ZooKeys 596: 129–141 (2016)
doi: 10.3897/zookeys.596.8399
http://zookeys.pensoft.net

DATA PAPER

A peer-reviewed open-access journal
ZooKeys
Launched to accelerate biodiversity research

PCR primers for 30 novel gene regions in the nuclear genomes of Lepidoptera

Niklas Wahlberg[1,2], Carlos Peña[1], Milla Ahola[1],
Christopher W. Wheat[3], Jadranka Rota[1,2]

[1] *Department of Biology, University of Turku, 20014 Turku, Finland* [2] *Department of Biology, Lund University, 223 62 Lund, Sweden* [3] *Department of Zoology, Stockholm University, 106 91 Stockholm, Sweden*

Corresponding author: *Niklas Wahlberg* (niklas.wahlberg@biol.lu.se)

Academic editor: *E.van Nieukerken* | Received 7 March 2016 | Accepted 27 May 2016 | Published 8 June 2016

http://zoobank.org/65505D7F-9350-4593-98EA-2012CDFA1320

Citation: Wahlberg N, Peña C, Ahola M, Wheat CW, Rota J (2016) PCR primers for 30 novel gene regions in the nuclear genomes of Lepidoptera. ZooKeys 596: 129–141. doi: 10.3897/zookeys.596.8399

Abstract

We report primer pairs for 30 new gene regions in the nuclear genomes of Lepidoptera that can be amplified using a standard PCR protocol. The new primers were tested across diverse Lepidoptera, including nonditrysians and a wide selection of ditrysians. These new gene regions give a total of 11,043 bp of DNA sequence data and they show similar variability to traditionally used nuclear gene regions in studies of Lepidoptera. We feel that a PCR-based approach still has its place in molecular systematic studies of Lepidoptera, particularly at the intrafamilial level, and our new set of primers now provides a route to generating phylogenomic datasets using traditional methods.

Keywords

Molecular systematics, Lepidoptera, phylogenomics, phylogenetics

Introduction

Post-Sanger sequencing technologies have opened up vast possibilities for acquiring molecular data for inferring phylogenetic relationships among taxa using 100s to 1000s of loci (Lemmon and Lemmon 2013), from whole genome sequences (e.g. Jarvis et al. 2014), to whole transcriptome sequences (e.g. Misof et al. 2014), to the targeted capture of conserved regions in genomes (e.g. Prum et al. 2015). However, these approaches require high quality DNA or RNA extracted from samples that are fresh or have been stored appropriately. Unfortunately, this quality requirement fails to capitalize on the wealth of material collected and cataloged in museums around the world. Recognizing this, many researchers are currently attempting to develop protocols to allow the extraction and sequencing of large amounts of DNA from old museum samples (e.g. Timmermans et al. 2016), but these methods are primarily limited to mitochondrial DNA.

For the past two decades, the standard protocol in insect molecular systematics has been to extract genomic DNA from one or two legs of dried individuals, often several years old, generally yielding very low concentrations of DNA. Today, millions of such genomic DNA extracts exist, each taken from suboptimally stored specimens, generated by individual researchers and large facilities such as the Canadian Centre for DNA Barcoding. These extracts have been used to PCR amplify specific gene regions, followed by Sanger sequencing. This standard approach has traditionally been restricted to fewer than 10 gene regions due to the lack of universal primers for more regions. Given this extensive DNA resource and the inability of the aforementioned methods to be easily applied to them, here we present an approach for using these extracts in the pursuit of phylogenomic insights.

As DNA sequencing technologies continue to evolve, the molecular systematist must judiciously choose which tools are best suited to the questions they wish to address. While genome scale data are certainly useful, such data are expensive, difficult to analyze and ultimately only a small fraction is utilized. Perhaps most importantly, such large scale datasets are likely only necessary for resolving deeper evolutionary events, such as the relationships among orders of insects (Misof et al. 2014) or superfamilies of e.g. Lepidoptera (Bazinet et al. 2013; Kawahara and Breinholt 2014). In contrast, datasets on the order of tens of genes have been highly useful for resolving relationships at the intrafamilial level, as has been repeatedly shown for e.g. lepidopteran families (Wahlberg et al. 2009, 2014; Kaila et al. 2011; Kawahara et al. 2011; Sihvonen et al. 2011; Zahiri et al. 2011, 2012; Zwick et al. 2011; Regier et al. 2012a, 2012b; Rota and Wahlberg 2012; Sohn et al. 2013). Thus, datasets generated with PCR-based methods have been and continue to be very insightful. However, in many such studies, some nodes are poorly supported with the scale of data available and more sequence data is needed. But, while it would be very interesting to sequence e.g. transcriptomes for the same species sampled, financial and practical constraints preclude such attempts. Rather, what is most likely to help resolve many of these ambiguities in a cost effective and timely fashion are more high quality loci that can be amplified with PCR across a range of DNA quality.

Whole genome sequences can now be used to search for suitable gene regions for primer design (e.g. Wahlberg and Wheat 2008). Such suitable gene regions are considered to be protein-coding genes that are single copy and have an exon that is longer than 500 bp. Long exons are needed as intron lengths can vary thousand fold between taxa, sometimes even between close relatives (Zhang and Hewitt 2003). Protein-coding genes are also preferred for inferring phylogenetic relationships as their alignments are generally unambiguous and conserved regions can be found for primer design.

Here we design and test PCR primers for long exon regions of single copy, protein-coding genes across Lepidoptera based on publicly available whole genome sequences of the order. The new gene regions are shown to be phylogenetically informative for Lepidoptera and can be used to complement the eight gene regions that have become standard in Lepidoptera phylogenetics (Wahlberg and Wheat 2008).

Material and methods

Single copy, protein-coding genes with exons longer than 500 bp were found while manually curating the set of genes listed in Misof et al. (2014) that were pulled out of eight publicly available Lepidoptera genomes using *tblastn*: *Bombyx mori* (NCBI accession GCA_000151625), *Plutella xylostella* (GCA_000330985), *Manduca sexta* (GCA_000262585), *Danaus plexippus* (GCA_000235995), *Heliconius melpomene* (GCA_000313835), *Melitaea cinxia* (GCA_000716385), *Chilo suppressalis* (GCA_000636095), and *Spodoptera frugiperda* (GCA_000753635). Sequences from all eight genomes were then used to design universal primers. We used the Python library primer-designer v0.2.0 (Peña 2015) to submit batches of FASTA alignments to the website primer4clades (Contreras-Moreira et al. 2009) in order to retrieve candidate primers for each gene sequence. Primer selection was based on high quality and amplicon length between 200 and 500 bp.

As in Wahlberg and Wheat (2008), universal tails were added to all primers to facilitate sequencing. Primers were aliquoted to a standard concentration of 10 μM for use. Primers were tested on a set of 24 species of Lepidoptera that represent major lineages within the order (Table 1). The DNA extracts of these specimens were previously used in the study by Mutanen et al. (2010). They mainly come from small amounts of tissue (such as legs) preserved in 100% EtOH (details of preservation and extraction methods can be found in Mutanen et al. 2010). The PCR reactions for all samples were done using the MyTaq^{TM} HS Red Mix (Bioline) in a final volume of 12.5 μl per sample. For each reaction we used 4 μl of MQ-H_2O, 6.25 μl of 2x MyTaq HS Red Mix, 0.625 μl of both forward and reverse primers and 1 μl of extracted DNA. All primers were tested with a standard thermal cycling profile of 95 °C for 7 minutes, then 40 cycles of 94 °C for 30 seconds, 55 °C for 30 seconds, 72 °C for 2 minutes, with a final extension period of 72 °C for 10 minutes. A standard cycling profile was chosen to simplify procedures and allow for large scale testing. No optimization of PCR reactions was attempted, as the goal was to find primers that work under exactly the same conditions, allowing the efficient processing of large numbers of samples in the laboratory without having to keep track of specific protocols for specific primer pairs. Success of PCR was visualized on agarose gels and successful PCR products were cleaned enzymatically and sent to Macrogen Europe (Amsterdam) for Sanger sequencing.

Sequences were trimmed of primer sequences and aligned by eye with reference to amino acid sequence in BioEdit 7 (Hall 1999) using the sequence from *Bombyx mori* as a reference for each gene. Aligned sequences were stored and curated using VoSeq (Peña and Malm 2012).

Table 1. Taxa used to test the primers for amplifying the new gene regions. The last column summarizes the number of new gene regions sequenced for each specimen. See Suppl. material 1 for information about which gene regions were successful.

Voucher code	Family	Genus	Species	Number of new genes sequenced
MM00058	Micropterigidae	*Micropterix*	*aureatella*	11
MM00867	Nepticulidae	*Ectoedemia*	*occultella*	18
MM00943	Tischeriidae	*Tischeria*	*ekebladella*	18
MM02175	Psychidae	*Taleporia*	*tubulosa*	22
MM00030	Gracillariidae	*Gracillaria*	*syringella*	26
MM00306	Yponomeutidae	*Yponomeuta*	*evonymellus*	27
MM00510	Tortricidae	*Tortrix*	*viridana*	22
MM00014	Schreckensteiniidae	*Schreckensteinia*	*festaliella*	26
MM02524	Epermeniidae	*Epermenia*	*illigerella*	24
MM03096	Pterophoridae	*Stenoptilia*	*veronicae*	22
MM00913	Alucitidae	*Alucita*	*hexadactyla*	19
MM03941	Choreutidae	*Choreutis*	*pariana*	21
MM00021	Urodidae	*Wockia*	*asperipunctella*	17
MM00116	Cossidae	*Cossus*	*cossus*	28
MM00125	Sesiidae	*Synanthedon*	*scoliaeformis*	29
MM00312	Zygaenidae	*Adscita*	*statices*	26
MM00034	Hesperiidae	*Pyrgus*	*malvae*	24
MM00042	Elachistidae	*Ethmia*	*pusiella*	25
MM00051	Pyralidae	*Pyralis*	*farinalis*	24
MM00027	Drepanidae	*Thyatira*	*batis*	28
MM00032	Geometridae	*Cyclophora*	*punctaria*	26
MM00394	Endromidae	*Endromis*	*versicolora*	29
MM01170	Noctuidae	*Apamea*	*crenata*	27
MM02696	Lasiocampidae	*Poecilocampa*	*populi*	24

Results

We selected a total of 48 gene regions (see Supplementary material for alignments) for primer design, of which 30 successfully amplified (Suppl. material 1 and 2) with the designed primers (Table 3). Only two gene regions were successfully amplified from all 24 test samples (ArgKin and DDX23), but the majority were successfully amplified from 20 or more samples (Table 2). The least successful gene region was LeuZip, which was sequenced from only 9 samples. No samples amplified all 30 gene regions (Table 1); the average number of successful gene regions was about 23. The least successful sample was *Micropterix* (11 out of 30 gene regions sequenced), which is not surprising as the primer design was based on ditrysian species, while *Micropterix* is likely to be the sister group to all the rest of Lepidoptera (Kristensen et al. 2015; Regier et al. 2015). The new gene regions give a total of 11,043 bp of data. The average amplicon length is 368 bp (ranging from 178 to 729 bp).

Table 2. Basic information about the new gene regions amplified and sequenced in this study, along with the traditional eight genes used in many previous studies for comparison.

Gene name	Length (bp)	Number of specimens successful	Variable (%)	Pars. Inf. (%)	Conserved (%)	Freq. A (%)	Freq. T (%)	Freq. C (%)	Freq. G (%)	GeneID from *Bombyx* genome
AFG3a	336	22	39.3	37.5	60.7	28.0	27.7	20.2	24.1	BGIBMGA010088
AFG3b	300	11	47.3	39.7	52.7	34.9	20.9	20.7	23.6	BGIBMGA010088
ANK13C	330	20	49.1	38.8	50.9	33.0	28.5	16.4	22.2	BGIBMGA007536
ArgK	388	24	44.6	33.5	55.4	22.9	19.0	32.1	26.1	BGIBMGA005812
Ca-ATPase	444	23	37.2	30.2	62.8	24.9	21.0	30.1	24.0	BGIBMGA000408
Ca2	410	18	44.9	38.5	55.1	33.2	23.6	18.5	24.7	BGIBMGA006603
chitinase	405	18	47.2	40.5	52.8	25.7	27.4	23.8	23.2	BGIBMGA008709
Cullin5	327	22	48.3	41.0	51.7	33.4	28.7	17.5	20.5	BGIBMGA011511
CycY	375	18	39.7	35.5	60.3	29.9	31.4	17.2	21.6	BGIBMGA005969
DDX23	303	24	46.9	43.2	53.1	40.4	22.6	13.8	23.2	BGIBMGA003429
Exp1	729	15	43.6	35.8	56.4	31.4	28.2	19.5	21.0	BGIBMGA010657
FCF1	173	17	49.7	42.8	50.3	32.4	27.7	16.2	23.7	BGIBMGA010318
GLYP	384	14	52.3	44.0	47.7	27.2	24.8	25.1	22.9	BGIBMGA010361
KRR1	283	16	47.0	39.2	53.0	35.4	26.4	18.0	20.2	BGIBMGA005381
LeuZip	372	9	49.5	35.8	50.5	36.4	24.8	18.0	20.9	BGIBMGA003300
MK6	255	20	52.2	45.1	47.8	32.8	28.1	18.6	20.6	BGIBMGA005641
MMP41	285	21	56.5	48.1	43.5	31.1	30.6	19.7	18.6	BGIBMGA007574
MPP2	330	21	44.9	40.3	55.2	29.0	29.4	22.6	19.1	BGIBMGA008312
NC	573	15	48.9	39.6	51.1	32.2	29.1	17.0	21.7	BGIBMGA005035
Nex9	420	21	60.5	47.4	39.5	33.1	24.8	19.0	23.2	BGIBMGA001032
PolII	360	22	43.9	39.4	56.1	30.1	25.3	19.7	24.8	BGIBMGA004994
ProSup	432	22	58.8	47.5	41.2	25.6	27.8	21.0	25.6	BGIBMGA004645
PSb	366	23	54.4	45.9	45.6	24.8	23.9	26.7	24.7	BGIBMGA000201
SARAH	381	16	56.4	44.9	43.6	29.2	27.8	23.3	19.7	BGIBMGA011095
Ssu72	249	23	55.0	48.2	45.0	36.0	28.1	16.0	19.9	BGIBMGA000925

Gene name	Length (bp)	Number of specimens successful	Variable (%)	Pars. Inf. (%)	Conserved (%)	Freq. A (%)	Freq. T (%)	Freq. C (%)	Freq. G (%)	GeneID from *Bombyx* genome
TIF3Cb	324	13	50.6	40.1	49.4	24.7	22.1	28.9	24.3	BGIBMGA012851
TIF6	336	18	50.0	42.6	50.0	24.4	21.4	25.5	28.8	BGIBMGA009830
UDPG6DH	405	21	49.1	41.0	50.9	30.1	27.4	20.9	21.6	BGIBMGA012188
VPS4	432	15	40.7	35.4	59.3	28.9	28.9	20.1	22.1	BGIBMGA005930
WD40	339	21	42.5	38.6	57.5	30.1	31.4	19.3	19.2	BGIBMGA006243
Genes from Wahlberg and Wheat (2008) for comparison										
CAD	826	24	52.4	42.7	47.6	35.9	28.3	14.6	21.2	
COI	1476	23	44.4	33.0	55.6	31.1	40.0	14.9	14.0	
EF1a	1047	21	34.9	27.2	65.1	25.4	23.0	27.6	24.0	
GAPDH	691	12	38.9	30.8	61.1	23.6	25.8	27.3	23.3	
IDH	722	23	48.2	41.1	51.8	31.2	27.1	19.8	21.9	
MDH	407	23	47.9	41.3	52.1	27.4	25.8	22.7	24.1	
RpS5	603	20	38.5	34.3	61.5	25.4	24.9	24.4	25.3	
wingless	400	20	58.5	48.5	41.5	21.7	18.3	28.9	31.0	

Table 3. Primers for 30 new gene regions with universal tails (T7promoter-TAATACGACTCAC-TATAGGG to forward primers and T3-ATTAACCCTCACTAAAGGG to reverse primers) attached to the 5’ end. F = Forward, R = Reverse. Gene names from Table 2.

Gene	Primer
AFG3a_F	TAATACGACTCACTATAGGGTGTGAAGAAGCTAAGatwgaratyatggartt
AFG3a_R	ATTAACCCTCACTAAAGGGTGTTGTTGTATTAAAAccrtccatytchac
AFG3b_F	TAATACGACTCACTATAGGGTGCTCAAGACGACCtdaaraaratmac
AFG3b_R	ATTAACCCTCACTAAAGGGCCTGTACCTTCCACGaaytcytcrtamgt
ANK13C_F	TAATACGACTCACTATAGGGCAAATACAAAATTTTATATGGAAytdaartgggaytt
ANK13C_R	ATTAACCCTCACTAAAGGGGCAACTGTTTCTTTTCTAtcytcwcgraadatcca
ArgK_F	TAATACGACTCACTATAGGGyGAyCCsATCATyGAGGACTACCA
ArgK_R	ATTAACCCTCACTAAAGGGAGrTGGTCCTCCTCrTTGCACCAvAC
Ca2_F	TAATACGACTCACTATAGGGAAACAGTGGACtgyttgaaraarttcaayg
Ca2_R	ATTAACCCTCACTAAAGGGGGTGTGTTGTCGATGaaraayttrtgraa
Ca-ATPase_F	TAATACGACTCACTATAGGGGAAtacgarccbgaaatgggwaargt
Ca-ATPase_R	ATTAACCCTCACTAAAGGGcdccrtgrgcggggtcgttraagtg
chitinase_F	TAATACGACTCACTATAGGGGGTGGGTGCTtayttygtngaatgggg
chitinase_R	ATTAACCCTCACTAAAGGGTGTCCACAccrtcraaraayttcca
Cullin5_F	TAATACGACTCACTATAGGGTGTTAGTTAAAGATGCTTTTATGgaygaycchmg
Cullin5_R	ATTAACCCTCACTAAAGGGTCTTAACCATTCAaccatrtcytcttcyttytc
CycY_F	TAATACGACTCACTATAGGGgattatgayaartataatccwgaacayaaaca
CycY_R	ATTAACCCTCACTAAAGGGcattgcytcyaatttytgtgcyctttcytt
DDX23_F	TAATACGACTCACTATAGGGACAAAAGATAAAGAACGTgargargargchat
DDX23_R	ATTAACCCTCACTAAAGGGTGATCTTTTTCAgaccartghckrtcatccca
Exp1_F	TAATACGACTCACTATAGGGgthaataaaytdtttgaattyatgcatga
Exp1_R	ATTAACCCTCACTAAAGGGggrtaytcttcaaartctttrttdatcat
FCF1_F	TAATACGACTCACTATAGGGACTGGACATCGtdcarartatgatggayt
FCF1_R	ATTAACCCTCACTAAAGGGTTGTAGCCACGATGtarcayttrtgytg
GLYP_F	TAATACGACTCACTATAGGGACTGCGACAAGAAtayttyatgtgygcbgc
GLYP_R	ATTAACCCTCACTAAAGGGTTCACTCGTTTTTCACCTtcytcytcdat
KRR1_F	TAATACGACTCACTATAGGGaatgcktggrctatgaaratwcc
KRR1_R	ATTAACCCTCACTAAAGGGtdataatrtcrcatccwatttcrtc
LeuZip_F	TAATACGACTCACTATAGGGTGCCTGTCACAAaaygaytggaaryt
LeuZip_R	ATTAACCCTCACTAAAGGGTTTGACCAGGGTTTttdgcrtarttraa
MK6_F	TAATACGACTCACTATAGGGTTAGAGAAGGTGATgtntggathtgyatgga
MK6_R	ATTAACCCTCACTAAAGGGTTCTTTCTGGTGCCATGtanggyttrca
MMP41_F	TAATACGACTCACTATAGGGGAAAACTGGGGTGCTAAagtdtayttyaaya
MMP41_R	ATTAACCCTCACTAAAGGGTCACTTTGtttttrttytchccaaawgtcat
MPP2_F	TAATACGACTCACTATAGGGCACTTCCGAATCccdtggttycartaycc
MPP2_R	ATTAACCCTCACTAAAGGGCCACAGCAGCTGTGtaytcyttdccraa
NC_F	TAATACGACTCACTATAGGGgatgaagaaaaycchaaraarttytt
NC_R	ATTAACCCTCACTAAAGGGacwatdgaccartggaarttcatdgc
Nex9_F	TAATACGACTCACTATAGGGTGCAACTGCAAgartttgtngaytggatg
Nex9_R	ATTAACCCTCACTAAAGGGCCCAGTCGTATTTAggytgbtcntcatacat
PolII_F	TAATACGACTCACTATAGGGCTGAAACACCTACAatggcbathgaytgggt
PolII_R	ATTAACCCTCACTAAAGGGGCTGTAGGGTTCCATttdgcrtgytcytt
ProSup_F	TAATACGACTCACTATAGGGGACAACAATCGACtggcayccnaayaa

Gene	Primer
ProSup_R	ATTAACCCTCACTAAAGGGCTGTCCAGTgactggaayttyttcatdgc
PSb_F	TAATACGACTCACTATAGGGGCTGGGAGCTACTggvtgytggtgygaya
PSb_R	ATTAACCCTCACTAAAGGGAGATGCAGTCTCCAGTGTAGatrtcdckytc
SARAH_F	TAATACGACTCACTATAGGGGAAGATGGTATGCCTAATAtwcaycchaayat
SARAH_R	ATTAACCCTCACTAAAGGGGTTCACCTTCTTCACGAggytcccadccna
Ssu72_F	TAATACGACTCACTATAGGGCAGCTGACAGACCTaaytgttaygarttygg
Ssu72_R	ATTAACCCTCACTAAAGGGCCGATTGTAGCTTCTtcrtgrttrtcytg
TIF3Cb_F	TAATACGACTCACTATAGGGGAAAAATCGACCACCTGtaytayaarttyga
TIF3Cb_R	ATTAACCCTCACTAAAGGGGCCAGCAGTTCTTTAggyttnccvgtcatca
TIF6_F	TAATACGACTCACTATAGGGCTGTGCGAGTGcarttygaraayaataa
TIF6_R	ATTAACCCTCACTAAAGGGTGTGTCAGCCAGGatytcytchgtrtc
UDPG6DH_F	TAATACGACTCACTATAGGGCAGGAACTGTGTtgggtvtaygarcaytg
UDPG6DH_R	ATTAACCCTCACTAAAGGGTCTTGTGTCGCCTgtrttyttyttraa
WD40_F	TAATACGACTCACTATAGGGGATCCACTTCACAcaygcyaaraayac
WD40_R	ATTAACCCTCACTAAAGGGCCTgtccartcacaytcyttytcttg
VPS4_F	TAATACGACTCACTATAGGGTGATTCTGATGATCCAGAAaaraaraaryt
VPS4_R	ATTAACCCTCACTAAAGGGCATCCATATCAttvccdacaccttgcatytg

The variability in the new gene regions appears to be similar to the widely used nuclear gene regions reported in Wahlberg and Wheat (2008) (Table 2). The base content across most of the fragments is fairly even, with some of them having a small AT bias (e.g., Exp1, CycY, and TIF3Cb), but none having a larger percentage than for example CAD, one of the widely used gene regions (Wahlberg and Wheat 2008), which has 64.1% of As and Ts. The number of parsimony informative sites ranges from 30.2% in Ca-ATPase to a little over 48% in MMP41 and Ssu72. For comparison, the range of parsimony informative sites in the Wahlberg and Wheat (2008) gene regions is almost the same, from 27.2% (EF1alpha) to 48.5% (wingless).

Discussion

We report here primers for 30 new nuclear gene regions that can be used to complement existing molecular data for Lepidoptera systematics. Our primers were designed to amplify gene regions across the entire taxonomic array of Lepidoptera and to work on relatively degraded material by amplifying less than 500 bp segments of the genome. Many of these primers are being used successfully in our laboratory for projects on e.g. the nymphalid subfamily Limenitidinae (Dhungel and Wahlberg in prep.), the families Geometridae (Brehm et al. in prep.), Choreutidae (Rota et al. in prep.), Limacodidae (Dupont et al. in prep.) and Riodinidae (Seraphim et al. in prep.). The phylogenetic utility of the used gene regions will be reported in more detail in the forthcoming papers: in summary, they are providing similar resolution as the standard gene regions reported in Wahlberg and Wheat (2008) in preliminary maximum likelihood analyses with RAxML that have been conducted.

We would like to stress that the gene regions described here should be seen as complementary to the standard gene regions (Wahlberg and Wheat 2008) and could be used in the event that more data is needed. Potential users should consult Suppl. material 1 to see which primers worked for taxa they are interested in. Based on our experiences in the laboratory, we would recommend that researchers consider using primers for AFG3a, ArgKin, Ca-ATPase, DDX23, MMP41, MPP2, Nex9, PolII, ProSup, PSb, SSU72, UDPG6DH, and WD40, as these tend to amplify consistently, especially across Ditrysia.

More specifically, it seems that several fragments are not very suitable for nonditrysians (none of the three exemplars that we used amplified AFG3b, CHITINASE, KRR1, NC, SARAH, VPS4) and the utility of several other fragments for these groups needs to be further tested (Ca2, GLYP, MPP2, NEX9, POLII, TIF3CB, and UDPG6DH amplified in only one of the nonditrysians tested). On the other hand, 21 fragments amplified in four or more of the six exemplars of Macroheterocera (the exceptions being LeuZip, which amplified in only one of them, and TIF3Cb and VPS4, which amplified in three out of six). The situation is more complex across the lower ditrysians and apoditrysians, which can be expected since these groups are quite divergent (Mutanen et al. 2010; Regier et al. 2013). For these groups our recommendation is to try CHITINASE and MK6 in addition to the above-mentioned fragments that appear to work across Lepidoptera.

In this study, we have used traditional Sanger sequencing to acquire the DNA sequences. However, almost all of the amplicons are short enough to be multiplexed and sequenced on a NextGen sequencing platform, such as Illumina. The advantages would be quick generation of a large number of sequences for a large number of samples. On the other hand, many systematists do not have access to NextGen sequencers, or the bioinformatics knowhow to process the raw data into useable formats, in which case the traditional PCR-based Sanger sequencing approach is still appropriate.

The approach we have used is highly conservative, as we sought to find primer pairs that work under standard conditions. It would thus be possible to design primers for the 18 gene regions that did not work under our strict criteria, but would work under different conditions. It is also possible to design primers that would amplify a longer segment of DNA, although such primer pairs would require fresh samples with little degradation of genomic DNA. It would also be possible to find more gene regions with exon lengths more than 500 bp, although a PCR-based approach becomes less and less efficient as the number of reactions grows. It is quite likely that datasets comprising up to 20 gene regions are sufficient for most phylogenetic studies within families (Zwick et al. 2011). For more difficult phylogenetic problems, NextGen sequencing approaches are recommended.

Acknowledgements

We thank Andrew Mitchell for critical comments on the manuscript. This study was funded by a grant from Kone Foundation to JR and the Academy of Finland to NW.

References

Bazinet AL, Cummings MP, Mitter KT, Mitter CW (2013) Can RNA-Seq resolve the rapid radiation of advanced moths and butterflies (Hexapoda: Lepidoptera: Apoditrysia)? An exploratory study. PLoS ONE 8: e82615. doi: 10.1371/journal.pone.0082615

Contreras-Moreira B, Sachman-Ruiz B, Figueroa-Palacios I, Vinuesa P (2009) primers4clades: a web server that uses phylogenetic trees to design lineage-specific PCR primers for metagenomic and diversity studies. Nucleic Acids Research 37: W95–W100. doi: 10.1093/nar/gkp377

Hall TA (1999) BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucl Acids Symp Ser 41: 95–98.

Jarvis ED, Mirarab S, Aberer AJ, Li B, Houde P, Li C, Ho SYW, Faircloth BC, Nabholz B, Howard JT, Suh A, Weber CC, da Fonseca RR, Li JW, Zhang F, Li H, Zhou L, Narula N, Liu L, Ganapathy G, Boussau B, Bayzid MS, Zavidovych V, Subramanian S, Gabaldon T, Capella-Gutierrez S, Huerta-Cepas J, Rekepalli B, Munch K, Schierup M, Lindow B, Warren WC, Ray D, Green RE, Bruford MW, Zhan XJ, Dixon A, Li SB, Li N, Huang YH, Derryberry EP, Bertelsen MF, Sheldon FH, Brumfield RT, Mello CV, Lovell PV, Wirthlin M, Schneider MPC, Prosdocimi F, Samaniego JA, Velazquez AMV, Alfaro-Nunez A, Campos PF, Petersen B, Sicheritz-Ponten T, Pas A, Bailey T, Scofield P, Bunce M, Lambert DM, Zhou Q, Perelman P, Driskell AC, Shapiro B, Xiong ZJ, Zeng YL, Liu SP, Li ZY, Liu BH, Wu K, Xiao J, Yinqi X, Zheng QM, Zhang Y, Yang HM, Wang J, Smeds L, Rheindt FE, Braun M, Fjeldsa J, Orlando L, Barker FK, Jonsson KA, Johnson W, Koepfli KP, O'Brien S, Haussler D, Ryder OA, Rahbek C, Willerslev E, Graves GR, Glenn TC, McCormack J, Burt D, Ellegren H, Alstrom P, Edwards SV, Stamatakis A, Mindell DP, Cracraft J, Braun EL, Warnow T, Jun W, Gilbert MTP, Zhang GJ (2014) Whole-genome analyses resolve early branches in the tree of life of modern birds. Science 346: 1320–1331. doi: 10.1126/science.1253451

Kaila L, Mutanen M, Nyman T (2011) Phylogeny of the mega-diverse Gelechioidea (Lepidoptera): Adaptations and determinants of success. Molecular Phylogenetics and Evolution 61: 801–809. doi: 10.1016/j.ympev.2011.08.016

Kawahara A, Ohshima I, Kawakita A, Regier JC, Mitter C, Cummings MP, Davis DR, Wagner DL, de Prins J, Lopez-Vaamonde C (2011) Increased gene sampling strengthens support for higher-level groups within leaf-mining moths and relatives (Lepidoptera: Gracillariidae). BMC Evolutionary Biology 11: 182. doi: 10.1186/1471-2148-11-182

Kawahara AY, Breinholt JW (2014) Phylogenomics provides strong evidence for relationships of butterflies and moths. Proceedings of the Royal Society of London B Biological Sciences 281: 20140970. doi: 10.1098/rspb.2014.0970

Kristensen NP, Hilton DJ, Kallies A, Milla L, Rota J, Wahlberg N, Wilcox SA, Glatz RV, Young DA, Cocking G, Edwards T, Gibbs GW, Halsey M (2015) A new extant family of primitive moths from Kangaroo Island, Australia and its significance for understanding early Lepidoptera evolution. Systematic Entomology 40: 5–16. doi: 10.1111/syen.12115

Lemmon EM, Lemmon AR (2013) High-throughput genomic data in systematics and phylogenetics. Annual Review of Ecology, Evolution, and Systematics 44: 99–121. doi: 10.1146/annurev-ecolsys-110512-135822

Misof B, Liu SL, Meusemann K, Peters RS, Donath A, Mayer C, Frandsen PB, Ware J, Flouri T, Beutel RG, Niehuis O, Petersen M, Izquierdo-Carrasco F, Wappler T, Rust J, Aberer AJ, Aspock U, Aspock H, Bartel D, Blanke A, Berger S, Bohm A, Buckley TR, Calcott B, Chen JQ, Friedrich F, Fukui M, Fujita M, Greve C, Grobe P, Gu SC, Huang Y, Jermiin LS, Kawahara AY, Krogmann L, Kubiak M, Lanfear R, Letsch H, Li YY, Li ZY, Li JG, Lu HR, Machida R, Mashimo Y, Kapli P, McKenna DD, Meng GL, Nakagaki Y, Navarrete-Heredia JL, Ott M, Ou YX, Pass G, Podsiadlowski L, Pohl H, von Reumont BM, Schutte K, Sekiya K, Shimizu S, Slipinski A, Stamatakis A, Song WH, Su X, Szucsich NU, Tan MH, Tan XM, Tang M, Tang JB, Timelthaler G, Tomizuka S, Trautwein M, Tong XL, Uchifune T, Walzl MG, Wiegmann BM, Wilbrandt J, Wipfler B, Wong TKF, Wu Q, Wu GX, Xie YL, Yang SZ, Yang Q, Yeates DK, Yoshizawa K, Zhang Q, Zhang R, Zhang WW, Zhang YH, Zhao J, Zhou CR, Zhou LL, Ziesmann T, Zou SJ, Li YR, Xu X, Zhang Y, Yang HM, Wang J, Wang J, Kjer KM, Zhou X (2014) Phylogenomics resolves the timing and pattern of insect evolution. Science 346: 763–767. doi: 10.1126/science.1257570

Mutanen M, Wahlberg N, Kaila L (2010) Comprehensive gene and taxon coverage elucidates radiation patterns in moths and butterflies. Proceedings of the Royal Society of London B Biological Sciences 277: 2839–2848. doi: 10.1098/rspb.2010.0392

Peña C (2015) primer-designer: Designs primers from FASTA files using the primers4clades website. https://github.com/carlosp420/primer-designer [accessed 14.5.2015]

Peña C, Malm T (2012) VoSeq: a Voucher and DNA Sequence Web Application. PLoS ONE 7: e39071. doi: 10.1371/journal.pone.0039071

Prum RO, Berv JS, Dornburg A, Field DJ, Townsend JP, Lemmon EM, Lemmon AR (2015) A comprehensive phylogeny of birds (Aves) using targeted next-generation DNA sequencing. Nature 526: 569–573. doi: 10.1038/nature15697

Regier JC, Brown JW, Mitter C, Baixeras J, Cho S, Cummings MP, Zwick A (2012a) A molecular phylogeny for the leaf-roller moths (Lepidoptera: Tortricidae) and its implications for classification and life history evolution. PLoS ONE 7: e35574. doi: 10.1371/journal.pone.0035574

Regier JC, Mitter C, Kristensen NP, Davis DR, van Nieukerken EJ, Rota J, Simonsen TJ, Mitter KT, Kawahara AY, Yen S-H, Cummings MP, Zwick A (2015) A molecular phylogeny for the oldest (non-ditrysian) lineages of extant Lepidoptera, with implications for classification, comparative morphology and life history evolution. Systematic Entomology 40: 671–704. doi: 10.1111/syen.12129

Regier JC, Mitter C, Solis MA, Hayden JE, Landry B, Nuss M, Simonsen TJ, Yen SH, Zwick A, Cummings MP (2012b) A molecular phylogeny for the pyraloid moths (Lepidoptera: Pyraloidea) and its implications for higher-level classification. Systematic Entomology 37: 635–656. doi: 10.1111/j.1365-3113.2012.00641.x

Regier JC, Mitter C, Zwick A, Bazinet AL, Cummings MP, Kawahara AY, Sohn J-C, Zwickl DJ, Cho S, Davis DR, Baixeras J, Brown J, Parr C, Weller S, Lees DC, Mitter KT (2013) A large-scale, higher-level, molecular phylogenetic study of the insect order Lepidoptera (moths and butterflies). PLoS ONE 8: e58568. doi: 10.1371/journal.pone.0058568

Rota J, Wahlberg N (2012) Exploration of data partitioning in an eight-gene dataset: phylogeny of metalmark moths (Lepidoptera, Choreutidae). Zoologica Scripta 41: 536–546. doi: 10.1111/j.1463-6409.2012.00551.x

Sihvonen P, Mutanen M, Kaila L, Brehm G, Hausmann A, Staude HS (2011) Comprehensive molecular sampling yields a robust phylogeny for geometrid moths (Lepidoptera: Geometridae). PLoS ONE 6: e20356. doi: 10.1371/journal.pone.0020356

Sohn JC, Regier JC, Mitter C, Davis D, Landry J-F, Zwick A, Cummings MP (2013) A molecular phylogeny for Yponomeutoidea (Insecta, Lepidoptera, Ditrysia) and its implications for classification, biogeography and the evolution of host plant use. PLoS ONE 8: e55066. doi: 10.1371/journal.pone.0055066

Timmermans MJT, Viberg C, Martin G, Hopkins K, Vogler AP (2016) Rapid assembly of taxonomically validated mitochondrial genomes from historical insect collections. Biological Journal of the Linnean Society 117: 83–95. doi: 10.1111/bij.12552

Wahlberg N, Leneveu J, Kodandaramaiah U, Peña C, Nylin S, Freitas AVL, Brower AVZ (2009) Nymphalid butterflies diversify following near demise at the Cretaceous/Tertiary boundary. Proceedings of the Royal Society of London B Biological Sciences 276: 4295–4302. doi: 10.1098/rspb.2009.1303

Wahlberg N, Rota J, Braby MF, Pierce NE, Wheat CW (2014) Revised systematics and higher classification of pierid butterflies (Lepidoptera: Pieridae) based on molecular data. Zoologica Scripta 43: 641–650. doi: 10.1111/zsc.12075

Wahlberg N, Wheat CW (2008) Genomic outposts serve the phylogenomic pioneers: designing novel nuclear markers for genomic DNA extractions of Lepidoptera. Systematic Biology 57: 231–242. doi: 10.1080/10635150802033006

Zahiri R, Holloway JD, Kitching IJ, Lafontaine JD, Mutanen M, Wahlberg N (2012) Molecular phylogenetics of Erebidae (Lepidoptera, Noctuoidea). Systematic Entomology 37: 102–124. doi: 10.1111/j.1365-3113.2011.00607.x

Zahiri R, Kitching IJ, Lafontaine JD, Mutanen M, Kaila L, Holloway JD, Wahlberg N (2011) A new molecular phylogeny offers hope for a stable family-level classification of the Noctuoidea (Lepidoptera). Zoologica Scripta 40: 158–173. doi: 10.1111/j.1463-6409.2010.00459.x

Zhang D-X, Hewitt GM (2003) Nuclear DNA analyses in genetic studies of populations: practice, problems and prospects. Molecular Ecology 12: 563–584. doi: 10.1046/j.1365-294X.2003.01773.x

Zwick A, Regier JC, Mitter C, Cummings MP (2011) Increased gene sampling yields robust support for higher-level clades within Bombycoidea (Lepidoptera). Systematic Entomology 36: 31–43. doi: 10.1111/j.1365-3113.2010.00543.x

Supplementary material 1

Table S1

Authors: Niklas Wahlberg, Carlos Peña, Milla Ahola, Christopher W. Wheat, Jadranka Rota

Data type: NCBI accession numbers

Explanation note: Details of the success of sequencing of the new gene regions. GenBank accession number indicates successful sequencing, dash indicates unsuccessful amplification.

Supplementary material 2

Sequences used for designing primers

Authors: Niklas Wahlberg, Carlos Peña, Milla Ahola, Christopher W. Wheat, Jadranka Rota

Data type: Reference sequences

Explanation note: A zip-file containing reference sequences for all 48 gene regions used for designing primers.