Application of DNA Barcoding in Species Identification of Pieridae Family from Entopia Penang Butterfly Farm

Ainol Azifa Mohd Faik[1]*, Lee Jee Whu[2], Lee Ping Chin[1], Awang Muhammad Sagaf Abu Bakar[3], Matthew Huan Yew Meng[4] and Lee Yoke Kuen[4]

[1] Faculty of Science and Natural Resources, Universiti Malaysia Sabah,
Jalan UMS, 88400 Kota Kinabalu, Sabah, Malaysia
[2] Institute for Research in Molecular Medicine (INFORMM) Universiti Sains Malaysia,
11800 USM, Penang Malaysia
[3] Veterinary Diagnostics Laboratory, P.O. Box 59, 89457 Tanjung Aru, Kota Kinabalu, Sabah,Malaysia
[4] Butterfly House (Pg) Sdn. Bhd., No. 830, Jalan Teluk Bahang,
11050 Penang, Malaysia; www.entopia.com

Misidentification of butterfly species can be contributed by certain factor(s) such as mimicry, seasonal sampling (dry/wet season), cryptic and sex dimorphism. DNA barcoding is a technique used for rapid identification of butterfly species at a molecular level based on the analysis of standardized mitochondrial DNA region. A 648 base-pair of *cytochrome oxidase c subunit I* (*COI*) is highly effective in identifying vertebrate and invertebrate species. Individuals of the same species will generally have very similar *COI* base sequence. *COI* is used as a DNA barcode reference due to maternal inheritance, compact structure, lack of genetic recombination and relatively fast evolutionary rate. Therefore, it is widely used in phylogenetic and evolution studies. The DNA barcode of identified butterfly species was submitted to Barcode of Life Data System (BOLD). Thirteen species belonging to Pieridae family from Entopia Penang Butterfly Farm were chosen for this study. With reference to some samples, morphological identification method was not sufficient. There were 10 butterflies that were discovered to be different based on DNA barcoding compared to morphological identification. Triplicates of *Eurema brigitta* (n=3) were found to be identified as *Eurema sari sodalis*. Morphologically identified *Eurema hecabe* (n=3) were discovered as *Eurema hecabe contubernalis*, *Eurema andersoni*, and *Eurema simulatrix* respectively. One sample of *Appias indra* (n=3), was identified as *Appias cardena perakana*. *Appias libythea* (n=3) were actually identified as *Appias olferna olferna*. From this result, the application of DNA barcoding is indeed a useful tool in further justifying the collected butterfly species.

Keywords: Mimicry, Cryptic, Sex dimorphisms, Cytochrome oxidase c subunit 1, DNA barcoding, Pieridae, BOLD Systems.

I. INTRODUCTION

Penang Butterfly Farm (PBF) was founded in 1986 and was later rebranded to Entopia in 2005. The name Entopia came about from the combination of two words "entomology" which means "the branch of zoology concerned with the study of insects" (Oxford Dictionaries, 2018) and "utopia" which means "an imagined place or state of things in which everything is perfect" (Oxford Dictionaries, 2018). Entopia has the idea of building a DNA barcode reference library for

*Corresponding author's e-mail:

their catalogued butterflies. This reference library could help to identify the butterflies at the DNA level. In this study, we attempted to create a DNA barcode reference library for the Pieridae family in the Barcode of Life Data System (BOLD) by applying DNA barcoding and analyzing their genetic differences.

Butterflies belong to the Kingdom Animalia and Phylum Arthropoda and were categorized as subphylum Hexapoda or six-legged arthropods. Butterflies are further grouped into the Class Insecta followed by the order Lepidoptera which means scaled wings (Kelly, 2016). Butterflies are categorized in this order due to the possession of two pairs of membranous wings clothed with overlapping scales. Butterflies belong to a single superfamily, the Papilionoidea with seven families: Papilionidae, Hedylidae, Hesperiidae, Pieridae, Nymphalidea, Riodinidae and Lycaenidae. Each butterfly will then fall into a unique genus and species (Enchanted Learning, 2017).

Pierids of the family Pieridae are medium-sized butterflies and are predominantly white or yellow. They have the tips of the legs (claws) that are forked, the forelegs that are full-sized and fully functional. The wings of many Pierid species reflect and absorb ultraviolet light in specific patterns, helping them to identify their potential mates of the same species and many species show sexual dimorphism (Idaho Museum of Natural History, 2016; Nanda & Feuerstein, 2006; Shinichi *et al.*, 2017). Several species of Pierids are seasonally different (Brakefield & Larsen, 1984).

Cytochrome c oxidase or complex IV is the terminal protein complex in the electron transport chain of mitochondrial oxidative phosphorylation and is made of 13 protein subunits. Only three of the subunits are encoded in the mitochondrial genome (COX I-III) and the other ten protein subunits are encoded in nuclear genome. *Cytochrome C oxidase I* (*COI*) gene is one of these three genes in mitochondrial genome that encodes cytochrome c oxidase subunit I (Linacre & Tobe, 2013). *COI* is a widely accepted marker for developing DNA barcodes for species identification and biodiversity analysis. There are approximately 648 nucleotides of *COI* molecular sequence that can determine the identity of unknown species. Due to their maternal inheritance, compact structure, lack of genetic recombination and relatively fast evolutionary rate, mitogenomes have been used widely in molecular phylogenetics and evolution studies (Cong & Grishin, 2016).

Classification or most generally known as taxonomy is the description and naming of species and the placement of them within a genera and highest taxa. It has become more difficult and challenging for humans to observe, record and analyze morphological characteristics or traits because the communication between the individuals occurs through chemical, behavioral or other ephemeral signals which last for a very short time (Shelly, 2014). Therefore, a different tool is required in order to increase the taxonomic accuracy. In this case, DNA barcoding offers a new additional taxonomic approach to assist in species identification with increased accuracy (Chakravarthy, 2015).

DNA barcoding as a molecular tool to help in the identification of species using *COI*, was introduced by Hebert and his colleagues (Hertz *et al.*, 2011). The Barcode of Life Data System (BOLD) was launched in 2005 as a workbench and repository with the support of a growing community of researchers, focusing on building a DNA barcode library for all eukaryotic life. BOLD Systems provides an integrated bioinformatics platform that supports all the processes of the analytical pathway from specimen collection to identification of species in barcode library.

II. MATERIALS AND METHODS

A. Species Collection

A total of 13 dried butterfly species provided by Entopia Penang Butterfly Farm (EPBF) were examined. Detailed specimen information of these species is available on BOLD under the project name COILP (Lepidoptera: Pieridae). Each species was identified morphologically by Entopia personel as *Catopsilia pomona*, *C. pomona* (subspecies), *C. pyranthe*, *C. scylla, Eurema brigitta, E. hecabe, Appias (Phrissura) aegis, A. albino, A. indra, A. lalassis, A. libythea, A. lyncida* and *A. nero*. There were two to three replicates for each species and were further stored at -20 °C.

B. DNA Extraction, Amplification of *COI* region and Sequencing

The legs of each butterfly sample were removed using forceps. The body was kept in a small zip lock bag and recorded. Specimens of the donor butterflies were stored in the fridge at -20 °C. Legs of each butterfly species were crushed using a micropestle in a microfuge containing 50 µl of 1X DNA Lysis Solution (BioLyse DNA Isolation Kit, Biosatria Sdn. Bhd., Malaysia) supplemented with 0.5 µl of 10 mg/ml Proteinase-K. The crushed tissue sample were incubated at 55 °C for 10 minutes in a shaking thermomixer (Eppendorf) and then cooled to room temperature for 5 minutes. Next, the microfuge tube containing lysate was centrifuged at 13 000 rpm for 10 minutes in a bench top centrifuge. The supernatant containing DNA (20 µl) was added into a new microfuge tube containing 80 µl of ultrapure water.

C. Amplification of *COI* Region Using Polymerase Chain Reaction (PCR)

In this study, the forward primer, COIF_650 (5'GGT CAA CAA ATC ATA AAG ATA TTG G 3') and reverse primer COIR_650 (5'TAA ACT TCA GGG TGA CCA AAA AAT CA3') were used (Folmer *et al.*, 1994). The total volume of PCR mixture was 30 µl with a final of 1.5 mM $MgCl_2$, 0.25 µM primer, 0.2 mM dNTP and 1.25 Unit of Taq Polymerase. The amplified product was resolved with 2 % agarose gel electrophoresis. Target band of PCR product of about 700 bp were gel purified (Agarose Gel Extraction Kit, Bioteke Corporation, China) and sent for DNA sequencing (First Base Laboratories Sdn. Bhd., Malaysia).

D. Data Analysis

Chromatograms of the sequences were verified and trimmed using Seqman (DNASTAR, Inc., Madison, WI, USA). The consensus sequence was constructed and aligned using Molecular Evolutionary Genetics Analysis Version 7 (MEGA 7) (Kumar *et al.*, 2015). Basic Local Alignment Search Tool (BLAST) in National Center for Biotechnology Information (NCBI) was applied to search regions of local similarity between query sequence and sequence from database. Data deposited in BOLD were compared using the Distance Summary and the presence of the barcode gap was determined by maximum intraspecific divergence plotted against nearest neighbour distance using the Kimura 2-parameter (K2P) model (Kimura, 1980). Pairwise distances were calculated using the same K2P model. MEGA 7.0 was also used to construct phylogenetic tree (including one outgroup) based on the Maximum Likelihood approach using Tamura and Nei model with 1000 bootstrap replicates (Tamura & Nei, 1993). The mitochondrial *COI* sequence of *Homo sapiens* (GI KP126163) was used as outgroup.

III. RESULTS AND DISCUSSION

A. Result Analysis of DNA sequences

Good quality *COI* DNA sequences were obtained after DNA sequencing with an average length of 550 bp. All the sequences matched with the *COI* sequence in the NCBI database ranging from 95%-100% similarity (Table 1). Based on the BLAST result, there were 10 butterflies that were discovered to be different species compared to morphological identification. Triplicates of *Eurema brigitta* (n=3) were found to be identified as *E. sari sodalis*. Morphologically identified *E. hecabe* (n=3) were discovered as *E. hecabe contubernalis*, *E. andersoni*, and *E. simulatrix* respectively. One sample of *Appias indra* (n=3), was identified as *A. cardena perakana. A. libythea* (n=3) were actually identified as *A. olferna olferna*.

Table 1. BLAST result of butterfly samples with NCBI database

Morphologically Characterized	**NCBI**		
	E-Value	**ID (%)**	**Species Name**
***Catopsilia pomona* (n=3)**	0.0	100	*Catopsilia pomona*
***Catopsilia pomona* subsp. (n=3)**	**0.0**	**99**	***Catopsilia pomona***
***Catopsilia pyranthe* (n=3)**	**0.0**	**100**	***Catopsilia pyranthe pyranthe***

***Catopsilia scylla*(n=3)**	**0.0**	**100**	***Catopsilia scylla scylla***
***Eurema brigitta* (n=3)**	**0.0**	**99**	***Eurema sari sodalis***
***Eurema hecabe* 1**	**0.0**	**98**	***Eurema hecabe contubernalis***
***Eurema hecabe* 2**	**0.0**	**100**	***Eurema andersoni***
***Eurema hecabe* 3**	**0.0**	**99**	***Eurema simulatrix***
***Appias(Phrissura) aegis* (n=3)**	0.0	99	*Phrissura aegis*
***Appias albina* (n=3)**	0.0	100	*Appias albina*
***Appias indra*(n=2)**	0.0	95	*Appias indra*
***Appias indra* 3**	**0.0**	**95**	***Appias cardena perakana***
Appias lalassis	0.0	100	*Appias lalassis*
***Appias libythea*(n=3)**	**0.0**	**100**	***Appias olferna olferna***
***Appias lyncida* (n=3)**	**0.0**	**100**	***Appias lyncida vasava***
***Appias nero* (n=3)**	**0.0**	**100**	***Appias nero nero***

B. Distance Summary and Barcode Gap Analysis

Intraspecific divergences ranged from 0.0 to 1.2% with an average of 0.21%. High intraspecific divergences (>1%) were detected in *Appias olferna olferna*, *A. albina* and *Eurema sari sodalis*. Interspecific divergences ranged from 0.0 to 15.63% (Table 2). *Appias cardena perakana* was found to have the lowest interspecific divergence (<1%). A significant barcode gap will be exhibited when the sequence divergence within species should be lower than the sequence divergence between species (Trivedi *et al.*, 2016). In this study, sequence divergence within species was referred to as maximum intraspecific distance while sequence divergence between species was referred to as interspecific distance or nearest neighbour (NN) distance. The scatterplot in Figure 1 is provided to confirm the existence and magnitude of the barcode gap. It shows the overlap of the max intraspecific distances vs the interspecific (NN) distances. Points above the line in this scatterplot indicate species with Barcode Gap. The nearest neighbour (NN) distance exceeded the maximum intraspecific in almost all species of the Pieridae family and only one case where the NN distance was zero that was between *A. cardena* and *A. indra*.

Table 2. Inter- and Intraspecific divergences according to the different taxonomic levels within the COI sequences of the Pieridae butterfly family.

Comparison	Min Dist (%)	Mean Dist (%)	Max Dist (%)	SE Dist (%)
Intraspecific genetic divergence (Within species)	0.00	0.21	1.20	0.01
Interspecific genetic divergence (Between species)	0.00	10.53	15.63	0.01

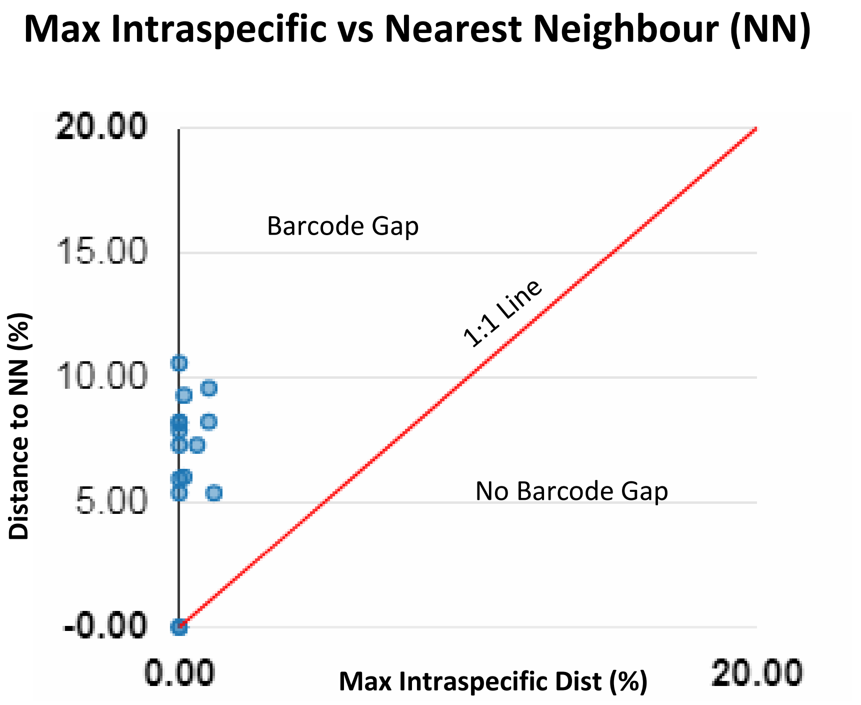

Figure 1. Comparison of maximum intraspecific sequence divergence with minimum interspecific sequence divergence for Pieridae butterfly family. Points below the 1:1 line indicate that a barcode gap is absent. Points above the 1:1 line indicate that a barcode gap is present.

C. Phylogenetic Tree of Nucleotide Sequences

The evolutionary history was inferred using the Maximum Likelihood (ML) method. The tree with the highest log likelihood (-3887.0785) is shown (Figure 2). The ML tree showed that all species with replicates recovered as monophyletic with high bootstrap values (>82%). *Catopsilia pomona* are very closely related to each other as they are the same species. Hence, they are grouped in the same clade. *Eurema andersoni, E. simulatrix* and *E. hecabe contubernalis* which were previously identified as *E. hecabe* were not grouped under the same clade. This is because *E. simulatrix* is a totally different species with *E. andersoni* with genetic difference of 10.4% and has 6.1% genetic diversity with *E. hecabe contubernalis*. These differences were clearly shown in the ML tree (Figure 2) where *E. simulatrix* and *E. hecabe contubernalis* were closely related compared to *E. andersoni*. This might indicate that *E. simulatrix* and *E. hecabe contubernalis*, both of the two different species might have adapted to the same environment. Natural selection primarily affects only those genes that directly are involved in the environmental adaptation. The gene flow or genetic exchange between the two species (*E. simulatrix* and *E. hecabe contubernalis*) may allow them to share allelles in common, created by mutations (Fish, 20111)

Morphologically characterized triplicates of *Catopsilia pyranthe* were having subspecies known as *C. pyranthe pyranthe* while misidentified triplicates of *E. brigitta* belong to subspecies *E. sari sodalis*.

Nevertheless, all of these misidentified triplicates were found to be monophyletic with their replicates with strong bootstrap value of >95%. *Phrissura aegis* synonym to *A. aegis,* has high interspecific divergence within the Appias genus, ranging from 10.9% to 13.7% and was shown in ML tree. Mutation and genetic drift ensure that the isolated same species become increasingly different genetically. Hence, genetic diversity between the isolated same species increases with increasing time. The gradual adaptation to their local environments and the chance of drifting away from the ancestral type both progress until the same species are no longer able to interbreed when they come into contact (Scott, 1986).

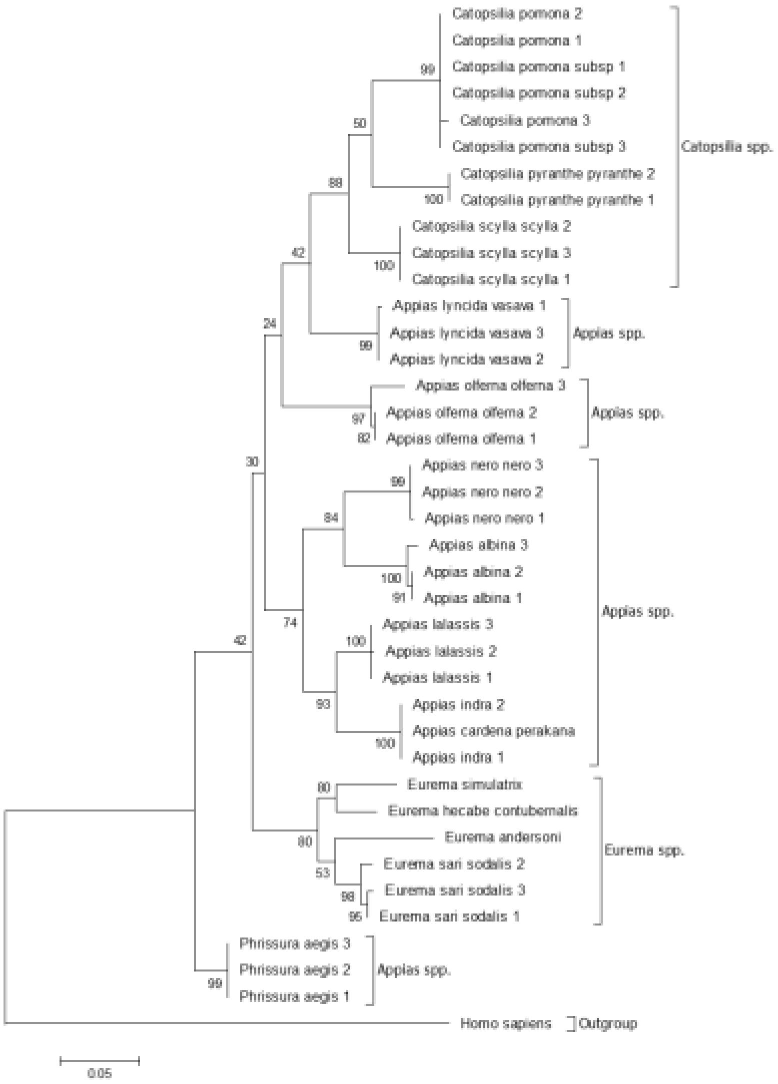

Figure 2. Molecular Phylogenetic analysis by Maximum Likelihood method. The evolutionary history was inferred by using the Maximum Likelihood method based on the Tamura-Nei model. The percentage of trees in which the associated taxa clustered together is shown next to the branches.

IV. SUMMARY

From this study, it was found that morphological identification could be misleading as the species identification through BLAST by using molecular data, pairwise distances and phylogenetic tree had opposed the morphological identification of few species. Therefore, identification of butterfly should include both phenotype and molecular methods. For this work, all data from the species belonging to Pieridae family were deposited into BOLD system. Research involving other butterfly families from Entopia is being carried out and this reference library is expected to help Entopia to easily determine the species of butterflies.

V. ACKNOWLEDGEMENTS

This research was supported by UMS under Grant SLB0122-SG-2016 (SGPUMS).

[1] Brakefield, P.M. & Larsen, T.B. (1984). *The Evolutionary Significance of Dry and Wet Season forms in Some Tropical Butterflies, Biological Journal of the Linnean Society*, vol. 22, pp. 1-12.

[2] Chakravarthy, A.K. (2015). *New Horizons in Insect Science:* Towards Sustainable Pest Management**,** New Delhi.

[3] Cong, Q. & Grishin, N. (2016). *The complete mitochondrial genome of Lerema accius and its Phylogenetic Implications*, *Peer J*, vol. 4, pp. 1546-1590.

[4] *Enchanted Learning, Classification of Butterflies and Moths*. (2017). from: http://www.enchantedlearning.com/subjects/butterflies/Classification.shtml.

[5] Fish, J.M. (2011). *Race and Intelligence: Separating Science from Myth*. Lawrence Erlbaum Associates, Inc, Mahwah (eds).

[6] 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. Mar. Biol. Biotechnol*, vol. 3, pp. 294–299.

[7] Hertz, P.E., Mcmillan, B. & Russell, P.J. (2011). *Biology: The Dynamic Science*. Second Ed. Brooks/Cole, United States of America.

[8] Idaho Museum of Natural History, *Family Pieridae, the Whites and Sulphurs*. (2016). from: http://imnh.isu.edu/digitalatlas/bio/insects/butrfly/fampier/fampie.htm.

[9] Kelly, J. (2016). *Levels of Classification of Butterflies, from:* http://animals.mom.me/levels-classification-butterflies-6338.html.

[10] Kimura, M. (1980). A simple method for estimating evolutionary rate of base substitutions through comparative studies of nucleotide sequences, *Journal of Molecular Evolution* vol. 16, pp. 111-120.

[11]Kumar, S., Stecher, G. & Tamura, K. (2015). MEGA7: Molecular Evolutionary Genetics Analysis version 7.0. *Molecular Biology and Evolution*

[12] Linacre, A.M.T. & Tobe, S.S. (2013). Wildlife DNA Analysis: Applications in Forensic Science, *John Wiley & Sons, United States of America*.

[13] Nanda, A. & Feuerstein, S. (2006). *Oracle PL/SQL for DBAs*. O' Really Media, Inc., Sebastopol.

[14] Oxford Dictionaries, s.v. “entomology”. (2018). from: https://en.oxforddictionaries.com/definition/entomology.

[15] Oxford Dictionaries, s.v. “utopia”. (2018). from: https://en.oxforddictionaries.com/definition/utopia.

[16] Scott, J.A. (1986). The Butterflies of North America: A Natural History and Field Guide. *Stanford University Press, California*.

[17] Shelly, T., Epsky, N., Jang, E.B., Flores, J.R. & Vargas, R. (2014). Trapping and the Detection, Control, and Regulation of Tephritid Fruit Flies: Lures, *Area-Wide Programs, and Trade Implications. United States of America*.

[18] Shinichi, N., Thamara, Z., Blanca, H., Andrew, F.E., Neild, J.P.W.H., Gerardo, L., Lauren, A., Holian, M.E. & Keith, R. (2017). Willmott. Remarkable sexual dimorphism, rarity and cryptic species: a revision of the ‘aegrota species group’ of the Neotropical butterfly genus Caeruleuptychia Forster, 1964 with the description of three new species (Lepidoptera, Nymphalidae, Satyrinae), *Insect Systematics & Evolution*.

[19] Tamura, K. & Nei, M. (1993). Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees, *Molecular Biology and Evolution*, vol. 10, pp. 512-526.

[20] Trivedi, S., Ansari, A.A., Ghosh, S.K. & Rehman, H. (2016). DNA Barcoding in Marine Perspectives: Assessment and Conservation of Biodiversity. Switzerland, *Springer International Publishing, Switzerland*