Mol. Cells *33*, 9-17, January 31, 2012
DOI/10.1007/s10059-012-2151-2

Molecules and Cells

A DNA Barcode Library for Korean Chironomidae (Insecta: Diptera) and Indexes for Defining Barcode Gap

Sungmin Kim[1], Kyo-Hong Song[2], Han-Il Ree[3], and Won Kim[4,*]

Non-biting midges (Diptera: Chironomidae) are a diverse population that commonly causes respiratory allergies in humans. Chironomid larvae can be used to indicate freshwater pollution, but accurate identification on the basis of morphological characteristics is difficult. In this study, we constructed a mitochondrial cytochrome *c* oxidase subunit I (COI)-based DNA barcode library for Korean chironomids. This library consists of 211 specimens from 49 species, including adults and unidentified larvae. The interspecies and intraspecies COI sequence variations were analyzed. Sophisticated indexes were developed in order to properly evaluate indistinct barcode gaps that are created by insufficient sampling on both the interspecies and intraspecies levels and by variable mutation rates across taxa. In a variety of insect datasets, these indexes were useful for re-evaluating large barcode datasets and for defining COI barcode gaps. The COI-based DNA barcode library will provide a rapid and reliable tool for the molecular identification of Korean chironomid species. Furthermore, this reverse-taxonomic approach will be improved by the continuous addition of other speceis' sequences to the library.

INTRODUCTION

Freshwater non-biting midges (Diptera: Chironomidae) include many species that cause human allergies (Baur, 1992; Lewis, 1956). Their larvae are important indicator organisms for water pollution because of their ability to thrive under a variety of freshwater conditions (Aagaard et al., 2004; Wright, 1984), including exposure to heavy metals, organic pesticides, and other xenobiotics (Al-Shami et al., 2010; Martinez et al., 2004). Although the correct identification of these species is of fundamental importance for subsequent studies, species identification is problematic, especially for immature specimens, larvae, damaged specimens, and those with morphological deformities (Pfenninger et al., 2007; Velle et al., 2005). In such cases, a molecular-based approach can be used for species identification (Tautz et al., 2002).

DNA barcoding is a molecular diagnostic method that employs a short DNA sequence to rapidly and accurately identify a species. This method has revitalized traditional taxonomy and allows for a better understanding of organisms and their relationships (Hebert and Gregory, 2005; Hebert et al., 2003; Kim et al., 2011; Yoo et al., 2006). However, DNA barcoding methods may not provide adequate resolution to identify recently diverged species, species complexes, or groups with a slow evolutionary rate (Kerr et al., 2007; Radulovici et al., 2010). Such cases require suitable markers for accurate species identification (Park et al., 2007). Another problem with DNA barcoding methods is that ambiguous barcode gaps may be formed by insufficient sampling at both the interspecies and intraspecies levels (Meier et al., 2008; Meyer and Paulay, 2005; Wiemers and Fiedler, 2007) and by the variable mutation rates across taxa (Galtier et al., 2009).

The present study aims to prepare a DNA barcode library that can accurately identify Korean chironomid species of various developmental stages. First, the mitochondrial cytochrome *c* oxidase subunit I (COI) gene was chosen as a genetic marker for DNA barcoding, because COI is suitable for studying evolution and for discriminating cryptic chironomid species (Carew et al., 2005; Ekrem et al., 2007). Second, a COI-based Korean chironomid barcode library was constructed, and both interspecies and intraspecies sequence variations were analyzed. Third, sophisticated indexes were developed to properly define barcode gaps between species, and the utility of these indexes was evaluated in various insect barcode datasets. Finally, the use of the reverse-taxonomic approach was discussed.

MATERIALS AND METHODS

Sampling of chironomid species

Chironomid adults attracted to the lights of stores and restaurants were aspirated at 16 locations (Table 1). The collected insects were stored in 75% ethanol. The antennae, head, wings,

[1]Interdisciplinary Program in Bioinformatics, Seoul National University, Seoul 151-747, Korea, [2]Wildlife Genetic Resources Center, National Institute of Biological Resources, Incheon 404-708, Korea, [3]Department of Environmental Medical Biology, Institute of Tropical Medicine, and Korean National Arthropods of Medical Importance Resource Bank, Yonsei University College of Medicine, Seoul 120-752, Korea, [4]School of Biological Sciences, Seoul National University, Seoul 151-747, Korea
*Correspondence: wonkim@plaza.snu.ac.kr

Received July 20, 2011; revised November 1, 2011; accepted November 1, 2011; published online December 1, 2011

Keywords: chironomidae, DNA barcode gap, DNA barcode library, DNA barcoding, reverse taxonomy

abdomen, and hypopygium were later dissected using fine needles under a stereomicroscope and were mounted in Hoyer's solution. All specimens were morphologically identified into species, and some specimens from each species were used for PCR analysis and DNA sequencing. The thoraces or legs were used for molecular studies, and the other body parts were mounted on slides for morphological studies.

Larval chironomids were collected at the Geumsa wetland, located in Geumsa-ri, Geumsa-myeon, Yeoju-gun, and Gyeonggi-do, Korea. The larvae were washed through a 1-mm sieve, preserved in water for 5 days to remove gut substances (Pfenninger et al., 2007), and stored in 75% ethanol. The larval head capsule and the body segment were stored in 75% ethanol for future morphological examination. A small part of the larval body was used for DNA sequencing. All specimens and specimen IDs were deposited into the Arthropods of Medical Importance Resource Bank, Department of Environmental Medical Biology, Yonsei University, Seoul, Korea.

DNA extraction and amplification

Genomic DNA was extracted from all samples using the Qiagen DNeasy Blood and Tissue kit. Standard PCR amplification and DNA sequencing protocols were used to sequence a fragment of the COI. Primers used for PCR were LCO1490 (5′-GGT CAA CAA ATC ATA AAG ATA TTG G-3′) and HCO2198 (5′-TAA ACT TCA GGG TGA CCA AAA AAT CA-3′) (Folmer et al., 1994). The 5′ region of COI was amplified using the following thermal cycling program: 94°C for 5 min, 35 cycles at 94°C for 0.5 min, 48°C for 1 min, and 72°C for 1.5 min, followed by a final extension at 72°C for 10 min. The 25 μl PCR reaction mixture included 14.7 μl of ultrapure water, 5 μl of 5× PCR buffer, 1 μl of each primer (10 μM), 1 μl of dNTPs (10 mM), 0.3 μl of Taq polymerase (5 U), and 2 μl of the DNA template. PCR products were purified using QIAquick PCR Purification Kit (Qiagen). Sequencing reactions were resolved on an ABI 3730 automated DNA sequencer.

DNA barcode analysis and tree building

All sequences were deposited into GenBank (Table 1) and aligned using MUSCLE software which implements several iterative refinements steps (Edgar, 2004). Interspecies and intraspecies sequence divergences were quantified using the Kimura-2 parameter (K2P) distance model (Kimura, 1980), and a neighbor-joining (NJ) tree was constructed by MEGA v4.0.2 (Tamura et al., 2007) with 10,000 bootstrap replicates.

Development of indexes for defining DNA barcode gap

Sophisticated indexes were developed to define the DNA barcode gap between species. The equations used were as follows:

$$D_{minInter} = \min(D_{GeneticModel}(X_i, Y_i)) \quad (1)$$

and

$$D_{maxIntra} = \max(D_{GeneticModel}(X_i, X_j)),\ i \neq j \quad (2)$$

where $X = \{X_1, X_2, \ldots, X_n\}$ is the set of each target species, and $Y = \{Y_1, Y_2, \ldots, Y_m\}$ is the set of non-target species that include all of the species in a barcode except the target species. $D_{GeneticModel}$ is the function used to calculate several genetic distances, such as Jukes-Cantor (Jukes and Cantor, 1969), Kimura 2-Parameter (Kimura, 1980), Tamura 3-Parameter (Tamura, 1992), and Tamura-Nei (Tamura and Nei, 1993). Each genetic distance model produced the same conclusion. We used the D_{K2P} function, which applies the K2P distance model that is commonly used for barcode analysis. X_i represents the sequences of the target species that belong to the same species, and Y_j represents sequences of non-target species. $D_{minInter}$ is the minimum interspecies distance between the target species and the non-target species. $D_{maxIntra}$ is the maximum intraspecies distance within the target species.

$$D_M = \frac{\sum_{i=0}^{n}\sum_{j=0}^{m} D_{K2P}(X_i, Y_j)}{n \times m} \quad (3)$$

D_M is the mean value of interspecies distances.

$$Index_{BarcodeGap} = \frac{|D_M - D_{minInter}|}{|D_M - D_{maxIntra}|} \quad (4)$$

$Index_{BarcodeGap}$ is the equation used to properly assess the DNA barcode gap for each target species.

$$Index_{Cutoff} = \frac{|GroupMax(D_M) - GroupMin(D_{minInter})|}{|GroupMax(D_M) - GroupMax(D_{maxIntra})|},\ \text{if } Index_{Cutoff} \geq \text{an empirical value, then an empirical value is } Index_{Cutoff} \quad (5)$$

$Index_{Cutoff}$ is the optimal cutoff value. In the formula, a function of GroupMax represents a maximum value determined on a box plot graph in a group of data, and a function of GroupMin represents a minimum value determined on a box plot graph in a group of data. The cutoff value varies in accordance with the taxon group and size of the barcode dataset. When the maximum intraspecies and the minimum interspecies distances overlap, this indicates an absence of DNA barcode gap. In this case, $Index_{Cutoff}$ is greater than the empirical value determined after analyzing the available insect barcode datasets.

$$\begin{cases} Index_{BarcodeGap} \leq Index_{Cutoff} : \text{Presence of DNA barcode gap} \\ Index_{BarcodeGap} > Index_{Cutoff} : \text{Absence of DNA barcode gap} \end{cases} \quad (6)$$

These conditions of $Index_{BarcodeGap}$ and $Index_{Cutoff}$ are used to assess whether DNA barcode gap exists in each species. For example, if the $Index_{BarcodeGap} > Index_{Cutoff}$, and the $D_{maxIntra}$ of the target species is greater than the average $D_{maxIntra}$ of the other species, then the target species is considered a cryptic species. On the other hand, if the $Index_{BarcodeGap} > Index_{Cutoff}$, and the $D_{minInter}$ of the target species less than the average $D_{minInter}$ of the other species, then the target species may be misidentified or may require more sensitive molecular markers for identification.

RESULTS

DNA barcode analysis of Korean chironomids

The 5′ end of the COI gene was sequenced in a total of 213 specimens (185 adults and 28 larvae, Table 1). Amplified DNA was 658 bp in length, excluding the primer regions. The sequences were translated to detect nuclear mitochondrial pseudo-genes (NUMTs) that are common in eukaryotes, including indels, frame-shift mutations, and in-frame stop codons (Bensasson et al., 2001; Song et al., 2008). Deletions were discovered in 2 sequences from *Dicrotendipes pelochloris*, and these sequences were excluded from the analysis. A high level of intraspecies K2P variation was discovered in *Dicrotendipes septemmaculatus* (0-0.131) and *Tanytarsus yoni* (0-0.120), while a low level of interspecies K2P variation was observed between *Cricotopus sylvestris* and *Cricotopus tricinctus* (0.009-0.015) and between *Chironomus* sp. larva_3 and *Procladius*

Table 1. Summary of the sequence dataset: collection location, number of sequences, and GenBank accession numbers

Taxon name	Collection location	No. of sequences	GenBank accession number (COI)	Voucher ID
Ablabesmyia longistyla	UCG, GCG	6	JN887039-JN887044	425, 431, 432, 433, 501, 506
Ablabesmyia monilis	UCG	1	JN887045	360
Chironomus dorsalis	GYG, GCG	6	JF412058, JF412059, JN887046-JN887049	449, 450, 481-484
Chironomus flaviplumus	SGYG	3	JF412075-JF412077	224-226
Chironomus fujiprimus	GYG,	4	JF412078-JF412081	376-378, 448
Chironomus javanus	HJC	4	JF412082-JF412085	315-317, 319
Chironomus kiiensis	UCG	4	JF412086-JF412089	276-279
Chironomus nipponensis	UCG, GYG, GCG	12	JF412090-JF412097, JN887050-JN887053	272-275, 280, ch20-ch22, 490, 494-496
Chironomus sp. larva_2	GYG	4	JF412060-JF412063	ch10-ch12, ch32
Chironomus sp. larva_3	GYG	3	JF412064-JF412066	ch26-ch28
Chironomus sp. larva_5	GYG	1	JF412068	ch18
Chironomus sp. larva_7	MG	4	JF412071-JF412074	sd1-sd4
Chrionomus plumosus	JSJ, GCG	4	JF412098-JF412100, JN887054	335-337, 493
Cladotanytarsus vanderwulpi	GYG	8	JF412101-JF412108	387-391, 397-399
Cricotopus bicinctus	UCG, GGS	4	JN887055-JN887058	359, 452-454
Cricotopus bimaculatus	UCG	3	JN887059-JN887061	355, 356, 358
Cricotopus sp.	GCG	1	JN887062	502
Cricotopus sylvestris	GYG, BJC	4	JF412070, JN887063-JN887065	ch25, 485, 487, 489
Cricotopus tricinctus	GYG	3	JN887066-JN887068	392-394
Dicrotendipes pelochloris	SYJ	4	JF412109-JF412112	320, 321, 323, 324
Dicrotendipes septemmaculatus	HJC, SYJ, UCG	8	HQ846344-HQ846351	297-300, 325-327, 357
Einfeldia dissidens	GYG	4	JF412113-JF412116	ch31, ch33, 381, 382
Eukiefferiella sp.	GCG	1	JN887069	499
Glyptotendipes tokunagai	GYG, JSJ	11	JF412117-JF412127	ch4-ch6, ch15, ch29, ch30, 338-342
Limnochironomus nervosus	HJC	4	JF412128-JF412131	301-304
Monodiamesa bathyphila	UCG	1	JN887070	421
Orthocladius yugashimaeusis	JNS	3	JN887071-JN887073	435, 436, 438
Parachironomus arcuatus	SB	2	JF412132, JF412133	409, 410
Paracladopelma camptolabis	UCG, GCG	3	JF412134, JN887074, JN887075	424, 500, 507
Parakiefferiella bathophila	UCG, GCG	13	JN887076-JN887088	361, 367-369, 372-375, 383-386, 511
Paratanytarsus haisooni	UCG	2	JF412135, JF412136	286, 287
Paratanytarsus inopertus	GCG	1	JN887089	508
Paratrichocladius rufiventris	JNS	4	JN887090-JN887093	477-480
Pentapedilum kaswiense	JSJ	2	JF412137, JF412138	348, 349
Polypedilum masudai	UCG, SYJ	8	JF412139-JF412146	288-292, 331-333
Polypedilum nubifer	UCG	5	JF412147-JF412151	281-285
Polypedilum scalaenus	JSJ	4	JF412152-JF412155	343, 345-347
Polypedilum ureshinoensis	HJC	1	JF412156	308
Polypedilum yonsanensis	GYG	5	JF412157-JF412161	ch7-ch9, 417, 418
Procladius choreus	UCG	4	JN887094-JN887097	293-296
Stictochironomus sinsauensis	UCG	6	JF412162-JF412167	420, 426-430
Tanypus punctipennis	SGYG	3	JN887098-JN887100	214-216
Tanytarsus ahyoni	JNS, GCG	2	JF412168, JN887101	434, 521
Tanytarsus kiseogi	HJC, GYG	5	JF412169-JF412173	305-307, 309, ch34
Tanytarsus sp.	GCG	5	JN887102-JN887106	512-514, 523, 524
Tanytarsus talcahashii	GCG	1	JN887107	522
Tanytarsus tamagotoi	UCG, GCG	14	JF412174-JF412185, JN887108, JN887109	350-354, 362-366, 422, 423, 509, 510
Tanytarsus yoni	GCG	6	JN887110-JN887115	515-520
Tokunagayusurika akamusi	HDC	2	JN887116, JN887117	455, 456

Gwanak-ro, Gwanak-gu, Seoul (GGS); Ichon-dong, Yongsan-gu, Seoul (IYS); Junggye-dong, Nowon-gu, Seoul (JNS); Geumsa-myeon, Yeoju-gun, Gyeonggi-do (GYG); Sin-gal-dong, Giheung-gu, Yongin-si, Gyeonggi-do (SGYG); Geunhwa-dong, Chuncheon-si, Gangwon-do (GCG); Udu-dong, Chuncheon-si, Gangwon-do (UCG); Hapdeok-eup, Dangjin-gun, Chungcheongnam-do (HDC); Baegun-myeon, Jecheon-si, Chungcheongbuk-do (BJC); Hansu-myeon, Jecheon-si, Chungcheongbuk-do (HJC); Samho-eup, Yeongam-gun, Jeollanam-do (SYJ); Jeonju-si, Jeollabuk-do (JJ); Changn-yeong-gun, Gyeongsangnam-do (CG); Miryang-si, Gyeongsangnam-do (MG); Saha-gu, Busan (SB); Jungmun-dong, Seogwipo-si, Jeju-do (JSJ)

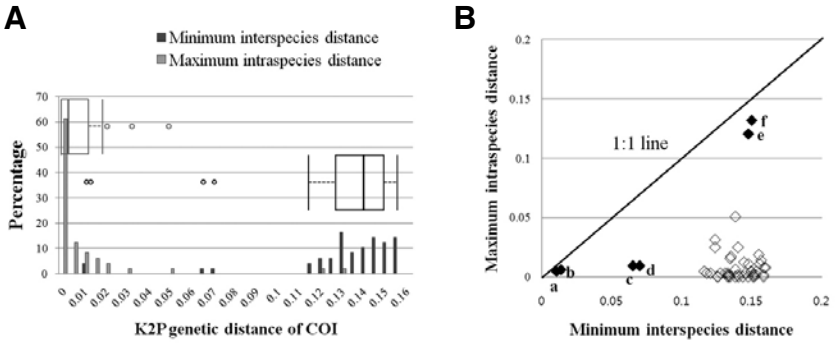

Fig. 1. (A) Distribution and box plots of Kimura 2-parameter (K2P) genetic distance of mitochondrial cytochrome *c* oxidase I (COI) among 49 species. (B) Minimum interspecies and maximum intraspecies distances for COI gene regions were calculated by the K2P genetic distance model. The bottom region of the 1:1 line represents the presence of a DNA barcode gap. Diamonds (black) near the 1:1 line indicate controversial species (a: *Cricotopus tricinctus*; b: *Cricotopus sylvestris*; c: *Chironomus* sp. larva_3; d: *Procladius choreus*; e: *Tanytarsus yoni*; f: *Dicrotendipes septemaculatus*).

choreus (0.065-0.075). The average minimum interspecies- and maximum intraspecies distances were 0.140 (n = 183, standard deviation = 0.016) and 0.006 (n = 183, standard deviation = 0.000), respectively, for all species except controversial species, including *Dicrotendipes septemmaculatus*, *Tanytarsus yoni*, *Cricotopus sylvestris*, *Cricotopus tricinctus*, *Chironomus* sp. larva_3, and *Procladius choreus*. There was an approximately 20-fold gap between the minimum interspecies- and maximum intraspecies variation (Fig. 1A). We also mapped the distribution of minimum interspecies- and maximum intraspecies distances of each species. The COI clearly discriminated each species group with consistent sequence lengths and significant interspecies K2P variation (0.175-0.213) (Fig. 1B). The NJ trees showed clear groupings at the species level (Fig. 2). Sixteen among 28 larvae were grouped with known species in the chironomid barcode library, namely, *Glyptotendipes tokunagai* (n = 6), *Chironomus nipponensis* (n = 3), *Einfeldia dissidens* (n = 2), *Tanytarsus kiseogi* (n = 1), *Polypedilum yonsanensis* (n = 3), and *Cricotopus sylvestris* (n = 1). The remaining 12 larval specimens were unidentified, namely, *Chironomus* sp. larva_2 (n = 4), *Chironomus* sp. larva_3 (n = 3), *Chironomus* sp. larva_5 (n = 1), and *Chironomus* sp. larva_7 (n = 4). There were no technical difficulties in amplifying DNA from specimens of diverse developmental stages.

Evaluation of the utility of $Index_{BarcodeGap}$ and $Index_{Cutoff}$

We used 7 insect datasets downloaded from the BOLD system to evaluate the newly developed indexes: Profile 100 Insect Families (SBGB), Ephemeroptera of North America (EPCHU) (Zhou et al., 2009), Trichoptera of Churchill 2002-2007 (PHCAD, TRCAP, MHTRI, DSTRI, and CUCAD) (Zhou et al., 2009), Chilean Smicridea (CLSMD) (Pauls et al., 2010), Caddisflies of New Zealand (NZCAD), Moths of Ontario (PMMO), and BOLD chironomids (GBDPC, COTW, and CAUS). NUMTs were excluded according to the criteria of Song et al. (2008). Only sequences longer than 500 bp were included. We then calculated $Index_{BarcodeGap}$ and $Index_{Cutoff}$ (Supplementary File 1). We determined an empirical value (Equation 5), 0.850, which represents the absence of DNA barcode gaps (Fig. 3). The presence of an unambiguous barcode gap between species was assessed on the basis of the conditions of $Index_{BarcodeGap}$ in relation to $Index_{Cutoff}$. Of the 850 species assessed, we found 70 controversial species groups with ambiguous barcode gaps from EPCHU (n = 1), Trichoptera of Churchill (n = 1), CLSMD (n = 1), NZCAD (n = 9), PMMO (n = 36), and BOLD chironomids (n = 22). Of these groups, 12 cryptic species were predicted (Table 2) (Fig. 4). Among BOLD chironomids, 8 cryptic species were predicted. We applied the distance conditions of $Index_{BarcodeGap}$ and $Index_{Cutoff}$ to the Korean chironomid barcode library. As a result, *Dicrotendipes septemmaculatus* and *Tanytarsus yonis*, which showed the largest intraspecies variations in the previous analysis, were predicted as cryptic species (Table 2).

DISCUSSION

COI barcode library of Korean chironomids

The current Korean chironomid barcode library comprises 211 sequences from 49 species. In this library, there is an approximately 20-fold gap between minimum interspecies- and maximum intraspecies distance in 43 species, with the exception of 6 controversial species. The 20-fold gap indicates a high degree of species-level resolution, especially compared to the 10-fold gap previously suggested for discriminating species (Hebert et al., 2003). Barcode analysis of the BOLD chironomid dataset, consisting 125 species, also showed consistent genetic differences between species (0.171-0.254), with the exception of 22 controversial species. Therefore, although COI-based identification of Chironomidae is a promising approach, it is not equally effective in all species (e.g., the closely related *Cricotopus sylvestris* and *Cricotopus tricinctus* species in the Chironomidae family). Genetic variation can also differ considerably among different taxa (Kerr et al., 2007). For example, the average ratios of minimum interspecific distance to maximum intraspecific distance were 7.534, 5.945, 15.449, 6.826, 6.483, and 6.475 for Ephemeroptera of North America, Trichoptera of Churchill, Chilean Smicridea, Caddisflies of New Zealand, Moths of Ontario, and BOLD chironomids, respectively (Fig. 4). Moreover, Moths of Ontario were characterized by low interspecies variation and considerable overlap between minimum interspecies- and maximum intraspecies variation (Fig. 4). Thus, in this case, molecular markers in addition to COI may be required to increase the resolution of identification. Alternatively, new sophisticated indexes for defining barcode gap must be used, as described below.

Developing the new indexes and detecting cryptic species

Insufficient interspecies and intraspecies sampling (Wiemers and Fiedler, 2007) and variable mutation rates across taxa (Galtier et al., 2009) can create difficulties in interpreting and analyzing DNA barcode gap. We developed new indexes that enable rapid and accurate assessment of DNA barcode gap. These new indexes were essential for the following reasons. First, although several other statistical approaches have been proposed (Kim et al., 2010; Nielsen and Matz, 2006; Steinke et al., 2005), these have proven too complex to apply and translate. Second, the range of overlap between interspecies- and intraspecies variation depends on the taxonomic group (Meier

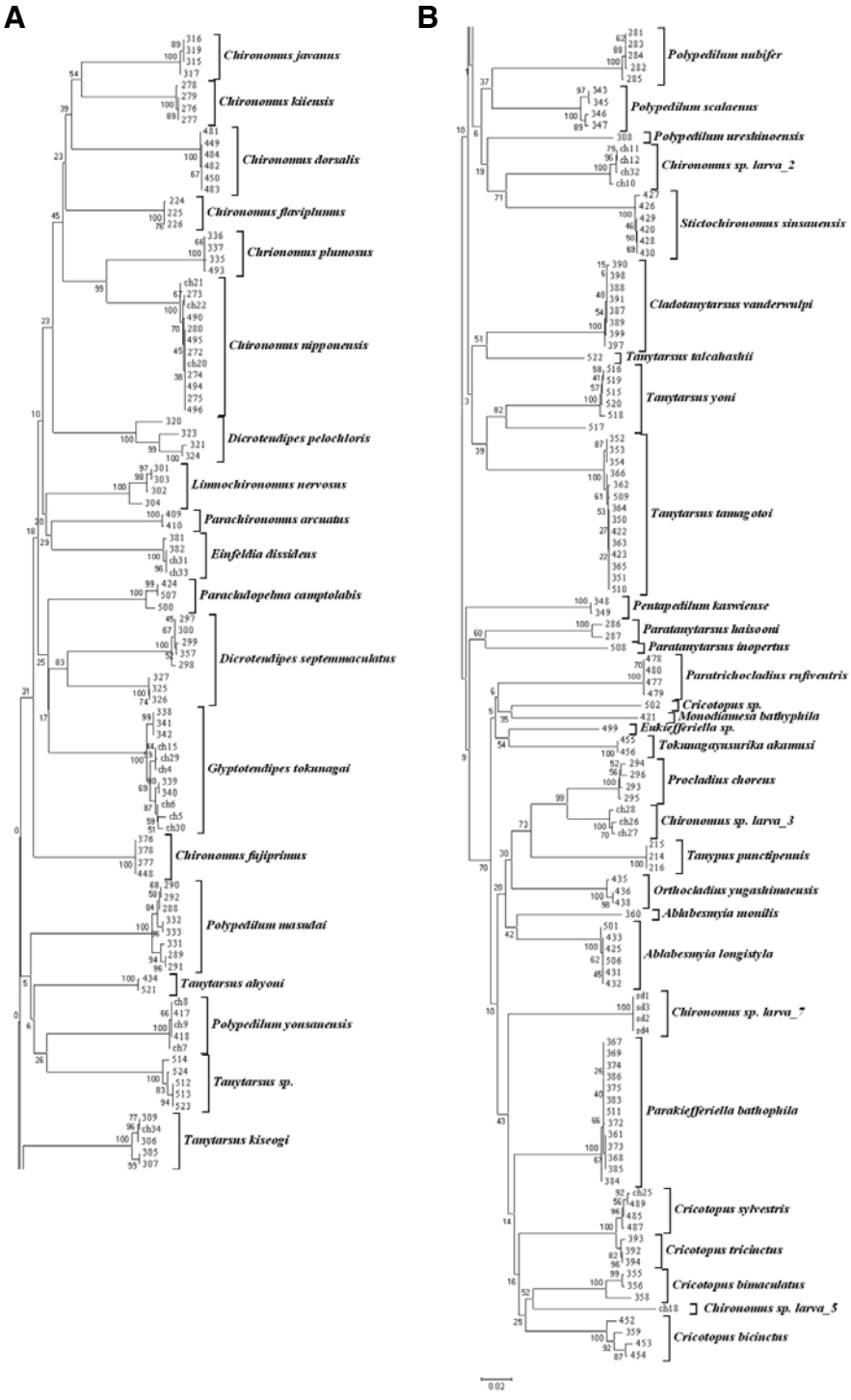

Fig. 2. Neighbor-joining (NJ) tree of K2P distances based on mtDNA COI. The bootstrap value of 10,000 replicates is shown on each node.

et al., 2006). With the rate of mitochondrial evolution reported by Hebert et al. (2003), a 10-fold gap is not always accurate. Third, data can be interpreted differently even with similar ratios of interspecies variation to intraspecies variation. For example, in the Korean chironomid dataset, *Dicrotendipes pelochloris* has sufficient variation among species (a ratio of 2.735, an interspecies distance of 0.138, an intraspecies distance of 0.050, and an $Index_{BarcodeGap}$ of 0.341), but *Cricotopus tricinctus* has low variation among species (a ratio of 2.344, an interspecies distance of 0.010, an intraspecies distance of 0.004, and an $Index_{BarcodeGap}$ of 0.967). In order to solve these problems, we defined the $D_{M,}$ which is derived from the difference between the target species and other species (Equation 3). The value of D_M relies on the genetic variation of the target species and the dataset. Lastly, a DNA barcode gap may be misinterpreted when the intraspecies variation is zero. However, both the D_M-$D_{minInter}$ and the D_M-$D_{maxIntra}$ equations resolve this problem (Equation 4). We calculated $Index_{BarcodeGap}$ and $Index_{Cutoff,}$

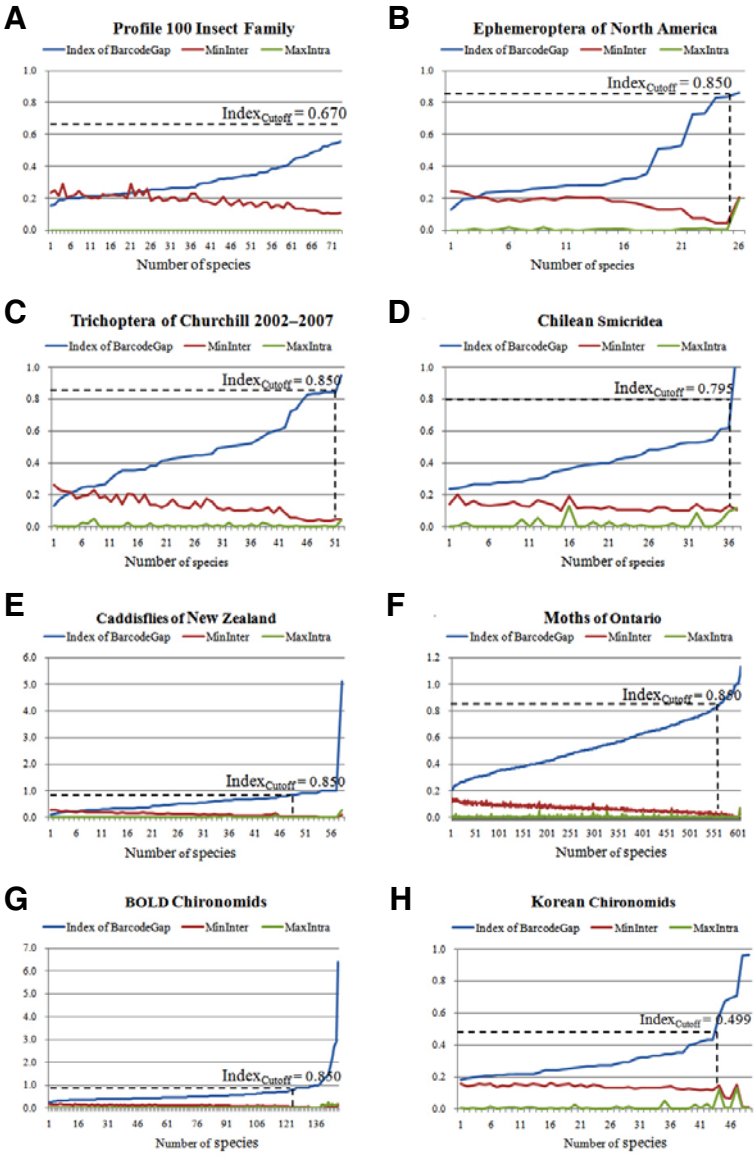

Fig. 3. Optimal cutoff value ($Index_{Cutoff}$), minimum interspecies distance (MinInter), and maximum intraspecies distance (MaxIntra) in the barcode datasets. These values are sorted by $Index_{Cutoff}$ in ascending order.

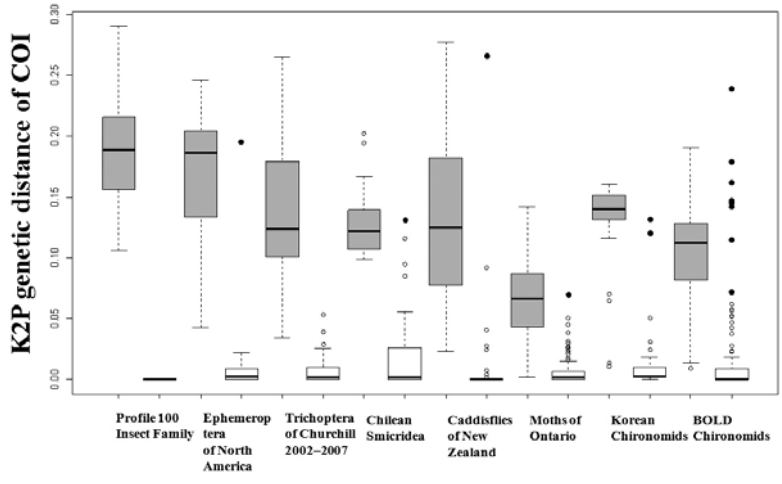

Fig. 4. Box plots depicting differences between the minimum interspecies variation (gray box plots) and the maximum intraspecies variation (white box plots) using the K2P distance model. The black dots indicate predicted cryptic species in each group.

Table 2. Summary of COI genetic divergence using the K2P distance model to predict cryptic species in insect datasets

Name of dataset	Cryptic species	Minimum interspecies distance	Maximum intraspecies distance	Mean inter-species distance	$Index_{BarcodeGap}$	$Index_{Cutoff}$
Ephemeroptera of North America	*Acentrella turbida*	0.204	0.196	0.258	0.862	0.850
Chilean Smicridea	*Smicridea annulicornis*	0.109	0.116	0.174	1.120	0.795
Caddisflies of New Zealand	*Neurochorema confusum*	0.110	0.266	0.240	5.121	0.925
Moths of Ontario	*Xanthorhoe labradorensis*	0.066	0.070	0.122	1.07	0.850
Korean Chironomids	*Dicrotendipes septemmaculatus*	0.150	0.132	0.194	0.709	0.499
	Tanytarsus yoni	0.148	0.121	0.186	0.584	0.499
BOLD Chironomids	*Chironomus bernensis*	0.033	0.146	0.205	2.910	0.850
	Chironomus dorsalis	0.108	0.179	0.192	6.401	0.850
	Chironomus oppositus	0.000	0.118	0.189	2.670	0.850
	Chironomus plumosus	0.042	0.115	0.211	1.749	0.850
	Echinocladius martini	0.107	0.162	0.207	2.218	0.850
	Paraphaenocladius impensus	0.123	0.142	0.201	1.315	0.850
	Riethia stictoptera	0.118	0.239	0.190	1.430	0.850
	Tanytarsus inaequalis	0.135	0.145	0.212	1.158	0.850

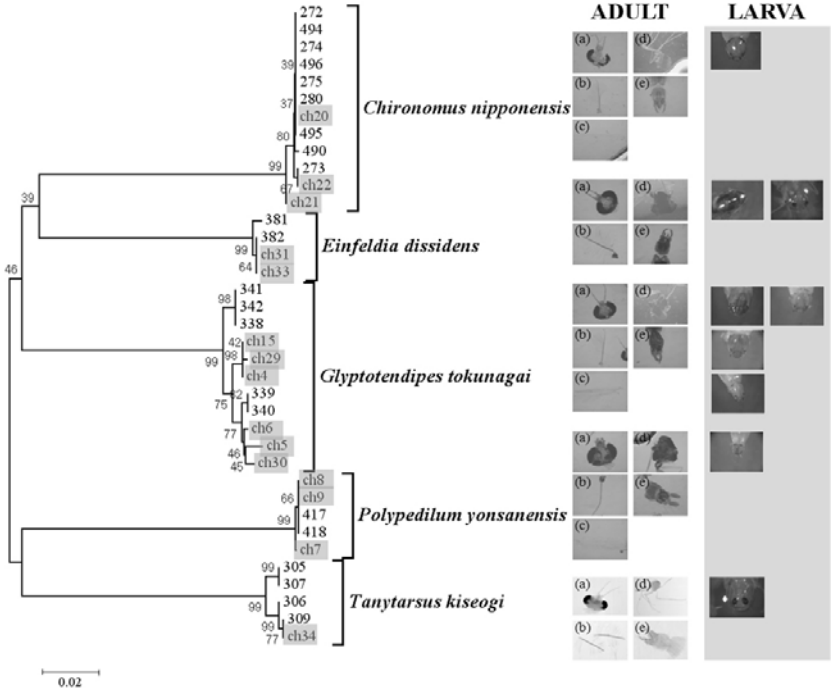

Fig. 5. Example of reverse taxonomy. Data include each clade in a NJ tree, bootstrap values on the branches, and the morphological characteristics of the adults [(a) head, (b) antennae, (c) wing, (d) abdomen, and (e) hypopygium] and larvae (gray box, showing the head capsule).

and tested their utility in 7 insect barcode datasets. The performance of $Index_{BarcodeGap}$ and $Index_{Cutoff}$ was better than using fold-change for defining the DNA barcode gap. With these indexes, we were able to detect 14 cryptic species among 70 controversial species groups (Table 2). In the case of *Acentrella turbida* of EPCHU and *Smicridea annulicornis* of CLSMD, the cryptic species had already been described in previous studies (Pauls et al., 2010; Zhou et al., 2009). The remaining 56 controversial species could have been misidentified or may require more sensitive molecular markers for better identification.

Reverse taxonomy using the DNA barcode library

Due to the heteromorphic life cycles of chironomids, deformities caused by freshwater pollution, and huge species diversity, traditional approaches are insufficient for species identification. In contrast, DNA barcoding enables species identification and

clarification of morphological characteristics. This process is also known as reverse taxonomy (Blaxter et al., 2005; Markmann and Tautz, 2005). DNA barcoding is a highly efficient approach that can provide clues for ecological correlations by linking the adult and immature stages, which are not taxonomically advanced. Moreover, the approach can be used to assess environmental conditions (Aagaard et al., 2004) and to infer past environments from larval head capsule fossils (Velle et al., 2005). In the present study, we successfully performed a reverse-taxonomic approach using COI signature sequences of the barcode library. Sixteen among 28 larvae were successfully identified, and thus, larval morphological characteristics (e.g., larval head capsule) could be explored (Fig. 5). The remaining 12 larvae were unidentified, due to the absence of adult sequence data in the current barcode library. In studies that employ DNA barcoding and pyrosequencing technology (Deagle et al., 2010; Zeale et al., 2010), a reverse- taxonomic approach might be ideal for understanding the diversity of organisms.

Future direction of the DNA barcode library

The COI-based DNA barcode library described in the present study is useful for the identification of Korean chironomid species regardless of developmental stages. When used with newly developed indexes for defining barcode gaps, the library also allowed us to identify cryptic species. The present study is the first approach for the molecular identification of Korean chironomids based on DNA barcoding. Therefore, this DNA barcode library should be improved by continuously adding new sequence data from additional species. This is particularly important, because approximately 80 species of chironomid midges have been reported in Korea (Ree, 2009), and many more species are expected to be identified.

Note: Supplementary information is available on the Molecules and Cells website (www.molcells.org).

ACKNOWLEDGMENTS

We thank Tae-Kun Seo for valuable comments on the manuscript. This work was supported by a grant from the Eco-Technopia 21 Project, which was funded by the Ministry of Environment (2008-05002-0065-0) and by the Brain Korea 21 Research Fellowship from the Ministry of Education, Science and Technology, Republic of Korea.

REFERENCES

Aagaard, K., Solem, J.O., Bongard, T., and Hanssen, O. (2004). Studies of aquatic insects in the Atna River 1987-2002. Hydrobiologia *521*, 87-105.

Al-Shami, S., Rawi, C.S.M., Nor, S.A.M., Ahmad, A.H., and Ali, A. (2010). Morphological deformities in *Chironomus* spp. (Diptera: Chironomidae) larvae as a tool for impact assessment of anthropogenic and environmental stresses on three rivers in the Juru River System, Penang, Malaysia. Environ. Entomol. *39*, 210-222.

Baur, X. (1992). Chironomid midge allergy. Arerugi *41*, 81-85.

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

Blaxter, M., Mann, J., Chapman, T., Thomas, F., Whitton, C., Floyd, R., and Abebe, E. (2005). Defining operational taxonomic units using DNA barcode data. Phil. Trans. R. Soc. B *360*, 1935-1943.

Carew, M.E., Pettigrove, V., and Hoffmann, A.A. (2005). The utility of DNA markers in classical taxonomy: using cytochrome oxidase I markers to differentiate Australian *Cladopelma* (Diptera: Chironomidae) midges. Ann. Entomol. Soc. Am. *98*, 587-594.

Deagle, B.E., Chiaradia, A., McInnes, J., and Jarman, S.N. (2010). Pyrosequencing faecal DNA to determine diet of little penguins: is what goes in what comes out? Conserv. Genet. *11*, 2039-2048.

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

Ekrem, T., Willassen, E., and Stur, E. (2007). A comprehensive DNA sequence library is essential for identification with DNA barcodes. Mol. Phylogenet. Evol. *43*, 530-542.

Folmer, O., Black, M., Hoeh, W., Lutz, R., and Vrijenhoek, R. (1994). DNA primers for amplification of mitochondrial cytochrome *c* oxidase subunit I from diverse metazoan invertebrates. Mol. Mar. Biol. Biotech. *3*, 294-299.

Galtier, N., Nabholz, B., Glemin, S., and Hurst, G.D.D. (2009). Mitochondrial DNA as a marker of molecular diversity: a reappraisal. Mol. Ecol. *18*, 4541-4550.

Hebert, P.D.N., and Gregory, T.R. (2005). The promise of DNA barcoding for taxonomy. Syst. Biol. *54*, 852-859.

Hebert, P.D.N., Ratnasingham, S., and deWaard, J.R. (2003). Barcoding animal life: cytochrome c oxidase subunit 1 divergences among closely related species. Proc. R. Soc. Lond. B *270*, 96-99.

Jukes, T.H., and Cantor, C.R. (1969). Evolution of protein molecules. In Mammalian Protein Metabolism. Munro H.H, ed. (New York: Academic Press), pp. 21-132.

Kerr, K.C.R., Stoeckle, M.Y., Dove, C.J., Weigt, L.A., Francis, C.M., and Hebert, P.D.N. (2007). Comprehensive DNA barcode coverage of North American birds. Mol. Ecol. Notes *7*, 535-543.

Kim, S., Eo, H.S., Koo, H., Choi, J.K., and Kim, W. (2010). DNA Barcode-based molecular identification system for fish species. Mol. Cells *30*, 507-512.

Kim, S., Kim, C.B., Min, G.S., Suh, Y., Bhak, J., Woo, T., Koo, H., Choi, J.K., Shin, M.k., Jung, J., et al. (2011). Korea Barcode of Life Database System (KBOL). Animal Cells & Systems Published in Proceedings.

Kimura, M. (1980). A simple method for estimating evolutionary rate of base substitution through comparative studies of nucleotide sequences. J. Mol. Evol. *16*, 111-120.

Lewis, D.J. (1956). Chironomidae as a pest in the Northern Sudan. Acta Trop. *13*, 142-158.

Markmann, M., and Tautz, D. (2005). Reverse taxonomy: an approach towards determining the diversity of meiobenthic organisms based on ribosomal RNA signature sequences. Philos. Trans. R. Soc. B Biol. Sci. *360*, 1917-1924.

Martinez, E.A., Moore, B.C., Schaumloffel, J., and Dasgupta, N. (2004). Teratogenic versus mutagenic abnormalities in chironomid larvae exposed to zinc and lead. Arch. Environ. Contam. Toxicol. *47*, 193-198.

Meier, R., Shiyang, K., Vaidya, G., and Ng, P.K.L. (2006). DNA barcoding and taxonomy in Diptera: a tale of high intraspecific variability and low identification success. Syst. Biol. *55*, 715-728.

Meier, R., Zhang, G., and Ali, F. (2008). The use of mean instead of smallest interspecific distances exaggerates the size of the "barcoding gap" and leads to misidentification. Syst. Biol. *57*, 809-813.

Meyer, C.P., and Paulay, G. (2005). DNA barcoding: error rates based on comprehensive sampling. PLoS Biol. *3*, e422.

Nielsen, R., and Matz, M. (2006). Statistical approaches for DNA barcoding. Syst. Biol. *55*, 162-169.

Park, M.H., Sim, C.J., Baek, J., and Min, G.S. (2007). Identification of genes suitable for DNA barcoding of morphologically indistinguishable Korean Halichondriidae sponges. Mol. Cells *23*, 220-227.

Pauls, S.U., Blahnik, R.J., Zhou, X., Wardwell, C.T., and Holzenthal, R.W. (2010). DNA barcode data confirm new species and reveal cryptic diversity in Chilean *Smicridea* (Trichoptera: Hydropsychidae). J. North Am. Benthol. Soc. *29*, 1058-1074.

Pfenninger, M., Nowak, C., Kley, C., Steinke, D., and Streit, B. (2007). Utility of DNA taxonomy and barcoding for the inference of larval community structure in morphologically cryptic *Chironomus* (Diptera) species. Mol. Ecol. *16*, 1957-1968.

Radulovici, A.E., Archambault, P., and Dufresne, F. (2010). DNA barcodes for marine biodiversity: moving fast forward? Diversity *2*, 450-472.

Ree, H.I. (2009). One new and six unrecorded species of chironomidae (Insecta: Diptera) in Korea. Korean J. Syst. Zool. *25*, 49-59.

Song, H., Buhay, J.E., Whiting, M.F., and Crandall, K.A. (2008). Many species in one: DNA barcoding overestimates the number

of species when nuclear mitochondrial pseudogenes are coamplified. Proc. Natl. Acad. Sci. USA *105*, 13486-13491.
Steinke, D., Vences, M., Salzburger, W., and Meyer, A. (2005). TaxI: a software tool for DNA barcoding using distance methods. Phil. Trans. R. Soc. B *360*, 1975-1980.
Tamura, K. (1992). Estimation of the number of nucleotide substitutions when there are strong transition-transversion and G+ C-content biases. Mol. Biol. Evol. *9*, 678-687.
Tamura, K., and Nei, M. (1993). Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. Mol. Biol. Evol. *10*, 512-526.
Tamura, K., Dudley, J., Nei, M., and Kumar, S. (2007). MEGA4: molecular evolutionary genetics analysis (MEGA) software version 4.0. Mol. Biol. Evol. *24*, 1596-1599.
Tautz, D., Arctander, P., Minelli, A., Thomas, R.H., and Vogler, A.P. (2002). DNA points the way ahead in taxonomy. Nature *418*, 479-479.
Velle, G., Brooks, S.J., Birks, H.J.B., and Willassen, E. (2005). Chironomids as a tool for inferring Holocene climate: an assessment based on six sites in southern Scandinavia. Quaternary Sci. Rev. *24*, 1429-1462.
Wiemers, M., and Fiedler, K. (2007). Does the DNA barcoding gap exist? a case study in blue butterflies (Lepidoptera: Lycaenidae). Front. Zool. *4*, 8-16.
Wright, J.F. (1984). The chironomid larvae of a small chalk stream in Berkshire, England. Ecol. Entomol. *9*, 231-238.
Yoo, H.S., Eah, J.Y., Kim, J.S., Kim, Y.J., Min, M.S., Paek, W.K., Lee, H., and Kim, C.B. (2006). DNA barcoding Korean birds. Mol. Cells *22*, 323-327.
Zeale, M.R.K., Butlin, R.K., Barker, G.L.A., Lees, D.C., and Jones, G. (2010). Taxon-specific PCR for DNA barcoding arthropod prey in bat faeces. Mol. Ecol. Notes *11*, 236-244.
Zhou, X., Adamowicz, S.J., Jacobus, L.M., DeWalt, R.E., and Hebert, P.D.N. (2009). Towards a comprehensive barcode library for arctic life- Ephemeroptera, Plecoptera, and Trichoptera of Churchill, Manitoba, Canada. Front. Zool. *6*, 30.