*Syst. Biol.* 71(6):1504–1523, 2022

https://doi.org/10.1093/sysbio/syac043
Advance Access publication June 16, 2022

# Detecting and Removing Sample Contamination in Phylogenomic Data: An Example and its Implications for Cicadidae Phylogeny (Insecta: Hemiptera)

CHRISTOPHER L. OWEN[1,*], DAVID C. MARSHALL[2], ELIZABETH J. WADE[3], RUSS MEISTER[2], GEERT GOEMANS[2], KRUSHNAMEGH KUNTE[4], MAX MOULDS[5], KATHY HILL[2], M. VILLET[6], THAI-HONG PHAM[7,8], MICHELLE KORTYNA[9], EMILY MORIARTY LEMMON[10], ALAN R. LEMMON[11], AND CHRIS SIMON[2]

[1]*Systematic Entomology Laboratory, USDA-ARS, c/o National Museum of Natural History, Smithsonian Institution, Washington, DC, USA;* [2]*Department of Ecology and Evolutionary Biology, University of Connecticut, Storrs, CT 06269, USA;* [3]*Department of Natural Science and Mathematics, Curry College, Milton, MA 02186, USA;* [4]*National Centre for Biological Sciences, Tata Institute of Fundamental Research, GKVK Campus, Bellary Road, Bangalore 560 065, India;* [5]*Department of Entomology, Australian Museum Research Institute, 1 William Street, Sydney NSW 2010, Australia;* [6]*Department of Biology, Rhodes University, Grahamstown 6140, South Africa;* [7]*Mientrung Institute for Scientific Research, Vietnam National Museum of Nature, Vietnam Academy of Science and Technology, Hue, Vietnam;* [8]*Graduate School of Science and Technology, Vietnam Academy of Science and Technology, Hanoi, Vietnam;* [9]*Department of Biological Science, Florida State University, 319 Stadium Drive, Tallahassee, FL, USA;* [10]*Department of Biological Science, Florida State University, 319 Stadium Drive, Tallahassee, FL 32306, USA, and* [11]*Department of Scientific Computing, Florida State University 400 Dirac Science Library, Tallahassee, FL 32306, USA*

**Correspondence to be sent to: Systematic Entomology Laboratory, USDA-ARS, c/o National Museum of Natural History, Smithsonian Institution, Washington, DC, USA;*
*E-mail: christopher. owen@usda. gov.*

*Received 8 January 2021; reviews returned 23 May 2022; accepted 7 June 2022*
*Associate Editor: Jacob Esselstyn*

*Abstract*.—Contamination of a genetic sample with DNA from one or more nontarget species is a continuing concern of molecular phylogenetic studies, both Sanger sequencing studies and next-generation sequencing studies. We developed an automated pipeline for identifying and excluding likely cross-contaminated loci based on the detection of bimodal distributions of patristic distances across gene trees. When contamination occurs between samples within a data set, a comparison between a contaminated sample and its contaminant taxon will yield bimodal distributions with one peak close to zero patristic distance. This new method does not rely on *a priori* knowledge of taxon relatedness nor does it determine the causes(s) of the contamination. Exclusion of putatively contaminated loci from a data set generated for the insect family Cicadidae showed that these sequences were affecting some topological patterns and branch supports, although the effects were sometimes subtle, with some contamination-influenced relationships exhibiting strong bootstrap support. Long tip branches and outlier values for one anchored phylogenomic pipeline statistic (*AvgNHomologs*) were correlated with the presence of contamination. While the anchored hybrid enrichment markers used here, which target hemipteroid taxa, proved effective in resolving deep and shallow level Cicadidae relationships in aggregate, individual markers contained inadequate phylogenetic signal, in part probably due to short length. The cleaned data set, consisting of 429 loci, from 90 genera representing 44 of 56 current Cicadidae tribes, supported three of the four sampled Cicadidae subfamilies in concatenated-matrix maximum likelihood (ML) and multispecies coalescent-based species tree analyses, with the fourth subfamily weakly supported in the ML trees. No well-supported patterns from previous family-level Sanger sequencing studies of Cicadidae phylogeny were contradicted. One taxon (*Aragualna plenalinea*) did not fall with its current subfamily in the genetic tree, and this genus and its tribe Aragualnini is reclassified to Tibicininae following morphological re-examination. Only subtle differences were observed in trees after the removal of loci for which divergent base frequencies were detected. Greater success may be achieved by increased taxon sampling and developing a probe set targeting a more recent common ancestor and longer loci. Searches for contamination are an essential step in phylogenomic analyses of all kinds and our pipeline is an effective solution. [Auchenorrhyncha; base-composition bias; Cicadidae; Cicadoidea; Hemiptera; phylogenetic conflict.]

Contamination of a genetic sample with DNA from one or more nontarget species is a continuing concern of molecular phylogenetic studies, both Sanger sequencing and next-generation sequencing (NGS) studies. Contamination causes problems for Sanger studies when DNA from a nontargeted taxon enters the pipeline and is amplified via the polymerase chain reaction (PCR), sequenced, and misidentified as belonging to the target taxon (e.g., Thomas et al. 1989; Derr et al. 1992; Austin et al. 1997; Zhang and Hewitt 2003). Sanger studies lessen, but do not eliminate, this risk through positive and negative PCR controls and thorough examination of basic statistics such as nucleotide bias/amino acid substitutions and examination of chromatograms for heterogeneous signal. PCR amplification of nontarget loci from the template genome (i.e., paralogs) is a related but separate issue.

NGS shotgun sequencing library methods do not have contamination controls analogous to the positive and negative PCR reactions of Sanger studies, and NGS data sets typically consist of many large alignments that are processed by automated pipelines, making it easier for contaminated sequences to go unrecognized. Many researchers send DNA/RNA to sequencing centers where evidence of contamination has been documented (Salter et al. 2014; Ballenghien et al. 2017). In addition, the PCR amplification step in some library prep kits can increase contaminant sequences to detectable levels. Consequently, examples of NGS contamination cases have begun to accumulate. Longo et al. (2011) found that >20% of NCBI Trace Archives, UCSC, and Ensembl nonhuman primate databases contained human DNA. Francois et al. (2020) reported that 35% of the arthropod genomes in the Ensembl genome database contained

varying levels of nontargeted sequences. Merchant et al. (2014) showed that *Neisseria gonorrhoeae* TCDC-NG08107 included partial sequences of cow and sheep. Other studies have mistaken contamination for horizontal gene transfer in rotifers and tardigrades (Wilson et al. 2018; Bemm et al. 2016). Contamination may be underreported because labs are reluctant to admit the problem or fail to check for it.

### *Removing Contamination in Phylotranscriptomic Studies*

Contamination has been identified in both phylogenomic studies and shotgun genome sequencing studies. Thus far, most inquiries into phylogenomic contamination have been in phylotranscriptomic studies. For example, Philippe et al. (2011) identified contamination in transcriptome studies in the form of chimeric sequences (paralogs were also discussed). More recently, contamination has been identified in phylotranscriptomic studies focusing on characiform fish (Betancur-R et al. 2019) and cnidarians (Kayal et al. 2018). Sequences can be designated to the incorrect taxon when multiplexing unique libraries on a single sequencing run on NGS platforms. The unique barcode associated with each sample can be assigned to an incorrect sample due to read misalignment (van der Valk et al. 2020), a phenomenon called "index hopping." As a result, a small proportion of sequencing reads barcoded for a given taxon may have misassigned sequences from other samples in the study. Another source of contamination is chimeras which can be formed by PCR-induced recombination during the library preparation step and during the downstream assembly when unrelated short reads from the same taxon are assembled (Yang and Smith 2013).

Some insect phylotranscriptomic studies attempt to exclude contamination by identifying similar sequences among taxa and then removing the less abundant (lower coverage) transcript. For example, Peters et al. (2017) identified transcripts that shared 98% sequence identity for at least 180 bp and then retained one transcript or removed both based on relative abundance. The software CroCo was developed for phylotranscriptomic studies and performs a similar identification of contamination (Simion et al. 2018). These approaches assume that a contaminant (e.g., paralog from the same taxon or ortholog from a nontargeted taxon) will occur at a low copy number relative to the noncontaminant transcript. "Branch length correlation analysis" has been used in other studies to find data artifacts (Simion et al. 2017; Arcila et al. 2021). This test aims to identify contamination that produces a long branch in gene trees. The branch length correlation analysis estimates a species tree using a concatenated data matrix and compares each gene tree while accounting for missing taxa and compares the terminal branch lengths in each tree to produce a ratio. If the ratio is greater than 5, the sequence is inspected by hand and potentially removed.

### *Removing Contamination in Reduced Representation Nextgen Sequencing Studies*

In anchored hybrid enrichment (AHE), ultraconserved element (UCE), and restriction-site associated DNA sequencing (RAD-seq) phylogenomic projects, most potential contamination is removed using sequence similarity. For example, Breinholt et al. (2018) used a 99% sequence similarity across 95% of the sequences as a cutoff and sequences fitting these criteria were removed. More recently, Prous et al. (2020) used sequence similarity to identify ~20 double digest RAD-seq loci in which cross-species contamination had occurred. Heterozygosity is also used to identify potential contamination. Lemmon et al. (in preparation) calculated the heterozygosity for all sites of assembled homologs/orthologs and removed those sites that contained greater than two alleles. At the Center for Anchored Phylogenomics (Florida State University), a three-pronged strategy is followed depending on the phylogenetic level involved (Box 1).

#### BOX 1. THREE FORMS OF CONTAMINATION MITIGATION/DETECTION OPERATING ON DIFFERENT LEVELS.

Because of the long length of the capture probes, contaminants can be excluded in a sequence capture experiment, so long as the contaminant is <82% similar in sequence in the probe region (Bossert and Danforth 2018). More of the contaminant is expected to be retained when the contaminant is more similar to the probe sequence. Detecting the contaminant can be very simple when the contaminant is well outside the group of study and genetically distant but can also be very difficult or impossible to detect when genetically similar (i.e., within the same genus as the study samples). Different strategies of removing contaminants can be used depending on the phylogenetic relatedness of the contaminants.

1) Assembly cluster coverage filter: best for removing very distant contaminants that do not enrich well but still to some degree. The sequences from these divergent contaminants get put into a different assembly cluster easily. They stand out because they have such low coverage relative to the rest of the assembly clusters.
2) Number of homologs: best for moderately distant contaminants. The number of homologs (assembly clusters passing the filter in 1) can be a good indicator of contamination, especially for shallow-scale projects. Contaminants picked up here are similar enough to the targets to be at roughly the same

coverage level but still different enough to be in separate assembly clusters. When an individual has substantially more assembly homolog clusters than other individuals, it is an indication of contamination or hybridization (i.e., a hybrid) but could also be a whole genome duplication (quite unlikely on shallow scales). Note that hybrids formed by very divergent parents can also fall into this category, but this is much more likely in plants. We demonstrate in this manuscript that using patristic distances and the method developed within can successfully identify and remove this type of contamination.

3) Heterozygosity: best for shallow contaminants and cross-contaminants. When the contaminant is quite similar to the study taxa (i.e., within the study group), the contaminating reads do not get put into their own assembly cluster; instead, they get mixed into an assembly cluster with target reads. This shows up as more heterozygous sites, and if allele phasing is performed, this can suggest polyploidy. This is the most difficult type to deal with.

Although sequence similarity can identify contamination in phylogenomic studies, it is not applicable and scalable to all phylogenomic studies based on the different types of contamination or the age of the taxa in the experiment (Box 1). Using sequence similarity will probably capture the contamination produced by index hopping, but it will miss contamination from taxa not included in the experiment. As discussed above, contamination can occur in nearly any stage of a phylogenomics project, so it is not inconceivable to think that experiments may include sequence data from taxa not included in the experiment. Sequence similarity as an indicator of phylogenomic contamination does not scale well for taxa that diverged recently and share a young most recent common ancestor. Most phylogenomic studies recycle ortholog sets, which are typically made from highly divergent taxa that share a most recent common ancestor hundreds of millions of years ago (e.g., Simao et al. 2015). Young taxa and old protein-coding orthologs typically result in few polymorphisms, which may lead to the removal of most, if not all, of the data due to high sequence similarity. Given the many types of contamination in phylogenomic experiments and the pitfalls of using sequence similarity, it is imperative to develop phylogenomic contamination identification methods that overcome these potential pitfalls.

The methodological study described here was motivated by a problem we encountered with phylogenetic inference that was caused by probable cross-contamination of different samples in the taxon set. This contamination was not detected by the upstream approaches in the AHE pipeline (Box 1), which were based in part on sequence similarity. In response, we developed a supplementary tool that applies the sequence similarity criterion in combination with additional tests based on expected bimodal phylogenetic signal across gene trees—allowing the removal of such contamination without reference to taxonomic assumptions.

Here, we present this novel method to remove contamination from bait-captured phylogenomic data and examine the effect of its removal on phylogenomic relationships in the Cicadidae (Insecta: Hemiptera). Specifically, we explore the performance of the new contamination-identification pipeline and how contamination affects topology, branch lengths, and nodal supports in both concatenated and multispecies coalescent phylogenomic analyses. We also examine the effects of compositional heterogeneity on phylogenetic results. We assess the usefulness of this AHE data set for confirming the monophyly of each cicada subfamily and the relationships among them, especially the rapid radiations suggested by earlier studies with limited Sanger sequencing (Marshall et al. 2018; Simon et al. 2019) and mtDNA genome data (Łukasik et al. 2019) data. The sample of 102 cicada species includes highly divergent taxa across cicada subfamilies and tribes as well as closely related congeners inside each tribe and is the first phylogenomic study of Cicadidae.

### *The Study Organisms*

Cicadas are plant-sucking hemipteran insects (Fig. 1) from suborder Auchenorrhyncha (containing cicadas and the large plant-sucking "hoppers"). Cicadas are known for life cycles up to 17 years and mating signals that are among the loudest sounds produced by terrestrial animals (Myers 1929; Claridge 1985; Williams and Simon 1995). Cicadas are agricultural pests for crops such as fruit orchards (Logan and Alspatch 2007), grapes (Mehdipour et al. 2016), and sugarcane (Ito and Nagamine 1981). Interest in cicadas as model organisms has expanded because of their obligate associations with the bacterial endosymbionts *Sulcia* and *Hodgkinia* (McCutcheon et al. 2009; Van Leuven et al. 2014; Campbell et al. 2018, Łukasik et al. 2018), their fungal parasites and symbionts (Cooley et al. 2018; Matsuura et al. 2018; Boyce et al. 2019; Lovett et al. 2020), new phylogenomic studies of periodical cicadas (Fujisawa et al. 2018; Du et al. 2019), and applications of cicada wing nanostructure to create water and bacterial resistant materials (Zhang et al. 2006; Xie et al. 2008; Hasan et al. 2013; Zada et al. 2016; Lin et al. 2018, Linklater et al. 2020).

Deep-level phylogenetic relationships within Cicadidae have been examined in two Sanger sequencing studies (Marshall et al. 2018; Simon et al. 2019). These revealed discordance between the genetic relationships and higher classification, and two new subfamilies were

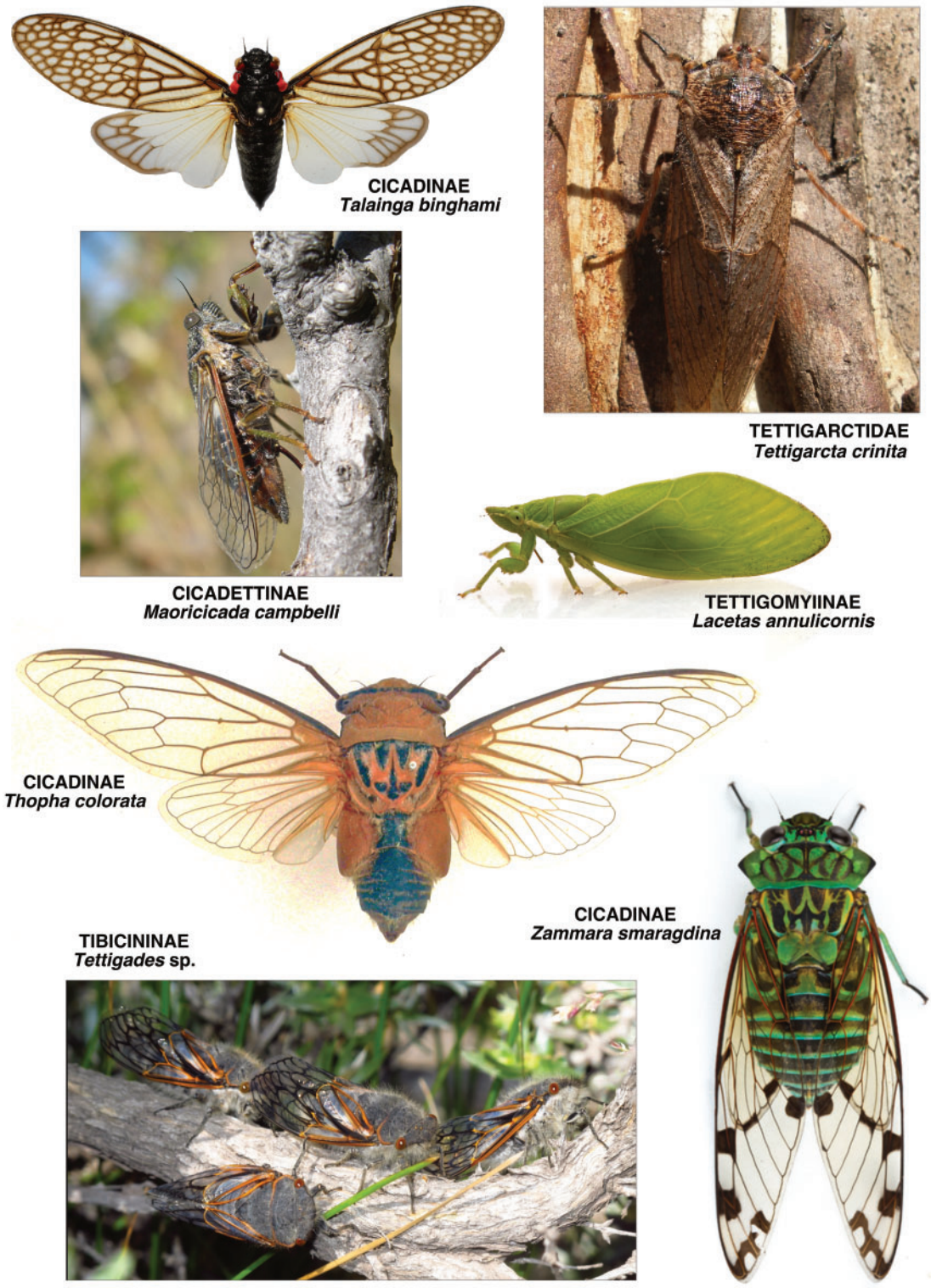


FIGURE 1. Cicadidae from four subfamilies and the sister family Tettigarctidae. Photo credits: *Maoricicada*, K. Hill; *Tettigades*, P. Łukasik; *Talainga* and *Thopha*, M. Moulds; *Zammara* and *Lacetas*, P. Naskrecki; *Tettigarcta*, C. Simon. Images not to scale.

proposed. Although many nodes were well supported, the relationships among the subfamilies Cicadinae, Cicadettinae, and Tettigomyiinae remained unresolved, along with those of major lineages within the Cicadinae, which includes the generally larger-bodied cicadas. Here, we generate phylogenomic trees in the hope of resolving relationships among taxa involved in these comparatively rapid radiations.

## Materials and Methods

### *Taxon Sampling*

We sequenced 100 ingroup and two outgroup cicada species (Table S1 of the Supplementary material available on Dryad at http://dx.doi.org/10.5061/dryad.tht76hdz1), with the aim of determining tribe and subfamily relationships. The ingroup taxon sample includes four of the five Cicadidae subfamilies, 44 of 59 tribes (Marshall et al. 2018; Simon et al. 2019; Sanborn 2018; 2021a, 2021b; Sanborn et al. 2020; Hill et al. 2021; Moulds et al. 2021) and 91 described ingroup genera. Twenty-four sampled genera have not been previously represented in family-level genetic studies. Cicadidae includes approximately 460 extant genera and 3118 species (Catalog of Life, accessed 13 May 2019). The two outgroup species from Tettigarctidae together with Cicadidae form superfamily Cicadoidea. Cicadas were identified by the authors and other expert collaborators using original literature (authorships in Table S1 of the Supplementary material available on Dryad) and reference cicada collections.

### *AHE Sequencing*

Sequence reads were collected and assembled at the Center for Anchored Phylogenomics (www.anchoredphylogeny.com/), and the scripts and code used to perform the AHE bioinformatics pipeline are publicly available in the supplemental material of the study by Granados Mendoza et al. (2020). Illumina libraries were prepared following Meyer and Kircher (2010), with modifications noted in Prum et al. (2015). In brief, extracted DNA was fragmented to ∼200–400 nt by sonication on a Covaris ultrasonicator. Common adapters containing 8 bp indexes were then added to the ends of the fragments. After quantification via Qubit, libraries were pooled in equal concentration in groups of ∼16. Each library pool was then enriched using one of two probe sets described in Dietrich et al. (2017) and Simon et al. (2019); these probe sets target broadly overlapping loci in Paraneoptera and Auchenorrhyncha, respectively. Enrichment pools were combined and sequenced on Illumina HiSeq lanes with a PE150 protocol with single-8 bp indexing. The total sequencing effort was 1.1 Gb per sample.

Reads passing the Casava high-chastity filter were demultiplexed (with no mismatches tolerated), prior to read merging, in which overlapping read pairs were merged to correct for sequencing errors and to remove adapter sequence (Rokyta et al. 2012). The reads were assembled using the quasi-*de novo* assembly approach described by Hamilton et al. (2016). In brief, reads were mapped to divergent references (in this case Cercopidae: *Philaenus spumarius*, Cicadellidae: *Ponana quadralaba*, Cicadidae: *Pauropsalta* sp., and Cicadellidae: *Graphocephala fennahi*), with reference sequences being obtained from the probe regions (Dietrich et al. 2017; Simon et al. 2019). Assembly clusters with low coverage (less than 55 reads) were not utilized downstream. This filter helps remove low-level contaminants. For the remainder of the clusters, consensus sequences were constructed with ambiguities being called if base composition at variable sites were more likely caused by sequencing error instead of heterozygosity.

Orthology was assessed for homologs at each locus using a neighbor-joining approach that relied on alignment-free distance matrices (see Hamilton et al. 2016 for details). Orthologous sets representing at least 50% of the taxa were processed downstream. Alignments of orthologous sequences estimated in MAFFT v7. 023b (Katoh and Standley 2013) were trimmed/masked using an automated process, in which misaligned regions are identified and masked, then sites with substantial amounts of missing data (in this case >50%) are removed. Processed alignments were manually inspected in Geneious R9 (Biomatters Ltd.; Kearse et al. 2012).

### *Phylogenetic Analysis*

We used overlapping AHE capture probe sets from Paraneoptera (lice, thrips, and hemipteran bugs) and Auchenorrhyncha (plant hoppers, leaf hoppers, cicadas and relatives). We used NCBI-BLAST v2.9.0+ blastn v2.9.0 (Camacho et al. 2009) to combine the data sets and determine orthologs. Although tree-based orthology is preferred over sequence similarity (Smith and Pease 2017), we used sequence similarity to save time. We removed all gaps from the alignments, combined them, and processed them as one throughout. We aligned each ortholog using MAFFT v7 (Katoh and Standley 2013) followed by manual curation in Mesquite v3.6 (Maddison and Maddison 2019). For evaluating molecular models of evolution, we identified the protein-coding and noncoding (i.e., intron and untranslated) regions of each ortholog, using the coding sequences of a publicly available cicada transcriptome (*Okanagana villosa*: GAWQ00000000.2) and blastx analysis with default settings, followed by hand curation in Mesquite. These sequence coordinates were used as subsets for PartitionFinder v2 (Guindon et al. 2010; Lanfear et al. 2017) to determine the best-fit models of evolution and partition scheme according to the Bayesian Information Criterion (BIC).

We used Garli v2.01 (Zwickl 2006) to estimate a maximum likelihood (ML) gene tree for each locus. Each ML gene tree search included 10 independent searches from random starting trees. We employed 100 nonparametric bootstrap replicates (Felsenstein 1985) to evaluate the branch support (BS). Each bootstrap replicate included five independent searches from random starting trees.

We next used ASTRAL-III (Sayyari and Mirarab 2016; Zhang et al. 2018) to estimate the Cicadidae species tree because it uses the multispecies coalescent to account for incomplete lineage sorting (Degnan and Rosenberg 2009; Edwards 2009). We used the best estimate ML tree for each locus as input with BS < 20 collapsed. We collapsed branches with low bootstrap support because it has

been shown that removing branches with low bootstrap support in gene trees can improve species tree estimation (Zhang et al. 2017). Node support was assessed for the ASTRAL-III phylogeny using local posterior probability (LPP; Sayyari and Mirarab 2016).

We explored subsampling loci based on relative rates of evolution because it has been shown that slowly and rapidly evolving loci may cloud phylogenetic signal for deep-time relationships (e.g., Townsend 2007; Regier et al. 2010; Cummins and McInerney 2011). We approximated the relative rate for each locus using the tree length (Oakley et al. 2013) and estimated ML and ASTRAL-III species phylogenies without 5% (21) of the slowest loci, without 5% (21) of the fastest loci, and without both of these subsets.

We also explored removing loci based on compositional heterogeneity, because taxa with similar compositional patterns can be inferred as more closely related than their true evolutionary relationship (e.g., Foster 2004; Jermiin et al. 2004; Meade and Pagel 2008; Crotty et al. 2020). We used p4 (Foster 2004) to estimate a null distribution of nucleotide composition for each locus using the ML phylogeny and 1000 replicates with simulated sequence data under the GTR+I+$G_4$ model (no partitions) and then applied a threshold of $P < 0.05$ to remove loci in exploratory ASTRAL-III and concatenated phylogenetic analyses (see below).

We also estimated the Cicadidae phylogeny by concatenating all loci for ML analysis. We estimated the best-fit BIC evolutionary models and partitioning scheme using ModelFinder (Kalyaanamoorthy et al. 2017) in IQ-TREE v2. 0-rc-1 (Minh et al. 2020) with loci as subsets using the command *-m TESTMERGEONLY*. This command uses the "greedy" algorithm found in PartitionFinder2. We estimated the ML concatenated species tree using the partitions and models identified above (Chernomor et al. 2016) from 200 independent tree searches with 100 tree searches from parsimony starting trees and 100 tree searches from random starting trees. We estimated the nodal supports of the best ML species tree using 100 nonparametric bootstrap replicates (Felsenstein 1985).

We further explored the support for the relationships between the four Cicadidae subfamilies using gene trees and quartet frequencies, with a Docker v2.1.04 (Merkel 2014) container version of DiscoVista v1.0 (Sayyari et al. 2018). Quartet frequencies are estimated by sampling each nonterminal edge in the phylogeny. Each nonterminal edge leads to four taxa/clades and a four-taxon tree has three topologies. Quartet frequencies calculate the frequencies of each alternative topology for each internal edge.

We generated tree and alignment statistics using Python modules ETE v3 (Huerta-Cepas et al. 2016), Dendropy v4.4.0 (Sukumaran and Holder 2010), and BioPython v1.75 (Cock et al. 2009). Trees for illustrations were generated using FigTree v1.4.4 (Rambaut 2006–2018) and modified by hand in Affinity Designer v1.8.4 (Serif [Europe] Ltd.).

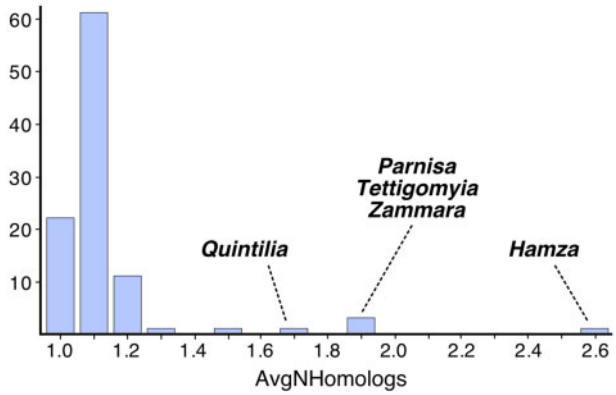


FIGURE 2. Histogram of values of the AHE pipeline statistic Average Number of Homologs (*AvgNHomologs*) (Table S2 of the Supplementary material available on Dryad) from the Cicadidae data set showing outlier values which were correlated with the presence of cross-sample contamination. X-axis shows upper value limit of each category.

### *Contamination Identification/Removal Analyses*

In preliminary analyses, we noticed that two taxa were recovered in unexpected phylogenetic positions that differed from those observed in earlier Sanger studies (see Discussion section). Examination of the AHE statistics revealed that these taxa (*Parnisa*, *Hamza*) were among 6–7 that exhibited outlier values of *AvgNHomologs* (Table S2 of the Supplementary material available on Dryad). These species had values ranging from 1.43 to 2.55, while the remaining taxa had values of 0.93–1.22 (Fig. 2). *AvgNHomologs* $\approx$ 1 implies no gene duplication in the history of the target genes (or that any duplications occurred so long ago that only one copy was enriched), which was our expectation since these orthologs are well conserved across the arthropod tree of life (e.g., Haddad et al. 2018). No genome duplications have been documented for Hemipterans. Outlier values of *AvgNHomologs* can also indicate the presence of contamination if the contaminant is not closely related to the template, or extreme heterozygosity due to hybridization. Examination of individual gene trees suggested that the genes in *Parnisa* and *Hamza* sorted into two cliques (Holland et al. 2010) supporting contrasting positions. For example, in some gene trees *Parnisa* (from South America, tribe Parnisini) was strongly supported as sister to *Kikihia rosea* (New Zealand, tribe Cicadettini), with the sequences often identical. The remainder of the genes placed *Parnisa* as a deep lineage among other Cicadettinae genera, as observed in an earlier Sanger-based phylogeny (Marshall et al. 2018). Manual checking of the other six outlier taxa suggested the same situation of 1) a bimodal support pattern with two gene-tree cliques or 2) an identical or near-identical sequence match to another taxon in the data set for one clique.

The recurring pattern of an identical or near-identical taxon match in one of two gene cliques led us to suspect that cross-contamination of samples in the data set had occurred, perhaps during laboratory dissections. Contamination with DNA from taxa outside the study was unlikely since a near match to a sampled taxon was

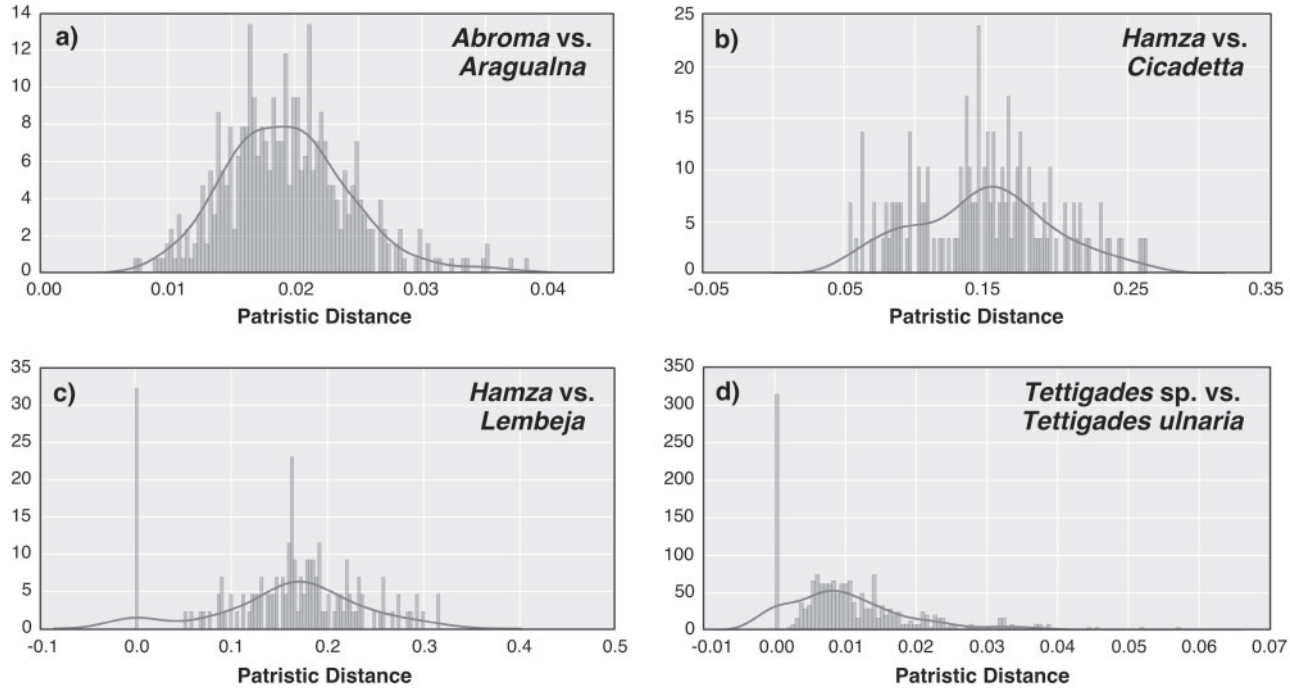


FIGURE 3. Example histograms of patristic distance values between focal taxa across all AHE loci showing patterns used to detect sample cross-contamination. Patristic distance values are obtained after rescaling all gene trees to a total gene-tree distance of 1.0 subst./site. a) Unimodal distribution expected from common phylogenetic signal across gene trees. b) Multimodal distribution suggesting conflicting sets of genes. No loci show near zero patristic distance so the contamination is not directly between these two taxa. c) Multimodal distribution with one peak close to zero probably caused by cross-contamination between these taxa. To remove contamination, we remove gene copies with patristic distances <0.01 from both matrices. d) Histogram from two species from very closely related genera. This bimodal distribution is probably an artifact caused by "binning" of very small genetic distances. Because of the discrete nature of genetic substitution and the short length of most of our AHE loci, distances slightly above zero are not possible. In our decision criteria, no loci are deleted if the second peak is also <0.01.

always involved. (See Discussion section for an explanation of how poor orthology inference was excluded.) Seeing the potential offered by the consistent association of bimodal gene-tree support and identical sequence matches, we sought to develop a method to identify and remove such cross-contaminated data without having to delete the affected taxa entirely, and without having to rely on *a priori* assumptions of taxon relatedness. This method would be an improvement from those using sequence similarity alone (see introduction).

Our contamination identification method uses ML gene trees and histograms of relative patristic distances between taxon pairs (distances obtained from gene trees rescaled to a common total length of 1.0). We assume that, given a large sample of loci across the genome, the rescaled patristic distances between two taxa should produce a unimodal distribution (Fig. 3a). Those taxa that are (partially) contaminated will produce a multimodal rather than unimodal distribution of patristic distances. When contamination occurs between samples, one of the peaks will appear at or near zero patristic distance in a comparison between a contaminated sample and its contaminant taxon (Fig. 3c). When a contaminated taxon is compared to another that is not the source of its contamination, the bimodal distribution will not have a peak close to zero patristic distance (Fig. 3b).

We developed a set of Python scripts to identify and remove gene sequences that are potentially involved in cross-contamination according to the above criteria. First, we estimated the ML tree for each locus (described above) and rescaled it to a total length of 1.0. Next, for each tree, we collected the pairwise patristic distances for all taxon pairs. Then, for each taxon pair, we plotted a histogram of the pairwise rescaled patristic distances for all loci and computed the kernel density using Matplotlib (Hunter 2007) and seabornv0.8.1 (https://github.com/mwaskom/seaborn/tree/v0.8.1) for visual inspection. Next, we determined whether the distribution is unimodal or multimodal using the Python v2 module PeakUtils v1.3.1 (https://bitbucket.org/lucashnegri/peakutils/src/master/). The PeakUtils parameter *min_len* sets the minimum length between distribution peaks. Another parameter, *thresh*, allows PeakUtils to identify peaks relative to the largest peak using a scalar value between 0 and 1. After many iterations of testing, we determined the best combination of *min_len* and *thresh* values to be 5.0 and 0.2, respectively. If the distribution of rescaled patristic distances was determined to be multimodal, then the peak locations were checked to determine if one was close to zero patristic distance. If this second condition was true, then for each gene in the near-zero peak (defined as a gene with a patristic distance < 0.01), we removed the sequence for that gene from its matrix in both involved taxa. Figure 4 illustrates the three decision rules of the pipeline. Thus, the method eliminates

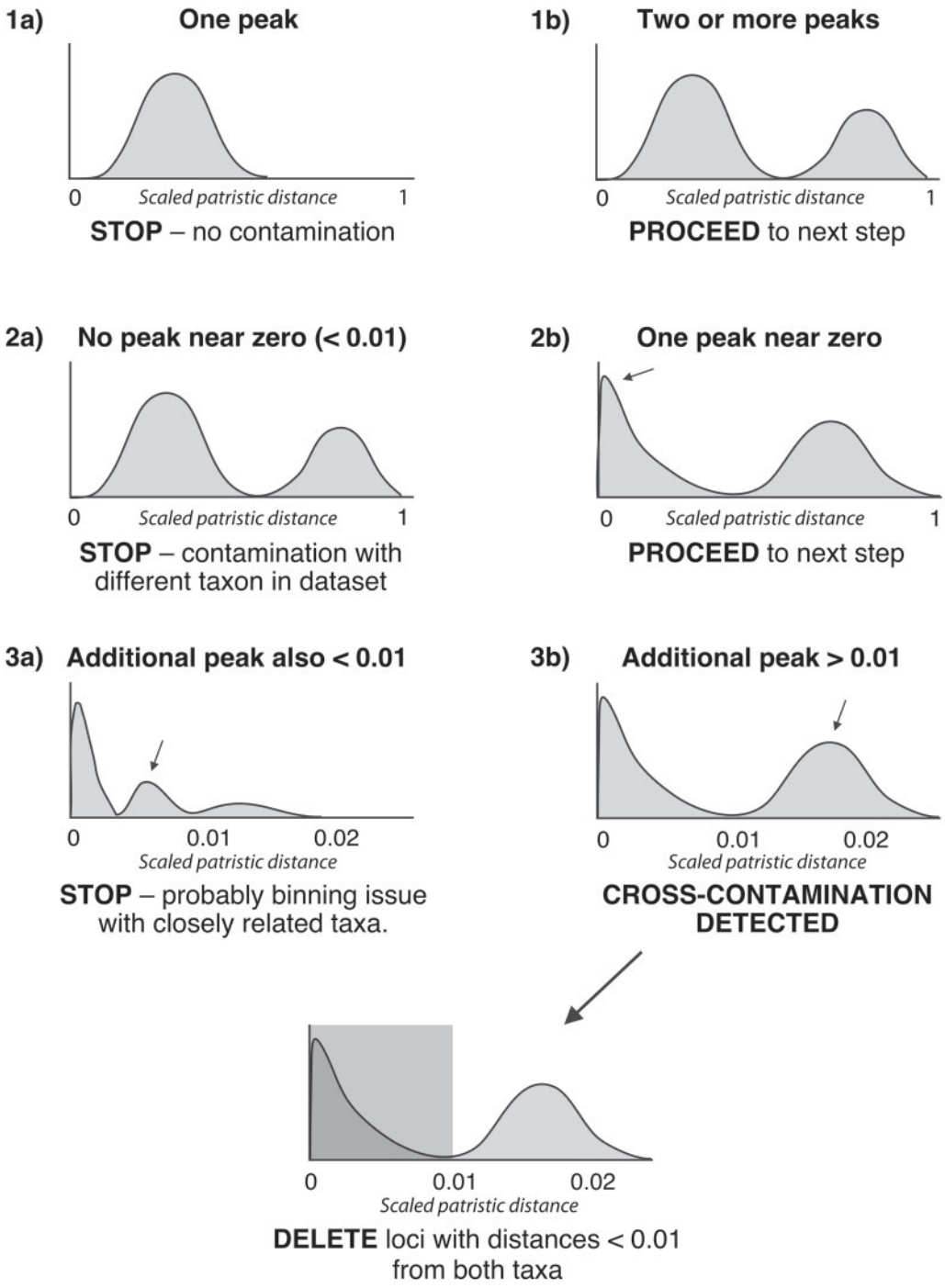


FIGURE 4. Decision rules involved in the contamination-detection pipeline. Step 1 determines if the histogram of rescaled patristic gene-tree distances between two taxa is unimodal or multimodal. Step 2 determines if one of the two peaks is at or near zero (<0.01 patristic distance). Step 3 determines if an additional peak is also less than 0.01 patristic distance, in which case the multimodal distribution is likely an artifact of close inter-taxon distance (see text). If 1b, 2b, and 3b are satisfied, gene sequences forming the near-zero peak are flagged as potentially involved in contamination and deleted from both taxa in the data matrix.

potentially contaminated sequences without appeal to *a priori* taxonomic knowledge and without requiring the removal of taxa from the study.

Because substitutions are discrete events and our gene lengths are not long, we found that a bimodal distribution of patristic distances can have two peaks below the critical value of 0.01. This is an expected artifact when two very closely related taxa are compared (Fig. 3d). As a result, we identified sequences as contaminated only when just one of the multimodal peaks had a patristic distance value less than 0.01.

After we removed all putatively contaminated sequences from the data matrices, we performed evolutionary model fitting and ML and bootstrap analyses again, as described above. Before these final analyses were conducted, a full phylogenetic analysis (IQ-TREE ML + ASTRAL-III, without the effects of nucleotide bias accommodated) was conducted with the taxa with outlier values for *AvgNHomologs* removed from the data set entirely. After deciding to proceed with a novel method for removal of contamination, we restored the outlier taxa to the data set and the contamination procedures above were applied.

## Results

### *Alignment Statistics*

The final sequence data belong to 429 loci (summary statistics are found in Table S2 of the Supplementary material available on Dryad). On average, 95 (SD = 6) taxa were sampled for each locus. Fifty-two loci (12%) contained all 102 taxa; therefore, the data set was decisive (Sanderson et al. 2010) and capable of distinguishing among alternative tree topologies based on our observed patterns of taxon occupancy in the data matrices for each gene. Although the data set was decisive, no taxon occurred in all 429 loci. *Chlorocysta suffusa* appears in most loci (427), while three taxa occurred in only 140 loci (*Parnisa* sp., *Zammara* c.f. *erna*, and *Azanicada zuluensis*). The average aligned sequence length for each locus was 268 bp (SD = 184) with the longest being 1819 bp and the shortest 101 bp. The mean number of parsimony informative sites for each locus was 96 (SD = 79), while the average number of constant sites among loci was 59 (SD = 8).

### *Contamination Identification Analyses*

The contamination identification pipeline identified potential contamination in 97 unique loci (23%) but only six taxon pairs: 1) *Z. erna* & *Platypleura octoguttata* (34 loci), 2) *Durangona tigrina* & *Hamza ciliaris* (18 loci), 3) *Quintilia wealei* & *Talcopsaltria olivei* (37 loci), 4) *Tettigomyia vespiformis* & *Odopoea insignifera* (30 loci), 5) *Parnisa* sp. & *K. rosea* (53 loci), and 6) *Lembeja paradoxa* & *H. ciliaris* (14 loci). Four of the 97 loci with contamination (4%) were identified as having four taxon pairs that were contaminated, while 42% (41/97) of contaminated loci contained only one pair of contaminated taxa.

Examination of gene trees by eye found other potential examples of contamination as determined by exact or near-exact well-supported discordant matches, but all of these involved small numbers of loci (1–4, usually 1), and could not be detected by the pipeline because they did not create detectable second peaks in the histograms of pairwise distances.

### *Evolutionary Models, Gene Trees, and Species Trees for Data before Removal of Cross-Contamination*

The concatenated alignment length was 115,063 bp and the PartitionFinder analysis favored 46 partitions, each with evolutionary model GTR+$G_4$. The BIC in PartitionFinder favored two partitions for most loci (56%), while a single model was favored by the BIC for five loci (1%). The most common best-fitting model among all partitions was HKY+I+$G_4$. This model was favored in 110 partitions, while the TRN+$G_4$ was the next most common model among partitions (82). Four models of molecular evolution only appeared once among partitions in all loci: K81, TVM+I, GTR, TVM, and F81+I+G.

The ML topology of the concatenated phylogeny largely favored monophyletic subfamilies, with just one genus apparently misclassified (*Aragualna plenalinea*) (Fig. 5a) (see Supplementary material: Refinement of Cicadidae Taxonomy available on Dryad). The most recent common ancestors of Cicadinae, Cicadettinae, and Tibicininae were all supported by BS = 100; while Tettigomyiinae was supported only weakly (BS = 58). Tibicininae was well supported (BS = 100) as a sister to all other subfamilies, but the relationships between the remaining three subfamilies were not resolved. The 11 taxa that were identified as affected by contamination (six pairs, two with one species in common) remained in their respective subfamilies.

The median bootstrap support for all gene trees that contain contamination was 41%, while the range of average bootstrap values among gene trees was 17–76%. The number of gene trees with contamination and an average bootstrap score ≥70 was 7 (2%). The mean tree length between gene trees was 2.45 (SD = 0.80). Gene tree lengths ranged from 0.78 to 5.34.

The ASTRAL-III topology of the contaminated data set (Fig. 6a) differed slightly from the ML phylogeny produced by IQ-TREE using the concatenated locus alignment. Cicadinae, Cicadettinae, and Tibicininae were all well supported by bootstrap support of 100%; however, Tettigomyiinae was estimated to be paraphyletic (Fig. 6a), rather than weakly monophyletic, with *Q. wealei* grouping with subfamily Cicadinae (LPP = 1.0) and sister to the remaining taxa of that clade (BS = 100). The rest of Tettigomyiinae was positioned as the sister lineage to the Cicadettinae with weak support (LPP = 0.8) (Fig. 6a). The locations of the contaminated taxa in both the ASTRAL-III and IQ-TREE ML concatenated trees are generally the same except for *Q. wealei* (as discussed above) and *H. ciliaris*, which was placed in a deeper position within its clade of African, Indian, and SE Asian Platypleurini in the ASTRAL-III tree compared to its position in the IQ-TREE ML tree (Figs. 5a and 6a).

### *Evolutionary Models, Gene Trees, and Species Trees after Removal of Cross-Contamination*

We use the phrase "contaminated data" to refer to the 372 sequences, belonging to 6 taxon pairs, that were flagged by the contamination identification pipeline and removed from 186 locus occurrences (97 unique loci). Removal of these 372 sequences from the overall 42,000 total increased the missing data by less than 1%, and

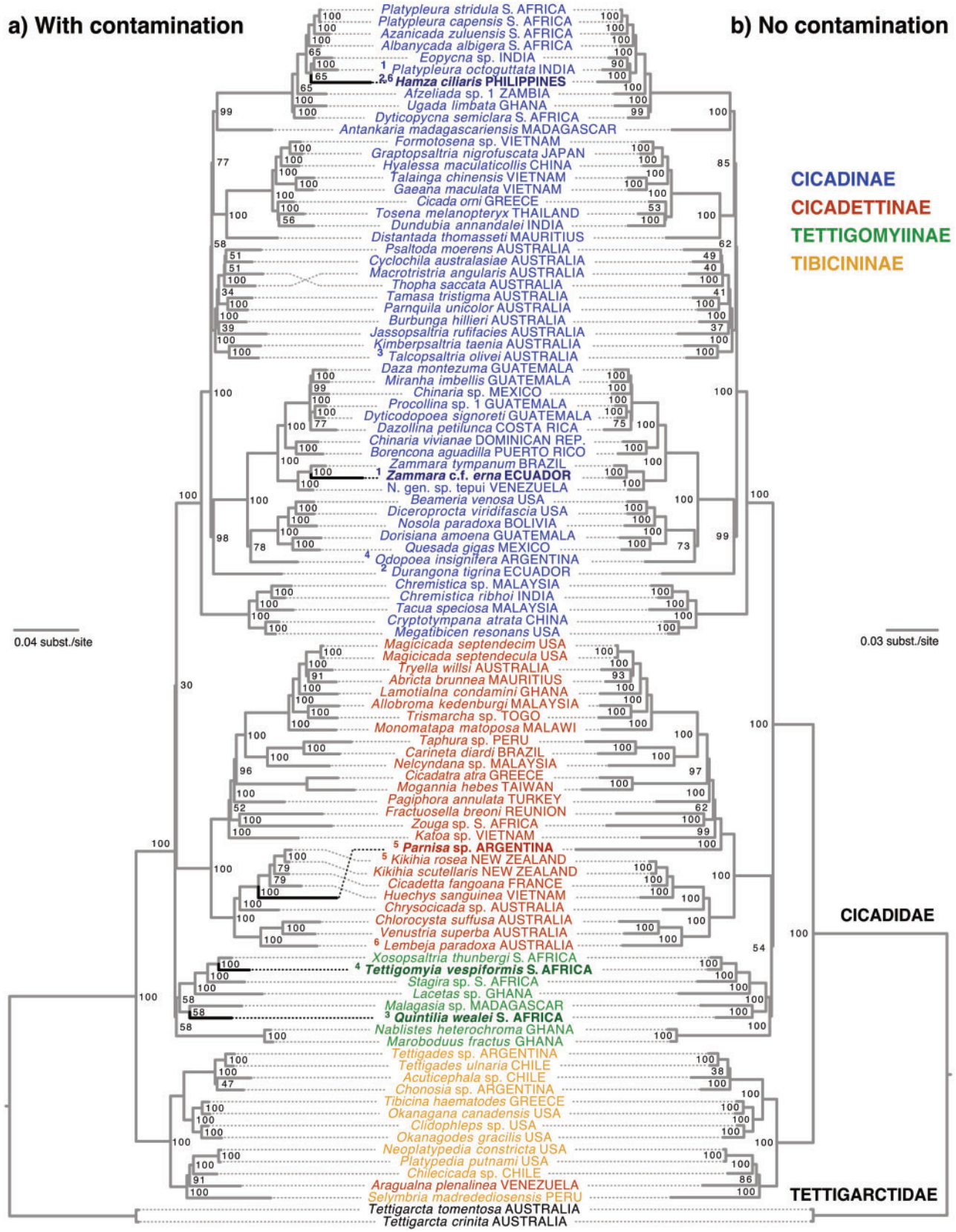


FIGURE 5. Comparison of IQ-TREE ML phylograms of the concatenated Cicadidae data set a) before and b) after removal of sequences involved in cross-contamination. Base-composition heterogeneity not corrected. Superscripts indicate taxon-pairs involved in contamination, with the contaminated taxon in bold font and with a highlighted branch. Support values shown are bootstrap percentages.

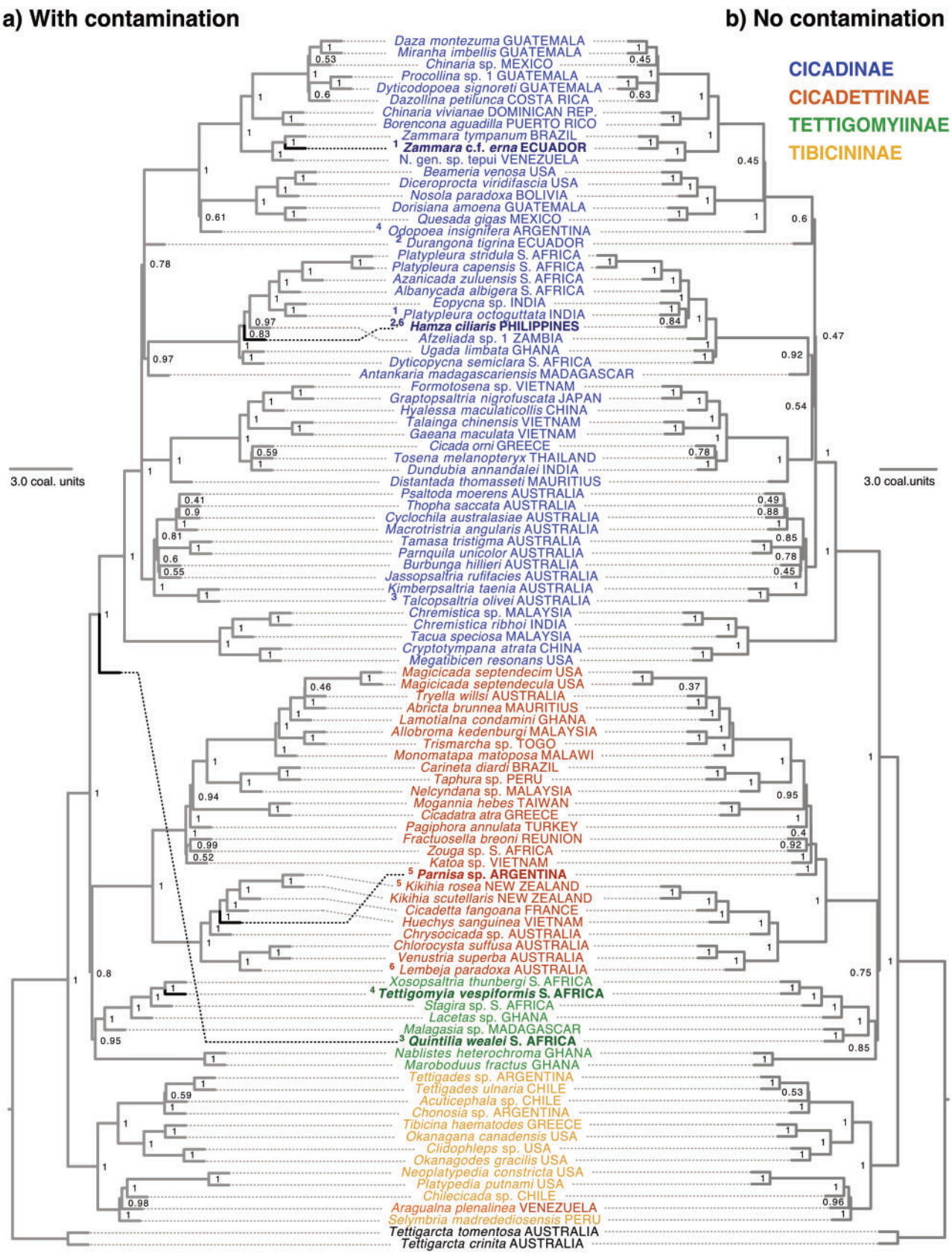


FIGURE 6. Comparison of ASTRAL-III phylograms of the Cicadidae data set a) before and b) after removal of sequences involved in cross-contamination. Base-composition heterogeneity not corrected. Superscripts indicate taxon-pairs involved in contamination, with the contaminated taxon in bold font and with a highlighted branch. Supports shown are local posterior probabilities.

the "contamination removed" data set remained decisive based on the observed taxon occupancy patterns. The number of loci (429) and the lengths of the concatenated alignment (115,063 bp) and individual loci remained the same. The following taxon occupancy reductions were made: *Parnisa* sp. 140 to 87 loci, *K. rosea* 401 to 348, *Q. wealei* 141 to 104, *T. olivei* 422 to 385, *Z. erna* 140 to 106, *P. octoguttata* 420 to 386, *T. vespiformis* 141 to 111, *O. insignifera* 413 to 383, *D. tigrina* 425 to 407, *H. ciliaris* 141 to 109, and *L. paradoxa* 412 to 398. The removal of contamination did not change the mean number of parsimony informative sites among loci (96) or the mean percentage of constant sites among loci (59%). After removing contamination, the BIC-favored models of evolution and partitioning schemes were similar to those estimated for the "contaminated" data set. HKY+I+$G_4$ was also the most common model among partitions, appearing in 117 cases. The least common model for the partitions did change.

The concatenated IQ-TREE ML species tree with contamination removed supported all subfamilies as monophyletic with BS = 100% (Fig. 5b). Tibicininae remained well-supported (BS = 100) as a sister to all other subfamilies. The relationships of Cicadettinae and Tettigomyiinae to Cicadinae remained unresolved, although Tettigomyiinae and Cicadettinae were weakly supported as sister taxa (BS = 54).

The removal of contamination changed the location and/or related bootstrap supports for some of the involved taxa in the ML analysis, in all but one case for only one taxon of a pair (Fig. 5b). *Parnisa* sp. moved away from *Kikihia* to a well-supported (BS = 100) deeply divergent position in a different subclade of Cicadettinae. The positions of *Q. wealei* and *T. olivei* did not change, but the support for the sister relationship between *Q. wealei* and *Malagasia* sp. increased from 58% to 100% as did support values for supporting nodes. No topology changes were noted for *Z. erna* or *P. octoguttata*, although support for *P. octoguttata* as sister to *Eopycna* sp. decreased from 100% to 90%. However, the tip branch for *Z. erna* shortened to more closely match lengths for other taxa in its clade. The position of *H. ciliaris* remained unchanged in the IQ-TREE tree but three supporting bootstraps increased after contamination removal from 65% to 100%, much like the *Tettigomyia* case. An unusually long tip branch for *Hamza* was also shortened to match others in its clade. No topology changes, bootstrap changes, or long branches were observed for *D. tigrina* or *L. paradoxa*, each of which had been diagnosed as involved with *Hamza* contamination. No clear changes were observed for *T. vespiformis* and *O. insignifera*.

The ASTRAL-III species tree inference of the contamination-removed data set largely agreed with the IQ-TREE contamination-removed ML tree except for the weakly resolved relationships within Cicadinae (Fig. 6b). Tibicininae, Cicadettinae, and Cicadinae all appeared with LPP supports of 1.0, while Tettigomyiinae was supported at LPP = 0.85.

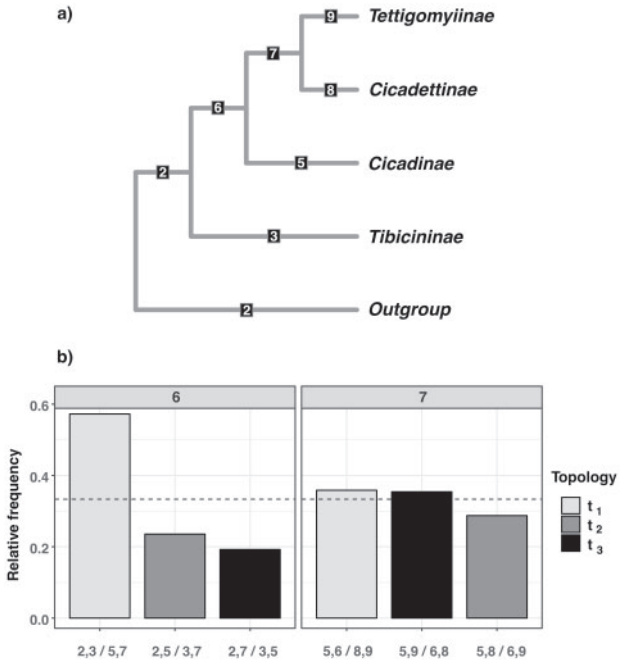


FIGURE 7. Support for alternative subfamily relationships evaluated using DiscoVista quartet sampling analysis of gene trees from the Cicadidae AHE data set. a) Species tree of subfamilies with branches numbered. b) Relative frequency of gene trees supporting three alternative arrangements of Branch 6 (left) and Branch 7 (right).

The placement and support of contaminated taxa in the ASTRAL-III analysis of "noncontaminated" data agreed with the IQ-TREE ML analysis of concatenated noncontaminated data (Figs. 5b and 6b). *Parnisa* sp. resolved as a deeper lineage in Cicadettinae, not closely related to *K. rosea*, and *H. ciliaris* moved to a position close to *Eopycna* and *P. octoguttata*.

IQ-TREE concatenated ML and ASTRAL-III coalescent analyses conducted on the data set after removing the six taxa with the most extreme *AvgNHomologs* values yielded nearly identical trees to those obtained with these taxa included after the removal of contaminated loci using our pipeline (Fig. S1 of the Supplementary material available on Dryad).

In the DiscoVista quartet sampling analysis, the frequency of "noncontaminated" gene trees strongly supported Tibicininae sister to all other Cicadidae subfamilies (Fig. 7). However, the relationships Tettigomyiinae + Cicadinae and Tettigomyiinae + Cicadettinae were present in nearly equal frequency in the "noncontaminated" gene trees. The sister relationship of Tettigomyiinae + Cicadettinae appeared in 35.85% of the gene trees, whereas Tettigomyiinae + Cicadinae was in 35.41% of gene trees.

### *Phylogenomic Analyses with Noncontaminated Data after Removing Compositionally Biased Data*

Filtering for compositional heterogeneity using p4 identified 61 loci. Removal of these lowered the total

number of loci to 368, which is 86% of the original total (429). This affected taxon occupancy, and the number of taxa in the "no contamination and no bias data set" ranged from 69 to 367, with a mean of 337 taxa per gene.

The IQ-TREE concatenated-data ML phylogeny estimated after contamination removal and correction for nucleotide bias was congruent in most relationships with the IQ-TREE ML phylogeny with only contamination removed (Fig. S2 of the Supplementary material available on Dryad). Topological changes only involved a small change with the placement of *Katoa* sp. within the Cicadettinae. Node support changes varied throughout the species tree with modest increases and decreases; however, all subfamilies were supported after contamination was removed and base composition was accounted for.

As with the IQ-TREE analyses, the ASTRAL-III analysis of the data set after removal of contamination and correction for nucleotide bias differed only subtly in topology and BS from the tree obtained after contamination removal alone (Fig. S3 of the Supplementary material available on Dryad). Three of the four subfamilies remained well supported (LPP = 1.00) after removing the compositionally biased loci. The support for Tettigomyiinae decreased LPP = 0.85 to LPP = 0.57 (Fig. S3 of the Supplementary material available on Dryad). Direct comparison of the IQ-TREE concatenated ML and ASTRAL-III coalescent trees obtained after removal of both contaminated and compositionally heterogeneous loci showed that supports tended to be stronger in the concatenated data tree, with a few exceptions (Fig. S4 of the Supplementary material available on Dryad).

The ASTRAL-III analyses exploring the removal of loci based on relative rates had a large effect on the topology at the subfamily rank. Specifically, Tettigomyiinae was estimated as polyphyletic when 5% of the slowest, 5% of the fastest, and 5% of the fastest and slowest loci were removed (Fig. S5 of the Supplementary material available on Dryad). In each species phylogeny, *Nablistes heterochroma* + *Maroboduus fractus* were sister to Cicadinae; however, the node was unsupported in each analysis (Fig. S5 of the Supplementary material available on Dryad). The remaining subfamilies were estimated as monophyletic and supported, but within each subfamily the relationships and support differed relative to the ASTRAL-III analysis with only contamination removed (Fig. 6b).

## Discussion

### *Removal of Cross-Contaminated Loci is Possible without Knowledge of Taxon Relatedness*

The contamination identification scripts flagged six putative examples of sample cross-contamination involving 11 taxa and 14–53 loci in each case; one taxon (*Hamza*) was flagged in comparison with two other taxa. Based on the examination of individual gene trees, we believe that the method identified all cases in which more than approximately four loci were involved, keeping in mind that the pipeline cannot distinguish contamination between extremely closely related taxa with rescaled patristic distances less than 0.01. We found approximately 20 other cases of potential cross-contamination, judging by identical or near-identical sequence matches for taxa otherwise believed to be more distantly related based on current classification (combined with much greater distances in the other loci). Fifteen of these involved just one locus, and the remainder involved 2–4 loci. Some cases involved taxa other than the 11 that were flagged by the contamination-removal scripts. Contamination involving only a few loci is undiagnosable by our contamination method, even with distantly related taxa, because the peak utility algorithm cannot diagnose a small peak height near zero. As the true evolutionary distance decreases, the number of contaminated loci must increase to be detected. Tuning the parameters of the peak detection algorithm to capture just one or a few contaminated loci leads to the undesirable removal of many loci for closely related taxon pairs. A similar problem occurs with the Breinholt et al. (2018) method when a small percent sequence divergence is used as a cutoff to remove potential contamination. We left these scattered cases of contamination in the data set because we did not want to exclude loci based on the current Cicadidae taxonomy, which is still being refined (Marshall et al. 2018; Sanborn et al. 2020). We regard the effect of rare erroneous loci to be comparable to rare miscalls in nucleotide sequences. Contamination by one or a few loci seems unlikely to substantially alter multispecies coalescent tree topology or BS patterns; however, studies have shown that a few loci can alter the topology and BSs in ML concatenated analyses (Brown and Thomson 2017; Shen et al. 2017).

Interesting questions remain regarding the process(es) and pattern of contamination in the data set. For example, we identified contamination related to 11 taxa, involving only 372 sequences out of more than 42,000. Why contamination affected so few sequences remains a mystery. No previous study finding cross-contamination among taxa in an NGS experiment has investigated the distribution of contamination among loci or addressed why nontargeted DNA affected only a small part of the assemblies (Longo et al. 2011; Merchant et al. 2014; Salter et al. 2014; Bemm et al. 2016; Ballenghien et al. 2017; Wilson et al. 2018; Francois et al. 2020). Although we found a solution to identify the pattern of contamination that worked for this data set, we hope that researchers will explore both the pattern of contamination and the processes that lead to it in phylogenomic experiments.

### *Modest Amounts of Contamination Can Alter the Topology and Bootstrap Scores*

Struck (2013) demonstrated that paralogous sequences can alter bootstrap supports and topology, and similar results should be expected from cross-contamination. Our results show that cross-contamination of samples can affect IQ-TREE ML and ASTRAL-III coalescent

analyses by "pulling" the contaminated species closer to the contaminating one in the species tree topology. In the most extreme case we identified, the 53 of 140 (37.8%) *Parnisa* loci introduced from *K. rosea* pulled *Parnisa* away from a deep position within subfamily Cicadettinae into a more distal position sister to the tribe Cicadettini which contains *Kikihia* (Figs. 5 and 6). It was this relationship that alerted us to a potential problem, because a sister–taxon relationship between *Parnisa* (from South America) and *Kikihia* (Australasian) would contradict the historical biogeography and tribal relationships implied by Sanger trees (Marshall et al. 2016, 2018; Simon et al. 2019). In the second most extreme case, 37 loci of 141 (26.2%) for the African species *Q. wealei* (subfamily Tettigomyiinae) were contaminated by sequences from *T. olivei* (Australia). This pulled *Quintilia* into a position sister to subfamily Cicadinae, which contains *Talcopsaltria*, but only in the ASTRAL-III analysis (Figs. 5 and 6). In the IQ-TREE analysis (Fig. 5), *Quintilia* held the same position found in the analysis without contamination, but only by a slight margin—the four nodes separating it from Cicadinae were supported by bootstraps of only 30–58%.

Our other examples show that the effect of contamination can be subtle, even with a substantial fraction of loci contaminated. For example, *Hamza* was contaminated by 18 and 14 gene sequences from two other species, one from a different subfamily. While the nodes supporting it suffered degraded support, its position was unchanged in the IQ-TREE ML analyses with concatenated data (Fig. 5). In the ASTRAL-III tree (Fig. 6), *Hamza* was pulled into a slightly deeper position in its clade. *Zammara* c.f. *erna* and *T. vespiformis* are even more surprising. While these taxa were contaminated by 34 and 30 loci, respectively, again from distantly related species, their unchanged positions were strongly supported by 100% bootstraps in all analyses (although some shifts in support occurred at nodes several steps deeper in the tree from these species). This might be explained in both cases by the presence of uncontaminated close relatives present in the tree, which hold the contaminated taxon in place despite considerable conflict. Returning to the *Parnisa* example, the subtlety of the problem is well illustrated by the 100% bootstraps supporting its incorrect position in the original IQ-TREE and ASTRAL-III trees (Figs. 5 and 6), despite the contamination of 37% of its loci.

Our results strongly suggest that the AHE pipeline statistics can be used to flag even phylogenetically cryptic cases of cross-contamination. As explained in the Results section, in three of the five taxon sets diagnosed as potentially contaminated, the removal of putatively contaminated data from the analysis substantially changed the topology and/or improved the supports for one member of the pair (*Parnisa, Quintilia*, and *Hamza*) (Figs. 5 and 6). In one additional case, a highly divergent long tip branch was observed for one member (*Z. erna*). These long branches might occur because substantial numbers of site mismatches are incorrectly modeled as apomorphic substitutions. These four taxa are among the six noted in the Results section as having the most extreme values for *AvgNHomologs* (1.68–2.55, compared to values up to 1.4 for the remainder) (Fig. 2, Table S2 of the Supplementary material available on Dryad). In the remaining case of diagnosed contamination, *T. vespiformis* and *O. insignifera*, it was *Tettigomyia* that exhibited the outlier *AvgNHomologs* value, suggesting that the change in its tip branch length between Figs. 5a and b was indeed an effect of contamination. We suspect that the changes in support observed for *O. insignifera* after contamination removal are a complicated outcome of the removal of contamination from *Z. erna*. Again, these five taxa are the ones we initially confirmed as likely to be contaminated by the hand curation of gene trees. *Azanicada zuluensis*, which has the sixth most extreme value for *AvgNHomologs*, shows little sign of contamination in hand curation or according to bimodal gene tree distance histograms. Because this taxon was present in only 140 gene matrices, we suspect that poor DNA quality and possibly problems with orthologs could be involved. Alternatively, this taxon could be contaminated with a relative that is too similar for the contamination-removal technique or hand curation to detect.

In addition to the contamination alone affecting the tree topology and bootstrap support by itself, there are other factors likely at play that could contribute to the effect of contamination on both topology and BS. Recently, Shen et al. (2017) demonstrated that a few genes or sites in an alignment can heavily influence a phylogenetic analysis. In our study, we believe that this is also true with reference to contaminated phylogenomic data when gene trees offer little resolution. For example, the ASTRAL-III phylogeny shows that Tettigomyiinae moves sister to Cicadinae and is not monophyletic when contamination is left in the data set (Fig. 6). This is probably driven by the fact that even when contamination is removed, the DiscoVista quartet frequencies analysis shows nearly equal prevalence in the gene trees of Tettigomyiinae relative to each of the other subfamilies (Fig. 7). Including contaminated data in the ASTRAL-III analysis renders Tettigomyiinae polyphyletic with the contaminated *Q. wealei* sister to Cicadinae. One way to limit the effects of undetected contamination is to have informative gene trees, but we recognize that potential rapid radiations, and short branches that accompany them, may be vulnerable to undetected contamination regardless of the strength of the signal across the gene trees.

The study by Shen et al. (2017) benefits from the fact that the data sets examined have historically been estimated many times with many different data types, but in this study, and in most phylogenomic data sets, some taxa are being included in the tree for the first time. We have previous cicada molecular (Marshall et al. 2018; Simon et al. 2019) and morphological (Moulds 2005) phylogenetic estimates at the subfamily rank, and only limited work evaluating tribal relationships (Marshall et al. 2016, 2018; Sanborn et al. 2020; Hill et al. 2021). Not having other phylogenetic hypotheses for some

or many taxa makes it difficult to propose alternative phylogenetic hypotheses in the face of contamination.

### *On the Potential Use of Mitochondrial DNA to Detect Contamination*

Bossert and Danforth (2018) noted that UCE and AHE experiments using universal capture baits can capture nontargeted sequences and ultimately introduce contamination into phylogenomic studies. They present the idea of using mtDNA from bycatch to flag the potential for contamination. Specifically, they propose checking for heterogeneous cytochrome oxidase subunit 1 (COI) signal (i.e., multiple "barcodes"), as an indication that a taxon and all of its sequences has been contaminated. Our results suggest that this method may lead to false negatives because we have shown that contamination may not lead to contaminant sequences being detectable for all captured loci. Moreover, some contaminated taxa may have two COI sequences, while others that are contaminated may have one copy, but otherwise be contaminated at other loci. In insects, this solution is complicated further because of the presence of nuclear copies of mitochondrial DNA segments (Bensasson et al. 2001; Song et al. 2008). Distinguishing between the real COI sequence, a COI paralog, and contamination seems nearly impossible without a reference genome or controlled Sanger sequencing experiments in which the real COI sequence and the COI paralog can be sequenced and compared.

Another way to use COI bycatch sequence data to determine whether the correct species was sequenced could be to compare the sequence to NCBI GenBank and confirm that the same organism, or a closely related taxon, is most similar. This approach also has associated caveats. First, one could compare the COI sequence to GenBank using sequence similarity (e.g., blastn), but this assumes that the same taxon or a closely related species is deposited in the sequence database. Many taxa in this study have never been sequenced for any gene. Furthermore, there are many studies demonstrating that identifying species, especially insects, using sequence similarity of one gene alone is not accurate even if the species does have a COI sequence in GenBank (e.g., Meiklejohn et al. 2019).

Phylogenetic analyses are also sometimes used to compare the bycatch COI data in a target-capture experiment to the nuclear capture data phylogeny to identify contamination or sample mix-up. This assumes that there will be congruence between the mtDNA and capture data; however, studies suggest many cases of evolutionary mitochondrial and nuclear discordance (e.g., Campbell et al. 2020; Prous et al. 2020).

### *Alternative Explanations for Gene-Tree Discordance*

Could processes other than specimen contamination create the patterns we observed in the Cicadidae AHE data set? Incorrect orthology assessment can create discordance between cliques of gene trees, but we argue that this explanation is unlikely here. Approximately 87 gene trees place Parnisa in a deep position apart from tribe Cicadettini. The remaining 53 loci place *Parnisa* as a tip branch within that tribe, identical or nearly identical to *K. rosea* from New Zealand. For those trees where *Parnisa* is identical to *Kikihia*, the simplest explanation based on incorrect orthology would be that, for those two taxa, the pipeline has assembled a nearly invariant nonhomolog. But, this explanation does not easily account for the consistent position of the *Parnisa* + *K. rosea* taxon pair across these trees, within the correct tribe for *Kikihia* (Cicadettini). Extensive introgression of *Kikihia* genes into *Parnisa*, because of recent hybridization, could create the pattern we observed, but this is impossible since the species are found on different continents.

### *AHE Loci Together Inform Deep-Level Cicadidae Relationships, But Not Individually*

In addition to exploring Cicadidae relationships and testing the subfamily classification with AHE loci, we assessed their utility for future phylogenomic experiments, especially within lower-ranking groups such as tribes and genera. We found that individual gene trees are uninformative, with average bootstrap scores less than 70% for nearly every gene tree even after we removed contaminated gene sequences (Supplementary Data available on Dryad). This is not surprising given other studies showing that individual gene trees rarely match the species tree (Salichos and Rokas 2013). The average bootstrap score among gene trees was 42%, while eight gene trees had average bootstrap support of ≥70%. The eight loci with relatively high average bootstrap scores were among the top 13 longest loci, which suggests that locus length is a driving factor behind the relatively higher average bootstrap values (see also Betancur-R et al. 2014). AHE gene trees that are weakly uninformative due to their short exon length are typical of invertebrate taxa at this time scale (e.g., Owen et al. 2020). Vertebrates have much longer exons (e.g., Ranwez et al. 2007) and consequently more highly supported gene trees. The average length of the Cicadidae AHE loci was 268 bp; therefore, one improvement to our probe data set would be a redesigned probe set that targets longer loci. This will be feasible only if the flanking regions (i.e., untranslated regions and introns) are not too divergent for this taxonomic scale. The probe set used here targets conserved loci that share a most recent common ancestor with Arthropoda (Haddad et al. 2018). Instead, a possible solution would be to build an AHE probe set where the orthologs were determined from a more recent common ancestor like Cicadidae, Cicadoidea, or Cicadomorpha. Others have noticed this trend too and are designing hybrid capture loci so that they are longer. For example, Karin et al. (2019) designed a hybrid capture probe set for squamates and demonstrated that the longer-length loci (>1500 bp)

outperformed other AHE and UCE data sets even with 56% or fewer loci. Within insects, Owen et al. (2020) developed a phylogenomic resource for Hemiptera that included identifying long-conserved exons within protein-coding orthologs. They identified 406 exons that were ≥600 bp among eight hemipteran genomes analyzed. The concatenated and multispecies coalescent phylogenies produced using the 406 exons matched the topologies and BSs of those phylogenies produced using all 3872 orthogroups with coding sequence and amino acids. This further suggests the potential of using fewer but more informative loci in phylogenomic experiments (Shen et al. 2016; Brown and Thomson 2017; Mongiardino Koch 2021).

Although our concatenated phylogenomic analyses strongly support the monophyly of all subfamilies except Tettigomyiinae, the branch lengths supporting the subfamilies suggest a rapid radiation. In all our IQ-TREE ML and ASTRAL-III coalescent trees, after the well-supported initial divergence of Tibicininae, the next lineages arise with very short stems, whether Tettigomyiinae is monophyletic or not. Specifically, the branch leading to the Tettigomyiinae is short relative to the branches subtending the other Cicadidae subfamilies (Fig. 5b, Fig. S4a of the Supplementary material available on Dryad). This short branch leading to the Tettigomyiinae was also seen in the earlier five-gene Sanger data set (Marshall et al. 2018). Developing and implementing a hybrid-capture data set based on longer loci may add support to the branch by adding additional informative sites contributing to the length of the branch; however, it is not a certain outcome. Adding additional tettigomyiine taxa will also be critically important and will be the focus of future studies.

In addition to referencing how well the Cicadidae AHE data perform to resolve the relationships among subfamilies, it is also important to critique the resolution near the phylogeny tips. Specifically, we wanted to determine how much genetic information was available to discern closely related species. Our phylogeny contains multiple shallow congeneric species (Fig. 5b, Fig. S4 of the Supplementary material available on Dryad). They are *Platypleura capensis* & *Platypleura stridula*, *Magicicada septendecim* & *Magicicada septendecula*, *Kikihia rosea* & *Kikihia scutellaris*, *Tettigades* sp. and *Tettigades ulnaria*, and *Tettigarcta crinita* & *Tettigarcta tomentosa*. In each of these cases, the concatenated phylogeny with contamination removed returned patristic distances between each pair <1% (Fig. 5b, Fig. S4a of the Supplementary material available on Dryad). The lack of genetic divergence between these taxa suggests that these loci are not evolving fast enough to produce larger genetic distances between congeneric taxa. This suggests that an alternative set of loci may be needed to reconstruct the evolutionary history of recently diverged species. A note of caution is in order here because our contamination identification methods cannot distinguish cross-contamination between very closely related species. However, the only cases that we found of taxa with patristic distances <1% all involve taxa that are congeneric or belong to genera known to be minimally divergent morphologically, such as *Platypedia*/*Neoplatypedia* and the two *Tettigades* taxa.

## Supplementary Material

Data available from the Dryad Digital Repository: http://dx.doi.org/10.5061/dryad.tht76hdz1.

## Data Availability

Data and scripts are available at Dryad Digital Repository: 10.5061/dryad.tht76hdz1.

## Competing interest

Mention of trade names or commercial products in this publication is solely for the purpose of providing specific information and does not imply recommendation or endorsement by the USDA; USDA is an equal opportunity provider and employer.

## Acknowledgments

The authors gratefully acknowledge assistance with specimen collection from a global network of collaborators: A. Mohagan, D. Mohagan, A. Sanborn, A. Bell, A. de Boer, F. Camposeco, J.-H. Chen, S. Chiswell, M. Claridge, S. Cowan, D. de le Cruz, J. Cooley, J. Cryan, S. Drosopoulos, T. Erwin, M. Gogala, J. Heath, M. Heath, M. Humphrey, H. Karube, O. Griffiths, F. Leguat, Z. Lei, E. M. Luque, T. McCabe, J. Monzon, B. Moulds, J. Olive, A. Prakash, B. Price, V. Sarkar, M. Schouten, J. Sueur, G. Svenson, D. Takiya, K. Prosenc Trilar, T. Trilar, J. Urban, M. Whiting, J. Xiang, J. Yoshimura, and the Bishop Museum. Jason Cryan and Julie Urban provided numerous essential samples from sites in Africa, South America, and Southeast Asia. Jian-Hong Chen, C. Dietrich, T. Trilar, and K. P. Trilar also provided samples from important regions. Mike Whiting and Brigham Young University's Monte L. Bean Life Science Museum contributed a large collection that will support ongoing research. Michel Boulard, Chris Dietrich, Hans Duffels, Masami Hayashi, Young June Lee, and Allen Sanborn assisted with specimen identification. We are grateful to Michelle Kortyna, Chris Zdyrski, Jake Cherry, and Sean Holland at Florida State University's Center for Anchored Phylogenomics for assistance with molecular data collection and analysis. We thank Eric R.L. Gordon and Mark Stukel for comments on the manuscript. Specimens collected by the authors were obtained in several countries and we wish to thank the many contacts and officials who assisted our research through the permitting process

(in alphabetical order by country): Argentina (Administracion de Parques Nacionales); Australia (Michelle Scott, Department of the Environment and Water Resources; Michelle Nissen, Department of Environment and Heritage Protection, QLD; Danny Stefoni, Department of Environment Regulation, WA); Chile (Administracion de Parques Nacionales); China (Z. Lei, Institute of Plant Protection, Chinese Academy of Agricultural Sciences); Costa Rica, Ministry of Environment and Energy and the National Institute of Biodiversity (I.J. Guevara Sequeira and H. Ramirez Murillo), permit nos 128-2003-OFAU and 2529201; Ghana, Wildlife Division, Forestry Commission (V. Attah), permit nos WD/A.185/Vol.6/22 and 005833; Madagascar, Ministry of the Environment, Forests, and Tourism (L.H. Rasoavahiny), permit no. 271/08/MEFT/SG/DGEF/DSAP/SSE; Malaysia, Economic Planning Unit, UPE (Munirah Abd. Manan), permit nos. 40/200/19 SJ.1040 (permit ID no. 1389), 40/200/19/1481 (permit ID no. 1933), and 40/200/19/1476; New Zealand (Department of Conservation, Te Papa Atawhai); Peru, Instituto Nacional de Recursos Naturales (G. Suarez de Freitas and A. Morizaki Taura) and Museo de Historia Natural, Universidad Nacional Mayor de San Marcos (G. Lamas Müller); Philippines, Protected Areas and Wildlife Bureau (Theresa Mundita S. Lim).

Research and collection permits for the Indian part of this project were issued by the state forest departments in Kerala (permit no. WL 10-3781/2012 dated 18/12/2012, and GO (RT) No. 376/2012/F and WLD dated 26/07/2012), Karnataka (permit no. 227/2014-2015 dated 2015/04/16), Goa (permit no. 2/21/GEN/WL and ET(S)/2013- 14/387 dated 2013/06/20), Nagaland (permit no. CWL/GEN/240/522-39, dated 14/08/2012), Meghalaya (permit no. FWC/G/173/Pt-II/474- 83, dated 27/05/2014), Arunachal Pradesh (permit no. CWL/G/13(95)/2011-12/Pt-III/ 2466-70, dated 16/02/2015), and West Bengal (permit no. 2115(9)/WL/4K-1/13/BL41, dated 06/11/2013), for which we thank the Principal Chief Conservator of Forest, Deputy Conservators of Forest, Wildlife Wardens and field officers of those states.

## Funding

This work was supported by NSF [DEB1655891, DEB0955849, DEB0720664, DEB0529679, and DEB0089946]. The Indian research component was partially funded by a USAID PEER Science Program grant [AID-OAA-A-11-00012], a Ramanujan Fellowship (Department of Science and Technology, Govt. of India) and a National Centre for Biological Sciences (NCBS) research grant to K.K. Logistical support and facilities for voucher deposition were provided by the University of Connecticut (UConn) Biodiversity Research Collection and the Museum and Field Stations Facility at the NCBS; the latter also supported the sequencing of the Indian material. The UConn Biocomputing Facility supported data analysis. The present study was also supported by the Vietnam Academy of Science and Technology (VAST) under the grant number NCXS02.04/22-23.

## References

Arcila D., Hughes L.C., Meléndez-Vazquez B., Baldwin C.C., White W.T., Carpenter K.E., Williams J.T., Santos M.D., Pogonoski J.J., Miya M., Ortí G. 2021. Testing the utility of alternative metrics of branch support to address the ancient evolutionary radiation of tunas, stromateoids, and allies (Teleostei: Pelagiaria). Syst. Biol. 70(6):1123–1144.

Austin J.J., Ross A.J., Smith A.B., Fortey R.A., Thomas R.H. 1997. Problems of reproducibility: does geologically ancient DNA survive in amber-preserved insects? Proc.R. Soc. Lond. B 264: 467–474.

Ballenghien M., Faivre N., Galtier N. 2017. Patterns of cross-contamination in a multispecies population genomic project: detection, quantification, impact, and solutions. BMC Biol. 15:1–16.

Bemm F., Weib C.L., Schultz J., Forster F. 2016. Genome of a tardigrade: Horizontal gene transfer or bacterial contamination? Proc. Natl. Acad. Sci. USA 113:E3054–E3056.

Bensasson D., Zhang D., Hartl D.L., Hewitt G.M. 2001. Mitochondrial pseudogenes: evolution's misplaced witnesses. Trends Ecol. Evol. 16:314–321.

Betancur-R R., Arcila D., Vari R.P., Hughes L.C., Oliveira C., Sabaj M.H., Orti G. 2019. Phylogenomic incongruence, hypothesis testing, and taxonomic sampling: the monophyly of characiform fishes. Evolution 73:329–345.

Betancur-R R., Naylor G.J. P., Orti G. 2014. Conserved genes, sampling error, and phylogenomic inference. Syst. Biol. 63:257–262.

Bossert S., Danforth B.N. 2018. On the universality of target-enrichment baits for phylogenomic research. Methods Ecol. Evol. 9(6):1453–1460.

Boyce G.R., Gluck-Thaler E., Slot J. C., Stajich J.E., Davis W.J., James T.Y., Cooley J.R., Panaccione D.G., Eilenberg J., De Fine Licht H., Macias A.M., Berger M.C., Wickert K.L., Stauder C.M., Spahr E.J., Maust M.D., Metheny A.M., Simon C., Kritsky G., Hodge K.T., Humber R.A., Gullion T., Short D.P.G., Kijimoto T., Mozgai D., Arguedas N., Kasson M.T. 2019. Psychoactive plant- and mushroom-associated alkaloids from two behavior-modifying cicada pathogens. Fungal Ecol. 41:147–164.

Breinholt J.W., Earl C., Lemmon A.R., Lemmon E.M., Xiao L., Kawahara A.Y. 2018. Resolving relationships among the megadiverse butterflies and moths with a novel pipeline for anchored phylogenomics. Syst. Biol. 67:78–93.

Brown, J.M., Thomson, R.C. 2017. Bayes factors unmask highly variable information content, bias, and extreme influence in phylogenomic analyses. Syst. Biol. 66(4):517–530.

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

Campbell E.O., Gage E.V., Gage R.V., Sperling F.A. 2020. Single nucleotide polymorphism-based species phylogeny of greater fritillary butterflies (Lepidoptera: Nymphalidae: Speyeria) demonstrates widespread mitonuclear discordance. Syst. Entomol. 45(2):269–280.

Campbell M.A., Łukasik P., Meyer M.M., Buckner M., Simon C., Veloso C., Michalik A., McCutcheon J.P. 2018. Changes in endosymbiont complexity drive host-level compensatory adaptations in cicadas. mBio 9:e02104–18.

Chernomor O., Von Haeseler A., Minh B.Q. 2016. Terrace aware data structure for phylogenomic inference from supermatrices. Syst. Biol. 65(6):997–1008.

Claridge M.F. 1985. Acoustic signals in the Homoptera: behavior, taxonomy, and evolution. Annu. Rev. Entomol. 30:297–317.

Cock P.J. A, Antao T., Chang J.T., Chapman B.A., Cox C.J., Dalke A., Friedberg I., Hamelryck T., Kauff F., Wilczynski B., et al. 2009. Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics 25:1422–1423.

Cooley J.R., Marshall D.C., Hill K.B.R. 2018. A specialized fungal parasite (*Massospora cicadina*) hijacks the sexual signals of periodical cicadas (Hemiptera: Cicadidae: *Magicicada*). Sci. Rep. 8:1432–1437.

Crotty S.M., Minh B.-Q., Bean N.G., Holland B.R., Tuke J., Jermiin L.S., Von Haeseler A. 2020. GHOST: recovering historical signal from heterotachously evolved sequence alignments. Syst. Biol. 69:249–264.

Cummins C.A., McInerney J.O. 2011. A method for inferring the rate of evolution of homologous characters that can potentially improve phylogenetic inference, resolve deep divergence and correct systematic biases. Syst. Biol. 60(6):833–844.

Degnan J.H., Rosenberg N.A. 2009. Gene tree discordance, phylogenetic inference and the multispecies coalescent. Trends Ecol. Evol. 24(6):332–340.

Derr J.N., Davis S.K., Woolley J.B., Wharton R.A. 1992. Reassessment of the 16S rRNA nucleotide sequence from members of the parasitic hymenoptera. Mol. Phylogenet. Evol. 1(4):338–341.

Dietrich C.H., Allen J.M., Lemmon A.R., Lemmon E.M., Takiya D.M., Evangelista O., Johnson K.P. 2017. Leafhopper and treehopper (Hemiptera: Cicadomorpha: Membracoidea) phylogeny: the limits of phylogenomics? Insect Syst. Divers. 1:57–72.

Du Z., Hasegawa H., Cooley J.R., Simon C., Yoshimura J., Cai W., Sota T., Li H. 2019. Mitochondrial genomics reveals shared phylogeographic patterns and demographic history among three periodical cicada species groups. Mol. Biol. Evol. 36:1187–1200.

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

Felsenstein J. 1985. Confidence intervals on phylogenetics: an approach using bootstrap. Evolution 39:783–791.

Foster P.G. 2004. Modeling compositional heterogeneity. Syst. Biol. 53:485–495.

Francois C.M., Durand F., Figuet E., Galtier N. 2020. Prevalence and implications of contamination in public genome resources: a case study of 43 reference arthropod assemblies. G3: Genes, Genomes, Genetics 10:721–730.

Fujisawa T., Koyama T., Kakishima S., Cooley J.R., Simon C., Yoshimura J., Sota T. 2018. Triplicate parallel life cycle divergence despite gene flow in periodical cicadas. Commun. Biol. 1:26:1–14.

Granados Mendoza C., Jost M., Hágsater E., Magallón S., van den Berg C., Lemmon E.M., Lemmon A.R., Salazar G.A., Wanke S., 2020. Target nuclear and off-target plastid hybrid enrichment data inform a range of evolutionary depths in the orchid genus *Epidendrum*. Front. Plant Sci. 10:1761.

Guindon S., Dufayard J.F., Lefort V., Anisimova M., Hordijk W., Gascuel O. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0 Syst. Biol. 59:307–321.

Haddad S., Shin S., Lemmon A.R., Lemmon E.M., Svacha P., Farrell B., Slipinski A., Windsor D., McKenna D.D. 2018. Anchored hybrid enrichment provides new insights into the phylogeny and evolution of longhorned beetles (Cerambycidae). Syst. Entomol. 43: 68–89.

Hamilton C.A., Lemmon A.R., Lemmon E.M., Bond J.E. 2016. Expanding anchored hybrid enrichment to resolve both deep and shallow relationships within the spider tree of life. BMC Evol. Biol. 16(1):212.

Hasan J., Crawford R.J., Ivanova E.P. 2013. Antibacterial surfaces: the quest for a new generation of biomaterials. Trends Biotechnol. 31:295–304.

Hill K.B. R, Marshall D.C., Marathe K., Moulds M.S., Lee Y.J., Pham T-H., Mohagan A.B., Sarkar V., Price B.W., Duffels J.P., Schouten M., de Boer A.J., Kunte K., Simon C. 2021. The molecular systematics and diversification of a taxonomically unstable group of primarily Asian cicada tribes related to Cicadini Latreille, 1802 (Hemiptera: Cicadidae). Invertebr. Syst. 35:570–601.

Holland B.R., Spencer H.G., Worthy T.H., Kennedy M. 2010. Identifying cliques of convergent characters: concerted evolution in the cormorants and shags. Syst. Biol. 59(4):433–445.

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

Hunter J.D. 2007. Matplotlib: a 2D graphics environment. Comput. Sci. Eng. 9:90.

Ito Y., Nagamine M. 1981. Why a cicada, *Mogannia minuta* Matsumura, became a pest of sugarcane: an hypothesis based on the theory of 'escape'. Ecol. Entomol. 6:273–283.

Jermiin L.S., Ho S.Y. W., Ababneh F., Robinson J., Larkum A.W. 2004. The biasing effect of compositional heterogeneity on phylogenetic estimates may be underestimated. Syst. Biol. 53:638–643.

Kalyaanamoorthy S., Minh B.Q., Wong T.K., Von Haeseler A., Jermiin L.S. 2017. ModelFinder: fast model selection for accurate phylogenetic estimates. Nat. Methods 14(6):587–589.

Karin B.R., Gamble T., Jackman T.R. 2019. Optimizing phylogenomics with rapidly evolving long exons: comparison with anchored hybrid enrichment and ultraconserved elements. Mol. Biol. Evol. 37:904–922.

Katoh K., Standley D.M. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol. Biol. Evol. 30:772–780.

Kayal E., Bentlage B., Pankey M.S., Ohdera A.H., Medina M., Plachetzki D.C., Collins A.G., Ryan J.F. 2018. Phylogenomics provides a robust topology of the major cnidarian lineages and insights on the origins of key organismal traits. BMC Evol. Biol. 18:68.

Kearse M., Moir R., Wilson A., Stones-Havas S., Cheung M., Sturrock S., Buxton S., Cooper A., Markowitz S., Duran C., Thierer T. 2012. Geneious basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics 28(12):1647–1649.

Lanfear R., Frandsen P.B., Wright A. M., Senfeld T., Calcott B. 2017. PartitionFinder 2: new methods for selecting partitioned models of evolution for molecular and morphological phylogenetic analyses. Mol. Biol. Evol. 34:772–773.

Lin N., Berton P., Moraes C., Rogers R.D., Tufenkji N. 2018. Nanodarts, nanoblades, and nanospikes: mechano-bactericidal nanostructures and where to find them. Adv. Colloid Interface Sci. 252:55–68.

Linklater D.P., Baulin V.A., Juodkazis S., Crawford R.J., Stoodley P., Ivanova E.P. 2020. Mechano-bactericidal actions of nanostructured surfaces. Nat. Rev. Microbiol. 19:8–22.

Logan D.P., Alspatch P.A. 2007. Negative association between chorus cicada, *Amphipsalta zelandica*, and armillaria root disease in kiwifruit.N. Z. Plant Protect. 60:235–240.

Longo M.S., O'Neill M.J., O'Neill R.J. 2011. Abundant human DNA contamination identified in non-primate genome databases. PLoS One 6:e16410.

Lovett B., Macias A., Stajich J.E., Cooley J.R. Eilenberg J., de Fine Licht H.H., Kasson M.T. 2020. Behavioral betrayal: how select fungal parasites enlist living insects to do their bidding. PLoS Pathog. 16:e1008598.

Łukasik, P., Chong R.A., Nazario K., Matsuura Y., Bublitz D., Campbell M.A., Meyer M., Van Leuven J.T., Pessacq P., Veloso C., Simon C., McCutcheon J.P. 2019. One hundred mitochondrial genomes of cicadas.J. Hered. 110:247–256.

Łukasik P., Nazario K., Van Leuven J.T., Campbell M.A., Meyer M., Michalik A., Pessacq P., Simon C., Veloso C., McCutcheon J.P. 2018. Multiple origins of interdependent endosymbiotic complexes in a genus of cicadas. Proc. Natl. Acad. Sci. USA 115:229–432.

Maddison W.P., Maddison D.R. 2019. Mesquite: a modular system for evolutionary analysis. Version 3.60. Available from: http://www.mesquiteproject.org.

Marshall D.C., Hill K.B. R., Moulds M.S., Vanderpool D., Cooley J. R., Mohagan A.B., Simon C. 2016. Inflation of molecular clock rates and dates: molecular phylogenetics, biogeography, and diversification of a global cicada radiation from Australasia (Hemiptera: Cicadidae: Cicadettini). Syst. Biol. 65:16–34.

Marshall D.C., Moulds M.S., Hill K.B.R., Price B.W., Wade E.J., Owen C.L., Goemans G., Marathe K., Sarkar V., Cooley J.R., et al. 2018. A molecular phylogeny of the cicadas (Hemiptera: Cicadidae) with a review of tribe and subfamily classification. Zootaxa 4424:1–64.

Matsuura Y., Moriyama M., Łukasik P., Vanderpool D., Tanahashi M., Meng X-Y., McCutcheon J.P., Fukatsu T. 2018. Recurrent symbiont recruitment from fungal parasites in cicadas. Proc. Natl. Acad. Sci. USA 115:E5970–E5979.

McCutcheon J.P., McDonald B.R., Moran N.A. 2009. Origin of an alternative genetic code in the extremely small and GC-rich genome of a bacterial symbiont. PLoS Genetics 5:e1000565.

Meade A., Pagel M. 2008. A phylogenetic mixture model for heterotachy. In: Pontarotti P., editor. Evolutionary biology from concept to application. Germany: Springer.p. 29–41.

Mehdipour M., Zamanian H., Farazmand H., Hosseini-Gharalari A. 2016. Disruption of reproductive behavior of grapevine cicada, *Cicadatra alhageos*, by acoustic signals playback. Entomol. Exp. Appl. 158:210–216.
Meiklejohn K.A., Damaso N., Robertson J.M. 2019. Assessment of BOLD and GenBank – their accuracy and reliability for the identification of biological materials. PLoS One 14(6):p. e0217084.
Merchant S., Wood D.E., Salzberg S.L. 2014. Unexpected cross-species contamination in genome sequencing projects. PeerJ 2:e675.
Merkel D. 2014. Docker: lightweight Linux containers for consistent development and deployment. Linux J. 239:2.
Meyer M., Kircher M. 2010. Illumina sequencing library preparation for highly multiplexed target capture and sequencing. Cold Spring Harb. Protoc. 2010(6):pdb-prot5448. doi:10.1101/pdb.prot5448.
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. Mol. Biol. Evol. 37(5):1530–1534.
Mongiardino Koch, N. 2021. Phylogenomic subsampling and the search for phylogenetically reliable loci. Mol. Biol. Evol. 38(9):4025–4038.
Moulds M.S. 2005. An appraisal of the higher classification of cicadas (Hemiptera: Cicadoidea) with special reference to the Australian fauna. Rec. Aust. Mus. 57:375–446.
Moulds M.S., Marshall D.C., Popple L.W. 2021. Kimberpsaltriini, a new tribe for a new Australian cicada allied to *Talcopsaltria* Moulds (Hemiptera: Cicadoidea: Cicadidae). Austr. Entomol. 48:149–160.
Myers J.G. 1929. Insect singers: a natural history of the cicadas. London; George Routledge and Sons. 304 pp.
Oakley T.H., Wolfe J.M., Lindgren A.R., Zaharoff A.K. 2013. Phylotranscriptomics to bring the understudied into the fold: monophyletic ostracoda, fossil placement, and pancrustacean phylogeny. Mol. Biol. Evol. 30(1):215–233.
Owen C.L., Stern D.B., Hilton S.K., Crandall K.A. 2020. Hemiptera phylogenomic resources: tree-based orthology prediction and conserved exon identification. Mol. Ecol. Resour. 20(5):1346–1360.
Peters R.S., Krogmann L., Mayer C., Donath A., Gunkel S., Meusemann K., Kozlov A., Podsiadlowski L., Petersen M., Lanfear R., Diez P.A. 2017. Evolutionary history of the Hymenoptera. Curr. Biol. 27:1013–1018.
Philippe H., Brinkmann H., Lavrov D.V., Littlewood D.T. J., Manuel M., Worheide G., Baurain D. 2011. Resolving difficult phylogenetic questions: why more sequences are not enough. PLoS Biol. 9:e1000602.
Prous M., Lee K.M., Mutanen M. 2020. Cross-contamination and strong mitonuclear discordance in *Empria* sawflies (Hymenoptera, Tenthredinidae) in the light of phylogenomic data. Mol. Phylogenet. Evol. 143:106670.
Prum R.O., Berv J.S., Dornburg A., Field D.J., Townsend J.P., Lemmon E.C., and Lemmon A.R. 2015. A fully resolved, comprehensive phylogeny of birds (Aves) using targeted next generation DNA sequencing. Nature 526:569–573.
Rambaut A. 2006–2018. Figtree v1.4.4. Available from: https://github.com/rambaut/figtree/releases (accessed 29 August 2020).
Ranwez V., Delsuc F., Ranwez S., Belkhir K., Tilak M.K., Douzery E.J. 2007. OrthoMaM: a database of orthologous genomic markers for placental mammal phylogenetics. BMC Evol. Biol. 7(1):241.
Regier J.C., Shultz J.W., Zwick A., Hussey A., Ball B., Wetzer R., Martin J.W., Cunningham C.W. 2010. Arthropod relationships revealed by phylogenomic analysis of nuclear protein-coding sequences. Nature 463(7284):1079–1083.
Rokyta D.R., Lemmon A.R., Margres M.J., and Arnow K. 2012. The venom-gland transcriptome of the eastern diamondback rattlesnake (*Crotalus adamanteus*). BMC Genomics 13:312.
Salichos L., Rokas A. 2013 Inferring ancient divergences requires genes with strong phylogenetic signals. Nature 497:327–331.
Salter S.J., Cox M.J., Turek E.M., Calus S.T., Cookson W.O., Moffat M.F., Turner P., Parkhill J., Loman N.J., Walker A.W. 2014. Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. BMC Biol. 12:87.
Sanborn A.F. 2018. The cicada genus *Procollina* Metcalf, 1952 (Hemiptera: Cicadidae): redescription including fourteen new species, with a key to the species of the subtribe Dazina Kato, 1932 rev. stat., the description of the Aragualnini n. tribe, and one new combination. Zootaxa 4389:1–65.
Sanborn A.F. 2021a. The cicadas (Hemiptera: Cicadoidea: Cicadidae) of Madagascar including a new tribe, five new genera, twelve new species, four new species synonymies, five revised species status, ten new combinations, new tribal assignments for four genera, one new subtribe synonymy, a checklist and key to the species. Zootaxa 4937:1–079.
Sanborn A.F. 2021b. A new species, genus and tribe of cicada (Hemiptera: Cicadoidea: Cicadidae: Tibicininae) from Chile with a list of Chilean cicada fauna. Zootaxa 4952:87–100.
Sanborn A.F., Marshall D.C., Moulds M.S., Puissant S., Simon C. 2020. Redefinition of the cicada tribe Hemidictyini Distant, 1905, status of the tribe Iruanini Boulard, 1993 rev. stat., and the establishment of Hovanini n. tribe and Sapantangini n. tribe (Hemiptera: Cicadidae). Zootaxa 4747:133–155.
Sanderson M.J., McMahon M.M., Steel M. 2010. Phylogenomics with incomplete taxon coverage: the limits to inference. BMC Evol. Biol. 10:155.
Sayyari E., Mirarab S. 2016. Fast coalescent-based computation of local branch support from quartet frequencies. Mol. Biol. Evol. 33:1654–1668.
Sayyari E., Whitfield J.B., Mirarab S. 2018. DiscoVista: interpretable visualizations of gene tree discordance. Mol. Phylogenet. Evol. 122:110–115.
Shen X.X., Hittinger C.T., Rokas, A. 2017. Contentious relationships in phylogenomic studies can be driven by a handful of genes. Nat. Ecol. Evol. 1(5):1–10.
Shen X.X., Salichos L., Rokas A. 2016. A genome-scale investigation of how sequence, function, and tree-based gene properties influence phylogenetic inference. Genome Biol. Evol. 8:2565–2580.
Simao F.A., Waterhouse R.M., Ioannidis P., Kriventseva E.V., Zdobnov E.M. 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics 31:3210–3212.
Simion P., Belkhir K., Francois C., Veyssier J., Rink J.C., Manuel M., Philippe H., Telford M.J. 2018. A software tool 'CroCo' detects pervasive cross-species contamination in next generation sequencing data. BMC Biol. 16:1–9.
Simion P., Philippe H., Baurain D., Jager M., Richter D.J., Di Franco A., Roure B., Satoh N., Queinnec E., Ereskovsky A. Lapebie P. 2017. A large and consistent phylogenomic dataset supports sponges as the sister group to all other animals. Curr. Biol. 27(7):958–967.
Simon C., Gordon E.R., Moulds M.S., Cole J., Haji D., Lemmon A.R., Lemmon E.M., Kortyna M., Nazario K., Wade E.J., Meister R., Goemans G., Chiswell S.M., Pessacq P., Veloso C., McCutcheon J.P., Łukasik P. 2019. Off-target capture data, endosymbiont genes and morphology reveal a relict lineage sister to all other singing cicadas. Biol.J. Linn. Soc. 128:865–886.
Smith S.A., Pease J.B. 2017. Heterogeneous molecular processes among the causes of how sequence similarity scores can fail to recapitulate phylogeny. Brief. Bioinformatics 18:451–457.
Song H., Buhay J.E., Whiting M.F., Crandall K.A. 2008. Many species in one: DNA barcoding overestimates the number of species when nuclear mitochondrial pseudogenes are coamplified. Proc. Natl. Acad. Sci. USA 105(36):13486–13491.
Struck T.H. 2013. The impact of paralogy on phylogenomic studies – a case study on annelid relationships. PLoS One 8:e62892.
Sukumaran J., Holder M.T. 2010. DendroPy: a Python library for phylogenetic computing. Bioinformatics 26:1569–1571.
Thomas R.H., Schaffner W., Wilson A. C., and Pääbo S. 1989. DNA phylogeny of the extinct marsupial wolf. Nature 340:465–467.
Townsend J.P. 2007. Profiling phylogenetic informativeness. Syst. Biol. 56(2):222–231.
van der Valk T., Vezzi F., Ormestad M., Dalen L., Guschanski K. 2020. Index hopping on the Illumina HiseqX platform and its consequences for ancient DNA studies. Mol. Ecol. Resour. 20(5):1171–1181.

Van Leuven J.T., Meister R.C., Simon C., McCutcheon J.P. 2014. Sympatric speciation in a bacterial endosymbiont results in two genomes with the functionality of one. Cell 158:1270–1280.

Williams K.S., Simon C. 1995. The ecology, behavior, and evolution of periodical cicadas. Annu. Rev. Entomol. 40:269–295.

Wilson C.G., Nowell R.W., Barraclough T.G. 2018. Cross-contamination explains "inter and intraspecific horizontal genetic transfers". Curr. Biol. 28:2436–2444.

Xie G., Zhang G., Lin F., Zhang J., Liu Z., Mu S. 2008. The fabrication of subwavelength anti-reflective nanostructures using a bio-template. Nanotechnology 19:1–5.

Yang Y., Smith S.A. 2013. Optimizing de novo assembly of short-read RNA-seq data for phylogenomics. BMC Genomics 14(1): 328.

Zada I., Zhang W., Li Y., Sun P., Cai N., Gu J., Liu Q., Su H., Zhang D. 2016. Angle dependent antireflection property of TiO2 inspired by cicada wings. Appl. Phys. Lett. 109:153701.

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

Zhang C., Sayyari E., Mirarab S. 2017. ASTRAL-III: increased scalability and impacts of contracting low support branches. In: Meidanis J., Nakhleh L., editors. Comparative genomics. RECOMB-CG 2017. Lecture Notes in Computer Science. Vol. 10562. Cham: Springer.p. 53–75.

Zhang D.-X., Hewitt, G.M. 2003. Nuclear DNA analyses in genetic studies of populations: practice, problems and prospects. Mol. Ecol. 12:563–584.

Zhang G., Xhang J., Xie G., Liu Z., Shao H. 2006. Cicada wings: a stamp from nature for nanoimprint lithography. Small 2:1440–1443.

Zwickl D.J. 2006. Genetic algorithm approaches for the phylogenetic analysis of large biological sequence datasets under the maximum likelihood criterion [PhD dissertation]. Austin, TX: The University of Texas. 115 pp.