Limnology (2020) 21:319–325
https://doi.org/10.1007/s10201-020-00618-y

RESEARCH PAPER

Identifying a "pseudogene" for the mitochondrial DNA COI region of the corixid aquatic insect, *Hesperocorixa distanti* (Heteroptera, Corixidae)

Koki Yano[1] · Masaki Takenaka[1] · Toshimasa Mitamura[2] · Koji Tojo[1,3,4]

Received: 18 July 2019 / Accepted: 23 March 2020 / Published online: 9 April 2020

Abstract
DNA barcoding has been actively used as a method for species identification, and it will become an increasingly important method in the future. However, DNA barcoding can occasionally encounter a major problem, namely "nuclear mitochondrial DNA pseudogenes (NUMTs)". In this paper, we report the discovery of a pseudogene in the mitochondrial COI region from aquatic heteropterans. PCR with Folmer's universal primer set produced two different lengths of DNA sequences: a standard 658 bp sequence and a slightly shorter 652 bp sequence. The presence of an "indel" and "stop codon" in the shorter sequence made them separable. We recommend using the "HCOoutout" primer as a downward primer for amplification of the COI region of aquatic heteropterans, especially corixid water bugs.

Keywords DNA barcoding · Hemiptera · mtDNA · NUMTs · Phylogeny · Pseudogene

Introduction

In recent years, the DNA barcoding technique has been widely used for species identification. Initially, it was used as an auxiliary tool in species identification for organisms that are difficult to differentiate based solely on their morphological characteristics. However, due to a decrease in the number of taxonomists with the necessary skill set to identify organisms based on morphological characteristics alone, it is expected that the use of DNA barcoding as an identification method will continue to increase in the future (Ratnasingham and Hebert 2007, 2013). In addition to identify species, such genetic analysis of the DNA barcoding region is also highly valuable in the detection of cryptic species and the reassessment of a species' diversity (Tojo et al. 2017; Okamiya et al. 2018; Saito et al. 2018; Yano et al. 2019).

In the aquatic insects of interest in this study, the standard DNA barcoding region is the mitochondrial DNA (mtDNA) COI region (658 bp). In invertebrates including insects, the primer set by Folmer et al. (1994) and Hebert et al. (2003) is widely used, and this primer set is well known as the "universal primer" for invertebrates. In some species groups, however, amplification of DNA fragments with the universal primer set is difficult (e.g. marine invertebrates). In such cases, another modified primer set applied to the same COI region is also often used (Geller et al. 2013). Due to the high versatility of the above two primer sets, and the ease of detecting both inter and intra species' genetic polymorphisms in the mtDNA COI region, these primer sets have been adopted in many studies targeting invertebrates including insects, as the first step in their molecular phylogenetic analyses (Folmer et al. 1994; Hebert et al. 2003; Geller et al. 2013). In fact, the "BOLD System (Barcode of Life Data System)", which is operated by an international organization, has registered more than 3.5 million COI region data sets to date (as of Feb 26th, 2020; URL https://biodiversi

Handling Editor: Wataru Makino.

✉ Koji Tojo
ktojo@shinshu-u.ac.jp

[1] Division of Mountain and Environmental Science, Interdisciplinary Graduate School of Science and Technology, Shinshu University, Matsumoto, Japan

[2] Fukushima Agricultural Technology Centre, Hama-Dori Research Centre, Fukushima, Japan

[3] Department of Biology, Faculty of Science, Shinshu University, Asahi 3-1-1, Matsumoto, Nagano 390-8621, Japan

[4] Institute of Mountain Science, Shinshu University, Matsumoto, Japan

tygenomics.net/projects/bold/), and users can use them to identify a wide range of taxa using only sequence data of the COI region (Collins and Cruickshank 2013).

In this study, the sequences of the mtDNA COI region of corixid aquatic insects (Heteroptera, Corixidae) were analyzed using Folmer's universal primer set (Folmer et al. 1994), and it was found that pseudogenes could also be amplified. Although the problems caused by pseudogenes, especially NUMTs: nuclear sequences of mitochondrial origin (nuclear mtDNA pseudogenes: Bensasson et al. 2001), have been reported widely in eukaryotes, their detection among heteropteran insects observed in this study is the first case. Therefore, for these aquatic bugs, we would like to highlight the risks involved with the DNA barcoding method when performing genetic analyses with such a widely used primer set (Folmer et al. 1994; Hebert et al. 2003; Geller et al. 2013) without careful data checking. We would also like to propose a method to avoid the amplification of pseudogenes in this group.

Materials and methods

To analyze the COI barcoding region, the total genomic DNA was extracted and subsequently purified from nine *Hesperocorixa* aquatic bugs; two specimens of *Hesperocorixa kolthoffi* (Specimen collection site: Arige, Wakamatsu-ku, Kitakyushu City, Fukuoka Prefecture, Japan, GenBank accession number (acc. no.) LC528377, LC528378; Fuchu, Sakaide City, Kagawa Prefecture, Japan, acc. no. LC528379, LC528380), one specimen of *Hesperocorixa distanti distanti* (Shippu, Atsuta-ku, Ishikari City, Hokkaido Prefecture, Japan, acc. no. LC528370, LC528371), and four specimens of *Hesperocorixa hokkensis* (Miyakozawa Wet-land, Oyama, Tsuruoka City, Yamagata Prefecture, Japan, acc. no. LC528372, LC528373; Nanko Park, Shirakawa City, Fukushima Prefecture, Japan, acc. no. LC528374–LC528376). Using the total genomic DNA as a template, the mtDNA COI region (658 bp) was amplified by PCR with Folmer's universal primer set (Folmer et al. 1994; Fig. 1, Table 1). Subsequently, the DNA nucleotide sequences of the mtDNA COI region were analyzed by the "direct sequence" method. The method used for the total genomic DNA extraction and purification of PCR products can be found in Suzuki et al. (2013, 2014, 2019; Takenaka and Tojo 2019; Takenaka et al. 2019).

PCR with Folmer's universal primer set and that with LCO1490 and HCOoutout were done under the same thermal and chemical conditions outlined in previous studies (Saito and Tojo 2016a, b; Saito et al. 2016, 2018). Since the HCOoutout primer has a different design to Folmer's universal primer, it completely covers the DNA barcoding region, and it is also clear that the HCOoutout primer is useful to analyze various invertebrate taxa (Pickett et al. 2006; Clouse and Wheeler 2014; Sanchez and Cassis 2018). Regarding the PCR of this tissue specimen, rTaq (TOYOBO, Osaka) was used as a DNA polymerase. As for PCR, a 2720 Thermal Cycler (Applied Biosystems, Tokyo) was used. All primer sets and relationships used in this study are described in Fig. 1 and Table 1.

Phylogenetic analysis was performed using the maximum-likelihood method (ML; Felsenstein 1981) with MEGA ver. 6.06 (Tamura et al. 2013) and the Bayesian method using BEAST ver. 2.5.2 (Bouckaert et al. 2019). Nodal support was measured with 1000 bootstrap replicates (Felsenstein 1985), and the posterior probabilities in

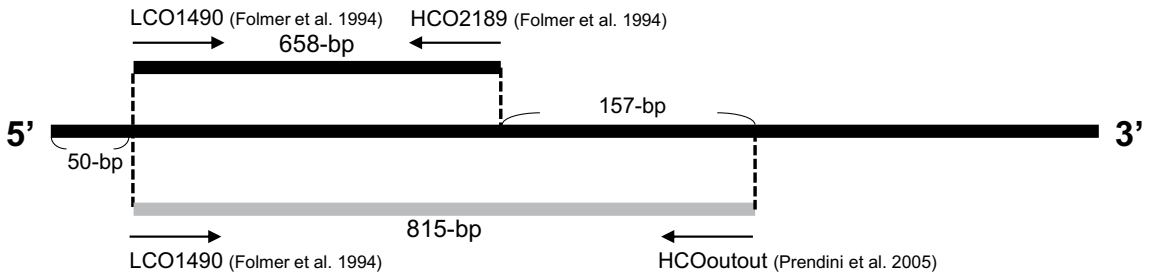


Fig. 1 Primer locations within the whole mtDNA COI region

Table 1 PCR primers used in this study

Gene	Primer name	Type	Sequence (5′–3′)	Length	References
mtDNA COI	Universal	Forward	GGTCAACAAATCATAAAGATATTGG	658 bp	Folmer et al. (1994)
	Universal	Reverse	TAAACTTCAGGGTGACCAAAAAATCA	658 bp	Folmer et al. (1994)
	HCOoutout	Reverse	GTAAATATATGRTGDGCTC	815 bp	Prendini et al. (2005)

BEAST 2.5.2. Bayesian analysis used 50 million Markov Chain Monte Carlo (MCMC) cycles with a sampling frequency of 1000. To obtain a consensus tree, data from the initial 10% of cycles were discarded as burn-in. Prior to ML and Bayesian phylogenetic estimations, the program KAKUSAN4 (Tanabe 2007) was used. TN93 + G + I was chosen as the best-fit model.

Results and discussion

Two types of DNA sequence with different lengths were obtained by the direct sequence results from Folmer's universal primer set (Fig. 2). One of the nucleotide sequences was of standard length (658 bp; e.g. acc. no. LC528377, *H. kolthoffi*) and the other was a slightly shorter sequence (652 bp; e.g. acc. no. LC528370, *H. distanti distanti*, acc. no. LC528374 *H. d. hokkensis*). This shorter sequence (652 bp) was considered to be a pseudogene, and not the targeted sequence of the mtDNA COI region for this analysis. On the other hand, *H. d. hokkensis* has four haplotypes, and one of them had exactly the same sequence as the haplotype of *H. d. distanti*. The haplotype of *H. kolthoffi* is a singleton. In addition, some "double peaks" were detected in two specimens (acc. no. LC528370 and LC528374) amplified by the universal primer set. For sites where double peaks were detected, the base with the larger peak was used for sequence comparison (Fig. 2).

We counted matches and mismatches to calculate the similarity between these different length nucleotide sequences, and it was relatively high (*ca*. 85%). However, the true COI sequence and putative pseudogene sequences were not monophyletic. Within the short sequences, the "stop codon" and "indel" were detected (Fig. 3). The phylogenetic analysis of these data sets is shown in Fig. 4. The short-typed sequences of *H. d. distanti* and *H. d. hokkensis* (acc. no.

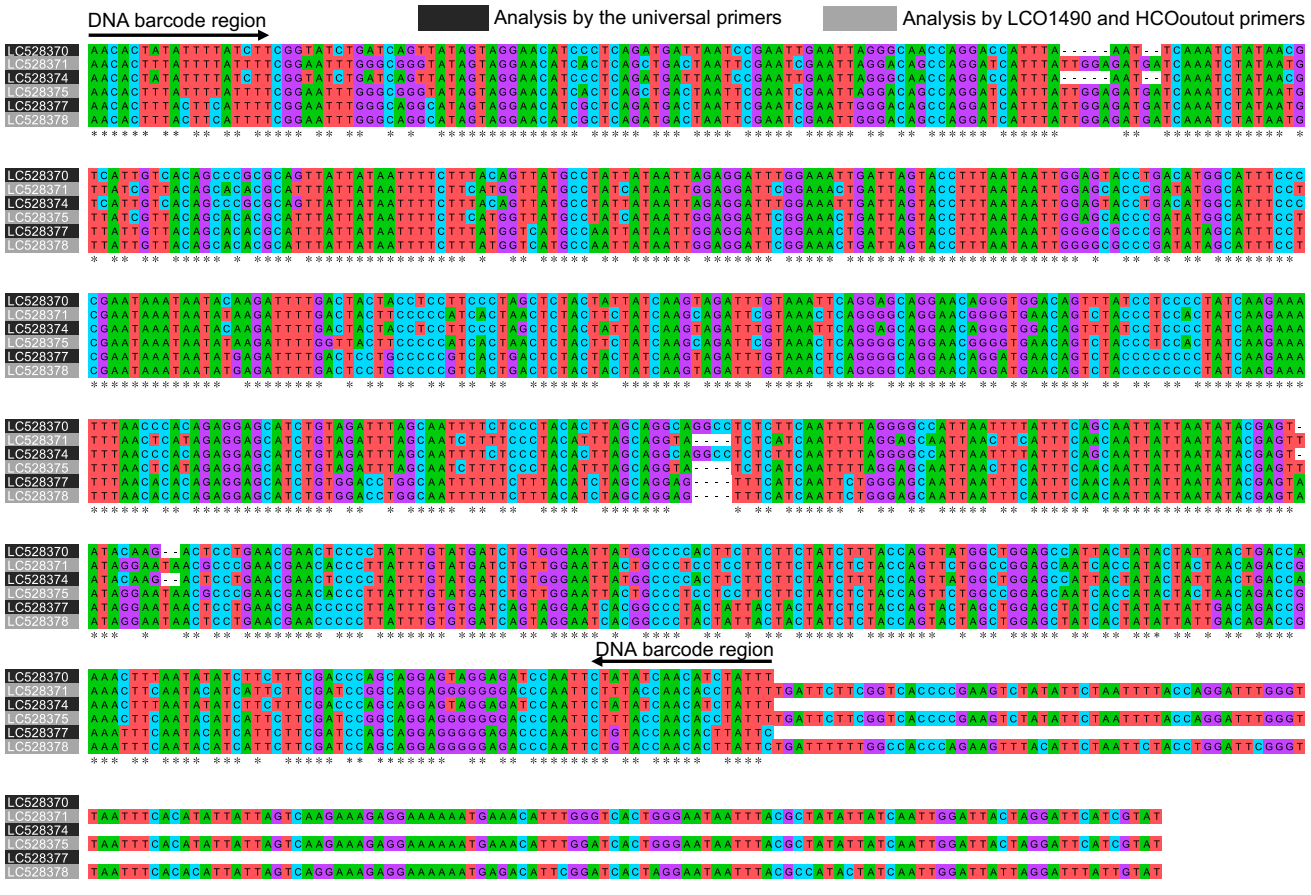


Fig. 2 Alignment of the DNA sequences of the mtDNA COI region of *Hesperocorixa* aquatic bugs and associated pseudogenes. Numbers in the sequence names are unique specimen numbers. Three sequences with black backgrounds were amplified with Folmer's universal primer set, and three sequences with gray backgrounds were amplified with the LCO1490 and HCOoutout primer set. Matching sites between all sequences are indicated with an asterisk. The sequences of the LC528370 and LC528371 are the results of analyzing the same sample with universal primer set (LC528370) and LCO1490 and HCOoutout primer set (LC528371). The same applies to the relationship between LC528374-75 and LC528377-78

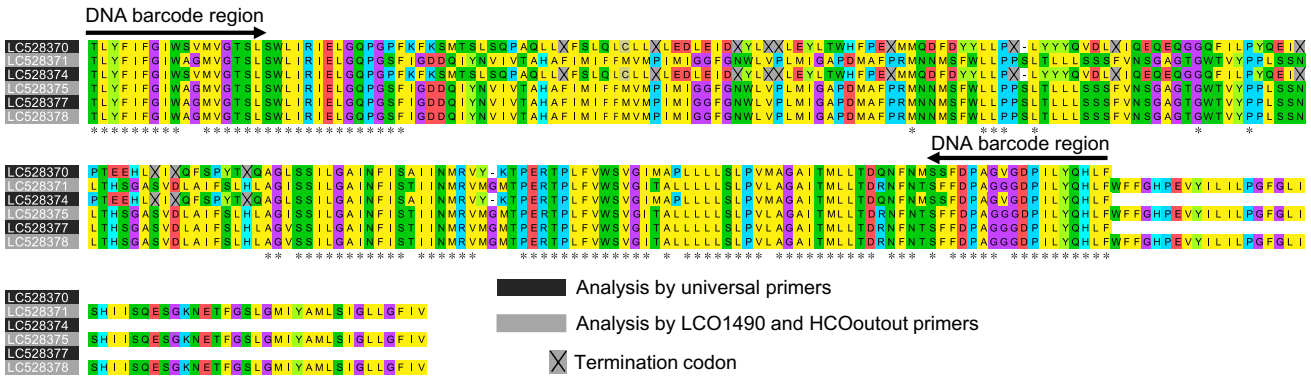


Fig. 3 Alignment of the protein sequences translated from the mtDNA COI sequences. Numbers in the sequence names are unique specimen numbers. Three sequences with black backgrounds were amplified with Folmer's universal primer set, and three sequences with gray backgrounds were amplified with the LCO1490 and HCOoutout primer set. Matching sites between all sequences are indicated with asterisks. The sequences of the LC528370 and LC528371 are the results of analyzing the same sample with universal primer set (LC528370) and LCO1490 and HCOoutout primer set (LC528371). The same applies to the relationship between LC528374-75 and LC528377-78

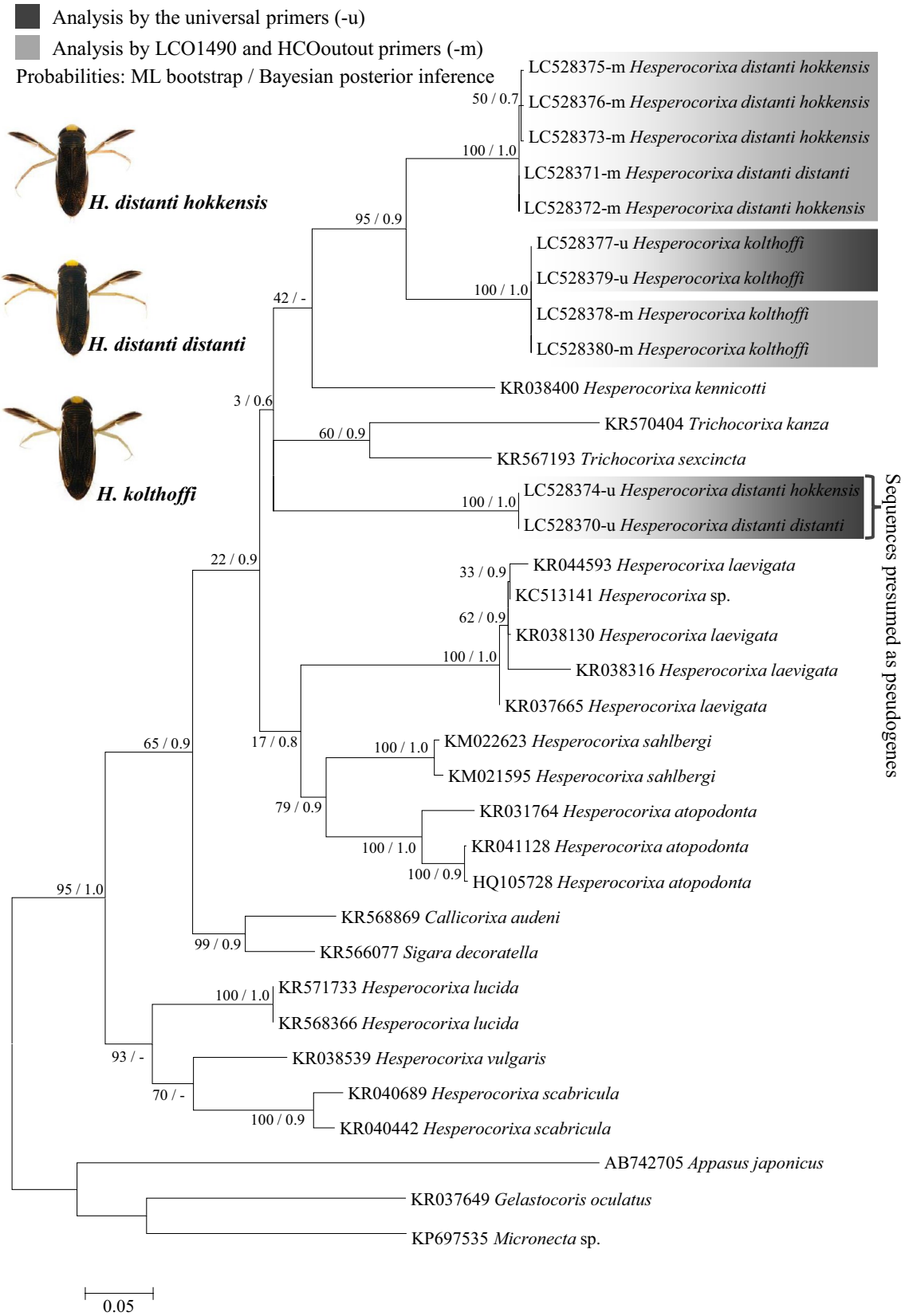
LC528370, LC528374) analyzed with Folmer's universal primer set were positioned outside the clade consisting of the same species and closely related species of aquatic Heteropterans (i.e., the short sequences were outside the clade of this species group).

The PCR product amplified with the LCO1490 and HCOoutout primer set included the same-length sequence (658 bp) as the ordinary COI region. These results revealed that in this species, at least two regions were amplified by Folmer's universal primer set. Since one of these was not detected by long PCR analysis, these results suggest that there may be a pseudogene outside the mitochondrial genome, or at the very least, outside the original COI region.

The existence of such "pseudogene(s)" in the mtDNA COI region is widely known in some insect orders [e.g., Diptera, Hymenoptera and Lepidoptera (Hazkani-Covo et al. 2010); Orthoptera (e.g. Song et al. 2014); Psocoptera (Chen et al. 2014); Coleoptera (King et al. 2015); Hemiptera *s. str.* (Tay et al. 2017)]. However, there has been no such report among Heteropterans (Heteroptera *s. str.*); therefore, this is the first time this has been reported. These findings highlight the need to consider the possible presence of pseudogenes when adopting the DNA barcoding method to identify species in this group. We think the "HCOoutout" primer set for DNA barcoding of Heteropterans will be an effective way to avoid contamination by pseudogenes.

In addition, DNA-based identification has recently become a popular method in the other types of studies: for example, an increasing number of studies have reported rapid investigation of biological community structures in a habitat by means of metabarcoding and/or metagenomic analysis. More recently, environmental DNA (eDNA) analyses that can investigate a taxonomic group of organisms inhabiting ponds, lakes, and rivers by merely sampling water from them are now frequently carried out (Minamoto et al. 2012; Bista et al. 2017; Doi et al. 2017). To carry out eDNA analyses more reliably and comprehensively, it is very important to accumulate more error-free DNA sequence information in the various genetic regions of the target taxa (e.g. the mtDNA COI, 16S rRNA, and cyt b region, and the nuclear histone H3 region).

Fig. 4 Phylogenetic tree reconstructed from the COI gene (658 bp)▶ and the pseudogenes (652 bp) of Corixidae. Maximum-likelihood bootstrap confidence and Bayesian posterior inference are indicated at nodes. Sequence names are the GenBank accession number plus taxon name. Primer sets used to amplify the sequences are indicated with suffixes as -u (Folmer's universal) and -m (modified). Corixid pictures were taken by T. Mitamura, one of the authors (*H. d. hokkensis*), and by Mr. Kei Hirasawa (*H. d. distanti* and *Hesperocorixa kolthoffi*). The sequences of the LC528370 and LC528371 are the results of analyzing the same sample with universal primer set (LC528370) and LCO1490 and HCOoutout primer set (LC528371). The same applies to the relationship between LC528374-75 and LC528377-78


Analysis by the universal primers (-u)
Analysis by LCO1490 and HCOoutout primers (-m)
Probabilities: ML bootstrap / Bayesian posterior inference
H. distanti hokkensis
H. distanti distanti
H. kolthoffi
LC528375-m Hesperocorixa distanti hokkensis
LC528376-m Hesperocorixa distanti hokkensis
LC528373-m Hesperocorixa distanti hokkensis
LC528371-m Hesperocorixa distanti distanti
LC528372-m Hesperocorixa distanti hokkensis
LC528377-u Hesperocorixa kolthoffi
LC528379-u Hesperocorixa kolthoffi
LC528378-m Hesperocorixa kolthoffi
LC528380-m Hesperocorixa kolthoffi
KR038400 Hesperocorixa kennicotti
KR570404 Trichocorixa kanza
KR567193 Trichocorixa sexcincta
LC528374-u Hesperocorixa distanti hokkensis
LC528370-u Hesperocorixa distanti distanti
Sequences presumed as pseudogenes
KR044593 Hesperocorixa laevigata
KC513141 Hesperocorixa sp.
KR038130 Hesperocorixa laevigata
KR038316 Hesperocorixa laevigata
KR037665 Hesperocorixa laevigata
KM022623 Hesperocorixa sahlbergi
KM021595 Hesperocorixa sahlbergi
KR031764 Hesperocorixa atopodonta
KR041128 Hesperocorixa atopodonta
HQ105728 Hesperocorixa atopodonta
KR568869 Callicorixa audeni
KR566077 Sigara decoratella
KR571733 Hesperocorixa lucida
KR568366 Hesperocorixa lucida
KR038539 Hesperocorixa vulgaris
KR040689 Hesperocorixa scabricula
KR040442 Hesperocorixa scabricula
AB742705 Appasus japonicus
KR037649 Gelastocoris oculatus
KP697535 Micronecta sp.
50 / 0.7
100 / 1.0
95 / 0.9
100 / 1.0
42 / -
3 / 0.6
60 / 0.9
100 / 1.0
22 / 0.9
33 / 0.9
62 / 0.9
100 / 1.0
65 / 0.9
17 / 0.8
100 / 1.0
79 / 0.9
100 / 1.0
100 / 0.9
95 / 1.0
99 / 0.9
100 / 1.0
93 / -
70 / -
100 / 0.9
0.05

Acknowledgements We express our thanks to Mr. Kei Hirasawa (Aquamarine Inawashiro Kingfishers Aquarium) for providing two corixids pictures. We are indebted to Mr. Kenji Takashino and Mr. Hirofumi Fujimoto, for their cooperation with the field research and collection of specimens. We are sincerely grateful to Dr. Catherine Docherty, Shinshu University and Birmingham University, for her many valuable suggestions for this study and critical reading of the manuscript. This study was supported by JSPS KAKENHI (JP16K14807 to KT), and grants from the River Environment Fund (27-1215-013, 2017-5211-025 to KT) of River and Watershed Environment Management.

References

Bensasson D, Zhang D, Hartl D, Hewitt G (2001) Mitochondrial pseudogenes: evolution's misplaced witnesses. Trends Ecol Evol 16:314–321

Bista I, Carvalho GR, Walsh K, Seymour M, Hajibabaei M, Lallias D, Christmas M, Creer S (2017) Annual time-series analysis of aqueous eDNA reveals ecologically relevant dynamics of lake ecosystem biodiversity. Nat Commun 8:14087

Bouckaert R, Vaughan TG, Barido-Sottani J, Duchêne S, Fourment M, Gavryushkina A, Heled J, Jones G, Kühnert D, De Maio N, Matschiner M, Mendes FK, Müller NF, Ogilvie HA, Du Plessis L, Popinga A, Rambaut A, Rasmussen D, Siveroni I, Suchard MA, Wu C-H, Xie D, Zhang C, Stadler T, Drummond AJ (2019) BEAST 2.5: an advanced software platform for Bayesian evolutionary analysis. PLoS Comp Biol 15:e1006650

Chen S, Wei D, Shao R, Shi J, Dou W, Wang J (2014) Evolution of multipartite mitochondrial genomes in the booklice of the genus *Liposcelis* (Psocoptera). BMC Genomics 15:861

Clouse RM, Wheeler WC (2014) Descriptions of two new, cryptic species of *Metasiro* (Arachnida: Opliones: Cyphophthalmi: Neogoveidae) from South Carolina, USA, including a discussion of mitochondrial mutation rates. Zootaxa 3814:177–201

Collins RA, Cruickshank RH (2013) The seven deadly sins of DNA barcoding. Mol Ecol Resour 13:969–975

Doi H, Katano I, Sakata Y, Souma R, Kosuge T, Nagano M, Ikeda K, Yano K, Tojo K (2017) Detection of an endangered aquatic heteropteran using environmental DNA in a wetland ecosystem. Roy Soc Open Sci 4:170568

Felsenstein J (1981) Evolutionary trees from DNA sequences: a maximum likelihood approach. J Mol Evol 17:368–376

Felsenstein J (1985) Confidence limits on phylogenies: an approach using the bootstrap. Evolution 39:783–791

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. Mol Marine Biol Biotechnol 3:294–299

Geller J, Meyer C, Parker M, Hawk H (2013) Redesign of PCR primers for a mitochondrial cytochrome *c* oxidase subunit I for marine invertebrates and application in all-taxa biotic surveys. Mol Ecol Resour 13:851–861

Hazkani-Covo E, Zeller RM, Martin W (2010) Molecular poltergeists: mitochondrial DNA copies (*numts*) in sequenced nuclear genomes. PLoS Genet 6:e1000834

Hebert PDN, Ratnasingham S, deWaard JR (2003) Barcoding animal life: cytochrome c oxidase subunit 1 divergences among closely related species. Proc R Soc Lond B 270:S96–S99

King JK, Riegler M, Thomas RG, Spooner-Hart RN (2015) Phylogenatic placement of Australian carrion beetles (Coleoptera: Silphidae). Austral Entomol 54:366–375

Minamoto T, Yamanaka H, Takahara T, Honjo MN, Kawabata Z (2012) Surveillance of fish species composition using environmental DNA. Limnology 13:193–197

Okamiya H, Sugawara H, Nagano M, Poyarkov NA (2018) An integrative analysis reveals a new species of lotic *Hynobius* salamander from Japan. PeerJ 6:e5084

Pickett KM, Carpenter JM, Wheeler WC (2006) Systematics of *Polistes* (Hymenoptera: Vespidae), with a phylogenetic consideration of Hamilton's haplodiploidy hypothesis. Ann Zool Fennici 43:390–406

Prendini L, Weygoldt P, Wheeler WC (2005) Systematics of the Damon variegatus group of Asian whip spiders (Chelicerata: Amblypygi): evidence from behaviour, morphology and DNA. Org Div Evol 5:203–236

Ratnasingham S, Hebert PDN (2007) BOLD: the barcode of life data system (http://www.barcodinglife.org). Mol Ecol Notes 7:355–364

Ratnasingham S, Hebert PDN (2013) A DNA-based registry for all animal species: the barcode index number (BIN) system. PLoS ONE 8:e66213

Saito R, Tojo K (2016a) Complex geographic- and habitat-based niche partitioning of an East Asian habitat generalist mayfly *Isonychia japonica* (Ephemeroptera: Isonychiidae) with reference to differences in genetic structure. Freshw Sci 35:712–723

Saito R, Tojo K (2016b) Comparing spatial patterns of population density, biomass, and genetic diversity patterns of the habitat generalist mayfly *Isonychia japonica* Ulmer (Ephemeroptera: Isonychiidae) in the Chikuma-Shinano River basin. Freshw Sci 35:724–737

Saito R, Jo J, Sekiné K, Bae YJ, Tojo K (2016) Phylogenetic analyses of the isonychiid mayflies (Ephemeroptera: Isonychiidae) in the northeast palearctic region. Entomol Res 46:246–259

Saito R, Kato S, Kuranishi RB, Nozaki T, Fujino T, Tojo K (2018) Phylogeographic analyses of the *Stenopsyche* caddisflies (Trichoptera, Stenopsychidae) of the Asian Region. Freshw Sci 37:562–572

Sanchez JA, Cassis G (2018) Towards solving the taxonomic impasse of the biocontrol plant bug subgenus *Dicyphus* (*Dicyphus*) (Insecta: Heteroptera: Miridae) using molecular, morphometric and morphological partitions. Zool J Linn Soc 184:330–406

Song H, Buhay JE, Crandall KA (2008) Many species in one: DNA barcoding overestimates the number of species when nuclear mitochondrial pseudogenes are coamplified. PNAS 105:13486–13491

Song H, Moulton MJ, Whiting MF (2014) Rampant nuclear insertion of mtDNA across diverse lineages within Orthoptera (Insecta). PLoS ONE 9:e110508

Suzuki T, Tanizawa T, Sekiné K, Kunimi J, Tojo K (2013) Morphological and genetic relationship of two closely related giant water bugs: *Appasus japonicas* Vuillefroy and *Appasus major* Esaki (Heteroptera: Belostomatidae). Biol J Linn Soc 110:615–643

Suzuki T, Kitano T, Tojo K (2014) Contrasting genetic structure of closely related giant water bugs: phylogeography of *Appasus japonicaus* and *Appasus major* (Insecta: Heteroptera, Belostmatidae). Mol Phylogenet Evol 72:7–16

Suzuki T, Suzuki N, Tojo K (2019) Parallel evolution of an alpine type ecomorph in a scorpionfly: Independent adaptation to high-altitude environments in multiple mountain locations. Mol Ecol 28:3225–3240

Takenaka M, Tojo K (2019) Ancient origin of a dipteromimid mayfly family endemic to the Japanese Islands and its genetic differentiation across tectonic faults. Biol J Linn Soc 126:555–573

Takenaka M, Tokiwa T, Tojo K (2019) Concordance between molecular biogeography of *Dipteromimus tipuliformis* and geological history in the local fine scale (Ephemeroptera, Dipteromimidae). Mol Phylogenet Evol 139:106547

Tamura K, Stecher G, Peterson D, Filipski A, Kumar S (2013) MEGA6: molecular evolutionary genetics analysis version 6.0. Mol Biol Evol 30:2725–2729

Tanabe AS (2007) Kakusan: a computer program to automate the selection of a nucleotide substitution model and the configuration of a mixed model on multilocus data. Mol Ecol Notes 7:962–964

Tay WT, Elfekih S, Court LN, Gordon KH, Delatte H, De Barro PJ (2017) The trouble with MEAM2: implications of pseudogenes on species delimitation in the globally invasive *Besimia tabaci* (Hemiptera: Aleyrodidae) cryptic species complex. Genome Biol Evol 9:2732–2738

Tojo K, Sekiné K, Takenaka M, Isaka Y, Komaki S, Suzuki T, Schoville SD (2017) Species diversity of insects in Japan: their origins and diversification processes. Entomol Sci 20:357–381

Yano K, Takenaka M, Tojo K (2019) Genealogical position of Japanese populations of the globally distributed mayfly *Cloeon dipterum* and related species (Ephemeroptera, Baetidae): a molecular phylogeographic analysis. Zool Sci 36:479–489