Development of mitochondrial DNA cytochrome *c* oxidase subunit I primer sets to construct DNA barcoding library using next-generation sequencing

Seikan Kurata[‡], Shota Mano[§], Naoyuki Nakahama[|], Shun K Hirota[¶], Yoshihisa Suyama[#], Motomi Ito[¤]

‡ Tomakomai Experimental Forest, Hokkaido University, Tomakomai, Japan
§ Faculty of Bioresource Sciences, Akita Prefectural University, Akita, Japan
| Institute of Natural and Environmental Sciences, University of Hyogo, hyogo, Japan
¶ Botanical Gardens, Osaka Metropolitan University, Osaka, Japan
# Graduate School of Agricultural Science, Tohoku University, Miyagi, Japan
¤ Komaba museum, University of Tokyo, Tokyo, Japan

Abstract

Insects are one of the most diverse eukaryotic groups on the planet, with one million or more species present, including those yet undescribed. The DNA barcoding system has been developed, which has aided in the identification of cryptic species and undescribed species. The mitochondrial cytochrome *c* oxidase I region (mtDNA COI) has been utilised for the barcoding analysis of insect taxa. Thereafter, next-generation sequencing (NGS) technology has been developed, allowing for rapid acquisition of massive amounts of sequence data for genetic analyses. Although NGS-based PCR primers designed to amplify the mtDNA COI region have been developed, their target regions were only a part of COI region and/or there were taxonomic bias for PCR amplification. As the mtDNA COI region is a traditional DNA marker for the DNA barcoding system, modified primers for this region would greatly contribute to taxonomic studies. In this study, we redesigned previously developed PCR primer sets that targetted the mtDNA COI barcoding region to

improve amplification efficiency and to enable us to conduct sequencing analysis on NGS. As a result, the redesigned primer sets achieved a high success rate (> 85%) for species examined in this study, covering four insect orders (Coleoptera, Lepidoptera, Orthoptera and Odonata). Thus, by combining the primers with developed primer sets for 12S or 16S rRNA regions, we can conduct more detailed taxonomic, phylogeographic and conservation genetic studies using NGS.

Keywords

biodiversity, DNA barcoding, mtDNA COI, insect, next-generation sequencing

Introduction

Biodiversity can be categorised into three levels: species diversity (richness), genetic diversity and ecosystem diversity. Species richness provides a straightforward method for describing community and regional diversity (Magurran 1988, Gotelli and Colwell 2001). An estimated 10 million species live on Earth (May 1988, May 1990, Stork 1993, Gaston 1991, Costello et al. 2011, Costello et al. 2013), although the exact number is unknown. In recent years, climate change due to anthropogenic effects has increased the risk of extinction for many living species. In addition, many species were likely exterminated before being described, even if the number of living organisms is underestimated. Insects are amongst the most diverse eukaryotic groups on the planet and at least one million species have been described (Stork 2018, Takenaka et al. 2023). Nevertheless, the 1–2% of insect species may be cryptic (Stork 2018) and many new and cryptic species have been described recently (e.g. Takenaka et al. (2023)).

The DNA barcoding system was developed to identify species through DNA sequencing (Hebert et al. 2003, Hebert and Gregory 2005). This innovative system facilitates rapid, accurate, automatable species identification using short standardised gene regions as internal species tags (Hebert and Gregory 2005). DNA barcoding and DNA sequencing approaches have contributed to the detection of cryptic and undescribed species (e.g. Burns et al. (2008), Rebijith et al. (2013)).

Hebert et al. (2003) identified the mitochondrial cytochrome *c* oxidase I (mtDNA COI) region as the core target region for DNA barcoding because it robustly detects moderate genetic differences amongst species as a marker in taxonomic and phylogenetic studies. The 12S and 16S ribosomal RNA loci are also used to identify specimens (e.g. Marquina et al. (2018), Takenaka et al. (2023)). Using genetic differences in the COI region, phylogeographic and conservation genetic studies of many insect species have been conducted (e.g. Buckley et al. (2009), Nakahama et al. (2018), Çıplak et al. (2022)), allowing access to substantial reference sequences in the International Nucleotide Sequence Database Collaboration (INSDC), which involves the DNA Data Bank of Japan (DDBJ), European Molecular Biology Laboratory (EMBL) and the National Center for Biotechnology Information (NCBI). The COI region is more suitable for DNA barcoding than

other mtDNA loci, specifically the 658-bp sequence amplified by the primer pair LCO1490 and HCO2198 established by Folmer et al. (1994), which has been widely used for various insect taxa. With the development of next-generation sequencing (NGS), it has become possible to obtain large numbers of short sequence reads (ca. 300 bp) rapidly. Therefore, the modification of primer sets to amplify the COI region efficiently across a diverse range of insect taxa and to conduct NGS sequencing would enhance reference sequences in the region for DNA barcoding.

Although primer sets have been developed to amplify the COI region for NGS-based analysis (i.e. DNA metabarcoding; for example, Leray et al. (2013); Meier et al. (2015); Elbrecht and Leese (2017)), contributing to the identification of insect species, these primer sets are not suitable for enhancing reference sequences in the COI region because they amplify only part of the COI region. Furthermore, insect species have not been identified for which these primer sets are suitable. Although Zhou et al. (2013), Liu et al. (2017) and Yang et al. (2020) developed new NGS-based pipelines of the mtDNA COI region, the specimens sampled in each analysis showed taxonomic bias, i.e. specimens were selected only from Lepidoptera, Diptera and Neuroptera, respectively. Therefore, there is a need to develop primer sets for NGS-based analysis that can amplify the full length of the COI barcoding region and are applicable to many insect taxa. Novel primer sets for NGS-based analysis that amplify the 12S and 16S ribosomal RNA loci have been developed and used to detect cryptic species in some insect taxa, including 11 orders, 42 families and 70 species (Takenaka et al. 2023). More recently, multiplexed phylogenetic marker sequencing (MPM-seq) was developed (Suyama et al. 2021), to enable the simultaneous detection multi-locus sequences. Thus, the development of primer sets targetting insect mtDNA could contribute to research on taxonomy, phylogeography and conservation genetics. In this study, we redesigned previously developed primer sets that targeted the entire COI barcoding region, which will enhance COI reference sequences for insect taxa.

Materials and Methods

Development of primer sets

To identify polymorphic sites in the primer annealing regions of different insect taxa, we downloaded COI sequences from the NCBI database (Suppl. material 1; 33 species, 29 families and 14 orders). These 33 sequences were aligned using MAFFT v.7.310-1 (Katoh 2002, Katoh and Standley 2013). Polymorphic sites within the COI region across the 33 species were visualised using MEGA-X (Kumar et al. 2018), which provided information for the modification of the primer pair to include mixed bases (e.g. A/G: R, A/T/C:H). Given that the total sequence amplified by this primer pair exceeds 500 bp, the resultant sequence reads are unsuitable for NGS-based sequence analysis. To address this problem, we integrated an intermediate primer pair, mlCOIintF and mlCOIintR (Leray et al. 2013), to distinguish the first half of the COI region in this study (1-319 bp, Fig. 1, Table 1). An approximately 240-bp portion of the COI region, which is moderately conserved amongst species (Leray et al. 2013), was selected as the annealing site for the forward primer of the second half of the COI region (262–658 bp, Fig. 1, Table 1). The sites were also modified

as described above, giving rise to two primer pairs; thus, modified LCO1490 and mlCOIintR amplify the first half of the COI region and the new forward primer COmfd_F and modified HCO2198 amplify the second half of the COI region (Table 1), with expected amplification products of approximately 350 bp.

Table 1.

Information of newly-modified primer sets in this study.

Primer name	Primer sequence (5' –3')[1]	Target region
LCO1490	TCWACWAAYCAYAARGAYATYGG	1–319 bp of the mtDNA COI
COmfd_R	GGDGGRTANAHHGTTCAHCCNGTHCC	
COmfd_F	CCNCGRHTRAAYAAYATRAGWTTYTG	262–658 bp of the mtDNA COI
HCO2198	ACTTCDGGRTGNCCAAARAAYCA	
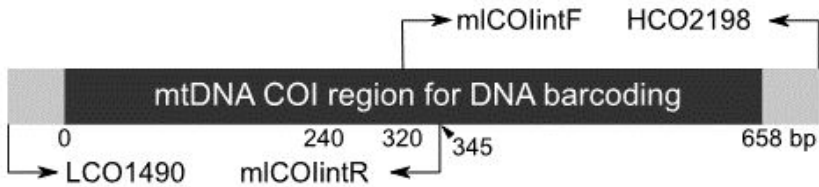
Figure 1. doi

Imaged position of the primers in this study.

Sample collection

Between April to October 2022, we collected 96 specimens comprising 96 species, 48 families and 11 orders (Table 2). Each specimen was preserved by freezing.

Table 2.

Specimen samples and the result of sequencing analysis.

Specimen samples information			Number of reads[2]		Sampling location		
Order	Family	species	1-319 bp	262-658 bp	Locality name	Latitude	Longitude
Coleoptera	Carabidae	*Carabus vanvolxemi* Putzeys, 1875	2839	3520	Japan: Mt. Moriyoshi	39.99°N	140.60°E

Specimen samples information			Number of reads*2		Sampling location		
Order	Family	species	1-319 bp	262-658 bp	Locality name	Latitude	Longitude
	Carabidae	*Carabus insulicola* Chaudoir, 1869	4035	2954	Japan: Akita, Yamamoto, Mitane	40.06°N	140.11°E
	Carabidae	*Craspedonotus tibialis* Schaum, 1863	149	102	Japan: Akita, Akita, Shimoshinjo	39.80°N	140.05°E
	Carabidae	*Cylindera ovipennis* Bates, 1883	131	215	Japan: Akita, Akita, Nibetsu	39.80°N	140.22°E
	Carabidae	*Scarites terricola* Bonelli, 1813	6141	2384	Japan: Akita, Minamiakita, Ogata	40.01°N	139.95°E
	Carabidae	*Pterostichus* Bonelli, 1810	3746	1547	Japan: Akita, Akita, Shimoshinjo	39.80°N	140.05°E
	Ceranbycidae	*Acalolepta luxuriosa* Bates, 1873	130	24	Japan: Akita, Minamiakita, Ogata	40.01°N	139.95°E
	Ceranbycidae	*Anoplophora chinensis* Forster, 1771	18024	15712	Japan: Akita, Akita, Nibetsu	39.80°N	140.22°E
	Ceranbycidae	*Batocera lineolata* Chevrolat, 1852	7707	2946	Japan: Akita, Minamiakita, Ogata	40.01°N	139.95°E
	Ceranbycidae	*Prionus insularis* Motschulsky, 1857	980	610	Japan: Akita, Minamiakita, Ogata	40.01°N	139.95°E
	Coccinellidae	*Aiolocaria hexaspilota* Hope, 1831	23	24	Japan: Aktia, Akita, Toyoiwatoyomaki	39.64°N	140.11°E
	Dytiscidae	*Cybister chinensis* Motschulsky, 1854	109	2	Japan: Akita, Noshiro, Asanai	40.15°N	140.04°E
	Dytiscidae	*Graphoderus adamsii* Clark, 1864	57	9	Japan: Akita, Noshiro, Asanai	40.15°N	140.04°E
	Dytiscidae	*Rhantus suturalis* MacLeay, 1825	-	-	Japan: Akita, Akita, Shimoshinjo	39.80°N	140.05°E
	Hydrophilidae	*Hydrochara affinis* Sharp, 1873	8020	7971	Japan: Akita, Minamiakita, Ogata	40.01°N	139.95°E
	Hydrophilidae	*Hydrophilus acuminatus* Motschulsky, 1854	8335	11208	Japan: Akita, Minamiakita, Ogata	40.01°N	139.96°E
	Lucanidae	*Dorcus rectus* Motschulsky, 1858	326	835	Japan: Akita, Minamiakita, Ogata	40.01°N	139.95°E

Specimen samples information			Number of reads*2		Sampling location		
Order	Family	species	1-319 bp	262-658 bp	Locality name	Latitude	Longitude
	Lucanidae	*Prosopocoilus inclinatus* Motschulsky, 1857	1184	2856	Japan: Akita, Minamiakita, Ogata	40.01°N	139.95°E
	Rhynchophoridae	*Sipalinus gigas* Fabricius, 1775	2783	1162	Japan: Akita, Minamiakita, Ogata	40.01°N	139.95°E
	Scarabaeidae	*Anomala schoenfeldti* Ohaus, 1915	536	337	Japan: Akita, Akita, Shimoshinjo	39.80°N	140.05°E
	Scarabaeidae	*Polyphylla albolineata* Motschulsky, 1861	1525	3223	Japan: Akita, Minamiakita, Ogata	40.01°N	139.95°E
	Scarabaeidae	*Protaetia orientalis* Gory & Percheron, 1833	5218	7267	Japan: Akita, Akita, Shimoshinjo	39.80°N	140.05°E
	Scarabaeidae	*Trypoxylus dichotomus* Linnaeus, 1771	1673	1417	Japan: Akita, Minamiakita, Ogata	40.01°N	139.95°E
	Tenebrionidae	*Cryphaeus amurensis* Heyden, 1884	-	-	Japan: Akita, Yamamoto, Mitane	40.06°N	140.11°E
Hymenoptera	Apidae	*Apis mellifera* Linnaeus, 1758	66	1727	Japan: Akita, Akita, Shimoshinjo	39.80°N	140.05°E
	Apidae	*Bombus diversus* Smith, 1869	608	727	Japan: Aomori, Shimokita, Higashidori	41.37°N	141.44°E
	Apidae	*Bombus terrestris* Linnaeus, 1758	-	-	Japan: Hokkaido, Hakodate, Bandai	41.79°N	140.73°E
	Pompilidae	*Auplopus carbonarius* Scopoli, 1763	-	-	Japan: Mt. Moriyoshi	39.99°N	140.60°E
	Scoliidae	*Scolia histrionica* Smith, 1873	-	-	Japan: Aomori, Shimokita, Higashidori	41.37°N	141.44°E
	Vespidae	*Oreumenes decoratus* Smith, 1852	443	268	Japan: Yamagata, Tsuruoka, Obari	38.56°N	139.86°E
	Vespidae	*Parapolybia indica* Saussure, 1854	1612	1106	Japan: Akita, Akita, Kanaashi	39.81°N	140.07°E
	Vespidae	*Polistes chinensis antennalis* Yamane, 1972	26	9	Japan: Akita, Akita, Shimoshinjo	39.80°N	140.05°E

Specimen samples information			Number of reads*2		Sampling location		
Order	Family	species	1-319 bp	262-658 bp	Locality name	Latitude	Longitude
	Vespidae	*Polistes jokahamae* Radoszkowski, 1887	3867	3411	Japan: Akita, Akita, Shimoshinjo	39.80°N	140.05°E
	Vespidae	*Polistes rothneyi* Cameron, 1900	626	338	Japan: Akita, Akita, Shimoshinjo	39.80°N	140.05°E
	Vespidae	*Vespa analis* Fabricius, 1775	1090	868	Japan: Akita, Minamiakita, Ogata	40.01°N	139.95°E
	Vespidae	*Vespa ducalis* Smith, 1852	1572	1669	Japan: Yamagata, Shinjo, Tsunozawa	38.71°N	140.26°E
	Vespidae	*Vespa mandarinia* Smith, 1852	7871	12385	Japan: Akita, Akita, Shimoshinjo	39.80°N	140.05°E
Lepidoptera	Agaristidae	*Sarbanissa subflava* Moore, 1877	4388	2931	Japan: Akita, Minamiakita, Ogata	40.01°N	139.95°E
	Callidulidae	*Pterodecta felderi* Bremer, 1864	3011	8868	Japan: Yamagata, Tsuruoka, Obari	38.56°N	139.86°E
	Drepanidae	*Hypsomadius insignis* Butler, 1877	858	1338	Japan: Akita, Minamiakita, Ogata	40.01°N	139.95°E
	Heterogeneidae	*Parasa sinica* Moore, 1877	3105	1538	Japan: Akita, Minamiakita, Ogata	40.01°N	139.95°E
	Lycaenidae	*Lycaena phlaeas* Linnaeus, 1761	48	-	Japan: Akita, Akita, Shimoshinjo	39.80°N	140.05°E
	Lycaenidae	*Hypaurotis fujisanus* Matsumura, 1910	1124	307	Japan: Yamagata, Tsuruoka, Obari	38.56°N	139.86°E
	Lymantriidae	*Euproctis similis* Fuessly, 1775	36	19	Japan: Akita, Minamiakita, Ogata	40.01°N	139.95°E
	Notodontidae	*Pterostoma gigantina* Staudinger, 1892	1110	65	Japan: Akita, Minamiakita, Ogata	40.01°N	139.95°E
	Nymphalidae	*Nymphalis canace* Linnaeus, 1763	2009	1002	Japan: Aichi, Tahara, Nishiyama	34.60°N	137.05°E
	Nymphalidae	*Minois dryas* Scopoli, 1763	1000	503	Japan: Aomori, Kitatsugaru, Nakadomari	41.19°N	140.34°E
	Nymphalidae	*Neope niphonica* Butler, 1881	10303	8532	Japan: Mt. Moriyoshi	39.99°N	140.60°E
	Nymphalidae	*Ochlodes ochracea* Bremer, 1861	1295	2339	Japan: Iwate, Kunohe, Kunohe	40.14°N	141.39°E

Specimen samples information			Number of reads[*2]		Sampling location		
Order	Family	species	1-319 bp	262-658 bp	Locality name	Latitude	Longitude
	Nymphalidae	*Vanessa indica* Herbst, 1794	12816	5480	Japan: Yamagata, Tsuruoka, Wasada	38.57°N	139.56°E
	Papilionidae	*Papilio xuthus* Linnaeus, 1767	1019	1520	Japan: Aichi, Tahara, Nishiyama	34.60°N	137.05°E
	Papilionidae	*Parnassius citrinarius* Butler, 1866	8379	11203	Japan: Yamagata, Tsuruoka, Wasada	38.57°N	139.56°E
	Pieridae	*Colias erate* Esper, 1805	6181	1115	Japan: Akita, Yurihonjo, Tsuchiya	39.39°N	140.07°E
	Pieridae	*Pieris melete* Menetries, 1857	4541	6119	Japan: Iwate, Kunohe, Kunohe	40.14°N	141.39°E
	Saturniidae	*Saturnia japonica* Moore, 1872	4500	8871	Japan: Akita, Minamiakita, Ogata	40.01°N	139.95°E
	Sphindidae	*Agrius convolvuli* Linnaeus, 1758	5921	10323	Japan: Akita, Akita, Shimoshinjo	39.80°N	140.05°E
	Sphindidae	*Smerinthus planus* Walker, 1856	4596	703	Japan: Akita, Minamiakita, Ogata	40.01°N	139.95°E
Hemiptera	Acanthosomatidae	*Sastragala esakii* Hasegawa, 1959	137	33	Japan: Akita, Akita, Nibetsu	39.80°N	140.22°E
	Belostomatidae	*Appasus japonicus* Vuillefroy, 1864	201	641	Japan: Aktia, Akita, Toyoiwatoyomaki	39.64°N	140.11°E
	Belostomatidae	*Appasus major* Esaki, 1934	88	93	Japan: Aktia, Akita, Toyoiwatoyomaki	39.64°N	140.11°E
	Cicadellidae	*Bothrogonia ferruginea* Fabricius, 1787	150	26	Japan: Yamagata, Tsuruoka, Wasada	38.57°N	139.56°E
	Cicadidae	*Graptopsaltria nigrofuscata* Motschulsky, 1866	6556	7179	Japan: Akita, Minamiakita, Ogata	40.01°N	139.95°E
	Cicadidae	*Platypleura kaempferi* Fabricius, 1794	3304	3833	Japan: Akita, Minamiakita, Ogata	40.01°N	139.95°E
	Cicadidae	*Yezoterpnosia nigricosta* Motschulsky, 1866	17050	18002	Japan: Mt. Moriyoshi	39.99°N	140.60°E
	Coreidae	*Leptoglossus occidentalis* Heidemann, 1910	404	74	Japan: Akita, Akita, Shimoshinjo	39.80°N	140.05°E

Specimen samples information			Number of reads[*2]		Sampling location		
Order	Family	species	1-319 bp	262-658 bp	Locality name	Latitude	Longitude
	Corixidae	*Hesperocorixa hokkensis* Matsumura, 1905	-	-	Japan: Akita, Minamiakita, Ogata	40.01°N	139.95°E
	Gerridae	*Aquarius paludum* Fabricius, 1794	-	-	Japan: Akita, Akita, Shimoshinjo	39.80°N	140.05°E
	Notonectidae	*Notonecta triguttata* Motschulsky, 1861	621	416	Japan: Akita, Akita, Shimoshinjo	39.80°N	140.05°E
	Pentatomidae	*Palomena angulosa* Motschulsky, 1861	-	-	Japan: Akita, Akita, Shimoshinjo	39.80°N	140.05°E
	Pentatomidae	*Pentatoma japonica* Distant, 1882	-	8	Japan: Iwate, Kunohe, Kunohe	40.14°N	141.39°E
	Reduviidae	*Agriosphodrus dohrni* Stal, 1862	-	-	Japan: Akita, Akita, Shimoshinjo	39.80°N	140.05°E
	Reduviidae	*Ectrychotes andreae* Thunberg, 1784	65	-	Japan: Akita, Akita, Shimoshinjo	39.80°N	140.05°E
	Reduviidae	*Velinus nodipes* Uhler, 1860	-	-	Japan: Akita, Akita, Shimoshinjo	39.80°N	140.05°E
Orthoptera	Acrididae	*Acrida cinerea* Thunberg, 1815	5446	9947	Japan: Akita, Akita, Shimoshinjo	39.80°N	140.05°E
	Acrididae	*Aiolopus thalassinus* Fabricius, 1781	13476	7387	Japan: Akita, Akita, Shimoshinjo	39.80°N	140.05°E
	Acrididae	*Locusta migratoria* Linnaeus, 1758	2739	452	Japan: Aichi, Tahara, Nishiyama	34.60°N	137.05°E
	Acrididae	*Oedaleus infernalis* Saussure, 1884	1095	441	Japan: Akita, Akita, Shimoshinjo	39.80°N	140.05°E
	Gryllotalpidae	*Gryllotalpa orientalis* Burmeister, 1838	2842	3479	Japan: Akita, Minamiakita, Ogata	40.01°N	139.95°E
	Tettigoniidae	*Ruspolia dubia* Redtenbacher, 1891	271	317	Japan: Akita, Minamiakita, Ogata	40.01°N	139.96°E
Diptera	Asilidae	*Neoitamus angusticornis* Loew, 1858	-	-	Japan: Akita, Akita, Shimoshinjo	39.80°N	140.05°E

Specimen samples information			Number of reads[*2]		Sampling location		
Order	Family	species	1-319 bp	262-658 bp	Locality name	Latitude	Longitude
	Asilidae	*Neoitamus* Osten Sacken, 1878	16	5	Japan: Mt. Moriyoshi	39.99°N	140.60°E
	Asilidae	*Promachus yesonicus* Bigot, 1887	3003	2058	Japan: Akita, Akita, Shimoshinjo	39.80°N	140.05°E
	Tabanidae	*Tabanus chrysurus* Loew, 1858	144	79	Japan: Iwate, Morioka, Yabukawa	39.85°N	141.45°E
Odonata	Coenagrionidae	*Paracercion hieroglyphicum* Brauer, 1865	871	928	Japan: Akita, Katagami, Ten-nou	39.91°N	140.02°E
	Gomphidae	*Davidius nanus* Selys, 1869	1009	1060	Japan: Iwate, Kunohe, Kunohe	40.14°N	141.39°E
	Lestidae	*Lestes sponsa* Hansemann, 1823	31	3	Japan: Iwate, Morioka, Yabukawa	39.85°N	141.45°E
	Libellulidae	*Rhyothemis fuliginosa* Selys, 1883	2078	2880	Japan: Akita, Minamiakita, Ogata	40.01°N	139.95°E
	Libellulidae	*Sympetrum darwinianum* Selys, 1883	87	107	Japan: Akita, Minamiakita, Ogata	40.01°N	139.95°E
	Libellulidae	*Sympetrum frequens* Selys, 1883	999	1698	Japan: Akita, Akita, Sotoasahikawa	39.75°N	140.10°E
	Libellulidae	*Sympetrum infuscatum* Selys, 1883	268	712	Japan: Akita, Akita, Shimoshinjo	39.80°N	140.05°E
	Libellulidae	*Sympetrum kunckeli* Selys, 1884	449	1340	Japan: Akita, Akita, Shimoshinjo	39.80°N	140.05°E
Dermaptera	Labiduridae	*Labidura riparia* Pallas, 1773	-	39	Japan: Akita, Akita, Shimoshinjo	39.80°N	140.05°E
Neuroptera	Myrmeleontidae	*Myrmeleon bore* Tjeder, 1941	160	41	Japan: Akita, Minamiakita, Ogata	40.01°N	139.95°E
Mantodea	Mantidae	*Statilia maculata* Thunberg, 1784	1328	156	Japan: Akita, Minamiakita, Ogata	40.01°N	139.95°E
	Mantidae	*Tenodera aridifolia* Stoll, 1813	46	-	Japan: Akita, Akita, Shimoshinjo	39.80°N	140.05°E
Blattaria	Blattidae	*Periplaneta japonica* Karny, 1908	395	631	Japan: Akita, Akita, Shimoshinjo	39.80°N	140.05°E

PCR amplification and sequencing analysis

Genomic DNA was extracted using DNeasy Blood & Tissue Kits (QIAGEN, Hilden, Germany) and the total DNA concentration was quantified with a NanoDrop ND-1000 (Thermo Scientific, Waltham, USA). Polymerase chain reaction (PCR) was performed for each specimen according to the manufacturer's protocol, in a final volume of 10 µl that included 5–10 ng of DNA, 1.0 µl Ex Taq Buffer, 0.2 µmol/l of primers, 0.8 µl of dNTP mixture (2.5 mM of each dNTP), 2 U of Takara Ex Taq polymerase (Takara Bio, Otsu, Japan) and sterile distilled water up to 10 µl. The PCR thermal cycling conditions were an initial 1 min denaturation at 94°C; 35 cycles of 94°C for 30 s, 52°C for 30 s, 72°C for 1 min, with a final 20-min extension at 72°C. The PCR product was verified using a MultiNA microchip electrophoresis system (SHIMADZU, Kyoto, Japan). Before sequencing, the PCR products (i.e. LCO1490–COmfd_R and COmfd_F–HCO2198) were pooled for each specimen sample and all 96 samples were prepared for sequencing. Subsequent paired-end sequencing was conducted using 2 × 250 bp cycle run on an Illumina MiSeq Sequencer (Illumina, San Diego, USA) and with the MiSeq Reagent Nano Kit v.2 (500 cycles).

Assembling sequences and phylogenetic analysis

Data preprocessing, quality control and identification of representative sequences were conducted using Claident v.0.2.2019.05.10 (Tanabe and Toju 2013), as described by Suyama et al. (2021).

Before quality control and data analysis using Claident, non-demultiplexed fastq files (261 bp) were produced from BCL files using *bcl2fastq* v.1.8.4 (Illumina). During this step, non-demultiplexed fastq reads were sorted, based on index reads (index1: 9 bp, index2: 5 bp) and the last position of the raw reads was trimmed (--use-bases-mask Y260n,I9,I5,Y260n). Subsequently, the non-multiplexed fastq reads were demultiplexed using the Claident command *clsplitseq*, which specifies indices and primer sequences and the quality threshold of the index sequence was set to 30 (--minqualtag=30). As the modified primers included 0–3 Ns for accommodation, the option (--truncateN=enable) was used. Files output from *clsplitseq* were deposited in the DDBJ (accession no. DRA017438).

As paired-end sequencing was used, it was possible to identify overlaps between forward and reverse reads. The *clconcatpair* command with the --mode=OVL argument was used to generate concatenated reads from the forward and reverse sequences. Any low-quality reads were filtered out using the *clfilterseq* command with settings --maxplowequal=0.1 --minqual=27, to remove positions with quality lower than Q27.

Further cleaning to remove noisy and chimeric sequences was performed using the following *clcleanseqv* parameters: --derepmode=FULLLENGTH --primarymaxnmismatch=0 --secondarymaxnmismatch=1 --pnoisycluster=0.5. Representative sequences for each sample were identified using the *clclasseqv* command with a 99% identity threshold (--minident=0.99). However, the sequences output by *clclasseqv* remained separated into two regions.

Overlap was detected and the sequences from the two loci were merged using the EMBOSS programme (Rice et al. 2000). The merged sequences were aligned using MAFFT v.7.310-1, and phylogenetic analysis was conducted using the Maximum-Likelihood method in IQ-TREE (Nguyen et al. 2014). The best substitution model was selected using the ModelFinder Plus option (-m MFP) and the GTR+F+I+G4 model was identified as giving the best fit according to the Bayesian Information Criterion (BIC). Additionally, the ultrafast bootstrap approximation and Shimodaira–Hasegawa approximate likelihood ratio test (SH-aLRT) were set to 1000 replicates to assess branch reliability (-bb 1000 and -alrt 1000). Three Collembola sequences (accession nos.: JN970939.1, MF916630.1 and KY829298.1) were included as outgroup. Phylogenetic analysis using the neighbour-joining method (Saitou and Nei 1987) was also conducted with MEGA-X under the substitution model of the Jukes–Cantor model. The consensus trees were visualised and edited using FigTree v.1.4.3 (Rambaut 2016). A BLAST search was also conducted using reference sequences in GenBank.

Table 3.

Success rate of sequence analysis.

Order	No. of specimen samples[3]	No. of samples[4]	Success rate (%)
Coleoptera	24	22	91.7
Hymenoptera	14	10	71.4
Lepidoptera	20	19	95
Hemiptera	16	9	56.3
Orthoptera	6	6	100
Diptera	4	3	75
Odonata	9	8	88.9
Others	5	3	–
Total	96	80	82.3

Results and Discussion

We redesigned two primer sets (LCO1490–COmfd_R and COmfd_F–HCO2198) that amplified the DNA barcoding region of the mtDNA COI region of insect taxa (see Table 1). The original barcoding primers for the mtDNA COI region are unsuitable for NGS-based analysis due to their excessive sequence length. Therefore, we modified the original primers (Folmer et al. 1994) and designed internal primers such as mlCOIintF/R (Leray et al. 2013) to amplify lengths appropriate for NGS-based analysis. The sequence reads were divided into two parts: a 319-bp fragment from the PCR amplification products of LCO1490–COmfd_R and a 397-bp fragment from COmfd_F–HCO2198. Our PCR amplification trial and sequence analysis indicated that our modified primer sets successfully amplified the mtDNA COI region, demonstrating the effectiveness of these

primer sets for taxonomic, phylogeographic and conservation genetic studies of insect taxa.

Figure 2. doi

The Maximum-Likelihood (ML) phylogenetic tree, based of mtDNA COI region which originated from sequence reads of the modified primer sets. The numbers on the *major* branches represents bootstrap values from the ultrafast bootstrap replications and SH-aLRT methods, respectively. The horizontal scale bar under the tree represents evolutionary distance between specimen taxa. Headers of scientific names represent abbreviations of the order names (e.g. Col. and Hym.).

Universality and efficiency of the modified primer sets

Using the modified COI primer sets for DNA barcoding, we conducted PCR to amplify samples from 96 species, encompassing 48 families and 11 orders and performed NGS sequencing analysis. These primers successfully amplified and sequenced the target mtDNA COI regions of 80 species from 41 families in 11 orders. Notably, the primer sets had high success rates for Coleoptera, Lepidoptera, Orthoptera and Odonata (Table 3). Despite moderate sequencing success rates and limited specimen samples, the primer sets also showed promise as effective barcoding primers for Hymenoptera, Hemiptera and Diptera.

After the clustering step in Claident, the modified barcoding primers generated 12–18024 (median, 1102; average, 2774) and 2–18002 (median, 1060; average, 2720) reads using LCO1490–COmfd_R and COmfd_F–HCO2198, respectively (Table 2, Suppl. material 2). There was a pronounced bias in the number of reads between LCO1490–COmfd_R and COmfd_F–HCO2198 (Table 2). However, the high degree of variability in the mtDNA COI region (Leray et al. 2013) suggests that this bias may be due to variability at the primer annealing sites.

To assess whether the modified barcoding primer sets could differentiate various insect taxa, we conducted phylogenetic analysis. The resulting phylogenetic tree showed that related insect taxa clustered within the same lineages (Fig. 2, Suppl. material 3). However, three orders were paraphyletic: Coleoptera, Hemiptera and Orthoptera (Fig. 2). The phylogenetic relationships amongst orders cannot be fully elucidated using a single locus, particularly when only short fragment sequences are available. Takenaka et al. (2023) also reported paraphyly of Coleoptera and Hemiptera. Therefore, we conclude that these phylogenetic results are not *major* issues within the scope of this study. Nevertheless, our results suggest that we obtained accurate sequences, as related species were identified as candidate sequences in BLAST searches (Suppl. material 4) and the primer sets appear to be suitable for insect barcoding analyses. We also directly compared additional Chironomid NGS assemblies whose DNA template libraries were similar to the sequence data from the Chironomid DNA Barcode Database (https://www.nies.go.jp/yusurika/en/contents/search.php). Although we performed NGS sequencing analysis of 16 Chironomid samples, we obtained complete mtDNA COI assemblies from 13 Chironomid specimen samples (Suppl. material 5). Comparing these 13 Chironomid NGS assemblies against the database, we detected no assembly errors (Suppl. material 5).

Future utilisation of the modified COI primer sets

The goal of our research was to modify the existing mtDNA COI primer set established by Folmer et al. (1994) for use in NGS sequencing analysis to enhance DNA barcoding reference databases. The original primer set of Folmer et al. (1994) is foundational to the DNA barcoding system and has been widely used in taxonomic, phylogeographic and conservation genetic studies. Recently, Leese et al. (2020) developed new primer sets for the mtDNA COI region tailored for NGS-based analysis. However, the sequences obtained

with these new primers were shorter than those produced by the original primer set, leading to concerns that they might not be as effective for enhancing DNA barcoding references. In this study, we modified the original primer set of Folmer et al. (1994). The modified primer sets are anticipated to greatly enhance DNA barcoding references, although these primers are not compatible with some insect taxa, as indicated in Table 2. Shokralla et al. (2015) also developed NGS-based universal primer sets for the mtDNA COI region; however, their success rate was ca. 70% for all insect taxa. Although the current study examined somewhat limited specimen samples, the success rate was 80% (Table 3). Notably, the success rate for Coleoptera was significantly higher than that reported by Shokralla et al. (2015), whereas our success rate was similar to that of Liu et al. (2013). However, because the primer sequences differ, the modified primer sets should be useful as supplemental primer sets for those of Liu et al. (2013) and Shokralla et al. (2015).

Advances in NGS system have led to various NGS applications in ecological studies. Suyama et al. (2021) introduced the multiplexed phylogenetic marker sequencing (MPM-seq) technique, which enables the simultaneous acquisition of genetic information using multiple primer sets. Takenaka et al. (2023) developed innovative primer sets targetting the mtDNA 16S and 12S rRNA regions, contributing to the discovery of cryptic species and previously undescribed species. As these primer sets generate short sequences, they are also suited for NGS-based analysis (Takenaka et al. 2023). Thus, using our modified primer sets for mtDNA COI, 16S rRNA and 12S rRNA in conjunction with MPM-seq allows more comprehensive taxonomic, phylogeographic and conservation genetic studies.

The mtDNA COI region is considered challenging for designing new NGS-based primer sets due to the high polymorphism rate (Deagle et al. 2014, Takenaka et al. 2023). This complexity has led to low sequencing success rates for insect taxa such as Hymenoptera and Hemiptera (Table 3). In this study, we designed primers manually by visualising primer annealing sites for 33 insect taxa (Suppl. material 1), without performing *in silico* analysis. We anticipate that further modified primer sequences will enhance the success rates of PCR amplification and sequencing analysis. Nevertheless, due to the presence of mixed bases in the primer sequences, which may lead to the amplification of non-target loci, the primer sequences must be redesigned with caution.

Acknowledgements

Funding program

Grant title

Author contributions

Seikan Kurata, Naoyuki Nakahama and Motomi Ito conceived the ideas. Shun K. Hirota and Yoshihisa Suyama performed sequencing analysis. Seikan Kurata, Shota Mano and Shun K. Hirota contributed to interpret the analysis results and data curration. All authors contributed to write the manuscript.

Conflicts of interest

The authors declare that they have no conflicts of interest.

References

- Buckley T, Marske K, Attanayake D (2009) Identifying glacial refugia in a geographic parthenogen using palaeoclimate modelling and phylogeography: the New Zealand stick insect *Argosarchus horridus* (White). Molecular Ecology 18 (22): 4650-4663. https://doi.org/10.1111/j.1365-294x.2009.04396.x
- Burns J, Janzen D, Hajibabaei M, Hallwachs W, Hebert PN (2008) DNA barcodes and cryptic species of skipper butterflies in the genus *Perichares* in Area de Conservación Guanacaste, Costa Rica. Proceedings of the National Academy of Sciences 105 (17): 6350-6355. https://doi.org/10.1073/pnas.0712181105
- Çıplak B, Yahyaoğlu Ö, Uluar O, Doğan Ö, Başibüyük HH, Korkmaz EM (2022) Phylogeography of Pholidopterini: Revising molecular clock calibration by Mid-Aegean Trench. Insect Systematics & Evolution 53 (5): 515-535. https://doi.org/10.1163/1876312x-bja10033
- Costello M, Wilson S, Houlding B (2011) Predicting total global species richness using rates of species description and estimates of taxonomic effort. Systematic Biology 61 (5). https://doi.org/10.1093/sysbio/syr080
- Costello M, May R, Stork N (2013) Response to comments on "Can we name earth's species before they go extinct?". Science 341 (6143): 237-237. https://doi.org/10.1126/science.1237381
- Deagle B, Jarman S, Coissac E, Pompanon F, Taberlet P (2014) DNA metabarcoding and the cytochrome *c* oxidase subunit I marker: not a perfect match. Biology Letters 10 (9). https://doi.org/10.1098/rsbl.2014.0562
- Elbrecht V, Leese F (2017) Validation and development of COI metabarcoding primers for freshwater macroinvertebrate bioassessment. Frontiers in Environmental Science 5 (11). https://doi.org/10.3389/fenvs.2017.00011
- Folmer O, Black M, Hoeh W, Lutz R, Vrijenhoek R (1994) DNA primers for amplification of mitochondrial cytochrome *c* oxidase subunit I from diverse metazoan invertebrates. Molecular Marine Biology and Biotechnology 3 (5): 294-299.

- Gaston K (1991) The magnitude of global insect species richness. Conservation Biology 5 (3): 283-296. https://doi.org/10.1111/j.1523-1739.1991.tb00140.x
- Gotelli N, Colwell R (2001) Quantifying biodiversity: procedures and pitfalls in the measurement and comparison of species richness. Ecology Letters 4 (4): 379-391. https://doi.org/10.1046/j.1461-0248.2001.00230.x
- Hebert PN, Cywinska A, Ball S, deWaard J (2003) Biological identifications through DNA barcodes. Proceedings of the Royal Society of London. Series B: Biological Sciences 270 (1512): 313-321. https://doi.org/10.1098/rspb.2002.2218
- Hebert PN, Gregory TR (2005) The promise of DNA barcoding for taxonomy. Systematic Biology 54 (5): 852-859. https://doi.org/10.1080/10635150500354886
- Katoh K (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Research 30 (14): 3059-3066. https://doi.org/10.1093/nar/gkf436
- Katoh K, Standley DM (2013) MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. Molecular Biology and Evolution 30 (4): 772-780. https://doi.org/10.1093/molbev/mst010
- Kumar S, Stecher G, Li M, Knyaz C, Tamura K (2018) MEGA X: Molecular Evolutionary Genetics Analysis across computing platforms. Molecular Biology and Evolution 35 (6): 1547-1549. https://doi.org/10.1093/molbev/msy096
- Leese F, Sander M, Buchner D, Elbrecht V, Haase P, Zizka VA (2020) Improved freshwater macroinvertebrate detection from environmental DNA through minimized nontarget amplification. Environmental DNA 3 (1): 261-276. https://doi.org/10.1002/edn3.177
- Leray M, Yang JY, Meyer CP, Mills SC, Agudelo N, Ranwez V, Boehm JT, Machida RJ (2013) A new versatile primer set targeting a short fragment of the mitochondrial COI region for metabarcoding metazoan diversity: application for characterizing coral reef fish gut contents. Frontiers in Zoology 10 (1). https://doi.org/10.1186/1742-9994-10-34
- Liu S, Li Y, Lu J, Su X, Tang M, Zhang R, Zhou L, Zhou C, Yang Q, Ji Y, Yu D, Zhou X (2013) SOAPBarcode: revealing arthropod biodiversity through assembly of Illumina shotgun sequences of PCR amplicons. Methods in Ecology and Evolution 4 (12): 1142-1150. https://doi.org/10.1111/2041-210x.12120
- Liu S, Yang C, Zhou C, Zhou X (2017) Filling reference gaps via assembling DNA barcodes using high-throughput sequencing – moving toward barcoding the world. GigaScience 6: 1-8. https://doi.org/10.17504/protocols.io.j99cr96
- Magurran A (1988) Ecological diversity and its measurement. Princeton University Press, Princeton. [ISBN 978-94-015-7360-3] https://doi.org/10.1007/978-94-015-7358-0
- Marquina D, Andersson A, Ronquist F (2018) New mitochondrial primers for metabarcoding of insects, designed and evaluated using in silico methods. Molecular Ecology Resources 19 (1): 90-104. https://doi.org/10.1111/1755-0998.12942
- May RM (1988) How Many Species Are There on Earth? Science 241 (4872): 1441-1449. https://doi.org/10.1126/science.241.4872.1441
- May RM (1990) How many species? Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences 330 (1257): 293-304. https://doi.org/10.1098/rstb.1990.0200
- Meier R, Wong W, Srivathsan A, Foo M (2015) $1 DNA barcodes for reconstructing complex phenomes and finding rare species in specimen-rich samples. Cladistics 32 (1): 100-110. https://doi.org/10.1111/cla.12115

- Nakahama N, Uchida K, Ushimaru A, Isagi Y (2018) Historical changes in grassland area determined the demography of semi-natural grassland butterflies in Japan. Heredity 121 (2): 155-168. https://doi.org/10.1038/s41437-018-0057-2
- Nguyen L, Schmidt H, von Haeseler A, Minh BQ (2014) IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. Molecular Biology and Evolution 32 (1): 268-274. https://doi.org/10.1093/molbev/msu300
- Rambaut A (2016) FigTree v1.4.3., a graphical viewer of phylogenetic trees. Release date: 2016-10-04. URL: http://tree.bio.ed.ac.uk/software/figtree/
- Rebijith KB, Asokan R, Kumar NKK, Krishna V, Chaitanya BN, Ramamurthy VV (2013) DNA barcoding and elucidation of cryptic aphid species (Hemiptera: Aphididae) in India. Bulletin of Entomological Research 103 (5): 601-610. https://doi.org/10.1017/s0007485313000278
- Rice P, Longden I, Bleasby A (2000) EMBOSS: The European Molecular Biology Open Software Suite. Trends in Genetics 16 (6): 276-277. https://doi.org/10.1016/s0168-9525(00)02024-2
- Saitou N, Nei M (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. Molecular Biology and Evolution 4 (4): 406-425. https://doi.org/10.1093/oxfordjournals.molbev.a040454
- Shokralla S, Porter T, Gibson J, Dobosz R, Janzen D, Hallwachs W, Golding GB, Hajibabaei M (2015) Massively parallel multiplex DNA sequencing for specimen identification using an Illumina MiSeq platform. Scientific Reports 5 (1). https://doi.org/10.1038/srep09687
- Stork N (1993) How many species are there? Biodiversity and Conservation 2 (3): 215-232. https://doi.org/10.1007/bf00056669
- Stork N (2018) How Many Species of Insects and Other Terrestrial Arthropods Are There on Earth? Annual Review of Entomology 63 (1): 31-45. https://doi.org/10.1146/annurev-ento-020117-043348
- Suyama Y, Hirota S, Matsuo A, Tsunamoto Y, Mitsuyuki C, Shimura A, Okano K (2021) Complementary combination of multiplex high-throughput DNA sequencing for molecular phylogeny. Ecological Research 37 (1): 171-181. https://doi.org/10.1111/1440-1703.12270
- Takenaka M, Yano K, Suzuki T, Tojo K (2023) Development of novel PCR primer sets for DNA barcoding of aquatic insects, and the discovery of some cryptic species. Limnology 24 (2): 121-136. https://doi.org/10.1007/s10201-022-00710-5
- Tanabe A, Toju H (2013) Two new computational methods for universal DNA barcoding: A benchmark using barcode sequences of bacteria, archaea, animals, fungi, and land plants. PLOS One 8 (10). https://doi.org/10.1371/journal.pone.0076910
- Yang C, Zheng Y, Tan S, Meng G, Rao W, Yang C, Bourne D, O'Brien P, Xu J, Liao S, Chen A, Chen X, Jia X, Zhang A, Liu S (2020) Efficient COI barcoding using high throughput single-end 400 bp sequencing. BMC Genomics 21 (1). https://doi.org/10.1186/s12864-020-07255-w
- Zhou X, Li Y, Liu S, Yang Q, Su X, Zhou L, Tang M, Fu R, Li J, Huang Q (2013) Ultra-deep sequencing enables high-fidelity recovery of biodiversity for bulk arthropod samples without PCR amplification. GigaScience 2 (1). https://doi.org/10.1186/2047-217x-2-4

Supplementary materials

Suppl. material 1: Sequence information for designing PCR primers `doi`

Authors: Seikan Kurata, Shota Mano, Naoyuki Nakahama, Shun K. Hirota, Yoshihisa Suyama, Motomi Ito
Data type: Table
Brief description: We designed the PCR primers by aligning the sequences.
Download file (18.99 kb)

Suppl. material 2: The number of reads after quality control or data analyses using Claident `doi`

Authors: Seikan Kurata, Shota Mano, Naoyuki Nakahama, Shun K. Hirota, Yoshihisa Suyama, Motomi Ito
Data type: Table
Download file (37.78 kb)

Suppl. material 3: Phylogenetic tree using the neighbour-joining method `doi`

Authors: Seikan Kurata, Shota Mano, Naoyuki Nakahama, Shun K. Hirota, Yoshihisa Suyama, Motomi Ito
Data type: Figure
Brief description: The neighbour-joining (NJ) tree based of mtDNA COI region which originated from sequence reads of the modified primer sets.
Download file (404.29 kb)

Suppl. material 4: The number of reads from first half (LCO1490–COmfd_R: 1-319) and second half (COmfd_F–HCO2198: 262-658) of the COI region, and the results of BLAST search. `doi`

Authors: Seikan Kurata, Shota Mano, Naoyuki Nakahama, Shun K. Hirota, Yoshihisa Suyama, Motomi Ito
Data type: Table
Brief description: The number of reads from first half (LCO1490–COmfd_R: 1-319) and second half (COmfd_F–HCO2198: 262-658) of the COI region. There is a great difference for the number of reads between the most dominant and the second largest read. In addition, the results of the BLAST search were shown on the right side in the Table.
Download file (34.00 kb)

Suppl. material 5: Comparison results of Chironomid NGS assemblies against the sequence data from the Chironomid DNA Barcode Database by the National Institute for Environmental Studies (NIES). `doi`

Authors: Seikan Kurata, Shota Mano, Naoyuki Nakahama, Shun K. Hirota, Yoshihisa Suyama, Motomi Ito

Data type: Table

Brief description: Comparison results of Chironomid NGS assemblies against the sequence data from Sanger sequencing.

Download file (19.07 kb)

Endnotes

[*1]   Illumina adapter sequences (F: CGCTCTTCCGATCTCTG; R: TGCTCTTCCGATCT GAC) were added to the 5' end of the primer sequence and 0–3Ns were inserted between adapter sequences and primer sequences.

[*2]   The number of reads from first half (LCO1490–COmfd_R: 1-319) and second half (COmfd_F–HCO2198: 262-658) of the COI region, which were clustered and counted by Claident and the unit is base pair (bp). The symbol of "–" indicates that suitable fragment reads for the mtDNA COI region were not detected.

[*3]   The number of specimen samples for sequencing analysis.

[*4]   The number of samples which pass the PCR trials and NGS-based sequencing analysis.