MOLECULAR ECOLOGY
RESOURCES
Molecular Ecology Resources (2010) **10**, 615–627 doi: 10.1111/j.1755-0998.2009.02823.x

DNA BARCODING

Assessing the effects of primer specificity on eliminating numt coamplification in DNA barcoding: a case study from Orthoptera (Arthropoda: Insecta)

MATTHEW J. MOULTON, HOJUN SONG and MICHAEL F. WHITING
Department of Biology, 401 WIDB, Brigham Young University, Provo, UT 84602, USA

Abstract

DNA barcoding is a diagnostic method of species identification based on sequencing a short mitochondrial DNA fragment of cytochrome oxidase I (COI), but its ability to correctly diagnose species is limited by the presence of nuclear mitochondrial pseudogenes (numts). Numts can be coamplified with the mitochondrial orthologue when using universal primers, which can lead to incorrect species identification and an overestimation of the number of species. Some researchers have proposed that using more specific primers may help eliminate numt coamplification, but the efficacy of this method has not been thoroughly tested. In this study, we investigate the taxonomic distribution of numts in 11 lineages within the insect order Orthoptera, by analysing cloned COI sequences and further test the effects of primer specificity on eliminating numt coamplification in four lineages. We find that numts are coamplified in all 11 taxa using universal (barcoding) primers, which suggests that numts may be widespread in other taxonomic groups as well. Increased primer specificity is only effective at reducing numt coamplification in some species tested, and only eliminates it in one species tested. Furthermore, we find that a number of numts do not have stop codons or indels, making it difficult to distinguish them from mitochondrial orthologues, thus putting the efficacy of barcoding quality control measures under question. Our findings suggest that numt coamplification is a serious problem for DNA barcoding and more quality control measures should be implemented to identify and eliminate numts prior to using mitochondrial barcodes for species diagnoses.

Keywords: DNA barcoding, numt, Orthoptera, primer specificity

Received 28 July 2009; revision received 31 October 2009; accepted 23 November 2009

Introduction

DNA barcoding is a method designed to identify species rapidly using a short, standardized gene region, Cytochrome *c* oxidase subunit I (COI), as a species tag (Hebert *et al.* 2003; Hebert & Gregory 2005). A 658-bp fragment of COI is amplified using universal primers, known as Folmer primers, which are effective at amplifying this region from metazoan mitochondria (Folmer *et al.* 1994; Hebert *et al.* 2003). A database of over 500 000 of these COI fragments, known as 'barcodes', has been established for nearly 38 000 species (Ratnasingham & Hebert 2007) and is currently being used to help rapidly assign individuals to known species and highlight potential new species (Hebert & Gregory 2005; Witt *et al.* 2006; Gomez *et al.* 2007). However, the use of DNA barcoding as an effective tool for species identification faces many challenges associated with mitochondrial DNA (mtDNA) including reduced effective population size, introgression, maternal inheritance, recombination, inconsistent mutation rate and heteroplasmy (see Rubinoff *et al.* 2006, for a complete discussion of these topics).

In addition to these challenges, primers intended to amplify the COI orthologue may also coamplify nuclear mitochondrial pseudogenes (numts) – copies of mitochondrial genes that are incorporated into the nuclear genome (Gellissen *et al.* 1983; Lopez *et al.* 1994; Zhang & Hewitt 1996b; Sorenson & Quinn 1998; Bensasson *et al.* 2001a). After nuclear integration, numts may accumulate

Correspondence: Matthew J. Moulton, Fax: 801-422-0090;
E-mail: matthewjmoulton@gmail.com

mutations (including in-frame stop codons and indels) and become nonfunctional (Bensasson *et al.* 2001a). Numts have been identified in a large number of eukaryotic lineages and vary greatly in number and size among species (Bensasson *et al.* 2001a; Richly & Leister 2004). Some researchers suggest that numts can be readily identifiable by characteristic mutations and be removed from analyses (Hebert *et al.* 2004a), but some numts lack such characteristics and are difficult to identify (Song *et al.* 2008). Some species that are known to contain a large number of numts (including insects) are especially problematic when conventional PCR methods coamplify many numts (Arctander 1995; Zhang & Hewitt 1996a, 1997; Williams & Knowlton 2001), which can lead to incorrect species identification via DNA barcoding and an overestimation of the number of species, even when quality control measures are in place (Song *et al.* 2008). One of the remedies to avoid numt coamplification, as suggested by Song *et al.* (2008), is to use more specific primers. However, it is not clear whether more specific primers will solve the problem of numt coamplification, as the efficacy of this method has not been thoroughly tested to date.

Careful primer design is critical to ensure amplification of the correct fragment in PCR (Weissensteiner *et al.* 2004). The ability of a primer to anneal to a DNA fragment depends on a number of factors including primer length, degeneracy, G + C content, melting temperature and the degree to which a primer matches the complementary fragment sequence (Hughes & Moody 2007). The use of more specific primers (especially taxon-specific primers that are exact complementary matches of the target sequence) has proven useful when sequencing complete mitochondrial genomes for use in phylogenetic inference (Cameron *et al.* 2007; Fenn *et al.* 2007; Sheffield *et al.* 2008) and for isolating small mtDNA fragments for some DNA barcoding analyses (Hebert *et al.* 2004a,b; Vences *et al.* 2005; Tedersoo *et al.* 2008). However, a DNA barcode is intended to identify species without having any a priori knowledge about the organism. Therefore, typical DNA barcoding methods use universal primers (such as the Folmer primers), even though such primers have been shown to coamplify numts (Bensasson *et al.* 2000; Williams & Knowlton 2001; Antunes & Ramos 2005; Benesh *et al.* 2006; Song *et al.* 2008).

The purpose of this study was to address the following questions. What is the taxonomic distribution of numts within the insect order Orthoptera? What is the effect of using more specific primers on eliminating numt coamplification in DNA barcoding? To what extent do primers of differing specificity coamplify different types of numts? To what degree can quality control measures correctly identify numts for removal prior to barcoding analyses? We discuss the results of this case study and the implications these results have on DNA barcoding efforts.

Materials and methods

Taxon sampling

We selected 11 taxa to investigate the taxonomic distribution of COI-like numts present in Orthoptera (Table 1). This taxon sampling represents 10 families from both Ensifera and Caelifera whose mitochondrial genomes have been completely sequenced (Fenn *et al.* 2007, 2008; Flook *et al.* 1995; Song *et al.* 2008; H. Song, unpublished data). We can be confident that we know the orthologous COI sequence because these complete genome sequences were generated using long PCR such that it is not likely that they contain numt sequences. The complete mitochondrial genomes were either downloaded from GenBank (accession numbers: *Schistocerca americana* – EU589056; *Locusta migratoria* – X80245; *Myrmecophilus manni* – EU938370; *Anabrus simplex* – EF373911) or obtained from H. Song (unpublished data) (*Prionotropis hystrix*, *Ellipes minutus*, *Tristira magellanica*, *Trigonopteryx* sp., *Physemacris variolosa*, *Lentula* sp. and *Lithidiopsis* sp.).

Table 1 List of taxa used in this study, including voucher number for voucher specimens deposited at Brigham Young University

Family	Subfamily	Genus	Species	Voucher no.
Pamphagidae	Prionotropisinae	*Prionotropis*	*hystrix*	OR151
Tridactylidae		*Ellipes*	*minutus*	OR153
Tristiridae	Tristirinae	*Tristira*	*magellanica*	OR204
Trigonopterygidae	Trigonopteryginae	*Trigonopteryx*	sp.	OR290
Pneumoridae		*Physemacris*	*variolosa*	OR293
Lentulidae	Lentulinae	*Lentula*	sp.	OR295
Lithidiidae	Lithidiinae	*Lithidiopsis*	sp.	OR316
Acrididae	Oedipodinae	*Locusta*	*migratoria*	Loc001
Acrididae	Cyrtacanthacridinae	*Schistocerca*	*americana*	Sch015
Tettigoniidae	Tettigoniinae	*Anabrus*	*simplex*	OR034
Myrmecophilidae	Myrmecophilinae	*Myrmecophilus*	*manni*	OR022

Table 2 List of primers used in this study, including sequences and properties

Primer	Sequence (5′–3′)	T_M	G + C%
LCO1490 (Folmer J)	GGTCAACAAATCATAAAGATATTGG	52.9	32.1
Orthoptera-specific J	TC.............G..C.....	53.4	36.0
Schistocerca J	TC........C.....G........	53.0	36.0
Locusta J	TC........C..C..G..C.....	57.7	44.0
Myrmecophilus J	TC....................C..	51.0	32.0
Anabrus J	TT.....T........G..C.....	51.4	32.0
HCO2198 (Folmer N)	TAAACTTCAGGGTGACCAAAAAATCA	55.3	34.6
Orthoptera-specific N	........T..........G.....	56.4	38.5
Schistocerca N	........T.....T.....G.....	56.4	38.5
Locusta N	........T..........G.....	56.4	38.5
Myrmecophilus N	..T.....T..A........G.....	53.7	34.6
Anabrus N	..G.............G..G.....	59.0	46.2

Dots indicate no nucleotide difference from the Folmer primers.

We extracted the Folmer region of COI from the mitochondrial genome of each taxon to use as an orthologous reference sequence. We selected 4 of the 11 taxa for a more detailed study of the effects of primer specificity on eliminating numt coamplification: *S. americana*, *L. migratoria*, *M. manni* and *A. simplex*. The Folmer regions of COI from three additional orthopteran lineages were included in phylogenetic analyses for taxonomic reference: (i) *Gryllotalpa orientalis* (Gryllotalpidae, Gryllotalpinae; AY660929.1); (ii) *Oxya chinensis* (Acrididae, Oxyinae; NC_010219.1); and (iii) *Ruspolia dubia* (Tettigoniidae, Conocephalinae; NC_009876.1). The Folmer regions of COI from four polyneopteran taxa were used as outgroups: (i) *Periplaneta fuliginosa* (Blattidae, Blattinae; NC_006076); (ii) *Reticulitermes santonensis* (Rhinotermitidae, Heterotermitinae; EF206315.1); (ii) *Tamolanica tamolana* (Mantidae, Mantinae; NC_007701.1); and (iv) *Sclerophasma paresisense* (Mantophasmatidae, Mantophasmatinae; NC_007702.1).

Primer design

We amplified the barcoding region of the COI gene from the 11 taxa using the Folmer primers in order to explore the distribution of numts within Orthoptera. The Folmer primers were designed from 11 diverse metazoan lineages in order to have no degenerate bases and are considered to be the most universal primers to amplify COI within metazoa (Folmer *et al.* 1994). The Folmer major (J) primer binds to the light chain of COI at *Drosophila yakuba* 5′-nucleotide number 1490 and the minor (N) primer binds to the 5′-nucleotide number 2198 of the heavy chain.

In order to study the effect of primer specificity on eliminating numt coamplification, we designed specific primers in addition to the Folmer primers. Based on all published and unpublished sequences of orthopteran COI available to us, we designed Orthoptera-specific primers that are more specific nucleotide matches of orthopteran mitochondrial orthologues and taxon-specific primers to be exact matches of the mitochondrial orthologue of each species. All primers designed for this study anneal in the same positions as the Folmer primers (Table 2).

PCR, sequencing and cloning

DNA from the 11 taxa was extracted using QIAGEN DNeasy kit and voucher specimens were deposited in the Insect Genomics Collection at Brigham Young University (Table 1). The Folmer region of COI was amplified via PCR using Elongase Enzyme mix (Invitrogen Corporation) and forward and reverse primer sets (Table 2). We used this high-fidelity enzyme mix to reduce polymerase error (error rate of 0.015% or 0.0987 bp per Folmer sequence; Leroux *et al.* 1997). The same PCR-cycling conditions described in Song *et al.* (2008) were adopted for this study. We obtained a total of 19 PCR amplicons (11 from all species using the Folmer primers and one from each, *S. americana*, *L. migratoria*, *M. manni* and *A. simplex* using Orthoptera-specific and taxon-specific primers) that were filter-cleaned using PrepEase Purification 96-well plate kit (USB Corporation) and cloned using the TOPO TA Cloning Kit (Invitrogen Corporation). About 50–120 colonies were picked from each cloning reaction and placed into 1× TE buffer (Fig. 1). Each colony of cells was lysed at 96 °C for 10 min to allow DNA to escape into the buffer solution. DNA from each colony was amplified via PCR using M13 primers included in the cloning kit (Invitrogen Corporation) and filter-cleaned using the protocol above. We used BigDye (version 3.1) chain terminating chemistry (Applied Biosystems Incorporated) to sequence the PCR amplicons for each colony and the 19 PCR

(a)

	Caelifera									Ensifera	
	P. hystrix	*E. minutus*	*T. magellanica*	*Trigonopteryx* sp.	*P. variolosa*	*Lentula* sp.	*Lithidiopsis* sp.	*S. americana*	*L. migratoria*	*M. manni*	*A. simplex*
Clones generated	50	50	50	50	50	50	50	100	100	100	100
Unique haplotypes	22	7	16	15	16	15	24	74	32	26	41
Sequences that match ortholog	24	34	27	22	28	28	18	1	57	57	27
Haplotypes with stop codons	3	4	2	4	2	2	10	37	4	1	5
Haplotypes with indels	1	4	2	6	2	3	14	31	1	0	12
Haplotypes with point mutations	22	6	16	15	15	16	27	74	32	26	41
Putative ortholog haplotypes	6	2	3	1	5	4	2	18	21	23	14
Heteroplasmy haplotypes	0	0	0	1	0	0	0	7	5	0	3
Numt haplotypes	16	5	13	13	11	11	22	49	6	3	24
Unique species identified using DNA barcoding methodology	4	1	4	5	4	2	9	6	1	1	10

(b)

	S. americana			*L. migratoria*			*M. manni*			*A. simplex*		
	Folmer	Ort.-spec	Tax.-spec	Folmer	Ort.-spec	Tax.-spec	Folmer	Ort.-spec	Tax.-spec	Folmer	Ort.-spec	Tax.-spec
Clones generated	100	120	100	100	100	100	100	100	100	100	100	100
Unique haplotypes	74	94	63	32	34	31	26	21	20	41	27	32
Sequences that match ortholog	1	1	1	57	38	76	57	64	64	27	64	52
Haplotypes with stop codons	37	61	45	4	4	0	1	1	2	5	6	4
Haplotypes with indels	31	56	45	1	5	0	0	1	1	12	5	6
Haplotypes with point mutations	74	95	63	32	34	32	26	23	20	41	27	32
Putative ortholog haplotypes	18	5	3	21	28	29	23	19	18	14	18	23
Heteroplasmy haplotypes	7	2	5	5	2	2	0	0	0	3	0	3
Numt haplotypes	49	87	55	6	4	0	3	2	2	24	9	6
Unique species identified using DNA barcoding methodology	6	16	9	1	1	1	1	1	1	10	3	1

Fig. 1 Sequence Data Summary. Summary of data obtained from nucleotide and amino acid sequences. Folmer primers co-amplify numts in all species tested to varying degrees. Tables include number of species that may be identified from a single individual due to coamplification of numts that exhibit >3% sequence divergence from the orthologue and do not have stop codons or indels. a) Data for all taxa using the Folmer primers. b) Data for four taxa listed by taxon and primer used to generate haplotypes (Folmer, Orthoptera-specific [Ort.-spec], and taxon-specific [Tax.-spec]).

amplicons obtained before cloning (to simulate actual DNA barcoding analyses). Each sequencing reaction was filter-cleaned and fractionated on an automated sequencing machine (ABI3730xl; Applied Biosystems Incorporated).

Table 3 List of GenBank accession numbers for haplotype nucleotide sequences obtained for this study

Taxon	GenBank accession numbers
Schistocerca americana	EU589119–EU589148, GU116114–GU116181, GU115905–GU115908, GU122627–GU122783, GU115857
Locusta migratoria	GU122341–GU122437
Myrmecophilus manni	GU122438–GU122504
Anabrus simplex	GU122241–GU122340
Prionotropis hystrix	GU122505–GU122527
Ellipes minutus	GU122528–GU122535
Tristira magellanica	GU122536–GU122552
Trygonopteryx sp.	GU122553–GU122568
Physemacris variolosa	GU122569–GU122585
Lentula sp.	GU122586–GU122601
Lithidiopsis sp.	GU122602–GU122626

Sequence analysis

Sequence data were compiled in Sequencher 4.7 (Gene Codes Corporation), vector and primer sequences were removed, and contigs were assembled to identify unique haplotypes. Haplotype nucleotide sequences are deposited in GenBank (Table 3). Each haplotype was blasted using MEGABLAST option against the nucleotide collection (nr/nt), available on the NCBI website (http://blast.ncbi.nlm.nih.gov/Blast.cgi?CMD=Web&PAGE_TYPE=BlastHome) to identify cloning error. Only haplotypes that blasted within the correct suborder with *E*-values ≤1.00E-30 were included in this study. We used Sequencher in order to quantify indels and point mutations by comparing each haplotype to its mitochondrial orthologue reference sequence. The number of stop codons in each haplotype was calculated using the Sequence Statistics tool available on MOSAS (Manipulation, Organization, Storage, and Analysis of Sequences; http://mosas.byu.edu/). Nucleotide sequences were translated into amino acid sequences using MEGA4 (Tamura *et al*. 2007).

We recognize that there may be potential sources of error in generating nucleotide data as described above (Williams & Knowlton 2001; Frey & Frey 2004; Song

et al. 2008), and therefore we followed explicit criteria in categorizing haplotypes. We identified putative orthologues as those haplotypes that satisfied all of the following criteria: (i) no stop codons; (ii) no indels; and (ii) three or fewer nucleotide substitutions (attributable to error) that cause one or fewer amino acid sequence differences compared to the orthologue. We identified numts as those haplotypes that satisfied at least one of the following criteria: (i) one or more stop codons; (ii) one or more indels; and (iii) any number of point mutations that cause two or more amino acid sequence differences compared to the orthologue. We identified heteroplasmy as those haplotypes that met all the criteria for putative orthologues except they contained greater than three nucleotide differences, but had identical amino acid sequences compared to the orthologue (silent substitutions not attributable to error).

Phylogenetic methods

We assembled a total of 10 data sets in order to explore how primer specificity might affect data generation and subsequent phylogenetic analyses (Table 4). Three 'primer-type' data sets were assembled to assess the effects of primer specificity on eliminating numt coamplification. Data set 1 included haplotypes of (i) *S. americana*; (ii) *L. migratoria*; (iii) *M. manni*; and (iv) *A. simplex* generated using the Folmer primers. This data set served as a means to understand how numts generated using Folmer primers were distributed among four orthopteran species and as a reference data set to compare with the other data sets. Data sets 2 and 3 included haplotypes of the four taxa above generated using the Orthoptera- and taxon-specific primers respectively. Four 'taxon' data sets were assembled to test whether primers of differing specificity preferentially amplify different types of numts. These four data sets included all haplotypes generated from each of the four taxa above, using all primer types (4 – *S. americana*, 5 – *L. migratoria*, 6 – *M. manni*, 7 – *A. simplex*). Finally, three 'sanitized' data sets were assembled to test whether quality control measures applied against numts could yield correct inference in a DNA barcoding framework. We excluded any haplotype that could be readily identified as a numt, based on the presence of stop codons or indels, from being assembled into these sanitized data sets. We then assembled all remaining haplotypes into three data sets based on the primers that were used to generate those haplotypes (8 – Folmer sanitized, 9 – Orthoptera-specific sanitized and 10 – taxon-specific sanitized).

Haplotype sequence lengths ranged from 254 to 766 bp across all data sets and were aligned using MUSCLE (Edgar 2004) with default parameters, which yielded 10 aligned data sets ranging in length from 658 to 1101 bp. The primer-type and taxon data sets (1–7) were analysed in both parsimony and Bayesian framework (gaps treated as missing). The parsimony analyses were performed using New Technology Search algorithms (sectorial search, ratchet, drift and tree fusing) in TNT (Goloboff *et al.* 2008) and bootstrap support values were calculated from 5000 replicates with 100 random-addition TBR replicates each. The Bayesian analyses were performed using MrBayes 3.1.2 (Ronquist & Huelsenbeck 2003) after identifying the best-fit models of nucleotide evolution for each data set under the AIC criteria given by MrModelTest 2.2 (program distributed by J.A.A. Nylander, Evolutionary Biology Centre, Uppsala University). The best-fit model was identified as GTR + G for the primer-type data sets (1–3) as well as the *S. americana* and *L. migratoria* data sets (4–5) and as GTR + I + G for the *M. manni* and *A. simplex* data sets (6–7). For all data sets, each Bayesian analysis consisted of running four simultaneous chains for 30 million generations and sampling every 1000 generations over four identical runs. Analyses were performed on a 64-node cluster of 512 Intel Xeon (E5345) processors at the Department of Biology, Brigham Young University. A majority rule consensus and posterior probabilities were calculated from the sampled trees after burn-in using TRACER 4.1 (http://tree.bio.ed.ac.uk/software/tracer/). The sanitized data sets (8–17) were

Table 4 List of data set used for phylogenetic analysis by type, name and number

Data set type	Data set number and name	Taxa/primers used to generate haplotypes
Primer type	1. Folmer	Four taxa/Folmer
Primer type	2. Orthoptera-specific	Four taxa/Orthoptera-specific
Primer type	3. Taxon-specific	Four taxa/taxon-specific
Taxon	4. *Schistocerca americana*	*Schistocerca americana*/all primers
Taxon	5. *Locusta migratoria*	*Locusta migratoria*/all primers
Taxon	6. *Myrmecophilus manni*	*Myrmecophilus manni*/all primers
Taxon	7. *Anabrus simplex*	*Anabrus simplex*/all primers
Sanitized	8. Folmer	All taxa/Folmer
Sanitized	9. Orthoptera-specific	Four taxa/Orthoptera-specific
Sanitized	10.Taxon-specific	Four taxa/taxon-specific

Sanitized data sets contain sequences without stop codons or indels.

analysed using neighbour-joining methods under a Kimura 2-parameter model in MEGA4 (Tamura *et al.* 2007), as typically utilized in barcoding studies. We also calculated haplotype sequence divergence in MEGA4 and determined the number of clusters that would be considered unique species under the DNA barcoding standard of ≥3% nucleotide sequence divergence (Hebert *et al.* 2003).

Results

Taxonomic distribution of numts in Orthoptera

Fig. 1 shows a summary of the results of nucleotide and amino acid data generated for each taxon using each primer type. We found that Folmer primers coamplified numts from all 11 taxa examined and that there were consistent proportions of point mutations across the entire length of numt sequences compared to their orthologues. We found numts containing stop codons in every species examined as well as numts containing indels from every species except *Myrmecophilus manni*. We also found that the degree to which numts were coamplified using the Folmer primers varied widely among taxa. For instance, we identified the greatest amount of numts in *Schistocerca americana* where 49 of 74 haplotypes (66%) were numts, but the least amount of numts in *M. manni* where 3 of 26 haplotypes (12%) were numts. Folmer primers were most effective at amplifying orthologous sequences in *M. manni* where 23 of 26 haplotypes (88%) were orthologues, but they were the least effective at amplifying orthologous sequences in *S. americana* where 18 of 74 haplotypes (24%) were orthologues.

Effect of primer specificity on numt coamplification

In order to understand the relationships between haplotypes sequenced, we reconstructed phylogenies from the primer-type data set (1–3) in a parsimony framework (Fig. 2) and by using Bayesian methods. The Bayesian trees provide topologies largely congruent with parsimony trees and support our conclusions; these topologies are not presented here in an effort to reduce the number and complexity of the figures. In all phylogenies, haplotypes from ensiferan and caeliferan taxa formed monophyletic groups and correctly grouped with reference taxa (labelled as Orthoptera outgroup). Also, haplotypes from each taxon formed a monophyletic clade in all phylogenies with the exception of a few numt haplotypes from *Locusta migratoria* and *Anabrus simplex* that nested inside the *S. americana* clade. These haplotype groupings are not likely to be the result of cross-contamination as each haplotype sequence yielded a correct BLAST result. We also found that some numt haplotypes, which were identified based on the presence of indels, grouped with the putative orthologues in the phylogeny (Fig. 2). We found that this was caused when these numts were aligned and gaps were inserted and treated as missing data in phylogeny reconstruction so that they grouped with the orthologues.

From the data in Fig. 1b and the phylogenies in Fig. 2, we deduce that the effect of increasing primer specificity on eliminating numt coamplification varies unpredictably between taxa. For *S. americana*, we found that neither Orthoptera- nor taxon-specific primers reduced the proportion of numt haplotypes coamplified and were also ineffective at increasing the proportion of sequences that exactly match the mitochondrial orthologue. For *L. migratoria*, Orthoptera-specific primers were effective at reducing the proportion of numt haplotypes and taxon-specific primers were effective at eliminating numt coamplification completely. However, only taxon-specific primers were effective at increasing the proportion of sequences that exactly matched the orthologue. For *M. manni*, we found that more specific primers help reduce numt coamplification, but there was no difference in the amount of numts coamplified between the Orthoptera- and taxon-specific primers. However, more specific primers did help increase the proportion of haplotypes that exactly matched the orthologue in this species. For *A. simplex*, both Orthoptera- and taxon-specific primers were effective at reducing the total proportion of numt haplotypes coamplified and increasing the proportion of haplotypes that exactly matched the orthologue. However, a much smaller difference was observed in the proportion of numts between Orthoptera- and taxon-specific primers than between Orthoptera-specific primers and Folmer primers.

We explored how nucleotide sequence characteristics of haplotypes (stop codons, indels and point mutations) were distributed across the phylogeny by mapping these characteristics onto the phylogeny reconstructed from data set 1 (Fig. 3). As expected, we found large polytomous clades of haplotypes that were categorized as putative orthologues grouping with the orthologue reference sequence. Some clades that lie near to the putative orthologue clade contained haplotypes that were identified as heteroplasmy, but are labelled as putative orthologues on the phylogenies. These heteroplasmy haplotypes had identical amino acid sequences as the orthologue, did not contain indels or stop codons, but did contain >3 nucleotide differences from the orthologue (silent substitutions). Within *S. americana*, there were numt haplotypes that group within heteroplasmy clades. These are possibly numts of heteroplasmy and can be readily identified as numts based on nucleotide and amino acid sequences. In contrast, there were some numt haplotypes within *A. simplex* that did not have stop codons or indels, but had a high number of point mutations that were not silent

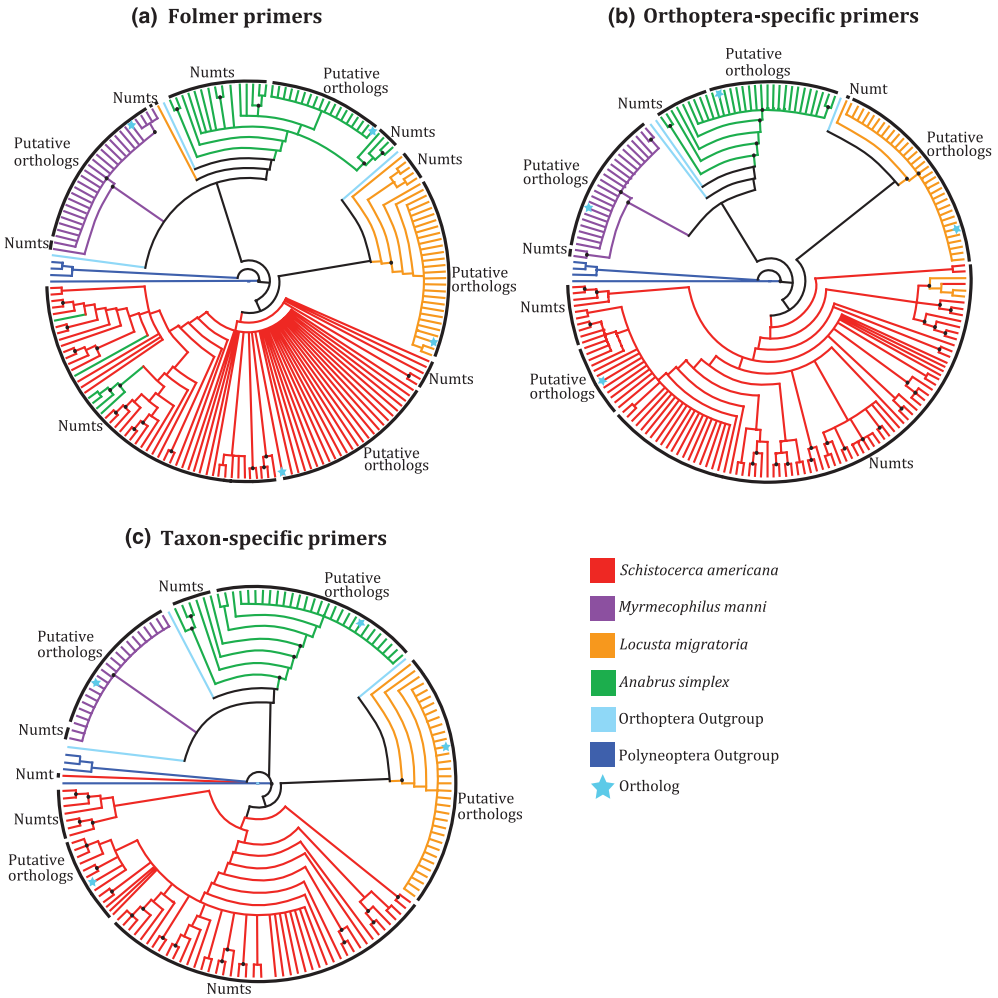
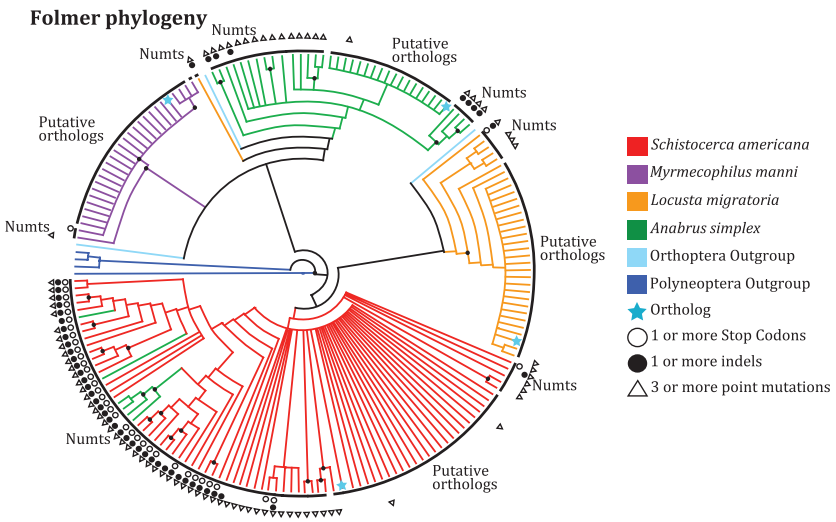


Fig. 2 Phylogenetic Analysis. Phylogenies reconstructed in a parsimony framework from haplotypes generated using primers of varying specificity. Dots on nodes indicate bootstrap value of ≥50. Phylogeny of haplotypes generated using: (a) Folmer primers (189 terminals); strict consensus of 568 MPTs (L=7802, CI=0.23, RI=0.75), (b) Orthoptera-specific primers (196 terminals); strict consensus of 439 MPTs (L=4035, CI=0.35, RI=0.83) and (c) Taxon-specific primers (162 terminals); strict consensus of 6 MPTs (L=3052, CI=0.39, RI=0.86).

Fig. 3 Folmer Phylogeny. Phylogeny reconstructed in a parsimony framework from haplotypes generated using Folmer primers. The number of in-frame stop codons, indels, and point mutations present in each haplotype are mapped on the topology.



substitutions. We identified these as numts based on their nonconserved amino acid sequences and by their distant groupings from the putative orthologues as inferred from the phylogeny.

Pattern of primer-specific numt generation inferred from phylogeny

We reconstructed phylogenies of all four taxa reconstructed in a parsimony framework from haplotypes generated using all three primer sets (data sets 4–7) (Fig. 4) and found congruent results from phylogenies reconstructed using the same data sets in a Bayesian framework (phylogenies not shown). In all phylogenies, the haplotypes of each taxon were recovered as monophyletic except for one divergent numt haplotype in the *L. migratoria* phylogeny. The vast majority of clades were composed of haplotypes generated using multiple primer types, suggesting that different primers coamplified similar types of numts. Furthermore, the distribution of numt haplotypes coamplified by all primer types across the entirety of each tree suggests that all primer types were equally capable of coamplifying numts incorporated into the nuclear genome recently and more anciently.

DNA barcoding analyses

We performed neighbour-joining analyses (as typically used in DNA barcoding) on the sanitized data sets (8–10) in which all haplotypes with indels or stop codons were removed prior to analysis (following the quality control recommendation of Song *et al.* 2008). We show the topologies reconstructed from haplotypes of four species (*S. americana*, *L. migratoria*, *M. manni* and *A. simplex*) in Fig. 5 and report the results of the phylogenies reconstructed from the other species (*P. hystrix*, *E. minutus*, *T. magellanica*, *Trigonopteryx.* sp., *P. variolosa*, *Lentula* sp. and *Lithidiopsis* sp.) in Fig. 1a. We found some haplotypes that we identified as numts from nucleotide and amino acid data exhibited more than 3% sequence divergence from the orthologue and would be classified as unique species under barcoding standards. We also found cases in *A. simplex* where heteroplasmy haplotypes exhibit more than 3% sequence divergence from the orthologue and would also be classified as unique species under barcoding standards, even though the amino acid sequences were identical. When using the Folmer primers, multiple species would be identified from a single individual in all species tested except for *L. migratoria* and *M. manni*. Some of the most extreme cases we found are in

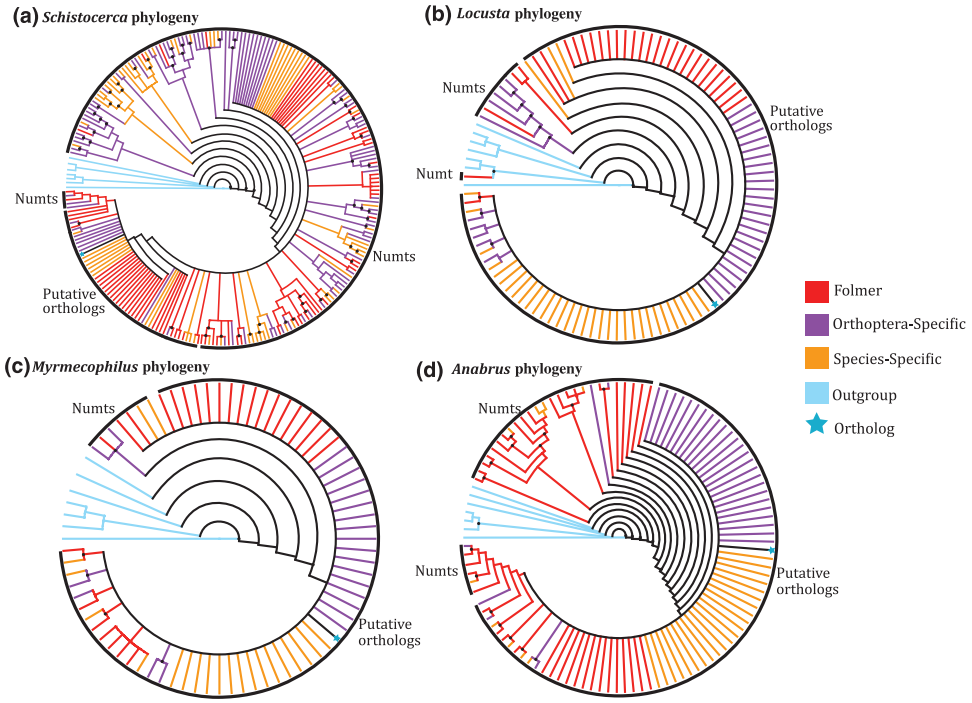


Fig. 4 Taxon Phylogenies. Phylogenies reconstructed in a parsimony framework from haplotypes generated using primers of varying specificity. Dots on nodes indicate bootstrap value of ≥50. Phylogeny of haplotypes generated from: (a) *Schistocerca americana* (242 terminals); strict consensus of 208 MPTs (L=5050, CI=0.29, RI=0.60), (b) *Locusta migratoria* (109 terminals); strict consensus of 11 MPTs (L=1722, CI=0.60, RI=0.68), (c) *Myrmecophilus manni* (79 terminals); strict consensus of 2 MPTs (L=1029, CI=0.64, RI=0.53) and (d) *Anabrus simplex* (112 terminals); strict consensus of 34 MPTs (L=1453, CI=0.47, RI=0.70).

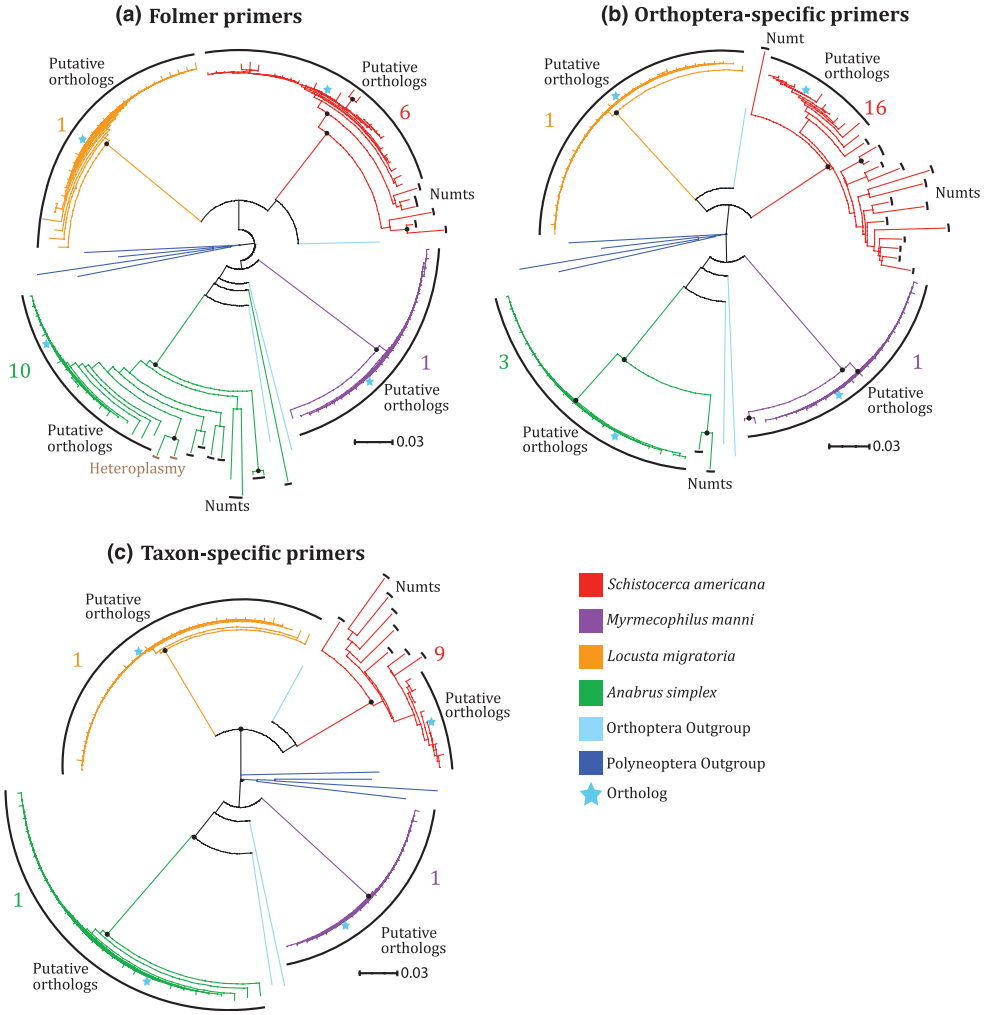


Fig. 5 Barcoding Analysis. Tree topologies reconstructed in a neighbour-joining framework from a subset of haplotypes that lack indels and stop codons. Dots on nodes indicate bootstrap value ≥90. Coloured numbers next to clades represent the number of species that would be identified from a single individual due to coamplification of >3% divergent numts without stop codons or indels. Topology of haplotypes generated using: (a) Folmer primers (131 terminals), (b) Orthoptera-specific primers (113 terminals) and (c) Taxon-specific primers (106 terminals).

S. americana (six unique species), *Lithidiopsis* sp. (nine unique species) and *A. simplex* (ten unique species). Orthoptera-specific primers were successful at reducing misidentification in *A. simplex*, but increased misidentification in *S. americana* (Fig. 5b). Taxon-specific primers are successful at eliminating misidentification of *A. simplex* completely, but still fail to correctly identify *S. americana* as a single species (Fig. 5c).

In addition to these analyses, we directly sequenced the original PCR amplicons (before cloning) in order to simulate what might happen in a standard barcoding analysis. Excerpts of the chromatogram data are presented in Fig. 6. Although the dominant peaks are identical to the orthologous sequence, there is a large amount of background noise. This noise is likely the result of coamplification of nonorthologous sequences. The noise also persists regardless of the specificity of primers used to generate the sequences.

Discussion

Taxonomic distribution of numts in Orthoptera

Our study clearly demonstrates that numts are prevalent in Orthoptera across a wide diversity of lineages. Previous studies have identified numts from a single family, Acrididae (Gellissen *et al.* 1983; Zhang & Hewitt 1996b; Bensasson *et al.* 2000; Song *et al.* 2008), but here we document that the presence of numts is a widespread phenomenon occurring in at least 10 different families belonging to six superfamilies and two suborders (Fig. 1a). Orthopteran species are known to

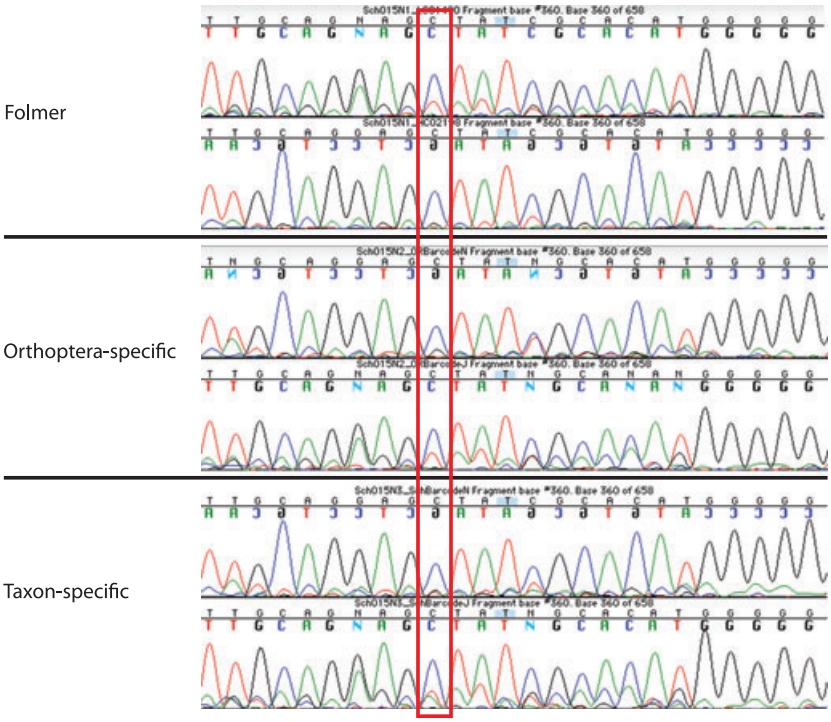


Fig. 6 Chromatogram Data. Chromatogram excerpts of DNA sequences obtained directly from *S. americana* using primers of different specificity without cloning. Background noise present is likely due to coamplification of non-orthologous sequences. The red box highlights one instance where multiple background peaks are present in all sequences regardless of primers used to obtain those sequences.

have some of the largest nuclear genomes of all metazoan lineages (Bensasson *et al*. 2001b) and this observation has been correlated with the particular prevalence of numts in this order (Bensasson *et al*. 2001a). Within Caelifera, we find numts not only in modern lineages such as Acrididae, but also in ancient lineages such as Tridactylidae, Trigonopterygidae and Pneumoridae. We also find numts in two diverse lineages within Ensifera. This observation indicates that the nuclear integration of mtDNA is not an isolated incident, but an ongoing event in the lineage in general going back at least to the Permian (~260 Ma), which is when the two suborders split (Sharov 1968). Our findings suggest that the taxonomic distribution of numts is more widespread than is generally acknowledged and it is still not clear the taxonomic distribution of numts in other insect and arthropod lineages. It is also unclear how PCR-cycling conditions may affect the proportion of numts coamplified within this and other taxonomic groups.

We document that there are many different types of numts in the nuclear genome of a given species. All 11 orthopteran species have more than one numt haplotype and the phylogenetic analyses of these numts show that they can form several distinct clades. This indicates that the past nuclear integration events in a given lineage are preserved in the nuclear genome and that we are able to coamplify these numts with universal primers, although there appears to be a variation in the abundance of different types of numts.

Effect of primer specificity on numt coamplification

Orthoptera-specific primers reduced numt coamplification in three of the four species, but were not effective at eliminating numt coamplification in any species. Taxon-specific primers were more effective at reducing numt coamplification in all species, but were only effective at eliminating numt coamplification in one species, *Locusta migratoria* (Fig. 1b). Increasing primer specificity appears to only be effective at eliminating numt coamplification in lineages with relatively few numts. However, in lineages with many numts, like *Schistocerca americana* and *Anabrus simplex*, more specific primers are only effective at reducing, but not eliminating numt coamplification. These findings are significant for barcoding and phylogenetic analyses alike and show that more specific primers will not guarantee that numts will not be coamplified. More caution needs to be taken when using mitochondrial genes in studies to ensure that the orthologue is amplified.

We find that some numts amplified from multiple species group together on phylogenies (Fig. 2). Precautionary measures in the lab and correct BLAST results for these haplotypes confirm that this was not due to cross-contamination. This phenomenon can be seen when some numt haplotypes from *A. simplex* and *L. migratoria* group with numt haplotypes within the *S. americana* clade (Fig. 2a,b). Although one might argue that this finding could suggest that some numts of ancient origin can be coamplified (meaning that the numts were incorporated

before the divergence of Caelifera and Ensifera), we do not think this is the case. Rather, numts from *A. simplex*, *L. migratoria* and *S. americana* that exhibit high sequence divergence from their respective orthologues will likely group together due to high levels of homoplasy. We do find strong evidence for coamplification of recent numts on our phylogenies and find that these numts can be coamplified even when using the most specific primers. Coamplification of numts of recent origin is likely due to the fact that mutations in the primer-binding regions have not accumulated sufficiently to prevent primer annealing.

It is important to be aware that, although some haplotypes within our designation of the putative orthologue clades have more than three point mutations, these haplotypes are identified as heteroplasmy because they have identical amino acid sequences as the orthologue (Fig. 3). The varying degree to which numts are present in different lineages suggests that numts may be incorporated many times and be evolving at different rates making it difficult to identify all numts present within an organism. The presence of heteroplasmy and numts of heteroplasmy make correct identification of numts even more difficult.

Pattern of primer-specific numt coamplification inferred from phylogeny

We set out to investigate whether different primer sets would each capture the entire numt diversity as revealed in this study. We found that most clades on all four phylogenies presented are composed of haplotypes amplified using different primer sets, implying that these distinct sets of primers coamplify the diversity of numts (Fig. 4), although there are still some cases of preferential coamplification. These findings are important in that they demonstrate that a single set of primers can coamplify most, if not all, the numt diversity.

Implications for DNA barcoding analyses

Ideally, using DNA barcoding methodology, one will sequence the Folmer region from an individual and diagnose it as one species based on its similarity to known barcode sequences in a database. However, our analyses show (in agreement with Song *et al.* 2008) that a single individual exhibits sufficient diversity among the numts and heteroplasmy haplotypes present, that the individual might mistakenly be diagnosed as multiple species (Fig. 5). Furthermore, we show that the suggestion that increased primer specificity may eliminate numt coamplification cannot be substantiated. We recognize that designing more specific primers for each species would also be very expensive (see Cameron *et al.* 2006 for a discussion of the costs associated with DNA barcoding) and would require some prior identification of an organism eliminating the utility of using DNA barcoding for species identification.

DNA barcoders typically do not clone PCR amplicons from individuals in order to generate their sequence data. However, we argue that there is sufficient background noise within chromatogram data (Fig. 6) to suggest that numts could be preferentially amplified and sequenced by chance, even without cloning. In fact, Song *et al.* (2008) were able to preferentially amplify and sequence numts from several crayfish specimens using conventional PCR methods without cloning. Furthermore, increasing primer specificity does not seem to reduce the amount of background noise present, and therefore does not reduce the chance of mistakenly using a numt sequence as a barcode.

Numts are a major obstacle for single-gene analyses such as DNA barcoding – especially when increasing primer specificity is ineffective at eliminating numts from certain taxonomic groups. We recognize that any phylogenetic analysis using a single gene would encounter similar problems, but that is precisely one of the reasons why modern systematists reject such single-gene studies. As in the case of *S. americana*, numts are pervasive and even numts without characteristic mutations can be coamplified when using taxon-specific primers. We demonstrate that numts are prevalent within 11 diverse lineages of Orthoptera and may be just as prevalent in other taxonomic groups. The more we search for numts, the more common they appear to be (Richly & Leister 2004; Antunes & Ramos 2005) and the presence of numts may be more of a rule than an exception.

Categorizing putative orthologues, numts and heteroplasmy before analysis can be difficult and definitions of these terms can be convoluted. We emphasize that it can be especially difficult to precisely distinguish some numts that do not have stop codons and indels from heteroplasmy. In order to make this distinction, it is useful to examine the amino acid translation data. This method is not used in typical barcoding analyses, and we suggest that current barcoding methods incorporate the use of amino acid sequences. However, we recognize that this approach is only applicable to protein-coding genes, and previous studies have shown that identifying numts or heteroplasmy in ribosomal genes is difficult (Olson & Yoder 2002).

The high number of numts identified within certain individuals and the broad presence of numts in individuals within Orthoptera are surprising. It is not unimaginable that numts are just as prevalent in other lineages and that increasing primer specificity will similarly not help DNA barcoding overcome the problems of species misidentification or overestimating the number of species.

We suggest that more studies need to be performed in order to investigate numt distribution across a broader taxonomic sampling and to assess the extent to which numt coamplification influences the results of DNA barcoding analyses. It is time that the proponents of DNA barcoding recognize the large impact numt coamplification can have on species misidentification (Hebert *et al.* 2004a) and seek for ways to eliminate misleading results due to numts.

Acknowledgements

This work was funded by National Science Foundation Grants EF-0531665 to MFW and DEB-0816962 to HS and MFW and by the Office of Research and Creative Activities at Brigham Young University. We thank Kevin Hiatt for his help with data collection, analysis and manuscript preparation. We thank Nathan Sheffield for writing perl scripts, which made these analyses more feasible. We also thank three anonymous reviewers for their suggestions to improve this manuscript.

References

Antunes A, Ramos MJ (2005) Discovery of a large number of previously unrecognized mitochondrial pseudogenes in fish genomes. *Genomics*, **86**, 708–717.

Arctander P (1995) Comparison of a mitochondrial gene and a corresponding nuclear pseudogene. *Proceedings of the Royal Society of London Society of London, Series B: Biological Sciences*, **262**, 13–19.

Benesh DP, Hasu T, Suomalainen LR, Valtonen ET, Tiirola M (2006) Reliability of mitochondrial DNA in an acanthocephalan: the problem of pseudogenes. *International Journal for Parasitology*, **36**, 247–254.

Bensasson D, Zhang DX, Hewitt GM (2000) Frequent assimilation of mitochondrial DNA by grasshopper nuclear genomes. *Molecular Biology and Evolution*, **17**, 406–415.

Bensasson D, Zhang D, Hartl DL, Hewitt GM (2001a) Mitochondrial pseudogenes: evolution's misplaced witnesses. *Trends in Ecology & Evolution*, **16**, 314–321.

Bensasson D, Petrov DA, Zhang DX, Hartl DL, Hewitt GM (2001b) Genomic gigantism: DNA loss is slow in mountain grasshoppers. *Molecular Biology and Evolution*, **18**, 246–253.

Cameron S, Rubinoff D, Will K (2006) Who will actually use DNA barcoding and what will it cost? *Systematic Biology*, **55**, 844–847.

Cameron SL, Lambkin CL, Barker SC, Whiting MF (2007) A mitochondrial genome phylogeny of Diptera: whole genome sequence data accurately resolve relationships over broad timescales with high precision. *Systematic Entomology*, **32**, 40–59.

Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, **32**, 1792–1797.

Fenn JD, Cameron SL, Whiting MF (2007) The complete mitochondrial genome sequence of the Mormon cricket (Anabrus simplex: Tettigoniidae: Orthoptera) and an analysis of control region variability. *Insect Molecular Biology*, **16**, 239–252.

Fenn JD, Song H, Cameron SL, Whiting MF (2008) A preliminary mitochondrial genome phylogeny of Orthoptera (Insecta) and approaches to maximizing phylogenetic signal found within mitochondrial genome data. *Molecular Phylogenetics and Evolution*, **49**, 59–68.

Flook PK, Rowell CH, Gellissen G (1995) The sequence, organization, and evolution of the Locusta migratoria mitochondrial genome. *Journal of Molecular Evolution*, **41**, 928–941.

Folmer O, Black M, Hoeh W, Lutz R, Vrijenhoek R (1994) DNA primers for amplification of mitochondrial cytochrome c oxidase subunit I from diverse metazoan invertebrates. *Molecular Marine Biology & Biotechnology*, **3**, 294–299.

Frey JE, Frey B (2004) Origin of intra-individual variation in PCR-amplified mitochondrial cytochrome oxidase I of Thrips tabaci (Thysanoptera: Thripidae): mitochondrial heteroplasmy or nuclear integration? *Hereditas*, **140**, 92–98.

Gellissen G, Bradfield JY, White BN, Wyatt GR (1983) Mitochondrial DNA sequences in the nuclear genome of a locust. *Nature*, **24**, 774–786.

Goloboff PA, Farris JS, Nixon KC (2008) TNT, a free program for phylogenetic analysis. *Cladistics*, **24**, 774–786.

Gomez A, Wright PJ, Lunt DH *et al.* (2007) Mating trials validate the use of DNA barcoding to reveal cryptic speciation of a marine bryozoan taxon. *Proceedings of the Royal Society of London, Series B: Biological Sciences*, **274**, 199–207.

Hebert PD, Gregory TR (2005) The promise of DNA barcoding for taxonomy. *Systematic Biology*, **54**, 852–859.

Hebert PD, Cywinska A, Ball SL, deWaard JR (2003) Biological identifications through DNA barcodes. *Proceedings of the Royal Society of London, Series B: Biological Sciences*, **270**, 313–321.

Hebert PD, Penton EH, Burns JM, Janzen DH, Hallwachs W (2004a) Ten species in one: DNA barcoding reveals cryptic species in the neotropical skipper butterfly Astraptes fulgerator. *Proceedings of the National Academy of Sciences, USA*, **101**, 14812–14817.

Hebert PD, Stoeckle MY, Zemlak TS, Francis CM (2004b) Identification of Birds through DNA Barcodes. *Public Library of Science Biology*, **2**, e312.

Hughes S, Moody A (2007) *PCR, illustrated edition*. Scion Publishing, Bloxham, UK.

Leroux C, Issel CJ, Montelaro RC (1997) Novel and dynamic evolution of equine infectious anemia virus genomic quasispecies associated with sequential disease cycles in an experimentally infected pony. *Journal of Virology*, **71**, 9627–9639.

Lopez JV, Yuhki N, Masuda R, Modi W, O'Brien SJ (1994) Numt, a recent transfer and tandem amplification of mitochondrial DNA to the nuclear genome of the domestic cat. *Journal of Molecular Evolution*, **39**, 174–190.

Olson LE, Yoder AD (2002) Using secondary structure to identify ribosomal numts: cautionary examples from the human genome. *Molecular Biology and Evolution*, **19**, 93–100.

Ratnasingham S, Hebert PDN (2007) *BOLD Identification System*. Available at: http://www.barcodinglife.org/views/login.php [Last accessed 1 July 2009]

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

Ronquist F, Huelsenbeck JP (2003) MRBAYES 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics*, **19**, 1572–1574.

Rubinoff D, Cameron S, Will K (2006) A genomic perspective on the shortcomings of mitochondrial DNA for ''barcoding'' identification. *Journal of Heredity*, **97**, 581–594.

Sharov AG (1968) Phylogeny of the Orthopteroidea. *Akademiya Nauk SSSR Trudy Paleontologicheskogo Instituta*, **118**, 1–216.

Sheffield NC, Song H, Cameron SL, Whiting MF (2008) A comparative analysis of mitochondrial genomes in Coleoptera (Arthropoda: Insecta) and genome descriptions of six new beetles. *Molecular Biology & Evolution*, **25**, 2499–2509.

Song H, Buhay JE, Whiting MF, Crandall KA (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 USA*, **105**, 13486–13491.

Sorenson MD, Quinn TW (1998) Numts: a challenge for avian systematics and population biology. *Auk*, **115**, 214–221.

Tamura K, Dudley J, Nei M, Kumar S (2007) MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. *Molecular Biology & Evolution*, **24**, 1596–1599.

Tedersoo L, Jairus T, Horton BM *et al*. (2008) Strong host preference of ectomycorrhizal fungi in a Tasmanian wet sclerophyll forest as revealed by DNA barcoding and taxon-specific primers. *New Phytologist*, **180**, 479–490.

Vences M, Thomas M, van der Meijden A, Chiari Y, Vieites DR (2005) Comparative performance of the 16S rRNA gene in DNA barcoding of amphibians. *Frontiers in Zoology*, **2**, 5.

Weissensteiner T, Griffin HG, Griffin AM (2004) *PCR Technology: Current Innovations*, 2nd edn. CRC Press, Boca Raton, FL.

Williams ST, Knowlton N (2001) Mitochondrial Pseudogenes Are Pervasive and Often Insidious in the Snapping Shrimp Genus Alpheus. *Molecular Biology & Evolution*, **18**, 1484–1493.

Witt JD, Threloff DL, Hebert PD (2006) DNA barcoding reveals extraordinary cryptic diversity in an amphipod genus: implications for desert spring conservation. *Molecular Ecology*, **15**, 3073–3082.

Zhang DX, Hewitt GM (1996a) Highly conserved nuclear copies of the mitochondrial control region in the desert locust Schistocerca gregaria: some implications for population studies. *Molecular Ecology*, **5**, 295–300.

Zhang DX, Hewitt GM (1996b) Nuclear integrations: challenges for mitochondrial DNA markers. *Trends in Ecology & Evolution*, **11**, 247–251.

Zhang DX, Hewitt GM (1997) Insect mitochondrial control region: a review of its structure, evolution and usefulness in evolutionary studies. *Biochemical Systematics and Ecology*, **25**, 99–120.