Do pseudogenes pose a problem for metabarcoding marine animal communities?

Jessica Schultz[1] and Paul Hebert[1]

[1]University of Guelph

September 27, 2021

Abstract

Because DNA metabarcoding typically employs sequence diversity among mitochondrial amplicons to estimate species composition, nuclear mitochondrial pseudogenes (NUMTs) can inflate diversity. This study quantifies the incidence and attributes of NUMTs derived from the 658 bp barcode region of cytochrome c oxidase I (COI) in 156 marine animal genomes. The number of NUMTs meeting four length criteria (>150 bp, >300 bp, >450 bp, >600 bp) was determined, and they were examined to ascertain if they could be recognized by their possession of indels or stop codons. In total, 389 NUMTs <100 bp were detected, with an average of 2.49 per species (range = 0–50) and a mean length of 336 bp +/- 208 bp. Among NUMTs lacking diagnostic features, 52.5% were [?]300 bp, 63.9% were [?]450 bp, and 76.2% were [?]600 bp. Studies examing 150 bp amplicons inflate the OTU count by 1.57x compared to the true species count and increase perceived intraspecific variation at COI by 1.19x (when sequence variants with >2% sequence divergence are recognized as different OTUs). There was a weak positive correlation between genome size and NUMT count but no variation among phyla, trophic groups or life history traits. While bioinformatic advances will improve NUMT detection, the best defense involves targeting long amplicons and developing reference databases that include both mitochondrial sequences and their NUMT derivatives.

Introduction

The interplay between the nuclear and mitochondrial genomes is a defining feature of eukaryotes. One aspect of these interactions is the transfer of segments of mitochondrial DNA (mtDNA) through the nuclear membrane and their subsequent incorporation into the nuclear genome during the repair of double-stranded breaks (Blanchard & Schmidt, 1996; du Buy & Riley, 1967; Lopez, Yuhki, Masuda, Modi, & O'Brien, 1994). The resultant pseudogenes, termed NUMTs, are prevalent throughout the animal kingdom (Bensasson, Zhang, Hartl, & Hewitt, 2001; Hazkani-Covo, Zeller, & Martin, 2010; Richly & Leister, 2004), but the factors responsible for their considerable variation in abundance among species is uncertain (Bensasson et al., 2001; Hazkani-Covo et al., 2010; Richly & Leister, 2004; Song, Buhay, Whiting, & Crandall, 2008). Some studies have reported a correlation between genome size and NUMT counts (Bensasson et al., 2001; Hazkani-Covo et al., 2010), but others have not (Gerstein & Zheng, 2006; Richly & Leister, 2004; Wang, Liu, Miao, Huang, & Xiao, 2020). NUMT frequency has also been linked to environmental factors (Gerstein & Zheng, 2006; Ricchetti, Tekaia, & Dujon, 2004; Song, Jiang, Yuan, Guo, & Miao, 2013), cellular stress (Bensasson et al., 2001; Hazkani-Covo et al., 2010), and population dynamics (Antunes & Ramos, 2005; Deceliere, Charles, & Biémont, 2005)

The evolutionary implications of NUMTs have attracted considerable attention (Balakirev & Ayala, 2003). Because rates of nucleotide substitution are typically much slower in nuclear than mitochondrial genomes (Bensasson et al., 2001; Lopez, Culver, Stephens, Johnson, & O'Brien, 1997), each NUMT is a relict of its mitochondrial ancestor. As a result, they provide useful genetic markers (Harrison et al., 2002; Ricchetti et

al., 2004) which can help to resolve phylogenies (Ko et al., 2015), identify cases of hybridization (Machida & Lin, 2017), and clarify genome dynamics (Matzen da Silva et al., 2011; Zhang & Hewitt, 1996). Although their evolution is slowed, NUMT sequences are not frozen, and, because they are not transcribed (Bensasson et al., 2001; Boore, 1999), there is no selection for a functional gene product. As a result, NUMTs accumulate indels leading to frameshifts, premature stop codons, and sequence changes coding for amino acid substitutions that would compromise functionality of the gene product (Bensasson et al., 2001; Morgan et al., 2013; Zhang & Hewitt, 1996). NUMTs with such changes are readily diagnosed (Buhay, 2009), but those lacking these features (Bensasson et al., 2001) create complexities if they are mistaken for mtDNA (Hazkani-Covo et al., 2010; Zhang & Hewitt, 1996).

PCR-based approaches run a particular risk of encountering this interpretational complexity because all gene regions with sequence congruence to a particular primer set are recovered whether they derive from the target mitochondrial gene or its NUMTs. Past studies have shown that undiagnosed NUMTs can lead to erroneous phylogenetic trees, distorting evolutionary relationships (Calabrese et al., 2017; Haran, Koutroumpa, Magnoux, Roques, & Roux, 2015). In addition, NUMTs can create complexities for DNA barcoding (Hebert, Cywinska, Ball, & DeWaard, 2003) as it employs sequence diversity in mtCOI as a basis for specimen identification and species discovery in the animal kingdom. In this case, unrecognized NUMTs can inflate both apparent species or haplotype richness and measures of intraspecific variation at COI (Buhay, 2009; Creedy et al., 2019; Song et al., 2008). The extent of these risks depends upon the analytical protocol. When PCR targets DNA extracts from single specimens, most amplicons derive from mtCOI because its copy number is far higher than those of any NUMT(s) (Bogenhagen, 2012; Quiros, Goyal, Jha, & Auwerx, 2017). Because they represent a small fraction of the amplicon pool in this situation, NUMTs rarely impede recovery of the mt COI sequence with Sanger analysis (Hebert, Penton, Burns, Janzen, & Hallwachs, 2004). However, because HTS platforms characterize single molecules, they do recover NUMTs, even when they comprise a small proportion of the amplicon pool. When analysis targets DNA extracts from single individuals, NUMT recovery can be minimized by excluding sequences with low copy numbers. However, NUMTs pose a greater threat for metabarcoding studies because this protocol involves the amplification of DNA extracts from bulk collections (Andújar, Arribas, Yu, Vogler, & Emerson, 2018; Andújar et al., 2020; Elbrecht, Vamos, Steinke, & Leese, 2018; Liu, Clarke, Baker, Jordan, & Burridge, 2019). In this case, NUMTs from species representing a large proportion of the sample biomass can be abundant, creating the prospect of interpretational errors when unrecognized (Buhay, 2009; Kunz, Tay, Elfekih, Gordon, & De Barro, 2019).

Because of this impact of species biomass, sequence count is a weak criterion for excluding COI NUMTs recovered through metabarcoding (Andújar et al., 2018, 2020; Elbrecht et al., 2018; Liu et al., 2019). Stringent quality filters are also of no value because NUMT and mtCOI templates should be equally susceptible to PCR and sequencing errors (Andújar et al., 2020; Elbrecht et al., 2018). Divergence values also cannot be used as a diagnostic feature (Baeza & Fuentes, 2013; Matzen da Silva et al., 2011) because NUMTs and mtDNA show highly variable divergences (Andújar et al., 2020). Two other diagnostic approaches are more useful: 1) many NUMTs are a truncated version of the COI gene (Richly & Leister, 2004), and 2) they often possess diagnostic sequence changes (i.e., indels that lead to frameshifts or premature stop codons). Recognition of these diagnostic features is optimized when the full 658 bp barcode region is recovered, but most HTS platforms generate considerably shorter reads (Reuter, Spacek, & Snyder, 2016; Rhoads & Au, 2015). NUMTs create interpretational complexity when they meet two criteria:1) their length exceeds that of the target amplicon, and 2) they lack frameshift indels or premature stop codons. In such cases, NUMTs with divergences exceeding the threshold (e.g., 2%) used to define operational taxonomic units (OTUs; a species proxy) or recognize species (Alberdi, Aizpurua, Gilbert, & Bohmann, 2018; Ratnasingham & Hebert, 2013) will lead to overestimation of the OTU or species count while NUMTs with lower divergences will inflate estimates of intraspecific COI variation.

NUMTs are likely to pose a particular challenge for metabarcoding studies in marine environments because their resident species are both phylogenetically diverse and poorly represented in DNA barcode reference libraries (Leray & Knowlton, 2016; Radulovici, Archambault, & Dufresne, 2010). For example, a single marine habitat can include representatives from two thirds of all animal phyla (Zeppilli et al., 2015), and

most of these species will be undescribed (Appeltans et al., 2012). As a result, marine metabarcoding studies often encounter many sequences that cannot be connected to a species represented in the barcode reference library. Do these sequences represent newly encountered species or NUMTs from known taxa? The risk of encountering NUMTs is certain as they occur in diverse marine lineages including cnidarians (Song et al., 2013), crustaceans (Bensasson et al., 2001; Nguyen, Murphy, & Austin, 2002; Williams & Knowlton, 2001), echinoderms (Jacobs et al., 1983), fishes (Antunes & Ramos, 2005; Morgan et al., 2013), nematodes (Derycke, Vanaverbeke, Rigaux, Backeljau, & Moens, 2010), and tunicates (Ahmed & Ali, 2016). Despite this fact, there has not been a comprehensive effort to ascertain the incidence and attributes of NUMTs in marine taxa, information that would be useful in evaluating the interpretational complexity introduced by them and how best to defend against such impacts.

The present study begins to provide this information by examining the prevalence of COI NUMTs in marine animals and considers their implications for both estimates of richness and measures of intraspecific variation at COI. Analysis began with the quantification of COI NUMTs derived from any segment of the full (circa 1500 bp) COI gene in 85 marine species to ascertain if certain regions are more prone to incorporation in the nuclear genome. Analysis then focused on determining the incidence and attributes of NUMTs derived from the 658 bp barcode region of COI in the genomes of 156 marine species. Particular effort was directed toward ascertaining the impact of target sequence length on the NUMT count, and on the incidence of those with diagnostic traits (indels, stop codons). Because read lengths vary with analytical protocol and HTS platform, we considered four commonly recovered amplicon lengths (150 bp, 300 bp, 450 bp, 600 bp) (Hebert et al., 2018; Reuter et al., 2016; Rhoads & Au, 2015). Finally, we examined the correlation between the incidence of NUMTs and genome size, and investigated if NUMT counts varied among phyla, trophic groups, or life history traits.

Materials and Methods

Data collection

We examined the incidence of COI NUMTs in the genomes of marine animals on the NCBI genome browser (Clark et al. 2016). To identify candidate genomes, we compared taxonomic names in the World Register of Marine Species (WoRMS; Horton et al. 2020) with the NCBI genome browser (https://www.ncbi.nlm.nih.gov/genome/browse). All genomes for marine invertebrates were downloaded together with those for at least one species per order of marine vertebrates. When more than one genome was available for a species, the reference genome (if available) or the most recent assembly was selected. In addition, we downloaded the COI sequence from the mitochondrial genome of each species and used Aliview (Larson 2014) to extract the 658 bp recovered by primers targeting the barcode region (Hebert et al. 2003). When available, the reference sequence for the full COI gene was also retained. When a COI sequence was unavailable on GenBank, the Barcode of Life Database (BOLD; Ratnasingham and Hebert 2007) was searched for a sequence.

NUMT search and identification

We conducted BLAST searches for mitochondrial COI against the genome sequence available for each species using the 658 bp barcode region as the query. Using Geneious Prime (version 2020.2.1), we conducted a BLASTn search with a maximum of 1000 hits and a maximum expectation value of e = 0.0001 to generate a list of hits. We excluded BLASTn hits <100 bp in length, or those with both 100% coverage and [?] 99.8% ID as these likely represented a mitochondrial sequence inadvertently included in the nuclear assembly.

The remaining hits were considered putative COI NUMTs and summary information (hit length, GC content, query coverage, percent similarity, e-value) were exported to Excel (Supp. Table 1). Using MUSCLE in Geneious Prime, each hit was aligned with the mitochondrial COI sequence for that species to visually search

for any insertions and deletions. Each sequence was then translated using the appropriate mitochondrial code to determine if premature stop codons were present. The presence of indels or premature stop codons (IPSCs) at any position along the sequence was recorded for six hit length categories: 100–200 bp, 200–300 bp, 300–400 bp, 400–500 bp, 500–600 bp and 600–700 bp.

Since using a longer COI query length could reveal additional NUMTs beyond the 658 bp barcode region, we conducted a second BLAST search among the invertebrates in our dataset using the full-length ([?]1500 bp) COI sequence when available. We used the same BLASTn parameters and strategy for identifying IPSCs as for the 658 bp query, and retained all hits >150 bp. This analysis made it possible to ascertain if certain regions of COI were more prone to incorporation into NUMTs by mapping hits to the reference COI sequence and then quantifying the coverage at each nucleotide position. We then plotted the coverage for all species in a particular phylum to determine the frequency with which each nucleotide position of COI appeared in a NUMT.

NUMT diagnosis

To quantify the incidence of NUMTs that presented an interpretational threat to metabarcoding, we ascertained the proportion of hits lacking diagnostic features at four sequence lengths (150, 300, 450, 600 bp) commonly recovered with HTS (Hebert et al., 2018; Reuter et al., 2016; Rhoads & Au, 2015) using the 658 bp COI query. HTS platforms generating short reads will recover all NUMTs, but only those with IPSCs in the target region will be diagnosed. For instance, platforms that generate 150 bp reads will capture all NUMTs [?]150 bp, but they will only be recognizable as NUMTs if they possess IPSCs within the first 150 bp. Accordingly, we considered hits diagnosable if they contained IPSCs in the sequence region recovered by the platform (i.e., 150, 300, 450, 600 bp). The mean number of both diagnosable and non-diagnosable hits per species was compared among the four sequence length categories using a Kruskal-Wallis rank sum test.

In addition, we examined the impact of NUMT divergence values on metabarcoding results, employing a standard sequence divergence threshold (2%) for delineating OTUs. Divergence values are only important for non-diagnosable NUMTs that are not excluded from downstream analysis. We therefore ascertained the proportion of undiagnosable NUMTs that would either inflate the OTU count (>2% divergence) or intraspecific barcode variation (<2% divergence).

Patterns of NUMT abundance among species

We examined the relationship between the number of hits ([?]100 bp) and both genome size and the quality of the assembly (contig N50) reported on NCBI using Spearman's rank correlations. In addition, we examined if species in certain phyla or those with differing ecological or life history traits were more likely to possess NUMTs. For these comparisons, we used results from the COI barcode query length (658 bp) and included all hits [?]100 bp. We compared the average number of hits per species among phyla and trophic groups using a Kruskal-Wallis rank sum tests, and between the presence and absence of other life history traits using Wilcoxon rank sum tests..

Ecological and life history information were primarily compiled from the Encyclopedia of Life (http://eol.org; accessed 12 Sept 2020), but additional information was obtained from the primary literature to fill gaps (see Supp. Data Table 2 for specific references). The traits examined included trophic category, mode of reproduction (asexual/sexual), hermaphroditism and colonialism. We recognized six trophic categories (predator/carnivore, grazer/herbivore, parasite, suspension feeder, omnivore, other) on the basis of adult feeding habits. The 'suspension feeder' category included passive suspension feeders, active filter feeders, and mucous net feeders. The 'omnivore' category included species with more than one equally prevalent trophic level or guild. 'Other' included a chemosymbiotroph, a surface deposit feeder, and two detritivores.

Results

Marine animal genomes

The 348 marine species on the NCBI Genome Browser included 124 invertebrates and 224 vertebrates (157 fishes, 33 mammals, 26 birds, 8 reptiles). We downloaded genomes for 188 of these species including all 124 invertebrates and 64 vertebrates. Thirty-two of these genomes (19 invertebrates, 13 vertebrates) were subsequently excluded from study, most because no COI sequence was available on GenBank or BOLD, but one species (*Hofstenia miamia*) was excluded because no FASTA file was available. Among the remaining 156 species, 85 invertebrates had a full ([?]1500 bp) mitochondrial COI sequence available. At least one species from each available vertebrate order was represented. Only partial (487–657 bp) COI sequences were available for 13 species and these were included in the 658 bp category.

Hit length and distribution

The 658 bp COI query sequence revealed 389 putative NUMTs [?]100 bp in the 156 genomes with 72 (46.2%) of the species possessing at least one (Supp. Table 1). The NUMT count averaged 2.49 +- 7.06 (SD) per genome, and ranged from 0–50. Hit lengths varied from 100–729 bp and averaged 336 bp +- 208 bp (Figures 1). Most hits (37.3%) were short (100–200 bp), but almost a quarter (24.4%) were 600–700 bp. Among the 389 hits, 282 (72.5%) contained IPSCs while 107 lacked them (Figures 1).

Forty of the 85 invertebrate genomes with a full-length COI sequence contained one or more NUMTs [?]150 bp. In total, 449 NUMTs were revealed with the full-length COI query (Supp. Table 1) with their lengths averaging 409 bp +- 284 bp (mean +- SD; Figure 2), but many (58.6%) were less than 300 bp (Figure 2). Most of these NUMTs (358, 79.7%) contained IPSCs, but 91 did not (Figure 2). Hits were not evenly distributed along COI as nucleotide positions showed more than 2-fold variation in the incidence of their inclusion in NUMTs (57–126 coverage for a particular nucleotide position) (Figure 3).

NUMT diagnosis

Longer read lengths reduced the number of NUMTs that were recovered (Figure 4; Kruskal-Wallis: $X^2 = 13.05$, df = 3, p = 0.005, n = 156) and the number without an IPSC (Figure 4; Kruskal-Wallis: $X^2 = 19.23$, df = 3, $p < 0.001$, n = 156). Removing these diagnosable hits significantly reduced the hit count for three length categories (Figure 4; Wilcoxon rank sum tests: *300 bp* : W = 10,405, p = 0.01; *450 bp* : W = 10,706, p = 0.02; *600 bp* : W = 10,550, p = 0.005, n = 156), but not for*150 bp* (Figure 4; Wilcoxon rank sum test: W = 10,787, p = 0.047, n = 156). Among those NUMTs lacking an IPSC, 52.5% were excluded with a read length of 300 bp, 63.9% with read length of 450 bp, and 76.2% with a read length of 600 bp (Table 1).

NUMTs with IPSCs possessed an average sequence divergence of 21.9% ± 8.8% from the mtCOI sequence in their parent species with divergences ranging from 0.3–36.0% (Figures 5 & 6). By comparison, NUMTs lacking IPSCs possessed an average divergence of 10.8% ± 9.3% (range = 0–30.5%) (Figures 5 & 6). Among the [?]150 bp hits which lacked an IPSC, 73.0% (89/122) had divergence values >2% so they could inflate the OTU count while another 30 with divergence values <2% could inflate the amount of barcode variation within their source species. The other three hits showed 0% divergence from mtCOI so would have no impact. Accordingly, studies targeting short amplions could increase OTU counts by 1.57x and intraspecific barcode variation by 1.19x.

Patterns of NUMT abundance among species

Genome sizes varied more than than 2000-fold from 3.03 Mb in the demosponge *Aplysina aerophoba* to 6,700 Mb in the ridgetail prawn*Palaemon carinicauda* (Figure 7; Supp. Table 2). There was a weak positive correlation between the hit count and genome size (Figure 7; Spearman's rank correlation: $\rho = 0.33$, p

<0.0001, n = 156). Contig N50s ranged 117,000 fold. There was a weak negative correlation between hit frequency and contig N50 across its 117,000 fold range (198 for *Ophionereis fasciata* to 23 x10^6 for*Chanos chanos* ((Supp. Table 2; Figure 7; Spearman's rank correlation: ρ = -0.19, p = 0.02, n = 156).

Arthropods and molluscs had the most COI hits [?]100 bp (Figure 8A), but mean counts did not differ significantly among phyla (Figure 8A; Kruskal-Wallis: X^2= 26.05, df = 16, p = 0.053, n = 156). When hits with IPSCs were removed, there was also no difference in mean hits among phyla (Figure 8B; Kruskal-Wallis: X^2 = 19.37, df = 16, p = 0.25, n = 156). Most phyla were represented by four or fewer representatives (Figure 8) so the power of this test was very limited. Among phyla with more representatives, echinoderms and cnidarians contained the highest percent of hits [?]100 bp without IPSCs at 66.7% and 33.3%, respectively (Supp. Table 2).

The number of hits did not differ among taxa in different trophic categories (Figure 9; Kruskal-Wallis: X^2= 9.03, df = 5, p = 0.11, n = 156), but the parasitic salmon louse,*Lepeophtheirus salmonis* , had the highest NUMT count (Table 2). NUMT incidence was also unrelated to any life history characteristic examined including asexual reproduction (Figure 10A; Wilcoxon rank: W = 1885.5, p = 0.052, n = 156), sexual reproduction (Figure 10B; Wilcoxon rank: W = 233.5, p = 0.96, n = 156), hermaphroditism (Figure 10C; Wilcoxon rank: W = 2448, p = 0.35, n = 156) or colonialism (Figure 10D; Wilcoxon rank: W = 786, p = 0.66, n = 156).

Discussion

This study evaluated the incidence and attributes of NUMTs in the genomes of 156 species of marine animals and considered the interpretational complexities introduced for biodiversity assessments using metabarcoding. Considering the 85 invertebrate species with a full-length COI sequence, most (58.4%) NUMTs were < 300 bp. This was also true for NUMTs derived from the 658 bp barcode region, where 51.2% were <300 bp. When hits with IPSCs were excluded, 72.5% of NUMTs were removed. On the other hand, when reads <300 bp were retained, NUMTs often lacked these diagnostic features. When reads were clustered into OTUs based on a 2% divergence threshold, these undiagnosable NUMTs inflated apparent species richness by up to 1.57x relative to the true number of species and barcode variation in these species by up to 1.19x. No significant difference in NUMT incidence was detected among species in different phyla, trophic categories or breeding systems, but the strength of these tests was limited by the low number of genomes available for analysis. Based on current information, NUMTs pose a substantial challenge for metabarcoding in marine environments, particularly when short amplicons are targeted as is often the case in dietary studies (Berry et al., 2017; da Silva et al., 2019).

Hit length and distribution

NUMTs did not appear to be evenly distributed along the COI gene, but it seems unlikely that certain regions are more prone to incorporation. NUMTs arise from two processes: novel insertions into the nuclear genome and post-insertion duplication and translocation (Bensasson et al., 2001; Hazkani-Covo et al., 2010; Richly & Leister, 2004). The uneven distribution we observed may be driven by NUMT duplications following translocation from the mitochondrial genome rather than by hotspots prone to NUMT integration (Calabrese et al., 2017). Confirming which mechanism is responsible for a particular NUMT requires detailed examination of mutation patterns and NUMT attributes in allied species (Behura, 2007; Bensasson et al., 2001).

The apparent biomodality of sequence lengths for NUMTs identified by both the 658 bp and 1500 bp queries is likely artefactual. While short NUMTs certainly predominate, the secondary peak of long NUMTs reflects the fact that any NUMT extending beyond the query length is included in the longest size category. If NUMTs were evaluated for entire mitochondrial genomes rather just the barcode region, the bimodal pattern would undoubtedly disappear.

NUMT diagnosis

Our results indicate that the impacts of NUMTs can be greatly reduced by targeted longer amplicons because short NUMTs are excluded and more IPSCs are exposed. As a result, the exclusion of sequences with IPSCs is a key bioinformatics step (Buhay, 2009; Kunz et al., 2019). In our analysis, screening for IPSCs failed to significantly reduce the NUMT count for the shortest read length (150 bp), but many NUMTs were found to possess an IPSCwith longer reads. Porter and Hajibabai (2021) further showed that short (˜300 bp) NUMTs are also less likely to contain other diagnostic features such as inappropriate amino acid substitutions that can identified via bioinformatic pipelines. Accordingly, NUMTs pose the highest risk to eDNA studies where amplicons range from 50-400 bp (Langlois, Allison, Bergman, To, & Helbing, 2021), dietary analyses where amplicons are 70-230 bp (Berry et al., 2017; da Silva et al., 2019) and marine metabarcoding studies which commonly target a 313 bp region of COI (Leray et al., 2013), while studies targeting the full 658 bp barcode region face a lower risk of NUMT misinterpretation. However, the analysis of longer amplicons does not eliminate the problem as 29.9% of >600 bp NUMTs lacked an IPSC, reinforcing a pattern seen in other taxa (Antunes & Ramos, 2005; Kunz et al., 2019; Richly & Leister, 2004). Furthermore, NUMTs of any length can be problematic, even when their incidence in the template pool is low, if primers have higher affinity for them than for the mitochondrial target (Kim, Lee, & Ju, 2013; Kunz et al., 2019).

Unrecognized NUMTs misclassified as distinct taxa (species, OTUs, ESVs, haplotypes) inflate diversity estimates generated by metabarcoding. If all NUMTs with >2% divergence in our dataset were mistaken for a distinct OTU, alpha diversity would have been inflated by1.57x. Conversely, if all NUMTs with <2% divergence were assumed to reflect haplotype diversity within a taxon, barcode variation would have been inflated by 1.19x. If intraspecific variation is not under investigation, grouping NUMTs within an OTU will lessen overestimation of taxon richness by eliminating those NUMTs with the lowest divergence which are also least likely to posess an IPSC. However, the latter approach may not outweigh the benefits of using ESVs such as improved resolution and their value as intrinsic units of biological diversity (Callahan, McMurdie, & Holmes, 2017). Whether OTUs or ESVs are used, sequence arrays containing NUMTs can still indicate beta diversity if NUMTs are either uncommon or consistently recovered (Porter & Hajibabaei, 2021).

Patterns of NUMT abundance among species

The incidence of NUMTs is influenced by diverse factors. Some are associated with genome size (e.g., frequency of double-stranded breaks, duplication via repetitive elements), but others are not (e.g., mitochondrial damage, cellular stress) so the linkage between the incidence of NUMTs and genome size is unpredictable (Antunes & Ramos, 2005; Hazkani-Covo et al., 2010; Ricchetti et al., 2004; Richly & Leister, 2004; Gerstein and Zheng, 2006;Wang et al., 2020). We detected a significant positive correlation between genome size and NUMT count, but the low coefficient of determination ($\rho = 0.33$) helps to explain the variable associations noted in earlier studies. As NUMT frequency varies by marker (Baeza & Fuentes, 2013), a search for their incidence across the entire mitochondrial genome would better evaluate the linkage between the prevalence of NUMTs and genome size.

The quality of genome assemblies, as indicated by contig N50, varied 117,000-fold among the 156 species examined in our study. High quality assemblies possessed significantly fewer NUMTs than lower quality assemblies, a result contrary to previous findings (Hazkani-Covo et al., 2010; Richly & Leister, 2004). Typically, NUMT counts are elevated in high quality assemblies because of their inadvertent exclusion from draft assemblies. For example, Hazkani-Covo et al. (2010) reported that NUMT counts doubled between low and high quality assemblies of 18 eukaryotes, including a 20-fold increase in *Drosophila melanogaster* and a 3-fold rise in *Takifugu rubripes* . Since NUMT detection is sensitive to search strategy, genome curation, and level of completion, estimates of NUMT prevalence will need to be updated as the quality of genome assemblies improves. Because the quality of current genome assemblies for most marine animals are low (mean contig N50: 340,000; median: 16,000), our NUMT counts may be substantial underestimates.

Other potential sources of errors in NUMT enumeration include the recovery of COI from bacterial symbionts

or other contaminants, heteroplasmy, or PCR/sequencing errors (Baeza & Fuentes, 2013; Song et al., 2008). For example, a recent genome assembly for a nematode inadvertently included many sequences of bacterial origin (Schiffer et al., 2019; Thorne, Kagoshima, Clark, Marshall, & Wharton, 2014). Our study revealed many NUMTs in several molluscs, including bivalves and cephalopods which employ doubly uniparental inheritance of mitochondrial DNA (Doucet-Beaupré et al., 2010; Shizas, 2012; Zouros, Ball, Saavedra, & Freeman, 1994) and other mechanisms that add similar diversity (Strugnell & Lindgren, 2007). Because of such factors, causes aside from NUMTs can contribute to perceived COI diversity (Shizas, 2012).

While our analyses did not reveal significant differences in NUMT counts among phyla, the highest numbers were found in arthropods, reinforcing evidence that their genomes often possess many NUMTs (Song et al., 2008). This fact is well-documented for decapods (Baeza & Fuentes, 2013; Gíslason, Svavarsson, Halldórsson, & Pálsson, 2013; Kim et al., 2013; Williams & Knowlton, 2001; Yuan et al., 2017), a group with high NUMT counts in our study. The capacity to rigorously test for variation in the incidence of NUMTs among phyla was constrained because most (12/17) had only a few genomes available for analysis.

Although no association was evident between NUMT counts and trophic category or breeding system, a parasitic copepod, *Lepeophtheirus salmonis* , had the highest NUMT count (50). In bacteria, pseudogenes are more common among parasites (Lawrence & Hendrix, 2001), and some parasitoid wasps contain many NUMTs, possibly linked to a high incidence of structural rerarrangements in their mitochondria (Yan et al., 2019). However, many marine organisms have complex life history patterns that include multiple trophic levels and reproductive strategies, complicating the detection of linkages with genome dynamics. Patterns may be apparent once more genomes have been sequenced.

Mitigating the impacts of NUMTs

Because NUMTs are often difficult to recognize, it is worth considering approaches that would minimize their impacts on metabarcoding. Methodological shifts offer one solution. For example, density gradient centrifugation can provide highly purified mitochondrial DNA for PCR (Cristescu, 2019; Deiner et al., 2017; Tsuri et al., 2021), but it requires freshly collected samples (Zhou et al., 2013). Reverse transcription PCR (RT-PCR) offers another powerful option as sequences are only recovered from mitochondrial DNA because NUMTs are not transcribed, and this method can be employed on properly preserved samples. (Bensasson et al., 2001; Porter & Hajibabaei, 2021; Song et al., 2008; D. Wang et al., 2019).

Aside from adjusting wet lab protocols, NUMTs can be excluded by more rigorous bioinfomatic screening (Porter & Hajibabaei, 2021). Tree-based methods that identify NUMTs based on their divergence from the mitochondrial copy can aid studies on a particular taxonomic group (Creedy et al., 2019), but are ineffective for work on communities because taxonomic diversity is high. Read count thresholds can remove 90-95% of NUMTs while retaining most legitimate mtDNA variants (Andújar et al., 2020). Pipelines that incorporate amino acid analysis into their filtering parameters can further improve the recognition of NUMTs (Kunz et al., 2019; Nugent et al., 2020; Porter & Hajibabaei, 2020; Porter & Hajibabaei, 2021). For example, Coil (Nugent et al., 2020) and METAWORKS (Porter & Hajibabaei, 2020) employ profile hidden Markov model (PHMM) analysis to compare the features of an unknown sequence with a translated profile built from a multiple sequence alignment for each taxonomic group to highlight frameshift errors. Because the COI protein has highly conserved amino acids at certain positions, sequences that lead to an amino acid substitution at these sites are likely to be NUMTs (Buhay, 2009; Hebert et al., 2003; Kunz et al., 2019).

Despite the potential of these defensive measures to minimize the impact of NUMTs, comprehensive, well-curated DNA barcode reference libraries are needed to support metabarcoding studies in marine environments (Bucklin, Steinke, & Blanco-Bercial, 2011; Leray & Knowlton, 2016). At present, NUMTs are typically discarded, but their retention in the reference library would aid data interpretation (Matzen da Silva et al., 2011), especially those derived from common, large-bodied species. There are two paths to develop a well-parameterized reference library that includes both COI sequences and their NUMT derivatives. The first involves the assembly of high quality nuclear and mitochondrial genome sequences for all marine species.

While this is an important goal (Lewin et al. 2018), its completion for all marine species will be a multi-decadal task. A much faster path involves the collection of fresh specimens and their analysis via RT-PCR to obtain mtCOI sequences, followed by PCR to amplify both mtCOI and NUMTs from the same specimen. This approach has the potential to characterize thousands of species in a single HTS run. Although the analysis of multiple markers can help to resolve ambiguities when NUMTs complicate results (Antunes & Ramos, 2005; Kunz et al., 2019; Richly & Leister, 2004; Song et al., 2008; van der Loos & Nijland, 2020), the challenge in developing a well-parameterized reference library for COI reinforces the need to focus on this task rather than on broadening analysis to additional genes.

DNA metabarcoding has gained rapid adoption due to its capacity to document biodiversity patterns in unprecedented detail at ever-decreasing costs (Taberlet, Coissac, Pompanon, Brochmann, & Willerslev, 2012; van der Loos & Nijland, 2020). At a time when marine biodiversity is in rapid decline (Dulvy, Sadovy, & Reynolds, 2003; Halpern, Selkoe, Micheli, & Kappel, 2007; WWF, 2020), it is critical that the metabarcoding results used to inform conservation and management decisions be reliable. Hence, it is essential to develop the informatics tools and analytical approaches to minimize the impacts of NUMTs on biodiversity assessments. Since marine organisms exhibit tremendous variation in the structure and organization of their genomes (Burger, Jackson, & Waller, 2012; Lavrov & Pett, 2016), further exploration and acknowledgment of NUMT prevalence is not only necessary for reliable metabarcoding data, but will likely inform evolutionary insights regarding genome dynamics across the tree of life (Zhang & Hewitt, 1996).

Acknowlegements

We carried out this work at the University of Guelph which resides on the treaty lands and territory of the Mississaugas of the Credit. We recognize that this gathering place where we work and learn is home to many past, present, and future Indigenous Peoples. NSERC supported this study through an Alexander Graham Bell Canada Graduate Scholarship to JAS and by a Discovery Grant to PDNH.

References

Ahmed, S. N., & Ali, A. J. H. (2016). Numts: An impediment to DNA barcoding of Polyclinids, Tunicata. *Mitochondrial DNA, 27* (5), 3395–3398. https://doi.org/10.3109/19401736.2015.1018238

Alberdi, A., Aizpurua, O., Gilbert, M. T. P., & Bohmann, K. (2018). Scrutinizing key steps for reliable metabarcoding of environmental samples. *Methods in Ecology and Evolution, 9* (1), 134–147. https://doi.org/10.1111/2041-210X.12849

Andújar, C., Arribas, P., Yu, D. W., Vogler, A. P., & Emerson, B. C. (2018). Why the COI barcode should be the community DNA metabarcode for the Metazoa. *Molecular Ecology, 27* (20), 3968–3975. https://doi.org/10.1111/mec.14844

Andújar, C., Creedy, T. J., Arribas, P., López, H., Salces-Castellano, A., Pérez-Delgado, A., ... Emerson, B. C. (2020). Validated removal of nuclear pseudogenes and sequencing artefacts from mitochondrial metabarcode data. *BioRxiv Preprint* , 1–37. https://doi.org/10.1101/2020.06.17.157347

Antunes, A., & Ramos, M. J. (2005). Discovery of a large number of previously unrecognized mitochondrial pseudogenes in fish genomes.*Genomics, 86* (6), 708–717. https://doi.org/10.1016/j.ygeno.2005.08.002

Appeltans, W., Ahyong, S. T., Anderson, G., Angel, M. V., Artois, T., Bailly, N., ... Costello, M. J. (2012). The magnitude of global marine species diversity. *Current Biology, 22* (23), 2189–2202. https://doi.org/10.1016/j.cub.2012.09.036

Baeza, J. A., & Fuentes, M. S. (2013). Exploring phylogenetic informativeness and nuclear copies of mitochondrial DNA (numts) in three commonly used mitochondrial genes: Mitochondrial phylogeny of peppermint,

cleaner, and semi-terrestrial shrimps (Caridea:*Lysmata* , *Exhippolysmata* and *Mergui*).*Zoological Journal of the Linnean Society, 168* , 699–722. https://doi.org/10.1111/zoj.12044

Balakirev, E. S., & Ayala, F. J. (2003). Pseudogenes: Are they "junk" or functional DNA? *Annual Review of Genetics, 37* , 123–151. https://doi.org/10.1146/annurev.genet.37.040103.103949

Behura, S. K. (2007). Analysis of nuclear copies of mitochondrial sequences in honeybee (*Apis mellifera*) genome. *Molecular Biology and Evolution, 24* (7), 1492–1505. https://doi.org/10.1093/molbev/msm068

Bensasson, D., Zhang, D. X., Hartl, D. L., & Hewitt, G. M. (2001). Mitochondrial pseudogenes: Evolution's misplaced witnesses. *Trends in Ecology and Evolution, 16* (6), 314–321. https://doi.org/10.1016/S0169-5347(01)02151-6

Berry, T. E., Osterrieder, S. K., Murray, D. C., Coghlan, M. L., Richardson, A. J., Grealy, A. K., ... Bunce, M. (2017). DNA metabarcoding for diet analysis and biodiversity: A case study using the endangered Australian sea lion (*Neophoca cinerea*). *Ecology and Evolution, 7* (14), 5435–5453. https://doi.org/10.1002/ECE3.3123

Blanchard, J. L., & Schmidt, G. W. (1996). Mitochondrial DNA migration events in yeast and humans: Integration by a common end-joining mechanism and alternative perspectives on nucleotide substitution patterns. *Molecular Biology and Evolution, 13* (3), 537–548. https://doi.org/10.1093/oxfordjournals.molbev.a025614

Bogenhagen, D. F. (2012). Mitochondrial DNA nucleoid structure.*Biochimica et Biophysica Acta - Gene Regulatory Mechanisms, 1819* (9–10), 914–920. https://doi.org/10.1016/j.bbagrm.2011.11.005

Boore, J. L. (1999). Animal mitochondrial genomes. *Nucleic Acids Research, 27* (8), 1767–1780. https://doi.org/10.1093/nar/27.8.1767

Bucklin, A., Steinke, D., & Blanco-Bercial, L. (2011). DNA barcoding of marine Metazoa. *Annual Review of Marine Science, 3* (1), 471–508. https://doi.org/10.1146/annurev-marine-120308-080950

Buhay, J. E. (2009). "COI-like" sequences are becoming problematic in molecular systematic and DNA barcoding studies. *Journal of Crustacean Biology, 29* (1), 96–110. https://doi.org/10.1651/08-3020.1

Burger, G., Jackson, C. J., & Waller, R. F. (2012). Unusual mitochondrial genomes and genes. In C. E. Bullerwell (Ed.),*Organelle genetics: Evolution of organelle genomes and gene expression* (pp. 44–77). Heidelberg: Springer-Verlag Berlin. https://doi.org/10.1007/978-3-642-22380-8

Calabrese, F. M., Balacco, D. L., Preste, R., Diroma, M. A., Forino, R., Ventura, M., & Attimonelli, M. (2017). NumtS colonization in mammalian genomes. *Scientific Reports, 7* (1), 1–10. https://doi.org/10.1038/s41598-017-16750-2

Callahan, B. J., McMurdie, P. J., & Holmes, S. P. (2017). Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. *ISME Journal, 11* , 2639–2643. https://doi.org/10.1038/ismej.2017.119

Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., & Sayers, E. W. (2016). GenBank. *Nucleic Acids Research, 44* (D1), D67–D72. https://doi.org/10.1093/nar/gkv1276

Creedy, T. J., Norman, H., Tang, C. Q., Qing Chin, K., Andujar, C., Arribas, P., ... Vogler, A. P. (2019). 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

Cristescu, M. E. (2019). Can environmental RNA revolutionize biodiversity science? *Trends in Ecology and Evolution, 34* (8), 694–697. https://doi.org/10.1016/j.tree.2019.05.003

da Silva, L. P., Mata, V. A., Lopes, P. B., Pereira, P., Jarman, S. N., Lopes, R. J., & Beja, P. (2019). Advancing the integration of multi-marker metabarcoding data in dietary analysis of trophic generalists.

Molecular Ecology Resources, 19 (6), 1420–1432. https://doi.org/10.1111/1755-0998.13060

Deceliere, G., Charles, S., & Biémont, C. (2005). The dynamics of transposable elements in structured populations. *Genetics, 169* , 467–474. https://doi.org/10.1534/genetics.104.032243

Deiner, K., Bik, H. M., Mächler, E., Seymour, M., Lacoursière-Roussel, A., Altermatt, F., ... Bernatchez, L. (2017). Environmental DNA metabarcoding: Transforming how we survey animal and plant communities.*Molecular Ecology, 26* (21), 5872–5895. https://doi.org/10.1111/mec.14350

Derycke, S., Vanaverbeke, J., Rigaux, A., Backeljau, T., & Moens, T. (2010). Exploring the use of cytochrome oxidase c subunit 1 (COI) for DNA barcoding of free-living marine nematodes. *PLoS ONE, 5* (10), e13716. https://doi.org/10.1371/journal.pone.0013716

Doucet-Beaupré, H., Breton, S., Chapman, E. G., Blier, P. U., Bogan, A. E., Stewart, D. T., & Hoeh, W. R. (2010). Mitochondrial phylogenomics of the Bivalvia (Mollusca): Searching for the origin and mitogenomic correlates of doubly uniparental inheritance of mtDNA. *BMC Evolutionary Biology, 10* (1), 1–19. https://doi.org/10.1186/1471-2148-10-50

du Buy, H. G., & Riley, F. L. (1967). Hybridization between the nuclear and kinetoplast DNAs of *Leishmania enriettii* and between nuclear and mitochondrial DNAs of mouse liver. *Proceedings of the National Academy of Sciences of the United Stated of America, 57* (3), 790–797. https://doi.org/10.1073/pnas.57.3.790

Dulvy, N. K., Sadovy, Y., & Reynolds, J. D. (2003). Extinction vulnerability in marine populations. *Fish and Fisheries, 4* (1), 25–64. https://doi.org/10.1046/j.1467-2979.2003.00105.x

Elbrecht, V., Vamos, E. E., Steinke, D., & Leese, F. (2018). Estimating intraspecific genetic diversity from community DNA metabarcoding data.*PeerJ, 6* , e4644. https://doi.org/10.7717/peerj.4644

Encyclopedia of Life. (2020). Retrieved September 12, 2020, from http://eol.org/

Geneious Prime version 2020.2.1. (2020). https://geneious.com/

Gerstein, M., & Zheng, D. (2006). The real life of pseudogenes.*Scientific American, 295* (2), 48–55. https://doi.org/10.1038/scientificamerican0806-48

Gíslason, O. S., Svavarsson, J., Halldórsson, H. P., & Pálsson, S. (2013). Nuclear mitochondrial DNA (numt) in the atlantic rock crab*Cancer irroratus* Say, 1817 (Decapoda, Cancridae).*Crustaceana, 86* (5), 537–552. https://doi.org/10.1163/15685403-00003191

Halpern, B. S., Selkoe, K. A., Micheli, F., & Kappel, C. V. (2007). Evaluating and ranking the vulnerability of global marine ecosystems to anthropogenic threats. *Conservation Biology, 21* (5), 1301–1315. https://doi.org/10.1111/j.1523-1739.2007.00752.x

Haran, J., Koutroumpa, F., Magnoux, E., Roques, A., & Roux, G. (2015). Ghost mtDNA haplotypes generated by fortuitous numts can deeply disturb infra-specific genetic diversity and phylogeographic pattern.*Journal of Zoological Systematics and Evolutionary Research, 53* (2), 109–115. https://doi.org/10.1111/jzs.12095

Harrison, P. M., Hegyi, H., Balasubramanian, S., Luscombe, N. M., Bertone, P., Echols, N., ... Gerstein, M. (2002). Molecular fossils in the human genome: Identification and analysis of the pseudogenes in chromosomes 21 and 22. *Genome Research, 12* (2), 272–280. https://doi.org/10.1101/gr.207102

Hazkani-Covo, E., Zeller, R. M., & Martin, W. (2010). Molecular poltergeists: Mitochondrial DNA copies (numts) in sequenced nuclear genomes. *PLoS Genetics, 6* (2), e1000834. https://doi.org/10.1371/journal.pgen.1000834

Hebert, P. D. N., Braukmann, T. W. A., Prosser, S. W. J., Ratnasingham, S., DeWaard, J. R., Ivanova, N. V, ... Zakharov, E. V. (2018). A Sequel to Sanger: Amplicon sequencing that scales. *BMC Genomics, 19* (219), 1–14. https://doi.org/10.1101/191619

Hebert, P. D. N., Cywinska, A., Ball, S. L., & DeWaard, J. R. (2003). Biological identifications through DNA barcodes. *Proceedings of the Royal Society B: Biological Sciences, 270* (1512), 313–321. https://doi.org/10.1098/rspb.2002.2218

Hebert, P. D. N., Penton, E. H., Burns, J. M., Janzen, D. H., & Hallwachs, W. (2004). Ten species in one: DNA barcoding reveals cryptic species in the neotropical skipper butterfly *Astraptes fulgerator* . *Proceedings of the National Academy of Sciences of the United States of America, 101* (41), 14812–14817. https://doi.org/10.1073/pnas.0406166101

Horton, T., Kroh, A., Ahyong, S., Bailly, N., Boyko, C. B., Brandão, S. N., ... Zhao, Z. (2020). World Register of Marine Species (WoRMS). Retrieved November 12, 2019, from http://www.marinespecies.org

Jacobs, H. T., Posakony, J. W., Grula, J. W., Roberts, J. W., Xin, J.-H., Britten, R. J., & Davidson, E. H. (1983). Mitochondrial DNA sequences in the nuclear genome of *Strongylocentrotus purpuratus* .*Journal of Molecular Biology, 165* , 609–632.

Jeffery, N. (2015). Genome size diversity and evolution in the Crustacea (PhD thesis). University of Guelph, Canada. Retrieved from http://atrium.lib.uoguelph.ca.subzero.lib.uoguelph.ca/xmlui/handle/10214/9216

Kim, S. J., Lee, K. Y., & Ju, S. J. (2013). Nuclear mitochondrial pseudogenes in *Austinograea alayseae* hydrothermal vent crabs (Crustacea: Bythograeidae): Effects on DNA barcoding. *Molecular Ecology Resources, 13* (5), 781–787. https://doi.org/10.1111/1755-0998.12119

Ko, Y. J., Yang, E. C., Lee, J. H., Lee, K. W., Jeong, J. Y., Park, K., ... Yim, H. S. (2015). Characterization of cetacean Numt and its application into cetacean phylogeny. *Genes and Genomics, 37* (12), 1061–1071. https://doi.org/10.1007/s13258-015-0353-7

Kunz, D., Tay, W. T., Elfekih, S., Gordon, K. H. J., & De Barro, P. J. (2019). Take out the rubbish – Removing NUMTs and pseudogenes from the*Bemisia tabaci* cryptic species mtCOI database. *BioRxiv Preprint* , 1–19. https://doi.org/10.1101/724765

Langlois, V. S., Allison, M. J., Bergman, L. C., To, T. A., & Helbing, C. C. (2021). The need for robust qPCR-based eDNA detection assays in environmental monitoring and species inventories. *Environmental DNA, 3* (3), 519–527. https://doi.org/10.1002/edn3.164

Larsson, A. (2014). AliView: A fast and lightweight alignment viewer and editor for large datasets. *Bioinformatics, 30* (22), 3276–3278. https://doi.org/10.1093/bioinformatics/btu531

Lavrov, D. V., & Pett, W. (2016). Animal mitochondrial DNA as we do not know it: Mt-genome organization and evolution in nonbilaterian lineages.*Genome Biology and Evolution, 8* (9), 2896–2913. https://doi.org/10.1093/gbe/evw195

Lawrence, J. G., Hendrix, R., & Casjens, S. (2001) What are the pseudogenes in bacterial genomes? *Trends in Microbiology, 9* (11), 535–540. https://doi.org/10.1016/S0966-842X(01)02198-9

Leray, M., & Knowlton, N. (2016). Censusing marine eukaryotic diversity in the twenty-first century. *Philosophical Transactions of the Royal Society B: Biological Sciences, 371* (1702), 20150331. https://doi.org/10.1098/rstb.2015.0331

Leray, M., Yang, J. Y., Meyer, C. P., Mills, S. C., Agudelo, N., Ranwez, V., ... Machida, R. J. (2013). A new versatile primer set targeting a short fragment of the mitochondrial COI region for metabarcoding metazoan diversity: Application for characterizing coral reef fish gut contents. *Frontiers in Zoology, 10* (34), 1–14. https://doi.org/10.1186/1742-9994-10-34

Lewin, H. A., Robinson, G. E., Kress, W. J., Baker, W. J., Coddington, J., Crandall, K. A., ... Zhang, G. (2018). Earth BioGenome Project: Sequencing life for the future of life. *Proceedings of the National Academy of Sciences of the United States of America, 115* (17), 4325–4333. https://doi.org/10.1073/PNAS.1720115115

Liu, M., Clarke, L. J., Baker, S. C., Jordan, G. J., & Burridge, C. P. (2019). A practical guide to DNA metabarcoding for entomological ecologists. *Ecological Entomology, 45* (3), 373–385. https://doi.org/10.1111/een.12831

Lobo, J., Costa, P. M., Teixeira, M. AL, Ferreira, M. S., Costa, M. H., & Costa, F. O. (2013). Enhanced primers for amplification of DNA barcodes from a broad range of marine metazoans. *BMC Ecology, 13* (1), 34. https://doi.org/10.1186/1472-6785-13-34

Lopez, J. V., Culver, M., Stephens, J. C., Johnson, W. E., & O'Brien, S. J. (1997). Rates of nuclear and cytoplasmic mitochondrial DNA sequence divergence in mammals. *Molecular Biology and Evolution, 14* (3), 277–286. https://doi.org/10.1093/oxfordjournals.molbev.a025763

Lopez, Jose V., Yuhki, N., Masuda, R., Modi, W., & O'Brien, S. J. (1994). Numt, a recent transfer and tandem amplification of mitochondrial DNA to the nuclear genome of the domestic cat. *Journal of Molecular Evolution, 39* (2), 174–190. https://doi.org/10.1007/BF00163806

Machida, R. J., & Lin, Y. Y. (2017). Occurrence of mitochondrial CO1 pseudogenes in *Neocalanus plumchrus* (Crustacea: Copepoda): Hybridization indicated by recombined nuclear mitochondrial pseudogenes. *PLoS ONE, 12* (2), 1–11. https://doi.org/10.1371/journal.pone.0172710

Matzen da Silva, J., Creer, S., dos Santos, A., Costa, A. C., Cunha, M. R., Costa, F. O., ... Carvalho, G. R. (2011). Systematic and evolutionary insights derived from mtDNA COI barcode diversity in the Decapoda (Crustacea: Malacostraca). *PLoS ONE, 6* (5), e19449. https://doi.org/10.1371/journal.pone.0019449

Morgan, J. A. T. T., Macbeth, M., Broderick, D., Whatmore, P., Street, R., Welch, D. J., & Ovenden, J. R. (2013). Hybridisation, paternal leakage and mitochondrial DNA linearization in three anomalous fish (Scombridae). *Mitochondrion, 13* (6), 852–861. https://doi.org/10.1016/j.mito.2013.06.002

Moulton, M. J., Song, H., & Whiting, M. F. (2010). Assessing the effects of primer specificity on eliminating numt coamplification in DNA barcoding: A case study from Orthoptera (Arthropoda: Insecta). *Molecular Ecology Resources, 10* (4), 615–627. https://doi.org/10.1111/j.1755-0998.2009.02823.x

Nguyen, T. T. T., Murphy, N. P., & Austin, C. M. (2002). Amplification of multiple copies of mitochondrial Cytochrome *b* gene fragments in the Australian freshwater crayfish, *Cherax destructor* Clark (Parastacidae: Decapoda). *Animal Genetics, 33* (4), 304–308. https://doi.org/10.1046/j.1365-2052.2002.00867.x

Nugent, C. M., Elliott, T. A., Ratnasingham, S., & Adamowicz, S. J. (2020). Coil: An R package for cytochrome *c* oxidase I (COI) DNA barcode data cleaning, translation, and error evaluation. *Genome, 63* , 291–305. https://doi.org/10.1139/gen-2019-0206

Porter, T.M., & Hajibabaei, M. (2020). METAWORKS: A flexible, scalable bioinformatic pipeline for multi-marker biodiversity assessments. *BioRxiv Preprint* , 1–32. https://doi.org/10.1101/2020.07.14.202960

Porter, T. M., & Hajibabaei, M. (2021). Profile hidden Markov model sequence analysis can help remove putative pseudogenes from DNA barcoding and metabarcoding datasets. *BMC Bioinformatics, 22* (256), 1–20. https://doi.org/10.1186/s12859-021-04180-x

Quiros, P. M., Goyal, A., Jha, P., & Auwerx, J. (2017). Analysis of mtDNA/nDNA ratio in mice. *Current Protocols in Mouse Biology, 7* (1), 47–54. https://doi.org/10.1002/cpmo.21

Radulovici, A. E., Archambault, P., & Dufresne, F. (2010). DNA barcodes for marine biodiversity: Moving fast forward? *Diversity, 2* (4), 450–472. https://doi.org/10.3390/d2040450

Ratnasingham, S., & Hebert, P. D. N. (2007). BOLD: The Barcode of Life Data System (www.barcodinglife.org). *Molecular Ecology Notes, 7* (3), 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* (7), e66213. https://doi.org/10.1371/journal.pone.0066213

Reuter, J. A., Spacek, D., & Snyder, M. P. (2016). High-throughput sequencing technologies. *Molecular Cell, 58* (4), 586–597. https://doi.org/10.1016/j.molcel.2015.05.004.High-Throughput

Rhoads, A., & Au, K. F. (2015). PacBio sequencing and its applications. *Genomics, Proteomics & Bioinformatics, 13* , 278–289. https://doi.org/10.1016/j.gpb.2015.08.002

Ricchetti, M., Tekaia, F., & Dujon, B. (2004). Continued colonization of the human genome by mitochondrial DNA. *PLoS Biology, 2* (9), 1313–1324. https://doi.org/10.1371/journal.pbio.0020273

Richly, E., & Leister, D. (2004). NUMTs in sequenced eukaryotic genomes. *Molecular Biology and Evolution, 21* (6), 1081–1084. https://doi.org/10.1093/molbev/msh110

Schiffer, P. H., Danchin, E. G. J., Burnell, A. M., Creevey, C. J., Wong, S., Dix, I., ... Blaxter, M. (2019). Signatures of the evolution of parthenogenesis and cryptobiosis in the genomes of Panagrolaimid nematodes. *IScience, 21* , 587–602. https://doi.org/10.1016/j.isci.2019.10.039

Shizas, N. V. (2012). Misconceptions regarding nuclear mitochondrial pseudogenes (Numts) may obscure detection of mitochondrial evolutionary novelties. *Aquatic Biology, 17* , 91–96. https://doi.org/10.3354/ab00478

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 of the United States of America ,105* (36), 13486–13491. https://doi.org/10.1073/pnas.0803076105

Song, S., Jiang, F., Yuan, J., Guo, W., & Miao, Y. (2013). Exceptionally high cumulative percentage of NUMTs originating from linear mitochondrial DNA molecules in the *Hydra magnipapillata*genome. *BMC Genomics, 14* (447), 1–13. https://doi.org/10.1186/1471-2164-14-447

Strugnell, J. M., & Lindgren, A. R. (2007). A barcode of life database for the Cephalopoda? Considerations and concerns. *Reviews in Fish Biology and Fisheries, 17,* 337–344. https://doi.org/10.1007/s11160-007-9043-0

Taberlet, P., Coissac, E., Pompanon, F., Brochmann, C., & Willerslev, E. (2012). Towards next-generation biodiversity assessment using DNA metabarcoding. *Molecular Ecology, 21* (8), 2045–2050. https://doi.org/10.1111/j.1365-294X.2012.05470.x

Thorne, M. A. S., Kagoshima, H., Clark, M. S., Marshall, C. J., & Wharton, D. A. (2014). Molecular analysis of the cold tolerant Antarctic nematode, *Panagrolaimus davidi* . *PLoS ONE, 9* (8), 104526. https://doi.org/10.1371/journal.pone.0104526

Tsuri, K., Ikeda, S., Hirohara, T., Shimada, Y., Minamoto, T., & Yamanaka, H. (2021). Messenger RNA typing of environmental RNA (eRNA): A case study on zebrafish tank water with perspectives for the future development of eRNA analysis on aquatic vertebrates. *Environmental DNA, 3* (1), 14–21. https://doi.org/10.1002/edn3.169

van der Loos, L., & Nijland, R. (2020). Biases in bulk: DNA metabarcoding of marine communities and the methodology involved. *Authorea Preprints* , mec.15592. https://doi.org/10.22541/AU.158497077.79519807

Wang, D., Xiang, H., Ning, C., Liu, H., Liu, J. F., & Zhao, X. (2019). Mitochondrial DNA enrichment reduced NUMT contamination in porcine NGS analyses. *Briefings in Bioinformatics, 21* (4), 1368–1377. https://doi.org/10.1093/bib/bbz060

Wang, J. X., Liu, J., Miao, Y. H., Huang, D. W., & Xiao, J. H. (2020). Tracking the distribution and burst of nuclear mitochondrial DNA sequences (Numts) in fig wasp genomes. *Insects, 11* (680), 1–15. https://doi.org/10.3390/insects11100680

Williams, S. T., & Knowlton, N. (2001). Mitochondrial pseudogenes are pervasive and often insidious in the snapping shrimp genus*Alpheus* . *Molecular Biology and Evolution, 18* (8), 1484–1493.

https://doi.org/10.1093/oxfordjournals.molbev.a003934

WWF. (2020). *Living planet report 2020: Bending the curve of biodiversity loss* . (R. E. A. Almond, M. Grooten, & T. Peterson, Eds.). Gland, Switzerland: WWF.

Yan, Z., Fang, Q., Tian, Y., Wang, F., Chen, X., Werren, J. H., & Ye, G. (2019). Mitochondrial DNA and their nuclear copies in the parasitic wasp *Pteromalus puparum* : A comparative analysis in Chalcidoidea.*International Journal of Biological Macromolecules, 121* , 572–579. https://doi.org/10.1016/j.ijbiomac.2018.10.039

Yuan, J., Gao, Y., Zhang, X., Wei, J., Liu, C., Li, F., & Xiang, J. (2017). Genome sequences of marine shrimp *Exopalaemon carinicauda*Holthuis provide insights into genome size evolution of caridea.*Marine Drugs, 15* (7), 213. https://doi.org/10.3390/md15070213

Zeppilli, D., Sarrazin, J., Leduc, D., Arbizu, P. M., Fontaneto, D., Fontanier, C., ... Fernandes, D. (2015). Is the meiofauna a good indicator for climate change and anthropogenic impacts? *Marine Biodiversity, 45* (3), 505–535. https://doi.org/10.1007/s12526-015-0359-z

Zhang, D. X., & Hewitt, G. M. (1996). Nuclear integrations: Challenges for mitochondrial DNA markers. *Trends in Ecology and Evolution, 11* (6), 247–251. https://doi.org/10.1016/0169-5347(96)10031-8

Zhou, X., Li, Y., Liu, S., Yang, Q., Su, X., Zhou, L., ... Huang, Q. (2013). Ultra-deep sequencing enables high-fidelity recovery of biodiversity for bulk arthropod samples without PCR amplification.*GigaScience, 2* (1). https://doi.org/10.1186/2047-217X-2-4

Zouros, E., Ball, A. O., Saavedra, C., & Freeman, K. R. (1994). Mitochondrial DNA inheritance. *Nature, 368* (6474), 818. https://doi.org/10.1038/368818a0

Data accessibility and benefit-sharing statement

Raw data, including the sequences and associated data for all NUMT hits, are contained in Supplementary Data Table 1. Supplementary Data Table 2 provides GenBank accession numbers or Barcode Index Numbers, as applicable, for all genomes and COI query sequences employed in our analysis.

Author contributions

JAS and PDNH conceived and designed the study, wrote, and edited the manscript. JAS also conducted the research and analyzed the data. Both authors contributed critically to the drafts and gave final approval for publication.

Tables and figures

Table 1. Total number of hits, diagnosable hits, and undiagnosable hits for four target sequence lengths of COI. Hits were diagnosable if they contained indels or premature stop codons (IPSC) in the target sequence (i.e., in the first 150, 300, 450 or 600 bp).

	Target length (bp)	**Target length (bp)**
	[?]150	**[?]300**
A) Total hits	A) Total hits	A) Total hits
Count	309	161
Mean	1.98	1.03

	Target length (bp)	**Target length (bp)**
SD	5.30	2.82
B) Diagnosable hits (IPSC present in region)	B) Diagnosable hits (IPSC present in region)	B) Diagnosable hits (IPS
Count	187	103
Mean	1.20	0.66
SD	4.05	2.36
C) Undiagnosable hits (IPSC absent in region)	C) Undiagnosable hits (IPSC absent in region)	C) Undiagnosable hits (I
Count	122	58
Mean	0.78	0.37
SD	1.71	0.85
>2% Divergent	89	38
<2% Divergent	30	18

Table 2. Details on the 25 marine animals with the greatest number of COI hits [?] 100 bp. 156 genomes were searched using the 658 bp barcode region of the mt COI gene.

Phylum	Species	Genome size (Mb)	Contig N50	Trophic group*	Hermaphroditism	Se
Arthropoda	*Lepeophtheirus salmonis*	665.13	16,673	Pa	0	1
Mollusca	*Octopus vulgaris*	1,772.96	490,217	P/C	0	1
Mollusca	*Octopus bimaculoides*	2,338.19	5,532	P/C	0	1
Chordata	*Balaenoptera musculus*	1,876.09	5,963,936	SF	0	1
Arthropoda	*Eriocheir sinensis*	1,549.19	45,088	P/C	0	1
Arthropoda	*Palaemon carinicauda*	6,699.72	696	Om	0	1
Arthropoda	*Strigamia maritima*	176.21	24,745	P/C	0	1
Chordata	*Petromyzon marinus*	1,130.42	170,712	Pa	0	1
Arthropoda	*Penaeus japonicus*	1,660.27	700	P/C	0	1
Mollusca	*Cumia reticulata*	67.1	890	P/C	0	1
Arthropoda	*Penaeus monodon*	1,632.39	1,782	Om	0	1
Chordata	*Ursus maritimus*	2,301.38	46,506	P/C	0	1
Chordata	*Carcharodon carcharias*	3,915.28	46,102	P/C	0	1
Chordata	*Hippocampus comes*	493.78	39,546	P/C	0	1
Arthropoda	*Limulus polyphemus*	1,828.27	11,441	P/C	0	1
Arthropoda	*Tachypleus tridentatus*	2,167.47	1,644,156	P/C	0	1
Porifera	*Amphimedon queenslandica*	166.7	11,817	SF	1	1
Arthropoda	*Parhyale hawaiensis*	2,752.56	10,438	G/H	0	1
Arthropoda	*Portunus trituberculatus*	990.24	788	P/C	0	1
Chordata	*Trichechus manatus*	3,103.81	37,750	G/H	0	1
Echinodermata	*Lytechinus variegatus*	1061.2	9,657	Om	0	1
Mollusca	*Mytilus galloprovincialis*	1,500.15	2,627	SF	1	1
Arthropoda	*Penaeus vannamei*	1,663.58	86,864	Om	1	1
Chordata	*Salpa thompsoni*	318.75	636	SF	1	1
Chordata	*Eptatretus burgeri*	2,608.38	7,991	P/C	1	1

* Pa = parasite; P/C = predator/carnivore; SF = suspension feeder; Om = omnivore; G/H = grazer/herbivore

+ Number of hits corrected for genome size (per 100 Mb); ++ Combined length of all COI NUMTs

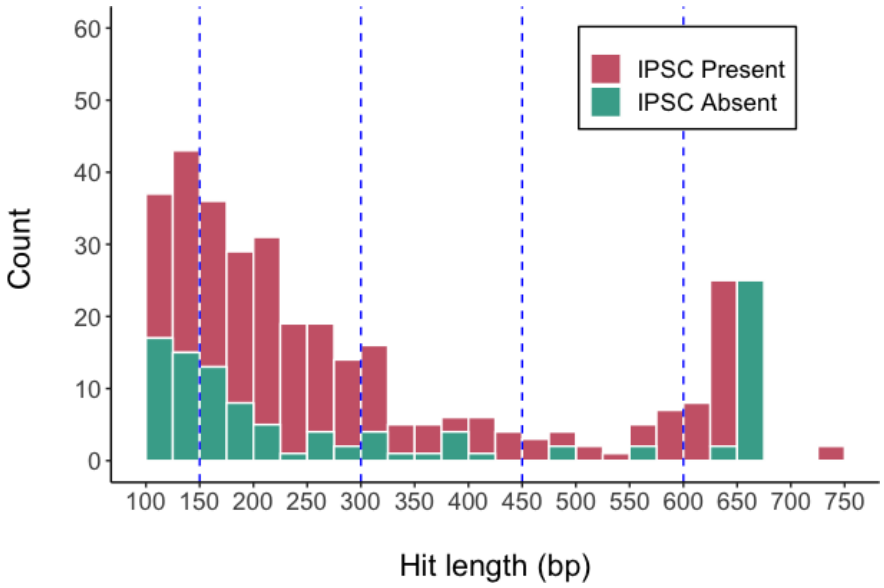

Figure 1 . The length distribution of COI NUMTs derived from the 658 bp barcode region of COI based upon searches in 156 marine animal genomes. The dotted vertical lines indicate read length cut-offs for standard HTS platforms (i.e., 150, 300, 450, 600 bp). A few NUMTs were longer than the 658 bp query length due to their possession of an insertion. n = 389 NUMTs [?]100 bp.

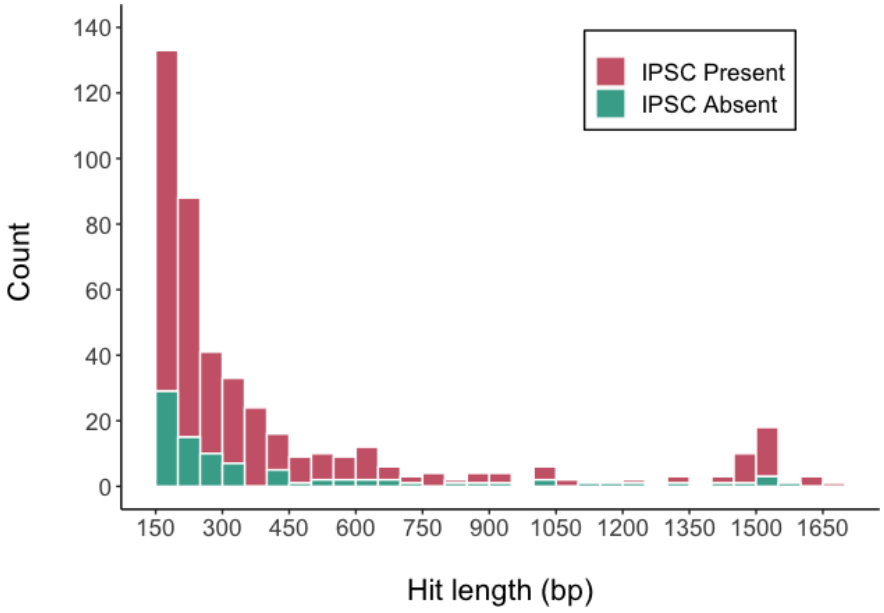

Figure 2 . The length distribution of COI NUMTs in the nuclear genomes of 85 marine invertebrate species where a full-length mitochondrial COI reference sequence was available to enable the search. n = 449 hits [?]150 bp.

Hosted file

`image3.emf` available at https://authorea.com/users/435441/articles/538460-do-pseudogenes-pose-a-problem-for-metabarcoding-marine-animal-communities

Figure 3 . Distribution of NUMTs along the mitochondrial COI gene for phyla of marine animals. 'Other' includes Annelida, Brachiopoda, Cephalorhyncha, Cnidaria, Ctenophora, Hemichordata, Nematoda, Nemertea, Platyhelminthes, Porifera, and Rotifera. Analysis was restricted to 85 genomes where a full-length mitochondrial COI reference sequence was available. n = 449 NUMTs [?]150 bp.

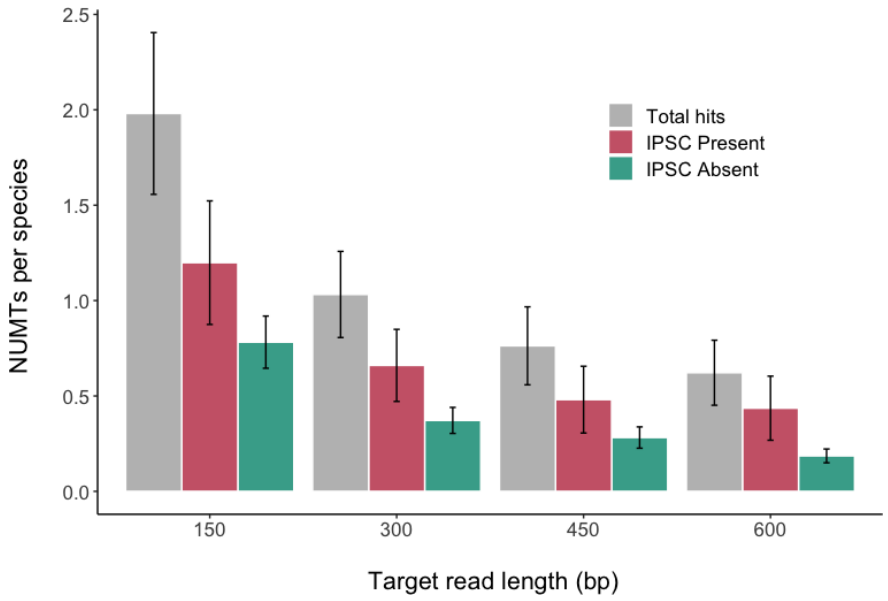

Figure 4 . The mean number of COI NUMTs for 156 marine animal species for four sequence lengths reovered by standard HTS platforms. Indels and premature stop codons (IPSC) were considered present if they occurred in the sequence region recovered by the platform (i.e., 150, 300, 450, 600 bp). Standard error bars are shown.

Hosted file

`image5.emf` available at https://authorea.com/users/435441/articles/538460-do-pseudogenes-pose-a-problem-for-metabarcoding-marine-animal-communities

Figure 5. Percent sequence divergence between the COI barcode sequence and NUMTs detected in 156 marine animal genomes for six length categories. Hits with indels or premature stop codons (IPSC) are in green while those lacking them are in red. The dotted vertical line in each panel indicates the 2% divergence threshold often used to delineate species. n = the number of hits in each length category.

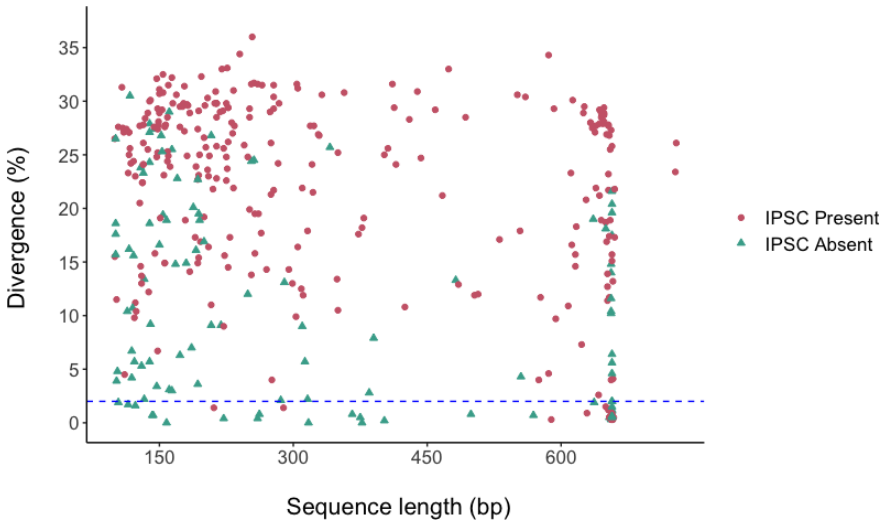

Figure 6 . Sequence divergence (%) between NUMTs and their mitochondrial counterpart as a function of sequence length for COI NUMTs with/without an IPSC (indel and premature stop codon). n = 389 hits [?]100 bp. Dashed line shows 2% divergence.

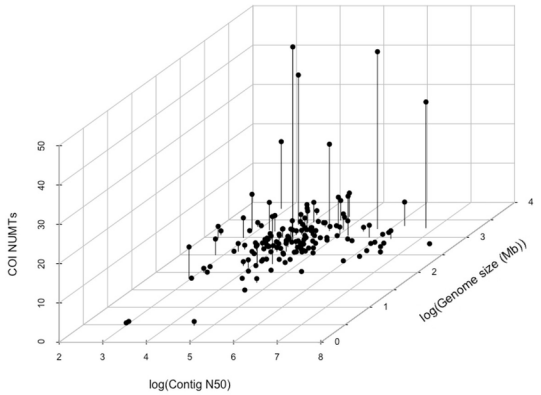

Figure 7 . The relationship between genome size, genome quality (contig N50), and the total number of COI NUMTs per genome for 156 marine animal species using the 658 bp barcode region of COI as a reference. n = 389 NUMTs [?]100 bp.

Hosted file

`image8.emf` available at https://authorea.com/users/435441/articles/538460-do-pseudogenes-pose-a-problem-for-metabarcoding-marine-animal-communities

Figure 8 . The mean number of COI NUMTs [?]100 bp by phylum for 156 marine animal species, including A) the total number of hits and B) the number without diagnostic features. The number in brackets indicates the number of genomes examined for each phylum.

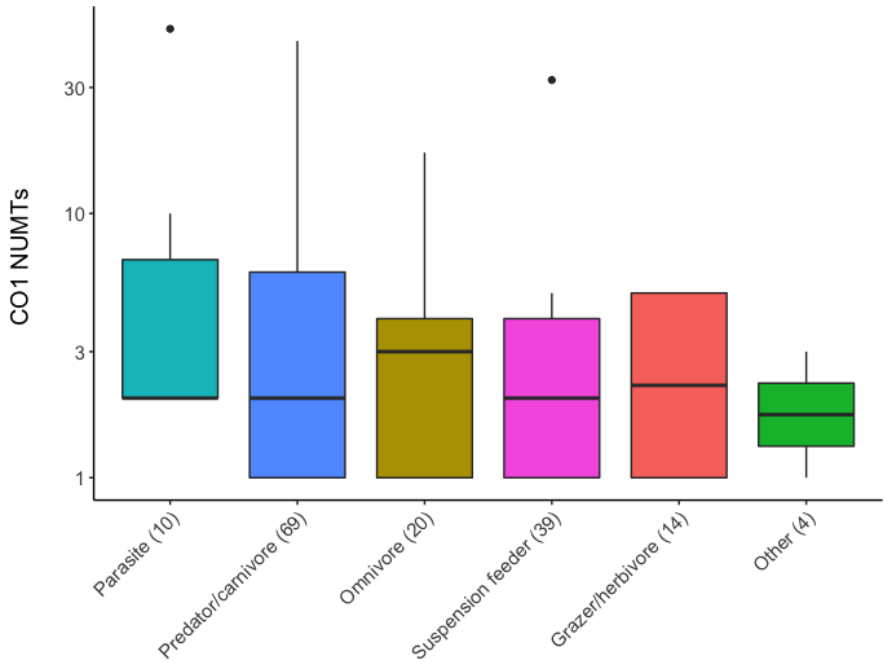

Figure 9 . Variation in the mean number of COI NUMTs [?]100 bp amomg 156 marine animal species with differing adult feeding habits. The number in brackets indicates the number of genomes in each category.

Hosted file

`image10.emf` available at https://authorea.com/users/435441/articles/538460-do-pseudogenes-pose-a-problem-for-metabarcoding-marine-animal-communities

Figure 10 . The mean number of COI NUMTs [?]100 bp for 156 marine animal species versus the presence or absence of A) asexual reproduction, B) sexual reproduction, C) hermaphroditism and D) colonialism. The numbers in brackets indicate the number of genomes in each category.