Eurasia Specialized Veterinary Publication

Entomological Research Letters

ISSN: 3108-639X

2025, Volume 5, Issue 2, Page No: 84-95
Copyright CC BY-NC-SA 4.0
Available online at: www.esvpub.com/

Sampling Design, Genomic Data Types, and Reproducibility Practices in the Expansion of Large-Scale Insect Phylogenomics

Ana Rodrigues[1]*, Tiago Martins[1], Bruno Lopes[2]

[1]Department of Phylogenomic Sampling and Data Types, Faculty of Biosciences, University of Lisbon, Lisbon, Portugal.
[2]Department of Reproducibility in Large-Scale Phylogenomics, Faculty of Life Sciences, University of Porto, Porto, Portugal.

***E-mail** ✉ ana.rodrigues@fc.ulisboa.pt

ABSTRACT

Large-scale phylogenomics has transformed insect systematics by expanding both taxonomic coverage and the number of loci available for inference. Expansion, however, has not produced a uniform methodological landscape. Sampling depth varies among lineages and taxonomic levels; genomes, transcriptomes, and targeted-enrichment datasets differ in locus recovery and reuse potential; missing data arise through several non-equivalent processes; and large matrices can retain gene-tree conflict or analytical sensitivity. This scoping review maps these dimensions while treating reproducibility as distinct from data volume or nominal node support. The reviewed evidence shows that museum collections can expand otherwise difficult sampling, whole-genome and transcriptomic resources support complementary phylogenomic strategies, and target capture can operate from deep to shallow phylogenetic scales. Yet probe divergence, specimen history, matrix filtering, laboratory conditions, and analytical choice can alter the recoverable evidence. We therefore propose that insect phylogenomic studies be interpreted through separate sampling, data-generation, matrix-composition, conflict, and reproducibility dimensions rather than a single notion of genomic completeness. This separation clarifies where increasing scale strengthens inference and where additional validation remains necessary.

Keywords: Insect phylogenomics, Taxon sampling, Whole-genome sequencing, Transcriptomics, Target capture, Reproducibility.

Received: 15 October 2025
Revised: 03 December 2025
Accepted: 04 December 2025

How to Cite This Article: Rodrigues A, Martins T, Lopes B. Sampling Design, Genomic Data Types, and Reproducibility Practices in the Expansion of Large-Scale Insect Phylogenomics. Entomol Appl Sci Lett. 2025;5(2):84-95.
https://doi.org/10.51847/50BwZ5cu1x

Introduction

Genome-scale sequencing has changed insect systematics from an enterprise frequently constrained by a small number of markers into one capable of recovering hundreds to thousands of loci across deep and shallow evolutionary scales. This expansion has created genuine opportunities, but it has also increased the number of design decisions capable of influencing the final phylogeny. Reviews of insect phylogenomics consequently emphasize both the expanded information content of genomic datasets and persistent problems involving sampling, orthology, matrix construction, model choice, and interpretation [1, 2].

The expansion is also taxonomically uneven. A large-scale assessment of published insect phylogenies identified substantial differences in phylogenetic coverage among lineages, showing that accumulation of phylogenetic information does not automatically generate proportional representation of insect diversity [3]. Sampling therefore remains part of the inferential problem rather than merely a logistical preliminary.

© 2025 Entomological Research Letters

Data volume likewise cannot be equated with analytical convergence. In Neuropterida, comparison of ultraconserved elements (UCEs), universal single-copy orthologs, and mitochondrial genomic data produced method- and data-dependent disagreement in parts of the inferred history [4]. Such cases make it necessary to separate the amount of sequence information from the concordance of the histories extracted from it.

Reproducibility adds a further dimension. Repeated molecular phylogenetic computations can differ slightly under particular software and analytical conditions, although computational nondeterminism is not necessarily the dominant source of phylogenetic disagreement [5]. Reproducibility in insect phylogenomics must therefore encompass more than rerunning a tree program: sampling provenance, input data, filtering, locus definition, software specification, deposited intermediate products, and interpretive boundaries all matter.

This review maps those dimensions without constructing an unsupported universal quality score. Its central proposed distinction is between sampling sufficiency, genomic-data availability, matrix recoverability, conflict visibility, and analytical reproducibility. These dimensions may covary, but the evidence does not justify treating them as interchangeable.

Materials and Methods

The review followed scoping-review reporting principles derived from PRISMA-ScR, with emphasis on transparent identification, eligibility assessment, data charting, and explicit separation of review findings from interpretive synthesis [6]. Searches covered publications dated 2017–2025 and were last verified on 25 August 2026. Search logic combined insect or major insect-order terms with phylogenomics, genome-scale phylogeny, whole-genome shotgun, transcriptome, target capture, anchored hybrid enrichment, UCE, and ultraconserved element. Secondary combinations incorporated taxon sampling, museum, museomics, missing data, occupancy, orthology, paralogy, discordance, workflow, code, repository, and reproducibility.

The documented screening universe comprised 35 unique DOI-verified journal records retained during iterative scholarly discovery through publisher/journal platforms, PubMed-indexed records where available, DOI-linked verification, and citation chaining. Because relevance, publication type, publication year, journal eligibility, and duplicate/version status were evaluated during retrieval rather than after importing a pooled bibliographic export, a separate post-export duplicate-removal stage and a separate title/abstract-exclusion stage were not applicable. Thirty-five reports were sought and successfully retrieved sufficiently for eligibility and evidence verification; no logged report was unavailable. All 35 logged records proceeded to evidence charting. Exploratory search hits rejected before meeting the logging criteria were not assigned retrospective counts and are excluded from the review-flow arithmetic.

Data charting separated taxonomic breadth, taxonomic depth, specimen source, genomic data type, target design, locus recovery, matrix treatment, missingness, conflict diagnostics, workflow reporting, repository availability, and reproducibility artifacts. “Not reported” was retained as a reporting state rather than inferred as evidence that a practice was absent.

Results and Discussion

Sampling expansion has occurred through both broader living-taxon representation and renewed access to historical collections. High-throughput extraction applied to thousands of museum insect specimens demonstrates that specimen-rich collections can become practical sources for large phylogenomic projects when historical-DNA processing itself is scaled [7]. This increases potential taxonomic reach but does not imply equivalent DNA quality across specimen ages, preservation histories, tissues, or taxa.

Strategic taxon addition can also alter resolution directly. In leafhoppers, adding poorly studied lineages and previously unavailable mitogenomes improved resolution in parts of the inferred phylogeny [8]. The relevant inference is not that mitochondrial genomes are sufficient for all phylogenetic depths, but that absence of key lineages can constrain the structure being inferred.

Large order-level studies illustrate a second form of expansion. Broad hemipteroid sampling combined thousands of genes with public deposition of sequence and analysis products [9]. Polyneopteran phylogenomics similarly used large gene sets to address historically difficult deep relationships [10]. In Coleoptera, extensive taxonomic and genomic sampling enabled simultaneous phylogenetic and comparative evolutionary analyses [11]. Transcriptome-scale sampling across Lepidoptera likewise recovered thousands of protein-coding loci across

broad superfamily representation [12]. These studies demonstrate scale, but their different taxa, data sources, and analytical objectives prevent a simple cross-order ranking of performance.

Recent studies also show why "taxon coverage" requires more than one denominator. Dense Nomada sampling illustrates lineage-focused phylogenomics at systematic-monograph scale [13], whereas a large Bembidiini dataset retained difficult or method-sensitive relationships despite recovery of more than 1,800 genes [14]. Preliminary Pentatomidae phylogenomics, meanwhile, expanded genomic evidence without eliminating substantial family-level sampling needs [15]. Coverage should therefore be described by taxonomic level, geographic representation, specimen availability, and relationship of the sampled set to the biological question.

Genome and transcriptome data

Whole-genome shotgun (WGS) data increasingly permit large ortholog sets to be assembled without restricting sequencing to a predefined target panel. In diving beetles, WGS-based phylogenomics recovered thousands of orthologs across broad taxon sampling and resolved much of the tree, yet some relationships remained dependent on analytical method [16]. Genomic abundance and topological stability are therefore separable properties.

A related distinction appears when taxonomic representation is unusually complete. Genomic sampling across all extant Hydrophiloidea–Histeroidea families clarified major relationships while retaining substantial conflict within part of the phylogeny [17]. Broad coverage can thus reduce one source of uncertainty without removing heterogeneous histories among genes.

More loci can also influence downstream taxonomic procedures in non-monotonic ways. Whole-genome phylogenomics of Stylops showed that species-delimitation results depended on both method and locus sampling, with increasing locus numbers not producing one invariant taxonomic answer [18]. Genomic scale should consequently not be interpreted as automatic taxonomic decisiveness.

Transcriptomes offer a different sampling route. In Sternorrhyncha, genome-derived and transcriptome-derived samples differed in UCE recovery, and alternative inference frameworks did not produce identical results at every higher-level relationship [19]. Transcriptomic recovery is conditioned by expression and assembly as well as evolutionary conservation, whereas genome-derived recovery is not restricted to expressed sequence in the same manner.

Contemporary insect systematics increasingly combines phylogenomic information with morphology, natural history, biogeography, and other evidence classes [20]. This broadens interpretive context but also makes provenance important: a genomic topology, morphological distinction, and ecological association remain different observations even when ultimately evaluated together.

Matrix composition further mediates genomic inference. Swallowtail-butterfly WGS phylogenomics used a large gene matrix for topology estimation but a reduced, more complete matrix for analyses with different data requirements [21]. Missingness can therefore be partly created or reduced by analytical design rather than being solely an intrinsic property of specimens.

Finally, transcriptomic and target-derived resources can sometimes be integrated. In Apidae, transcriptome and UCE data were combined to extend taxonomic and locus coverage across samples with different genomic resources [22]. Such integration is feasible only when locus homology and matrix construction remain explicit.

Because these studies expose different ways in which genomic abundance becomes usable phylogenetic evidence, the state-and-response ledger below separates the source of genomic information from the inferential state produced after locus recovery and matrix construction.

Table 1. Genomic-data states that separate sequence abundance from phylogenomic usability

Genome/transcriptome state	Taxonomic with observed basis	Uncertainty within its limit
High-density WGS ortholog recovery	Broad beetle sampling Recovered ortholog set and topology Thousands of loci can resolve much of a large phylogeny	Some nodes remain method-sensitive Proposed — code genomic abundance separately from topological stability
Broad family representation with genomic conflict	Higher-level Coleoptera Taxon representation and gene conflict Complete higher-level sampling can coexist with discordant gene histories	Preferred topology does not erase conflict Proposed — sampling sufficiency and conflict visibility are separate states

Whole-genome delimitation sensitivity	Species-rich Stylops Locus number and delimitation output Taxonomic output changes with locus and method choice	Species counts are algorithm-dependent Do not equate additional loci with automatically stronger delimitation
Genome-derived versus transcriptome-derived UCE recovery	Sternorrhyncha Target-locus recovery Input source changes recoverable locus composition	Expression and assembly constrain transcriptomes Record missingness relative to data source
Integrative phylogenomic evidence	Insect systematics broadly Multiple evidence classes Genomic data increasingly coexist with morphological, geographic, and natural-history evidence	Evidence classes need not be concordant Preserve provenance before integrative interpretation
Purpose-specific matrix reduction	Papilionidae WGS Matrix completeness Fewer genes may provide greater completeness for selected analyses	Lower missingness is obtained by discarding information Treat matrix construction as an analytical decision
Cross-source locus integration	Apidae Homologous loci from transcriptome and UCE resources Heterogeneous resources can expand usable sampling	Homology and missingness must be checked explicitly Integration is conditional, not automatic interchangeability

Target-capture data

Target capture occupies a different region of the sampling–data landscape because the genomic search space is specified before sequencing. Anchored phylogenomics in Lepidoptera established that targeted enrichment can recover hundreds of homologous regions across deeply divergent members of a megadiverse insect order [23]. Its scalability derives from repeated acquisition of a defined locus class, but the same design decision constrains what can be recovered.

Bait construction is therefore part of the analytical workflow rather than an invisible laboratory detail. Methods for identifying conserved genomic elements and designing bait sets make target selection explicit and reproducible, while also emphasizing that sequence conservation alone does not establish orthology [24]. A target may hybridize successfully yet still require paralogy and homology assessment before species-history inference.

Probe portability has measurable phylogenetic boundaries. Comparisons among beetle UCE probe sets showed declining utility as phylogenetic distance from probe-design taxa increased and demonstrated additional effects of design breadth and strategy [25]. The evidence supports a conditional relationship between design and recovery, not a universal divergence threshold.

Museum material expands the potential target-capture sampling pool but introduces another source of heterogeneity. UCE phylogenomics of long-horned bees recovered useful genomic information from severely degraded historical specimens when capture and curation were adapted to that material [26]. In Odonata spanning approximately 120 years of museum history, specimen age and body size were poor standalone predictors of locus recovery, whereas capture differed between museum collections [27]. Preservation history and collection context therefore cannot be replaced by specimen age as a simple quality surrogate.

Laboratory parameters can further shift the matrix that reaches phylogenetic analysis. In leaf-footed bugs, lower hybridization temperatures increased recovery under divergent bait–target conditions but reduced some specificity or depth measures and increased signals consistent with potential paralog recovery [28]. Increasing the quantity of recovered targets can therefore carry a countervailing validation cost.

Recent applications demonstrate that target capture spans very different phylogenetic scales. An Orthoptera-focused UCE resource recovered a large targeted locus set for order-level inference [29]. A tailored panel in Noteridae recovered thousands of loci for much shallower diversification among cryptic aquatic-beetle lineages [30]. Target capture is thus not intrinsically a deep-phylogeny technology; its effective scale depends on panel design and sampling.

Sampling completeness within target-capture studies must still be stated at the appropriate level. Miltogramminae phylogenomics covered a substantial proportion of genera while sampling a much smaller fraction of described species, illustrating why genus-level and species-level completeness cannot be merged [31]. In a New Zealand cicada radiation, anchored nuclear data increased resolution but strong support coexisted with gene/site conflict and mito-nuclear discordance [32]. Neither dense target recovery nor high node support uniquely diagnoses the biological cause of discordance.

The selection flow is placed here because target-capture evidence completes the principal data-acquisition classes examined before the review turns to matrix incompleteness. The flow therefore records exactly how the evidence set entering those comparisons was logged and retained.

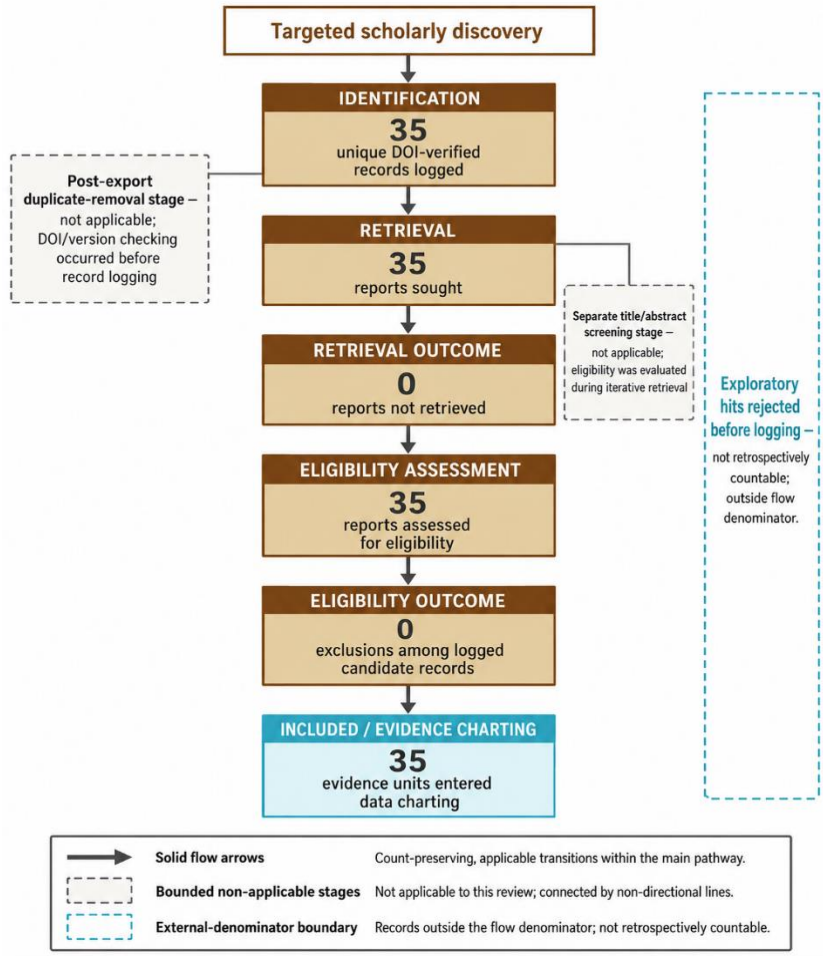


Figure 1. Logged-record boundaries in the phylogenomic evidence search and eligibility process

Missing data

Missing data in insect phylogenomics are not one biological or computational phenomenon. Transcriptome-derived matrices can lack loci because relevant genes were not expressed or assembled, whereas genome-derived matrices may show different recovery patterns [19]. WGS matrices may also be deliberately filtered to increase occupancy for a particular analysis [21]. In target-capture datasets, missingness can emerge because probe sets are phylogenetically distant from target taxa [25], because historical specimens yield heterogeneous DNA recovery [26], or because laboratory capture conditions alter the balance between recovery and specificity [28]. These mechanisms should not be pooled into an undifferentiated "percentage missing."

Matrix incompleteness must also be distinguished from phylogenomic conflict. Incongruence can originate from biological history as well as analytical processes, including incomplete lineage sorting, introgression, gene duplication and loss, model inadequacy, and other sources of heterogeneity [33]. Missing loci, low support, and conflicting well-sampled loci therefore represent different inferential states.

The proposed synthesis used here treats missingness as a pathway-specific property: its interpretation depends on what generated the absent information, whether absence is structured across taxa or loci, and whether the same

region of the tree remains unstable when data completeness changes. This distinction prevents an arbitrary completeness threshold from substituting for causal or analytical diagnosis.

The mechanism–boundary matrix is needed because the same visible outcome—an incomplete matrix—can arise through scientifically different processes that require different tests.

Table 2. Distinguishing pathways to incomplete phylogenomic matrices from conflict and analytical uncertainty

Boundary condition and decisive check	Missing-data with evidential standing	Biological/taxonomic con with observed basis	Interpretation or consequence
Transcript expression and assembly are not genome sequencing **Compare recovery by data source before attributing absence to taxon quality**	Data-source-dependent locus absence Evidence-derived	Genome versus transcriptome inputs Different recoverable locus sets	Matrix completeness depends partly on source architecture
Information is discarded to satisfy another analytical requirement **Report both filtering rule and analysis-specific rationale**	Purposeful occupancy filtering Evidence-derived	Large WGS matrices Reduced matrix contains fewer loci but greater completeness	Missingness can be altered intentionally for analysis
No universal divergence threshold is established **Evaluate probe portability within the focal clade**	Probe-divergence-associated absence Evidence-derived	Target capture across divergent taxa Declining target recovery with bait–target distance	Missingness may be phylogenetically structured
Age alone is an insufficient quality proxy **Record collection and preservation context where available**	Specimen-mediated recovery loss Evidence-derived	Historical museum material Uneven recovery among specimens or collections	Absence can reflect preservation history
Protocol effects are bait- and taxon-dependent **Validate recovery and specificity jointly**	Capture-condition-dependent recovery Evidence-derived	Target-enrichment laboratory workflow Recovery changes alongside specificity/depth or paralog indicators	More recovered sequence need not mean cleaner orthology
Mechanism cannot be assigned from discordance alone **Test alternative biological and analytical explanations**	Well-sampled but conflicting signal Evidence-derived	Rapid radiations or heterogeneous gene histories Discordant genes, sites, data classes, or topologies	Conflict is not equivalent to missingness
Classification has not been prospectively validated **Test whether pathway-aware reporting improves sensitivity analyses**	Pathway-specific missingness classification Proposed synthesis	Cross-study synthesis Missing entries linked to their generating process	Avoids one generic completeness score

Workflow transparency

Phylogenomic reproducibility depends on whether another investigator can reconstruct how biological material became an analytical matrix and how that matrix became an inferred tree. Merely naming a program does not preserve the decisions embedded in filtering, locus selection, software versions, parameters, scripts, and computing environments. Guidance for biological code sharing therefore emphasizes documentation, version control, persistent archiving, licensing, and linkage between code and underlying data [34].

Model specification is an additional layer. Formal machine-readable descriptions can expose probabilistic assumptions that otherwise remain dispersed across prose, configuration files, and software defaults [35]. Such specification does not demonstrate that a biological model is adequate, but it improves the visibility of what was actually assumed.

Transparency should consequently be treated as a chain of reconstructable analytical states rather than a binary code-available/code-unavailable classification. Some insect studies provide raw reads, assemblies, matrices, or other analytical products, but different deposited objects support different levels of reconstruction. The proposed distinction is therefore among named workflow, parameterized workflow, reconstructable workflow, and archived executable workflow. These categories describe reporting states; they are not a validated quality score.

The cross-scale grid below distinguishes what must remain visible as a workflow moves from biological material to computational inference.

Table 3. Workflow states that determine whether insect phylogenomic inference can be reconstructed rather than merely described

Boundary or failure mode and decisive check	Scale of the recorded ob with observed basis	Interpretive consequence
Software named only ***Evidence-based distinction***		
Version, defaults, and parameters may remain unknown **Determine whether the analysis can be reconstructed from prose alone**	Analysis program Program identity	Broad analytical class is recognizable
Versioned software ***Proposed audit state***		
External dependencies may still differ **Reinstall stated version and dependencies**	Program/environment Software release recorded	Narrows implementation uncertainty
Parameterized workflow ***Proposed audit state***		
Manual preprocessing can remain hidden **Recreate input-to-output transformation**	Command/model level Filters, options, models, thresholds, partitions	Decision path becomes inspectable
Scripted workflow ***Evidence-informed state***		
Scripts may depend on unavailable data or software **Execute against deposited inputs**	Multi-step computation Scripts or workflow files	Processing order becomes recoverable
Archived executable workflow ***Proposed synthesis***		
Re-execution still does not validate biological assumptions **Independent rerun plus output comparison**	Code + version + persistent archive Versioned code, documentation, persistent identifier, environment information	Highest practical reconstructability among proposed states
Explicit model specification ***Evidence-derived***		
Explicitness does not guarantee model adequacy **Compare model specification with biological question and sensitivity analyses**	Statistical model Machine-readable or fully declared assumptions	Model structure can be audited separately from implementation

Repository and code availability

Repository practice should be separated into the objects being preserved. Public raw sequence data permit renewed assembly or locus extraction; processed alignments and matrices permit closer replication of the published analysis; tree files preserve inferred outputs; and scripts or workflow descriptions document transformations among these objects. Examples from large insect phylogenomic studies show that raw and processed resources can be deposited publicly, but the presence of one artifact does not demonstrate that every analytical transition is recoverable [9]. Recent WGS studies likewise demonstrate the value of linking large genomic matrices to accessible data products while retaining method-sensitive interpretations [16].

Target-design resources constitute another reproducibility object. Open bait-design procedures make the origin of conserved targets inspectable [24], whereas taxonomic analyses based on large genomic matrices may require both sequence deposits and sufficient documentation of how loci entered downstream delimitation or tree

inference [18]. Code-sharing guidance therefore supports distinguishing persistent, documented, versioned resources from a transient link or a prose statement that software was used [34].

A further caution concerns missing reporting. Failure to locate code or a processed matrix is evidence about reported availability, not proof that the material never existed. The proposed repository architecture therefore codes raw data, derived data, executable logic, software environment, model declaration, and persistent archival identity separately.

These distinctions require a visual representation because reproducibility depends on handoffs among nested objects rather than on a single repository endpoint.

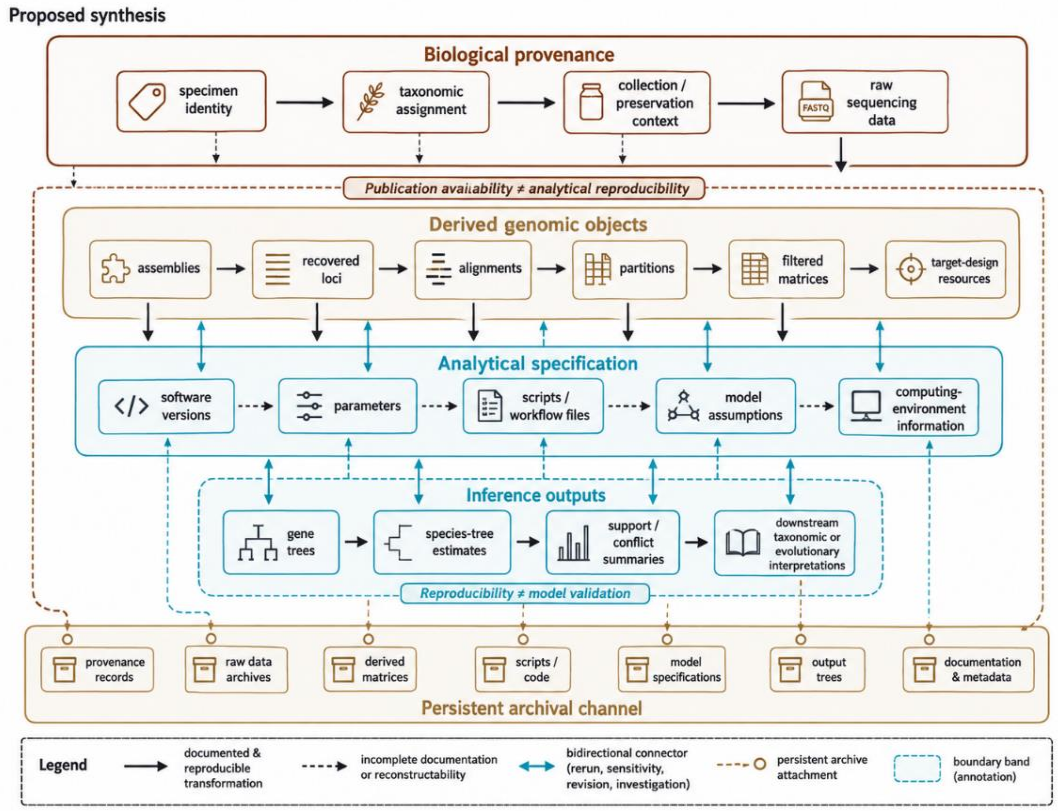


Figure 2. Analytical handoffs that determine whether a phylogenomic result remains reusable after publication

Reproducibility landscape

Reproducibility is therefore multidimensional. Computational repeatability concerns whether the same analysis generates materially similar results when rerun [5]. Workflow reproducibility concerns whether inputs, transformations, parameters, and software can be reconstructed. Repository reproducibility concerns whether the necessary objects remain persistently obtainable [34]. Interpretive reproducibility concerns whether readers can distinguish observed conflict from the explanation assigned to it; phylogenomic incongruence may have several biological or analytical origins [33].

These dimensions need not move together. A large WGS matrix can be openly archived yet contain relationships that vary among analytical approaches [16]. Genome-scale species delimitation can be technically repeatable while remaining sensitive to method and locus sampling [18]. Strongly supported target-capture trees can likewise coexist with gene/site conflict or mito-nuclear discordance [32]. Reproducibility should thus not be treated as synonymous with node support, matrix size, or data availability.

The conditionality matrix below formalizes the article's proposed distinction between reconstructability and inferential robustness without turning either into a numerical score.

Table 4. Failure modes produced when different dimensions of phylogenomic reproducibility are mistaken for one another

Interpretation within its limit	Failure mode if overinte and decisive check	Observable basis
Repeatable computation		
Evidence-derived		
Numerical implementation is comparatively stable **Model and data assumptions may still be wrong**	Repeatability mistaken for biological correctness Sensitivity analysis across defensible models	Similar output from repeated computation
Public raw data only		
Proposed synthesis		
Primary evidence can be reprocessed **Processing choices may be unavailable**	Deposit mistaken for complete workflow transparency Archive processing logic and derived matrices	Sequence reads deposited
Public matrix without reconstructable preprocessing		
Proposed synthesis		
Published tree can be approximately rerun **Filtering and orthology decisions are hidden**	Matrix provenance becomes opaque Link matrix to locus-selection and filtering steps	Alignment or supermatrix available
Reconstructable workflow with model sensitivity		
Evidence-informed		
Analysis can be recreated **Alternative models may yield different relationships**	Successful rerun mistaken for robust inference Report model and matrix sensitivity	Inputs, code, versions, parameters available
High-support but conflict-rich phylogeny		
Evidence-derived		
Dominant signal is identifiable **Gene histories may remain heterogeneous**	Support mistaken for historical unanimity Report concordance/conflict diagnostics	Strong node support plus discordant loci/sites/data classes
Open and conflict-aware analysis		
Proposed synthesis		
Reuse and interpretation are both facilitated **Reproducibility does not prove mechanism**	Remaining uncertainty may still be biological or taxonomic Preserve alternative explanations and validation boundaries	Persistent artifacts plus explicit sensitivity/conflict reporting

Methodological gaps

The first unresolved problem is sampling asymmetry. Published insect phylogenetic coverage remains uneven [3], so further increases in sequencing capacity will not by themselves correct taxonomic imbalance. Sampling audits should distinguish species, genus, family, geography, specimen source, and availability of historical material rather than reporting one overall taxon count.

A second gap is comparability among genomic data types. Neuropteridan analyses demonstrate that different genomic partitions and analytical choices can support conflicting relationships [4]. Genome- and transcriptome-derived UCE datasets likewise differ in recoverability and can respond differently to inference method [19]. Comparative benchmarking should therefore ask when different data types converge, where they diverge, and whether disagreement originates in locus biology, recovery, sampling, filtering, or analytical assumptions.

Target-capture transferability remains similarly conditional. Probe-set utility changes with phylogenetic distance and design breadth [25]. Better reporting of bait provenance, target-to-sampled-taxon divergence, failed recovery, paralog filtering, and laboratory modifications would make apparent missingness more interpretable across studies.

Finally, conflict diagnosis remains a major inferential gap. Phylogenomic incongruence has multiple plausible biological and analytical origins [33]. Treating discordance as noise, or assigning it automatically to incomplete lineage sorting, introgression, or methodological error, exceeds what topology disagreement alone can establish. Future insect phylogenomics would benefit from designs that combine explicit conflict diagnostics with taxon expansion, alternative genomic partitions, model sensitivity, and reproducible workflow preservation.

The broader methodological priority is therefore not maximal data accumulation but traceable expansion: adding taxa, loci, and analytical complexity while preserving enough provenance to determine why an inference changes. This is the principal proposed synthesis of the review and should be tested prospectively rather than treated as an established standard.

Conclusion

Large-scale insect phylogenomics has expanded the evidential reach of systematics, but genomic scale has not made study design, matrix construction, conflict, or reproducibility interchangeable concerns. Taxon sampling can remain uneven even in data-rich clades; genomes, transcriptomes, and target-capture resources impose different recovery constraints; missing data can arise through biologically and analytically different pathways; and high support can coexist with discordance or method sensitivity. The central contribution of this review is a proposed separation of sampling sufficiency, genomic-data availability, matrix recoverability, conflict visibility, workflow reconstructability, and inferential robustness. Treating these as distinct but interacting dimensions provides a more defensible basis for comparing large insect phylogenomic studies than either locus count or nominal openness alone. Future progress depends on increasing phylogenomic scale while preserving the provenance needed to explain what changed, why it changed, and which conclusions remain conditional.

Acknowledgments: None

Conflict of Interest: None

Financial Support: None

Ethics Statement: None

References

1. Young AD, Gillung JP. Phylogenomics—principles, opportunities and pitfalls of big-data phylogenetics. Syst Entomol. 2020;45(2):225-47. doi:10.1111/syen.12406
2. Johnson KP. Putting the genome in insect phylogenomics. Curr Opin Insect Sci. 2019;36:111-7. doi:10.1016/j.cois.2019.08.002
3. Chesters D. Where are the biggest gaps in phylogenetic coverage of insect diversity? Syst Entomol. 2025;50(1):221-36. doi:10.1111/syen.12652
4. Zhang R, Wang L, Tian S, Liu Y, Jiang Y, Zhou X, et al. Inconsistent performance of multi-type genomic data in phylogenomics of neuropteridan insects, with solutions toward conflicting results. Syst Entomol. 2025;50(4):855-75. doi:10.1111/syen.12684
5. Kumar S, Tao Q, Lamarca AP, Tamura K. Computational reproducibility of molecular phylogenies. Mol Biol Evol. 2023;40(7). doi:10.1093/molbev/msad165
6. Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, et al. PRISMA extension for scoping reviews (PRISMA-ScR): Checklist and explanation. Ann Intern Med. 2018;169(7):467-73. doi:10.7326/M18-0850
7. Holmquist A, Tavris H, Kim G, Esposito L, Fisher B, Lam A. Towards large-scale museomics projects: A cost-effective and high-throughput extraction method for obtaining historical DNA from museum insect specimens. Mol Ecol Resour. 2025;25(8). doi:10.1111/1755-0998.14117
8. Liang Z, Zhang J, Dietrich CH, Du Y, Dai W. Sampling poorly studied lineages improves resolution in mitogenome-based phylogenies of leafhoppers (Hemiptera: Cicadellidae). Insect Syst Divers. 2024;8(5):4. doi:10.1093/isd/ixae026

9. Johnson KP, Dietrich CH, Friedrich F, Beutel RG, Wipfler B, Peters RS, et al. Phylogenomics and the evolution of hemipteroid insects. Proc Natl Acad Sci U S A. 2018;115(50):12775-80. doi:10.1073/pnas.1815820115
10. Wipfler B, Letsch H, Frandsen PB, Kapli P, Mayer C, Bartel D, et al. Evolutionary history of Polyneoptera and its implications for our understanding of early winged insects. Proc Natl Acad Sci U S A. 2019;116(8):3024-9. doi:10.1073/pnas.1817794116
11. McKenna DD, Shin S, Ahrens D, Balke M, Beza-Beza C, Clarke DJ, et al. The evolution and genomic basis of beetle diversity. Proc Natl Acad Sci U S A. 2019;116(49):24729-37. doi:10.1073/pnas.1909655116
12. Kawahara AY, Plotkin D, Espeland M, Meusemann K, Toussaint EFA, Donath A, et al. Phylogenomics reveals the evolutionary timing and pattern of butterflies and moths. Proc Natl Acad Sci U S A. 2019;116(45):22657-63. doi:10.1073/pnas.1907847116
13. Straka J, Benda D, Policarová J, Astapenková A, Wood TJ, Bossert S. A phylogenomic monograph of West-Palaearctic Nomada (Hymenoptera: Apidae). Insect Syst Divers. 2024;8(1):1. doi:10.1093/isd/ixad024
14. Pflug JM, Schmidt J, Sproul JS, Weng YM, Maruyama M, Henrichs B, et al. Phylogenomics of the major lineages of Bembidion and related ground beetles (Coleoptera: Carabidae: Bembidiini). Insect Syst Divers. 2024;8(5):9. doi:10.1093/isd/ixae025
15. Genevcius BC. Dissecting Pandora's box: Preliminary phylogenomic insights into the internal and external relationships of stink bugs (Hemiptera: Pentatomidae). Insect Syst Divers. 2024;8(4):1. doi:10.1093/isd/ixae024
16. Bergsten J, Nylander JAA, Ospina OE, Lemmon AR, Miller KB. Whole genome shotgun phylogenomics resolve the diving beetle tree of life. Syst Entomol. 2025;50(4):940-74. doi:10.1111/syen.12685
17. Kusy D, Motyka M, Simon-Pražák J, Lackner T, Prokin A, Fikáček M. Phylogenomics resolves the relationships among Hydrophiloidea-Histeroidea families (Coleoptera) and challenges the single colonization of aquatic habitats. Syst Entomol. 2025;50(4):813-35. doi:10.1111/syen.12679
18. Lähteenaro M, Straka J, Forshage M, Hovmöller R, Nakase Y, Nilsson AL, et al. Phylogenomic species delimitation of the twisted-winged parasite genus Stylops (Strepsiptera). Syst Entomol. 2024;49(2):294-313. doi:10.1111/syen.12618
19. 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
20. Ribeiro TMA, Espíndola A. Integrated phylogenomic approaches in insect systematics. Curr Opin Insect Sci. 2024;61:101150. doi:10.1016/j.cois.2023.101150
21. Allio R, Scornavacca C, Nabholz B, Clamens AL, Sperling FAH, Condamine FL. Whole genome shotgun phylogenomics resolves the pattern and timing of swallowtail butterfly evolution. Syst Biol. 2020;69(1):38-60. doi:10.1093/sysbio/syz030
22. Bossert S, Murray EA, Almeida EAB, Brady SG, Blaimer BB, Danforth BN. Combining transcriptomes and ultraconserved elements to illuminate the phylogeny of Apidae. Mol Phylogenet Evol. 2019;130:121-31. doi:10.1016/j.ympev.2018.10.012
23. 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
24. Faircloth BC. Identifying conserved genomic elements and designing universal bait sets to enrich them. Methods Ecol Evol. 2017;8(9):1103-12. doi:10.1111/2041-210X.12754
25. Gustafson GT, Glynn RD, Short AEZ, Tarasov S, Gunter NL. To design, or not to design? Comparison of beetle ultraconserved element probe set utility based on phylogenetic distance, breadth, and method of probe design. Insect Syst Divers. 2023;7(4):4. doi:10.1093/isd/ixad014
26. Freitas FV, Branstetter MG, Franceschini-Santos VH, Dorchin A, Wright KW, López-Uribe MM, et al. UCE phylogenomics, biogeography, and classification of long-horned bees (Hymenoptera: Apidae: Eucerini), with insights on using specimens with extremely degraded DNA. Insect Syst Divers. 2023;7(4):3. doi:10.1093/isd/ixad012
27. Goodman A, Tolman E, Uche-Dike R, Abbott J, Breinholt JW, Bybee S, et al. Assessment of targeted enrichment locus capture across time and museums using odonate specimens. Insect Syst Divers. 2023;7(3):5. doi:10.1093/isd/ixad011

28. Forthman M, Gordon ERL, Kimball RT. Low hybridization temperatures improve target capture success of invertebrate loci: A case study of leaf-footed bugs (Hemiptera: Coreoidea). R Soc Open Sci. 2023;10(6):230307. doi:10.1098/rsos.230307
29. Jasso-Martínez JM, Castañeda-Osorio R, Mil-Salazar DA, Pintos-Garduño KF. Phylogeny of Orthoptera (Insecta) using targeted ultraconserved element data. Mol Phylogenet Evol. 2025;211:108402. doi:10.1016/j.ympev.2025.108402
30. Baca SM, Gustafson GT, DeRaad DA, Alexander A, Hime PM, Short AEZ. A shallow-scale phylogenomics approach reveals repeated patterns of diversification among sympatric lineages of cryptic Neotropical aquatic beetles (Coleoptera: Noteridae). Syst Entomol. 2025;50(1):82-101. doi:10.1111/syen.12643
31. Johnston NP, Pape T, Piwczyński M, Wallman JF, Wiegmann BM, Cassel BK, et al. Anchored phylogenomics and revised classification of the Miltogramminae (Diptera: Sarcophagidae). Syst Entomol. 2024;49(1):138-55. doi:10.1111/syen.12609
32. 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
33. 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
34. Abdill RJ, Talarico E, Grieneisen L. A how-to guide for code sharing in biology. PLoS Biol. 2024;22(9). doi:10.1371/journal.pbio.3002815
35. Drummond AJ, Chen K, Mendes FK, Xie D. LinguaPhylo: A probabilistic model specification language for reproducible phylogenetic analyses. PLoS Comput Biol. 2023;19(7). doi:10.1371/journal.pcbi.1011226