Estimating phylogenetic relationships despite discordant gene trees across loci: the species tree of a diverse species group of feather mites (Acari: Proctophyllodidae)

LACEY L. KNOWLES[1]* *and* PAVEL B. KLIMOV[1]*

[1] *University of Michigan, Department of Ecology and Evolutionary Biology, Museum of Zoology, 1109 Geddes Ave., Ann Arbor, Michigan 48109-1079, USA*

(*Submitted 14 December 2010; Revised 4 February 2011; accepted 7 February 2011; first published online 20 April 2011*)

SUMMARY

With the increased availability of multilocus sequence data, the lack of concordance of gene trees estimated for independent loci has focused attention on both the biological processes producing the discord and the methodologies used to estimate phylogenetic relationships. What has emerged is a suite of new analytical tools for phylogenetic inference – species tree approaches. In contrast to traditional phylogenetic methods that are stymied by the idiosyncrasies of gene trees, approaches for estimating species trees explicitly take into account the cause of discord among loci and, in the process, provides a direct estimate of phylogenetic history (i.e. the history of species divergence, not divergence of specific loci). We illustrate the utility of species tree estimates with an analysis of a diverse group of feather mites, the *pinnatus* species group (genus *Proctophyllodes*). Discord among four sequenced nuclear loci is consistent with theoretical expectations, given the short time separating speciation events (as evident by short internodes relative to terminal branch lengths in the trees). Nevertheless, many of the relationships are well resolved in a Bayesian estimate of the species tree; the analysis also highlights ambiguous aspects of the phylogeny that require additional loci. The broad utility of species tree approaches is discussed, and specifically, their application to groups with high speciation rates – a history of diversification with particular prevalence in host/parasite systems where species interactions can drive rapid diversification.

Key words: Acari, gene tree, feather mites, *pinnatus*-group, *Proctophyllodes*, species tree.

INTRODUCTION

Phylogenetic approaches can provide unprecedented insights into the patterns of species relatedness, as well as on the biological processes generating molecular divergence among species, by incorporating models of genetic processes into the phylogenetic estimation procedure. For example, phylogenetic estimates have been improved by what is now the routine use of nucleotide models of molecular evolution in phylogenetic methods (Felsenstein, 2004). Likewise phylogenetic procedures that incorporate models of other biological processes underlying patterns of molecular divergence among species – species tree approaches (Knowles and Kubatko, 2010) – also can significantly improve phylogenetic estimates.

Notable among the insights that species tree approaches offer is the phylogenetic resolution of some notoriously difficult scenarios for historical reconstruction (Maddison and Knowles, 2006; Carstens and Knowles, 2007; Brumfield *et al*. 2008; Kubatko and Gibbs, 2010). This includes cases involving recently diverged species (e.g. species A and B in Fig. 1), as well as cases of rapid speciation (e.g. the short internal branches in the species tree separating species E from the ancestor that gave rise to the sister taxa C and D in Fig. 1). Under such situations gene tree discord is expected (Takahata, 1989; Maddison, 1997) and species relationships may be obscured by the deep coalescence (i.e. the failure of gene lineages to coalesce within a species lineage before subsequent speciation events; see Fig. 1). Likewise, when the discord among gene trees is not taken into account when estimating a phylogeny, the general reliability of the inference becomes questionable (Kubatko and Degnan, 2007; Degnan and Rosenberg, 2009; Huang and Knowles, 2009) and the interpretation of support for particular relationships becomes problematic (Mossel and Vigoda, 2005; Cranston *et al*. 2009). Lastly, accurate phylogenies can be estimated with fewer loci with a species tree approach relative to the failure to incorporate an appropriate model (e.g. by concatenating data despite differences in the gene trees of independent loci). This means that investigations into the amount of data required for accurate phylogenetic estimates have been overestimated and reflect the problems

* Corresponding authors: University of Michigan, Museum of Zoology 1109 Geddes Ave., Ann Arbor, Michigan 48109-1079 USA. LLK: Phone: (734) 763-5603. Fax: (734) 763-4080. E-mail: knowlesl@umich.edu. PBK: Phone: (734) 763-4354. Fax: (734) 763-4080. E-mail: pklimov@umich.edu

 © Cambridge University Press 2011
doi:10.1017/S003118201100031X

https://doi.org/10.1017/S003118201100031X Published online by Cambridge University Press

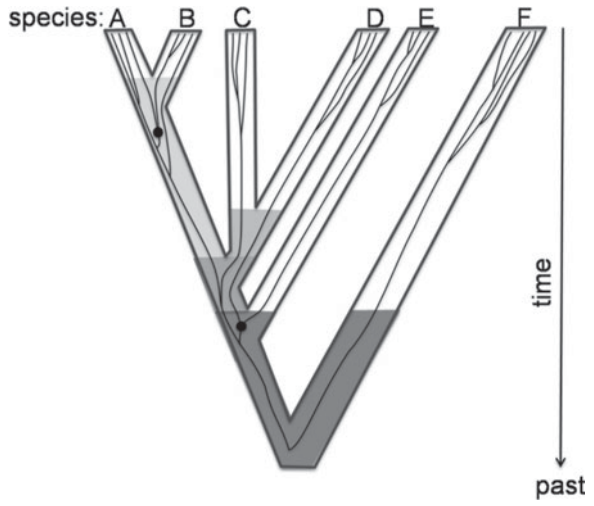

Fig. 1. Species tree with contained gene tree showing the deep coalescence of gene lineages (marked with circles). Incongruence between a gene tree and the underlying species tree may occur in not only recently derived species, but also can extend back in time along the internal species-tree branches (i.e. those shown in shades of grey). The deep coalescence of gene lineages in the more distant past (i.e. the short internal branches in the species tree separating species E from the ancestor that gave rise to the sister taxa C and D) arises from the random loss of gene lineages by genetic drift of the two or more gene lineages that coexisted in the past, even though multiple ancestral gene lineages no longer persist in species C, D and E because they did not diverge recently. In other words, even in monophyletic species (e.g. species C, D and E), deep coalescence in the past leads to gene tree-species tree discord in the present (from Knowles, 2009).

associated with concatenating data-sets, as opposed to an inherent property of estimating species relationships *per se* (see Maddison and Knowles, 2006; Liu and Pearl, 2007).

Despite the relative recency of species tree approaches (i.e. Maddison and Knowles, 2006), there has been a proliferation of methods (reviewed in Degnan and Rosenberg, 2009; Liu *et al*. 2009; Knowles and Kubatko, 2010). While much of the research on obtaining direct estimates of species trees has been driven by computational developments, these methodological changes do not represent the inception of new core phylogenetic concepts. In spite of the fact that estimating species trees involves a fundamental shift in how molecular data are used and interpreted, the target is still the phylogeny. The estimation of a species tree puts the focus on the object of systematic interest – species relationships.

Here we estimate a species tree for a group of very diverse feather mites – the *pinnatus* species group (genus *Proctophyllodes*) (see Fig. 2). Species of the genus *Proctophyllodes* are common, worldwide-distributed ectosymbionts of passeriform birds (rarely others) that inhabit the underside of wing and tail flight feathers at all developmental stages (Atyeo and Braasch, 1966). The mites feed on uropygial gland secretions and other material trapped in this oil (e.g. aging feather fragments, sloughed cells from the skin), but do not appear to cause damage to the bird feathers or skin (Atyeo and Braasch, 1966; Blanco *et al*. 1999; Hartup *et al*. 2004).

We focus on the *pinnatus*-group of feather mites, with 35 described species, because of several aspects of the history of diversification. Preliminary phylogenetic analyses reveal that the taxa in the *pinnatus*-group of feather mites are characterized by relatively short internodes compared to the other described species of *Proctophyllodes* (Fig. 3). Secondly, the *pinnatus*-group has a large number of constituent species (i.e. it is among the most diverse group of the ten species groups in the genus *Proctophyllodes*) (Atyeo and Braasch, 1966; Badek *et al*. 2008), which means that the speciation rate is higher compared to other groups of similar age. Such historical scenarios – the formation of many species over a relatively short period of time – are expected to be characterized by widespread discord among gene trees (reviewed in Knowles and Kubatko, 2010). As such, the *pinnatus*-group of feather mites is an ideal group to analyze using species-tree approaches that model the discord among loci, rather than ignore it (as with analyses of concatenated data).

MATERIALS AND METHODS

Taxon Sampling

A total of 21 species (of the 35 described species from the *pinnatus*-species group, genus *Proctophyllodes*) and 3 outgroup species were sequenced in this study (Table 1). Of these, 21 ingroup species, 27 individuals collected from 26 bird hosts were sequenced such that the study also included samples across hosts for a given species (Table 1). Genomic DNA was extracted according to previously described protocols (Klimov and OConnor, 2008).

Sequence Data

Four nuclear genes: the ribosomal loci 18S (1767 nt aligned) and 28S (3677 nt), and the protein-coding loci elongation-factor 1α, EF1-α (1215 nt), and heat-shock protein-70, HSP70-5 (1713 nt) were sequenced. Alignments of the ribosomal loci conformed with alignments of a large mite data-set (based on 543 sequences of mites; data not published) using as reference the secondary structures of *Apis mellifera* (Gillespie *et al*. 2006) and *Saccharomyces cerevisiae* available from the Comparative RNA Web site (Cannone *et al*. 2002). Individual sequences, especially hairpin-stem loops, were further evaluated in the program mfold v.3.1, which folds rRNA based on free energy minimization (Mathews *et al*. 1999; Zuker, 2003), using the default settings. Although

https://doi.org/10.1017/S003118201100031X Published online by Cambridge University Press

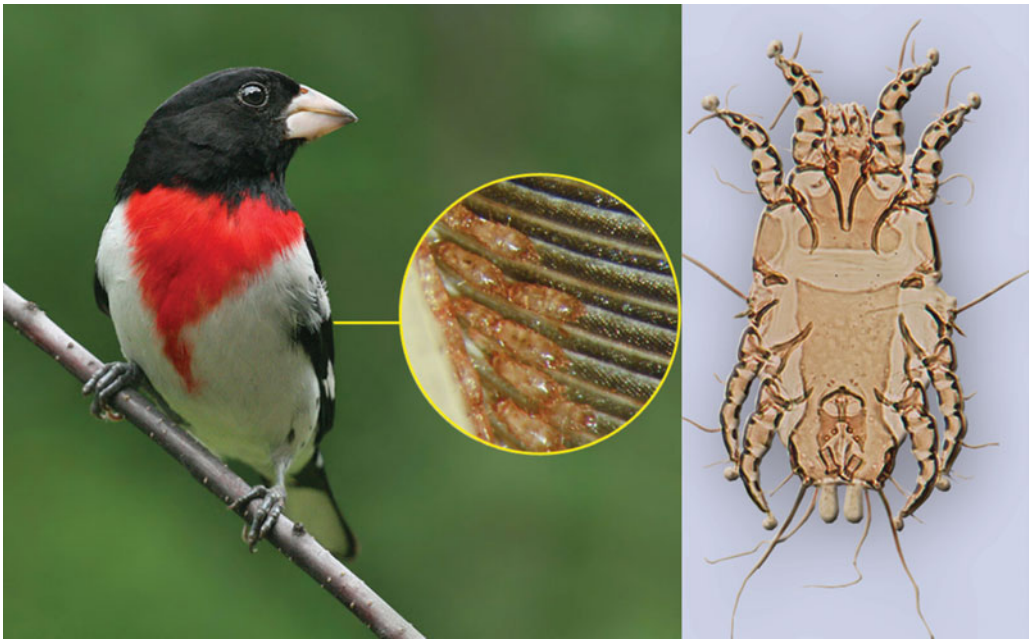

Fig. 2. Rose-breasted grosbeak, *Pheucticus ludovicianus* (left), the host of the mite *Proctophyllodes pheuctici* (right) showing the typical location of mites on the wing primary feathers. The bird photo was downloaded from http://pixdaus.com (public domain); the photos of the mites are from Barry OConnor (UMMZ vouchers BMOC 08-0320-003 and BMOC 07-0626-001; reproduced with permission).

alignments of exons of protein-coding genes were unambiguous (i.e. they did not contain gaps), a few regions containing introns were excluded from the phylogenetic analyses to avoid errors associated with possible mis-assignment of homology arising from gaps in the sequences. A total of 7982 nt were analyzed (18S: 1722 nt; 28S: 3485 nt; EF1-α: 1077 nt; HSP70: 1698 nt). Primer sequences and PCR protocols are given in Supplement 1 (The online supplementary material can be viewed at http://journals.cambridge.org/par). All sequences were deposited in GenBank (Accession Nos. HM165035 through HM165154) (Table 1).

Phylogenetic analyses

Best-fit models of nucleotide substitution were identified for each locus using Akaike information criterion values in the program Modeltest (Posada and Buckley, 2004) and used in the Bayesian estimates of species trees and gene trees.

Species and gene trees were estimated in *BEAST v.1.5.4 (Drummond and Rambaut, 2007). The program *BEAST (Heled and Drummond, 2010) was used, as opposed to BEST (see Liu *et al*. 2009), because of computational differences that make *BEAST more efficient (for details, see Heled and Drummond, 2010). These programs do not accommodate recombination. Recombination could potentially reduce the support for species relationships by introducing additional uncertainty in the estimated gene trees (note that all phylogenetic methods for estimating gene trees would be similarly affected by violating the assumption of no recombination).

Several short exploratory runs were conducted in *BEAST to fine-tune the parameter-specific settings for MCMC search. The species tree was estimated from two separate MCMC analyses which were run for 2×10^8 generations with parameters sampled every 1000 steps (discarding a burn-in of $1{\cdot}2\times10^5$ generations). Independent runs were combined using the program LogCombiner v.1.4.6 (Drummond and Rambaut, 2007). The program Tracer v1.5 (Rambaut and Drummond, 2009) was then used to determine if individual chains mixed well and the analyses had converged by graphing the trace plots of multiple MCMC chains started from random starting positions. Effective sample size of each parameter exceeded 200, except for the 28S alpha parameter where it was fluctuating at around 140. For each independent run, posterior probabilities for each node were compared to further ensure convergence. Tree topologies were visualized using the programs TreeAnnotator v.1.5.4 (Drummond and Rambaut, 2007) and FigTree v1.3.1 (Rambaut, 2009).

RESULTS

Estimates of the gene trees for the four nuclear loci (Fig. 4) show that among the 21 species in the *pinnatus*-group of feather mites there is considerable discord across the gene trees. Much of the discord is concentrated among the short branches separating the deeper divergence events as opposed to those separating the terminal branches. The relationships among sister taxa/clades are generally consistent across the individual gene trees. The primary

https://doi.org/10.1017/S003118201100031X Published online by Cambridge University Press

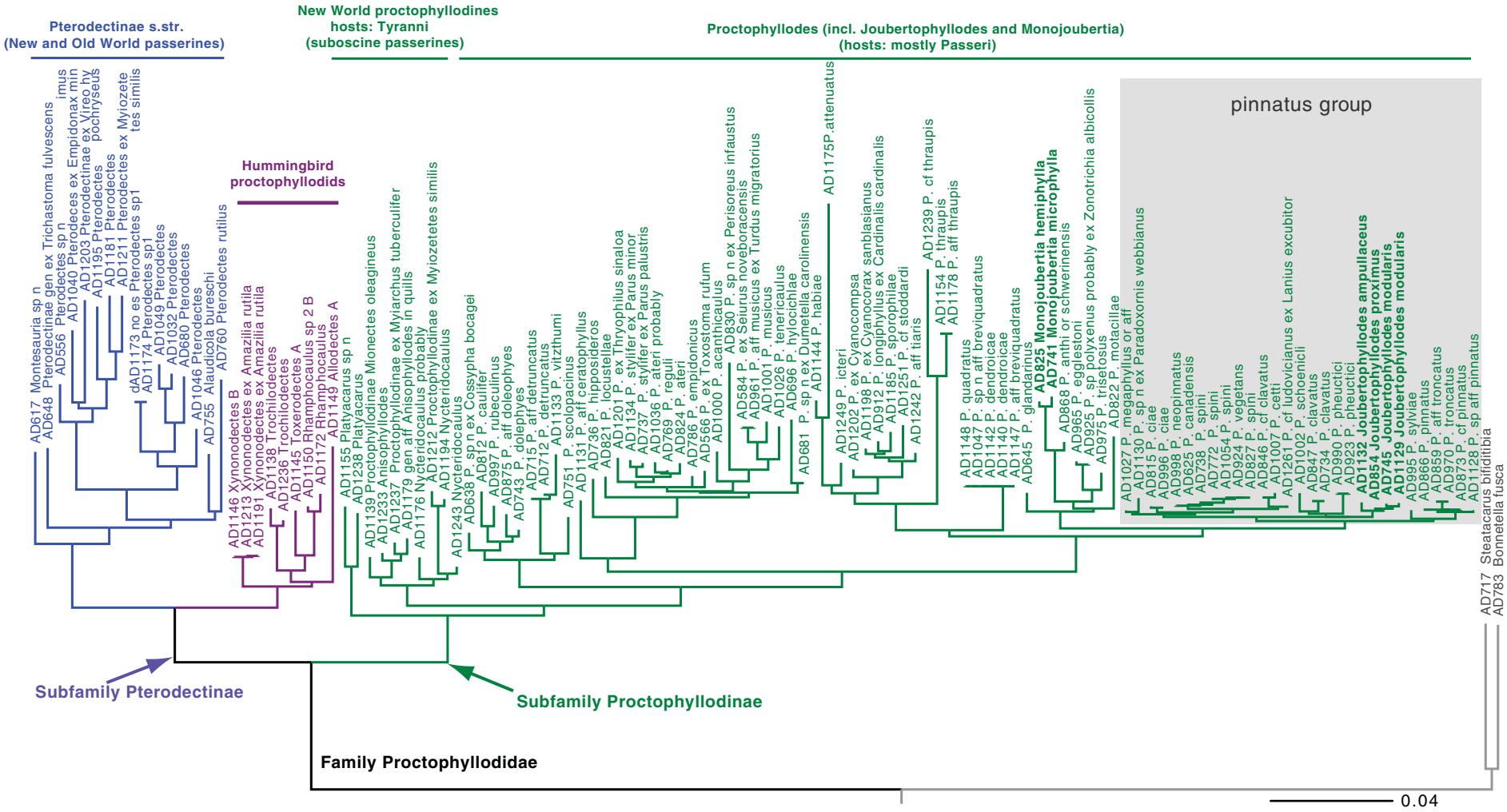

Fig. 3. Schematic showing the phylogenetic position of the *Proctophyllodes pinnatus*-group within the subfamily of the Proctophyllodinae (based on analyses of 18S, 28S, EF1-α, SRP54, and HSP70; Klimov and OConnor, unpublished). Note the relatively short internodes separating species within the *pinnatus* species group that suggests species-tree methods are particularly relevant to resolving phylogenetic relationships because of expected gene tree discord (reviewed in Knowles and Kubatko, 2010).

https://doi.org/10.1017/S003118201100031X Published online by Cambridge University Press

https://doi.org/10.1017/S003118201100031X Published online by Cambridge University Press

Table 1. Taxonomic sampling and GenBank reference numbers of the *Proctophyllodes pinnatus*-group and outgroups

					GenBank			
id	Species	Country	Host	Museum voucher	18S	28S	EF1-α	HSP70
645	*P. glandarinus*	USA	*Bombycilla cedrorum*	BMOC 06-0602-001	HM165035	HM165065	HM165095	HM165125
925	*P. polyxenus*	USA	*Zonotrichia albicollis*	BMOC 07-0914-003	HM165036	HM165066	HM165096	HM165126
965	*P. egglestoni*	USA	*Agelaius phoeniceus*	BMOC 07-0423-004	HM165037	HM165067	HM165097	HM165127
745	*J. modularis*	Russia	*Prunella modularis*	BMOC 06-1119-094	HM165038	HM165068	HM165098	HM165128
1129	*J. modularis*	Russia	*Emberiza spodocephala*	BMOC 08-0608-002	HM165039	HM165069	HM165099	HM165129
854	*J. proximus*	Kazakhstan	*Emberiza schoeniclus*	BMOC 07-1015-037	HM165040	HM165070	HM165100	HM165130
1132	*J. ampullaceus*	Russia	*Emberiza aureola*	BMOC 08-0608-005	HM165041	HM165071	HM165101	HM165131
1027	*P. megaphyllus*	Kazakhstan	*Prunella atrogularis*	BMOC 07-1119-088	HM165042	HM165072	HM165102	HM165132
1130	*P.* sp. n.	Russia	*Paradoxornis webbianus*	BMOC 08-0608-003	HM165043	HM165073	HM165103	HM165133
846	*P. cf. clavatus*	Russia	*Acrocephalus schoenobaenus*	BMOC 06-1119-068	HM165044	HM165074	HM165104	HM165134
847	*P. clavatus*	Kazakhstan	*Sylvia curruca*	BMOC 07-1119-005	HM165045	HM165075	HM165105	HM165135
734	*P. clavatus*	Russia	*Sylvia borin*	BMOC 06-1119-001	HM165046	HM165076	HM165106	HM165136
1007	*P. cetti*	Kazakhstan	*Cettia cetti*	BMOC 07-1119-041	HM165047	HM165077	HM165107	HM165137
995	*P. sylviae*	Russia	*Sylvia atricapilla*	BMOC 06-1119-011	HM165048	HM165078	HM165108	HM165138
924	*P. vegetans*	USA	*Carpodacus mexicanus*	BMOC 07-0921-001	HM165049	HM165079	HM165109	HM165139
738	*P. spini*	Russia	*Carduelis spinus*	BMOC 06-1119-009	HM165050	HM165080	HM165110	HM165140
772	*P. spini*	USA	*Carduelis tristis*	BMOC 06-1125-001	HM165051	HM165081	HM165111	HM165141
827	*P. spini*	USA	*Carduelis pinus*	BMOC 07-1121-003	HM165052	HM165082	HM165112	HM165142
1054	*P. spini*	USA	*Carduelis pinus*	BMOC 07-1121-003	HM165053	HM165083	HM165113	HM165143
625	*P. canadensis*	USA	*Sitta canadensis*	BMOC 06-0612-021	HM165054	HM165084	HM165114	HM165144
859	*P. aff. troncatus*	Kazakhstan	*Passer hispaniolensis*	BMOC 07-1119-023	HM165055	HM165085	HM165115	HM165145
970	*P. troncatus*	USA	*Passer domesticus*	BMOC 07-0409-001	HM165056	HM165086	HM165116	HM165146
866	*P. pinnatus*	Kazakhstan	*Carduelis chloris*	BMOC 07-1015-086	HM165057	HM165087	HM165117	HM165147
873	*P. cf. pinnatus*	Kazakhstan	*Carduelis cannabina*	BMOC 07-1015-107	HM165058	HM165088	HM165118	HM165148
1128	*P. aff. pinnatus*	Russia	*Carduelis sinica*	BMOC 08-0608-001	HM165059	HM165089	HM165119	HM165149
815	*P. ciae*	Kazakhstan	*Emberiza leucocephalos*	BMOC 07-1015-070	HM165060	HM165090	HM165120	HM165150
996	*P. ciae*	Russia	*Emberiza citrinella*	BMOC 06-1119-014	HM165061	HM165091	HM165121	HM165151
998	*P. neopinnatus*	Russia	*Loxia curvirostra*	BMOC 06-1119-039	HM165062	HM165092	HM165122	HM165152
1261	*P. cf. ludovicianus*	USA	*Lanius excubitor*	BMOC 08-0124-001	HM165063	HM165093	HM165123	HM165153
1002	*P. schoenicli*	Russia	*Emberiza schoeniclus*	BMOC 06-1119-255	HM165064	HM165094	HM165124	HM165154

P. – *Proctophyllodes*; J. – *Joubertophyllodes* (= *Proctophyllodes*, synonymy not yet published); *Proctophyllodes glandarinus, P. polyxenus*, and *P. egglestoni* are outgroups. The following multiple samples from specific species were included in the study: 745, 1129 (*J. modularis*); 847, 734 (*P. clavatus*); 738, 772, 827, 1054 (*P. spini*); 815, 996 (*P. ciae*).

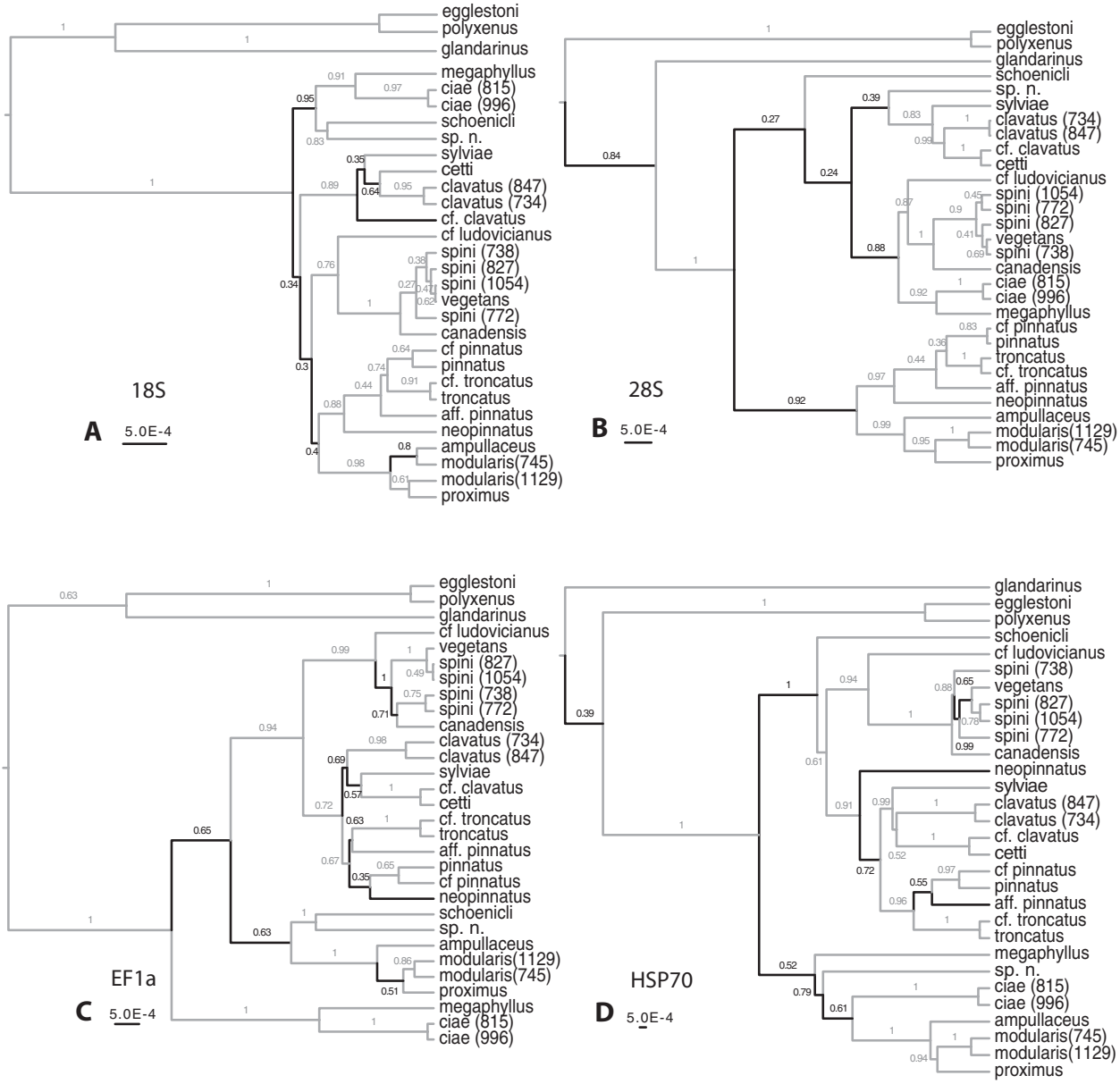

Fig. 4. Estimated gene trees for the nuclear loci: (A) 18S, (B) 28S, (C) EF1-α, and (D) HSP70. Branches that are not congruent with the species tree are shown in black, whereas congruent branches across the independent loci are shown in grey. Posterior probabilities are shown for each node.

exception occurs with the taxa: *P. neopinnatus, P. cf clavatus*, and *P. aff. pinnatus*.

Many of the branches are resolved in the Bayesian estimate of the species tree with relatively strong support (Fig. 5). This includes cases where a clade has relatively strong support in the species tree, even though within any of the individual gene trees (except for the 28S gene tree; Fig. 4) the support is very low (e.g. the clade of *P.* cf. *ludovicianus, P. canadensis, P. spini*, and *P. vegetans*). However, the deeper nodes remain ambiguous.

Despite the low support among the some of earliest splits within the *pinnatus*-group of feather mites, there are some interesting relationships that are hypothesized. For example, the species tree places the 5 taxa, *P. megaphyllus, P. ciae, P. ampullaceus, P. modularis*, and *P. proximus*, in a clade. Yet in the individual gene trees, these taxa occur in different clades, and in some cases distantly related clades (e.g. the gene trees of 28S and 18S, Fig. 4). Consideration of certain morphological apomorphies pertaining to the male genitalia in the *pinnatus*-group (Atyeo and Braasch, 1966; Badek *et al*. 2008), revealed that they are also shared with *P. megaphyllus, P. ciae, P. ampullaceus* thus justifying the placement of these three taxa in the *pinnatus*-group rather than previous assignment to a separate genus, *Joubertophyllodes*.

DISCUSSION

Confronted with the question of how to analyze data when gene trees estimated from independent loci often differ, the increased availability of multi-locus data (ironically) leaves empiricists in a state of limbo

https://doi.org/10.1017/S003118201100031X Published online by Cambridge University Press

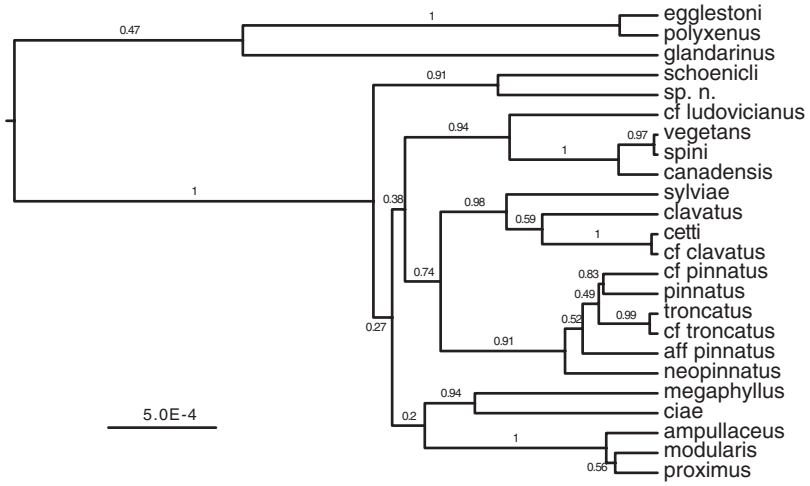

Fig. 5. Bayesian estimate of the species tree for the *Proctophyllodes pinnatus*-group with posterior probability values for species relationships shown.

as the awareness and popularity of species-tree approaches grows (e.g. Carstens and Knowles, 2007; Edwards *et al*. 2007; Belfiore *et al*. 2008; Brumfield *et al*. 2008; Joly *et al*. 2009; Kubatko and Gibbs, 2010; Linnen, 2010; McCormack *et al*. 2011; Oneal *et al*. 2010). Species-tree approaches clearly can provide a powerful framework for estimating phylogenetic relationships (reviewed in Knowles, 2009; Liu *et al*. 2009; Knowles and Kubatko, 2010). Moreover, by incorporating a model of the coalescent process (along with a model of the nucleotide substitution process), these new methods can provide accurate phylogenetic estimates despite widespread incomplete lineage sorting and discordance among the gene trees from independent loci (Maddison and Knowles, 2006; Eckert and Carstens, 2008; McCormack *et al*. 2009; Mossel and Roch, 2010). However, the species relationships may not all be resolved simply by applying a species tree approach, as in the case of the *pinnatus*-group of feather mites considered here (Fig. 5). Does this mean that we should forego species tree approaches and opt for the simpler approach of concatenating the data from discordant loci (i.e. combining the sequences across loci into a single nucleotide matrix for analysis)?

There are actually a number of reasons not to concatenate the data. First, concatenation itself will not necessarily lead to better-resolved species relationships. When data are concatenated across independent loci with differing gene trees, there is no way to know whether the estimated tree matches the underlying species tree (Maddison, 1997; Maddison and Knowles, 2006). The addition of nucleotide sequences to concatenated data may actually be positively misleading (i.e. the estimated tree will not match the true species tree even with the addition of concatenated data – Kubatko and Degnan, 2007), and higher support for such trees may simply be an artifact of constraining the data to fit a single tree when in reality there is a mixture of trees (see Mossel and Vigoda, 2005; Cranston *et al*. 2009).

The species tree estimated for the *pinnatus*-group of feather mites could definitely be improved upon. Among the well-resolved nodes within clades is ambiguity among the relationships of the clades (Fig. 5). Although less than ideal, there are several noteworthy aspects about the resolved nodes, as well as the unresolved nodes (which are discussed in further detail below). First, without the species-tree analysis, it is not possible to infer clades within the diverse species group based on estimates of the individual gene trees. Not only do the hypothesized clades differ across loci, but the constituent members of putative clades also varies among the gene trees. For example, consider the clade identified in the species tree comprised of the species *P. megaphyllus, P. ciae, P. ampullaceus, P. modularis*, and *P. proximus*. This clade is not consistently identified in the individual gene trees. Yet, this clade estimated in the species tree is corroborated by a morphological character shared with (and unique to) the species in the *pinnatus*-group – shared derived genitalic characters in the males. This phylogenetic estimate suggests that other characters associated with mating in this clade of taxa (i.e. the enlargement of posterior legs in males) that had been used to place *P. ampullaceus, P. modularis*, and *P. proximus* in a different genus, *Joubertophyllodes,* is an example of rapid morphological evolution of characters misinterpreted under a traditional taxonomic paradigm as evidence of clade distinct from the *pinnatus*-group. The species-tree estimate based on the molecular data highlights an alternative set of morphological characters which is diagnostic to the *pinnatus*-group (including "*Joubertophyllodes*"), proposing a classification of these mites based on their phylogenetic affinities rather than on a divergent, but admittedly arbitrary, character. Second, the species tree has fairly strong support for many of the relationships (see below for a discussion

https://doi.org/10.1017/S003118201100031X Published online by Cambridge University Press

of the nodes without strong support). Notwithstanding the theoretical problems associated with concatenating data across discordant gene trees discussed above, from an empirical perspective, the phylogenetic relationships estimated in the species tree (Fig. 5) are a step forward in trying to resolve the enigmatic phylogeny of this diverse group of feather mites and clearly demonstrated the need to revise the current taxonomic classification to include *Jouberto-phyllodes* into the *pinnatus*-group of the genus *Proctophyllodes*.

What underlies the poor resolution among some branches in the species tree?. The unresolved branches in the species tree could reflect characteristics of the history of diversification itself, aspects of the data-set used to estimate the phylogeny, or an interaction of these two properties. We discuss the likelihood of each of these in detail and what each would imply about possible strategies to improve the phylogenetic estimate.

Phylogenetic estimation is difficult because of the diversification history. The history of diversification itself might be particularly difficult if the rate of speciation in the *pinnatus*-group of feather mites is high (i.e. if the group has undergone rapid diversification). This is suggested by the large number of taxa (i.e. it is among the most diverse species groups in the genus *Proctophyllodes* – Badek *et al*. 2008). Relative to other clades in the genus, the internodes separating the species also appear to be short (Fig. 3) (although aspects of the mutational processes can not be ruled out as discussed below). These patterns suggest the history of diversification itself in the *pinnatus*-group might contribute to the poor phylogenetic resolution. This is because with rapid diversification there is insufficient time for the sorting of ancestral gene lineages (i.e. the fixation of a gene lineage within a species lineage by genetic drift) before subsequent speciation events, which gives rise to widespread discord across the gene trees of independent loci (Maddison, 1997; Degnan and Rosenberg, 2009; Knowles, 2009).

Simulation studies show that, all else being equal (i.e. similar levels of molecular divergence across species for a given number of taxa), diversification histories characterized by more gene tree discord will require sequence data from more independent loci to obtain accurate phylogenetic estimates (e.g. Maddison and Knowles, 2006; Knowles and Chan, 2008; McCormack *et al*. 2009). Consequently, one possible solution to resolving the enigmatic relationships within the *pinnatus*-group of feather mites is to simply collect more data.

What type of data to collect is actually more nuanced than the choices when simply concatenating data and will depend very much on the history of diversification. For example, for very recently diverged taxa, collecting sequence data from more individuals can actually increase the accuracy of species tree estimates (Maddison and Knowles, 2006; McCormack *et al*. 2009). However, if the species exhibit reciprocally monophyletic gene trees, there is no additional phylogenetic genetic information contained in the sequences of multiple individuals. Instead, increasing the number of loci sampled will improve phylogenetic accuracy. Note that if a preponderance of deep coalescence (Fig. 1) has generated gene tree discord among loci, the key is to collect data from more loci, not simply more base-pairs for the sampled loci included in the study. This is because with histories characterized by widespread gene tree discord, only with the sampling of additional loci can the relative probabilities of different gene trees become apparent, and thereby provide the relevant information for distinguishing among alternative species trees (Maddison and Knowles, 2006). For example, different gene trees (i.e. different topologies) might be possible under different species trees. However, the relative probabilities of the gene trees (and hence the frequency that a certain topology is observed across loci) will differ across alternative species trees (Degnan and Salter, 2005). In other words, because the species tree itself specifies the probability of observing different gene trees (Takahata, 1989; Degnan and Salter, 2005), the distribution of observed gene trees contains information about the actual phylogenetic history (i.e. the underlying species tree).

Aspects of the collected data make phylogenetic estimation difficult. In addition to the contribution of deep coalescence to gene tree discord (Fig. 1), aspects of the data may make phylogenetic estimation difficult. Limited amounts of sequence variation could lead to poorly resolved gene trees. Likewise, the estimated gene trees may be compromised by errors in their estimation. With regards to the unresolved branches in the estimated species tree for the feather mites, these explanations seem generally unlikely. First, many of the nodes in the species tree do indeed have strong support (Fig. 5). This includes branches separating recently derived species. If limited sequence variation was causing problems, it is exactly in the cases of recent speciation events where there has not been sufficient time to accumulate mutations that one would expect to see poor resolution of species relationships. However, this is not the case in the estimated species tree of the *pinnatus*-group. This argues against errors in the estimation of the gene trees as a general explanation for the unresolved nodes in the estimated species tree for the *pinnatus*-group. Moreover, the Bayesian approach used to estimate the species tree in our study also accounts for uncertainty in the estimated gene trees.

https://doi.org/10.1017/S003118201100031X Published online by Cambridge University Press

Even if gene trees are estimated accurately, the estimated gene tree may not actually match the underlying genealogical history of loci. What is observed in empirical data-sets are estimated gene trees – that is, the product of the underlying genealogical history of gene lineages and the mutational process that provides the information for estimating that genealogical history. This mutational variance can cause difficulties with obtaining accurate estimates of species trees (see Huang and Knowles, 2009; Huang *et al*. 2010) because the probabilities provided by the coalescent models used in species tree approaches are based on the unobservable quantity – the actual underlying gene genealogy – rather than the data input of these methods – the estimated gene tree. This leaves open the possibility that a difference between the underlying gene genealogy and the estimated gene tree might affect the accuracy of species tree estimates (Huang *et al*. 2010). As with the nuanced effects of sampling design on the accuracy of species tree estimates (e.g. McCormack *et al*. 2009), the relative contribution of mutational variance versus coalescent variance (i.e. the discord across loci caused by the deep coalescence of gene lineages; Fig. 1) differs depending on the underlying history of species diversification, the sampling design and the total sampling effort (Huang and Knowles, 2009; Huang *et al*. 2010). Although increasing the number of loci will increase the accuracy of species tree estimates (Knowles, 2010), there is also a greater contribution of mutational variance (i.e. the mismatch between the estimated gene tree and the underlying gene genealogy) relative to coalescent variance (i.e. gene tree discord caused by the deep coalescence of gene lineages) with increased sampling of loci. One possible (but as yet unexplored) solution would be to identify loci with higher rates of evolution. This would in principle increase the probability that branches in the genealogy experience mutations, so it is less likely the estimated gene tree would differ from the underlying gene genealogy.

CONCLUSIONS

There are many theoretical reasons that species tree approaches should increase the accuracy of phylogenetic estimates. The benefits of adopting a species tree approach, as exemplified in this study, can also be realized in practice. Many of the clades within the diverse *pinnatus*-group of feather mites in the estimated species tree are well supported, despite the discord in the constituent gene trees across loci (i.e. comparing Fig. 4 and 5). The lack of resolution among some of the branches in this empirical study (most notably among the short internodes separating the clades) is consistent with the expected phylogenetic difficulties when species have diversified rapidly. Luckily, species tree approaches not only provide a means for estimating such recalcitrant phylogenies, but simulation studies offer helpful guides for making decisions about how to effectively deal with such difficult historical scenarios. Given that interactions between host and parasite can accelerate speciation rates, species tree approaches should be especially useful for estimating the phylogenetic histories of such groups.

ACKNOWLEDGEMENTS

We thank Dr. S. Mironov (Zoological Institute, Russian Academy of Sciences, Saint Petersburg) for taxonomic discussion and Barry OConnor for sharing his photo of mites of the *pinnatus*-group. This work was supported by grants from the National Science Foundation (DEB-0918218 to L. Lacey Knowles and DEB-0613769 to Barry OConnor) and the Russian Ministry of Education and Science (02.740.11.5139) to Pavel B. Klimov. The molecular work of this study was conducted in the Genomic Diversity Laboratory of the University of Michigan Museum of Zoology.

REFERENCES

Atyeo, W. T. and Braasch, N. L. (1966). The feather mite genus *Proctophyllodes* (Sarcoptiformes: Proctophyllodidae). *Bulletin of the University of Nebraska State Museum* **5**, 1–354.

Badek, A., Dabert, M., Mironov, S. V. and Dabert, J. (2008). A new species of the genus *Proctophyllodes* (Analgoidea: Proctophyllodidae) from Cetti's warbler *Cettia cetti* (Passeriformes: Sylviidae) with DNA Barcode Data. *Annales Zoologici* **58**, 397–402.

Belfiore, N. M., Liu, L. and Moritz, C. (2008). Multilocus phylogenetics of a rapid radiation in the genus *Thomomys* (Rodentia: Geomyidae). *Systematic Biology* **57**, 294–310.

Blanco, G., Seoane, J. and de la Puente, J. (1999). Showiness, nonparasitic symbionts, and nutritional condition in a passerine bird. *Annales Zoologici Fennici* **36**, 83–91.

Brumfield, R. T., Liu, L., Lum, D. E. and Edwards, S. V. (2008). Comparison of species tree methods for reconstructing the phylogeny of bearded manakins (Aves: Pipridae, *Manacus*) from multilocus sequence data. *Systematic Biology* **57**, 719–731.

Cannone, J., Subramanian, S., Schnare, M., Collett, J., D'Souza, L., Du, Y., Feng, B., Lin, N., Madabusi, L., Muller, K., Pande, N., Shang, Z., Yu, N. and Gutell, R. (2002). The Comparative RNA Web (CRW) Site: an online database of comparative sequence and structure information for ribosomal, intron, and other RNAs. *BMC Bioinformatics* **3**, 2.

Carstens, B. C. and Knowles, L. L. (2007). Estimating species phylogeny from gene-tree probabilities despite incomplete lineage sorting: an example from *Melanoplus* grasshoppers. *Systematic Biology* **56**, 400–411.

Cranston, K. A., Hurwitz, B., Ware, D., Stein, L. and Wing, R. A. (2009). Species trees from highly incongruent gene trees in rice. *Systematic Biology* **58**, 489–500.

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

Degnan, J. H. and Salter, L. A. (2005). Gene tree distributions under the coalescent process. *Evolution* **59**, 24–37.

Drummond, A. J. and Rambaut, A. (2007). BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evolutionary Biology* **7**, 214.

Eckert, A. J. and Carstens, B. C. (2008). Does gene flow destroy phylogenetic signal? The performance of three methods for estimating species phylogenies in the presence of gene flow. *Molecular Phylogenetics and Evolution* **49**, 832–842.

Edwards, S. V., Liu, L. and Pearl, D. K. (2007). High-resolution species trees without concatenation. *Proceedings of the National Academy of Sciences, USA* **104**, 5936–5941.

Felsenstein, J. (2004). *Inferring Phylogenies*, Sinauer Associates, Sunderland, MA, USA.

Gillespie, J. J., Johnston, J. S., Cannone, J. J. and Gutell, R. R. (2006). Characteristics of the nuclear (18S, 5·8S, 28S and 5S) and mitochondrial

https://doi.org/10.1017/S003118201100031X Published online by Cambridge University Press

(12S and 16S) rRNA genes of *Apis mellifera* (Insecta: Hymenoptera): structure, organization, and retrotransposable elements. *Insect Molecular Biology* **15**, 657–686.

Hartup, B. K., Stott-Messick, B., Guzy, M. and Ley, D. H. (2004). Health survey of house finches (*Carpodacus mexicanus*) from Wisconsin. *Avian Diseases* **48**, 84–90.

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

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

Huang, H. and Knowles, L. L. (2009). What is the danger of the anomaly zone for empirical phylogenetics? *Systematic Biology* **58**, 527–536.

Joly, S., McLenachan, P. A. and Lockhart, P. J. (2009). A statistical approach for distinguishing hybridization and incomplete lineage sorting. *American Naturalist* **174**, E54–E70.

Klimov, P. B. and OConnor, B. M. (2008). Origin and higher-level relationships of psoroptidian mites (Acari: Astigmata: Psoroptidia): evidence from three nuclear genes. *Molecular Phylogenetics and Evolution* **47**, 1135–1156.

Knowles, L. L. (2009). Estimating species trees: methods of phylogenetic analysis when there is incongruence across genes. *Systematic Biology* **58**, 463–467.

Knowles, L. L. (2010). Sampling strategies for species-tree estimation. In *Estimating Species Trees: Practical and Theoretical Aspects* (ed. Knowles, L. L. and Kubatko, L. S.), Wiley-Blackwell, Hoboken, NJ, USA.

Knowles, L. L. and Chan, Y.-H. (2008). Resolving species phylogenies of recent evolutionary radiations. *Annals of the Missouri Botanical Garden* **95**, 224–231.

Knowles, L. L. and Kubatko, L. S. (2010). Estimating species trees: an introduction to concepts and models. In *Estimating Species Trees: Practical and Theoretical Aspects* (ed. Knowles, L. L. and Kubatko, L. S.), pp. 1–12. Wiley-Blackwell, Hoboken, NJ, USA.

Kubatko, L. and Degnan, J. (2007). Inconsistency of phylogenetic estimates from concatenated data under coalescence. *Systematic Biology* **56**, 17–24.

Kubatko, L. S. and Gibbs, H. L. (2010). Estimating species relationships and taxon distinctiveness in *Sistrurus* rattlesnakes using multi-locus data. In *Estimating Species Trees: Practical and Theoretical Aspects* (ed. Knowles, L. L. and Kubatko, L. S.), pp. 193–206. Wiley-Blackwell, Hoboken, NJ, USA.

Linnen, C. (2010). Species-tree estimation for complex divergence Histories: A case study in *Neodiprion* sawflies. In *Estimating Species Trees: Practical and Theoretical Aspects* (ed. Knowles, L. L. and Kubatko, L. S.), pp. 145–192. Wiley-Blackwell, Hoboken, NJ, USA.

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

Liu, L., Yu, L., Kubatko, L., Pearl, D. K. and Edwards, S. V. (2009). Coalescent methods for estimating phylogenetic trees. *Molecular Phylogenetics and Evolution* **53**, 320–328.

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

Maddison, W. P. and Knowles, L. L. (2006). Inferring phylogeny despite incomplete lineage sorting. *Systematic Biology* **55**, 21–30.

Mathews, D. H., Sabina, J., Zuker, M. and Turner, D. H. (1999). Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *Journal of Molecular Biology* **288**, 911–940.

McCormack, J. E., Heled, J., Delaney, K. S., Peterson, A. T. and Knowles, L. L. (2010). Calibrating divergence times on species trees versus gene trees: Implications for speciation history of *Aphelocoma* jays. *Evolution* **65**, 184–202.

McCormack, J. E., Huang, H. and Knowles, L. L. (2009). Maximum likelihood estimates of species trees: how accuracy of phylogenetic inference depends upon the divergence history and sampling design. *Systematic Biology* **58**, 501–508.

Mossel, E. and Roch, S. (2010). Incomplete lineage sorting: consistent phylogeny estimation from multiple loci. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* **7**, 166–171.

Mossel, E. and Vigoda, E. (2005). Phylogenetic MCMC algorithms are misleading on mixtures of trees. *Science* **309**, 2207–2209.

Oneal, E., Otte, D. and Knowles, L. L. (2010). Testing for biogeographic mechanisms promoting divergence in Caribbean crickets (genus *Amphiacusta*). *Journal of Biogeography* **37**, 530–540.

Posada, D. and Buckley, T. R. (2004). Model selection and model averaging in phylogenetics: advantages of Akaike information criterion and Bayesian approaches over likelihood ratio tests. *Systematic Biology* **53**, 793–808.

Rambaut, A. (2009). FigTree. Available online at *http://tree.bio.ed.ac.uk/software/figtree/*.

Rambaut, A. and Drummond, A. J. (2009). Tracer v1.5. Available from *http://beast.bio.ed.ac.uk/Tracer*.

Takahata, N. (1989). Gene genealogy in three related populations: consistency probability between gene and population trees. *Genetics* **122**, 957–966.

Zuker, M. (2003). Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Research* **31**, 3406–3415.

https://doi.org/10.1017/S003118201100031X Published online by Cambridge University Press