DiscoMark: Nuclear marker discovery from orthologous sequences using low coverage genome data

Harald Detering*[1,2,3], Sereina Rutschmann*[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 77 markers. Among these, we randomly 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. Recently, tools have been developed for (1) specific primer design such as for automated primer design from transcriptome data (SCRIMER, Morkovsky *et al.* 2015), for individual degenerate primers (GEMI, Sobhy *et al.* 2012; PRIMER3, Untergasser *et al.* 2012; CEMASUITE, Lane *et al.* 2015), for highly variable DNA targets (PrimerDesign, Brodin *et al.* 2013; PrimerDesign-M, Yoon and Leitner 2015), viral genomes (PriSM, Yu *et al.* 2015), multiple primer design (BatchPrimer3, You *et al.* 2008; PrimerView, O'Halloran 2015) and (2) the discovery of specific markers, including single nucleotide polymorphism (SNP) markers (PolyMarker, Ramirez-Gonzalez *et al.* 2015), and putative single copy nuclear loci (MarkerMiner, Chamala *et al.* 2015). In addition, the challenge of developing new markers lies both in the discovery of conserved regions, the design of primer pairs and an estimation of their suitability as phylogenetic markers.

Our aim was to develop a flexible, user-friendly program that works with FASTA-formatted files of putative orthologous sequences from whole-genome or whole-transcriptome data, identified conserved regions and and designs primers based on these multiple sequence alignments. Here we present DiscoMark (=Discovery of Markers), a program for the discovery of phylogenetically suitable nDNA markers and design of primer pairs. The program can be used to easily screen for nDNA markers and design primers that can be used for Sanger sequencing as well as high-throughput sequencing. The program is structured into several steps that can be individually optimized by the user and run independently. In terms of input the program can be applied on large and small sets of taxa, including both closely and distantly related species. Ideally, orthologous sequences in combination with a whole-genome reference sequence are used. Thus, exon/intron boundaries can be inferred using the reference for each marker. Under the default settings, the program will design several primer pairs that anneal in conserved regions. The visualization of the alignments with potential primers allows the user to choose between primers targeting exons or introns (e.g. exon-primed intron-crossing (EPIC) markers). Additionally, information about the suitability as phylogenetic markers is provided by an estimate of the number of SNPs per marker and the applicability across species. Finally, we demonstrate the utility of DISCOMARK for (1) closely related species (i.e. *Cloeon dipterum* s.l. species complex) using whole-genome data, and (2) distantly related species (i.e. insect order Ephemeroptera) using whole-genome data derived from genome sequencing projects.

Materials and Methods

DISCOMARK implementation

The program DISCOMARK is written in Python and is developed to design primer pairs in conserved regions of predicted orthologous genes. Orthologs are most suited for phylogenetic studies. The ortholog identification step is not part of the DISCOMARK workflow but DISCOMARK is designed to directly work with the output of several ortholog prediction programs, e.g. HAMSTR (Ebersberger *et al.* 2009), or Orthograph (https://github.com/mptrsen/Orthograph, last accessed March 25, 2016). Orthologous groups may be derived from genomic or transcriptomic sequencing data. In addition to the orthologous genes, genomic data such as whole-genome sequencing data can be provided to DISCOMARK as a guide to detect exon/intron boundaries. DISCOMARK performs seven steps, combining Python scripts with widely used bioinformatics programs (Fig. 1). The steps: (1) combine orthologous groups of sequences, (2) align sequences of each orthologous group using MAFFT v.7.205 (Katoh and Standley 2013), (3) trim sequence alignments with TRIMAL v.1.4 (Capella-Gutierrez *et al.* 2009), (4) align sequences against a reference (e.g. whole-genome dataset from the same or closely related taxa) with BLASTN v.2.2.29 (Altschul *et al.* 1997; Camacho *et al.* 2009) and re-alignment using MAFFT, (5) design primer pairs on single-gene alignments using a modified version of PRIFI (Fredslund *et al.* 2005), adapted by us into a Python package that uses BioPython v.1.65 and Python v.3.4.3, (6) check primer specificity with BLASTN, and (7) generate output in several formats (visual HTML report, tabular data and FASTA files of the primers). The results of each step can be inspected in the respective output folders.

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

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

3 Trim alignments. In order to remove poorly aligned regions, sequence alignments are trimmed using TRIMAL. The program TRIMAL analyzes the distribution of gaps and mismatches in the alignment and discard alignment positions and sequences of low quality. By default, DISCOMARK calls TRIMAL with the '-strictplus' method. The preset is used by TRIMAL to derive the specific thresholds for alignment trimming (minimum gap score, minimum residue similarity score, conserved block size). Since alignment trimming largely depends on the input data and influences the downstream results, TRIMAL can also be run with different settings (e.g. '-gappyout', '-strict', '-automated1'; but see Capella-Gutierrez *et al.* (2009). Alternatively, there is also the option to deactivate the alignment trimming with the DISCOMARK option '--no-trim' or use alternative trimming programs such as GBLOCKS (Castresana 2000; Talavera and Castresana 2007) or GUIDANCE2 (Landan and Graur 2008; Sela *et al.* 2015).

4 Blast and alignment to reference. In this step a genomic reference sequence for each input ortholog is identified and added to the trimmed alignment. This step is particularly important when working with coding sequences which do not contain intron sequences; thus, a genomic sequence is needed to infer intron/exon boundaries. Working with coding sequences is advisable for more distantly related taxa which may include intron length polymorphisms, or to target EPIC markers. Any whole-genome data set (from one of the included taxa or a closely related taxa) can be used as reference for mapping the ortholog sequences. Here, mapping means that the input sequences are compared to the reference sequences, which are defined by the user using the local alignment program BLASTN. The best locally aligning reference sequence (the one that yields the longest alignment among all input sequences) for each orthologous group is added to the corresponding sequence alignment. Reference sequences are cut to 100 base pairs (bp) upstream and downstream of the first, respectively last, BLAST hit to avoid alignment length inflation. Then, the extended alignments are re-aligned with MAFFT. The reference alignment step is optional; however, the inclusion of whole-genome data is essential for estimating intron/exon boundaries. Given that information, the focus of target sequences to be amplified can be on entire exon markers, EPIC markers, or a combination.

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

6 Check marker specificity. To ensure the specificity of the designed primer pairs, we compare their sequences against the NCBI database ('refseq_mrna'). Primer sequences are searched in the NCBI database ('refseq_mrna') using the online BLASTN interface. The default search settings are restricted to human and bacterial targets using the Entrez query 'txid2[ORGN] OR txid9606[ORGN]' because these are most likely to be present as contaminants in sequencing libraries. The result hits of the BLAST search are indicated to the user in the HTML output.

7 Visualize results. As final step, the program produces a HTML report containing the list of designed primers, an alignment viewer and plots visualizing the discovered set of markers. Besides the primer sequences the report lists several features such as the melting temperatures, predicted sequence length, and the number of taxa amplified by each primer set. Selected primer pairs and primer lists can be downloaded as FASTA or CSV files, respectively. In order to provide a measure of the suitability of the markers for phylogenetic reconstruction the program calculates the number of SNPs between a primer pair by comparing the aligned input sequences against each other. The number of SNPs between each primer pair is visualized in relation to the estimated product length (see Fig. 2 for an example) and reported in the tabular output. Furthermore, the report highlights the species coverage achieved by the discovered markers, i.e. how many species' sequences each primer set is expected to amplify, as an estimate of how universal each primer set can be applied. Additionally, functional annotations are reported, if available, to guide the user in the selection of markers of interest. Annotations can be supplied in form of a tab-delimited file with the '-a' option. In principle, any kind of annotations can be used depending on the desired research objective. In our usage scenarios, we used gene ontology (GO) terms which were retrieved by mapping the gene IDs contained in the HAMSTR core ortholog set via the UniProt website (http://www.uniprot.org/, last accessed December 20, 2015).

Usage cases

Closely related species - Cloeon dipterum s.l. species complex. To test the suitability of DISCOMARK for closely related species, we designed primer pairs for the species complex of *Cloeon dipterum* s.l. (Ephemeroptera: Baetidae). The species complex consists of several closely related species, including *Cloeon peregrinator* GATTOLLIAT & SARTORI, 2008 from Madeira (Gattolliat *et al.* 2008; Rutschmann *et al.* 2014; Table 1). As input to design the primer pairs data we used whole-genome sequencing data of *Cloeon dipterum* L. 1761 (Baetidae; Sequence Read Archive SRP050093) and expressed sequence tags (EST) of *Baetis* sp. (Baetidae; FN198828-FN203024). The sequence reads of *C. dipterum* were trimmed and *de novo* assembled using NEWBLER v.2.5.3 (454 Life Science Corporation) under the default settings for large datasets. Ortholog sequences prediction of both data sets was performed with HAMSTR v.9 using the insecta_hmmer3-2 core reference taxa set (http://www.deep-phylogeny.org/hamstr/download/datasets/hmmer3/insecta_hmmer3-2.tar.gz, last accessed December 20, 2015), including 1,579 orthologous genes. We ran the program DISCOMARK with default settings ('python run_project.py -i input/Cloeon -i input/Baetis -r input/reference/Cloeon.fa -a input/co2go.ixosc.csv -d output/cloeon_baetis'), using the predicted orthologs from HAMSTR and the whole-genome *Cloeon*-data as reference (step 4). The Pearson correlation between the number of SNPs between primer pairs and corresponding estimated product length was calculated using the function cor within the stats package for R (R Development Core Team, 2016). A t-test for significance was performed using the function cor.test.

From the total of designed primer pairs (77 markers, 338 primer pairs, see results) we selected eight and amplified them for four species of the *C. dipterum* species complex (Table 1) in the laboratory. We used standardized polymerase chain reactions (PCR; 35-40 PCR cycles with annealing temperature of 55°C), followed by Sanger sequencing. Forward and reverse sequences were assembled and edited with GENEIOUS R7 v.7.1.3 (Biomatters Ltd.), indicating ambiguous positions following the IUPAC nucleotide codes. Heterozygous sequences were decoded with CODONCODEALIGNER v.3.5.6 (CodonCode Corporation) using the find and split heterozygous function. Multiple sequence alignments were created for all sequences per marker. The predicted orthologous sequences of *Baetis* sp. were used as reference to infer the exon-intron splicing boundaries (canonical and non-canonical splice site pairs). The final sequence alignments were checked for the occurrence of stop codons and indels, and split into exon and intron parts using a custom Python script (https://github.com/srutschmann/python_scripts, last accessed March 28, 2016). Sequence alignments were phased using the program PHASE v.2.1.1 (Stephens *et al.* 2001; Stephens and Donnelly 2003) with a cutoff value of 0.6 (Harrigan *et al.* 2008; Garrick *et al.* 2010), whereby input and output files were formatted using the Perl scripts included in SeqPHASE (Flot 2010). Heterozygous sites that could not be resolved were coded as ambiguity codes for subsequent analyses. After phasing, all alignments were re-aligned with MAFFT. The number of variable and informative sites, and the nucleotide diversity per exon alignment was calculated with a custom script.

To investigate the heterogeneity of each marker´s DNA sequences, we reconstructed haplotype networks, using Fitchi (Matschiner 2015). As input for each marker we inferred a gene tree using the program RAxML v.8 (Stamatakis 2014) with the GTRCAT model and 1,000 bootstrap replicates under the rapid bootstrap algorithm. The phylogenetic relationships were calculated with Bayesian inference, using MrBayes v.3.2.3 (Ronquist *et al.* 2012) based on a concatenated nDNA matrix that consisted of the exon sequences from all 15 nDNA markers. The best-fitting model of molecular evolution for each sequence alignment was selected via a BIC criterion in jModelTest v.2.1 (Guindon and Gascuel 2003; Darriba *et al.* 2012). We calculated 10^6 generations with random seed, a burn-in of 25% and four MCMC chains. As an outgroup we used the predicted orthologous sequences of *Baetis* sp..

Distantly related species - insect order Ephemeroptera. In this test case, we used contigs derived from whole-genome sequencing projects of the species *Baetis* sp. (Baetidae; BioProject PRJNA219528), *Ephemera danica* MÜLLER 1764 (Ephemeridae; BioProject PRJNA219552), *Eurylophella* sp. (Ephemerellidae; BioProject PRJNA219556), and *Isonychia bicolor* WALKER 1853 (Isonychiidae; BioProject PRJNA219568). The contigs from each species were used for ortholog predicting with HAMSTR v.13.2.4 (http://sourceforge.net/projects/hamstr/files/hamstr.v13.2.4.tar.gz, last accessed December 20, 2015). We ran DISCOMARK with the default settings, using the *Baetis* sp. data as reference ('python run_project.py -i input/Baetis -i input/Ephemera -i input/Eurylophella -i input/Isonychia -r input/references/Baetis.fa -a input/co2go.ixosc.csv -d output/mayflies').

Results

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

DISCOMARK identified a total of 804 nDNA markers and 77 alignments with 338 primer pairs for orthologous sequences of both species (*Baetis* sp. and *C. dipterum* s.l.). Ortholog prediction yielded 403 orthologous sequences for the *Baetis* sp. EST-data and 1,211 for *C. dipterum*. For the individual species, DISCOMARK identified 790 markers for *C. dipterum* and 123 for *Baetis* sp. The lengths of the markers including both species were between 201 and 925 bp with median length of 451.5 bp. The number of SNPs per marker ranged from zero to 37 (median: 5) with an average of one SNP per 68 bp. Marker length and number of SNPs were correlated with a Pearson's correlation coefficient of 0.35 (Pearson's product-moment correlation $P < 0.001$). The total run time for this data set on a local Linux machine (quad-core Intel i5, 8 GB RAM) was 24 min.

The haplotype networks based on the eight selected markers showed a clear structure for all markers, including two markers with shared haplotypes for the two species from the U.S. and Madeira (Fig. 3 and Fig. S1, Supporting information). The length of the concatenated sequence alignment of the eight markers was 3,530 bp (2,526 bp exon sequence, Table S1, Supporting information). The exon sequence matrix contained 78 variable sites, 27 informative sites, and was 92.6% complete. The nucleotide diversity ranged between 0.009 and 0.028 (median: 0.013). Phylogenetic tree reconstruction based on these eight markers resulted in a phylogeny with fully resolved nodes (Bayesian posterior probability (PP) ≥ 95%; Fig. 3). The species *C. dipterum* sp1 was found as outgroup to a clade containing the species *C. dipterum* sp2 from Switzerland and the two species from the U.S and Madeira. The latter two species formed a monophyletic clade.

Distantly related species - insect order Ephemeroptera

In total, we found 22 orthologs with a total of 48 primer pairs for all four species (Table S2, Supporting information). The input files per species (i.e. putative orthologous sequences) ranged from 1,445 to 1,523. We detected 41 markers that covered three of the species (99 primer pairs), 81 markers covering two species (210 primer pairs), and 117 markers that covered any single species (478 primer pairs). For the individual species, *Baetis* sp. had the most markers available (214) of the single- and multi-species markers. There were 138 markers for *Eurylophella* sp., 107 markers for *I. bicolor*, and 88 markers for *E. danica*. The lengths for all markers covering all four species varied between 216 and 997 bp with median of 398.5 bp, containing between 39 and 298 SNPs per marker (Fig. 2,) with a SNP every 4.1 bp on average. Marker length and number of SNPs were correlated with a Pearson's correlation coefficient of 0.97 (Pearson's product-moment correlation $P < 0.001$). Run time for this data set on a Linux client (quad-core Intel i5, 8 GB RAM) was 46 min.

Discussion

To our knowledge, the program DiscoMark is the first stand-alone program with the aim of designing primer pairs based on multiple sequence alignments on a genome-wide scale. The visual output gives guidance on the suitability of each marker (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 scale' (i.e. for closely related species many markers with intermediate number of SNPs and for distantly related species fewer markers with generally higher number of SNPs 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.

Markers development within the order Ephemeroptera

The usage of DiscoMark adds an extensive set of new potential nDNA markers to the ones that have been used to date for mayfly phylogenies based on 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 phylogenetic reconstructions are still mostly based on the information of mitochondrial DNA markers (e.g. Rutschmann *et al.* 2014; Macher *et al.* 2016). The availability of more genome data will be very valuable in order to increase the number of markers suitable for phylogenetic studies. The use of the larger marker set for *C. dipterum* developed here resulted in a fully resolved phylogenetic tree in contrast to Rutschmann *et al.* (2014). The availability of more markers promote fine-scaled phylogenetic studies, which are needed to resolve the phylogenetic relationships of so-called morphologically cryptic species that can not be resolved with standard markers (Dijkstra *et al.* 2014).

Acknowledgements

We are thankful to our research groups, in particular the Phylogenomics Lab at the University of Vigo for constructive discussion that improved this project. Research was partially supported by the Leibniz Association (PAKT für Forschung und Innovation "FREDIE" project) and the Swiss National Science Foundation (Early PostDoc.Mobility grant P2SKP3_15869 to S.R.). This is publication number ### of the Berlin Center for Genomics in Biodiversity Research.

References

Altschul SF, Madden TL, Schaffer AA *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, **25**, 3389-3402.

Baird NA, Etter PD, Atwood TS *et al.* (2008) Rapid SNP Discovery and Genetic Mapping Using Sequenced RAD Markers. *PLoS ONE*, **3**, e3376.

Brodin J, Krishnamoorthy M, Athreya G *et al.* (2013) A multiple-alignment based primer design algorithm for genetically highly variable DNA targets. *BMC Bioinformatics*, **14**, 255.

Camacho C, Coulouris G, Avagyan V *et al.* (2009) BLAST+: architecture and applications. *BMC Bioinformatics*, **10**, 421.

Capella-Gutierrez S, Silla-Martinez JM, Gabaldon T (2009) trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics*, **25**, 1972-1973.

Castresana J (2000) Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Molecular Biology and Evolution*, **17**, 540-552.

Chamala S, García N, Godden GT *et al.* (2015) MarkerMiner 1.0: new application for phylogenetic marker development using angiosperm transcriptomes. *Applications in Plant 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.

Macher JN, Salis RK, Blakemore KS, Tollrian R, Matthaei CD *et al.* (2016) Multiple- stressor effects on stream invertebrates: DNA barcoding reveals contrasting responses of cryptic mayfly species. *Ecological Indicators*, **61**, 159-169.

Matschiner M (2015) Fitchi: haplotype genealogy graphs based on the Fitch algorithm. *Bioinformatic*s, doi:10.1093/biooinformatics/btv717.

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.

Vuataz L, Sartori M, Gattolliat JL, Monaghan MT (2013) Endemism and diversification in freshwater insects of Madagascar revealed by coalescent and phylogenetic analysis of museum and field collections. *Molecular Phylogenetics and Evolution*, **66**, 979-991.

Vuataz L, Sartori M, Wagner A, Monaghan MT (2011) Toward a DNA taxonomy of Alpine *Rhithrogena* (Ephemeroptera: Heptageniidae) using a mixed Yule-coalescent analysis of mitochondrial and nuclear DNA. *PLoS ONE*, **6**, e19728.

Yoon H, Leitner T (2015) PrimerDesign-M: a multiple-alignment based multiple-primer design tool for walking across variable genomes. *Bioinformatics*, **31**, 1472-1474.

You FM, Huo N, Gu YQ, Luo MC, Ma Y *et al.* (2008) BatchPrimer3: a high throughput web application for PCR and sequencing primer design. *BMC Bioinformatics*, **9**, 253.

Yu L, Barakat E, Di Francesco J, Herzig HP (2015) Two-dimensional polymer grating and prism on Bloch surface waves platform. *Optics Express*, **23**, 31640-31647.

Zeng L, Zhang Q, Sun R, Kong H, Zhang N *et al.* (2014) Resolution of deep angiosperm phylogeny using conserved nuclear genes and estimates of early divergence times. *Nature 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.

Author Contributions

S.R., H.D., S.S., and M.T.M. conceived the study. S.R. coordinated the project and performed the empirical analyses. H.D. implemented the program in Python. S.R. and H.D. drafted the manuscript. S.S. gave guidance for the ortholog prediction. J.F. provided the code of the PriFi web tool. All authors gave helpful comments to the manuscript and approved the final version.

Tables

Table 1 List of species used for the usage examples of the closely related species; *Cloeon dipterum* s.l. species complex.

Species	**Voucher**	**Acc. no. *cox1***	**Geographical origin**
Cloeon dipterum sp1	SR21B07	KJ631626	Switzerland
Cloeon dipterum sp2	SR21B06	KJ631625	Switzerland
Cloeon dipterum sp3	US	KU757184	U.S.
Cloeon peregrinator	SR23A10	KU757122	Madeira

Figures

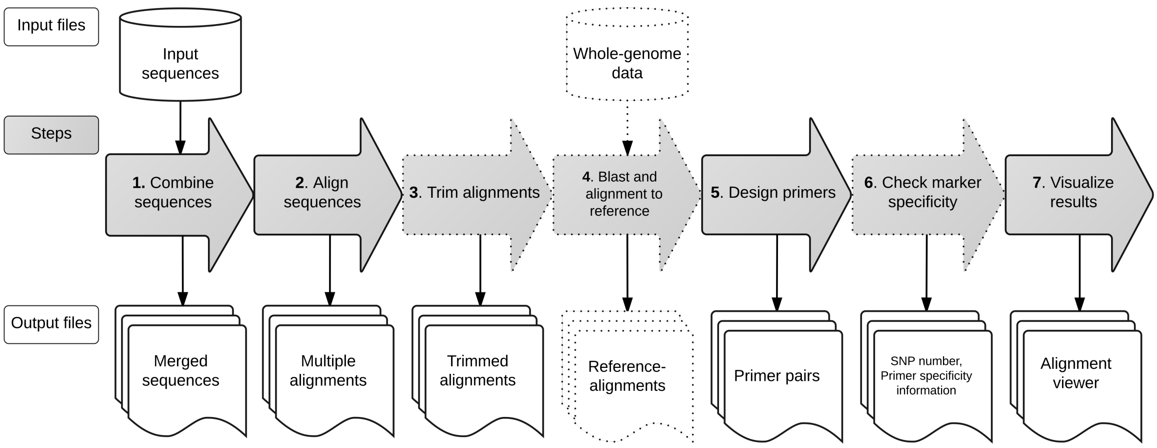


Fig. 1 Overview of the DiscoMark workflow and processing steps. Arrows with a broken outline indicate optional steps (for details see Materials and Methods section).

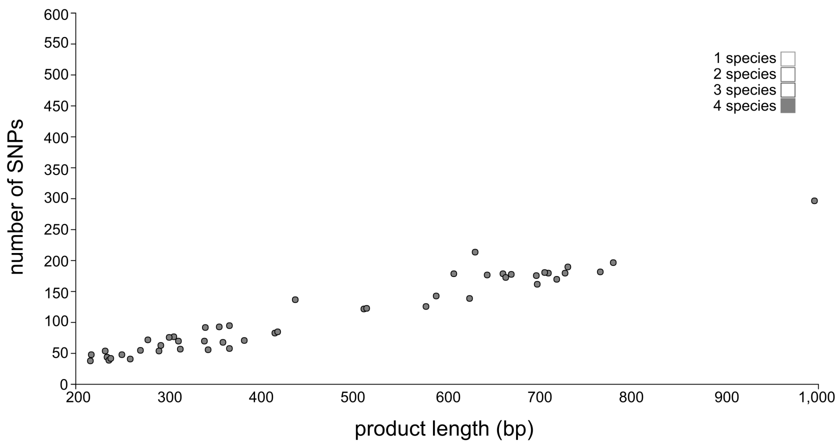
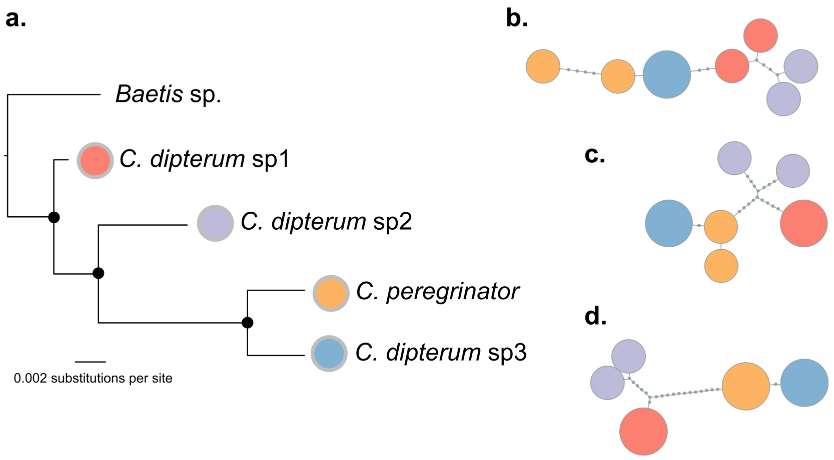


Fig. 2 Visualization of DiscoMark results: Scatter plot displaying the number of single nucleotide polymorphisms (SNPs) versus product length for each marker of the four mayfly species: *Baetis* sp., *Eurylophella* sp., *Ephemera danica*, and *Isonychia bicolor*. Shown are the markers for all four species (for details see Materials and Methods section).



495

496 **Fig. 3** Phylogenetic reconstruction and haplotype networks for the empirical data. **a**,
497 Phylogenetic reconstruction of four representatives of the species complex *Cloeon dipterum*
498 s.l., including *C. peregrinator*, based on the exon sequences of the eight newly developed
499 nuclear DNA markers (2,526 base pairs). Bayesian inference was used to reconstruct the tree
500 based on the concatenated supermatrix alignment. Bayesian posterior probabilities ≥ 95% are
501 indicated by filled circles. *Baetis* was used as an outgroup. Scale bar represents substitutions
502 per site. **b-d**, Haplotype networks of three amplified markers, **b**, marker 412045, **c**, marker
503 412741, **d**, marker 412048 (full set of haplotype networks is available in Fig. S1, Supporting
504 information). Circles are proportional to haplotype frequencies. Small circles along the branch
505 indicate missing or unsampled haplotypes. Colors correspond to the four putative species.