Eurasia Specialized Veterinary Publication

Entomological Research Letters

ISSN: 3108-639X

2024, Volume 4, Issue 1, Page No: 146-155
Copyright CC BY-NC-SA 4.0
Available online at: www.esvpub.com/

Gene-Tree Conflict, Missing Data, and Model Misspecification Demand Explicit Uncertainty Rather Than False Certainty in Insect Phylogenomics

Chen Hao[1]*, Liu Fang[1], Zhao Lin[2], Wei Zhang[3]

[1]Department of Gene-Tree Conflict and Missing Data, Faculty of Life Sciences, Zhejiang University, Hangzhou, China.
[2]Department of Model Misspecification and Uncertainty, Faculty of Biosciences, Nanjing University, Nanjing, China.
[3]Department of Phylogenomic Inference and Explicit Uncertainty, Faculty of Sciences, University of Melbourne, Melbourne, Australia.

***E-mail** ✉ chen.hao@zju.edu.cn

ABSTRACT

Phylogenomic datasets can produce apparently decisive insect phylogenies while retaining substantial uncertainty about the histories represented by individual loci, the adequacy of sampled data, and the assumptions linking gene trees to species trees. This article develops a framework for distinguishing uncertainty generated by biological gene-tree heterogeneity, missing taxa or loci, orthology error, model misspecification, and downstream species-tree inference. The central argument is that conventional branch support cannot function as a universal summary of these distinct uncertainties. Gene-tree discordance may represent incomplete lineage sorting, introgression, duplication and loss, or analytical error; missingness may interact with taxon representation and phylogenetic information rather than operate as a simple percentage threshold; and apparent orthology can be disrupted by differential gene loss without making all multicopy data unusable. Model adequacy introduces a further boundary between strongly supported inference and adequately represented evolutionary processes. The proposed uncertainty-source architecture therefore separates the origin of uncertainty, its observable manifestation, the inference stage it can affect, and the degree to which it remains unresolved after species-tree estimation. The framework does not assign numerical certainty grades or identify a single correct topology. Instead, it establishes a basis for reporting when confidence is localized to a statistical procedure while biological or analytical alternatives remain unresolved. Explicitly preserving those alternatives should make insect phylogenomic conclusions more interpretable, reproducible, and appropriately conditional.

Keywords: Gene-tree discordance, Species-tree inference, Orthology, Missing data, Model adequacy, Concordance factors

Received: 18 May 2024
Revised: 20 June 2024
Accepted: 22 June 2024

How to Cite This Article: Hao C, Fang L, Lin Z, Zhang W. Gene-Tree Conflict, Missing Data, and Model Misspecification Demand Explicit Uncertainty Rather Than False Certainty in Insect Phylogenomics. Entomol Res Lett. 2024;4(1):146-55. https://doi.org/10.51847/4z2DKQ8g8a

Introduction

Genome-scale sequencing has transformed insect systematics without eliminating disagreement among inferred histories. Phylogenomic incongruence can originate from biological processes, including incomplete lineage sorting, introgression, recombination, and gene duplication and loss, but also from contamination, orthology mistakes, gene-tree estimation error, and inadequate evolutionary models [1]. These causes operate at different analytical levels. A discordant locus may therefore be an expected genealogical realization, an observation affected by limited information, or an artefact of data construction. Treating all three outcomes simply as “conflict” obscures the evidential distinction that matters for species-tree interpretation.

© 2024 Entomological Research Letters

The same problem applies to support. Bootstrap proportions and posterior probabilities quantify particular forms of uncertainty conditional on an analysis, but increasing sequence length can make these values approach apparent decisiveness without demonstrating that model assumptions are adequate or that genomic histories are concordant [2]. Recent insect phylogenomics makes this distinction concrete. Whole-genome analyses of *Neodiprion* recovered a strongly supported species tree while documenting extensive genomic discordance associated with variation in phylogenetic information, missingness, GC content, recombination, and other genomic properties [3]. New Zealand cicadas likewise contained strongly supported relationships with substantial gene- and site-level conflict [4], whereas rapid radiation in *Parnassius* butterflies produced extensive discordance interpreted primarily in relation to incomplete lineage sorting in that sampled system [5].

These studies do not imply that strong support is intrinsically misleading. They show instead that support for a fitted tree and uncertainty about the historical processes producing the data are not equivalent quantities. This article therefore proposes an uncertainty-source architecture for insect phylogenomics. Its purpose is not to choose among competing histories automatically, but to preserve distinctions between observed conflict, plausible cause, analytical sensitivity, species-tree consequence, and residual uncertainty. The resulting framework treats certainty as stage-specific rather than as a property that can be compressed safely into a single branch value.

Materials and Methods

This article uses a comparative framework rather than a new phylogenomic dataset or statistical analysis. Published evidence was organized according to four questions: where an uncertainty source enters the analytical chain; what observable pattern can reveal it; which downstream inference it can affect; and whether available evidence permits attribution to a specific cause. The architecture deliberately separates data-integrity uncertainty, genealogical heterogeneity, sampling and locus-definition uncertainty, orthology uncertainty, model inadequacy, and uncertainty introduced when estimated gene trees are summarized into species trees. These classes are analytical categories proposed here, not empirically calibrated error grades.

A central methodological rule is that similar topological conflict can have different causes. Comparative work on phylogenomic discordance demonstrates that incomplete lineage sorting, reticulation, and analytical error may require multiple diagnostics rather than a single decisive test [6]. The framework therefore treats causal assignment as an inference step that can remain unresolved. Failure to discriminate among alternatives is recorded as uncertainty rather than converted into a categorical historical explanation.

Technical provenance forms a separate upstream layer. In cicada phylogenomics, sample contamination altered inferred relationships and demonstrated that strongly supported topology does not by itself verify sample identity or sequence provenance [7]. Such errors differ conceptually from true genealogical discordance because they precede biological interpretation. Their inclusion in the architecture prevents downstream support from retrospectively legitimizing compromised input data.

Analytical specification is treated similarly. Adephagan beetle phylogenomics showed that model choice and excessive data trimming can alter deep relationships and generate phylogenetic artefacts [8]. Missingness, taxon sampling, and model choice can also interact rather than contribute independent effects, as demonstrated in comparative phylogenomic analyses outside insects [9]. Accordingly, the proposed architecture does not assign uncertainty sources to isolated boxes. It permits interactions among data composition, gene-tree estimation, and model fit while maintaining a boundary between evidence that a result is sensitive and evidence explaining why it is sensitive.

Results and Discussion

Gene-tree heterogeneity is the first uncertainty class for which disagreement can be biologically expected rather than interpreted as analytical failure. Deep relationships within Aphididae illustrate this problem: gene-tree discordance, introgression signals, and conditions associated with the gene-tree anomaly zone jointly complicated reconstruction of family-level history [10]. The relevant result is not that one mechanism universally explains aphid discordance, but that a species tree can summarize genomic histories produced by more than one evolutionary process.

Conflict can also be structured across the genome. Chromosome-aware analyses of assassin bugs found heterogeneous historical signal among genomic partitions, demonstrating that disagreement need not be

distributed uniformly among loci [11]. A branch supported by many loci overall may therefore contain localized genomic regions with different histories. Such structure becomes analytically important when interpreting whether conflict represents random estimation noise, linked evolutionary history, or another process requiring targeted testing.

Methodological treatment adds another layer. Transcriptomic analyses of Auchenorrhyncha recovered alternative, strongly supported relationships under different coding and analytical strategies while identifying compositional heterogeneity as relevant to some contentious nodes [12]. This establishes an important boundary: high support for alternative topologies under different reasonable analyses indicates analytical sensitivity, not simultaneous confirmation of incompatible histories. Neither support value alone identifies the source of disagreement.

Expected incomplete-lineage-sorting discordance can be distinguished statistically from patterns that depart from a simple multispecies-coalescent expectation. Quartet-based simplex approaches provide tests for whether observed gene-tree distributions are compatible with such expectations [13]. However, departure from the expectation does not uniquely diagnose introgression, contamination, hidden paralogy, or model error. In the proposed architecture, therefore, an excess-discordance signal is a diagnostic state rather than a historical conclusion. Gene-tree heterogeneity is resolved only when evidence can distinguish among plausible biological and analytical alternatives; otherwise its unresolved origin must propagate forward.

Missing taxa and missing loci

Missing data are often described as a scalar property of a matrix, but their inferential effect depends on what is missing and where the absence occurs. Expanded molecular and taxonomic representation in planthoppers altered the evidential context for deep Fulgoromorpha relationships, illustrating that incomplete taxon sampling can shape phylogenetic hypotheses independently of the nominal size of a sequence matrix [14]. Denser taxon representation is not a guarantee of correctness, yet the location of unsampled lineages can change long-branch structure, character distribution, and the comparisons available for discriminating alternatives.

Missing loci likewise should not be reduced to a universal completeness threshold. Sequence-length heterogeneity can affect tree estimation differently depending on alignment and placement strategy, with some methods tolerating fragmentary sequences better than others [15]. Thus, two matrices with comparable proportions of absent characters may carry very different phylogenetic information. The proposed architecture records missingness together with its distribution across taxa and loci, rather than treating matrix occupancy as a self-interpreting quality score.

A further complication is that a named molecular gene is not necessarily an independent genealogical unit. Coalescent reasoning depends on assumptions about recombination and the historical boundaries of independently inherited regions [16]. Missing-locus accounting therefore presupposes that the units being counted are meaningful for the inference model. This distinction is especially important when long loci span recombination histories or when organellar and nuclear data are analyzed under assumptions of comparable genealogical independence. The framework consequently separates absence of data from uncertainty about the definition of the locus itself.

Orthology error

Orthology inference introduces uncertainty before estimated gene trees can be interpreted as histories of the same evolutionary entity. A locus that appears as one copy per sampled species is not necessarily a true ortholog set. Differential loss after duplication can produce pseudoorthologs whose single-copy representation conceals paralogous history [17]. The relevant risk depends on duplication and loss history, however; single-copy status is therefore insufficient evidence of orthology but is not evidence that pseudoorthology must have occurred.

The converse boundary is equally important. Paralogy does not make species-tree inference uniformly unreliable. Simulations and empirical comparisons show that several incomplete-lineage-sorting-aware species-tree methods can remain robust to paralogs under tested conditions, even though gene-tree estimation error may still be substantial [18]. Analyses using broader gene families further demonstrate that multicopy data can increase usable genomic information without necessarily producing major topological disruption in high-quality genome datasets [19]. These findings prevent the framework from equating orthology uncertainty with mandatory exclusion.

Methodological responses consequently form a continuum. Multicopy gene-family trees can be decomposed before species-tree estimation, providing one strategy for retaining information otherwise removed by strict single-copy filtering [20]. Other methods explicitly incorporate paralogy into quartet-based species-tree

estimation [21]. Neither approach eliminates uncertainty in gene-family history, because errors in input gene trees, incomplete family sampling, or mismatched assumptions can remain.

Differential loss is particularly important because its effect is history-dependent. Modeling after whole-genome duplication shows that the location and timing of gene loss can determine whether pseudoorthology biases species-tree inference and whether increasing locus number ameliorates the problem [22]. The proposed framework therefore distinguishes orthology assignment uncertainty, multicopy accommodation, and gene-loss-induced pseudoorthology rather than treating them as one binary pass/fail condition.

Because orthology problems differ not only in origin but also in whether they should trigger exclusion, reanalysis, or explicit model accommodation, a compact comparison is needed to preserve those distinctions without inventing an orthology-quality score.

Table 1. Gene-family-history states that change the interpretation of apparent orthology in phylogenomic species-tree inference

Historical origin			Downstream propagation	Analytical response		
Gene-family history	**Observed genomic evidence**	**Duplication/loss context**	**Species-tree consequence**	**Analytical test**	**Residual uncertainty**	**Standing**
Latent duplication and loss						
Apparent single-copy orthology	Single-copy occupancy without demonstrated family-history congruence	One sampled copy per taxon, but ancestral duplication/loss history unresolved	Apparent orthology may conceal pseudoorthology	Compare family-tree history with species-tree expectations rather than treating copy count as proof	Risk depends on duplication and differential-loss history	Evidence-supported distinction
Pseudoortholog formation	Gene-family topology compatible with retention of different ancestral copies	Ancestral duplication followed by lineage-specific loss	A locus may support a history inconsistent with the intended orthologous comparison	Test duplication/loss scenarios and sensitivity to suspected families	Frequency and topology are parameter- and history-dependent	Evidence-supported mechanism
Retained multicopy families	Multicopy gene-family trees instead of forced single-copy ortholog sets	Multiple homologous copies remain in some taxa	Additional data can remain phylogenetically informative	Compare strict filtering with multicopy-aware inference	Robustness depends on gene-tree accuracy and inference method	Evidence-supported alternative
Decomposition and paralogy-aware analysis						
Decomposition before species-tree analysis	Decomposed homolog trees retaining broader gene-family information	Multicopy trees can be partitioned into usable subtrees	Avoids discarding all multicopy families while preserving a species-tree input structure	Evaluate sensitivity to decomposition and gene-tree uncertainty	Input-family-tree error can propagate through decomposition	Evidence-supported analytical strategy
Paralogy-aware species-tree inference	Multicopy quartets or homolog relationships analyzed directly	Duplication/loss represented within the inference procedure	Paralogy becomes a modeled source of gene-tree variation rather than an automatic exclusion criterion	Compare paralogy-aware and restricted-ortholog analyses	Model assumptions and gene-family estimation remain consequential	Evidence-supported analytical strategy
Unresolved provenance						
Unresolved orthology provenance	Conflicting family trees, incomplete copies, or unstable assignments	Available evidence cannot distinguish true orthology, pseudoorthology,	Species-tree support should not erase unresolved gene-family provenance	Preserve orthology provenance as downstream uncertainty	No categorical historical assignment is justified	Proposed reporting state

	or benign paralogy

The same evidence also implies that orthology error develops through time rather than appearing only at the final alignment stage. **Figure 1** therefore separates historical duplication and loss from later observation, orthology assignment, and species-tree processing, while retaining the possibility that multicopy histories can be accommodated rather than automatically discarded.

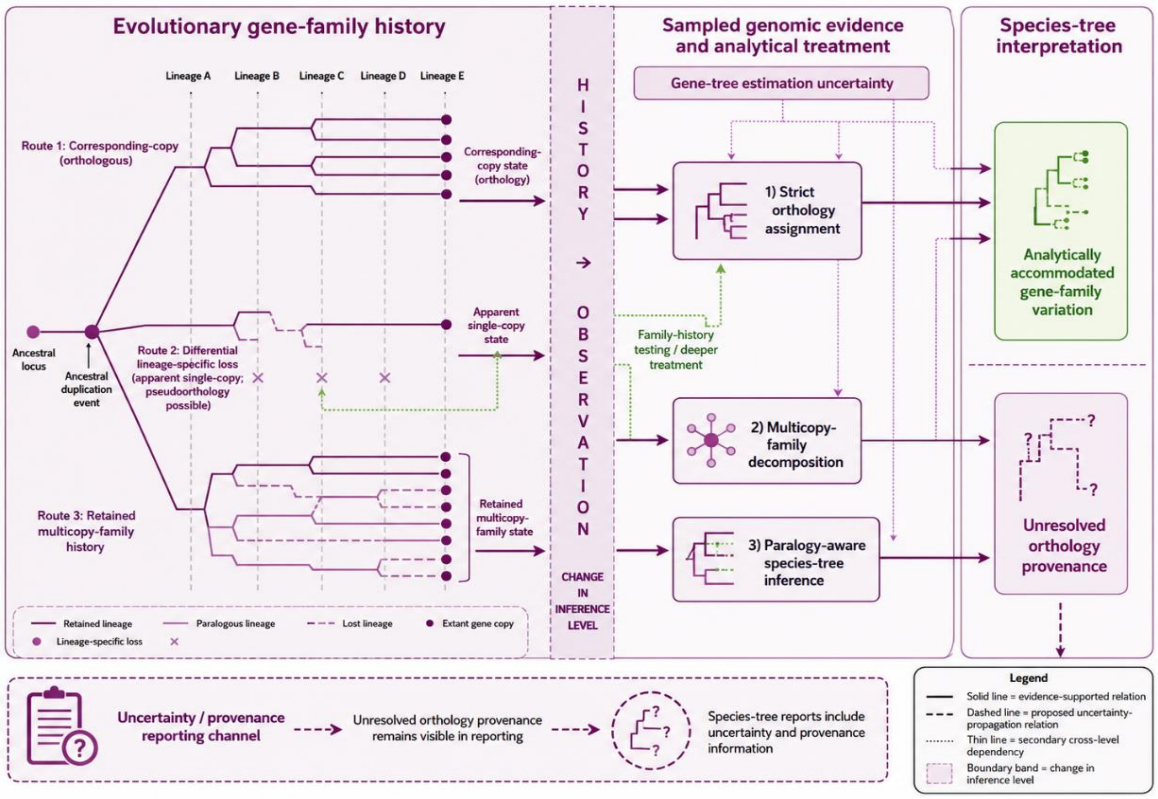


Figure 1. Historical gene-family divergence creates distinct routes from true orthology, pseudoorthology, and retained paralogy to species-tree uncertainty

Model misspecification

Even correctly sampled and correctly assigned loci can yield misleading inference when the evolutionary process represented by the model differs materially from the process generating the sequences. Heterotachy provides one example: evolutionary rates can change across lineages and sites in ways that violate simpler homogeneous assumptions. Mixture-model approaches such as GHOST demonstrate that historical signal can sometimes be recovered when heterotachous evolution is modeled explicitly [23]. This evidence does not establish that any particular mixture model is universally appropriate; it establishes model heterogeneity as an uncertainty source that cannot be diagnosed from branch support alone.

Substitution saturation creates a different failure mode. Repeated substitutions can reduce or distort historical information, particularly over deep timescales. Comparative analyses across phylogenomic datasets found that excluding loci identified as saturated improved inference in tested cases, while also showing that saturation is not uniformly distributed across loci or datasets [24]. Saturation filtering should therefore be interpreted as a data- and model-sensitive diagnostic, not as an automatic preprocessing rule.

More generally, statistical rejection of a model and demonstrated corruption of a particular topology are not identical statements. Comparative model testing across phylogenomic datasets showed frequent inadequacy of substitution or concatenation assumptions and better fit of multispecies-coalescent formulations in evaluated conditions, yet some loci that violated a model had limited influence on downstream species-tree estimation [25]. The proposed architecture therefore records both model inadequacy and topological sensitivity. A failed adequacy test establishes a reason for caution; it does not independently identify which branch is wrong.

Disagreement between inference paradigms should be treated similarly. Comparative work dissecting concatenation- and quartet-based results shows that conflict between methods can reflect the distribution and strength of phylogenetic signal among genes and sites rather than a simple contest in which one method is intrinsically trustworthy [26]. Consequently, the architecture rejects "method agreement" as the sole validation criterion. Concordant results obtained under similarly misspecified assumptions may remain overconfident, whereas disagreement among plausible analyses can reveal where model choice, gene-tree estimation, or uneven signal requires investigation. In insect phylogenomics, explicit uncertainty begins precisely at this boundary: analytical sensitivity is evidence about the stability of an inference, not proof of the historical mechanism that produced the instability.

Support metrics and their limits

High branch support is meaningful only in relation to the quantity it estimates. Gene concordance factors and site concordance factors measure how frequently decisive genes or sites support a focal branch, thereby exposing heterogeneity that bootstrap values may not display [27]. A low gene concordance factor is not automatically evidence of biological discordance, because short loci or weakly informative gene trees can also reduce agreement. Concordance therefore adds a second evidential axis rather than replacing conventional support.

Support statistics also differ in their assumptions and interpretation. Historical and methodological synthesis shows that bootstrap, posterior-probability, concordance, and related measures do not answer the same inferential question [28]. Treating them as interchangeable "confidence" values erases this distinction. Concordance-based approaches can additionally use the distribution of gene-wise agreement to inform relationships or rates, but their interpretation remains conditional on gene-tree quality and the structure of the dataset [29]. Recent insect radiations reinforce the practical point: high conventional support can coexist with substantial underlying gene or site conflict [3–5].

Propagation of uncertainty to species trees

Species-tree estimation does not reset upstream uncertainty to zero. Quartet-based analyses show that weighting gene-tree contributions by their estimated uncertainty can improve species-tree accuracy under tested conditions [30]. This result is important because an estimated gene tree is itself an inference, not an error-free observation.

A stronger form of propagation retains distributions over gene-tree uncertainty rather than supplying only one estimated tree per locus. QuCo demonstrates that co-estimating species trees while integrating gene-tree distributions can improve accuracy in simulations and an empirical bee example [31]. Neither result establishes a universally superior estimator, but together they show that uncertainty can be carried forward rather than discarded at the gene-tree/species-tree boundary [30, 31].

These alternatives create an evidential reconciliation problem: different pipelines may report the same final topology while preserving very different amounts of upstream uncertainty. **Table 2** distinguishes those propagation states without converting them into an unsupported certainty grade.

Table 2. Species-tree propagation states that determine whether upstream gene-tree uncertainty is retained, transformed, or hidden

Compression		
Propagation dimension	**Point-estimate compression**	
Upstream condition	One estimated gene tree supplied per locus	
Species-tree treatment	Gene-tree uncertainty is collapsed before summary inference	
Interpretive consequence	Final support can appear precise while uncertainty in locus histories is no longer visible	
Sensitivity test	Compare sensitivity to alternative gene trees or support-aware weighting	
Residual issue	Gene-tree estimation error is not represented directly	
Standing	Evidence-supported concern	

Uncertainty retained		
Propagation dimension	**Uncertainty-weighted contribution**	**Distribution-aware propagation**
Upstream condition	Gene-tree branches or quartets carry estimated reliability information	Multiple plausible gene trees are retained
Species-tree treatment	Weakly supported contributions are down-weighted	Species-tree inference integrates across gene-tree distributions

Interpretive consequence	Species-tree estimation acknowledges unequal confidence among inferred gene histories	Upstream uncertainty remains represented during species-tree estimation
Sensitivity test	Compare weighted and unweighted results across difficult branches	Evaluate whether conclusions persist across posterior gene-tree variation
Residual issue	Weighting remains conditional on the error model and support measure	Computational and model assumptions remain
Standing	Evidence-supported strategy	Evidence-supported strategy

Topology stability and sensitivity		
Propagation dimension	**Topology-stable, provenance-uncertain**	**Topology-sensitive propagation**
Upstream condition	Alternative propagation methods recover the same species-tree branch	Species-tree relationship changes when gene-tree uncertainty is represented differently
Species-tree treatment	Agreement is retained without assigning a historical cause	Downstream inference becomes conditional on upstream uncertainty treatment
Interpretive consequence	Stable topology may coexist with unresolved biological or analytical explanation	Branch should not be presented as method-independent certainty
Sensitivity test	Report stability separately from source attribution	Target branch-specific sensitivity and source diagnostics
Residual issue	Stability is not causal diagnosis	Correct historical explanation may remain unresolved
Standing	Proposed reporting state	Proposed reporting state

The central propagation problem is spatial as well as statistical: uncertainty can be retained, compressed, or relabelled at successive inferential boundaries. **Figure 2** makes those transitions explicit while keeping topological stability separate from historical attribution.

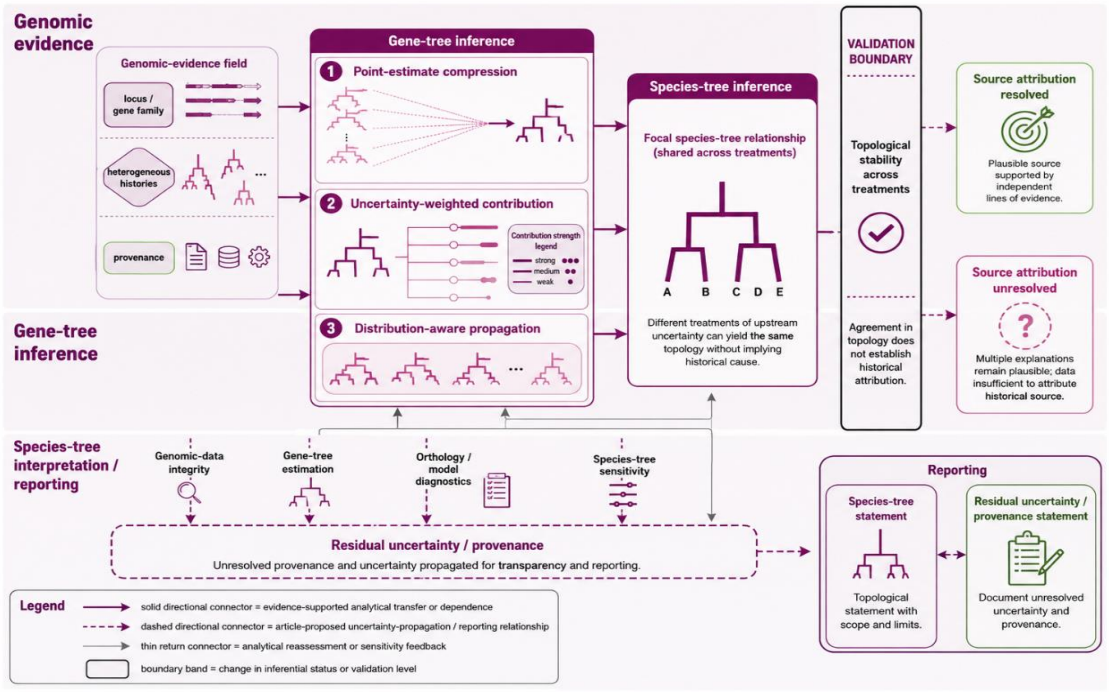


Figure 2. Where phylogenomic certainty is retained, compressed, or re-labelled across gene-tree and species-tree inference

Uncertainty reporting architecture

The proposed reporting architecture separates quantities that are commonly compressed into one statement of confidence. At minimum, a contentious branch should distinguish conventional branch support from gene- or site-level concordance because these metrics describe different dimensions of evidential agreement [27–29]. It

should also state when important alternative causes of discordance remain unresolved; using multiple diagnostic approaches does not guarantee unique source attribution [6].

A second reporting layer records provenance and propagation. Technical failures such as contamination can precede evolutionary interpretation and should therefore remain visible when they materially affect a result [7]. The proposed architecture consequently distinguishes input-data integrity, gene-tree heterogeneity, missingness or locus-definition concerns, orthology provenance, model sensitivity, treatment of gene-tree uncertainty during species-tree estimation, and residual alternative explanations. These fields are not a numerical certainty score. They are a structured declaration of what has been resolved, what is analytically stable, and what remains historically uncertain.

A major implication is that genomic scale and evidential independence should not be treated as synonymous. Reanalysis of contentious phylogenomic relationships shows that a small number of influential genes can determine a concatenated topology, while an influential locus need not be erroneous merely because its influence is disproportionate [32]. Earlier cross-dataset work likewise demonstrated that a handful of genes or sites can dominate support for some difficult branches, whereas more established control relationships are less sensitive [33]. Large matrices can therefore contain concentrated rather than broadly replicated phylogenetic signal.

The proposed framework responds by preserving uncertainty at the level where it originates. Gene-tree disagreement is not automatically error; missingness is not reducible to matrix occupancy; paralogy is neither automatically fatal nor automatically benign; and model rejection does not itself identify an incorrect branch. This organization follows the broader recognition that phylogenomic incongruence is multicausal [1]. Its specific contribution is to require unresolved source attribution to remain attached to downstream species-tree statements instead of disappearing when a topology receives high support.

The framework is deliberately not a calibrated estimator. It does not produce a probability that a branch is true, diagnose introgression from discordance alone, or prescribe one universal filtering or species-tree method. Its utility therefore requires prospective evaluation. Appropriate validation would test whether structured uncertainty reporting improves reproducibility of analytical choices, identifies branches whose apparent certainty depends on hidden assumptions, and helps independent investigators distinguish topological stability from historical explanation. Such validation is necessary because reform of phylogenomic support should add interpretable information rather than simply multiply statistics [2].

Conclusion

Insect phylogenomics requires confidence statements that preserve the origin and propagation of uncertainty rather than allowing a strongly supported species tree to erase unresolved upstream alternatives. The proposed architecture separates biological gene-tree heterogeneity, sampling and locus-definition uncertainty, orthology provenance, model adequacy, and gene-tree-to-species-tree propagation. It treats agreement among methods as evidence of stability rather than automatic proof of historical cause, and disagreement as information requiring diagnosis. Explicit uncertainty is therefore a more precise statement of what the evidence supports, what remains conditional, and what still requires validation.

Acknowledgments: None

Conflict of Interest: None

Financial Support: None

Ethics Statement: None

References

1. Steenwyk JL, Li Y, Zhou X, Shen XX, Rokas A. Incongruence in the phylogenomics era. Nat Rev Genet. 2023;24(12):834-50. doi:10.1038/s41576-023-00620-x
2. Thomson RC, Brown JM. On the need for new measures of phylogenomic support. Syst Biol. 2022;71(4):917-20. doi:10.1093/sysbio/syac002

3. Herrig DK, Ridenbaugh RD, Vertacnik KL, Everson KM, Sim SB, Geib SM, et al. Whole genomes reveal evolutionary relationships and mechanisms underlying gene-tree discordance in Neodiprion sawflies. Syst Biol. 2024;73(5):839-60. doi:10.1093/sysbio/syae036
4. Stukel M, Porczak AE, Gordon ERL, Vailionis J, Haji D, Buckley TR, et al. Phylogenomics improves the phylogenetic resolution and provides strong evidence of mito-nuclear discordance in two genera of a New Zealand cicada (Hemiptera: Cicadidae) species radiation. Syst Entomol. 2024;49(2):237-57. doi:10.1111/syen.12613
5. He B, Zhao Y, Su C, Lin G, Wang Y, Li L, et al. Phylogenomics reveal extensive phylogenetic discordance due to incomplete lineage sorting following the rapid radiation of alpine butterflies (Papilionidae: Parnassius). Syst Entomol. 2023;48(4):585-99. doi:10.1111/syen.12592
6. Morales-Briones DF, Kadereit G, Tefarikis DT, Moore MJ, Smith SA, Brockington SF, et al. Disentangling sources of gene tree discordance in phylogenomic data sets: Testing ancient hybridizations in Amaranthaceae s.l. Syst Biol. 2021;70(2):219-35. doi:10.1093/sysbio/syaa066
7. Owen CL, Marshall DC, Wade EJ, Meister R, Goemans G, Kunte K, et al. Detecting and removing sample contamination in phylogenomic data: An example and its implications for Cicadidae phylogeny (Insecta: Hemiptera). Syst Biol. 2022;71(6):1504-23. doi:10.1093/sysbio/syac043
8. Vasilikopoulos A, Balke M, Kukowka S, Pflug JM, Martin S, Meusemann K, et al. Phylogenomic analyses clarify the pattern of evolution of Adephaga (Coleoptera) and highlight phylogenetic artefacts due to model misspecification and excessive data trimming. Syst Entomol. 2021;46(4):991-1018. doi:10.1111/syen.12508
9. Prasanna AN, Gerber D, Kijpornyongpan T, Aime MC, Doyle VP, Nagy LG. Model choice, missing data, and taxon sampling impact phylogenomic inference of deep Basidiomycota relationships. Syst Biol. 2020;69(1):17-37. doi:10.1093/sysbio/syz029
10. Owen CL, Miller GL. Phylogenomics of the Aphididae: Deep relationships between subfamilies clouded by gene tree discordance, introgression and the gene tree anomaly zone. Syst Entomol. 2022;47(3):470-86. doi:10.1111/syen.12542
11. Knyshov A, Gordon ERL, Masonick PK, Castillo S, Forero D, Hoey-Chamberlain R, et al. Chromosome-aware phylogenomics of assassin bugs (Hemiptera: Reduvioidea) elucidates ancient gene conflict. Mol Biol Evol. 2023;40(8):msad168. doi:10.1093/molbev/msad168
12. Skinner RK, Dietrich CH, Walden KKO, Gordon E, Sweet AD, Podsiadlowski L, et al. Phylogenomics of Auchenorrhyncha (Insecta: Hemiptera) using transcriptomes: Examining controversial relationships via degeneracy coding and interrogation of gene conflict. Syst Entomol. 2020;45(1):85-113. doi:10.1111/syen.12381
13. Allman ES, Mitchell JD, Rhodes JA. Gene tree discord, simplex plots, and statistical tests under the coalescent. Syst Biol. 2022;71(4):929-42. doi:10.1093/sysbio/syab008
14. Bucher M, Condamine FL, Luo Y, Wang M, Bourgoin T. Phylogeny and diversification of planthoppers (Hemiptera: Fulgoromorpha) based on a comprehensive molecular dataset and large taxon sampling. Mol Phylogenet Evol. 2023;186:107862. doi:10.1016/j.ympev.2023.107862
15. Smirnov V, Warnow T. Phylogeny estimation given sequence length heterogeneity. Syst Biol. 2021;70(2):268-82. doi:10.1093/sysbio/syaa058
16. Doyle JJ. Defining coalescent genes: Theory meets practice in organelle phylogenomics. Syst Biol. 2022;71(2):476-89. doi:10.1093/sysbio/syab053
17. Smith ML, Hahn MW. The frequency and topology of pseudoorthologs. Syst Biol. 2022;71(3):649-59. doi:10.1093/sysbio/syab097
18. Yan Z, Smith ML, Du P, Hahn MW, Nakhleh L. Species tree inference methods intended to deal with incomplete lineage sorting are robust to the presence of paralogs. Syst Biol. 2022;71(2):367-81. doi:10.1093/sysbio/syab056
19. Smith ML, Vanderpool D, Hahn MW. Using all gene families vastly expands data available for phylogenomic inference. Mol Biol Evol. 2022;39(6):msac112. doi:10.1093/molbev/msac112
20. Willson J, Roddur MS, Liu B, Zaharias P, Warnow T. DISCO: Species tree inference using multicopy gene family tree decomposition. Syst Biol. 2022;71(3):610-29. doi:10.1093/sysbio/syab070
21. Zhang C, Scornavacca C, Molloy EK, Mirarab S. ASTRAL-Pro: Quartet-based species-tree inference despite paralogy. Mol Biol Evol. 2020;37(11):3292-307. doi:10.1093/molbev/msaa139

22. Xiong H, Wang D, Shao C, Yang X, Yang J, Ma T, et al. Species tree estimation and the impact of gene loss following whole-genome duplication. Syst Biol. 2022;71(6):1348-61. doi:10.1093/sysbio/syac040
23. Crotty SM, Minh BQ, Bean NG, Holland BR, Tuke J, Jermiin LS, et al. GHOST: Recovering historical signal from heterotachously evolved sequence alignments. Syst Biol. 2020;69(2):249-64. doi:10.1093/sysbio/syz051
24. Duchêne DA, Mather N, Van Der Wal C, Ho SYW. Excluding loci with substitution saturation improves inferences from phylogenomic data. Syst Biol. 2022;71(3):676-89. doi:10.1093/sysbio/syab075
25. Jiang X, Edwards SV, Liu L. The multispecies coalescent model outperforms concatenation across diverse phylogenomic data sets. Syst Biol. 2020;69(4):795-812. doi:10.1093/sysbio/syaa008
26. Shen XX, Steenwyk JL, Rokas A. Dissecting incongruence between concatenation- and quartet-based approaches in phylogenomic data. Syst Biol. 2021;70(5):997-1014. doi:10.1093/sysbio/syab011
27. Minh BQ, Hahn MW, Lanfear R. New methods to calculate concordance factors for phylogenomic datasets. Mol Biol Evol. 2020;37(9):2727-33. doi:10.1093/molbev/msaa106
28. Simon C. An evolving view of phylogenetic support. Syst Biol. 2022;71(4):921-8. doi:10.1093/sysbio/syaa068
29. Walker JF, Smith SA, Hodel RGJ, Moyroud E. Concordance-based approaches for the inference of relationships and molecular rates with phylogenomic data sets. Syst Biol. 2022;71(4):943-58. doi:10.1093/sysbio/syab052
30. Zhang C, Mirarab S. Weighting by gene tree uncertainty improves accuracy of quartet-based species trees. Mol Biol Evol. 2022;39(12):msac215. doi:10.1093/molbev/msac215
31. Rabiee M, Mirarab S. QuCo: Quartet-based co-estimation of species trees and gene trees. Bioinformatics. 2022;38(Suppl 1):i413-21. doi:10.1093/bioinformatics/btac265
32. Walker JF, Brown JW, Smith SA. Analyzing contentious relationships and outlier genes in phylogenomics. Syst Biol. 2018;67(5):916-24. doi:10.1093/sysbio/syy043
33. Shen XX, Hittinger CT, Rokas A. Contentious relationships in phylogenomic studies can be driven by a handful of genes. Nat Ecol Evol. 2017;1:0126. doi:10.1038/s41559-017-0126