

SPECIAL ISSUE: SEQUENCE CAPTURE

Target gene enrichment in the cyclophyllidean cestodes, the most diverse group of tapeworms

HAO YUAN,* JIAMEI JIANG,* FRANCISCO AGUSTÍN JIMÉNEZ,† ERIC P. HOBERG,‡
JOSEPH A. COOK,§ KURT E. GALBREATH¶ and CHENHONG LI*

*Key Laboratory of Exploration and Utilization of Aquatic Genetic Resources, Ministry of Education, Shanghai Ocean University, Shanghai 201306, China, †Department of Zoology, Southern Illinois University, Carbondale IL 62901, USA, ‡US National Parasite Collection and Animal Parasitic Disease Laboratory, USDA Agricultural Research Service, BARC East No. 1180, 10300 Baltimore Avenue, Beltsville MD 20705, USA, §Museum of Southwestern Biology and Department of Biology, University of New Mexico, Albuquerque NM 87131, USA, ¶Department of Biology, Northern Michigan University, Marquette MI 49855, USA

Abstract

The Cyclophyllidea is the most diverse order of tapeworms, encompassing species that infect all classes of terrestrial tetrapods including humans and domesticated animals. Available phylogenetic reconstructions based either on morphology or molecular data lack the resolution to allow scientists to either propose a solid taxonomy or infer evolutionary associations. Molecular markers available for the Cyclophyllidea mostly include ribosomal DNA and mitochondrial loci. In this study, we identified 3641 single-copy nuclear coding loci by comparing the genomes of *Hymenolepis microstoma*, *Echinococcus granulosus* and *Taenia solium*. We designed RNA baits based on the sequence of *H. microstoma*, and applied target enrichment and Illumina sequencing to test the utility of those baits to recover loci useful for phylogenetic analyses. We captured DNA from five species of tapeworms representing two families of cyclophyllideans. We obtained an average of 3284 (90%) of the targets from the test samples and then used captured sequences (2 181 361 bp in total; fragment size ranging from 301 to 6969 bp) to reconstruct a phylogeny for the five test species plus the three species for which genomic data are available. The results were consistent with the current consensus regarding cyclophyllidean relationships. To assess the potential for our method to yield informative genetic variation at intraspecific scales, we extracted 14 074 single nucleotide polymorphisms (SNPs) from alignments of four *Arostrilepis macrocirrosa* and two *A. cooki* and successfully inferred their relationships. The results showed that our target gene tools yield data sets that provide robust inferences at a range of taxonomic scales in the Cyclophyllidea.

Keywords: cestodes, gene capture, next-generation sequencing, nuclear markers, phylogenetics, phylogeography

Received 7 November 2015; revision received 2 February 2016; accepted 17 March 2016

Introduction

The Order Cyclophyllidea encompasses more than 3100 species of tapeworms, which are classified into 15 families, including the economically important Anoplocephalidae, Hymenolepididae and Taeniidae (Schmidt 1986; Khalil *et al.* 1994; Georgiev 2003). Some of the species contained in the latter family, such as *Taenia solium* (pork tapeworm), *T. saginata* (beef tapeworm) and *Echinococcus* spp. are etiological agents of debilitating zoonotic infectious diseases that affect people across the globe (White 2000; Gottstein *et al.* 2015). Significantly, *T. solium* is identified as the most important food-borne

parasite of humans throughout the world (Robertson *et al.* 2013). Furthermore, some species of *Hymenolepis* are known to infect humans (Bogitsh *et al.* 2013; Muehlenbachs *et al.* 2015), whereas members of Anoplocephalidae include important parasites of domestic ungulate stock (Dorchies 1999). The cyclophyllideans are found in numerous vertebrate species, primarily among birds and mammals but also in amphibians and reptiles, and the rate of new species discovery remains high (Makarikov *et al.* 2011, 2013), indicating that our current understanding of their diversity is incomplete (Georgiev 2003).

Phylogenetic inference among the Cyclophyllidea remains challenging despite advances over the past 15 years exploring comparative morphology and total evidence from integrated morphological and molecular data (e.g. Hoberg *et al.* 1999, 2001). Molecular data are

Correspondence: Kurt Galbreath, Fax: +1 906 2271063; E-mail: kgalbrea@nmu.edu and Chenhong Li, Fax: +86 21 61900263; E-mail: chli@shou.edu.cn

now widely used in studying phylogenetic relationships among cyclophyllideans (e.g. von Nickisch-Rosenegk *et al.* 1999; Littlewood *et al.* 2008; Nakao *et al.* 2013; Haukisalmi *et al.* 2014), as well as species identification (Ramnath *et al.* 2014; Terefe *et al.* 2014), new species description (Greiman & Tkach 2012; Greiman *et al.* 2013; Makarikov *et al.* 2013) and historical biogeography (Wickström *et al.* 2003; Galbreath *et al.* 2013; Yanagida *et al.* 2014). Most molecular markers used in the cyclophyllideans include ribosomal DNA (rDNA) (Haukisalmi *et al.* 2010; Greiman *et al.* 2013; Makarikov *et al.* 2013; Ramnath *et al.* 2014) and mitochondrial genes (von Nickisch-Rosenegk *et al.* 1999; Wickström *et al.* 2003; Haukisalmi *et al.* 2014). These common markers are widely used in molecular systematics at deep taxonomic scales within the Cestoda (Mariaux 1998; Olson *et al.* 2001; Hardman & Hardman 2006; Littlewood *et al.* 2008; Waeschenbach *et al.* 2012), but they have important limitations. For example, ribosomal DNA may evolve too slowly to resolve recently diverged lineages (Littlewood *et al.* 2008). To increase phylogenetic resolution at fine taxonomic scales some investigators have sequenced long stretches of mitochondrial DNA (mtDNA, Waeschenbach *et al.* 2012). Indeed, the whole mitochondrial genome has been interrogated to find the best strategy of character sampling for inferring phylogenetic relationships in cestodes (Hardman & Hardman 2006; Littlewood *et al.* 2008). However, because mtDNA is nonrecombining, all mtDNA loci are linked and provide only a single perspective on evolutionary history that may not represent the true species history due to stochastic error associated with genealogical processes (Brito & Edwards 2009). Thus, the limited number of independent loci available to cestode systematists has precluded resolution of both deep and shallow nodes in phylogenies (Philippe *et al.* 1994; Mossel & Steel 2005). Lack of independent loci also limits accurate estimation of phylogeographic and demographic parameters (Lee & Edwards 2008).

Recent development in target gene enrichment and next-generation sequencing technology made capturing and sequencing thousands of loci tractable for phylogenetic studies in nonmodel organisms (Faircloth *et al.* 2012; Lemmon *et al.* 2012; McCormack *et al.* 2012; Li *et al.* 2013). The target genes could be ultraconserved elements (UCE) or conserved coding sequences across a wide range of species, and target genes could be enriched using baits designed on a distantly related species; therefore, genome sequence from the species of interest is not necessary for designing the baits. Those approaches have been successfully adopted in studying a variety of non-model vertebrate organisms, including mammals, birds, snakes, turtles and sharks (Crawford *et al.* 2012; Faircloth *et al.* 2012; Lemmon *et al.* 2012; McCormack *et al.* 2012; Li

et al. 2013). In addition, baits can be designed from single-copy nuclear coding sequences. For example, baits designed on exons from the elephant shark (*Callorhynchus milii*) were used to enrich target sequences from species of 13 different orders of chondrichthyans (Li *et al.* 2013). Or multiple versions of each probe spanning the known diversity were applied to capture 1552 exon targets of the Ophiuroidea (Hugall *et al.* 2016). Although target gene enrichment and the next-generation sequencing approach could be used to produce a wealth of useful data more efficiently than employing the traditional PCR and Sanger sequencing method, no attempts have been made to try these protocols on cestodes.

In this study, we identified 3641 single-copy nuclear coding loci comparing the genomes of *Hymenolepis microstoma*, *Echinococcus granulosus* and *T. solium* (Tsai *et al.* 2013). We designed RNA baits based on the sequence of *H. microstoma*, and tested the utility of those markers for phylogenetic inference with five species of cyclophyllidean cestodes representing the genera *Schizorchis* (Anoplocephalidae) and *Arostrilepis* (Hymenolepididae). We also extracted 14 074 single nucleotide polymorphism sites (SNPs) from alignments of six individual worms, four *A. macrocirrosa* and two *A. cooki*, and inferred their relationships. Our goals were to develop a set of genetic markers and to optimize data analysis pipelines that can be used to enrich and sequence thousands of single-copy nuclear loci for studying phylogenetic and phylogeographic questions in the cestodes.

Materials and methods

Samples and DNA extraction

Nine individuals from five species of cyclophyllideans were sampled, including two species of *Schizorchis* and three species of *Arostrilepis*. *Schizorchis* is a genus of anoplocephalid tapeworms that appears to be exclusively associated with pikas (genus *Ochotona*), small lagomorphs that occur in North America and Asia (Rausch & Smirnova 1984). The diversity of *Schizorchis* remains incompletely understood (Tinnin *et al.* 2008; Galbreath & Hoberg 2012). In contrast, the genus *Arostrilepis* is a group of hymenolepidid tapeworms that is distributed widely across the Holarctic in association with numerous rodent genera, including voles (*Myodes*, *Microtus*, *Arvicola*), lemmings (*Lemmus*, *Synaptomys*), mice (*Peromyscus*) and pocket gophers (*Thomomys*). This cestode genus was once thought to include only a single species, *Hymenolepis horrida* (now *A. horrida*) (Rausch 1952), but now is understood to represent a complex of at least twelve independent nominal species (Makarikov *et al.* 2011, 2013). The complex is further represented by several additional genetic lineages that

are currently recognized but not described (Hoberg *et al.* 2012; K.E. Galbreath, A.A. Makarikov & E.P. Hoberg, unpubl. data). To date, however, phylogenetic resolution of these two species complexes has been limited due to a lack of informative molecular markers (Galbreath & Hoberg 2012; Galbreath *et al.* 2013; Makarikov *et al.* 2013). Thus, these species groups represent excellent systems on which to test our new method of identifying and sequencing numerous nuclear loci. Because we are also interested in applying our methods at shallow phylogeographic scales in cestodes, we also sampled two individuals of *A. cooki* within continental North America and four individuals of *A. macrocirrosa* collected from both sides of the Bering Strait to evaluate intraspecific genetic variation (Table 1; Fig. 1).

Specimens for this study were obtained from the Beringian Coevolution Project, an international collaboration that is sampling and archiving mammals and their parasites from across the Holarctic region (Cook *et al.* 2005; Hoberg *et al.* 2012). All cestode samples were preserved in the field in 70% ethanol and archived at -20°C . Genomic DNA was extracted from tissue samples consisting of 5–10 proglottids using a Qiagen DNeasy Blood and Tissue kit (Qiagen, Shanghai, China). The concentration of the DNA extraction was measured using NanoDrop 3300 Fluorospectrometer (Thermo Fisher Scientific, Waltham, MA, USA).

Target gene markers, baits design and synthesis

Single-copy conserved coding sequences were selected as our gene targets by comparing the genome sequences of *E. granulosus*, *T. solium* and *H. microstoma* (Tsai *et al.* 2013) using a bioinformatics tool, EvolMarkers (Li *et al.* 2012). *Hymenolepis microstoma* was used as the query and *E. granulosus* and *T. solium* were used as the subjects in the comparison. The default settings of EvolMarkers were used except that the minimum length of each target locus was set to 301 bp. The selected target sequences were single-copy and could be found in all three genomes. The fasta files of the target sequences of the three species can be found in Appendix S1 (Supporting information).

Baits of 120 bp were designed on the target sequences of *H. microstoma* with $3\times$ tiling (i.e. 80 bp overlapping between adjacent baits). Simple repeats or low complexity regions, which might compromise capture efficacy, were masked through RepeatMasker.org. There were 39 009 full repeat-masked baits. The design and synthesis of biotinylated RNA baits was completed at MYcroarray [Ann Arbor, MI, USA, catalogue number: 140916-Tapeworm, (39 009 probes, 120mer)].

Library preparation, target enrichment and Illumina sequencing

Total amount of DNA used for library preparation ranged from 300 to 1400 ng for each sample (Table 1). The DNA was sheared to ~ 250 bp using a Covaris M220 Focused-ultrasonicatorTM (Covaris, Inc., MA, USA). Blunt-end repair, adapter ligation, fill-in and prehybridization PCR followed the protocol of Li *et al.* (2013).

The hybridization steps followed MYbaits-manual-v3 (MYcroarray; Ann Arbor) with some modifications. In the hybridization mix, instead of adding 5.5 μL of RNA baits, 1 μL of the baits and 4.5 μL nuclease-free water was added in our experiment to reduce costs. We have previously shown this modification to be equally effective to the original recipe. Similarly Hugall *et al.* (2016) also diluted their baits in gene-capture study. Each sample was enriched twice to increase the percentage of on-target reads. That is, the product from the first enrichment experiment was used as the library for a second enrichment as suggested by Li *et al.* (2013). After the second enrichment, each sample was amplified with indexed primers so that all samples could be pooled together for sequencing and then the sequenced reads could be distinguished by their index. The concentration of products following final enrichment was measured using a NanoDrop 3300 Fluorospectrometer.

All final enriched and indexed products were pooled in equimolar concentrations, and the concentration of the pooled sample was measured using a NanoDrop 3300 Fluorospectrometer. The pooled sample was then sequenced on an Illumina HiSeq 2500 sequencer (Illumina, Inc, San Diego, CA, USA), with the nine samples multiplexed with 91 samples representing other projects. In total, the 100 samples took only one lane of the HiSeq flow cell. The steps involved in designing baits and conducting bench work are shown in Fig. 2 Phase 1.

Reads assembly

The raw reads were parsed to separate files for each species according to the index on the adapter. Adapter sequences and low quality sites ($Q < 20$) were trimmed from the reads using *Cutadapt-1.1* (Martin 2011) contained within the wrapper tool 'trim_galore_v0' (http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/). Then, the reads were parsed to corresponding locus and PCR duplicates were filtered using a custom Perl script (*rmrep.pl* and *bandp.pl*, Appendix S2). Next, the filtered reads were assembled into contigs by *Trinity v20140717* (Grabherr *et al.* 2011) using the default parameter settings, and *Geneious v7.1.5* (Kearse *et al.* 2012) was used to merge the overlapped contigs into longer ones. In the next step, all contigs from each gene of every

Table 1 Sampling localities, sequencing results of the nine individuals of the genus *Arostrilepis* and *Schizorchis*

Species	Sample ID	Museum ID*	Locality	DNA (ng)†	Raw reads (bp)‡	Trimmed reads (bp)§	Filtered reads (bp)¶	No. captured loci (%)**	On-target reads (%)††	Seq. depth‡‡
<i>A. macrocirrosa</i>	AF36167c4	UAM:Mamm:75266 MSB:Para:1277	Seward Peninsula, Alaska	500	8 163 560	8 137 054	3 047 444	3467 (95)	31.20	59
<i>A. macrocirrosa</i>	AF55136	UAM:Mamm:57557 MSB:Para:1305	Wrangell-St. Elias National Park, Alaska	500	25 581 840	25 497 618	7 847 690	3377 (93)	3.12	16
<i>A. macrocirrosa</i>	H105c1	No voucher	Tunguska River, Siberia	500	41 079 204	40 936 600	17 239 532	3180 (87)	1.37	14
<i>A. macrocirrosa</i>	AF38880c1	UAM:Mamm:79956	Kolyma River, Siberia	500	18 095 380	18 034 896	5 272 582	3410 (94)	7.67	27
<i>A. cooki</i>	IF6751	MSB:Mamm:158109 MSB:Para:1245	Meziadin Junction, British Columbia	500	28 987 006	28 891 378	9 317 094	3445 (95)	4.67	27
<i>A. cooki</i>	AZ7	MSB:Mamm:155593	Jubilee Lake, Oregon	300	21 977 568	21 905 270	6 537 788	3149 (86)	0.89	8
<i>A. beringiensis</i>	Lem2	No voucher	Indigirka River Delta, Siberia	1400	38 725 584	38 601 442	15 043 904	3416 (94)	2.98	28
<i>S. ochotonae</i>	KG470c1	CUMV:Mamm:20750 MSB:Para:20944	Pemberton, British Columbia	500	34 270 204	34 159 096	13 532 216	3206 (88)	1.17	5
<i>S. caballeri</i>	IF9737c3	UAM:Mamm:93191 MSB:Para:20987	Lake Clark National Park, Alaska	500	45 891 212	457 330 46	20 256 224	2904 (80)	5.61	15

*Museum catalog numbers for host and parasite specimens where available. Museums include University of Alaska Museum (UAM), Museum of Southwestern Biology (MSB), and Cornell University Museum of Vertebrates (CUMV). Host records in mammal collections are denoted by 'Mamm' and parasite collections are denoted by 'Para'.

†The amount of DNA used for library preparation.

‡The total basepairs of the raw reads.

§The number of basepairs of the reads after trimming.

¶The number of basepairs of the reads after removing PCR-duplicates.

**Number of loci that could be assembled in each species.

††Percentage of on-target reads calculated without counting the PCR-duplicates.

‡‡Average sequence depth without counting the PCR-duplicates.

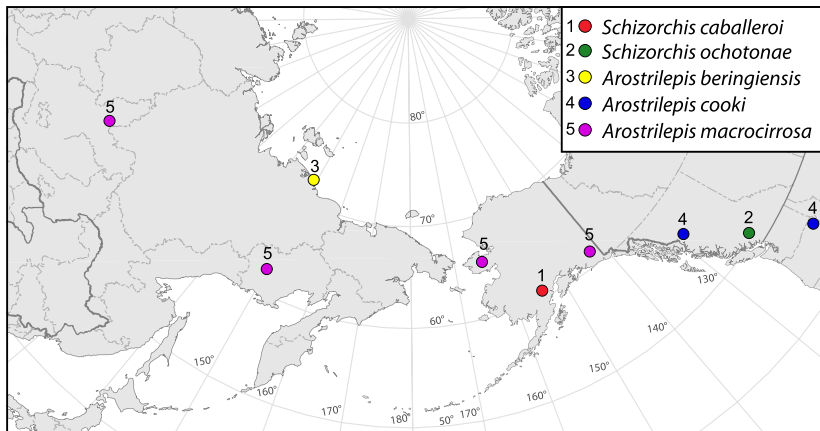


Fig. 1 Sampling localities of the nine specimens of the two genera, *Schizorchis* and *Arotrilepis*. Different coloured circles depict various species.

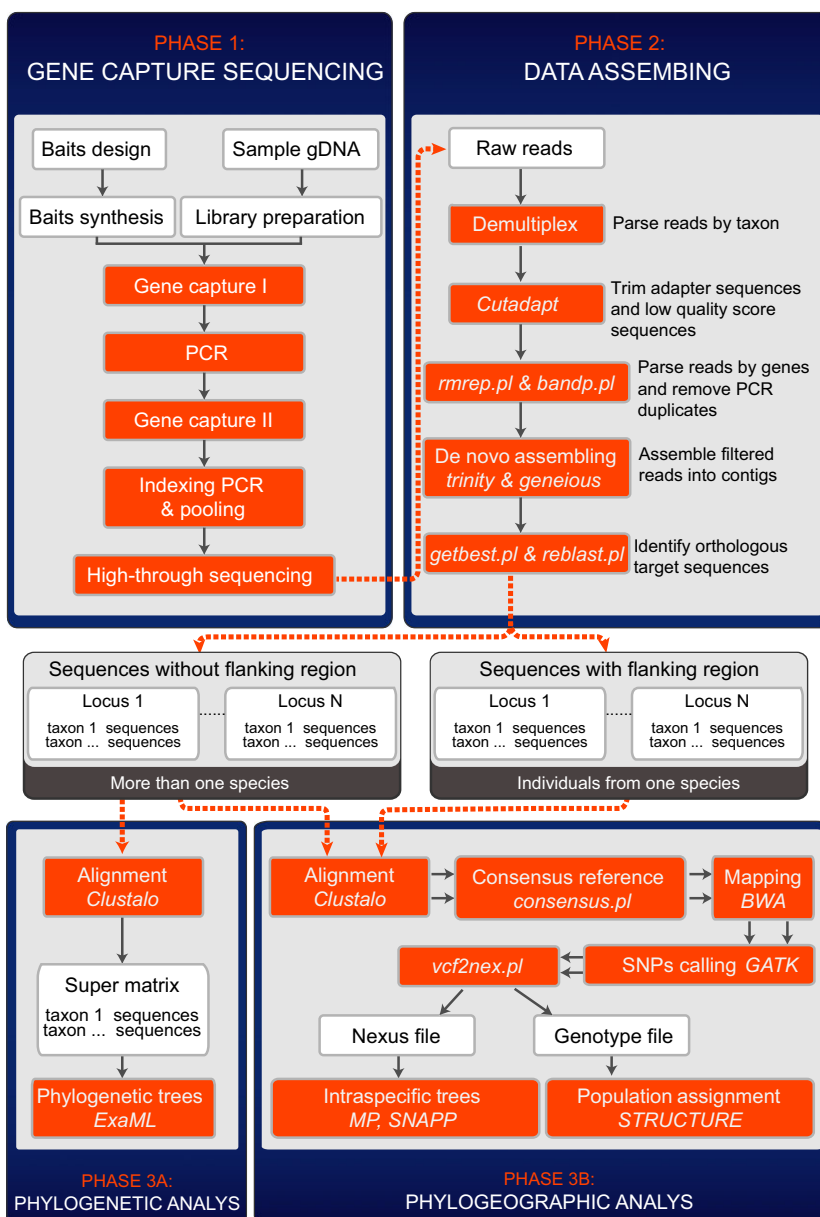


Fig. 2 The workflow represents three main phases in the target gene enrichment process: hybridization-based enrichment and sequencing, data assembling, and phylogenetic and phylogeographic analysis.

species were compared with the bait sequence using the Smith-Waterman algorithm (Smith & Waterman 1981) to identify the assembled contig that is most similar to the bait sequence. To verify the orthology of the retrieved sequence, the putative target sequence was aligned to the genome of *H. microstoma* using BLAST (Altschul *et al.* 1990). If the best hit from the BLAST results was not in the target region of *H. microstoma*, that particular sequence was considered to be paralogous and discarded without further analysis. Custom Perl scripts were used to automate the Smith-Waterman comparison and the BLAST search (*getbest.pl* and *reblast.pl*, Appendix S2). The pipeline of data assembly is illustrated in Fig. 2 Phase 2.

Sequence statistics, phylogenetic inference and phylogeographic analysis

For phylogenetic analysis, only coding regions of the captured sequences were used because the flanking non-coding regions of many loci were hard to align and the site homology was difficult to determine (Fig. 2 Phase 3a). The retrieved sequences of each locus were aligned using *Clustal Omega v1.1.1* (Sievers *et al.* 2011) with default parameter settings. Sequence statistics of each locus were calculated on the alignments. The length of the alignment, proportion of GC content, and average uncorrected pairwise distance (p-dist) were calculated using custom Perl scripts (*gc.pl* and *pdistance.pl*, Appendix S2). The consistency index (CI) and retention index (RI) were calculated using *PAUP* 4.0b10* (Swofford 2003). Phylogenetic trees were reconstructed using the maximum likelihood method implemented in *ExaML v3* (Kozlov *et al.* 2015) under the GTRGAMMA model with 1000 bootstrap replicates to assess nodal support.

For phylogeographic analysis, both the coding and noncoding regions of the captured sequences were examined. Average lengths of noncoding regions were calculated by subtracting the length of contigs without flanking region from the length of contigs with flanking region, which was counted using a custom Perl script (*len.pl*, Appendix S2). We applied different strategies for calling SNPs. To investigate variation between *A. macrocirrosa* and *A. cooki*, we made consensus sequences only from the coding region of each marker. This allowed us to avoid alignment problems associated with highly divergent noncoding regions. In intraspecific comparisons, we examined genetic variation based on both the coding and noncoding regions of all individual samples (*consensus.pl*, Appendix S2).

The trimmed fastq sequences were mapped to the reference using *BWA v0.7.5a* (Li & Durbin 2009) with default parameter settings. Base quality score recalibration, local realignment, SNP and INDEL discovery and genotyping across all samples simultaneously using

standard hard filtering parameters were carried using GATK-3.2.2 (McKenna *et al.* 2010). GATK Best Practices recommendations were followed (DePristo *et al.* 2011; Van der Auwera *et al.* 2013). A custom Perl script (*vcf2nex.pl*, Appendix S2) was used to convert the SNP vcf file to Nexus and genotype data file formats (Fig. 2 Phase 3b). For analyses assuming linkage equilibrium among loci, one of the best SNP sites (highest SNP calling score, fewer missing data) was chosen for each target region for the analyses.

A maximum parsimony tree was reconstructed for the four *A. macrocirrosa* and two *A. cooki* using the SNP data with 1000 bootstraps using *PAUP* 4.0b10*. We also inferred relationships among *Arostrilepis* samples using the SNAPP (Bryant *et al.* 2012) template implemented in BEAST 2.1.3 (Bouckaert *et al.* 2014). SNAPP was run for 1 000 000 iterations with sampling every 100th iteration. The first half of the trees was discarded as burn-in.

In order to test if the SNP data could help to identify the number of genetically distinct groups, we analysed the *A. macrocirrosa* data using Structure 2.3.4 (Pritchard *et al.* 2000). We ran three iterations for 100 000 generations (using a 100 000 burnin) for each value of K (number of population clusters) ranging from 2 to 4. To identify the number of population clusters that captures the major structure in the data we used Structure Harvester (Earl & vonHoldt 2012) to calculate the peak value for delta K (Evanno *et al.* 2005). The steps for phylogeographic analysis are summarized in Fig. 2 Phase 3b.

Results

Sequencing results

The raw reads obtained from the Illumina sequencer ranged from 8 163 560 bp to 45 891 212 bp for different samples (Table 1). Trimming the low quality reads and adapters resulted in <1% of the reads being removed. PCR duplicates accounted for more than half of the reads. All samples had sequences captured in at least 80% of target loci. The average number of loci captured was 3284 (90%). The percentage of unique on-target reads was 6.5% on average, with the lowest value of 0.89% in *A. cooki*, and the highest value of 31.2% in *A. macrocirrosa*. The average sequence depth counting only unique on-target reads was 22×, with the highest depth (59×) in *A. macrocirrosa* and the lowest depth (5×) found in *S. ochotonae* (Table 1).

Of the assembled contigs of all samples, 232 failed the homology check and were discarded as possible paralogs. Of the 3641 total target loci, 3639 were captured in at least one sample. Two loci (*Taenia_solium*. TsM1_s164.100374.101060 and *Taenia_solium*. TsM1_s1721.11728.10646) had no captured data probably

due to the low efficacy of the corresponding baits, whereas 1693 loci were captured for all the individuals. Alignment of sequences from the five captured species to the three reference genomes revealed that the length of the sequenced targets varied from 301 to 6969 bp, and the average length was 599 bp (Fig. 3, Appendix S3). The average GC content ranged from 42% to 51%. The average p-dist among the eight species ranged from 0.10 to 0.38 substitutions per site, with an overall average of 0.23. The average consistency index was 0.90 (0.76–1), and the average retention index was 0.88 (0.66–1). The length of noncoding region flanking the targets was centred on 500 bp, with a range between 1 and 1082 bp (Fig. 3, Appendix S3).

Phylogenetic analysis

The maximum likelihood phylogeny based on the 3639 loci (Fig. 4) has all nodes well-resolved with bootstrap support of 100. The five test species were grouped with congeners and split into two clades. The *Arostrilepis* species were grouped with *H. microstoma*. The *Schizorchis* species were placed as sister to the Taeniidae (Fig. 4).

Phylogeographic analysis

We extracted 14 074 SNPs from the bam file mapped four *A. macrocirrosa* and two *A. cooki*. One of the best

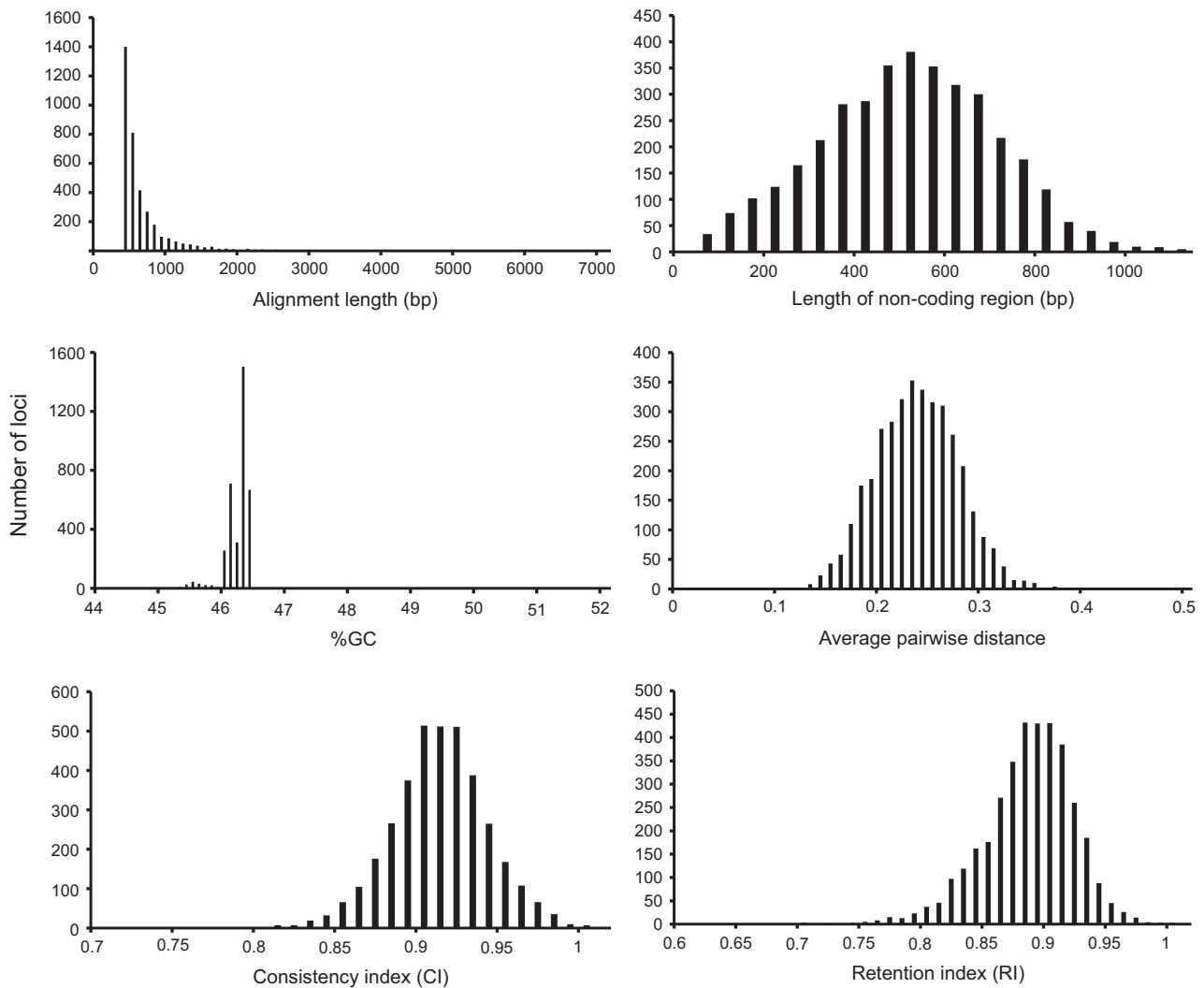


Fig. 3 Average length of the coding region, average length of the noncoding flanking region, average GC content, average pairwise distance, retention index and consistency index distribution of 3,639 loci. The length of noncoding region was calculated from all nine samples list in Table 1. All other statistics were calculated from the alignment of five species tested plus three species with genome sequence available.

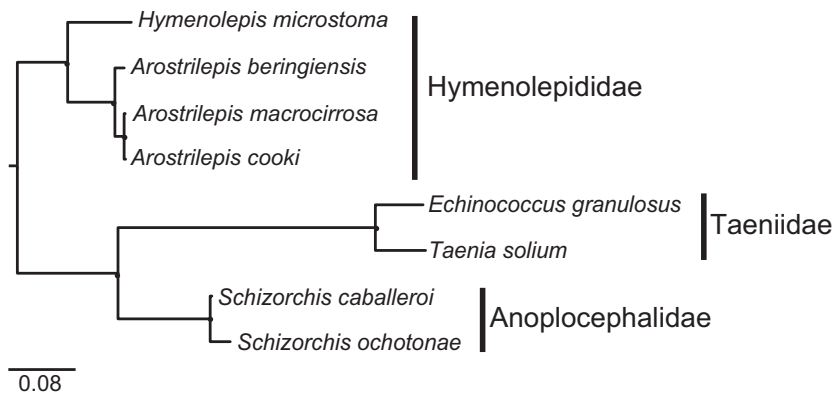


Fig. 4 ML tree reconstructed from the 3639 loci for eight species in Cyclophyllidea using ExaML. All nodes gain a 100 bootstrap support.

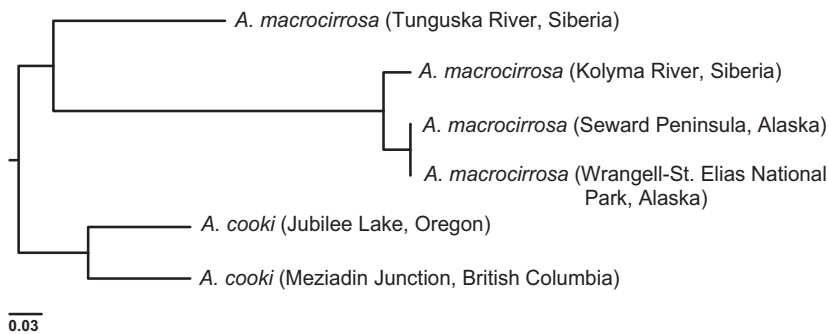


Fig. 5 Interrelationships of four *A. macrocirrosa* from Alaska and Siberia reconstructed on 14 074 SNPs using SNAPP. Two *A. cooki* were used for outgroup.

SNP sites was chosen for each of the 3639 loci. From the SNAPP analysis using *A. cooki* as the outgroup, we found the two *A. macrocirrosa* individuals sampled from Alaska were sister and closely related to an individual sampled from eastern Siberia. The sampled individual of *A. macrocirrosa* from central Siberia was derived from a deeper split in the phylogeny (Fig. 5). The maximum parsimony analysis showed the same relationships among the *A. macrocirrosa* individuals (Fig. S1).

We found 17 175 SNPs from the bam file mapping only the four *A. macrocirrosa* to target reference containing both coding regions and the flanking noncoding region. There were 5508 SNPs in the coding region (4 SNPs per kb) and 11 667 SNPs in the noncoding region (5 SNPs per kb). We chose one best SNP site that had the least number of missing data and the best calling score from each target locus. We obtained 2359 SNPs and used them in the Structure analysis to infer the number and distribution of genetic groups in *A. macrocirrosa*. The evaluation of delta K revealed that three population clusters provided the best estimate of population structure. These three clusters included 1) the two individuals collected from Alaska, 2) the specimen from eastern Siberia (Kolyma River), and 3) the specimen from central Siberia (Tunguska River). Under $K = 2$, the sample collected from eastern Siberia was grouped with the Alaskan samples.

Discussion

Target gene enrichment

We collected sequence data from 3639 of the 3641 total target loci. A common set of 1693 loci was captured in all of the samples. Only two loci failed the enrichment in every sample, which could be due to secondary structure in the target regions preventing hybridization between the baits and the targets, or malfunction in the synthesis of RNA baits for the target regions. We sampled species representing only two of the 15 families of the Cyclophyllidea: Anoplocephalidae (*Schizorchis*) and Hymenolepididae (*Arostrilepis*). Our taxon sampling was based on the research interests of the authors and clearly does not fully represent the diversity of the whole order. Thus, the usefulness of the baits employed here for other clades within the Cyclophyllidea remains to be tested empirically. However, our success in using baits designed on *H. microstoma* (Hymenolepididae) to sequence loci from samples representing a different family (Anoplocephalidae) indicates that the method shows promise for bridging large taxonomic gaps between available genomic data and focal taxa. Predictably, the number of loci captured in both *Schizorchis* species was low relative to the *Arostrilepis* species, suggesting that phylogenetic distance between the bait sequences and

the target sequences does influence the efficacy of target enrichment (Hedtke *et al.* 2013; Li *et al.* 2013; Hugall *et al.* 2016). Nevertheless, both *Schizorchis* species have more than 80% target loci captured (Table 1), so our target loci and enrichment protocols are demonstrated to be broadly useful among cyclophyllidean cestodes. In the future, we shall test these target loci in other cestode clades. With advancements in cestode genomics (Olson *et al.* 2012), we expect that there will be more cestode genomes or transcriptomes from different orders available for designing baits, which should increase the gene capture efficacy in different clades.

The number of captured loci is not correlated with either the number of trimmed reads or the number of filtered unique reads (Fig. S2). Therefore, merely increasing the number of reads would not improve the capture results in our experiment. Further decreasing the hybridization and washing temperature might help in obtaining more target sequences, especially for the targets that are divergent from the baits. We found a high number of PCR duplicates in all samples, resulting from the amplification steps for library preparation and hybridization. The PCR duplicates cannot provide new information and are detrimental in SNP calling, so we suggest reducing the number of cycles used for amplifying the library and for the indexing PCR in the future.

Variation of the markers and phylogeny

The range of uncorrected pairwise distances calculated for captured loci show that the markers we targeted are potentially suitable for resolving phylogenetic relationships at both shallow and deep taxonomic scales. The average consistency index (0.9) and the average retention index (0.88) are relatively high, indicating these target loci have low homoplasy and are good molecular markers for phylogenetic reconstructions of cestodes. The ML tree (Fig. 4) based on these loci is consistent with previous studies (von Nickisch-Rosenegk *et al.* 1999). We only showed an example of a maximum likelihood analysis based on concatenated data. If we had multiple samples for each species, a species tree approach, such as that implemented using MP-EST (Liu *et al.* 2010) could be used. Species tree methods assume linkage equilibrium among loci. The position of each locus on the three genomes (Appendix S3) could be used to determine the linkage among loci, or the linkage equilibrium between adjacent loci could be tested empirically.

Intraspecific variation and phylogeography

We targeted coding sequences, but we also collected sequence data from the flanking regions as 'by catch'.

We sheared the genomic DNA into ~250 bp for library preparation, and we could obtain sequences from flanking regions of around 500 bp total (including flanking regions that are both upstream and downstream of the target region). Thus, if more flanking region sequence was preferred, one could break the genomic DNA into longer pieces when making libraries, but inserts >1 kb are not recommended for sequencing on Illumina instruments. Our expectation was that the flanking sequences should be more variable than the coding sequences because they are noncoding and therefore presumably less constrained by purifying selection. However, our observation was that the noncoding flanking region only had slightly more SNP sites (5 SNPs per kb) than the coding region (4 SNPs per kb).

Both the intraspecific trees and the Structure analysis indicated that the sample of *A. macrocirrosa* collected from eastern Siberia is more closely related to samples collected from Alaska than to samples collected from central Siberia. Similar phylogeographic patterns also have been reported in small mammals (Fedorov *et al.* 1999; Galbreath & Cook 2004) as well as other organisms (e.g., peatmoss, Shaw *et al.* 2014). Admittedly, with only one individual sample per site our sampling was inadequate to rigorously address specific phylogeographic questions, but these data present an excellent illustration of the potential utility for our data processing and analysis pipeline in phylogeographic investigations. Besides the analyses we demonstrated that there are many other tools that could utilize the power of thousands of SNPs for different goals, such as Bayes factor delimitation (*with genomic data; BFD*) for species delimitation (Leaché *et al.* 2014), or Approximate Bayesian computation (ABC) for estimating demographic parameters (Li & Jakobsson 2012; Excoffier *et al.* 2013).

Conclusion

As an alternative to using traditional mtDNA or rDNA markers, we developed tools for sequencing thousands of single-copy nuclear coding sequences in cyclophyllideans by applying target enrichment and next generation sequencing. With multiplexed sequencing and optimized cost-saving steps in bench work, the cost of obtaining thousands of loci in hundreds of samples is no longer prohibitive (approximately \$160 per sample = \$80 for baits + \$30 for sequencing + \$50 for library prep and other costs). Data collected using this method can be applied to a range of phylogenetic and phylogeographic questions, and the potential of using genome-scale data to investigate the history of cyclophyllideans extends far beyond that which we have demonstrated.

Acknowledgements

This work was supported by the Program for Professor of Special Appointment (Eastern Scholar) at Shanghai Institutions of Higher Learning and the Innovation Program of Shanghai Municipal Education Commission to CL. This study is a contribution of the Beringian Coevolution Project, supported in part by the National Science Foundation (DEB 0196095, 0415668, 1256943) and organized by JAC, EPH and KEG. It was also supported in part by Department of Interior National Park Service support to JAC and a Doctoral Dissertation Improvement Grant to KEG (DEB 0506042). FAJ thanks the office of the Vice Chancellor for Research, SIU. We appreciate contributions of cestode specimens by Heikki Henttonen, Voitto Haukisalml and Antti Lavikainen.

References

- Altschul S, Gish W, Miller W, Myers E, Lipman D (1990) Basic local alignment search tool. *Journal of Molecular Biology*, **215**, 403–410.
- Bogitsh BJ, Carter CE, Oeltmann TN (2013) *Human Parasitology*, 4th edn. Academic Press, Amsterdam.
- Bouckaert R, Heled J, Kuhnert D *et al.* (2014) BEAST 2: a software platform for Bayesian evolutionary analysis. *PLoS Computational Biology*, **10**, e1003537.
- Brito PH, Edwards SV (2009) Multilocus phylogeography and phylogenetics using sequence-based markers. *Genetica*, **135**, 439–455.
- Bryant D, Bouckaert R, Felsenstein J, Rosenberg NA, RoyChoudhury A (2012) Inferring species trees directly from biallelic genetic markers: bypassing gene trees in a full coalescent analysis. *Molecular Biology and Evolution*, **29**, 1917–1932.
- Cook JA, Hoberg EP, Koehler A *et al.* (2005) Beringia: intercontinental exchange and diversification of high latitude mammals and their parasites during the Pliocene and Quaternary. *Mammal Study*, **30**, S33–S44.
- Crawford NG, Faircloth BC, McCormack JE *et al.* (2012) More than 1000 ultraconserved elements provide evidence that turtles are the sister group of archosaurs. *Biology Letters*, **8**, 783–786.
- DePristo MA, Banks E, Poplin R *et al.* (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics*, **43**, 491–498.
- Dorchies P (1999) Moniezia expansa: importance and economical consequences. *Revue De Medecine Veterinaire*, **150**, 107–110.
- Earl DA, vonHoldt BM (2012) STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. *Conservation Genetics Resources*, **4**, 359–361.
- Evanno G, Regnaut S, Goudet J (2005) Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Molecular Ecology*, **14**, 2611–2620.
- Excoffier L, Dupanloup I, Huerta-Sanchez E, Sousa VC, Foll M (2013) Robust demographic inference from genomic and SNP data. *PLoS Genetics*, **9**, e1003905.
- 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.
- Fedorov VB, Goropashnaya A, Jarrell GH, Fredga K (1999) Phylogeographic structure and mitochondrial DNA variation in true lemmings (Lemmus) from the Eurasian Arctic. *Biological Journal of the Linnean Society*, **66**, 357–371.
- Galbreath KE, Cook JA (2004) Genetic consequences of Pleistocene glaciations for the tundra vole (*Microtus oeconomus*) in Beringia. *Molecular Ecology*, **13**, 135–148.
- Galbreath KE, Hoberg EP (2012) Return to Beringia: parasites reveal cryptic biogeographic history of North American pikas. *Proceedings Biological Sciences/The Royal Society*, **279**, 371–378.
- Galbreath KE, Ragaliauskaitė K, Kontrimavichus L, Makarikov AA, Hoberg EP (2013) A widespread distribution for *Arostrilepis tenuicirrosa* (Eucestoda: Hymenolepididae) in *Myodes voles* (Cricetidae: Arvicolinae) from the Palearctic based on molecular and morphological evidence: historical and biogeographic implications. *Acta parasitologica/Witold Stefanski Institute of Parasitology, Warszawa, Poland*, **58**, 441–452.
- Georgiev BB (2003) Cestoda (Tapeworms). In: *Lower Metazoans and Lesser Deuterostomes. Grzimek's Animal Life Encyclopedia* (eds Thoney DA, Schlager N), pp. 225–243. Gale, Detroit.
- Gottstein B, Stojkovic M, Vuitton DA *et al.* (2015) Threat of alveolar echinococcosis to public health – a challenge for Europe. *Trends in Parasitology*, **31**, 407–412.
- Grabherr MG, Haas BJ, Yassour M *et al.* (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology*, **29**, 644–652.
- Greiman SE, Tkach VV (2012) Description and phylogenetic relationships of *Rodentolepis gnoskei* n. sp. (Cyclophyllidae: Hymenolepididae) from a shrew *Suncus varilla minor* in Malawi. *Parasitology International*, **61**, 343–350.
- Greiman SE, Tkach VV, Cook JA (2013) Description and molecular differentiation of a new *Staphylocystoides* (Cyclophyllidae: Hymenolepididae) from the dusky shrew *Sorex monticolus* in Southeast Alaska. *The Journal of parasitology*, **99**, 1045–1049.
- Hardman M, Hardman LM (2006) Comparison of the phylogenetic performance of neodermatan mitochondrial protein-coding genes. *Zoologica Scripta*, **35**, 655–665.
- Haukisalml V, Hardman LM, Henttonen H (2010) Taxonomic review of cestodes of the genus *Catenotaenia* Janicki, 1904 in Eurasia and molecular phylogeny of the Catenotaeniidae (Cyclophyllidae). *Zootaxa*, **2489**, 1–33.
- Haukisalml V, Hardman LM, Hoberg EP, Henttonen H (2014) Phylogenetic relationships and taxonomic revision of *Paranoplocephala* Luhe, 1910 sensu lato (Cestoda, Cyclophyllidae, Anoplocephalidae). *Zootaxa*, **3873**, 371–415.
- Hedtke SM, Morgan MJ, Cannatella DC, Hillis DM (2013) Targeted enrichment: maximizing orthologous gene comparisons across deep evolutionary time. *PLoS One*, **8**, e67908.
- Hoberg EP, Jones A, Bray RA (1999) Phylogenetic analysis among the families of the Cyclophyllidae (Eucestoda) based on comparative morphology, with new hypotheses for co-evolution in vertebrates. *Systematic Parasitology*, **42**, 51–73.
- Hoberg E, Mariaux J, Brooks D (2001) Phylogeny among orders of the Eucestoda (Cercaromorphae): Integrating morphology, molecules and total evidence. In: *Interrelationships of the Platyhelminthes. Systematics Association Special Vol. Ser. 60* (eds Littlewood D, Bray R), pp. 112–126. Taylor and Francis, London, UK.
- Hoberg E, Galbreath K, Cook J, Kutz S, Polley L (2012) Northern host-parasite assemblages: History and biogeography on the borderlands of episodic climate and environmental transition. In: *Advances in Parasitology* (eds Rollinson D, Hays SI), pp. 1–97. Elsevier, Amsterdam, NL.
- Hugall AF, O'Hara TD, Hunjan S, Nilsen R, Moussalli A (2016) An exon-capture system for the entire class Ophiuroidea. *Molecular biology and evolution*, **33**, 281–294.
- Kearse M, Moir R, Wilson A *et al.* (2012) Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics*, **28**, 1647–1649.
- Khalil LF, Jones A, Bray RA (1994) *Keys to the Cestode Parasites of Vertebrates*. CAB International, Willingford, UK.
- Kozlov AM, Aberer AJ, Stamatakis A (2015) ExaML version 3: a tool for phylogenomic analyses on supercomputers. *Bioinformatics*, **31**, 2577–2579.
- Leaché AD, Fujita MK, Minin VN, Bouckaert RR (2014) Species delimitation using genome-wide SNP data. *Systematic Biology*, **63**, 534–542.
- Lee JY, Edwards SV (2008) Divergence across Australia's Carpentarian barrier: statistical phylogeography of the red-backed fairy wren (*Malurus melanocephalus*). *Evolution*, **62**, 3117–3134.
- Lemmon AR, Emme SA, Lemmon EM (2012) Anchored hybrid enrichment for massively high-throughput phylogenomics. *Systematic Biology*, **61**, 727–744.

- Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, **25**, 1754–1760.
- Li S, Jakobsson M (2012) Estimating demographic parameters from large-scale population genomic data using Approximate Bayesian Computation. *BMC Genetics*, **13**, 22.
- Li C, Riethoven JJ, Naylor GJ (2012) EvolMarkers: a database for mining exon and intron markers for evolution, ecology and conservation studies. *Molecular Ecology Resources*, **12**, 967–971.
- Li C, Hofreiter M, Straube N, Corrigan S, Naylor GJ (2013) Capturing protein-coding genes across highly divergent species. *BioTechniques*, **54**, 321–326.
- Littlewood DTJ, Waeschenbach A, Nikolov PN (2008) In search of mitochondrial markers for resolving the phylogeny of cyclophyllidean tapeworms (Platyhelminthes, Cestoda) — a test study with Davaineidae. *Acta Parasitologica*, **53**, 133–144.
- Liu L, Yu L, Edwards SV (2010) A maximum pseudo-likelihood approach for estimating species trees under the coalescent model. *BMC Evolutionary Biology*, **10**, 302.
- Makarikov AA, Gulyaev VD, Kontrimavichus VL (2011) A redescription of *Arostrilepis horrida* (Linstow, 1901) and descriptions of two new species from Palaearctic microtine rodents, *Arostrilepis macrocirrosa* sp. n. and *A. tenuicirrosa* sp. n. (Cestoda: Hymenolepididae). *Folia Parasitologica*, **58**, 108–120.
- Makarikov AA, Galbreath KE, Hoberg EP (2013) Parasite diversity at the Holarctic nexus: species of *Arostrilepis* (Eucestoda: Hymenolepididae) in voles and lemmings (Cricetidae: Arvicolinae) from greater Beringia. *Zootaxa*, **3608**, 401–439.
- Mariaux J (1998) A molecular phylogeny of the Eucestoda. *The Journal of parasitology*, **84**, 114–124.
- Martin M (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet Journal*, **17**, 10–12.
- McCormack JE, Faircloth BC, Crawford NG *et al.* (2012) Ultraconserved elements are novel phylogenomic markers that resolve placental mammal phylogeny when combined with species-tree analysis. *Genome Research*, **22**, 746–754.
- McKenna A, Hanna M, Banks E *et al.* (2010) The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, **20**, 1297–1303.
- Mossel E, Steel M (2005) How much can evolved characters tell us about the tree that generated them? In: *Mathematics of Evolution and Phylogeny* (ed. Gascuel O), pp. 384–412. Oxford University Press, Oxford.
- Muehlenbachs A, Bhatnagar J, Agudelo CA *et al.* (2015) Malignant transformation of *Hymenolepis nana* in a human host. *New England Journal of Medicine*, **373**, 1845–1852.
- Nakao M, Lavikainen A, Yanagida T, Ito A (2013) Phylogenetic systematics of the genus *Echinococcus* (Cestoda: Taeniidae). *International Journal for Parasitology*, **43**, 1017–1029.
- von Nickisch-Roseneck M, Lucius R, Loos-Frank B (1999) Contributions to the phylogeny of the Cyclophyllidae (Cestoda) inferred from mitochondrial 12S rDNA. *Journal of Molecular Evolution*, **48**, 586–596.
- Olson PD, Littlewood DT, Bray RA, Mariaux J (2001) Interrelationships and evolution of the tapeworms (Platyhelminthes: Cestoda). *Molecular Phylogenetics and Evolution*, **19**, 443–467.
- Olson PD, Zarowiecki M, Kiss F, Brehm K (2012) Cestode genomics — progress and prospects for advancing basic and applied aspects of flatworm biology. *Parasite Immunology*, **34**, 130–150.
- Philippe H, Chenuil A, Adoutte A (1994) Can the Cambrian explosion be inferred through molecular phylogeny? *Development*, **120**, S15–S25.
- Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. *Genetics*, **155**, 945–959.
- Ramnath JD, Dutta AK, Das B, Tandon V (2014) Molecular characterization of the Indian poultry nodular tapeworm, *Raillietina echinobothrida* (Cestoda: Cyclophyllidae: Davaineidae) based on rDNA internal transcribed spacer 2 region. *Journal of parasitic diseases: official organ of the Indian Society for Parasitology*, **38**, 22–26.
- Rausch R (1952) Studies on the helminth fauna of Alaska. XI. Helminth parasites of microtine rodents; taxonomic considerations. *The Journal of parasitology*, **38**, 415–444.
- Rausch RL, Smirnova LV (1984) The genus *Schizorchis* (Cestoda, Anoplocephalidae) in Eurasian pikas, *Ochotona* spp. (Lagomorpha, Ochotonidae), with descriptions of new species. *Transactions of the American Microscopical Society*, **103**, 144–156.
- Robertson L, van der Giessen W, Batz M, Kojima M, Cahill S (2013) Have foodborne parasites finally become a global concern? *Trends in Parasitology*, **29**, 101–103.
- Schmidt GD (1986) *CRC Handbook of Tapeworm Identification*. CRC Press, Boca Raton, Florida.
- Shaw AJ, Golinski GK, Clark EG *et al.* (2014) Intercontinental genetic structure in the amphi-Pacific peatmoss *Sphagnum miyabeae* (Bryophyta: Sphagnaceae). *Biological Journal of the Linnean Society*, **111**, 17–37.
- Sievers F, Wilm A, Dineen D *et al.* (2011) Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular Systems Biology*, **7**, 539.
- Smith TF, Waterman MS (1981) Identification of common molecular subsequences. *Journal of Molecular Biology*, **147**, 195–197.
- Swofford DL (2003) *PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods) Version 4*. Sinauer Associates, Sunderland, Massachusetts.
- Terefe Y, Hailemariam Z, Menkir S *et al.* (2014) Phylogenetic characterisation of *Taenia* tapeworms in spotted hyenas and reconsideration of the “Out of Africa” hypothesis of *Taenia* in humans. *International Journal for Parasitology*, **44**, 533–541.
- Tinnin DS, Gardner SL, Ganzorig S (2008) Helminths of small mammals (Chiroptera, Insectivora, Lagomorpha) from Mongolia with a description of a new species of *Schizorchis* (Cestoda: anoplocephalidae). *Comparative Parasitology*, **75**, 107–114.
- Tsai IJ, Zarowiecki M, Holroyd N *et al.* (2013) The genomes of four tapeworm species reveal adaptations to parasitism. *Nature*, **496**, 57–63.
- Van der Auwera AG, Carneiro MO, Hartl C *et al.* (2013) From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Current protocols in bioinformatics/editorial board, Andreas D. Baxevanis. . . [et al.]* **11**, 11 10 11–11 10 33.
- Waeschenbach A, Webster BL, Littlewood DT (2012) Adding resolution to ordinal level relationships of tapeworms (Platyhelminthes: Cestoda) with large fragments of mtDNA. *Molecular Phylogenetics and Evolution*, **63**, 834–847.
- White AC Jr (2000) Neurocysticercosis: updates on epidemiology, pathogenesis, diagnosis, and management. *Annual Review of Medicine*, **51**, 187–206.
- Wickström LM, Haukisalmi V, Varis S *et al.* (2003) Phylogeography of the circumpolar *Paranoplocephala arctica* species complex (Cestoda: Anoplocephalidae) parasitizing collared lemmings (*Dicrostonyx* spp.). *Molecular Ecology*, **12**, 3359–3371.
- Yanagida T, Carod JF, Sako Y *et al.* (2014) Genetics of the pig tapeworm in Madagascar reveal a history of human dispersal and colonization. *PLoS One*, **9**, e109002.

C., F.A. and K.E. conceived the project. H. and J. performed data collection and analyses. All authors contributed to writing.

Data Accessibility

DNA Sequences used for baits designing: Appendix S1. Perl scripts: Appendix S2. Summary information about the 3639 nuclear loci: Appendix S3. DRYAD entry doi: the vcf, nexus and Structure input file were deposited at DRYAD <http://dx.doi.org/10.5061/dryad.7k3 h4>.

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Fig. S1 Maximum parsimony tree based on 14 074 SNPs.

Fig. S2 High number of ○ trimmed reads, or □ filtered reads did not lead to more captured target loci.

Appendix S1 Maximum parsimony tree based on 14,074 SNPs. All nodes received 100 bootstrap supports.

Appendix S2 Custom Perl scripts used in data analyses.

Appendix S3 Sequence statistics of the 3,639 nuclear loci tested in the cyclophyllid cestodes.