Data Article

Pairwise sequence comparison data of the DNA barcodes of aquatic insects

Koji Inai¹, Kei Wakimura¹, Mikio Kato¹,²,*

¹Riverine Metagenomics Research Group, Faculty of Liberal Arts and Sciences, Osaka Prefecture University, 1-1 Gakuencho, Naka-ku, Sakai 599-8531 JAPAN
²Department of Biological Science, Graduate School of Science, Osaka Prefecture University, 1-1 Gakuencho, Naka-ku, Sakai 599-8531 JAPAN

Article history:
Received 28 August 2020
Accepted 31 August 2020
Available online 8 September 2020

Keywords:
Aquatic insect
COI
DNA barcode
histone H3
sequence comparison
sequence similarity

A B S T R A C T

This study compared the DNA sequences of cytochrome c oxidase subunit I (COI) and histone H3 of Ephemeroptera, Odonata, Plecoptera, and Trichoptera in a pairwise manner, and calculated the sequence similarities based on uncorrected P-distance (number of identical sites in both sequences per total number of the sites compared). Datasets of annotated sequences, the source organisms of which are identified at the species level in taxonomy, were retrieved from INSD (GenBank/ENA/DDBJ) as of the end of May 2020. Similarity scores of the pairwise comparison were sorted by the combinations of taxonomic groups; intraspecific variations, intrageneric-interspecific divergences, intrafamily-intergeneric divergences, and intraorder-interfamily divergences for Ephemeroptera, Odonata, Plecoptera, and Trichoptera. Similarity scores at the cumulative relative frequency points (1%, 5%, 10%, and median) may be used as the threshold to differentiate between the taxonomic groups based on sequence match. This is often done in the characterization of morphologically-unidentified specimens using

DOI of original article: 10.1016/j.egg.2020.100065
* Corresponding Author.
E-mail address: mkato@b.s.osakafu-u.ac.jp (M. Kato).

https://doi.org/10.1016/j.dib.2020.106284
2352-3409/© 2020 The Author(s). Published by Elsevier Inc. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/)
barcode sequences, in the metabarcoding analysis of the local fauna, and environmental DNA analysis.

© 2020 The Author(s). Published by Elsevier Inc. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/)

**Specifications Table**

| Subject | Genetics |
|---------|----------|
| Specific subject area | DNA taxonomy and molecular phylogeny of aquatic insects |
| Type of data | Pairwise sequence comparison of the DNA barcodes of aquatic insects summarized in Tables and histograms (Figures) |
| How data were acquired | Sequence similarity was calculated as the uncorrected P-distance (number of identical sites in both sequences per the total number of the sites compared) by pairwise sequence comparison using an ad-hoc software. |
| Data format | Raw |
| Parameters for data collection | Well-annotated data (source organisms of which are identified precisely at the species level; not obfuscated with sp.) of COI and histone H3 were retrieved from INSD. Partial coding regions (COI: 658-bp, histone H3: 328-bp) were subjected to pairwise comparison. Short sequences were also analyzed, but homologously aligned sequences shorter than the half of their standard length were excluded from the analyses. |
| Description of data collection | Pairwise sequence comparison was performed, and similarity scores were sorted by respective combinations of the taxonomic groups (intraspecific, interspecific, intergeneric, and interfamily comparisons). Computing was performed using an ad hoc software written by one of the authors (K.I.), the source code file of which is attached to this manuscript. File name: sequence_comparison.R |
| Data source location | Primary data sources: List of INSD accession numbers of the sequence data used for the analyses is attached to this manuscript. File names: List of INSD accession numbers_COI.txt, List of INSD accession numbers_H3.txt |
| Data accessibility | Provided with the article |
| Related research article | Kei Wakimura, Yasuhiro Takemon, Shin-ichi Ishiwata, Kazumi Tanida, Eman M. Abbas, Koji Inai, Akikazu Taira, Aki Tanaka, Mikio Kato A reference collection of Japanese aquatic macroinvertebrates *Ecological Genetics and Genomics* 17, 2020, 100065 |

**Value of the Data**

- Species identification based on DNA sequences is a key step in metabarcoding studies and environmental DNA (eDNA) analysis. We provide a standard of sequence similarity critical values that enables us to assign a taxonomic identity by sequence match.
- Sequence data of obfuscated organism names (named with sp.) in INSD will be assigned to the specific taxon via similarity search based on the criteria presented in these data. This provides benefits to scientists investigating biogeography and taxonomy.
- Because many aquatic insects have a preference in micro- and macro-habitats, they are recognized as indicator species for that specific environment. Based on the critical values presented here, the specimens obtained are identified using DNA sequencing, and evaluation of the aquatic environment via metabarcoding will be enabled.
- These data will also reveal that misidentification of specimens and/or confusion of nomenclature exists in the INSD sequencing data and that the presence of cryptic species is also
plausible. Analyzing the species showing higher intraspecific distances may clarify the cryptic species, and this may attract the attention of insect taxonomists.

1. Data Description

Source code of the program used for pairwise sequence comparison is attached to the manuscript (file name: sequence_comparison.R). INSD accession numbers of the nucleotide sequences used in the analyses are listed in the text files (file names: List of INSD accession numbers_COI.txt, List of INSD accession numbers_H3.txt). Raw data of pairwise sequence comparison are compressed in a supplementary material attached to this paper.

Figures 1A-1H are drawn based on the pairwise sequence comparison data, and display the frequency distribution of the sequence similarity data pairs (scale at left) (shown in %). Fig. 1A and 1E show the histograms of cytochrome c oxidase subunit I (COI) and histone H3 of Ephemeroptera, respectively. Fig. 1B and 1F show the histograms of COI and histone H3 of Odonata, respectively. Fig. 1C and 1G show the histograms of COI and histone H3 of Plecoptera, respectively. Fig. 1D and 1H show the histograms of COI and histone H3 of Trichoptera, respectively. Each panel includes the histograms showing intraspecific variation, interspecific divergence, intergeneric divergence, and interfamily divergence. Blue lines indicate cumulative relative frequency curves (shown in %, scale at right). Vertical lines of yellow, red, green, and light blue represent the scores of sequence similarity at 1%, 5%, 10%, and the median of cumulative relative frequencies, respectively.

Tables 1A and 1B list the number of sequencing data of COI and histone H3 used in the analyses, respectively. Numbers of taxa (species, genera, and families), to which the sequencing data belong, are also shown.

Table 2A and 2B summarize the sequence comparison results of COI and histone H3, respectively. Critical scores of sequence similarity at each acceptance region (corresponding to the vertical lines in Fig. 1) are shown.

Table 3 shows the results of sequence-matching between the INSD dataset and the sequencing data of specimens named with sp. in Ephemeroptera, which we collected [3]. There was 42 out of 132 specimens attributed to the known species based on the median of intraspecific variations as the threshold, which was 98.2% and 98.7% for COI and histone H3, respectively.

2. Experimental Design, Materials and Methods

Characterization and identification of living organisms via specific DNA sequences (barcodes) has become popular since Hebert et al. [1] had established the technique and procedure. Intraspecific variations and interspecific divergence of barcode sequences are, however, often problematic because of the complexity in taxonomy. Barcoding has poor outcomes in incompletely sampled groups [2]. We have been making efforts to collect Japanese aquatic insects for identification analysis based on morphology and characterization via DNA barcodes [3].

Morphology-based identification is often difficult for specimens at immature developmental stages and those damaged during collection. Also, specific names cannot be given to undescribed taxa. Similarity search for DNA barcodes in INSD is effectively applied to resolve these problems if the reference sequencing data of precisely-identified species are available in INSD. We examined the intraspecific variations and interspecific, intergeneric, and interfamily divergences of DNA barcodes for Ephemeroptera, Odonata, Plecoptera, and Trichoptera using the INSD dataset (as of May 2020) in order to propose critical values for the DNA-based assignment of taxonomic identity.
Figure 1. Pairwise comparison of INSD data.
Figure 1. Continued

B Odonata COI

Intraspecific comparison

Intrageneric-interspecific comparison

Intrafamily-intergeneric comparison

Intraorder-interfamily comparison

Figure 1. Continued
Figure 1. Continued
Figure 1. Continued
Figure 1. Continued
F Odonata histone H3

Intraspecific comparison

Intrageneric-interspecific comparison

Intrafamily-intergeneric comparison

Intraorder-interfamily comparison

Figure 1. Continued
Figure 1. Continued
Figure 1. Continued
Table 1A
COI sequences retrieved from INSD (GenBank/ENA/DDBJ) as of May 2020

|                           | Ephemeroptera | Odonata | Plecoptera | Trichoptera |
|---------------------------|---------------|---------|------------|-------------|
| Number of DNA sequences   | 12522         | 9238    | 6984       | 34592       |
| Number of species         | 781           | 1365    | 629        | 3347        |
| Number of genera          | 175           | 401     | 168        | 461         |
| Number of families        | 31            | 36      | 16         | 45          |

Table 1B
Histone H3 sequences retrieved from INSD (GenBank/ENA/DDBJ) as of May 2020

|                           | Ephemeroptera | Odonata | Plecoptera | Trichoptera |
|---------------------------|---------------|---------|------------|-------------|
| Number of DNA sequences   | 630           | 442     | 519        | 108         |
| Number of species         | 200           | 311     | 191        | 59          |
| Number of genera          | 105           | 149     | 111        | 34          |
| Number of families        | 30            | 30      | 16         | 12          |

Table 2A
Threshold of sequence similarity (%) of COI

| Sequence comparison     | Acceptance region | Ephemeroptera | Odonata | Plecoptera | Trichoptera |
|-------------------------|-------------------|---------------|---------|------------|-------------|
| Intraspecific           | median            | 98.2          | 99.3    | 99.4       | 99.6        |
| Interspecific (intrageneric) | 1% of data pairs | 94.2          | 99.7    | 99.0       | 94.5        |
|                         | 5% of data pairs  | 88.2          | 97.7    | 94.9       | 89.8        |
|                         | 10% of data pairs | 86.1          | 94.9    | 92.3       | 88.4        |
|                         | median            | 81.6          | 88.0    | 85.7       | 84.0        |
| Intergeneric (intrafamily) | 1% of data pairs | 84.7          | 89.9    | 91.0       | 88.0        |
|                         | 5% of data pairs  | 83.2          | 87.1    | 88.3       | 86.5        |
|                         | 10% of data pairs | 82.5          | 86.1    | 86.0       | 85.6        |
|                         | median            | 80.1          | 83.3    | 81.9       | 81.1        |
| Interfamily (intraorder) | 1% of data pairs | 83.2          | 85.7    | 84.1       | 83.4        |
|                         | 5% of data pairs  | 81.6          | 84.4    | 83.0       | 80.8        |
|                         | 10% of data pairs | 80.7          | 83.7    | 82.4       | 79.5        |
|                         | median            | 77.2          | 81.1    | 80.3       | 76.0        |

Table 2B
Threshold of sequence similarity (%) of histone H3

| Sequence comparison     | Acceptance region | Ephemeroptera | Odonata | Plecoptera | Trichoptera |
|-------------------------|-------------------|---------------|---------|------------|-------------|
| Intraspecific           | median            | 98.7          | 99.3    | 100.0      | 99.8        |
| Interspecific (intrageneric) | 1% of data pairs | 100.0         | 100.0   | 99.8       | 100.0       |
|                         | 5% of data pairs  | 99.9          | 99.9    | 99.1       | 98.0        |
|                         | 10% of data pairs | 99.7          | 99.6    | 98.7       | 97.2        |
|                         | median            | 94.5          | 98.2    | 97.2       | 90.8        |
| Intergeneric (intrafamily) | 1% of data pairs | 97.7          | 98.7    | 97.3       | 93.5        |
|                         | 5% of data pairs  | 96.3          | 96.9    | 96.2       | 90.5        |
|                         | 10% of data pairs | 94.9          | 95.5    | 95.6       | 89.2        |
|                         | median            | 90.8          | 87.3    | 92.1       | 84.4        |
| Interfamily (intraorder) | 1% of data pairs | 93.4          | 94.4    | 93.0       | 91.9        |
|                         | 5% of data pairs  | 92.3          | 91.9    | 91.6       | 89.5        |
|                         | 10% of data pairs | 91.6          | 90.2    | 90.9       | 88.4        |
|                         | median            | 88.4          | 83.7    | 87.9       | 83.5        |

Well-annotated data (source organisms of which are identified at the species level; not obfuscated with sp.) of COI and histone H3 genes were retrieved from INSD. Partial coding regions (COI: 658-bp, histone H3: 328-bp) were subjected to pairwise comparison, and the scores of sequence similarity were sorted by respective combinations of the taxonomic groups. Short se-
### Table 3
Identification of ambiguous specimens of Ephemeroptera by DNA barcodes.

| Gene      | Accession No. | Specimens with laboratory ID | Reference data showing highest similarity | Sequence similarity |
|-----------|---------------|------------------------------|------------------------------------------|---------------------|
| Histone H3 | MK774487      | Ameletus sp. OPU_BS_2017-221-OC-EP | GQ433995 Ameletus montanus              | 325/328; 99.1%      |
| Histone H3 | MK774515      | Ameletus sp. OPU_BS_2017-222-OC-EP | GQ433995 Ameletus montanus              | 327/328; 99.7%      |
| Histone H3 | KF562981      | Ameletus sp. OPU_BS_A2012-262 | AY870291 Ameletus costalis              | 327/328; 99.7%      |
| COI        | KP970695      | Baetis sp. MK-2015a OPU_BS_B2013-101 | KP970701 Acentrella sibirica            | 652/658; 99.1%      |
| COI        | KP970699      | Baetis sp. MK-2015c OPU_BS_B2013-134 | KF563060 Nigrobaetis chooratus         | 654/658; 99.4%      |
| Histone H3 | JQ655111      | Baetis sp. OPU_BS_B2010-23 | KF562972 Baetis thermicus               | 327/328; 99.7%      |
| Histone H3 | JQ650129      | Baetis sp. OPU_BS_B2011-19 | MH260739 Nigrobaetis taiwanensis        | 324/328; 98.8%      |
| Histone H3 | JQ650161      | Cincticostella sp. OPU_BS_C2011-113 | KF562982 Ephicerella longicrada         | 324/328; 98.8%      |
| Histone H3 | KF563005      | Cincticostella sp. OPU_BS_C2011-117 | KF562982 Ephicerella longicrada         | 328/328; 100%       |
| COI        | KF563023      | Cinygmula sp. OPU_BS_C2011-90 | MK774298 Paracymgula zhiltzovae         | 656/658; 99.7%      |
| COI        | KF563024      | Cinygmula sp. OPU_BS_C2011-91 | MK774290 Paracymgula zhiltzovae         | 656/658; 99.7%      |
| Histone H3 | MK774329      | Drunella sp. OPU_BS_2017-165-YS-EP | MK774351 Drunella ishiyama              | 657/658; 99.8%      |
| Histone H3 | MK774449      | Drunella sp. OPU_BS_2017-165-YS-EP | MK774458 Drunella ishiyama              | 328/328; 100%       |
| COI        | MK774330      | Drunella sp. OPU_BS_2017-166-YS-EP | MK774351 Drunella ishiyama              | 658/658; 100%       |
| Histone H3 | MK774450      | Drunella sp. OPU_BS_2017-166-YS-EP | MK774456 Drunella ishiyama              | 325/328; 99.1%      |
| Histone H3 | MK774512      | Drunella sp. OPU_BS_2017-208-YS-EP | JQ650124 Drunella ishiyama              | 328/328; 100%       |
| Histone H3 | MK774514      | Drunella sp. OPU_BS_2017-323-YS-EP | JQ650124 Drunella ishiyama              | 328/328; 100%       |
| COI        | KP970696      | Ecdyonurus sp. MK-2015c OPU_BS_E2013-130 | MK774290 Paracymgula zhiltzovae         | 656/658; 99.7%      |
| Histone H3 | MK774489      | Ecdyonurus sp. OPU_BS_2018-038-IN-EP | JQ650162 Afronurus yoshidae            | 328/328; 100%       |
| Histone H3 | KP970755      | Epeorus sp. MK-2015 OPU_BS_E2012-235 | MK774420 Epeorus ikanonis              | 328/328; 100%       |
| Histone H3 | KP970756      | Epeorus sp. MK-2015 OPU_BS_E2012-236 | MK774420 Epeorus ikanonis              | 328/328; 100%       |
| Histone H3 | KF562963      | Ephemeralla sp. OPU_BS_E2012-129 | MK774468 Teloganopsis punctijetae       | 328/328; 100%       |
| COI        | KF563053      | Epheremerella sp. OPU_BS_E2012-86 | KP970723 Epheremerella notata           | 656/658; 99.7%      |
| COI        | KF563054      | Epheremerella sp. OPU_BS_E2012-90 | KP970723 Epheremerella notata           | 655/658; 99.5%      |
| Histone H3 | KP970729      | Epheremerella sp. OPU_BS_E2013-51 | KF562983 Epheremerella atagosana        | 327/328; 99.7%      |

(continued on next page)
Table 3 (continued)

| Gene     | Accession No. | Specimens with laboratory ID             | Reference data showing highest similarity | Sequence similarity |
|----------|---------------|------------------------------------------|------------------------------------------|---------------------|
| COI      | KP970725      | Ephemerella sp. OPU_BS_E2014-68          | MH260769 Ephemerella imanishi             | 654/658; 99.4%      |
| COI      | KP970726      | Ephemerella sp. OPU_BS_E2014-69          | MH260769 Ephemerella imanishi             | 653/658; 99.2%      |
| COI      | KP970727      | Ephemerella sp. OPU_BS_E2014-71          | MH260769 Ephemerella imanishi             | 655/658; 99.5%      |
| COI      | KP970728      | Ephemerella sp. OPU_BS_E2014-73          | MH260769 Ephemerella imanishi             | 653/658; 99.2%      |
| COI      | KP970715      | Ephemerellidae sp. OPU_BS_E2014-4        | MK774364 Teloganopsis punctisetae         | 658/658; 100%       |
| COI      | MK774406      | Heptageniidae sp. OPU_BS_2017-159-EP     | MK774314 Rhithrogena japonica             | 658/658; 100%       |
| Histone H3 | MK774446    | Heptageniidae sp. OPU_BS_2017-159-DB-EP  | MK774418 Rhithrogena japonica             | 328/328; 100%       |
| COI      | MK774367      | Paraleptophlebia sp. OPU_BS_2017-209-EP  | GU354161 Paraleptophlebia chocolata      | 654/658; 99.4%      |
| Histone H3 | MK774469    | Paraleptophlebia sp. OPU_BS_2017-209-YS-EP | GQ433999 Paraleptophlebia chocolata      | 328/328; 100%       |
| COI      | KF563018      | Paraleptophlebia sp. OPU_BS_P2010-362    | GU354161 Paraleptophlebia chocolata      | 656/658; 99.7%      |
| Histone H3 | JQ650148    | Paraleptophlebia sp. OPU_BS_P2010-362    | GQ433999 Paraleptophlebia chocolata      | 328/328; 100%       |
| Histone H3 | KF562978    | Paraleptophlebia sp. OPU_BS_P2012-248    | JQ650134 Paraleptophlebia spinosa        | 327/328; 99.7%      |
| Histone H3 | KF562979    | Paraleptophlebia sp. OPU_BS_P2012-249    | JQ650134 Paraleptophlebia spinosa        | 327/328; 99.7%      |
| Histone H3 | KF970731    | Paraleptophlebia sp. OPU_BS_P2013-106    | GQ433999 Paraleptophlebia chocolata      | 328/328; 100%       |
| Histone H3 | KF970732    | Paraleptophlebia sp. OPU_BS_P2013-200    | GQ433999 Paraleptophlebia chocolata      | 328/328; 100%       |
| Histone H3 | KF970733    | Paraleptophlebia sp. OPU_BS_P2013-202    | GQ433999 Paraleptophlebia chocolata      | 328/328; 100%       |
| Histone H3 | KF563013    | Rhithrogena sp. OPU_BS_R2010-16          | KP970686 Epeorus aesculus                | 653/658; 99.2%      |
| Histone H3 | JQ650146    | Rhithrogena sp. OPU_BS_R2010-16          | MH260744 Epeorus aesculus                | 324/328; 98.8%      |
| COI      | KF563014      | Rhithrogena sp. OPU_BS_R2010-19          | KP970686 Epeorus aesculus                | 654/658; 99.4%      |
| Histone H3 | KF970721    | Siphlonurus sp. OPU_BS_S2014-10          | KF563047 Siphlonurus sanukensis          | 653/658; 99.2%      |
| Histone H3 | KF970745    | Siphlonurus sp. OPU_BS_S2014-10          | KF562970 Siphlonurus sanukensis          | 328/328; 100%       |
| COI      | KF970746      | Siphlonurus sp. OPU_BS_S2014-11          | KF562970 Siphlonurus sanukensis          | 328/328; 100%       |
| Histone H3 | KF970718    | Siphlonurus sp. OPU_BS_S2014-7           | KF562970 Siphlonurus sanukensis          | 328/328; 100%       |
| Histone H3 | KF970742    | Siphlonurus sp. OPU_BS_S2014-7           | KF562970 Siphlonurus sanukensis          | 328/328; 100%       |
| COI      | KF970720      | Siphlonurus sp. OPU_BS_S2014-9           | KF562970 Siphlonurus sanukensis          | 328/328; 100%       |
| Histone H3 | KF970744    | Siphlonurus sp. OPU_BS_S2014-9           | KF562970 Siphlonurus sanukensis          | 328/328; 100%       |
quences in INSD were also analyzed, but the homologously aligned sequences shorter than the half of their standard length were excluded.

For intraspecific comparison, each of the sequence pairs was counted to achieve frequency distribution of the similarity scores. For the comparisons between higher taxonomic levels, the frequency was normalized to each pair of the species; e.g. in the case of comparison between species-X (2 sequences available) and species-Y (3 sequences available), there are six combinations of sequence pairs, and then 1/6 is taken as the respective similarity score. This normalization was done to avoid the excess contribution of the species, for which multiple sequencing data exist in INSD.

Declaration of Competing Interest

This work was supported by Japan Society for the Promotion of Science (JSPS) KAKENHI (Grant Number: 17K07541). The authors declare that they have no known competing financial interests or personal relationships which have, or could be perceived to have, influenced the work reported in this article.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.dib.2020.106284.

References

[1] P.D.N. Hebert, A. Cywinska, S.L. Ball, J.R. de Waard, Biological identifications through DNA barcodes, Proc. R. Soc. Lond. B 270 (2003) 313–321.
[2] C.P. Meyer, G. Paulay, DNA barcoding: Error rates based on comprehensive sampling, PLoS Biol 3 (2005) e422.
[3] K. Wakimura, Y. Takemon, S. Ishiwata, K. Tanida, E. M. Abbas, K. Inai, A. Taira, A. Tanaka, M. Kato, A reference collection of Japanese aquatic macroinvertebrates. Ecol. Genet. Genom. 17, 2020, 100065.