Interrogating 1000 Insect Genomes for NUMTs: A Risk Assessment for Species Scans

Paul Hebert[1], Dan Bock[2], and Sean Prosser[3]

[1]University of Guelph
[2]University of British Columbia
[3]Biodiversity Institute of Ontario

July 27, 2022

Abstract

The nuclear genomes of most animal species include segments of the mitogenome, but the count of these NUMTs varies greatly. This study examines the incidence of NUMTs derived from a 658 bp region of the cytochrome c oxidase I (COI) gene as a proxy for other coding regions of the mitochondrial genome. Analysis focuses on the most diverse group of terrestrial organisms, insects, because COI-based identification systems play a key role in clarifying their diversity, an essential antecedent to genome sequencing. Nearly 10,000 COI NUMTs [?] 100 bp were detected in the genomes of 1,002 insect species with a range from 0–443. NUMT counts were similar among congeners, but differences among genera in a family were often large with genome size explaining 56% of the mitogenome-wide variation in counts. While many of these NUMTs possessed an indel or premature stop codon allowing their exclusion, the others could complicate species diagnosis as they averaged 10.1% divergence from their mitochondrial homologue. The count of NUMTs varies widely among insect lineages, peaking in groups that employ direct development or incomplete metamorphosis. They can raise the apparent species count by up to 22% when the 658 bp barcode region is examined while shorter targets (300 bp, 150 bp) elevate exposure (58–111%) to "ghost" species. As a result, NUMTs represent a particular complication for protocols (e.g., eDNA, metabarcoding) which employ short amplicons for biodiversity assessments.

Interrogating 1000 Insect Genomes for NUMTs: A Risk Assessment for Species Scans

Paul D N Hebert*

Dan G Bock

Sean WJ Prosser

* Correspondence: phebert@uoguelph.ca

Centre for Biodiversity Genomics

University of Guelph

Guelph ON N1G 2W1

Abstract

The nuclear genomes of most animal species include segments of the mitogenome, but the count of these NUMTs varies greatly. This study examines the incidence of NUMTs derived from a 658 bp region of the cytochrome *c* oxidase I (COI) gene as a proxy for other coding regions of the mitochondrial genome. Analysis focuses on the most diverse group of terrestrial organisms, insects, because COI-based identification systems play a key role in clarifying their diversity, an essential antecedent to genome sequencing. Nearly 10,000 COI NUMTs ≥ 100 bp were detected in the genomes of 1,002 insect species with a range from 0–443. NUMT counts were similar among congeners, but differences among genera in a family were often large with genome size explaining 56% of the mitogenome-wide variation in counts. While many of these NUMTs possessed an indel or premature stop codon allowing their exclusion, the others could complicate species diagnosis as they averaged 10.1% divergence from their mitochondrial homologue. The count of NUMTs varies widely among insect lineages, peaking in groups that employ direct development or incomplete metamorphosis. They can raise the apparent species count by up to 22% when the 658 bp barcode region is examined while shorter targets (300 bp, 150 bp) elevate exposure (58–111%) to "ghost" species. As a result, NUMTs represent a particular complication for protocols (e.g., eDNA, metabarcoding) which employ short amplicons for biodiversity assessments.

Keywords: biodiversity, cytochrome *c* oxidase 1, DNA barcoding, genome size, OTU, pseudogene

Introduction

The nuclear genomes of most animal species contain segments of the mitogenome (Bensasson *et al*. 2001a) captured during the repair of double-strand breaks associated with meiotic recombination (Yu and Gabriel 1999, Ricchetti *et al*. 1999). Many of these NUMTs (nuclear DNA sequences of mitochondrial origin) are short, but some include much of the mitochondrial genome (Richly and Leister 2004). Their prevalence reflects both recurrent integration events and subsequent duplication and diversification. For example, more than 750 NUMTs, ranging in length from 100 bp to 16,106 bp, comprise 0.01% of the human genome (Richly and Leister 2004, Dayama *et al.* 2014). A third of them have arisen through distinct insertion events; the rest likely reflect duplications following integration (Tourmen *et al*. 2002, Hazkani-Covo *et al*. 2003, Pamilo *et al*. 2007). Extensive variation is apparent in their age; some entered the nuclear genome tens of millions of years ago while others are recent (Dayama *et al*. 2014, Gunbin *et al*. 2017). While mechanisms of NUMT insertion are not fully characterized (Hazkani-Covo *et al*. 2010), their incorporation seems to follow the entanglement of mtDNA with nDNA during cell division (Henze and Martin 2001) as densities are highest near centromeres (Viljakainen *et al*. 2010, Michalovova *et al*. 2013). Although most NUMTs are not transcribed, some appear to regulate gene activity (Chatre and Ricchetti 2011) while others impact the phenotype by disrupting gene function; such cases are, for example, responsible for several human diseases (Hazkani-Covo *et al*. 2010). While NUMTs can offer novel phylogenetic insights (Thalmann *et al.* 2004) because sequence change is slowed 3x–4x after their transfer into the nuclear genome (Perna and Kocher 1996), they also represent a complication for identification systems that employ mitochondrial markers for species discrimination (Song *et al.* 2008, Creedy *et al.* 2020, Francoso *et al.* 2019).

NUMT counts differ markedly among animal lineages and are positively correlated with size of the nuclear genome (Hazkani-Covo *et al*. 2010), but they do vary among closely related taxa. For example, species of *Apis* (honeybee) have many more NUMTs than most other members of their family (Pamilo *et al*. 2007). In taxa with multiple NUMTs, sequence divergence from mtCOI often shows considerable variation reflecting their different timing of incorporation. Those with > 2% sequence divergence pose complexity to approaches using mitochondrial markers for species identification, such as the COI region employed for DNA barcoding (Hebert *et al*. 2003). While NUMTs with an IPSC (indel or premature stop codon) can be identified and filtered, those lacking these features are readily mistaken for the target mitochondrial marker, inflating estimates of diversity in contexts ranging from studies of dietary composition (Dunshea *et al*. 2008) to species richness (Song *et al*. 2008). To evaluate their impact on such applications, we utilized public nuclear and mitochondrial sequence data to examine the prevalence of COI-derived NUMTs in 1,002 insect species. Among these taxa, 668 possessed a nuclear assembly derived from high coverage data, making it possible to estimate genome size, and to examine the relationship between genome size and NUMT abundance/attributes. Analysis of this dataset also allowed evaluation of their impacts on the varied analytical approaches that employ mitochondrial markers, especially COI, for biodiversity assessments.

Materials and methods

Nuclear genome dataset

Analysis began with extraction of metadata for all nuclear genome assemblies for the 1,479 insect species in NCBI's Genome database (https://www.ncbi.nlm.nih.gov/genome/) using the assembly-stats option of the 'ncbi-genome-download' package (https://github.com/kblin/ncbi-genome-

download). Sequence coverage, contig N50, and assembly level (i.e., contig, scaffold, chromosome) were recorded, and this information was used to select a representative assembly for each species when several were available. Specifically, we favoured chromosome over scaffold over contig assemblies. When a species had multiple genomes with the same assembly level, we chose the one with the highest coverage. We next used 'taxize' R (Chamberlain and Szocs 2013) to record the membership of each species in an insect order and family. Thirty-three of the 1,479 assemblies were subsequently excluded because of data problems: 15 derived from bacterial endosymbionts (see Supplementary Materials – Nuclear genome sizes), 13 lacked a species identification, 2 were hybrids, 2 were incomplete, and 1 had an assembly error. The other 1,446 assemblies were downloaded between 11/29/21–12/2/21 using 'ncbi-genome-download'.

COI barcode dataset from BOLD

We examined BOLD (Ratnasingham and Hebert 2007) to ascertain if COI barcodes were available for these 1,446 species. Because it is synchronized with GenBank, BOLD provides simultaneous access to both sources of COI barcodes. When coverage was available for a species, all COI records > 645 bp were downloaded. For sequences > 665 bp, the barcode region was excised using Aliview (Larsson 2014). If more than one Barcode Index Number (BIN; Ratnasingham and Hebert 2013) was associated with a binomen (as expected for unrecognized species complexes), the dominant BIN was used so long as it represented > 65% of the records. Those flagged as contaminants, those with stop codons, and those marked as problematic were omitted. After applying these filters, COI barcodes were recovered from 783 (54.1%) of the 1,446 species.

Mitogenome dataset

105 We searched for the mitogenome of these 1,446 species to provide additional COI barcodes and
106 as a basis for examining if the incidence of NUMTs for the COI barcode region was similar to that
107 for other segments of the mitogenome. We first used 'ncbi-acc-download'
108 (https://github.com/kblin/ncbi-acc-download) to obtain mitogenomes from the NCBI Organelle
109 Genome Resources (https://www.ncbi.nlm.nih.gov/genome/organelle/. On 12/1/21, this repository
110 included mitogenomes for 2,897 insect species. Of these, 391 overlapped with our 1,446 species
111 while mitogenomes for another 13 species were archived with their nuclear genome assembly.
112 Among these 404 NCBI-sourced mitogenomes, 219 were annotated, while 185 were not. As a final
113 step, because genome assemblies can possess 'overlooked' mitogenomes (Vieira and Prosdocimi
114 2019), we screened all nuclear assemblies to identify scaffolds likely to represent unannotated
115 mitogenomes (see Supplementary Materials –Identification of new mitogenomes). All
116 mitogenomes lacking an annotation, whether derived from NCBI or from mitogenome mining,
117 were annotated using the MITOS server (http://mitos.bioinf.uni-leipzig.de/index.py; Bernt *et al*.
118 2013). We then filtered the presumptive mitogenomes, retaining only those with all 13 protein-
119 coding genes found in animal mitogenomes (Boore 1999) and with the standard gene order (see
120 Supplementary Materials – Mitogenome filtering and annotation). These filters produced
121 mitogenomes for 440 species (30.4% of the 1,446 total species), of which 332 were from NCBI
122 and 108 were newly recovered from nuclear assemblies.

123

124 **Combined COI barcode dataset**

125 The COI barcode dataset needed for NUMT detection was assembled by combining the
126 mitogenome and BOLD datasets as follows. For the 440 species with full-length mitogenomes, we
127 used BEDTools getfasta (v.2.30.0; Quinlan 2014) to extract the full-length COI sequence and then

employed Aliview to isolate the 658 bp barcode region. All 440 mitogenome-derived COI barcodes were then run through the BOLD Identification tool (http://boldsystems.org/index.php/IDS_OpenIdEngine) to verify their derivation from the correct species. This step resulted in the removal of 21 mitogenome-derived barcodes and their source mitogenomes as they were either misidentified or derived from contamination. Finally, we incorporated BOLD-derived sequences for 583 species to create a barcode dataset with coverage for 1,002 (419 + 583) of the 1,446 target species (69.3%). These sequences are available as a dataset (DS-NUMTINS) on BOLD dx.doi.org/10.5883/DS-NUMTINS.

NUMT abundance, density, and size distribution

Before analysis, each nuclear genome was filtered to exclude residual mitochondrial DNA sequences. First, we searched for and removed scaffolds, irrespective of their size, that included the term 'mitochondrion' in their FASTA header. Second, we removed all unannotated scaffolds that we identified as a mitogenome.

We then interrogated the nuclear genomes of these 1,002 species for NUMTs derived from the barcode region using BLASTn searches that employed the COI barcode from each species as the query. BLAST parameters included a maximum expectation value (-evalue = 0.0001) and a percent identity > 60% (-perc_identity 60) to the query. In practice, > 99% of the NUMTs recovered through this approach showed ≥ 65% identity to the query sequence. We only considered BLAST hits ≥ 100 bp in subsequent analyses for two reasons. First, when matches involve sequences < 100 bp, the average BLAST E-value approaches the threshold (10^{-6}) considered reliable for DNA-based homology matches (Pearson 2013). Second, most studies which employ DNA for species

identification (e.g., Hellberg *et al.* 2019, Nithaniyal *et al*. 2021, Rinkert *et al*. 2021) target amplicons ≥ 100 bp so results are unaffected by shorter NUMTs.

We processed the BLASTn results to remove hits with 100% query coverage (± 1 bp) that were also very similar (ID ≥ 99%) to the query COI barcode sequence. We reasoned that such sequences were likely to represent segments of the mitochondrial genome still present in the nuclear data despite our mitigation efforts. The remaining hits were presumed to be valid, enabling a count of COI NUMTs for each species. To investigate the length distribution of NUMTs exceeding the 658 bp COI barcode, we repeated the prior steps using full-length COI sequences (ca. 1,500 bp) as the query, employing records derived from the 419 species with mitogenomes (**Table S7**).

NUMT counts for COI versus other regions of the mitogenome

We next determined if the incidence of NUMTs for the COI barcode region was similar to those for other coding regions of the mitogenome. This analysis employed the fasta_windows_v1.1.sh script (https://github.com/kdillmcfarland/sliding_windows/) to partition each of the 419 mitogenomes into 15–22 non-overlapping fragments matching the COI barcode (i.e., window size = 658 bp; slide size = 658 bp), and including the other 12 protein-coding genes and the two rRNA genes. They were extracted from the full-length mitogenomes using the annotation files and BEDTools getfasta as described for COI above. While the annotation files recovered the 14 genes for most mitogenomes, some *de novo* annotations were incomplete, reducing the apparent length of a few mitogenomes (see Supplementary Materials – Mitogenome filtering and annotation). BLAST was used to assess the number of NUMTs derived from each fragment in each species as described for COI barcodes. We then generated a mean NUMT count for the set of fragments from each species to create a mitogenome-wide average and compared it with the NUMT count for the

barcode region using a linear model in R v. 4.1.0 (R Development Core Team 2011) and log_2-transformed values for both metrics. To confirm the relationship between these two variables was not impacted by heavy sampling of certain insect genera, the analysis was repeated with a dataset containing one representative per genus.

Patterns of NUMT variation across insect taxa

We examined the impact of the quality of nuclear genome assemblies (sequence coverage, assembly level) on NUMT counts. A Wilcoxon rank-sum test in R was used to compare NUMT counts from low and high coverage assemblies (see Supplementary Materials – Nuclear genome sizes) while the relationship between NUMT counts and contig N50 was evaluated using Spearman's rank correlation in R. As NUMT counts typically increase with genome size (Hazkani-Covo *et al*. 2010), we used Spearman's coefficient in R to examine the strength of this correlation for the 668 insect species whose high coverage assembly allowed estimation of their genome size.

To visualize variation in NUMT counts among the 668 species and its relationship to genome size, we built circular cladograms based on COI barcodes in raxmlGUI v2.0.7 (Edler *et al*. 2020) for the five major orders and for the pooled 12 minor orders. We then used the R package "ggtree" (Yu *et al*. 2017) to overlay bars showing NUMT counts and genome size on the four cladograms. To test for differences in NUMT counts among orders, we used Kruskal-Wallis rank sum tests in R. Because sample sizes for most orders were low, we restricted this analysis to the five major orders.

NUMT diagnosis and impacts on species scans

NUMTs are typically diagnosed via screens for indels or premature stop codons (IPSCs; Bensasson *et al*. 2001a). To determine if the NUMTs identified in our analysis were diagnosable, we first searched for indels. Specifically, we screened each NUMT for frameshift indels (i.e., those not in a multiple of three) using a custom R script. To identify premature stop codons, we uploaded all NUMTs to BOLD where they were translated and subsequently screened for invertebrate mitochondrial stop codons.

To determine the impact of NUMTs on species scans, we considered five length categories (100–150 bp, 151–300 bp, 301–450 bp, 451–600 bp, 601–658+ bp) recovered by HTS platforms (Quail *et al.* 2012, Hebert *et al.* 2018, McCombie *et al*. 2019). These categories are hereafter designated as C1, C2, C3, C4, and C5. We focused attention on a subset of C5 NUMTs (C5*) that were long enough to span the barcode region (651–661 bp) and that possessed >2% divergence from mtCOI in their source species.

Because NUMTs lacking IPSCs can impact species diagnosis, we compared the proportion of diagnosable NUMTs for each category using a homogeneity chi-square test in R. We also compared the length and nucleotide divergence for diagnosable/non-diagnosable NUMTs of COI using Spearman's rank sum tests. We employed a 2% sequence divergence threshold to categorize NUMTs lacking IPSCs into either distinct Operational Taxonomic Units (OTUs) that would inflate the species count (NUMTs > 2% divergence) or into haplotypes that would be grouped with their parent species, inflating its intraspecific COI variation (NUMTs < 2% divergence). To ascertain their impact on estimates of species richness, we used the RESL algorithm (Ratnasingham and Hebert 2013) to generate OTU counts for the NUMT array derived from the 668 species for NUMTs with three lengths (150 bp, 300 bp, 658 bp).

Impact of analytical protocols on NUMT exposure

Analytical protocols can influence exposure to NUMTs when they examine differing numbers or lengths of amplicons. Efforts to expand the DNA barcode reference library always focus on acquiring a 658 bp barcode. When a single amplicon is targeted, NUMT exposure is determined by the number of C5* NUMTs. NUMT exposure can similarly be determined for eDNA and metabarcoding protocols by considering NUMTs in several length categories.

Barcode library construction examines a single amplicon with high-quality DNA extracts, but extracts with degraded DNA require the examination of multiple amplicons. Work on lightly degraded extracts typically examines two C3 amplicons (300–450 bp) that jointly cover the barcode region (Hebert *et al.* 2013). Binding sites for their primers can potentially occur in any NUMT with a length > 300 bp. Studies on heavily degraded DNA extracts, such as those from century-old museum specimens, often examine 100–150 bp amplicons (D'Ercole *et al.* 2021, Prosser *et al.* 2016) so all five categories must be considered.

Any COI NUMT can contain the binding sites for the primers used to recover a segment of the barcode region so long as its length exceeds that of the target amplicon. For example, a 300 bp segment of COI cannot be recovered from C1 and C2 NUMTs, but it might be included in C3–C5 with the likelihood of its inclusion being determined by the category's fractional coverage of the full 658 bp barcode region. Consequently, the NUMT exposure for any category is:

Exposure = mean length of category/length of the barcode region

As a result, the number of C3 NUMTs which will be amplified by a primer set targeting a 300 bp region of COI = # C3 NUMTs multiplied by their exposure (375 bp/658 bp = 0.57). Exposure rises to 0.80 (525/658) for C4, and to 0.96 (625/658) for C5. As two C3 amplicons must be analyzed to recover a full-length barcode, the total NUMT exposure is doubled, and the resultant assembly has

three possible compositions (2 mtCOI sequences, 2 NUMTs, mtCOI/NUMT chimera). When analysis targets 100–150 bp amplicons, exposure varies 5-fold among the length categories (C1 = 0.19, C2 = 0.34, C3 = 0.57, C4 = 0.80, C5 = 0.96). In this case, total exposure involves summing the values for the five categories.

Results

NUMT counts: Impact of sequence coverage and assembly contiguity

BLASTn detected 16,584 (≥ 20 bp) and 9,826 (≥ 100 bp) NUMTs derived from the COI barcode region among the 1,002 species (17 orders, 149 families, 591 genera) with both a genome assembly and DNA barcode sequence. **Table S1** lists these hits together with their key attributes (length, similarity to query sequence). Most of these species (987/1,002) had a coverage estimate for their genome assembly. These values varied by six orders of magnitude and were bimodal with the low distribution possessing a mean/median coverage of 1.02x/1.07x while the high distribution had a mean/median of 124.1x/76.0x **(Figure S1**). Given this bimodality, the break point (5x) between the distributions was used to designate the nuclear assembly for each species as either low coverage (hereafter LC) or high coverage (hereafter HC). The 15 species lacking an estimate were assigned as LC.

The number of COI NUMTs showed marked variation among taxa; 162 of the 1,002 species had none, while the others possessed from 1 to 443 (**Figure 1**). Among the 668 HC species, the number in each $\log_2$ interval from 0–32 NUMTs per genome showed less than two-fold variation, followed by a halving of the species number with each subsequent doubling in the NUMT count. The 334 LC species showed a similar pattern, but the highest NUMT values were missing, leading to a lower average NUMT count (4.1 versus 12.6) (**Table 1**, **Figure S2**). However, 97.9% of LC

species (327/334) were Lepidoptera versus 28.4% in the HC set. The difference in average NUMT count between coverage classes was greatly reduced (LC = 4.1; HC = 5.7) when analysis compared members of this order and became insignificant when analysis only examined the six families in both datasets (Sign test, $P = 0.22$, **Table S3**). Genome contiguity (contig N50) did not impact counts in the LC (Spearman's rank correlation: $p = 0.09$, $P = 0.12$, n = 334) or HC (Spearman's rank correlation: $p = -0.01$, $P = 0.88$, n = 668) assemblies. Because genome size estimates obtained from assembly length were determined to be unreliable for LC species (see Supplementary Materials – Nuclear genome sizes; **Figure S1, Figure S3**), detailed analysis focused on the HC set (**Figure 2**). In total, they possessed 8,423 NUMTs ≥ 100 bp with counts ranging from 0–443 per species (**Table 1** and **Table S2**). As 126 of the HC species lacked NUMTs, the others possessed an average of 15.5.

Lengths and diagnosis of COI NUMTs

When analysis considered NUMTs recovered with 658 bp barcode queries, lengths varied from 100–754 bp in the 1,002 species (**Table 1**). Most were short; 30% were < 150 bp, 71% were < 300 bp, and 88% were < 600 bp. NUMTs recovered using a full-length COI query sequence from 283 of the HC species ranged from the low cut-off (100 bp) to circa 1,550 bp, the length of the gene (**Figure 3**). The secondary peak near the upper value was an artifact reflecting the fact that some NUMTs included COI together with upstream and/or downstream gene regions. Ignoring this peak, the length distribution of COI NUMTs closely approximated a Pereto distribution (alpha = 1).

Sequence similarity of the 8,423 NUMTs to their COI barcode query ranged from 64–100% (**Figure 2**). Two-thirds possessed IPSCs, but this percentage varied among the five length

287 categories (X^2 = 190.0; $P < 10^{-5}$, df = 4), increasing from 57% in those 100–150 bp to 77% for
288 those 451–600 bp (**Table 2** and **Figure 3**). The percentage of NUMTs > 600 bp with an IPSC
289 declined to 64%, likely reflecting their lower divergence from mtCOI than the other length
290 categories (4.8% versus 10.9%). Considering all NUMT lengths, sequence divergence from the
291 mtCOI query was greater for those with IPSCs (18.6%) than for those without (10.0%) (**Table**
292 **S1**). Among the 5,607 NUMTs with an IPSC, 3,571 possessed both diagnostic features; 1,528 only
293 had an indel, and 508 only possessed a stop codon.
294
295 **NUMT counts for COI relative to mitogenome-wide counts**
296 The NUMT count for the COI barcode region was a strong predictor of the mean count for
297 other mitogenome coding regions (R^2 = 0.72) in the HC species (**Figure 4**). This relationship was
298 unchanged when analysis considered one species per genus (R^2 = 0.71). Moreover, the slope of the
299 regression was close to 1.0 indicating that NUMT counts for COI matched those for other coding
300 regions in the mitochondrial genome.
301
302 **Variation in genome sizes and COI NUMT counts among insect taxa**
303 Considering all HC taxa, the count of COI NUMTs was positively correlated with genome
304 size (R^2 = 34%), and r-squared rose when counts for the entire mitogenome (R^2 = 56%) were
305 considered (**Figure 5**). Congeneric species showed limited variation in both genome size and
306 NUMT count (**Figure 6**), but there was a 100-fold difference in mean counts among the 17 insect
307 orders (**Table 3**). This variation was associated with a key developmental variable as the mean
308 NUMT count was 4-fold higher (39.6 versus 9.3) in species with incomplete than complete
309 metamorphosis (**Figure 7**), a highly significant difference (Wilcoxon rank-sum test, $P = 4.47 \times 10^{-}$

310 ^{9}). NUMT counts also showed significant variation among the five major orders (Kruskal-Wallis:
311 $X^2 = 43.66$, df = 4, $P = 7.52 \times 10^{-9}$, n = 638). Hemiptera, the only one employing incomplete
312 metamorphosis, had the highest count (23.0), but the others showed considerable variation as the
313 mean for Hymenoptera (17.8) was 3x that for Lepidoptera (5.3) and twice those for Diptera (8.0)
314 and Coleoptera (9.8) (**Table 3**).

315 The extent of intra-ordinal variation could only be assessed for the five major insect orders
316 (**Figure S4**). Among them, Hemiptera had the most variable NUMT counts (CV = 1.97), followed
317 closely by Diptera and Hymenoptera while Coleoptera and Lepidoptera showed less variation.
318 **Figure 8** provides an overview of the patterns of variation in genome size and NUMT counts for
319 the 668 species.

320

321 **NUMTs and DNA-based identifications**

322 **Figure 9** displays three key attributes (length, sequence divergence from mtCOI,
323 presence/absence of IPSC) for each NUMT detected in the two species with the greatest genome
324 size difference in the five major insect orders. These paired comparisons show consistently higher
325 NUMT counts in species with large genome sizes. **Figure S5** expands this representation of counts
326 and attributes to all 668 HC species. Among their 8,423 NUMTs, 5,607 had an IPSC while the
327 other 2,816 (**Table 2**) included all five length categories: 1,092 C1 (100–150 bp), 978 C2 (151–
328 300 bp), 238 C3 (301–450 bp), 135 C4 (451–600 bp), and 373 C5 (600–658+ bp). Most (2,545)
329 of these NUMTs occurred as a single copy in the genome, but others were represented by up to
330 ten copies: (n = 81 (2 copies); n = 12 (3 copies); n = 6 (4 copies); n =2 (5 copies); n = 2 (6 copies);
331 n = 1 (8 copies); n = 1 (9 copies); n = 1 (10 copies).

When analysis employs primers for the full barcode region, only C5* NUMTs can inflate the species count. Among the 373 C5 NUMTs, 226 in 113 species were C5*. Most of these species possessed just one or two C5*, but two had ten (**Figure 10, Figure S6**). In the 69 species with a single C5*, the NUMTs showed a wide range of divergence (2.1–24.2%) from mtCOI and the same pattern extended to species with several C5*. A ML tree indicated that the C5* NUMT(s) in each species typically showed closest affinity to its mtCOI counterpart (**Figure 11**). Species with several C5* often possessed several similar or identical NUMTs dispersed in their genome. For example, all 10 in *Mimumesa dahlbomi* showed little sequence divergence from each other (0.26%) while 9 of 10 in *Zaprionus ornatus* were identical. Because of these cases of close sequence similarity, RESL assigned the 226 C5* NUMTs to 139 OTUs, a conversion percentage of 65%. By comparison, RESL assigned the 668 mtCOI sequences from their source species to 632 OTUs, a 95% conversion percentage. If a study recovered all C5* NUMTs, the OTU count for HC species would be inflated by 22% [(139 + 632)/632]. RESL indicated that NUMTs in shorter length categories (150 bp, 300 bp) showed a conversion percentage of roughly 67%, similar to that for C5* (**Figure S7**).

Genome sizes – Towards a risk registry for NUMTs

Because it is a good predictor of NUMT count, all genome size data for insect species was assembled. The resulting compilation included 1,838 species representing 26 of 27 insect orders, and 229 of their 1,000 component families **(Table S4**). Mean genome size varied 60-fold from 130 Mb in Strepsiptera to 7,737 Mb in Orthoptera (**Table S5**). The average genome size was > 1,600 Mb for the three orders with direct development, > 800 Mb for 8 of 11 orders with incomplete metamorphosis, and < 800 Mb for 11 of 12 orders with complete metamorphosis (**Table S5**). While

355 congeneric species had similar genome sizes (**Figure 6**), those in different families within an order
356 often showed marked divergence. For example, among the nine orders represented by at least five
357 families, the ratio of high/low genome sizes varied 8-fold (22.3–Coleoptera, 2.9–Lepidoptera)
358 (**Table S6**). A plot of mean genome size against the number of described species in each insect
359 order further indicated that those with the highest species counts all possessed a small genome size
360 (**Figure 12**).
361
362 **Analytical Protocols – Towards a risk registry for NUMTs**
363 NUMT exposure varies fivefold among the three analytical protocols targeting a single
364 amplicon (**Table 4**). Library construction with a 658 bp amplicon could encounter up to 226 C5*
365 amplicons, 34% of the species count. By comparison, studies targeting 300 bp or 150 bp amplicons
366 could recover 578 and 1,118 NUMTs respectively, 87% and 167% of the species count. Because
367 about a third of the NUMTs in each length category have identical or similar sequences, the count
368 of distinct OTUs would show less inflation – 22%, 58%, and 111% respectively. Efforts to
369 assemble a complete barcode sequence from 2–5 amplicons elevate the risk of NUMT exposure,
370 but the extent of OTU inflation cannot be predicted because the NUMT count, their relative
371 frequencies, and sequence divergences will determine the number and composition of chimeric
372 sequences.
373
374 ## Discussion
375 The presence of NUMTs in insect genomes has been known for 40 years (Gellissen *et al*. 1983),
376 but details on their abundance and attributes have only slowly gained clarity. Early studies revealed
377 that NUMTs range widely in size (Richly and Leister 2004), that NUMT counts vary among taxa

(Pamilo *et al.* 2007), and that sequence change slows after nuclear integration (Lopez *et al.* 1997, Bensasson *et al.* 2001a). Because of the latter property, NUMTs can illuminate deep time events (Mishmar *et al.* 2004, Miraldo *et al.* 2012). However, they can also obscure the present, especially for approaches that employ mitochondrial gene regions as a basis for specimen identification and species discovery (Buhay 2009, Andujar *et al*. 2020). This complexity arises because DNA-based biodiversity assessments employ primers that amplify the target region in diverse taxa so they also amplify NUMTs within their nuclear genomes.

Although past work has revealed NUMTs in many insect lineages (Bensasson *et al*. 2001b, Pamilo *et al.* 2007, Viljakainen *et al.* 2010, Jordal and Kambestad 2014, Francosco *et al.* 2019, Yan *et al*. 2019), no prior study has systematically characterized their abundance and attributes. In addressing this gap, the present study confronted some limitations. A third of nuclear assemblies were derived from too low coverage data to allow the estimation of genome size. Among those with adequate coverage, 20% lacked a mitogenome or corresponding COI sequence although they undoubtedly resided in the sequence data from the nuclear assembly (Vieira and Prodoscimi 2019). As genome sequencing programs expand, the joint assembly of mitochondrial and nuclear genomes should be expected.

Variation in NUMT counts

Despite data constraints, this study has provided a good overview of NUMT counts and distribution across the class Insecta. COI NUMTs were detected in all but one of the 17 orders examined, and it (Neuroptera) was only represented by two species. Among the 668 HC species, 126 lacked COI NUMTs (≥ 100 bp), but the others averaged 15.6 with counts ranging from 1 to 443. Pereira *et al.* (2021) suggested that some mitochondrial segments might be incorporated less

401 frequently into the nuclear genome than the COI barcode region, but our mitogenome-wide scan
402 did not support this hypothesis. In accord with expectations (Hazkani-Covo *et al*. 2010), NUMT
403 counts were positively correlated with genome size, and R^2 rose from 34% to 56% when analysis
404 extended from the COI barcode to the entire mitogenome. As the barcode region only represents
405 a small segment of the mitochondrial genome, this increase was expected, but it does mean that
406 the prediction of COI NUMT counts from genome size will be imprecise. However, given this
407 correlation and the mean count of 13 NUMTs for the barcode region, insect genomes likely possess
408 an average of about 325 NUMTs (as the barcode region represents 4% of the mitogenome).
409 Although larger sample sizes are required to tighten confidence estimates, our analysis revealed
410 a 100-fold difference in mean COI NUMT counts among the 17 insect orders with genome data.
411 This variation largely reflected the interplay between a deterministic factor, genome size, and a
412 stochastic process, the inclusion of COI versus another mitochondrial region in the NUMT array
413 for a species. While the latter factor impedes the prediction of NUMT counts for individual species,
414 it did not obscure an important generalization. NUMT counts are much higher in insect species
415 with direct development or incomplete metamorphosis than in those employing complete
416 metamorphosis, reflecting genome size differences. As a consequence, NUMT counts were usually
417 low in the four orders that comprise > 90% of all insect species – Coleoptera, Diptera,
418 Hymenoptera, and Lepidoptera (Stork 2018). As Orthoptera has, by far, the largest mean genome
419 size of the 27 orders, it is no surprise that the first insect NUMT was discovered in it (Gellissen *et*
420 *al.* 1983), and that many subsequent studies highlighting the complexities introduced by NUMTs
421 focused on it (Bensasson *et al.* 2001b, Song *et al.* 2008, Song *et al.* 2014, Kaya and Ciplak 2018,
422 Pereira *et al.* 2021). While the present results confirm that COI NUMTs occur in most insect
423 genomes, they do indicate that they are much less common in the most species-rich orders.

The current results provide a first sense of taxa within each major order with elevated NUMT counts, but more data is needed to allow a lineage-by-lineage threat assessment. For example, the species of Coleoptera examined in this study displayed little variation in NUMT counts, but the family with the largest genome size (e.g., Phengodidae) was not represented. Similarly, the low NUMT count for Lepidoptera reflected results from just a third of its families, and the sole species from a basal lineage (Adelidae) was a high outlier. Among 37 families of Hymenoptera, the Cynipidae possessed much larger genome sizes than the others, and its members showed elevated NUMT counts. However, some species in other families (e.g., Apidae, Formicidae) had high NUMT counts despite a small genome size, indicating that risk assessments will require consideration of other factors. Importantly, representatives of the most species-rich families of insects (e.g., Braconidae, Cecidomyiidae, Chironomidae, Ichneumonidae, Phoridae) all had low NUMT counts.

NUMT attributes and recognition

Aside from documenting their prevalence and distribution, this study has clarified two attributes that determine the influence of NUMTs on measures of species diversity – their lengths and the fraction with an IPSC. Nearly 50% of the COI NUMTs detected in this study were too short (< 100 bp) to impact most biodiversity assessments, but species did possess an average of 12.6 NUMTs above this length threshold. With a mean length of 272 bp, just 5% spanned the 658 bp barcode region and two thirds had an IPSC. As a result, only 113 of the 668 species possessed a NUMT that could elevate the apparent species count. This incidence is likely representative of other protein-coding segments of the mitogenome, but studies employing 12S or 16S rRNA will be more exposed (2x–3x) to NUMTs because an IPSC filter cannot be applied.

447

448 **Impact of NUMTs on DNA-based identification workflows**

449 The much higher copy number of mtCOI should act to reduce exposure to NUMTs. On
450 average, diploid insect nuclear genomes are about 60,000x larger than their mitochondrial
451 counterparts (1,000 Mb versus 16 Kb). In extracts prepared from whole insects, mtDNA typically
452 comprises less than 0.5% of the total DNA (Zhou *et al.* 2013). Presuming two copies of each
453 NUMT per nuclear genome, mitochondrial gene regions will enter PCR with a 150x higher count
454 (60,000 x .005/2). While this difference favours their recovery, variation in amplification can erase
455 it (e.g., a NUMT with a 20% higher PCR efficiency will dominate the final amplicon pool after 35
456 cycles). The risk of recovering a mix of mtCOI and its NUMT amplicons will extend to every
457 species whose nuclear genome includes binding sites for the primers being used. As a
458 consequence, NUMTs pose risks to all workflows underpinning DNA-based biodiversity
459 assessments – from construction of the DNA barcode reference library to its use for inferring the
460 species composition of environmental samples whether by metabarcoding or eDNA. If all NUMTs
461 were recovered, OTU counts would be inflated by 22% if analysis targeted 658 bp amplicons, by
462 58% at 300 bp, and by 111% at 150 bp. These inflation factors presume that our analysis recovered
463 all NUMTs were discovered in the 1,002 species, but many polymorphic NUMTs will have been
464 overlooked as their detection requires the analysis of multiple individuals per species (Lang *et al.*
465 2012, Dayama *et al.* 2014)

466

467 ## Conclusions

468 This study indicates that the interpretational complexities introduced by NUMTs for studies of
469 insect biodiversity vary with taxonomy, analytical protocol, and target gene region. From a

taxonomic perspective, impacts are greater for species with large genome sizes, primarily those that with direct development or incomplete metamorphosis. Because these orders comprise <10% of insect species, the overall exposure to NUMTs is reduced, but the present study detected 226 C5* NUMTs that could increase the perceived species count by 22%. Protocols targeting shorter amplicons raised the inflation value to as much as 111% because they are more abundant and less likely to possess an IPSC. Finally, the gene region employed for the DNA barcoding can impact exposure. Ribosomal genes (12S, 16S) increase NUMT exposure by 2x–3x because they cannot be filtered via IPSC scans. If used in protocols that target short amplicons (e.g., 150 bp), they could produce a 3-fold inflation in perceived taxon diversity.

The present study only considered insects, but a similar analysis on marine invertebrates generated congruent results (Schultz and Hebert 2022). The complexities introduced by NUMTs can be managed by extending informatics platforms and modifying analytical approaches. As a first step, BOLD should create a structured database for all C5* NUMTs. Based on this study, their inclusion would only increase the overall sequence count by about 30%. As well, analytical protocols such as long PCR and RT-PCR can discriminate NUMTs from their mtCOI counterparts (Schultz and Hebert 2022).

While this study has clarified the threats posed by NUMTs, empirical work is needed to understand their actual recovery and the factors that influence it. For example, does the higher initial template concentration of mtCOI reduce NUMT recovery? Is the ratio of NUMT/mtCOI reads stable or, if not, what explains the variation? These investigations have begun (Andujar et al 2020), but they must be expanded to both understand and mitigate the impacts of NUMTs on biodiversity assessments.

Acknowledgements

We thank Suz Bateson for aid with graphics while Sujeevan Ratnasingham and Thomas Braukmann provided insightful comments on earlier drafts of this manuscript. The research was enabled by an award from the Government of Canada's New Frontiers in Research Fund (NFRF), [NFRFT-2020-00073]. Additionally, this work was funded by the Government of Canada through Genome Canada and Ontario Genomics (OGI-208), as part of BIOSCAN–Canada Large Scale Applied Research Program. PDNH gratefully acknowledges support from the Canada Research Chairs program.

References

Andujar, C., Creedy, T.J., Arribas, P., López, H., Salces-Castellano, A., Pérez-Delgado, A.J., Vogler, A.P., & Emerson, B.C. (2020). Validated removal of nuclear pseudogenes and sequencing artefacts from mitochondrial metabarcode data. *Molecular Ecology Resources, 21*, 1772–1787. doi: 10.1111/1755-0998.13337

Bensasson, D., Zhang, D.-X., Hartl, D.L., & Hewitt, G.M. (2001a). Mitochondrial pseudogenes: evolution's misplaced witnesses. *Trends in Ecology and Evolution, 16*, 314–321.

Bensasson, D., Zhang, D.-X., & Hewitt, G.M. (2001b). Frequent assimilation of mitochondrial DNA by grasshopper nuclear genomes. *Molecular Biology and Evolution, 17*, 406–415.

Bernt, M., Donath, A., Juhling, F., Externbrink, F., Florentz, C., Fritzsch, G., Putz, J., Middendorf, M., Stadler, P.F. (2013). MITOS: improved de novo metazoan mitochondrial genome annotation. *Molecular Phylogenetics and Evolution, 69*, 313–319.

Boore, J.L. (1999). Animal mitochondrial genomes. *Nucleic Acids Research, 27*, 1767–1780.

Buhay, J.E. (2009). COI-like sequences are becoming problematic in molecular systematic and DNA barcoding studies. *Journal of Crustacean Biology, 29*, 96–110.

Chamberlain, S.A., & Szocs, E. (2013). taxize: taxonomic search and retrieval in R. *F1000Research*, *2*, 191.

Chatre, L., & Ricchetti, M. (2011). Nuclear mitochondrial DNA activates replication in *Saccharomyces cervisiae*. *PLOS ONE, 6*, e17235.

Creedy, T. J., Norman, H., Tang, C. Q., Qing Chin, K., Andujar, C., Arribas, P., ... Vogler, A. P. (2020). A validated workflow for rapid taxonomic assignment and monitoring of a

national fauna of bees (Apiformes) using high throughput DNA barcoding. *Molecular Ecology Resources*, 20, 40–53. https://doi.org/10.1111/1755-0998.13056

Dayama, G., Emery, S.B., Kidd, J.M., & Mills, R.E. (2014). The genomic landscape of polymorphic human nuclear mitochondrial insertions. *Nucleic Acids Research, 42*, 12640–12649.

D'Ercole, J., Prosser, S.W.J., & Hebert, P.D.N. (2021). A SMRT approach for targeted amplicon sequencing of museum specimens (Lepidoptera) – patterns of nucleotide divergence. *PeerJ, 9*, e10420.

Dunshea, G., Barros, N.B., Wells, R.S., Gales, N.J., Hindell, M.A., & Jarman, S.N. (2008). Pseudogenes and DNA-based diet analyses: a cautionary tale from a relatively well sampled predator-prey system. *Bulletin of Entomological Research, 98*, 239–248.

Edler, D., Klein, J., Antonelli, A., & Silvestro, D. (2020). raxmlGUI 2.0: A graphical interface and toolkit for phylogenetic analyses using RAxML. *Methods in Ecology and Evolution, 12*, 373–377.

Francoso, E., Zuntini, A.R., Ricardo, P.C., Silva, J.P.N., Brito, R., Oldroyd, B.P., & Arias, M.C. (2019). Conserved numts mask a highly divergent mitochondrial-COI gene in a species complex of Australian stingless bees *Tetragonula* (Hymenoptera: Apidae). *Mitochondrial DNA Part A, 30*, 806–817.

Gellissen, G., Bradfield, J.Y., White, B.N., & Wyatt, G.R. (1983). Mitochondrial DNA sequences in the nuclear genome of a locust. *Nature, 301*, 631–634.

544 Gunbin, K., Peshkin, L., Popadin, K., Annis, S., Ackermann, R.R., & Khrapko, K. (2017).
545 Integration of mtDNA pseudogenes coincides with speciation of the human genus. A
546 hypothesis. *Mitochondrion, 34*, 20–23.

547 Hazkani-Covo, E., Zeller, R.M. & Martin, W. (2010). Molecular poltergeists: mitochondrial DNA
548 copies (*numts*) in sequenced nuclear genomes. *PLOS Genetics, 6*, e1000834.

549 Hazkani-Covo, E., Sorek, R., & Graur, D. (2003). Evolutionary dynamics of large *Numts* in the
550 human genome: rarity of independent insertions and abundance of post-insertion
551 duplications. *Journal of Molecular Evolution, 56*, 169–174.

552 Hebert, P.D.N., Braukmann, T.W.A., Prosser, S.W.J., Ratnasingham, S., deWaard, J.R., Ivanova,
553 N.V., Janzen, D.H., Hallwachs, W., Naik, S., Sones, J.E., & Zakharov, E.V. (2018). A
554 Sequel to Sanger: Amplicon sequencing that scales. *BMC Genomics, 19*, 219.

555 Hebert, P.D.N., deWaard, J.R., Zakharov, E.V., Prosser, S.W.J., Sones, J.E., McKeown, J.T.A.,
556 Mantle, D.B., & La Salle, J. (2013). A DNA 'Barcode Blitz': Rapid digitization and
557 sequencing of a natural history collection. *PLOS ONE, 8*, e68535.

558 Hebert, P.D.N., Cywinska, A., Ball, S.L., & deWaard, J.R. (2003). Biological identifications
559 through DNA barcodes. *Proceedings of the Royal Society B: Biological Sciences*, *270*,
560 313–321.

561 Hellberg, R.S., Isaacs, R.B., & Hernandez, E.L. (2019). Identification of shark species in
562 commercial products using DNA barcoding. *Fisheries Research, 210*, 81–88.

563 Henze, K., & Martin, W. (2001). How do mitochondrial genes get into the nucleus? *Trends in*
564 *Genetics, 17*, 383–387.

565 Jordal, B.H., & Kambestad, M. (2014). DNA barcoding of bark and ambrosia beetles reveals
566 excessive NUMTs and consistent east-west divergence across Palearctic forests.
567 *Molecular Ecology Resources, 14*, 7–17.

568 Kaya, S., & Ciplak, B. (2018). Possibility of numt co-amplification from gigantic genome of
569 Orthoptera: testing efficiency of standard PCR protocol in producing orthologous COI
570 sequences. *Heliyon, 4*, e000929.

571 Lang, M., Sazzini, M., Calabrese, F.M., Simone, D., Boattini, A., Romeo, G., Luiselli, D.,
572 Attimonelli, M., & Gasparre, G. (2012). Polymorphic NumtS trace human population
573 relationships. *Human Genetics, 131*, 757–771.

574 Larsson, A. (2014). AliView: A fast and lightweight alignment viewer and editor for large datasets.
575 *Bioinformatics, 30*, 3276–3278. https://doi.org/10.1093/bioinformatics/btu531.

576 Lopez, J.V., Culver, M., Stephens, J.C., Johnson, W.E., & O'Brien, S.J. (1997). Rates of nuclear
577 and cytoplasmic mitochondrial DNA sequence divergence in mammals. *Molecular*
578 *Biology and Evolution, 14*, 277–286.

579 McCombie, W.R., McPherson, J.D., & Mardis, E.R. (2019). Next-generation sequencing
580 technologies. *Cold Spring Harbor Perspectives in Medicine, 1*, 9. doi:
581 10.1101/cshperspect.a036798.

582 Michalovova, M., Vyskot, B., & Kejnovsky, E. (2013). Analysis of plastid and mitochondrial
583 DNA insertions in the nucleus (NUPTS and NUMTS) of six plant species: size, relative
584 age, and chromosomal location. *Heredity, 111*, 314–320.

585 Miraldo, A., Hewitt, G.M., Dear, P.H., Paulo, O.S., & Emerson, B.C. (2012). Numts help to
586 reconstruct the demographic history of the ocellated lizard (*Lacerta lepida*) in a secondary
587 contact zone. *Molecular Ecology, 21*, 1005–1018.

588 Mishmar, D., Ruiz-Pesinim, E., Brandon, M., & Wallace, D.C. (2004). Mitochondrial DNA-like
589 sequences in the nucleus (NUMTs): insights into our African origins and the mechanism
590 of foreign DNA integration. *Human Mutation, 23*, 125–133.

591 Nithaniyal, S., Majumder, S., Umapathy, S., & Parani, M. (2021). Forensic application of DNA
592 barcoding in the identification of commonly occurring poisonous plants. *Journal of*
593 *Forensic and Legal Medicine, 78*, 102126. doi: 10.1016/j.jflm.2021.102126.

594 Pamilo, P., Viljakainen, L., & Vihavainen, A. (2007). Exceptionally high density of NUMTs in
595 the honeybee genome. *Molecular Biology and Evolution, 24*, 1340–1346.

596 Pearson, W.R. (2013). An introduction to sequence similarity ("homology") searching. *Current*
597 *Protocols in Bioinformatics, 42*, 3.1.1-3.1.8. doi: 10.1002/0471250953.bi0301s42.

598 Pereira, R.J., Ruiz-Ruano, F.J., Thomas, C.J.E., Perez-Ruiz, M., Jimenez-Bartolome, M., Liu, S.,
599 de la Torre, J., & Bella, J.L. (2021). Mind the *numt*: Finding informative mitochondrial
600 markers in a giant grasshopper genome. *Journal of Zoological Systematics and*
601 *Evolutionary Research, 59*, 635–645.

602 Perna, N.T., & Kocher, T.D. (1996). Mitochondrial DNA: Molecular fossils in the nucleus.
603 *Current Biology, 6*, 128–129.

Prosser, S., deWaard, J.R., Miller, S.E., & Hebert, P.D.N. (2016). DNA barcodes from century-old type specimens using next generation sequencing. *Molecular Ecology Resources, 16*, 487–497.

Quail, M., Smith, M. E., Coupland, P., Otto, T. D., Harris, S. R., Connor, T. R., Bertoni, A., Swerdlow, H. P., & Gu, Y. (2012). A tale of three next generation sequencing platforms: Comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genomics*, *13*, 341. https://doi.org/10.1186/1471-2164-13-341

Quinlan, A.R. (2014). BEDTools: the swiss-army tool for genome feature analysis. *Current Protocols in Bioinform*atics, *47*, 11.12.1–11.12.34.

R Development Core Team. (2011). R: A Language and Environment for Statistical Computing (http://www.R-project.org/). R Foundation for Statistical Computing.

Ratnasingham, S., & Hebert, P.D.N. (2007). BOLD: The Barcode of Life Data System (www.barcodinglife.org). *Molecular Ecology Notes, 7*, 355–364. https://doi.org/10.1111/j.1471-8286.2007.01678.x.

Ratnasingham, S., & Hebert, P.D.N. (2013). A DNA-based registry for all animal species: The Barcode Index Number (BIN) system. *PLoS One, 8*, e66213. https://doi.org/10.1371/journal.pone.0066213.

Ricchetti, M., Fairhead, C., & Dujon, B. (1999). Mitochondrial DNA repairs double-strand breaks in yeast chromosomes. *Nature*, *402*, 96–100.

Richly, E., & Leister, D. (2004). NUMTs in sequenced eukaryotic genomes. *Molecular Biology and Evolution, 21*, 1081–1084.

625 Rinkert, A., Misiewicz, T.M., Carter, B.E., Salmaan, A., & Whittall, J.B. (2021). Bird nests as
626 botanical time capsules: DNA barcoding identifies the contents of contemporary and
627 historical nests. *PLoS One, 16*, e0257624.

628 Schultz, J.A., & Hebert, P.D.N. (2022). Do pseudogenes pose a problem for metabarcoding marine
629 animal communities? *Molecular Ecology Resources*, 00, 1-18,
630 https://doi.org/10.1111/1755-0998.13667.

631 Song, H., Moulton, M.J., & Whiting, M.F. (2014). Rampant nuclear insertion of mtDNA across
632 diverse lineages within the Orthoptera (Insecta) *PLOS ONE, 9*, 2110508.

633 Song H., Buhay, J.E., Whiting, M.F., & Crandall, K.A. (2008). Many species in one: DNA
634 barcoding overestimates the number of species when nuclear mitochondrial pseudogenes
635 are coamplified. *Proceedings of the National Academy of Sciences USA, 105*, 13486–
636 13491.

637 Stork, N.E. (2018). How many species of insects and other terrestrial arthropods are there on earth?
638 *Annual Review of Entomology, 63*, 31–45.

639 Thalmann, O., Hebler, J., Poinar, H.N., Paabo, S., & Vigilant, L. (2004). Nuclear insertions help
640 and hinder inference of the evolutionary history of gorilla mt DNA. *Molecular Ecology,*
641 *14*, 179–188.

642 Tourmen, Y., Baris, O., Dessen, P., Jacques, C., Malthiery, Y., & Reynier, P. (2002). Structure
643 and chromosomal distribution of mitochondrial pseudogenes. *Genomics, 80*, 71–77.

644 Vieira, G.A., & Prosdocimi, F. (2019). Accessible molecular phylogenomics at no cost: obtaining
645 14 new mitogenomes for the ant subfamily Pseudomyrmecinae from public data. *PeerJ,*
646 *7*, e6271.

647 Viljakainen, L., Oliveira, D.C.S.G., Werren, J.H., & Behura, S.K. (2010). Transfers of
648 mitochondrial DNA to the nuclear genome in the wasp *Nasonia vitripennis*. *Insect*
649 *Molecular Biology, 19*, 27–35

650 Yan, Z., Fang, Q., Tian, Y., Wang, F., Chen, X., Werren, J.H., & Ye, G. (2019). Mitochondrial
651 DNA and their nuclear copies in the parasitic wasp *Pteromalus puparum*: A comparative
652 study in the Chalcidoidea. *International Journal of Biological Macromolecules, 121*, 572–
653 579.

654 Yu, G., Smith, D., Zhu, H., Guan, Y., & Lam, T.T. (2017). ggtree: an R package for visualization
655 and annotation of phylogenetic trees with their covariates and other associated data.
656 *Methods in Ecology and Evolution, 8*, 28–36. doi: 10.1111/2041-210X.12628.

657 Yu, X., & Gabriel, A. (1999). Patching broken chromosomes with extranuclear cellular DNA.
658 *Molecular Cell, 4*, 873–881.

659 Zhou, X., Li, Y., Liu, S., Yang, Q., Su, X., Zhou, L., Tang, M., Fu, R., Li, J., & Huang, Q. (2013).
660 Ultra-deep sequencing enables high-fidelity recovery of biodiversity from bulk arthropod
661 samples without PCR amplification. *GigaScience, 2*, 4.

Conflict of Interest

The authors declare no conflict of interest.

Data Availability Statement

The COI sequences used to query the 1,002 insect genomes, together with information on the source specimen for each record, are available as a dataset (DS-NUMTINS) on BOLD dx.doi.org/10.5883/DS-NUMTINS. Supplementary Figures and Tables in the Supporting Information document provide much of the data, but three Supplementary Tables and a Supplementary Figure are attached to the project dataset on BOLD (www.boldsystems.org). They are also directly available at the following URLs: **Table S1**– https://bit.ly/3wUdFOr-TableS1; **Table S2** – https://bit.ly/3PHSXdt-TableS2; **Table S4** – https://bit.ly/3sZkf5r-TableS4; **Figure S5** – https://bit.ly/38PVkKA-FigureS5). All custom scripts employed in data analysis are available on Zenodo at https://doi.org/10.5281/zenodo.6584411.

Author Contributions

PDNH designed the study, secured the funding, and led assembly of the manuscript. DGB and SWJP led data acquisition/analysis and composed key sections of the manuscript.

680 **Table 1**: *NUMT attributes for 668 insect species with high (≥ 5x) and 334 species with low (<*

681 *5x) coverage nuclear assemblies. Fifteen species with uncertain coverage were assigned to the*

682 *low category.*

Coverage	n	# NUMTs	Count Mean/Range	Length Mean/Range (bp)	Proportion with IPSC
High	668	8,423	12.6; 0–443	271 ± 177; 100–754	0.67
Low	334	1,380	4.1; 0–49	254 ± 164; 100–709	0.46

683

684 **Table 2**: *Number of NUMTs with/without IPSCs in five length categories and mean sequence*
685 *divergence between these NUMTs and their mitochondrial COI homologue. Analysis considered*
686 *the 668 species with a high coverage genome.*

687

	With IPSC		Without IPSC		
Size Range (bp)	**# NUMTs**	**Mean Divergence (%)**	**# NUMTs**	**Mean Divergence (%)**	**Total**
100–150	1,453	19.1 ± 6.7	1,092	12.01 ± 7.1	2,545
151–300	2,401	19.6 ± 7.5	978	10.8 ± 7.3	3,379
301–450	693	19.2 ± 8.1	238	7.9 ± 6.5	931
451–600	446	18.2 ±9.3	135	7.2 ± 5.9	581
601–661	614	13.0 ± 8.5	373	4.8 ± 4.1	987
Total	**5,607**	**18.6 ± 7.9**	**2,816**	**10.1 ± 7.2**	**8,423**

688

689 **Table 3**: *Mean genome size and counts for two lengths of COI NUMTs. Analysis considers 668*
690 *species with genome assemblies ≥ 5x belonging to 17 orders. * Orders developing via incomplete*
691 *metamorphosis. Other orders develop via complete metamorphosis.*

692

Order	n	Mean Genome Size (Mb)	Mean Count ≥ 100 bp	Mean Count ≥ 658 bp
Orthoptera*	5	2391	140.6	4.4
Phasmatodea*	2	2318	138.5	6.0
Blattodea*	5	1558	84.0	1.2
Odonata*	2	1146	32.5	0.0
Plecoptera*	2	371	30.5	0.5
Hemiptera*	49	660	23.0	1.0
Ephemeroptera*	2	327	1.0	0.5
Siphonaptera	1	776	51.0	0.0
Megaloptera	1	768	28.0	0.0
Hymenoptera	131	330	17.8	0.8
Coleoptera	54	562	9.8	0.7
Diptera	213	284	8.0	0.6
Strepsiptera	1	156	7.0	0.0
Lepidoptera	190	529	5.3	0.3
Trichoptera	7	791	5.0	1.3
Thysanoptera	1	416	1.0	0.0
Neuroptera	2	549	0.0	0.0
Total	**668**	**453**	**12.6**	**0.6**

693

694 **Table 4**: *Impact of analytical protocol on exposure to non-diagnosable NUMTs (i.e., those without*
695 *an IPSC) for the 668 HC species. For NUMT # see Table 2. Total/species = # of amplicons x*
696 *NUMT exposure/668. See Materials section for explanation of the exposure value.*

697

Protocol	# Amplicons	Length	# NUMT x Exposure	Total/Species
Barcode Library	1	651–661	C5* = 226 x 1 **TOTAL = 226**	226/668 = 0.34
Barcode Library	2	300–450	C3 = 238 x 0.57 + C4 = 135 x 0.80 + C5 = 349 x 0.96 **TOTAL = 578**	578/668 x 2 = 1.74
Barcode Library	5	1001–50	C1 = 1092 x 0.19 C2 = 978 x 0.34 C3 = 238 x 0.57 C4 = 135 x 0.80 C5 = 349 x 0.96 **TOTAL = 1118**	1118/668 x 5 = 8.35
eDNA	1	100–150	C1 = 1092 x 0.19 C2 = 978 x 0.34 C3 = 238 x 0.57 C4 = 135 x 0.80 C5 = 349 x 0.96 **TOTAL = 1118**	1118/668 = 1.67
Metabarcoding	1	300–450	C3 = 238 x 0.57 + C4 = 135 x 0.80 + C5 = 349 x 0.91 **TOTAL = 578**	0.87

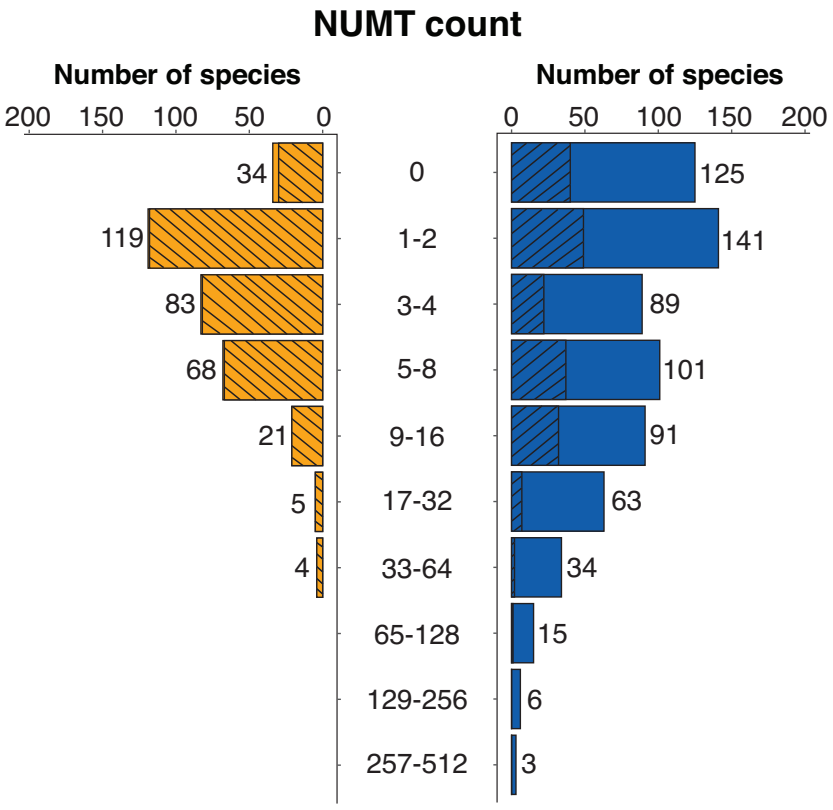
698

699 **Figure 1**: *Number of NUMTs (≥ 100 bp) derived from the barcode region of COI for 1,002 insect*

700 *species. NUMT counts are plotted separately for species with low (334) and high (668) coverage*

701 *assemblies. Orange = low; blue = high; slashed bars = Lepidoptera. Analysis considered species*

702 *with both a DNA barcode sequence and a nuclear assembly with a coverage estimate.*



703

Figure 2: *Plot of the 8,423 COI NUMTs (≥ 100 bp) identified in high coverage nuclear genomes from 668 insect species. The length of each NUMT is shown as well as its sequence divergence from mitochondrial COI. Values > 658 bp arise through insertions while those < 658 bp reflect deletions or the original incorporation of a truncated fragment. Green = NUMT with a frameshift indel and/or a stop codon. Red = NUMT lacking these features.*

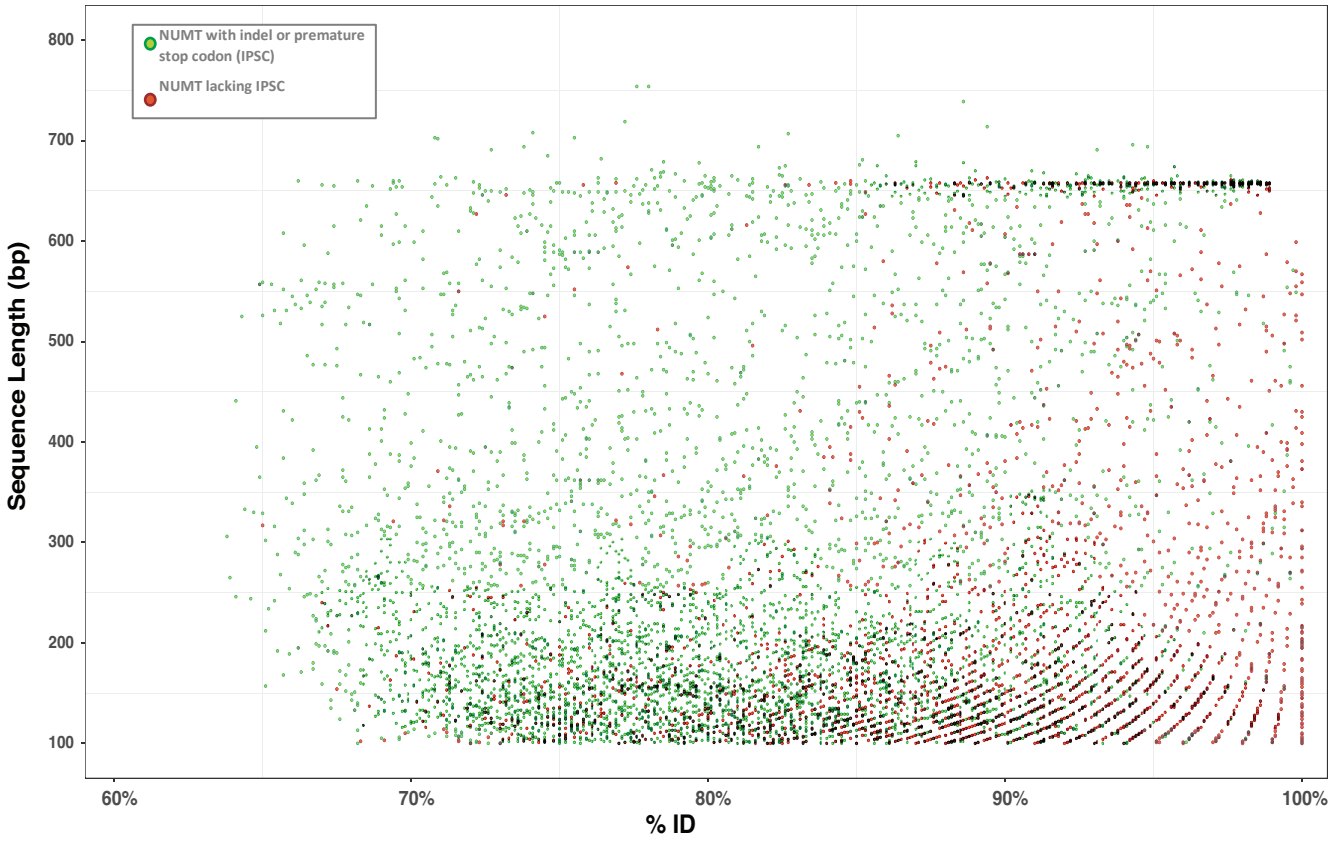


Figure 3: *Length distribution of COI NUMTs for 283 insect species as revealed by using a full-length (ca. 1,500 bp) COI query. Lengths only show the region corresponding to COI; the secondary peak circa 1,500 bp reflects NUMTs that extend beyond COI and those with an internal insertion. The proportion of NUMTs with a frameshift indel or premature stop codon (IPSC) is shown for each length category.*

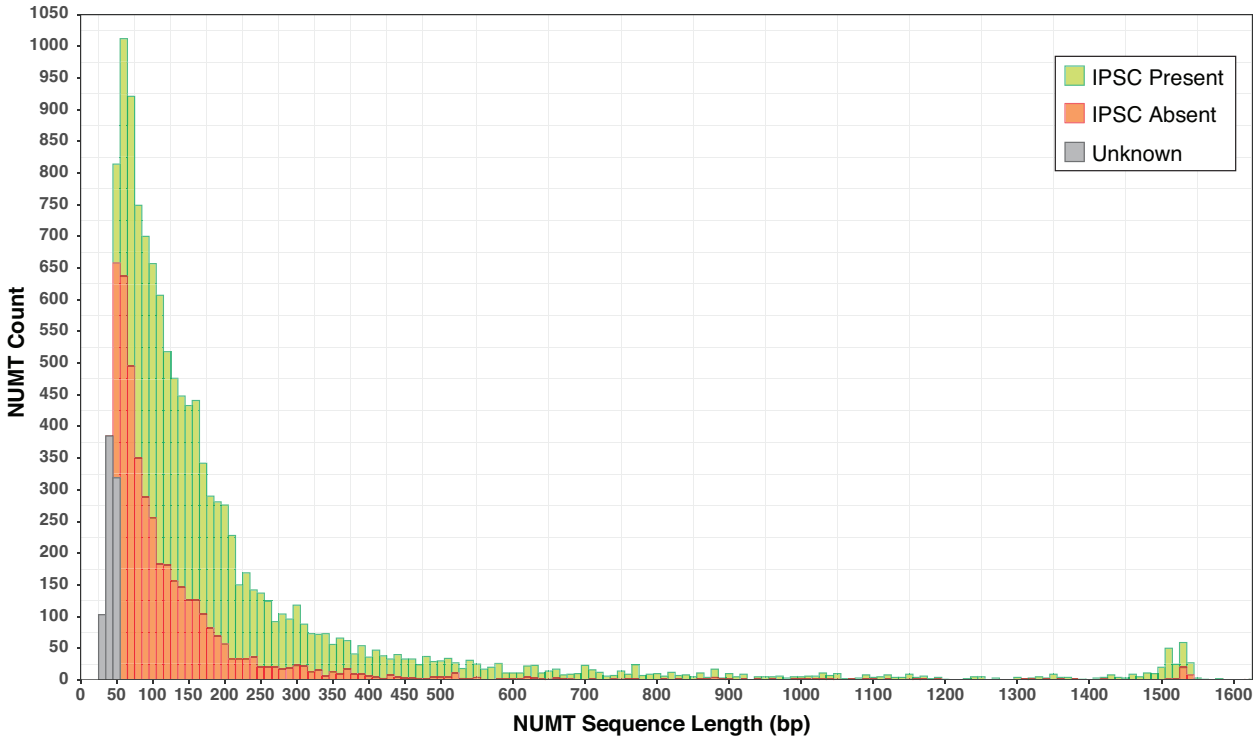


Figure 4: *Correlation between NUMT counts for the COI barcode and average mitogenome-wide NUMT counts for species with a mitogenome and 5x nuclear assembly. 77 species lacking NUMTs were excluded from analysis. (a) 242 species; (b) One species from each of the 191 genera. The red dashed line has an intercept of 0 and a slope of 1.*

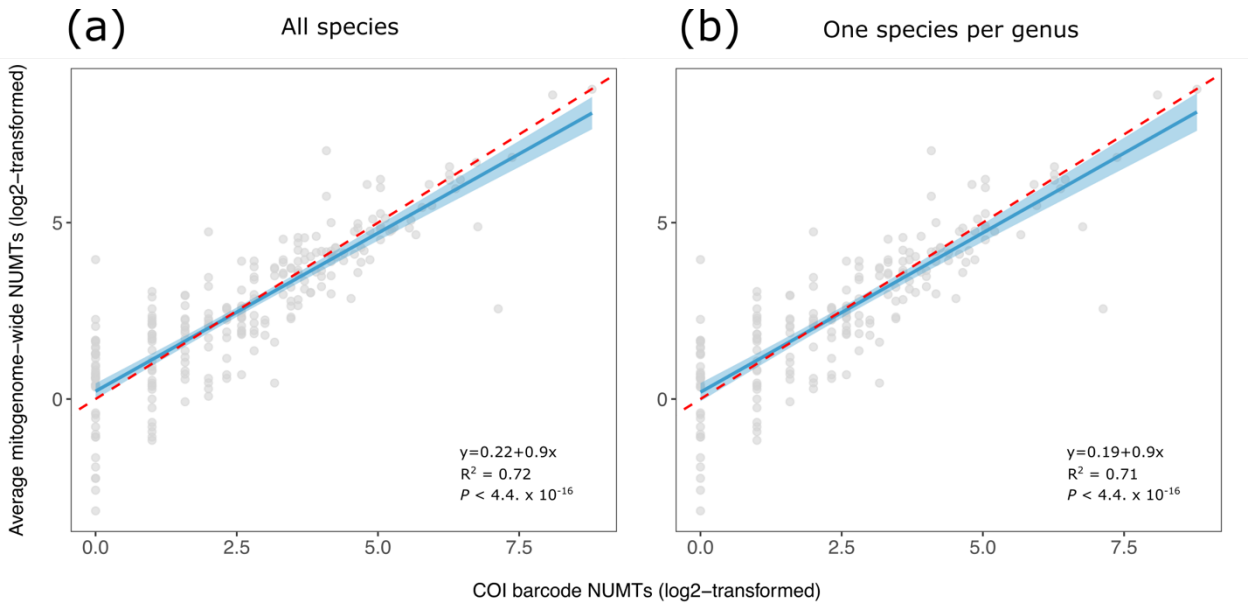
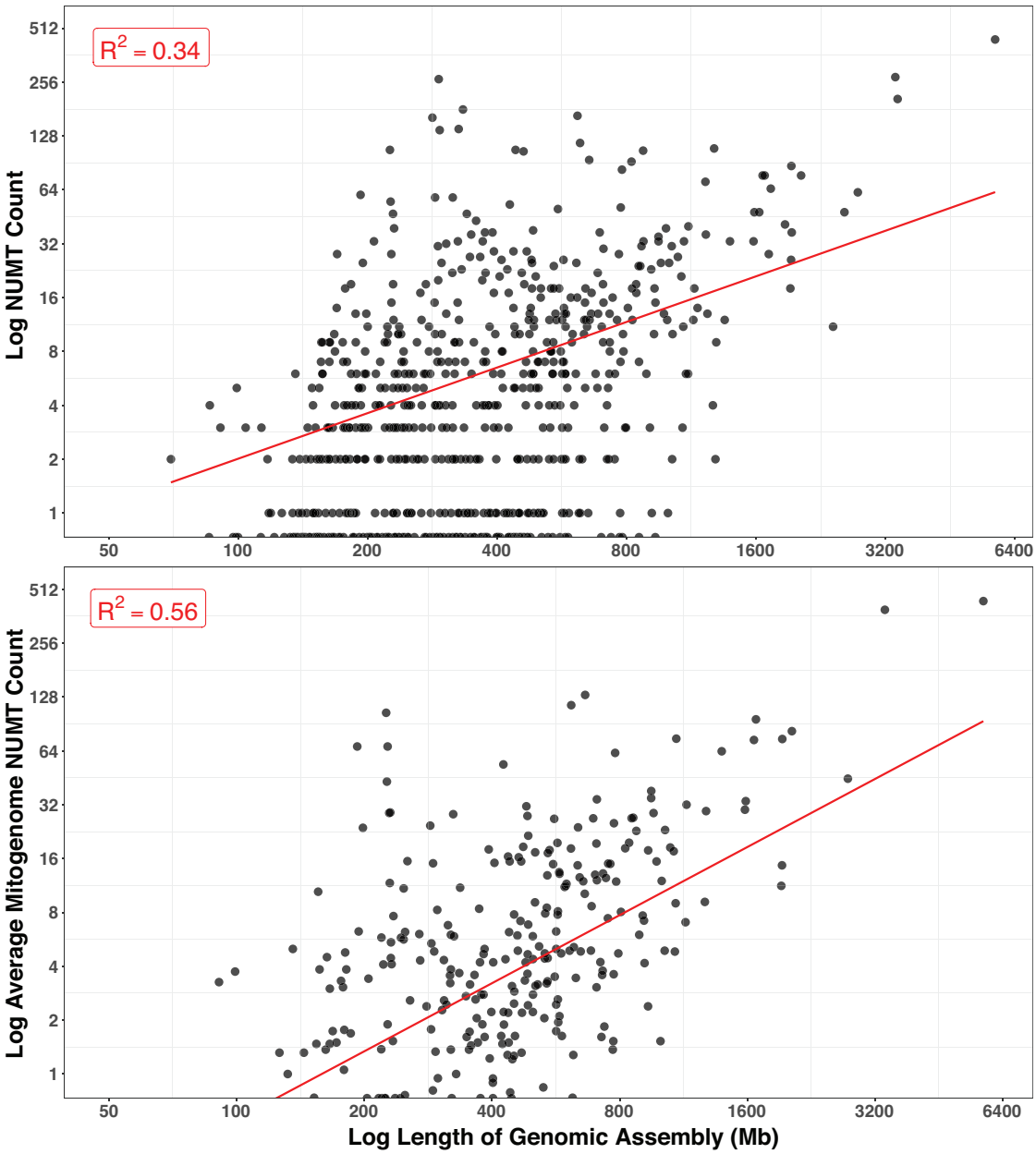
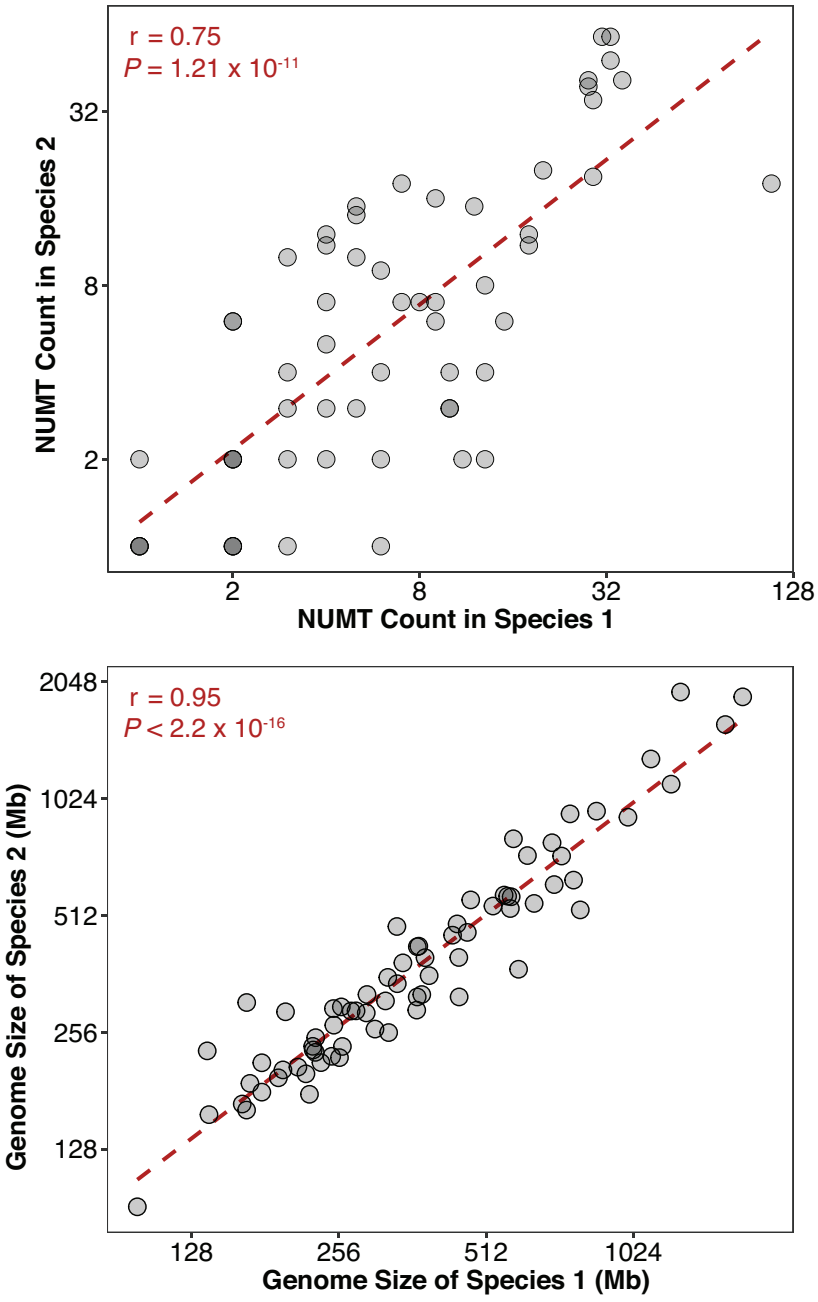


Figure 5: *Correlation between log assembly length and log$_2$ count of COI NUMTs (≥ 100 bp) for species with ≥ 5x coverage. Above: NUMT counts for COI barcode region (n = 668). Below: Mean NUMT counts for 658 bp segments of entire mitogenome (n =391).*



725 **Figure 6**:*Correlation in log_2 NUMT (≥ 100 bp) counts and log_2 genome size for 72 pairs of*

726 *congeneric species.*



727

Figure 7: *Box plots comparing the COI NUMT count (rank-transformed) in species with complete or incomplete metamorphosis for five insect orders represented by more than 50 species and for composites of other orders with fewer species.*

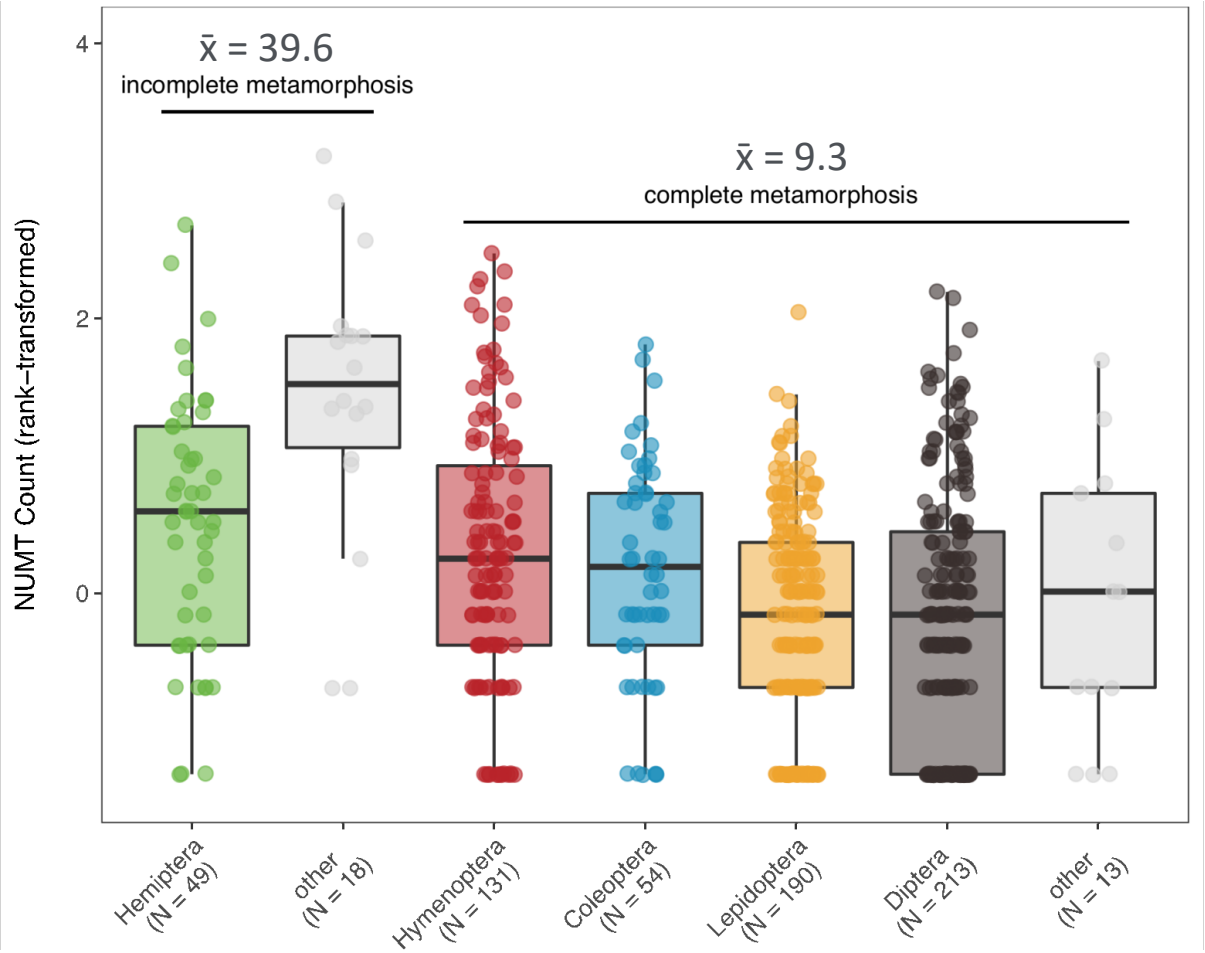


Figure 8: *Circular cladograms for insect species with high coverage nuclear genomes based on sequence divergence in the 658 bp COI barcode region. Bars at the tip of each node indicate NUMT count or genome size. Upper Panel – 637 species in five major insect orders. Lower panel – 31 species in 12 minor orders with those employing incomplete metamorphosis first and those with complete metamorphosis next.*

Figure 9: *Bivariate plots showing the length and sequence divergence from mitochondrial COI for each NUMT in ten species with the smallest and largest reported genome size for each of the five major insect orders. Gray indicates NUMTs with an IPSC (indel and/or premature stop codon). Other colours indicate five length categories of NUMTs lacking these features.*

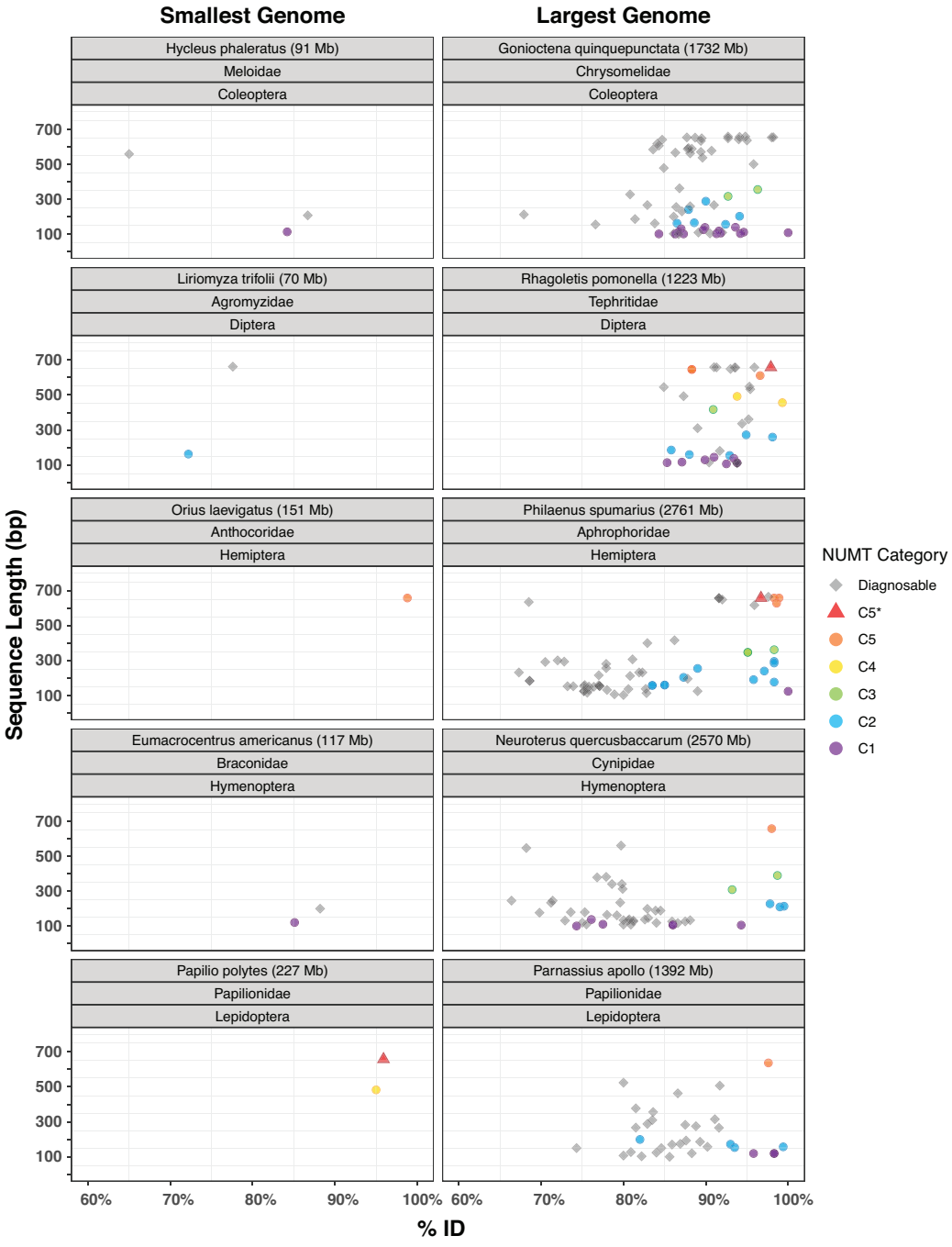
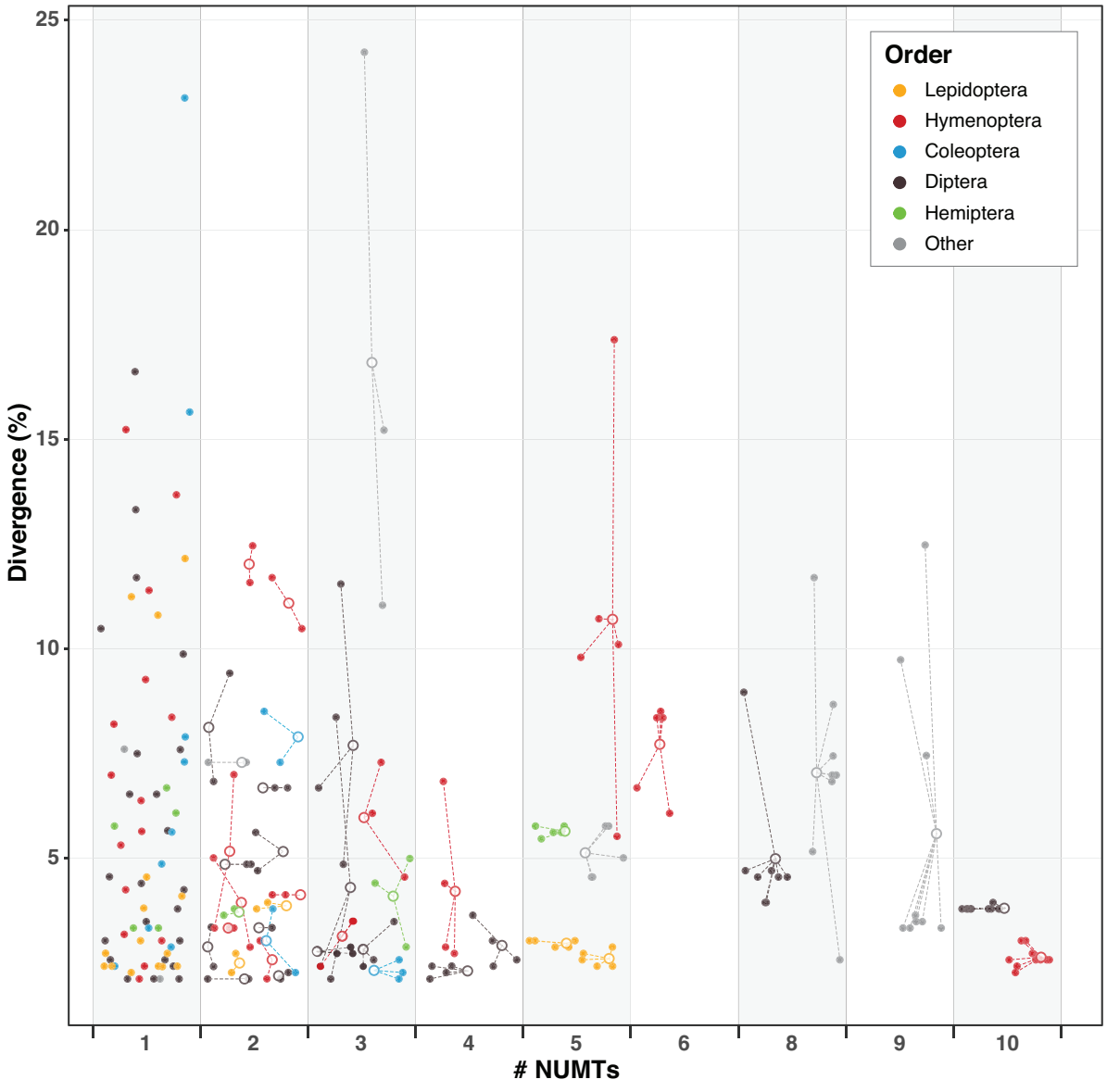
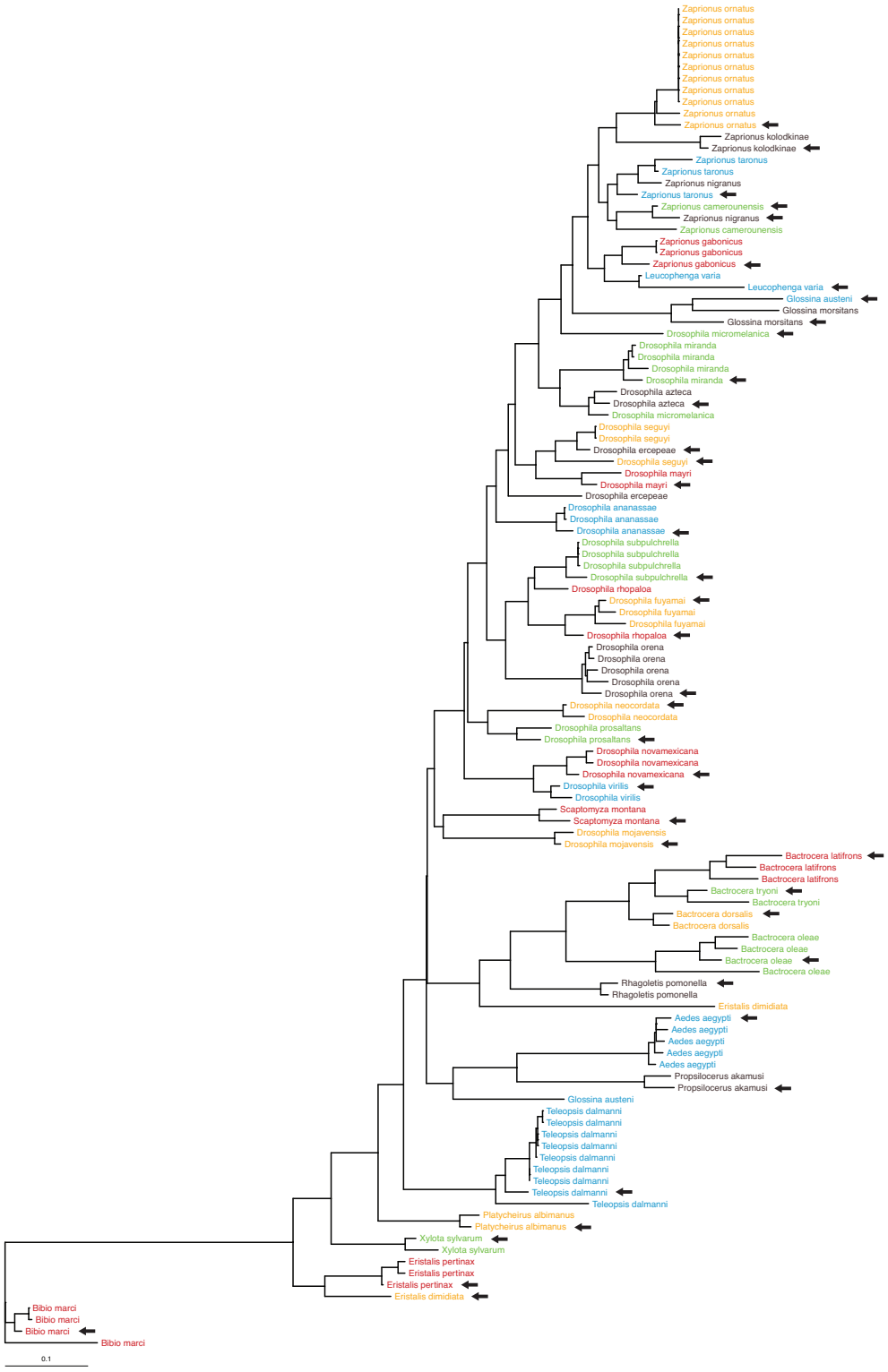


Figure 10: *Plot of C5* NUMTs (i.e., those with no IPSC, > 2% divergence from mitochondrial homologue, length = 651–661 bp) in 113 species. All C5* NUMTs (solid circles) from a species are connected via dotted lines to a point representing the mean divergence (open circles) for that species (this connection is absent in species with only one C5* NUMT). The other 555 HC species lacked C5* NUMTs. The other orders include Orthoptera (lanes 5 & 9), Phasmatodea (lane 8), Blattodea (lane 1=7% divergence, lane 2 & 3), and Plecoptera (lane 1= 2% divergence).*

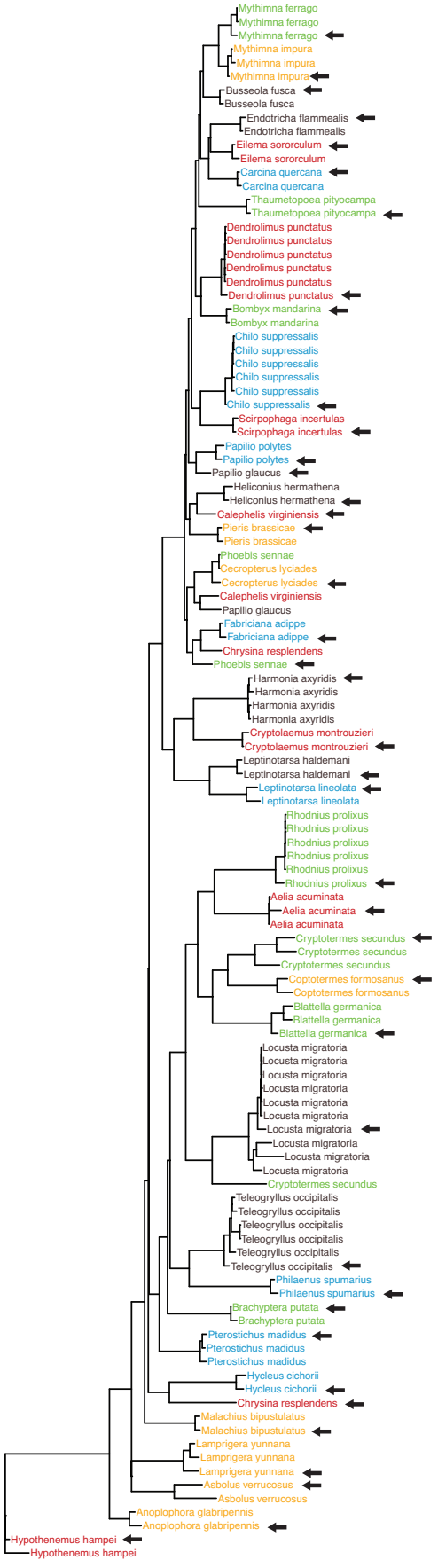


760 **Figure 11**: *Maximum likelihood tree displaying sequence similarities between 226 C5* NUMTs*

761 *and the mtCOI from their 113 source species. Arrows indicate mtCOI.*



762


Mythimna ferrago
Mythimna ferrago
Mythimna ferrago
Mythimna impura
Mythimna impura
Mythimna impura
Busseola fusca
Busseola fusca
Endotricha flammealis
Endotricha flammealis
Eilema sororculum
Eilema sororculum
Carcina quercana
Carcina quercana
Thaumetopoea pityocampa
Thaumetopoea pityocampa
Dendrolimus punctatus
Dendrolimus punctatus
Dendrolimus punctatus
Dendrolimus punctatus
Dendrolimus punctatus
Dendrolimus punctatus
Bombyx mandarina
Bombyx mandarina
Chilo suppressalis
Chilo suppressalis
Chilo suppressalis
Chilo suppressalis
Chilo suppressalis
Chilo suppressalis
Scirpophaga incertulas
Scirpophaga incertulas
Papilio polytes
Papilio polytes
Papilio glaucus
Heliconius hermathena
Heliconius hermathena
Calephelis virginiensis
Pieris brassicae
Pieris brassicae
Phoebis sennae
Cecropterus lyciades
Cecropterus lyciades
Calephelis virginiensis
Papilio glaucus
Fabriciana adippe
Fabriciana adippe
Chrysina resplendens
Phoebis sennae
Harmonia axyridis
Harmonia axyridis
Harmonia axyridis
Harmonia axyridis
Cryptolaemus montrouzieri
Cryptolaemus montrouzieri
Leptinotarsa haldemani
Leptinotarsa haldemani
Leptinotarsa lineolata
Leptinotarsa lineolata
Rhodnius prolixus
Rhodnius prolixus
Rhodnius prolixus
Rhodnius prolixus
Rhodnius prolixus
Rhodnius prolixus
Aelia acuminata
Aelia acuminata
Aelia acuminata
Cryptotermes secundus
Cryptotermes secundus
Cryptotermes secundus
Coptotermes formosanus
Coptotermes formosanus
Blattella germanica
Blattella germanica
Blattella germanica
Locusta migratoria
Locusta migratoria
Locusta migratoria
Locusta migratoria
Locusta migratoria
Locusta migratoria
Locusta migratoria
Locusta migratoria
Locusta migratoria
Locusta migratoria
Cryptotermes secundus
Teleogryllus occipitalis
Teleogryllus occipitalis
Teleogryllus occipitalis
Teleogryllus occipitalis
Teleogryllus occipitalis
Teleogryllus occipitalis
Philaenus spumarius
Philaenus spumarius
Brachyptera putata
Brachyptera putata
Pterostichus madidus
Pterostichus madidus
Pterostichus madidus
Hycleus cichorii
Hycleus cichorii
Chrysina resplendens
Malachius bipustulatus
Malachius bipustulatus
Lamprigera yunnana
Lamprigera yunnana
Lamprigera yunnana
Asbolus verrucosus
Asbolus verrucosus
Anoplophora glabripennis
Anoplophora glabripennis
Hypothenemus hampei
Hypothenemus hampei
0.1

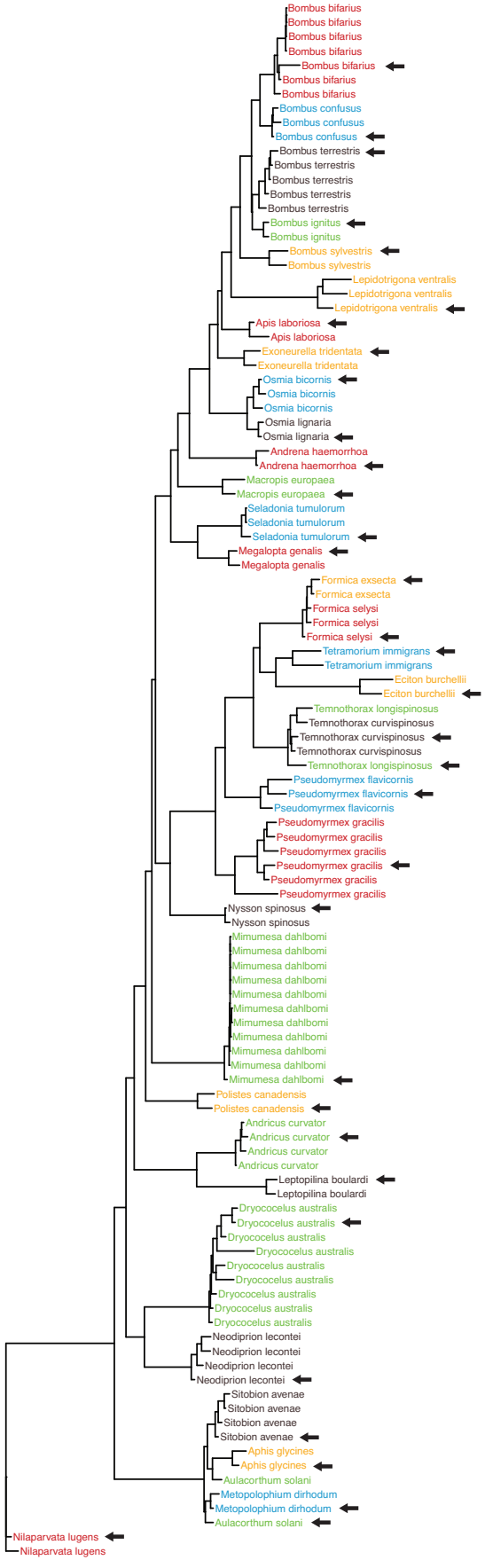

Bombus bifarius
Bombus bifarius
Bombus bifarius
Bombus bifarius
Bombus bifarius
Bombus bifarius
Bombus bifarius
Bombus bifarius
Bombus confusus
Bombus confusus
Bombus confusus
Bombus terrestris
Bombus terrestris
Bombus terrestris
Bombus terrestris
Bombus terrestris
Bombus ignitus
Bombus ignitus
Bombus sylvestris
Bombus sylvestris
Lepidotrigona ventralis
Lepidotrigona ventralis
Lepidotrigona ventralis
Apis laboriosa
Apis laboriosa
Exoneurella tridentata
Exoneurella tridentata
Osmia bicornis
Osmia bicornis
Osmia bicornis
Osmia lignaria
Osmia lignaria
Andrena haemorrhoa
Andrena haemorrhoa
Macropis europaea
Macropis europaea
Seladonia tumulorum
Seladonia tumulorum
Seladonia tumulorum
Megalopta genalis
Megalopta genalis
Formica exsecta
Formica exsecta
Formica selysi
Formica selysi
Formica selysi
Tetramorium immigrans
Tetramorium immigrans
Eciton burchellii
Eciton burchellii
Temnothorax longispinosus
Temnothorax curvispinosus
Temnothorax curvispinosus
Temnothorax curvispinosus
Temnothorax longispinosus
Pseudomyrmex flavicornis
Pseudomyrmex flavicornis
Pseudomyrmex flavicornis
Pseudomyrmex gracilis
Pseudomyrmex gracilis
Pseudomyrmex gracilis
Pseudomyrmex gracilis
Pseudomyrmex gracilis
Pseudomyrmex gracilis
Nysson spinosus
Nysson spinosus
Mimumesa dahlbomi
Mimumesa dahlbomi
Mimumesa dahlbomi
Mimumesa dahlbomi
Mimumesa dahlbomi
Mimumesa dahlbomi
Mimumesa dahlbomi
Mimumesa dahlbomi
Mimumesa dahlbomi
Mimumesa dahlbomi
Mimumesa dahlbomi
Polistes canadensis
Polistes canadensis
Andricus curvator
Andricus curvator
Andricus curvator
Andricus curvator
Leptopilina boulardi
Leptopilina boulardi
Dryococelus australis
Dryococelus australis
Dryococelus australis
Dryococelus australis
Dryococelus australis
Dryococelus australis
Dryococelus australis
Dryococelus australis
Dryococelus australis
Neodiprion lecontei
Neodiprion lecontei
Neodiprion lecontei
Neodiprion lecontei
Sitobion avenae
Sitobion avenae
Sitobion avenae
Sitobion avenae
Aphis glycines
Aphis glycines
Aulacorthum solani
Metopolophium dirhodum
Metopolophium dirhodum
Aulacorthum solani
Nilaparvata lugens
Nilaparvata lugens
0.1

Figure 12: *Variation in mean genome size and species counts for 27 insect orders. As genome size has not been determined for any Raphidioptera, an estimate was obtained by averaging values for Megaloptera and Neuroptera, the other two lineages in the superorder Neuropterida (Engel et al. 2018). Brown = no metamorphosis. Red = incomplete metamorphosis. Blue = complete metamorphosis.*