Insect Systematics and Diversity, 2026, **10**, ixag021
DOI: https://doi.org/10.1093/isd/ixag021
Research

ENTOMOLOGICAL SOCIETY OF AMERICA
SHARING INSECT SCIENCE GLOBALLY

OXFORD

Methods

Unsupervised machine learning for species discovery in *Eurytoma* and *Phylloxeroxenus* (Hymenoptera: Eurytomidae) parasitoids of oak gall wasps

Christian L. Weinrich*[,1], Kristýna Bubeníková[1,2], Sofia I. Sheikh[1], MaKella J. Steffensen[1], Anna K.G. Ward[1], Yuanmeng Miles Zhang[3], and Andrew A. Forbes[1]

[1]Department of Biology, University of Iowa, Iowa City, IA, USA
[2]Department of Zoology, Faculty of Science, Charles University, Prague, Czech Republic
[3]Institute of Ecology and Evolution, University of Edinburgh, Edinburgh, UK

*Corresponding author. Department of Biology, University of Iowa, 129 E Jefferson St, Iowa City, IA 52245, USA (Email: christian-weinrich@uiowa.edu).

Subject Editor: Julian Dupuis

Species discovery (inferring species limits de novo, without a priori hypotheses) from genetic data has become more common as molecular tools have expanded and has been a helpful initial step in tackling the taxonomic impediment for small insects. Often species discovery involves a single locus (eg *mitochondrial cytochrome oxidase I* [*mtCOI*]), but the accessibility of techniques for large sub-genomic sequencing projects (1000 s of loci) makes it possible to approach molecular species discovery with more robust datasets. Here, we test unsupervised machine learning (UML) methods for species discovery on a set of ultra-conserved element loci for a large collection of parasitic wasps reared from North American oak galls, all initially thought to be in genus *Eurytoma* Illiger. UML methods produced species hypotheses that largely aligned with those that emerge from a commonly used *mtCOI*-based species partitioning method, and that also tended to match existing species descriptions. Results revealed a new genus-level association with oak galls (*Phylloxeroxenus* Ashmead) hidden among the *Eurytoma*, 2 distinct lineages of *Eurytoma*, including a new lineage of *Eurytoma* more closely related to the South American genus *Kavayva* Zhang, Gates, and Silvestre, evidence for one or more cryptic *Eurytoma* species, and a mix of generalist and specialist host ranges. We make recommendations for how best to employ UML methods to similar datasets.

Keywords: Cynipini, parasitic wasps, taxonomic impediment, UCEs

Introduction

Parasitic wasps are tremendously species-rich but relatively under-studied animals, with estimated millions of undiscovered and undescribed species globally (Burke and Sharanowski 2024). The impediment for discovering, delimiting, and ultimately describing species is more extreme for parasitic wasps than for many other animals (LaSalle and Gauld 1992). While reproductive isolating barriers—the traits that reduce or prevent gene flow between species—can often be studied directly in larger-bodied organisms, most parasitic wasp groups are prohibitively small and therefore rarely studied in nature, making assessment of species limits via ecological assays or behavioral observation difficult. Morphological differences are critical for species description but can be similarly problematic for delimiting parasites because it may be unclear whether morphological variation is intra- vs interspecific, while at the same time, one may have many hundreds or even thousands of specimens to interrogate (Godfray et al. 1999). Morphological description also requires taxon-specific expertise and/or accessible keys, which, for some parasitic wasp groups, are lacking. These difficulties require systemic change, but current problems impeding new species descriptions do not preclude continuing efforts toward new species discovery, particularly when those efforts employ molecular data.

Discovery of new putative species using molecular markers is not a new endeavor. *Mitochondrial cytochrome oxidase I* (*mtCOI*) is now routinely used as an animal "DNA barcode" (Hebert et al. 2003) for parasitoid species discovery (Smith et al. 2008, Fagan-Jeffries et al. 2018, Ward et al. 2020, Awad et al. 2025, Dietenberger et al. 2025) and is also used alongside morphology for new species descriptions (eg Fernandez-Triana et al. 2023; sometimes controversially [Sharkey et al. 2021]). *MtCOI* sequencing has been useful in species discovery because it offers an initial look at potential species-level differences and has long been a relatively low-cost molecular approach. But today, sequencing hundreds or thousands of loci, or even entire genomes, for large numbers of individuals is increasingly affordable and technologically straightforward. Furthermore, these same large genomic datasets are often collected for the purpose of addressing

Received: 14 October 2025. **Revised:** 25 February 2026. **Accepted:** 26 April 2026

evolutionary or ecological questions separate from species discovery. Considering that the primary downsides of using *mtCOI* for initial species discovery are linked to it being just a single locus (Funk and Omland 2003, Llopart et al. 2005), it seems logical to use sub-genomic datasets for species discovery as a replacement for (or complement to) single-locus data, particularly if those datasets have already been collected.

There are several methods to analyze multi-locus datasets for genetic species delimitation (that is, testing a priori species hypotheses using genetic data). However, many of them are not well-suited to the task of species discovery (ie inferring species hypotheses de novo) and may be especially challenging for a group like parasitic wasps. Many species delimitation programs such as BPP (Flouri et al. 2018) are not suitable for discovery because they require an a priori population assignment that has a sizable and direct effect on the outcome of the final delimitation. Some others, such as SODA (Rabiee and Mirarab 2021), require multiple individuals per species, which still necessitates an implicit a priori species concept and is difficult to guarantee when specimens are rare and hard to collect. Programs that use a multi-species coalescent model (MSC) to infer species limits cannot distinguish between true species divergence and intra-specific population structure (Jackson et al. 2017, Sukumaran and Knowles 2017). MSC methods are also known to over-split continuous geographic clines (Chambers and Hillis 2020). Consequently, these programs are prone to over-split taxa with detectable population structure, a probable scenario for parasitic wasps given their typically low population sizes, limited dispersal, and widespread but patchy distributions (Quicke 2012). Moreover, many of these methods rely on Markov Chain Monte Carlo simulations, which are computationally expensive, and may have additional, non-trivial requirements (eg an ultrametric tree; Fujisawa and Barraclough 2013) that are also computationally expensive and difficult to obtain. This often limits the size of datasets, restricting the total number of loci, number of samples, or both. As a result, these methods may fail to take full advantage of multi-locus/sub-genomic data.

Both as a supplement to and replacement for species delimitation programs, a number of research projects have turned to unsupervised machine learning (UML) methods for clustering specimens based on genetic data (eg Derkarabetian et al. 2019, Mussmann et al. 2020, Martin et al. 2021). Machine learning refers to algorithms that can improve in their ability to perform a task by training on representative data without user input. UML refers to algorithms that do not rely on labeled data provided by the user for training. Consequently, some of the UML algorithms that have been used for species *delimitation* are appealing for species *discovery* because these algorithms do not require a priori hypotheses about the samples, populations, or model of species divergence.

In this study, we test the utility of several candidate UML approaches for species discovery in an assemblage of North American eurytomid wasp parasites reared from galls induced by oak gall wasps. We focus on wasps in the genus *Eurytoma* Illiger (Hymenoptera: Eurytomidae). Genus *Eurytoma* is an excellent candidate for exploring novel species discovery methods because taxonomic resources are wanting, and there may be many undescribed species. The primary North American *Eurytoma* key (Bugbee 1967) is notorious for its tricky taxonomy and lack of supporting figures. Moreover, the subfamily to which *Eurytoma* belongs (Eurytominae) exhibits considerable morphological homoplasy (Lotfalizadeh et al. 2007), and the generic concepts for *Eurytoma* and other Eurytomidae require revision (Zhang et al. 2025). As a case in point, some samples in our study turned out to be in the genus *Phylloxeroxenus* Ashmead, which superficially resembles *Eurytoma* and had not previously been recorded from North American oak galls (see Results). Most recent studies involving North American *Eurytoma* have employed *mtCOI* barcodes in species discovery (eg Zhang et al. 2014, Forbes et al. 2016).

Our research group is interested in the evolutionary histories of several genera of oak-gall-associated parasitic wasps, including *Eurytoma*, and to this end, we have been sequencing ultra-conserved elements (UCEs; Faircloth et al. 2012) for unidentified wasps reared from known gall hosts. Thus, we have a large UCE dataset for more than 100 individual eurytomids reared from a diversity of oak galls collected across the continental United States but limited a priori sense of which or how many species they represent. Here, we apply 3 UML methods to UCE data for these wasps and evaluate the results to discover putative species. We then evaluate the morphology of representatives of each putative species using existing keys and, for a subset of samples, compare our results to a more standard *mtCOI* species discovery workflow.

Methods

Sample Rearing/Selection

All eurytomid wasps used in this study were reared from galls induced by gall wasps on oak trees (*Quercus*). Oak gall wasps (Hymenoptera: Cynipidae, Cynipini) oviposit into meristematic tissue of oak trees and induce the formation of a gall, inside which the gall wasp larva(e) develop (Stone et al. 2002, Egan et al. 2018). Gall morphology, location, and oak species association are all specific to the gall wasp species, and in North America, oak gall wasp species richness is estimated to exceed 700 species, many of which remain undescribed (Melika et al. 2021). The oak gall wasps, and/or their galls, are host to several common genera of parasitic wasps, all of which are even less well known than their host gall wasps (Ward et al. 2022a). *Eurytoma* reared from oak galls are assumed to be parasites of the gall wasp larva, but could alternatively be feeding on gall tissue or on other gall inhabitants.

Galls were collected from trees, and galls induced by oak gall wasps of the same species from the same tree were placed together in cups and given a collection-event-specific identifier code (Ward et al. 2022b). We placed rearing cups in an incubator set to mimic seasonal temperature, humidity, and light–dark cycles and monitored cups daily for up to 2 years. Emerging insects were captured and preserved for genetic work in individual tubes with 95% EtOH. Collections are described in more detail in Ward et al. (2022a). We collected >4,000 individual *Eurytoma* wasps from galls induced by >60 species of oak gall wasp (Ward et al. 2022a). Because we did not know a priori which or how many species were in this collection, we selected samples for sequencing by choosing individuals that spanned the diversity of host galls, host tree species, and geographical locations. This resulted in a sample of 102 wasps (Supplementary Table S1).

We photographed the lateral habitus and a forewing for each specimen and then destructively extracted DNA from whole wasp bodies using a modified CTAB approach (Chen et al. 2010). Though we did not preserve bodies of the specific specimens from which DNA was extracted, other wasps had often emerged from the same galls on the same or a nearby day and

had the same morphology as those we had photographed and extracted. We used these individuals as morphological surrogates for destructively extracted wasps. Surrogate wasps were pinned and deposited into the collection of the University of Iowa Museum of Natural History (Supplementary Table S2). We recommend that groups interested in replicating our methods not follow our lead in destructively extracting DNA and instead use a non-destructive technique.

We prepared UCE libraries using Kapa HyperPlus v5.19 kits (Kapa Biosystems Inc., Wilmington, Massachusetts, United States) with an enzymatic approach to the initial fragmentation step (Branstetter et al. 2017). After library amplification, we used a double-sided size-select bead clean to enrich for fragments between 300 and 500 bp (SPRI-Based size Selection, Beckman Coulter Inc.) and verified fragment sizes on an Agilent 2100 Bioanalyzer (Agilent, Santa Clara, California, United States). We pooled and hybridized libraries using the Hym v2P bait set (Branstetter et al. 2017) and confirmed enrichment using relative qPCR. We sequenced the library on one lane of a NovaSeq6000 SP flowcell (2 × 150 bp; Illumina, Inc., San Diego, California, United States) at the University of Iowa Institute of Human Genomics.

UCE Processing

UCE sequences were processed using phyluce v1.7.3 (Faircloth 2016). Reads were adapter and quality trimmed using illumiprocessor (Faircloth 2013), a wrapper around Trimmomatic (Bolger et al. 2014). Trimmed reads were assembled into contigs using SPAdes v3.13.0 (Prjibelski et al. 2020).

Unphased UCE Data

After contig assembly, UCE loci were extracted, aligned, and used to create an unphased data matrix following Zhang et al. (2025). Briefly, loci were aligned via MAFFT (Katoh and Standley 2013), and alignments were internally trimmed via Gblocks (Castresana 2000) with the following settings: b1 = 0.5, b2 = 0.5, b3 = 12, and b4 = 7. All newly generated samples had >1,500 recovered loci.

In addition to the 102 focal eurytomid samples sequenced and analyzed for this study, raw reads for 55 specimens representing each of the major clades in the Eurytomidae tree created by Zhang et al. (2025) were downloaded from SRA, then processed and assembled with the same UCE workflow. The UCE loci for the additional samples were then combined with UCE loci for the 102 focal samples and used to create an unphased 80% completeness data matrix with 1,395 loci included. This unphased UCE matrix was used to construct the maximum likelihood phylogeny.

Maximum Likelihood Phylogeny

The unphased 80% completeness UCE data matrix (1,395 loci, 1,025,134 bp) was used to construct a maximum likelihood phylogeny using IQ-TREE v2.3.6 (Minh et al. 2020). IQ-TREE was run with the ModelFinder (Kalyaanamoorthy et al. 2017) option, which automatically determines the best model for nucleotide evolution, 1000 ultrafast bootstrap (Hoang et al. 2018) replicates to assess support along with "-bnni" to reduce the risk of overestimation, and a Shimodaira–Hasegawa approximate likelihood-rate test (SHaLRT, Guindon et al. 2010) with 1,000 replicates. Only nodes with support values of UFB ≥ 95 and SH-aLRT ≥ 80 were considered robust. The phylogeny was rooted on subfamily Rileyinae (*Rileya* Ashmead and *Neorileya* Ashmead).

Based on visual examination of the tree, the focal eurytomid wasps fall into 3 distinct clades: clade "A," which is most closely related to the samples from the Zhang et al. (2025) *Phylloxeroxenus* clade; clade "B," which is most closely related to *Kavayva* Zhang, Gates, and Silvestre; and clade "C," which is nested within the *Eurytoma sensu stricto* clade from Zhang et al. (2025). In addition to these alphabetic labels, we arbitrarily numbered sub-clades within B and C for ease of discussion and because STRUCTURE *K* evaluation requires some a priori population assignment (see below).

For all subsequent analyses, we created 4 sample partitions, 1 containing all gall-derived samples ("ABC"), and 1 for each clade alone ("A" only, "B" only, and "C" only). Although a separate sample partition was created for clade A, because its position on the tree and the results from STRUCTURE suggest A comprises a single population/species, we did not perform separate UML analyses on A.

Phased UCE Data

Trimmed reads were re-mapped to assembled UCE contigs using the phyluce Mapping Snakemake workflow, which relies on SAMtools (Li et al. 2009) and BWA-MEM (Li 2013). Mapped reads for the 102 focal wasps were haplotype-phased via the phyluce phasing workflow. The phasing workflow is a reimplementation of the haplotype phasing approach described in Andermann et al. (2019). Briefly, reads are aligned to assembled loci, and allelic variants are identified and phased using SAMtools (Li et al. 2009) and Pilon (Walker et al. 2014).

Following examination of the maximum likelihood tree, we constructed matrices of 75%, 95%, and 100% completeness using phased UCEs from each of the 4 sample partitions (ABC, A, B, and C) using a series of custom Python scripts, resulting in 12 different datasets for downstream analysis. Here, completeness refers to UCE loci where at least *X*% of specimens had data at that locus. Because we assume male specimens are haploid, we dropped any loci from male specimens if they had been resolved as heterozygous after phasing. Phased UCE loci were used to produce the SNP datasets.

SNP Extraction

SNPs were extracted from the phased UCE data matrices using SNP-sites (Page et al. 2016). SNP-sites was run with the -c option, which discards columns containing gaps or ambiguous nucleotides. The extracted SNPs were saved as FASTA alignments. Using a custom Python script, a single SNP from each UCE locus was selected to avoid linkage between SNP sites from within the same locus. The first SNP from each locus was chosen (as opposed to a random SNP) to improve reproducibility. Selected SNPs were then saved in various formats for downstream species delimitation analyses. This process was repeated for all 12 of the phased UCE data matrices. The SNP data matrices were used as the input for the species discovery methods (STRUCTURE and UML methods).

Species Discovery

While UML methods are appealing for the task of species discovery because they do not require a priori species hypotheses, it should be noted that the methods do not implement explicitly biological models and are not guaranteed to produce

species-level clusters. Because of this, and because phylogenetic signal is inherently hierarchical, we first chose to perform UML on the full set of focal specimens (ABC datasets) and then on subsets of samples comprised of individual monophyletic clades that contained apparent phylogenetic substructure (B only and C only datasets).

STRUCTURE

Extracted SNPs were converted to STRUCTURE-compatible input format using a custom Python script. For STRUCTURE analysis, males were treated as diploids with 1 allele marked as missing. STRUCTURE was run with a burn-in period of 100,000 steps and 1,000,000 MCMC steps after burn-in. STRUCTURE runs were repeated 5 times for each *K* value, with *K* values ranging from 2 to 10. STRUCTURE results were analyzed using Evanno's delta *K* (Evanno et al. 2005) as provided by the StructureSelector web server (Li and Liu 2018). For the purposes of sorting output, specimens were assigned to their alphanumeric sub-clades (eg B1, B2, C1, C2, etc.).

Unsupervised Machine Learning

In the context of species discovery/delimitation, UML is used to perform dimensionality reduction, taking samples in a high-dimensionality genetic space and producing a lower-dimensionality latent space representation such that samples are easier to visualize and cluster. Models are trained on the original high-dimensional genetic data, and "learning" thus refers to the iterative process of reducing the amount of information that is lost by moving from the original dimensionality to the latent space. Here we use 3 UML approaches, random forests (RF; Breiman 2001), *t*-distributed stochastic neighbor embedding (*t*-SNE; van der Maaten and Hinton 2008), and variational autoencoders (VAE Kingma and Welling 2022), for the task of species discovery from UCE data.

For UML analyses, males were treated as haploids in preliminary runs. However, this resulted in the UML methods primarily clustering samples based on sex. To prevent this, males were instead treated as though they were diploids, which were homozygous at all sites. Because the UML methods do not implement explicitly biological models, we do not expect the treatment of males as homozygous diploids has meaningfully altered the result, other than to prevent males and females from clustering separately.

All UML analyses used the SNPs extracted from the phased UCEs. No additional information, such a guide tree or a priori species groups, are provided to the UML models. UML analyses were performed on ABC, B, and C SNP datasets (A-only datasets were not included). Besides the sample partitioning based on inclusion of different clades and the VAE train-test split described below, datasets were not further partitioned. All UML analyses were replicated 10 times per dataset.

RF and t-SNE

RF is a type of ensemble learning that uses bootstrap aggregation ("bagging") to sample the original data (Breiman 1996). Each bootstrap sample is used to produce a classification tree under a null hypothesis of no genetic structure. The "out of bag" data points are then processed with the classification tree to produce a pairwise, binary-state proximity matrix. This is repeated for each bootstrap, and a final proximity matrix is produced by summing the distance for each pair across all bootstraps and dividing by the total number of bootstraps.

t-SNE is a method of nonlinear dimensionality reduction that begins by creating a similarity matrix by modeling the similarity of pairs of data points in the original high-dimensional space as probabilities in a Gaussian distribution such that more similar points have higher probability. Next, points are randomly placed in the latent space, and the similarity of data points is modeled under a *t*-distribution. The algorithm then iteratively shifts points about in the latent space to minimize the divergence between the 2 similarity matrices.

RF and *t*-SNE were trained on the phased SNP data via an R script based on the script written by Derkarabetian et al. (2019) and modified by Martin et al. (2021). Further modifications to the script include removal of the PCA/DAPC analyses, a different function to perform classical multidimensional scaling (cMDS), and a different approach to *K* selection and clustering (described below). RF predictions were based on majority votes from 10,000 decision trees (ntrees = 1e4). RF results were ordinated using both classical and isotonic multidimensional scaling (isoMDS). *T*-SNE replicates were run for a maximum of 20,000 iterations each with a perplexity value of 10, chosen based on the results of a perplexity grid search approach used by Martin et al. (2021).

Variational autoencoders

A VAE is composed of 2 neural networks, the encoder and the decoder, which are trained simultaneously. The encoder takes the original data and reduces it to a Gaussian distribution in latent space. The decoder attempts to reconstruct the original data from the latent space distributions. The divergence between the original data and the reconstruction is used to score the latent space representation and tune the neural networks.

For VAE analysis, SNPs were converted to a one-hot encoding via a custom Python script. VAE was implemented using a Python script originally written by Derkarabetian et al. (2019), but with the early stopping callback modifications made by Martin et al. (2021). Each replicate was performed with a maximum of 100,000 epochs. For each VAE replicate, samples from the data were randomly subdivided by the script into "train" and "test" sets, wherein "train" sets were used to train the model, and the "test" sets were withheld. The test sets were then used to assess the performance of the model and check for overfitting to the train sets.

K selection and clustering

In the versions of the script written by Derkarabetian et al. and Martin et al., RF and *t*-SNE outputs were clustered using 2 algorithms, partitions around medoids (PAM) and hierarchical clustering (HC), and *K* selections were based on 3 different statistical indices, highest mean silhouette width, gap statistic, or Bayesian information criterion. However, in preliminary runs of the analyses, we found that automating the *K* selection based on these indices often led to inconsistent and incorrect results, yet manual selection based on visual inspection of the index plots was time-consuming and introduced opportunity for human error and bias. Instead, we took an alternative approach based on the NbClust package in R (Charrad et al. 2014). NbClust provides the ability to perform clustering (via *k*-means or one of several HC methods) over several *K* values

and then evaluate each *K* value using up to 30 different cluster number indices in a single function. We found that automated *K* selection was more reliable for individual indices when using NbClust, and that the addition of many more indices further improved the apparent consistency of *K* selection results.

For RF results, the proximity matrix was first ordinated using cMDS or isoMDS. Each ordination was then clustered and evaluated using NbClust using all 30 available indices (index = "alllong"). The final *K* value and clustering scheme for each ordination was the *K* value chosen by the greatest number of indices. For *t*-SNE, we used the same NbClust approach, but using 26 of the available indices (index = "all"), excluding the 4 indices with the longest computation time. For both RF and *t*-SNE, we repeated the NbClust step using 4 clustering methods: *k*-means and 3 HC methods, "average," "complete," and "centroid."

VAE results were clustered by DBSCAN (Ester et al. 1996) as implemented in the vae_dbscan.py script written by Martin et al. (2021). DBSCAN derives clusters using density-based clustering based on a distance threshold, ε, instead of an a priori *K* value. We used the default value for ε (2× the SD, averaged globally).

Comparison of UML Results with *mtCOI* and Morphology

Mitochondrial cytochrome oxidase I

For comparing UCE-based species discovery methods to more common single-locus-based methods, we also sequenced a 650 bp fragment of the *mtCOI* gene for 27 of the same samples. We used sample-specific combinations of indexed primers with distinct 13 bp oligonucleotides attached to the primers using a combination of LCO1490 (5′-GGTCAACAAATCATAAAGATATTGG-3′) and a degenerate version of HCO2198 (5′-TAAACTT CWGGGTGWCCAAAAAATCA-3′) (Folmer et al. 1994). The PCR conditions were an initial denaturation step at 95 °C for 2 min, followed by 30 cycles of 95 °C for 15 s, 48 °C for 15 s, and 72 °C for 30 s, and a final elongation at 72 °C for 1 min. We pooled amplicons and sequenced them with the Ligation Sequencing Kit V14 (SQK-LSK114) with a R10.4.1 Flongle Flow Cell on a MinION Mk1B sequencer (Oxford Nanopore, Oxford UK). Samples were processed using a lab standard demultiplexing pipeline that uses MiniBar (Krehenwinkel et al. 2019) for demultiplexing and primer trimming and VSEARCH (Rognes et al. 2016) for quality filtering, chimeral removal, and read clustering. Nine samples were sequenced on an ABI 3500 (Applied Biosystems, Foster City, California, United States). For an additional 5 samples, we extracted full or partial *mtCOI* sequence from sequenced UCE libraries using the phyluce command phyluce_assembly_match_contigs_to_barcodes (Faircloth 2016). The final set of samples for the *mtCOI* comparison (GenBank accessions: PZ015861-PZ015901) included all of the alphanumeric subclades except for C3.

mtCOI sequences were aligned with Geneious 8.1.9 using the Geneious Alignment Builder. The *mtCOI* alignment was used with ASAP, which produces a species partition based on HC of pairwise sequence differences for single-locus data (Puillandre et al. 2021). Results of ASAP are reported in Table 1. The *mtCOI* phylogeny was made using IQ-TREE v2.3.6 (Minh et al. 2020) with the ModelFinder (Kalyaanamoorthy et al. 2017) option and 1000 ultrafast bootstrap (Hoang et al. 2018) replicates to assess nodal support. IQ-TREE was also run with the -bnni and -czb options to reduce model violations and collapse zero-length branches, respectively.

Morphology

We used photographs of destructively extracted specimens, pinned bodies of surrogates (representatives from the same collections as the specimens that were sequenced), keys by DiGiulio (1997) and Bugbee (1967), and original species descriptions to match results from species discovery methods to existing morphological species. Morphological analyses were performed independently by authors KB and AAF and then results compared.

Results

ML Tree

The ML tree (Fig. 1) shows that the focal eurytomids fall into 3 distinct clades, clade "A," which is most closely related to *Phylloxeroxenus* samples from Zhang et al. (2025), clade "B," which is most closely related to genus *Kavayva*, and clade "C," which is nested within the *Eurytoma sensu stricto* clade. For ease of discussion, we further numbered sub-clades within each of the B and C clades.

STRUCTURE

ABC (All Samples)

Visual assessment of the STRUCTURE plots and Evanno's delta *K* support *K* = 3 in all completeness datasets (Supplementary Figs S1 to S9). The 3 clusters correspond to the A, B, and C clades.

A clade

Delta *K* values support either *K* = 6 (75% and 95% datasets; Supplementary Figs S12 and S15) or *K* = 7 (100%; Supplementary Fig. S18). However, these values are driven by sharp declines in mean lnP(*K*) at *K* = 5 or *K* = 6, and graphs of mean lnP(*K*) show only slight increases between *K* preceding the declines, suggesting that delta *K* may not be a useful metric for this dataset (Supplementary Figs S11, S14, and S17). Visual inspection of the STRUCTURE plots (Supplementary Figs S10, S13, and S16) shows that at *K* = 2, replicates show 1 of 2 patterns in all 3 completeness datasets: all samples are either partially assigned to both populations, or Eur_680_019_13B is assigned to a second population separate from all other A samples. Sample 680_019_13B is notable as the sample in A with the most missing loci and the longest branch on the ML phylogeny compared to the rest of A (Fig. 1). For all *K* higher than 2, additional populations result in spurious partial assignments across many samples. Together, we take this to support *K* = 1 for clade A.

B clade

STRUCTURE results for B are inconclusive. Although *K* values chosen by mean lnP(*K*) are higher (Supplementary Figs S20 and S23), Evanno's delta *K* supports *K* = 3 for the 75% and 95% B datasets (Supplementary Figs S21 and S24). Delta *K* for the 100% dataset supports *K* = 8, but, as with clade A, this is again attributable to a steep drop in mean lnP(*K*) between *K* = 8 and *K* = 9 (Supplementary Figs S26 and S27). There is a second peak in delta *K* at *K* = 4, but this is weakly

Table 1. Summary of results from UML, ASAP, and morphology, overlaying geography and ecology of samples

Clade (from UCE phylogeny)	Species hypothesis (UML/UCE)	ASAP (COI)	Morphology	US states	Host genera
A	A	A	*Phylloxeroxenus* sp.	IA, MO, OH, KY, TX	***Andricus***, ***Disholcaspis***, ***Melikaiella***, ***Neuroterus***, ***Phylloteras***
B1	B1	B1	*Eurytoma hecale* Walker 1843	PA, KY, IA, IL, WI, MO, TN	*Acraspis, Amphibolips, Callirhytis, Druon, Dryocosmus*, ***Kokocynips***, ***Melikaiella***, ***Philonix***
B2	B2+B3	B2	*Eurytoma querci* Fullaway 1912 (syn: *E. californica*)	OH, IA	*Andricus*, ***Disholcaspis, Neuroterus***
B3		B3	*Eurytoma californica* Ashmead 1887	CA	*Andricus*
B4	B4	B4	*Eurytoma prunicola* Walsh 1870	IA, OH, PA, TN	*Amphibolips*, ***Andricus***
B5	B5+B6	B5	*Eurytoma sphaera* Bugbee 1967	IA, IL	***Andricus***
B6		B6	*Eurytoma brevivena* Bugbee 1958	AZ, KY, OH	*Disholcaspis*
B7	B7+B8	B7	no morphological ID	AZ	***Disholcaspis***
B8		B8a/B8b	no morphological ID	IA, TX	***Callirhytis***
C1	C1	C1	*Eurytoma querciglobuliI* (Fitch 1859)	IA, IL, KY, MO, TN	*Disholcaspis*
C2	C2+C3+C4	C2a/C2b	*Eurytoma studiosa* Say 1836	IA, WI	*Acraspis, Andricus, Callirhytis*, ***Neuroterus***
C3		n/a (no *mtCOI* sequence)	*Eurytoma studiosa* Say 1836	ME	*Amphibolips*
C4		C4	*Eurytoma studiosa* Say 1836	AZ, CA	*Andricus, Callirhytis, Xanthoteras*
C5	C5	C5	*Eurytoma studiosa* Say 1836	IA, KY, MO, NY, PA, TN, TX	*Acraspis, Andricus, Amphibolips, Callirhytis, Druon*, ***Neuroterus***, ***Philonix***

"U.S. States" and "Host genera" indicate the locations of collections for samples in this paper and are not inclusive of collections of the same morphological species from other studies. Genus names in bold type are novel hosts for this species.

supported, and the plot of mean lnP(K) is essentially flat before the drop at $K = 9$.

At $K = 3$, visual inspection of STRUCTURE plots (Supplementary Figs S19, S22, and S25) shows B1-3 as primarily assigned to 1 population, B4-5 primarily assigned to a second, B7 + 8 assigned in split proportions to the first 2 populations, and all samples in B1-3 and B6-8 also having partial assignment to a third population. STRUCTURE plots show many spurious populations with low assignment across many or all samples, starting at $K = 2$ and increasing as K increases. We note that a recurring pattern seen in replicates at various K values and in all datasets is a B1, B2-3 + B5-6, B4, B7-8 split. This pattern is notable for its similarity to the B 100% UML results.

C clade

Mean lnP(K) and Evanno's delta K both support $K = 3$ for the 95% and 100% datasets (Supplementary Figs S32, S33, S35, and S36). However, Evannos's delta K supports $K = 6$ for the 75% dataset (Supplementary Fig. S30), and it appears that the high K for 75% is due to a sharp decrease in mean lnP(K) between $K = 6$ and $K = 7$ (Supplementary Fig. S29). Visual inspection of the STRUCTURE plots and the graphs for mean lnP(K) and delta K for values less than $K = 7$ suggest $K = 3$ is likely a more appropriate value for the 75% dataset (Supplementary Figs S28 to S30). At $K = 3$, the STRUCTURE plots for 95% and 100% (Supplementary Figs S31 and S34) show C1, C2, and C5 forming distinct populations. C3 and C4 are primarily assigned to the same population as C2 but also show partial assignment to the other clusters. This is true for 3/5 replicates for 75% as well, although 2 replicates show C1 + C5 as 1 population, C2 to 4 as a second, and partial assignment of all samples to a third.

UML Species Discovery

Overview

Results were variable across datasets and across methods (Figs 2 to 4). In general, datasets with lower completeness resulted in greater differences between methods and more variability across replicates within methods. Overall, C datasets produced results with the greatest agreement among different methods, and B produced results with the least agreement. The combined ABC datasets were intermediate. Of the 3 UML algorithms, RF was the most consistent across replicates, both in terms of individual cluster assignment and range of best K. VAE was least consistent and generally produced the highest K values. All clustering algorithms produced identical or highly similar results, although this varied by dataset, with 75% and 95% B dataset results varying more than others. When clustering methods deviated, K-means was slightly more likely to produce a different result compared to the HC methods. The ordination method appeared to have minimal effect on RF results, although minor differences for a particular ordination + clustering method can be seen in most datasets.

ABC

For all 3 completeness datasets, RF and VAE both support a separation of the A, B, and C clades (Fig. 2; Supplementary Figs S37 and S38). For 75% and 95%, t-SNE results cluster A (excluding sample 680_019_13B), C5, and 680_019_13B +

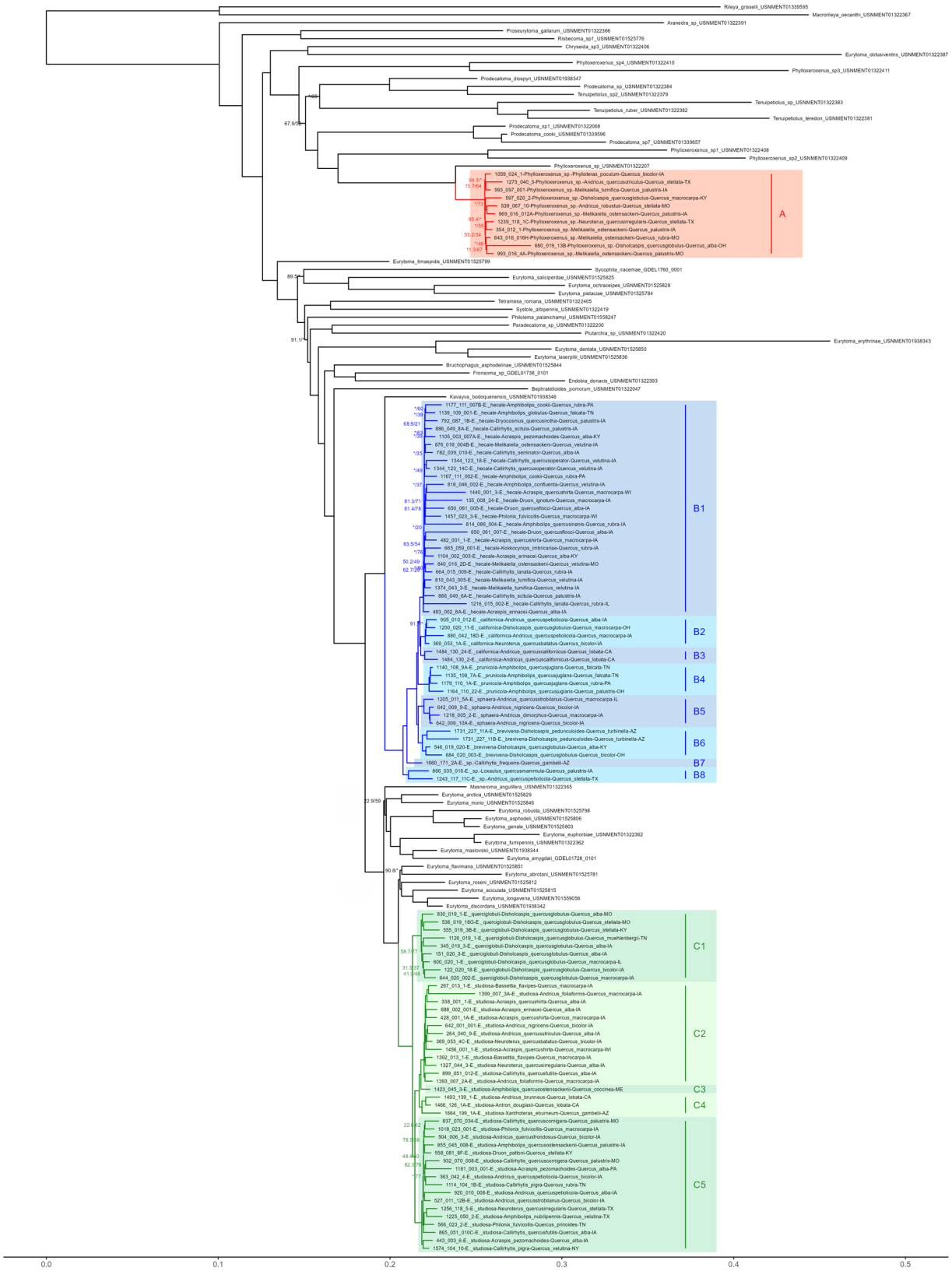

Fig. 1. Maximum likelihood UCE phylogeny. Phylogeny based on 80% completeness matrix of 102 focal *Eurytoma*, plus 55 Eurytomid samples from Zhang et al. (2025). Focal *Eurytoma* tip labels include numeric specimen ID, morphological ID, host gall, host tree, and state. Node support is based on 1000 UFB replicates and 1000 SHaLRT replicates. Bootstrap values are displayed as "[UFB]/[SHaLRT]" and only show UFB scores ≤95 and SHaLRT scores ≤80.

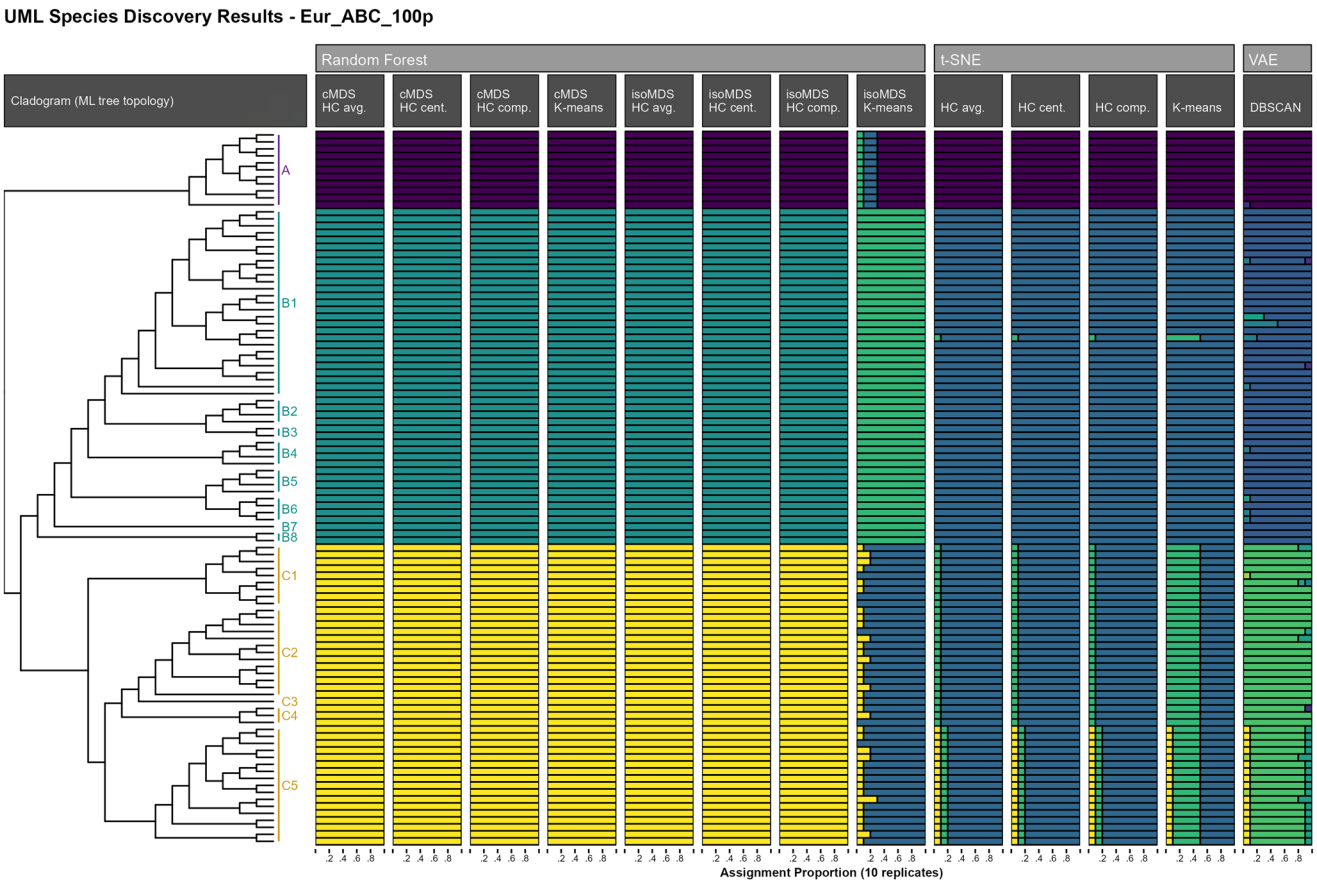

Fig. 2. UML species discovery results—all samples (A + B + C), 100% data completeness. The cladogram shows the ML tree topology and the position of the alphanumerically labeled clades. Each column represents the results of a UML dimensionality reduction plus clustering method. Horizontal bars within each column represent the proportion of times an individual was assigned to a given cluster across replicates, with different colors representing different clusters. Horizontal bars are aligned to the tips of the cladogram. Within each method (column), clusters were compared pairwise across replicates via normalized mutual information (NMI), and the replicate with the highest average NMI was chosen as the reference replicate. Cluster labeling (coloring) was then standardized to the reference replicate via the Hungarian method. Cluster labeling is not standardized between different methods.

B1-8 + C1-4 as the major division, with some individual replicates suggesting further divisions within that pattern (Supplementary Figs S37 and S38). At 100% completeness, *t*-SNE separates A and B+C, although about half of the *K*-means replicates support the A, B, and C division (Fig. 3).

B clade

B datasets produced results that are difficult to interpret, especially for the lower completeness datasets where there is little agreement between methods and across replicates (Supplementary Figs S39 and S40). At 100%, there is better agreement between methods and across replicates, but the most common clusterings in the RF results are B1, B2-3 + B5-8, B4 or B1, B2-3 + B5-6, B4 + B7-8, both of which contain polyphyletic groups given the ML tree topology (Fig. 3). The *t*-SNE results are similar, although B7 and 1 member of B8 tend to be placed with B1. VAE results are inconclusive.

C clade

The results of RF and *t*-SNE for all 3 datasets support a delimitation of C into 3 groups, C1, C2-4, and C5, with some methods suggesting a further separation of C4 or C3-4 from C2 (Fig. 4; Supplementary Figs S41 and S42). The latter delimitation is clearest in the cMDS + *K*-means results for the 100% dataset, but similarities can be seen in the lower completeness datasets as well (Supplementary Figs S41 and S42). Due to the higher *K* values and greater inconsistency across replicates, VAE results are harder to interpret, although they do not appear incompatible with either a C1, C2-4, C5 or C1, C2, C3-4, C5 clustering.

UML Synthesis

After analyzing the results of all clustering methods and the phylogeny, we identified a clustering that is at least partially supported by UML methods, is concordant with the tree topology, and approximates species-level differences. In general, we gave greater weight to the higher completeness datasets and to the HC results when there were inconsistent results between clustering methods. We also tended to give greater weight to RF when there were inconsistent results between UML methods because RF results were less likely to be polyphyletic and thus easier to reconcile with the tree (although see results for B). In the specific case of the B data, we chose to reconcile the 100% UML results with the tree topology by splitting polyphyletic clusters into monophyletic groups. Therefore, based on the results of phylogenetic, STRUCTURE, and UML analyses, we

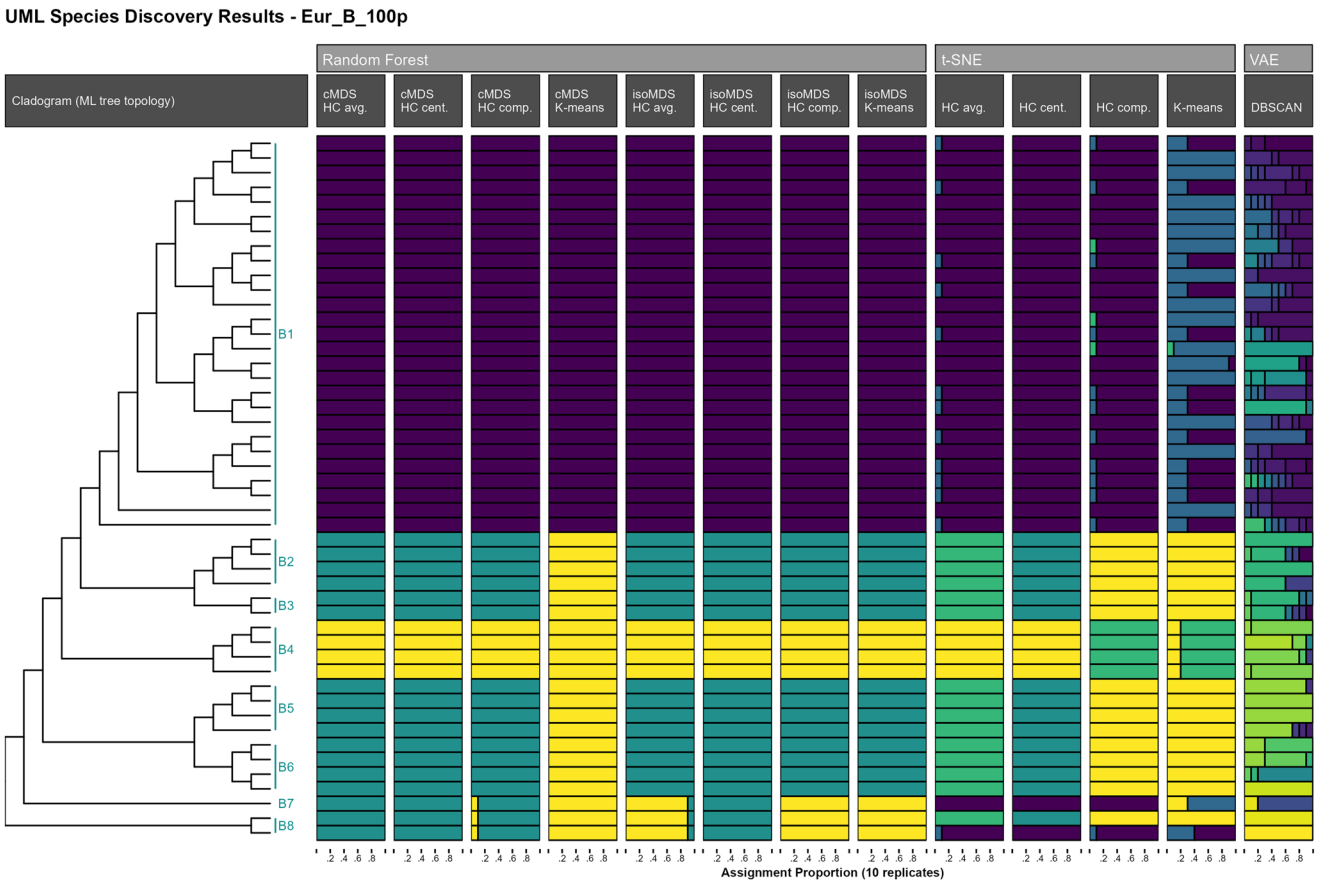

Fig. 3. UML species discovery results—B clade, 100% data completeness. The cladogram shows the ML tree topology and the position of the alphanumerically labeled clades. Each column represents the results of a UML dimensionality reduction plus clustering method. Horizontal bars within each column represent the proportion of times an individual was assigned to a given cluster across replicates, with different colors representing different clusters. Horizontal bars are aligned to the tips of the cladogram. Within each method (column), clusters were compared pairwise across replicates via normalized mutual information (NMI), and the replicate with the highest average NMI was chosen as the reference replicate. Cluster labeling (coloring) was then standardized to the reference replicate via the Hungarian method. Cluster labeling is not standardized between different methods.

propose an initial species hypothesis comprised of the following clusters: A, B1, B2 + B3, B4, B5 + B6, B7 + B8, C1, C2 to 4, and C5.

Morphology and *mtCOI*

Representative wasps from all clades that we examined matched existing species descriptions, though we noted several apparent differences between our specimens and descriptions of types and found that the Bugbee (1967) key was ambivalent on several characters. All wasps in clade A ran to *Phylloxeroxenus* in the DiGiulio (1997) key. This identification is supported by the placement of clade A wasps sister to *Phylloxeroxenus* from the Zhang et al. (2025) UCE dataset (Fig. 1). There are no prior records of *Phylloxeroxenus* associated with North American oak galls (UCD Community 2023). Other collections were *Eurytoma*. A full report on morphology can be found in Supplemental Document S2.

UML methods typically matched results from the ASAP (*mtCOI*) approach (Table 1), with 4 major exceptions, all of which involved ASAP splitting clades that UML methods lumped. First, UML combined wasps in the B2 and B3 clades, while ASAP split them. Clades B2 and B3 wasps best fit the morphologies, respectively, of *Eurytoma querci* Fullaway and *Eurytoma californica* Ashmead, 2 species that were synonymized by Grissell (1973). UML methods support this synonymy while ASAP does not. Second, ASAP also split clades B5 and B6 (morphologically *Eurytoma sphaera* Bugbee and *Eurytoma brevivena* Bugbee), while UML methods generally lumped them. Third, ASAP split apart C2 and C4 (no C3 *mtCOI* sequences were available), while UML methods generally combined C2 + C3 + C4. Wasps from clades C2 to C5 all keyed to the named species *Eurytoma studiosa* Say, such that whichever species discovery method is followed *E. studiosa* appears to consist of at least 2 species. Finally, wasps in clades B7 and B8 were split by ASAP but lumped by UML. These had no morphological ID due to a lack of surrogate material.

Discussion

UML methods produced species hypotheses that generally aligned with results gleaned from a parallel species discovery effort using *mtCOI*, and that also generally matched existing morphological species definitions. UML methods thus appear to be useful for situations where researchers have low confidence in morphological IDs, and when—as is increasingly common—a sub-genomic sequencing dataset has already been collected for other research purposes. We here discuss variation in outcomes across methods, make general recommendations for future study, and offer thoughts on what these results mean

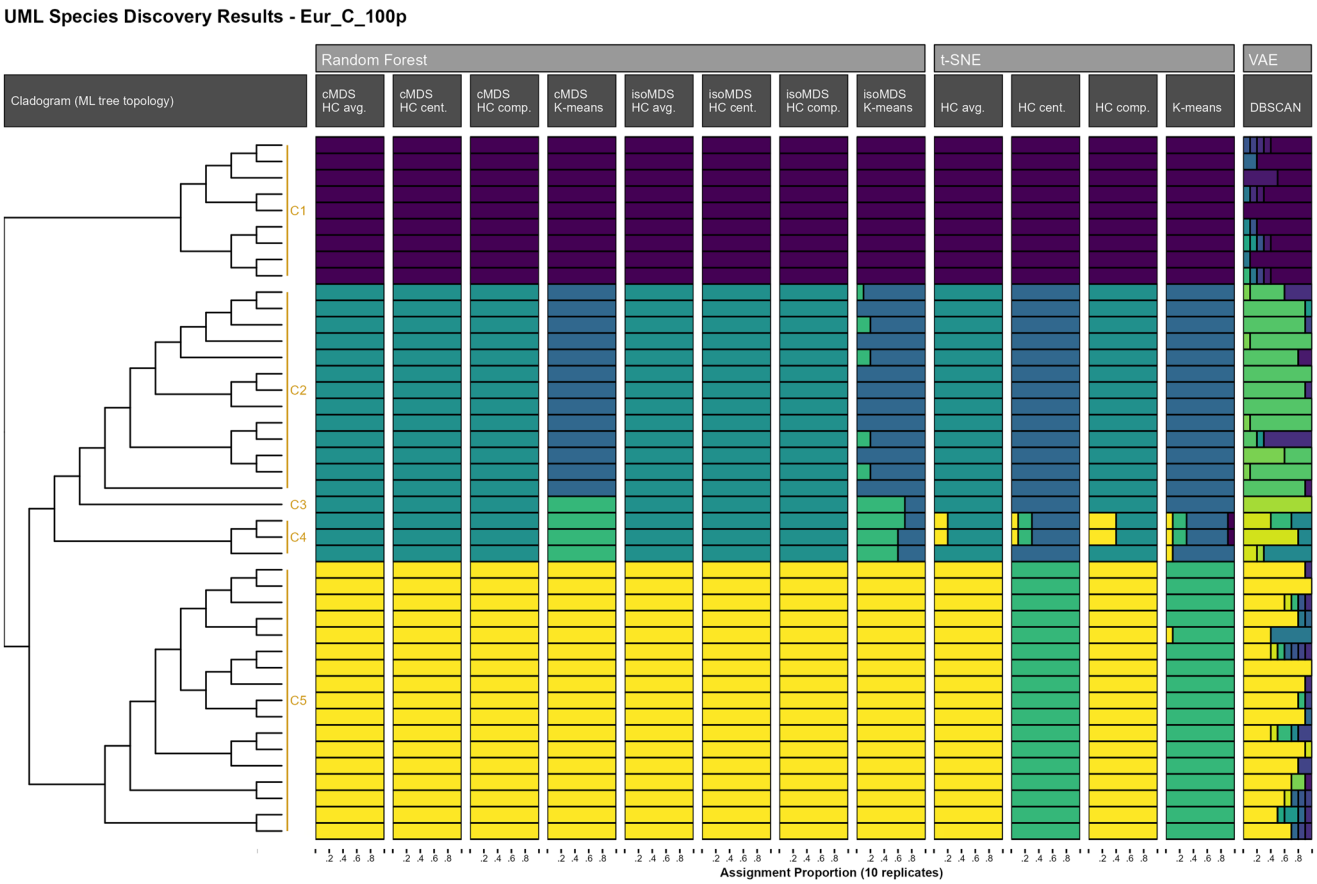

Fig. 4. UML Species Discovery Results—C clade, 100% data completeness. The cladogram shows the ML tree topology and the position of the alphanumerically labeled clades. Each column represents the results of a UML dimensionality reduction plus clustering method. Horizontal bars within each column represent the proportion of times an individual was assigned to a given cluster across replicates, with different colors representing different clusters. Horizontal bars are aligned to the tips of the cladogram. Within each method (column), clusters were compared pairwise across replicates via normalized mutual information (NMI), and the replicate with the highest average NMI was chosen as the reference replicate. Cluster labeling (coloring) was then standardized to the reference replicate via the Hungarian method. Cluster labeling is not standardized between different methods.

for the study of the diversity of North American *Eurytoma* parasites of oak galls.

Comparison of Methods

Of all UML methods, RF more consistently produced results that were concordant with tree topology, had less variability between replicates, and were more easily interpretable. Clustering and ordination methods both had relatively minor effects on the outcome of RF results, although the clustering method (particularly *K*-means vs HC) had a slightly larger effect. Martin et al. (2021) found that RF (particularly the cMDS ordination with PAM clustering and highest mean silhouette width *K selection*) produced fewer spurious clusters than other clustering methods, but that it only detected the deepest phylogenetic bifurcation in the data. Interestingly, the authors suggested that RF may detect further divisions when performed separately on divergent subtrees. Indeed, we found that RF only picked up deep bifurcations in the ABC datasets (Fig. 2) but identified further divisions when repeated on the B and C sample partitions alone (Figs 3 and 4).

Except for the ABC dataset lumping B and C, our *t*-SNE results for all 3 100% datasets showed otherwise remarkable agreement with other methods. This stands in contrast to other projects using this method for delimitation that found *t*-SNE was prone to over-splitting and often produced phylogenetically spurious clusters (Mussmann et al. 2020, Martin et al. 2021, but see Derkarabetian et al. 2019). Conversely, we did observe phylogenetically spurious groups in lower completeness ABC datasets and over-splitting in lower completeness B datasets, especially for B1. This is consistent with the data filtering results of Martin et al. (2021) in suggesting that *t*-SNE performs better on datasets with less missing data.

However, we also note that none of our *t*-SNE results showed the "horizontal striping" seen in Martin et al. where all *t*-SNE replicates show the exact same spurious clusters, which is peculiar given that *t*-SNE is a randomly seeded algorithm with a non-deterministic outcome. We attribute this behavior to an unintentional bug in the original script, which did not set a new random number generation seed each run, and we believe our modification to the script to explicitly set the random seed at the start of each replicate both addresses this bug and explains why our *t*-SNE results show more variation between replicates.

In the B and C datasets, VAE results were inconsistent across replicates and demonstrated a tendency to over-split compared to other methods. However, VAE performed well on the ABC data, demonstrating clear agreement with other methods and greater consistency across replicates. The results for B and C contrast with previous projects where

VAE produced results that were highly consistent across replicates and yielded clusters that were more concordant with phylogeny compared to other UML methods (Derkarabetian et al. 2019, Martin et al. 2021). It is unclear why this is the case, but is plausibly due to differences in the underlying data, such as sample selection, as well as biological differences between study systems. Given the ABC results, one interpretation is that VAE is better suited to detect clusters when there is greater genetic separation, such as when evaluating deeper splits in phylogeny or when species are well-isolated. Shallower splits in phylogeny, reticulation, incomplete lineage sorting, or gradient variation across geographic range may adversely affect VAE performance. Additionally, the results of many UML methods, including VAE, are sensitive to hyperparameter settings, and finding optimal settings can be a major challenge. Future work using *t*-SNE or VAE for species discovery should consider implementing grid search methods to tune hyperparameters as Martin et al. (2021) did for their *t*-SNE perplexity.

Effects of Missing Data

We compared the effects of different amounts of missing data by creating datasets with different levels of completeness (75%, 95%, and 100%) for each of the sample subsets (ABC, B, and C). Here, completeness is defined as the minimum number of samples a UCE locus must be present in to be retained in the final data matrix. While each locus must be present in X% of samples, filtering is agnostic to which samples they are present in, and samples can thus have a different number of missing loci on a per-individual basis.

Datasets with lower completeness showed more phylogenetically spurious clusters, greater inconsistency across replicates, greater discordance between methods, and higher *K* values (over-splitting) than datasets with greater completeness, but the extent to which this was true varied by dataset. We found that more missing data had minimal effect on the results of C, but the lower completeness datasets for B are nearly uninterpretable. Mussmann et al. (2020) found that RF and *t*-SNE both tended to cluster individuals with proportionally high missing data together, and Martin et al. (2021) observed more noise in the form of spurious cluster formation and inconsistency across replicates as missing data increased. One possible interpretation of these results is that more missing data introduces more noise that obscures the phylogenetic signal. However, it may also be that missing data are nonrandomly distributed and recovered UCE loci with high completeness are prone to ascertainment bias. In that case, overly stringent filtering could potentially lead to obfuscation of signals like introgression or hybridization.

While other studies using these methods for delimitation have used completeness cutoffs near 70% (Derkarabetian et al. 2019, Mussmann et al. 2020, Martin et al. 2021), we find that the 75% data matrix produced results that were less consistent and harder to interpret than the more complete datasets, especially for clade B. Why our results appear to be more affected by missingness is unclear, although it may be due to peculiarities of the different study systems, such as relatively more complex genetic structure in our focal wasps, or differences in sampling, such as more unequal collection across species and/or geographic ranges.

Resolution at Different Scales

The ABC datasets showed clear separation between the 3 clades but did not indicate any finer phylogenetic resolution. However, analyzing the B and C datasets separately produced results with finer resolution within each clade. Based on the ML phylogeny, it appears that each clade (A, B, and C) is more proximate to a genus-level grouping compared to the combined ABC dataset, which contains several genera. Thus, when analyzing the whole dataset, UML methods resolved genera but not species. Although this is not a problem unique to UML methods, these results highlight the importance of careful sample selection, data curation, and iterative analyses to better target the desired level of phylogenetic hierarchy.

The C 100% dataset alone produced remarkably concordant results across the RF and *t*-SNE results for all completeness datasets. We note that while the final delimitation clearly delineate C1, C2-4, and C5, individuals from C3 and C4 are occasionally placed in separate populations. This is similar to the STRUCTURE results at $K = 3$, which show C3 and C4 as jointly assigned to the separate C2 and C5 populations. We note that C2 samples were collected from Iowa, C3 was collected from Maine, and C4 samples were collected from California and Arizona. Given the different sampling locations, one plausible interpretation is that C2-4 is a single species, and the placement of C3 and C4 into separate clusters by some methods is a consequence of isolation by distance.

Compared to C, B produced more phylogenetically spurious groups and less consistent results, even in the higher completeness datasets. In the 100% data, we found that B grouped B2-3 with B5-6 or B5-8, and B4 was consistently delineated as either a separate population or grouped together with B7-8. The pattern of separating B4 from the rest of B2-6 was also observed in several STRUCTURE plot runs. There are many possible explanations for why these polyphyletic groupings seem so well supported. First, it is possible that the phylogeny is incorrect; however, this seems unlikely given the support values for the topology. Although the branching pattern within each alphanumeric subclade is not well-supported, the placement of subclades relative to each other is. UML methods could, however, be detecting phenomena such as hybridization, introgression, or incomplete lineage sorting that are obscured by the phylogeny. As discussed above, the lower completeness datasets could also be suffering from too much missing data, whereas the 100% dataset may have lost information by filtering non-randomly missing loci or may be too small in terms of the total number of SNPs. A high completeness dataset, but with a greater number of SNPs/UCE loci, may show improved results.

Another possibility is that B4 wasps contain some shared, derived genetic signature that better enables UML methods to differentiate them from the other B samples. Here, we note that where other B subclades are collected from various gall species, B4 samples were all collected from the same gall host type across different geographic locations. This is suggestive of host specialization, seen elsewhere in the *Disholcaspis* gall specialists of clade C1. A specialist species might have less genetic variance than more generalist clades if specialization involves strong selection that reduces variance (either genome-wide or regionally). Conversely, even if the remaining B clades are all reproductively isolated from one another, a more generalist habit

may translate into insufficient signal to separate them in the current datasets.

Finally, it is also important to consider the UML programs themselves. Because these programs are not trying to fit clusters under a phylogenetic framework or an explicitly biological model, it is somewhat expected that some clusters will not always be concordant with phylogeny. Second, the performance of the algorithms is dependent on the quality of training they receive. Although B contained a similar number of individual samples to C, those samples belong to a larger number of species. Additionally, species are not represented equally, with B1 making up more than half of the samples in B. Both features are likely to adversely affect the learning of UML methods. Separately, there may also be more complex genetic structure within populations in B relative to C, which could require more data to provide adequate training.

Diversity of North American Oak Gall-Associated *Eurytoma* and *Phylloxeroxenus*

While a major goal of this research was to assess the performance of UML methods, we also learned about the animals themselves. One major outcome of this work is the finding that at least 1 wasp species from genus *Phylloxeroxenus* is a common and widespread associate of oak galls in North America. This discovery could be construed simply as arising from a sampling oversight by our group (ie the inclusion of wasps from the wrong genus in a study focused on *Eurytoma*), but the fact that this is clearly a common association that had not previously been discovered in association with oak galls despite massive collection efforts (eg Kinsey 1930, Weld 1959, Ward et al. 2022a) suggests that these wasps have been consistently lumped with *Eurytoma*. We submit this as another clear symptom of the current taxonomic impediment for *Eurytoma*, a paraphyletic genus with poor taxonomic resources in great need of revision (Zhang et al. 2025).

We also find evidence that *E. studiosa* consists of at least 2 species. All molecular methods separate C2 + C3 + C4 from C5 (all morphologically *E. studiosa*), and some UML methods further separate members of C3 + C4 from C2. While genetic differences among C2, C3, and C4 could be ascribed to geographic distance (our collections were from the midwestern, eastern, and southwestern/pacific United States, respectively), the geographic range of C5 overlaps with that of C2, demonstrating clear evidence of reproductive isolation between these 2 "*E. studiosa*" species. Future work should approach a revision of this apparent species complex.

Other morphological species were lumped by UML methods. Clades B2 and B3 wasps were morphologically identified as *E. querci* and *E. californica*, respectively. These species were both originally described from wasps collected in California, United States, but were later synonymized by Grissell (1973) under *E. californica*. UML results agree with the synonymy, but ASAP did not, and wasp in the 2 clades ran to different species in Bugbee's (1967) (admittedly problematic) key. It would be worthwhile to collect more specimens, including from some of the original hosts of *E. querci* in CA, to further interrogate these relationships. Clades B5 and B6, respectively, *E. sphaera* and *E. brevivena*, were also combined by UML methods. Clade B6 wasps were collected across a large geographic area (Arizona to Kentucky/Ohio), while clade B5 wasps were collected just in Iowa and Illinois. Without additional collections, it not clear whether these should be maintained as 2 distinct species or synonymized.

It is also worth noting that clade "B" does not group with the rest of *Eurytoma* s.s. but instead forms a sister relationship with *Kavayva*, a South American seed-feeding genus with distinct morphology and host (Zhang et al. 2021). This is particularly surprising given the morphological, ecological, and geographic similarities between clades "B" and "C," highlighting the limited understanding of North American oak gall parasitoids. A taxonomic revision of North American *Eurytoma* is necessary to resolve this polyphyly.

A broader goal of our research group's work is to understand the evolutionary histories of many different parasitic wasp genera associated with North American oak galls. Several of these genera have proved to be much more species-rich and host-specific than previously thought, often limited to attacking galls induced by just 1 or a few species of gall wasp and/or on a small subset of oak tree species (Ward et al. 2020, 2024, Sheikh et al. 2022, Zhang et al. 2022). This first molecular look at the oak gall-associated *Eurytoma* suggests that some species are considerably less specialized than species in other gall-associated parasitoid genera. Several species in this study (eg B1, C2 + C3 + C4, and including A1, the new species of *Phylloxeroxenus*) were reared from oak galls induced by >8 different gall wasps and on diverse oak tree species. These species host ranges echo those of some polyphagous generalist Palearctic *Eurytoma*, some with >75 known gall hosts (Askew et al. 2013). Other species in our collections do appear more specialized (eg B4 and C1). This intra-generic variation in specialization is similar to the mix of specialization, oligophagy, and generalism found in another eurytomid oak gall parasitoid genus, *Sycophila* Walker (Zhang et al. 2022). As such, North American Eurytomidae may make a useful model for studying the underlying causes of specialization versus generalism. An obvious starting point for this future work would be to look at ecological differences (mode of attack, relative ability to feed on insect or gall material, timing of attack, etc.) and to place those differences in a phylogenetic context.

Conclusions

Our results demonstrate that UML methods are a viable approach for de novo species discovery without a priori population assignment. This approach may be useful when multi-locus data are available from taxonomy-agnostic sequencing projects or when it is preferable to single-locus approaches, and/or when taxonomic/morphological resources (including expertise) are limited or prove difficult to work with. Although discovery is just the first step in delimiting and ultimately describing new species, it provides a necessary foundation to begin asking questions about the biology, ecology, and life history of the focal organisms.

Specimen Collection Statement

Insect Systematics and Diversity supports compliance with the Nagoya Protocol. The authors attest that all legal and regulatory requirements, including export and import collection permits, have been followed for the collection of specimens from source populations at any international, national, regional, or other geographic level for all relevant field specimens collected as part of this study.

Acknowledgements

The authors thank R. Bagley, W. Carr, S. Devine, A. Driscoe, K. McElroy, D. McGarry, K. Neely, M. Shakally, E. Tvedte, and J. Verry for help with gall collection and/or insect rearing.

Author Contributions

Christian L. Weinrich (Conceptualization [equal], Investigation [lead], Software [lead], Visualization [equal], Writing—original draft [equal], Writing—review & editing [equal]), Kristýna Bubeníková (Investigation [supporting], Validation [equal], Writing—original draft [supporting], Writing—review & editing [equal]), Sofia I. Sheikh (Investigation [equal], Resources [equal], Writing—review & editing [equal]), MaKella J. Steffensen (Investigation [equal], Visualization [supporting], Writing—original draft [equal], Writing—review & editing [equal]), Anna K.G. Ward (Funding acquisition [equal], Investigation [equal], Resources [equal], Writing—review & editing [equal]), Yuanmeng Miles Zhang (Investigation [equal], Resources [equal], Writing—review & editing [equal]), and Andrew A. Forbes (Conceptualization [equal], Funding acquisition [equal], Resources [equal], Supervision [equal], Visualization [equal], Writing—original draft [equal], Writing—review & editing [equal])

Supplementary Material

Supplementary material is available at *Insect Systematics and Diversity* online.

Funding

Collections, sequencing, and analyses were supported by grants from the National Science Foundation to AAF (2418250), from the American Genetic Association to AKGW, and to SIS from the Center for Global and Regional Environmental Research. KB acknowledges financial support for her stay in the Forbes lab by the Fulbright US Scholar Program, which is sponsored by the US Department of State and the Fulbright Commission in the Czech Republic. YMZ was supported by the European Union's Horizon 2020 Research and Innovation Programme under Marie Skłodowska-Curie grant agreement 101024056.

Conflicts of Interest

None declared.

Data Availability

UCE sequence reads are available on SRA (BioProject: PRJNA1339037). *mtCOI* sequences are available on GenBank (accessions: PZ015861-PZ015901). Scripts and results files are available on Dryad (DOI: 10.5061/dryad.vq83bk477).

References

Andermann T, Fernandes AM, Olsson U, et al. 2019. Allele phasing greatly improves the phylogenetic utility of ultraconserved elements. *Syst. Biol.* 68:32–46. https://doi.org/10.1093/sysbio/syy039

Askew RR, Melika G, Pujade-Villar J, et al. 2013. Catalogue of parasitoids and inquilines in cynipid oak galls in the West Palaearctic. *Zootaxa* 3643:1–133. https://doi.org/10.11646/zootaxa.3643.1.1

Awad J, Reinisch R, Moser M, et al. 2025. Untangling host specialization in a "double dark taxa" system. *Ann. Entomol. Soc. Am.* 118: 206–219. https://doi.org/10.1093/aesa/saaf003

Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30:2114–2120. https://doi.org/10.1093/bioinformatics/btu170

Branstetter MG, Danforth BN, Pitts JP, et al. 2017. Phylogenomic insights into the evolution of stinging wasps and the origins of ants and bees. *Curr. Biol.* 27:1019–1025. https://doi.org/10.1016/j.cub.2017.03.027

Breiman L. 1996. Bagging predictors. *Mach. Learn.* 24:123–140. https://doi.org/10.1007/BF00058655

Breiman L. 2001. Random forests. *Mach. Learn.* 45:5–32. https://doi.org/10.1023/A:1010933404324

Bugbee RE. 1967. Revision of chalcid wasps of genus *Eurytoma* in America North of Mexico. *Proc. U. S. Natl. Mus.* 118:433–552.

Burke GR, Sharanowski BJ. 2024. Parasitoid wasps. *Curr. Biol.* 34:R483–R488. https://doi.org/10.1016/j.cub.2024.03.038

Castresana J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol. Biol. Evol.* 17:540–552. https://doi.org/10.1093/oxfordjournals.molbev.a026334

Chambers EA, Hillis DM. 2020. The multispecies coalescent over-splits species in the case of geographically widespread taxa. *Syst. Biol.* 69:184–193. https://doi.org/10.1093/sysbio/syz042

Charrad M, Ghazzali N, Boiteau V, et al. 2014. NbClust: an R package for determining the relevant number of clusters in a data set. *J. Stat. Soft.* 61:1–36. https://doi.org/10.18637/jss.v061.i06

Chen H, Rangasamy M, Tan SY, et al. 2010. Evaluation of five methods for total DNA extraction from western corn rootworm beetles. *PLoS One*. 5:e11963. https://doi.org/10.1371/journal.pone.0011963

Derkarabetian S, Castillo S, Koo PK, et al. 2019. A demonstration of unsupervised machine learning in species delimitation. *Mol. Phylogenet. Evol.* 139:106562. https://doi.org/10.1016/j.ympev.2019.106562

Dietenberger M, Jechow A, Sann M, et al. 2025. Shedding light on dark taxa: exploring a cryptic diversity of parasitoid wasps affected by artificial light at night. *Sci. Rep.* 15:6237. https://doi.org/10.1038/s41598-025-88111-3

DiGiulio JA. 1997. Eurytomidae. In: Gibson GAP, Huber JT, Woolley JB, editors. *Annotated keys to the genera of Neartic Chalcidoidea (Hymenoptera)*. National Research Council of Canada Monograph Publishing Program. p. 477–495.

Egan SP, Hood GR, Martinson EO, et al. 2018. Cynipid gall wasps. *Curr. Biol.* 28:R1370–R1374. https://doi.org/10.1016/j.cub.2018.10.028

Ester M, Kriegel H-P, Sander J, et al. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In: Simoudis E, Han J, Fayyad U, editors. KDD'96: *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining; Portland, Oregon*. AAAI Press. p. 226–231.

Evanno G, Regnaut S, Goudet J. 2005. Detecting the number of clusters of individuals using the software structure: a simulation study. *Mol. Ecol.* 14:2611–2620. https://doi.org/10.1111/j.1365-294X.2005.02553.x

Fagan-Jeffries EP, Cooper SJB, Bertozzi T, et al. 2018. DNA barcoding of microgastrine parasitoid wasps (Hymenoptera: Braconidae) using high-throughput methods more than doubles the number of species known for Australia. *Mol. Ecol. Resour.* 18:1132–1143. https://doi.org/10.1111/1755-0998.12904

Faircloth BC. 2013. illumiprocessor: a trimmomatic wrapper for parallel adapter and quality trimming [Computer software]. http://dx.doi.org/10.6079/J9ILL

Faircloth BC. 2016. PHYLUCE is a software package for the analysis of conserved genomic loci. *Bioinformatics* 32:786–788. https://doi.org/10.1093/bioinformatics/btv646

Faircloth BC, McCormack JE, Crawford NG, et al. 2012. Ultraconserved elements anchor thousands of genetic markers spanning multiple

evolutionary timescales. *Syst. Biol.* 61:717–726. https://doi.org/10.1093/sysbio/sys004
Fernandez-Triana JLL, Shimbori EMM, Whitfield JBB, et al. 2023. A revision of the parasitoid wasp genus *Alphomelon* Mason with the description of 30 new species (Hymenoptera, Braconidae). *Zookeys*. 1175:5–162. https://doi.org/10.3897/zookeys.1175.105068
Flouri T, Jiao X, Rannala B, et al. 2018. Species tree inference with BPP using genomic sequences and the multispecies coalescent. *Mol. Biol. Evol.* 35:2585–2593. https://doi.org/10.1093/molbev/msy147
Folmer O, Black M, Hoeh W, et al. 1994. DNA primers for amplification of mitochondrial cytochrome c oxidase subunit I from diverse metazoan invertebrates. *Mol. Mar. Biol. Biotechnol.* 3:294–299.
Forbes AA, Hall MC, Lund J, et al. 2016. Parasitoids, hyperparasitoids, and inquilines associated with the sexual and asexual generations of the gall former, *Belonocnema treatae* (Hymenoptera: Cynipidae). *Ann. Entomol. Soc. Am.* 109:49–63. https://doi.org/10.1093/aesa/sav112
Fujisawa T, Barraclough TG. 2013. Delimiting species using single-locus data and the Generalized Mixed Yule Coalescent Approach: a revised method and evaluation on simulated data sets. *Syst. Biol.* 62:707–724. https://doi.org/10.1093/sysbio/syt033
Funk DJ, Omland KE. 2003. Species-level paraphyly and polyphyly: frequency, causes, and consequences, with insights from animal mitochondrial DNA. *Annu. Rev. Ecol. Evol. Syst.* 34:397–423. https://doi.org/10.1146/annurev.ecolsys.34.011802.132421
Godfray HCJ, Lewis OT, Memmott J. 1999. Studying insect diversity in the tropics. *Philos. Trans. R Soc. Lond. B Biol. Sci.* 354:1811–1824. https://doi.org/10.1098/rstb.1999.0523
Grissell EE. 1973. New species of Eurytoma associated with Cynipidae. *Pan-Pac. Entomol.* 49:354–362.
Guindon S, Dufayard J-F, Lefort V, et al. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst. Biol.* 59:307–321. https://doi.org/10.1093/sysbio/syq010
Hebert PDN, Cywinska A, Ball SL, et al. 2003. Biological identifications through DNA barcodes. *Proc. Biol. Sci.* 270:313–321. https://doi.org/10.1098/rspb.2002.2218
Hoang DT, Chernomor O, von Haeseler A, et al. 2018. UFBoot2: improving the ultrafast bootstrap approximation. *Mol. Biol. Evol.* 35:518–522. https://doi.org/10.1093/molbev/msx281
Jackson ND, Carstens BC, Morales AE, et al. 2017. Species delimitation with gene flow. *Syst. Biol.* 66:799–812. https://doi.org/10.1093/sysbio/syw117
Kalyaanamoorthy S, Minh BQ, Wong TKF, et al. 2017. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat. Methods*. 14:587–589. https://doi.org/10.1038/nmeth.4285
Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* 30:772–780. https://doi.org/10.1093/molbev/mst010
Kingma DP, Welling M. 2022. Auto-encoding variational bayes. https://doi.org/10.48550/arXiv.1312.6114, preprint: not peer reviewed.
Kinsey AC. 1930. The gall wasp genus *Cynips* : a study in the origin of species. *Inndiana Univ. Stud.* 16:1–577.
Krehenwinkel H, Pomerantz A, Henderson JB, et al. 2019. Nanopore sequencing of long ribosomal DNA amplicons enables portable and simple biodiversity assessments with high phylogenetic resolution across broad taxonomic scale. *Gigascience*. 8:giz006. https://doi.org/10.1093/gigascience/giz006
LaSalle J, Gauld I. 1992. Parasitic Hymenoptera and the biodiversity crisis. *Redia Firenze* 74:315–334.
Li H. 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. https://doi.org/10.48550/arXiv.1303.3997, preprint: not peer reviewed.
Li H, Handsaker B, Wysoker A, et al.; 1000 Genome Project Data Processing Subgroup. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25:2078–2079. https://doi.org/10.1093/bioinformatics/btp352
Li Y, Liu J. 2018. StructureSelector: a web-based software to select and visualize the optimal number of clusters using multiple methods. *Mol. Ecol. Resour.* 18:176–177. https://doi.org/10.1111/1755-0998.12719
Llopart A, Lachaise D, Coyne JA. 2005. Multilocus analysis of introgression between two sympatric sister species of *Drosophila: Drosophila yakuba* and *D. santomea*. *Genetics* 171:197–210. https://doi.org/10.1534/genetics.104.033597
Lotfalizadeh H, Delvare G, Rasplus J-Y. 2007. Phylogenetic analysis of Eurytominae (Chalcidoidea: Eurytomidae) based on morphological characters. *Zool. J. Linn. Soc* 151:441–510. https://doi.org/10.1111/j.1096-3642.2007.00308.x
Martin BT, Chafin TK, Douglas MR, et al. 2021. The choices we make and the impacts they have: machine learning and species delimitation in North American box turtles (Terrapene spp.). *Mol. Ecol. Resour.* 21:2801–2817. https://doi.org/10.1111/1755-0998.13350
Melika G, Nicholls JA, Abrahamson WG, et al. 2021. New species of Nearctic oak gall wasps (Hymenoptera: Cynipidae, Cynipini). *Zootaxa* 5084:1–131. https://doi.org/10.11646/zootaxa.5084.1.1
Minh BQ, Schmidt HA, Chernomor O, et al. 2020. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. *Mol. Biol. Evol.* 37:1530–1534. https://doi.org/10.1093/molbev/msaa015
Mussmann SM, Douglas MR, Oakey DD, et al. 2020. Defining relictual biodiversity: conservation units in speckled dace (Leuciscidae: Rhinichthys osculus) of the Greater Death Valley ecosystem. *Ecol. Evol.* 10:10798–10817. https://doi.org/10.1002/ece3.6736
Page AJ, Taylor B, Delaney AJ, et al. 2016. SNP-sites: rapid efficient extraction of SNPs from multi-FASTA alignments. *Microb. Genom.* 2:e000056. https://doi.org/10.1099/mgen.0.000056
Prjibelski A, Antipov D, Meleshko D, et al. 2020. Using SPAdes de novo assembler. *Curr. Protoc. Bioinformatics*. 70:e102. https://doi.org/10.1002/cpbi.102
Puillandre N, Brouillet S, Achaz G. 2021. ASAP: assemble species by automatic partitioning. *Mol. Ecol. Resour.* 21:609–620. https://doi.org/10.1111/1755-0998.13281
Quicke DLJ. 2012. We know too little about parasitoid wasp distributions to draw any conclusions about latitudinal trends in species richness, body size and biology. *PLoS One*. 7:e32101. https://doi.org/10.1371/journal.pone.0032101
Rabiee M, Mirarab S. 2021. SODA: multi-locus species delimitation using quartet frequencies. *Bioinformatics* 36:5623–5631. https://doi.org/10.1093/bioinformatics/btaa1010
Rognes T, Flouri T, Nichols B, et al. 2016. VSEARCH: a versatile open source tool for metagenomics. *PeerJ*. 4:e2584. https://doi.org/10.7717/peerj.2584
Sharkey MJ, Janzen DH, Hallwachs W, et al. 2021. Minimalist revision and description of 403 new species in 11 subfamilies of Costa Rican braconid parasitoid wasps, including host records for 219 species. *Zookeys*. 1013:1–665. https://doi.org/10.3897/zookeys.1013.55600
Sheikh SI, Ward AKG, Zhang YM, et al. 2022. *Ormyrus labotus* (Hymenoptera: Ormyridae): another generalist that should not be a generalist is not a generalist. *Insect Syst. Divers* 6:8. https://doi.org/10.1093/isd/ixac001
Smith MA, Rodriguez JJ, Whitfield JB, et al. 2008. Extreme diversity of tropical parasitoid wasps exposed by iterative integration of natural history, DNA barcoding, morphology, and collections. *Proc. Natl. Acad. Sci. U S A* 105:12359–12364. https://doi.org/10.1073/pnas.0805319105
Stone GN, Schönrogge K, Atkinson RJ, et al. 2002. The population biology of oak gall wasps (Hymenoptera: Cynipidae). *Annu. Rev. Entomol.* 47:633–668. https://doi.org/10.1146/annurev.ento.47.091201.145247
Sukumaran J, Knowles LL. 2017. Multispecies coalescent delimits structure, not species. *Proc. Natl. Acad. Sci. USA*. 114:1607–1612. https://doi.org/10.1073/pnas.1607921114
UCD Community. 2023. Universal Chalcidoidea Database. https://ucd.chalcid.org/.

van der Maaten L, Hinton G. 2008. Visualizing data using t-SNE. *J. Mach. Learn. Res.* 9:2579-2605. https://www.semanticscholar.org/paper/Visualizing-Data-using-t-SNE-Maaten-Hinton/1c46943103bd7b7a2c7be86859995a4144d1938b

Walker BJ, Abeel T, Shea T, et al. 2014. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One*. 9:e112963. https://doi.org/10.1371/journal.pone.0112963

Ward AKG, Bagley RK, Egan SP, et al. 2022a. Speciation in Nearctic oak gall wasps is frequently correlated with changes in host plant, host organ, or both. *Evolution* 76:1849–1867. https://doi.org/10.1111/evo.14562

Ward AKG, Busbee RW, Chen RA, et al. 2022b. The arthropod associates of 155 North American cynipid oak galls. *Zool. Stud.* 61:e57. https://doi.org/10.6620/ZS.2022.61-57

Ward AKG, Sheikh SI, Forbes AA. 2020. Diversity, host ranges, and potential drivers of speciation among the inquiline enemies of oak gall wasps (Hymenoptera: Cynipidae). *Insect Syst. Divers* 4:3. https://doi.org/10.1093/isd/ixaa017

Ward AKG, Zhang YM, Brown GE, et al. 2024. Speciation in kleptoparasites of oak gall wasps often correlates with shifts into new tree habitats, tree organs, or gall morphospace. *Evolution* 78:174–187. https://doi.org/10.1093/evolut/qpad202

Weld LH. 1959. *Cynipid galls of the eastern United States*.

Zhang YM, Delvare G, Blaimer BB, et al. 2025. Phasing in and out of phytophagy: phylogeny and evolution of the family Eurytomidae (Hymenoptera: Chalcidoidea) based on ultraconserved elements. *Syst. Entomol* 50:780–793. https://doi.org/10.1111/syen.12682

Zhang YM, Gates MW, Shorthouse JD. 2014. Testing species limits of Eurytomidae (Hymenoptera) associated with galls induced by *Diplolepis* (Hymenoptera: Cynipidae) in Canada using an integrative approach. *Can. Entomol.* 146:321–334. https://doi.org/10.4039/tce.2013.70

Zhang YM, Gates MW, Silvestre R, et al. 2021. Description of Kavayva, gen. nov., (Chalcidoidea, Eurytomidae) and two new species associated with Guarea (Meliaceae), and a review of New World eurytomids associated with seeds. *JHR*. 86:101–121. https://doi.org/10.3897/jhr.86.71309

Zhang YM, Sheikh SI, Ward AKG, et al. 2022. Delimiting the cryptic diversity and host preferences of *Sycophila* parasitoid wasps associated with oak galls using phylogenomic data. *Mol. Ecol.* 31:4417–4433. https://doi.org/10.1111/mec.16582