Eurasia Specialized Veterinary Publication

Entomological Research Letters

ISSN: 3108-639X

2025, Volume 5, Issue 1, Page No: 142-152
Copyright CC BY-NC-SA 4.0
Available online at: www.esvpub.com/

Reproducible Phylogenomic Taxonomy Requires Traceable Links from Insect Genomes to Species Hypotheses and Formal Names

Sara Al-Fahd[1]*, Noura Al-Khalifa[1], Omar Al-Turki[2]

[1]Department of Phylogenomic Taxonomy and Traceable Links, College of Biosciences, King Saud University, Riyadh, Saudi Arabia.
[2]Department of Species Hypotheses and Formal Nomenclature, College of Agriculture, King Fahd University of Petroleum and Minerals, Dhahran, Saudi Arabia.

***E-mail** ✉ s.alfahd@ksu.edu.sa

ABSTRACT

Phylogenomics has expanded the scale of insect systematics, but taxonomic reproducibility requires more than rerunning a tree-building pipeline. Auditability is weakened when specimen identity, sequence processing, orthology, alignment, tree estimation, species-hypothesis formation, and nomenclatural action are disconnected across files and repositories. This article develops a proposed traceability-centered methodology in which each inferential transition remains linked to the biological and computational objects from which it was derived. The approach separates genome quality from taxonomic identity, orthology hypotheses from source sequences, gene trees from species histories, branch support from genomic concordance, and candidate species hypotheses from formal names. It also identifies conditions under which apparently robust trees remain taxonomically insufficient because they depend on processing choices, alignment decisions, model assumptions, discordance, or incomplete specimen anchoring. The methodology is an auditable architecture rather than a universal decision rule: it prescribes no species thresholds and does not claim validated taxonomic performance. Its purpose is to preserve the evidence chain through which genomic observations become progressively stronger taxonomic propositions.

Keywords: Phylogenomics, Insect systematics, Specimen vouchers, Orthology inference, Species delimitation, Taxonomic reproducibility

Received: 27 Month 2025
Revised: 26 Month 2025
Accepted: 29 Month 2025

How to Cite This Article: Al-Fahd S, Al-Khalifa N, Al-Turki O. Reproducible Phylogenomic Taxonomy Requires Traceable Links from Insect Genomes to Species Hypotheses and Formal Names. Entomol Res Lett. 2025;5(1):142-52. https://doi.org/10.51847/FB0aiOqYhB

Introduction

Genome-scale data have transformed insect systematics, yet molecular scale does not eliminate inferential instability. Neuropteridan analyses show that alternative genomic data types and analytical treatments can yield conflicting relationships, while extensive ultraconserved-element sampling in ants still leaves uncertainty at difficult nodes [1, 2]. Reproducibility therefore cannot begin with the final topology while treating preceding decisions as biologically neutral.

The stakes increase when phylogenetic structure becomes species-level interpretation. Molecular delimitation has revealed substantial cryptic insect diversity, showing that lineage-delimitation decisions can materially change recognized diversity [3]. Genomic resolution, however, does not by itself establish which lineages warrant formal names.

Specimen identity creates a further boundary. Type genomics offers a route for linking genomic evidence directly to name-bearing material [4]. Such linkage is not equivalent to species delimitation, but it exposes the central

© 2025 Entomological Research Letters

problem addressed here: a taxonomic proposition should remain traceable to the organism, sequence transformations, phylogenetic inference, and species hypothesis from which it arose.

This article therefore proposes a traceability-centered architecture for phylogenomic taxonomy. Vouchers, raw sequences, processed datasets, orthology assignments, alignments, gene trees, species-tree estimates, species hypotheses, and names are treated as distinct but linkable objects. The architecture is methodological synthesis, not a validated universal standard.

Genome and specimen provenance

Genomic taxonomy begins with an organism rather than an accession. Voucher specimens make that origin retrievable and allow later reassessment of identification, morphology, locality, contamination, or taxonomic interpretation [5]. A voucher is not proof of correct identification; its value is that the biological object remains available for audit. Provenance should therefore connect specimen identifier, repository, tissue or extraction identifier, and sequence record.

Museum genomics extends this principle through historical collections, which can provide genomic material from older, rare, geographically unavailable, or nomenclaturally important specimens, although preservation and DNA recovery vary [6]. Curatorial history and molecular history should therefore remain connected when historical material enters phylogenomic analysis.

Type-derived sequences provide a particularly strong nomenclatural anchor because they associate molecular data with name-bearing material. Explicit identification of such records can improve sequence–name traceability in public databases [7]. Nevertheless, a type sequence answers what genomic material is attached to a name; it does not independently decide whether populations constitute separate species.

The proposed provenance record accordingly distinguishes collector/event information, repository and catalog number, type status where relevant, subsample, extraction, library, sequencing run, assembly, and deposited sequence. These need not use one universal metadata schema, but their relationships should remain resolvable. When a physical voucher is absent, that absence should remain visible rather than being hidden behind an accession number.

Sequence processing provenance

A specimen does not pass unchanged into phylogenetic inference. Reference-genome selection and filtering thresholds can jointly alter downstream phylogenomic results, making preprocessing an inferentially relevant stage [8]. Reproducibility should therefore preserve the exact parameters and versions connecting raw data to the dataset used for orthology or tree inference.

Genome-quality measures belong in this record without acquiring taxonomic meanings they do not possess. BUSCO provides standardized assessment of assembly completeness and redundancy through conserved ortholog sets [9]. Such metrics characterize technical completeness; they do not verify specimen identity or establish species membership. Technical-quality provenance and taxonomic-identity provenance should remain separate even when attached to the same assembly.

Planthopper phylogenomics illustrates how taxon sampling, gene selection, data class, and inference strategy jointly define the analysis that produces an insect tree [10]. The relevant provenance object is therefore not merely "the alignment," but a versioned derivative with recoverable parent data, exclusions, filtering logic, and downstream destination.

Computational reproducibility also requires software versions, dependencies, parameters, workflows, and environments sufficient for analyses to be rerun [11]. Rerunnability remains distinct from scientific validity: an unsuitable choice can be reproduced exactly. The proposed processing layer therefore links each transformed dataset to its parent, transformation, software context, parameterization, and successor object.

Orthology and alignment decisions

Phylogenomic matrices depend on hypotheses about comparable characters. Orthology is inferred from gene-family relationships and duplication histories rather than directly observed as an immutable genome property [12]. Its provenance should therefore include method, version, input data, orthogroup outputs, exclusion rules, and mappings from retained loci to source sequences.

Alignment adds another transformation. Trimming can remove misleading sites, but benchmark evidence shows that its effect depends on dataset and method; greater removal is not synonymous with greater accuracy [13].

Preserving precursor and final alignments makes it possible to determine whether a disputed relationship depends on treatment of ambiguous positions.

Sequence representation is similarly conditional. Simulation and empirical analyses of deep phylogeny show that DNA is not intrinsically inferior to protein data under all circumstances; performance depends on evolutionary depth, model fit, and analytical treatment [14]. Nucleotide, amino-acid, codon, partitioned, or recoded analyses should therefore be recorded as explicit analytical alternatives rather than silent defaults.

The proposed architecture treats orthology assignments and alignments as separate inferential objects. Each retained locus should be traceable backward to source sequence and forward to the tree analysis in which it participates. Excluded paralogs, loci, taxa, or sites need not dominate the manuscript, but their disposition should remain recoverable.

Tree inference

A phylogenetic topology is an estimate produced under a defined procedure. Maximum-likelihood software such as IQ-TREE 2 makes model specification, search settings, software version, and support procedures part of the provenance of the resulting tree [15]. Depositing a Newick file without those links reproduces an output, not necessarily the analytical route that generated it.

Inference level must also remain explicit. ASTRAL-III estimates species trees from gene trees under assumptions intended to accommodate incomplete lineage sorting [16]. Gene trees, concatenated trees, and species-tree estimates are therefore distinct inferential objects. Even a species-tree estimate remains an evolutionary inference rather than a species delimitation or nomenclatural act.

Branch support and genomic agreement are also different. Concordance factors describe the extent to which genes or sites support competing relationships and therefore convey information that conventional branch-support measures do not [17]. A highly supported branch can still sit over heterogeneous genomic signal.

Discordance should likewise retain competing explanations. Empirical phylogenomics shows that gene-tree conflict may reflect analytical artifact as well as biological history [18]. The biological details of a vertebrate example cannot be transferred directly to insects, but the methodological principle is relevant: discordance is an observation requiring diagnosis, not automatic evidence for one mechanism.

These distinctions motivate a proposed tree-to-taxonomy boundary. Before a branch is carried forward as a species hypothesis, its record should identify the dataset, matrix construction, inference level, model, support measure, and material evidence of conflict. No universal numerical threshold is implied. The objective is to explain why superficially similar trees may justify different downstream taxonomic interpretations.

Because prose tends to flatten those conditional states, **Table 1** separates situations in which an apparently robust branch remains stable, ambiguous, or insufficient for species-level interpretation.

Table 1. When phylogenomic branches remain taxonomically conditional despite apparent analytical support

Processing, orthology, support, and gene-tree sensitivity				
Phylogenomic comparison dimension	**Processing-sensitive topology**	**Orthology- or alignment-dependent branch**	**Gene-tree/species-tree level mismatch**	**High support with limited concordance**
Taxonomic context with observed basis	Relationships vary with reference choice, filtering, or matrix construction Alternative processed datasets recover different placements or support	Homology assessment is difficult or matrix construction changes Placement shifts with orthology, trimming, representation, or locus exclusion	Multilocus data contain discordant locus histories Gene trees and species-tree estimates differ or support different resolutions	Large matrices contain heterogeneous genomic signal Conventional support is high but concordance reveals competing histories
Downstream taxonomic claim	Branch remains conditional on upstream data construction	Taxonomic interpretation remains contingent on character construction	State the relationship at the level actually estimated	Preferred topology should not be equated with uniform genomic agreement
Uncertainty that remains	Effect is dataset-specific	No single method removes uncertainty universally	Species-tree estimates inherit	Concordance is not a probability of correctness

			model and gene-tree assumptions	
Robustness test and proposed state	Re-run defensible preprocessing alternatives and preserve dataset lineage Evidence-supported condition	Retain locus-to-source mapping and compare defensible matrices Evidence-supported condition	Preserve inferential level and test sensitivity Evidence-supported condition	Report support and concordance for their distinct meanings Evidence-supported condition
Discordance diagnosis and tree-to-taxonomy transition				
Phylogenomic comparison dimension	**Discordance with competing causes**	**Tree-to-taxonomy transition**		
Taxonomic context with observed basis	Conflict may reflect history or analytical artifact Alternative gene trees covary with information, treatment, or plausible processes	A lineage is considered for species-hypothesis formation Dataset, model, level, support, concordance, and conflict provenance are documented		
Downstream taxonomic claim	Do not assign a biological cause without discriminating evidence	Lineage becomes an explicit candidate hypothesis, not a formal taxon		
Uncertainty that remains	Similar conflict patterns can arise differently	No universal support or divergence threshold is prescribed		
Robustness test and proposed state	Compare analytical and biological explanations Evidence-supported condition	Record the transition as a new linked inferential object Proposed boundary rule		
Support-concordance stress test				
Phylogenomic comparison dimension	**Stable topology across documented alternatives**			
Taxonomic context with observed basis	Relationship persists across defensible analyses Similar placement survives predefined sensitivity analyses			
Downstream taxonomic claim	Stability reduces concern about one analytical artifact but does not establish a species			
Uncertainty that remains	Sampling, structure, gene flow, and species concepts remain unresolved			
Robustness test and proposed state	Carry lineage forward for species-hypothesis testing Proposed interpretive classification			

Species-hypothesis formation

A phylogenetic lineage becomes taxonomically meaningful only after the inferential question changes from "what relationship is estimated?" to "what biological entity is being hypothesized?" Multispecies-coalescent delimitation can recover population structure rather than species, and explicit migration can alter the interpretation of lineage separation [19, 20]. Consequently, neither reciprocal monophyly nor a coalescent partition should be treated as a self-executing species decision. The same branch may represent deep population structure, geographically restricted gene flow, recent divergence, or a lineage that merits species-level treatment; the tree alone does not adjudicate among those interpretations.

One defensible response is to treat delimited units as hypotheses that must survive comparison with alternatives. Integrative work on *Desmognathus* demonstrates the utility of beginning with phylogeographic partitions and

testing them against additional evidence rather than treating the initial genomic partition as the endpoint [21]. This logic transfers methodologically to insects without transferring salamander-specific thresholds or biological expectations. The important transferable principle is procedural: candidate lineages should enter an explicit hypothesis-testing stage rather than crossing directly from a phylogeny into nomenclature.

Uncertainty can also be retained formally. Probabilistic approaches make species assignment conditional on stated criteria, models, and competing hypotheses [22]. For the present methodology, this supports an explicit species-hypothesis object containing the candidate lineage, evidence used to define it, model or criterion under which it was recovered, plausible alternatives, and unresolved uncertainty. Importantly, probability, support, concordance, and diagnostic distinctness remain different quantities; none is converted here into a universal species threshold.

In insects, the *Acromyrmex octospinosus* complex provides a direct demonstration that phylogenomic signal can acquire different taxonomic meaning when combined with population, morphological, geographic, and historical evidence [23]. More broadly, integrative taxonomy increasingly combines heterogeneous "taxon-omic" evidence, but decisions about compatibility, weighting, and conflict remain methodological rather than automatic [24]. Unsupervised learning can assist discovery of candidate partitions across multivariate evidence, yet its clusters remain dependent on selected features and analytical design [25]. Machine-assisted clustering therefore belongs upstream of the taxonomic decision, as a way of generating or comparing hypotheses rather than authorizing names.

Evidence classes should also remain identifiable after integration. Morphological diagnosability, geographic coherence, genomic differentiation, reproductive evidence, ecological association, and historical nomenclatural information need not agree, and disagreement can itself be informative. The proposed architecture therefore avoids a single composite score. Instead, each species hypothesis carries an evidence profile that records concordant, conflicting, absent, and method-dependent observations. This prevents strong evidence in one domain from silently substituting for evidence that was never collected in another.

A further requirement is lineage-identity continuity. The candidate species hypothesis should be traceable to the exact individuals or vouchers included in the genomic partition and to the tree object from which that partition was interpreted. If taxon sampling changes, the hypothesis receives a new version rather than silently inheriting the meaning of an earlier partition. This is particularly important in insect groups with geographically sparse sampling, historical museum material, or cryptic diversity, where the limits of a candidate lineage may change as additional populations are incorporated.

The proposed species-hypothesis stage therefore sits between tree inference and nomenclature. It receives lineages, support, concordance, population structure, gene-flow considerations, candidate partitions, and independent evidence, but it outputs a testable taxonomic proposition rather than a name. A hypothesis can be retained, revised, merged, split, or left unresolved without rewriting upstream genomic evidence. Conversely, contradictory evidence should feed backward to the relevant analytical stage when it suggests that the apparent lineage depends on processing, orthology, alignment, taxon sampling, or tree-model decisions.

The distinction is sufficiently central that it is represented as a nested evidence field rather than a linear pipeline in **Figure 1**; the figure makes explicit both the upstream feedback routes and the boundary beyond which a candidate lineage cannot move merely because it is genomically well supported.

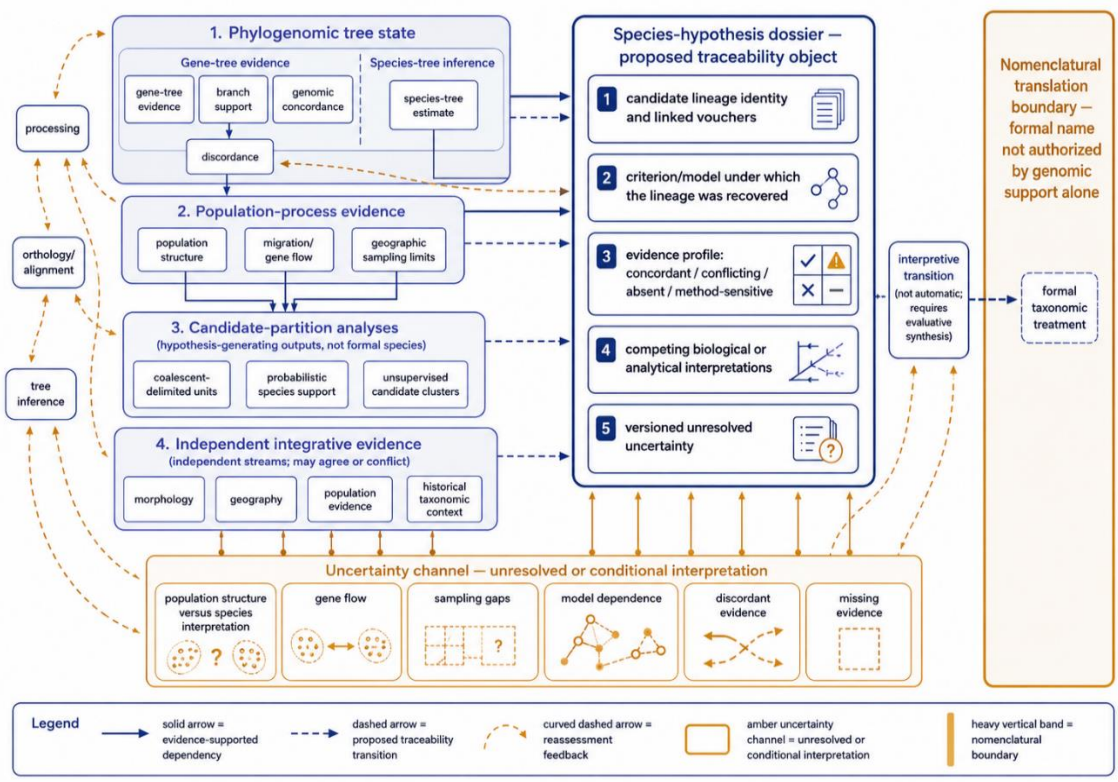


Figure 1. The evidence boundary between phylogenomic lineages and taxonomically actionable species hypotheses

Nomenclatural translation

Species delimitation and formal taxonomy are related but non-identical operations. Contemporary discussions of next-generation taxonomy emphasize that a delimited lineage may be treated differently depending on species concept, evidential context, nomenclatural history, and practical taxonomic judgment [26]. The transition from species hypothesis to name therefore requires information not contained in a tree: whether an existing name applies, whether type material anchors that name to the focal lineage, whether competing synonyms or historical concepts exist, and whether diagnostic evidence is adequate for the contemplated taxonomic act.

The relevant translation object is not simply a species label attached to a terminal branch. It is a documented interpretation linking a candidate lineage to the nomenclatural evidence that justifies a particular treatment. When type-derived sequence information exists, it can narrow that interpretation, but absence of such information does not license assignment by genomic proximity alone. Conversely, genomic differentiation from a sequenced type does not by itself establish which unsampled populations belong inside or outside the resulting species concept.

This translation can fail in more than one direction. A genomic lineage may be strongly supported yet remain nomenclaturally unresolved because type identity is uncertain. Conversely, an existing name can be securely anchored while the genomic boundary surrounding it remains poorly sampled. Integrative revision of the *Acromyrmex octospinosus* complex illustrates why genomic, morphological, geographic, and nomenclatural evidence must be reconciled rather than treated as interchangeable [23]. Historical names can also embody taxonomic concepts that predate genomic sampling, so the act of matching a lineage to a name may require reinterpretation of older descriptions, types, localities, and synonymies rather than a simple database lookup.

The proposed methodology therefore treats nomenclatural translation as its own recorded transformation. The input is a species hypothesis with explicit uncertainty; the output may be assignment to an existing name, resurrection or synonymy where independently justified, description of a new taxon where evidential and nomenclatural requirements are met, or deliberate retention of unresolved status. Importantly, “no nomenclatural action” is a legitimate output. Deferral preserves a testable hypothesis without converting uncertainty into premature formal stability.

Table 2 separates these states because identical genomic evidence can lead to different nomenclatural consequences depending on name-bearing material, historical usage, diagnostic evidence, and sampling completeness.

Table 2. Nomenclatural translation states that prevent genomic lineages from becoming names by default

Translation state	Type or name anchor	Genomic evidence entering translation	Permissible process and consequence	Residual uncertainty, audit record, and state
Matched anchor				
Candidate lineage with matching type anchor	Type or reliably linked name-bearing material	Species hypothesis and diagnostic evidence are concordant with the existing name concept	Test identity against type-linked evidence and historical usage Apply an existing name if the total evidence supports that interpretation	Population limits or unsampled variation may remain Type or voucher identifier, evidence comparison, decision rationale Evidence-supported transition logic
Ambiguous anchor or multiple lineages under one name				
Candidate lineage with ambiguous type association	Existing name is plausible but its type relationship is unresolved	Genomic lineage is recoverable but nomenclatural identity is uncertain	Resolve type identity before changing formal usage Retain provisional or unresolved treatment	Genomic support cannot substitute for name anchoring Competing name hypotheses and missing evidence Proposed audit state
Multiple genomic lineages under one historical name	One name currently spans differentiated candidate lineages	Genomic partitions require integrative evaluation	Test morphology, geography, population evidence, and type placement Revise only where evidence supports distinct taxonomic treatment	Structured populations may mimic species partitions Evidence streams, conflicts, type assignment, sampling limits Evidence-supported problem; proposed record structure
Named taxa without genomic separation				
Existing named taxa lack genomic separation	Separate names exist but sampled genomes do not recover the expected partition	Genomic evidence conflicts with current classification	Examine sampling, identification, gene flow, model dependence, and non-genomic evidence Maintain, synonymize, or defer only after broader taxonomic assessment	Absence of genomic separation is not automatically synonymy Negative evidence, conflicting evidence, and sampling limits Proposed audit state
Inadequate anchor and deferred translation				
Candidate lineage lacks adequate nomenclatural anchor	Distinct genomic hypothesis but no secure existing name assignment	Species-level evidence may be substantial	Search nomenclatural history and type material; obtain additional evidence where necessary New description or unresolved status depending on evidence	A lineage is not automatically a new species Name-availability assessment and diagnostic basis Proposed boundary rule
Translation deferred	Evidence streams remain materially contradictory or incomplete	Candidate species hypothesis is retained but not stabilized	Preserve hypothesis and uncertainty without formal taxonomic change No nomenclatural action	Delay avoids converting unresolved inference into formal certainty Explicit unresolved questions and validation needs Proposed boundary rule

Data, code, and voucher linkage

The traceability problem becomes operational only when identifiers can cross repositories. Voucher and type-associated records make sequence identity more auditable because a molecular record can be reconnected to

biological material [5, 7]. The same principle should extend to raw reads, assemblies, orthology outputs, alignments, trees, scripts, software environments, species-hypothesis records, and nomenclatural acts. A public accession without the links among these objects preserves data availability but not necessarily the complete inferential chain.

Three kinds of linkage are needed. Biological linkage connects sequence-bearing material to a voucher, repository, collecting event, and type status where relevant. Computational linkage connects every derived file to its input, software environment, parameters, and version. Interpretive linkage connects a tree or candidate partition to the species hypothesis and later taxonomic treatment that used it. Losing any one of these linkages creates a different reproducibility failure: the organism cannot be re-examined, the analysis cannot be reconstructed, or the reasoning that converted an output into a taxonomic proposition cannot be recovered.

Semantic interoperability can strengthen the third linkage for non-genomic evidence. Formal models and ontologies for insect anatomy demonstrate how morphological terminology can be represented in machine-readable, consistently referenced form [27]. Such formalization does not establish homology or species boundaries, but it can make the evidential objects used in integrative taxonomy easier to discover, compare, and connect to specimen records.

The proposed architecture therefore favors persistent identifiers, explicit parent–child relationships among analytical objects, versioned code and environments, repository-stable voucher identifiers, and cross-links among genomic and taxonomic records. A provenance graph need not reside in one database: resolvable identifiers can connect distributed repositories. Missing links should be represented as missing rather than silently inferred, because the visibility of an evidential gap is itself part of reproducibility.

Reproducibility architecture

The proposed architecture can be summarized as an auditable chain of transformations rather than a chain of increasingly authoritative outputs. Its objects are: biological specimen; genomic observation; processed dataset; orthology and alignment hypothesis; gene-tree or species-tree estimate; species hypothesis; nomenclatural interpretation; and formal name where justified. Each transition records the evidence consumed, analytical operation performed, alternatives considered, output identifier, and unresolved uncertainty. The architecture therefore preserves both data lineage and inference lineage.

A minimal transition record would contain six logical elements: source-object identifiers; transformation or interpretive action; method, software, or decision criterion; resulting-object identifier; uncertainty or competing interpretation; and responsible provenance, such as repository, version, or taxonomic publication. These are proposed information requirements, not a new mandatory metadata standard. Existing repositories can implement them through cross-links rather than duplicating their underlying records.

Two rules keep this architecture from becoming a taxonomic scoring system. First, reproducibility is not validity. Executable workflows improve computational rerunnability [11], but they cannot make an unsuitable specimen identification, reference choice, model, or species concept correct. Second, evidence strength is stage-specific. High phylogenetic branch support does not erase discordance [17], integrative evidence can alter the interpretation of a genomic lineage [23], and species delimitation remains distinct from nomenclatural translation [26]. A later stage may therefore question an earlier interpretation without invalidating the observation itself.

The architecture is deliberately reversible. If a later morphological observation, newly sequenced type, expanded geographic sample, or model comparison contradicts a species hypothesis, the record should identify which upstream transition must be reconsidered. Revision need not overwrite the earlier state; versioned hypotheses can preserve how the taxonomic interpretation changed and why. Ontology-based identifiers can further stabilize the language used for morphological evidence across such revisions [27]. This creates an audit trail of taxonomic reasoning rather than a succession of disconnected final answers.

Boundary conditions remain important. Historical specimens may provide limited DNA; type material may be unavailable or unsuitable for destructive sampling; repositories differ in identifier persistence and machine readability; older names may lack precise locality information; genomic sampling can be geographically incomplete; and some species hypotheses may remain difficult under continuing gene flow or recent divergence. The proposed architecture does not resolve those limitations. It records where they enter the inference chain so that downstream certainty is not detached from upstream constraints.

A final distinction concerns the granularity of traceability. Not every exploratory script, discarded tree, or intermediate file needs equal prominence in a publication, but any decision that can change the biological or

nomenclatural interpretation should remain recoverable. This includes changes in taxon inclusion, locus retention, orthology assignment, alignment treatment, model selection, species-delimitation criterion, evidence weighting, and name application. The proposed architecture therefore prioritizes consequential transitions rather than attempting to archive every computational event as equivalent scientific evidence.

This also provides a practical way to represent disagreement. Two taxonomists may accept the same vouchers, sequences, and tree yet differ at the species-hypothesis or nomenclatural stage because they apply different species concepts or interpret conflicting evidence differently. A traceable record does not eliminate that disagreement; it identifies where it occurs. Reproducibility in this setting means that another researcher can reconstruct both the shared evidence and the interpretive divergence. That distinction is particularly important for taxonomic revisions, where transparency about why conclusions differ may be more scientifically useful than concealing disagreement behind a single final classification.

Implementation can therefore vary among museums, genomic repositories, code platforms, sequence databases, and taxonomic registries. The methodological minimum is logical rather than platform-specific: every taxonomically consequential assertion should be traceable backward to its evidential objects and forward to the interpretation it generated. A useful implementation should allow a reader to start with a formal name and recover the species hypothesis, tree, alignment, orthology decisions, processed and raw sequence objects, and voucher information, while also allowing movement in the opposite direction from specimen to current taxonomic interpretation.

Validation priorities follow directly from that goal. Future implementations should test whether independent researchers can reconstruct cross-repository links, rerun computational stages where feasible, locate the relevant voucher or type information, distinguish evidence-supported from proposed transitions, identify conflicts that were retained rather than suppressed, and recover why a species hypothesis did or did not become a formal name. Such testing would evaluate traceability and reproducibility; it would not, by itself, establish that the taxonomic conclusion is biologically correct.

Conclusion

Reproducible phylogenomic taxonomy requires more than reproducible tree estimation. It requires continuity between the physical insect, genomic data, analytical transformations, phylogenetic objects, species hypothesis, and nomenclatural decision. The proposed architecture preserves these as distinct linked stages so that uncertainty, conflict, and revision remain visible. It prescribes no universal species threshold and does not make genomic resolution synonymous with taxonomic certainty. Its central requirement is stricter than simple data deposition: every formal taxonomic interpretation should remain recoverably connected to the evidence, analytical choices, alternative explanations, and inferential transitions that produced it.

Acknowledgments: None

Conflict of Interest: None

Financial Support: None

Ethics Statement: None

References

1. 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
2. Borowiec ML, Zhang YM, Neves K, Ramalho MO, Fisher BL, Lucky A, et al. Evaluating UCE data adequacy and integrating uncertainty in a comprehensive phylogeny of ants. Syst Biol. 2025;74(5):700-22. doi:10.1093/sysbio/syaf001
3. Li X, Wiens JJ. Estimating global biodiversity: The role of cryptic insect species. Syst Biol. 2023;72(2):391-403. doi:10.1093/sysbio/syac069

4. Letsch H, Greve C, Hundsdoerfer AK, Irisarri I, Moore JM, Espeland M, et al. Type genomics: A framework for integrating genomic data into biodiversity and taxonomic research. Syst Biol. 2025;74(6):1029-44. doi:10.1093/sysbio/syaf040
5. Buckner JC, Sanders RC, Faircloth BC, Chakrabarty P. The critical importance of vouchers in genomics. Elife. 2021;10:e68264. doi:10.7554/eLife.68264
6. Card DC, Shapiro B, Giribet G, Moritz C, Edwards SV. Museum genomics. Annu Rev Genet. 2021;55:633-59. doi:10.1146/annurev-genet-071719-020506
7. Renner SS, Scherz MD, Schoch CL, Gottschling M, Vences M. Improving the gold standard in NCBI GenBank and related databases: DNA sequences from type specimens and type strains. Syst Biol. 2024;73(2):486-94. doi:10.1093/sysbio/syad068
8. Rick JA, Brock CD, Lewanski AL, Golcher-Benavides J, Wagner CE. Reference genome choice and filtering thresholds jointly influence phylogenomic analyses. Syst Biol. 2024;73(1):76-101. doi:10.1093/sysbio/syad065
9. Manni M, Berkeley MR, Seppey M, Simão FA, Zdobnov EM. BUSCO update: Novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. Mol Biol Evol. 2021;38(10):4647-54. doi:10.1093/molbev/msab199
10. Deng J, Stroiński A, Szwedo J, Ghanavi HR, Yapar E, Castillo Franco D, et al. Phylogenomic insights into the relationship and the evolutionary history of planthoppers (Insecta: Hemiptera: Fulgoromorpha). Syst Entomol. 2025;50(3):495-518. doi:10.1111/syen.12666
11. Grüning B, Chilton J, Köster J, Dale R, Soranzo N, van den Beek M, et al. Practical computational reproducibility in the life sciences. Cell Syst. 2018;6(6):631-5. doi:10.1016/j.cels.2018.03.014
12. Emms DM, Kelly S. OrthoFinder: Phylogenetic orthology inference for comparative genomics. Genome Biol. 2019;20(1):238. doi:10.1186/s13059-019-1832-y
13. Steenwyk JL, Buida TJ 3rd, Li Y, Shen XX, Rokas A. ClipKIT: A multiple sequence alignment trimming software for accurate phylogenomic inference. PLoS Biol. 2020;18(12):e3001007. doi:10.1371/journal.pbio.3001007
14. Kapli P, Kotari I, Telford MJ, Goldman N, Yang Z. DNA sequences are as useful as protein sequences for inferring deep phylogenies. Syst Biol. 2023;72(5):1119-35. doi:10.1093/sysbio/syad036
15. 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
16. 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
17. Lanfear R, Hahn MW. The meaning and measure of concordance factors in phylogenomics. Mol Biol Evol. 2024;41(11):msae214. doi:10.1093/molbev/msae214
18. 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
19. Sukumaran J, Knowles LL. Multispecies coalescent delimits structure, not species. Proc Natl Acad Sci U S A. 2017;114(7):1607-12. doi:10.1073/pnas.1607921114
20. Kornai D, Jiao X, Ji J, Flouri T, Yang Z. Hierarchical heuristic species delimitation under the multispecies coalescent model with migration. Syst Biol. 2024;73(6):1015-37. doi:10.1093/sysbio/syae050
21. Pyron RA, O'Connell KA, Duncan SC, Thom G, Brumfield RT, Beamer DA. Speciation hypotheses from phylogeographic delimitation yield an integrative taxonomy for seal salamanders (Desmognathus monticola). Syst Biol. 2023;72(1):179-97. doi:10.1093/sysbio/syac065
22. Hua X, Moritz C. A phylogenetic approach to delimitate species in a probabilistic way. Syst Biol. 2025;74(5):723-39. doi:10.1093/sysbio/syaf004
23. Mera-Rodríguez D, Fernández-Marín H, Rabeling C. Phylogenomic approach to integrative taxonomy resolves a century-old taxonomic puzzle and the evolutionary history of the Acromyrmex octospinosus species complex. Syst Entomol. 2025;50(3):469-94. doi:10.1111/syen.12665
24. Karbstein K, Kösters L, Hodač L, Hofmann M, Hörandl E, Tomasello S, et al. Species delimitation 4.0: Integrative taxonomy meets artificial intelligence. Trends Ecol Evol. 2024;39(8):771-84. doi:10.1016/j.tree.2023.11.002

25. Pyron RA. Unsupervised machine learning for species delimitation, integrative taxonomy, and biodiversity conservation. Mol Phylogenet Evol. 2023;189:107939. doi:10.1016/j.ympev.2023.107939
26. Vences M, Miralles A, Dufresnes C. Next-generation species delimitation and taxonomy: Implications for biogeography. J Biogeogr. 2024;51(9):1709-22. doi:10.1111/jbi.14807
27. Girón JC, Tarasov S, González Montaña LA, Matentzoglu N, Smith AD, Koch M, et al. Formalizing invertebrate morphological data: A descriptive model for cuticle-based skeleto-muscular systems, an ontology for insect anatomy, and their potential applications in biodiversity research and informatics. Syst Biol. 2023;72(5):1084-100. doi:10.1093/sysbio/syad025