DNA barcodes successfully delimit morphospecies in a super-diverse insect genus

Chao Song[1], Xiao-Long Lin[2, *], Qian Wang[3] & Xin-Hua Wang[1]

[1] College of Life Sciences, Nankai University, 300071, Tianjin, China

[2] Department of Natural History, NTNU University Museum, Norwegian University of Science and Technology, NO-7491, Trondheim, Norway

[3] Tianjin key Laboratory of Aqua-Ecology & Aquaculture, Fisheries of College, Tianjin Agricultural University, 300384, Tianjin, China

[*] Corresponding author. E-mail: lin880224@gmail.com

Running title: DNA barcodes delimit morphospecies in an insect genus

Abstract

Polypedilum Kieffer (Diptera: Chironomidae), with 520 currently known species worldwide, can be extremely difficult to identify to species-level based on morphology. We used 3,670 cytochrome *c* oxidase subunit I (COI) barcodes to explore the efficiency of the COI barcodes to differentiate between species in a super-diverse aquatic insect genus. The Barcode of Life Database (BOLD) presented 286 BIN-clusters in *Polypedilum*, representing 163 morphospecies, of which 93 were contributed from our lab. Molecular operational taxonomic units (OTUs) ranged from 158 to 345, based on Automatic Barcode Gap Discovery (ABGD), the Barcode Index Number (BIN), Bayesian Poisson tree processes (bPTP), generalized mixed Yule coalescent (GMYC), j-MOTU, Multi-rate Poisson tree processes (mPTP), neighbor-joining (NJ) tree, and Pre-Threshold Clustering. In comparison, GMYC, bPTP, mPTP and BIN suggested more species than warranted by morphology. While ABGD, j-MOTU, NJ, Pre-Threshold Clustering and ABGD yielded a conservative number of species when setting higher thresholds. Nine species complexes with deep intraspecific divergences indicated 18 potentially cryptic species, which require further taxonomic research including complete life histories as well as nuclear genetic data to be resolved. The discrimination of *Polypedilum* species by DNA barcodes proved successful in 94.4% of all studied morphological species.

Introduction

Since Hebert, Ratnasingham, and deWaard (2003) developed the idea of using a standardized short DNA fragment as a barcode for species identification, large amounts of funds and numerous local or global projects have been carried out. By January 28, 2018, more than 274,998 species (over 5,948,000 records) were registered in the Barcode of Life Data Systems (BOLD). Despite the drawbacks of the methods (Bertheau, Schuler, Krumbock, Arthofer, & Stauffer, 2011; Whitworth, Dawson, Magalon, & Baudry, 2007; Wiemers & Fiedler, 2007), DNA barcoding is now well-established as a tool in taxonomy and ecology and applied in a wide variety of areas, such as food production regulation (Becker, Hanner, & Steinke, 2011), biodiversity conservation (Francis et al., 2010), detection of invasive species and other nature management (Ardura, Linde, Moreira, & Garcia-Vazquez, 2010; Armstrong & Ball, 2005). Due to the revolution in sequencing technology going from Sanger sequencing to high-throughput platforms, combined with new bioinformatic pipelines and computational infrastructure, biodiversity studies are transitioning from the barcoding of individuals to the metabarcoding of communities (Cristescu, 2014; Lobo, Shokralla, Costa, Hajibabaei, & Costa, 2017; Stoeck, Kochems, Forster, Lejzerowicz, & Pawlowski, 2018). With rapid population of next generation sequencing, metadata flooded into public database. Even though, metabarcoding is restricted to a comprehend taxonomic reference libraries based on high quality identifications (Taberlet, Coissac, Pompanon, Brochmann, & Willerslev, 2012).

Different algorithms may produce different operational taxonomic units (OTUs) and suggest different species boundaries. The choice of analytical method for the analysis of DNA barcode data can be of great importance in biodiversity assessments. Among a range of methods, distance-based and tree-based approaches are commonly used in barcoding studies (Austerlitz et al., 2009; Birch, Walsh, Cantrill, Holmes, & Murphy, 2017; Chu, Tong, & Chan, 1999). The instance-based methods usually require a threshold similarity value to distinguish intraspecific and interspecific variations (Meier, Shiyang, Vaidya, & Ng, 2006). For example, Automatic Barcode

Gap Discovery (ABGD) detects a gap in the distribution of divergences that corresponds to differences between intraspecific and interspecific distances. When the gap is absent, the method does not work well for species delimitation (Puillandre, Lambert, Brouillet, & Achaz, 2012). The widely used BOLD system, the Barcode Index Number (BIN) system employs varied distance metrics to generate a neighbor-joining (NJ) tree and established as a persistent registry for life OUTs in the Barcode of Life Data System (BOLD, www.bold.system.org)(Ratnasingham & Hebert, 2007, 2013). While the jMOTU was designed to generate OTUs using a range of cutoff values selected by the user (Jones, Ghoorah, & Blaxter, 2011). The Pre-Threshold Clustering method uses the program TaxonDNA (http://taxondna.sf.net/) to test the viability of threshold values for distinguishing intra- from interspecific variability for different threshold set (Meier et al., 2006).

Tree-based methods apply the "phylogenetic species concept", which defines a species as the smallest resolvable separately evolving lineage or the smallest diagnosable cluster (Baum & Donoghue, 1995; Eldredge & Cracraft, 1980). The methods consider the phylogenetic signal of the sequences and attain higher classification accuracy. However, neighbor-joining tree is used to construct phylogenetic trees based on evolutionary distance data (Saitou & Nei, 1987). The Generalized Mixed Yule Coalescent (GMYC) delimits distinct genetic clusters by optimizing the set of nodes that define the transitions between inter-and intra-specific processes (Fujisawa & Barraclough, 2013; Pons et al., 2006). The Poison Tree Process model (PTP) consider the number of substitutions between branching and speciation that are independent events (Zhang, Kapli, Pavlidis, & Stamatakis, 2013). While the Multi-rate Poisson tree processes (mPTP) incorporates different levels of intraspecific genetic diversity deriving from differences in either the evolutionary history or sampling of each species (Kapli et al., 2017).

The insect family Chironomidae, commonly known as non-biting midges, is the most diverse and abundant invertebrate group in freshwater ecosystems (Milošević, Simić, Stojković, & Živić, 2012). Their wide range of habitat and environmental preferences make them good indicators of aquatic ecosystem change and water

quality (Nicacio & Juen, 2015; Sæther, 2000). Subfossils of chironomid larval head capsules deposited in lake sediments are well-established as useful for quantitative estimates of past temperature changes (Eggermont & Heiri, 2012). However, species-level identification of chironomids based on morphology is both taxonomically challenging and time consuming (Nzelu et al., 2015) and chironomids are therefore often excluded from biodiversity assessments and monitoring.

Polypedilum Kieffer, 1912 is the largest genus of Chironomidae, containing more than 520 described species worldwide (P. Ashe pers. comm. 2017). Immatures of *Polypedilum* can occur in both standing and flowing waters, even in those at high latitude and altitude (Zhang, Song, Qi, & Wang, 2016). Moreover, three *Polypedilum* species, including the famous sleeping midge *P. vanderplanki*, are known for their capability of anhydrobiosis and tolerance of complete dehydration (Cornette et al., 2017; Cranston, 2014; Hinton, 1960). *Polypedilum* may also be among the most abundant invertebrates in eutrophic ponds, reaching densities of up to 1,200 larvae per square meter (Int Panis, Bervoets, & Verheyen, 1995).

However, species delimitation of *Polypedilum* based on morphology is a great challenge. Firstly, most species are described based on adult males only and immature stages are yet to be associated (Song, Wang, Zhang, Sun, & Wang, 2016; Yan et al., 2017; Zhang et al., 2016). Secondly, great phenotypic variation (e.g. wing spots location and numbers), and potential cryptic species complexes are common in this genus (Cranston, Martin, & Spies, 2016). Thus, DNA barcode data should be of great value both in identification of species and analyses of species boundaries (Lin, Stur, & Ekrem, 2017; Yang et al., 2012). However, few barcode studies have included many *Polypedilum* species and before our study, only 54 species were recorded in the BOLD.

The objectives of this study are to contribute the COI DNA barcode reference library of nonbiting midges, assess cryptic species diversity and explore species boundaries of *Polypedilum* non-biting midges.

Materials and Methods

Taxon sampling and data collection

The majority of specimens were collected from China over the last decade, but some also originate from Europe (e.g. Czech Republic, Germany and Norway). Specimens were identified by main taxonomic revisions and species descriptions (e.g. Adeoye & Sæther, 2008; Lin, Qi, Zhang, & Wang, 2013; Oyewo & Sæther, 1998; Sæther, 2000, 2001; Sæther, Andersen, Pinho, & Mendes, 2010; Sæther & Sundal, 1998; Townes, 1945; Vårdal, Bjorlo, & Sæther, 2002; Yamamoto & Yamamoto, 2015; Yan et al., 2017; Zhang et al., 2016; Zhang & Wang, 2005). In addition to our own data, *Polypedilum* COI barcodes, longer than 500 base pairs and without stop codons, were searched and added to the dataset named "*Polypedilum* DNA barcodes (DS-POLYCOI)" on January 13, 2017 in BOLD. DOI: dx.doi.org/10.5883/BOLD:AAW3949. In total, 3,670 COI barcodes (File S1–2) were included, of which 347 barcodes of 93 identified species were from our lab at the College of Life Sciences, Nankai University, Tianjin, China; 266 barcodes were from the GenBank, and the remaining 3,057 barcodes were available from various projects in BOLD.

DNA extraction, PCR amplification, sequencing and alignment

All sampled adults were preserved in 75–85% ethanol, larvae in 100% ethanol, and stored at 4°C in the dark prior to the extraction. The targeted taxa were sorted and dissected under a stereo microscope. Thorax and one pair of legs were used for genomic DNA extraction. All extraction procedure followed the QIAGEN DNA Blood and Tissue kit protocol provided by the manufacturer. Chinese voucher specimens are deposited in the College of Life Sciences, Nankai University, Tianjin, China.

The standard 658 bp mitochondrial COI barcode region was amplified using the universal primers LCO1490 and HCO2198 (Folmer, Black, Hoeh, Lutz, & Vrijenhoek, 1994). Polymerase Chain Reaction (PCR) amplifications were done in a 25 µl volume including 12.5 µl 2×Es Taq MasterMix (CoWin Biotech Co., Beijing,

China), 0.625 µl of each primer, 2 µl of template DNA and 9.25 µl deionized H_2O. Alternatively, 50 µl volume containing 5 µl DNA template, 5 µl 1× PCR Buffer (containing $MgCl_2$), 1 µl 0.2 mM of dNTPs, 1.25 µl 0.25 µM of each primer, 1.5 units of TaqPlus Polymerase and 36.2 µl of deionized H_2O. PCR was performed on a PowerCylcer Gradient SL (Biometra Gmbh, Göttingen, Germany), with an initial denaturation step of 95°C for 4 min followed by 40 cycles of 94°C for 45 s, 52°C for 45 s, 72°C for 1 min, and one cycle at 72°C for 10 min. PCR products were electrophoresed in 1.0% agarose gel, purified and sequenced in both directions with ABI 3730 or ABI 3730XL capillary sequencers at Shanghai Sangon Biotechnology Co., Ltd., Beijing, China, or Beijing Genomics Institute Co., Ltd, Beijing, China.

Raw sequences were assembled and edited in BioEdit v.7.2.5 (http://bioedit.software.informer.com/). The sequences were aligned using the Muscle algorithm (Edgar, 2004) and checked for stop codons in MEGA v.7 on the amino acids (Kumar, Stecher, & Tamura, 2016). Barcode sequences were uploaded to the BOLD along with images and related information of voucher specimens. Sequence names were edited in Mesquite v.3.2 (Maddison & Maddison, 2017). Haplotype networks for some species complexes were constructed using TCS in POPART (Clement, Posada, & Crandall, 2000; Leigh & Bryant, 2015). The nucleotide compositions and pairwise genetic distances were calculated in MEGA using the K2P model (Kimura, 1980).

Putative species estimation

a) *Distance-based approaches*

A Neighbor joining tree was constructed in MEGA using K2P substitution model, treating gaps / missing data with "pairwise deletion", running 500 bootstrap replicates. Automatic barcode gap discovery analysis (ABGD) was implemented on the website (wwwabi.snv.jussieu.fr/public/abgd/abgdweb.html, Puillandre et al., 2012), using relative gap width (X = 1.0) and intraspecific divergence (*P*) values between 0.005 and 0.100 with the K2P model. All other settings were default. BIN assignments of our registered DNA dataset in BOLD was registered on July 13, 2017.

The analytical package (jMOTU) was used to generate OTUs, with parameters ranging from 1–100, with Low BLAST identify filter value 95%, following the default settings (Jones et al., 2011). Pre-thresholds clustering were run with thresholds set in a range from 0.5–11.5% in TaxonDNA or Speciesdentifier v1.8 (Meier et al., 2006).

b) ***Tree-based approach***

A reduced dataset, containing 1,087 sequences, was generated from manual deletion of the highly similar sequences based on an UPGMA tree. The input ultrametric tree for GMYC was constructed using BEAST v1.8.2 (Drummond, Suchard, Xie, & Rambaut, 2012). Settings were as follows: strict clock, MCMC chain using 100 million generations, TN93 substitution model, Yule speciation model. Other parameters available from the authors. The MCMC log on posterior values were examined in Tracer 1.6 (Drummond et al., 2012; Rambaut, Suchard, Xie, & Drummond, 2014) and a burn-in with 30% was set to get an optimal consensus tree. ST-GMYC was applied using the *splits* package in with the guides available on Tomochika's webpage (https://tmfujis.wordpress.com/2013/04/23/how-to-run-gmyc/). The PTP analyses used a rooted phylogenetic input tree constructed with raxmlGUI v1.3 using 500 non-parametric replicates, and the GTR + G + I nucleotide substitutions model (Silvestro & Michalak, 2012). The Bayesian Poisson tree processes (bPTP) analyses were run on the web server (http://species.h-its.org/ptp) with 500,000 MCMC generations, a burn-in of 0.1 and other parameters as default. mPTP analyses were implemented on the web server (http://mptp.h-its.org) using the multi rate Poison tree process model, and followed default settings.

Results

The complete dataset consisted of 3,670 barcodes, ranging from 596 to 658 bp in length (Fig. 1A). In total, there were 371 variable sites (56.4%), of which 322 (86.8%) were parsimony informative. The sequences were heavily AT-biased (66.9%), especially in the third position where 88.8% were A or T (Table 1).

Table 1. Variable and informative sites and average nucleotide composition in the 3,670 COI barcode sequences.

Nucleotide Position	Variable sites (%)	Informative sites (%)	Thymine (%)	Cytosine (%)	Adenine (%)	Guanine (%)
1st	26.5	24.1	26.0	16.9	29.2	27.9
2nd	12.6	7.1	43.7	26.8	12.8	16.7
3rd	60.9	68.8	47.6	8.4	41.2	2.7
Total	56.4	48.5	39.1	17.4	27.8	15.8

There were 50 single lineages, where morphospecies were represented by one sequence. The average intraspecific pairwise distance was 1.68% and the maximum intraspecific divergence was up to 18.3% found in *Polypedilum cultellatum* (Goetghebuer), even higher than the average interspecific divergence of our dataset (15.8%). Some similar cases of deep intraspecific divergence were found in *P. convexum* Johannsen (ranging from 0–16.0%), *P. convictum* (Walker) (ranging from 0–17.6%) and *P. unifascium* (Tokunaga) (ranging from 0–13.9%). In total, there are 40 morphospecies (24.6% of the studied species) with average intraspecific divergence of 2–3%, 25 species (15.4%) with 3–4%, 16 species (9.8%) with 4–5%, 11 species (6.8%) with 5–8% and two species higher than 8%. Disregarding these potential cryptic species complexes, the maximum intraspecific divergence was up to 10.8% found in *P. masudai* (Tokunaga). The maximum interspecific genetic divergence (25.1%) was observed between *P.* sp. 3BD and *P. yongsanensis* Ree et Kim, while the minimum interspecific divergence was low to 0.5%, between *P. quadriguttatum* Kieffer and *P. simulans* Townes (vouchers not examined). Not considering the above case, the minimum divergence between examined species of *Polypedilum* was 5.9% between *P. fanjingensis* Zhang & Wang and *P. kasumiense* Sasa.

Putative species estimation

Based on the NJ tree (File S3), 3,670 DNA barcodes of 162 morphospecies clustered into 180 clades. Except for a few morphospecies [*P. convictum*, *P. cultellatum*, *P. japonicum*, *P. unifascium* and *P. scalaenum* (Schrank)], most of the candidate species formed monophyletic clusters. Nevertheless, distant geographic populations of a few species grouped into nested clades with deep intraspecific divergences, such as *P. pullum* (Zetterstedt) and *P. tsukubaense* (Sasa).

The reduced dataset, including 1,087 COI sequences, yielded 158–286 putative species. The ABGD analysis, using a 4–7% maximum intraspecific divergence, yielded 158–170 OTUs (Fig. 2), which was close to the numbers of the *a priori* identified morphospecies (162). Applying the Pre-threshold clustering method with hierarchical pre-thresholds from 1% to 10%, gave 84–350 OTUs (Fig. 3). Setting higher initial threshold values from 5% to 8%, gave 162–192 OTUs. The analytical package jMOTU gave 162–193 OTUs (Fig. 4) when using single clustering and selected cutoff ranging from 40 (6%) to 53 (8%).

In BOLD, 3,659 of 3,670 barcodes were assigned to 286 BINs, of which 92 BINs with one record, of 59 BINs with two records. Generally, BIN-discordance with morphological identifications were found in 37 species. For instance, 35 morphospecies were assigned two or more BINs: *P. convexum*, *P. masudai*, and *P. unifascium* have seven BINs; *P. cultellatum* and *P. leei* have five BINs; and *P. convictum* and *P. kyotoense* have four BINs. One case of BIN sharing, with a single BIN comprising several different names was found for BOLD:ACS6046 (*P.* sp. S1A and *P.* sp. S1B).

The single-threshold General Mixed Yule-coalescent calculations (ST-GMYC) yielded 242 entities with a confidence interval ranging from 230–249 (Fig. 5). The mPTP model produced a more conservative number of clusters (198) (File S4) compared with the bPTP method, which gave 288–345 clusters (File S5). Depending on the applied method, the numbers of significantly different putative species ranged from 158 to 348. Arranging the used methods by increasing conservativeness give the following bPTP < BINs < GMYC < mPTP < Pre-threshold

Clustering < jMOTU< ABGD. Integrating the results of all methods with morphological boundaries indicates that an average threshold of 5–8% is appropriate to delineate *Polypedilum* species using COI DNA barcodes.

Discussion

Distance methods vs Phylogeny methods

In the present study, based on a dataset of 162 *Polypedilum* morphospecies, phylogeny-based methods produced 25–78% more than distance-based methods when setting a higher threshold. Methods based on genetic distance are more sensitive to a similarity threshold that indicate a gap that between the lowest interspecific distance and highest intraspecific distance. For example, BINs in BOLD have been proven a very good reflection of traditional taxonomy in many animal groups (Young, Behan-Pelletier, & Hebert, 2012). Due to the low intra-cluster distance (2.2%) at the Initial Clustering step of RESL methodology (Ratnasingham & Hebert, 2013), when applying to the genus of *Polypedilum*, BINs system showed over-splitted species numbers in comparison with the current taxonomy. Principally, when the lowest interspecific distance (Fig. 1B) exceeds the highest intraspecific distance (Fig. 1C), the barcode exists (Figs 1D; 6) (Meier, Zhang, & Ali, 2008). In such a case, value of the "threshold" will be well defined, which will facilitate to discovery of cryptic species or new species.

Owing to different species with different population size and divergence time, a universal threshold that fits all taxa does not exist (Yang & Rannala, 2017). Most DNA barcoding studies try to define such a fixed "threshold" value, nevertheless the threshold value is somewhat subjective and arbitrary (Yang & Rannala, 2017). For example, Hebert, Stoeckle, Zemlak, and Francis (2004) proposed the interspecific divergences at least 10 times as larger as the intraspecific divergence known as the "10×rule", Rossini et al. (2016) with 2%, (Smith, Fisher, & Hebert, 2005) with 3% and Dowton, Meiklejohn, Cameron, and Wallman (2014) with 4%. Apparently, the "threshold" are exclusively implied to different taxonomic groups (Havermans, Nagy,

Sonet, De Broyer, & Martin, 2011). For example, a threshold 2–3% was suggested for some groups of Hymenoptera, Ephemeroptera, Plecoptera and Trichoptera (Monaghan, Balke, Gregory, & Vogler, 2005; Schmidt, Schmid-Egger, Moriniere, Haszprunar, & Hebert, 2015; Webb et al., 2012; Zhou, Jacobus, DeWalt, Adamowicz, & Hebert, 2010), 3–5% for some Diptera species (Lin, Stur, & Ekrem, 2015; Nzelu et al., 2015), and 6–8% for the family Hydropsychidae of Trichoptera (Pauls, Blahnik, Zhou, Wardwell, & Holzenthal, 2010). As for dipteral genus *Polypedilum*, the threshold values 5–8% were set so that we could get a high confidence interval of putative species.

The second problem of distance-based methods is ignoring evolutionary relationships (Kapli et al., 2017). Tree based methods were not affected by such thresholds, because they use phylogenetic relationships to accurate barcode assignment. Nevertheless, the accuracy of phylogeny-based methods strongly depend on the input phylogenetic trees. Take GMYC for example, an ultrametric input tree is needed when delimiting species in phylogeny placements. The ultrametric tree most are constructed by BEAST using Markov Chain Monte Carlo Bayesian sampling methods, which was compute-intensive and potentially error-prone process. Meanwhile very few studies can clearly specify the priors set in the MCMC analysis (Yang & Rannala, 2017). Besides, GMYC and PTP methods are more sensitive to unbalanced sample size of datasets (Zhang et al., 2013). Under-sampling of rare species or over-sampled species with small intraspecific variation will compromise species delimitation (Zhang et al., 2013). In our dataset, 50 singletons (species with only one barcode) accounting for 30.8% of our morphospecies might skew the results (Lim, Balke, & Meier, 2012). To avoid the over-sample species with small intraspecific variations, we manually deleted the highly similar sequences based on an UPGMA tree. Tree-based approaches seemed over-splitted when applied to our dataset. On the other hand, the entities delimited by phylogeny-based methods are putative species, for the phylogenetic trees used in analysis are gene tress not species trees. The phylogenetic trees inferred on single COI gene could also have some problems, such as incomplete lineage sorting, hybridization or in recent speciation

events (Mutanen et al., 2016). Therefore, conflicting results should be indicators for species boundaries, which deserve detailed inspection. Consequently, there is no standard algorithm or input parameters to apply that can best recover actual species boundaries for all organisms (Ratnasingham & Hebert, 2013). Especially when dealing with large-scale dataset, it is necessary to incorporate an integrative approach for delimiting species.

Geographic population diversity vs Cryptic species diversity

Our results demonstrated rich species diversity among the genus of *Polypedilum*, for the estimated numbers of OTUs much more than the pre-identified morphospecies. Based on molecular phylogenetic analysis, species-level para- and polyphyly in DNA barcode bene trees as a result of deep intra-specific divergence, which may indicate the presence of cryptic species. For instance, *P. cultellatum* grouped into five distinct clades in the phylogenetic trees (Fig. 7B) and divided into five BINs in BOLD. Among the five distinct clusters, there are three from China, one from Czech Republic and one from Japan and South Korea. To further explore its phylogeographic relationships, the haplotype network analysis was well shown (Fig. 7A), and different populations corresponded with respective clades in NJ tree. Besides, morphology differences were not observed between the four clades (clade from Japan and South Korea not available). Consequently, DNA barcodes probably revealed four potential cryptic species within *P. cultellatum* species complex. Several similar cases were illustrated as well. Such as, the species *P. convictum* formed eight BINs in BOLD and seven clades in NJ tree. Among the seven clades, there were five from North America (America + Canada), one from Asia (China), and one clade from Europe (Germany). *P. convexum* formed seven BINs and four clades in NJ tree, two clades from China, one clade from Japan and one from Australia. *P. unifascium* (nine BINs and five Clades), of two clades from China, one clade from Japan, and one from China + Korea. According to the shifting balance theory, many species comprise small, partially isolated populations (Slatkin, 1987). Therefore, geographical populations with deep intraspecific divergence, are more likely to be reveal cryptic species, which indicate18 potential cryptic species within *Polypedilum*. On one hand, traditional

species classification relies on qualitative or quantitative morphological differences, which could lead to confusion between sibling taxa or cryptic species and incorrect species ascription (Tamar et al., 2015). DNA-based methods facilitate to differentiate closely related species, or reveal the presence of distinct taxa that morphologically indistinguishable. Consequently, increasing the samples of the geographic populations of the studied species is expected to disclose more potential cryptic species. On the other hand, COI based DNA barcodes have proven successful in delimiting *Polypedilum* species, while some morphospecies recognized as paraphyletic corresponding to gene trees. Therefore, multiple evidences, with nuclear genes and thorough morphological data could provide a better resolution to further illuminate intra- and interspecific relationships of *Polypedilum* species.

Nevertheless, it is necessary to designate specific OTU that bears the species name. The first option is to barcode the holotype or paratype specimens, which is not feasible for chironomids for all the type specimens are slide mounted. An alternative solution is to barcode the fresh specimens sampled from the type locality. Nevertheless, there are shortages for the solutions: (1) species distributions may change (2) several OTUs could occur in one collection site (Porco et al., 2012). It is urgent to validate these cryptic specie with additional nuclear markers, as well as morphological and ecological data. Besides, voucher specimen's information from public database such as GenBank or BOLD were not always correct. The corresponding strategy are re-examination of voucher specimens for further validation, especially for some subtle morphological variations (e.g. body colorations, wing setations and spotting patterns).

New species vs Taxonomic synonyms

Recent taxonomic studies especially referring to new species would provide DNA barcode as new "characters" for species description and delimitation of chironomids (Lin et al., 2017; Yan et al., 2017). Molecular data provide a more objective view when defining "species", such as scale of intra- or interspecific distances. Comparatively, morphological variations or phenotypic changes are usually difficult to define as intra- or interspecific boundaries. Song et al. (2016) discussed that

whether some characters such as wing spotting patterns and adult male genitalia (anal point projection length variations) be diagnosis to separate species.

In this study, some morphological traits were discussed with DNA barcodes. For instance, Ree and Kim (1981) described *Polypedilum yongsanensis* as a new species for their highly similarities. While Orel, Kang, and Makarchenko (2017) and Sasa & Kikuchi (1995) synonymized *P. yongsanensis* with *P. nubeculosum* for their highly similarities. However, the minimum interspecific genetic COI divergence between these two populations is up to 16.1%, suggesting them as different species. Moreover, we found that *P. nubeculosum* distributed in Europe and Inner Mongolia region of China, while *P. yongsanensis* distributed in Oriental China, South Korea and Japan.

In contrast to conspecific species with deep genetic divergence, there are certain species complexes with low interspecific divergence and formed same clades in the phylogenetic trees. Under this situation, we should examine the morphological characters of voucher specimens, and then reconsider whether some characters as diagnosis to delimit species.

Therefore, we carefully discussed all the anomalous cases in our studies. (1) *Polypedilum griseoguttatum* Kieffer (vouchers not available) vs *P. masuda*i (Fig. 8). Vårdal et al. (2002) originally described *P. griseoguttatum* and showed high similarity *P. masudai*, except the absence of short anal point projections, which is the key characters in the *Polypedilum* species taxonomy. Therefore, the record *P. griseoguttatum* (KJ530963) probably be a misidentification. (2) *Polypedilum simulans* vs *P. quadriguttatum* vs *P.* sp.18SC (Fig. 9). These three morphospecies are closely related to each other by having highly similar genitalia in adult males, but could be distinguished by having different wing patterns and leg colors [e.g. *P.* sp.18SC (wing with spot, legs yellowish), *P. quadriguttatum* (wing with spots, partial legs brown or yellowish) and *P. simulans* (wing without spots, legs yellowish)]. We have checked the specimens of *P. simulans* from Europe, yet specimen of *P. quadriguttatum* was not accessible (await to be further tested). According to Song et al. (2016), *P.* sp.18SC could be different species from *P. simulans* for having wing spots. However, these two species obviously formed into a monophyletic group in the NJ tree, which

conflicts Song et al. (2016). (3) *P*. sp. S1A vs *P*. sp. S1B vs *P*. sp. S1C vs *P*. sp. S1H (Fig. 10). These four named species apparently clustered in a group on the NJ tree with low 3.8–6.2% COI divergence, which could be generally dealt as four species in one. Carew & Hoffmann (2017) suggested that these groups should be distinct taxa based on nuclear genes and morphological data. (4) *P*. *scalaenum* vs *P*. *unifascium* (Fig. 11). Song et al. (2016) discussed this case and concluded that specimens were misidentified as *P*. *scalaenum*. Judging from the above results, DNA barcodes could reveal potential misidentifications or junior synonyms.

Conclusion and Prospects

In general, DNA barcodes can successfully delimit *Polypedilum* non-biting midges with 94.4% match with morphospecies. Comparing the performances of different analytical tools, methods of Neighbor joining tree and ABGD fit well on the *Polypedilum* COI barcodes dataset. Considering the distance-based approaches, 5–8% threshold on average is tentatively suggested for species delimitation of *Polypedilum* non-biting midges. Unusual deep intraspecific divergences in some species complexes were detected, indicating potential morphological misidentifications or cryptic species. It is urgent to corporate additional nuclear genes, more morphological traits from different life stages, and ecological data for further research.

Acknowledgments

We are grateful to Dr. Torbjørn Ekrem (NTNU University Museum, Trondheim, Norway) for his valuable comments and effort to improve the manuscript. We are also grateful to Dr. Hong-Qu Tang, (Jinan University, Guangzhou, China), Dr. Wen-Bin Liu (Tianjin Normal University, Tianjin, China) and Bing-Jiao Sun (Nankai University, Tianjin, China) for their selfless assistance in samples collection, and Dr. Elisabeth Stur (NTNU University Museum, Trondheim, Norway) for her loaning specimens. We would like to thank China Scholarship Council (CSC) and National Natural Science Foundation of China (NSFC: 31272284, 31301908, 31460572,

31672324 & 31672264).

References

Adeoye, E., & Sæther, O. A. (2008). Revision of *Polypedilum* (*Pentapedilum*) Kieffer and *Ainuyusurika* Sasa et Shirasaki (Diptera: Chironomidae). *Zootaxa, 1953*, 1-145.

Ardura, A., Linde, A. R., Moreira, J. C., & Garcia-Vazquez, E. (2010). DNA barcoding for conservation and management of Amazonian commercial fish. *Biological Conservation, 143*(6), 1438-1443. http://doi:10.1016/j.biocon.2010.03.019

Armstrong, K. F., & Ball, S. L. (2005). DNA barcodes for biosecurity: invasive species identification. *Philosophical Transactions of the Royal Society B-Biological Sciences, 360*(1462), 1813-1823. http://doi:10.1098/rstb.2005.1713

Austerlitz, F., David, O., Schaeffer, B., Bleakley, K., Olteanu, M., Leblois, R., . . . Laredo, C. (2009). DNA barcode analysis: a comparison of phylogenetic and statistical classification methods. *BMC Bioinformatics, 10*. http://doi:10.1186/1471-2105-10-S14-S10

Baum, D. A., & Donoghue, M. J. (1995). Choosing among Alternative Phylogenetic Species Concepts. *Systematic Botany, 20*(4), 560-573. http://doi:10.2307/2419810

Becker, S., Hanner, R., & Steinke, D. (2011). Five years of FISH-BOL: brief status report. *Mitochondrial DNA, 22(S1)*, 3-9. http://doi:10.3109/19401736.2010.535528

Bertheau, C., Schuler, H., Krumbock, S., Arthofer, W., & Stauffer, C. (2011). Hit or miss in phylogeogSyst Biolraphic analyses: the case of the cryptic NUMTs. *Molecular Ecology Resources, 11*(6), 1056-1059. http://doi:10.1111/j.1755-0998.2011.03050.x

Birch, J. L., Walsh, N. G., Cantrill, D. J., Holmes, G. D., & Murphy, D. J. (2017). Testing efficacy of distance and tree-based methods for DNA barcoding of grasses (Poaceae tribe Poeae) in Australia. *PLoS One, 12*(10), e0186259. http://doi:10.1371/journal.pone.0186259

Chu, K. H., Tong, J. G., & Chan, T. Y. (1999). Mitochondrial cytochrome oxidase I sequence divergence in some Chinese species of *Charybdis* (Crustacea : Decapoda : Portunidae). *Biochemical Systematics and Ecology, 27*(5), 461-468.

Clement, M., Posada, D., & Crandall, K. A. (2000). TCS: a computer program to estimate gene genealogies. *Molecular Ecology, 9*(10), 1657-1659. http://doi:10.1046/j.1365-294x.2000.01020.x

Cornette, R., Yamamoto, N., Yamamoto, M., Kobayashi, T., Petrova, N. A., Gusev, O., . . . Okuda, T. (2017). A new anhydrobiotic midge from Malawi, *Polypedilum pembai* sp n. (Diptera: Chironomidae), closely related to the desiccation tolerant midge, *Polypedilum vanderplanki* Hinton. *Systematic*

Entomology, 42(4), 814-825. http://doi:10.1111/syen.12248
Cranston, P. S. (2014). A new putatively cryptobiotic midge, Polypedilum ovahimba sp nov (Diptera: Chironomidae), from southern Africa. *Austral Entomology, 53*(4), 373-379. http://doi:10.1111/aen.12090
Cranston, P. S., Martin, J., & Spies, M. (2016). Cryptic species in the nuisance midge *Polypedilum nubifer* (Skuse) (Diptera: Chironomidae) and the status of *Tripedilum* Kieffer. *Zootaxa, 4079*(4), 429-447. http://doi:10.11646/zootaxa.4079.4.3.
Cristescu, M. E. (2014). From barcoding single individuals to metabarcoding biological communities: towards an integrative approach to the study of global biodiversity. *Trends in Ecology & Evolution, 29*(10), 566-571. http://doi:10.1016/j.tree.2014.08.001
Dowton, M., Meiklejohn, K., Cameron, S. L., & Wallman, J. (2014). A Preliminary Framework for DNA Barcoding, Incorporating the Multispecies Coalescent. *Systematic Biology, 63*(4), 639-644. http://doi:10.1093/sysbio/syu028
Drummond, A. J., Suchard, M. A., Xie, D., & Rambaut, A. (2012). Bayesian Phylogenetics with BEAUti and the BEAST 1.7. *Molecular Biology and Evolution, 29*(8), 1969-1973. http://doi:10.1093/molbev/mss075
Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research, 32*(5), 1792-1797. http://doi:10.1093/nar/gkh340
Eggermont, H., & Heiri, O. (2012). The chironomid-temperature relationship: expression in nature and palaeoenvironmental implications. *Biological reviews of the Cambridge Philosophical Society, 87*(2), 430-456. http://doi:10.1111/j.1469-185X.2011.00206.x
Eldredge, N., & Cracraft, J. (1980). *Phylogenetic Patterns and the Evolutionary Process: Method and Theory in Comparative Biology*. New York: Columbia University Press.
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.
Francis, C. M., Borisenko, A. V., Ivanova, N. V., Eger, J. L., Lim, B. K., Guillen-Servent, A., . . . Hebert, P. D. (2010). The role of DNA barcodes in understanding and conservation of mammal diversity in southeast Asia. *PLoS One, 5*(9), e12575. http://doi:10.1371/journal.pone.0012575
Fujisawa, T., & Barraclough, T. G. (2013). Delimiting species using single-locus data and the Generalized Mixed Yule Coalescent approach: a revised method and evaluation on simulated data sets. *Systematic Biology, 62*(5), 707-724. http://doi:10.1093/sysbio/syt033
Havermans, C., Nagy, Z. T., Sonet, G., De Broyer, C., & Martin, P. (2011). DNA barcoding reveals new insights into the diversity of Antarctic species of *Orchomene* sensu lato (Crustacea: Amphipoda: Lysianassoidea). *Deep-Sea Research Part Ii-Topical Studies in Oceanography, 58*(1-2), 230-241.

http://doi:10.1016/j.dsr2.2010.09.028
Hebert, P. D. N., Ratnasingham, S., & deWaard, J. R. (2003). Barcoding animal life: cytochrome c oxidase subunit 1 divergences among closely related species. *Proceedings of the Royal Society B-Biological Sciences, 270*, S96-99. http://doi:10.1098/rsbl.2003.0025
Hebert, P. D. N., Stoeckle, M. Y., Zemlak, T. S., & Francis, C. M. (2004). Identification of birds through DNA barcodes. *Plos Biology, 2*(10), 1657-1663. http://doi:10.1371/journal.pbio.0020312
Hinton, H. E. (1960). Cryptobiosis in the larva of *Polypedilum vanderplanki* Hint. (Chironomidae). *Journal of Insect Physiology, 5*(3-4), 286-300. http://doi:10.1016/0022-1910(60)90011-1
Int Panis, L., Bervoets, L., & Verheyen, R. (1995). The spatial distribution of *Caenis horaria* (L., 1758) (Caenidae, Ephemeroptera) in a pond in Niel (Belgium). *Bulletin et annales de la Société entomologique de Belgique, 131*, 47-51.
Jones, M., Ghoorah, A., & Blaxter, M. (2011). jMOTU and Taxonerator: Turning DNA Barcode Sequences into Annotated Operational Taxonomic Units. *PLoS One, 6*(4), e19259. http://doi:10.1371/journal.pone.0019259.
Kapli, P., Lutteropp, S., Zhang, J., Kobert, K., Pavlidis, P., Stamatakis, A., & Flouri, T. (2017). Multi-rate Poisson tree processes for single-locus species delimitation under maximum likelihood and Markov chain Monte Carlo. *Bioinformatics, 33*(11), 1630-1638. http://doi:10.1093/bioinformatics/btx025
Kimura, M. (1980). A Simple Method for Estimating Evolutionary Rates of Base Substitutions through Comparative Studies of Nucleotide-Sequences. *Journal of Molecular Evolution, 16*(2), 111-120. http://doi:10.1007/Bf01731581
Kumar, S., Stecher, G., & Tamura, K. (2016). MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. *Molecular Biology and Evolution, 33*(7), 1870-1874. http://doi:10.1093/molbev/msw054
Leigh, J. W., & Bryant, D. (2015). POPART: full-feature software for haplotype network construction. *Methods in Ecology and Evolution, 6*(9), 1110-1116. http://doi:10.1111/2041-210x.12410
Lim, G. S., Balke, M., & Meier, R. (2012). Determining Species Boundaries in a World Full of Rarity: Singletons, Species Delimitation Methods. *Systematic Biology, 61*(1), 165-169. http://doi:10.1093/sysbio/syr030
Lin, X. L., Qi, X., Zhang, R. L., & Wang, X. H. (2013). A new species of *Polypedilum* (*Uresipedilum*) Oyewo & Sæther, 1998 from Zhejiang Province of Oriental China (Diptera, Chironomidae). *Zookeys*(320), 43-49. http://doi:10.3897/zookeys.320.5147
Lin, X. L., Stur, E., & Ekrem, T. (2015). Exploring Genetic Divergence in a Species-Rich Insect Genus Using 2790 DNA Barcodes. *PLoS One, 10*(9), e0138993. http://doi:10.1371/journal.pone.0138993
Lin, X. L., Stur, E., & Ekrem, T. (2017). DNA barcodes and morphology reveal unrecognized species in Chironomidae (Diptera). *Insect Systematics & Evolution*. http://doi:10.1163/1876312X-00002172
Lobo, J., Shokralla, S., Costa, M. H., Hajibabaei, M., & Costa, F. O. (2017). DNA

metabarcoding for high-throughput monitoring of estuarine macrobenthic communities. *Scientific Reports, 7*(1), 15618. http://doi:10.1038/s41598-017-15823-6

Maddison, W. P., & Maddison, D. R. (2017). Mesquite: a modular system for evolutionary analysis. *Version 3.31 http://mesquiteproject.org*.

Meier, R., Shiyang, K., Vaidya, G., & Ng, P. K. L. (2006). DNA barcoding and taxonomy in diptera: A tale of high intraspecific variability and low identification success. *Systematic Biology, 55*(5), 715-728. http://doi:10.1080/10635150600969864

Meier, R., Zhang, G. Y., & Ali, F. (2008). The Use of Mean Instead of Smallest Interspecific Distances Exaggerates the Size of the "Barcoding Gap" and Leads to Misidentification. *Systematic Biology, 57*(5), 809-813. http://doi:10.1080/10635150802406343

Milošević, D., Simić, V., Stojković, M., & Živić, I. (2012). Chironomid faunal composition represented by taxonomic distinctness index reveals environmental change in a lotic system over three decades. *Hydrobiologia, 683*(1), 69-82. http://doi:10.1007/s10750-011-0941-8

Monaghan, M. T., Balke, M., Gregory, T. R., & Vogler, A. P. (2005). DNA-based species delineation in tropical beetles using mitochondrial and nuclear markers. *Philosophical Transactions of the Royal Society B-Biological Sciences, 360*(1462), 1925-1933. http://doi:10.1098/rstb.2005.1724

Mutanen, M., Kivela, S. M., Vos, R. A., Doorenweerd, C., Ratnasingham, S., Hausmann, A., . . . Godfray, H. C. (2016). Species-Level Para- and Polyphyly in DNA Barcode Gene Trees: Strong Operational Bias in European Lepidoptera. *Systematic Biology, 65*(6), 1024-1040. http://doi:10.1093/sysbio/syw044

Nicacio, G., & Juen, L. (2015). Chironomids as indicators in freshwater ecosystems: an assessment of the literature. *Insect Conservation and Diversity, 8*(5), 393-403. http://doi:10.1111/icad.12123

Nzelu, C. O., Cáceres, A. G., Arrunátegui-Jiménez, M. J., Lañas-Rosas, M. F., Yañez-Trujillano, H. H., Luna-Caipo, D. V., . . . Kato, H. (2015). DNA barcoding for identification of sand fly species (Diptera: Psychodidae) from leishmaniasis-endemic areas of Peru. *Acta Tropica, 145*, 45-51. http://doi:10.1016/j.actatropica.2015.02.003

Oyewo, E. A., & Sæther, O. A. (1998). Revision of Afrotropical *Polypedilum* Kieffer subgen. *Uresipedilum* Sasa et Kikuchi, 1995 (Diptera : Chironomidae), with a review of the subgenus. *Annales De Limnologie-International Journal of Limnology, 34*(3), 315-362. http://doi:10.1051/limn/1998028

Pauls, S. U., Blahnik, R. J., Zhou, X., Wardwell, C. T., & Holzenthal, R. W. (2010). DNA barcode data confirm new species and reveal cryptic diversity in Chilean Smicridea (Smicridea) (Trichoptera:Hydropsychidae). *Journal of the North American Benthological Society, 29*(3), 1058-1074. http://doi:10.1899/09-108.1

Pons, J., Barraclough, T. G., Gomez-Zurita, J., Cardoso, A., Duran, D. P., Hazell,

S., . . . Vogler, A. P. (2006). Sequence-based species delimitation for the DNA taxonomy of undescribed insects. *Systematic Biology, 55*(4), 595-609.

Porco, D., Bedos, A., Greenslade, P., Janion, C., Skarzynski, D., Stevens, M. I., . . . Deharveng, L. (2012). Challenging species delimitation in Collembola: cryptic diversity among common springtails unveiled by DNA barcoding. *Invertebrate Systematics, 26*(5-6), 470-477. http://doi:10.1071/Is12026

Puillandre, N., Lambert, A., Brouillet, S., & Achaz, G. (2012). ABGD, Automatic Barcode Gap Discovery for primary species delimitation. *Molecular Ecology, 21*(8), 1864-1877. http://doi:10.1111/j.1365-294X.2011.05239.x

Rambaut, A., Suchard, M. A., Xie, D., & Drummond, A. J. (2014). Tracerv1.6, Available: http://beast.bio.ed.ac.uk/ Tracer.2014. .

Ratnasingham, S., & Hebert, P. D. N. (2007). BOLD: The Barcode of Life Data System (www.barcodinglife.org). *Molecular Ecology Notes, 7*(3), 355-364. http://doi:10.1111/j.1471-8286.2007.01678.x

Ratnasingham, S., & Hebert, P. D. N. (2013). A DNA-Based Registry for All Animal Species: The Barcode Index Number (BIN) System. *PLoS One, 8*(7), e66213. http://doi:10.1371/journal.pone.0066213

Ree, H. I., & Kim, M. S. (1981). Studies on Chironomidae (Diptera) in Korea. 1.Taxonomical study on adults of Chironomidae. *Proceedings of College of Natural Science, 6*.

Rossini, B. C., Oliveira, C. A. M., de Melo, F. A. G., Bertaco, V. D., de Astarloa, J. M. D., Rosso, J. J., . . . Oliveira, C. (2016). Highlighting *Astyanax* Species Diversity through DNA Barcoding. *PLoS One, 11*(12), e0167203. http://doi:10.1371/journal.pone.0167203

Sæther, O. A. (2000). Zoogeographical patterns in Chironomidae (Diptera). *SIL Proceedings, 1922-2010, 27*(1), 290-302. http://doi:10.1080/03680770.1998.11901242

Sæther, O. A. (2001). Revision of the Nearctic species of the genus *Polypedilum* Kieffer (Diptera: Chironomidae) in the subgenera *P.* (*Polypedilum*) and *P.* (*Uresipedilum*) Oyewo and Sæther. *Journal of the North American Benthological Society, 20*(1), 156-157. http://doi:10.2307/1468198

Sæther, O. A., Andersen, T., Pinho, L. C., & Mendes, H. F. (2010). The problems with *Polypedilum* Kieffer (Diptera: Chironomidae), with the description of *Probolum* subgen. n. *Zootaxa*(2497), 1-36. http://doi:10.5281/zenodo.195747

Sæther, O. A., & Sundal, A. (1998). *Cerobregma*, a new subgenus of *Polypedilum* Kieffer, with a tentative phylogeny of subgenera and species groups within *Polypedilum* (Diptera: Chironomidae). *Journal of the Kansas Entomological Society, 71*(3), 315-382.

Saitou, N., & Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution, 4*(4), 406-425. http://doi:10.1093/oxfordjournals.molbev.a040454

Schmidt, S., Schmid-Egger, C., Moriniere, J., Haszprunar, G., & Hebert, P. D. (2015). DNA barcoding largely supports 250 years of classical taxonomy: identifications for Central European bees (Hymenoptera, Apoidea *partim*).

Molecular Ecology Resources, 15(4), 985-1000. http://doi:10.1111/1755-0998.12363

Silvestro, D., & Michalak, I. (2012). raxmlGUI: a graphical front-end for RAxML. *Organisms Diversity & Evolution, 12*(4), 335-337. http://doi:10.1007/s13127-011-0056-0

Slatkin, M. (1987). Gene Flow and the Geographic Structure of Natural-Populations. *Science, 236*(4803), 787-792.

Smith, M. A., Fisher, B. L., & Hebert, P. D. N. (2005). DNA barcoding for effective biodiversity assessment of a hyperdiverse arthropod group: the ants of Madagascar. *Philosophical Transactions of the Royal Society B-Biological Sciences, 360*(1462), 1825-1834.

Song, C., Wang, Q., Zhang, R. L., Sun, B. J., & Wang, X. H. (2016). Exploring the utility of DNA barcoding in species delimitation of *Tripodura* (*Tripodura*) non-biting midges (Diptera: Chironomidae). *Zootaxa, 4079*(5), 534-550. http://doi:10.11646/zootaxa.4079.5.2.

Stoeck, T., Kochems, R., Forster, D., Lejzerowicz, F., & Pawlowski, J. (2018). Metabarcoding of benthic ciliate communities shows high potential for environmental monitoring in salmon aquaculture. *Ecological Indicators, 85*, 153-164. http://doi:10.1016/j.ecolind.2017.10.041

Taberlet, P., Coissac, E., Pompanon, F., Brochmann, C., & Willerslev, E. (2012). Towards next-generation biodiversity assessment using DNA metabarcoding. *Molecular Ecology, 21*(8), 2045-2050. http://doi:10.1111/j.1365-294X.2012.05470.x

Tamar, K., Carranza, S., den Bosch, H. I., Sindaco, R., Moravec, J., & Meiri, S. (2015). Hidden relationships and genetic diversity: Molecular phylogeny and phylogeography of the Levantine lizards of the genus *Phoenicolacerta* (Squamata: Lacertidae). *Molecular Phylogenetics and Evolution, 91*, 86-97. http://doi:10.1016/j.ympev.2015.05.002

Townes, H. K. (1945). The Nearctic Species of Tendipedini - Diptera, Tendipedidae (= Chironomidae). *American Midland Naturalist, 34*(1), 1-206. http://doi:10.2307/2421112

Vårdal, H., Bjorlo, A., & Sæther, O. A. (2002). Afrotropical *Polypedilum* subgenus *Tripodura*, with a review of the subgenus (Diptera: Chironomidae). *Zoologica Scripta, 31*(4), 331-402. http://doi:10.1046/j.1463-6409.2002.00096.x

Webb, J. M., Jacobus, L. M., Funk, D. H., Zhou, X., Kondratieff, B., Geraci, C. J., . . . Hebert, P. D. N. (2012). A DNA Barcode Library for North American Ephemeroptera: Progress and Prospects. *PLoS One, 7*(5), e38063. http://doi:10.1371/journal.pone.0038063

Whitworth, T. L., Dawson, R. D., Magalon, H., & Baudry, E. (2007). DNA barcoding cannot reliably identify species of the blowfly genus *Protocalliphora* (Diptera: Calliphoridae). *Proceedings of the Royal Society B-Biological Sciences, 274*(1619), 1731-1739. http://doi:10.1098/rspb.2007.0062

Wiemers, M., & Fiedler, K. (2007). Does the DNA barcoding gap exist? - a case study in blue butterflies (Lepidoptera: Lycaenidae). *Frontiers in Zoology, 4*, 8.

http://doi:10.1186/1742-9994-4-8
Yamamoto, N., & Yamamoto, M. (2015). A revised subgeneric position for *Polypedilum* (*Probolum*) *simantokeleum*, with description of a new *Uresipedilum* species in Japan (Diptera: Chironomidae). *Zootaxa, 3999*(3), 439-445. http://doi:10.11646/zootaxa.3999.3.9
Yan, C. C., Song, C., Liu, T., Zhao, G. J., Hou, Z. Y., Cao, W., & Wang, X. H. (2017). Two new and one newly recorded species of *Polypedilum* Kieffer 1912 with DNA barcodes from Oriental China (Chironomidae: Diptera). *Zootaxa, 4238*(1), 109-118. http://doi:10.11646/zootaxa.4238.1.8
Yang, Z., Landry, J.-F., Handfield, L., Zhang, Y., Alma Solis, M., Handfield, D., . . . Hebert, P. D. N. (2012). DNA barcoding and morphology reveal three cryptic species of *Anania* (Lepidoptera: Crambidae: Pyraustinae) in North America, all distinct from their European counterpart. *Systematic Entomology, 37*(4), 686-705. http://doi:10.1111/j.1365-3113.2012.00637.x
Yang, Z. H., & Rannala, B. (2017). Bayesian species identification under the multispecies coalescent provides significant improvements to DNA barcoding analyses. *Molecular Ecology, 26*(11), 3028-3036. http://doi:10.1111/mec.14093
Young, M. R., Behan-Pelletier, V. M., & Hebert, P. D. N. (2012). Revealing the Hyperdiverse Mite Fauna of Subarctic Canada through DNA Barcoding. *PLoS One, 7*(11), e48755. http://doi:10.1371/journal.pone.0048755
Zhang, J., Kapli, P., Pavlidis, P., & Stamatakis, A. (2013). A general species delimitation method with applications to phylogenetic placements. *Bioinformatics, 29*(22), 2869-2876. http://doi:10.1093/bioinformatics/btt499
Zhang, R. L., Song, C., Qi, X., & Wang, X. H. (2016). Taxonomic review on the subgenus *Tripodura* Townes (Diptera: Chironomidae: *Polypedilum*) from China with eleven new species and a supplementary world checklist. *Zootaxa, 4136*(1), 1-53. http://doi:10.11646/zootaxa.4136.1.1
Zhang, R. L., & Wang, X. H. (2005). *Polypedilum* (*Cerobregma*) Sæther & Sundal from China (Diptera: Chironomidae). *Aquatic Insects, 27*(1), 47-55. http://doi:10.1080/01650420400019262
Zhou, X., Jacobus, L. M., DeWalt, R. E., Adamowicz, S. J., & Hebert, P. D. N. (2010). Ephemeroptera, Plecoptera, and Trichoptera fauna of Churchill (Manitoba, Canada): insights into biodiversity patterns from DNA barcoding. *Journal of the North American Benthological Society, 29*(3), 814-837. http://doi:10.1899/09-121.1

Figure legends

Figure 1. Summary of the 3,670 aligned sequences of *Polypedilum*. (A) Sequence lengths; (B) inter-specific distances; (C) intra-specific distances; (D) combined intra- and interspecific distances.

Figure 2. Number of OTUs under ABGD online with the prior intraspecific

divergence based on 3,670 DNA barcodes of *Polypedilum*.

Figure 3. The Number of OTUs based on DNA barcodes of *Polypedilum* using Pre-Threshold Clustering at different thresholds.

Figure 4. Number of OTUs based on DNA barcodes of *Polypedilum* defined by different cutoff values generated from jMOTU.

Figure 5. Results of the species delimitation analysis for the *Polypedilum* according to the GMYC single-threshold model on the DNA barcodes dataset with 1,087 individuals. (A) Lineage-through-time plot based on the ultrametric tree obtained from COI sequences. The sharp increase in branching rate, corresponding to the transition from interspecific to intraspecific branching events, is indicated by a red vertical line. The x-axes (both in panels A and B) show substitutions per nucleotide site; (B) likelihood function produced by GMYC to estimate the peak of transition between cladogenesis (interspecific diversification) and allele intraspecific coalescence along the branches; (C) ultrametric tree with 1,087 individuals obtained in BEAST setting coalescent prior and strict clock model. Red clusters and black lines (singletons) indicate putative species calculated by the model.

Figure 6. Histogram of pairwise K2P distances generated from ABGD online.

Figure 7. *Polypedilum cultellatum* species complex. (A). TCS network based on the mitochondrial cytochrome *c* oxidase subunit I (COI) dataset of the *Polypedilum cultellatum* species complex. Different colors correspond to the different collection sites; (B) neighbor joining tree of the *P. cultellatum* species complex based on K2P distances in DNA barcodes. Numbers on branches represent bootstrap support (>70%) based on 1000 replicates; scale represents K2P genetic distances.

Figure 8. Neighbor joining subtree based on DNA barcodes of *Polypedilum griseoguttatum* and *P. masuda*i. Numbers on branches represent bootstrap support (>70%) based on 500 replicates; scale represents K2P genetic distances.

Figure 9. Neighbor joining subtree based on DNA barcodes of *Polypedilum simulans*, *P. quadriguttatum* and *P.* sp.18SC. Numbers on branches represent bootstrap support (>70%) based on 500 replicates; scale represents K2P genetic distances.

Figure 10. Neighbor joining subtree based on DNA barcodes of *Polypedilum* sp. S1A,

P. sp. S1B, *P*. sp. S1C and *P*. sp. S1H. Numbers on branches represent bootstrap support (>70%) based on 500 replicates; scale represents K2P genetic distances.

Figure 11. Neighbor joining subtree based on DNA barcodes of *Polypedilum scalaenum* and *P*. *unifascium*. Numbers on branches represent bootstrap support (>70%); scale represents K2P genetic distances.

Supplementing information

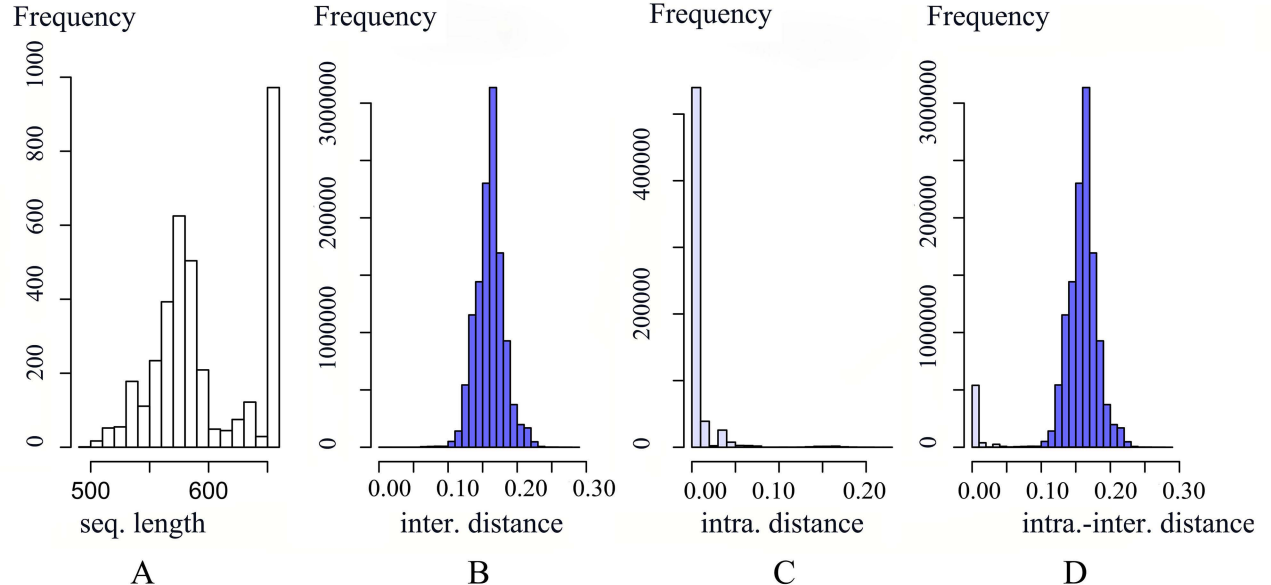
File S1. Dataset of 3,670 *Polypedilum* COI barcode sequences.

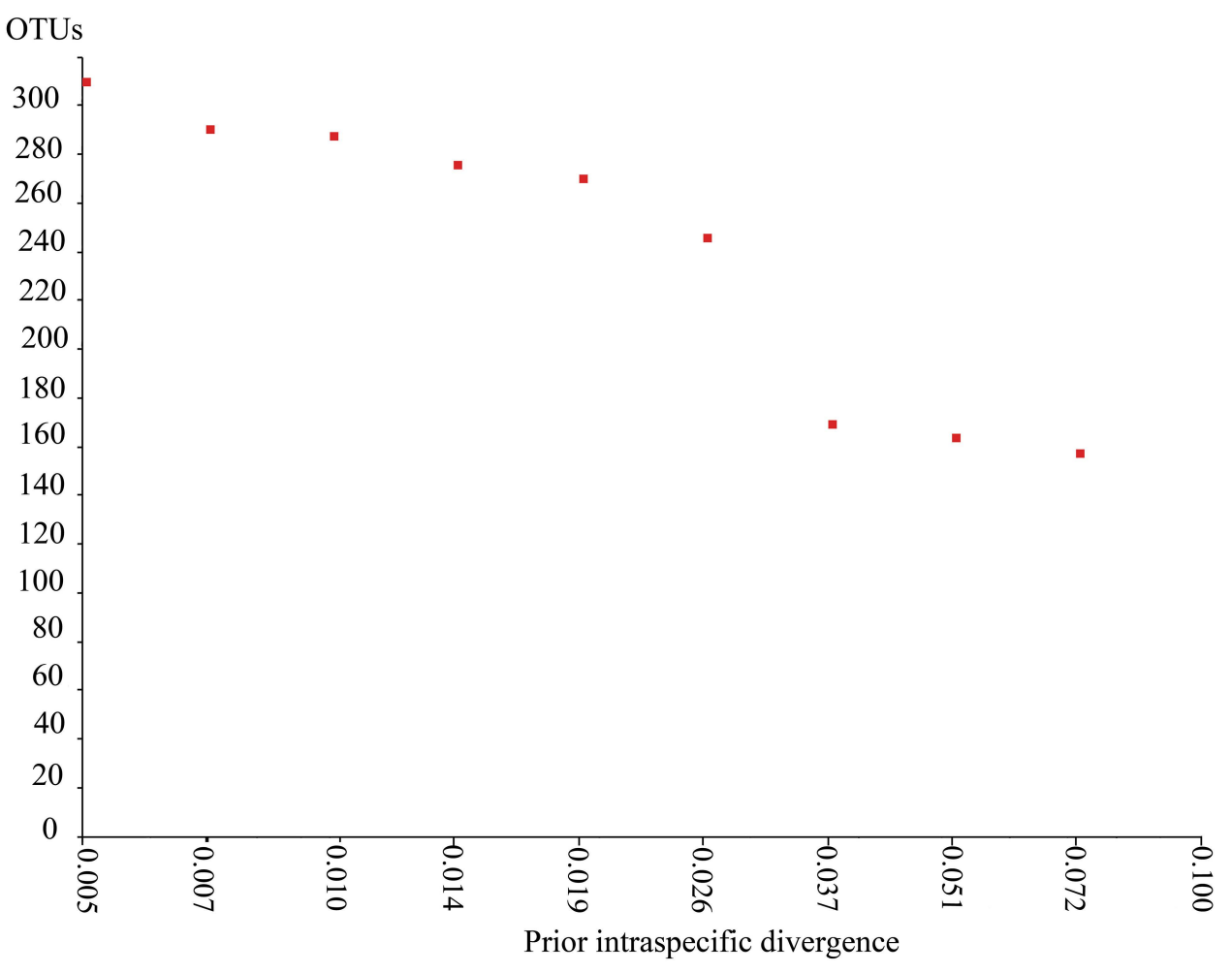
File S2. BOLD sample ID, GenBank Accession Numbers of studied *Polypedilum*.

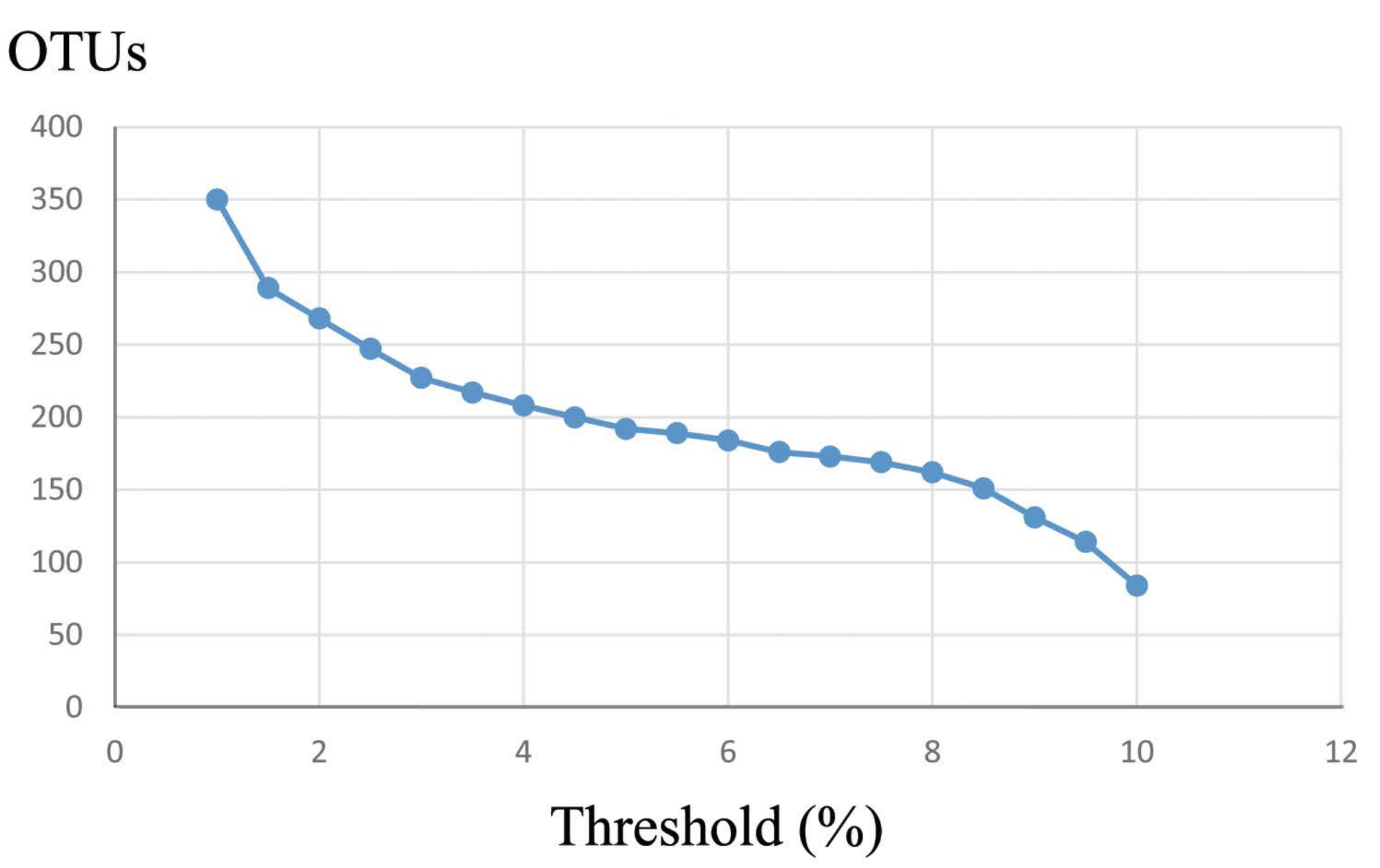
File S3 Neighbor joining bootstrap consensus tree for 3,670 DNA barcodes of *Polypedilum*. Numbers on branches are bootstrap support using 500 bootstrap replicates.

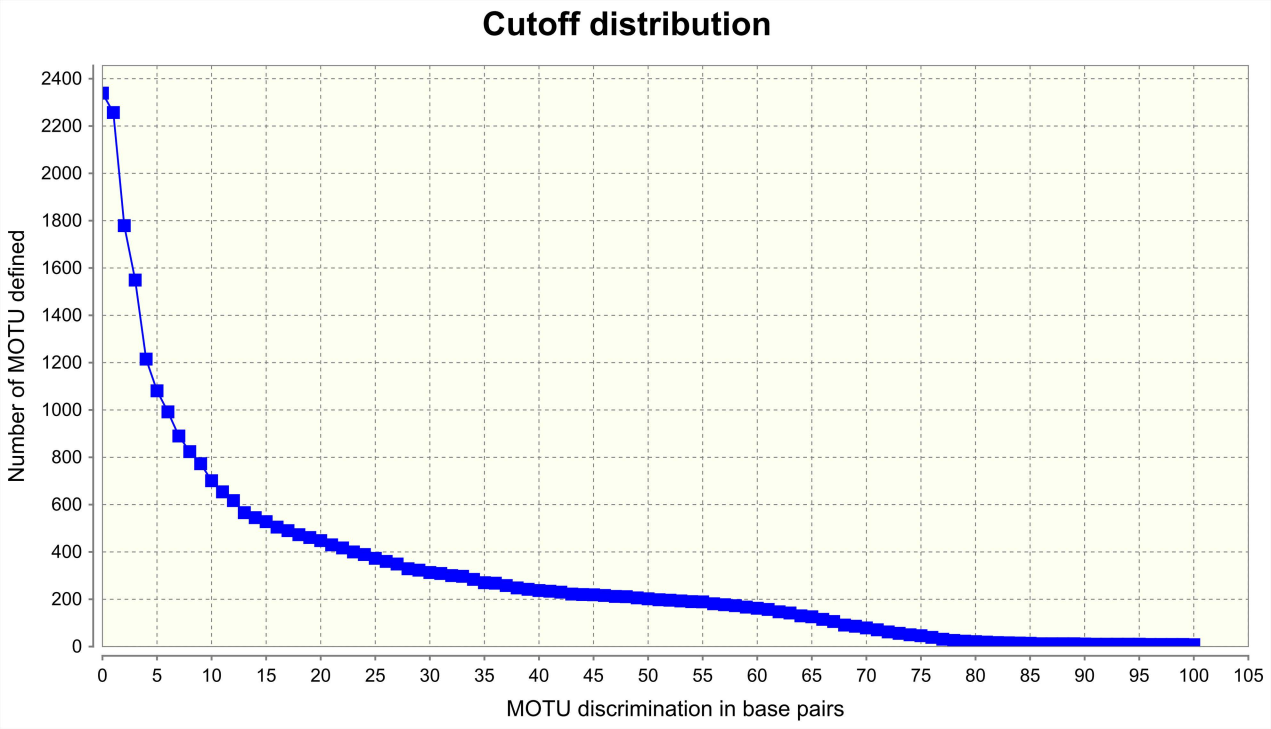
File S4. Maximum likelihood tree based on DNA barcodes using bPTP model.

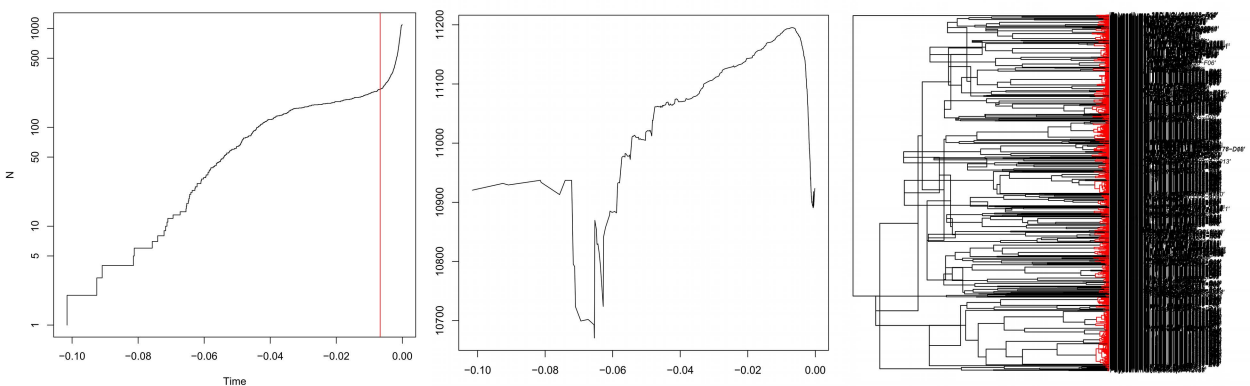
File S5. Maximum likelihood tree based on DNA barcodes using mPTP model.

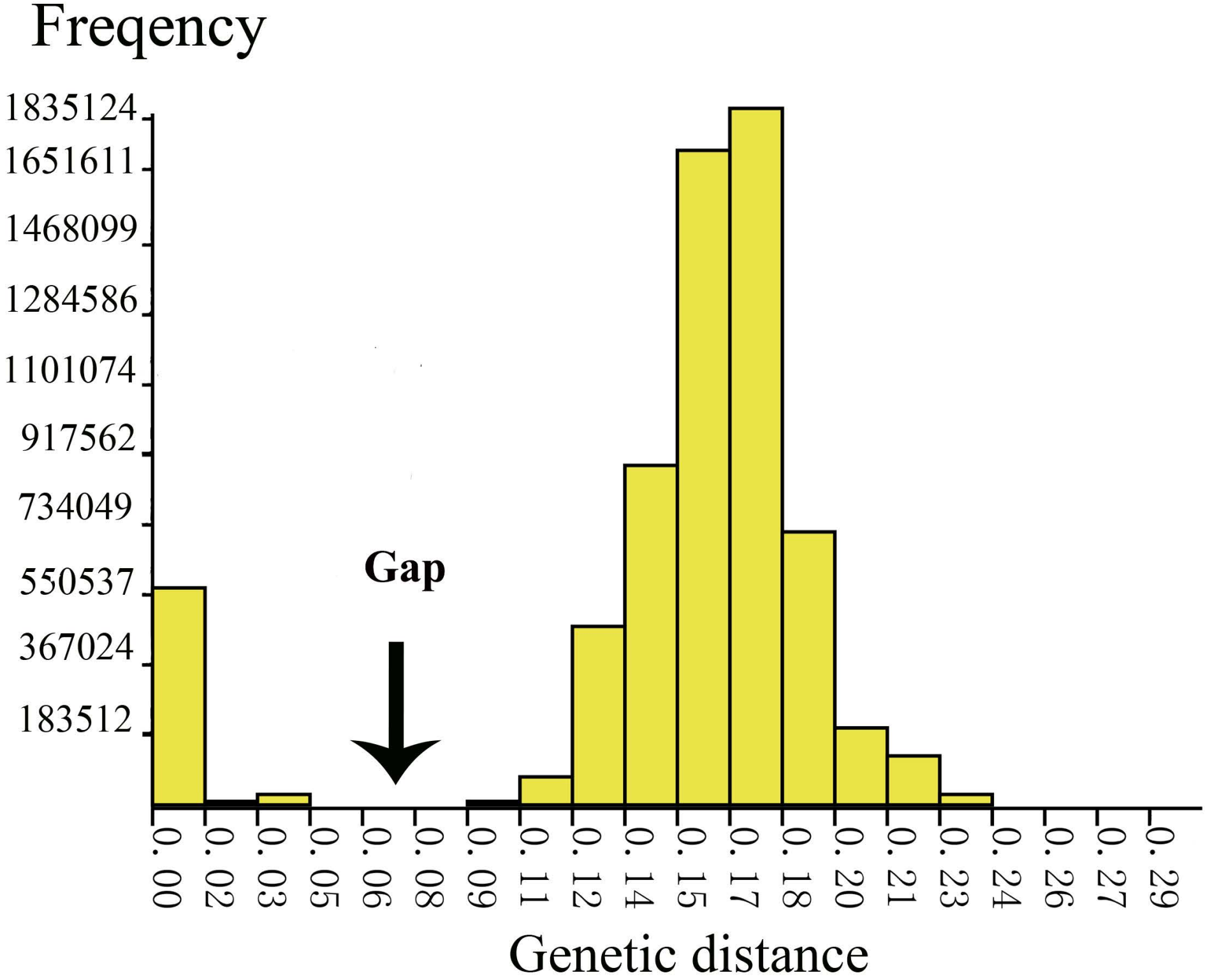

Frequency
0
200
400
600
800
1000
500
600
seq. length
A
Frequency
0
1000000
2000000
3000000
0.00
0.10
0.20
0.30
inter. distance
B
Frequency
0
200000
400000
0.00
0.10
0.20
intra. distance
C
Frequency
0
1000000
2000000
3000000
0.00
0.10
0.20
0.30
intra.-inter. distance
D

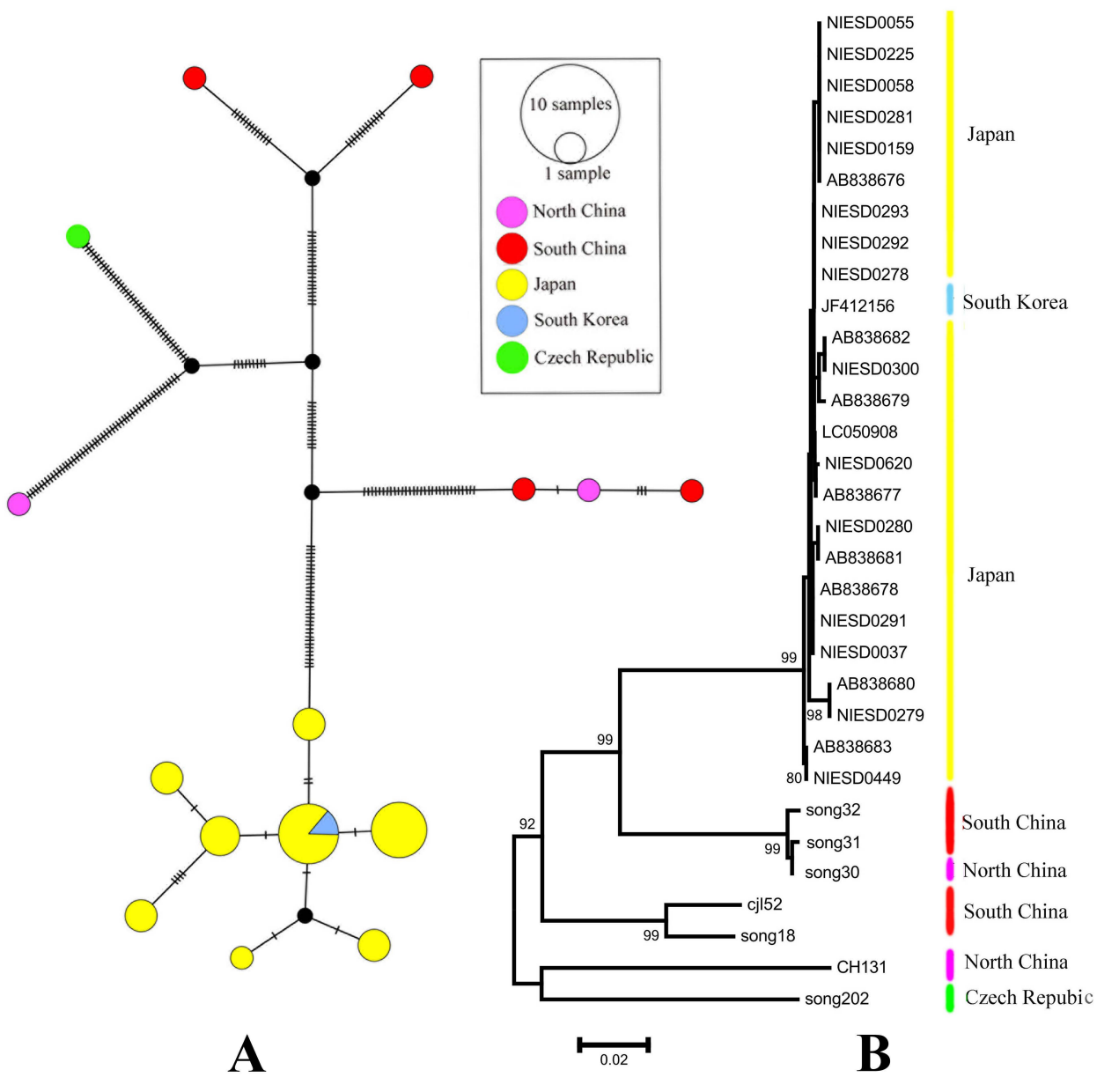

OTUs
300
280
260
240
220
200
180
160
140
120
100
80
60
40
20
0
0.005
0.007
0.010
0.014
0.019
0.026
0.037
0.051
0.072
0.100
Prior intraspecific divergence

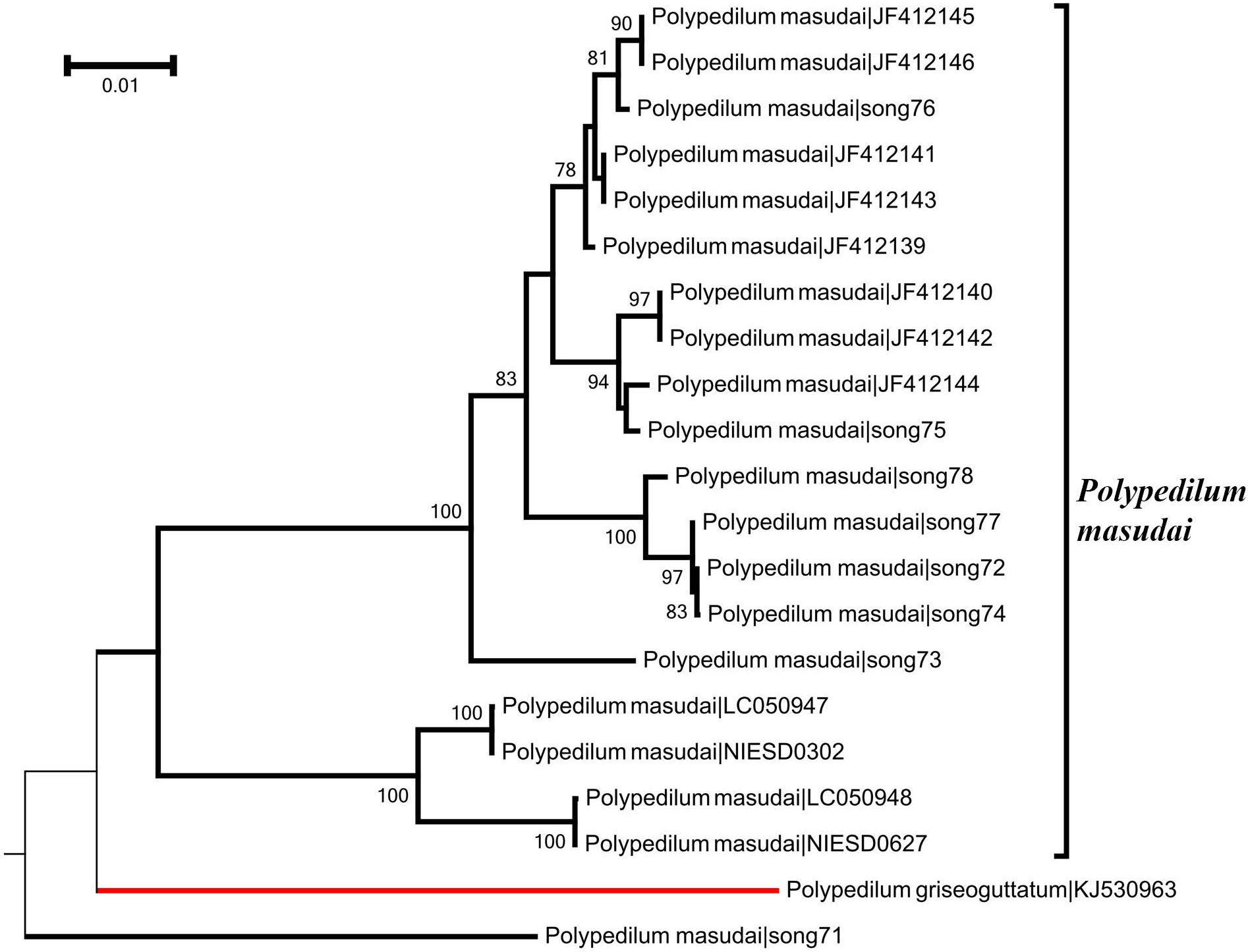

OTUs
400
350
300
250
200
150
100
50
0
0
2
4
6
8
10
12
Threshold (%)

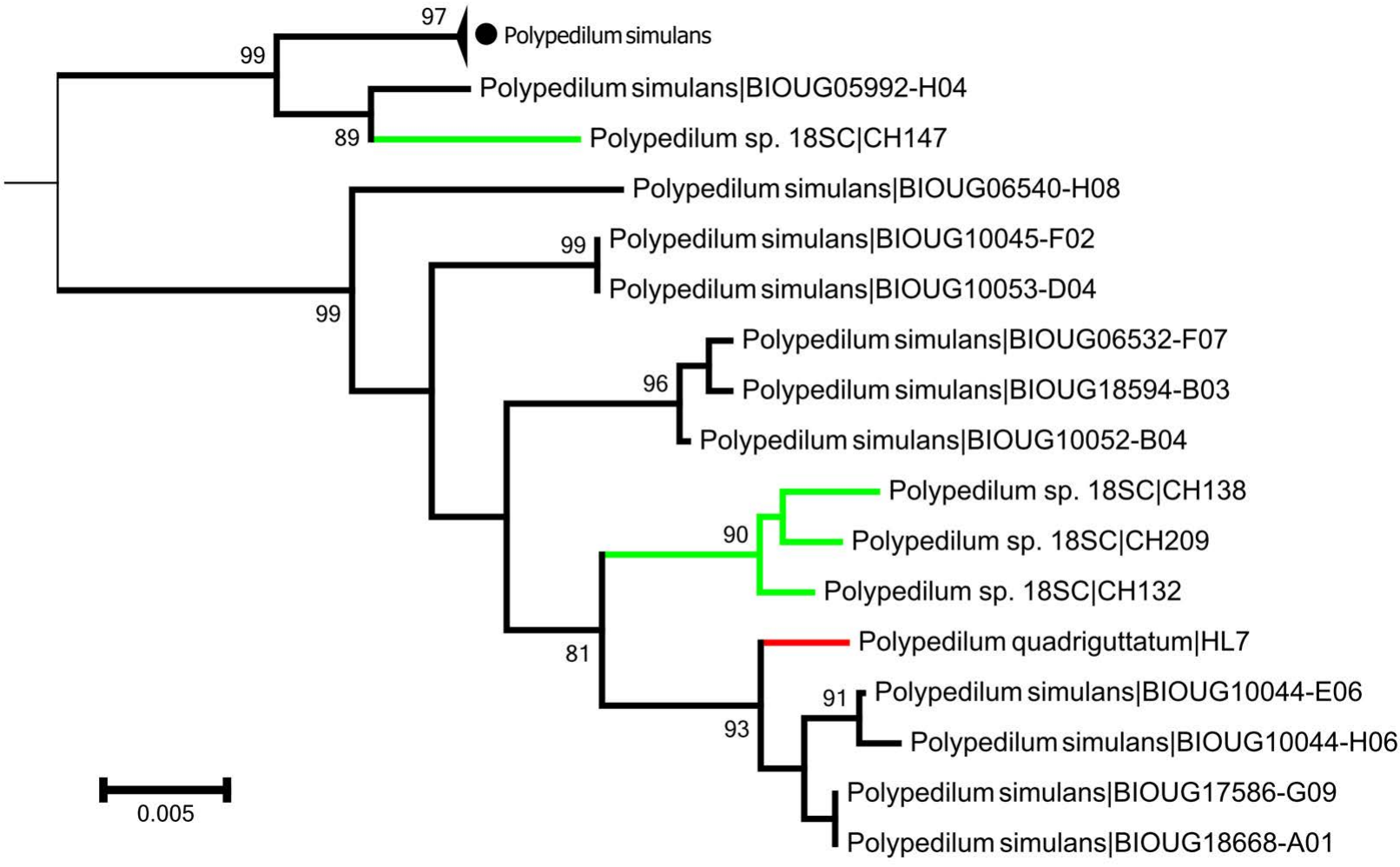

Cutoff distribution
Number of MOTU defined
2400
2200
2000
1800
1600
1400
1200
1000
800
600
400
200
0
0
5
10
15
20
25
30
35
40
45
50
55
60
65
70
75
80
85
90
95
100
105
MOTU discrimination in base pairs

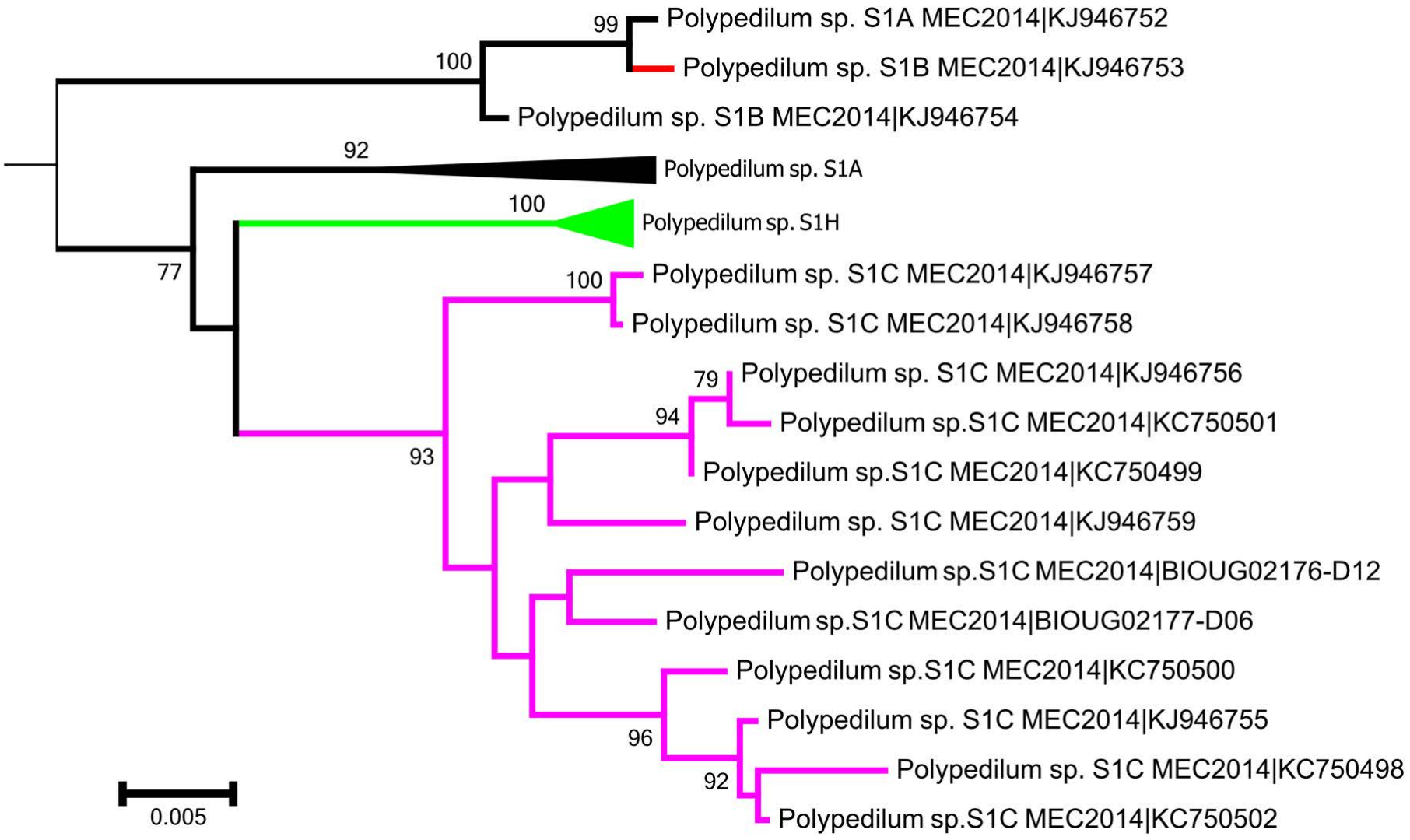

N
1000
500
100
50
10
5
1
−0.10
−0.08
−0.06
−0.04
−0.02
0.00
Time
A
11200
11100
11000
10900
10800
10700
−0.10
−0.08
−0.06
−0.04
−0.02
0.00
Time
B
C

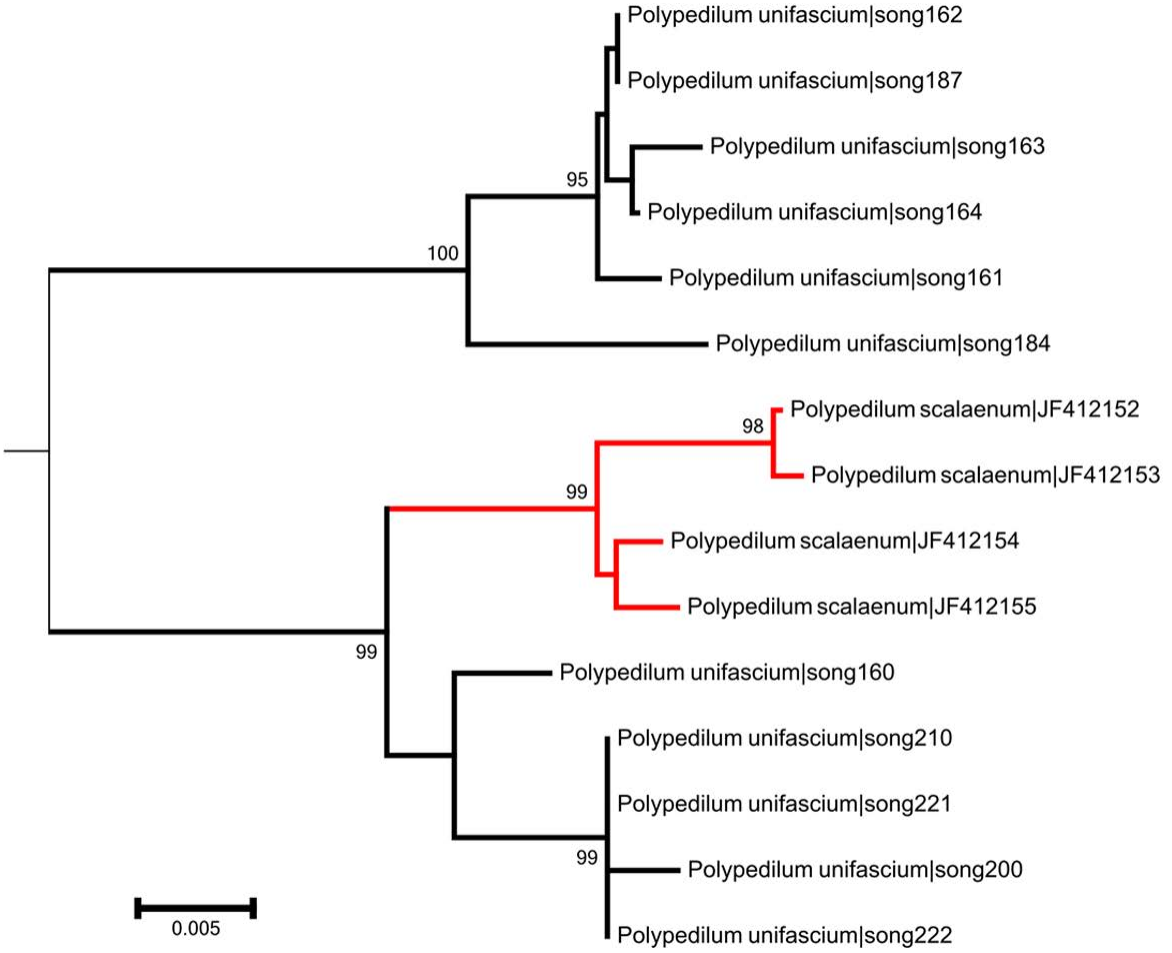

Freqency
1835124
1651611
1468099
1284586
1101074
917562
734049
550537
367024
183512
Gap
0.00
0.02
0.03
0.05
0.06
0.08
0.09
0.11
0.12
0.14
0.15
0.17
0.18
0.20
0.21
0.23
0.24
0.26
0.27
0.29
Genetic distance


10 samples
1 sample
North China
South China
Japan
South Korea
Czech Republic
NIESD0055
NIESD0225
NIESD0058
NIESD0281
NIESD0159
AB838676
NIESD0293
NIESD0292
NIESD0278
JF412156
AB838682
NIESD0300
AB838679
LC050908
NIESD0620
AB838677
NIESD0280
AB838681
AB838678
NIESD0291
NIESD0037
AB838680
NIESD0279
AB838683
NIESD0449
song32
song31
song30
cjl52
song18
CH131
song202
99
98
99
80
92
99
99
Japan
South Korea
Japan
South China
North China
South China
North China
Czech Republic
0.02
A
B


0.01
Polypedilum masudai|JF412145
Polypedilum masudai|JF412146
Polypedilum masudai|song76
Polypedilum masudai|JF412141
Polypedilum masudai|JF412143
Polypedilum masudai|JF412139
Polypedilum masudai|JF412140
Polypedilum masudai|JF412142
Polypedilum masudai|JF412144
Polypedilum masudai|song75
Polypedilum masudai|song78
Polypedilum masudai|song77
Polypedilum masudai|song72
Polypedilum masudai|song74
Polypedilum masudai|song73
Polypedilum masudai|LC050947
Polypedilum masudai|NIESD0302
Polypedilum masudai|LC050948
Polypedilum masudai|NIESD0627
Polypedilum griseoguttatum|KJ530963
Polypedilum masudai|song71
90
81
78
97
94
83
100
100
97
83
100
100
100
Polypedilum masudai


97
Polypedilum simulans
99
Polypedilum simulans|BIOUG05992-H04
89
Polypedilum sp. 18SC|CH147
Polypedilum simulans|BIOUG06540-H08
99
Polypedilum simulans|BIOUG10045-F02
Polypedilum simulans|BIOUG10053-D04
99
Polypedilum simulans|BIOUG06532-F07
96
Polypedilum simulans|BIOUG18594-B03
Polypedilum simulans|BIOUG10052-B04
Polypedilum sp. 18SC|CH138
90
Polypedilum sp. 18SC|CH209
Polypedilum sp. 18SC|CH132
81
Polypedilum quadriguttatum|HL7
91
Polypedilum simulans|BIOUG10044-E06
93
Polypedilum simulans|BIOUG10044-H06
Polypedilum simulans|BIOUG17586-G09
Polypedilum simulans|BIOUG18668-A01
0.005


99
Polypedilum sp. S1A MEC2014|KJ946752
100
Polypedilum sp. S1B MEC2014|KJ946753
Polypedilum sp. S1B MEC2014|KJ946754
92
Polypedilum sp. S1A
100
Polypedilum sp. S1H
77
100
Polypedilum sp. S1C MEC2014|KJ946757
Polypedilum sp. S1C MEC2014|KJ946758
79
Polypedilum sp. S1C MEC2014|KJ946756
94
Polypedilum sp.S1C MEC2014|KC750501
93
Polypedilum sp.S1C MEC2014|KC750499
Polypedilum sp. S1C MEC2014|KJ946759
Polypedilum sp.S1C MEC2014|BIOUG02176-D12
Polypedilum sp.S1C MEC2014|BIOUG02177-D06
Polypedilum sp.S1C MEC2014|KC750500
96
Polypedilum sp. S1C MEC2014|KJ946755
92
Polypedilum sp. S1C MEC2014|KC750498
Polypedilum sp.S1C MEC2014|KC750502
0.005


Polypedilum unifascium|song162
Polypedilum unifascium|song187
Polypedilum unifascium|song163
95
Polypedilum unifascium|song164
100
Polypedilum unifascium|song161
Polypedilum unifascium|song184
Polypedilum scalaenum|JF412152
98
Polypedilum scalaenum|JF412153
99
Polypedilum scalaenum|JF412154
Polypedilum scalaenum|JF412155
99
Polypedilum unifascium|song160
Polypedilum unifascium|song210
Polypedilum unifascium|song221
99
Polypedilum unifascium|song200
Polypedilum unifascium|song222
0.005