Running head: LEPIDOPTERA ANCHORED HYBRID ENRICHMENT

Resolving relationships among the megadiverse butterflies and moths with a novel pipeline for Anchored Phylogenomics

Jesse W. Breinholt[1,2,§], Chandra Earl [1], Alan R. Lemmon[3], Emily Moriarty Lemmon[4], Lei Xiao[1], Akito Y. Kawahara[1§]

[1]Florida Museum of Natural History, University of Florida, Gainesville, FL 32611, USA

[2]RAPiD Genomics, Gainesville, FL 32601, USA

[3]Dept. of Scientific Computing, Florida State University, Tallahassee, FL 32306, USA

[3]Dept. of Biological Science, Florida State University, Tallahassee, FL 32306, USA

[§]Corresponding authors

Email addresses:

JWB: jessebreinholt@gmail.com

CE: sunray1@ufl.edu

ARL: alemmon@fsu.edu

EML: chorusfrog@bio.fsu.edu

LX: lxiao@ufl.edu

AYK: kawahara@flmnh.ufl.edu

Abstract

The advent of next-generation sequencing technology has allowed for the collection of large portions of the genome for phylogenetic analysis. Hybrid enrichment and transcriptomics are two techniques that leverage next-generation sequencing and have shown much promise. However, methods for processing hybrid enrichment data are still limited. We developed a pipeline for anchored hybrid enrichment (AHE) read assembly, orthology determination, contamination screening, and data processing for sequences flanking the target "probe" region. We apply this approach to study the phylogeny of butterflies and moths (Lepidoptera), a megadiverse group of more than 157,000 described species with poorly understood deep-level phylogenetic relationships. We introduce a new, 855 locus anchored hybrid enrichment kit for Lepidoptera phylogenetics and compare resulting trees to those from transcriptomes. The enrichment kit was designed from existing genomes, transcriptomes and expressed sequence tag (EST) data and was used to capture sequence data from 54 species from 23 lepidopteran families. Phylogenies estimated from AHE data were largely congruent with trees generated from transcriptomes, with strong support for relationships at all but the deepest taxonomic levels. We combine AHE and transcriptomic data to generate a new Lepidoptera phylogeny, representing 76 exemplar species in 42 families. The tree provides robust support for many relationships, including those among the seven butterfly families. The addition of AHE data to an existing transcriptomic dataset lowers node support along the Lepidoptera backbone, but firmly places taxa with AHE data on the phylogeny. To examine the efficacy of AHE at different taxonomic levels, phylogenetic analyses were also conducted on a sister group representing a more recent divergence, the Saturniidae and Sphingidae. These analyses utilized sequences from the probe region and data flanking it, nearly doubled the size of the dataset; all resulting trees were well supported. We hope that our data processing pipeline, hybrid enrichment gene set, and approach of combining AHE data with transcriptomes will be useful for the broader systematics community.

Keywords: exon, hybrid enrichment, phylogenetics, sequence capture, target capture, transcriptomics, ultraconserved elements

Phylogenomics has significantly changed how we approach evolutionary questions. Studies that utilize less than ten genes are becoming less common; methods that collect hundreds or thousands of loci are now cost-efficient and have proven useful in constructing robust phylogenies. Although full genomes may be the ideal source of phylogenomic data, they remain unavailable for many non-model taxa. Two data collection approaches have risen to the forefront for use in deep-level phylogenomics: transcriptomics (e.g., Borner et al. 2014; Breinholt and Kawahara 2013; Garrison et al. 2016; González et al. 2015; Meusemann et al. 2010; Misof et al. 2014; Oakley et al. 2012; Simon et al. 2012) and hybrid enrichment (Cronn et al. 2012; Faircloth et al. 2012; Jones and Good 2016; Lemmon et al. 2012; Lemmon and Lemmon 2013; McCormack et al. 2013b).

Expressed mRNAs are used to generate transcriptomes, and therefore there is no need for the foreknowledge of targeted gene regions. Transcriptomic methods require fresh or properly stored tissue, which can severely limit the number of taxa that can be included in a phylogenetic study (Cronn et al. 2012; Lemmon and Lemmon 2013; McCormack et al. 2013b). Hybrid enrichment requires prior knowledge of the desired targets and use DNA probes to hybridize and selectively remove targets from a genome (Cronn et al. 2012; Jones and Good 2016; Lemmon and Lemmon 2013; McCormack et al. 2013b). Hybrid enrichment techniques also allow researchers to use ethanol-preserved tissues, stored DNA extractions (Faircloth et al. 2012; McCormack et al. 2013b), and in some cases, old museum specimens (Bi et al. 2013; Guschanski et al. 2013), increasing the potential number of taxa that can be included in a phylogenomic study.

Anchored phylogenomics is a hybrid enrichment approach developed to capture and enrich data from moderately conserved anchor regions from genomes of distantly related taxa (Lemmon et al. 2012; Lemmon and Lemmon 2013). Data are enriched by designing probes from several lineages and densely tiling probes across the anchored regions (Lemmon et al. 2012). Published studies using anchored phylogenomics have focused primarily on vertebrates (Brandley et al. 2015; Eytan et al. 2015; Lemmon et al. 2012; Lemmon and Lemmon 2013; Peloso et al. 2016; Prum et al. 2015; Pyron et al. 2014; Ruane et al. 2015), although a few recent studies have focused on arthropods, such as spiders (Hamilton et al. 2016) and flies (Young et al. 2016). Ultraconserved elements (UCEs) is another hybrid enrichment sequencing technique that has also been shown to work efficiently in some insect groups, such as Hymenoptera (Blaimer et al. 2016; Faircloth et al. 2015). Several reviews have examined the utility of transcriptomics and hybrid enrichment for phylogenomics (Cronn et al. 2012; Jones and Good 2016; Lemmon and Lemmon 2013; McCormack et al. 2013b), but these papers have not thoroughly compared the phylogenetic utility of hybrid enrichment and transcriptomic data on the same group of taxa.

Butterflies and moths (Lepidoptera) constitute one of the most speciose insect orders, with more than 157,000 described species (van Nieukerken et al. 2011). Lepidoptera include some of the most important model organisms for questions related to ecology and evolutionary biology (Roe et al. 2009); therefore, understanding their phylogenetic relationships is of fundamental importance. Lepidopteran species diversity is highest in the Ditrysia, a clade constituting approximately 98% of all described butterflies and moths (van Nieukerken et al. 2011). Until recently, a common set of eight to eleven mitochondrial and nuclear genes (Wahlberg and Wheat 2008) and a set of up to 27 protein-coding genes (Cho et al. 2011; Kawahara et al. 2011; Regier et al. 2013; 2015; Sohn et al. 2013; Zwick et al. 2011) were used as a standard for the majority of lepidopteran phylogenetic studies. However, studies that sought to resolve relationships among superfamilies, even with nearly 20 genes, were hampered by weak node support (Regier et al. 2013; 2015). Phylogenetic analyses using transcriptomes have begun to provide stronger support for these deep relationships (Bazinet et al. 2013; Bazinet et al. 2016; Breinholt and Kawahara 2013; Kawahara and Breinholt 2014). There is great promise in the use of transcriptomics for Lepidoptera phylogenetics, but phylotranscriptomics can be restricted by limitations associated with tissue freshness, preservation methods, and sequencing cost. Hybrid enrichment is less sensitive to tissue quality and quantity, is generally more cost-efficient, and therefore allows for the potential inclusion of a vast number of Lepidoptera specimens stored in ethanol, or kept dry in museum collections.

The focus of this study is to present an anchored hybrid enrichment (AHE) data processing pipeline for the broader systematics community, make available an AHE probe set for butterfly and moth phylogenetics, and compare the phylogenetic utility of AHE and transcriptomic data. The new probe set was used to sequence an exemplar set of 55 lepidopteran species, sampled from across the order. We evaluate the utility of the AHE data for phylogenomics across two taxonomic levels: across the entire order, the origin of which is estimated to be in the Jurassic, and on the Sphingidae and Saturniidae, a sister-group that is estimated to have a Paleogene orgin, approximately 50 mya (Misof et al. 2014). Because the AHE probe set targets exons from protein-coding genes, it can be included in existing transcriptomic datasets, and allows for an objective test of whether the addition of locus-rich transcriptomic data to a smaller AHE dataset can lead to a well-supported phylogeny.

MATERIALS AND METHODS

Probe Design

We targeted exons from Lepidoptera transcriptomes and genomes to create an anchored hybridization probe set (Bi et al. 2012; Hugall et al. 2015). We used HaMStR v8b (Ebersberger et al. 2009) with the InsectaHMMERv3-2 core ortholog set (1,579 core orthologs) setting *Bombyx mori* as the reference, and implementing the "-representative" option to search for orthologous genes in five Lepidoptera genomes that were available at the time of this study: *Bombyx mori* (Xia et al. 2004), *Danaus plexippus* (Zhan et al. 2011), *Heliconius melpomene* (The Heliconius Genome Consortium 2012), *Manduca sexta* (Agricultural Pest Genomics Resource Database, http://www.agripestbase.org), and *Plutella xylostella* (You et al. 2013). Orthologous genes were aligned by amino acids using TranslatorX (Abascal et al. 2010) and MAFFT v7.029b (Katoh and Standley 2013). Exon boundaries of genes in from these five model species were identified by mapping raw genomic reads to the corresponding transcriptome sequences using ShallowMapper4, a Java program written by ARL (Available from the Dryad Digital Repository: http://datadryad.org, doi:10.5061/dryad.355s2, http://dx.doi.org/10.5061/dryad.355s2). An initial set of probe loci were identified that fit within exons of these five species and follow the conservation and uniqueness properties, as defined by Lemmon et al. (2012). We generated reference kmers with the five model species' alignments to search for these genes in 23 transcriptomes and expressed sequence tags (ESTs). Quickscan5, a Java script by ARL (http://dx.doi.org/10.5061/dryad.355s2) used the kmers to map contig sequences from the 23 sequences to the candidate locus set. The 23 taxa were subsequently added to the design kit for a total of 28 "reference" taxa representing 22 lepidopteran families (taxon names and their corresponding GenBank SRA numbers are listed in Supplementary File 1: Table S1). We selected 855 loci for capture that were present in at least 70% of the reference taxa, resulting in an average of 650 loci per taxon (Supplementary File 1: Table S2). These loci had an average probe length of 254 bp and average pairwise similarity of 77% (Supplementary File 1: Table S3). The 855 loci correspond to 590 genes from *Bombyx mori* with one to seven anchored phylogenomic loci per gene (Supplementary File 1: Table S3). After final locus selection, we used a 3x density probe tiling strategy and included 57,138 probes in the SureSelect Target Enrichment XT kit (Agilent Technologies, Santa Clara, USA), largely following the method of Lemmon et al. (2012). Throughout this manuscript, we refer to this locus set as the "Lep1" probe set, which can be accessed from the Dryad Digital Repository (http://dx.doi.org/10.5061/dryad.355s2; Supplementary File 2: Probe design file). Three of the 23 transcriptomes used in the probe design were new and were generated and assembled following methods outlined in Kawahara and Breinholt (2014). They have been submitted to the SRA GenBank database, and have the following accession numbers: SRR1794084 (*Apatelodes pithala* [Apatelodidae], BioSample SAMN03333886), SRR1794032 (*Caloptilia triadicae* [Gracillariidae], BioSample SAMN03333594), and SRR1794082 (*Urbanus proteus* [Hesperiidae], BioSample

SAMN03333596). Assembled transcriptomes for these three taxa are available from the Dryad Digital Repository (http://dx.doi.org/10.5061/dryad.355s2).

Taxon sampling and data generation

This study included 110 taxa, including 55 samples sequenced for AHE, 45 transcriptomes, 5 genomes, and 5 EST samples (Supplementary File 1: Table S5). The 55 species sequenced using AHE included 1 trichopteran outgroup and 54 Lepidoptera species spanning 18 superfamilies and 22 families across the order (Supplementary File 1: Table S4). These taxa were chosen for AHE sequencing to examine the capture success of the Lep1 probe set and its efficacy in resolving deep and shallow relationships across Lepidoptera. Forty-three of the 54 Lepidoptera species belong in families that are represented by one of the 28 reference taxa; the remaining 11 species are in families without a reference taxon (Choreutidae, Dryadaulidae, Erebidae, Gelechiidae, Limacodidae, Micropterigidae, Neopseustidae, Papilionidae, Pieridae, Psychidae, and Tortricidae). Monophyly of Lepidoptera is firmly established by an impressive suite of morphological and molecular data (Regier et al. 2009; 2013; Kristensen et al. 2007; Misof et al. 2014). Thus, a single non-lepidopteran outgroup, *Hydropsyche rossi* (Trichoptera: Hydropsychidae) was enriched and sequenced with the Lep1 kit and was included in this study to root trees.

DNA was extracted using the OmniPrep Genomic DNA Extraction Kit (G-Biosciences: Catalog #786-136, St. Louis, MO, USA), from tissues that were collected and preserved in 100% EtOH and stored at -80ºC. Remaining tissue and wing vouchers for the specimens used in this study are stored at the McGuire Center for Lepidoptera &

Biodiversity, University of Florida, Gainesville, FL, USA, following our published protocol (Cho et al. 2016). DNA extracts were processed at Florida State University's Center for Anchored Phylogenomics (www.anchoredphylogeny.com), Tallahassee, FL, USA. Genomic DNA extractions were fragmented to 175-275 bp inserts using a Covaris sonicator. Illumina sequencing adapters containing 8 bp sample-specific indexes were ligated to these inserts (Lemmon et al. 2012). Samples were pooled into groups of 16 and the Lep1 Agilent Custom SureSelect kit was used to isolate regions of interest by hybridization and enrichment. Enriched libraries were sequenced on one lane of paired-end, 150 bp, Illumina HiSeq 2500. After sequencing, Illumina reads were de-multiplexed (separated by barcode indexes with zero mismatches tolerated).

Processing anchored hybid enrichment data

Anchored enrichment data were processed with an eight-step pipeline that uses established programs and a series of custom scripts written in Python (Fig. 1). The first six steps are outlined below. The final two steps are discussed in more detail in the 'Data matrix construction' section.

Clean raw reads (Fig. 1: step 1)

Paired-end raw Illumina reads were cleaned and adapters were removed using Trim Galore! ver. 0.4.0 (www.bioinformatics.babraham.ac.uk/projects/trim_galore/) allowing a minimum read size of 30 bp and trimming to remove bases with a Phred score below 20. (For the number of reads before and after quality trimming, see Supplementary File 1: Table S4).

Assembly (Fig. 1: step 2)

For each AHE locus, cleaned reads were assembled using an iterative baited assembly (IBA) process. Assembly was implemented with a custom Python script, IBA.py (http://dx.doi.org/10.5061/dryad.355s2) that uses USEARCH v7.0 (Edgar 2010) to select raw reads with high similarity to the probe region from the reference taxa. IBA then builds a *de novo* assembly with Bridger v2014-12-01 (Chang et al. 2015) from the selected reads. IBA uses the *de novo* assembly as the bait in subsequent IBA iterations to extend sequences outside of the probe region. IBA assembles each locus independently and screens each assembly to ensure it hits the targeted probe of the reference taxa. For each taxon that we sequenced using AHE, we used three IBA iterations with a kmer size of 25 and enforced a minimum of 10x kmer coverage of assembled sequences. Bridger assembled isoforms when an alternate splice path with the minimum kmer coverage was supported; these likely represent different copies of DNA in heterozygous individuals. Therefore, these isoforms were combined together to form a consensus sequence later in the pipeline.

Alignment (Fig. 1: step 3)

Assembled sequences of each locus were added to a reference taxon alignment using MAFFT v7.245 (Katoh and Standley 2013) with the commands, "–addlong" and "–adjustdirectionaccurately". These commands were implemented to adjust sequences in the opposite direction of the reference alignment and to add the assembled sequences that include data on either side of the probe region to the reference alignment. Alignments were trimmed to the probe region, using a custom Python script, extract_probe_region.py (http://dx.doi.org/10.5061/dryad.355s2) by splitting the alignment into three parts: the head, probe, and tail sections. This script uses the reference taxon included in the alignment to define the probe region. Throughout this manuscript, we refer to the head and tail data regions as the flanks.

Orthology (Fig. 1: step 4)

Sequences trimmed to the probe region (as described in step 3, above) were used to accurately determine the location of that sequence on the *Bombyx mori* genome. NCBI BLASTN (Camacho et al. 2009) was used to map sequences to the *B. mori* genome, allowing a maximum of three target hits (-max_target_seqs 3) and three hits per target (-max_hsps 3). To assure that each sequence had a single definitive hit to the genome, we compared bit scores of the blast hits for each sequence by filtering the blast results with a Python script (s_hit_checker.py: http://dx.doi.org/10.5061/dryad.355s2). A bit score that was ≥ 90% of the best bit score was considered too close to differentiate as a single hit to the genome, and that sequence was removed. We then re-blasted sequences with a single decisive hit to the *B. mori* genome using BLASTN to determine its location.

To ensure orthology, we used a Python script (ortholog_filter.py: http://dx.doi.org/10.5061/dryad.355s2) to select single hit sequences that mapped to the same location on the *B. mori* genome as the *B. mori* probe. Although our probe regions were exons, we chose genome mapping over transcriptome orthology programs that use proteomes to help determine orthology of transcriptomic data (e.g., orthoMCL (Li et al. 2003), HaMStR (Ebersberger et al. 2009), Orthograph (https://github.com/mptrsen/Orthograph)) for three reasons: 1) captured data are from the genome and not the proteome, therefore pseudogenes, nonfunctional duplications, and other genomic regions can be captured and sequenced, 2) if there is any cross-contamination from other samples sequenced on the same Illumina lane, these programs could assemble a single gene by combining exons from different samples, and 3) transcriptome orthology programs are designed to assemble coding genes, leading to the processing of exons without potentially valuable flanking data.

Alignment and isoform consensus (Fig. 1: step 5)

Sequences were split into locus-specific alignments using a Python script (split.py: http://dx.doi.org/10.5061/dryad.355s2) and then aligned with MAFFT. The isoforms generated by the Bridger assembler were turned into a single strict consensus sequence by processing alignments for each locus in FASconCAT-G (Kück and Longo 2014).

Contamination removal (Fig. 1: step 6)

To identify possible sequence contamination, we used USEARCH to blast sequences against themselves. We identified sequence pairs that were 99% identical across 95% of the total sequence length. We used a Python script (contamination_filter.py: http://dx.doi.org/10.5061/dryad.355s2) to parse the BLAST output and identify hits from sequence pairs belonging to distantly related families, and these sequences were removed. We then used a Python script (remove_duplicates.py: http://dx.doi.org/10.5061/dryad.355s2) to identify sequences for each taxon that had more than one sequence per locus. These sequences could be terminal duplications or contamination from taxa in the same family. To be conservative, these sequences were removed. Sequences that passed step 6 were considered orthologs. Only AHE loci that were represented in alignments by at least 75% of sampled taxa were included in the final datasets.

Datasets

We constructed six datasets in order to assess the phylogenetic utility of the Lep1 probe set (Fig. 2). Dataset completeness was estimated in ALISTAT (Misof et al. 2014). The six datasets were:

Dataset 1: acrossLEP_AHE (23 taxa, 557 loci [90,238 bp], 24% missing data): taxa included in this dataset were from lineages across Lepidoptera (plus the outgroup), consisting only of captured AHE data.

Dataset 2: acrossLEP_AHE+PARTtrans (75 taxa, 557 loci [90,238 bp], 25% missing data): taxa from lineages across Lepidoptera, consisting of 23 taxa with AHE sequence capture data (dataset 1), 26 Lep1 reference taxa (two reference taxa were excluded due to low locus coverage), and 26 transcriptomes of Kawahara and Breinholt (2014). The *Bicyclus anynana* sequence that was included in Kawahara and Breinholt (2014) was excluded from this dataset because many of its Lep1 loci could not be recovered from the limited amount of available EST data.

Dataset 3: acrossLEP_AHE+ALLtrans (76 taxa, 2,948 loci [2,522,806 bp], 69% missing data): taxa from across Lepidoptera; this dataset combined the 557 AHE loci with the reference sequences and the 2,696 gene transcriptome dataset of Kawahara and Breinholt (2014). These datasets were merged, and taxa captured for AHE data shared 305 loci with the 2,696 genes. The remaining 252 AHE loci that were absent from the larger datasets were taken from available raw read data, added to the dataset, and aligned.

Dataset 4: shallow_probe+flanks (48 taxa, 749 loci [281,241 bp], 18% missing data): taxa from Bombycoidea and relatives, consisting of loci from AHE data (35 taxa), the Lep1 probe set (7 taxa), and transcriptomes (6 taxa) from Kawahara and Breinholt (2014), for a total of 749 loci.

Dataset 5: shallow_probe (48 taxa, 749 loci [166,766 bp], 19% missing data): the same taxa and loci as dataset 4, but without data flanking the probe region.

Dataset 6: shallow_flanks (35 taxa, 749 loci [114,475 bp], 12% missing data): the 35 AHE taxa and loci from dataset 4, but excluding data from the probe region. *Bombyx mori* and thirteen taxa that were sequenced for transcriptomes were excluded from this dataset because they lacked sequence flanking the probe region.

Data matrix construction

In this section, we detail how each dataset was constructed. Summary statistics for each dataset are provided in Table 1; additional details can be found in Supplementary File 1: Tables S8 and S9. Pairwise identity values listed in the supplementary tables were calculated with Geneious v 8.0.3 (Biomatters 2014).

Dataset 1: acrossLEP_AHE: Using sequences that passed steps 1-6 of the pipeline, full-length assemblies were collected for each locus and then aligned to reference sequences using MAFFT (with the "–addlong" and "–adjustdirectionaccurately" functions) (Fig. 1: step 7). For each locus, we visually screened and inspected the level of conservation of sequences outside the probe region. We found that across Lepidoptera, there was little conservation outside of the probe region; thus, alignments were trimmed to the probe region for phylogenomic analysis (Fig. 1: step 8a).

Dataset 2: acrossLEP_AHE+PARTtrans: This dataset combined AHE data, sequences from Lep1 reference taxa, and transcriptomes from Kawahara and Breinholt (2014). This dataset was created to assess the effect of nearly tripling taxon sampling for the same set of 557 Lep1 loci that were in dataset 1. In order to add sequences to dataset 1 from taxa in the study of Kawahara and Breinholt (2014), we began with raw Illumina transcriptome reads. We assembled each AHE locus using a modified version of IBA, which assembled and trimed loci to the probe region (IBA_trans.py: (http://dx.doi.org/10.5061/dryad.355s2). After assembly, these data were processed in the pipeline (steps 1-6) and added to dataset 1.

Dataset 3: acrossLEP_AHE+ALLtrans: This dataset included 557 AHE Lep 1 loci, sequences from Lep1 reference taxa, and the 2,696-locus transcriptomes from Kawahara and Breinholt (2014). In total, 305 of the 557 AHE-captured Lep1 loci overlapped with the transcriptomes. To merge datasets, AHE loci were added to the transcriptome alignments using MAFFT, with the "–add" function. When both a transcriptome and

AHE sequence was present from the same taxon, we made a strict consensus for that taxon in FASconCAT-G. The remaining 252 AHE loci and transcriptomic loci of Kawahara and Breinholt (2014) were concatenated to the dataset using FASconCAT-G.

Dataset 4: shallow_probe+flanks: This dataset contained bombycoid and lasiocampid sequence data from three sources: 749 AHE captured Lep 1 loci, the reference sequences used to create the Lep1 probe set, and transcriptomes from Kawahara and Breinholt (2014). Since the probe region was designed to fit within exons, the flanking regions were nearly all introns. Across Lepidoptera, the introns were mostly unalignable, but among bombycoids, these flanking regions were relatively conserved. Incorporating these regions into a data matrix can be challenging because the targeted exons can be located at different places in each genome, can contain transposable elements, and because other non-conserved insertions can be abundant. MAFFT was used to align conserved data and two custom Python scripts were used to remove unalignable flanking regions from AHE sequences. For sequences that passed steps 1-6, including isoforms, the MAFFT commands "--allowshift --unalignlevel 0.8 --reorder –leavegappyregion" was used to produce a global alignment that allows alignable regions to be located (Fig. 1: step 7). The custom script (alignment_DE_trim.py: http://dx.doi.org/10.5061/dryad.355s2) was used to trim alignment columns according to density (number of sequences with data/total number of sequences in the alignment) and entropy (based on nucleotide entropy that ranges from 0-2, estimated with equation 1 of Xia et al. (2003)). This script was used to remove columns with < 60% density (set low to account for reference taxa and transcriptomes that only have probe region data), and sites that are estimated to be nearly random with entropy > 1.5 (Fig. 1: step 8b). Using a second script (flank_dropper.py: http://dx.doi.org/10.5061/dryad.355s2), problematic flanking sequences were removed (Fig. 1: step 8b). This script generates a 50% consensus sequence of the alignment and estimates a basic distance for the head and tail regions, separately, by scoring each position as a match (0), gap (0), or mismatch (1) to the consensus sequence. Scores for each taxon were normalized by the total number of gaps it has in that region. The script turns sequence data in the head and tail regions into gaps if the score is above the set number of standard deviations above the mean (i.e., very different from the consensus). Two standard deviations above the mean was used as a numerical cutoff for both head and tail regions. FASconCAT-G was used to make a strict consensus sequences of isoforms.

Dataset 5: shallow_probe: This dataset was constructed by trimming dataset 4 to the probe region (i.e., removing the flanking head and tail regions; Fig. 1: step 8a) with a custom Python script (Extract_probe_region.py: http://dx.doi.org/10.5061/dryad.355s2).

Dataset 6: shallow_flanks: After running extract_probe_region.py on dataset 5, output files from the head and tail regions of all loci were concatenated to construct dataset 6. (Dataset 6 consisted only of the flanking regions from the 35 target captured taxa from dataset 4). Since the 13 taxa from dataset 4 were from transcriptomic data only, they did not have intronic flanking sequences, and were excluded.

Synonymous and non-synonymous signal

Previous phylogenetic studies have shown that nucleotide saturation can cause misleading results (Betancur et al. 2013; Breinholt and Kawahara 2013; Regier et al. 2009; 2013; Soltis et al. 2002; Song et al. 2010; Zwick et al. 2012). To examine the extent of saturation for Lep1 loci, we made a saturation plot for each codon position in dataset 2 with DAMBE v5.3.16 (Xia et al. 2003) (Fig. S1). This plot shows that the first codon position is partially saturated and the third codon position is significantly saturated across all Lep1 loci. For nucleotide datasets 1-3, the degen v1.4 Perl script (Zwick et al. 2012) was used to exclude synonymous signal that can contribute to saturation, and the third codon position was removed. Nucleotide datasets with the synonymous signal excluded by the degen v1.4 script and the third codon position removed, are herein termed "degen12" datasets. Datasets 1-2 were also analyzed as amino acids. For dataset 3 (acrossLEP_AHE+ALLtrans), ALISCORE v. 2.0 (Kück et al. 2010; Misof and Misof 2009) and ALICUT v. 2.2 (Kück 2011) were run prior to running the degen v1.4 Perl script (Zwick et al. 2012). Due to the size and computation time required to analyze dataset 3, this dataset was not analyzed as amino acids. Amino acids were also not analyzed for datasets 4-6 (Table 1) since these datasets include both coding (probe) and non-coding regions (flanks).

Model selection and phylogenetic analysis

Due to the large size of the data matrices and the number of data partitions, we limited phylogenetic analyses to maximum likelihood (ML) and a coalescent-based species-tree method (ASTRAL) that can efficiently handle large datasets. For ML analyses, each dataset was partitioned by site entropy using the k-means algorithm in

PartitionFinder, using the commands "--raxml --kmeans entropy --all-states --min-subset-size 1000" (Frandsen et al. 2015; Lanfear et al. 2012). Partitioning-by-site was chosen because it is significantly faster than other methods that estimate partitions among loci (i.e., the rcluster method, (Lanfear et al. 2014)), and because initial partitioning tests resulted in nearly identical topologies for k-means and rcluster methods. The AICc score, as calculated in IQ-TREE (Prum et al. 2015), was used to find the optimal model for each partition estimated in PartitionFinder. Due to the matrix size of dataset 3, data were partitioned by codon position following the partitioning scheme of Kawahara and Breinholt (2014). For all ML analyses, IQ-TREE was used with the "–spp" option that allows partition-specific rates. To find the most likely tree, 100 separate ML searches were run, as well as 100 searches using the "–t RANDOM" function, which, after initial model optimization on a parsimony tree, uses 100 random tree topologies as starting trees for each search. One hundred non-parametric bootstrap replicates were initially calculated in IQ-TREE. We then used the "-I autoFC" option in RAxML to test whether each analysis fulfilled the bootstrap stopping criterion of Pattengale et al. (2009). If the bootstrap tree set failed to meet the criterion, an additional 50 replicates were estimated until it passed the bootstrap stopping criterion

We used the program ASTRAL v 4.7.3 (Mirarab et al. 2014) on datasets 2 and 4. ASTRAL is a coalescent-based species-tree method known to account for high levels of gene tree conflict due to incomplete lineage sorting. ASTRAL uses input gene trees and is scalable to very large datasets (Mirarab et al. 2014). The best model of evolution and 1000 ultrafast bootstrap approximation replicates (Minh et al. 2013) was calcuated for each gene in IQ-TREE before estimating an ASTRAL species tree. To calculate species-tree node support, 500 bootstrap replicates were run in ASTRAL using the site-only bootstrap estimation option (Seo 2008) by sampling from the ultrafast bootstrap replicate trees estimated in IQ-TREE.

RESULTS

Lepidoptera anchored hybrid enrichment probe set

The newly designed Lepidoptera probe set, Lep1, is available from the Dryad Digital Repository (http://dx.doi.org/10.5061/dryad.355s2). The Lep1 probe set had varying success in capturing target sequences across the 55 exemplar species. The number of loci captured ranged from 233 to 814 with an average of 728 loci (numbers estimated from IBA assemblies, Supplementary File 1: Table S4). Across Lepidoptera, the number of loci captured per taxon was correlated with the patristic tree distance to the nearest reference taxon and nearly significant to the number of loci of the nearest Lep1 reference taxon (multivariable least squares regression analysis, $P < 0\ .0001$, $R^2 = 0.68$, tree distance $p < 0.001$; number of loci in the nearest reference taxon, $p < 0.0504$). In general, species in lineages closer to the base of Lepidoptera yielded fewer loci (233 to 411) and species in the Ditrysia had higher capture success, ranging from 472 to 810 loci (Fig. 3).

The number of Lep1 loci found in the five genomes was high, as expected (the number of loci captured in parentheses, from highest to lowest): *Bombyx mori* (855), *Danaus plexippus* (852), *Manduca sexta* (841), *Heliconius melpomene* (805), and *Plutella xylostella* (728) (Supplementary File 1: Table S7). Success in assembling the Lep1 loci from the 48 transcriptomes and ESTs, including reference taxa in the Lep1 kit, varied from 243 to 843 loci with an average of 693 loci per taxon (Supplementary File 1: Table S7), and 40 of 48 species (83.3%) had ≥ 70% of the 855 targeted genes present (Supplementary File 1: Table S7). Two reference species (*Spodoptera frugiperda*, *Spodoptera littoralis*) included in the Lep1 probe design were not included in the phylogenetic analysis due to their low number of Lep1 loci.

Target capture success and dataset completeness

Dataset 1 (acrossLEP_AHE): The number of loci captured ranged from 193 to 469 with an average of 333 (Supplementary File 1: Table S8). The concatenated 557-locus alignment totaled 45,119 amino acid residues and the degen12 dataset was 90,238 bp (76% dataset completeness). Dataset 2 (acrossLEP_AHE+PARTtrans): Dataset 2 included the 557 Lep1 loci for 23 taxa as in dataset 1, but also included 48 reference taxa that had up to 557 loci taken from transcriptomes and genomes. The number of loci obtained from the five genomes were: *Bombyx mori* (557), *Danaus plexippus* (545), *Manduca sexta* (539), *Heliconius melpomene* (513), and *Plutella xylostella* (458) (Supplementary File 1: Table S8). The 48 reference taxa had 144 to 557 Lep1 loci with an average of 451 per taxon (Supplementary File 1: Table S8). The concatenated 557-locus alignment totaled 45,119 amino acid residues and the degen12 dataset was 90,238 bp in length (75% dataset completeness). Dataset 3 (acrossLEP_AHE+ALLtrans): Dataset 3 totaled 2,948 loci for 76 taxa, which included the 557 Lep1 loci from dataset 1 and 2,696 loci from Kawahara and Breinholt (2014). The concatenated 2,948-locus degen12 alignment totaled 2,522,806 bp (31% dataset completeness). Dataset 4 (shallow_probe+flanks): This dataset contained 48 bombycoids and the number of loci captured ranged from 647 to 735 with an average of 708 loci (Supplementary File 1: Table S9). The concatenated 749-locus alignment totaled 281,241 bp (82% dataset completeness). The majority of the missing data were from the transcriptomes lacking data outside of the probe region; the transcriptome sequences had 42-58% missing data. Taxa sequenced using the Lep 1 kit for AHE had fewer missing data (2-17%), excluding the lasiocampid outgroup, which had 22% of its data missing. Dataset 5 (shallow_probe region): The dataset 5 alignment totaled 166,766 bp (81% dataset completeness). Taxa represented by transcriptome sequences were missing 3-30% of its data; taxa sequenced using the Lep 1 kit for AHE only had 2-12% missing data (Supplementary File 1: Table S9). Dataset 6 (shallow_flanks): Dataset 6 was 114,475 bp in sequence length, and had 88% dataset completeness. Taxa sequenced using the Lep 1 kit for AHE only had 4-24% missing data, excluding the lasiocampid outgroup, which had 38% missing data (Supplementary File 1: Table S9).

Partitioning, model selection, and phylogenetic analyses

PartitionFinder divided nucleotide datasets into the following number of partitions: dataset 1 (8 partitions), dataset 2 (10 partitions), dataset 4 (32 partitions), and datasets 5 and 6 (24 partitions each). For amino acid datasets, dataset 1 was divided into 8 partitions and dataset 2 was split into 11 partitions (Supplementary File 1: Table S10). Due to its size, dataset 3 was partitioned by the first and second codon position, following Kawahara and Breinholt (2014). Models chosen by IQ-TREE using the AICc are reported in Supplementary File 1: Table S10. The bootstrap stopping criterion of Pattengale et al.

(2009) calculated 150-400 bootstrap replicates to be sufficient for all datasets (Supplementary File 1: Table S11).

The ML analyses for dataset 1 did not provide strong support for inter-superfamilial relationships and both the nucleotide and amino acid datasets resulted in similar tree topologies (Figs. 4a, S2). When the 557 AHE loci from 52 transcriptomes were added to dataset 1, the expanded taxon set (dataset 2) resulted in an ML tree that had higher support towards the tips of the tree and along the Lepidoptera backbone (only three nodes with < 70% BP along the backbone) (Figs. 4, S3).

While the expanded taxon sampling of dataset 2 improved bootstrap support for many nodes (Figs. 4b, 5a), it left key nodes at deep parts of the tree weakly supported (e.g., Callidulidae + Thyrididae, Dalceridae + Limacodidae + Megalopygidae, Cossidae, Gelechioidea (Gelechiidae + Lecithoceridae), Papilionoidea, Pterophoridae + Urodidae). Many deep splits in the ASTRAL species tree (Fig. S4) were also poorly supported (BS < 50%). The ML and species-tree analyses of dataset 2 had no nodes that conflicted with high support (BS ≥ 80%; Figs. 4b, S4). However, the analysis of dataset 2 resulted in an ML tree with strong support for Macroheterocera (BP ≥ 90% for all but 7 nodes) and its interfamilial relationships, including strong support for Mimallonidae as the sister group to this clade (BS = 99%; Fig. 4b). Relationships among butterfly families were also well supported (all but two nodes had BS = 100%; Figs. 4b, 5a, S2, S3).

ML analyses of dataset 3 resulted in trees that placed taxa with AHE data only in positions that were largely congruent with phylogenies from previously published studies 2011(e.g. Bazinet et al. 2013; Kawahara and Breinholt 2014; Mutanen et al. 2010; Regier et al. 2013). Trees from dataset 3 had generally higher support for relationships across the backbone compared to trees from datasets 1 and 2. However, node support along the backbone was lower by comparison to the 2,696 locus, transcriptome phylogeny of Kawahara and Breinholt (2014) (Figs. 4ab, 5ab).

ML trees from all "shallow" Bombycoidea analyses (datasets 4-6) had generally high bootstrap support, with all but five nodes with > 90% BS (Fig. 6a-c). The ML analyses from these datasets all provided strong support for the sister-group relationship of Saturniidae and Sphingidae, with Bombycidae as the sister-group to this clade (Fig. 6a-c). The ASTRAL analysis of dataset 4 (Fig. S5) also supported Bombycidae as the sister group to Saturniidae + Sphingidae (BS ≥ 98%). For ML and ASTRAL analyses, the majority of nodes in Sphingidae and Saturniidae was well supported. Macroglossinae was monophyletic (BS = 100%) and *Langia* was the sister group to Sphinginae + remaining Smerinthinae (BS ≥ 99%). Most tribes and subtribes within Macroglossinae and Smerinthinae were paraphyletic (Figs. 6a-c, S5). For both the ML and ASTRAL analyses, Oxyteninae was the most distant saturniid subfamily in the tree and its placement was well supported (BS = 100%). Both analyses also supported the Hemileucinae as the sister group to Ceratocampinae, and Attacini as the sister group to Saturniini (Figs. 6a-c, S5). Overall, the ML and ASTRAL trees (Figs. 6a-c, S5) were not in conflict, although there were a few relationships that differed, such as the positions of two species pairs, *Manduca sexta* + *Ceratomia amyntor*, and *Mimas tiliae* + *Pachysphinx occidentalis*.

Discussion

The number of studies that use anchored phylogenomics has risen dramatically in the last several years, and there is a clear need for reliable data processing methods for the systematic biology community. The data processing pipeline in this study has been written in such a way that the scripts will work for any taxonomic group for which there is a genome available from a closely related taxon. The pipeline was designed to process AHE data but the approach could also be used to process data enriched using ultraconserved elements (UCEs), although an established pipeline is already publically available to process such data (see http://ultraconserved.org/#software). Our novel-processing pipeline in combination with publically available anchored probe sets make anchored phylogenomics more accessible to the phylogenomics community.

Our new Lep1 probe set worked well and captured > 600 genes for all but three species sampled in the diverse clade Ditrysia. The highest capture success came from taxa in the Macroheterocera and Papilionoidea. We predict that capture success was high for these two groups because they had the greatest number of reference transcriptomes and genomes that were included in the initial probe design. Exemplars from the Myoglossata captured at least 395 AHE loci, demonstrating that the kit can capture reasonably well across Lepidoptera. The correlation between capture efficiency and the phylogenetic proximity to the nearest reference taxon suggests that increasing lineage representation in the probe design could further enhance capture efficiency and utility of the Lep1 probe set. This new genomic resource offers the ability to do phylogenomic studies for most lepidopteran groups by facilitating the collection of hundreds of loci.

ML trees generated from the 557 AHE loci across-Lepidoptera datasets (datasets 1, 2) were unable to provide strong support for many nodes along the Lepidoptera backbone, but none of the well-supported nodes were in conflict with the better resolved transcriptome-based trees of Bazinet et al. (2013) and Kawahara and Breinholt (2014).

The multispecies coalescent analyses also did not provide strong support across the Lepidoptera backbone and had no strongly supported nodes in conflict with the ML trees from the same dataset (see Supplementary File 3 for further discussion of the multispecies coalescent results). While studies that utilized hybridization capture data in other taxonomic groups have resulted in many well-resolved phylogenies (Brandley et al. 2015; Eytan et al. 2015; Hamilton et al. 2016; Lemmon et al. 2012; McCormack et al. 2013a; Peloso et al. 2016; Prum et al. 2015; Pyron et al. 2014; Ruane et al. 2015; Smith et al. 2014; Young et al. 2016), the results obtained here indicate that more data loci and/or taxa may be needed to confidently resolve backbone relationships among superfamilies of Lepidoptera. Recent phylotranscriptomic studies also suggested much more data is needed to resolve difficult Lepidoptera relationships (Bazinet et al. 2013;2016; Breinholt and Kawahara 2013; Kawahara and Breinholt 2014) and the lack of strong support from the AHE loci is likely due to the amount of data that could be captured with the Lep1 probe set (see Supplementary File 3 for a comparison of data sets size and informative sites). Improving the Lep1 probe kit by increasing lineage representation may ameliorate the limitations for deep-scale studies allowing for a more efficient capture across Lepidoptera.

Analysis of dataset 3 consisting of AHE loci and the 2,696 gene phylotranscriptomic dataset from Kawahara and Breinholt (2014) resulted in a topology congruent with the transcriptome-based higher lepidopteran trees estimated by Kawahara and Breinholt (2014) and Bazinet et al. (2013). Taxa sampled for AHE loci only were generally placed with support in parts of the tree that were consistent with morphology-based hypotheses, but overall branch support for the backbone of the tree decreased when

AHE loci were added to the much larger transcriptomic dataset. This decreased support is consistent with previous studies that utilized a dataset with large blocks of missing data (Cho et al. 2011; Jiang et al. 2014; Kawahara et al. 2013; Kawahara et al. 2016; Ponce et al. 2015; Simmons 2012;2014; Zwick et al. 2011). Although support dropped across the backbone, most shallow relationships were well supported (Figs. 5a, b) and comparable to bootstrap support values of Kawahara and Breinholt (2014). These results show that leveraging AHE methods to increase taxon sampling and integrate data with a more sequence-rich backbone is feasible (Jiang et al. 2014) and can be used to place taxa confidently.

The Lepidoptera AHE data provided strong support for relationships within superfamilies (see Supplementary File 3 for a more thorough discussion on lepidopteran relationships). For instance, relationships among the seven butterfly families was well supported and consistent with published butterfly relationships based on a smaller set of genes (Heikkilä et al. 2012). Relationships among and within some of the most taxonomically diverse families in the Bombycoidea (i.e., Bombycidae, Saturniidae and Sphingidae, (defined as the SBS group, Regier et al. 2008), In Bombycoidea, AHE data were conserved well beyond the exons (probe region) and into introns (flanks), significantly increasing the amount of available data for phylogenetic analyses. For example, dataset 4, which included both the probe region and its flanks was 114,475 base pairs longer than dataset 5, which only included the probe region. Trees resulting from the probe region alone (Fig. 6b) and the flanking regions only (Fig. 6c) were highly congruent with each other and with the dataset that includes both regions (Fig. 6a). The few topological differences between these trees (Fig. 6a-c) appear to be attributable to differences in taxon sampling, as taxa from transcriptomes did not have any data in the flanking regions and were removed from dataset 6 (Fig. 6c). The high congruence in these three datasets (datasets 4-6, Fig. 6a-c) indicates that the processing scripts that trim the alignment by density and nucleotide entropy, and remove spurious data from flanking regions, successfully remove problematic data and can generate informative datasets with congruent signal from the probe and flanking regions.

In summary, we make the following suggestions and hope they will prove valuable to studies that utilize AHE data in the quest to assemble the many levels of the tree of life: 1) The pipeline described here will perform better when using a well-assembled genome. Since the pipeline relies on a genome for determining orthologous loci we encourage users to try several different reference genomes when available. We further suggest the quality and completeness of the genome assembly may be more important to a good orthology assessment than the evolutionary distance from the targeted group. 2) The single hit and genome mapping ortholog approach, such as the one presented here should be used to screen loci before including them in a target enrichment kit design. Such an approach would increase the effectiveness of our post-sequencing AHE pipeline and reduce the loss of loci due to gene duplication or very similar genomic regions in a reference genome. 3) Even if a set of AHE loci fail to estimate strong support for deep relationships they can be strategically combined with transcriptomes and genomes to “tile in” taxa for deep-level phylogenomic studies. Such an approach would allow the placement of taxa in phylogenomic trees when fresh material needed to generate a transcriptome is not available. One suggestion to account for the possible reduction in bootstrap support in combined datasets is to enforce a backbone constraint estimated soley from the transcriptomic data when trying to place taxa with AHE data. 4) Studies that focus on more recent evolutionary time scales should take advantage of data flanking the probe region. These regions can significantly increase the amount of data for analysis. The data flaking the probe region are known to be useful and informative at shallow divergences (Lemmon and Lemmon 2012) and could help increases accuracy and stability of estimated relationship. The data from the probe and the flanking regions can further be used to look at stability and reliability of the resulting phylogenies by analyzing them separately.

SUPPLEMENTARY MATERIAL

Supplementary material, including data files and/or online-only appendices, can be found in the Dryad data repository at http://datadryad.org, doi:10.5061/dryad.355s2 and TreeBASE https://treebase.org/base, http://purl.org/phylo/treebase/phylows/study/TB2:S20274.

ACKNOWLEDGEMENTS

We thank J. Daniels, D.R. Davis, T. Emmel, R. Fessenden, D. Halbritter, J. Hansen, J.E. Hayden, G. Hill, N. Homziak, P.R. Houlihan, I. Kitching, S.-M. Lee, L. Reeves, and K. Willmott for assisting with taxon sampling for this study. S. Epstein and FLMNH volunteers provided assistance with specimen preparation. M. Kortyna and H. Ralicki assisted with data collection at the Center for Anchored Phylogenomics at Florida State University. A. Bazinet, M. Cummings, M. Espeland, P. Frandsen, C. Hamilton, K. Kjer, D. McKenna, C. Mitter, K. Mitter, D. Plotkin, D. Triant and A. Zwick provided insightful comments. We thank the reviewers and editors for comments that significantly improved the manuscript. J. Kramer provided bioinformatic assistance. M. Gitzendanner, A. Moskalenko, and C. Taylor of the University of Florida High-Performance Computing Center for providing support and priority access to allow completion of the many large phylogenomic analyses, and The Florida State University Research Computing Center also provided computational assistance. The Danum Valley Field Centre (Sabah, Malaysia) and the Nouragues Research Station (French Guiana) provided facilities and resources that allowed for specimen collection.

FUNDING

This work was supported in part by NSF grants DEB-1354585 and IOS-1121739 to AYK, DEB-1557007 and DEB-1541500 to AYK and JWB, DEB-1120516 to EML, and IIP-1313554 to ARL and EML. The University of Florida, Florida Museum of Natural History, the National Geographic Society (grant numbers 9107-12 and 9686-15), and a Centre National de la Recherche Scientifique (CNRS) grant supported costs associated with fieldwork.

References

Abascal F, Zardoya R, Telford MJ. 2010. TranslatorX: multiple alignment of nucleotide sequences guided by amino acid translations. Nucleic Acids Res., 38:W7-W13

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.

Bazinet AL, Mitter KT, Davis DR, Van Nieukerken EJ, Cummings MP, Mitter C. 2016. Phylotranscriptomics resolves ancient divergences in the Lepidoptera. Syst. Entomol. doi:10.1111/syen.12217

Betancur RR, Li C, Munroe TA, Ballesteros JA, Orti G. 2013. Addressing gene tree discordance and non-stationarity to resolve a multi-locus phylogeny of the flatfishes (Teleostei: Pleuronectiformes). Syst. Biol., 62:763-785.

Bi K, Linderoth T, Vanderpool D, Good JM, Nielsen R, Moritz C. 2013. Unlocking the vault: next-generation museum population genomics. Mol. Ecol., 22:6018-6032.

Bi K, Vanderpool D, Singhal S, Linderoth T, Moritz C, Good J. 2012. Transcriptome-based exon capture enables highly cost-effective comparative genomic data collection at moderate evolutionary scales. BMC Genomics, 13:403.

Biomatters. 2014. Geneious v 8.0.3 Available from: http://www.geneious.com.

Blaimer BB, Lloyd MW, Guillory WX, Brady SG. 2016. Sequence capture and phylogenetic utility of genomic ultraconserved elements obtained from pinned insect specimens. PLoS ONE, 11:e0161531.

Borner J, Rehm P, Schill RO, Ebersberger I, Burmester T. 2014. A transcriptome approach to ecdysozoan phylogeny. Mol. Phylogenet. Evol., 80:79-87.

Brandley MC, Bragg JG, Singhal S, Chapple DG, Jennings CK, Lemmon AR, Lemmon EM, Thompson MB, Moritz C. 2015. Evaluating the performance of anchored hybrid enrichment at the tips of the tree of life: a phylogenetic analysis of Australian Eugongylus group scincid lizards. BMC Evol. Biol., 15:1-14.

Breinholt JW, Kawahara AY. 2013. Phylotranscriptomics: Saturated third codon positions radically influence the estimation of trees based on next-gen data. Genome Biol Evol, 5:2082-2092.

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009. BLAST+: architecture and applications. BMC Bioinformatics, 10:1-9.

Chang Z, Li G, Liu J, Zhang Y, Ashby C, Liu D, Cramer C, Huang X. 2015. Bridger: a new framework for de novo transcriptome assembly using RNA-seq data. Genome Biology, 16:30.

Cho S, Epstein SW, Mitter K, Hamilton CA, Plotkin D, Mitter C, Kawahara AY. 2016. Preserving and vouchering butterflies and moths for large-scale museum-based molecular research. PeerJ, 4:e2160.

Cho S, Zwick A, Regier JC, Mitter C, Cummings MP, Yao J, Du Z, Zhao H, Kawahara AY, Weller S, Davis DR, Baixeras J, Brown JW, Parr C. 2011. Can deliberately incomplete gene sample augmentation improve a phylogeny estimate for the advanced moths and butterflies (Hexapoda: Lepidoptera)? Syst. Biol., 60:782-796.

Cronn R, Knaus BJ, Liston A, Maughan PJ, Parks M, Syring JV, Udall J. 2012. Targeted enrichment strategies for next-generation plant biology. Am. J. Bot., 99:291-311.

Ebersberger I, Strauss S, von Haeseler A. 2009. HaMStR: Profile hidden markov model based search for orthologs in ESTs. BMC Evol. Biol., 9:157.

Edgar RC. 2010. Search and clustering orders of magnitude faster than BLAST. Bioinformatics, 26:2460-2461.

Eytan RI, Evans BR, Dornburg A, Lemmon AR, Lemmon EM, Wainwright PC, Near TJ. 2015. Are 100 enough? Inferring acanthomorph teleost phylogeny using Anchored Hybrid Enrichment. BMC Evol. Biol., 15:1-20.

Faircloth BC, Branstetter MG, White ND, Brady SG. 2015. Target enrichment of ultraconserved elements from arthropods provides a genomic perspective on relationships among Hymenoptera. Molecular Ecology Resources, 15:489-501.

Faircloth BC, McCormack JE, Crawford NG, Harvey MG, Brumfield RT, Glenn TC. 2012. Ultraconserved elements anchor thousands of genetic markers spanning multiple evolutionary timescales. Syst. Biol., 61:717-726.

Frandsen PB, Calcott B, Mayer C, Lanfear R. 2015. Automatic selection of partitioning schemes for phylogenetic analyses using iterative k-means clustering of site rates. BMC Evol. Biol., 15:1-17.

Garrison NL, Rodriguez J, Agnarsson I, Coddington JA, Griswold CE, Hamilton CA, Hedin M, Kocot KM, Ledford JM, Bond JE. 2016. Spider phylogenomics: untangling the Spider Tree of Life. PeerJ, 4:e1719.

González VL, Andrade SCS, Bieler R, Collins TM, Dunn CW, Mikkelsen PM, Taylor JD, Giribet G. 2015. A phylogenetic backbone for Bivalvia: an RNA-seq approach. Proceedings of the Royal Society of London B: Biological Sciences, 282.

Guschanski K, Krause J, Sawyer S, Valente LM, Bailey S, Finstermeier K, Sabin R, Gilissen E, Sonet G, Nagy ZnT, Lenglet G, Mayer F, Savolainen V. 2013. Next-generation museomics disentangles one of the largest primate radiations. Syst. Biol., 62:539-554.

Hamilton CA, Lemmon AR, Lemmon EM, Bond JE. 2016. Expanding anchored hybrid enrichment to resolve both deep and shallow relationships within the spider tree of life. BMC Evol. Biol., 16:212.

Heikkilä M, Kaila L, Mutanen M, Pena C, Wahlberg N. 2012. Cretaceous origin and repeated tertiary diversification of the redefined butterflies. Proc Biol Sci, 279:1093-1099.

Hugall AF, O'Hara TD, Hunjan S, Nilsen R, Moussalli A. 2015. An exon-capture system for the entire class Ophiuroidea. Mol. Biol. Evol., 33:281-294.

Jiang W, Chen S-Y, Wang H, Li D-Z, Wiens JJ. 2014. Should genes with missing data be excluded from phylogenetic analyses? Mol. Phylogenet. Evol., 80:308-318.

Jones MR, Good JM. 2016. Targeted capture in evolutionary and ecological genomics. Mol. Ecol., 25:185-202.

Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol. Biol. Evol., 30:772-780.

Kawahara A, Ohshima I, Kawakita A, Regier J, Mitter C, Cummings M, Davis D, Wagner D, 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 Evol. Biol., 11:1-14.

Kawahara AY, Breinholt JW. 2014. Phylogenomics provides strong evidence for relationships of butterflies and moths. Proceedings of the Royal Society B: Biological Sciences, 281.

Kawahara AY, Breinholt JW, Ponce FV, Haxaire J, Xiao L, Lamarre GPA, Rubinoff D, Kitching IJ. 2013. Evolution of Manduca sexta hornworms and relatives: Biogeographical analysis reveals an ancestral diversification in Central America. Mol. Phylogenet. Evol., 68:381-386.

Kawahara AY, Plotkin D, Ohshima I, Lopez-Vaamonde C, Houlihan PR, Breinholt JW, Kawakita A, Xiao LEI, Regier JC, Davis DR, Kumata T, Sohn J-C, De Prins J, Mitter C. 2016. A molecular phylogeny and revised higher-level classification for the leaf-mining moth family Gracillariidae and its implications for larval host-use evolution. Syst. Entomol.

Kristensen NP, Scoble M, Karsholt O. 2007. Lepidoptera phylogeny and systematics: the state of inventorying moth and butterfly diversity. In: Zhang ZQ, Shear WA editors. Linnaeus Tercentenary: Progress in Invertebrate Taxonomy, p. 699-747.

Kück P. 2011. ALICUT: a Perlscript which cuts ALISCORE identified RSS. Department 597 of Bioinformatics, Zoologisches Forschungsmuseum A. Koenig (ZFMK), Bonn, Germany, version 2.3.

Kück P, Longo GC. 2014. FASconCAT-G: extensive functions for multiple sequence alignment preparations concerning phylogenetic studies. Frontiers in Zoology, 11:1-8.

Kück P, Meusemann K, Dambach J, Thormann B, von Reumont BM, Wägele JW, Misof B. 2010. Parametric and non-parametric masking of randomness in sequence alignments can be improved and leads to better resolved trees. Frontiers in Zoology, 7:1-12.

Lanfear R, Calcott B, Ho S, Guindon S. 2012. PartitionFinder: combined selection of partitioning schemes and substitution models for phylogenetic analyses. Mol. Biol. Evol., 29:1695-1701.

Lanfear R, Calcott B, Kainer D, Mayer C, Stamatakis A. 2014. Selecting optimal partitioning schemes for phylogenomic datasets. BMC Evol. Biol., 14:82.

Lemmon AR, Emme S, Lemmon EM. 2012. Anchored hybrid enrichment for massively high-throughput phylogenomics. Syst. Biol., 61:727-744.

Lemmon AR, Lemmon EM. 2012. High-throughput Identification of Informative nuclear loci for shallow-scale phylogenetics and phylogeography. Syst. Biol., 61:745-761.

Lemmon EM, Lemmon AR. 2013. High-Throughput genomic data in systematics and phylogenetics. Annual Review of Ecology, Evolution, and Systematics, 44:99-121.

Li L, Stoeckert CJ, Roos DS. 2003. OrthoMCL: Identification of ortholog groups for eukaryotic genomes. Genome Res., 13:2178-2189.

McCormack JE, Harvey MG, Faircloth BC, Crawford NG, Glenn TC, Brumfield RT. 2013a. A Phylogeny of Birds Based on Over 1,500 Loci Collected by Target Enrichment and High-Throughput Sequencing. PLoS ONE, 8:e54848.

McCormack JE, Hird SM, Zellmer AJ, Carstens BC, Brumfield RT. 2013b. Applications of next-generation sequencing to phylogeography and phylogenetics. Mol. Phylogenet. Evol., 66:526-538.

Meusemann K, von Reumont BrM, Simon S, Roeding F, Strauss S, Kück P, Ebersberger I, Walzl M, Pass G, Breuers S, Achter V, von Haeseler A, Burmester T, Hadrys H, Wägele JW, Misof B. 2010. A phylogenomic approach to resolve the arthropod tree of life. Mol. Biol. Evol., 27:2451-2464.

Minh BQ, Nguyen MA, von Haeseler A. 2013. Ultrafast approximation for phylogenetic bootstrap. Mol. Biol. Evol., 30:1188-1195.

Mirarab S, Reaz R, Bayzid MS, Zimmermann T, Swenson MS, Warnow T. 2014. ASTRAL: genome-scale coalescent-based species tree estimation. Bioinformatics, 30:i541-i548.

Misof B, Liu S, 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, Aspöck U, Aspöck H, Bartel D, Blanke A, Berger S, Böhm A, Buckley TR, Calcott B, Chen J, Friedrich F, Fukui M, Fujita M, Greve C, Grobe P, Gu S, Huang Y, Jermiin LS, Kawahara AY, Krogmann L, Kubiak M, Lanfear R, Letsch H, Li Y, Li Z, Li J, Lu H, Machida R, Mashimo Y, Kapli P, McKenna DD, Meng G, Nakagaki Y, Navarrete-Heredia JL, Ott M, Ou Y, Pass G, Podsiadlowski L, Pohl H, von Reumont BM, Schütte K, Sekiya K, Shimizu S, Slipinski A, Stamatakis A, Song W, Su X, Szucsich NU, Tan M, Tan X, Tang M, Tang J, Timelthaler G, Tomizuka S, Trautwein M, Tong X, Uchifune T, Walzl MG, Wiegmann BM, Wilbrandt J, Wipfler B, Wong TKF, Wu Q, Wu G, Xie Y, Yang S, Yang Q, Yeates DK, Yoshizawa K, Zhang Q, Zhang R, Zhang W, Zhang Y, Zhao J, Zhou C, Zhou L, Ziesmann T, Zou S, Li Y, Xu X, Zhang Y, Yang H, Wang J, Wang J, Kjer KM, Zhou X. 2014. Phylogenomics resolves the timing and pattern of insect evolution. Science, 346:763-767.

Misof B, Misof K. 2009. A monte carlo approach successfully identifies randomness in multiple sequence alignments: a more objective means of data exclusion. Syst. Biol., 58:21-34.

Mutanen M, Wahlberg N, Kaila L. 2010. Comprehensive gene and taxon coverage elucidates radiation patterns in moths and butterflies. Proceedings of the Royal Society B: Biological Sciences.

Oakley TH, Wolfe JM, Lindgren AR, Zaharoff AK. 2012. Phylotranscriptomics to bring the understudied into the fold: Monophyletic Ostracoda, fossil placement and Pancrustacean phylogeny. Mol. Biol. Evol., 30:215-233.

Pattengale N, Alipour M, Bininda-Emonds OP, Moret BE, Stamatakis A. 2009. How many bootstrap replicates are necessary? In: Batzoglou S editor. Research in Computational Molecular Biology, Springer Berlin Heidelberg, p. 184-200.

Peloso PLV, Frost DR, Richards SJ, Rodrigues MT, Donnellan S, Matsui M, Raxworthy CJ, Biju SD, Lemmon EM, Lemmon AR, Wheeler WC. 2016. The impact of anchored phylogenomics and taxon sampling on phylogenetic inference in narrow-mouthed frogs (Anura, Microhylidae). Cladistics, 32:113-140.

Ponce FV, Breinholt JW, Hossie T, Barber JR, Janzen DH, Hallwachs W, Kawahara AY. 2015. A molecular phylogeny of *Eumorpha* (Lepidoptera: Sphingidae) and the evolution of anti-predator larval eyespots. Syst. Entomol., 40:401-408.

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.

Pyron RA, Hendry CR, Chou VM, Lemmon EM, Lemmon AR, Burbrink FT. 2014. Effectiveness of phylogenomic data and coalescent species-tree methods for resolving difficult nodes in the phylogeny of advanced snakes (Serpentes: Caenophidia). Mol. Phylogenet. Evol., 81:221-231.

Regier J, Zwick A, Cummings M, Kawahara A, Cho S, Weller S, Roe A, Baixeras J, Brown J, Parr C, Davis D, Epstein M, Hallwachs W, Hausmann A, Janzen D, Kitching I, Solis MA, Yen S-H, Bazinet A, Mitter C. 2009. Toward reconstructing the evolution of advanced moths and butterflies (Lepidoptera: Ditrysia): an initial molecular study. BMC Evol. Biol., 9:280.

Regier JC, Grant MC, Mitter C, Cook CP, Peigler RS, Rougerie R. 2008. Phylogenetic relationships of wild silkmoths (Lepidoptera: Saturniidae) inferred from four protein-coding nuclear genes. Syst. Entomol., 33:219-228.

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 (nonditrysian) lineages of extant Lepidoptera, with implications for classification, comparative morphology and life-history evolution. Syst. Entomol., 40:671-704.

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.

Roe AD, Weller SJ, Baixeras J, Brown J, Cummings MP, Davis D, Kawahara AY, Parr C, Regier JC, Rubinoff D. 2009. Evolutionary framework for Lepidoptera model systems. Genetics and Molecular Biology of Lepidoptera:1-24.

Ruane S, Raxworthy CJ, Lemmon AR, Lemmon EM, Burbrink FT. 2015. Comparing species tree estimation with large anchored phylogenomic and small Sanger-sequenced molecular datasets: an empirical study on Malagasy pseudoxyrhophiine snakes. BMC Evol. Biol., 15:1-14.

Seo TK. 2008. Calculating bootstrap probabilities of phylogeny using multilocus sequence data. Mol. Biol. Evol., 25:960-971.

Simmons MP. 2012. Radical instability and spurious branch support by likelihood when applied to matrices with non-random distributions of missing data. Mol. Phylogenet. Evol., 62:472-484.

Simmons MP. 2014. A confounding effect of missing data on character conflict in maximum likelihood and Bayesian MCMC phylogenetic analyses. Mol. Phylogenet. Evol., 80:267-280.

Simon S, Narechania A, DeSalle R, Hadrys H. 2012. Insect phylogenomics: Exploring the source of incongruence using new transcriptomic data. Genome Biol Evol, 4:1295-1309.

Smith BT, Harvey MG, Faircloth BC, Glenn TC, Brumfield RT. 2014. Target capture and massively parallel sequencing of ultraconserved elements for comparative studies at shallow evolutionary time scales. Syst. Biol., 63:83-95.

Sohn J-C, Regier JC, Mitter C, Davis D, Landry J-Fo, 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.

Soltis PS, Soltis DE, Savolainen V, Crane PR, Barraclough TG. 2002. Rate heterogeneity among lineages of tracheophytes: Integration of molecular and fossil data and evidence for molecular living fossils. Proceedings of the National Academy of Sciences, 99:4430-4435.

Song H, Sheffield NC, Cameron SL, Miller KB, Whiting MF. 2010. When phylogenetic assumptions are violated: base compositional heterogeneity and among-site rate variation in beetle mitochondrial phylogenomics. Syst. Entomol., 35:429-448.

The Heliconius Genome Consortium. 2012. Butterfly genome reveals promiscuous exchange of mimicry adaptations among species. Nature, 487:94-98.

van Nieukerken EJ, Kaila L, Kitching IJ, Kristensen NP, Lees DC, Minet J, Mitter C, Mutanen M, Regier JC, Simonsen TJ, Wahlberg N, Yen S-H, Zahiri R, Adamski D, Baixeras J, Bartsch D, Bengtsson BÅ, Brown JW, Bucheli SR, Davis DR, De Prins J, De Prins W, Epstein ME, Gentili-Poole P, Gielis C, Hättenschwiler P, Hausmann A, Holloway JD, Kallies A, Karsholt O, Kawahara A, Koster SJC, Kozlov M, Lafontaine JD, Lamas G, Landry J-F, Lee S, Nuss M, Penz C, Rota J, Schmidt, B. C. S, A., Sohn JC, Solis MA, Tarmann GM, Warren AD, Weller S, Yakovlev R, Zolotuhin V, Zwick A. 2011. Order Lepidoptera Linnaeus, 1758. In: Zhang, Z.-Q. (Ed.) Animal biodiversity: An outline of higher-level classification and survey of taxonomic richness. Zootaxa, 3148:212-221.

Wahlberg N, Wheat CW. 2008. Genomic outposts serve the phylogenomic pioneers: designing novel nuclear markers for genomic DNA extractions of Lepidoptera. Syst. Biol., 57:231-242.

Xia Q, Zhou Z, Lu C, Cheng D, Dai F, Li B, Zhao P, Zha X, Cheng T, Chai C, Pan G, Xu J, Liu C, Lin Y, Qian J, Hou Y, Wu Z, Li G, Pan M, Li C, Shen Y, Lan X, Yuan L, Li T, Xu H, Yang G, Wan Y, Zhu Y, Yu M, Shen W, Wu D, Xiang Z, group Ga, Yu J, Wang J, Li R, Shi J, Li H, Li G, Su J, Wang X, Li G, Zhang Z, Wu Q, Li J, Zhang Q, Wei N, Xu J, Sun H, Dong L, Liu D, Zhao S, Zhao X, Meng Q, Lan F, Huang X, Li Y, Fang L, Li C, Li D, Sun Y, Zhang Z, Yang Z, Huang Y, Xi Y, Qi Q, He D, Huang H, Zhang X, Wang Z, Li W, Cao Y, Yu Y, Yu H, Li J, Ye J, Chen H, Zhou Y, Liu B, Wang J, Ye J, Ji H, Li S, Ni P, Zhang J, Zhang Y, Zheng H, Mao B, Wang W, Ye C, Li S, Wang J, Wong GK-S, Yang H. 2004. A draft sequence for the genome of the domesticated silkworm (*Bombyx mori*). Science, 306:1937-1940.

Xia X, Xie Z, Salemi M, Chen L, Wang Y. 2003. An index of substitution saturation and its application. Mol. Phylogenet. Evol., 26:1-7.

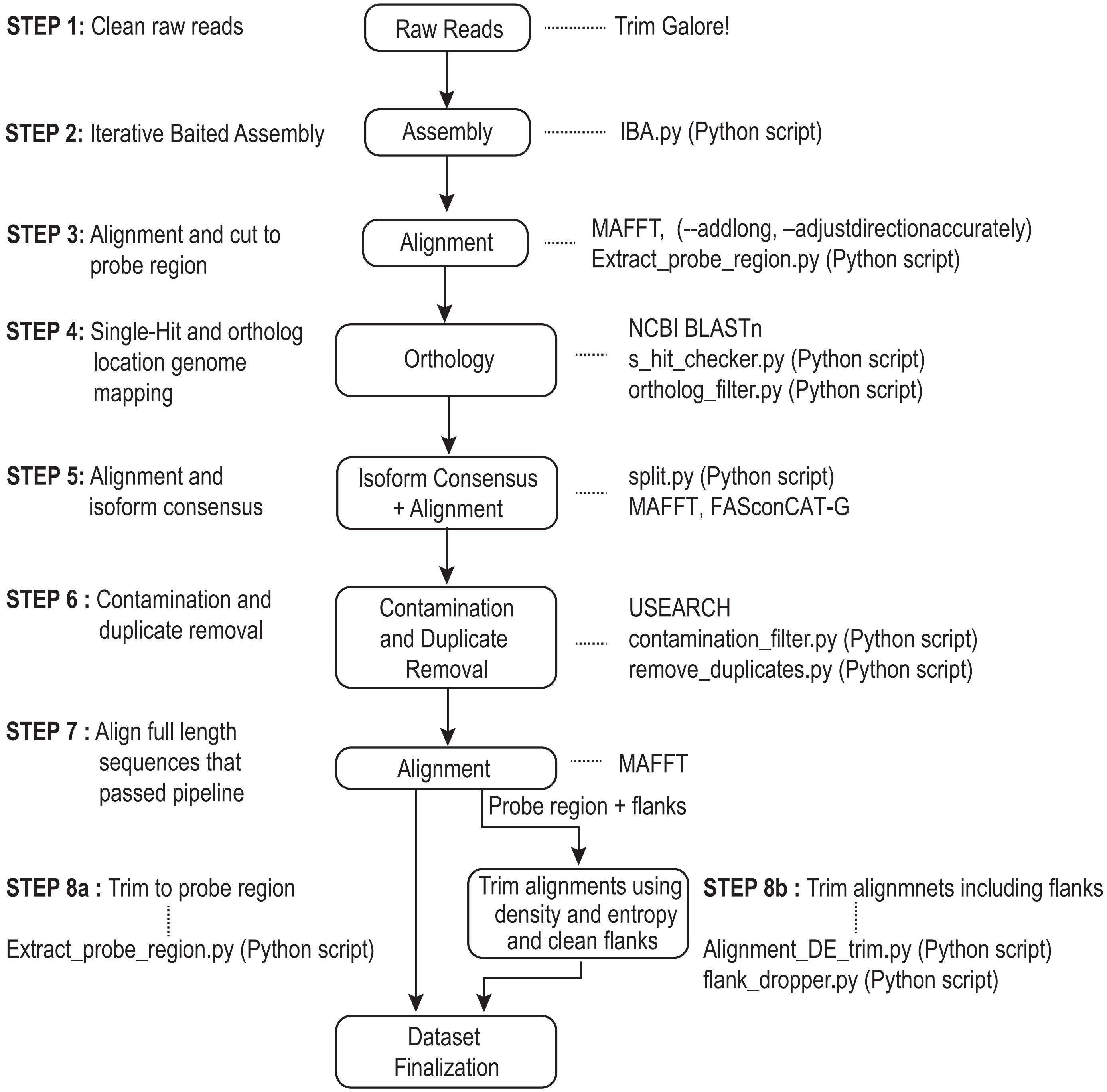
You M, Yue Z, He W, Yang X, Yang G, Xie M, Zhan D, Baxter SW, Vasseur L, Gurr GM, Douglas CJ, Bai J, Wang P, Cui K, Huang S, Li X, Zhou Q, Wu Z, Chen Q, Liu C, Wang B, Li X, Xu X, Lu C, Hu M, Davey JW, Smith SM, Chen M, Xia X, Tang W, Ke F, Zheng D, Hu Y, Song F, You Y, Ma X, Peng L, Zheng Y, Liang Y, Chen Y, Yu L, Zhang Y, Liu Y, Li G, Fang L, Li J, Zhou X, Luo Y, Gou C, Wang J, Wang J, Yang H, Wang J. 2013. A heterozygous moth genome provides insights into herbivory and detoxification. Nat. Genet., 45:220-225.

Young AD, Lemmon AR, Skevington JH, Mengual X, Stahls G, Reemer M, Jordaens K, Kelso S, Lemmon EM, Hauser M, De Meyer M, Misof B, Wiegmann BM. 2016. Anchored enrichment dataset for true flies (order Diptera) reveals insights into the phylogeny of flower flies (family Syrphidae). BMC Evol. Biol., 16:143.

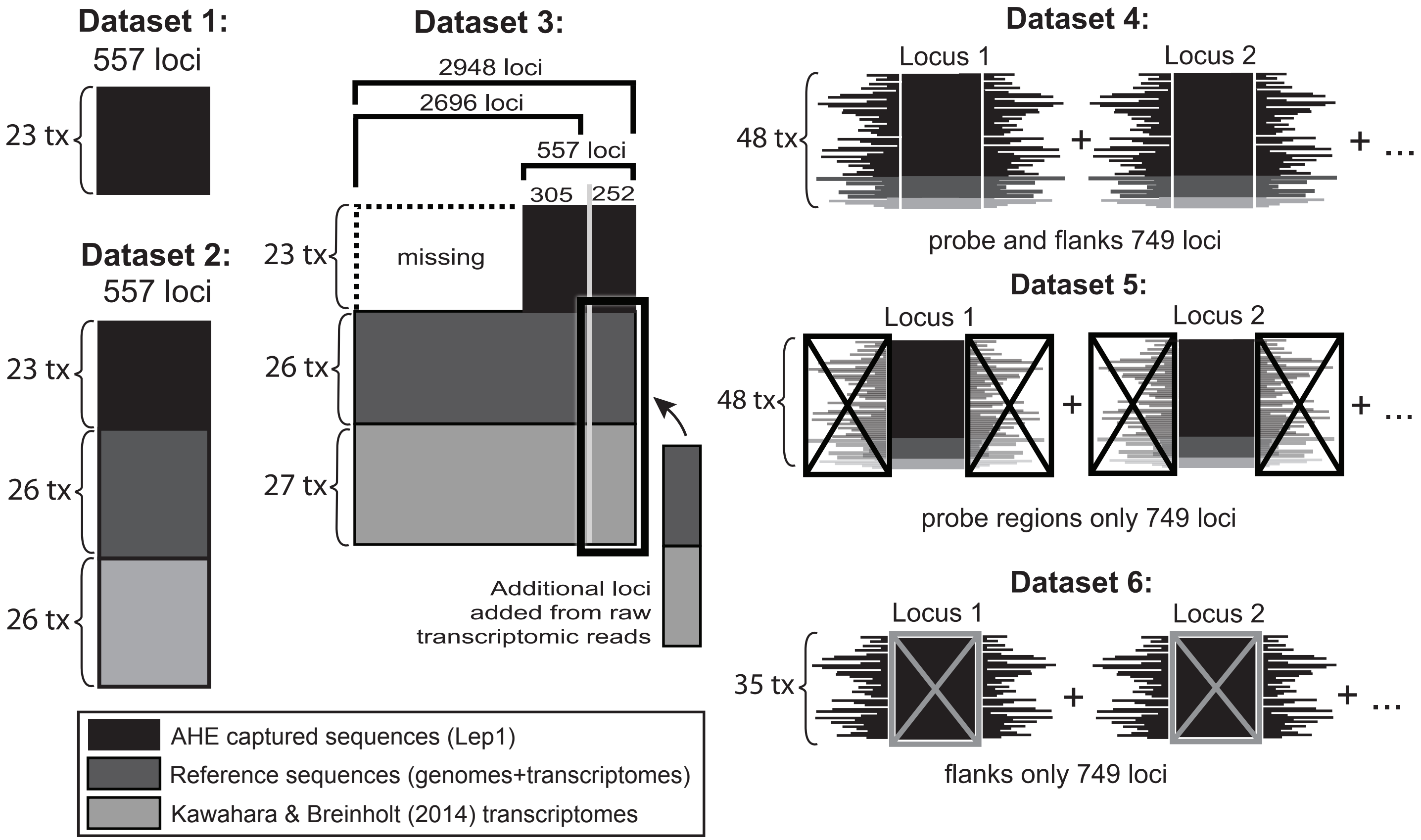
Zhan S, Merlin C, Boore JL, Reppert SM. 2011. The monarch butterfly genome yields Insights into long-distance migration. Cell, 147:1171-1185.

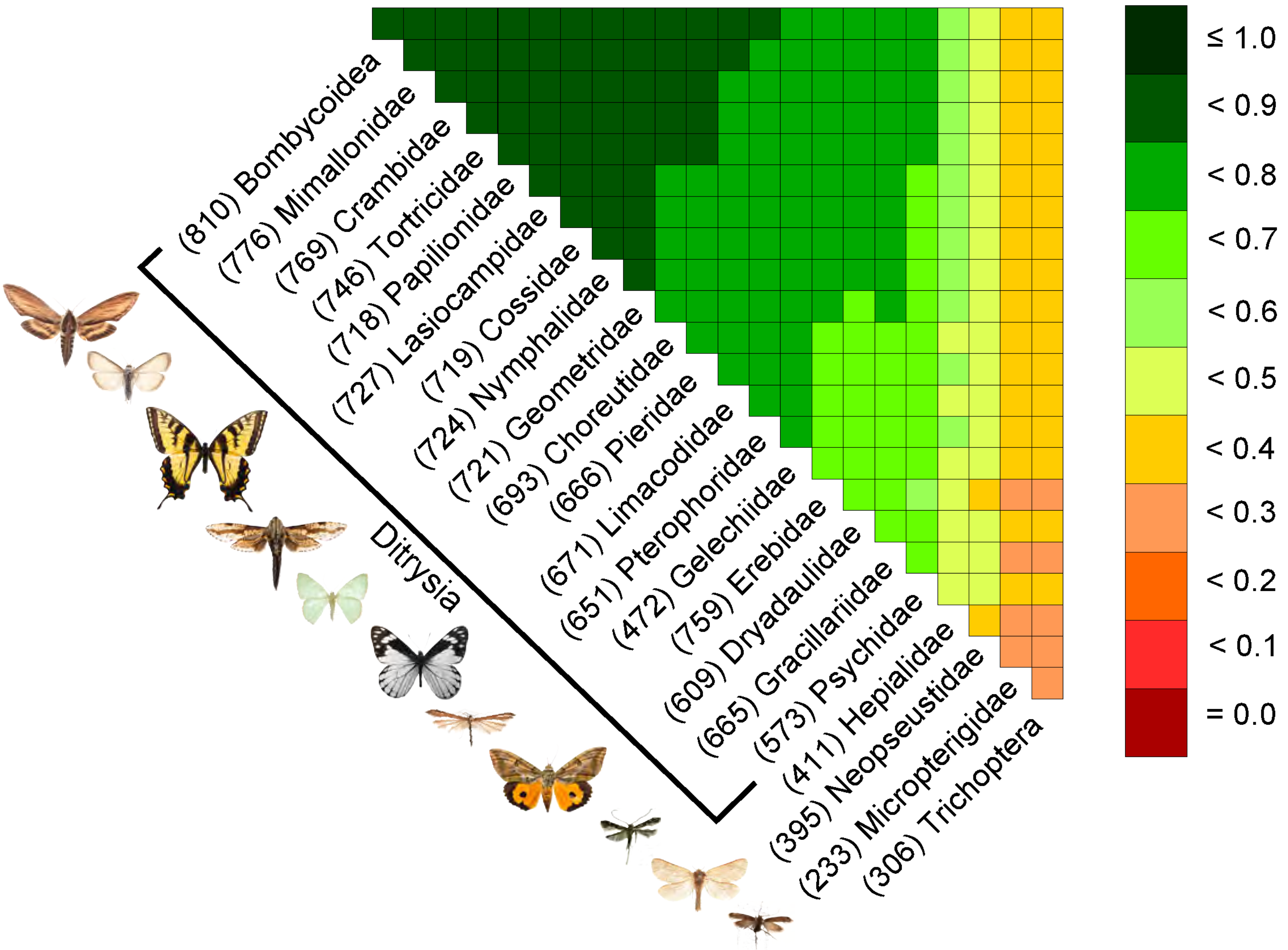
Zwick A, Regier JC, Mitter C, Cummings MP. 2011. Increased gene sampling yields robust support for higher-level clades within Bombycoidea (Lepidoptera). Syst. Entomol., 36:31-43.

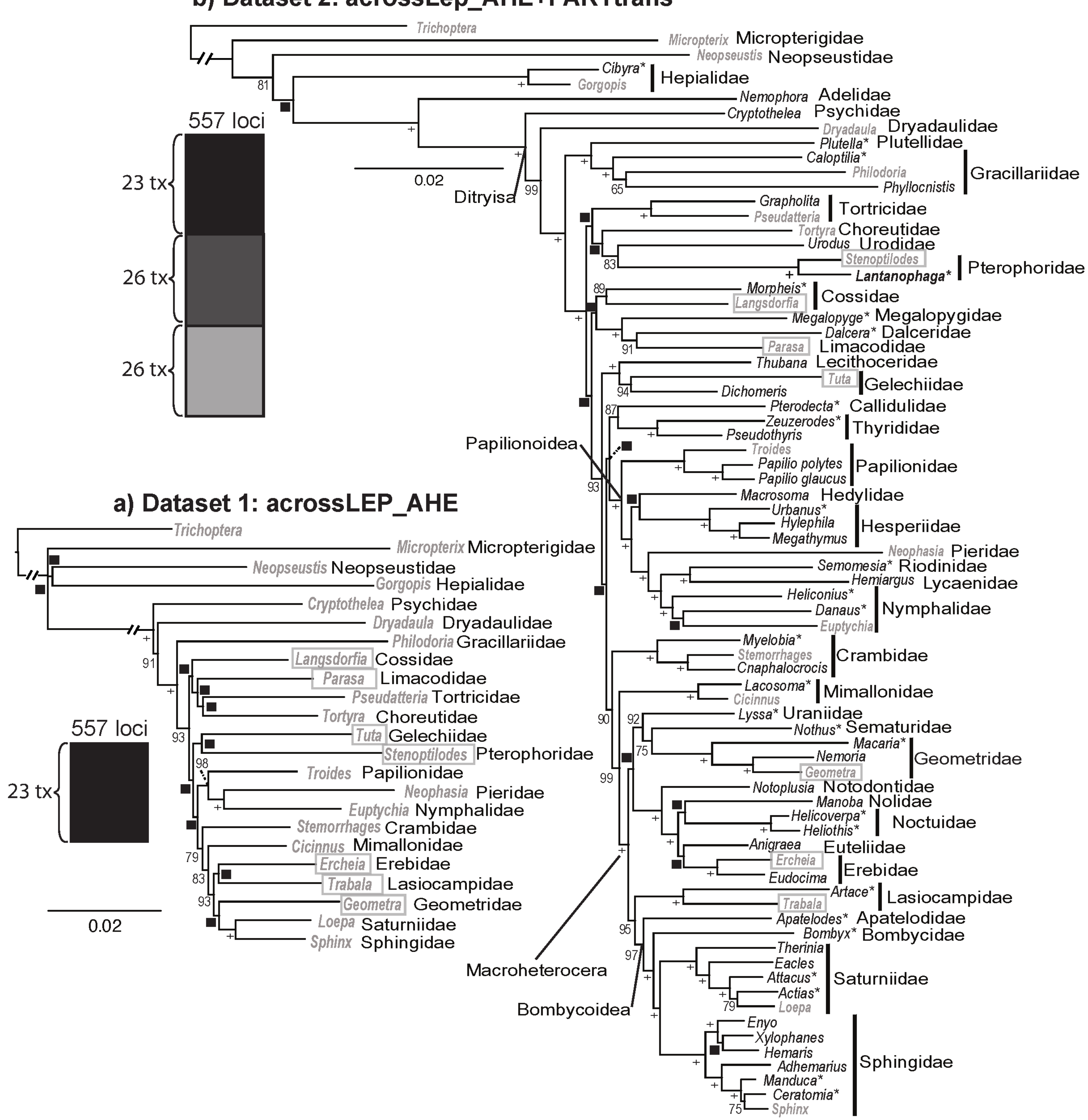
Zwick A, Regier JC, Zwickl DJ. 2012. Resolving discrepancy between nucleotides and amino acids in deep-level Arthropod phylogenomics: differentiating serine codons in 21-amino-acid models. PLoS ONE, 7:e47450.

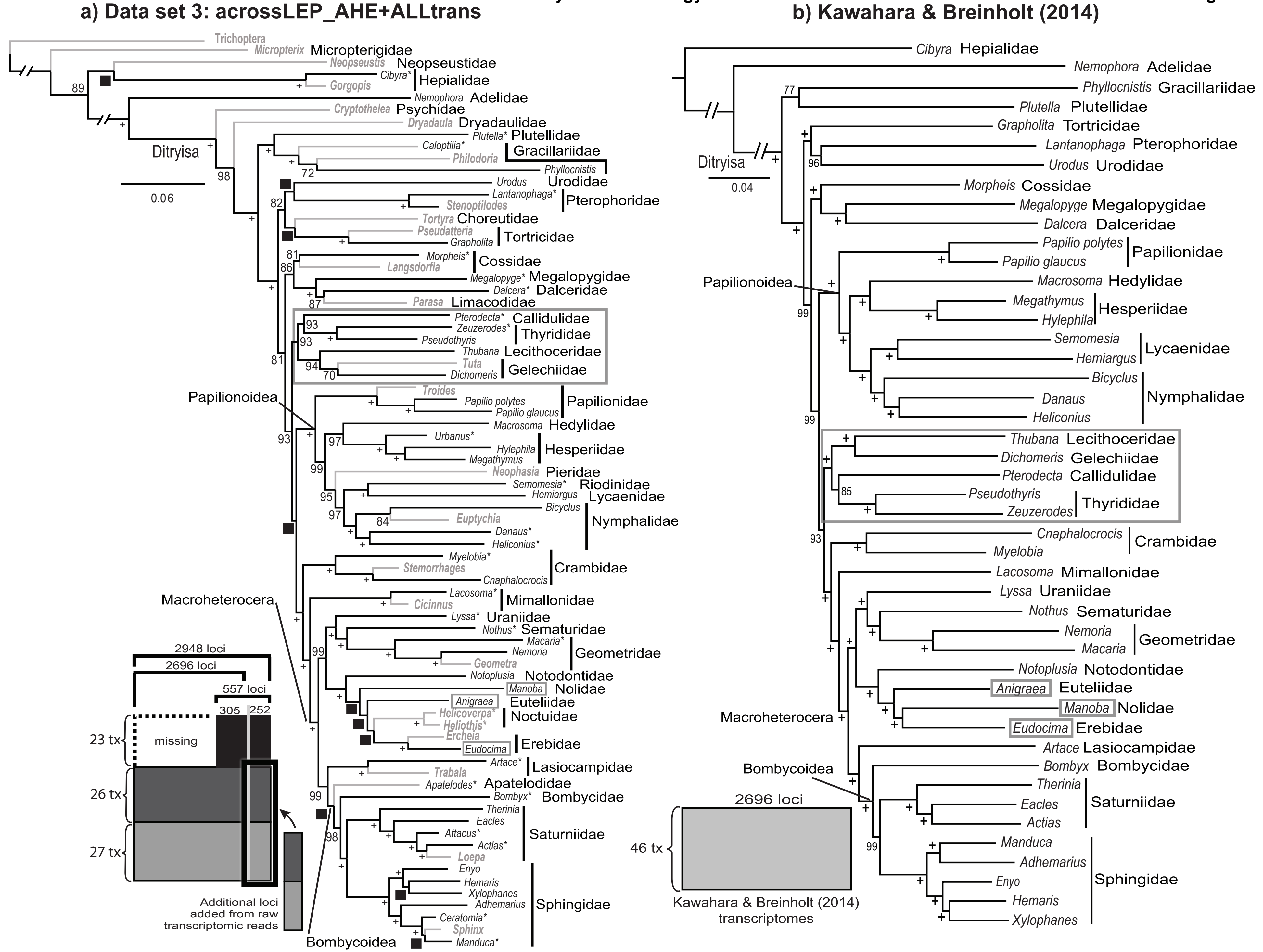

STEP 1: Clean raw reads
Raw Reads
Trim Galore!
STEP 2: Iterative Baited Assembly
Assembly
IBA.py (Python script)
STEP 3: Alignment and cut to probe region
Alignment
MAFFT, (--addlong, –adjustdirectionaccurately)
Extract_probe_region.py (Python script)
STEP 4: Single-Hit and ortholog location genome mapping
Orthology
NCBI BLASTn
s_hit_checker.py (Python script)
ortholog_filter.py (Python script)
STEP 5: Alignment and isoform consensus
Isoform Consensus + Alignment
split.py (Python script)
MAFFT, FASconCAT-G
STEP 6 : Contamination and duplicate removal
Contamination and Duplicate Removal
USEARCH
contamination_filter.py (Python script)
remove_duplicates.py (Python script)
STEP 7 : Align full length sequences that passed pipeline
Alignment
MAFFT
Probe region + flanks
STEP 8a : Trim to probe region
Extract_probe_region.py (Python script)
Trim alignments using density and entropy and clean flanks
STEP 8b : Trim alignmnets including flanks
Alignment_DE_trim.py (Python script)
flank_dropper.py (Python script)
Dataset Finalization

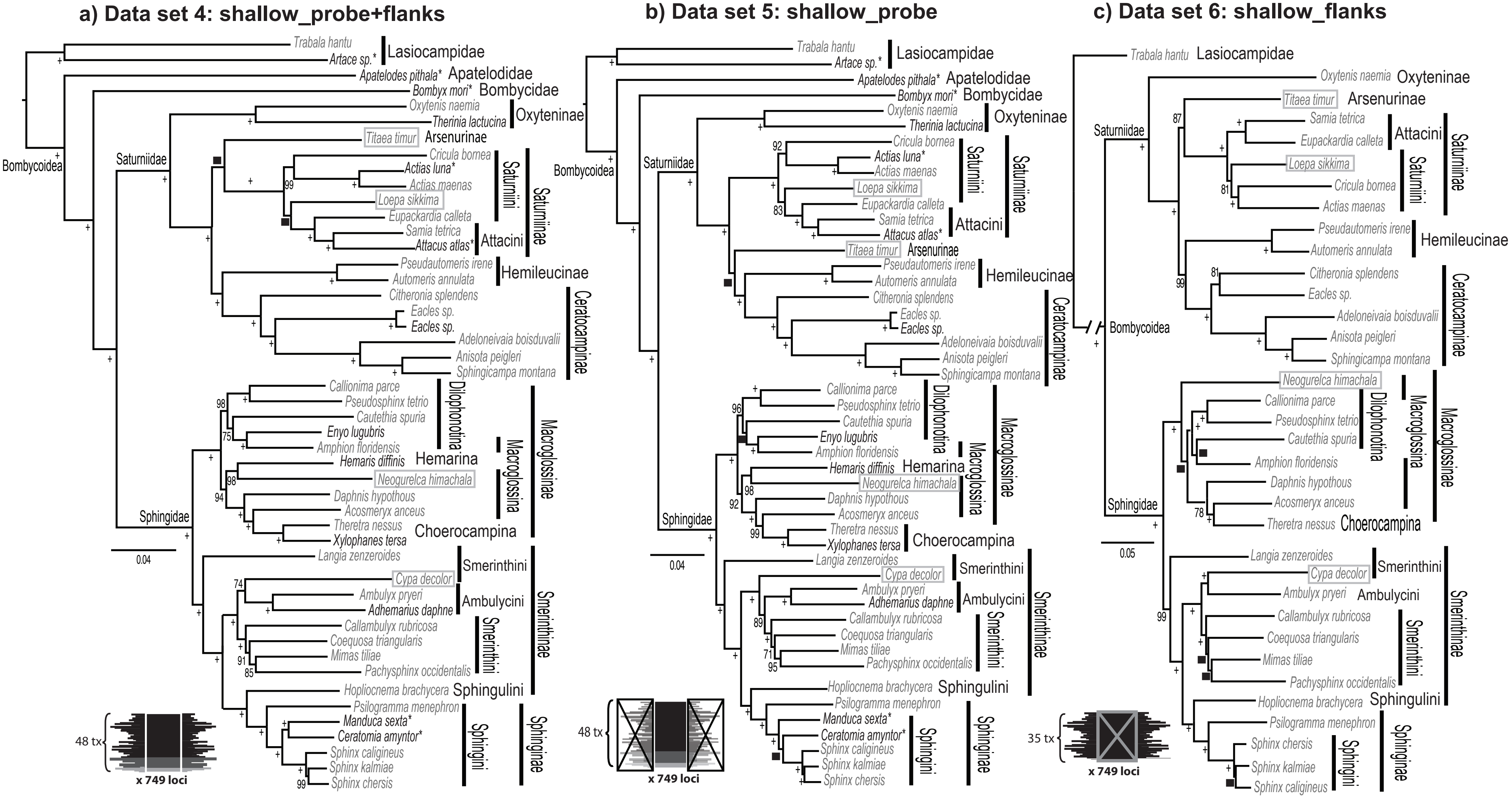
Dataset	# taxa	# loci	# bp	# AA	% C	ML	ASTRAL
1	23	557	90,238	45,119	76%	Fig. 4a	n/a
2	75	557	90,238	45,119	75%	Fig. 4b	Fig. S4
3	76	2948	2,522,806	n/a	31%	Fig. 5a	n/a
4	48	749	281,241	n/a	82%	Fig. 6a	Fig. S5
5	48	749	166,766	n/a	81%	Fig. 6b	n/a
6	36	749	114,475	n/a	88%	Fig. 6c	n/a


Dataset 1:
557 loci
23 tx
Dataset 2:
557 loci
23 tx
26 tx
26 tx
Dataset 3:
2948 loci
2696 loci
557 loci
305
252
23 tx
missing
26 tx
27 tx
Additional loci added from raw transcriptomic reads
Dataset 4:
Locus 1
Locus 2
48 tx
+
+ ...
probe and flanks 749 loci
Dataset 5:
Locus 1
Locus 2
48 tx
+
+ ...
probe regions only 749 loci
Dataset 6:
Locus 1
Locus 2
35 tx
+
+ ...
flanks only 749 loci
AHE captured sequences (Lep1)
Reference sequences (genomes+transcriptomes)
Kawahara & Breinholt (2014) transcriptomes


(810) Bombycoidea
(776) Mimallonidae
(769) Crambidae
(746) Tortricidae
(718) Papilionidae
(727) Lasiocampidae
(719) Cossidae
(724) Nymphalidae
(721) Geometridae
(693) Choreutidae
(666) Pieridae
(671) Limacodidae
(651) Pterophoridae
(472) Gelechiidae
(759) Erebidae
(609) Dryadaulidae
(665) Gracillariidae
(573) Psychidae
(411) Hepialidae
(395) Neopseustidae
(233) Micropterigidae
(306) Trichoptera
Ditrysia
≤ 1.0
< 0.9
< 0.8
< 0.7
< 0.6
< 0.5
< 0.4
< 0.3
< 0.2
< 0.1
= 0.0


b) Dataset 2: acrossLep_AHE+PARTtrans
557 loci
23 tx
26 tx
26 tx
0.02
Trichoptera
Micropterix Micropterigidae
Neopseustis Neopseustidae
Cibyra* Gorgopis Hepialidae
Nemophora Adelidae
Cryptothelea Psychidae
Dryadaula Dryadaulidae
Plutella* Plutellidae
Caloptilia* Philodoria Phyllocnistis Gracillariidae
Ditrysia
Grapholita Pseudatteria Tortricidae
Tortyra Choreutidae
Urodus Urodidae
Stenoptilodes Lantanophaga* Pterophoridae
Morpheis* Langsdorfia Cossidae
Megalopyge* Megalopygidae
Dalcera* Dalceridae
Parasa Limacodidae
Thubana Lecithoceridae
Tuta Dichomeris Gelechiidae
Pterodecta* Callidulidae
Zeuzerodes* Pseudothyris Thyrididae
Papilionoidea
Troides Papilio polytes Papilio glaucus Papilionidae
Macrosoma Hedylidae
Urbanus* Hylephila Megathymus Hesperiidae
Neophasia Pieridae
Semomesia* Riodinidae
Hemiargus Lycaenidae
Heliconius* Danaus* Euptychia Nymphalidae
Myelobia* Stemorrhages Cnaphalocrocis Crambidae
Lacosoma* Cicinnus Mimallonidae
Lyssa* Uraniidae
Nothus* Sematuridae
Macaria* Nemoria Geometra Geometridae
Notoplusia Notodontidae
Manoba Nolidae
Helicoverpa* Heliothis* Noctuidae
Anigraea Euteliidae
Ercheia Eudocima Erebidae
Artace* Trabala Lasiocampidae
Apatelodes* Apatelodidae
Bombyx* Bombycidae
Therinia Eacles Attacus* Actias* Loepa Saturniidae
Enyo Xylophanes Hemaris Adhemarius Manduca* Ceratomia* Sphinx Sphingidae
Macroheterocera
Bombycoidea
81 99 65 83 89 91 94 87 93 90 92 75 99 95 97 79 75
a) Dataset 1: acrossLEP_AHE
557 loci
23 tx
0.02
Trichoptera
Micropterix Micropterigidae
Neopseustis Neopseustidae
Gorgopis Hepialidae
Cryptothelea Psychidae
Dryadaula Dryadaulidae
Philodoria Gracillariidae
Langsdorfia Cossidae
Parasa Limacodidae
Pseudatteria Tortricidae
Tortyra Choreutidae
Tuta Gelechiidae
Stenoptilodes Pterophoridae
Troides Papilionidae
Neophasia Pieridae
Euptychia Nymphalidae
Stemorrhages Crambidae
Cicinnus Mimallonidae
Ercheia Erebidae
Trabala Lasiocampidae
Geometra Geometridae
Loepa Saturniidae
Sphinx Sphingidae
91 93 98 79 83 93


a) Data set 3: acrossLEP_AHE+ALLtrans
Trichoptera
Micropterix Micropterigidae
Neopseustis Neopseustidae
Cibyra* Hepialidae
Gorgopis
89
Nemophora Adelidae
Cryptothelea Psychidae
Dryadaula Dryadaulidae
Ditryisa
98
0.06
Plutella* Plutellidae
Caloptilia* Gracillariidae
Philodoria
72
Phyllocnistis
Urodus Urodidae
Lantanophaga* Pterophoridae
Stenoptilodes
82
Tortyra Choreutidae
Pseudatteria
Grapholita Tortricidae
81
86
Morpheis* Cossidae
Langsdorfia
Megalopyge* Megalopygidae
Dalcera* Dalceridae
87
Parasa Limacodidae
Pterodecta* Callidulidae
93
Zeuzerodes* Thyrididae
93
Pseudothyris
Thubana Lecithoceridae
94
Tuta
70
Dichomeris Gelechiidae
81
Papilionoidea
Troides
Papilio polytes Papilionidae
Papilio glaucus
Macrosoma Hedylidae
97
Urbanus*
Hylephila Hesperiidae
Megathymus
93
99
Neophasia Pieridae
95
Semomesia* Riodinidae
Hemiargus Lycaenidae
97
Bicyclus
84
Euptychia Nymphalidae
Danaus*
Heliconius*
Myelobia*
Stemorrhages Crambidae
Cnaphalocrocis
Macroheterocera
Lacosoma* Mimallonidae
Cicinnus
Lyssa* Uraniidae
Nothus* Sematuridae
99
Macaria*
Nemoria Geometridae
Geometra
Notoplusia Notodontidae
Manoba Nolidae
Anigraea Euteliidae
Helicoverpa* Noctuidae
Heliothis*
Ercheia
Eudocima Erebidae
Artace* Lasiocampidae
Trabala
Apatelodes* Apatelodidae
Bombyx* Bombycidae
99
Therinia
Eacles
Attacus* Saturniidae
Actias*
98
Loepa
Enyo
Hemaris
Xylophanes Sphingidae
Adhemarius
Ceratomia*
Sphinx
Manduca*
Bombycoidea
2948 loci
2696 loci
557 loci
305 252
23 tx
missing
26 tx
27 tx
Additional loci added from raw transcriptomic reads
b) Kawahara & Breinholt (2014)
Cibyra Hepialidae
Nemophora Adelidae
77
Phyllocnistis Gracillariidae
Plutella Plutellidae
Grapholita Tortricidae
Lantanophaga Pterophoridae
Ditryisa
96
Urodus Urodidae
0.04
Morpheis Cossidae
Megalopyge Megalopygidae
Dalcera Dalceridae
Papilio polytes
Papilio glaucus Papilionidae
Papilionoidea
Macrosoma Hedylidae
99
Megathymus
Hylephila Hesperiidae
Semomesia
Hemiargus Lycaenidae
Bicyclus
Danaus Nymphalidae
Heliconius
99
Thubana Lecithoceridae
Dichomeris Gelechiidae
Pterodecta Callidulidae
85
Pseudothyris
Zeuzerodes Thyrididae
93
Cnaphalocrocis
Myelobia Crambidae
Lacosoma Mimallonidae
Lyssa Uraniidae
Nothus Sematuridae
Nemoria
Macaria Geometridae
Notoplusia Notodontidae
Anigraea Euteliidae
Macroheterocera
Manoba Nolidae
Eudocima Erebidae
Artace Lasiocampidae
Bombycoidea
Bombyx Bombycidae
Therinia
Eacles Saturniidae
Actias
2696 loci
46 tx
99
Manduca
Adhemarius
Enyo Sphingidae
Hemaris
Xylophanes
Kawahara & Breinholt (2014) transcriptomes


a) Data set 4: shallow_probe+flanks
Trabala hantu
Artace sp.*
Lasiocampidae
Apatelodes pithala*
Apatelodidae
Bombyx mori*
Bombycidae
Oxytenis naemia
Therinia lactucina
Oxyteninae
Titaea timur
Arsenurinae
Cricula bornea
Actias luna*
Actias maenas
Loepa sikkima
Eupackardia calleta
Samia tetrica
Attacus atlas*
Saturniini
Attacini
Saturniinae
Pseudautomeris irene
Automeris annulata
Hemileucinae
Citheronia splendens
Eacles sp.
Eacles sp.
Adeloneivaia boisduvalii
Anisota peigleri
Sphingicampa montana
Ceratocampinae
Bombycoidea
Saturniidae
Callionima parce
Pseudosphinx tetrio
Cautethia spuria
Enyo lugubris
Amphion floridensis
Hemaris diffinis
Neogurelca himachala
Daphnis hypothous
Acosmeryx anceus
Theretra nessus
Xylophanes tersa
Dilophonotina
Hemarina
Macroglossina
Choerocampina
Macroglossinae
Langia zenzeroides
Cypa decolor
Ambulyx pryeri
Adhemarius daphne
Callambulyx rubricosa
Coequosa triangularis
Mimas tiliae
Pachysphinx occidentalis
Smerinthini
Ambulycini
Smerinthini
Smerinthinae
Hopliocnema brachycera
Sphingulini
Psilogramma menephron
Manduca sexta*
Ceratomia amyntor*
Sphinx caligineus
Sphinx kalmiae
Sphinx chersis
Sphingini
Sphinginae
Sphingidae
99
98
75
98
94
74
91
85
99
0.04
48 tx
x 749 loci
b) Data set 5: shallow_probe
Trabala hantu
Artace sp.*
Lasiocampidae
Apatelodes pithala*
Apatelodidae
Bombyx mori*
Bombycidae
Oxytenis naemia
Therinia lactucina
Oxyteninae
Cricula bornea
Actias luna*
Actias maenas
Loepa sikkima
Eupackardia calleta
Samia tetrica
Attacus atlas*
Saturniini
Attacini
Saturniinae
Titaea timur
Arsenurinae
Pseudautomeris irene
Automeris annulata
Hemileucinae
Citheronia splendens
Eacles sp.
Eacles sp.
Adeloneivaia boisduvalii
Anisota peigleri
Sphingicampa montana
Ceratocampinae
Bombycoidea
Saturniidae
Callionima parce
Pseudosphinx tetrio
Cautethia spuria
Enyo lugubris
Amphion floridensis
Hemaris diffinis
Neogurelca himachala
Daphnis hypothous
Acosmeryx anceus
Theretra nessus
Xylophanes tersa
Dilophonotina
Hemarina
Macroglossina
Choerocampina
Macroglossinae
Langia zenzeroides
Cypa decolor
Ambulyx pryeri
Adhemarius daphne
Callambulyx rubricosa
Coequosa triangularis
Mimas tiliae
Pachysphinx occidentalis
Smerinthini
Ambulycini
Smerinthini
Smerinthinae
Hopliocnema brachycera
Sphingulini
Psilogramma menephron
Manduca sexta*
Ceratomia amyntor*
Sphinx caligineus
Sphinx kalmiae
Sphinx chersis
Sphingini
Sphinginae
Sphingidae
92
83
96
98
92
99
89
71
95
0.04
48 tx
x 749 loci
c) Data set 6: shallow_flanks
Trabala hantu
Lasiocampidae
Oxytenis naemia
Oxyteninae
Titaea timur
Arsenurinae
Samia tetrica
Eupackardia calleta
Attacini
Loepa sikkima
Cricula bornea
Actias maenas
Saturniini
Saturniinae
Pseudautomeris irene
Automeris annulata
Hemileucinae
Citheronia splendens
Eacles sp.
Adeloneivaia boisduvalii
Anisota peigleri
Sphingicampa montana
Ceratocampinae
Saturniidae
Bombycoidea
Neogurelca himachala
Callionima parce
Pseudosphinx tetrio
Cautethia spuria
Amphion floridensis
Daphnis hypothous
Acosmeryx anceus
Theretra nessus
Dilophonotina
Macroglossina
Choerocampina
Macroglossinae
Langia zenzeroides
Cypa decolor
Ambulyx pryeri
Callambulyx rubricosa
Coequosa triangularis
Mimas tiliae
Pachysphinx occidentalis
Smerinthini
Ambulycini
Smerinthini
Smerinthinae
Hopliocnema brachycera
Sphingulini
Psilogramma menephron
Sphinx chersis
Sphinx kalmiae
Sphinx caligineus
Sphingini
Sphinginae
Sphingidae
87
81
99
81
78
99
0.05
35 tx
x 749 loci