0.50) *0.000*
Nt GTR+G tl50 h. α	0.30 (0.48)	0.23 (0.55)	0.89 (0.31)	8.0% (15.7%)	0.9 (2.7)	0.56 (0.50)	2.22 (0.30)	0.84 (0.37)	29.04 (14.59)	1.00 (0.00)
AA EMP+G tl50 h. αω	0.23 (0.43) 1.000	0.21 (0.50) 1.000	0.94 (0.24) 0.727	8.0% (6.5%) *0.000*	2.5 (2.0) *0.001*	0.32 (0.47) 1.000	3.10 (0.78) *0.000*	0.93 (0.26) 1.000	1.38 (1.09) *0.000*	0.34 (0.48) *0.000*
Nt GTR+G tl50 h. αω	0.33 (0.50)	0.30 (0.58)	0.86 (0.35)	13.6% (19.6%)	1.6 (3.1)	0.41 (0.49)	2.35 (0.38)	0.89 (0.31)	28.79 (13.98)	1.00 (0.00)
AA EMP+G tl100 h. ω	0.34 (0.48) 1.000	0.25 (0.52) 1.000	0.96 (0.20) 1.000	-19.9% (4.8%) *0.000*	-8.8 (2.5) *0.000*	0.00 (0.00) *0.000*	3.99 (0.54) *0.000*	1.00 (0.00) 1.000	2.13 (1.59) *0.000*	0.50 (0.50) *0.000*
Nt GTR+G tl100 h. ω	0.35 (0.49)	0.30 (0.55)	0.91 (0.29)	-6.4% (15.1%)	-1.8 (2.6)	0.69 (0.46)	3.25 (0.34)	1.00 (0.00)	32.05 (15.44)	1.00 (0.00)
AA EMP+G tl100 h. α	0.27 (0.38) 1.000	0.20 (0.44) 0.200	0.99 (0.10) 0.285	-4.2% (5.6%) *0.000*	-2.0 (2.5) 0.162	0.49 (0.50) 0.494	7.15 (0.95) *0.000*	1.00 (0.00) 1.000	2.84 (2.06) *0.000*	0.64 (0.48) *0.000*
Nt GTR+G tl100 h. α	0.35 (0.50)	0.35 (0.57)	0.92 (0.27)	-4.4% (18.1%)	-1.5 (2.3)	0.63 (0.49)	3.90 (0.44)	1.00 (0.00)	31.17 (15.08)	1.00 (0.00)
AA EMP+G tl100 h.αω	0.30 (0.49) 1.000	0.26 (0.56) 1.000	0.92 (0.27) 0.448	-9.4% (5.4%) *0.000*	-4.3 (2.6) *0.000*	0.16 (0.37) *0.000*	7.17 (1.23) *0.000*	1.00 (0.00) 1.000	2.18 (1.43) *0.000*	0.52 (0.50) *0.000*
Nt GTR+G tl100 h.αω	0.35 (0.58)	0.30 (0.59)	0.82 (0.39)	-6.4% (16.8%)	-1.8 (2.8)	0.69 (0.46)	4.02 (0.48)	1.00 (0.00)	30.20 (13.74)	1.00 (0.00)
AA EMP+G tl150 h. ω	0.32 (0.42) 1.000	0.23 (0.48) 1.000	0.95 (0.22) 1.000	-28.7% (5.1%) *0.000*	-15.3 (3.6) *0.000*	0.00 (0.00) *0.000*	6.47 (0.93) *0.000*	1.00 (0.00) 1.000	2.42 (1.80) *0.000*	0.55 (0.50) *0.000*
Nt GTR+G tl150 h. ω	0.36 (0.47)	0.29 (0.53)	0.91 (0.29)	-22.4% (11.7%)	-3.3 (3.5)	0.48 (0.50)	4.19 (0.35)	1.00 (0.00)	32.87 (15.43)	1.00 (0.00)
AA EMP+G tl150 h. α	0.25 (0.45) 1.000	0.21 (0.51) 1.000	0.96 (0.20) 0.497	-15.2% (5.8%) *0.000*	-7.9 (3.5) *0.000*	0.03 (0.17) *0.000*	10.93 (1.35) *0.000*	1.00 (0.00) 1.000	3.58 (2.07) *0.000*	0.74 (0.44) *0.000*
Nt GTR+G tl150 h. α	0.31 (0.53)	0.28 (0.58)	0.88 (0.33)	-19.7% (14.3%)	-2.9 (3.2)	0.56 (0.50)	5.69 (0.64)	1.00 (0.00)	32.99 (14.44)	1.00 (0.00)
AA EMP+G tl150 h.αω	0.29 (0.47) 1.000	0.26 (0.51) 1.000	0.89 (0.31) 1.000	-21.4% (5.7%) *0.000*	-11.8 (3.7) *0.000*	0.01 (0.10) *0.000*	10.43 (1.56) *0.000*	1.00 (0.00) 1.000	2.61 (1.76) *0.000*	0.61 (0.49) *0.000*
Nt GTR+G tl150 h.αω	0.35 (0.53)	0.33 (0.59)	0.83 (0.38)	-22.3% (16.7%)	-3.3 (3.3)	0.45 (0.50)	5.80 (0.71)	1.00 (0.00)	31.50 (14.77)	1.00 (0.00)

Table 2.4. Topology and tree length inference accuracy and posterior predictive test statistics using amino acid and nucleotide models. Simulation was done using three variants of a MutSel codon model: branch-heterogeneous dN/dS (ω), branch-heterogeneous ASRV (α), and branch-heterogeneous dN/dS and ASRV (αω). Displayed values are averages across 100 simulation/inference runs. Standard deviations are in parentheses. Benjamini-Yekutieli adjusted p-values for amino acid versus nucleotide comparisons are listed below the standard deviation of each amino acid entry. Adjusted p-values smaller than 0.05 are italicized and in red.

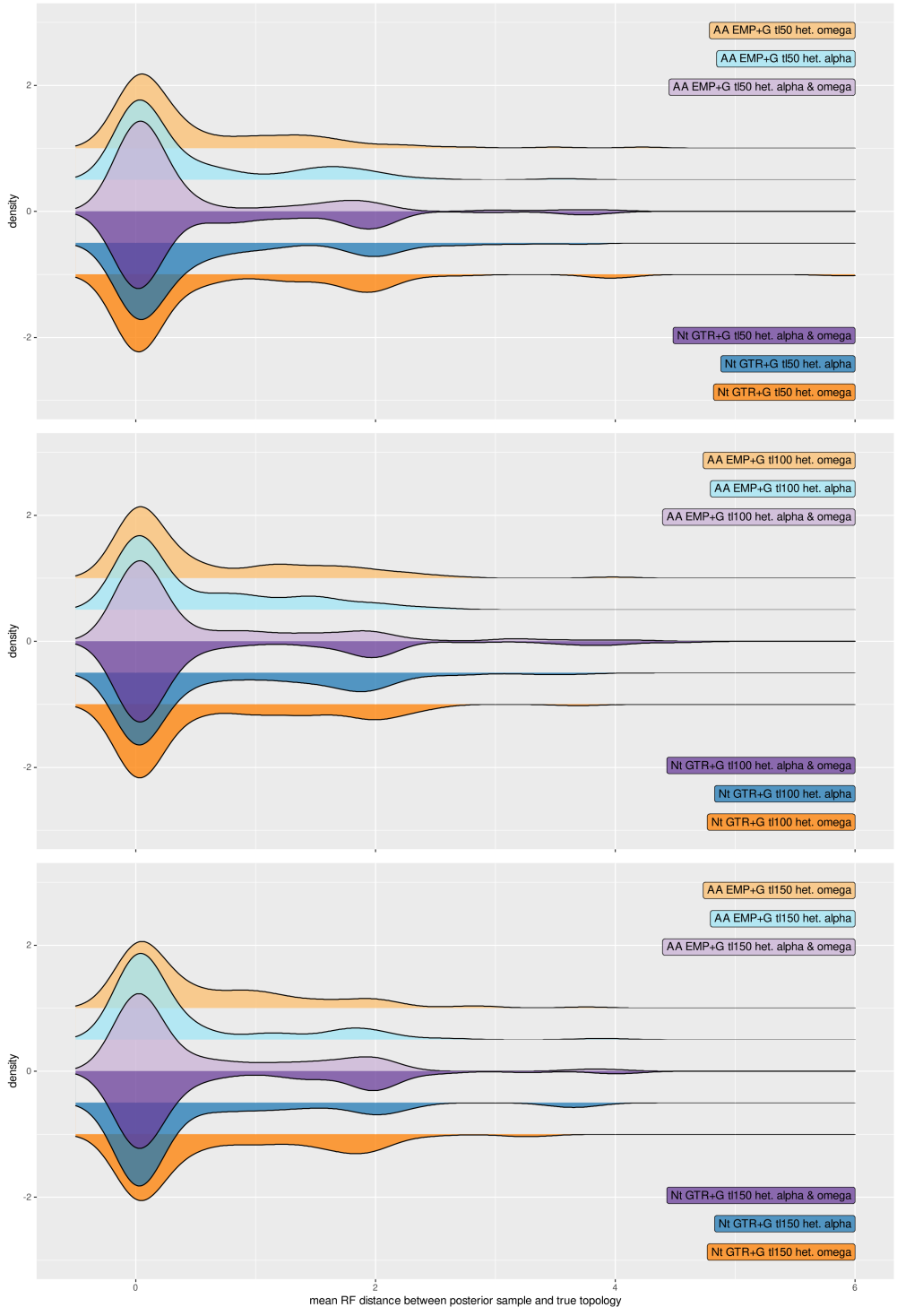

Figure 2.28. Distributions of the mean RF distance between every topology in the posterior sample and the true topology. Labels in the right corners indicate the data type, inference model, tree length (in expected number of substitutions per codon), and simulation model variant.

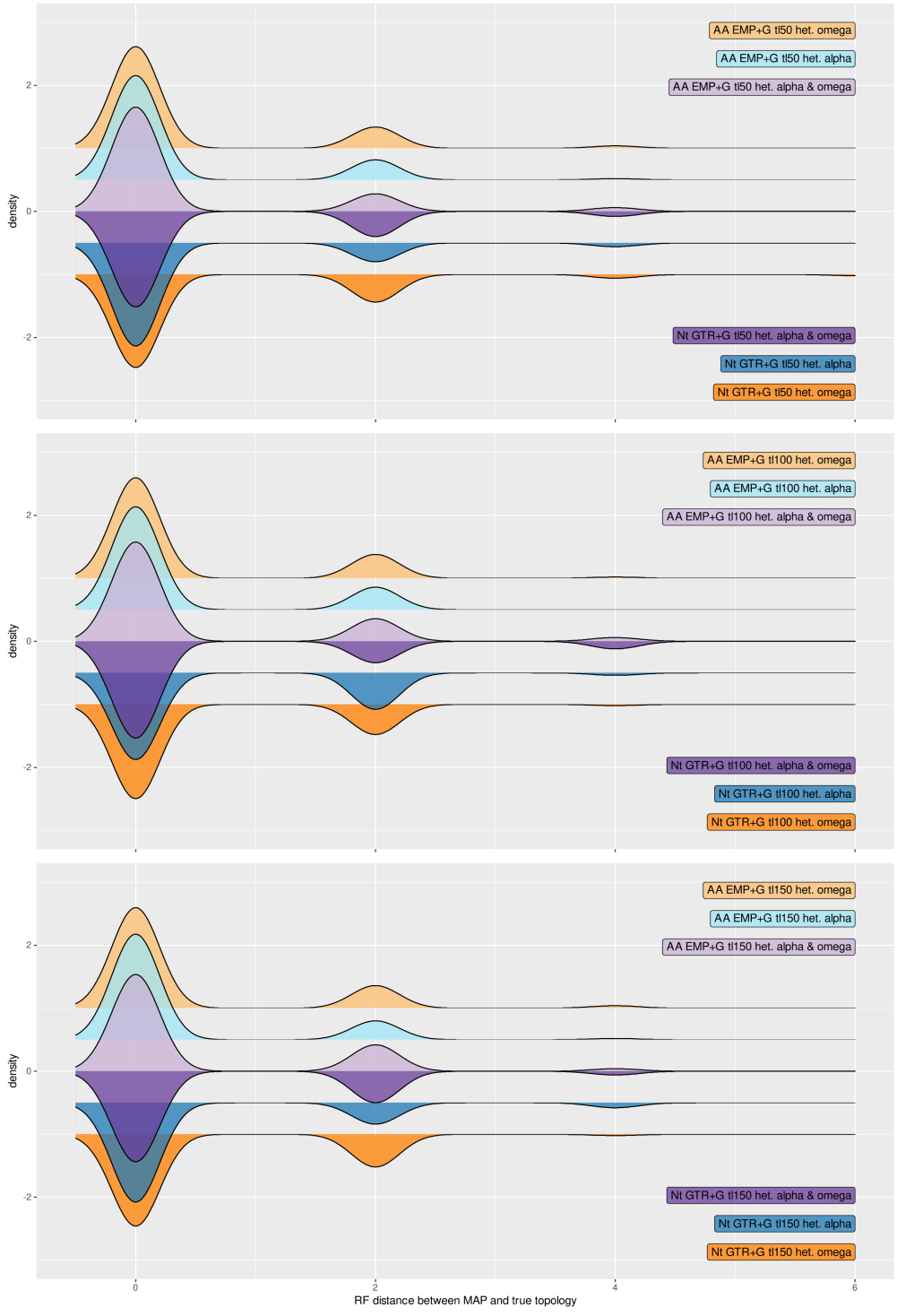

Figure 2.29. Distributions of the RF distance between the MAP topology and the true topology. Labels in the right corners indicate the data type, inference model, tree length (in expected number of substitutions per codon), and simulation model variant.

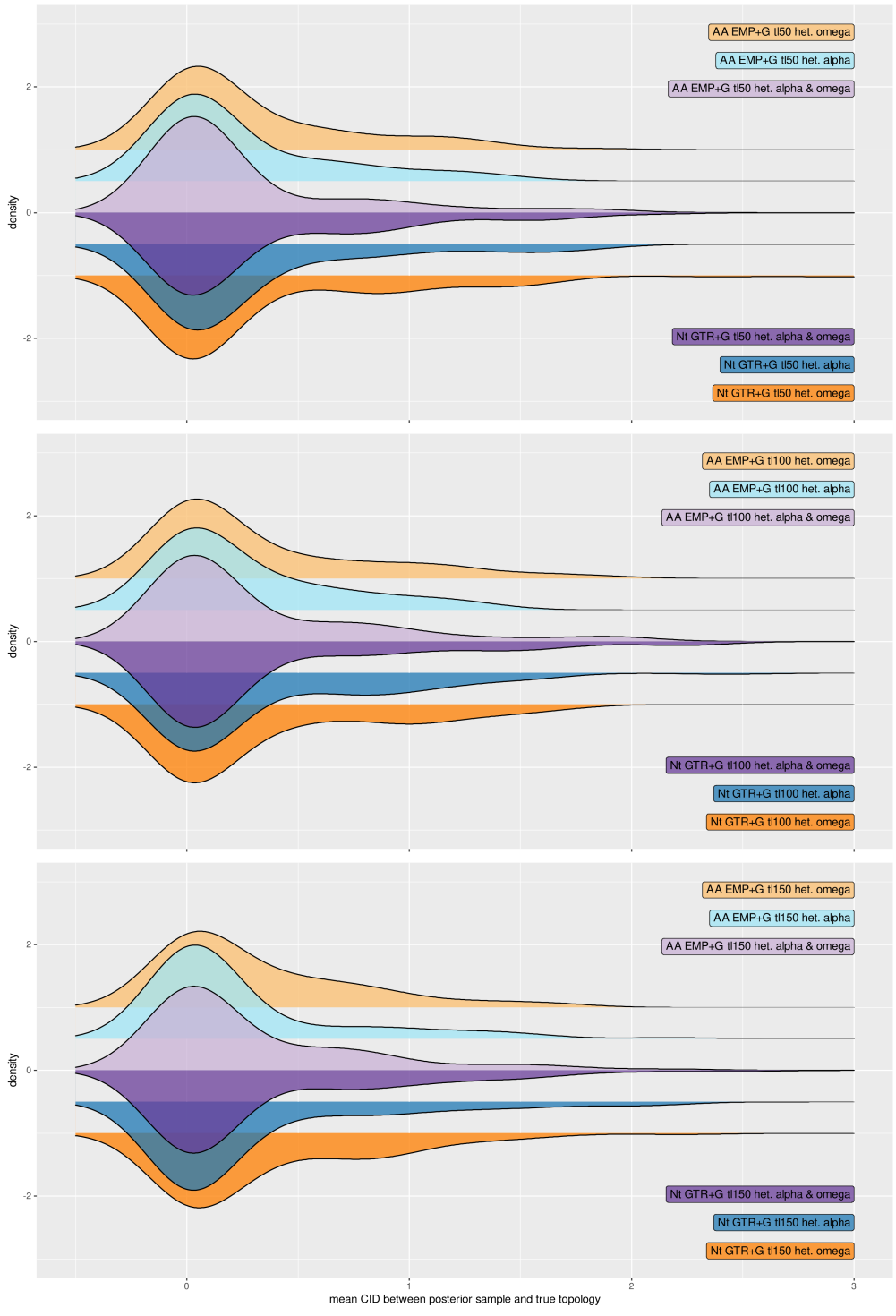

Figure 2.30. Distributions of the mean CID between every topology in the posterior sample and the true topology. Labels in the right corners indicate the data type, inference model, tree length (in expected number of substitutions per codon), and simulation model variant.

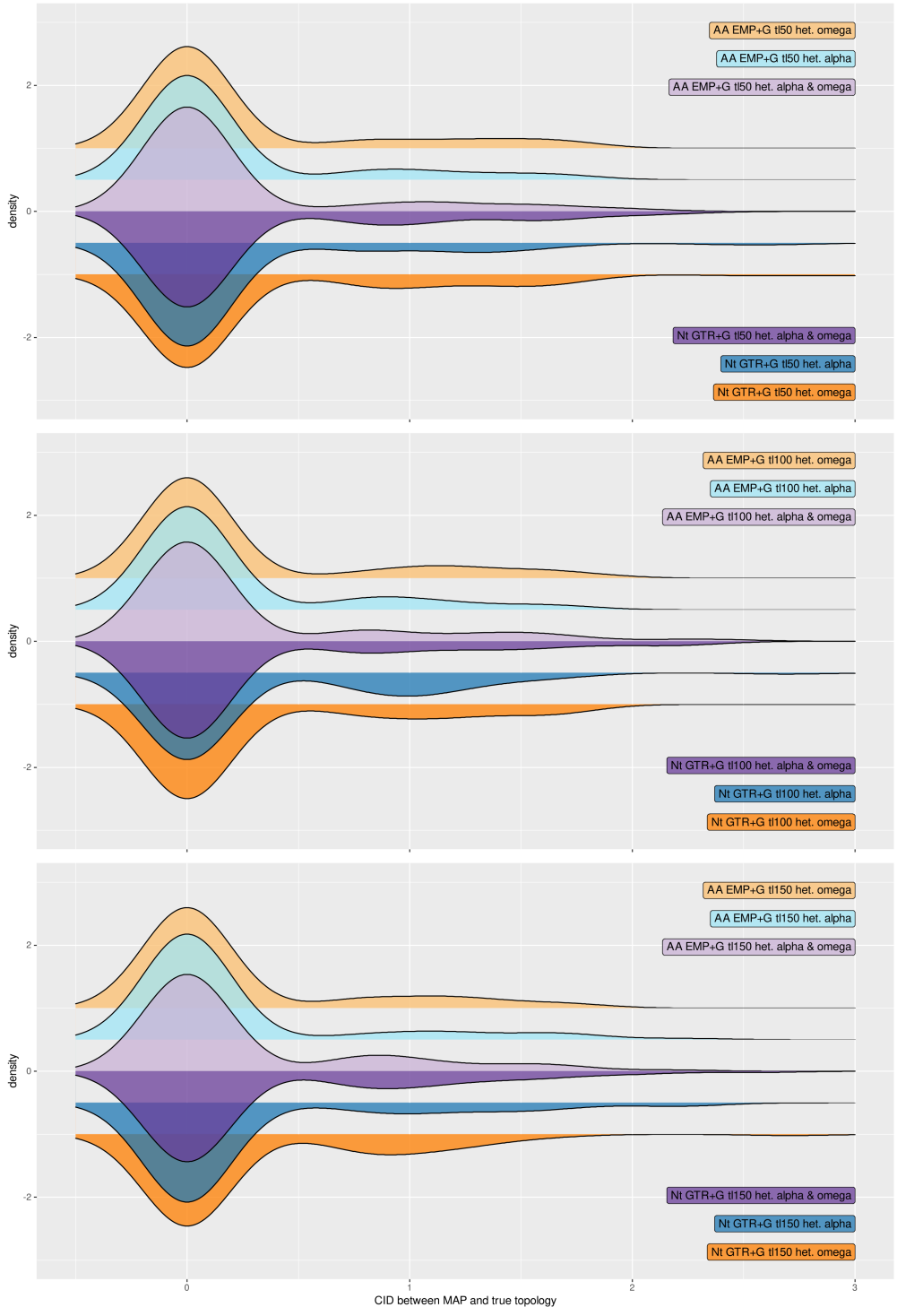

Figure 2.31. Distributions of the CID between the MAP topology and the true topology. Labels in the right corners indicate the data type, inference model, tree length (in expected number of substitutions per codon), and simulation model variant.

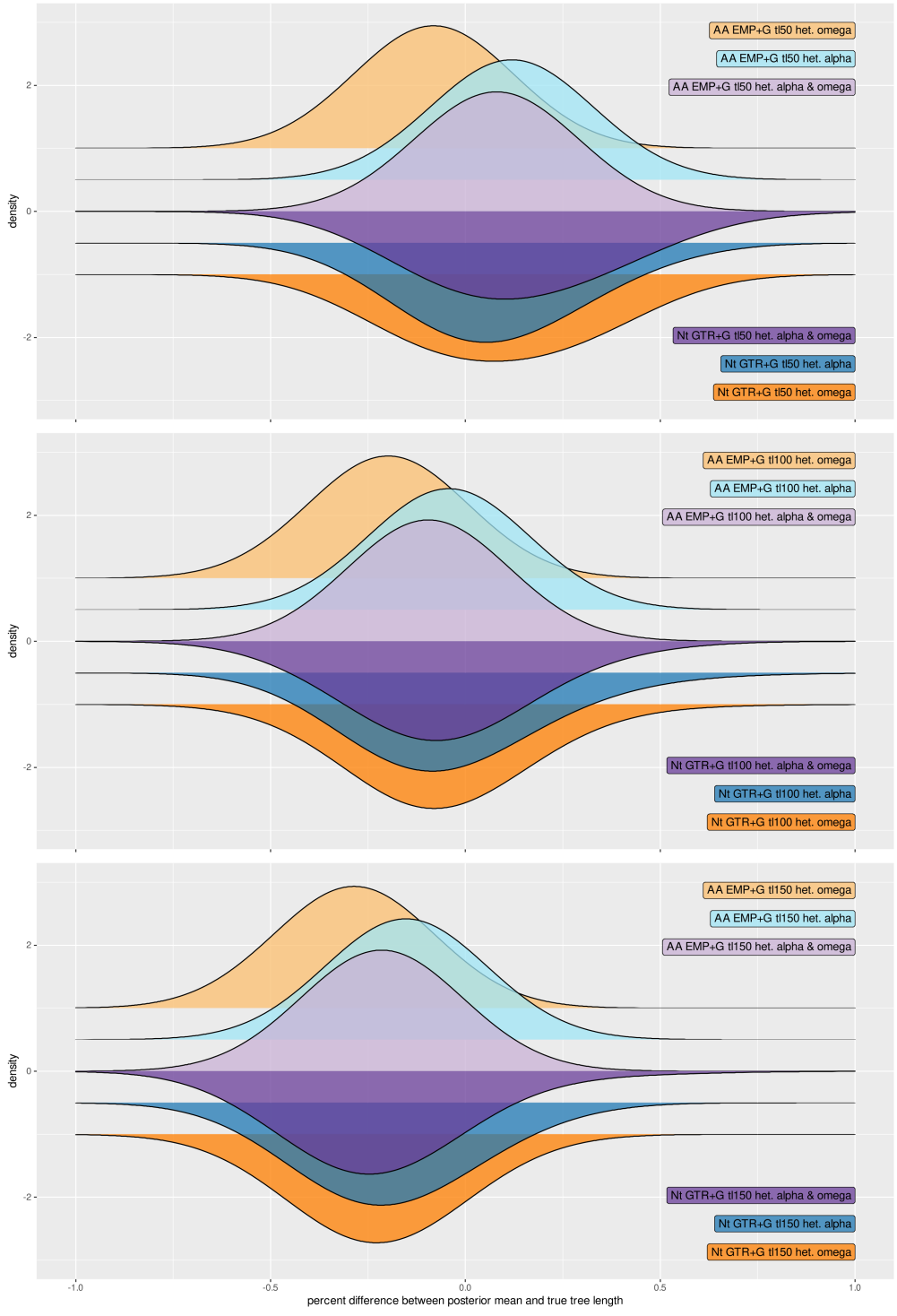

Figure 2.32. Distributions of the percent difference between estimated and true tree lengths; i.e. (estimated - true) ÷ true. X-axis labels are in decimal (1 corresponds to 100%). Labels in the right corners indicate the data type, inference model, tree length (in expected number of substitutions per codon), and simulation model variant.

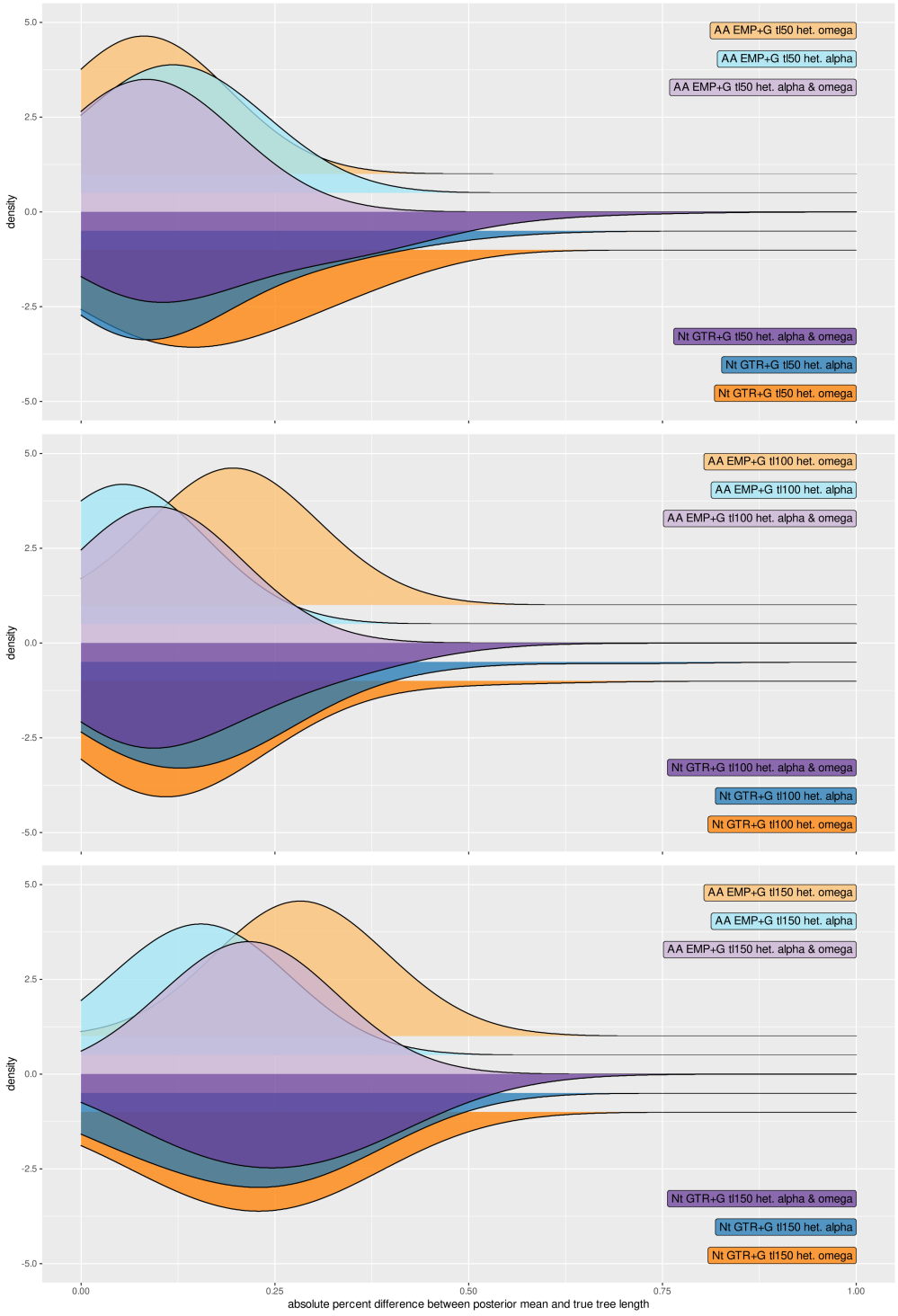

Figure 2.33. Distributions of the absolute value of the percent difference between estimated and true tree lengths; i.e. |(estimated - true) ÷ true|. X-axis labels are in decimal (1 corresponds to 100%). Labels in the right corners indicate the data type, inference model, tree length (in expected number of substitutions per codon), and simulation model variant.

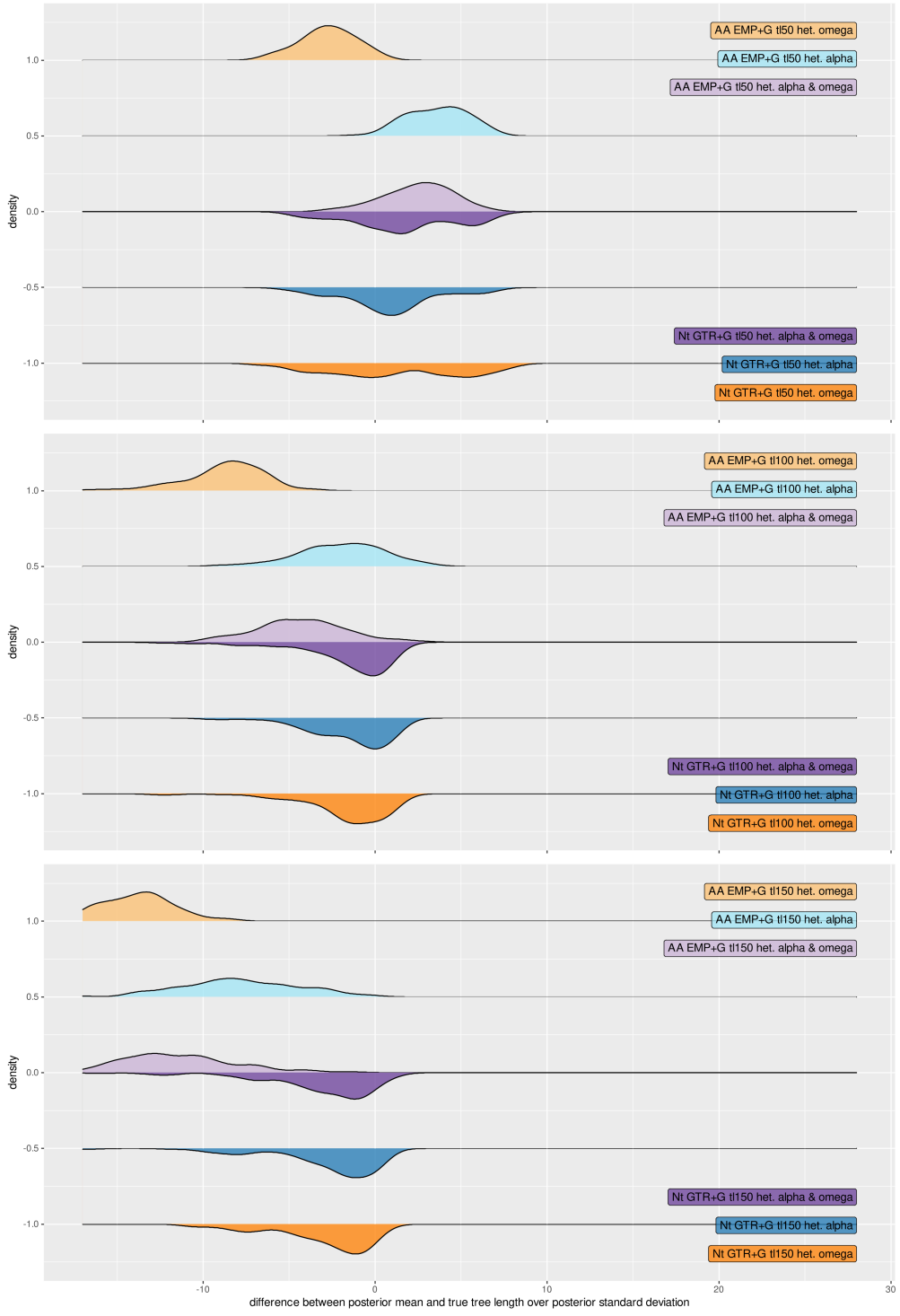

Figure 2.34. Distributions of the difference between estimated and true tree lengths as a multiple of the standard deviation of the posterior distribution. Labels in the right corners indicate the data type, inference model, tree length (in expected number of substitutions per codon), and simulation model variant.

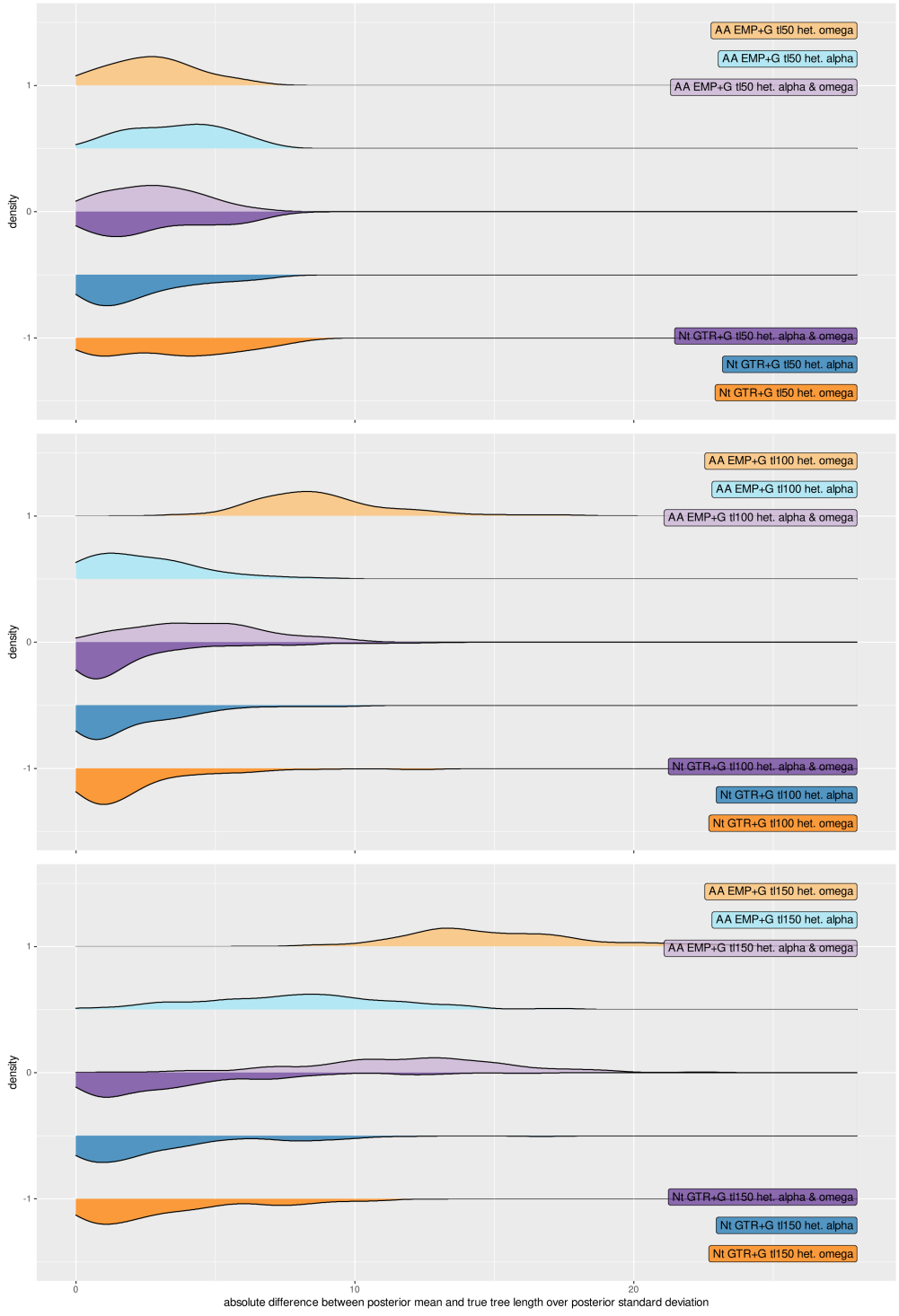

Figure 2.35. Distributions of the absolute value of the difference between estimated and true tree lengths as a multiple of the standard deviation of the posterior distribution. Labels in the right corners indicate the data type, inference model, tree length (in expected number of substitutions per codon), and simulation model variant.

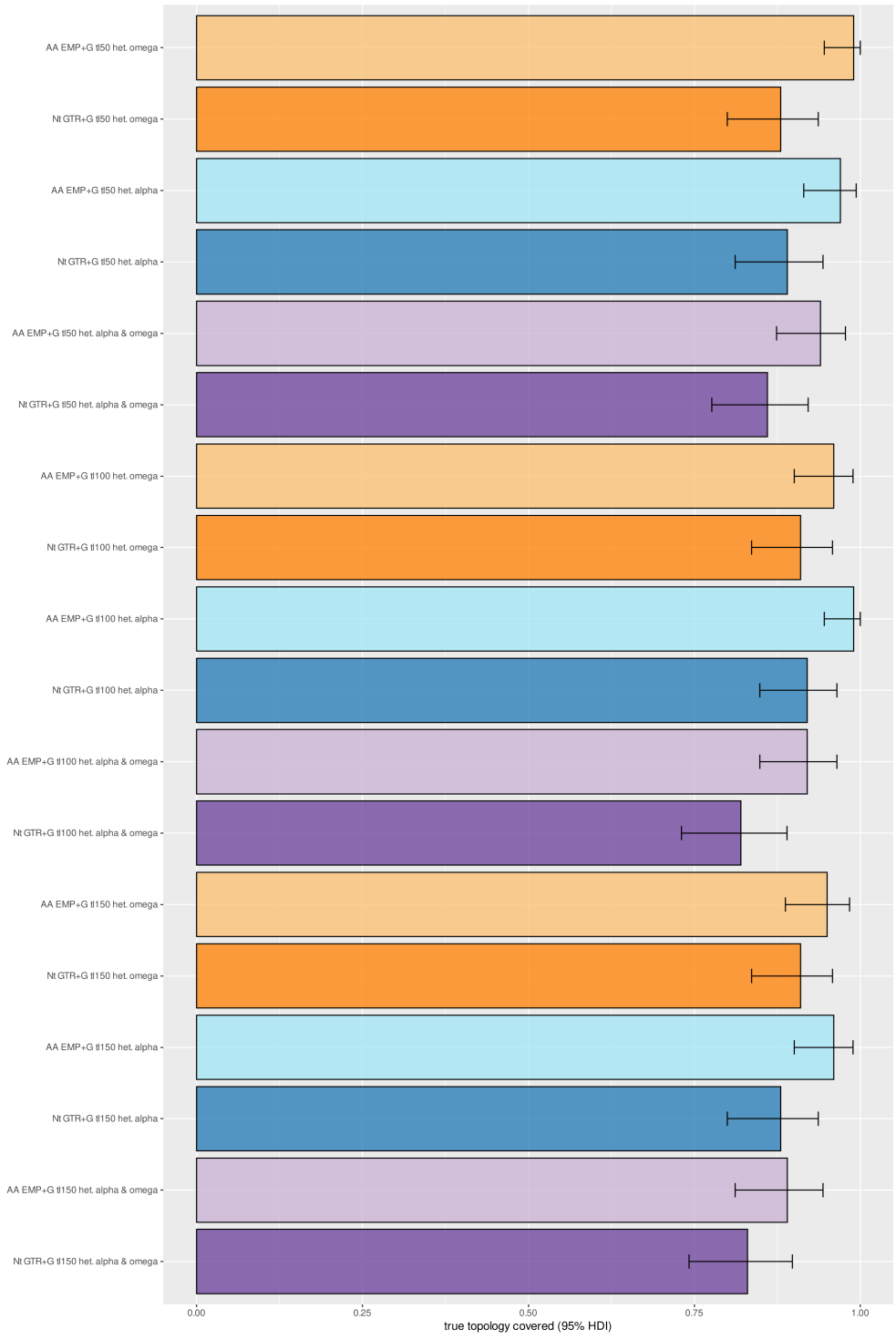

Figure 2.36. Fractions of instances when the true topology is covered by the 95% credible set. Labels to the left of each bar indicate the data type, inference model, tree length (in expected number of substitutions per codon), and simulation model variant.

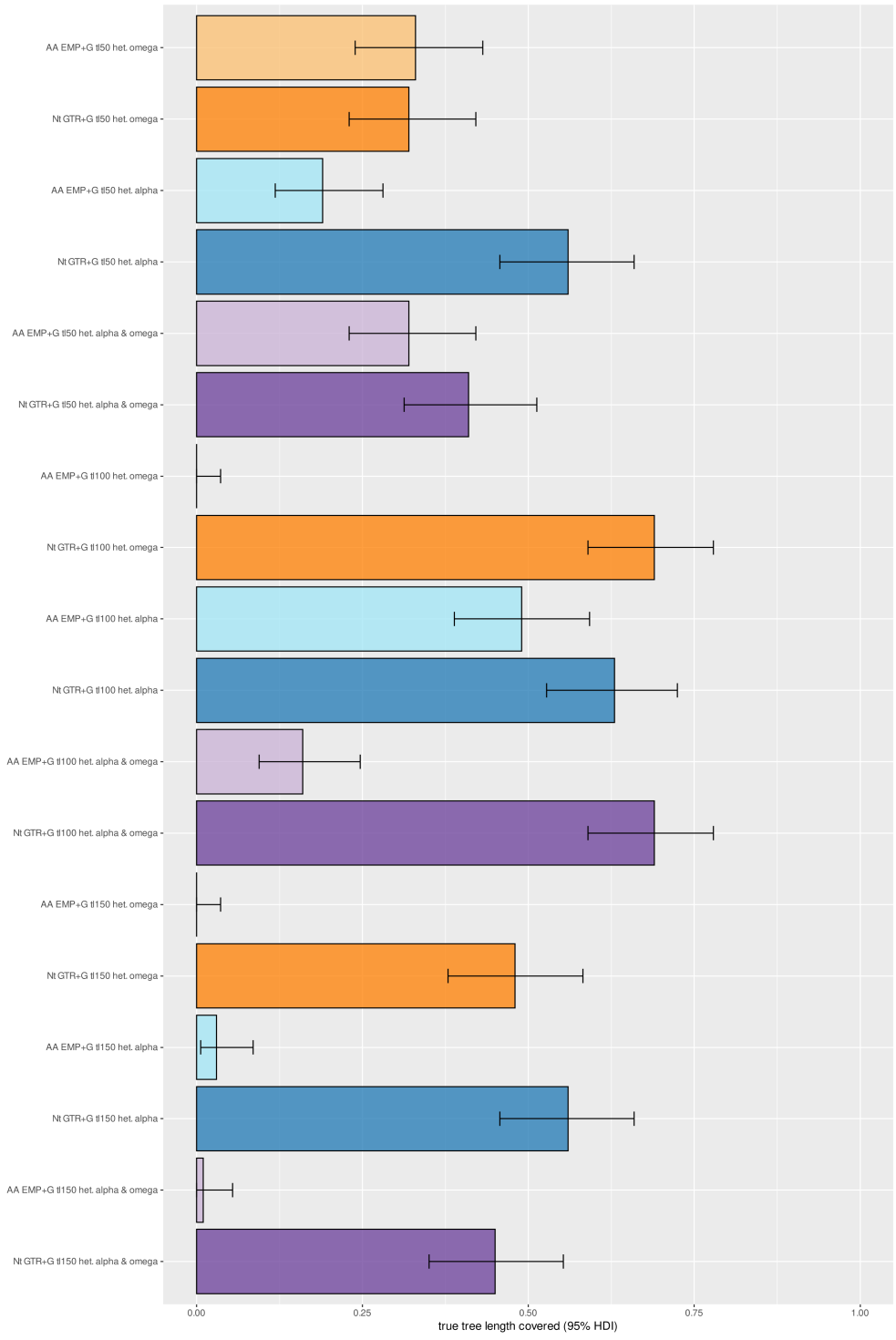

Figure 2.37. Fractions of instances when the true tree length is covered by the 95% credible interval. Labels to the left of each bar indicate the data type, inference model, tree length (in expected number of substitutions per codon), and simulation model variant.

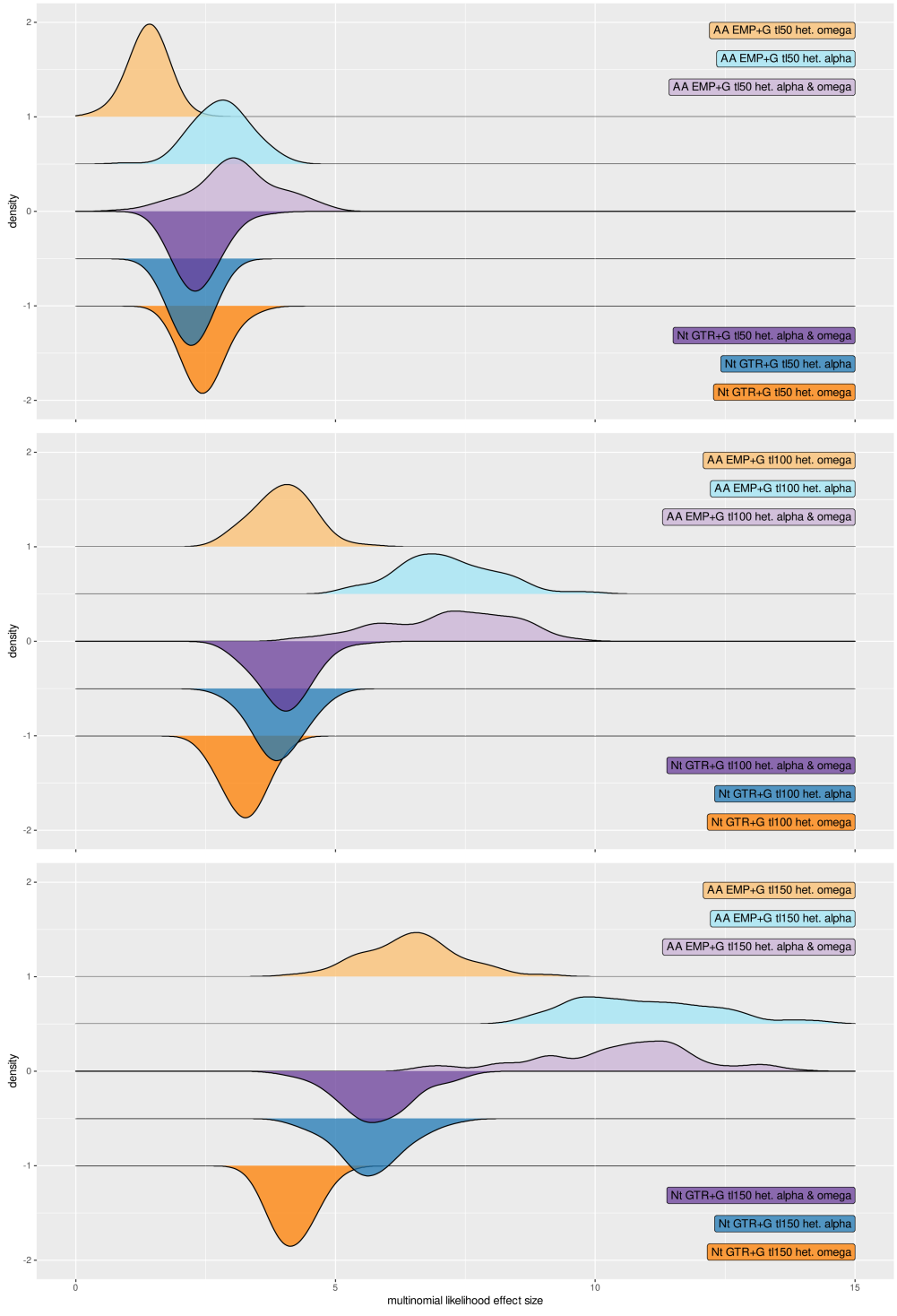

Figure 2.38. Distributions of the multinomial likelihood effect size. Labels in the right corners indicate the data type, inference model, tree length (in expected number of substitutions per codon), and simulation model variant.

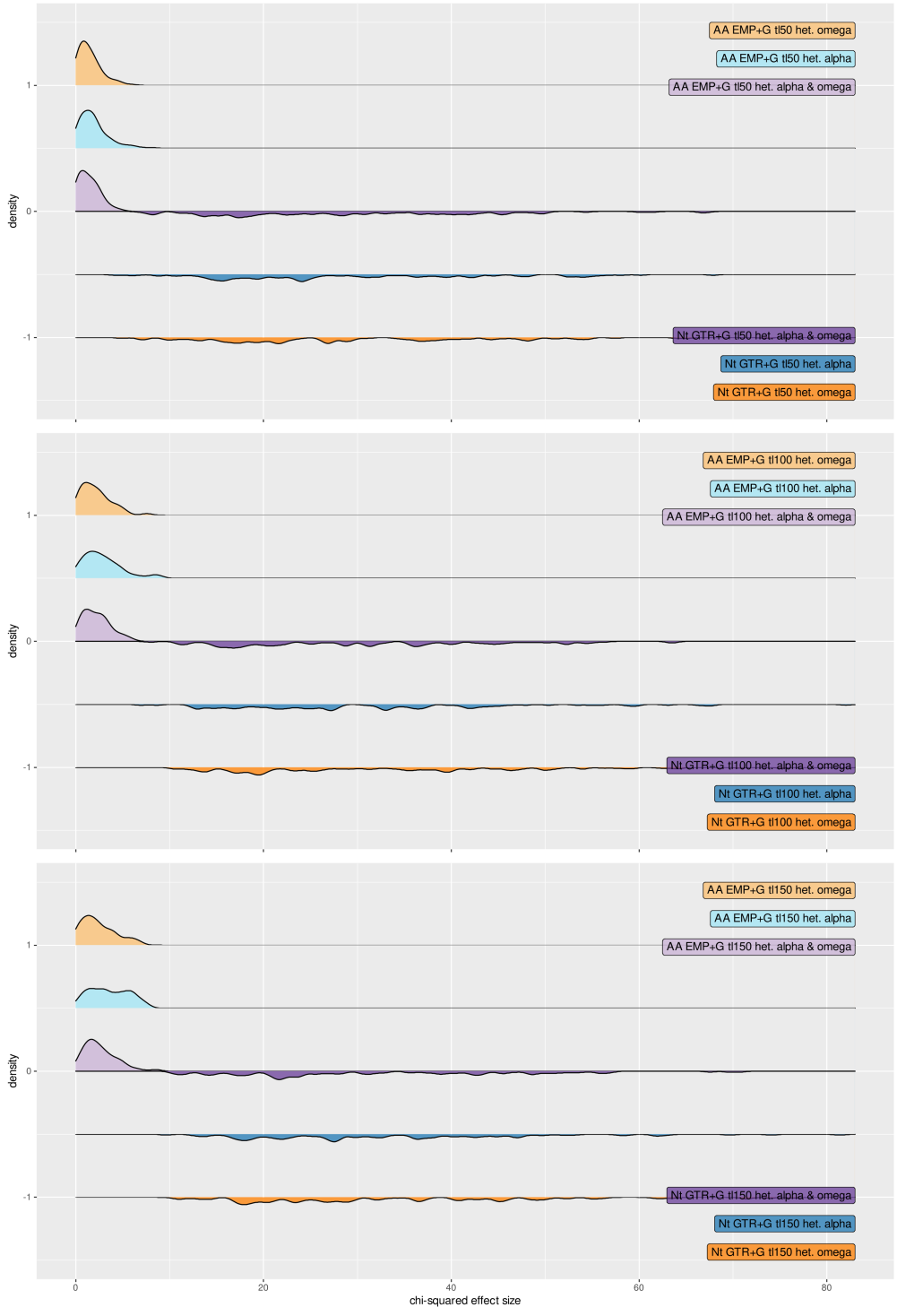

Figure 2.39. Distributions of the chi-squared effect size. Labels in the right corners indicate the data type, inference model, tree length (in expected number of substitutions per codon), and simulation model variant.

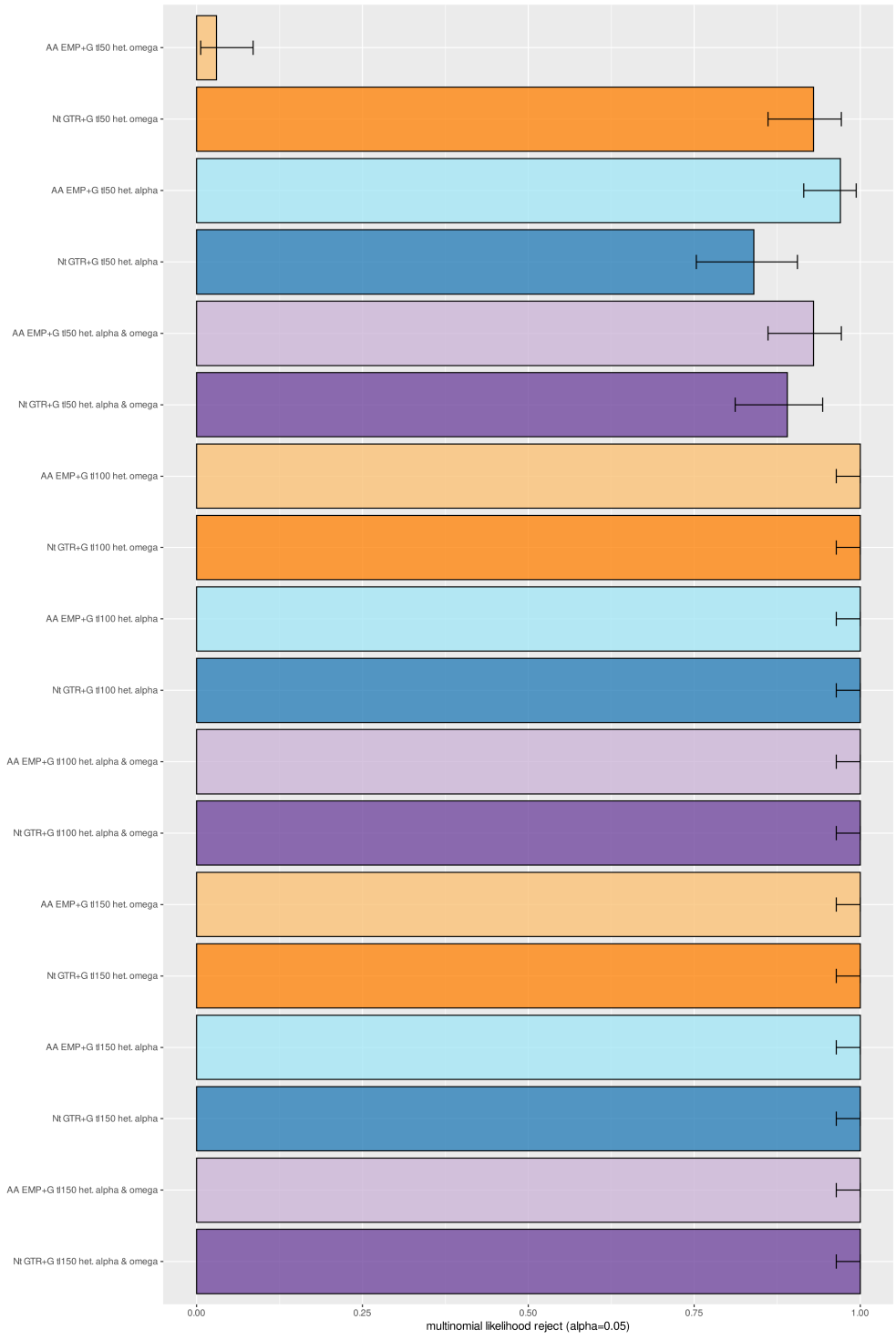

Figure 2.40. Fractions of instances when the inference model is rejected when using the multinomial likelihood statistic. Labels to the left of each bar indicate the data type, inference model, tree length (in expected number of substitutions per codon), and simulation model variant.

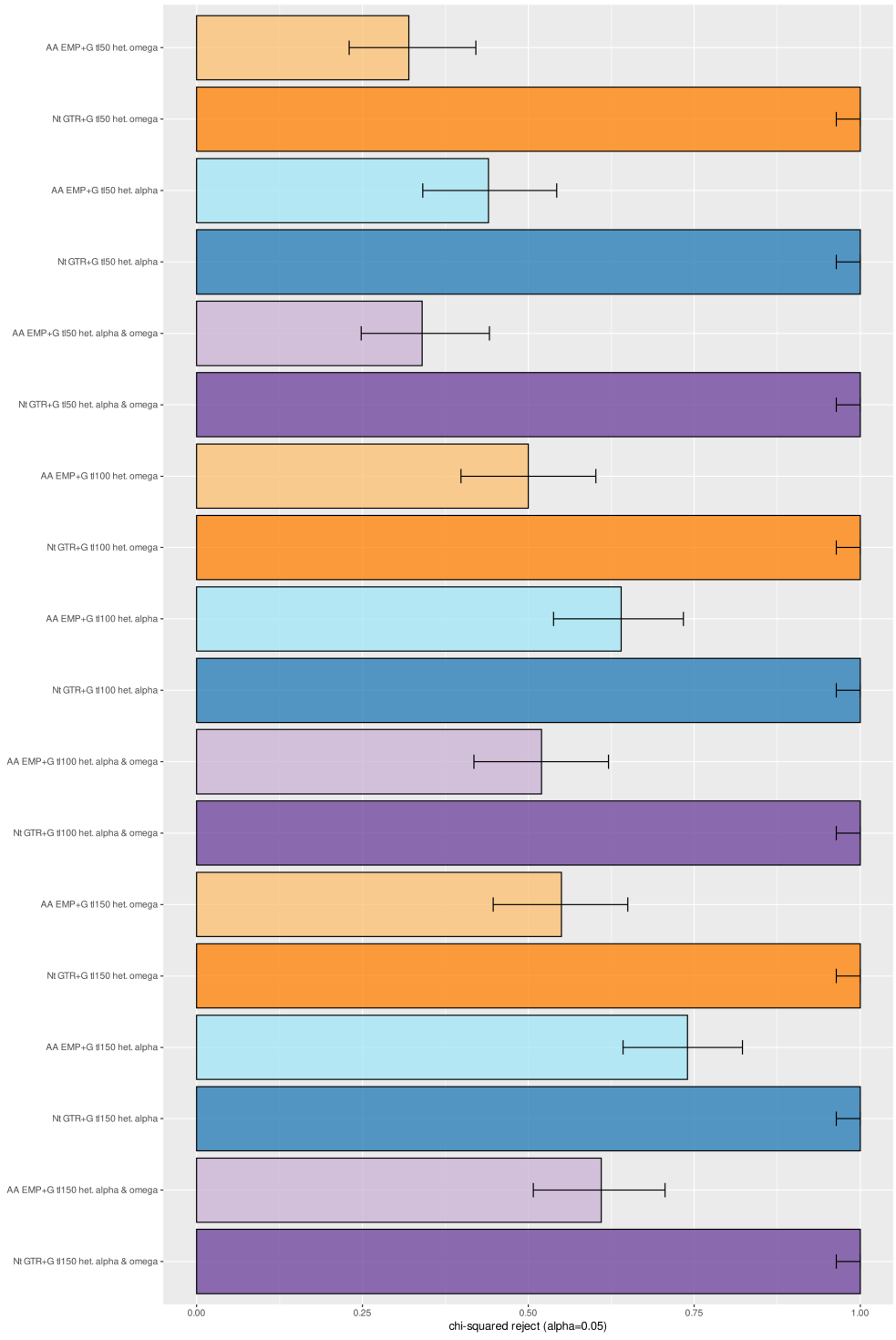

Figure 2.41. Fractions of instances when the inference model is rejected when using the chi-squared statistic. Labels to the left of each bar indicate the data type, inference model, tree length (in expected number of substitutions per codon), and simulation model variant.

Key to the Nearctic species of Scoliidae (Hymenoptera)

Khouri, Z.[1], Kimsey, L.S.[1]

[1]Bohart Museum of Entomology, University of California, Davis, CA, U.S.A.

Introduction

Members of the family Scoliidae, sometimes referred to as mammoth wasps, are a clade of aculeate Hymenoptera related to bradynobaenids, ants, and apoids (Johnson *et al*., 2013; Branstetter *et al*., 2017; Peters *et al*., 2018). There are approximately 560 described species with a distribution covering all continents except Antarctica (Osten, 2005). Mammoth wasps are parasitoids of scarabaeid beetle larvae (Clausen, 1940) and have consequently been evaluated and used as biological control agents (Illingworth, 1921; Wilson, 1960; DeBach, 1964). Despite this, the family remains relatively poorly studied, with an unstable taxonomy and the difficulty of specimen identification being obstacles to research (Day *et al.*, 1981; Elliot, 2011).

This publication aims to provide a key to the Nearctic scoliid fauna. All existing identification resources covering the region are limited in geographic scope (e.g. Porter (1981) treating the fauna of the Lower Rio Grande Valley) or taxonomic coverage (e.g. Bradley (1928a) revising the genus *Colpa*) or contain factual errors (e.g. MacKay (1987) incorrectly implying that the longer hind tibial spur in *Xanthocampsomeris limosa* is acute and black). Additionally, even the more recent literature on American mammoth wasps (e.g. Grissell, 2007) uses outdated taxonomic names, making it difficult for non-experts to make connections to other regional (e.g. Liu *et al*., 2021) or global (Osten, 2005) taxonomic treatments of the group.

We cover all species present in the Nearctic region as well as in the Neotropical part of the continental United States (i.e. southern Florida). However, due to the authors having more

limited access to material from northern and central Mexico, we expect the key to be less reliable there. Some covered species have very broad distributions and exhibit considerable geographically-correlated variation in color and sometimes other characters (e.g. *Scolia guttata* and *Pygodasis ephippium*). In such cases, character states used in the key may not accurately represent specimens collected outside the geographic scope of this study. We cover subspecies only in the cases of *Dielis plumipes* and *Scolia dubia*, where morphological differences between groups currently recognized as subspecies are comparable to those distinguishing species. Both subspecies of *S. dubia* are sympatric in the southern United States (MacKay, 1987), indicating they may be separate species.

This key was built using existing identification resources (Bartlett, 1912; Rohwer, 1927; Bradley, 1928a, b, 1957; Porter, 1981; MacKay, 1987; Grissell, 2007), the primary taxonomic literature, and specimens in the Bohart Museum of Entomology, the California Department of Food and Agriculture entomology collection, and the University of Florida entomology collection.

Note on taxonomic names

We use names compatible with the world checklist of Osten (2005). The existing keys cited above refer to all New World Campsomerini (excluding *Colpa*) as *Campsomeris*. The many subgenera of *Campsomeris* (see Bradley, 1957) have since been elevated to genus rank, and the genus *Campsomeris, sensu* Osten (2005), no longer contains any Nearctic species. Nearctic taxa previously in *Campsomeris* are now in *Dielis*, *Pygodasis*, and *Xanthocampsomeris*. Conversely, species formerly in *Trielis* and *Crioscolia* are now in *Colpa*, with *Trielis* being considered a junior synonym of *Colpa* (Day *et al*., 1981) and *Crioscolia* being treated as a subgenus. Some

species names included in the older keys (e.g. *Scolia consors* in MacKay, 1987) have since been synonymized and are not treated here. Refer to Osten (2005) for lists of synonyms.

Notes on problematic taxa

Scolia bicincta Fabricius, 1775:

Some specimens that match the description of *Scolia bicincta* and would be identified as such using existing resources such as Grissell (2007) differ from each other morphologically. Specifically, some individuals have a tubercle that comes to a central point on the second sternum of the metasoma (S2) and a deeply undercut, medially emarginate, transverse furrow near the base of the first metasomal sternum (S1), as seen in *Scolia mexicana* (Fig. 3.1A). We have not found any differences between these specimens and those of *S. mexicana* apart from color. Other individuals have a laterally extended tubercle on S2 that lacks a central point and a non-emarginate or weakly emarginate, shallow furrow on S1 (Fig. 3.1B). These latter specimens match the specimens in the Fabricius collection in the Natural History Museum of Denmark, University of Copenhagen (Fig. 3.2). Phylogenetic analysis of ultraconserved element data (Fig. 3.3, see Supporting Information for methods) recovered these specimens as forming a monophyletic group sister to *Scolia dubia*. On the other hand, the "non-typical" specimens formed a monophyletic group sister to the only specimen of *Scolia mexicana* included in the analysis.

Fabricius (1775) referred to the Banks collection in the Natural History Museum, London, when he described *S. bicincta*. Turner (1909) subsequently stated that the type is in the Banks collection. Bradley (1964a) disagreed, noting that the specimen in the Banks collection labeled as *S. bicincta* did not match the description of Fabricius (1775) as it has three light-colored bands on the abdomen, while the description refers to two bands. Bradley (1964a) further stated that he

had placed a neotype label on specimen number 71 from the Kiel Fabricius collection in the Natural History Museum of Denmark, University of Copenhagen, but did not formally designating a neotype. We have found that Bradley's neotype label is actually on a different specimen in the "Coll. Sehested & Tønder-Lund" tray. Specimen 71 in the "Coll. I. C. Fabricius" tray has since suffered pest damage, with the head mostly destroyed.

Given the morphological observations and phylogenetic results of the present study, the identity of the *S. bicincta* type is important. We have not been able to examine the specimen in the Banks collection. If that specimen matches the specimens in the Copenhagen Fabricius collection, both *S. bicincta* and *S. mexicana* remain valid species. Either the Banks specimen would be recognized as the holotype, or a lectotype might be chosen if the Copenhagen specimens are demonstrated to be from the same type series. If the latter cannot be established and the Banks specimen corresponds morphologically to *S. mexicana*, *S. mexicana* would become a junior synonym of *S. bicincta*, and a new species would need to be described to represent the group that is closely related to *S. dubia* and that includes the majority of specimens currently labelled as *S. bicincta* in collections. This would be undesirable, as it conflicts with the historical application of these names and is likely to cause confusion. In this case, designating a neotype that preserves prevailing usage might be justified (article 75.6 of the International Code of Zoological Nomenclature).

Pending clarification of the identity of the *S. bicincta* type, we conservatively treat specimens with *S. mexicana* morphology but yellow bands on the metasoma as *S. mexicana* in this key, as we are unable to find differences apart from color, and as these specimens do not, to our knowledge, occur in sympatry with typical *S. mexicana*. This extends the known range of *S. mexicana* north to the state of Maryland.

Scolia nobilitata Fabricius, 1805:

This is a very widespread species present on both sides of the 100th meridian (Bartlett, 1912; Grissell, 2007). Individuals vary greatly in color. Osten (2005) lists three subspecies and numerous synonyms.

Dielis pilipes (Saussure, 1852):

Bradley (1964b) noted that *Dielis pilipes* should be excluded from *Dielis* but did not provide any argumentation. This change was never implemented. We observe that *D. pilipes* lacks some characters consistently present in other *Dielis*, such as a medial longitudinal depression on the clypeus of the female and a deep, straight transverse furrow on the base of S1. The molecular phylogenies inferred in Chapter 1 recover *D. pilipes* as sister to *Xanthocampsomeris limosa* (the only species of *Xanthocampsomeris* included in the study) and confirm that *D. pilipes* in not closely related to other *Dielis*. However, we refrain from making taxonomic changes, pending a revision of *Dielis* and *Xanthocampsomeris*.

Dielis and *Xanthocampsomeris* males:

Male Campsomerini tend to be more morphologically uniform than their respective females. While subspecies of *Dielis plumipes* are easily distinguished in the case of female specimens, there are no consistent characters distinguishing males. We therefore do not include *D. plumipes* subspecies in the male key. Males of all *Dielis* species are very similar morphologically, and locality information may be crucial for correct identification. Refer to Bradley (1928b) for distribution maps. This issue is even more severe in the case of *Xanthocampsomeris*. The males of *Xanthocampsomeris hesterae* and *Xanthocampsomeris completa* have not been described, and identification is unreliable in geographic areas where both occur as well as where distributions

overlap with that of *Xanthocampsomeris limosa*. The genus is in need of revision, and molecular data may be necessary to correctly associate the sexes.

Notes on terminology

We follow the terminology of Michener (1944) for wing venation (Fig. 3.4). Numbered terga and sterna refer to terga and sterna of the metasoma; i.e. tergum 1 and sternum 1 (abbreviated as T1 and S1 respectively) are the first tergum and sternum of the metasoma and the second tergum and sternum of the true abdomen.

Key to sexes

1 **a.** Antenna with 12 segments; 6 visible metasomal terga; metasomal S8 without apical spines .....Female

b. Antenna with 13 segments; 7 visible metasomal terga; metasomal S8 with 3 apical spines .....Male

Key to females

1 **a.** Forewing with second recurrent vein (2m-cu) absent 2

b. Forewing with second recurrent vein present 9

2(1) **a.** Forewing with three submarginal cells (vein 1r-m present), vein 2r-m usually meets 1r-m 3

b. Forewing with two submarginal cells (vein 1r-m absent) 4

3(2) **a.** Metasomal T3-6 uniformly orange-red with orange-red setae; northern Mexico, USA Texas to California .....*Triscolia ardens*

b. Entire body orange; Baja California peninsula .....*Triscolia badia*

4(2) **a.** Metasomal T4-6 uniformly orange-red with orange-red setae, color of integument sometimes darkened but always contrasting with that of metasomal T1-2, mesosoma, and head, which are entirely black or dark brown without pale markings; IF yellow or white markings are present, they are limited to metasomal T3 5

b. Metasomal T4-6 NOT uniformly orange-red; markings variable or entirely absent 6

5(4) **a.** Metasomal T3 with two lateral yellow-white spots; USA Atlantic coast west to Arizona and Colorado, south to Texas, north to Illinois .....*Scolia dubia dubia*

b. Metasomal T3 without spots; Texas, New Mexico, and Arizona .....

.....*Scolia dubia haematodes*

6(4) **a.** Lateral margins of metasomal T6 and S6 conspicuously constricted when viewed dorsally or ventrally (Fig. 3.5A); head and mesosoma black; metasoma black, often with yellow-white lateral spots on any or all metasomal T1-5; most of Mexico, USA Texas and New Mexico .....*Scolia guttata*

b. Lateral margins of T6 and S6 not constricted (Fig. 3.5B), tapering gradually; markings variable 7

7(6) **a.** Pronotum with yellow, orange, or red markings; northern Mexico, continental USA

.....*Scolia nobilitata*

b. Pronotum uniformly black, without pale markings 8

8(7) **a.** Basal tubercle of metasomal S2 with single medial point (Fig. 3.1A.a); base of metasomal S1 with transverse furrow deeply undercut and strongly emarginate medially (Fig. 3.1A.b); usually entirely black, but some specimens with yellow lateral spots on metasomal terga and sometimes sterna, spots on metasomal T2 and T3 sometimes fused, forming medially emarginate bands; Mexico, USA Arizona, New Mexico, and Texas, north to Maryland[1] .....*Scolia mexicana*

b. Basal tubercle of metasomal S2 extended laterally, with two weak lateral points (Fig. 3.1B.a); base of metasomal S1 with transverse furrow NOT clearly undercut (Fig. 3.1B.b), medial emargination variable; metasomal T2 and T3 with yellow-white bands, bands without pronounced medial emarginations and covering almost the entire length of their respective tergum; USA east of 100th meridian .....*Scolia bicincta*

9(1) **a.** Forewing with three submarginal cells (vein 1r-m present); frons with smooth transverse furrow (Fig. 3.6A.a, B.a) 10

b. Forewing with two submarginal cells (vein 1r-m absent); frons without transverse furrow (Fig. 3.6B) 13

10(9) **a.** Area between antennal sockets and transverse furrow of frons forming a distinct elevated platform (Fig. 3.6B.b) 11

b. Area between antennal sockets and transverse furrow of frons not forming an elevated platform (Fig. 3.6A) 12

1 See "notes on problematic taxa" section above.

11(10) a. Vertex, mesosoma, and dorsal surface of metasoma black, with yellow or reddish markings; west of the Rocky Mountains (upper Sonoran life zone) .....*Colpa alcione*

b. Vertex, mesosoma, and metasoma rusty red, with yellow markings; Sonora, southern Arizona and California .....*Colpa flammicoma*

12(10) a. Scutellum with conspicuous longitudinal furrow (Fig. 3.7A); Sonora and Arizona north to Kansas .....*Colpa pollenifera*

b. Scutellum without longitudinal furrow; if trace of furrow present, it is evanescent posteriorly (Fig. 3.7B); Atlantic and gulf coast of USA, northern Mexico and southern USA Texas to southern California, Great Plains region north to North Dakota .....*Colpa octomaculata*

13(9) a. Metapleuron with conspicuous shelf-like area ventrad of metapleural flange, shelf with sharp, sometimes carinate, ventral margin (Fig. 3.8A) 14

b. Metapleuron without shelf-like area ventrad of metapleural flange (Fig. 3.8B) 19

14(13) a. Medial area of metanotum mostly smooth, with few irregularly distributed setae-bearing punctures; dorsal surface of propodeum with conspicuous medial triangular impunctate area, its apex extending to posterior apex of dorsal area of propodeum (Fig. 3.9A) 15

b. Medial area of metanotum uniformly covered with setae-bearing punctures; dorsal surface of propodeum without medial triangular impunctate area (Fig. 3.9B) OR, if impunctate area present, its apex not reaching posterior apex of dorsal area of propodeum 17

15(14) a. Metasomal T1 and T4 black, without markings; areas of metasomal T2-3 apicad of subapical transverse row of setae with orange color not or only slightly extending beyond callosities (Fig. 3.10A); wings evenly infuscate; tropical and subtropical South and Central America to approximately 20°N, Lesser Antilles, eastern Greater Antilles, southern Florida .....*Dielis dorsata*

b. At least metasomal T1 or T4 with yellow or orange markings; apical areas of metasomal T2-3 with yellow or orange color extending beyond callosities, callosities forming dark lateral notches in tergal bands (Fig. 3.10B); wings darker apically 16

16(15) a. Colored band on metasomal T3 with black medial notch in dorsal view about 1/3 as wide and 1/2 as deep as band; markings yellow; Greater Antilles, southern Florida .....*Dielis trifasciata*

b. Colored band on metasomal T3 with black medial notch much narrower and shallower than 1/3 and 1/2 the width and depth of the band respectively, colored band thus occupying almost entire dorsal surface of tergum; markings yellow-orange; Mexico, Hispaniola, USA central and southern California and southern Arizona .....*Dielis tolteca*

17(14) a. Medial posterior vertical surface of propodeum rugose (Fig. 3.11A, C); Great Plains .....*Dielis plumipes confluenta*

b. Medial posterior vertical surface of propodeum NOT rugose (Fig. 3.11B, D), mostly smooth but sometimes punctate close to dorsal margin 18

18(17) a. Posterior margin of dorsomedial surface of propodeum forming a broadly rounded lamelliform shelf extending laterally to the transverse propodeal lines (Fig. 3.12A); metasomal T3 with apical yellow band with broad shallow medial notch (depth of notch much less than 1/2 the depth of yellow band) and narrower lateral notches mesad of callosities; USA eastern Texas to North Carolina .....*Dielis plumipes fossulana*

b. Posterior margin of dorsomedial surface of propodeum forming medially tapering wedge-like projection that does not extend laterally to transverse propodeal lines (Fig. 3.12B); metasomal T3 with deeper medial notch about 1/2 the depth of yellow band and no lateral notches visible in dorsal view, although band gradually narrows laterally and callosities may form small lateral notches visible in lateral view; Massachusetts south to northern Georgia, west to eastern Kentucky .....*Dielis plumipes plumipes*

19(13) a. Yellow or orange markings limited to metasomal T2-3, integument black; apical setae of metasomal T4-5 black 20

b. Yellow or orange markings present at least on metasomal T1-3, integument variable; apical setae of metasomal T4-5 yellow or orange 21

20(19) a. Metasomal T2-3 almost entirely orange; propodeum with lateral carina strongly curved, sometimes almost angular (Fig. 3.13); base of metasomal S1 with transverse furrow strongly emarginate medially; northwestern South America, Central America, north to southern Arizona and southern Texas .....*Pygodasis ephippium*

b. Metasomal T2-3 each with pair of dorsolateral yellow spots; propodeum with lateral carina gently curved or almost straight (Fig. 3.8B.c); base of metasomal S1 with transverse furrow not emarginate medially; northern Texas to Massachusetts .....*Pygodasis quadrimaculata*

21(19) a. Forewing cell 1R1 setose only along stigma; anterior transverse furrow of metasomal S1 medially absent or indistinct; USA west of 100th meridian .....*Dielis pilipes*

b. At least anterior 1/2 of forewing cell 1R1 setose; anterior transverse furrow of metasomal S1 complete 22

22(21) a. Posterior margin of dorsomedial surface of propodeum with curved carina (Fig. 3.12C); Greater Antilles .....*Xanthocampsomeris tricincta*

b. Posterior margin of dorsomedial surface of propodeum without curved carina (Fig. 3.12D) 23

23(22) a. Forewing cells R, 1M, and 1Rs setose; Central America, Arizona, New Mexico, Texas .....*Xanthocampsomeris completa*

b. Forewing cells R, 1M, and 1Rs asetose 24

24(23) a. Longer hind tibial spurs acute to bluntly rounded, NOT expanded apically (Fig. 3.13C); metanotum with yellow spot; northern South America, Central America, Lesser Antilles, southern Texas .....*Xanthocampsomeris hesterae*

b. Longer hind tibial spurs spatulate (rounded AND expanded apically) (Fig. 3.13A-B); metanotum without yellow spot 25

25(24) a. Forewing cells 1R1 and 2R1 with at most the anterior two thirds densely setose; posterior surface of propodeum coarsely punctate; Mexico, Arizona .....*Xanthocampsomeris limosa*

b. Forewing cells 1R1 and 2R1 completely setose; posterior surface of propodeum impunctate; Greater Antilles, Florida .....*Xanthocampsomeris fulvohirta*

Key to males

1 **a.** Forewing with second recurrent vein (2m-cu) absent 2

b. Forewing with second recurrent vein present 9

2(1) **a.** Forewing with 3 submarginal cells (vein 1r-m present); California to Texas, northern Mexico 3

b. Forewing with 2 submarginal cells (vein 1r-m absent) 4

3(2) **a.** Metasomal T3-6 uniformly orange-red with orange-red setae; northern Mexico, USA Texas to California .....*Triscolia ardens*

b. Entire body orange; Baja California peninsula .....*Triscolia badia*

4(2) **a.** Integument of pronotum and metanotum always with yellow markings; yellow markings sometimes also present on head, scutellum, propodeum, and some or all terga of the metasoma; northern Mexico, continental USA .....*Scolia nobilitata*

b. Integument of mesosoma black, without pale markings 5

5(4) **a.** Metasomal T2-6 with orange-red posterior setal fringes 6

b. Metasomal T2-6 with black fringes, no orange-red setae or markings anywhere on the body 7

6(5) **a.** Metasomal T3 with 2 lateral yellow-white spots; USA Atlantic coast west to Arizona and Colorado, south to Texas, north to Illinois .....*Scolia dubia dubia*

b. Metasomal T3 without spots; Texas, New Mexico, and Arizona .....*Scolia dubia haematodes*

7(5) **a.** Basal tubercle of metasomal S2 with single medial point (Fig. 3.1A.a); base of metasomal S1 with transverse furrow deeply undercut and strongly emarginate medially (Fig. 3.1A.b); Mexico, USA Arizona, New Mexico, and Texas, north to Maryland[2] .....*Scolia mexicana*

b. Basal tubercle of metasomal S2 extended laterally, without medial point (Fig. 3.1B.a); base of metasomal S1 with transverse furrow NOT clearly undercut (Fig. 3.1B.b), medial emargination variable 8

8(7) **a.** Transverse tubercle of metasomal S2 with two distinct lateral points and central depression (Fig. 3.15); yellow markings on metasomal terga, if present, are non-contiguous lateral spots; wings with cyan or blue iridescence; most of Mexico, USA Texas and New Mexico .....*Scolia guttata*

b. Transverse tubercle of metasomal S2 at most with weak lateral points, without pronounced central depression (Fig. 3.1B.a); metasomal T2-3 with yellow-white bands, bands without pronounced medial emarginations and covering almost the entire length of their respective tergum; wings with bronze iridescence; USA east of 100th meridian .....*Scolia bicincta*

9(1) **a.** Forewing with 3 submarginal cells (vein 1r-m present); volsella with articulation between basal and apical parts; frons with smooth transverse furrow (Fig. 3.16A.a, 3.16B.a) 10

b. Forewing with 2 submarginal cells (vein 1r-m absent); basal and apical parts of volsella fused; frons without transverse furrow (Fig. 3.16C-D) 13

2 See "notes on problematic taxa" section above.

10(9) **a.** Antennae not or only slightly clavate (Fig. 3.17C); base of metasomal S1 without transverse furrow; mesopleural setae erect 11

b. Antennae clavate (Fig. 3.17D); base of metasomal S1 with transverse furrow; mesopleural setae appressed and shiny 12

11(10) **a.** Scutellum with conspicuous longitudinal furrow (Fig. 3.17A); Sonora and Arizona north to Kansas .....*Colpa pollenifera*

b. Scutellum without conspicuous longitudinal furrow (Fig. 3.17B); Atlantic and gulf coast of USA, northern Mexico and southern USA Texas to southern California, Great Plains region north to North Dakotas .....*Colpa octomaculata*

12(10) **a.** Face with discrete, oval, punctate, setose yellow area mesad of each antennal base (Fig. 3.16A.b); scape orange or yellow; metasomal integument red, shading into black posteriorly; Sonora, southern Arizona and California .....*Colpa flammicoma*

b. Face with oval, punctate, setose areas mesad of each antennal base fused and black (Fig. 3.16B.b); scape black, sometimes with yellow markings; metasomal integument black; west of the Rocky Mountains (upper Sonoran life zone) .....*Colpa alcione*

13(9) **a.** Hind tibial spurs black 14

b. Hind tibial spurs white or light-colored 15

14(13) a. Bands on metasomal terga yellow, deeply notched medially, sometimes separated into lateral spots; yellow markings always present on metasomal T2-3, sometimes present on T1-5; northern Texas to Massachusetts .....*Pygodasis quadrimaculata*

b. Bands on metasomal terga orange to orange-red, NOT notched medially, covering almost the entire surfaces of metasomal T2-3; metasomal T1 sometimes with small spot, no markings on T4-5; northwestern South America, Central America, north to southern Arizona and southern Texas .....*Pygodasis ephippium*

15(13) a. Forewing almost entirely setose .....*Xanthocampsomeris*

b. Apical area of forewing (apicad of veins) mostly asetose 16

16(15) a. Face with frontal line deeply impressed, frons with elevated impunctate area on either side of frontal line (Fig. 3.16C.c); yellow spots on lateral corners of pronotum in dorsal view; USA west of 100th meridian ..... *Dielis pilipes*

b. Face with frontal line shallow, frons without elevated impunctate area on either side of frontal line (Fig. 3.16D.c), or if an elevated area is present, it is punctate; pronotum in dorsal view without yellow markings 17

17(16) a. Clypeus entirely yellow except for a small central dark spot; Greater Antilles, southern Florida .....*Dielis trifasciata*

b. Clypeus NOT entirely yellow, with extensive black markings 18

18(17) a. Pronotum entirely black or with thin medial posterior yellow band; New Mexico, Colorado, Wyoming east to the Atlantic coast .....*Dielis plumipes*

b. Pronotum with broad yellow band extending laterally 19

19(18) a. Parameres with dense basal brush of setae; Mexico, Hispaniola, USA central and southern California and southern Arizona .....*Dielis tolteca*

b. Parameres without dense basal brush of setae; tropical and subtropical South and Central America to approximately 20°N, Lesser Antilles, eastern Greater Antilles, southern Florida .....*Dielis dorsata*

Conflicts of interest

The authors declare that there are no conflicts of interest.

Acknowledgments

We would like to thank M. Hauser (California Department of Food and Agriculture), S.L. Heydon (Bohart Museum of Entomology, University of California, Davis), and K. Williams (University of Florida; California Department of Food and Agriculture) for providing access to scoliid specimens at their respective institutions; L. Vilhelmsen (Natural History Museum of Denmark) and S. Ryder (Natural History Museum, London) for coordinating imaging of Fabrician specimens at their respective institutions; B.E. Boudinot, T. Zavortink, and F. Keller (University of California, Davis) for testing the key; T. Zavortink for providing access to specimens in his personal collection; N. Tam for providing a line drawing and for proofreading and commenting on the manuscript.

References

Bartlett, O. C. (1912). The North American digger wasps of the subfamily Scoliinae. *Annals of the Entomological Society of America, 5*(4), 293-340.

Bradley, J. C. (1928a). A revision of the New World species of Trielis a subgenus of Campsomeris (Hymenoptera: Scoliidae). *Transactions of the American Entomological Society (1890-), 54*(3), 195-214.

Bradley, J. C. (1928b). The species of Campsomeris (Hymenoptera-Scoliidae) of the plumipes group, inhabiting the United States, the Greater Antilles, and the Bahama Islands. *Proceedings of the Academy of Natural Sciences of Philadelphia*, *80*, 313-337.

Bradley, J. C. (1957). The taxa of Campsomeris (Hymenoptera: Scoliidae) occurring in the New World. *Transactions of the American Entomological Society (1890-)*, *83*(2), 65-77.

Bradley, J. C. (1964a). The Fabrician types of Scoliidae (Hymenoptera), with notes and an appendix by J. G. Betrem. *Spolia Zoologica Musei Hauniensis*, *21*, 1-38.

Bradley, J. C. (1964b). Further notes on the American taxa of Campsomeris (Hymenoptera: Scoliidae). *Entomological News*, *25*, 101-108.

Branstetter, M. G., Danforth, B. N., Pitts, J. P., Faircloth, B. C., Ward, P. S., Buffington, M. L., ... & Brady, S. G. (2017). Phylogenomic insights into the evolution of stinging wasps and the origins of ants and bees. *Current Biology*, *27*(7), 1019-1025.

Clausen, C. P. (1940). *Entomophagous insects*. McGraw-Hill book Company, Incorporated.

Day, M. C., Else, G. R., & Morgan, D. (1981). The most primitive scoliidae (Hymenoptera). *Journal of natural History*, *15*(4), 671-684.

DeBach, P. (1964). *Biological control of insect pests and weeds*. Reinhold Publishing Corporation, New York.

Elliott, M. G. (2011). Annotated catalogue of the Australian Scoliidae (Hymenoptera). *Technical Reports of the Australian Museum, Online*, *22*, 1-17.

Fabricius, J. C. (1775). *Systema entomologiae, sistens insectorum classes, ordines, genera, species, adjectis synonymis, locis, descriptionibus, observationibus*.

Grissell, E. E. (2007). Scoliid Wasps of Florida, Campsomeris, Scolia and Trielis spp. (Insecta: Hymenoptera: Scoliidae). *Institute of Food and Agricultural Sciences Extension Electronic Data Information Source*, 1-9. *See https://edis.ifas.ufl.edu/pdf/IN/IN74500.pdf*

Illingworth, J. F. (1921). Natural enemies of sugar-cane beetles in Queensland. *Queensland Bureau Sugar Experimental Station Division of Entomology Bulletin*, *13*, 1-47.

Johnson, B. R., Borowiec, M. L., Chiu, J. C., Lee, E. K., Atallah, J., & Ward, P. S. (2013). Phylogenomics resolves evolutionary relationships among ants, bees, and wasps. *Current Biology*, *23*(20), 2058-2062.

Liu, Z., Van Achterberg, C., He, J. H., & Chen, X. X. (2021). A checklist of Scoliidae (Insecta: Hymenoptera) from China. *Zootaxa*, *4966*(2), 101-126.

MacKay, W. P. (1987). The scoliid wasps of the southwestern United States (Hymenoptera: Scoliidae). *The Southwestern Naturalist*, 357-362.

Michener, C. D. (1944). Comparative external morphology, phylogeny, and a classification of the bees (Hymenoptera). Bulletin of the AMNH; v. 82, article 6.

Osten, T. (2005). Checkliste der Dolchwespen der Welt (Insecta: Hymenoptera, Scoliidae). *Bericht der Naturforschenden Gesellschaft Augsburg*, *62*, 1-62.

Peters, R. S., Niehuis, O., Gunkel, S., Bläser, M., Mayer, C., Podsiadlowski, L., ... & Krogmann, L. (2018). Transcriptome sequence-based phylogeny of chalcidoid wasps (Hymenoptera: Chalcidoidea) reveals a history of rapid radiations, convergence, and evolutionary success. *Molecular Phylogenetics and Evolution*, *120*, 286-296.

Porter, C. C. (1981). Scoliidae (Hymenoptera) of the lower Río Grande valley. *Florida Entomologist*, 441-453.

Rohwer, S. A. (1927). Some scoliid wasps from tropical America. *Journal of the Washington Academy of Sciences*, *17*(6), 150-155.

Turner, R. E. (1909). Remarks on some genera of the Scoliidae, with descriptions of new species. *Annals and Magazine of Natural History*, *3*(18), 476-486.

Wilson, F. (1960). A review of the biological control of insects and weeds in Australia and Australian New Guinea. *Tech. Commun. Commonw. Inst. Biol. Control, Ottawa*, (1).

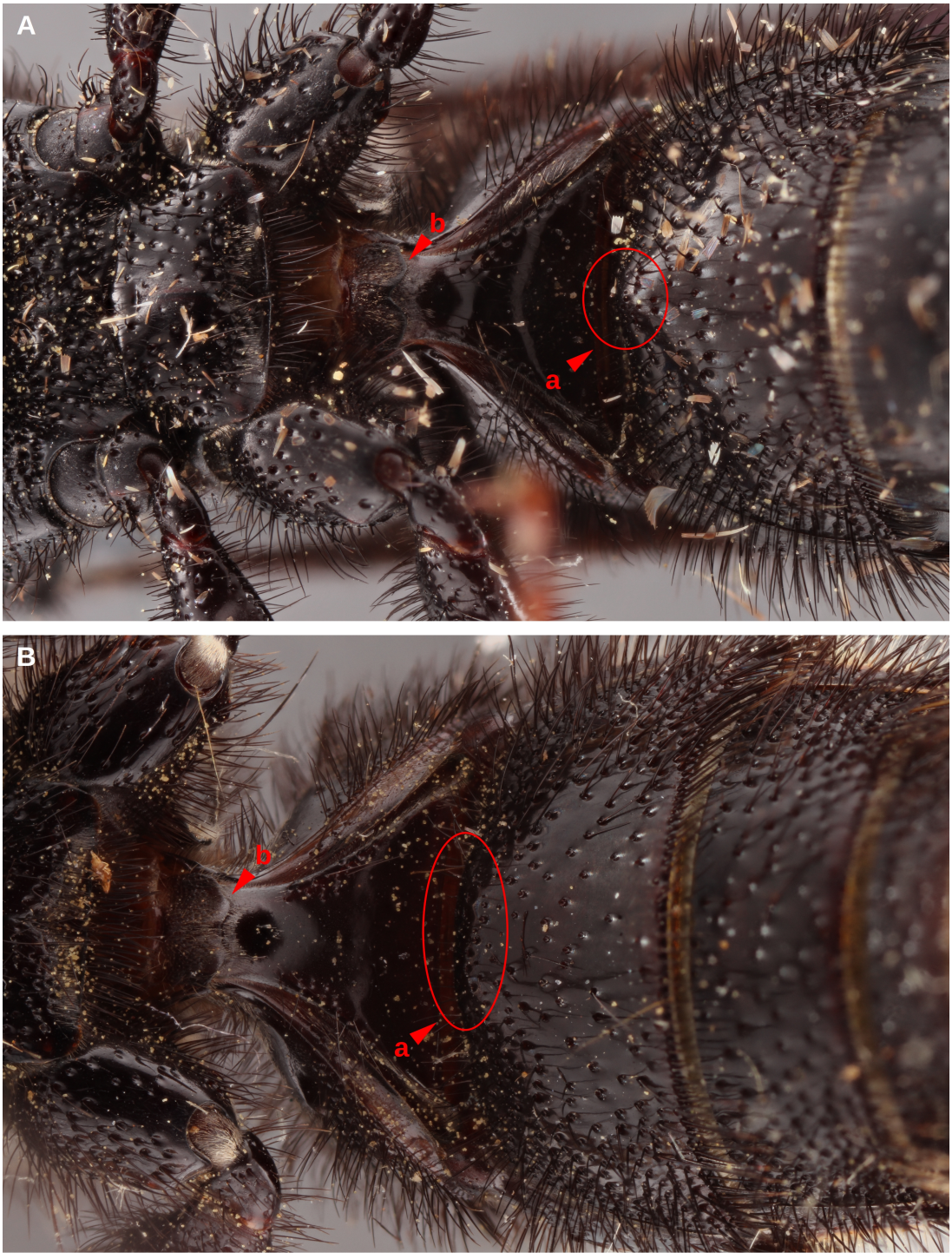

Figure 3.1. Anterior metasoma, ventral view. (A) *Scolia mexicana* ♀; (B) *Scolia bicincta* ♀. (a) Basal tubercle of metasomal S2; (b) basal transverse furrow of metasomal S1.

Figure 3.2. *Scolia bicincta* specimen 70 ♂ from the Fabricius collection, Natural History Museum of Denmark, University of Copenhagen. Anterior metasoma, ventral view. (a) Basal tubercle of metasomal S2; (b) basal transverse furrow of metasomal S1. (photo credit: Mikkel Høegh Post, the Natural History Museum of Denmark, used with permission)

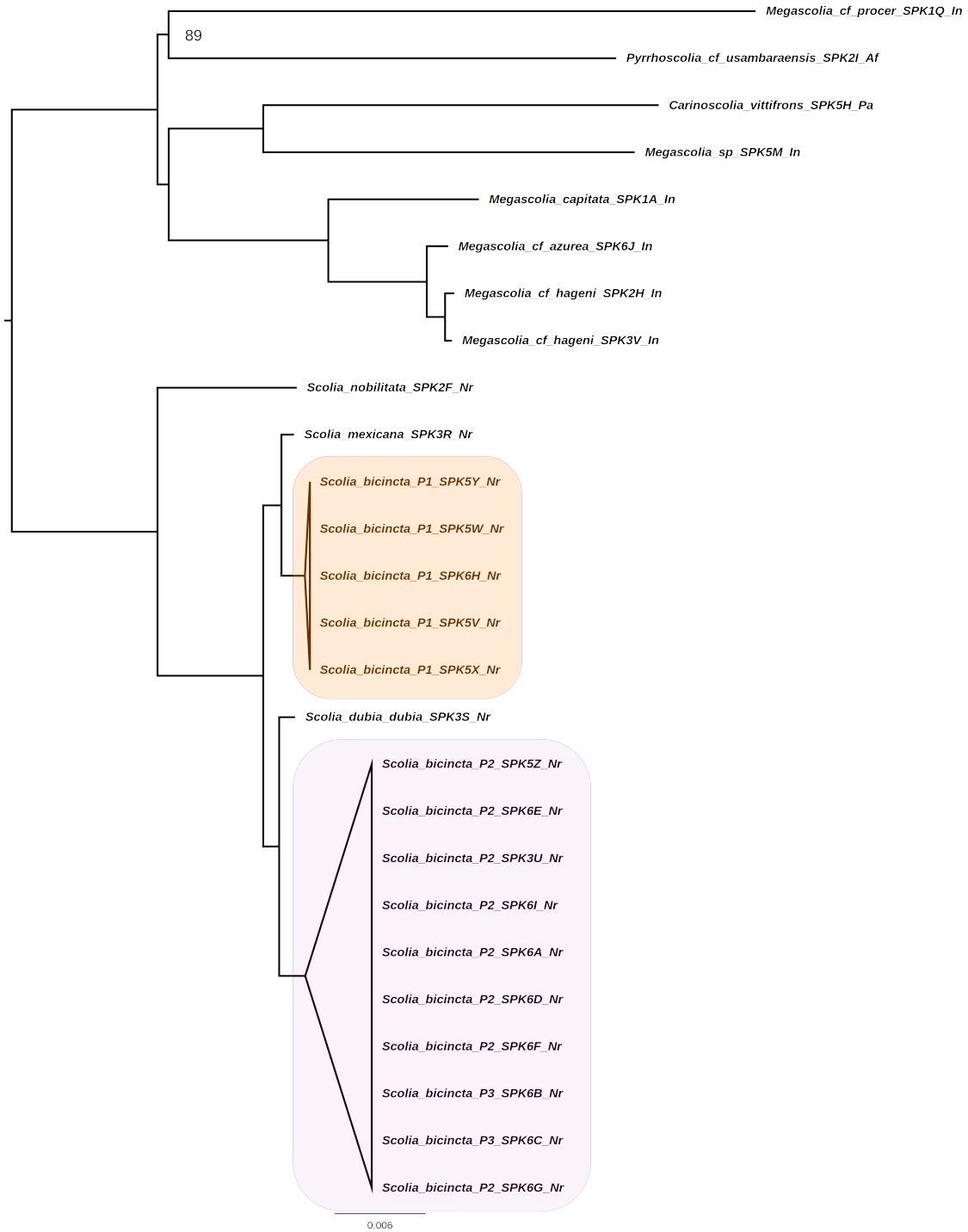

Figure 3.3. Maximum likelihood phylogeny using IQTREE of Nearctic *Scolia* based on 211 UCE loci with full taxon coverage. Specimens traditionally identified as *Scolia bicincta* fall into two clades: “typical” *S. bicincta* sister to *Scolia dubia* (purple) and a clade sister to *Scolia mexicana* (orange). *Megascolia-Carinoscolia-Pyrrhoscolia* clade used as outgroup. Branch lengths are in expected number of nucleotide substitutions per site. Support values based on 1000 Ultrafast Bootstrap replicates. Unlabeled nodes have maximal support. See Supporting Information for methods details.

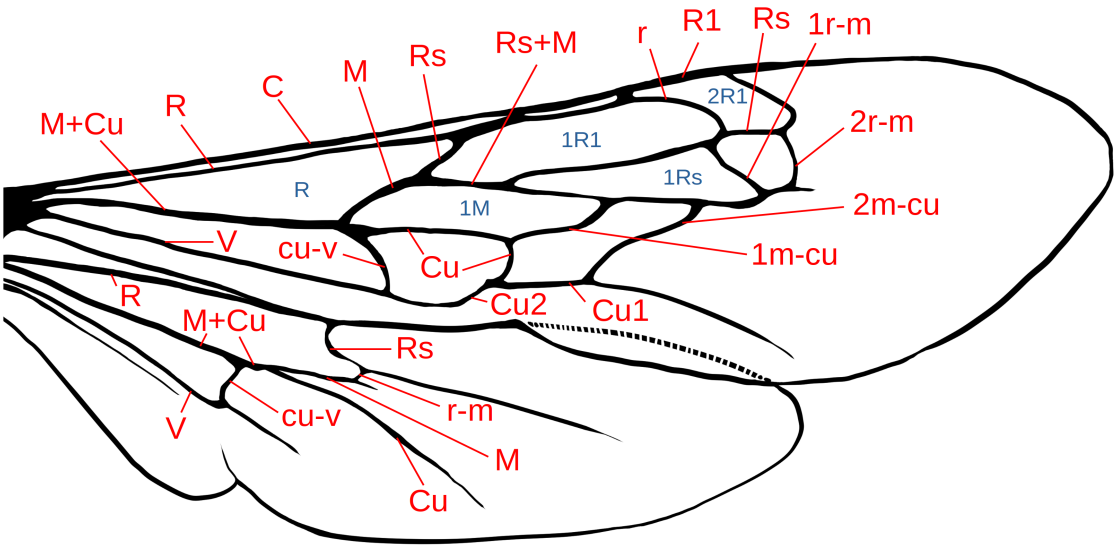

Figure 3.4. Wings of *Colpa*, representing “complete” scoliid wing venation. Veins 1r-m and/or 2m-cu are absent in some taxa. Vein names in red, cell names in blue. (drawing credit: Nicole Tam, used with permission)

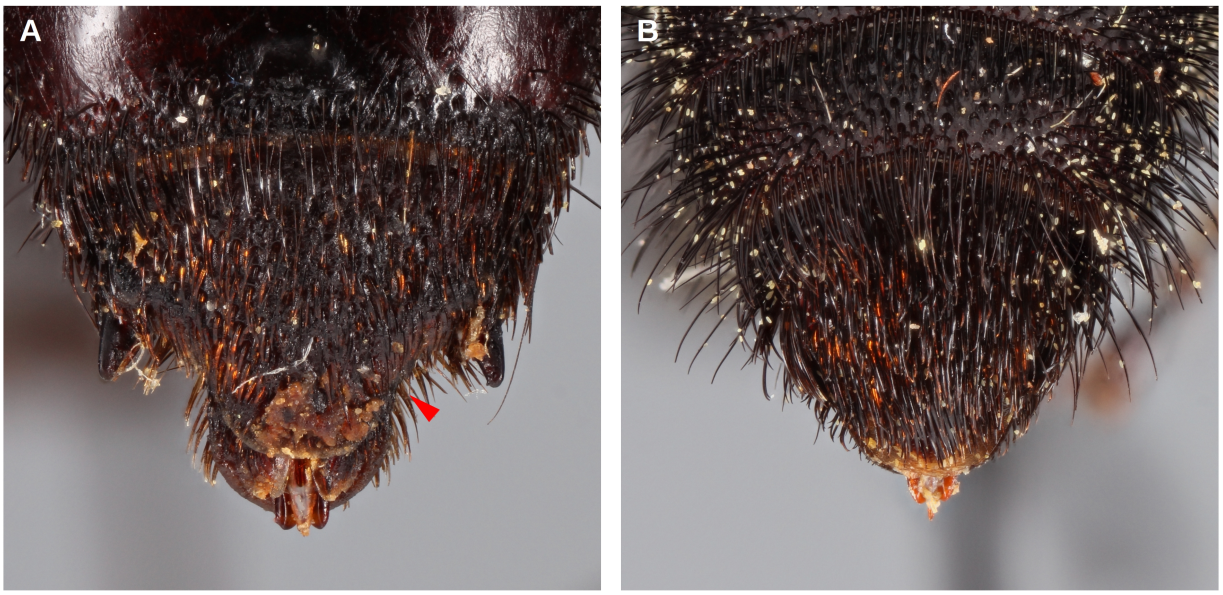

Figure 3.5. Apex of metasoma. (A) *Scolia guttata* ♀; (B) *Scolia mexicana* ♀.

Figure 3.6. Head. (A) *Colpa octomaculata* ♀; (B) *Colpa flammicoma* ♀; *Dielis dorsata* ♀.

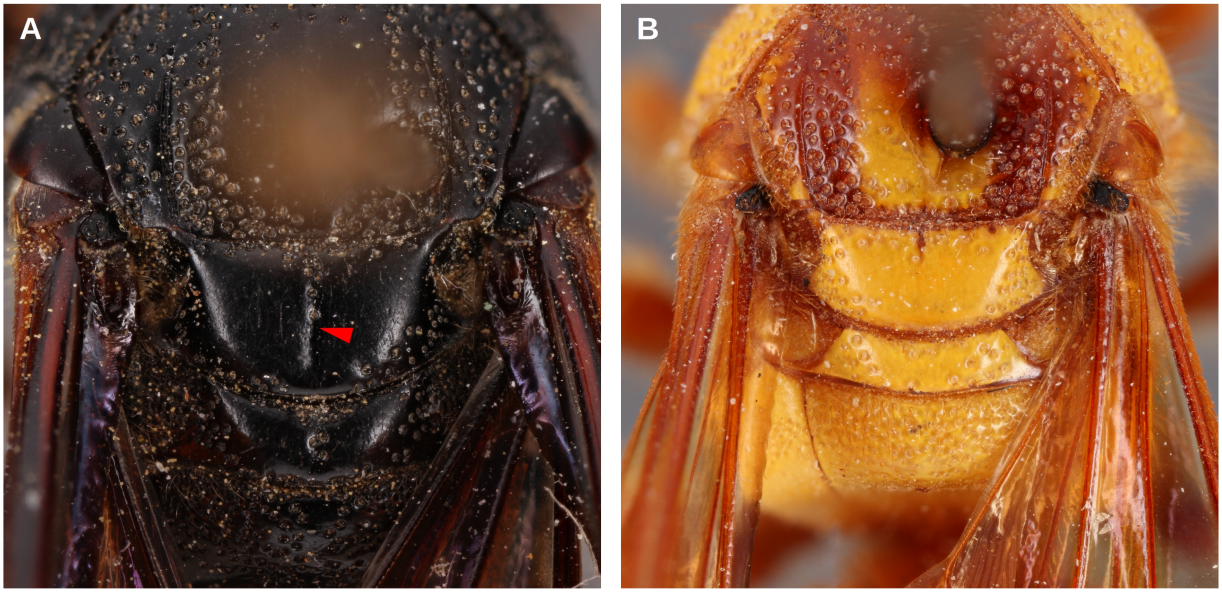

Figure 3.7. Scutellum and metanotum. (A) *Colpa pollenifera* ♀; (B) *Colpa octomaculata* ♀.

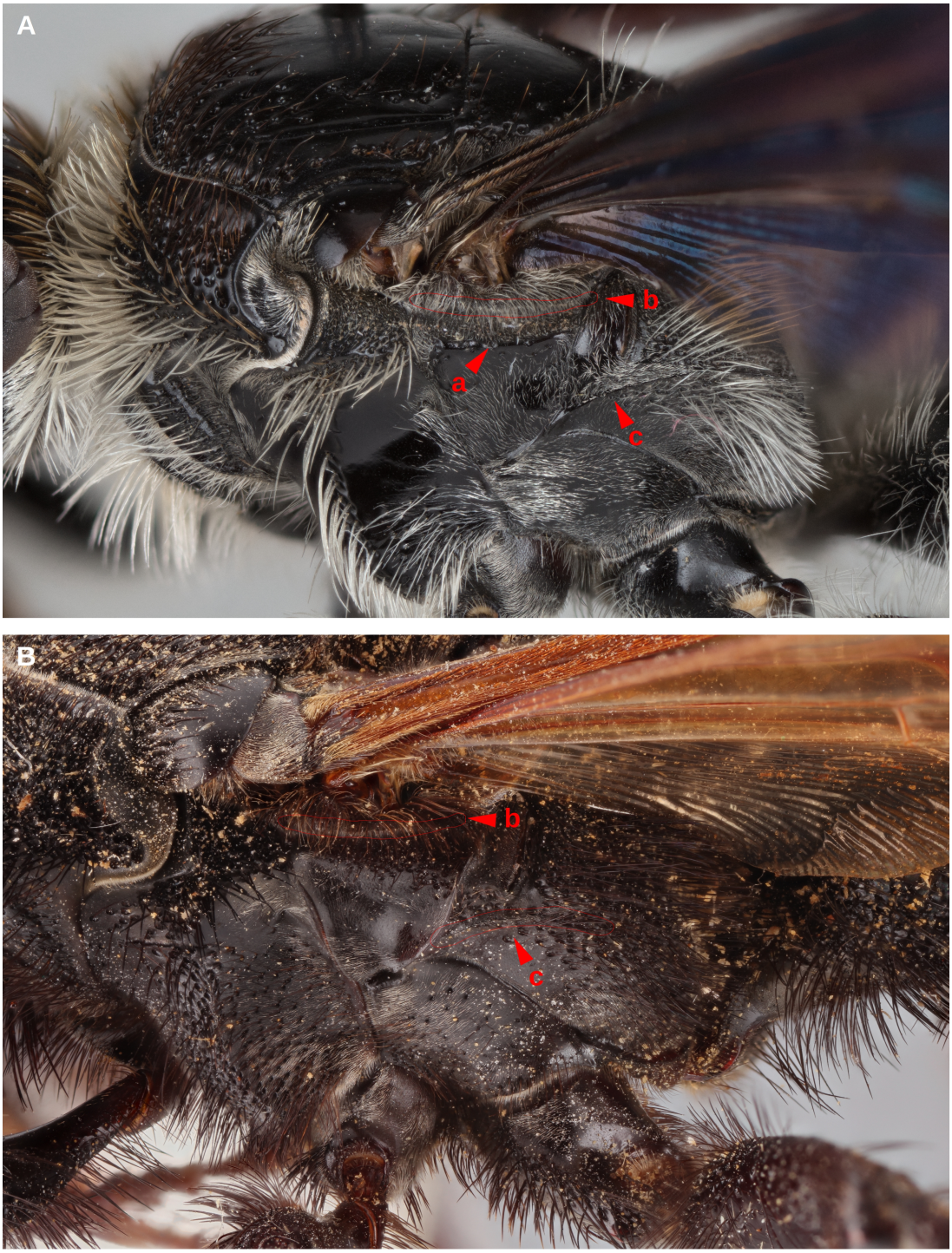

Figure 3.8. Mesosoma, lateral view. (A) *Dielis dorsata* ♀; (B) *Pygodasis quadrimaculata* ♀. (a) Superior longitudinal carina of metapleuron; (b) metapleural flange; (c) lateral carina of propodeum.

Figure 3.9. Metanotum and propodeum. (A) *Dielis trifasciata trifasciata* ♀; (B) *Dielis plumipes fossulana* ♀.

Figure 3.10. Metasoma, lateral view. (A) *Dielis dorsata* ♀; (B) *Dielis trifasciata* ♀.

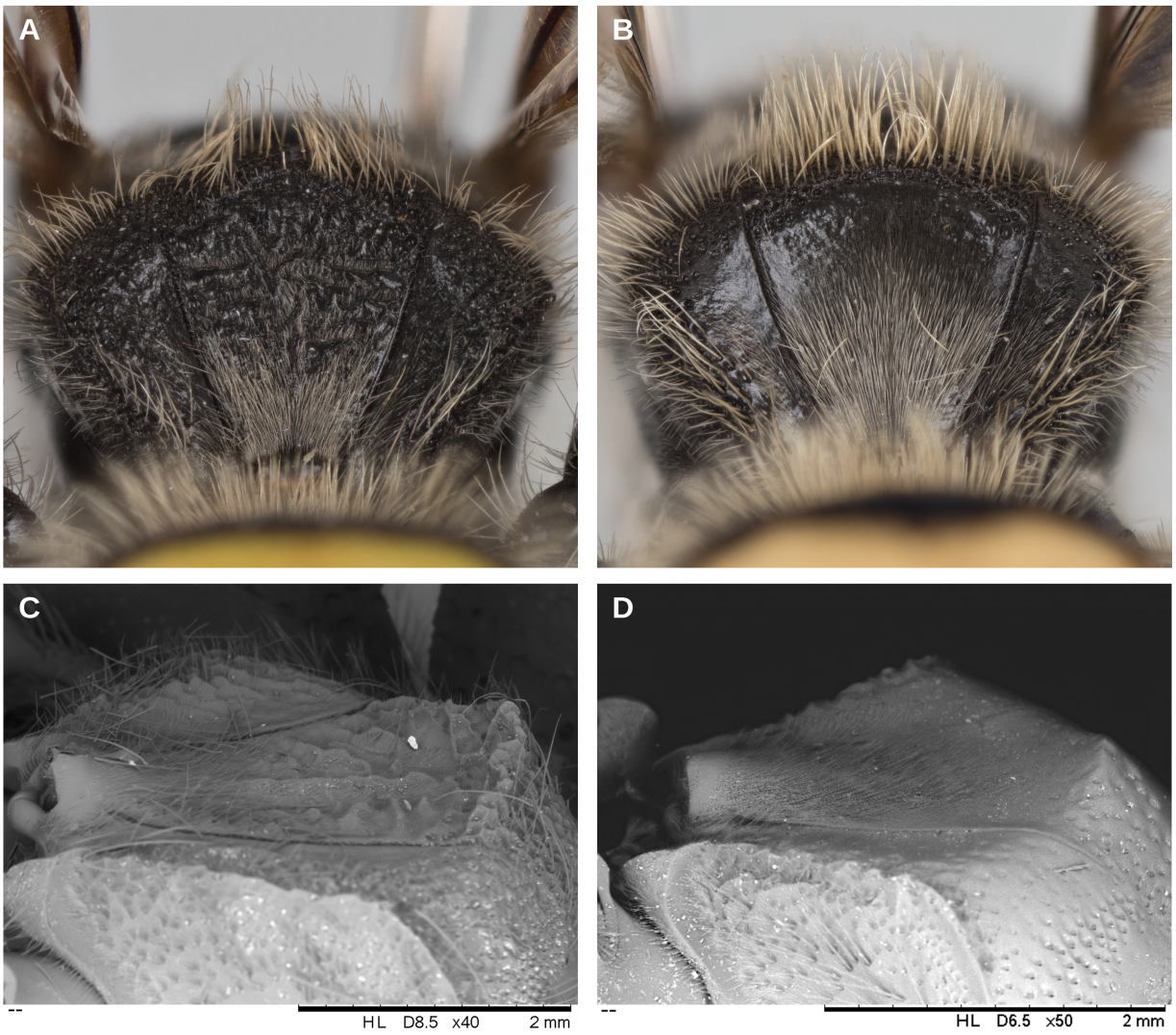

Figure 3.11. Posterior surface of propodeum. (A) *Dielis plumipes confluenta* ♀; (B) *Dielis plumipes fossulana* ♀; (C) *Dielis plumipes confluenta* ♀; (D) *Dielis dorsata* ♀.

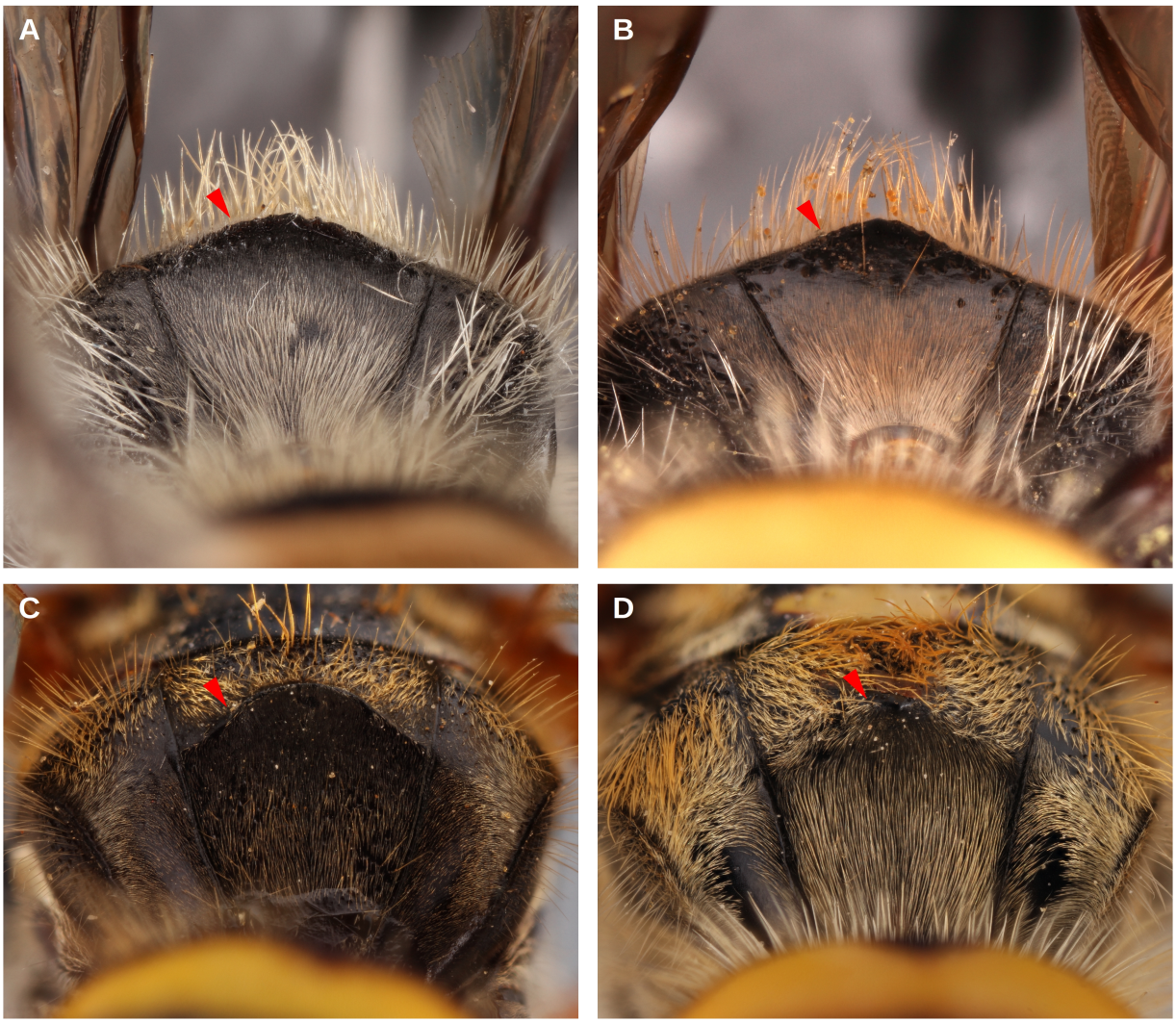

Figure 3.12. Posterior surface of propodeum. (A) *Dielis plumipes fossulana* ♀; (B) *Dielis plumipes plumipes* ♀; *Xanthocampsomeris tricincta* ♀; *Xanthocampsomeris hesterae* ♀.

Figure 3.13. *Pygodasis ephippium* ♀ metapleuron and propodeum, lateral view.

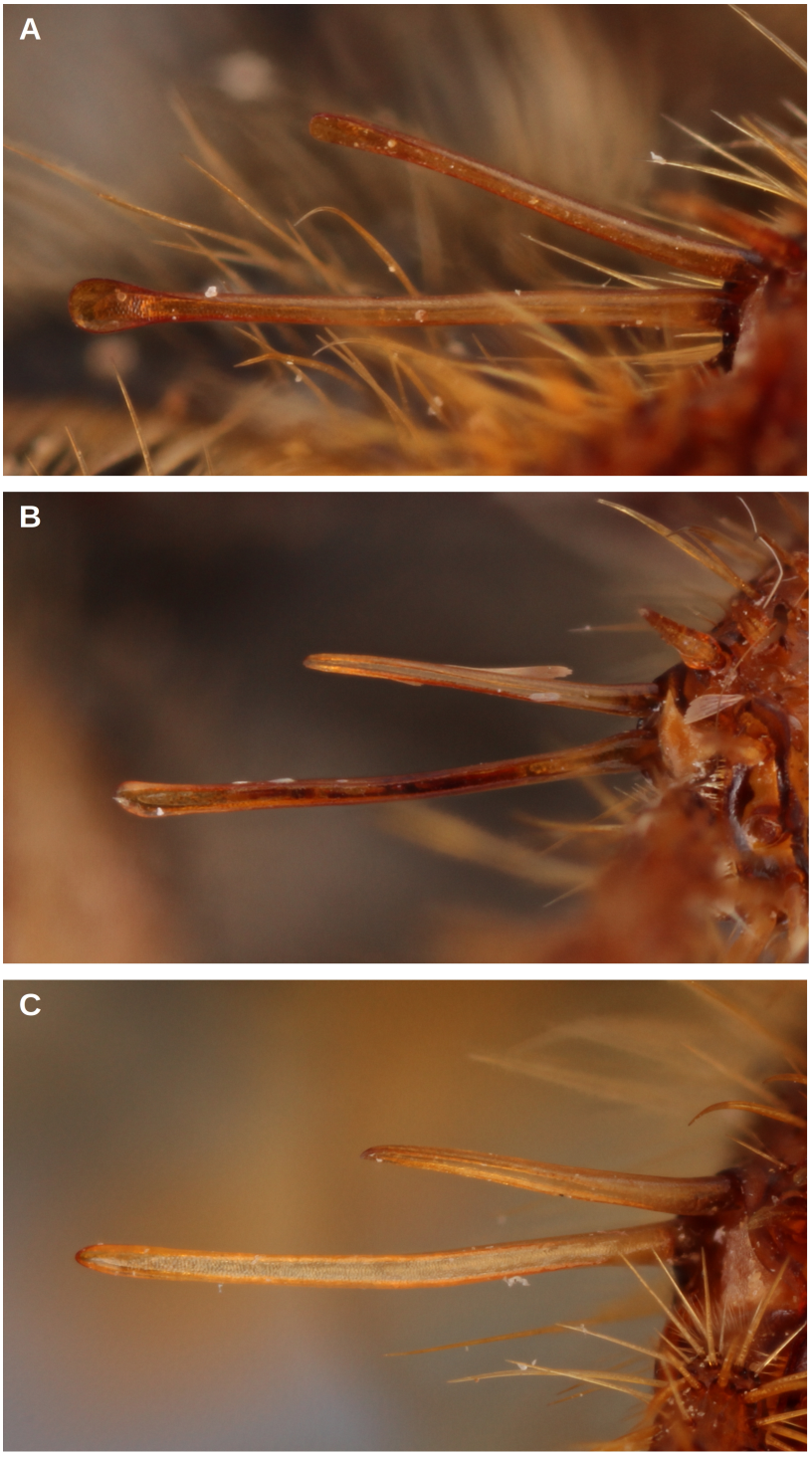

Figure 3.14. Hind tibial spurs. (A) *Xanthocampsomeris limosa* ♀; (B) *Xanthocampsomeris fulvohirta* ♀; (C) *Xanthocampsomeris hesterae* ♀.

Figure 3.15. *Scolia guttata* ♂. Anterior metasoma, ventral view.

Figure 3.16. Head. (A) *Colpa flammicoma* ♂; (B) *Colpa alcione* ♂; (C) *Dielis pilipes* ♂; (D) *Dielis plumipes* ♂. (a) Transverse furrow of frons; (b) interantennal area; (c) frontal line.

Figure 3.17. (A) Scutellum, *Colpa pollenifera* ♂; (B) scutellum, *Colpa octomaculata* ♂; (C) antenna, *Colpa octomaculata* ♂; (D) antenna, *Colpa alcione* ♂.

Supporting Information

Taxon and locus selection

We used 2404 ultraconserved element (UCE) loci (Faircloth *et al*., 2012; 2015) and 26 specimens from the dataset used in Chapter 1. Our ingroup included one specimen each of *Scolia dubia dubia*, *Scolia mexicana*, and *Scolia nobilitata*, as well as 15 specimens matching the description of *Scolia bicincta*. We used the *Megascolia* + *Carinoscolia* + *Pyrrhoscolia* clade (8 specimens) as an outgroup.

Alignment and trimming

We used MAFFT (Katoh & Standley, 2013) version 7.407 with the E-INS-i algorithm (Altschul, 1998) for multiple sequence alignment. We then performed edge-trimming using the phyluce_align_get_trimmed_alignments_from_untrimmed script from the phyluce package (Faircloth, 2016) version 1.6.8. Following trimming, we summarized alignment statistics using AMAS (Borowiec, 2016) and removed any alignments that had missing taxa or 15% or more missing data at the site level, retaining 223 loci. Preliminary phylogenetic analysis resulted in unexpectedly long terminal branches subtending some taxa, possibly due to sequencing and/or alignment error. We therefore used spruceup (Borowiec, 2019) version 2020.2.19 to mask parts of sequences that are potentially spurious.

Phylogenetic analysis

We estimated a maximum likelihood phylogeny using IQTREE (Minh *et al.*, 2020; Chernomor *et al.*, 2016) version 2.0.6. We first used matched-pairs tests of symmetry (Jermiin *et al*., 2017; Naser-Khdour *et al*., 2019) to remove alignments that likely violate SRH assumptions. We then partitioned by locus and chose the best-fitting model for each partition based on BIC (Schwarz,

1978) from among substitution models from the GTR (Tavaré, 1986) family and discretized gamma (Yang, 1994) and free-rates ASRV models. We used the edge-linked proportional partition model for branch lengths. We also performed 1000 ultrafast bootstrap replicates (Hoang *et al.* 2018), and used the --bnni option.

Supporting information references

Altschul, S. F. (1998). Generalized affine gap costs for protein sequence alignment. *Proteins: Structure, Function, and Bioinformatics*, *32*(1), 88-96.

Borowiec, M. L. (2016). AMAS: a fast tool for alignment manipulation and computing of summary statistics. *PeerJ*, *4*, e1660.

Borowiec, M. L. (2019). Spruceup: fast and flexible identification, visualization, and removal of outliers from large multiple sequence alignments. *Journal of Open Source Software*, *4*(42), 1635.

Chernomor, O., Von Haeseler, A., & Minh, B. Q. (2016). Terrace aware data structure for phylogenomic inference from supermatrices. *Systematic biology*, *65*(6), 997-1008.

Faircloth, B. C. (2016). PHYLUCE is a software package for the analysis of conserved genomic loci. *Bioinformatics*, *32*(5), 786-788.

Faircloth, B. C., Branstetter, M. G., White, N. D., & Brady, S. G. (2015). Target enrichment of ultraconserved elements from arthropods provides a genomic perspective on relationships among Hymenoptera. *Molecular ecology resources*, *15*(3), 489-501.

Faircloth, B. C., McCormack, J. E., Crawford, N. G., Harvey, M. G., Brumfield, R. T., & Glenn, T. C. (2012). Ultraconserved elements anchor thousands of genetic markers spanning multiple evolutionary timescales. *Systematic biology*, *61*(5), 717-726.

Hoang, D. T., Chernomor, O., Von Haeseler, A., Minh, B. Q., & Vinh, L. S. (2018). UFBoot2: improving the ultrafast bootstrap approximation. *Molecular biology and evolution*, *35*(2), 518-522.

Jermiin, L. S., Jayaswal, V., Ababneh, F. M., & Robinson, J. (2017). Identifying optimal models of evolution. In *Bioinformatics* (pp. 379-420). Humana Press, New York, NY.

Katoh, K., & Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular biology and evolution*, *30*(4), 772-780.

Minh, B. Q., Schmidt, H. A., Chernomor, O., Schrempf, D., Woodhams, M. D., Von Haeseler, A., & Lanfear, R. (2020). IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. *Molecular biology and evolution*, *37*(5), 1530-1534.

Naser-Khdour, S., Minh, B. Q., Zhang, W., Stone, E. A., & Lanfear, R. (2019). The prevalence and impact of model violations in phylogenetic analysis. *Genome biology and evolution*, *11*(12), 3341-3352.

Schwarz, G. (1978). Estimating the dimension of a model. *The annals of statistics*, *6*(2), 461-464.

Tavaré, S. (1986). Some probabilistic and statistical problems in the analysis of DNA sequences. *Lectures on mathematics in the life sciences*, *17*(2), 57-86.

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