MOLECULAR ECOLOGY
RESOURCES
Molecular Ecology Resources (2011) **11**, 521–529 doi: 10.1111/j.1755-0998.2010.02956.x

PERMANENT GENETIC RESOURCES ARTICLE

Developing EPIC markers for chalcidoid Hymenoptera from EST and genomic data

KONRAD LOHSE,* BARBARA SHARANOWSKI,† MARK BLAXTER,* JAMES A. NICHOLLS* and GRAHAM N. STONE*

**Institute of Evolutionary Biology, University of Edinburgh, Edinburgh EH9 3JT, UK, †Department of Entomology, University of Manitoba, 223 Animal Science Building, Winnipeg, Manitoba, R3T 2N2, Canada*

Abstract

Increasing numbers of phylogeographic studies make comparative inferences about the histories of co-distributed species. Although the aims of such studies are best achieved by jointly analysing sequences from multiple loci in a model-based framework, such data currently exist for few nonmodel systems. We used existing genomic data and expressed sequence tags (ESTs) for Hymenoptera and other insects to design intron-crossing primers for 40 loci, mainly ribosomal proteins, for chalcidoid parasitoids. Amplification success was scored on a range of taxa associated with two natural communities; oak galls and figs. Taxa were chosen at increasing distance from *Nasonia*, which was used for primer design, (i) within Pteromalids, (ii) within Chalcidoidea (Eupelmidae, Eulophidae, Eurytomidae, Ormyridae, Torymidae) and (iii) for a selection of distantly related gall and fig wasps (Cynipidae, Agaonidae). To assess the utility of these loci for phylogeographic and population genetic studies, we compared genetic diversity between Western Palaearctic refugia for two species. Our results show that it is feasible to design a large number of exon-primed-intron-crossing (EPIC) loci that may be informative about phylogeographic history within species but amplify across a large taxonomic range.

Keywords: Chalcidoidea, EPIC loci, introns, primer design

Received 10 September 2010; revision received 15 November 2010; accepted 19 November 2010

Despite the increasing realization that multilocus data are required to adequately resolve histories at or below the species level (Zhang & Hewitt 2003; Jennings & Edwards 2005; Carstens & Knowles 2007), the majority of phylogeographic analyses of nonmodel organisms are still primarily based on mitochondrial DNA. When nuclear data are included, they are often presented as an add-on used to 'corroborate' qualitative inferences made from mitochondrial genealogies, rather than being analysed jointly in a model-based framework. One reason for the relatively slow uptake of model-based approaches by phylogeographers is simply the practical difficulty of obtaining a sufficient number of informative loci in nonmodel organisms. A recent study using multiple loci to estimate divergence and migration across a phylogeographic barrier (Lee *et al.* 2009) in a quantitative framework (Nielsen & Wakeley 2001, Hey & Nielsen 2004) found that stable parameter estimation requires a minimum of five nuclear loci. The general challenge is to identify enough loci that have a mutation rate high enough to generate a detectable signal of population level processes, whose evolution is at least approximately clock-like, and for which phylogeographic signals have not been overwritten by the effects of recombination. Additionally, on a practical level, amplification across related taxa is desirable both to reduce the cost of primer development and to facilitate comparisons across multiple species. In many ways, this contradicts the requirement of high levels of intraspecific variation. For example, most of the loci commonly used in phylogenetic analyses or for DNA barcoding, such as the D2 region of the 28S ribosomal RNA gene, amplify readily across a wide range of taxa (Cook *et al.* 2002; Rokas *et al.* 2002; Stone *et al.* 2009), but show little or no diversity below the species level (Stone *et al.* 2007). Conversely, anonymous loci generally provide good resolution in the target species but do not cross-amplify well at all (Jennings & Edwards 2005; Carstens & Knowles 2006; Lee *et al.* 2009).

Introns in single-copy nuclear genes offer a potential escape from this conundrum (Creer 2007). They evolve faster than coding regions and so are likely to contain sufficient intraspecific diversity to reconstruct genealogies, but are flanked by conserved exons (hence the term EPIC — exon-primed, intron-crossing — for such loci), which can be used as priming sites ensuring amplification

Correspondence: Konrad Lohse, Fax: +44(0)131 650 8684; E-mail: konrad.lohse@gmail.com

across a reasonable taxonomic range (Lessa 1992; Palumbi & Baker 1994; Creer 2007). Although intron sequences have been used in phylogeographic analyses of vertebrates (Li *et al.* 2010; Gifford & Larson 2008; Peters *et al.* 2008; Lee *et al.* 2009) and fruit flies (Wilder & Hollocher 2003; Das *et al.* 2004), their use in nonmodel taxa is still rare and their potential for comparative multispecies studies remains to be explored.

Here, we develop EPIC loci for phylogeographic inference in chalcidoid parasitoid wasps (Hymenoptera: Chalcidoidea), species-rich components of most terrestrial communities and dominant natural enemies of many insect herbivores (Askew 1980; Godfray 1994; Bailey *et al.* 2009). The complications of length variation in introns, which in diploid organisms often necessitates a time-consuming cloning step, can be avoided in Hymenoptera simply by using haploid males for which sequences can be obtained directly. Our aim was to identify loci that provide resolution at and below the species level while amplifying across a taxonomically diverse set of chalcidoid taxa, allowing multilocus, multispecies analyses of natural parasitoid communities. To avoid having to design and optimize primers for each species individually, we took a large-scale approach utilizing pre-existing genomic resources. The strategy was to develop primers for a large number of highly conserved genes using alignments of expressed sequence tags (ESTs) and publicly available genomic data from Hymenoptera (including the chalcid *Nasonia vitripennis*) and other insects. If transcripts have abundant conserved sites across a wide range of taxa, there should be sufficient minimally degenerate priming sites to amplify from disparate taxa.

Amplification success of candidate loci was assessed in two diverse and well-studied natural communities centred around herbivorous gall wasps (Hymenoptera; Cynipidae) on oak (*Quercus*) (Stone *et al.* 2002; Hayward & Stone 2005) and fig wasps (Hymenoptera; Agaonidae) (Weiblen 2002; Machado *et al.* 2005), with particular focus on the chalcidoid parasitoids associated with them. Primers were screened at increasing taxonomic distance from *Nasonia* (Pteromalidae): (i) in different genera of Pteromalidae, (ii) in five other families within the Chalcidoidea (Eulophidae, Eupelmidae, Eurytomidae, Ormyridae, Torymidae) and (iii) for a selection of phylogenetically distant host taxa in both systems (Cynipidae and Agaonidae, respectively). In total, this screening set encompasses a diverse set of taxa including both pest species (Aebi *et al.* 2006) as well as groups frequently used as biological control agents (e.g. Sha *et al.* 2007; Mena-Correa *et al.* 2009).

The rationale of developing such a large set of nuclear loci, which co-amplify across most species within these assemblages is to maximize overlap of loci used in future multispecies comparisons and to minimize potential ascertainment bias that species-specific choices of loci may introduce. To assess the potential of these loci for phylogeographic inference, we screened multiple individuals of two widespread parasitoids of oak galls (*Cecidostiba fungosa* and *Mesopolobus amaenus*) and measured diversity across three major glacial refugia in the Western Palaearctic.

Methods

Choice of nuclear loci and EST libraries

Putative orthologous gene alignments, developed for a separate phylogenomic study of Hymenoptera (Sharanowski *et al.* 2010), were used to develop primers. EST alignments were constructed from cDNA libraries for six hymenopteran taxa: *Neodiprion sertifer* (Diprionidae), *Campoletis sonorensis* (Ichneumonidae), *Pelecinus polyturator* (Pelecinidae), *Pristaulacus strangliae* (Aulacidae), an unidentified ceraphronid (Ceraphronidae) and an unidentified eucoiliine (Figitidae). Sequences were also obtained from public databases (NCBI) from the following taxa: *N. vitripennis* (Hymenoptera: Pteromalidae), *Solenopsis invicta* (Hymenoptera: Formicidae), *Lysiphlebus testacipes* (Hymenoptera: Braconidae), *Tribolium castaneum* (Coleoptera: Tenebrionidae), *Myzus persicae* (Hemiptera: Aphididae), *Acyrthosiphon pisum* (Hemiptera: Aphididae) and *Locusta migratoria* (Orthoptera: Acrididae). All sequences were compared against three annotated model genomes; *Drosophila melanogaster* (Diptera: Drosophilidae), *Bombyx mori* (Lepidoptera: Bombycidae) and *Apis mellifera* (Hymenoptera: Apidae). For details on cDNA library construction, contig assemblies, orthology determination and alignment protocols, see methods in Sharanowski *et al.* (2010).

EST alignments for 76 genes meeting the orthology criterion were filtered to include at least four hymenopteran taxa. Additionally, only alignments with less than 25% average difference at nonsynonymous sites across all Hymenoptera were utilized. Although this is an arbitrary cut-off, restricting the number of nonsynonymous changes was intended to aid primer design by decreasing the amount of degeneracy required to achieve amplification across a broad range of taxa. The average numbers of nonsynonymous sites for alignments were calculated using the Nei–Gojobori method (Nei & Gojobori 1986) in MEGA 4 (Tamura *et al.* 2007).

Of the 40 EST alignments meeting the above-mentioned criteria, 27 were RPs. We focused primarily on introns in RP genes for three reasons: (i) RP genes are typically conserved across eukaryotes; (ii) most RP genes do not occur in multiple copies; and (iii) there is no evidence to suggest genetic linkage. We also designed primers spanning introns in 13 conserved regulatory genes that met the above-mentioned criteria: *RACK1, SUI, Tctp,*

Table 1 Primer sequence and CG identifier for 26 nuclear loci which amplified a product in at least one of the focal taxa (Table 2) and primer sequence for *Cox1*

Locus	Primer	CG	Forw	Rev
Ant_sesB	40Fb/Rb	16944	GCCAAYGTYATCMGDTACTTC	TACKGTRTCRAAKGGATAGGA
Bellwether	33Fb/Rb	3612	GAAGAGGAAGTWYGARTTRGGWC	TTCRTACCAYTGBCTGAADGG
Magonashi	38F/R	9401	CTACGTCGGHCACAARGGHAART	TCTTGAACDAGRTARTAAAARCATC
nAcRbeta	39F/R	11348	GAGACBGACATCACBTTCTACAT	AGNAGATAYTTGGCRATGAGY
nAcRbeta	39Fb/Rb	11348	ATYATGAARTCRAACGTHTGG	ATGTAGAAVGTGATGTCVGTCTC
NIp	31F/R	7917	CTYTTRGGWCCAGARGCYAA	GTDSCAAGDAGATKGTGTCC
Pros25	26F/R	5266	GAATATGCYTTRGCHGCNGT	GTAKGCDCCVGADGGATCAC
RACK1	18Fb/Rb	7111	GATGGGTYACBCAAATYG	ATACCTTGACDACNCGRTCC
Ran	32F/R	1404	TAYATTCARGGMCARTGYGC	GGRTCCATTGTRACTTCTGG
RpL10ab	19F/R	7283	TAYGATCCVCARAAGGACAARC	AGGAGHCCAGGRAATTTRCCR
RpL12	10F/R	7939	GTGTACAGRCCDAMRATCGT	AADCCAGTTGGNARCATRTG
RpL13a	6F/R	1475	ATGACKGGCTTCAGYRAWAAG	GACATRAACTTYADCTTGTTCCTG
RpL15	2F/R	17420	GGGTGCNACTTAYGGHAARC	GCGMAGYTCACGRTGYTTDTG
RpL27a	28Fb/R	15442	CAAYTTYGACAARTACCATCCWG	CCYTTKCCYARRAGTTTGTA
RpL37	27F/R	9091	GAARGGTACNTCVAGYTTTGG	GACCRGTDCCRGTRGTCTTCCT
RpL37a	36F/R	5827	CGHACVAAGAAGGTTGGAATCAC	GTYCTYTTGCAYCGYTTGC
RpL39	16F/R	3997	ATGTCGGCHCAYAARACKTT	CTTBARCTTGGTTCKYCTCCA
RpS12	23F/R	11271	ATGGATGTSAAYACMGCMCTS	AGGGGTHTCHTCACCRAART
RpS15	20Fb/R	8332	GAYCARCTYCTDGAYATGC	CKACCRTGYTTWACAGGYTT
RpS17	34Fb/Rb	3922	CGCTATYATTCCWASCAARC	CAATRATRTCRTGYTCCARAGC
RpS18	22F/R	8900	GTYATGTTYGCYATGACNGC	KRAGRCCCCAGTARTGWCG
RpS23	21F/R	8415	ACVMGVTGGAAGGCYAATCC	ATGACCYTTACGHCCRAATCC
RpS4	11F/R	11276	BAARGCATGGATGTTRGACA	GGTCWGGRTADCGRATRGT
RpS8	5F/R	7808	GAAGAGGAAGTWYGARTTRGGWC	TTCRTACCAYTGBCTGAADGG
sansfille	35F/R	4528	CHWTVAAAATGCGTGGWCAAG	CDGGGAAYTGATTRAACARCAT
SUI	24F/R	17737	CCTTTGCWGATGCAATCAAG	CCGTGVACCTTSAGYTGDTC
Tctp	25F/R	4800	AYGAGATGTTCTCNGAYAC	GATRTCCATDGATTCNCCRGT
Cox1	pF2/ 2413d	n/a	ACCIGTDATRATRGGDGGITTYGGDAA	GCTADYCAICTAAAAATYTTRATW CCD GT

Mp20, myofilin, NIp ran, bellwether, AntSesB, nAcRbeta, magonashi, sansfille, pros25 (Table 1).

Primer design

The 40 EST alignments were aligned against *D. melanogaster* genomic sequences in BioEdit using ClustalW (Thompson *et al.* 1994) and checked by eye. Primers were designed to match coding exon regions flanking known introns in *D. melanogaster*. We chose priming sites that were conserved across Hymenoptera and, whenever possible, across other insect sequences in the alignment. Initially, we matched primer sequences to *N. vitripennis*, the only chalcidoid in the set, then built in up to 54-fold degeneracy to increase amplification success across taxa. Sequences from the braconid wasp *L. testacipes* frequently proved too divergent to be included in this degeneracy. If possible multiple, often nested, primers were designed for each locus (Table 1).

Standard primer characteristics (melting temperature, scores for dimer formation, self-annealing and 3′ stability) were checked in FastPCR (Kalendar *et al.* 2009) and Primer3 (Untergasser *et al.* 2007) using default settings.

Cross-species screening

Whole genomic DNA was extracted from specimens stored in 98% ethanol following Nicholls *et al.* (2010). Within Pteromalidae, primers were tested on three parasitoid species associated with oak galls (*C. fungosa, M. amaenus, Caenacis lauta)* and three nonpollinating, parasitic figwasps (*Sycoscapter* sp., *Philotrypesis*, *Walkerella* sp.). *Nasonia vitripennis*, the only chalcidoid sequence included in the EST alignments, was used as a positive control. We also screened amplification success in one species from each of a further five families in the superfamily Chalcidoidea: *Torymus affinis* (Torymidae), *Omyrus nitidulus* (Ormyridae), *Eupelmus annulatus* (Eupelmidae), *Baryscapus pallidae* (Eulophidae) and *Eurytoma brunniventris* (Eurytomidae) (Table 2). These families are representative of parasitoids across many insect communities (Askew 1980). Finally, primers were tested on six species of gall wasps (Cynipidae) and three species of fig wasps (Agaonidae), representing two more distantly related families that are the focus of much ongoing research into community evolution (Table 2).

Table 2 Amplification success and product sizes of primers developed from hymenopteran expressed sequence tag libraries tested on hymenopteran taxa from two natural communities. Locus names are from FLYBASE according to the *Drosophila melanogaster* genomic region used in the alignment for primer design. Only primer pairs that amplified in at least one of the test species are shown. Primer pairs that failed to amplify a PCR product in a particular species are indicated by 0, combinations resulting in multiple bands by D and those not tested by dashes. Sequencing was only attempted in *Cecidostiba fungosa*, *Caenacis lauta* and *Mesopolobus amaenus*. If the exact product size could not be determined because of messy sequence at the ends, only the length of the readable sequence is shown (in bold). Product sizes in the other taxa were estimated on agarose gels

		Pteromalidae						Chalcidoidea					Cynipidae						Agaonidae		
LOCUS	Primers	*C. fungosa*	*C. lauta*	*M. amaenus*	*Sycoscapter sp.*	*Philotrypesis sp.*	*Walkerella sp.*	*Eurytoma brunniventris*	*Baryscapus pallidae*	*Torymus affinis*	*Eupelmus annulatus*	*Ormyrus nitidulus*	*Andricus quercustamuli*	*Dryocosmus kuriphilus*	*Andricus dentimitratus*	*Plagiotrochus quercusilicis*	*Pediaspis aceris*	*Diplolepis rosae*	*Plestodontes frogatti*	*Ceratosolen appendiculatus*	*Platyneura sp.*
AntSesB	40Fb/Rb	728	612	**592**	700	700	750	600	650	1300	900	750	850	1500	1500	1500	0	750	0	0	850
Bellwether	33Fb/Rb	576	D	595	D	600	D	D	450	D	D	600	D	D	0	0		D	0	D	D
magonashi	38F/R	350	—	—	—	—	—	300	0	0	0	1500	—	—	—	—	—	—	—	—	—
nAcRbeta	39F/R	289	279	279	350	350	350	550	0	0	300	600	0	0	0	0	0	0	1500	0	350
nAcRbeta	39Fb/Rb	488	485	—	600	600	600	500	800	0	0	950	0	0	0	0	0	0	0	1100	850
Nlp	31F/R	0	—	—	—	—	—	0	500	550	350	400	—	—	—	—	—	—	—	—	—
Pros25	26F/R	470	472	0	500	500	550	650	450	0	0	500	0	1000	D	0	0	0	0	0	0
Rack1	18Fb/Rb	**862**	**566**	**825**	850	0	850	1100	900	950	950	900	0	0	0	0	900	0	0	0	0
Ran	32F/R	499	499	469	600	600	600	500	500	550	500	600	550	1000	500	900	600	1000	600	450	650
RpL10ab	19F/R	**968**	**1028**	**987**	1000	1000	1000	450	1000	1050	0	950	1000	1000	0	1000	1000	1000	0	0	0
RpL12	10F/R	D	—	0	0	0	0	400	D	0	0	D	750	0	0	0	0	0	0	0	0
RpL13a	6F/R	864	933	0	0	0	0	1000	0	0	0	0	0	0	0	0	0	0	0	0	0
RpL15	2Fb/Rb	652	628	642	700	700	0	550	500	500	500	500	0	700	700	0	700	850	700	0	0
RpL27a	28Fb/R	609	554	583	800	650	700	600	600	0	0	600	0	0	0	1500	800	800	0	0	0
RpL37	27F/R	**903**	952	628	650	650	650	D	600	950	550	500	350	400	400	400	D	600	900	650	600
RpL37a	36F/R	220	222	232	250	250	250	200	250	200	200	250	750	0	0	0	0	0	250	250	250
RpL39	16F/R	585	564	592	0	0	0	650	600	700	0	600	0	600	0	0	0	0	600	0	0
RpS12	23F/R	800	—	—	—	—	—	0	800	0	0	750	—	—	—	—	—	—	—	—	—
RpS15	20Fb/R	761	765	800	0	650	800	500	0	0	0	0	0	0	0	0	650	800	0	0	0
RpS17	34Fb/Rb	861	900	0	900	600	800	600	0	0	0	0	0	1000	0	1500	0	650	0	0	0
RpS18	22F/R	819	843	836	900	1000	1000	1000	1000	0	1500	0	900	D	1000	1000	1000	1500	1500	0	0
RpS23	21F/R	268	268	268	300	300	300	250	250	300	250	200	350	0	D	0	0	300	300	300	300
RpS4	11F/R	782	769	761	800	800	800	800	750	800	850	D	800	0	0	0	0	0	0	800	0
RpS8	5F/R	**446**	454	460	550	550	0	500	450	500	450	0	700	700	0	700	800	0	550	0	0
sansfille	35F/R	447	450	434	0	0	0	0	0	0	0	450	0	0	0	0	0	0	0	0	0
SUI	24F/R	887	0	831	0	900	0	800	800	900	0	800	1500	1500	0	1500	1500	1500	0	900	900
Tctp	25F/R	494	498	462	0	0	0	600	500	0	0	500	1000	0	0	0	0	0	0	0	0
Total(*)		26	22	19	17	18	16	24	21	14	13	22	13	12	7	9	10	12	9	8	9

Polymerase chain reactions (PCR) were performed in 20 μL reactions using the following mix for all primer combinations: 2.0 μL 10× Bioline PCR buffer, 2.0 μL bovine serum albumin (10 mg/mL), 0.8 μL $MgCl_2$ (50 mM), 0.16 μL dNTPs (25 mM each), 0.1 μL Taq Polymerase (5 U/μL, Bioline), 0.2 μL of each primer (20 μM) and 1 μL DNA template.

A generic touchdown PCR protocol was used for all loci: 94 °C for 3 min, followed by cycles of 94 °C for 15 s, an annealing step of 40 s, 72 °C for 3 min, with a final extension at 72 °C for 10 min. The annealing temperature was varied as follows: the first 10 cycles decreased in 1 °C increments from 65 to 56 °C, followed by 30 cycles each with an annealing step at 55 °C.

Divergence, diversity and information content

To assess the utility of the new EPIC loci for intraspecific studies, we measured diversity within *C. fungosa* and *M. amaenus*. In each species, three male individuals, one each from three different Pleistocene refugia in southern Europe (Iberia, the Balkans and Asia Minor), were sequenced for all loci that amplified in the initial screen. Sequences were also obtained from a single male of *C. lauta*, a species closely related to *C. fungosa*. PCR products were sequenced directly in both directions using BigDye chemistry (Perkin Elmer Biosystems, Waltham, MA, USA) run on an ABI 3730 capillary sequencer in the GenePool Edinburgh. Chromatograms were checked by eye and complimentary reads aligned using Sequencher v. 4.8. Edited sequences were aligned in ClustalW and checked by eye. Exonic regions were determined by comparison to *D. melanogaster* protein sequences and checked for an open reading frame. As a comparison, we sequenced a 698 bp region of the frequently used mitochondrial cytochrome *c* oxidase subunit 1 gene (*Cox1*) for the above samples using primers COI_pF2 and COI_2413d (Table 1). These primers were designed for chalcidoid wasps and amplify a fragment largely overlapping the LCO/HCO region of *Cox1* (Folmer *et al*. 1994), but excluding a poly-T repeat at its 5′ end present in Chalcidoidea, which causes slippage during PCR resulting in uninterpretable sequence.

Average pairwise diversity (π) in the Western Palaearctic (both in *C. fungosa* and *M. amaenus*) and divergence (K) between *C. fungosa* and the closely related outgroup *C. lauta* were computed using DNAsp (Rozas and Rozas 1995). When choosing loci for intraspecific studies, it is crucial to avoid ascertainment bias. Selecting loci based on their diversity potentially confounds coalescence variance with differences in mutation rate between loci. For example, excluding loci which show no or low diversity in the focal species may upwardly bias age estimates of phylogeographic events. We therefore used divergence at each locus to obtain an unbiased measure of information content. The number of divergent sites between *C. fungosa* and *C. lauta* at each locus normalized by its mean across loci was taken as an overall measure of information content taking locus length into account. We also computed summaries for intron (K_{in}, π_{in}) and synonymous exon (K_s,π_s) sites separately. Because the latter are thought to evolve neutrally, this comparison should reveal the potential selective constraint on introns. Both *bellwether* and *SUI* failed to amplify in *C. lauta*, leaving 18 loci for which divergence and information content could be computed.

Results

Cross-species screening

Of the 40 loci tested, 32 successfully amplified a product in the positive control, *N. vitripennis* and of these 26 yielded a PCR product in at least one other chalcidoid taxon (Table 2). Amplification success differed markedly both between different Pteromalid taxa and between chalcidoid families. For example, fewer loci amplified in the fig-associated Pteromalids compared with the species attacking oak galls. Similarly, only 13 loci amplified in *E. annulatus* (Eupelmidae), whereas amplification success in *E. brunniventris* (Eurytomidae) (24 loci) and *O. nitidulus* (Ormyridae) (22 loci) was comparable to that in the three oak gall–associated Pteromalid species. Only nine loci (*AntSesB, bellwether, RACK1, ran, RpL15, RpL37, RpL37a, RpS23, RpS4*) amplified a product in all six chalcidoid families associated with oak galls. Amplification success was considerably lower both in the Cynipidae and Agaonidae compared to any of the chalcidoid parasitoids, which is expected given that the former are taxonomically much more distantly related to *Nasonia* (Table 2).

Product length varied widely between chalcidoid species with some combinations of primer pairs and taxa yielding PCR products in excess of 1000 bp, too long for direct sequencing (Table 2). Similarly, some fragments (*AntSesB* and *SUI*) were consistently larger in cynipids than in chalcidoids. Whether this variation is random or reflects genome-wide differences in intron length or indeed genome size between hymenopteran taxa remains to be explored. The fact that the majority of the loci that amplified in *T. affinis* were longer in this species than in any of the other five chalcidoids, does suggest some general genome-wide difference between taxa.

Divergence, diversity and information content

Sequences for *C. fungosa*, *C. lauta* and *M. amaenus* are deposited in GenBank (accession nos HM208872–HM209026

and HQ596410–HQ596457). Taken across loci, mean per site divergence between *C. fungosa* and *C. lauta* was higher at synonymous exon sites ($K_s = 13.9\%$) than in introns ($K_{in} = 7.3\%$). In contrast, average per site diversity was similar between synonymous sites ($\pi_s = 0.9\%, 1.1\%$) and introns ($\pi_{in} = 1.0\%, 1.0\%$) in *C. fungosa* and *M. amaenus*, respectively (Table 3). Loci differed considerably in their overall information content (Table 3). In *C. fungosa*, the most informative loci include *RpL37*, *nAcRbeta*, *RpL13a* and *RpS15*. Perhaps not surprisingly, those also tended to have rather high diversity in the introns (π_{in}), which in some cases was comparable to synonymous site diversity in *Cox1*. Conversely, the two loci with the lowest diversity in either *C. fungosa* (*RpL39, RpL37a*) or *M. amaenus* (*RpS23* and *RpS8*) had low or average information content (Table 3). Generally, average K_s was about three times lower for nuclear loci than for *Cox1*, and average levels of nuclear diversity both in *C. fungosa* and *M. amaenus* were much lower than synonymous diversity in *Cox1*. Levels of diversity observed at individual loci differed considerably between *C. fungosa* and *M. amaenus*, despite the fact that the mean values were similar for the two species. For example *RpL39*, which is monomorphic in *C. fungosa*, had above average diversity ($S = 7$) in *M. amaenus* and—on a similar spatial scale–has proven to be informative in the Torymid *Megastigmus stigmatizans* (Nicholls *et al.* 2010). This is expected because diversity at a particular locus is determined both by its mutation rate and the randomness of genetic drift.

In *C. fungosa*, 13 of 20 loci and in *M. amaenus* 7 of 16 loci contained (mainly single nucleotide) indels between the three refugial populations (Table 3). Indel variation matters in practice for two reasons. First, it provides additional characters for population genetic and phylogeographic inference. Secondly, it creates problems when aligning sequences from heterozygous diploid individuals necessitating an additional cloning step. However, given that the number of indels is expected to be lower within than between populations, our results suggest that direct sequencing of diploid females in these species may be possible in the majority of cases. This is confirmed by Lohse *et al.* (2010) who sequenced two diploid female

Table 3 Basic properties of nuclear loci in *Cecidostiba fungosa* and *Mesopolobus amaenus*. Length values exclude indels in the *C. fungosa–Caenacis lauta* alignment. Diversity across three major Pleistocene refugia and divergence between *C. fungosa* and *C. lauta* were calculated for introns (π_{in}, K_{in}) and synonymous exon sites (π_s, K_s) separately. Also shown are the number of introns (#In), the total number of polymorphic sites (S) and indels within species and, for *Cecidostiba fungosa*/*C. lauta*, a measure of information content (Info). Corresponding summaries for *COI* are provided

		Length			*C. fungosa*/*C. lauta*			Diversity (*C. fungosa*)				Diversity (*M. amaenus*)			
LOCUS	Primers	#In	Total	Intron	K_s	K_{in}	Info	π_s	π_{in}	S	Indel	π_s	π_{in}	S	Indel
AntSesB	40fb, 40rb	2	606	156	0.076	0.148	0.981	0.000	0.008	2	0	0.000	0.024	7	0
Bellwether	33fb, 33rb	1	550	216	n/a	n/a	n/a	0.000	0.003	2	3	n/a	n/a	n/a	n/a
nAcRbeta	39f, 39r, 39fb, 39rb	2	728	113	0.371	0.227	2.039	0.004	0.000	1	0	0.000	0.044	10	2
Rack1	18fb, 18rb	2	560	304	0.087	0.052	0.578	0.000	0.007	3	3	0.021	0.010	10	0
Ran	32f, 32r	1	499	202	0.090	0.091	0.659	0.011	0.003	2	0	0.000	0.009	3	1
RpL10ab	19f, 19r	2	955	807	0.072	0.043	1.001	0.000	0.003	3	1	0.044	0.006	9	1
RpL13a	6f, 6r	2	849	720	0.000	0.097	1.975	0.000	0.019	21	1	n/a	n/a	n/a	n/a
RpL15	2fb, 2rb	2	618	412	0.233	0.056	1.065	0.000	0.002	2	2	0.000	0.011	7	2
RpL27a	28fb, 28r	2	501	338	0.155	0.101	1.078	0.017	0.030	16	4	0.000	0.007	4	1
RpL37	27f, 27r	1	866	788	0.017	0.123	2.681	0.033	0.020	24	4	0.000	0.016	13	3
RpL37a	36f, 36r	1	220	91	0.408	0.069	0.436	0.000	0.000	0	0	0.000	0.013	2	1
RpL39	16f, 16r	1	463	444	0.000	0.086	1.055	0.000	0.000	0	0	0.000	0.009	7	0
RpS15	20fb, 20rb	1	739	475	0.073	0.091	1.308	0.058	0.035	30	4	n/a	n/a	n/a	n/a
RpS18	22f, 22r	2	813	562	0.072	0.052	1.011	0.020	0.005	6	0	n/a	n/a	n/a	n/a
RpS23	21f, 21r	1	268	79	0.119	0.127	0.408	0.016	0.042	6	1	0.016	0	1	0
RpS4	11f, 11r	2	754	431	0.094	0.083	1.290	0.000	0.000	1	1	0.000	0.008	7	0
RpS8	5f, 5r	1	422	242	0.060	0.034	0.311	0.029	0.008	6	2	0.000	0.003	1	0
sans_fille	35f, 35r	1	446	84	0.140	0.037	0.367	0.017	0.000	2	1	0.017	0.000	2	0
SUI	24f. 24r	1	823	636	n/a	n/a	n/a	0.000	0.006	6	2	0.000	0.006	6	0
Tctp	25f, 25r	2	493	148	0.134	0.088	0.670	0.000	0.014	3	0	0.040	0.018	8	0
Total		30	12173	7249						136	29			97	11
MEAN			608.6	362.5	0.139	0.073		0.009	0.010	6.8		0.011	0.010	6.1	
COI		n/a	698	n/a	0.353			0.090		24		0.209		54	

C. fungosa individuals for the 20 loci described here and found length heterozygosity necessitating a cloning step only in five cases.

Discussion

We have shown that EPIC markers can be developed relatively straightforwardly for nonmodel organisms using publicly available EST and genomic data. Our strategy of testing a large number of degenerate primers on a set of focal taxa avoids time-consuming, species-specific PCR optimization and efficiently identified a set of loci of likely value across six families of chalcidoid parasitoids and beyond. This approach to marker development should be feasible in any group of organism for which EST or transcriptomic data are available and has recently also been used in fish (Li *et al.* 2010). While in this study, alignment to the *Drosophila* genome allowed us to specifically target known introns, developing EPIC markers is possible even without a genomic reference, if one accepts that a fraction of primers spanning splice sites will not amplify.

We emphasize that numbers of loci available in candidate species within chalcidoid families could probably be increased by further taxon-specific PCR optimization or an additional cloning step. Although nuclear mutation rates are on average lower than those of mitochondria, this and previous studies (Lee *et al.* 2009) show that, because of coalescent and mutational variance, the same does not necessarily hold for levels of diversity observed at individual loci. We also do not find the dramatic difference between mitochondrial and nuclear divergence which has been reported for *Nasonia* sister species and attributed to *Wolbachia*-induced sweeps (Oliveira *et al.* 2008). Thus, despite their lower per site mutation rate on average, multiple EPIC loci such as the ones developed here, if analysed in a model-based framework, should be far more informative about within-species phylogeographic history than mitochondrial data. The *C. fungosa* data obtained in this study have been successfully used to resolve the relationships between refugial populations in this species (Lohse *et al.* 2010). This study used larger samples per population and tested for recombination within each locus using the four-gamete test. Encouragingly, only three loci showed evidence for recombination, and a mere 265 bp of sequence (around 2% of the data) had to be excluded to trim loci into nonrecombing blocks.

If patterns of divergence across loci in *C. fungosa* and *C. lauta* are at all representative, the most informative loci for within-species historical inferences in Chalcidoidea are likely to include *RpL37, nAcRbeta, RpL13a, RpS15, RpS4* and *AntSesB*. If, as recent power analyses suggest, between five and a dozen loci are sufficient to infer ancestral population parameters in divergence models reliably (Jennings & Edwards 2005), these EPIC loci should allow multilocus phylogeographic analysis across a broad taxonomic range of chalcidoid parasitoids and in turn, facilitate comparative phylogeographic analysis of natural chalcidoid assemblages. The observed variation in amplification success between families would suggest that it may be impossible to use a standard set of loci across taxa even if this may be desirable to avoid confounding true differences in species histories with locus-specific effects. However, as long as enough loci per species are sampled to capture the variance in genealogical history and outgroup comparisons are used to account for heterogeneity in mutation rates across loci, there is no a priori reason against using only partially overlapping sets of loci in multispecies comparisons. Given that the primers developed here are anchored in highly conserved coding regions and many of them amplify across a large taxonomic range, they may also be of use as genomic tools more broadly in the Hymenoptera and other insects. For example, some of the loci employed in this study (e.g. *RpL15, RpL27a, ran*) have previously been used as markers for QTL mapping in Lepidoptera (Papanicolaou *et al.* 2005). Furthermore, the exonic portions of many of these loci have proven informative for resolving deep-level relationships within the Hymenoptera (Sharanowski *et al.* 2010).

An important question is to what extent introns in highly conserved genes evolve neutrally. Generally, our finding of lower levels of divergence in introns compared to synonymous sites in *C. fungosa* is consistent with previous results from genome-wide studies in *Drosophila* suggesting that introns are under purifying selection, which may be particularly strong in highly conserved genes (Haddrill *et al.* 2005; Halligan & Keightley 2006). Similarly, correlations between intron length and divergence have been interpreted as evidence for selective constraints on regulatory elements present in long introns (Halligan & Keightley 2006). We tested for this in *C. fungosa* and found a negative but nonsignificant trend between intron length and K_{in} ($r = -0.265, P = 0.189$). This suggests that any correlation between intron length and selective constraint, if present in *C. fungosa*, is likely to be weak. Thus, it may be difficult to avoid potential biases arising from selective constraints by selecting short introns. On the contrary, because information content is a function of both intron length and K_{in}, the most informative loci in the present set are those containing long introns (Table 3). However, while selective constraints on introns or linked exons should not lead to systematic biases in estimates of ancestral population parameters, they do translate in a trade-off between maximizing information for phylogeographic inference and the ability to amplify across taxa. This has been demonstrated

previously in a study on birds (Lee *et al*. 2009) which found per site diversity in anonymous loci, presumably intergenic DNA, to exceed those in introns. On the other hand, using EPIC loci with known orthology and function for phylogeographic inference can be viewed as an improvement over anonymous loci for which orthology and function are generally unknown (Jennings & Edwards 2005). In general, with the increasing volumes of publicly available genome and transcriptome data, developing EPIC primers for nonmodel organisms is now straightforward and multilocus nuclear sequence data will surely become the standard in studies of population history and phylogeography.

Acknowledgements

Many thanks to James Cook, George Melika, Majide Tavakoli, Juli Pujade-Villar, Pablo Fuentes-Utrilla and Anna Moynihan for providing specimens. This work was supported by grants from the U.K. Natural Environment Research Council to GNS (NE/E014453/1 and NER/B/504406/1) and the National science Foundation to BS (Grant EF-0337220). KL is supported by a BBSRC studentship.

References

Aebi A, Schönrogge K, Melika G *et al*. (2006) Parasitoid recruitment to the globally invasive chestnut gall wasp *Dryocosmus kuriphilus*. In: *Galling Arthropods and Their Associates* (ed. Ozaki K, Yukawa J, Ohgushi T, Price PW), pp. 103–121. Springer, Tokyo.

Askew RR (1980) The diversity of insect communities in leaf mines and plant galls. *The Journal of Animal Ecology*, **49**, 145–152.

Bailey R, Schönrogge K, Cook JM *et al*. (2009) Host niches and defensive extended phenotypes structure parasitoid wasp communities. *PLoS Biology*, **7**, e1000179.

Carstens BC, Knowles LL (2006) Variable nuclear markers for melanoplus oregonensis identified from the screening of a genomic library. *Molecular Ecology Notes*, **6**, 683–685.

Carstens BC, Knowles LL (2007) Shifting distributions and speciation: species divergence during rapid climate change. *Molecular Ecology*, **16**, 619–627.

Cook JM, Rokas A, Pagel M, Stone GN (2002) Evolutionary shift between host oak section and host-plant organs in *Andricus* gallwasps. *Evolution*, **56**, 1821–1830.

Creer S (2007) Choosing and using introns in molecular phylogenetics. *Evolutionary Bioinformatics*, **3**, 99–108.

Das A, Mohanty S, Stephan W (2004) Inferring the population structure and demography of *Drosophila ananassae* from multilocus data. *Genetics*, **168**, 1975–1985.

Folmer O, Black M, Hoeh W, Lutz R, Vrijenhoek R (1994) DNA primers for amplification of mitochondrial cytochrome c oxidase subunit 1 from diverse metazoan invertebrates. *Molecular Marine Biology and Biotechnology*, **3**, 294–299.

Gifford ME, Larson A (2008) In situ genetic differentiation in a hispaniolan lizard (*Ameiva chrysolaema*): a multilocus perspective. *Molecular Phylogenetics and Evolution*, **49**, 277–291.

Godfray HJC (1994) *Parasitoids. Behavioural and Evolutionary Ecology*. Princeton University Press, Princeton, New Jersey.

Haddrill PR, Thornton KR, Charlesworth B, Andolfatto P (2005) Multilocus patterns of nucleotide variability and the demographic and selection history of *Drosophila melanogaster* populations. *Genome Research*, **15**, 790–799.

Halligan DL, Keightley PD (2006) Ubiquitous selective constraints in the *Drosophila* genome revealed by a genome-wide interspecies comparison. *Genome Research*, **16**, 875–884.

Hayward A, Stone GN (2005) Oak gall wasp communities: Evolution and ecology. *Basic and Applied Ecology*, **6**, 435–443.

Hey J, Nielsen R (2004) Multilocus methods for estimating population sizes, migration rates and divergence time, with applications to the divergence of *Drosophila pseudoobscura* and *D. persimilis*. *Genetics*, **167**, 747–760.

Jennings WB, Edwards SV (2005) Speciational history of Australian grass finches (*Poephila*) inferred from thirty gene trees. *Evolution*, **59**, 2033–2047.

Kalendar R, Lee D, Schulman AH (2009) FastPCR software for PCR primer and probe design and repeat search. *Genes, Genomes and Genomics*, **3**, 1–14.

Lee JY, Edwards SV, Webster M (2009) Divergence across Australia's Carpentarian barrier: statistical phylogeography of the red-backed fairy wren (*Malurus melanocephalus*). *Evolution*, **62**, 3117–3134.

Lessa EP (1992) Rapid surveying of DNA sequence variation in natural populations. *Molecular Biology and Evolution*, **9**, 323–330.

Li C, Riethoven JJ, Ma L (2010) Exon-primed intron-crossing (EPIC) markers for non-model teleost fishes. *BMC Evolutionary Biology*, **10**, 90. ISSN 1471-2148. doi: 10.1186/1471-2148-10-90.

Lohse K, Sharanowski B, Stone G (2010) Quantifying the Pleistocene history of the oak gall parasitoid *Cecidostiba fungosa*. *Evolution*, **64**, 2664–2681.

Machado CA, Robbins N, Gilbert MTP, Herre EA (2005) Critical review of host specificity and its coevolutionary implications in the fig/fig-wasp mutualism. *Proceedings of the National Academy of Sciences of the United States of America*, **102**(Suppl. 1), 6558–6565.

Mena-Correa J, Sivinski J, Anzures-Dadda A, Ramìrez-Romero R, Gates M, Aluja M (2009) Consideration of *Eurytoma sivinskii* (Gates and Grissell), a eurytomid (Hymenoptera) with unusual foraging behaviors, as a biological control agent of tephritid (Diptera) fruit flies. *Biological Control*, **53**, 9–17.

Nei M, Gojobori T (1986) Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Molecular Biology and Evolution*, **3**, 418–426.

Nicholls JA, Preuss S, Hayward A *et al*. (2010) Concordant phylogeography and cryptic speciation in two western Palaearctic oak gall parasitoid species complexes. *Molecular Ecology*, **19**, 592–609.

Nielsen R, Wakeley J (2001) Distinguishing migration from isolation: a Markov chain Monte Carlo approach. *Genetics*, **158**, 885–896.

Oliveira DCSG, Raychoudhury R, Lavrov DV, Werren JH (2008) Rapidly evolving mitochondrial genome and directional selection in mitochondrial genes in the parasitic wasp *Nasonia* (Hymenoptera: Pteromalidae). *Molecular Biology and Evolution*, **25**, 2167–2180.

Palumbi SR, Baker CS (1994) Contrasting population structure from nuclear intron sequence and mtDNA of humpback whales. *Molecular Biology and Evolution*, **11**, 426–435.

Papanicolaou A, Joron M, Mcmillan WO, Blaxter ML, Jiggins CD (2005) Genomic tools and cDNA derived markers for butterflies. *Molecular Ecology*, **14**, 2883–2897.

Peters JL, Zhuravlev YN, Fefelov I, Humphries EM, Omland KE (2008) Multilocus phylogeography of a holarctic duck: colonization of North America from Eurasia by gadwall (*Anas strepera*). *Evolution*, **62**, 1469–1483.

Rokas A, Nylander JA, Ronquist F, Stone GN (2002) A maximum-likelihood analysis of eight phylogenetic markers in gallwasps (Hymenoptera: Cynipidae): implications for insect phylogenetic studies. *Molecular Phylogenetics and Evolution*, **22**, 1055–7903.

Rozas J, Rozas R (1995) DnaSP, DNA sequence polymorphism: an interactive program for estimating population genetics parameters from DNA sequence data. *Bioinformatics*, **11**, 621–625.

Sha ZL, Zhu CD, Murphy RW, Huang DW (2007) *Diglyphus isaea* (Hymenoptera: Eulophidae): a probable complex of cryptic species that forms an important biological control agent of agromyzid leaf miners. *Journal of Zoological Systematics and Evolutionary Research*, **45**, 128–135.

Sharanowski BJ, Robbertse B, Walker J *et al.* (2010) Expressed sequence tags reveal Proctotrupomorpha (minus Chalcidoidea) as sister to Aculeata (Hymenoptera: Insecta). *Molecular Phylogenetics and Evolution*, **57**, 101–112.

Stone GN, Challis RJ, Atkinson RJ *et al.* (2007) The phylogeographical clade trade: tracing the impact of human-mediated dispersal on the colonization of northern Europe by the oak gallwasp *Andricus kollari*. *Molecular Ecology*, **16**, 2768–2781.

Stone GN, Hernandez-Lopez A, Nicholls JA *et al.* (2009) Extreme host plant conservatism during at least 20 million years of host plant pursuit by oak gallwasps. *Evolution*, **63**, 854–869.

Stone GN, Schönrogge K, Atkinson RJ, Bellido D, Pujade-Villar J (2002) The population biology of oak gall wasps (Hymenoptera: Cynipidae). *Annual Review of Entomology*, **47**, 633–668, doi: 10.1146/annurev.ento.47.091201.145247.

Tamura K, Dudley J, Nei M, Kumar S (2007) MEGA4: Molecular evolutionary genetics analysis (MEGA) software version 4.0. *Molecular Biology and Evolution*, **24**, 1596–1599.

Thompson JD, Higgins DG, Gibson TJ (1994) ClustalW – improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position specific gap-penalities and weight matrix choice. *Nucleic Acids Research*, **22**, 4673–4680.

Untergasser A, Nijveen H, Xiangyu R, Bisseling T, Geurts R, Leunissen JAM (2007) Primer3Plus, an enhanced web interface to primer3. *Nucleic Acids Research*, **35**, W71–W74.

Weiblen GD (2002) How to be a fig wasp. *Annual Review of Entomology*, **47**, 299–330.

Wilder JA, Hollocher H (2003) Recent radiation of endemic Caribbean *Drosophila* of the *dunni* subgroup inferred from multilocus DNA sequence variation. *Evolution*, **57**, 2566–2579.

Zhang DX, Hewitt GM (2003) Nuclear DNA analyses in genetic studies of populations: practice, problems and prospects. *Molecular Ecology*, **12**, 563–584.