2008-09-09

Many species in one: DNA barcoding overestimates the number of species when nuclear mitochondrial pseudogenes are coamplified

Michael F. Whiting
Michael_Whiting@byu.edu

Hojun Song

See next page for additional authors

Follow this and additional works at: http://scholarsarchive.byu.edu/facpub

Part of the Biology Commons

Original Publication Citation

BYU ScholarsArchive Citation
Whiting, Michael F.; Song, Hojun; Crandall, Keith A.; and Buhay, Jennifer E., "Many species in one: DNA barcoding overestimates the number of species when nuclear mitochondrial pseudogenes are coamplified" (2008). All Faculty Publications. Paper 170.
http://scholarsarchive.byu.edu/facpub/170

This Peer-Reviewed Article is brought to you for free and open access by BYU ScholarsArchive. It has been accepted for inclusion in All Faculty Publications by an authorized administrator of BYU ScholarsArchive. For more information, please contact scholarsarchive@byu.edu.
Many species in one: DNA barcoding overestimates the number of species when nuclear mitochondrial pseudogenes are coamplified

Hojun Song*1, Jennifer E. Buhay*2, Michael F. Whiting*, and Keith A. Crandall*

*Department of Biology, Brigham Young University, Provo, UT 84602; and ‡Belle W. Baruch Institute for Marine Sciences, University of South Carolina, Columbia, SC 29208

Edited by W. Ford Doolittle, Dalhousie University, Halifax, NS, Canada, and approved July 14, 2008 (received for review March 28, 2008)

Nuclear mitochondrial pseudogenes (numts) are nonfunctional copies of mtDNA in the nucleus that have been found in major clades of eukaryotic organisms. They can be easily coamplified with orthologous mtDNA by using conserved universal primers; however, this is especially problematic for DNA barcoding, which attempts to characterize all living organisms by using a short fragment of the mitochondrial cytochrome c oxidase I (COI) gene. Here, we study the effect of numts on DNA barcoding based on phylogenetic and barcoding analyses of numt and mtDNA sequences in two divergent lineages of arthropods: grasshoppers and crayfish. Single individuals from both organisms have numts of the COI gene, many of which are highly divergent from orthologous mtDNA sequences, and DNA barcoding analysis incorrectly overestimates the number of unique species based on the standard metric of 3% sequence divergence. Removal of numts based on a careful examination of sequence characteristics, including indels, in-frame stop codons, and nucleotide composition, drastically reduces the incorrect inferences of the number of unique species, but even such rigorous quality control measures fail to identify certain numts. We also show that the distribution of numts is lineage-specific and the presence of numts cannot be known a priori. Whereas DNA barcoding strives for rapid and inexpensive generation of molecular species tags, we demonstrate that the presence of COI numts makes this goal difficult to achieve when numts are prevalent and can introduce serious ambiguity into DNA barcoding.

© 2008 by The National Academy of Sciences of the USA

Author contributions: H.S., J.E.B., M.F.W., and K.A.C. designed research; H.S. and J.E.B. performed research; H.S. and J.E.B. analyzed data; and H.S., J.E.B., M.F.W., and K.A.C. wrote the paper.

The authors declare no conflict of interest.

Author contributions: H.S., J.E.B., M.F.W., and K.A.C. designed research; H.S. and J.E.B. performed research; H.S. and J.E.B. analyzed data; and H.S., J.E.B., M.F.W., and K.A.C. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Data deposition: The sequences reported in this paper have been deposited in the GenBank database (accession nos. EU589049–EU589148, EU583504–EU583573, EU583577–EU583678, EF207161–EF207162, and EF207165–EF207168).

Freely available online through the PNAS open access option.

†To whom correspondence should be addressed: Department of Biology, 692 Wadsworth Building, Brigham Young University, Provo, UT 84602. E-mail: hojun.song@byu.edu.

This article contains supporting information online at www.pnas.org/cgi/content/full/0803076105/DCSupplemental.

© 2008 by The National Academy of Sciences of the USA
from diverse taxa by using a limited set of primers is heralded as one of the attractive features of this marker (14). It is true that relatively conserved regions within mtDNA allow the design of “universal” primers, which can amplify mitochondrial fragments from an unknown species (15). However, conserved primers can be a double-edged sword when numts are present because they can coamplify numts in addition to the target mtDNA (7, 8). If the nuclear integration of numts was an ancient and sufficient sequence divergence accumulated in the orthologous mtDNA, the conserved primers would be more likely to amplify numts in preference to mtDNA, which could possibly result in unambiguous, paralogous sequences (8). Despite this serious problem, numts have been dismissed as a minor concern for DNA barcoding (16) and the issue of numts has not been adequately addressed.

In this study, we investigate the effect of including numts in DNA barcoding in two divergent lineages of arthropods, insects, and crustaceans, which are known to have especially large numbers of numts (8, 17–19). We also examine the effect of numts at different levels of divergence: subfamily-level (grasshoppers) and species- and population-level (crayfish). Herein, we show that both grasshopper and crayfish species included in the study have numts of the COI gene and barcoding methods would incorrectly infer that single individuals belong to multiple, unique species. The prevalence of numts appears to be both species-specific and population-specific and the pattern of numt distribution is considerably different between lower-level and higher-level divergence among taxa. Finally, we demonstrate the importance of data exploration in DNA barcoding practice by examining sequence characteristics of numts.

**Results and Discussion**

Coamplification of Numts with Orthologous mtDNA. Our results strongly suggest that a large number of paralogous haplotypes of various divergences are coamplified with the orthologous mtDNA sequences when conserved primers are used in both grasshoppers and crayfish, which can be identified by the presence of indels, point mutations, and in-frame stop codons [supporting information (SI) Table S1]. The majority of the coamplified paralogs can be easily considered nonfunctional numt haplotypes because of the presence of in-frame stop codons, which is especially evident in our crayfish data in which 97.3% of paralogs have stop codons. A large number of these numts have unusually high numbers of point mutations (mean = 65.52, n = 110), suggesting that nuclear integration of mtDNA would result in random accumulation of nucleotide changes. In the grasshopper data, however, there are many paralogs that cannot readily be categorized as numts because they lack in-frame stop codons and differ from the orthologous mtDNA sequences by one or two nucleotides. If the same haplotype that appears to be functional other than the ortholog is repeatedly found from a single individual, one can suspect heteroplasmacy (5). Indeed, heteroplasmacy seems to explain the presence of certain paralogous haplotypes in Schistocerca americana. However, there are many paralogs represented by single haplotypes with small nucleotide differences in all four grasshopper species. Although, it is unlikely that these are Taq polymerase errors during PCR because of the high fidelity polymerase we used (0.015% error rate or 0.0732 bp per reaction); we cannot rule out the possibility of PCR error amplified by additional cloning (5, 18). Also, heteroplasmacy might in fact be a plausible explanation for these haplotypes because we limited our study to only 30 clones per grasshopper species, thus not exploring the full extent of heteroplasmic diversity.

If the proportion of numts is high compared with the orthologous mtDNA fragments in a given PCR product, it is possible to generate unambiguous paralogous sequences (8). This is exacerbated when the conserved primers preferentially amplify numts because of relatively ancestral sequence similarity of numts or divergence within the primer regions of the orthologous mtDNA. In this case, typical indicators of different PCR products, such as multiple bands on gels and double peaks, background noise, and ambiguity in sequence chromatograms, will not be present; hence, paralogous sequences can be mistaken as orthologous mtDNA. In fact, this exact phenomenon was observed in 18 crayfish individuals of Orconectes barri and Orconectes australis from which numts were amplified and cleanly sequenced without cloning.

**Phylogenetic Analyses and Distribution of Numts.** For the grasshopper data, the parsimony and the Bayesian analyses both recovered the monophyly of the orthologous mtDNA and haplotypes for three of four species (Fig. 1A); although, the topology was different in the placement of *S. americana* and *Calliptamus italicus* clades between the two methods. In each species, the largest clade was the polytomous clade consisting of the mtDNA ortholog and several similar haplotypes. The remaining haplotypes formed highly structured clades within each species, and this pattern was especially evident in *Acrida willemsi* and *S. americana*. For crayfish data, the Bayesian analysis recovered a topology mostly congruent with the parsimony analysis (Fig. 1B) and both analyses found a large clade of numt haplotypes (84 in parsimony and 82 in Bayesian) and a small clade of numts (18 in both analyses), distinctively divergent from the clearly defined clades of the orthologous mtDNA of four species. These numt clades consisted of the haplotypes from *O. australis, O. barri*, and *Orconectes packardi*, which were not necessarily grouped either by the species or the populations. Only four numt haplotypes were placed among orthologs (one in *Orconectes inconstans*, one in *O. australis*, and two in *O. barri*). A clade consisting of three numt haplotypes of *O. barri* was robustly placed near the root of trees in both analyses. Among the orthologous mtDNA clades, three of four species formed monophyletic clades, with the exception of *O. australis* that had one large clade sister to *O. barri* and a small clade basal to the *australis + barri* clade. Based on the phylogenetic analyses, number of indels, point mutations, in-frame stop codons, and sequence divergence, it is possible to conclude: among grasshoppers, *Locusta migratoria* has three numts, *A. willemsi* has six, *C. italicus* has two, and *S. americana* has at least 11 numts; among crayfish, *O. australis* has 60, *O. barri* has 46, *O. inconstans* has one, and *O. packardi* has four numts. On average, 32.54% and 41.88% of haplotypes generated from grasshoppers and crayfish, respectively, were numts (Table S1). It is important to note that this is a conservative estimate of the number of numts per species because it is limited by the number of individuals and clones we generated.

Both the grasshopper and crayfish data suggest that there can be multiple types of numts present within single individuals that vary considerably in nucleotide composition, suggesting multiple independent transfer events from the mitochondrial genome to the nucleus (8, 17). Moreover, our data suggest that these independent nuclear integration events can give rise to a family of numts that can diverge at different substitution rates. For example, we found that the relationships among the numt haplotypes of *S. americana* are highly structured and a similar pattern was observed in other grasshopper and crayfish species. Not only can independent transfer occur multiple times, but it can occur at very different phylogenetic levels. Whereas many numts of *S. americana* are closely related to the orthologous mtDNA, two numt haplotypes form a strong clade with the mtDNA ortholog of *Anabrus simplex*, which belongs to a different suborder within Orthoptera. Similarly, three numt haplotypes of *O. barri* were placed near outgroups belonging to different crayfish genera. These findings collectively suggest that there could have been an ancient nuclear integration event and that enough time has passed for these numts to have accumu-
We conclude that the frequency of DNA transfer from mitochondria to the nucleus varies among species and might be determined by the presence of numts, which can clearly predate and postdate speciation events. The analysis of numts suggests that a positive correlation between the number of numts and nuclear genome size might exist and it is possible that the amount of numts in a given population is proportional to the number of individuals within a population 

\[\text{frequency of DNA transfer} \propto \text{number of numts} \propto \text{population size} \]

The distribution of numts in both datasets suggests that their prevalence may be lineage-specific. For example, we sequenced numts from individuals collected from 11 of 56 cave crayfish populations (Tables S2 and S3). Among the four species, numts were sequenced for 7 of 34 localities in *O. australis* (southernmost region of range, primarily caves in Alabama), 2 of 7 in *O. barri* (southernmost region of range, caves in Tennessee), 1 of 3 in *O. incomptus*, and 1 of 12 in *O. packardi*. This observed pattern does not necessarily mean that there is a nonrandom population-specific variation in the level of numt prevalence. Nuclear integration of mtDNA happens at the level of the individual (8) and a large population size can effectively dilute the amount of numts in a given population. In this case, the proportion of numts is much smaller than that of the orthologous mtDNA in a given individual, rendering the numt cocomplication less likely, even with a possibility of the presence of pleiomorphic numts in the nucleus. However, whether a population experiences genetic drift because of an extreme bottleneck, numts can be fixed in a few founders, resulting in a disproportionately high level of numts. Another intrinsic factor, nuclear genome size, might also play a role that results in uneven distribution of numts. Whereas all grasshopper species have numts, *A. willemsei* and *S. americana* have especially high numbers that are divergent from each other. Bensasson et al. (21) suggested that a positive correlation between the number of numts and nuclear genome size might exist and it is possible that these two species might have larger nuclear genomes than the others. Alternatively, inherent species-specific differences in the frequency of DNA transfer from mitochondria to the nucleus and in the rate of loss of numts in the nucleus have also been

![Diagram](image_url)
Rigorous Quality Control Against Numts Is Necessary in DNA Barcoding. According to the standard DNA barcoding protocol published by the Consortium for the Barcode of Life (CBOL, http://barcoding.si.edu), there appears to be a minimum amount of quality control involved in generating COI sequences. By using simple protocols, it is easy to generate molecular species tags rapidly and inexpensively, which is one of the main goals of DNA barcoding initiative (22). However, molecular evolutionary processes such as the nuclear integration of mtDNA present challenges to every step of this standard protocol, and without specific quality control measures in place, the integrity of DNA barcode sequences is seriously compromised (3). Despite obvious shortcomings because of numts, the proponents of DNA barcoding have argued that “numts have proven a minor limitation to using mitochondrial barcode in groups studied so far.” Hebert et al. (16) also suggested that the taxonomic implication of numts is small. Our study clearly demonstrates that coamplification of numts with the mtDNA orthologs is not only a major limitation of DNA barcoding, but also has significant taxonomic implications. Reported cases of numts are ever increasing and numts should not be treated as mere nuisance any more. In a showcase study of DNA barcoding, Hebert et al. (16) found that 13 individuals of skipper butterfly Astraptes fulgerator had heterozygous sequences and concluded that they were numts because the nonmitochondrial sequences of these 13 individuals (obtained by subtracting “typical sequences” from the ambiguous region) were highly similar among each other. The rationale behind this conclusion was that numts would be conserved among individuals because numts represent an ancient molecular event whereas the corresponding mtDNA sequences would be more variable. However, empirical studies of numts suggest that substitution occurs at different rates once mtDNA has been integrated into nucleus (23). Brower (24) argued against the claim by Hebert et al. (16) and suggested that the 13 haplotypes that shared heterozygosity were unlikely to be numts, but likely to be heteroplasmic mtDNA at best. We downloaded these questionable sequences from GenBank (AY666889, AY666943, AY666968, and AY667044) and looked for any sign of compositional bias and other indicators of numts. In comparison with other DNA barcodes Hebert et al. (16) used, these presumed numts had identical nucleotide composition and translated amino acid sequences as the orthologous mtDNA sequences and had no indels or in-frame stop codons. In other words, these presumed numts sequences were fully functional copies of mtDNA and are likely heteroplasmic mtDNA. With careful examination of the sequences in question, Hebert and colleagues could have easily avoided their incorrect inference.

How to Control for Numts. Among the few barcoding studies that did attempt to control for numts (25), researchers extracted DNA from tissues known to be rich in mitochondria and amplified slightly longer fragments (750 bp) based on the idea that numts are shorter than the barcode amplicon (26). However, muscle tissues still do contain nuclear DNA that can harbor numts and the size of numts can be highly variable and not necessarily smaller than the typical 700-bp size of DNA barcodes (10). Several methods have been suggested as means to avoid numt coamplification, including RT-PCR, long PCR, and mtDNA enrichment (8). However, these methods are often tedious, time-consuming and expensive, and their efficiency is often not high enough to totally avoid numts (27). Recent studies have questioned the universal utility of the Folmer region in DNA barcoding (28), and especially when a large number of numts are suspected in this region, it would be worthwhile to analyze additional markers other than COI gene. We recommend a number of steps that researchers could attempt to avoid:

- Examine the Folmer region in DNA barcoding (28), and especially when a large number of numts are suspected in this region, it would be worthwhile to analyze additional markers other than COI gene. We recommend a number of steps that researchers could attempt to avoid:

  - Examine the Folmer region in DNA barcoding (28), and especially when a large number of numts are suspected in this region, it would be worthwhile to analyze additional markers other than COI gene. We recommend a number of steps that researchers could attempt to avoid:

- Examine the Folmer region in DNA barcoding (28), and especially when a large number of numts are suspected in this region, it would be worthwhile to analyze additional markers other than COI gene. We recommend a number of steps that researchers could attempt to avoid:

- Examine the Folmer region in DNA barcoding (28), and especially when a large number of numts are suspected in this region, it would be worthwhile to analyze additional markers other than COI gene. We recommend a number of steps that researchers could attempt to avoid:

- Examine the Folmer region in DNA barcoding (28), and especially when a large number of numts are suspected in this region, it would be worthwhile to analyze additional markers other than COI gene. We recommend a number of steps that researchers could attempt to avoid:

- Examine the Folmer region in DNA barcoding (28), and especially when a large number of numts are suspected in this region, it would be worthwhile to analyze additional markers other than COI gene. We recommend a number of steps that researchers could attempt to avoid:

- Examine the Folmer region in DNA barcoding (28), and especially when a large number of numts are suspected in this region, it would be worthwhile to analyze additional markers other than COI gene. We recommend a number of steps that researchers could attempt to avoid:
that potentially have numts, this issue cannot simply be ignored. Otherwise, the number of single individuals that are inferred to be multiple species because of numt contamination may become the legacy of the DNA barcode movement.

Materials and Methods

Taxon Sampling. To study the evolution and distribution of numts at higher-level divergence, we included four grasshopper species belonging to four different subfamilies of Acrididae. To establish the orthology of mtDNA, we used the taxa whose partial or complete mitochondrial genomes have been sequenced: Acrida willemsei (Acridinae, EUS89053), Calliptamus italicus (Calliptaminae, EUS89054), Locusta migratoria (Oedipodinae, EUS89051), and Schistocerca americana (Cyrtauchenidae, EUS89055). We generated numts from single individuals per species and used the same individuals that the complete mitochondrial genomes were sequenced from with an exception of L. migratoria. As outgroups, we used the COI regions of a Mormon cricket Anabrus simplex (EUS89052), a cockroach Graptopsalbula portentosa (EUS89049), and a termite Mastotermes darwinensis (EUS89050). To study the numts at population- and species-level divergence, we included a total of 119 individuals of four closely related species belonging to the cave crayfish genus Orconectes, collected from 56 localities along the Cumberland Plateau of the Southern Appalachians: O. australis, O. barri, O. incompitus, and O. packardi. As outgroups, we included O. limosus (AF517105), Procambarus simulans (EUS83575), and three species of the genus Cambarus (C. gentryi [DQ41785], C. tenebrosus [EUS83576], and C. bartoni [EUS83574]). COI from the complete mtDNA genome of Cherax destructor (NC.011243) was used for reference. GenBank accession numbers for the haplotypes are EUS89057–EUS89148 (grasshoppers) and EUS83540–EUS83573, EUS83577–EUS83678, EF207161–EF207162, and EF207165–EF207168 (crayfish). Details about numt amplification can be found in SI Methods.

Characterization of Numts. To ensure the quality and identity, each haplotype was blasted by using MegabLAST option against the nucleotide collection (nr/nr) as implemented in the National Center for Biotechnology Information website (http://www.ncbi.nlm.nih.gov/BLAST/Blast.cgi). Only the haplotypes that had high similarity to the COI sequence were used for the further analyses. For example, the blast search revealed that eight cloned sequences (3 from L. migratoria and 5 from A. willemsei) were of either bacterial or unknown origins and these sequences were treated as cloning error and removed from further analyses. We characterized the haplotypes in Sequencer 4.6 for length, nucleotide composition, and number of in-frame stop codons. Putative indels and point mutations were estimated by comparing the haplotype sequences against the mtDNA orthologs. The number of unique haplotypes was determined and the sequence divergence from the orthologous mtDNA sequence for each species was calculated under Kimura 2-parameter (K2P) model in MEGA 3.1 (32), as routinely used in barcoding studies, despite this model being under fit relative to the data (see below).

Data Analysis. To study divergence pattern of numts, we performed phylogenetic analyses in both parsimony and Bayesian frameworks. For both grasshopper and crayfish datasets, the unique haplotypes and the mtDNA orthologs were aligned in MUSCLE (33) by using default parameters. For grasshopper data, we created a matrix of 69 terminals (7 mtDNA orthologs and 62 unique haplotypes) and 475 aligned nucleotides. For crayfish data, we created a matrix of 215 terminals (5 outgroups and 210 unique haplotypes of four species) and 663 aligned nucleotides. Within the parsimony framework, the aligned sequence data were analyzed with gaps treated as missing, by using search algorithms implemented in TNT (www.zmuc.dk/public/phylens). To assess support, we calculated standard bootstrap values based on 1,000 replicates (100 random-addition TBR replicates each) and Bremer support values, both in TNT. Within the Bayesian framework, we analyzed the datasets by using the program MrBayes v.3.1 (34) after selecting the best-fit models of nucleotide evolution under the AIC criteria by using MrModeltest 2.2 (program distributed by J.A.A. Nylander, Evolutionary Biology Centre, Uppsala University). The analyses consisted of running four simultaneous chains for 20 million generations for grasshopper data (GTR + G) and six simultaneous chains for 30 million generations for crayfish data (HKY + I + G), both sampling every 1,000 generations. Four independent identical Bayesian runs were performed to ensure convergence on similar results and the nodal support was assessed by using the posterior probability generated from a consensus tree of the sampled trees past burn-in determined by using Trace 1.4 (http://beast.bio.ed.ac.uk).

To study how the presence of numts might influence the inferences from DNA barcoding, we performed a neighbor-joining (NJ) analysis under K2P
model and calculated the sequence divergence among haplotypes on each dataset in MEGA 3.1 as typically used in barcoding studies (Z, 16). From the NJ analyses and sequence divergence data, we then determined the number of clusters that would be considered unique species under the DNA barcoding standard (3% or higher sequence divergence). Numts are known to accumulate in-frame stop codons because they become nonfunctional after nuclear integration and are no longer under selective pressure to conserve an ORF (8). Based on the presence of in-frame stop codons, it would be possible to identify numts from the data and remove them from the analyses. We applied this method to the grasshopper and crayfish datasets and performed barcoding analyses on the reduced datasets to test whether the correct inferences would be made after such a correction.

ACKNOWLEDGMENTS. We thank C. Shepherd and N. Sheffield for technical assistance; H. Park for illustration; J.W. Sites for comments on the manuscript; two anonymous reviewers for improving the quality of the manuscript. This work was supported by National Science Foundation (NSF) Grants EF-0531665 to MFW and EF-0531762 to KAC.