JOURNAL OF CRUSTACEAN BIOLOGY, 29(1): 96–110, 2009

"COI-LIKE" SEQUENCES ARE BECOMING PROBLEMATIC IN MOLECULAR SYSTEMATIC AND DNA BARCODING STUDIES

Jennifer E. Buhay

University of South Carolina, Belle W. Baruch Institute of Marine and Coastal Sciences, Columbia, South Carolina 29208, U.S.A. and IAP World Services, NOAA, 705 Convent Avenue, Pascagoula, Mississippi 39567, U.S.A. (jenbuhay@gmail.com)

ABSTRACT

The cytochrome c oxidase subunit I (COI) gene plays a pivotal role in a global effort to document biodiversity and continues to be a gene of choice in phylogenetic and phylogeographic studies. Due to increased attention on this gene as a species' barcode, quality control and sequence homology issues are re-emerging. Taylor and Knouft (2006) attempted to examine gonopod morphology in light of the subgeneric classification scheme within the freshwater crayfish genus *Orconectes* using COI sequences. However, their erroneous analyses were not only based on supposed mitochondrial sequences but also incorporated many questionable sequences due to the possible presence of numts and manual editing or sequencing errors. In fact, 22 of the 86 sequences were flagged as "COI-like" by GenBank due to the presence of stop codons and indels in what should be the open reading frame of a conservative protein-coding gene. A subsequent search of "COI-like" accessions in GenBank turned up a multitude of taxa across Crustacea from published and unpublished studies thereby warranting this illustrated discussion about quality control, pseudogenes, and sequence composition.

KEY WORDS: cytochrome c oxidase subunit I, molecular taxonomy, numt, protein-coding gene, pseudogene

DOI: 10.1651/08-3020.1

Downloaded from https://academic.oup.com/jcb/article/29/1/96/2548098 by guest on 23 April 2024

INTRODUCTION

The mitochondrial protein-coding gene, cytochrome c oxidase subunit I (COI), is a widely accepted marker for molecular identification to the species level across diverse taxa (examples of large scale projects: springtails, Hogg and Hebert, 2004; butterflies, Hebert et al., 2004a; birds, Hebert et al., 2004b; fishes, Ward et al., 2005; crustaceans, Costa et al., 2007). Approximately 700 nucleotides of COI molecular sequence can be used to query large COI datasets to help determine species' identity of unknown samples, a method known as "barcoding" (Hebert et al., 2003a, b). It is now possible to submit a sequence of unknown origin to the Consortium for the Barcoding of Life website (BOL: http://www.barcodinglife.org/views/idrequest.php) and within seconds, either the name of the species (if there is a reference sequence from that species accessioned into the database) or the name of the closest related taxa (if there is no reference sequence for species' comparison in the database) will appear on the query screen along with percent COI sequence similarity of the top 20 species' matches.

While the method of matching unknown molecular sequences to an online database is not new (for example, similar queries can be done with the Blast Search option in GenBank: http://www.ncbi.nlm.nih.gov/blast/Blast.cgi), barcoding and similar methods using "molecular taxonomy" (Dayrat, 2005) such as "DNA Surveillance" (http://www.cebl.auckland.ac.nz:9000/) are highly dependent on 1) accurate identification of species in the reference database for comparison (http://www.barcoding.si.edu/DNABarcoding.htm) and 2) accurate molecular sequences. The backbone of the BOL relies heavily on the gathering of molecular data from preserved and curated museum specimens (vouchers), representing a collaboration between members of the BOL Consortium including among others, the National Museum of Natural History (Smithsonian Institution: www.si.edu), and the National Institutes of Health's online repository GenBank (http://www.ncbi.nlm.nih.gov/Genbank).

Ultimately, the responsibility of accurate identification of animal specimens rests with the researchers who determine species' identity using a host of morphological characters, and hopefully, these researchers create photo-documentation and/or accession museum vouchers to enable cross-checking of their species' diagnosis by others. While some physical characters are "better than others" to place a species' label on an organism, morphologically cryptic species remain cryptic species without examination of non-morphological features such as genetic sequence data. Morphological features are sometimes useless and misleading when trying to determine the species' identity of various larval stages, females for which keys are virtually non-existent in many animal groups, or specimens mutilated from intensive collection methods (such as trawls). Similarly, diagnostic morphological characters are sometimes missed, such as internal anatomy which can be difficult to dissect out or color patterns which are lost in the preservation process. These are all issues driving traditional taxonomists to move beyond the strict diagnoses of species, our units of biodiversity, using solely morphological information and into the realm of modern molecular approaches for more robust species diagnoses (Paquin and Hedin, 2004; Sites and Marshall, 2004).

The ease and low cost of gathering molecular data has made it rather commonplace for systematic biologists to sequence a standard set of genes for phylogenetic studies, particularly mitochondrial genes with "universal" primers,

e.g., primers designed to amplify the same gene across most animal species such as Folmer et al., 1994 primers for COI. DNA extractions, PCR reactions, and cycle-sequencing methods are routine lab exercises in most university biochemistry and genetics courses, even in some high schools. However, analytical training in molecular evolution using nucleotide sequences does not always go hand in hand with the lab methods of gathering the data. The price of faulty sequence data and evolutionary analyses has so far been considered to be negligible, but a number of recent publications have revealed some snowballing problems that need to be corrected in past studies and prevented in future work.

A discussion of the fundamentals with molecular data may seem like old hat to most, but there are certainly important and underlying issues that seem to plague the inexperienced time and time again in the arena of crustacean molecular genetic data as it applies to systematic, phylogenetic, and barcoding approaches. For purposes of this discussion, I chose to focus on the mitochondrial COI gene because of the sheer number of issues flagged in GenBank and its critical role in the BOL project to document global biodiversity. BOL is also closely networked with similar initiatives, including Census of Marine Life (CoML: http://www.comlsecretariat.org/) and the International Barcoding of Life project (http://www.dnabarcoding.org). Until BOL and/or GenBank begins requiring original unedited chromatogram files from automated sequencers (which is a consideration of both organizations now) and/or phred scores (www.phrap.com/phred; program which assigns quality control scores to nucleotide base calls using original trace files from automated sequencers), the scientific community must judge the quality of the data and findings on an individual basis. Currently, it appears that quality control of genetic data at the researcher level is taking a backseat to quick authorship.

COI is a protein-coding gene, and as such, has an open reading frame. Open reading frames do not include stop codons or indels leading to gaps in the alignment which disrupt the translation of the DNA sequence into amino acids. While this may seem obvious to many, there are consistently crustacean sequences accessioned to GenBank for the COI gene that have stop codons present in the open reading frame or indels leading to stop codons (Table 1). These ''COI-like'' sequences range throughout Crustacea, including krill, crab, amphipod, crayfish, squat lobster, shrimp, isopod, barnacle, and copepod. GenBank staff now flag these questionable sequences as ''COI-like'' because the amino acid translations include stop codons and interruptions of the reading frame and it is not possible to determine if it is a sequence editing error (such as an accidental deletion of a base in the sequence or misreading a base in the chromatogram) or a pseudogene without the original chromatograms to examine. GenBank staff perform an amino acid translation (see the invertebrate table: http://www.ncbi.nlm.nih.gov/Taxonomy/Utils/wprintgc.cgi#SG5) when they receive the sequences and if translation is not possible due to indels and stop codons (TAA and TAG), they contact the person who submitted the sequences before making the data available to the public. If the submission is not corrected by the original submitter, then the translation is not possible and the accession is marked with ''COI-like.''

While some ''COI-like'' sequences may represent simple errors in manual editing and lack of quality control, others may actually be nuclear copies of mitochondrial derived genes (numts: Lopez et al., 1994) that crossed into the nuclear genome and became non-functional and therefore, non-coding. Pseudogenes can and do accumulate mutations, which may disrupt the open reading frame and persist due to the lack of evolutionary constraints. Numts are not new to molecular genetic studies (Zhang and Hewitt, 1996; Bensasson et al., 2001; Richly and Leister, 2004). Numts, behaving as junk DNA, commonly occur when portions of the mitochondrial genome jump into the nuclear genome by recombination or crossing over mechanisms. Schneider-Broussard and Neigel (1997) reported the first case of numts in crustaceans by examining length and composition in comparison to mitochondrial sequences. Some species have rampant numts throughout their nuclear genome (shrimp: Williams and Knowlton, 2001; bird: Sorenson and Quinn, 1998, cat: Cracraft et al., 1998; and primate: Hazkani-Covo and Graur, 2007). Numts have previously been reported in Australian crayfish (Nguyen et al., 2002) and COI numts are especially problematic and insidious for some cave crayfish species (Song et al., 2008).

Numt sequences confound phylogenetic analyses (see review by Arctander, 1995; Sorenson and Quinn, 1998) because the basic assumption of a comparison of homologous DNA is violated when numts are present. An understanding of molecular genetic data (in this case, protein coding sequence data) provides the first clues as to the presence of numts in a dataset as numts tend to have their open reading frame disrupted through mutation – such mutations are normally eliminated in functional copies of mtDNA because such mutations disrupt the function of the gene and result in an evolutionarily selective disadvantage; but in numts, such mutations are tolerated because they are nonfunctional copies of mtDNA in the nuclear genome. In the case of barcoding, the numt sequences of COI can be highly divergent from the actual COI sequences which presents a major problem because identification of species is based on sequence similarity (Song et al., 2008). Additionally, high genetic divergences are used to indicate possible new species that may be nested within species' complexes. This can be problematic when extremely divergent numt sequences are mistaken for the presence of cryptic species. The suggested threshold to detect new candidate species is 10× the average intraspecific sequence difference (K2P distance) of the animal group (Hebert et al., 2004b). For example, birds have 0.27% intraspecific difference and therefore a 2.7% threshold difference to separate distinct species (Hebert et al., 2004b). Numts can and certainly do accumulate mutations that exceed 10× intraspecific differences (Song et al., 2008).

In contrast from highly divergent sequences, one should recognize that younger species and species with hybrid zones will not exhibit sufficient variation to be determined as distinctly different using only barcodes. Regardless of the threshold value and the criticisms about using a percent sequence difference to indicate possible new species (Moritz

Downloaded from https://academic.oup.com/jcb/article/29/1/96/2548098 by guest on 23 April 2024

Table 1. Results of CoreNucleotide online search for sequences accessioned as COI by crustacean researchers but flagged by GenBank staff as "COI-like." Keywords for search were "similar cytochrome I crustacea" and "COI-like." Indels and stop codons were compared against similar published mitochondrial genomes. Note: Only one (of sometimes several) "coi-like" sequence of a species is presented.

Taxa	Data source	GenBank no.	Reading frame problem
Liberonautes rubigimanus	Plachetzki and Cumberlidge (unpublished)	AF399978	1 bp deletion at 442; Stop codons at 466, 472
Seychellum alluaudi	Plachetzki and Cumberlidge (unpublished)	AF399973	1 bp insertion at 28; Stop codons at 88, 103, 235, 427, 436, 445, 469, 475
Liberonautes chaperi	Plachetzki and Cumberlidge (unpublished)	AF399977	1 bp deletion at 409; Stop codons at base 415, 433
Eriocheir sinensis	Tang et al. (2003)	AF317334	1 bp insertion at 504; Stop codon at 572
Eriocheir japonica	Tang et al. (2003)	AF317329	1 bp deletion at 259; Stop codons at 268, 271, 313
Chiromantes haemotocheir	Tang et al. (2003)	AF317342	1 bp deletion at 506; Stop codons at 521, 545
Farfantepenaeus subtilis	de Francisco and Galetti Jr. (unpublished)	AY344198	Not COI; stop codons in every reading frame shift
Virilastacus retamali	Crandall and Rudolph (2007)	EF599152	Sloppy 5′ end; 1 bp insertion at 78; Stop codons at 112, 142, 274, 466, 487, 499
Virilastacus araucanius	Crandall and Rudolph (2007)	EF599156	No stop codons, but sequence is odd compared to reference
Virilastacus rucapihuelensis	Crandall and Rudolph (2007)	EF599149	Sloppy 5′ end; sloppy 3′ end
Samastacus spinifrons	Crandall and Rudolph (2007)	EF599159	1 bp deletion at 395; Stop codons at 437, 458, 476, 482, 593
Munidopsis quadrata	Costa et al. (2007)	DQ882093	No stop codons, but sequence is odd compared to reference
Euxinia maesticus	Costa et al. (2007)	DQ889170	No stop codons, but sequence is odd compared to reference
Mytilocypris ambiguosa	Costa et al. (2007)	DQ889162	No stop codons, but sequence is odd compared to reference
Exosphaeroma sp.	Costa et al. (2007)	DQ889151	No stop codons, but sequence is odd compared to reference
Diporeia hoyi	Costa et al. (2007)	DQ889144	No stop codons, but sequence is odd compared to reference
Orconectes propinquus	Costa et al. (2007)	DQ889165	No stop codons, but sequence is odd compared to reference
Mysis americana	Costa et al. (2007)	DQ889161	No stop codons, but sequence is odd compared to reference
Caligus sp. 2	Andrews et al. (unpublished)	EF452643	Stop codons at 324, 342
Thysanoessa raschii	Karlbom (unpublished)	EF015495	1 bp insertion at 294; Stop codons at 424, 523
Gennadas brevirostris	Karlbom (unpublished)	EF015494	1 bp insertion at 294; Stop codon at 424
Thysanopoda sp.	Karlbom (unpublished)	EF015493	1 bp insertion at 294; Stop codons at 403, 556
Sergestes arctica	Karlbom (unpublished)	EF015492	1 bp insertion at 294; Stop codons at 403, 424
Gnathophausia zoea	Karlbom (unpublished)	EF015490	3 bp deletion at 316; Stop codon at 490
Chiridus armatus	Vestheim et al. (2005)	AY660604	Sloppy 5′ end; 1 bp deletion at 596; Stop codon at 599; Sloppy 3′ end
Scopelocheirus schellenbergi	Blankenship and Yayanos (2005)	AY830432	Sloppy 5′ end; Stop codon at 85
Paramysis kroyeri	Cristescu and Hebert (unpublished)	AY529037	Stop codon at 292
Chthamalus dalli	Wares and Currier (unpublished)	AY795367	Not COI; stop codons in every reading frame shift
Fenneropenaeus indicus	Querci et al. (unpublished)	AY395245	Not COI; stop codons in every reading frame shift
Engaeus cisternarius	Hansen and Smolenski (2002)	AF482494	1 bp insertion at 30; Stop codons at 95, 122, 128, 146, 149, 179, 215, 230, 245, 254, 293, 296, 305, 320, 350, 383, 431, 440, 449, 458, 482, 488, 560, 584
Parastacoides tasmanicus inermis	Hansen and Smolenski (2002)	AF482493	Sloppy 5′ end; 1 bp insertion at 89; Stop codons at 117, 192, 441, 600
Parastacoides tasmanicus tasmanicus	Hansen and Smolenski (2002)	AF482492	Sloppy 5′ end; 1 bp insertion at 30; Stop codons at 95; 128, 146, 149, 161, 176, 209, 215, 230, 245, 254, 281, 305, 350, 431, 482, 488
Cambarus diogenes	Taylor and Knouft (2006)	AY701191	1 bp deletion around 1491
Orconectes meeki meeki	Taylor and Knouft (2006)	AY701213	3 bp insertion around 1075
Orconectes indianensis	Taylor and Knouft (2006)	AY701198	Extra T at 1448 and lots of ambiguity codes at 3′ end
Orconectes kentuckiensis	Taylor and Hardman (2002); Taylor and Knouft (2006)	AF474369	Same sequence as AY701196 probably accessioned again; 1 bp deletion at 1474
Orconectes acares	Taylor and Knouft (2006)	AY701227	Extra C at 1478
Orconectes barrenensis	Taylor and Knouft (2006)	AY701228	Extra C at 1480
Orconectes cristavarius	Taylor and Knouft (2006)	AY701230	Extra A at 1469
Orconectes forceps	Taylor and Knouft (2006)	AY701231	Mostly ambiguity codes (sloppy sequence); 1 bp deletion at 938; Stop codons at 988, 964
Orconectes macrus	Taylor and Knouft (2006)	AY701236	Extra G at 1499
Orconectes menae	Taylor and Knouft (2006)	AY701238	1 bp deletion at 1512
Orconectes neglectus chaenodactylus	Taylor and Knouft (2006)	AY701240	Sloppy 5′ end
Orconectes williamsi	Taylor and Knouft (2006)	AY701252	1 bp deletion at 1494
Orconectes chickasawae	Taylor and Knouft (2006)	AY701216	1 bp insertion at 1448
Orconectes cooperi	Taylor and Knouft (2006)	AY701218	1 bp deletion at 42; Stop codons at 76, 79, 91, 115, 133, 160, 163, 166, 184, 187, 199, 202, 214, 239, 253, 268, 283, 292, 319, 331, 343, 388, 421, 427, 448, 469, 478, 487, 514, 520, 598, 637, 730, 751, 805, 823, 877, 889, 964, 1018, 1051, 1060, 1066, 1111, 1114, 1141, 1207, 1222, 1285, 1291, 1378, 1381, 1390, 1426
Orconectes holti	Taylor and Knouft (2006)	AY701225	1 bp deletion at 53; Stop codons at 76, 79, 91, 97, 115, 133, 163, 184, 199, 202, 214, 238, 247, 253, 268, 283, 292, 331, 421, 427, 469, 478, 487, 520, 598, 622, 730, 751, 766, 805, 811, 823, 841, 877, 883, 889, 964, 1042, 1051, 1066, 1111, 1114, 1141, 1207, 1222, 1255, 1378, 1381, 1390, 1399, 1425

Downloaded from https://academic.oup.com/jcb/article/29/1/96/2548098 by guest on 23 April 2024

Table 1. Continued.

Taxa	Data source	GenBank no.	Reading frame problem
Orconectes immunis	Taylor and Knouft (2006)	AY701220	2 inserted unknown bases at 1479
Orconectes validus	Taylor and Knouft (2006)	AY721593	Extra C at 21; Stop codons at 31, 157, 289, 337, 502, 784, 949, 958, 991, 997, 1072, 1240, 1336, 1435, 1444
Orconectes compressus	Taylor and Knouft (2006)	AY701217	Extra T at 1456; extra A at 1474
Orconectes australis	Taylor and Knouft (2006)	AY701204	2 bp deletion at 52; 1 bp deletion at 1460; Stop codons at 127, 154, 286, 334, 376
Orconectes inermis	Taylor and Knouft (2006)	AY701201	1 bp deletion at 6; Stop codons at 76, 79, 91, 97, 115, 124, 184, 202, 214, 238, 268, 283, 292, 319, 331, 343, 346, 358, 388, 448, 469, 478, 487, 511, 598, 637, 730, 751, 766, 805, 811, 823, 841, 883, 889, 922, 964, 1018, 1042, 1066, 1111, 1114, 1207, 1222, 1255, 1285, 1291, 1378, 1381, 1390, 1399, 1426
Orconectes pellucidus	Taylor and Knouft (2006)	AY701203	Sloppy 5′ end
Procambarus acutus	Taylor and Hardman (2002); Taylor and Knouft (2006)	AY701194	Same sequence as AF474366 probably accessioned again; extra T at 1508

and Cicero, 2004; Meyer and Paulay, 2005), numts and quality control have been largely ignored issues that have major ramifications not just to barcoding approaches, but to everyone who uses free publicly-available "COI-like" data for their systematic revisions, phylogeographic studies, and genetic diversity estimates. Also, cases of ribosomal DNA numts do exist (Schneider-Broussard et al., 1998; Olsen and Yoder, 2002; Nguyen et al., 2002) but they are much harder to detect because they are structural genes, not protein-coding genes with obvious stop codons.

When numts are present, many nuclear copies of the gene along with the mtDNA gene copies are amplified during PCR and it is highly likely that a forward primer amplifies a different gene portion than a reverse primer. PCR gels have bright beautiful bands corresponding to the assumed size of the gene fragment, but sequencing reactions run forward and reverse directions separately. When these forward and reverse contigs are assembled in any manual editing program (such as Sequencher: Gene Codes Corporation; BioEdit: Hall, 1999; and MacClade: Maddison and Maddison, 2002), the contigs sometimes do not match, but by using a "reference" sequence (which can be marked in Sequencher for example), one can guess (using ambiguity codes) what the overlapping region sequence is, however, it is not a sequence that actually exists if the contigs are from two different regions, i.e., one contig from the mitochondrial genome and the other from the nuclear genome. The presence of numts and COI will most times result in messy chromatograms in both directions which are practically unreadable because many different PCR products are sequenced simultaneously (Fig. 1). Additionally, sections of the chromatogram may be readable but they may not be the target gene (Fig. 2) and therefore, the data should be discarded rather than trying to read around the messy sections using ambiguity codes.

By cloning the PCR products, it is possible to isolate the COI sequence from the numt sequences, however, cloning can be expensive and requires a different skillset and training. If the numts are insidious throughout the nuclear genome, it may be difficult to pin down the COI sequence. An example from *Orconectes barri* (Buhay and Crandall, 2008) provides a great illustration of the obvious differences between two sequences amplified by the Folmer et al. (1994) primers for COI cloned from the same crayfish individual: one clone is the target gene COI and the other clone is a COI numt. COI has an opening reading frame and by using a reference sequence (in the case of crayfish, a suitable reference is the COI sequence from the completed mitochondrial genome project of *Cherax destructor* Clark, 1936, GenBank NC_011243), the alignment reveals a high degree of similarity between the three crayfish sequences (*C. destructor* COI, *O. barri* COI, and *O. barri* numt) at the 5′ end (Fig. 3). COI is a very conserved gene with respect to amino acids and despite differences in the nucleotides (particularly at third position), those differences often do not translate into amino acid changes. But in cases where numts

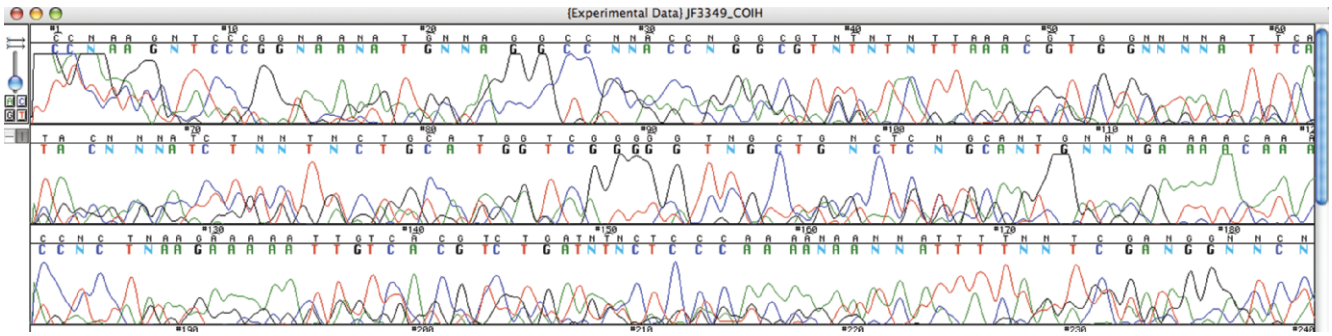

Fig. 1. Screenshot of sloppy COI chromatogram in Sequencher. This is an example of trying to sequence a PCR product that includes numts and COI.

Downloaded from https://academic.oup.com/jcb/article/29/1/96/2548098 by guest on 23 April 2024

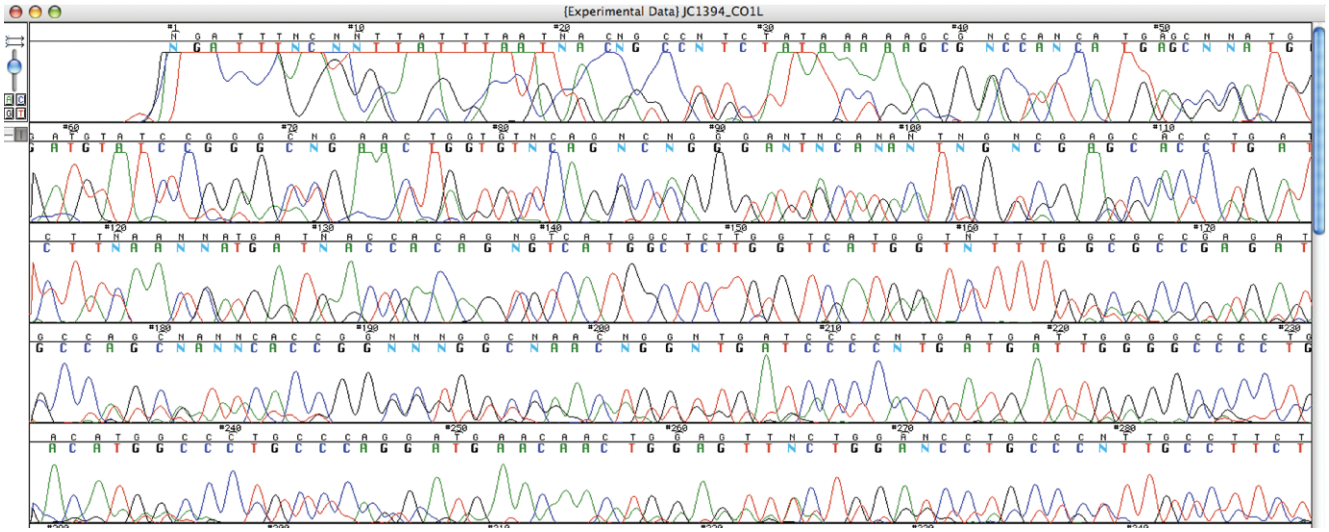

Fig. 2. Screenshot of messy but in some places readable COI chromatogram in Sequencher. Despite the readable sections, messy sequences should be discarded.

are present, there are obvious indels and stop codons which are not present in COI sequences. By comparing the same gene region of a numt and COI isolated from the same individual, there are shifts in the reading frame of the numt which lead to stop codons downstream (Fig. 4). The COI and the numt regions were amplified simultaneously with the Folmer et al. (1994) primers yet, the sequences are 11% different, and would represent two distinct lineages using the suggested barcoding threshold criteria of determining candidate taxa. This presents a huge problem when species such as *O. barri* and *O. australis* (Rhoades, 1941) have been found with at least 46 and 60 different numts of COI, respectively, and vary up to 12% from COI (Song et al., 2008). Therefore, gathering genetic data from the COI gene (or any gene for that matter) for systematic and phylogenetic studies is not as simple as aligning sequences from clean chromatograms.

The purpose of this paper is to educate researchers, reviewers, and students of the possible problems encountered in molecular evolutionary studies of crustaceans especially when the raw chromatograms are not inspected during peer review and the quality of the data is rarely

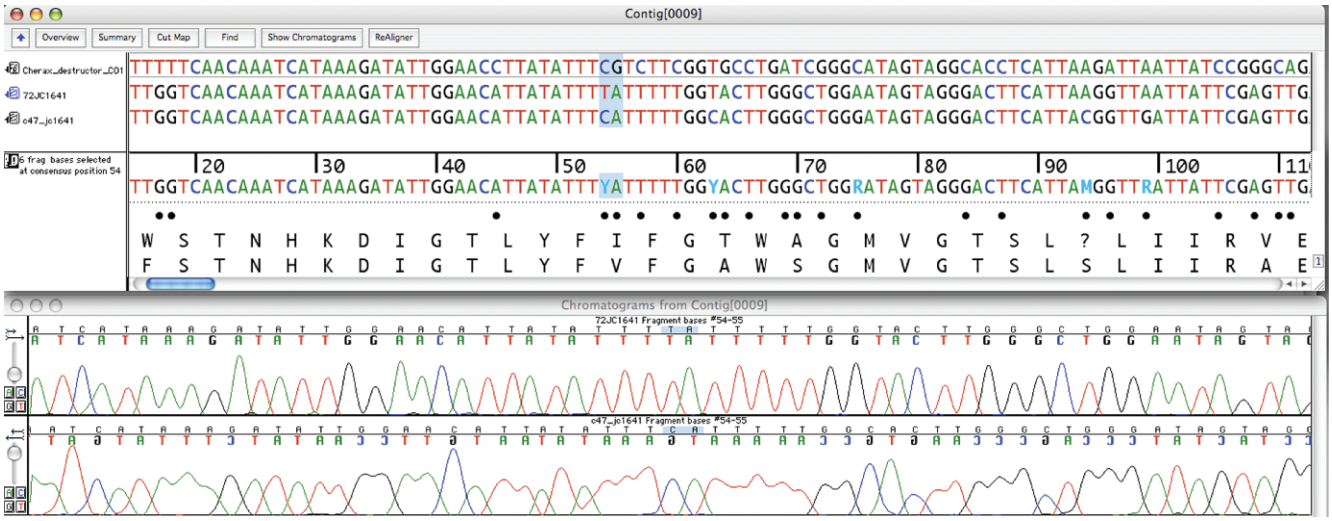

Fig. 3. Screenshot of 5′ end of sequence text, translations, and chromatograms in Sequencher. *Cherax destructor* is the reference COI crayfish sequence from the mitochondrial genome available in GenBank (NC_011243). *C. destructor* COI has a "CG" at bases 54-55. Sequence "72JC1641" is the cloned COI from individual JC1641 of *Orconectes barri* and the corresponding chromatogram is on top. This *O. barri* COI has a "TA" at bases 54-55. Sequence "c47_jc1641" is a cloned COI numt from individual JC1641 of *O. barri* and the corresponding chromatogram is on the bottom. This *O. barri* numt has a "CA" at bases 54-55. The consensus sequence of the *O. barri* COI and numt begins with TTGGTCAA under the 20 base pair marker. The big black dots are indicative of base changes for the consensus sequence relative to the reference sequence at the corresponding base positions. The translation for the consensus *O. barri* sequence is on top (begins with WSTNHKD. . .) and the translation for *C. destructor* sequence is the bottom text and begins with FSTNHKD. *Cherax* and *Orconectes* are from different families of crayfish (Parastacidae and Cambaridae, respectively), yet there is a high degree of similarity in the translated text. The majority of the base mutations at the 5′ end are synonymous.

Downloaded from https://academic.oup.com/jcb/article/29/1/96/2548098 by guest on 23 April 2024

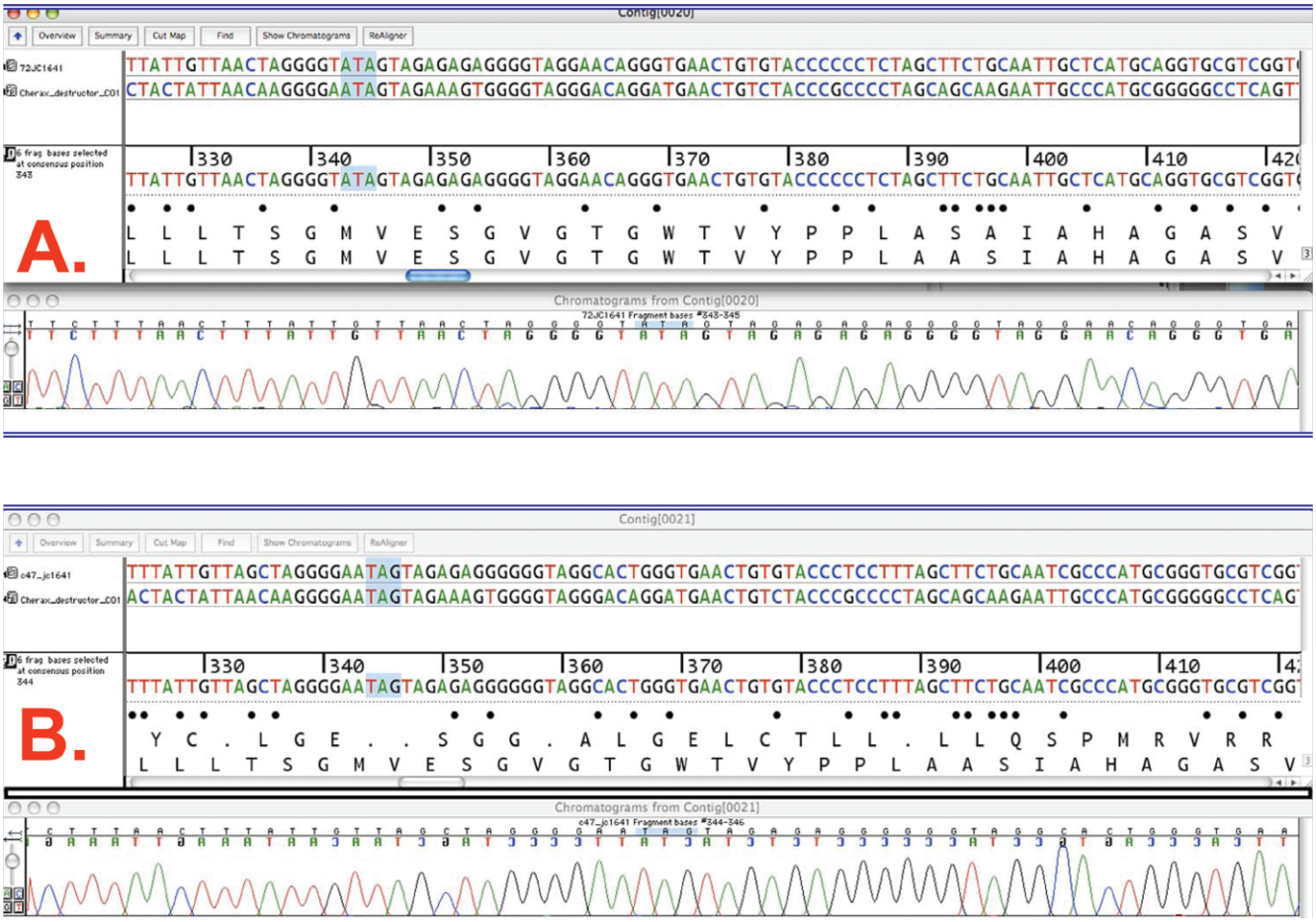

Fig. 4. Screenshots of COI and COI numt sequences, translations, and chromatograms in Sequencher. A. *Orconectes barri* COI sequence "72JC1641" with reference sequence *Cherax destructor* COI. There are no translation differences between *O. barri* and *C. destructor* despite base differences between the text window of bases 325-420. B. *Orconectes barri* COI numt sequence "c47_jc1641" with reference sequence *Cherax destructor* COI. There are obvious translation differences between the *O. barri* numt and *C. destructor* COI. Each "." represents a stop codon in the *O. barri* numt translation which begins with YC.LGE.SGG.AL while the *C. destructor* translation begins with LLLTSGMVESGV with no stop codons present between the nucleotide text of 325-420. The sequence text within the shown window matches up but there is a single base insertion upstream which forced the reading frame to shift translation by one base which leads to many stop codons (TAA and TAG) and a shift in the translated text above the chromatogram.

scrutinized. Because of the sheer number of "COI-like" sequences in their publicly available dataset, I present a re-analysis of Taylor and Knouft (2006) to illustrate how to recognize errors and red flags. Accessioning numts (rather than the intended target gene) and poor quality sequence data onto GenBank not only casts doubt on the accessioner's competence, but there is also the potential to ruin the systematic studies of others who unknowingly incorporate that data into their phylogenies and databases (for a famous example, human numts were thought to be dinosaur mtDNA; Zischler et al., 1995 reanalysis of Woodward et al., 1994). Indeed, the issue comes down to one of homology which traditionally is assumed and then tested in morphological studies, but assumed and then ignored in molecular (particularly single locus and mtDNA) studies.

Materials and Methods

The sequence dataset from GenBank for the Taylor and Knouft (2006) study (herein referred to as TK) with no changes (other than 5′ end alignment for comparison) is provided in Online Appendix 1 (see DOI p. 96). This dataset directly furnished to GenBank by the authors was flagged for further review, of which the authors ignored requests for translation information [see AY721593 *Orconectes validus* ("NCBI staff are still waiting for submitters to provide appropriate coding region information")]. The dataset was first accessioned in 2004 and since, no corrections have been submitted. After examining each sequence individually, I compiled reading frame information about these questionable sequences in Table 1. There are suspected numt sequences in their dataset, but because of the quantity (22 of 86 sequences flagged as "COI-like") and widespread locations of insertions and deletions, I do not feel confident labeling specific sequences as numts. Because of the obvious lack of quality control and obliviousness to reading frames in 25% of the sequences, I am also hesitant about the sequences considered to be COI.

I aligned the entire TK sequence set by eye to recreate their dataset for phylogenetic analysis (which included addition of many gaps) to then reconstruct the trees in TK, but with branch lengths which were not presented in their phylogenies but used extensively for their subsequent ecological and morphological analyses. I changed the ambiguity-coded bases (which included Y, M, W, K, R, and S) to N. Ideally, it would be best to use an alignment program (such as MUSCLE: Edgar, 2004 or MAFFT: http://align.genome.jp/mafft/). The total alignment is provided as Online Appendix 2 (see DOI p. 96). The appropriate model of evolution (similarly ignoring the lack of homology due to the presence of indels) was determined using ModelTest (Posada and Crandall, 1998) and those parameters were set in MrBayes (Ronquist and Huelsenbeck, 2003). MrBayes was run over 20 chains for 20 million generations.

Downloaded from https://academic.oup.com/jcb/article/29/1/96/2548098 by guest on 23 April 2024

The TK alignment was then split into two separate datasets (seemingly legitimate COI sequences and questionable "COI-like" sequences) which were analyzed for codon usage bias compared to the COI sequence (GenBank NC_011243) published as part of the complete mitochondrial genome of the freshwater crayfish *Cherax destructor* (Miller et al., 2004). Codon usage analysis (Sharp and Li, 1987; Wan et al., 2004) provides clues to amino acid composition, structure of the protein, and mutational frequency and distribution across the gene (Sueoka, 1999). Highlighting compositional differences between sequences can be an effective way to distinguish between mtDNA and numts (Bensasson et al., 2001). For the codon analysis, I excluded TK sequences with lots of unknown or missing bases and sequences which were extremely divergent. All gaps added to the COI sequences for the phylogenetic analyses were removed for codon usage analysis. I chose the same first and last base for each dataset, which resulted in the questionable sequence dataset (1447 bases) being longer than the reference *Cherax* sequence and COI dataset (1443 bases) because of indels. The starting base for the analyses was the 9th nucleotide in the dataset provided in Online Appendix 2 which corresponds to the first triplet (CGA). The last triplet of the analyzed datasets was TCT. This truncation also eliminated the sloppy 5′ and 3′ ends with indels. Using DNASP (Rozas et al., 2003), I assigned the coding region based on the COI sequence reading frame of *Cherax destructor*. All sequences were translated using the invertebrate mitochondrial code. I counted the frequency of each codon present in each sequence for comparison along with the high and low codon counts for each dataset.

Finally, I examined all the "COI-like" sequences currently available in GenBank across Crustacea with the most closely related published mitochondrial genome sequence of COI to determine the likely problem in the reading frame. I also illustrate sequence similarities of the Folmer et al. (1994) region of COI from various taxa with completed mtDNA genomes.

Results

The total alignment for the complete TK dataset for the phylogenetic analyses included 1541 bases, but the first two bases are outside the reading frame of COI. The start of COI is ACG (Miller et al., 2004). Therefore, the first base of the COI gene is actually the 3rd base in the TK dataset. The model determined for the 1541 base alignment using the Akaike Information Criterion (AIC) by Modeltest was the general time reversible model, GTR + I + G, which included number of substitution types = 6, rates = gamma, shape = 0.6194, and proportion of invariable sites = 0.4800. This was the same model reported by TK although they did not state what criterion they used (hierarchical likelihood ratio tests or AIC). Therefore, I do feel that this alignment (which included the gaps, indels, stop codons, and extra two non-COI bases at the beginning) is probably very similar to the alignment used by TK and hence, appropriate for recreation of their phylogenetic analyses. Posterior probability values from my Bayesian analysis (Fig. 5) are also very similar to TK but I only included significant nodal support (95% posterior probability: Huelsenbeck and Ronquist, 2001).

It appears that the questionable sequences are mainly basal and many cluster together, but the Bayesian tree does not clearly illuminate the problems contained in the sequence dataset. Numts of recent origin within a species often appear sister to functional sequences, while numts which arose earlier in the phylogeny often appear as basal clusters found across many species. Based on my personal experience with cave *Orconectes* (Song et al., 2008), I do suspect that the *O. inermis* (Cope, 1872) and *O. australis* sequences are indeed COI numts. The cave *Orconectes* have numts with a one base pair deletion at the 5′ end of the gene (within the first 200 bases), leading to many stop codons downstream. Many COI numts of the cave *Orconectes* also have a "CA" at the 5′ end of the sequence which does not directly code for a stop codon but does seem to be a diagnostic feature and a clue to stop codons and indels downstream (Fig. 3). The TK *O. inermis* sequence had 54 stop codons and the *O. australis* had 5 stop codons after 5′ nucleotide deletions (Table 1). Many of the other questionable sequences have single base deletions at the 3′ end, which do not result in stop codons. If the sequences were not sloppy and these deletions are correct, then these base changes represent frameshift mutations which modify the structures of the resulting molecule. Even though 3′ ends are thought to be less conservative in protein-coding genes, the 3′ ends for their seemingly COI sequences show little variation (see next paragraph about codon usage). Therefore, I feel these sequences are most likely the result of sequencing and editing errors (lack of quality control), but possibly, some could be numt sequences. Also, the COI gene was not sequenced in one piece and was not amplified using the Folmer et al. (1994) primers. It is possible that "COI" primers designed by Taylor actually amplify numt regions rather than COI.

Seventeen questionable ("COI-like") sequences and 41 supposed COI sequences from TK were individually examined for codon usage. The results of the codon analyses showed an obvious difference between the seemingly legitimate COI sequences and the questionable "COI-like" sequences compared to the *Cherax destructor* sequence (Fig. 6). I included two *Cambarus* species in the COI set and two cave *Orconectes* in the questionable set. The cave *Orconectes* are most closely related to *Cambarus* (Crandall and Fitzpatrick, 1996; Fetzner, 1996; Sinclair et al., 2003; Buhay and Crandall, 2005; Buhay and Crandall, 2008). I marked the high and low codon counts with bars to show the range of variation within each dataset. COI did not show wide-ranging variation even with the two *Cambarus* sequences and additionally, the COI set did not exhibit large differences from *Cherax destructor* (a crayfish from another family: Parastacidae). I did find that codon variation was much higher in the questionable sequences, along with the presence of stop codons which would have prevented translation of the protein. Mutational biases can greatly affect phylogenetic reconstruction and inferences (Collins et al., 1994; Griffiths, 1997). The compositional analysis of codons revealed that the mitochondrial COI gene is highly conserved even when compared to a different family and that sequence composition is a clue for separating numt and questionable "COI-like" sequences from mtDNA COI.

"COI-like" crustacean sequences (in addition to the ones from TK) were downloaded from GenBank (Table 1) and individually examined in MacClade using the COI sequences from completed mtDNA genome studies (Table 2) as references. Many of these sequences have stop codons which prevent translation, but some of the sequences have no obvious problems with a superficial survey. I contacted GenBank for specifics about sequences which were labeled "COI-like" but had no indels and stop codons. GenBank replied that there are "coding region annotation problems" and that the submitters did not reply to requests for information. Some sequences could not be aligned to the reference dataset and it was obvious that the sequences are

Downloaded from https://academic.oup.com/jcb/article/29/1/96/2548098 by guest on 23 April 2024

Fig. 5. Bayesian tree with branch lengths of the Taylor and Knouft (2006) aligned dataset. Sequences in bold with stars are questionable and were marked by GenBank staff as ''COI-like.'' Significant support (posterior probability $> 95\%$) is marked at the nodes.

Downloaded from https://academic.oup.com/jcb/article/29/1/96/2548098 by guest on 23 April 2024

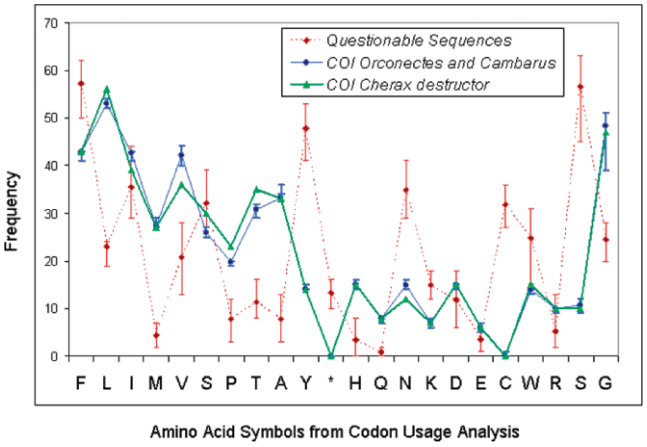

Fig. 6. Codon usage analysis for the Taylor and Knouft (2006) dataset. Questionable "COI-like" and possibly legitimate COI sequences along with the *Cherax destructor* sequence used for the analysis are provided in Table 2. The codon abbreviations for amino acids are standard and the codon determinations were based on the invertebrate mitochondrial translation table using DNASP. The mean codon counts along with the highest and lowest values for the dataset are given as points and "error" bars. The reference sequence (*Cherax destructor*) is the line with triangles, the possibly legitimate COI sequences from the TK dataset are the circles, and the questionable "COI-like" sequences from the TK dataset are the diamonds.

not COI because of the presence of stop codons at every frameshift. Overall, it appears that many "COI-like" sequences are possibly editing errors and that the submitters were not aware that protein-coding genes have open reading frames and it is necessary to check translations of the sequences.

Discussion

What are the Red Flags?

The aligned TK dataset was filled with questionable sequences, which may be due to sequence editing error, a lack of knowledge about molecular genetic sequence data, the presence of numts, or a combination of these issues. Sequences that were part of a previous paper (Taylor and Hardman, 2002) also contained 3′ gaps but were not reported or corrected (note: this was before GenBank staff began performing standard amino acid translations as part of the accession process), and ironically, Taylor accessioned them doubly into GenBank (*Orconectes kentuckiensis* AF474369 and AY701196; *Procambarus acutus* AF474366 and AY701194) along with *Orconectes juvenilis* (AF474352 and AY70233). These three individuals were reported in TK from the same localities with the same collection number as was provided in Taylor and Hardman (2002). The previous accessions were not marked with "COI-like" and it is possible that many other crustacean sequences are not flagged because GenBank only began providing the amino acid translation in 2003.

Many of the issues in TK should have been red flags to competent reviewers, even without seeing the sequence data or chromatograms in hand. First, the authors stated that uncorrected pairwise divergences ranged from 0.1% to 15.1% for species' pairs with a mean of 9.4% divergence for their data. They did not, however, present this information

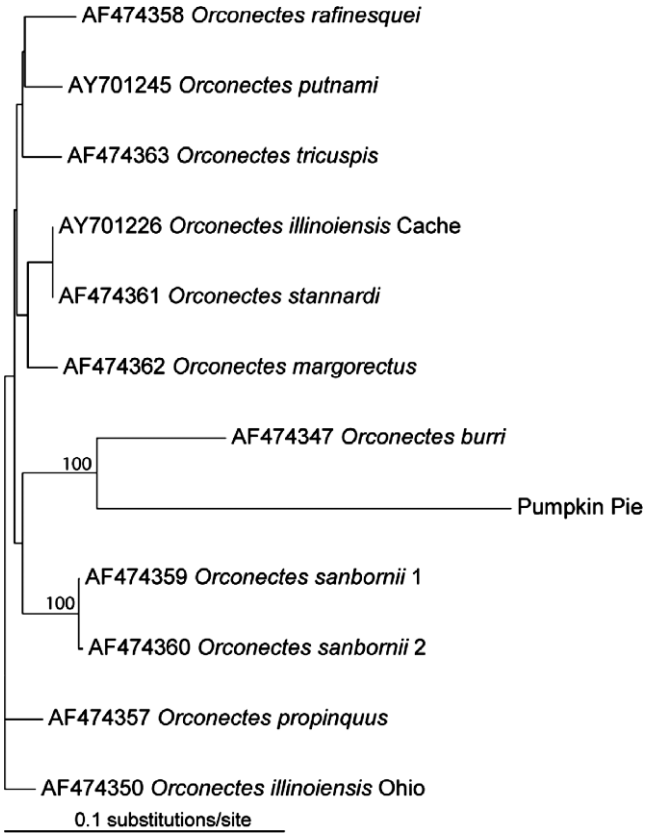

Fig. 7. An example of creating a phylogenetic tree in PAUP* using random text (Pumpkin Pie recipe). A recipe was copied and pasted directly into a nexus file subset of random sequences from Taylor and Knouft (2006). This nexus file was then opened in MacClade to truncate the sequences to the same length to accommodate the recipe. The nexus file was re-opened in PAUP* and executed. The "NJ" command was used to create a neighbor-joining tree and the "bootstrap" command was used for nodal support. All error messages were ignored. The point of the exercise was to show that anyone can insert any text into a nexus file and output a phylogenetic tree, even one with 100% "support" for a bogus clade. PAUP* issues warnings for spacing and symbol errors, not sequence text or homology. A thorough understanding of phylogenetic methods and theory is not required to push buttons in PAUP* to create a tree. This is a great exercise for students to perform when learning the ins and outs of PAUP* and any tree-building software.

as a table. Their mean estimate is much higher than previously reported for COI. Sinclair et al. (2003) reported COI divergences among species of the same genus at about 6% and among genera at about 11% after examining sequence data from hundreds of freshwater crayfish species around the world. Additionally, the codon usage examination revealed that the translated *Cherax destructor* COI sequence showed similar patterns in amino acid frequency to the COI sequences of *Cambarus* and *Orconectes*. The high estimates of sequence divergence for COI reported by TK should have been questioned by reviewers who were familiar with the conservation of this gene.

A second red flag was the lack of branch lengths on the TK phylogenetic trees. If the authors had presented such a phylogram (a phylogeny with branch lengths proportional to the amount of evolutionary change), some long branches (either due to numts which accumulate mutations in the absence of constraint or sequence editing errors in the absence of quality

Downloaded from https://academic.oup.com/jcb/article/29/1/96/2548098 by guest on 23 April 2024

Table 2. Crustacean species with published mitochondrial genomes, GenBank number, and taxonomic classification.

Crustacean	GenBank number	Taxonomy
Waterflea: *Daphnia pulex*	NC_000844	Branchiopoda; Diplostraca; Cladocera; Anomopoda; Daphniidae.
Tadpole shrimp: *Triops cancriformis*	NC_004465	Branchiopoda; Phyllopoda; Notostraca; Triopsidae.
Tadpole shrimp: *Triops longicaudatus*	NC_006079	Branchiopoda; Phyllopoda; Notostraca; Triopsidae.
Brine shrimp: *Artemia franciscana*	NC_001620	Branchiopoda; Sarsostraca; Anostraca; Artemiidae.
Fleshy prawn: *Fenneropenaeus chinensis*	NC_009679	Malacostraca; Eumalacostraca; Eucarida; Decapoda; Dendrobranchiata; Penaeoidea; Penaeidae.
Shrimp: *Litopenaeus vannamei*	NC_009626	Malacostraca; Eumalacostraca; Eucarida; Decapoda; Dendrobranchiata; Penaeoidea; Penaeidae.
Prawn: *Marsupenaeus japonicus*	NC_007010	Malacostraca; Eumalacostraca; Eucarida; Decapoda; Dendrobranchiata; Penaeoidea; Penaeidae.
Tiger prawn: *Penaeus monodon*	NC_002184	Malacostraca; Eumalacostraca; Eucarida; Decapoda; Dendrobranchiata; Penaeoidea; Penaeidae.
Crayfish: *Cherax destructor*	NC_011243	Malacostraca; Eumalacostraca; Eucarida; Decapoda; Pleocyemata; Astacidea; Parastacoidea; Parastacidae.
Crab: *Callinectes sapidus*	NC_006281	Malacostraca; Eumalacostraca; Eucarida; Decapoda; Pleocyemata; Brachyura; Eubrachyura; Heterotremata; Portunoidea; Portunidae.
Crab: *Portunus trituberculatus*	NC_005037	Malacostraca; Eumalacostraca; Eucarida; Decapoda; Pleocyemata; Brachyura; Eubrachyura; Heterotremata; Portunoidea; Portunidae.
Crab: *Geothelphusa dehaani*	NC_007379	Malacostraca; Eumalacostraca; Eucarida; Decapoda; Pleocyemata; Brachyura; Eubrachyura; Heterotremata; Potamoidea; Potamidae.
Giant crab: *Pseudocarcinus gigas*	NC_006894	Malacostraca; Eumalacostraca; Eucarida; Decapoda; Pleocyemata; Brachyura; Eubrachyura; Heterotremata; Xanthoidea; Eriphiidae.
Crab: *Eriocheir sinensis*	NC_006992	Malacostraca; Eumalacostraca; Eucarida; Decapoda; Pleocyemata; Brachyura; Eubrachyura; Thoracotremata; Grapsoidea; Varunidae.
Volcano shrimp: *Halocaridina rubra*	NC_008413	Malacostraca; Eumalacostraca; Eucarida; Decapoda; Pleocyemata; Caridea; Atyoidea; Atyidae; Caridellinae.
Giant prawn: *Macrobrachium rosenbergii*	NC_006880	Malacostraca; Eumalacostraca; Eucarida; Decapoda; Pleocyemata; Caridea; Palaemonoidea; Palaemonidae.
Spiny lobster: *Panulirus japonicus*	NC_004251	Malacostraca; Eumalacostraca; Eucarida; Decapoda; Pleocyemata; Palinura; Palinuroidea; Palinuridae.
Mantis shrimp: *Gonodactylus chiragra*	NC_007442	Malacostraca; Eumalacostraca; Hoplocarida; Stomatopoda; Unipeltata; Gonodactyloidea; Gonodactylidae.
Mantis shrimp: *Lysiosquillina maculata*	NC_007443	Malacostraca; Eumalacostraca; Hoplocarida; Stomatopoda; Unipeltata; Lysiosquilloidea; Lysiosquillidae.
Mantis shrimp: *Squilla empusa*	NC_007444	Malacostraca; Eumalacostraca; Hoplocarida; Stomatopoda; Unipeltata; Squilloidea; Squillidae.
Mantis shrimp: *Squilla mantis*	NC_006081	Malacostraca; Eumalacostraca; Hoplocarida; Stomatopoda; Unipeltata; Squilloidea; Squillidae.
Mantis shrimp: *Harpiosquilla harpax*	NC_006916	Malacostraca; Eumalacostraca; Hoplocarida; Stomatopoda; Unipeltata; Squilloidea; Squillidae.
Sea slater: *Ligia oceanica*	NC_008412	Malacostraca; Eumalacostraca; Peracarida; Isopoda; Oniscidea; Ligiidae.
Copepod: *Tigriopus californicus*	NC_008831	Maxillopoda; Copepoda; Neocopepoda; Podoplea; Harpacticoida; Harpactidae.
Copepod: *Tigriopus japonicus*	NC_003979	Maxillopoda; Copepoda; Neocopepoda; Podoplea; Harpacticoida; Harpactidae.
Salmon louse: *Lepeophtheirus salmonis*	NC_007215	Maxillopoda; Copepoda; Neocopepoda; Podoplea; Siphonostomatoida; Caligidae.
Barnacle: *Megabalanus volcano*	NC_006293	Maxillopoda; Thecostraca; Cirripedia; Thoracica; Sessilia; Balanidae.
Acorn barnacle: *Tetraclita japonica*	NC_008974	Maxillopoda; Thecostraca; Cirripedia; Thoracica; Sessilia; Tetraclitidae.
Sea firefly: *Vargula hilgendorfii*	NC_005306	Ostracoda; Myodocopa; Myodocopida; Cypridinoida; Cypridinidae.

control) would have been apparent and suspicious (Fig. 5). During re-analyses, several of the questionable sequences had unusually long branches in the phylograms. TK then used the branch length information (which included the 22 ''COI-like'' sequences) from both their parsimony and Bayesian phylogenies to examine genetic associations with habitat and morphology, but they unknowingly compounded the errors. This illustrates the necessity for thoughtful examination of sequences before performing further analyses and before making the sequences available to other researchers who may also be unaware of the errors and issues.

Faulty Data Makes Faulty Trees

As an example of ''what not to do'' but is commonly done, i.e., get a tree by pushing buttons in PAUP*: Swofford (2001), I used a subset of the TK dataset (eleven random taxa) and added my favorite recipe for pumpkin pie (imagine it is a numt sequence or junk DNA; Online Appendix 3, see DOI p. 96) to the nexus file without changes just by copying and pasting the text. After saving the nexus file, I opened the file in MacClade which beeped numerous error messages during the file opening. I pushed the ''ignore'' or ''cancel'' button for every error message (which dealt with such necessary things as teaspoon of salt and 2 eggs). Despite the repeated error messages, the ''sequence'' for pumpkin pie was inserted with the rest of the crayfish dataset which included IUPAC codes and ''?''s (Online Appendix 4, see DOI p. 96). I truncated the crayfish sequence dataset to include the complete pumpkin pie recipe. After saving the file in nexus format, I executed the

Downloaded from https://academic.oup.com/jcb/article/29/1/96/2548098 by guest on 23 April 2024

Fig. 8. Screenshot of COI sequences from complete mitochondrial genomes of crustaceans listed on Table 2. The window in MacClade shows the 5′ end of COI, approximately bases 148-234 represented as color-coded amino acids to display similarity. This region represents the 5′ end of the gene region amplified by the Folmer et al. (1994) primers.

file and produced not just a neighbor-joining tree complete with crayfish sequences and pumpkin pie, but I also bootstrapped the results to get 100% support for a sister relationship between pumpkin pie and *Orconectes burri* (Taylor and Sabaj, 1998) (Fig. 7). My point is that anyone can push buttons in PAUP* and get a tree or a distance matrix, even with completely bogus "sequences." It takes no molecular understanding to create a phylogenetic tree in

Fig. 9. Screenshot of COI sequences from complete mitochondrial genomes of crustaceans listed on Table 2. The window in MacClade shows the middle of the COI gene, approximately at bases 670-756, to represent the 3′ end of the region amplified by the Folmer et al. (1994) primers. Amino acids are colored to display similarity.

Downloaded from https://academic.oup.com/jcb/article/29/1/96/2548098 by guest on 23 April 2024

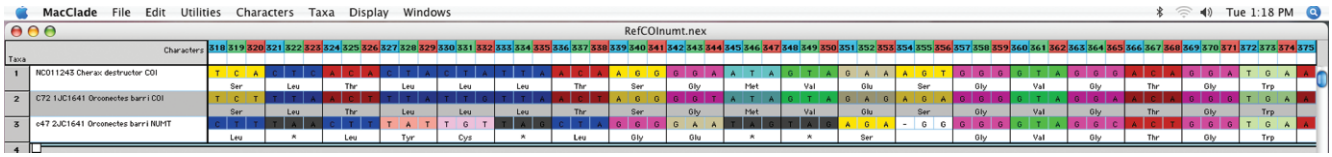

Fig. 10. Screenshot of two COI sequences and one numt sequence in MacClade. The top sequence is the reference COI sequence from *Cherax destructor*, the middle sequence (C72_1JC1641) is COI from *Orconectes barri*, and the bottom sequence (c47_2JC1641) is a COI numt from *O. barri*. Below each sequence is the translation which is done by setting the "genetic code" to "invertebrate" under the "utilities" menu at the top. Stop codons are marked with a "*" and the nucleotides are colored black. The *O. barri* numt has four stops before a gap at base 354 that I inserted to align the rest of the sequence with the *O. barri* COI above it. Indels are rampant in numt sequences of *O. barri* but the sequences retain some similarity to COI sequences.

PAUP*. Obviously, nodal support for a clade, clean chromatograms, and phylogenetic software do not provide bright flashing warnings that say "Error: numts present in your dataset" or "You've got issues: there are gaps in your supposed protein-coding gene." Even error messages from computer programs, crustacean peers, and molecular data technicians can be and are ignored. The devil is in the details, but how does one even realize errors and numts in their sequences in the first place?

The Million Dollar Question: Are Numts Present?

Without experience in handling molecular genetic data and training in molecular evolutionary analyses, a researcher might not recognize what exactly (if anything) is wrong if the sequence chromatograms were clean. I found that COI numts are common across the obligate cave species in the genus *Orconectes* and the numts (instead of the target COI) do sequence cleanly using the Folmer et al. (1994) primers. I also found that by cloning the PCR products of individuals with sloppy chromatograms, I can isolate the mtDNA sequences, but they are far outnumbered by numt sequences which can make a cheap easy project expensive and tedious quickly. By using abdominal (tail) tissue, which TK used, the chances of sequencing numts rather than mtDNA is higher than when using the mitochondrial-rich tissues, such as the gills from under the carapace. Therefore, my first suggestion to avoid numts is to use mitochondrial-rich tissue.

My next suggestion is to cross-check clean sequences with COI sequences from the published mitochondrial genomes of the most closely-related taxa to your study group. There are currently 29 complete sequences of the entire mitochondrial genomes for most groups of crustaceans, including branchiopod, decapod, stomatopod, isopod, copepod, ostracod, and barnacle (Table 2). These whole genomes are invaluable and overlooked resources which should be reviewed. Typically, publications of the genome will include the start and stop codons (initiation and termination triplets) of the gene and if any peculiar issues are present (such as the mitochondrial gene rearrangements in *Cherax destructor* compared to other crustaceans; Miller et al., 2004). The COI sequences from the mitochondrial genomes can be accessed through GenBank (they are also available as Online Appendix 5, see DOI p. 96), downloaded into alignment files and editing programs, and set as the reference sequence to aid in aligning new data. Doing this will help determine the 5′ start and will help with amino acid translation of the sequences to check for stop codons and breaks in the reading frame. The region covered by the Folmer et al. (1994) primers is conserved in amino acids even when compared across different crustacean families (Figs. 8, 9) and it should be relatively easy to correct sloppy 5′ and 3′ ends just by translating sequences and comparing them to references.

Who's on First: Are Stop Codons and Indels Present in the Reading Frame?

When chromatograms are available, Sequencher has the option to set the reading frame using a reference sequence (Fig. 4) which then displays the translated text below the nucleotide text. When indels are present which disrupt the open reading frame, there is an obvious shift present in the translated text relative to the reference sequence translated text. When a stop codon is present, Sequencher marks the stop codon with a "." rather than an amino acid code (such as "S" for serine or "V" for valine) in the translated text. Sequencher allows for simultaneous examination of the translated sequence along with the chromatogram to determine if there is a nucleotide ambiguity or a genuine break in the reading frame not due to sloppy sequence issues. For larger datasets, it may be easier once the editing of the raw data and chromatograms is done to review the alignment in MacClade to check for larger similarity patterns. MacClade is a useful program for checking alignments and datasets because triplets can be color coded with the nucleotides still visible within the colors (Fig. 10). Stop codons are by default black and are marked with a "*" rather than an abbreviation for an amino acid. It is a good idea to set the start of an alignment of a protein-coding gene with the first base of the codon triplet. In MacClade, this option will set the triplet pattern for the entire dataset.

When chromatograms are not available (as is the case with downloading sequence text from GenBank), MacClade can be used to check for inconsistencies in sequences relative to a reference sequence across large datasets (up to 1500 sequences at a time). Once the reading frame is set, it is easy to scan each sequence for stop codons, which may not occur at the same region for each sequence (Fig. 11). Whether the stop codons are due to sloppy sequences or editing cannot always be determined or separated from genuine numt sequences using only sequence texts. But it should be obvious that sequences with stop codons are not COI and that extremely different sequences should not be used in any phylogenetic, systematic, or barcoding study. Furthermore, these sequences should absolutely not be accessioned to GenBank as mtDNA COI because it causes confusion and exacerabates errors for cross-referenced databases, such as the BOL network.

Downloaded from https://academic.oup.com/jcb/article/29/1/96/2548098 by guest on 23 April 2024

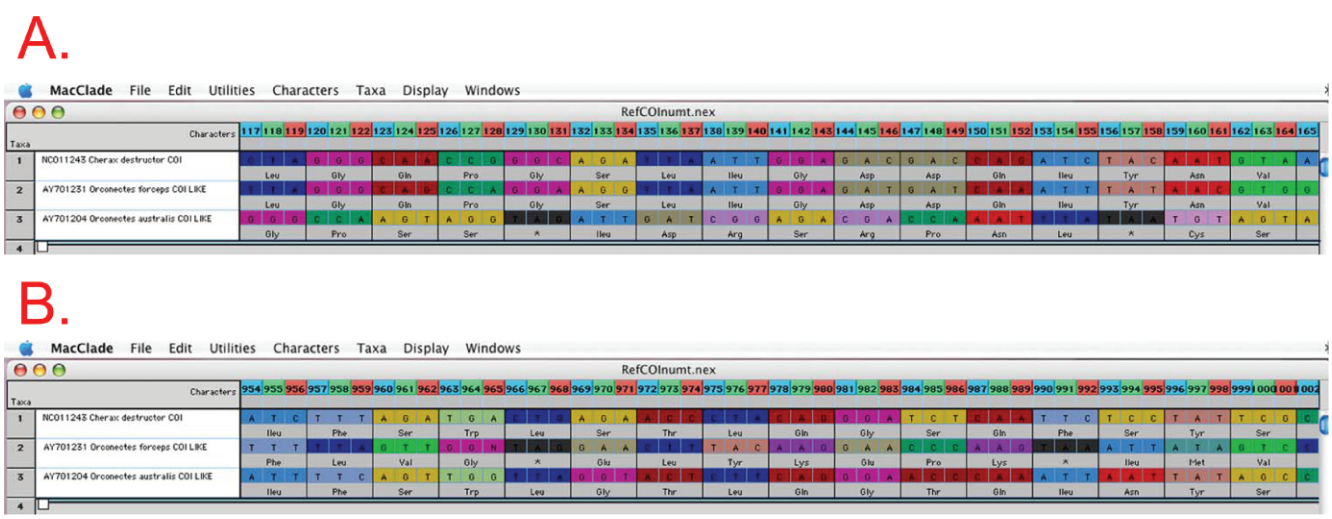

Fig. 11. Screenshot comparison of "COI-like" sequences with the reference COI sequence in MacClade to show translations. The top sequence is the reference COI sequence of *Cherax destructor*. The middle sequence is the "COI-like" sequence of *Orconectes forceps* (GenBank AY701231) and the bottom sequence is the "COI-like" sequence of *Orconectes australis* (GenBank AY701204) both from Taylor and Knouft (2006). A. Window of the 5′ end of the COI gene shows stop codons in *O. australis* at bases 129-131 and 156-158. B. Window of the middle region of the COI gene shows a stop codon in *O. forceps* at bases 990-992 but there are obvious differences leading up to the stop codon.

Accessioning BARCODE Sequences to GenBank

The BOL database is linked with GenBank and many other online databases, which becomes especially problematic when "COI-like" sequences are considered to be COI data. For many of these barcoded species, there are at best only a handful of sequences available to represent the lineage. It is the hope that more than three individuals have barcodes to characterize each species, but for now, the main goal of BOL is to gather COI data from as many species on the planet as possible. Individual specimens represent the majority of species in the database, which makes it critical to have both accurate species identification with accurate COI sequences to serve as references for queried unknowns.

When accessioning COI data to GenBank, the keyword "BARCODE" is used to designate cross-linkage with the BOL database. As an example, COI from *Orconectes barri* is already accessioned into GenBank (EF207164) using the KEYWORD "BARCODE" and "mitochondrion" as the SOURCE. There are now updated required elements for accessioning barcodes using Sequin (http://www.ncbi.nlm.nih.gov/Sequin/index.html) which include 1) voucher number, 2) collector, 3) collection locality, and 4) PCR primer sequences used for amplification and sequencing. This information fits under the "source" section of "FEATURES." An example of the new format is GenBank (DQ882094). This specific information is useful for queries of unknown sequences (molecular taxonomy) performed to detect the origin and identity of invasive and introduced species, for phylogeographic studies, and to answer questions about voucher identifications and lab protocols by a particular researcher.

What's a Numt to do?

Numt sequences are not totally useless pieces of genetic information. The study of pseudogenes provides important clues to gene duplication events, genome size, and ancestral gene functions (Zhang and Hewitt, 1996; Bensasson et al., 2001). For these reasons, numt sequences are accessioned onto GenBank, but most certainly not as protein-coding mitochondrial genes or BARCODES! Using Sequin, numt sequences should be marked as "genomic DNA" with a note of "numt" under the "source" section of "FEATURES." An example of a numt accession is EF011414. Because it is a numt, the translated text of the protein normally added by the GenBank staff as part of the accession process of protein-coding genes was not performed.

Conclusion

I hope this study sheds bright light on the issues of quality control and homology in molecular projects and I urge authors inexperienced with molecular evolutionary studies to make both the raw data and alignments available upfront for inspection by reviewers and editors. It is also absolutely necessary to correct (and possibly remove if not fixable) the "COI-like" sequences in GenBank if you were the accessioner of those sequences. Similarly, it may be necessary to issue erratums or retractions for published studies that are based largely on erroneous data. Until publicly-available molecular databases issue requirements for quality control, we must scrutinize and be wary of free data. It is commonplace to question identification of species by zoologists as a source of error in phylogenetic studies, but as this study has shown, molecular data errors are not just negligible issues anymore – they are cause for serious concern which must be addressed. Correcting faulty molecular data is not as simple as changing a species' ID label in a database, museum record, or on a phylogenetic tree – it is an issue that can involve some or all of the following: resequencing, re-editing and re-checking chromatograms, cloning, re-analyzing phylogenetic input and output, re-examining systematic conclusions, and re-accessioning GenBank sequences.

Downloaded from https://academic.oup.com/jcb/article/29/1/96/2548098 by guest on 23 April 2024

I urge authors, reviewers, and editors to be vigilant of questionable data such as high percent sequence divergences, accessioned sequences with lots of ambiguity codes, phylogenetic trees with unusual long branch lengths, and systematic relationships which make no sense (such as supposed sister species falling out in different clades of a tree). The use of other genes (Rubinoff, 2006) and many of the methods I illustrated should help determine the source of the error, whether it be numts, contaminated DNA, or poor sequence quality and editing. There is also a flowchart provided by Song et al. (2008) of laboratory and analytical methods used to avoid numt contamination and detect numts in a dataset – I urge all of you to please examine your sequences. Many of the methods for numt detection were addressed in detail in this manuscript. Beyond that, I urge the inexperienced to take coursework in molecular evolution and analyses to become knowledgable about the basics of genetic data underlying the methods of moving DNA around the lab and pushing buttons in computer software. I hope that by bringing these concerns to the forefront about ''COI-like'' data, that subsequent studies will take greater care to authenticate the data upon which crucial conservation conclusions are based.

Acknowledgements

Thank you to Keith Crandall and two anonymous reviewers for helpful comments on this manuscript. This is Belle W. Baruch Institute of Marine and Coastal Sciences Contribution No. 1480.

References

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

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

Blankenship, L. E., and A. A. Yayanos. 2005. Universal primers and PCR of gut contents to study marine invertebrate diets. Molecular Ecology 14: 891-899.

Buhay, J. E., and K. A. Crandall. 2005. Subterranean phylogeography of freshwater crayfishes shows extensive gene flow and surprisingly large population sizes. Molecular Ecology 14: 4259-4273.

———, and ———. 2008. Taxonomic revision of cave crayfish in the genus *Orconectes*, subgenus *Orconectes* (Decapoda: Cambaridae) along the Cumberland Plateau, including a description of a new species - *Orconectes barri*. Journal of Crustacean Biology 28: 57-67.

Clark, E. 1936. The freshwater and land crayfishes of Australia. Memoirs of the National Museum of Victoria 10: 5-58.

Collins, T. M., P. H. Wimberger, and G. J. P. Naylor. 1994. Compositional Bias, Character-State Bias, and Character-State Reconstruction Using Parsimony. Systematic Biology 43: 482-496.

Cope, E. D. 1872. On the Wyandotte Cave and its fauna. American Naturalist 6: 406-422.

Costa, F. O., J. R. deWaard, J. Boutillier, S. Ratnasingham, R. Dooh, M. Hajibabaei, and P. D. N. Hebert. 2007. Biological identifications through DNA barcodes: the case of the Crustacea. Canadian Journal of Fisheries and Aquatic Sciences 64: 272-295.

Cracraft, J., J. Felsenstein, J. Vaughn, and K. Helm-Bychowski. 1998. Sorting out tigers (*Panthera tigris*): Mitochondrial sequences, nuclear inserts, systematics, and conservation genetics. Animal Conservation 1: 139-150.

———, and J. F. Fitzpatrick, Jr. 1996. Crayfish molecular systematics: using a combination of procedures to estimate phylogeny. Systematic Biology 45: 1-26.

———, and E. H. Rudolph. 2007. A new species of burrowing crayfish *Virilastacus retamali* (Decapoda: Parastacidae) from the southern Chile peatland. Journal of Crustacean Biology 27: 502-512.

Dayrat, B. 2005. Towards integrative taxonomy. Biological Journal of the Linnean Society 85: 407-415.

Edgar, R. C. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Research 32: 1792-1797.

Fetzner, J. W., Jr. 1996. Biochemical systematics and evolution of the crayfish genus *Orconectes* (Decapoda: Cambaridae). Journal of Crustacean Biology 16: 111-141.

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

Griffiths, C. S. 1997. Correlation of Functional Domains and Rates of Nucleotide Substitution in Cytochrome b. Molecular Phylogenetics and Evolution 7: 352-365.

Hall, T. A. 1999. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucleic Acids Symposium Series 41: 95-98.

Hansen, B., and A. J. Smolenski. 2002. A preliminary examination of the molecular relationships of some crayfish species from the genus Parastacoides (Decapoda: Parastacidae), pp. 547-554. In, G. J. Whisson and B. Knott (eds.), Freshwater Crayfish 13: Proceedings of the 13th symposium of the International Association of Astacology.

Hazkani-Covo, E., and D. Graur. 2007. A comparative analysis of numt evolution in human and chimpanzee. Molecular Biology and Evolution 24: 13-18.

Hebert, P. D. N., A. Cywinska, S. L. Ball, and J. R. de Waard. 2003a. Biological identifications through DNA barcodes. Proceedings of the Royal Society B 270: 313-322.

———, S. Ratnasingham, and J. R. deWaard. 2003b. Barcoding animal life: cytochrome c oxidase subunit 1 divergences among closely related species. Proceedings of the Royal Society of London B 270: 596-599.

———, E. H. Penton, J. M. Burns, D. H. Janzen, and W. Hallwachs. 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.

———, M. Y. Stoeckle, T. S. Zemlak, and C. M. Francis. 2004b. Identification of birds through DNA barcodes. PLoS Biology 2: 1657-1663.

Hogg, I. D., and P. D. N. Hebert. 2004. Biological identification of springtails (Collembola: Hexapoda) from the Canadian Arctic, using mitochondrial DNA barcodes. Canadian Journal of Zoology 82: 749-754.

Huelsenbeck, J. P., and F. Ronquist. 2001. MrBayes: Bayesian inference of phylogeny. Bioinformatics 17: 754-755.

Lopez, J. V., N. Yuhki, R. Masuda, W. Modi, and S. J. O'Brien. 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.

Maddison, W. P., and D. R. Maddison. 2002. MacClade: Analysis of Phylogeny and Character Evolution. Version 4.0. Sinauer: Sunderland, Massachusetts.

Meyer, C. P., and G. Paulay. 2005. DNA Barcoding: Error Rates Based on Comprehensive Sampling. PLoS Biology 3: e422.

Miller, A. D., T. T. T. Nguyen, C. P. Burridge, and C. M. Austin. 2004. Complete mitochondrial DNA sequences of the Australian freshwater, *Cherax destructor* (Crustacea: Decapoda: Parasticidae): a novel gene order revealed. Gene 331: 65-72.

Moritz, C., and C. Cicero. 2004. DNA barcoding: promise and pitfalls. PLoS Biology 2, e354.

Nguyen, T. T. T., N. P. Murphy, and C. M. Austin. 2002. Amplification of multiple copies of mitochondrial Cytochrome b gene fragments in the Australian freshwater crayfish, *Cherax destructor* Clark (Parastacidae: Decapoda). Animal Genetics 33: 304-308.

Olsen, L. E., and A. D. Yoder. 2002. Using secondary structure to identify ribosomal numts: cautionary examples from the human genome. Molecular Biology and Evolution 19: 93-100.

Paquin, P., and M. Hedin. 2004. The power and perils of 'molecular taxonomy': a case study of eyeless and endangered *Cicurina* (Araneae: Dictynidae) from Texas caves. Molecular Ecology 13: 3239-3255.

Posada, D., and K. A. Crandall. 1998. MODELTEST: Testing the model of DNA substitution. Bioinformatics 14: 817-818.

Rhoades, R. 1941. Notes on some crayfishes from Alabama cave, with the description of a new species and a new subspecies. Proceedings of the United States National Museum 91: 141-148.

Downloaded from https://academic.oup.com/jcb/article/29/1/96/2548098 by guest on 23 April 2024

Richly, E., and D. Leister. 2004. NUMTs in sequenced eukaryotic genomes. Molecular Biology and Evolution 21: 1081-1084.

Ronquist, F., and J. P. Huelsenbeck. 2003. MRBAYES 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19: 1572-1574.

Rozas, J., J. C. Sanchez-DelBarrio, X. Messeguer, and R. Rozas. 2003. DnaSP, DNA polymorphism analyses by the coalescent and other methods. Bioinformatics 19: 2496-2497.

Rubinoff, D. 2006. DNA Barcoding Evolves into the Familiar. Conservation Biology 20: 1548-1549.

Schneider-Broussard, R., and J. E. Neigel. 1997. A large subunit mitochondrial ribosomal DNA sequence translocated to the nuclear genomes of two stone crabs. Molecular Biology and Evolution 14: 156-165.

———, D. L. Felder, C. A. Chlan, and J. E. Neigel. 1998. Tests of phylogeographic models with nuclear and mitochondrial DNA sequence variation in the stone crabs, *Menippe adina* and *M. mercenaria*. Evolution 52: 1671-1678.

Sharp, P. M., and W. H. Li. 1987. The codon Adaptation Index - a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Research 15: 1281-95.

Sinclair, E. A., J. W. Fetzner, Jr., J. E. Buhay, and K. A. Crandall. 2004. Proposal to complete a phylogenetic taxonomy and systematic revision for freshwater crayfish (Astacidae). Freshwater Crayfish 14: 21-29.

Sites, J. W., Jr., and J. C. Marshall. 2004. Operational criteria for delimiting species. Annual Review of Ecology, Evolution, and Systematics 35: 199-227.

Song, H., J. E. Buhay, M. F. Whiting, and K. A. Crandall. 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 U.S.A. 105: 13486-13491.

Sorenson, M. D., and T. W. Quinn. 1998. Numts: A Challenge for Avian Systematics and Population Biology. Auk 155: 214-221.

Sueoka, N. 1999. Two aspects of DNA base composition: G+C content and translation-coupled deviation from intra-strand rule of A = T and G = C. Journal of Molecular Evolution 49: 49-62.

Swofford, D. L. 2001. PAUP*: Phylogenetic Analysis Using Parsimony (*and other methods). Version 4. Sinauer Associates, Sunderland, Massachusetts.

Tang, B., K. Zhou, D. Song, G. Yang, and A. Dai. 2003. Molecular systematics of the Asian mitten crabs, genus *Eriocheir* (Crustacea: Brachyura). Molecular Phylogenetics and Evolution 29: 309-316.

Taylor, C. A., and M. H. Sabaj. 1998. A new crayfish of the genus Orconectes from the Blood River drainage of western Kentucky and Tennessee (Decapoda: Cambaridae). Proceedings of the Biological Society of Washington 111: 645-652.

———, and M. Hardman. 2002. Phylogenetics of the crayfish subgenus *Crockerinus*, genus *Orconectes* (Decapoda: Cambaridae), based on cytochrome oxidase I. Journal of Crustacean Biology 22: 874-881.

———, and J. H. Knouft. 2006. Historical influences on genital morphology among sympatric crayfishes: systematics and gonopod evolution in the genus *Orconectes* (Cambaridae). Biological Journal of the Linnean Society 89: 1-12.

Vestheim, H., B. Edvardsen, and S. Kaartvedt. 2005. Assessing feeding of a carnivorous copepod using species-specific PCR. Marine Biology 147: 381-385.

Wan, X.-F., D. Xu, A. Kleinhofs, and J. Zhou. 2004. Quantitative relationship between synonymous codon usage bias and GC composition across unicellular genomes. BMC Evolutionary Biology 4: 19.

Ward, R. D., T. S. Zemlak, B. H. Innes, P. R. Last, and P. D. N. Hebert. 2005. DNA barcoding Australia's fish species. Philosophical Transactions of the Royal Society B 360: 1847-1857.

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

Woodward, S. R., N. J. Weyand, and M. Bunnell. 1994. DNA sequence from Cretaceous period bone fragments. Science 266: 1229-1232.

Zhang, D-X., and G. M. Hewitt. 1996. 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.

Zischler, H., M. Hoss, O. Handt, A. von Haeseler, A. C. van der Kuyl, and J. Goudsmit. 1995. Detecting dinosaur DNA. Science 268: 1192-1193.

RECEIVED: 29 March 2008.
ACCEPTED: 22 July 2008.

Downloaded from https://academic.oup.com/jcb/article/29/1/96/2548098 by guest on 23 April 2024