


# Multiple full-length variants of the mitochondrial COI DNA barcode region are prevalent in north European sawflies

Marko Prous<sup>1,2</sup>  | Santtu Urpilainen<sup>2†</sup> | Paul D. N. Hebert<sup>3</sup> | Evgeny Zakharov<sup>3</sup> | Niina Kiljunen<sup>2</sup> | Marko Mutanen<sup>2</sup>

<sup>1</sup>Museum of Natural History, University of Tartu, Tartu, Estonia

<sup>2</sup>Ecology and Genetics Research Unit, University of Oulu, Oulu, Finland

<sup>3</sup>Centre for Biodiversity Genomics, University of Guelph, Guelph, Ontario, Canada

## Correspondence

Marko Prous, Museum of Natural History, University of Tartu, Tartu 51003, Estonia.  
Email: [mprous@ut.ee](mailto:mprous@ut.ee)

## Funding information

Estonian Research Council, Grant/Award Number: STP42; New Frontiers in Research Fund, Grant/Award Number: NFRFT-2020-00073; Canada Foundation for Innovation's Major Science Initiatives Fund, Grant/Award Number: 42450; Research Council of Finland

Editor: Christopher Owen

## Abstract

DNA barcoding, the use of a standard DNA fragment for species identification, has emerged as a major field of biodiversity research. The effectiveness of this approach rests on the premise that much less variation exists within species than between them. While exceptions occur, this has been demonstrated in many animal taxa where the COI gene is effective in species discrimination. Sawflies are an exception to this pattern because DNA barcodes often fail to distinguish congeneric species. Using high-throughput single-molecule DNA sequencing to recover COI sequences from thousands of sawflies, we found that single individuals often possess multiple, seemingly functional, full-length DNA barcodes (i.e., unrecognizable as nuclear pseudogenes)—a phenomenon not documented at similar prevalence in any animal taxon. While the evolutionary causes of multiple variants require further investigation, our observation is remarkable as it violates the one-barcode-one-specimen assumption. The presence of multiple variants of barcodes within individuals does not jeopardize the concept, but it introduces a complexity for species inventories based on metabarcoding. They will overestimate the species count when barcode-based operational species units are used as species proxies. Similarly, DNA barcode reference libraries must consider how best to deal with the high frequency of multiple intra-individual variants.

## KEYWORDS

DNA barcoding, heteroplasmy, hymenoptera, NUMTs, Symphyta

## INTRODUCTION

Since its introduction more than 20 years ago (Hebert et al., 2003), DNA barcoding, the use of short, standard fragments of DNA to assign biological specimens to a species, has transformed many areas of biodiversity research (DeSalle & Goldstein, 2019; Hebert et al., 2016; Smith et al., 2008). Due to its effectiveness in distinguishing cryptic taxa, DNA barcoding has facilitated taxonomic research (Janzen et al., 2017; Smith et al., 2007). Similarly, it has made it

possible to probe species interactions with unprecedented detail and accuracy (Hrček et al., 2011; Hrček & Godfray, 2015; Wirta et al., 2014) and has enabled efficient mapping of community diversity (Castaño et al., 2020; Creedy et al., 2022; van der Loos & Nijland, 2021). DNA barcoding can also accelerate species discovery in 'dark taxa', that is, poorly studied hyper-diverse groups of organisms (Page, 2016). Besides biological research, the approach is valuable in applied contexts such as forensics (Meiklejohn et al., 2021), biomonitoring (Pawłowski et al., 2021; Wang et al., 2019), detection of invasive species (Madden et al., 2019) and deterring marketplace fraud (Galimberti et al., 2019; Siozios et al., 2020). Innovative

† Deceased.

applications based on the concept are constantly emerging, such as the field of metabarcoding (Taberlet et al., 2018). Consequently, massive enterprises such as BIOSCAN (<https://ibol.org/programs/bioscan/>), BGE (<https://biodiversitygenomics.eu/>) and many national initiatives are presently building DNA barcode reference databases at local, regional or global scales.

In animals, the standard DNA barcode, a 658 bp segment of the mitochondrial COI (cytochrome *c* oxidase) gene (Hebert et al., 2003), was selected for several reasons. The mitogenome is haploid and shows little or no recombination (Saville et al., 1998), aiding data interpretation. Multiple copies of COI are present in each cell, facilitating its PCR amplification even from very small organisms. In addition, the evolution of COI is usually rapid enough to discriminate closely related species (Kerr et al., 2007; Lopez-Vaamonde et al., 2021; Pentinsaari et al., 2014; Roslin et al., 2022). However, DNA barcodes are less effective in distinguishing species in some animal lineages, including sawflies (Hymenoptera, Symphyta). This compromised performance has been linked to mitonuclear discordance (Linnen & Farrell, 2007; Liston et al., 2023; Prous et al., 2017, 2020) and to low barcode divergence among species (Prous et al., 2021).

The interpretation of DNA barcode data is sometimes complicated by NUMTs, non-coding copies of COI located in the nuclear genome. When such NUMTs are sufficiently long and retain binding sites for both primers used in PCR, they are co-amplified with mtCOI. Because NUMTs are generally less than 300 bp in length, the risk of exposure is greatest when short fragments of COI are amplified, for example in eDNA or metabarcoding studies (Song et al., 2008). While relatively few NUMTs span the barcode region, they do occur (Hebert et al., 2023). As NUMTs are not exposed to natural selection, they often accumulate frameshift mutations and stop codons which allow their discrimination from mtCOI, but long NUMTs without these diagnostic features have been detected in genomic assemblies (Hebert et al., 2023).

Heteroplasmy, the phenomenon where an individual harbours more than one mtDNA haplotype, represents a second potential source of interpretational complexity. While most organisms inherit a single mitogenome, exceptions have been documented in diverse taxa (Ladoukakis & Zouros, 2017), including plants (Levsen et al., 2016), bivalve molluscs (Ghiselli et al., 2019), lobsters (Chow et al., 2021), ants (Meza-Lázaro et al., 2018), beetles (Kastally & Mardulyn, 2017; Sriboonlert & Wonnapijit, 2019) and potentially sawflies (Liston et al., 2023; Prous et al., 2021). Divergent heteroplasmy (>1% divergence) based on whole mitochondrial genomes has been reported in the tuatara (Macey et al., 2021) and a beetle (Kastally & Mardulyn, 2017). Heteroplasmy is generally considered uncommon so it is not thought to significantly compromise eDNA or metabarcoding analyses, but large-scale studies on its prevalence have not been conducted. An important feature of heteroplasmy is that its presence is difficult to document using Sanger sequencing—until recently the standard method for barcode reference library construction. This barrier reflects the fact that heteroplasmy is easily misinterpreted as contamination in Sanger sequencing chromatograms as both are revealed by the presence of multiple peaks.

The sawfly fauna of northern Europe has been under intensive taxonomic reassessment for the last decade (Liston et al., 2017, 2022; Prous et al., 2017, 2025). As one component of this work, over 20,000 specimens have been DNA barcoded. This analysis revealed unusual patterns of barcode variation. First, barcode sharing was prevalent among closely related species and, second, many individuals of some species possessed two or more barcodes with deep divergence, many of which were examined for nuclear markers to evaluate possible cryptic diversity. The shift from Sanger sequencing to high-throughput sequencing (HTS) revealed an unexpected pattern: deep intraspecific variants were not only frequent within species, but also within individuals, suggesting that long NUMTs and/or heteroplasmy were prevalent in sawflies.

In this study, we investigate the frequency of sawflies possessing two or more COI variants by examining 6763 specimens across 88 genera of sawflies from North Europe, chiefly Finland. Our goals were (1) to ascertain if within-individual variability results from the co-amplification of recognizable NUMTs (frame shifts, stop codons) spanning the barcode region or as multiple seemingly functional mitochondrial variants within individuals, and (2) to assess the prevalence of these full-length within-individual barcode variants among sawflies. Although the presence of seemingly functional or almost functional intra-individual mitochondrial variants has been reported (Li et al., 2021; Ožana et al., 2022; Prous et al., 2025), our study is the first large scale and systematic investigation of the prevalence of long within-individual barcode variants based on a massive HTS data set.

## MATERIALS AND METHODS

The sawfly material was collected during many field excursions. Most specimens were net collected as adults, but several hundred were reared from larvae. Most specimens were collected by the authors, but a few by other collectors. All except six specimens were collected from Finland, that is, from Norway (3), Germany (1) and Austria (2). Full collection and taxonomic data of all specimens, including specimen photographs, can be retrieved from the public BOLD dataset <https://doi.org/10.5883/DS-SYMVARI>. Additional data (e.g., secondary COI variants that are not permitted in GenBank or BOLD) are included in the [Supplementary Material](#) and on Dryad (Mutanen et al., 2025).

Tissue pulling, specimen photography and data entry were carried out at the University of Oulu, Finland, by the authors, as part of the Finnish Barcode of Life (FinBOL) project activities. The sawfly samples were not ordered taxonomically for sequencing as they remained largely unidentified to species at that stage, but sawfly samples were mostly not mixed with non-sawflies in the 96-well lysis plates. DNA extraction, COI amplification and sequencing took place at the Centre for Biodiversity Genomics (CBG) at the University of Guelph, Canada, following the procedure outlined in Hebert et al. (2018). Sequencing was conducted using Pacific Biosciences Sequel I and II platforms. Details of laboratory analysis, such as primers used, are available in the LIMs reports for each record. LIMs reports are also available through the BOLD dataset (see above). As sequencing was conducted

as part of FinBOL's barcode reference library construction activities, amplicons from multiple specimens were pooled, often together with other arthropods. In each run, amplicons from either 4560 or 9120 specimens were pooled for analysis.

Our initial datasets were recovered using circular consensus sequences (CCSs) of mtCOI amplicons generated by Sequel I/Sequel II platforms at the CBG. This analytical path generates a read count for every CCS recovered from a specimen, a taxonomic assignment based on the reference library in BOLD, and a percent similarity to the closest reference sequence. This dataset, which was derived from the analysis of 6773 specimens representing about 480 species, included 14,090 sequence variants. After initial filtering (see [Supplementary Material](#)), 14,072 sequences from 6763 sawflies were retained for further analyses. Some identifications delivered by analysis with mBRAVE were based on fewer than 20 bases of overlap, which led to incorrect taxonomic assignments. As a result, the dataset (14,072 sequences) was re-identified against a reference sawfly dataset (see below). The main analytical steps were (1) exclusion of sequences reflecting cross-contamination or non-target sequences (e.g., endosymbionts, parasitoids), (2) exclusion of NUMTs with diagnostic features and (3) identification of the remaining specimens with multiple variants. Cases of contamination were defined as those with sequences matching specimens from a different genus. Diagnosable NUMTs were defined as sequences with stop codons and/or frameshift indels (1 or 2 bp insertions/deletions). We provide counts for each of these three categories of sequences (contaminants, diagnosable NUMTs, potentially functional sequences) based on either all CSSs or restricted to CSSs with at least three reads to minimize exposure to sequencing error or tag jumping.

The sawfly specimens were identified based on morphology by MP and MM to species or species group level. Although many of the specimens in some species groups could have been identified to species level, they were left at the group level for the purpose of this study due to the fact that mitochondrial COI does not allow reliably species level identifications in these groups (Prous et al., 2017, 2020, 2025). To control for contaminations among sawfly specimens, we used a sawfly reference dataset that contains published and unpublished validated mtCOI sequences connected to the Electronic World Catalogue of Symphyta (Taeger et al., 2018). Due to the lack of consistency in taxonomic names across databases (most linked to the recent generic revision of the Nematinae; Prous et al., 2014) as well as gaps or errors in identification, the names of specimens were corrected to a genus or species level according to the latest taxonomy before further analyses (Prous et al., 2025; Taeger et al., 2018).

Sequences derived from contamination were removed based on four filters. Sequences were first excluded if they were proteobacterial or derived from a different family than the source specimen. After this filtration, the remaining sequences were re-identified based on a reference dataset using BlastN (Camacho et al., 2009) with—value 1E-150—max\_target\_seqs 100 (to ensure finding the closest reference sequences), and excluded if the sequence-based genus assignment did not match the actual genus determined through morphological examination. Some additional cases of contamination (18 sequences

removed, all supported by only one read) were manually detected within genera, involving distantly related species or species groups not known to share barcodes.

NUMT detection was done in three steps: (1) detection of frameshifts with BlastN (Camacho et al., 2009) against the sawfly nucleotide reference dataset (as above), (2) detection of stop codons with BlastX (Camacho et al., 2009) against an amino acid (aa) reference dataset (translated sawfly reference dataset) and (3) manual checks of sequences for which the Blast results were ambiguous regarding frameshifts. Reads longer than the amplicon region ( $\geq 659$  bp) were retained if stop codons or frameshifts were only detected in the primer regions. Reads significantly shorter than expected amplicon length (e.g.,  $< 212$  amino acids; a 658 bp amplicon should include 219 aa) were checked to ascertain if they were obviously chimeric (one part of read identical or very similar to other sequences and the other part unalignable) and those detected were excluded. Although sequences with more than three aa deletions are probably NUMTs, they were retained for consistency (there were only five sequences with more than two aa deletions). The remaining sequences were (1) aligned with MAFFT (Katoh & Standley, 2013) at a nucleotide level to establish a common translation frame, (2) translated to amino acids using the invertebrate genetic code, (3) aligned again with MAFFT at an aa level and (4) using *pal2nal* (Suyama et al., 2006) to recover a nucleotide alignment based on the aa alignment. This enabled the detection of additional problematic sequences, including overlooked stop codons close to the ends of sequences, frameshift mutations and unusually long indels. See [Supplementary Material](#) for additional details.

As an additional quality control measure, we also reported the number of possible NUMTs/heteroplasmic variants by only considering sequences differing by more than one nucleotide substitution or indel. Although this criterion will undoubtedly exclude many true variants, those remaining are more reliable, particularly when supported by multiple reads. The more stringently filtered dataset (CSSs supported by  $> 2$  reads and excluding intra-individual variants differing by one nucleotide) containing specimens with seemingly functional intra-individual COI variants was manually examined for PCR chimeras. As a final validation step, we compared the variants detected in 14 specimens in the most stringently filtered dataset and an independently obtained (an additional tissue sample extracted and sequenced) dataset by Prous et al. (2025), available at <https://doi.org/10.5852/ejt.2025.977.2799.12791>.

Phylogenetic trees were generated for the most stringently filtered dataset and subsets of this (Figures 2–5) with FastTree (Price et al., 2009), using default parameters: Jukes-Cantor, CAT approximation with 20 rate categories; balanced Support: SH-like 1000. Prior to phylogenetic analyses, the primer sequences were trimmed. The tree based on the most stringently filtered dataset (Figure S1 in the [Supplementary Material](#)) was arbitrarily rooted between Tenthredinoidea and the other sawflies (Pamphiliidae, Xiphydriidae and Cephidae). The trees based on the data subsets (examples inside Pamphiliidae and Tenthredinidae, Figures 2–5) were rooted based on the full tree (Figure S1 in the [Supplementary Material](#)).

Analyses were done using custom scripts written in R (RCoreTeam, 2021) using packages *plyr* (Wickham, 2011), *dplyr* (Wickham et al., 2021), *xlsx* (Dragulescu & Arendt, 2020) and *tidyr* (Wickham, 2021).

For p-distance calculations, we used R package *ape* (Paradis & Schliep, 2019). The *ape* package was also used to calculate the ratio of non-synonymous (amino acid-changing, dN) to synonymous (silent, dS) substitutions for each pair of intra-individual variants in the most stringently filtered dataset. A chi-squared test in R (RCoreTeam, 2021) was used to check if the sex ratio differed between two groups of specimens: those with multiple apparently functional intra-individual COI variants and those with a single COI variant.

## RESULTS

### Characterization of the sawfly dataset

The contamination recognition step removed 2702 sequences (derived from 1939 specimens) and 273 specimens (Table 1). The removal of diagnosable NUMTs excised 3458 sequences (from 1768 specimens) and 317 specimens. The remaining dataset contained 7879 sequence variants derived from 6173 specimens. After excluding variants with <3 reads, the dataset contained 6064 sequence variants from 5474 specimens.

After the exclusion of contaminants and diagnosable NUMTs, 20.8% of specimens contained two or more COI variants (Table 1)

**TABLE 1** Number of specimens and COI sequence variants before and after consecutive filtering steps.

	Specimens	Variants
Before filtering steps	6763	14,072
Contaminants removed	273 (4.0%)	2702 (19.2%) <sup>a</sup>
Contaminants, read count >2	114 (1.7%)	439 (3.1%) <sup>b</sup>
NUMTs removed	317 (4.7%)	3458 (24.6%)
NUMTs, read count >2	178 (2.6%)	857 (6.1%)
After first filtering steps	6173	7879 <sup>c</sup>
With multiple variants	1282 (20.8%)	2988 (37.9%)
Sequences with at least 3 reads	5474	6064
With multiple variants	519 (9.5%)	1109 (18.3%)
Treating intra-individual variants differing by one nucleotide as one	497 (9.1%)	1058 (17.4%)
Additionally excluding PCR chimeras	497 (9.1%)	1039 (17.1%)

<sup>a</sup>Most “contaminants” (>60%) are endosymbionts, parasitoids and other non-sawflies.

<sup>b</sup>Most “contaminants” (>90%) with at least 3 reads are endosymbionts, parasitoids and other non-sawflies.

<sup>c</sup>An additional 33 variants were excluded here that were treated as different variants in the original data simply because of slight differences in overlap length (differentially cut sequences at 5' and 3' ends). These otherwise identical sequences were also excluded in subsequent steps. Of these identical sequences, the one with fewer reads was chosen for exclusion.

After excluding sequences with fewer than three reads, 9.5% of specimens possessed two or more variants. When only variants differing by more than one nucleotide were retained, 19.4% (all sequences) or 9.1% (sequences supported by three or more reads) of the specimens had more than one variant (Table 1). Finally, the removal of chimeras in the dataset with sequences supported by more than two reads left 497 specimens (9.1%) with 1039 variants (Table 1, Figure S1).

### Co-presence of Intra-individual variants in sawflies

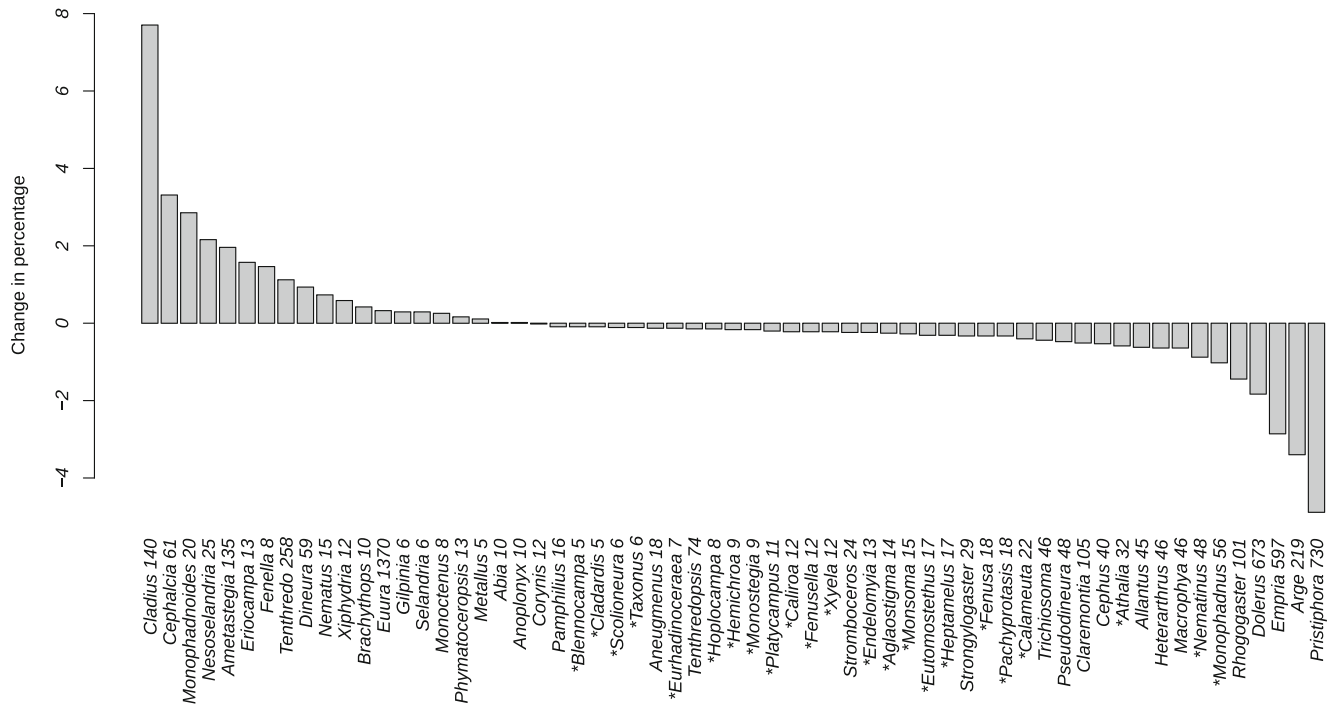
After the exclusion of diagnosable NUMTs, a fifth (1282/6173 or 20.8%) of the specimens in the sawfly dataset possessed intra-individual variants. Of the specimens with multiple variants, most had two variants (median variant count two), but as many as 10 were observed. The median read count for the major variant in these specimens was 18 (maximum = 126) while the median read count for the minor variant was two (maximum = 35). When present, the other variants had a median read count of one and a maximum of 15.

Once sequences with fewer than three reads were excluded (Table 1), 9.5% of the specimens (519/5474) in the sawfly dataset retained intra-individual variants. Most of these specimens had two variants (median variant count two), with a maximum of four. For the major variant, the median read count was 19 with a maximum of 110. For the next most common variant, the median read count was five with a maximum of 33. Minor variants had a median read count of four and a maximum of 15.

Considering only the dataset based on sequences supported by at least three reads, the following seven genera with at least three specimens with more than one variant showed a high incidence (>30% of specimens) of intra-individual variation (Table S2): *Fenella* (100%), *Monophadnoides* (80%), *Eriocampa* (69%), *Nesoselandria* (52%), *Cladius* (36%), *Cephalcia* (36%), *Nematus* (33%) and *Xiphydria* (33%).

Considering all 347 species or species groups, 110 included specimens with more than one intra-individual variant (dataset with sequences supported by at least three reads). Considering only species or species-groups represented by at least 3 specimens with more than one variant, 26 showed more than 30% of their members with two or more COI variants. They included *Fenella nigrita*, *Eriocampa dorpatica*, *Euura respondens* (all 100%), *Ametastegia tenera* (92%), *Dolerus subarcticus* (82%), *Monophadnoides rubi* (80%), *Euura obducta* (79%), *Tenthredo silensis* (67%), *Eriocampa ovata* (67%), *Ametastegia perla* (60%), *Nematus tulunensis* (60%), *Xiphydria prolongata* (60%), *Nesoselandria morio* (52%), *Cladius pectinicornis* (45%), *Tenthredo ferruginea* (43%), *Dolerus vestigialis* (41%), *Empria pallimacula* (37%), *Cephalcia* spp. (36%), *Claremontia brevicornis* (35%), *Cladius brullei* (33%), *Euura pavida* (33%), *Dineura viridiorata* (32%), *Cladius compressicornis* (32%), *Euura vaga* (32%), *Euura myosotidis* (31%) and *Tenthredo atra* group (30%).

The occurrence of intra-individual variants is clearly taxonomically widespread in sawflies, but in some genera no intra-individual variants were detected or they were much less frequent compared to the taxa above. Considering genera with at least five specimens and sequences supported by at least three reads, *Pristiphora*, *Arge*, *Empria* (except *E. pallimacula* and *E. pumila*), *Dolerus*, *Rhogogaster* and *Monophadnus*



**FIGURE 1** Shown for each sawfly genus is the percentage of specimens in the multiple variant dataset (excluding specimens with only single detected COI variant) minus the percentage of specimens in the whole dataset (including specimens with single and multiple variants). This figure only shows genera with at least five specimens and with sequences supported by at least three reads. Genera marked with '\*' contain only specimens with a single variant. The number of specimens in the whole dataset is shown after each genus name. See Dataset S1 for the underlying data.

**TABLE 2** Number of female and male specimens (excluding larvae) and variants in different datasets.

Dataset	# female	# male	p-value of chi-squared tests (specimens)	% female	# variants female	# variants male	p-value of chi-squared tests (variants)
all vs. multi	3851	2281	0.65	62.80	4878	2953	0.28
	792	484		62.07	1819	1156	
single vs. all	3059	1797	0.56	62.99	3059	1797	0.11
all_rc3 vs. multi_rc3	3450	2010	0.76	63.19	3785	2217	0.57
	310	187		62.37	645	394	
single_rc3 vs. all_rc3	3140	1823	0.73	63.27	3140	1823	0.49

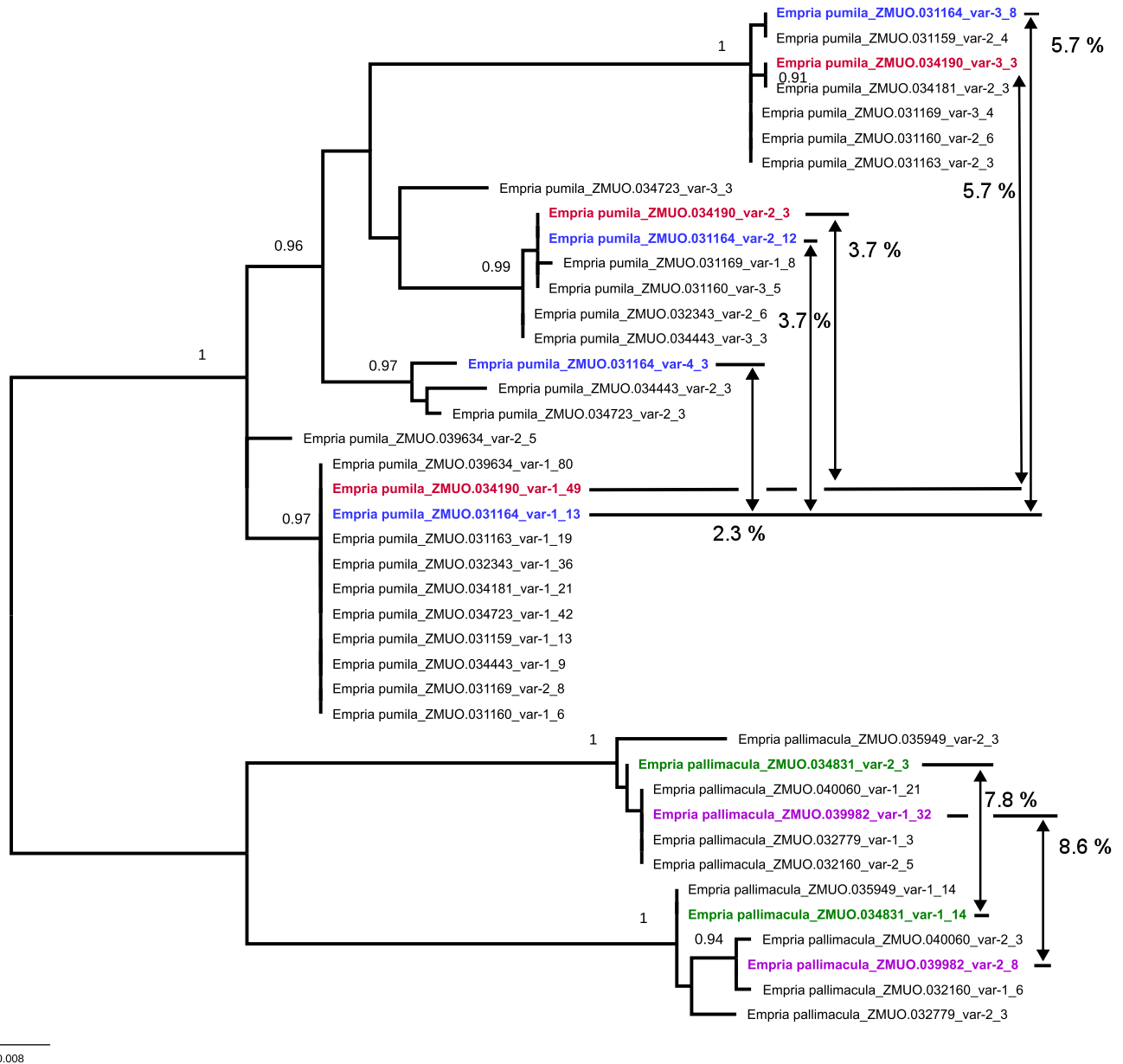
Note: 'all' includes specimens both with single and multiple apparently functional intra-individual COI variants. 'multi' dataset contains only specimens with multiple intra-individual variants and 'single' dataset contains only specimens with one variant. 'rc3' means that only sequences supported by at least three reads were considered. Possible PCR chimeras were excluded and variants differing by single nucleotide were treated as single variants in rc3 datasets. Chi-squared tests were performed to compare male and female numbers between multiple variant and all or single datasets.

(no secondary variants detected) were particularly under-represented in the multiple variant dataset (Figure 1). Tables S1 and S2 provide details on all species and genera with multiple intra-individual variants. Supplementary table in the Dataset S1 also includes genera where only a single COI variant was detected in all specimens.

Comparison of variants of 14 specimens (30 variants) present in the most stringently filtered dataset (497 specimens with 1039 variants, PacBio dataset) and an independent dataset (Nanopore dataset; Prous et al., 2025) revealed that only 12 (from 9 specimens) PacBio variants out of 30 are identical to variants from the corresponding specimens in the Nanopore dataset. However, the non-identical PacBio

variants (identity to Nanopore sequences 97.37%–99.85%, mostly 99.23%–99.85%) are almost all unique in the whole PacBio dataset (16 variants with identity of 98.91%–99.85% to the others), suggesting underestimation of intra-individual variants rather than contamination. The remaining two variants (out of 18 not detected with Nanopore sequencing) from two specimens (ZMUO.033566, ZMUO.040755) are identical to each other (read counts 25 and 18) and to some other conspecific (*Euura clitellata* s. str.) specimens in the PacBio dataset.

When specimens with multiple apparently functional variants were compared to specimens with a single variant, there were no statistically significant differences in sex ratios (Table 2). This was the



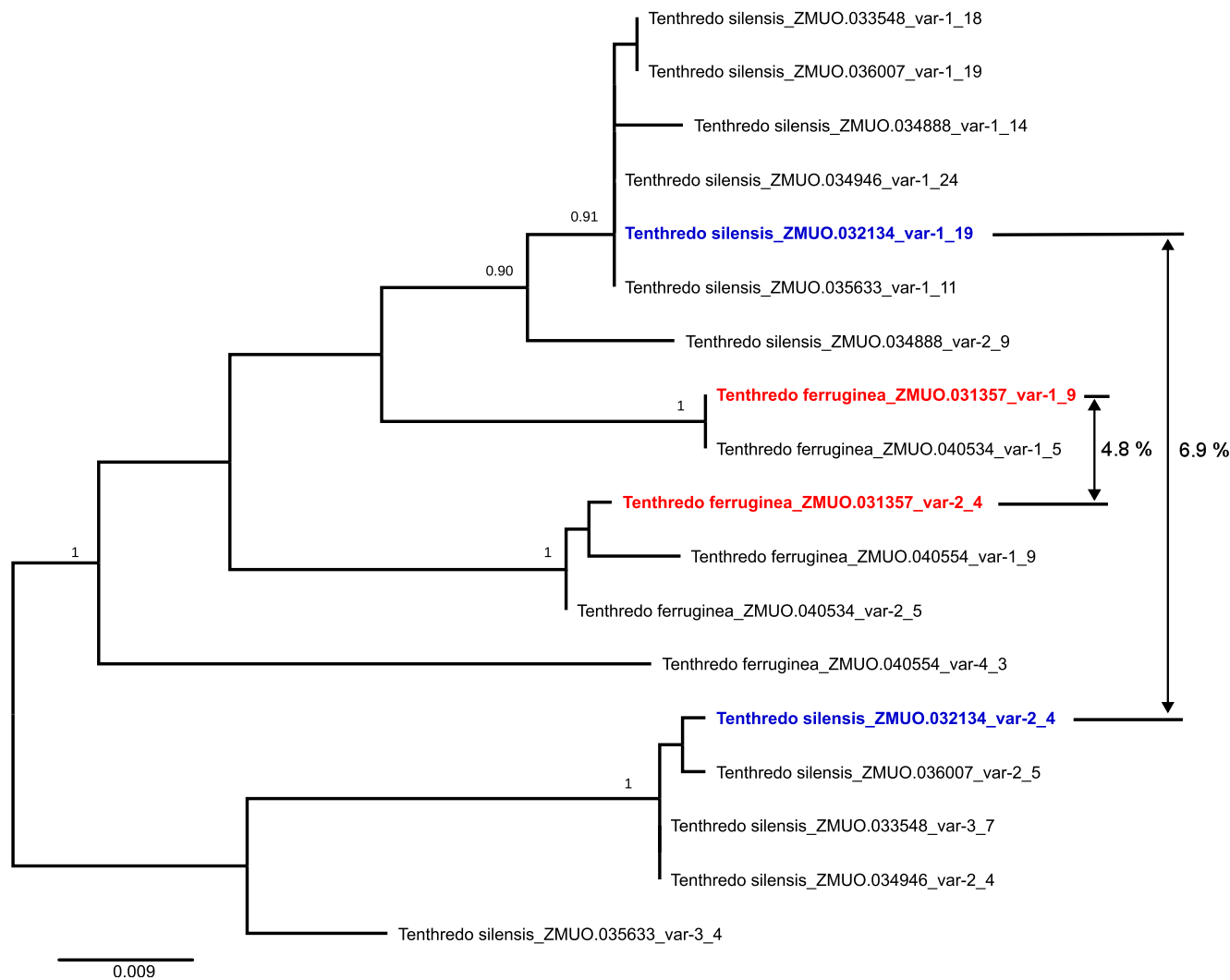
**FIGURE 2** Specimens of *Empria pallimacula* and *E. pumila* with seemingly functional intra-individual COI variants. SH-like branch support values above 0.9 are shown. *p*-distances between the variants are shown for selected individuals. Maximum inter-individual distance is 5.8% for *E. pumila* and 9.4% for *E. pallimacula*.

case for comparisons made at both the specimen and variant levels, whether or not sequences were supported by more than two reads (*p*-values 0.11–0.76, Table 2).

### Intra-individual divergences of sawflies

The maximum likelihood tree for the most stringently filtered dataset of specimens with intra-individual variants is provided as [Supplementary Material](#), but some remarkable examples are illustrated in Figures 2–5. Specimens of *Empria pumila* possessed up to four variants with divergences from 2.3% to 5.7% while two variants with

up to 8.6% divergence were detected in *E. pallimacula* (Figure 2). Two closely related species of *Tenthredo* (*T. ferruginea*, *T. silensis*) possessed intra-individual variants showing up to 6.9% divergence (Figure 3). Specimens of *Cephalcia* (number of species uncertain) are grouped into two main barcode clusters (except ZMUO.033841, whose variants are split between the two groups), but the clusters within both groups are rendered polyphyletic by intra-individual variants. The maximum genetic distances between intra-individual variants within these two main groups are 4.8% and 6.6% (Figure 4). There were also two groups within *Cladius pectinicornis*, which do not share specimens (Figure 5), but again, the variants from single individuals are distributed across within-group clusters. Within the



**FIGURE 3** Specimens of *Tenthredo ferruginea* and *T. silensis* with seemingly functional intra-individual COI variants. SH-like branch support values above 0.9 are shown. p-distances between the variants are shown for selected individuals. Maximum intra-individual distance is 6.9% for *T. silensis* and 6.0% for *Tenthredo ferruginea*. Maximum inter-individual distance is 7.8% between *T. ferruginea* and *T. silensis* and 7.7% within the same species (*T. silensis*).

groups, the maximum distances between intra-individual variants are 3.4% and 3.8%.

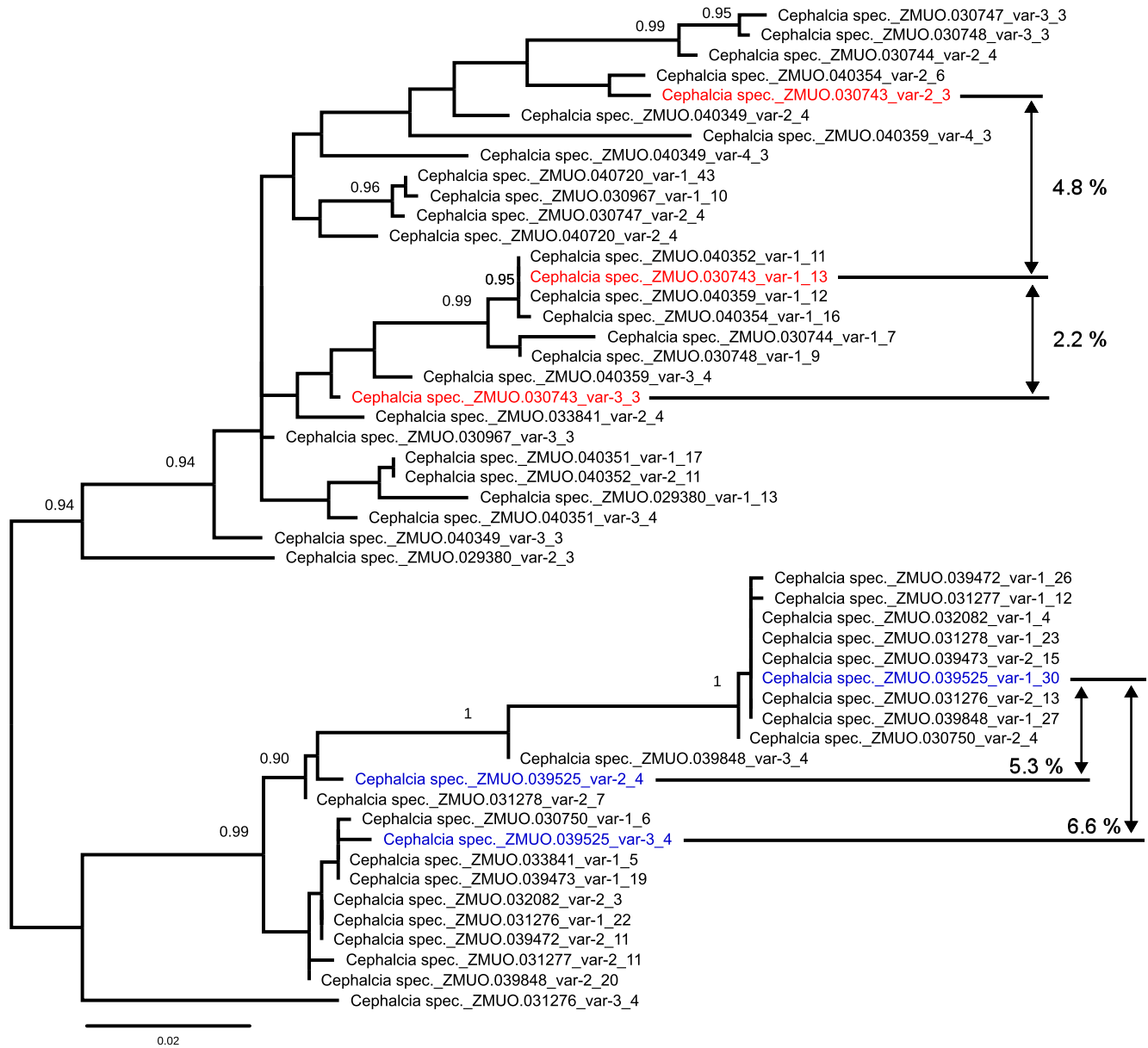
### Non-synonymous to synonymous substitution ratios (dN/dS) of intra-individual variants

dN/dS can be used to estimate the strength of selection acting on protein-coding sequences. A ratio close to one suggests relaxed selection (no difference between rates of synonymous and non-synonymous mutations), indicating that the gene might not be functional. Ratios less than one indicate purifying selection (non-synonymous mutation rate lower than synonymous rate) while ratios greater than one suggest positive selection (non-synonymous mutation rate higher than synonymous rate). Using the most stringently filtered dataset (497 specimens with 1039 variants), dN/dS was

calculated for intra-individual variants of each specimen. 96% of pairwise comparisons (567/589) give ratios below 0.5, and 71% are below 0.1, indicating that most variants have been under strong purifying selection. There are only eight pairwise comparisons (1%) with ratios above 0.9, two of them above 2 (5.2 for variants with 1.6% divergence and infinity for variants with only two differences that are non-synonymous). See [Dataset S1](#) for the table containing all within-specimen pairwise comparisons.

### DISCUSSION

With 5474 specimens and about 347 species or species groups of sawflies represented in the final validated dataset, our study represents the first large-scale investigation of intra-individual full-length mtCOI DNA barcode variants. Our detection of multiple variants



**FIGURE 4** Specimens of *Cephalcia* (number of species uncertain) with seemingly functional intra-individual COI variants. SH-like branch support values above 0.9 are shown. *p*-distances between the variants are shown for selected individuals, which happen to be the maximum intra-individual distances within the two main groups (4.8% and 6.6%).

might potentially arise from sequencing errors or overlooked contaminants, but sequencing errors were eliminated by excluding reads that were only recovered once. Bioinformatics pipelines do provide an efficient way to recognize sequence variation derived from contamination, particularly when the taxa in a sample are known and when comprehensive reference libraries are available. Although we could not always reliably detect contamination between different congeneric species, the residual contamination level after filtering steps must be very low or absent, particularly in the most stringently filtered dataset (497 specimens with 1039 variants). We removed only 18 variants out of about 7900 as contamination between congeners (cases which could be recognized), all supported by only one read. The intra-individual variants were frequently unique sequences and often

differed by more than two nucleotides from the others, excluding the possibility of cross-contamination in such cases. The widespread presence of intra-individual variants has been found repeatedly by other studies in a large sawfly genus *Euura* (about 27% of the specimens in this work) with longer amplicons and different sequencing technology (Liston et al., 2023; Prous et al., 2021, 2025). Independent DNA extraction and sequencing of some specimens (Prous et al., 2025) used in this study revealed significant underestimation of within specimen COI variant diversity (>50% more variants detected) instead of congeneric contamination. This is supported also by nuclear genes (Prous et al., 2025) with the expected number of variants for males (one) and females (one or two, except a few parthenogenetic species that may be triploid). Given the prevalence of long intra-individual

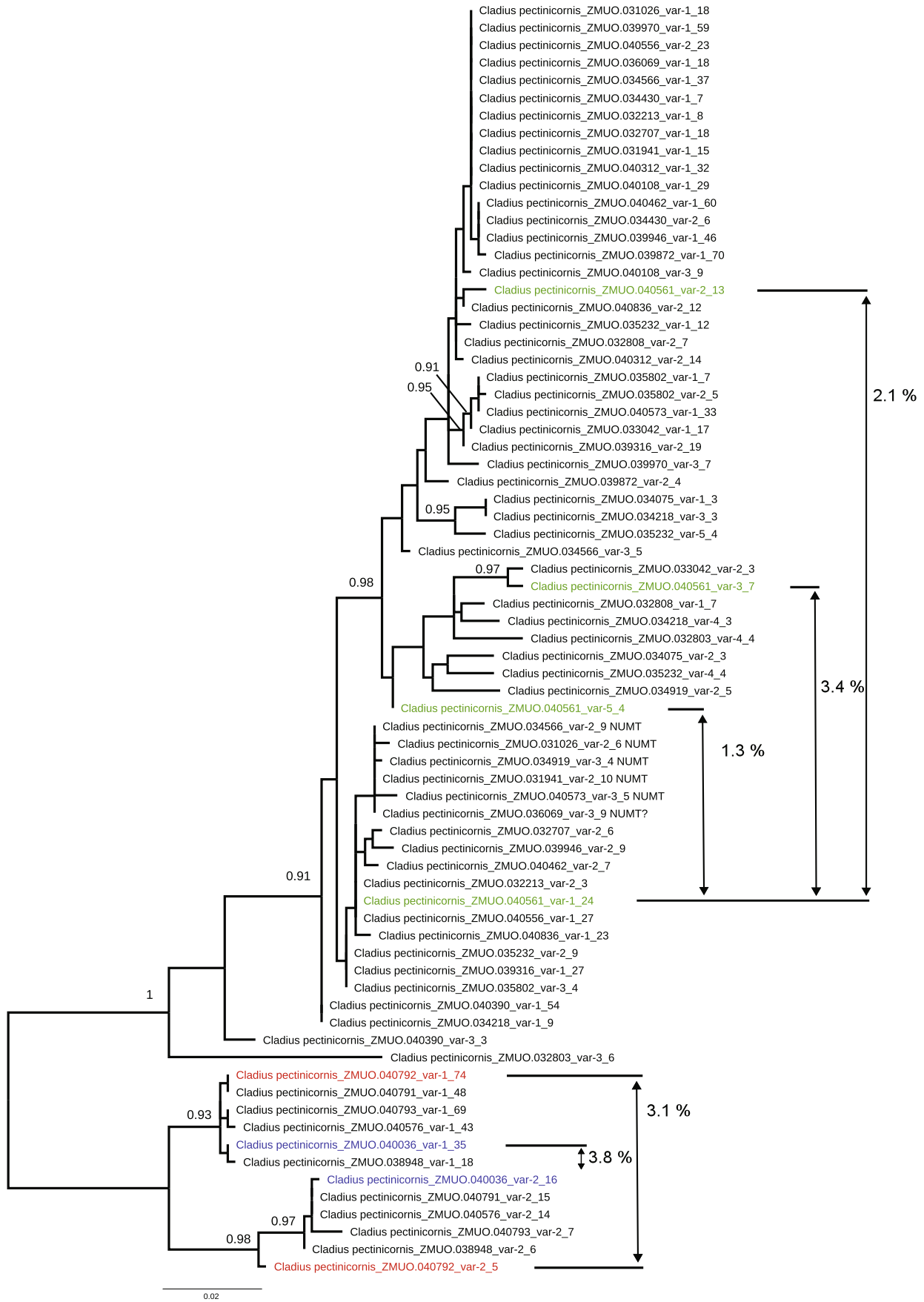


FIGURE 5 Legend on next page.

13653113, 2026, 1, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/syen.12031 by University Of Oulu KIP24081001 Library, Wiley Online Library on [26/02/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

variants, our observations have important implications for results gathered through metabarcoding or eDNA analysis because such intra-individual variants inflate the estimated species count. For example, the most stringently filtered dataset contains 110 species or species groups, but would increase by about 45 species (3% distance threshold) or even 70 species (2% distance threshold) if intra-individual variants were assumed to belong to different specimens. Variation of this type is certainly not restricted to sawflies, but its prevalence varies among taxonomic groups.

The possible biological explanations for the observed seemingly functional intra-individual COI variants are nuclear insertions of mitochondrial fragments (NUMTs) or heteroplasmy (multiple mtDNA haplotypes).

For example, NUMTs (although the exact nature of the intra-individual variation found here is still to be confirmed) are more prevalent in insect species with direct development or incomplete metamorphosis than in those employing complete metamorphosis, reflecting genome size differences (Hebert et al., 2023). Elucidating the prevalence of intra-individual variants across all taxa is an obvious next step. The emergence of high-throughput DNA sequencers enables their efficient detection as demonstrated here but may also provide avenues to mitigate its downsides.

Our analysis demonstrates that HTS produces many sequences that seemingly represent the target region but could actually be NUMTs. NUMTs are a well-recognized complication for DNA barcoding, and recognition of those with frameshift mutations or stop codons is straightforward. However, the diagnosis of shorter NUMTs is often challenging. A recent analysis of 1002 insect genomes revealed that NUMTs are ubiquitous in insects (Hebert et al., 2023). While it found that most NUMTs are short and that most NUMTs spanning the barcode region possess indels or premature stop codons, this was not always the case. Some long NUMTs lack diagnostic features, but they were uncommon (Hebert et al., 2023). By contrast, our dataset includes thousands of variants that remain unrecognizable as NUMTs, provided that the intra-individual variants really represent NUMTs. However, at least in the species-rich genus *Euura* (>25% of specimens in our dataset), 25% of about 1000 specimens show seemingly functional intra-individual variants even based on 1078–1087 bp COI amplicons (Prous et al., 2025). Moreover, non-synonymous to synonymous substitution ratios (dN/dS) of most intra-individual variants detected here (considering only those supported by at least three reads) indicate strong purifying selection (ratios below 0.1 in 71% of pairwise comparisons), at least in the recent past. There are only eight pairwise comparisons out of 589 (1%) with ratios above 0.9, indicating likely pseudogenes. Hence the key questions are: (a) Why are long intra-individual variants with no indels or stop codons, and with little or no indication of pseudogenization (low dN/dS) so prevalent in sawflies, and (b) in addition to NUMTs, might processes such as heteroplasmy explain their occurrence?

Some of the intra-individual variants which appear fully functional possess more than 2% sequence divergence, a threshold often used to flag putative cryptic species. Moreover, cases of apparently homologous variants occurring in different species were frequent in many genera, compromising the ability of DNA barcodes to separate these species. Further, the observed incidence of intra-individual variability is likely an underestimate because the number of mitochondrial variants varied among individuals. Therefore, any single high-throughput run is unlikely to recover all variants present in a species.

In Sanger sequencing, the presence of two or more co-amplified barcode variants generates double peaks in chromatograms, but there is no way to ascertain if they reflect contamination, heteroplasmy or NUMTs. Deeper insights into these situations only became possible with the emergence of single molecule sequencers capable of generating long reads (e.g., PacBio, Oxford Nanopore). Although the detection of co-amplified variants is now efficient, it remains difficult to discriminate heteroplasmic variants from long NUMTs which lack stop codons or frameshift mutations. The discrimination of such cases requires either transcriptomic analysis—because NUMTs are not transcribed while heteroplasmic variants are—or whole genome sequencing to localize NUMTs in the nuclear genome.

Can we exclude the possibility that the cases of deep divergence between multiple variants reflect heteroplasmy? Past studies have shown that most cases of heteroplasmy involve very low levels of divergence (Dowling, 2014; Leeuwen et al., 2008). The seemingly functional intra-individual COI variants of ~650 bp barcodes detected using PacBio are also found with Nanopore sequencing of ~1080 bp mtCOI fragment (Prous et al., 2025). This observation supports the heteroplasmy hypothesis, although Hebert et al. (2023) observed a few dozen NUMTs spanning over 1500 bp but lacking frameshift mutations or stop codons. As such, the detection of long variants in sawflies is not sufficient evidence to support the heteroplasmy hypothesis. It has been suggested that paternal leakage of mitochondria increases with increasing genetic divergence between the parents (e.g., in the case of interspecific hybrids) due to failure of paternal mtDNA elimination (Ladoukakis & Zouros, 2017; Mastrantonio et al., 2019). For example, divergent parent mitochondrial types are frequently detected in *Pelophylax* frogs but only in hybrids (Radojčić et al., 2015). In sawflies, mitonuclear discordance is common and may reflect mitochondrial introgression (Linnen & Farrell, 2007; Prous et al., 2020), which would be consistent with the prevalence of possible heteroplasmy detected in this study.

If the ‘functional’ intra-individual variants are NUMTs, they would be expected to be more prevalent in diploid females than haploid males if the presence/absence of NUMT is polymorphic within species. However, no difference in sex ratio was found between specimens with and without multiple variants (Table 2), a result that can be explained if the NUMTs are universally present within species.

**FIGURE 5** Specimens of *Cladius pectinicornis* with seemingly functional intra-individual COI variants. SH-like branch support values above 0.9 are shown. *p*-distances between the variants are shown for selected individuals.

To rule out the NUMT hypothesis, complete nuclear and mitochondrial genomes should be sequenced. For example, the chromosome level genome assembly for the sawfly *Tenthredo mesomela* ([https://www.ncbi.nlm.nih.gov/datasets/genome/GCA\\_943736025.1/](https://www.ncbi.nlm.nih.gov/datasets/genome/GCA_943736025.1/)), includes two nuclear scaffolds (OX031020, OX031013) with 98%–99% sequence identity to ~7000 bp segments of the mitogenome (GenBank accession OX031023), indicating the presence of long NUMT insertions. Another possibility to distinguish between heteroplasmy and NUMTs would be to screen for the multiple variants in different specimens and populations of the same species (Kastally & Mardulyn, 2017). Divergent NUMTs should be present in all specimens, while heteroplasmic variants should occur in only some of the individuals (Kastally & Mardulyn, 2017) due to strong genetic bottlenecks expected for mitochondria in germline cells.

As long intra-individual barcode variants are widespread in sawflies, it is important to consider how best to mitigate their effect both on the identification of single specimens and on metabarcoding-derived inferences about the species present in a sample. We recognize two concrete actions towards this goal. First, DNA barcoding should transition to single molecule sequencing instead of Sanger, because only the former can separate the multiple variants. The shift from Sanger to HTS is underway and, before long, all barcoding will be based on the latter approach as it is much less expensive. Second, as long as barcoding applications include PCR, intra-individual barcode variants will be co-amplified. Although the much higher copy number of mtCOI should reduce exposure to NUMTs (Hebert et al., 2023), our experience with sawflies is that any of these variants could realistically produce the highest read count, and therefore be selected as the specimen's barcode sequence, which is the default approach in most bioinformatics pipelines. As such, these co-amplified, seemingly functional variants persist in data sets. For a practical solution to this problem, we suggest that they be explicitly indexed as alternative barcodes and integrated into reference libraries, as otherwise such 'ghost' species will continue to be an issue in metagenomics datasets. Alternatively, DNA barcoding could be based on transcribed DNA which would enable the exclusion of non-transcribed NUMTs, but this is only possible for fresh material as RNA degrades quickly.

In conclusion, the analysis of 6000 sawfly specimens using high-throughput single-molecule sequencing platforms frequently recovered multiple long mtCOI variants from single sawflies. While the factors underlying this phenomenon require further investigation, their presence has significant implications for DNA barcoding and metabarcoding studies on this group. While the documented observations represent a complication for barcode library construction initiatives, single molecule-based sequencing platforms provide efficient ways to overcome it. Such ambitious enterprises should therefore increasingly turn to using high-throughput platforms.

#### AUTHOR CONTRIBUTIONS

M.P. and M.M. designed research. S.U., M.P., and E.Z. performed research. S.U., M.P. analysed data. S.U., M.P., and M.M. wrote the initial draft of the paper. M.P, M.M., and P.D.N.H. wrote the final version with contributions from E.Z. and N.K.

#### ACKNOWLEDGEMENTS

This study has been supported in part by the Research Council of Finland through research infrastructure funding to FinBIF consortium and in part by the New Frontiers in Research Fund (NFRFT-2020-00073) and by the Canada Foundation for Innovation's Major Science Initiatives Fund (MSIF 42450). MP received support from the Estonian Research Council grant STP42. We thank the following people who provided specimens for the genetic analyses: Ali Karhu, Andrew Liston, Esko Viitanen, Eva Söderholm, Ewald Altenhofer, Hannu Alen, Harry Nyström, Iina Eskelinen, Iiro Kakko, Jaakko Pohjoismäki, Juha Salokannel, Juhani Itämies, Jussi Vilen, Keijo Mattila, Lari Heikkinen, Lauri Kaila, Marianne Niemelä, Matias Mustonen, Mika Pajari, Mikko Penttinen, Pekka Pohjola, Pekka Raukko, Petri Ahlroth, Petri Metsälä, Riikka Jarkko, Sami Haapala, Teppo Mutanen, Tomi Mutanen, Tupu Vuorinen and Veli-Matti Mukkala. We thank Lari Heikkinen, Riikka Jarkko and Iina Eskelinen for taking care of the larvae and assisting in sample preparation. Several suggestions by the reviewers significantly helped to improve the manuscript. We also want to remember our co-author and PhD student Santtu Urpilainen who sadly passed away due to illness before the publication of this work. Open access publishing facilitated by Oulun yliopisto, as part of the Wiley - FinELib agreement.

#### FUNDING INFORMATION

Funding by the Research Council of Finland through research infrastructure funding to FinBIF consortium; in part by the New Frontiers in Research Fund (NFRFT-2020-00073) and by the Canada Foundation for Innovation's Major Science Initiatives Fund (MSIF 42450); Estonian Research Council grant STP42.

#### CONFLICT OF INTEREST STATEMENT

The authors declare no conflicting interest.

#### DATA AVAILABILITY STATEMENT

Complete specimen data, including high-resolution photographs, collection and taxonomic data, COI sequences (usually a dominant variant) and GenBank accession numbers for the specimens can be retrieved from the public BOLD dataset <https://dx.doi.org/10.5883/DS-SYMVARI>. Original PacBio CCS reads and their classifications (valid variants, contaminations, NUMTs, chimeras) are included as Dataset S1. Dataset S2 (nucleotide) and Dataset S3 (amino acid) include reference sawfly COI sequences used for re-identification of PacBio reads. Dataset S4 contains specimens with intra-individual COI variants (fasta alignment used to build the tree in Figure S1) after the most stringent filtering steps (sequences supported by at least three reads, with a minimum of two differences between the intra-individual variants, no contaminations, no variants identifiable as NUMTs, no chimeras).

#### ORCID

Marko Prous  <https://orcid.org/0000-0002-5329-7608>

#### REFERENCES

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K. et al. (2009) BLAST+: architecture and applications. *BMC*

- Bioinformatics*, 10(1), 421. Available from: <https://doi.org/10.1186/1471-2105-10-421>
- Castaño, C., Berlin, A., Brandström Durling, M., Ihrmark, K., Lindahl, B.D., Stenlid, J. et al. (2020) Optimized metabarcoding with Pacific biosciences enables semi-quantitative analysis of fungal communities. *New Phytologist*, 228(3), 1149–1158. Available from: <https://doi.org/10.1111/nph.16731>
- Chow, S., Yanagimoto, T. & Takeyama, H. (2021) Detection of heteroplasmy and nuclear mitochondrial pseudogenes in the Japanese spiny lobster *Panulirus japonicus*. *Scientific Reports*, 11(1), 21780. Available from: <https://doi.org/10.1038/s41598-021-01346-8>
- Creedy, T.J., Andújar, C., Meramveliotakis, E., Noguerales, V., Overcast, I., Papadopoulou, A. et al. (2022) Coming of age for COI metabarcoding of whole organism community DNA: towards bioinformatic harmonisation. *Molecular Ecology Resources*, 22(3), 847–861. Available from: <https://doi.org/10.1111/1755-0998.13502>
- DeSalle, R. & Goldstein, P. (2019) Review and interpretation of trends in DNA barcoding. *Frontiers in Ecology and Evolution*, 7, 1–11. Available from: <https://doi.org/10.3389/fevo.2019.00302>
- Dowling, D.K. (2014) Evolutionary perspectives on the links between mitochondrial genotype and disease phenotype. *Biochimica et Biophysica Acta (BBA)-General Subjects*, 1840(4), 1393–1403. Available from: <https://doi.org/10.1016/j.bbagen.2013.11.013>
- Dragulescu, A. & Arendt, C. (2020) *Xlsx: Read, Write, Format Excel 2007 and Excel 97/2000/XP/2003 Files. R package version 0.6.5*. Available at: <https://cran.r-project.org/package=xlsx>
- Galimberti, A., Casiraghi, M., Bruni, I., Guzzetti, L., Cortis, P., Berterame, N.M. et al. (2019) From DNA barcoding to personalized nutrition: the evolution of food traceability. *Current Opinion in Food Science*, 28, 41–48. Available from: <https://doi.org/10.1016/j.cofs.2019.07.008>
- Ghiselli, F., Maurizii, M.G., Reunov, A., Ariño-Bassols, H., Cifaldi, C., Pecci, A. et al. (2019) Natural Heteroplasmy and mitochondrial inheritance in bivalve Molluscs. *Integrative and Comparative Biology*, 59(4), 1016–1032. Available from: <https://doi.org/10.1093/icb/icz061>
- Hebert, P.D.N., Bock, D.G. & Prosser, S.W.J. (2023) Interrogating 1000 insect genomes for NUMTs: a risk assessment for estimates of species richness. *PLoS One*, 18(6), e0286620. Available from: <https://doi.org/10.1371/journal.pone.0286620>
- Hebert, P.D.N., Braukmann, T.W.A., Prosser, S.W.J., Ratnasingham, S., DeWaard, J.R., Ivanova, N.V. et al. (2018) A sequel to sanger: amplicon sequencing that scales. *BMC Genomics*, 19, 219. Available from: <https://doi.org/10.1186/s12864-018-4611-3>
- Hebert, P.D.N., Cywinska, A., Ball, S.L. & DeWaard, J.R. (2003) Biological identifications through DNA barcodes. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 270(1512), 313–321. Available from: <https://doi.org/10.1098/rspb.2002.2218>
- Hebert, P.D.N., Hollingsworth, P.M. & Hajibabaei, M. (2016) From writing to reading the encyclopedia of life. *Philosophical Transactions of the Royal Society, B: Biological Sciences*, 371(1702), 20150321. Available from: <https://doi.org/10.1098/rstb.2015.0321>
- Hrček, J. & Godfray, H.C.J. (2015) What do molecular methods bring to host–parasitoid food webs? *Trends in Parasitology*, 31(1), 30–35. Available from: <https://doi.org/10.1016/j.pt.2014.10.008>
- Hrček, J., Miller, S.E., Quicke, D.L.J. & Smith, M.A. (2011) Molecular detection of trophic links in a complex insect host-parasitoid food web. *Molecular Ecology Resources*, 11(5), 786–794. Available from: <https://doi.org/10.1111/j.1755-0998.2011.03016.x>
- Janzen, D.H., Burns, J.M., Cong, Q., Hallwachs, W., Dapkey, T., Manjunath, R. et al. (2017) Nuclear genomes distinguish cryptic species suggested by their DNA barcodes and ecology. *Proceedings of the National Academy of Sciences*, 114(31), 8313–8318. Available from: <https://doi.org/10.1073/pnas.1621504114>
- Kastally, C. & Mardulyn, P. (2017) Widespread co-occurrence of two distantly related mitochondrial genomes in individuals of the leaf beetle *Gonioctena intermedia*. *Biology Letters*, 13(11), 20170570. Available from: <https://doi.org/10.1098/rsbl.2017.0570>
- Katoh, K. & Standley, D.M. (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution*, 30(4), 772–780. Available from: <https://doi.org/10.1093/molbev/mst010>
- Kerr, K.C.R., Stoeckle, M.Y., Dove, C.J., Weigt, L.A., Francis, C.M. & Hebert, P.D.N. (2007) Comprehensive DNA barcode coverage of north American birds. *Molecular Ecology Notes*, 7(4), 535–543. Available from: <https://doi.org/10.1111/j.1471-8286.2006.01670.x>
- Ladoukakis, E.D. & Zouros, E. (2017) Evolution and inheritance of animal mitochondrial DNA: rules and exceptions. *Journal of Biological Research-Thessaloniki*, 24(1), 2. Available from: <https://doi.org/10.1186/s40709-017-0060-4>
- Leeuwen, T.V., Vanholme, B., Pottelberge, S.V., Nieuwenhuys, P.V., Nauen, R., Tirry, L. et al. (2008) Mitochondrial heteroplasmy and the evolution of insecticide resistance: non-Mendelian inheritance in action. *Proceedings of the National Academy of Sciences*, 105(16), 5980–5985. Available from: <https://doi.org/10.1073/pnas.0802224105>
- Levsen, N., Bergero, R., Charlesworth, D. & Wolff, K. (2016) Frequent, geographically structured heteroplasmy in the mitochondria of a flowering plant, ribwort plantain (*Plantago lanceolata*). *Heredity*, 117(1), 1–7. Available from: <https://doi.org/10.1038/hdy.2016.15>
- Li, T., Wang, Y., Sui, Z., Wang, T., Nian, J., Jiang, J. et al. (2021) Multiple mitochondrial haplotypes within individual specimens may interfere with species identification and biodiversity estimation by DNA barcoding and metabarcoding in fig wasps. *Systematic Entomology*, 46(4), 887–899. Available from: <https://doi.org/10.1111/syen.12500>
- Linnen, C.R. & Farrell, B.D. (2007) Mitonuclear discordance is caused by rampant mitochondrial introgression in Neodiprion (hymenoptera: Diprionidae) sawflies. *Evolution*, 61(6), 1417–1438. Available from: <https://doi.org/10.1111/j.1558-5646.2007.00114.x>
- Liston, A., Heibo, E., Prous, M., Vårdal, H., Nyman, T. & Vikberg, V. (2017) North European gall-inducing *Euura* sawflies (hymenoptera, Tenthredinidae, Nematinae). *Zootaxa*, 4302(1), 1–115. Available from: <https://doi.org/10.11646/zootaxa.4302.1.1>
- Liston, A., Mutanen, M., Heidema, M., Blank, S.M., Kiljunen, N., Taeger, A. et al. (2022) Taxonomy and nomenclature of some Fennoscandian sawflies, with descriptions of two new species (hymenoptera, Symphyta). *Deutsche Entomologische Zeitschrift*, 69(2), 151–218. Available from: <https://doi.org/10.3897/dez.69.84080>
- Liston, A., Vikberg, V., Mutanen, M., Nyman, T. & Prous, M. (2023) Palaearctic willow-catkin sawflies: a revision of the amentorum species group of *Euura* (hymenoptera, Tenthredinidae). *Zootaxa*, 5323(3), 349–395. Available from: <https://doi.org/10.11646/zootaxa.5323.3.2>
- Lopez-Vaamonde, C., Kirichenko, N., Cama, A., Doorenweerd, C., Godfray, H.C.J., Guiguet, A. et al. (2021) Evaluating DNA barcoding for species identification and discovery in European Gracillariid moths. *Frontiers in Ecology and Evolution*, 9, 1–16. Available from: <https://doi.org/10.3389/fevo.2021.626752>
- Macey, J.R., Pabinger, S., Barbieri, C.G., Buring, E.S., Gonzalez, V.L., Mulcahy, D.G. et al. (2021) Evidence of two deeply divergent co-existing mitochondrial genomes in the tuatara reveals an extremely complex genomic organization. *Communications Biology*, 4(1), 116. Available from: <https://doi.org/10.1038/s42003-020-01639-0>
- Madden, M.J.L., Young, R.G., Brown, J.W., Miller, S.E., Frewin, A.J. & Hanner, R.H. (2019) Using DNA barcoding to improve invasive pest identification at U.S. ports-of-entry. *PLoS One*, 14(9), e0222291. Available from: <https://doi.org/10.1371/journal.pone.0222291>
- Mastrantonio, V., Urbanelli, S. & Porretta, D. (2019) Ancient hybridization and mtDNA introgression behind current paternal leakage and

- heteroplasmy in hybrid zones. *Scientific Reports*, 9(1), 19177. Available from: <https://doi.org/10.1038/s41598-019-55764-w>
- Meiklejohn, K.A., Burnham-Curtis, M.K., Straughan, D.J., Giles, J. & Moore, M.K. (2021) Current methods, future directions and considerations of DNA-based taxonomic identification in wildlife forensics. *Forensic Science International: Animals and Environments*, 1, 100030. Available from: <https://doi.org/10.1016/j.fsiae.2021.100030>
- Meza-Lázaro, R.N., Poteaux, C., Bayona-Vásquez, N.J., Branstetter, M.G. & Zaldívar-Riverón, A. (2018) Extensive mitochondrial heteroplasmy in the neotropical ants of the *Ectatomma ruidum* complex (Formicidae: Ectatomminae). *Mitochondrial DNA Part A DNA Mapping, Sequencing, and Analysis*, 29(8), 1203–1214. Available from: <https://doi.org/10.1080/24701394.2018.1431228>
- Mutanen, M., Prous, M., Urpilainen, S., Hebert, P., Zakharov, E. & Kiljunen, N. (2025) Multiple full-length variants of the Mitochondrial COI DNA Barcode Region are prevalent in North European Sawflies. *Dryad*. Available from: <https://doi.org/10.5061/dryad.r4xgd2rz>
- Ožana, S., Dolný, A. & Pánek, T. (2022) Nuclear copies of mitochondrial DNA as a potential problem for phylogenetic and population genetic studies of Odonata. *Systematic Entomology*, 47(4), 591–602. Available from: <https://doi.org/10.1111/syen.12550>
- Page, R.D.M. (2016) DNA barcoding and taxonomy: dark taxa and dark texts. *Philosophical Transactions of the Royal Society, B: Biological Sciences*, 371(1702), 20150334. Available from: <https://doi.org/10.1098/rstb.2015.0334>
- Paradis, E. & Schliep, K. (2019) Ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics*, 35(3), 526–528. Available from: <https://doi.org/10.1093/bioinformatics/bty633>
- Pawlowski, J., Bonin, A., Boyer, F., Cordier, T. & Taberlet, P. (2021) Environmental DNA for biomonitoring. *Molecular Ecology*, 30(13), 2931–2936. Available from: <https://doi.org/10.1111/mec.16023>
- Pentinsaari, M., Hebert, P.D.N. & Mutanen, M. (2014) Barcoding beetles: a regional survey of 1872 species reveals high identification success and unusually deep interspecific divergences. *PLoS One*, 9(9), e108651. Available from: <https://doi.org/10.1371/journal.pone.0108651>
- Price, M.N., Dehal, P.S. & Arkin, A.P. (2009) FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. *Molecular Biology and Evolution*, 26(7), 1641–1650. Available from: <https://doi.org/10.1093/molbev/msp077>
- Prous, M., Blank, S.M., Goulet, H., Heibo, E., Liston, A., Malm, T. et al. (2014) The genera of Nematinae (hymenoptera, Tenthredinidae). *Journal of Hymenoptera Research*, 40, 1–69. Available from: <https://doi.org/10.3897/JHR.40.7442>
- Prous, M., Kramp, K., Vikberg, V. & Liston, A. (2017) North-Western Palaearctic species of *Pristiphora* (hymenoptera, Tenthredinidae). *Journal of Hymenoptera Research*, 59, 1–190. Available from: <https://doi.org/10.3897/jhr.59.12656>
- Prous, M., Lee, K.M. & Mutanen, M. (2020) Cross-contamination and strong mitonuclear discordance in Empria sawflies (hymenoptera, Tenthredinidae) in the light of phylogenomic data. *Molecular Phylogenetics and Evolution*, 143, 106670. Available from: <https://doi.org/10.1016/j.ympev.2019.106670>
- Prous, M., Liston, A., Monckton, S.K., Kramp, K., Vårdal, H., Vikberg, V. et al. (2025) West Palaearctic species of *Euura Newman*, 1837 (hymenoptera, Tenthredinidae). *European Journal of Taxonomy*, 977(1), 1–377. Available from: <https://doi.org/10.5852/ejt.2025.977.2799>
- Prous, M., Liston, A. & Mutanen, M. (2021) Revision of the west Palaearctic *Euura bergmanni* and *oligospila* groups (hymenoptera, Tenthredinidae). *Journal of Hymenoptera Research*, 84, 187–269. Available from: <https://doi.org/10.3897/jhr.84.68637>
- R Core Team. (2021) *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.
- Radojčić, J.M., Krizmanić, I., Kasapidis, P. & Zouros, E. (2015) Extensive mitochondrial heteroplasmy in hybrid water frog (*Pelophylax* spp.) populations from Southeast Europe. *Ecology and Evolution*, 5(20), 4529–4541. Available from: <https://doi.org/10.1002/ece3.1692>
- Roslin, T., Somervuo, P., Pentinsaari, M., Hebert, P.D.N., Agda, J., Ahlroth, P. et al. (2022) A molecular-based identification resource for the arthropods of Finland. *Molecular Ecology Resources*, 22(2), 803–822. Available from: <https://doi.org/10.1111/1755-0998.13510>
- Saville, B.J., Kohli, Y. & Anderson, J.B. (1998) mtDNA recombination in a natural population. *Proceedings of the National Academy of Sciences*, 95(3), 1331–1335. Available from: <https://doi.org/10.1073/pnas.95.3.1331>
- Siozios, S., Massa, A., Parr, C.L., Verspoor, R.L. & Hurst, G.D.D. (2020) DNA barcoding reveals incorrect labelling of insects sold as food in the UK. *PeerJ*, 8, e8496. Available from: <https://doi.org/10.7717/peerj.8496>
- Smith, M.A., Rodríguez, J.J., Whitfield, J.B., Deans, A.R., Janzen, D.H., Hallwachs, W. et al. (2008) Extreme diversity of tropical parasitoid wasps exposed by iterative integration of natural history, DNA barcoding, morphology, and collections. *Proceedings of the National Academy of Sciences*, 105(34), 12359–12364. Available from: <https://doi.org/10.1073/pnas.0805319105>
- Smith, M.A., Wood, D.M., Janzen, D.H., Hallwachs, W. & Hebert, P.D.N. (2007) DNA barcodes affirm that 16 species of apparently generalist tropical parasitoid flies (Diptera, Tachinidae) are not all generalists. *Proceedings of the National Academy of Sciences*, 104(12), 4967–4972. Available from: <https://doi.org/10.1073/pnas.0700050104>
- Song, H., Buhay, J.E., Whiting, M.F. & Crandall, K.A. (2008) Many species in one: DNA barcoding overestimates the number of species when nuclear mitochondrial pseudogenes are coamplified. *Proceedings of the National Academy of Sciences*, 105(36), 13486–13491. Available from: <https://doi.org/10.1073/pnas.0803076105>
- Sriboonlert, A. & Wonnapijit, P. (2019) Comparative mitochondrial genome analysis of the firefly, *Inflata indica* (Coleoptera: Lampyridae) and the first evidence of heteroplasmy in fireflies. *International Journal of Biological Macromolecules*, 121, 671–676. Available from: <https://doi.org/10.1016/j.ijbiomac.2018.10.124>
- Suyama, M., Torrents, D. & Bork, P. (2006) PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Research*, 34, W609–W612. Available from: <https://doi.org/10.1093/nar/gkl315>
- Taberlet, P., Bonin, A., Zinger, L. & Coissac, E. (2018) *Environmental DNA: for biodiversity research and monitoring*. Oxford, England: Oxford University Press. Available from: <https://books.google.fi/books?id=1e9IDwAAQBAJ>
- Taeger, A., Liston, A.D., Prous, M., Groll, E.K., Gehroldt, T. & Blank, S.M. (2018) ECatSym – Electronic World Catalog of Symphyta (Insecta, Hymenoptera). Program version 5.0 (19 Dec 2018), data version 40 (23 Sep 2018) Senckenberg Deutsches Entomologisches Institut (SDEI). Available at: <https://sdei.de/ecatsym/>
- van der Loos, L.M. & Nijland, R. (2021) Biases in bulk: DNA metabarcoding of marine communities and the methodology involved. *Molecular Ecology*, 30(13), 3270–3288. Available from: <https://doi.org/10.1111/mec.15592>
- Wang, P., Yan, Z., Yang, S., Wang, S., Zheng, X., Fan, J. et al. (2019) Environmental DNA: an emerging tool in ecological assessment. *Bulletin of Environmental Contamination and Toxicology*, 103(5), 651–656. Available from: <https://doi.org/10.1007/s00128-019-02720-z>
- Wickham, H. (2011) The Split-apply-combine strategy for data analysis. *Journal of Statistical Software*, 40(1), 1–29. Available from: <https://doi.org/10.18637/jss.v040.i01>

- Wickham, H. (2021) *Tidyr: Tidy Messy Data. R package version 1.1.3. [Computer software]*. Available at: <https://cran.r-project.org/package=tidyr>
- Wickham, H., François, R., Henry, L. & Müller, K. (2021) *Dplyr: A Grammar of Data Manipulation. R package version 1.0.7. [Computer software]*. Available at: <https://cran.r-project.org/package=dplyr>
- Wirta, H.K., Hebert, P.D.N., Kaartinen, R., Prosser, S.W., Várkonyi, G. & Roslin, T. (2014) Complementary molecular information changes our perception of food web structure. *Proceedings of the National Academy of Sciences*, 111(5), 1885–1890. Available from: <https://doi.org/10.1073/pnas.1316990111>

## SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

**Dataset S1.** Data tables containing original sawfly PacBio CCS reads, variants after contamination and NUMT filtering, sequence IDs for variants identified as contaminations, variants identified as NUMTs, variants differing by single nucleotide (substitution or indel), identical variants, variants (supported by at least three reads) identified as PCR chimeras.

**Dataset S2.** Reference sawfly COI sequences in fastA format used for reidentification of PacBio reads.

**Dataset S3.** Reference sawfly COI protein sequences in fastA used for reidentification of PacBio reads.

**Dataset S4.** Sawfly specimens with intraindividual COI variants (fastA alignment used to build the tree in Fig. S1) after the most stringent filtering steps (sequences supported by at least three reads, with minimum of two differences between the intraindividual variants, no contaminations, no variants identifiable as NUMTs, no chimeras).

**Data S1.** Supporting Information.

**How to cite this article:** Prous, M., Urpilainen, S., Hebert, P.D.N., Zakharov, E., Kiljunen, N. & Mutanen, M. (2026) Multiple full-length variants of the mitochondrial COI DNA barcode region are prevalent in north European sawflies. *Systematic Entomology*, 51(1), e70031. Available from: <https://doi.org/10.1111/syen.70031>