Eurasia Specialized Veterinary Publication

Entomological Research Letters

ISSN: 3108-639X

2024, Volume 4, Issue 2, Page No: 141-150
Copyright CC BY-NC-SA 4.0
Available online at: www.esvpub.com/

Discordance, Sampling Design, and Model Choice in Phylogenomic Inference Across Rapidly Diversifying Insect Radiations

Ivan Petrov[1], Olga Ivanova[1]*, Dmitry Smirnov[2]

[1]Department of Phylogenomic Discordance and Sampling Design, Faculty of Life Sciences, Lomonosov Moscow State University, Moscow, Russia.
[2]Department of Model Choice and Rapid Radiations, Faculty of Biosciences, Saint Petersburg State University, Saint Petersburg, Russia.

***E-mail** ✉ olga.ivanova@mgu.ru

ABSTRACT

Rapid insect radiations are a severe test of phylogenomic inference because large genomic matrices can yield strongly supported species-tree hypotheses while retaining extensive conflict among genes, sites, marker classes, or analytical models. This methodological review evaluates how taxon sampling, locus selection, missing data, orthology assessment, paralogy, gene-tree estimation, model choice, and introgression-aware analyses jointly condition interpretation of discordance. Evidence from aphids, alpine butterflies, cicadas, leafhoppers, sawflies, tephritid flies, hymenopterans, and broader methodological studies indicates that no single diagnostic resolves all short-internode radiations. The central synthesis proposed here is a design-conditional interpretation of discordance: biological explanations should be considered only after plausible pipeline-generated contributions have been stress-tested. In this view, support values summarize confidence under a specified analysis but do not by themselves distinguish incomplete lineage sorting, gene flow, hidden paralogy, missingness, model misspecification, or gene-tree error. Robust phylogenomic inference therefore depends on the stability of conclusions across defensible sampling and model perturbations and on concordance among evidence classes that make different assumptions. This approach preserves uncertainty where historical processes remain statistically non-identifiable and avoids treating a single preferred topology as complete resolution of a rapid radiation.

Keywords: Phylogenomics, Gene-tree discordance, Incomplete lineage sorting, Introgression, Orthology, Multispecies coalescent

Received: 30 October 2024
Revised: 19 December 2024
Accepted: 21 December 2024

How to Cite This Article: Petrov I, Ivanova O, Smirnov D. Discordance, Sampling Design, and Model Choice in Phylogenomic Inference Across Rapidly Diversifying Insect Radiations. Entomol Res Lett. 2024;4(2):141-50. https://doi.org/10.51847/5cKV6VMwFK

Introduction

Rapid diversification compresses successive lineage splits into short intervals, leaving limited opportunity for ancestral polymorphisms to sort and creating phylogenetic histories in which many loci need not share the same genealogy. Insects provide particularly revealing examples because dense genomic sampling has not removed conflict from several radiations. Aphid phylogenomics recovered deep relationships from thousands of orthologues yet retained strong gene-tree discordance, evidence of introgression, and branches compatible with an anomaly-zone setting; alpine *Parnassius* butterflies showed extensive discordance largely consistent with incomplete lineage sorting; and New Zealand cicadas combined improved nuclear resolution with persistent mito-nuclear incongruence [1–3]. These studies establish that genome scale and high branch support are not synonymous with genealogical unanimity.

© 2024 Entomological Research Letters

The interpretive difficulty is sharpened by whole-genome work in *Neodiprion* sawflies. Local concordance varied with informative-site content, missingness, GC composition, gene density, recombination-related variables, and introgression signals, demonstrating that properties of the data and properties of the evolutionary history can covary within the same radiation [4]. Such associations do not identify a single causal mechanism. They instead show why a discordant species-tree problem is also a design problem: observed conflict is filtered through which taxa were sampled, which loci were retained, how orthology was defined, how gene trees were estimated, and which species-history model was imposed.

This review therefore treats discordance as an inferential state rather than a diagnosis. The proposed synthesis separates three questions that are often collapsed: whether genomic evidence is internally heterogeneous, whether analytical choices amplify or suppress that heterogeneity, and which historical process best explains the residual pattern. The scope is methodological rather than taxonomically revisionary. The objective is not to identify a universally optimal pipeline, but to specify where conclusions about rapidly diversifying insect lineages become conditional on sampling design and model choice.

Taxon sampling

Taxon sampling affects both the biological representation of a radiation and the statistical context in which short branches are estimated. In Deltocephalinae, an anchored-hybrid-enrichment study comprising 730 terminals and a large character matrix resolved many relationships that had remained difficult, yet the shortest internal branches still showed conflict across analyses [5]. Dense sampling can therefore improve resolution without erasing a genuinely difficult diversification interval. This distinction matters because “more taxa” is not equivalent to “all lineages relevant to an ancestral diversification event have been represented.”

Broad hymenopteran studies make a related point at deeper timescales. Ultraconserved-element analyses of aculeate wasps, ants, and bees showed that the configuration of ingroup and outgroup representation can affect difficult deep placements [6], while a large phylogenomic analysis of Hymenoptera demonstrated the resolving power of extensive taxon and gene coverage across the order [7]. These studies support strategic breadth, but neither establishes a universal minimum taxon number for rapid radiations.

Near-complete sampling in a non-insect radiation clarifies another methodological boundary. In parrots, extensive taxonomic coverage still left gene-tree conflict that could be separated probabilistically into artifactual and biological components using information and noise diagnostics [8]. Its biological details cannot be transferred directly to insects, but its methodological lesson is relevant: taxonomic completeness does not make gene-tree estimation error disappear. For insect radiations, sampling adequacy should therefore be evaluated by representation of lineages bearing on contested internodes, outgroup adequacy, and sensitivity of focal relationships to taxon addition or removal rather than by raw species count alone.

Locus sampling

Locus sampling determines how much phylogenetic information is available and which genomic histories are allowed to contribute. In the *Anastrepha fraterculus* species complex, locus-resampling analyses showed that a comparatively small subset could reproduce much of the resolution obtained from the larger genomic dataset [9]. The result demonstrates diminishing information gains in that particular system, not a general locus threshold for insect phylogenomics. Rapid radiations differ in internode duration, recombination history, sequence variability, and gene-flow history, so an empirically adequate locus set in one clade may be inadequate in another.

Data provenance also constrains the available marker set. A Sternorrhyncha analysis using ultraconserved elements extracted from genomic and transcriptomic resources found marked differences in locus recovery between the two source types [10]. Thus, “locus number” is partly a property of sampling technology and data availability rather than simply an investigator-selected parameter. Anchored phylogenomic work in Lepidoptera similarly demonstrated that capture, assembly, contamination control, orthology assessment, and alignment are linked stages determining the matrix eventually presented to tree inference [11].

The multispecies coalescent adds a further conceptual requirement: a physical gene or captured region is not automatically an independent coalescent unit. Recombination can make a long region genealogically heterogeneous, whereas linkage can reduce effective independence among nominal loci [12]. Locus design should therefore be justified by information content and genealogical assumptions together. The proposed implication is to replace a simple “more loci are better” rule with a sensitivity question: does the focal relationship persist across defensible locus definitions, information filters, and subsampling regimes?

Missing data

Missing data influence phylogenomic inference through their distribution, association with informative sites, and interaction with model behavior. Work on deep eukaryote phylogeny showed that site-specific mixture models can be sensitive to missingness when influential outliers and mosaic signal are present [13]. Although that example is outside insects, it demonstrates why a fixed percentage-completeness threshold has no general statistical meaning.

Lepidopteran phylogenomics provides a more direct entomological boundary. Reanalysis of a difficult butterfly-and-moth backbone found that alternative data formulations, including nucleotide versus amino-acid treatment and different modeling choices, did not uniformly remove conflict [14]. Missingness therefore cannot be evaluated independently of composition, model adequacy, and gene-tree estimation. In *Neodiprion*, genomic windows with lower missingness tended to show greater site concordance, but missingness covaried with other genomic features, so the association cannot be converted into a causal claim [4].

For rapidly diversifying insects, the useful question is not whether missing data are present, but whether missingness is structured around particular taxa, loci, or contested branches and whether conclusions change when that structure is perturbed. The same nominal amount of missingness can consequently have different inferential meanings; **Table 1** distinguishes these configurations without assigning an unsupported completeness score.

Table 1. Missing-data configurations that alter the interpretation of discordance in insect phylogenomics

Taxonomic with observed basis	Consequence for inference	Interpretive limit and decisive check
Diffuse incompleteness		
Evidence-supported synthesis		
Broad multilocus matrices **Missing cells dispersed across taxa and loci**	May reduce usable information without selectively targeting one contested branch	Effect depends on retained signal and model behavior Compare full and occupancy-filtered matrices
Lineage-structured missingness		
Proposed diagnostic category		
Particular taxa or clades repeatedly underrepresented **Recurrent locus absence in the same lineage**	Can weaken placement of that lineage and alter adjacent short internodes	Absence may reflect source data, sequence divergence, or capture failure Perturb taxon occupancy and examine topology stability
Locus-structured missingness		
Evidence-supported synthesis		
Marker classes or source types differ in recovery **Unequal recovery among loci or genomic versus transcriptomic inputs**	Changes which genomic histories contribute to species-tree estimation	Recovery rate is not equivalent to locus reliability Repeat inference across locus-recovery classes
Missingness coupled to low information		
Evidence-supported synthesis		
Windows or loci carrying few informative sites **Reduced information and lower concordance occur together**	Apparent conflict may partly reflect difficult gene-tree estimation	Covariation does not establish causation Model information content and missingness jointly
Model-sensitive missingness		
Evidence-supported synthesis		
Matrices containing influential sites, mosaics, or compositional heterogeneity **Topology changes across data treatments or model classes**	Missingness can participate in model-dependent instability	Demonstrated effects are dataset-specific Compare defensible model and filtering alternatives before biological attribution

Orthology and paralogs

Orthology errors are especially consequential in rapid radiations because a confidently estimated gene tree can still represent the wrong correspondence among duplicated gene copies. Empirical work on complex target-capture data has shown that artifactual orthologues can arise through differential assembly and produce substantial topological heterogeneity [15]. The relevant lesson is not that target capture is intrinsically unreliable, but that single-copy appearance after assembly is not proof of evolutionary orthology.

Conversely, paralogy is not uniformly fatal to species-tree estimation. Simulation and empirical analyses found that some methods designed for incomplete lineage sorting can remain robust under specified levels and patterns

of paralogy [16]. The effect of pseudo-orthologues likewise depends on the history of duplication, loss, and lineage sorting rather than following a single direction [17]. These findings argue against an indiscriminate "discard every multicopy locus" rule.

Several methods make that conditionality explicit. OrthoFinder uses gene-tree-aware orthology inference rather than sequence similarity alone [18]. ASTRAL-Pro extends quartet-based species-tree inference to multicopy gene trees [19]. GeneRax models gene-family histories under duplication, transfer, and loss while accounting for a species tree [20], whereas SpeciesRax estimates species trees directly from gene families under a related reconciliation framework [21]. These methods do not collectively solve every source of phylogenomic discordance, particularly introgression or severe gene-tree error, but they widen the admissible analytical space beyond strict single-copy filtering.

The central proposed distinction is therefore among sequence acquisition, copy-history interpretation, and species-history inference. These stages are temporally ordered but analytically recursive: an early orthology decision determines which genealogies survive into later coalescent or concatenation analyses, whereas downstream instability can justify renewed inspection of suspicious gene families. **Figure 1** makes this dependency explicit while maintaining a boundary between copy-history uncertainty and biological interpretation of species-tree conflict.

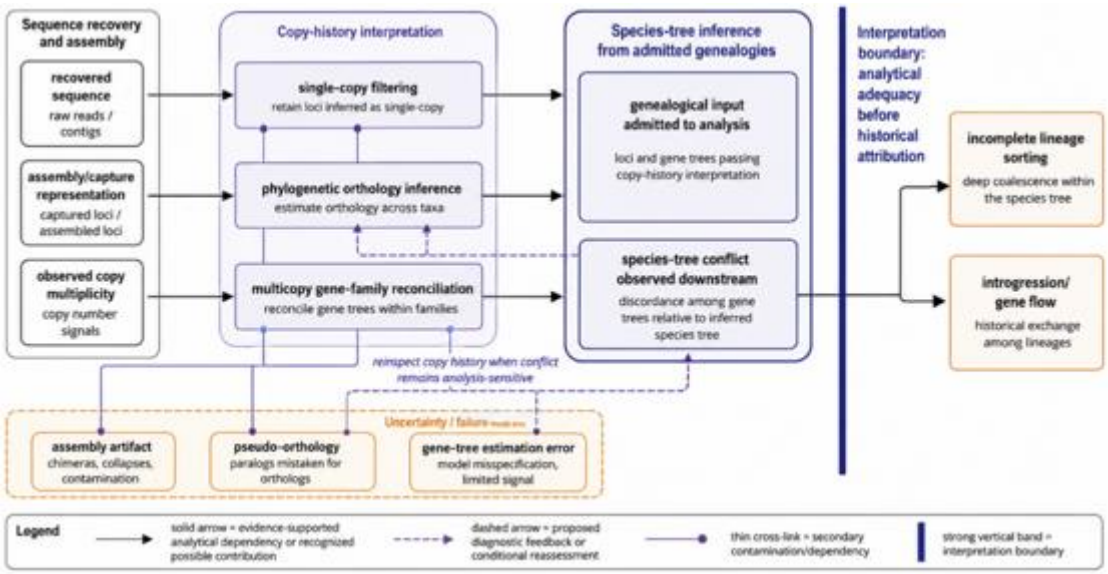


Figure 1. Copy-history uncertainty propagates through locus admission before species-tree discordance can be interpreted

Gene-tree discordance

Gene-tree discordance is the observable distribution of conflicting genealogical signal; it is not itself a mechanism. Rate heterogeneity among lineages can reduce agreement between estimated gene trees and species trees, adding an analytical contribution to conflict that can coexist with genuine biological heterogeneity [22]. Quartet-based concordance analysis provides a complementary perspective by asking whether observed discordance lies within expectations under a multispecies coalescent model. Deviations can indicate model inadequacy, estimation error, or reticulation, but they are not uniquely diagnostic of introgression [23].

Locus-wise likelihood comparisons further show why aggregate branch support is insufficient. Different loci can provide sharply different relative support for competing topologies even when a preferred species-tree branch has high overall support [24]. Conflict should consequently be represented as a distribution of evidence rather than collapsed into a binary resolved/unresolved label. Its consequences also extend beyond topology: incongruence between gene-tree and species-tree histories can bias divergence-time estimation when dating analyses condition on an inappropriate topology [25].

The proposed diagnostic interpretation distinguishes four analytically useful states: conflict compatible with expected lineage sorting; conflict dominated by weak or unstable gene-tree estimation; conflict structured around particular genomic or sampling properties; and residual conflict that persists after those analytical alternatives

have been challenged. These states are neither mutually exclusive biological mechanisms nor a certainty scale. Instead, they specify why superficially similar patterns of disagreement can justify different next analyses. Because identical-looking tree disagreement can arise under different conditions, the principal failure modes and their discriminating tests are more efficiently separated in a conditionality matrix than in a linear narrative.

Table 2. Conditional interpretations of gene-tree discordance and the failure modes that can mimic historical conflict

Interpretation consequence, failure mode or competing explanation, and discriminating test	Discordance state with evidential standing	Analytical setting with observed basis
Compatible with substantial ancestral lineage sorting	Short-internode concordance deficit	Rapid radiation with many alternative gene histories
Gene-tree error can produce a similar distribution	Evidence-supported synthesis	Low concordance concentrated around brief successive splits
Compare concordance with multispecies-coalescent expectations and gene-tree quality		
Part of the species-tree conflict may be analytical	Information-limited discordance	Loci or genomic windows carrying weak phylogenetic information
Rate heterogeneity, missingness, or inadequate optimization	Evidence-supported synthesis	Conflict associated with low informative-site content or unstable gene trees
Re-estimate after information and search perturbations		
Indicates heterogeneous genealogical or estimation conditions	Genomically structured discordance	Conflict varies systematically among genomic regions
Recombination, introgression, composition, or structured missingness	Evidence-supported synthesis	Concordance covaries with genomic properties or marker classes
Test predictors jointly and avoid single-variable causal attribution		
Demonstrates that one marker class does not capture the entire history	Marker-class discordance	Nuclear and mitochondrial histories disagree
Introgression, lineage sorting, lineage-specific transmission history, or estimation error	Evidence-supported synthesis	Strong but incompatible marker-specific topologies
Compare independent nuclear evidence and explicit historical models		
Justifies escalation to richer historical models rather than declaring complete resolution	Residual discordance after perturbation	Preferred relationship persists while gene/locus disagreement survives defensible reanalysis
Unsampled lineages and unmodeled heterogeneity remain possible	Proposed diagnostic category	Stable focal topology with persistent underlying conflict
Evaluate coalescent fit and introgression-aware alternatives		

Coalescent and concatenation models

Concatenation and multispecies-coalescent approaches answer related but non-identical questions. Concatenation treats aligned sites as contributing to a common topology under the selected sequence model, whereas multispecies-coalescent approaches explicitly allow gene histories to differ through ancestral lineage sorting. A broad comparison across phylogenomic datasets found that multispecies-coalescent models frequently provided a better account of heterogeneous gene histories than concatenation, but the result does not establish that a coalescent analysis is universally preferable for every dataset or branch [26]. Model adequacy depends on whether discordance is sufficiently represented by the evolutionary process assumed.

Summary coalescent methods also inherit uncertainty from their input gene trees. ASTRAL-III permits scalable inference from partially resolved gene trees, but simulation and empirical analyses show that input resolution and estimation quality affect downstream performance [27]. Consequently, a strongly supported summary species tree is not independent confirmation of the gene trees from which it was constructed. In rapid insect radiations, this dependency is important because precisely those short internodes that motivate coalescent analysis can also produce weakly estimated gene histories.

A single topology across all sites need not be the only alternative. Mixture-across-sites-and-trees models allow different subsets of sites to support distinct topological histories [28]. Such models can better represent heterogeneous genomic signal, but topological mixture alone does not identify the underlying biological process. Likewise, modern maximum-likelihood implementations provide extensive model, support, and search options [29], while automated substitution-model selection can identify a better-supported member of a specified

candidate set [30]. Neither procedure eliminates the possibility that the chosen model remains biologically incomplete.

Model comparison should therefore be conditional rather than hierarchical. Concatenation can reveal a dominant genome-wide signal; coalescent approaches can represent incomplete lineage sorting; and topology-mixture methods can accommodate site-level heterogeneity. Agreement across these analyses strengthens the stability of a relationship under different assumptions. Disagreement is more informative when it is treated as a diagnostic outcome requiring explanation rather than as a contest settled by whichever analysis returns the highest nominal support.

Introgression-aware methods

The multispecies coalescent usually assumes that lineages diverge without subsequent gene flow unless migration or hybridization is represented explicitly. Simulation work demonstrates that gene flow can alter coalescent-based species-tree inference, with the magnitude and direction of the effect depending on its timing and configuration [31]. Thus, discordance exceeding expectations under a no-gene-flow model is a reason to consider reticulation, not evidence that reticulation has already been demonstrated.

Explicit introgression models illustrate the additional information needed. Full-likelihood analysis of the *Heliconius* erato–sara group reconstructed a complex divergence history by comparing models that incorporated gene flow rather than relying only on a bifurcating species tree. At the same time, methodological analyses show that unsampled or extinct "ghost" lineages can change the interpretation of introgression statistics and can influence coalescent estimates of topology or divergence time [32–34]. These findings place taxon sampling directly inside introgression inference: a significant site-pattern statistic need not uniquely identify the sampled donor, recipient, or historical event.

Introgression-aware analysis should therefore combine several evidence classes whose assumptions differ. Site-pattern asymmetry can identify departures from a simple bifurcating expectation; explicit coalescent-with-gene-flow models can compare historical scenarios; topology mixtures or networks can represent heterogeneous genealogical histories; and genomic localization can test whether discordance is distributed in a pattern compatible with particular histories. None of these signals is uniquely diagnostic in isolation. Incomplete lineage sorting, gene-tree error, hidden paralogy, model misspecification, recombination, and unsampled lineages remain competing explanations whose plausibility depends on the dataset.

The proposed interpretive boundary is consequently stringent: gene flow becomes a preferred historical explanation only when evidence remains structured after sampling, orthology, gene-tree quality, and no-gene-flow model adequacy have been challenged and when multiple introgression-sensitive analyses converge without requiring mutually incompatible histories. **Table 3** separates these inferential states because the same nominal "introgression signal" can imply different evidential obligations.

Table 3. Introgression-aware evidence states and the historical inferences they can and cannot sustain

Introgression-aware stat with evidential standing	Process hypothesis addre with observed basis	Consequence for inference, principal boundary or failure mode, and required test or cross-check
Site-pattern asymmetry **Evidence-supported synthesis**	Gene flow relative to a simple bifurcating history Sampled taxon quartets or related configurations Excess allele-pattern asymmetry	Motivates reticulation testing Ghost lineages, ancestral structure, sampling configuration Evaluate alternative taxon configurations and unsampled-lineage scenarios
Coalescent model improvement with gene flow **Evidence-supported synthesis**	Divergence accompanied by migration or introgression Explicitly parameterized lineage histories Gene-flow model fits data better than a corresponding no-flow model	Supports a reticulate history under tested models Misspecified direction, timing, or missing lineage Compare competing historical models and parameterizations
Topology-mixture concentration **Evidence-supported synthesis**	Different genomic subsets support alternative histories Recombining genomic data Recurrent support for more than one topology	Demonstrates heterogeneous genealogical signal Mixture can arise from ILS or estimation error Compare topology frequencies with coalescent expectations and gene-tree quality

Spatially structured genomic discordance **Evidence-supported synthesis**	Introgression affecting restricted genomic regions Genome-scale windows or locus classes Discordance clustered rather than uniformly distributed	Can strengthen a localized gene-flow interpretation Selection, recombination, missingness, or information content may covary Model genomic predictors jointly and test alternative processes
Ghost-sensitive introgression signal **Evidence-supported synthesis**	Reticulation involving an unsampled or extinct lineage Incomplete taxon sampling Statistic changes interpretation under alternative unsampled-lineage scenarios	Weakens unique donor–recipient attribution Non-identifiability of the historical source Expand sampling where possible and compare ghost-lineage models
Residual convergent reticulation evidence **Proposed evidential state**	Conflict persists after analytical alternatives are stress-tested Rapid radiation with stable sampling and orthology results Multiple assumption-distinct analyses favor compatible gene-flow histories	Permits a stronger, still model-bounded introgression interpretation Unmodeled extinct lineages or processes remain possible Require convergence across independent diagnostics

Because introgression inference crosses a boundary from observed genomic heterogeneity to an historical process claim, its evidential dependencies are represented explicitly in **Figure 2** rather than as a direct arrow from discordance to hybridization.

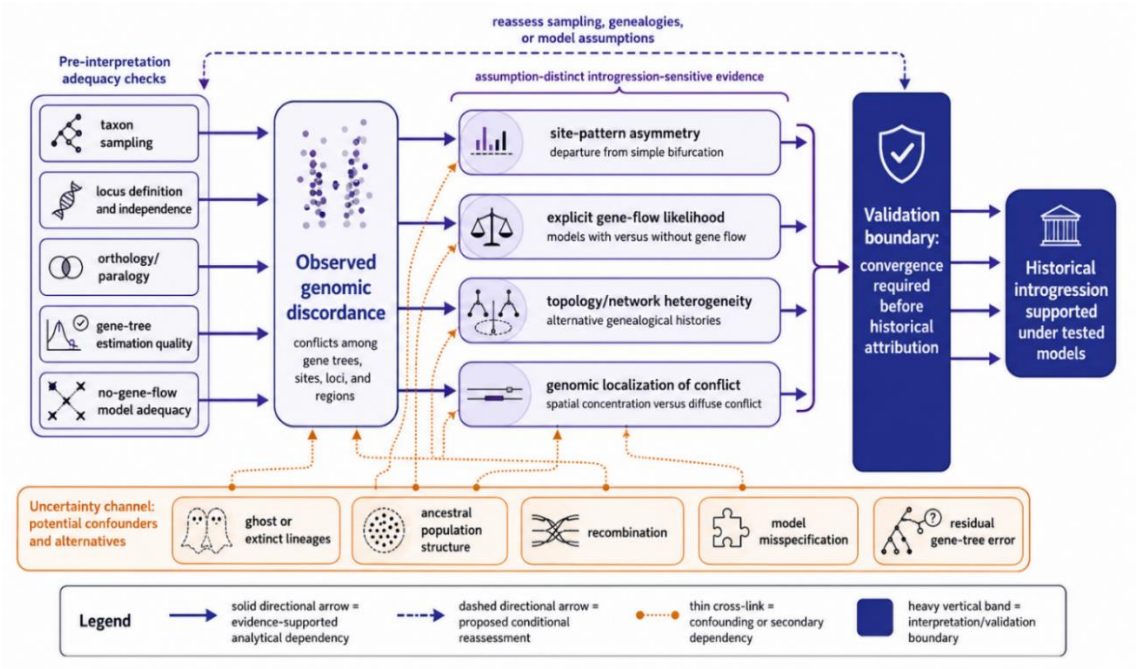


Figure 2. Historical gene-flow inference is bounded by analytical adequacy and unsampled-lineage uncertainty

Sensitivity and robustness

Robustness is not equivalent to repeatedly obtaining a high support value under nearly identical settings. Even the numerical optimization used to estimate gene trees can matter. A large evaluation of maximum-likelihood tree searching showed that increasing the number of searches could change estimated gene trees and, in some empirical datasets, alter downstream species-tree inference [35]. Search effort is therefore part of the analytical design rather than an invisible computational detail.

For rapid insect radiations, sensitivity analysis should perturb choices that plausibly affect the contested internodes: taxon representation, locus occupancy, locus definition, orthology treatment, gene-tree search effort, substitution models, filtering, coalescent versus concatenation assumptions, and introgression-aware alternatives. These perturbations are not equally relevant in every dataset. A relationship that moves only when poorly

supported loci are included poses a different problem from one that reverses under reasonable taxon additions or under equally defensible evolutionary models.

The proposed criterion is interpretive robustness: a conclusion is stronger when both the focal relationship and the explanation assigned to its discordance remain compatible across defensible analytical alternatives. Topological stability without mechanistic stability should be reported as such. Conversely, persistent instability is itself information about the limits of the available data rather than a defect that must be hidden by selecting one preferred pipeline.

Recommended analytical workflow

A useful precedent for resolving heterogeneous phylogenomic signal comes from multi-method analyses that combine gene-tree, species-tree, site-pattern, network, syntenic, and simulation evidence rather than attributing discordance from one diagnostic alone [36]. Transferring that principle to insects supports a proposed workflow organized around successive questions rather than a fixed software recipe.

First, define the historical question and contested internodes before maximizing data volume. Sampling should then be tested for representation of relevant lineages, outgroups, and marker classes. Second, establish locus admissibility by examining information content, missingness, independence assumptions, orthology, and multicopy history. Third, evaluate gene-tree quality and characterize discordance as a distribution across loci, sites, and genomic regions rather than as a single support statistic.

Only after these checks should species-history models be compared. Concatenation, coalescent summaries, topology mixtures, and explicit gene-flow models should be interpreted according to the processes each can and cannot represent. If discordance remains concentrated or systematically exceeds expectations under a no-gene-flow coalescent, introgression-aware analyses become warranted. Their interpretation should remain conditional on ghost lineages, ancestral structure, recombination, and sampling incompleteness.

Finally, the complete analysis should be perturbed across a small set of scientifically defensible alternatives chosen before interpretive preference is known. The manuscript's proposed decision rule is therefore qualitative: do not escalate a biological explanation merely because an analytical method can produce it; escalate when simpler analytical explanations have been challenged and assumption-distinct evidence converges. Where topology is stable but process attribution is not, report the topology and uncertainty separately. Where both topology and process remain unstable, the defensible outcome is unresolved history rather than forced resolution.

This workflow is not a universal algorithm and has not been prospectively validated across insect radiations. Its purpose is to make inferential dependencies explicit, preserve discordance as evidence rather than nuisance, and expose where additional taxa, loci, genomic resources, or alternative evolutionary models could actually change interpretation.

Conclusion

Rapidly diversifying insect lineages expose a central limitation of genome-scale phylogenetics: more sequence data can sharpen support without uniquely identifying species history or the process that generated discordance. Taxon representation, locus definition, missingness, orthology, paralogy, gene-tree estimation, and model assumptions all shape the conflict presented to downstream inference.

The principal contribution of this review is a proposed design-conditional interpretation of phylogenomic discordance. Under this view, incomplete lineage sorting, introgression, or another historical process should not be assigned directly from disagreement among trees. Analytical sources of discordance should first be challenged, competing species-history models should then be compared according to their assumptions, and historical interpretation should strengthen only when independent evidence classes converge.

This approach does not promise a universally resolved tree. Its more defensible objective is to distinguish relationships that are stable under reasonable analytical perturbation from those whose topology or historical explanation remains conditional. For rapid insect radiations, explicitly preserving that boundary between observed genomic conflict and inferred evolutionary history is itself an improvement in phylogenomic inference.

Acknowledgments: None

Conflict of Interest: None

Financial Support: None

Ethics Statement: None

References

1. 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
2. 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
3. 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
4. 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
5. Cao Y, Dietrich CH, Zahniser JN, Dmitriev DA. Dense sampling of taxa and characters improves phylogenetic resolution among deltocephaline leafhoppers (Hemiptera: Cicadellidae: Deltocephalinae). Syst Entomol. 2022;47(3):430-44. doi:10.1111/syen.12540
6. Branstetter MG, Danforth BN, Pitts JP, Faircloth BC, Ward PS, Buffington ML, et al. Phylogenomic insights into the evolution of stinging wasps and the origins of ants and bees. Curr Biol. 2017;27(7):1019-25. doi:10.1016/j.cub.2017.03.027
7. Peters RS, Krogmann L, Mayer C, Donath A, Gunkel S, Meusemann K, et al. Evolutionary history of the Hymenoptera. Curr Biol. 2017;27(7):1013-18. doi:10.1016/j.cub.2017.01.027
8. Smith BT, Merwin J, Provost KL, Thom G, Brumfield RT, Ferreira M, et al. Phylogenomic analysis of the parrots of the world distinguishes artifactual from biological sources of gene tree discordance. Syst Biol. 2023;72(1):228-41. doi:10.1093/sysbio/syac055
9. Congrains C, Dupuis JR, Rodriguez EJ, Norrbom AL, Steck G, Sutton B, et al. Phylogenomic analysis provides diagnostic tools for the identification of Anastrepha fraterculus (Diptera: Tephritidae) species complex. Evol Appl. 2023;16(9):1598-618. doi:10.1111/eva.13589
10. Liu D, Cui J, Liu Y, Niu M, Wang F, Zhao Q, et al. Ultraconserved elements from transcriptome and genome data provide insight into the phylogenomics of Sternorrhyncha (Insecta: Hemiptera). Cladistics. 2024;40(5):496-509. doi:10.1111/cla.12585
11. Breinholt JW, Earl C, Lemmon AR, Lemmon EM, Xiao L, Kawahara AY. Resolving relationships among the megadiverse butterflies and moths with a novel pipeline for anchored phylogenomics. Syst Biol. 2018;67(1):78-93. doi:10.1093/sysbio/syx048
12. Doyle JJ. Defining coalescent genes: Theory meets practice in organelle phylogenomics. Syst Biol. 2022;71(2):476-89. doi:10.1093/sysbio/syab053
13. Al Jewari CA, Baldauf SL. Conflict over the eukaryote root resides in strong outliers, mosaics and missing data sensitivity of site-specific (CAT) mixture models. Syst Biol. 2023;72(1):1-16. doi:10.1093/sysbio/syac029
14. Rota J, Twort V, Chiocchio A, Peña C, Wheat CW, Kaila L, et al. The unresolved phylogenomic tree of butterflies and moths (Lepidoptera): Assessing the potential causes and consequences. Syst Entomol. 2022;47(4):531-50. doi:10.1111/syen.12545
15. Frost LA, Bedoya AM, Lagomarsino LP. Artifactual orthologs and the need for diligent data exploration in complex phylogenomic datasets: A museomic case study from the Andean flora. Syst Biol. 2024;73(2):308-22. doi:10.1093/sysbio/syad076
16. 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

17. 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
18. Emms DM, Kelly S. OrthoFinder: Phylogenetic orthology inference for comparative genomics. Genome Biol. 2019;20(1):238. doi:10.1186/s13059-019-1832-y
19. 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
20. Morel B, Kozlov AM, Stamatakis A, Szöllősi GJ. GeneRax: A tool for species-tree-aware maximum likelihood-based gene family tree inference under gene duplication, transfer, and loss. Mol Biol Evol. 2020;37(9):2763-74. doi:10.1093/molbev/msaa141
21. Morel B, Schade P, Lutteropp S, Williams TA, Szöllősi GJ, Stamatakis A. SpeciesRax: A tool for maximum likelihood species tree inference from gene family trees under duplication, transfer, and loss. Mol Biol Evol. 2022;39(2):msab365. doi:10.1093/molbev/msab365
22. Vankan M, Ho SYW, Duchêne DA. Evolutionary rate variation among lineages in gene trees has a negative impact on species-tree inference. Syst Biol. 2022;71(2):490-500. doi:10.1093/sysbio/syab051
23. 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
24. Mount GG, Brown JM. Comparing likelihood ratios to understand genome-wide variation in phylogenetic support. Syst Biol. 2022;71(4):973-85. doi:10.1093/sysbio/syac014
25. Carruthers T, Sun M, Baker WJ, Smith SA, de Vos JM, Eiserhardt WL. The implications of incongruence between gene tree and species tree topologies for divergence time estimation. Syst Biol. 2022;71(5):1124-46. doi:10.1093/sysbio/syac012
26. 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
27. Zhang C, Rabiee M, Sayyari E, Mirarab S. ASTRAL-III: Polynomial time species tree reconstruction from partially resolved gene trees. BMC Bioinformatics. 2018;19(Suppl 6):153. doi:10.1186/s12859-018-2129-y
28. Wong TKF, Cherryh C, Rodrigo AG, Hahn MW, Minh BQ, Lanfear R. MAST: Phylogenetic inference with mixtures across sites and trees. Syst Biol. 2024;73(2):375-91. doi:10.1093/sysbio/syae008
29. Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A, et al. IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era. Mol Biol Evol. 2020;37(5):1530-4. doi:10.1093/molbev/msaa015
30. Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermiin LS. ModelFinder: Fast model selection for accurate phylogenetic estimates. Nat Methods. 2017;14(6):587-9. doi:10.1038/nmeth.4285
31. Long C, Kubatko L. The effect of gene flow on coalescent-based species-tree inference. Syst Biol. 2018;67(5):770-85. doi:10.1093/sysbio/syy020
32. Thawornwattana Y, Seixas FA, Yang Z, Mallet J. Full-likelihood genomic analysis clarifies a complex history of species divergence and introgression: The example of the erato-sara group of Heliconius butterflies. Syst Biol. 2022;71(5):1159-77. doi:10.1093/sysbio/syac009
33. Tricou T, Tannier E, de Vienne DM. Ghost lineages highly influence the interpretation of introgression tests. Syst Biol. 2022;71(5):1147-58. doi:10.1093/sysbio/syac011
34. Pang XX, Zhang DY. Impact of ghost introgression on coalescent-based species tree inference and estimation of divergence time. Syst Biol. 2023;72(1):35-49. doi:10.1093/sysbio/syac047
35. Liu C, Zhou X, Li Y, Hittinger CT, Pan R, Huang J, et al. The influence of the number of tree searches on maximum likelihood inference in phylogenomics. Syst Biol. 2024;73(5):807-22. doi:10.1093/sysbio/syae031
36. 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