DISCOMARK: Nuclear marker discovery from orthologous sequences using draft genome data

Sereina Rutschmann*[1,2,3], Harald Detering*[1,2,3], Sabrina Simon[4,5], Jakob Fredslund[6], Michael T. Monaghan[1,2]

Addresses:

[1]*Leibniz-Institute of Freshwater Ecology and Inland Fisheries (IGB), Müggelseedamm 301, 12587 Berlin, Germany*

[2]*Berlin Center for Genomics in Biodiversity Research, Königin-Luise-Straße 6-8, 14195 Berlin, Germany*

[3]*Department of Biochemistry, Genetics and Immunology, University of Vigo, 36310 Vigo, Spain*

[4]*Sackler Institute for Comparative Genomics, American Museum of Natural History, Central Park West and 79th St., New York, NY 10024, USA*

[5]*Biosystematics Group, Wageningen University, Droevendaalsesteeg 1, 6708 PB Wageningen, The Netherlands*

[6]*Alexandra Institute, Åbogade 34, 8200 Aarhus, Denmark*

Keywords:

marker discovery, mayfly, non-model organism, primer design, phylogenetics

Correspondence:

Sereina Rutschmann, Phylogenomics Lab, Department of Biochemistry, Genetics and Immunology, University of Vigo, 36310 Vigo, Spain. E-Mail: sereina.rutschmann@gmail.com

* these authors contributed equally

Running title:

DISCOMARK - Phylogenetic marker development

Abstract

High-throughput sequencing has laid the foundation for fast and cost-effective development of phylogenetic markers. Here we present the program DISCOMARK, which streamlines the development of nuclear DNA (nDNA) markers from whole-genome (or whole-transcriptome) sequencing data, combining local alignment, alignment trimming, reference mapping and primer design based on multiple sequence alignments in order to design primer pairs from input orthologous sequences. In order to demonstrate the suitability of DISCOMARK we designed markers for two groups of species, one consisting of closely related species and one group of distantly related species. For the closely related members of the species complex of *Cloeon dipterum* s.l. (Insecta, Ephemeroptera), the program discovered a total of 78 markers. Among these, we selected eight markers for amplification and Sanger sequencing. The exon sequence alignments (2,526 base pairs (bp)) were used to reconstruct a well supported phylogeny and to infer clearly structured haplotype networks. For the distantly related species we designed primers for several families in the insect order Ephemeroptera, using available genomic data from four sequenced species. We developed primer pairs for 23 markers that are designed to amplify across several families. The DISCOMARK program will enhance the development of new nDNA markers by providing a streamlined, automated approach to perform genome-scale scans for phylogenetic markers. The program is written in Python, released under a public license (GNU GPL v2), and together with a manual and example data set available at: https://github.com/hdetering/discomark.

Introduction

The inference of phylogenetic relationships has benefited profoundly from the availability of nuclear DNA (nDNA) sequences for an increasing number of organism groups. The development of new phylogenetic markers has provided unprecedented insight into the evolutionary relationships of non-model organisms in particular (Ellegren 2014). Large sets of nDNA markers (single copy genes) have recently been designed for taxonomic groups for which genomic resources were available, e.g. cichlid fish (Meyer *et al.* 2015), ray-finned fish (Near *et al.* 2012), reptiles (Ruane *et al.* 2014), birds (Kerr *et al.* 2014) and flowering plants (Zeng *et al.* 2014). However, for many taxonomic groups there are only a handful of nDNA markers available that are suitable for phylogenetic reconstruction. Other approaches, such as ultra-conserved element (UCE) sequencing (Faircloth *et al.* 2012), anchored hybrid enrichment (Lemmon and Lemmon 2012), restriction site-associated DNA (RAD) sequencing (Baird *et al.* 2008) or genotyping by sequencing (GBS, Elshire *et al.* 2011) have become popular for addressing specific questions in systematics or population genetics; however, these methods are still cost-intensive, require a comparatively high amount of starting DNA material and can depend on the availability of reference genomes (e.g. anchored hybrid enrichment). Consequently, standard Sanger sequencing approaches are still in high demand for various research questions.

Identification of novel phylogenetic markers has been a predominantly manual process, which impedes their large-scale development, and comprehensive primer design based on large sets of multiple sequence alignments remains challenging. Existing tools can generally be classified into those developed for primer design or for marker discovery. Primer design tools can design primer pairs for single loci (e.g. GEMI, Sobhy *et al.* 2012; PRIMER3, Untergasser *et al.* 2012; CEMASUITE, Lane *et al.* 2015) or multiple loci at once

70 (BATCHPRIMER3, You *et al.* 2008; PRIMERVIEW, O'Halloran 2015). Some programs are
71 specific for highly variable DNA targets (PRIMERDESIGN, Brodin *et al.* 2013; PRIMERDESIGN-
72 M, Yoon and Leitner 2015), viral genomes (PRISM, Yu *et al.* 2015), and transcriptome input
73 data (SCRIMER, Morkovsky *et al.* 2015). Marker discovery tools target single nucleotide
74 polymorphism (SNP) markers in polyploid organisms (POLYMARKER, Ramirez-Gonzalez *et*
75 *al.* 2015), and putative single copy loci from plant transcriptomes (MARKERMINER, Chamala
76 *et al.* 2015). The challenge of developing new phylogenetic markers lies both in the discovery
77 of conserved regions, the design of primer pairs and an estimation of the level of their
78 phylogenetic signal.

79 Here we present DISCOMARK (=Discovery of Markers), a flexible, user-friendly program
80 that identifies conserved regions and designs primers based on multiple sequence alignments
81 taken from FASTA-formatted files of putative orthologous sequences from whole-genome or
82 whole-transcriptome data. The program can be used to easily screen for phylogenetically
83 suitable nDNA markers and to design primers that can be used for Sanger sequencing as well
84 as high-throughput sequencing. The program is structured into several steps that can be
85 individually optimized by the user and run independently. In terms of input, the program can
86 be applied on large and small sets of taxa, including both closely and distantly related species.
87 Ideally, orthologous sequences in combination with a whole-genome reference sequence are
88 used. Thus, exon/intron boundaries can be inferred using the reference for each marker. Under
89 the default settings, the program will design several primer pairs that anneal in conserved
90 regions. The visualization of the alignments with potential primers allows the user to choose
91 between primers targeting exons or introns (e.g. exon-primed intron-crossing (EPIC)
92 markers). Additionally, information about the suitability as phylogenetic markers is provided
93 by an estimate of the number of SNPs per marker and the applicability across species. Finally,

94 we demonstrate the utility of DISCOMARK for (1) closely related species (i.e. *Cloeon dipterum*
95 s.l. species complex) using whole-genome data, and (2) distantly related species (i.e. insect
96 order Ephemeroptera) using whole-genome data derived from genome sequencing projects. In
97 order to generate genomic reference sequences we used draft whole genome sequencing at
98 shallow coverage followed by draft genome assembly. In one scenario (*C. dipterum* s.l.
99 species complex) we demonstrate that incomplete genomic data can be used for ortholog
100 prediction and primer design as well.

101 **Materials and Methods**

102 *DISCOMARK implementation*

103 The program DISCOMARK is written in Python and was developed to design primer pairs in
104 conserved regions of predicted orthologous genes. Orthologs are required for phylogenetic
105 studies. The ortholog identification step is not part of the DISCOMARK workflow but
106 DISCOMARK is designed to directly work with the output of several ortholog prediction
107 programs, e.g. HAMSTR (Ebersberger *et al.* 2009), or Orthograph
108 (https://github.com/mptrsen/Orthograph, last accessed March 25, 2016). Orthologous groups
109 may be derived from genome or transcriptome sequence data. In addition to the orthologous
110 genes, genomic data such as whole-genome sequencing data can be provided to DISCOMARK
111 as a guide to detect exon/intron boundaries. DISCOMARK performs seven steps, combining
112 Python scripts with widely used bioinformatics programs (Fig. 1). The steps: (1) combine
113 orthologous groups of sequences, (2) align sequences of each orthologous group using
114 MAFFT v.7.205 (Katoh and Standley 2013), (3) trim sequence alignments with TRIMAL v.1.4
115 (Capella-Gutierrez *et al.* 2009), (4) align sequences against a reference (e.g. whole-genome
116 dataset from the same or closely related taxa) with BLASTN v.2.2.29 (Altschul *et al.* 1997;

117 Camacho *et al.* 2009) and re-alignment using MAFFT, (5) design primer pairs on single-gene
118 alignments using a modified version of PRIFI (Fredslund *et al.* 2005), adapted by us into a
119 Python package that uses BioPython v.1.65 and Python v.3.4.3, (6) check primer specificity
120 with BLASTN, and (7) generate output in several formats (visual HTML report, tabular data
121 and FASTA files of the primers). The results of each step can be inspected in the respective
122 output folders.

123 *1. Combine sequences*. In the first step, the putative orthologous sequences of different taxa
124 are combined according to the orthologous groups. The input files are expected to be
125 nucleotide sequences in FASTA format. We recommend using putative orthologous exon
126 sequences (e.g. CDS) in combination with whole-genome data (e.g. a draft genome
127 assembly). Each input file is expected to contain the sequences of one orthologous group;
128 orthologs of each input taxon are to be organized into a taxon folder. Importantly, file names
129 represent the ortholog identifiers used to combine orthologous sequences of the various input
130 taxa; by default, ortholog prediction tools follow that convention.

131 *2. Align sequences*. Orthologous sequences combined according to the orthologous groups are
132 separately aligned with the multiple sequence alignment (MSA) program MAFFT. Alignment
133 parameters can be specified by the user via a configuration file (discomark.conf, located in the
134 program folder). Default parameters are the following: '--localpair --maxiterate 16 --
135 inputorder --preservecase --quiet' (L-INS-i alignment method). We chose MAFFT as multiple
136 alignment tool because it combines accuracy and efficiency and has been adopted widely in
137 the scientific community (Pais *et al.* 2014; Szitenberg *et al.* 2015).

138 *3. Trim alignments*. In order to remove poorly aligned regions, sequence alignments are
139 trimmed using TRIMAL. The program TRIMAL analyzes the distribution of gaps and

140 mismatches in the alignment and discard alignment positions and sequences of low quality.
141 By default, DISCOMARK calls TRIMAL with the '-strictplus' method. The preset is used by
142 TRIMAL to derive the specific thresholds for alignment trimming (minimum gap score,
143 minimum residue similarity score, conserved block size). Since alignment trimming largely
144 depends on the input data and influences the downstream results, TRIMAL can also be run with
145 different settings (e.g. '-gappyout', '-strict', '-automated1'; but see Capella-Gutierrez *et al.*
146 (2009). Alternatively, there is also the option to deactivate the alignment trimming with the
147 DISCOMARK option '--no-trim' or use alternative trimming programs such as GBLOCKS
148 (Castresana 2000; Talavera and Castresana 2007) or GUIDANCE2 (Landan and Graur 2008;
149 Sela *et al.* 2015).

150 *4. Blast and alignment to reference*. In this step a genomic reference sequence for each input
151 ortholog is identified and added to the trimmed alignment. This step is particularly important
152 when working with coding sequences which do not contain intron sequences; thus, a genomic
153 sequence is needed to infer intron/exon boundaries. Working with coding sequences is
154 advisable for more distantly related taxa which may include intron length polymorphisms, or
155 to target EPIC markers. Any whole-genome data set (from one of the included taxa or a
156 closely related taxa) can be used as reference for mapping the ortholog sequences. Here,
157 mapping means that the input sequences are compared to the reference sequences, which are
158 defined by the user using the local alignment program BLASTN. The best locally aligning
159 reference sequence (the one that yields the longest alignment among all input sequences) for
160 each orthologous group is added to the corresponding sequence alignment. Reference
161 sequences are cut to 100 base pairs (bp) upstream and downstream of the first, respectively
162 last, BLAST hit to avoid alignment length inflation. Then, the extended alignments are re-
163 aligned with MAFFT. Finally, the best BLAST hits of all sequences belonging to a marker are

164 compared and a 'uniq_ref' flag is set if they all map to the same reference sequence. The
165 reference alignment step is optional; however, the inclusion of whole-genome data is essential
166 for estimating intron/exon boundaries. Given that information, the focus of target sequences
167 to be amplified can be on entire exon markers, EPIC markers, or a combination.

168 *5. Design primers*. The single-gene alignments, after trimming, mapping and re-aligning to a
169 reference, are used as input to design primer pairs. We integrated the webtool PRIFI
170 (http://cgi-www.daimi.au.dk/cgi-chili/PriFi/main, last accessed December 20, 2015) as a
171 Python package that provides a comprehensive set of parameters. As default settings for
172 DISCOMARK we chose the following: estimated product length between 200-1,000 bp
173 ('OptimalProductLength = [400, 600, 800, 1000], MinProductLength = 200,
174 MaxProductLength = 1000'), maximum number of ambiguity positions within the primer
175 sequences ('MaxMismatches = 2'), primer length between 20-30 bp ('MinPrimerLength = 20,
176 MaxPrimerLength = 30, OptimalPrimerLength = [20, 25]'), melting temperature of the primer
177 pairs between 50-60°C ('MinTm = 50.0, MinTmWithMismatchesAllowed = 58.0,
178 SuggestedMaxTm = 60.0'), and we set the maximum number of primer pairs per alignment to
179 six (note: only settings different from the PRIFI default are mentioned above). The program
180 PRIFI was originally developed to design intron-spanning markers (but see Fredslund et al.
181 2005). Here we use it because it enables primer design based on MSA input. Parameters for
182 PRIFI can be specified in the DISCOMARK configuration file ('discomark.conf').

183 *6. Check marker specificity*. To ensure the specificity of the designed primer pairs, we
184 compare their sequences against the NCBI database ('refseq_mrna'). Primer sequences are
185 searched in the NCBI database ('refseq_mrna') using the online BLASTN interface. The
186 default search settings are restricted to human and bacterial targets using the Entrez query

187 'txid2[ORGN] OR txid9606[ORGN]' because these are most likely to be present as
188 contaminants in sequencing libraries. The result hits of the BLAST search are indicated to the
189 user in the HTML output.

190 *7. Visualize results*. As final step, the program produces a HTML report containing the list of
191 designed primers, an alignment viewer and plots visualizing the discovered set of markers
192 (Fig. 2). Besides the primer sequences the report lists several features such as the melting
193 temperatures, predicted sequence length, and the number of taxa amplified by each primer set
194 (Fig. 2a). Selected primer pairs and primer lists can be downloaded as FASTA or CSV files,
195 respectively. The report highlights the species coverage achieved by each discovered marker,
196 i.e. how many species' sequences each primer set is expected to amplify, as an estimate of
197 how universally each primer set can be applied. An identification of uniqueness within the
198 reference genome (as an estimator of single-copy status) is given within the flag 'uref',
199 starting whether all sequences belonging to a marker had their best BLAST hit to the same
200 reference sequence. Additionally, functional annotations are reported, if available, to guide
201 the user in the selection of markers of interest. Annotations can be supplied in form of a tab-
202 delimited file with the '-a' option. In principle, any kind of annotations can be used depending
203 on the desired research objective. In our usage scenarios, we used gene ontology (GO) terms
204 which were retrieved by mapping the gene IDs contained in the HAMSTR core ortholog set
205 via the UniProt website (http://www.uniprot.org/, last accessed December 20, 2015).
206 Common input ortholog sequences are visualized (Fig. 2b). In order to provide a measure of
207 the suitability of the markers for phylogenetic reconstruction the program calculates the
208 number of SNPs between a primer pair by comparing the aligned input sequences against
209 each other. The number of SNPs between each primer pair is visualized in relation to the
210 estimated product length (Fig. 2c) and reported in the tabular output.

211 *Usage cases*

212 *Closely related species - Cloeon dipterum s.l. species complex.* To test the suitability of
213 DISCOMARK for closely related species, we designed primer pairs for the species complex of
214 *C. dipterum* s.l. (Ephemeroptera: Baetidae). The species complex consists of several closely
215 related species, including *Cloeon peregrinator* GATTOLLIAT & SARTORI, 2008 from Madeira
216 (Gattolliat *et al.* 2008; Rutschmann *et al.* 2014). As input to design the primer pairs, we used
217 whole-genome sequencing data of *C. dipterum* L. 1761 and expressed sequence tags (EST) of
218 *Baetis* sp. (Table 1). The sequence reads of *C. dipterum* were trimmed and *de novo* assembled
219 using NEWBLER v.2.5.3 (454 Life Science Corporation) under the default settings for large
220 datasets. Ortholog sequences prediction of both data sets was performed with HAMSTR v.9
221 using the insecta_hmmer3-2 core reference taxa set (http://www.deep-
222 phylogeny.org/hamstr/download/datasets/hmmer3/insecta_hmmer3-2.tar.gz, last accessed
223 December 20, 2015), including 1,579 orthologous genes. We ran the program DISCOMARK
224 with default settings ('python run_project.py -i input/Cloeon -i input/Baetis -r
225 input/reference/Cloeon.fa -a input/co2go.ixosc.csv -d output/cloeon_baetis'), using the
226 predicted orthologs from HAMSTR and the whole-genome *Cloeon*-data as reference (step 4).
227 For comparison, we also ran DISCOMARK without a reference and also present these results.
228 The Pearson correlation between the number of SNPs located between primer pairs and
229 corresponding estimated product length was calculated using the function cor within the stats
230 package for R (R Development Core Team, 2016). A t-test for significance was performed
231 using the function cor.test.

232 From the total of designed primer pairs (78 markers, 445 primer pairs, see results) we
233 selected eight and amplified them for four species of the *C. dipterum* species complex (Table

234 1) in the laboratory. The eight markers were manually selected based on the following
235 criteria: best alignment (i.e. EPIC markers), longest product length, and most species covered.
236 We used standardized polymerase chain reactions (PCR; 35-40 PCR cycles with annealing
237 temperature of 55°C), followed by Sanger sequencing. Forward and reverse sequences were
238 assembled and edited with GENEIOUS R7 v.7.1.3 (Biomatters Ltd.), indicating ambiguous
239 positions following the IUPAC nucleotide codes. Sequences containing heterozygous indels
240 (i.e. alleles with length polymorphisms) were phased with CODONCODEALIGNER v.3.5.6
241 (CodonCode Corporation). For this we used the implemented 'Split Heterozygous Indels'
242 function. Multiple sequence alignments were created for all sequences for each marker. The
243 predicted orthologous sequences of *Baetis* sp. were used as reference to infer the exon-intron
244 splicing boundaries (canonical and non-canonical splice site pairs). The final sequence
245 alignments were checked for the occurrence of stop codons and indels, and split into exon and
246 intron parts using a custom Python script (https://github.com/srutschmann/python_scripts, last
247 accessed March 28, 2016). Sequence alignments were phased using the program PHASE
248 v.2.1.1 (Stephens *et al.* 2001; Stephens and Donnelly 2003) with a cutoff value of 0.6
249 (Harrigan *et al.* 2008; Garrick *et al.* 2010), whereby input and output files were formatted
250 using the Perl scripts included in SEQPHASE (Flot 2010). Heterozygous sites that could not
251 be resolved were coded as ambiguity codes for subsequent analyses. After phasing, all
252 alignments were re-aligned with MAFFT. The number of variable and informative sites, and
253 the nucleotide diversity per exon alignment was calculated with a custom script.

254 To investigate the heterogeneity of each marker´s DNA sequences, we reconstructed
255 haplotype networks, using FITCHI (Matschiner 2016). As input for each marker we inferred a
256 gene tree using the program RAXML v.8 (Stamatakis 2014) with the GTRCAT model and
257 1,000 bootstrap replicates under the rapid bootstrap algorithm. The phylogenetic relationships

258 were calculated with Bayesian inference, using MRBAYES v.3.2.3 (Ronquist *et al.* 2012) based
259 on a concatenated nDNA matrix that consisted of the exon sequences from all 15 nDNA
260 markers. The best-fitting model of molecular evolution for each sequence alignment was
261 selected via a BIC criterion in JMODELTEST v.2.1 (Guindon and Gascuel 2003; Darriba *et al.*
262 2012). We calculated 10^6 generations with random seed, a burn-in of 25% and four MCMC
263 chains. As an outgroup we used the predicted orthologous sequences of *Baetis* sp..

264 *Distantly related species - insect order Ephemeroptera.* In this test case, we used contigs
265 derived from whole-genome sequencing projects of the species *Baetis* sp., *Ephemera danica*
266 MÜLLER 1764, *Eurylophella* sp., and *Isonychia bicolor* WALKER 1853 (Table 1). The contigs
267 from each species were used for ortholog predicting with HAMSTR v.13.2.4
268 (http://sourceforge.net/projects/hamstr/files/hamstr.v13.2.4.tar.gz, last accessed December 20,
269 2015). We ran DISCOMARK with the default settings twice. For the first run, we used the
270 *Baetis* sp. data as reference ('python run_project.py -i input/Baetis -i input/Ephemera -i
271 input/Eurylophella -i input/Isonychia -r input/reference/Baetis.fa -a input/co2go.ixosc.csv -d
272 output/mayflies'). The second run was performed without a reference ('python run_project.py
273 -i input/Baetis -i input/Ephemera -i input/Eurylophella -i input/Isonychia -a
274 input/co2go.ixosc.csv -d output/mayflies_without_reference').

275 **Results**

276 *Closely related species - species complex of Cloeon dipterum s.l.*

277 Using a reference, DISCOMARK identified 78 nDNA markers with 445 primer pairs for
278 orthologous sequences of both species (*Baetis* sp. and *C. dipterum* s.l.). Ortholog prediction
279 yielded 403 orthologous sequences for the *Baetis* sp. EST-data and 1,211 for *C. dipterum*. For

280 the individual species, DISCOMARK identified 793 markers for *C. dipterum* and 123 for *Baetis*
281 sp. Markers including both species were between 200 and 931 bp with median length of 412
282 bp long. Their number of SNPs per marker ranged from zero to 82 (median: five) with an
283 average of one SNP per 68 bp. Marker length and number of SNPs of the common markers
284 were correlated with a Pearson's correlation coefficient of 0.28 (Pearson's product-moment
285 correlation $P < 0.001$). Without using a reference, DISCOMARK identified 73 markers with
286 460 primer pairs for orthologous sequences of both species. The total run time for this data set
287 on a local Linux machine (quad-core Intel i5, 8 GB RAM) was < 30 min.

288 The haplotype networks based on the eight selected markers showed a clear structure for
289 all markers, including two markers with shared haplotypes for the two species from the U.S.
290 and Madeira (Fig. 3 and Fig. S1, Supporting information). The length of the concatenated
291 sequence alignment of the eight markers was 3,530 bp (2,526 bp exon sequence, Table S1,
292 Supporting information). The exon sequence matrix contained 78 variable sites, 27
293 informative sites, and was 92.6% complete. The nucleotide diversity ranged between 0.009
294 and 0.028 (median: 0.013). Phylogenetic tree reconstruction based on these eight markers
295 resulted in a phylogeny with fully resolved nodes (Bayesian posterior probability (PP) ≥ 95%;
296 Fig. 3a).

297 The species *C. dipterum* sp1 was found as outgroup to a clade containing the species *C.*
298 *dipterum* sp2 from Switzerland and the two species from the U.S and Madeira. The latter two
299 species formed a monophyletic clade. The use of the marker set for *C. dipterum* developed
300 here resulted in a fully resolved phylogenetic tree in contrast to the mitochondrial tree of
301 Rutschmann *et al.* (2014). The phylogeny of the *Cloeon*-species complex presented here also

302 is in complete agreement with species tree reconstructions based on 59 nuclear DNA markers
303 (Rutschmann *et al.*, submitted).

304 *Distantly related species - insect order Ephemeroptera*

305 In total, we found 23 orthologs with a total of 53 primer pairs for all four species (Table S2,
306 Supporting information) for the first run with a reference. The input files per species (i.e.
307 putative orthologous sequences) ranged from 1,445 to 1,523. We detected 38 markers that
308 covered three of the species (99 primer pairs), 81 markers covering two species (258 primer
309 pairs), and 118 markers that covered any single species (684 primer pairs). For the individual
310 species, *Baetis* sp. had the most markers available (209) of the single- and multi-species
311 markers. There were 136 markers for *Eurylophella* sp., 104 markers for *I. bicolor*, and 87
312 markers for *E. danica*. The lengths for all markers covering all four species varied between
313 207 and 997 bp with median of 517 bp, containing between 39 and 298 SNPs per marker with
314 a SNP every four bp on average. Marker length and number of SNPs were correlated with a
315 Pearson's correlation coefficient of 0.96 (Pearson's product-moment correlation $P < 0.001$).
316 Without a reference, DISCOMARK identified 25 orthologs with 66 primer pairs for all four
317 species. Run time for this data set on a Linux client (quad-core Intel i5, 8 GB RAM) was < 50
318 min.

319 **Discussion**

320 DISCOMARK is the first stand-alone program of which we are aware that discovers putative
321 single-copy nDNA markers and designs primer pairs based on multiple sequence alignments
322 on a genome-wide scale. The visual output gives guidance on the suitability of each marker
323 i.e. variability within and between species measured as number of SNPs, and information

about the included species of each marker. Using this approach, primers can be specifically chosen to match the 'phylogenetic resolution', i.e. many markers with intermediate number of SNPs for closely related species, and fewer markers with generally higher number of SNPs for distantly related species can be selected. The automatic processing, including combining, aligning, trimming and blasting sequences of any nucleotide FASTA sequences together with the produced graphical output significantly facilitate the design of primer pairs for a large number of nDNA markers. Nevertheless, users retain a high degree of flexibility by the stepwise nature of the workflow. DISCOMARK is free, open-source software to assist the development of markers for non-model species on the genome scale. We demonstrated the efficacy of our approach for closely related species as well as for members of divergent families within an order of insects. Using a reference genome enabled resolution of intron-exon boundaries but is not a strict requirement for marker design. We strongly recommend carefully checking the selected primer pairs in the alignment viewer. The performance of DISCOMARK will largely depend on the properties of your input data (i.e. fidelity of ortholog prediction, genome complexity, divergence of species.)

Marker development within the order Ephemeroptera

The usage of DISCOMARK added an extensive set of new potential nDNA markers to the ones that have been used to date for mayfly (Ephemeroptera) phylogenies based on few individual genes (histone 3, elongation factor 1 alpha, phosphenolpyruvate carboxykinase (Vuataz *et al.* 2011; Pereira-da-Conceicoa *et al.* 2012; Vuataz *et al.* 2013)). Most recent tree reconstructions remain based on mitochondrial DNA markers (e.g. Rutschmann *et al.* 2014 (used three mitochondrial genes); Leys *et al.* 2016 (used cytochrome *c* oxidase subunit 1 (*cox1*) gene)). With DISCOMARK, the availability of more genome data will increase the number of markers

347 suitable for phylogenetic studies. This will promote more fine-scale phylogenetic studies,
348 which are needed to resolve more recent evolutionary events and the phylogenetic
349 relationships of morphologically cryptic species that can not be resolved with standard
350 markers (Dijkstra *et al.* 2014).

351 **Acknowledgements**

352 We thank Katrin Preuß and Susan Mbedi for their help with genome sequencing, and to our
353 research groups, in particular to David Posada and Sara Rocha, and to Eric Coissac and three
354 anonymous reviewers for constructive comments that improved the quality of this project.
355 Research was partially supported by the Leibniz Association (PAKT für Forschung und
356 Innovation "FREDIE" project), the Swiss National Science Foundation (Early
357 PostDoc.Mobility grant P2SKP3_15869 to S.R.), and a visiting fellowship from the Japan
358 Association for the Advancement of Science (L-15543 to M.T.M.). This is publication
359 number 39 of the Berlin Center for Genomics in Biodiversity Research.

360 **References**

361 Altschul SF, Madden TL, Schaffer AA *et al.* (1997) Gapped BLAST and PSI-BLAST: a new
362 generation of protein database search programs. *Nucleic Acids Research*, **25**, 3389-3402.
363 Baird NA, Etter PD, Atwood TS *et al.* (2008) Rapid SNP Discovery and Genetic Mapping
364 Using Sequenced RAD Markers. *PLoS ONE*, **3**, e3376.
365 Brodin J, Krishnamoorthy M, Athreya G *et al.* (2013) A multiple-alignment based primer
366 design algorithm for genetically highly variable DNA targets. *BMC Bioinformatics*, **14**,
367 255.
368 Camacho C, Coulouris G, Avagyan V *et al.* (2009) BLAST+: architecture and applications.
369 *BMC Bioinformatics*, **10**, 421.
370 Capella-Gutierrez S, Silla-Martinez JM, Gabaldon T (2009) trimAl: a tool for automated
371 alignment trimming in large-scale phylogenetic analyses. *Bioinformatics*, **25**, 1972-1973.
372 Castresana J (2000) Selection of conserved blocks from multiple alignments for their use in
373 phylogenetic analysis. *Molecular Biology and Evolution*, **17**, 540-552.
374 Chamala S, García N, Godden GT *et al.* (2015) MarkerMiner 1.0: new application for
375 phylogenetic marker development using angiosperm transcriptomes. *Applications in Plant*
376 *Sciences*, **4**, 1400115.

Dijkstra KD, Monaghan MT, Pauls SU (2014) Freshwater biodiversity and aquatic insect diversification. *Annual Review of Entomology*, **59**, 143-163.

Darriba D, Taboada GL, Doallo R, Posada D (2012) jModelTest 2: more models, new heuristics and parallel computing. *Nature Methods*, **9**, 772.

Ebersberger I, Strauss S, von Haeseler A (2009) HaMStR: profile hidden markov model based search for orthologs in ESTs. *BMC Evolutionary Biology*, **9**, 157.

Ellegren H (2014) Genome sequencing and population genomics in non-model organisms. *Trends in Ecology & Evolution*, **29**, 51-63.

Elshire RJ, Glaubitz JC, Sun Q *et al.* (2011) A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS ONE*, **6**, e19379.

Faircloth BC, McCormack JE, Crawford NG *et al.* (2012) Ultraconserved elements anchor thousands of genetic markers spanning multiple evolutionary timescales. *Systematic Biology*, **61**, 717-726.

Flot J-F (2010) seqphase: a web tool for interconverting phase input/output files and fasta sequence alignments. *Molecular Ecology Resources*, **10**, 162-166.

Fredslund J, Schauser L, Madsen LH, Sandal N, Stougaard J (2005) PriFi: using a multiple alignment of related sequences to find primers for amplification of homologs. *Nucleic Acids Research*, **33**, W516-520.

Garrick RC, Sunnucks P, Dyer RJ (2010) Nuclear gene phylogeography using PHASE: dealing with unresolved genotypes, lost alleles, and systematic bias in parameter estimation. *BMC Evolutionary Biology*, **10**, 118.

Gattolliat J-L, Hugher SJ, Monaghan MT, Sartori M (2008) Revision of Mdeiran mayflies (Insecta, Ephemeroptera). *Zootaxa*, **1957**, 69-80.

Guindon S, Gascuel O (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Systematic Biology*, **52**, 696-704.

Harrigan RJ, Mazza ME, Sorenson MD (2008) Computation vs. cloning: evaluation of two methods for haplotype determination. *Molecular Ecology Resources*, **8**, 1239-1248.

Katoh K, Standley DM (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution*, **30**, 772-780.

Kerr KCR, Cloutier A, Baker AJ (2014) One hundred new universal exonic markers for birds developed from a genomic pipeline. *Journal of Ornithology*, **155**, 561-569.

Landan G, Graur D (2008) Local reliability measures from sets of co-optimal multiple sequence alignments. *Pacific Symposium on Biocomputing*, 15-24.

Lane CE, Hulgan D, O'Quinn K, Benton MG (2015) CEMAsuite: open source degenerate PCR primer design. *Bioinformatics*, **31**, 3688-3690.

Lemmon AR, Lemmon EM (2012) High-throughput identification of informative nuclear loci for shallow-scale phylogenetics and phylogeography. *Systematic Biology*, **61**, 745-761.

Leys M, Keller I, Räsänen K, Gattolliat J-L, Robinson CT (2016) Distribution and population genetic variation of cryptic pecies of the Alpine mayfly *Baetis alpinus* (Ephemeroptera: Baetidae) in the Central Alps. *BMC Evolutionary Biology*, **16**:77.

Matschiner M (2016) Fitchi: haplotype genealogy graphs based on the Fitch algorithm. *Bioinformatic*s, **32**, 1250-1252.

Meyer BS, Matschiner M, Salzburger W (2015) A tribal level phylogeny of Lake Tanganyika cichlid fishes based on a genomic multi-marker approach. *Molecular Phylogenetics and Evolution*, **83**, 56-71.

Morkovsky L, Paces J, Ridl J, Reifova R (2015) Scrimer: designing primers from transcriptome data. *Molecular Ecology Resources*, **15**, 1415-1420.

Near TJ, Eytan RI, Dornburg A, Kuhn KL, Moore JA *et al.* (2012) Resolution of ray-finned fish phylogeny and timing of diversification. *Proceedings of the National Academy of Sciences*, **109**, 13698-13703.

O'Halloran DM (2015) PrimerView: high-throughput primer design and visualization. *Source Code for Biology and Medicine*, **10**, 8.

Pais FS, Ruy Pde C, Oliveira G, Coimbra RS (2014) Assessing the efficiency of multiple sequence alignment programs. *Algorithms for Molecular Biology*, **9**, 4.

Pereira-da-Conceicoa LL, Price BW, Barber-James HM, Barker NP, de Moor FC *et al.* (2012) Cryptic variation in an ecological indicator organism: mitochondrial and nuclear DNA sequence data confirm distinct lineages of *Baetis harrisoni* Barnard (Ephemeroptera: Baetidae) in southern Africa. *BMC Evolutionary Biology*, **12**, 26.

R Core Team (2016) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, available at: https://www.R-project.org (last acessed 26 March 2016).

Ramirez-Gonzalez RH, Uauy C, Caccamo M (2015) PolyMarker: A fast polyploid primer design pipeline. *Bioinformatics*, **31**, 2038-2039.

Ronquist F, Teslenko M, van der Mark P, Ayres DL, Darling A *et al.* (2012) MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Systematic Biology*, **61**, 539-542.

Ruane S, Bryson RW, Jr., Pyron RA, Burbrink FT (2014) Coalescent species delimitation in milksnakes (genus *Lampropeltis*) and impacts on phylogenetic comparative analyses. *Systematic Biology*, **63**, 231-250.

Rutschmann S, Gattolliat JL, Hughes SJ, Báez M, Sartori M *et al.* (2014) Evolution and island endemism of morphologically cryptic *Baetis* and *Cloeon* species (Ephemeroptera, Baetidae) on the Canary Islands and Madeira. *Freshwater Biology*, **59**, 2516-2527.

Sela I, Ashkenazy H, Katoh K, Pupko T (2015) GUIDANCE2: accurate detection of unreliable alignment regions accounting for the uncertainty of multiple parameters. *Nucleic Acids Research*, **43**, W7-14.

Sobhy H, Haitham S, Philippe C (2012) Gemi: PCR primers prediction from multiple alignments. *Comparative and Functional Genomics*, **2012**, 1-5.

Stamatakis A (2014) RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, **30**, 1312-1313.

Stephens M, Donnelly P (2003) A comparison of bayesian methods for haplotype reconstruction from population genotype data. *The American Journal of Human Genetics*, **73**, 1162-1169.

Stephens M, Smith NJ, Donnelly P (2001) A new statistical method for haplotype reconstruction from population data. *The American Journal of Human Genetics*, **68**, 978-989.

Szitenberg A, John M, Blaxter ML, Lunt DH (2015) ReproPhylo: An Environment for Reproducible Phylogenomics. *PLoS Computational Biology*, **11**, e1004447.

Talavera G, Castresana J (2007) Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Systematic Biology*, **56**, 564-577.

Untergasser A, Cutcutache I, Koressaar T *et al.* (2012) Primer3—new capabilities and interfaces. *Nucleic Acids Research*, **40**, e115.

470 Vuataz L, Sartori M, Gattolliat JL, Monaghan MT (2013) Endemism and diversification in
471 freshwater insects of Madagascar revealed by coalescent and phylogenetic analysis of
472 museum and field collections. *Molecular Phylogenetics and Evolution*, **66**, 979-991.
473 Vuataz L, Sartori M, Wagner A, Monaghan MT (2011) Toward a DNA taxonomy of Alpine
474 *Rhithrogena* (Ephemeroptera: Heptageniidae) using a mixed Yule-coalescent analysis of
475 mitochondrial and nuclear DNA. *PLoS ONE*, **6**, e19728.
476 Yoon H, Leitner T (2015) PrimerDesign-M: a multiple-alignment based multiple-primer
477 design tool for walking across variable genomes. *Bioinformatics*, **31**, 1472-1474.
478 You FM, Huo N, Gu YQ, Luo MC, Ma Y *et al.* (2008) BatchPrimer3: a high throughput web
479 application for PCR and sequencing primer design. *BMC Bioinformatics*, **9**, 253.
480 Yu L, Barakat E, Di Francesco J, Herzig HP (2015) Two-dimensional polymer grating and
481 prism on Bloch surface waves platform. *Optics Express*, **23**, 31640-31647.
482 Zeng L, Zhang Q, Sun R, Kong H, Zhang N *et al.* (2014) Resolution of deep angiosperm
483 phylogeny using conserved nuclear genes and estimates of early divergence times. *Nature*
484 *Communications*, **5**, 4956.

Data Accessibility

The program, user manual and example data sets are freely available at: https://github.com/hdetering/discomark (last accessed March 28, 2016). Scripts used for the analyses are available at: https://github.com/srutschmann/python_scripts (last accessed March 28, 2016). All DNA sequences from this study are available under GenBank accessions: KU987258-KU987260, KU987265-KU987268, KU987273-KU987276, KU987285-KU987288. GenBank accession numbers for sequences included in previous studies are the following: KU971838-KU971840, KU971851, KU972090-KU972092, KU972104, KU972490-KU972492, KU972503, KU973060-KU973061, KU973074. Sequence alignments and tree files are available at the Dryad repository (doi.org/10.5061/dryad.9sf96).

Author Contributions

S.R., H.D., and M.T.M. conceived the study. S.R. coordinated the project and performed the empirical work. S.R. and H.D. designed the program. H.D. implemented the program. S.R., H.D., and M.T.M wrote the manuscript. S.S. gave guidance for the ortholog prediction. J.F. provided the code of the PRIFI web tool. All authors provided comments and approved the final manuscript.

502 **Tables**

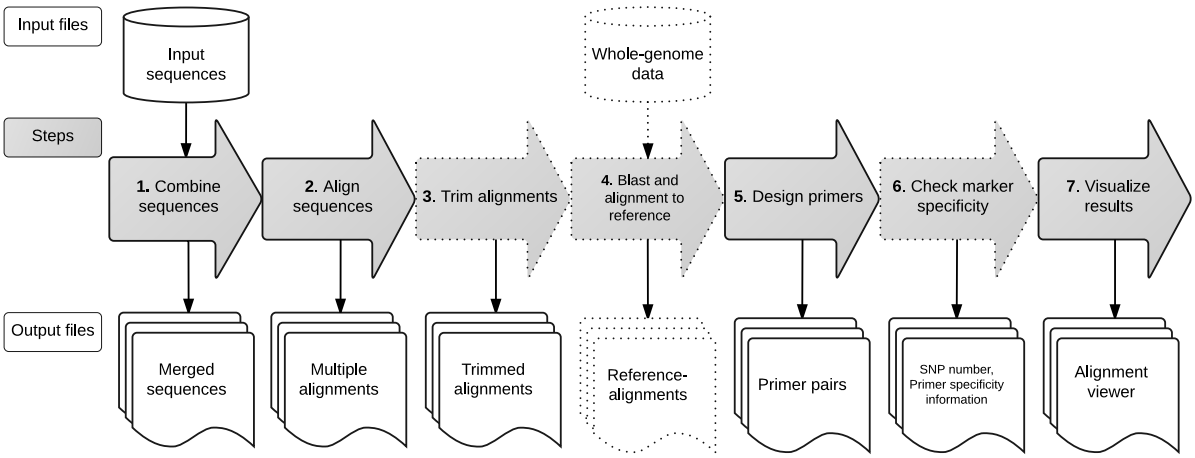
503 **Table 1** List of species used for the usage examples of the closely related species (*Cloeon*
504 *dipterum* s.l. species complex), and the distantly related species (insect order Ephemeroptera).
505 Voucher number, geographic origin, and GenBank accession numbers of cytochrome *c*
506 oxidase subunit 1 (*cox1*) gene are given for the specimens used for the marker amplification.
507 GenBank accession numbers of the genome data and the number of putative orthologs are
508 given for the species used to run DISCOMARK. For both data sets species used as reference are
509 indicated with *.

510

Species (Voucher, Origin)	**Family**	**Acc. Number**	**Orthologs**
Closely related species			
Baetis sp.	Baetidae	FN198828-FN203024	403
*Cloeon dipterum**	Baetidae	PRJNA268073	1,211
Cloeon dipterum sp1 (SR21B07, Switzerland)	Baetidae	KJ631626	-
Cloeon dipterum sp2 (SR21B06, Switzerland)	Baetidae	KJ631625	-
Cloeon dipterum sp3 (US, U.S.)	Baetidae	KU757184	-
Cloeon peregrinator (SR23A10, Madeira)	Baetidae	KU757122	-
Distantly related species			
Baetis sp.*	Baetidae	PRJNA219528	1,518
Ephemera danica	Ephemeridae	PRJNA219552	1,445
Eurylophella sp.	Ephemerellidae	PRJNA219556	1,523
Isonychia bicolor	Isonychiidae	PRJNA219568	1,457

511

512 **Figure legends**



513

514 **Fig. 1** Overview of the DISCOMARK workflow and processing steps. Arrows with a broken
515 outline indicate optional steps (for details see Materials and Methods section).

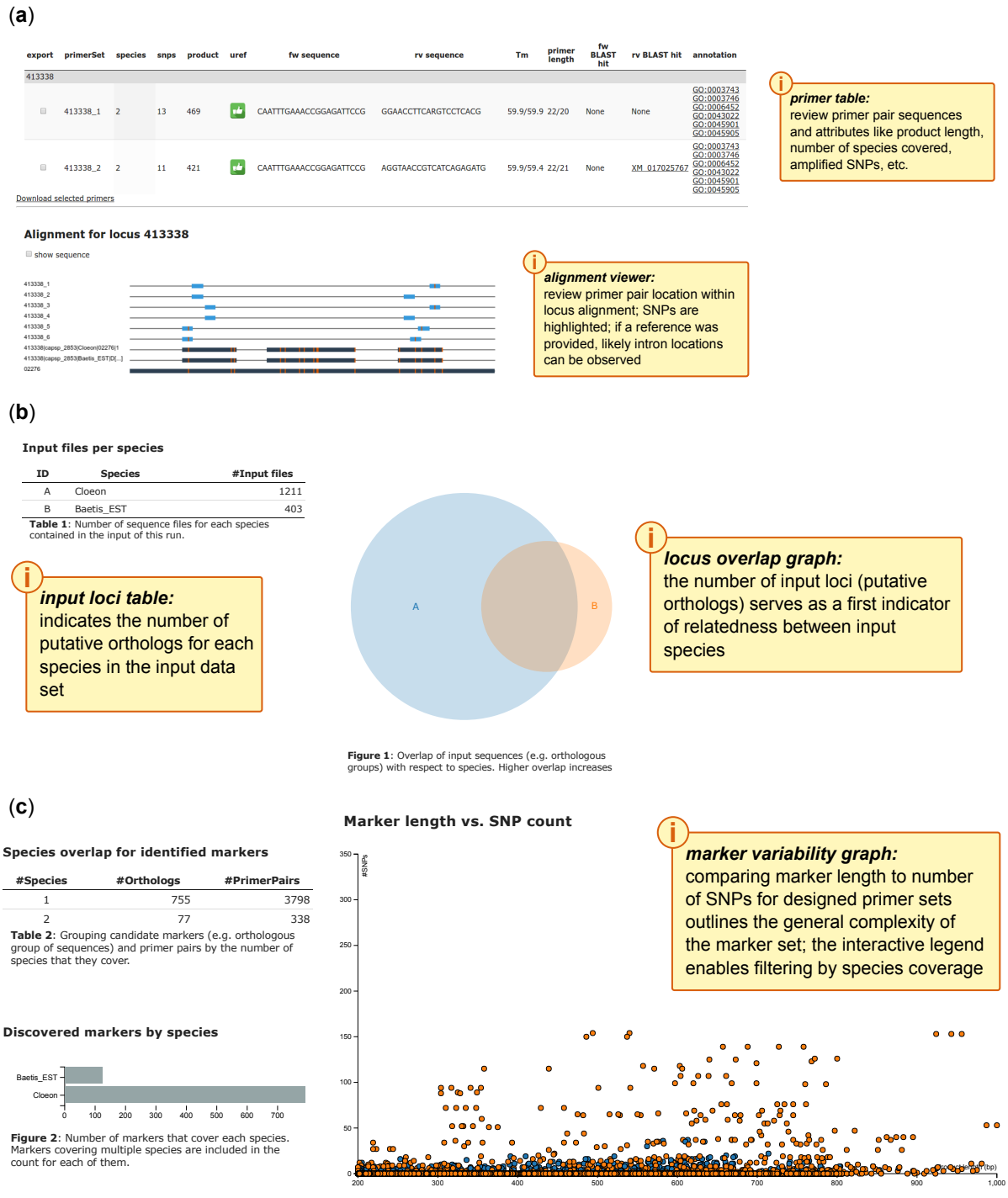


Fig. 2 Schematic visualization of DISCOMARK HTML output, (**a**), overview of designed primers and view of alignment, including suggested primer pairs, (**b**), overview of provided input files, including a graph with common putative orthologous sequences, (**c**) information about the output, including identified markers, discovered markers per species, and scatter plot displaying the number of single nucleotide polymorphisms (SNPs) versus product length for each primer pair.

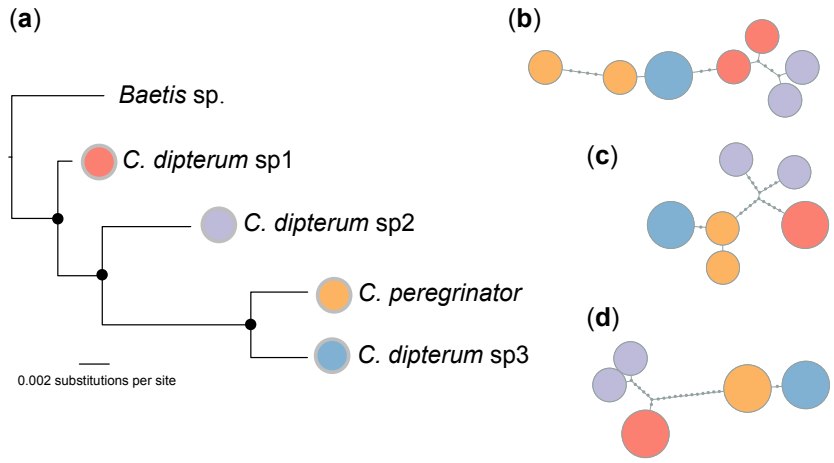


Fig. 3 Phylogenetic reconstruction and haplotype networks for the empirical data. (**a**), Phylogenetic reconstruction of four representatives of the species complex *Cloeon dipterum* s.l., including *C. peregrinator*, based on the exon sequences of the eight newly developed nuclear DNA markers (2,526 base pairs). Bayesian inference was used to reconstruct the tree based on the concatenated supermatrix alignment. Bayesian posterior probabilities ≥ 95% are indicated by filled circles. *Baetis* was used as an outgroup. Scale bar represents substitutions per site. (**b-d**), Haplotype networks of three amplified markers, (**b**), marker 412045, (**c**), marker 412741, (**d**), marker 412048 (full set of haplotype networks is available in Fig. S1, Supporting information). Circles are proportional to haplotype frequencies. Small circles along the branch indicate missing or unsampled haplotypes. Colors correspond to the four putative species.