Syst. Biol. 66(3):426–439, 2017

DOI:10.1093/sysbio/syw099
Advance Access publication January 25, 2017

Construction of a Species-Level Tree of Life for the Insects and Utility in Taxonomic Profiling

DOUGLAS CHESTERS*

*Key Laboratory of Zoological Systematics and Evolution, Institute of Zoology, Chinese Academy of Sciences, Beijing 100101, China; *Correspondence to be sent to: Key Laboratory of Zoological Systematics and Evolution, Institute of Zoology, Chinese Academy of Sciences, Beijing 100101, China; Email: dchesters@ioz.ac.cn*

Received 20 September 2015; reviews returned 3 November 2015; accepted 18 October 2016
Associate Editor: Vincent Savolainen

Abstract.—Although comprehensive phylogenies have proven an invaluable tool in ecology and evolution, their construction is made increasingly challenging both by the scale and structure of publically available sequences. The distinct partition between gene-rich (genomic) and species-rich (DNA barcode) data is a feature of data that has been largely overlooked, yet presents a key obstacle to scaling supermatrix analysis. I present a phyloinformatics framework for draft construction of a species-level phylogeny of insects (Class Insecta). Matrix-building requires separately optimized pipelines for nuclear transcriptomic, mitochondrial genomic, and species-rich markers, whereas tree-building requires hierarchical inference in order to capture species-breadth while retaining deep-level resolution. The phylogeny of insects contains 49,358 species, 13,865 genera, 760 families. Deep-level splits largely reflected previous findings for sections of the tree that are data rich or unambiguous, such as inter-ordinal Endopterygota and Dictyoptera, the recently evolved and relatively homogeneous Lepidoptera, Hymenoptera, Brachycera (Diptera), and Cucujiformia (Coleoptera). However, analysis of bias, matrix construction and gene-tree variation suggests confidence in some relationships (such as in Polyneoptera) is less than has been indicated by the matrix bootstrap method. To assess the utility of the insect tree as a tool in query profiling several tree-based taxonomic assignment methods are compared. Using test data sets with existing taxonomic annotations, a tendency is observed for greater accuracy of species-level assignments where using a fixed comprehensive tree of life in contrast to methods generating smaller de novo reference trees. Described herein is a solution to the discrepancy in the way data are fit into supermatrices. The resulting tree facilitates wider studies of insect diversification and application of advanced descriptions of diversity in community studies, among other presumed applications. [Data integration; data mining; insects; phylogenomics; phyloinformatics; tree of life.]

Supermatrix-based phylogenies rarely have more than 10,000 tips, which can give the impression of an upper limit on the number of taxa it is possible to analyze. This apparent limit often stems from difficulties in combining two different categories of publicly available sequence data (Sanderson et al. 2003; Wiens et al. 2005). "Phylogenomic" data sets provide a great many genes for a sparse sampling of model taxa, and are well-suited to resolving deep nodes, but lack breadth. "Species-level phylogenetic" data sets offer a few standard barcode genes for a great many taxa, which can be used to build trees with thousands of representative species, but without depth or resolution. There have been efforts to unify such partitions into single matrices by selecting an optimal point along the "species-rich" versus "gene-rich" continuum (Sanderson et al. 2003; Meusemann et al. 2010; Meyer 2011), but these fit poorly when scaled to the genomic level. Inability to effectively integrate key data partitions hampers downstream applications that require broad but reliable trees, such as community profiling based on metabarcoding or metagenomics of pooled or environmental DNA samples.

DNA-based profiling has broad applications, facilitating studies on biodiversity gradients (Stahlhut et al. 2013), conservation habitats (Hajibabaei et al. 2011), insectivore interactions (Clare et al. 2009; Pickett et al. 2012), pest–parasitoid interactions (Smith et al. 2011), comparative community ecology (Yu et al. 2012), and mechanisms behind spatial and temporal distributions (Andújar et al. 2015). In these contexts, placing DNA samples into a phylogenetic context gives various measures of diversity, including intrinsic species-level diversity (Blaxter et al. 2005; Pons et al. 2006), taxonomic diversity (Matsen et al. 2010; Berger et al. 2011; Filipski et al. 2015), ecological indices (Yu et al. 2012), phylogenetic diversity (Ives and Helmus 2010), and biodiversity variables (Pereira et al. 2013). The accuracy of tree-derived indices can be improved with the use of more comprehensive reference trees (compared with plot-level or regional reference trees e.g. Erickson et al. 2014). This increasing need for species-rich reference phylogenies necessitates protocols that address weak data integration, particularly of data-type. Further, supertree-like approaches (e.g., Hinchliff et al. 2015) are a partial solution only, since these DNA-based applications naturally require a framework that includes both phylogeny and sequence alignment.

The unrivalled diversity of insects (Grimaldi and Engel 2005) equates to equally large gaps in phylogenetic data (Rainford et al. 2014) and the lack of a resolved species-level tree of life, which limits hypothesis testing on diversification (Wiens et al. 2015). Transcriptomes (Misof et al. 2014; http://1kite.org/) have superseded complete genomes (Savard et al. 2006; Meusemann et al. 2010; Simon et al. 2012) as the leading data type in reconstructing the backbone of the insect tree. However, various factors have been identified which impact on the accuracy of reconstructions of these data. Artifactual relationships have been observed from supermatrices of combined proteome and transcriptome data (Letsch et al. 2012). Amino acids have been shown less prone to error from saturation, heterogeneous composition (Jeffroy

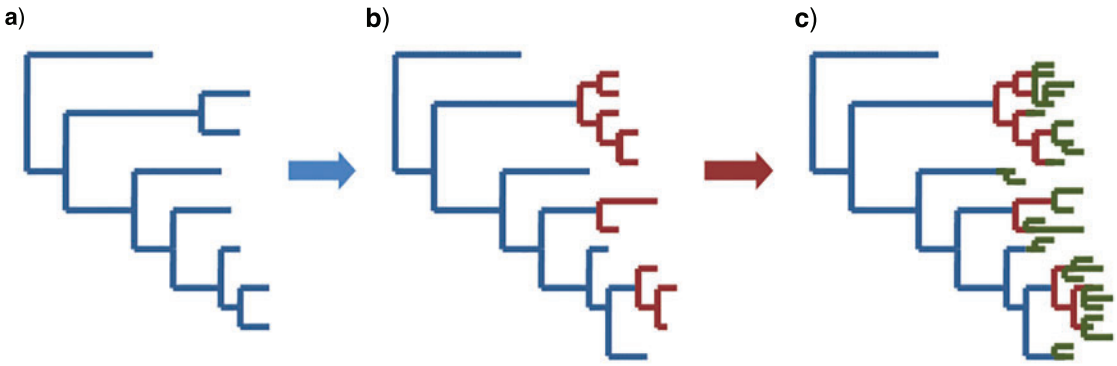

FIGURE 1. Successive, hierarchical construction of a species-level tree of life, reflecting structural and informational features of sequence data. a) nuclear phylogenomic backbone is filled internally using b) mitogenomic and c) species-rich data.

et al. 2006; Dávalos and Perkins 2008) or parameter choices (Simon et al. 2009). Long branch attraction can mislead inference (Bergsten 2005; Lemmon et al. 2009; Simmons 2012), as can segments of randomness in sequence alignment (Talavera and Castresana 2007; Kueck et al. 2010). The presence of error from the above factors might be indicated by replication of phylogenetic inference on matrices which have been alternatively reduced to address them (Whelan et al. 2015; Garrison et al. 2016).

Mitochondrial genomes (mitogenomes) have contributed to insect systematics at the intra-ordinal range (Cameron 2014), and are now accumulating at a greater rate with the successful application of next-generation sequencing to pooled insect samples (Timmermans et al. 2010; Tang et al. 2014; Timmermans et al. 2014). Similarly targeting systematics at the intra-ordinal range (Heraty et al. 2011; Regier et al. 2013; Bocak 2014), generation of "traditional markers" has received a particular boost in species-breadth from barcoding endeavors (Hebert et al. 2003). Parallel to such data generation, phyloinformatics-centric studies have emerged as a powerful approach to synthesizing publically available data (e.g. Peters et al. 2011; Hedtke et al. 2013; McMahon et al. 2015).

The limited resolving power which is characteristic of species-rich data sets may be addressed with imposition of basal constraints using information from other sources (Rainford et al. 2014), with character-rich genomes likely containing the required information content (Rokas et al. 2003; Dunn et al. 2008; Zwick et al. 2011). Despite this, reports in insect systematics are consistently restricted in data-type, usually to one of either nuclear genomic (Meusemann et al. 2010; Simon et al. 2012; Misof et al. 2014), mitogenomic (Cameron 2014), or species-rich markers (Heraty et al. 2011; Peters et al. 2011; Hedtke et al. 2013; Bocak 2014). A satisfactory integration is lacking, although largely within principles already established (e.g. Bininda-Emonds 2004; Kress et al. 2010; Smith et al. 2013). Hierarchical analysis is one solution to aligning tree construction with the structure of informative data (Fig. 1), with species-rich inference occurring within a genomic backbone. In addition to increasing the total amount of data utilized, partitions are used where required, with information-rich partitions resolving deep-level splits and broadly-sampled partitions resolving tip-level. This optimized framework enables upscaling of inference, providing more broad-level trees which are required in DNA-based profiling, while facilitating studies of diversification (certainly in the case of insects). Herein, I present an integrated phyloinformatics protocol called SOPHI (Structurally-Optimized-PHylo-Informatics) which permits the generation of species-level phylogenies far outscaling existing supermatrix-derived trees, the necessity of which is demonstrated in DNA-based community profiling.

METHODS

Figure 2 gives an overview of the steps required in the presented approach to constructing the species-level tree of life for insects. Software versions, command options and other technical details are provided in the supplementary document (http://dx.doi.org/10.5061/dryad.27114).

Constructing Backbone Supermatrices for Phylogenomics

A single representative transcriptome was downloaded for each insect order plus several outgroup species (from the three major orders of noninsect hexapods, including two distinct taxa from the presumed sister taxa Diplura). A total of 32 Transcriptomes were downloaded (http://www.ncbi.nlm.nih.gov/Traces/wgs/) and processed. Putative orthologs in the transcriptomes were identified using the "Insecta Core Orthologs" an established set of orthologous insect sequences generated through the Inparanoid-TC approach (Ebersberger et al. 2009), and downloaded from http://www.deep-phylogeny.org/hamstr/download/datasets/hmmer3/ The transcriptomes were queried using TBlastN (from Blast+; Camacho 2009) as to contain both orthologs and paralogs, and translated subject

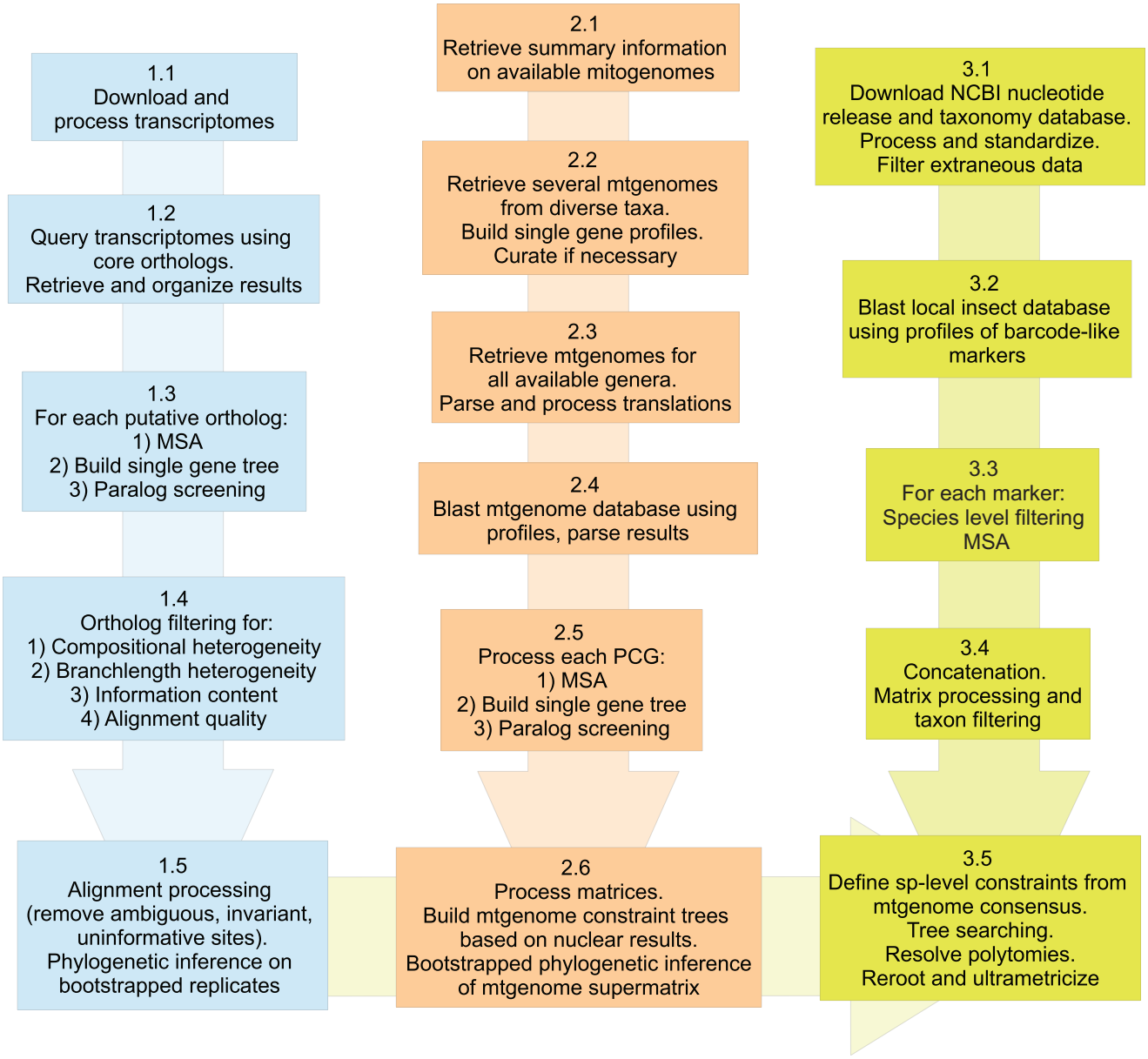

FIGURE 2. Flowchart of informatics steps in the sophi pipeline. Data are mined, processed, organized, and analyzed (proceeding top to bottom), whereas the phylogeny is constructed hierarchically from data most gene-rich to most species-rich (lower, left to right). Steps 1.1–1.5 use nuclear transcriptomes, 2.1–2.6 mitogenomes, and 3.1–3.5 species-rich markers, with step numbers corresponding to those in the pipeline (Supplementary Materials available on Dryad). MSA = multiple sequence alignment, PCG = protein coding gene, mtgenome = mitochondrial genome.

sequences parsed directly from Blast output files (parse_ortholog_results.pl; note, scripts named in parentheses were newly developed for the current study). This gave a preliminary set of homologous protein sequences which were then filtered using a treebased approach to remove paralogs, retaining a single ortholog for each species. Each homolog was aligned with Mafft (Katoh et al. 2002) using the most accurate algorithm (L-INS-i), and single gene trees built both using FastTree (Price et al. 2009) and RaxML (Stamatakis 2014) with the rapid bootstrap option. Homologs with their corresponding gene-trees were inputted into PhyloTreePruner (Kocot et al. 2013) for removal of paralogs. This software collapses gene-tree nodes with support less than a user supplied value, then the largest subtree identified and retained in which sequences from the same species are monophyletic (or part of the same polytomy). An advantage of this approach is that sequences are not erroneously discarded as paralogs when the corresponding trees are weak in resolution. Also note "in-paralogs" are not necessarily filtered by PhyloTreePruner, although these are thought not to contribute to topological error.

As reviewed earlier, several factors are known to mislead phylogenomic inference. Thus for subsequent tests of matrix filtration, each ortholog was first quantified as following; (i) a compositional homogeneity test (Foster et al. 2004; Nesnidal et al. 2010) was conducted via the posterior predictive model check as implemented in PhyloBayes (Lartillot et al. 2004). PhyloBayes was used with a constrained gene-tree, running an MCMC then calculating the statistic for global deviation over all taxa; (ii) variance in branch-lengths was calculated using a Perl script (calculate_branchlength_variance.pl); (iii) information content and tree-likeness was estimated with the software Mare (Meyer 2011), designed to maximize information content of matrices via both data presence and tree-likeness of partitions. Mare was set to ensure retention of all taxa, and various settings were applied for weighting of information content; (iv) after concatenation (concatenate_v2.pl), matrices were processed for alignment quality with Gblocks (Castresana 2000), which selects blocks of conserved positions with highly conserved flanking positions.

Reducing the likelihood of bias through ortholog filtration can produce conflicting results depending on details of the method employed (Garrison et al. 2016). Sensitivity of phylogenetic results to approach in matrix filtration was examined first by individually processing matrices, each to account for relative degree of one of the sources of topological bias (labeled Nucl1-4; Table 1), and with parameters adjusted as to ensure formation of equivalently sized matrices (measured after removal of invariant, ambiguous and parsimony uninformative sites using the script exclude_parsimony_uninformative.pl); a 73232 amino acid matrix was formed after filtration of orthologs with a compositional z-score greater than −1.1 (see Nesnidal et al. 2010); a 73051 amino acid matrix after filtering orthologs with branchlength variance deviating beyond the 65th percentile (Johnson et al. 2013); a 73730 amino acid matrix after running Mare with a reduced weighting of information content (alpha); and a 73466 amino acid matrix after processing with Gblocks (settings in Supplementary Methods available on Dryad).

Two further matrix reductions both accounted for multiple sources of bias (Nucl5 and Nucl6, see Table 1). From Nucl5: 197 orthologs were removed with a P value from the compositional homogeneity test of <0.05; 167 orthologs were removed which were distributional outliers for branch-length heterogeneity (Supplementary Fig. 1a available on Dryad). Ortholog removal was excessive under default parameters for MARE (Supplementary Fig. 1b available on Dryad), so information weighting was reduced (to 0.1), under which optimality criteria was maximal under the removal of 209 orthologs. For Nucl6, orthologs were grouped (binned) according to the topological variation across their respective gene trees, then ortholog bins filtered where displaying greater degrees of bias. This was since gene-tree variants with characters known to mislead inference suggests artifactual topologies. Tree-space was characterized using an algorithm that groups gene-trees where congruent, and splits where conflicting bipartitions are observed between pairs (Mirarab et al. 2014). The size of bins is determined by a threshold for determination of a conflicting node, for example at the threshold of 70% a pair of gene trees would be in conflict in the presence of an incompatible grouping with bootstrap support above this value. The level of three potential sources of bias (compositional and branch-length heterogeneity, and information content) was tested across gene-tree bins.

Table 1. Description of data-type, filtering treatment and topological constraints of the 11 phylogenies inferred

Label	Description
Nucl0	Nuclear orthologs unfiltered
Nucl1	Nuclear orthologs reduced for compositional heterogeneity
Nucl2	Nuclear orthologs reduced for branch-length heterogeneity
Nucl3	Nuclear orthologs reduced for information content
Nucl4	Nuclear orthologs reduced for alignment quality
Nucl5	Nuclear orthologs reduced for compositional and branch-length heterogeneity, and information content
Nucl6	Nuclear ortholog *bins* reduced for compositional and branch-length heterogeneity, and information content
Unconstr-Mt	Mitogenomes unconstrained
Nucl-Mt	Mitogenomes constrained to nuclear ortholog results.
Unconstr-Sp	Species-rich partition unconstrained
Nucl-Sp	Species-rich partition constrained to a consensus of trees from Nucl1-Nucl4
Nucl-Mt-Sp	Species-rich partition constrained to tree Nucl-Mt, as per Figure 1

Phylogenies were inferred for each treatment of the transcriptome data using RaxML under the best fit model as determined by that software. The level of support for each bipartition in the primary inferred topology was described by giving its prevalence in comparison to the next most frequent, alternative bipartition (Salichos and Rokas 2013). This "internode certainty" was calculated for bipartitions on trees from supermatrix analysis, using the distributions of topologies in gene-trees, using the implementation in RaxML adjusted for incomplete gene-trees (Kobert et al. 2016).

Building a Supermatrix from Mitochondrial Genomes

A local database was formed of mitogenome protein-coding sequences from all available insect genera. Complete mitogenomes were downloaded from NCBI (http://www.ncbi.nlm.nih.gov/nucleotide/). The retrieved data was filtered to retain a single entry per genus, then partial mitogenomes added for genera not otherwise included. Protein sequences were parsed (parse_translations_from_genbank_flatfile.pl), giving a database of 10282 sequences.

The database was organized to gene via reference to a set of profiles. Profiles were made from the high-quality annotations given to "refseq" references, one randomly selected from each insect order. Genes were automatically parsed from each, and placed into a separate file (13 files, one for each protein coding gene, using the script parse_translations_from_genbank_flatfile.pl). For each of these 13 reference profiles, homologs were extracted from the 10282 sequence database using BlastP with an evalue of 1e-10. As above, results were parsed (parse_ortholog_results.pl), each protein aligned using Mafft, single gene trees inferred, paralogs removed with PhyloTreePruner, and loci concatenated.

Building a Species-Level Supermatrix

For retrieving species-rich markers I first made a local Blast database of all available Insect DNA sequences, mined and processed according to the protocol described in Chesters and Zhu (2014). Very briefly, the NCBI Invertebrate nucleotide release (ftp.ncbi.nih.gov/genbank/gbinv*.seq.gz) and taxonomy database (ftp://ftp.ncbi.nih.gov/pub/taxonomy/) was downloaded, current species-level identifiers parsed and integrated with the sequence data, redundant sequence data removed using Usearch (Edgar 2010) and a Blast searchable database formed with MakeBlastDB (Camacho 2009).

Species-rich genes were retrieved from the insect database. The barcode fragment of mitochondrial cytochrome oxidase subunit I (COI) was retrieved independently, being by far the most widely sequenced in animals. COI's were retrieved (with BlastN) using as queries the 100 sequences of the original insect profile given in appendix B of Hebert et al. (2003). Several additional widely-sequenced loci were inferred according to Chesters and Zhu (2014). A computationally tractable set of sequences was sampled (randomly selecting four sequences from each family, numbering ~2000) and subject to an all-against-all Blast search followed by Markov clustering (van Dongen 2000). This delineation of genes was independently verified according to the sequence annotations, with a small number of differently named genes discarded, and ambiguous clusters (those an assemblance of unfamiliar names) discarded altogether (fetch_gene_names.pl). Several of the top ranking gene clusters (excepting the highest ranking, COI, which was retrieved previously) were aligned, checked by eye and trimmed, then each used as profiles for retrieval of species-rich homologs from the local insect database.

I assessed several candidate software packages for aligning species-level markers; BlastAlign (Belshaw and Katzourakis 2005), Clustal Omega (Sievers et al. 2011) with and without user-specified profile HMM's, Mafft (Katoh et al. 2002) under the FFT-NS-2 algorithm, the Mothur (Schloss et al. 2009) "align.seqs" function, Pynast (Caporaso et al. 2010), and SINA (Pruesse et al. 2012). Prior to alignment each marker was filtered; (i) only sequences with complete Linnaean binomials were retained, and (ii) a single exemplar sequence was selected for each species (the sequence with the least ambiguous sites, which I have found introduce further alignment errors in the current context). For COI, Pynast aligned against a curated COI profile generated the preferred alignment. Although performance was inconsistent, and alignments from various software were used for other markers (see Supplementary data available on Dryad for more details).

Inferring the Insect Tree of Life

Supermatrices were constructed for analysis, although the combined approach (e.g. Wiens 2005) when applied to genomic and species-rich data generates a matrix structurally distinct from typical phylogenetic benchmarks; conventional concatenation with ~50,000 species and ~1600 loci generates an unwieldable ~4.2 gigabyte supermatrix of >98% missing data (see Lemmon et al. 2009). Alternatively, missing data is reduced with separate matrices, with nuclear ortholog, mitogenome, and species-level partitions individually of 34.4%, 36.5%, and 84.8% missing data, respectively. A hierarchical approach to tree-construction fits these partitions, with a phylogenetic backbone from gene-rich partitions (Fig. 1), and topological constraints imposed during species-rich inferences (e.g. Jetz et al. 2012; Erickson et al. 2014; Rainford et al. 2014).

The three partitions were variously analyzed (Table 1, lower) using RaxML where computationally feasible, and FastTree otherwise. For mitogenome tree Nucl-Mt, uncertainty was taken into account with bootstrapped mitogenome matrices, each analyzed under a randomly selected constraint topology of bootstraps conducted on Nucl1-Nucl4. Three trees were generated for the species-level matrix, with tree Nucl-Mt-Sp utilizing the most backbone information (as in Fig. 1).

Setting Backbone Constraints during Tree Searches

As described above, some tree searches were conducted in which backbone constraints were enforced. Two types of constraint were applied based on the analysis of backbone trees (implemented in the script read_deep_level_constraints.pl). "Taxon constraints" forced monophyly to members of a specific taxon, whereas "relational constraints" forced monophyly to sister taxa. For example, based on recent insect reports, a taxon constraint would be Endopterygota (all and only decedents of a specific node on the backbone tree, are of this taxon), whereas a relational constraint would be Hymenoptera sister to all remaining members of the Endopterygota. Inferring reasonable relational constraints depends on correspondence of specific parent/child nodes of the taxonomic hierarchy to the taxonomic names assigned to the three branches splitting from a node of interest on the backbone tree. In

the example above for instance, the taxon Hymenoptera are a child of the taxon Endopterygota, all other members of which are descended from the sister branch in the backbone tree. Further, relational constraints must account for taxa absent on the backbone tree; where sister taxa of a relational constraint are absent in the backbone tree, these must "float" or remain independent of the constraint in downstream inference.

The making of constraints first required assignment of taxonomic names to each node of the backbone tree. According to the species descended from a given node, the most inclusive taxonomic name was assigned, and if present, any taxa intermediate to that assigned to the parent node (of the backbone tree). After taxon labeling of the backbone topology, relational constraints were then formed for each node as thus: (from the NCBI taxonomic hierarchy) retrieve child taxa derived from the taxon assigned to the node; then for each child taxon: if the taxon is present in the terminals descended from the backbone node then assign state 1; or if the taxon exists elsewhere on the backbone tree assign state 0 (constraints are given in Supplementary Materials available on Dryad, see file "constraints"). Finally the lineage of each member in the species-level supermatrix was read and used to assign a state (one of "1", "0", "-") for each constraint, then this was used in the FastTree inference.

Utilizing the Phylogeny in Characterization of Taxonomic Diversity in Query Data

Software for phylogeny-based profiling have been developed for cases where a comprehensive reference phylogeny is not available, although there are wider software options where they are. Several placement approaches were analyzed which represented a continuum between utilizing an existing tree of life, to those placing on local de novo reference trees. Placement to the complete tree was conducted using the scalable NJ/ML software FastTree. This software lacks a formal placement algorithm, so a wrapper script (fasttree_EP.pl) was developed to read a reference tree and generate the necessary input files for fixing references only during a standard tree search. Secondly, the ML evolutionary placement algorithm implemented in RaxML places queries to a subset of branches of a fixed and complete reference tree (Berger et al. 2011). The placement approach employed in PPlacer (Matsen et al. 2010) is very similar, although in preliminary tests PPlacer was found less robust on the current data sets. For local placement, BAGpipe (Papadopoulou et al. 2014) generates a phylogeny for each group of queries and SAP (Munch et al. 2008) a phylogeny individually for each query. BAGpipe first groups queries according to broad similarity, then for each query group references are retrieved and a phylogeny is built. A simplified version of BAGpipe was used in which steps originally included to address issues specific to length variable markers (much less relevant when using COI) were omitted. SAP retrieves references with subsequent tree building and taxonomic assignment steps, individually for each query. SAP assignment was not feasible under Bayesian options, thus the constrained NJ option was used. The reference environment for taxonomic assignment for both SAP and Bagpipe was standardized, each using 12 retrieved reference sequences for each query.

Because RaxML and FastTree perform only phylogenetic placements (SAP and Bagpipe additionally conduct taxonomic assignment), for these two methods assignments were made on the query-placed trees using the software bagpipe_phylo (Chesters et al. 2015). The four methods were optimized according to their default criteria for the most part; for FastTree and Bagpipe these were nearest reference leaf and distance to nearest reference leaf (for related measures, see Matsen et al. 2013); RaxML was distal length (length of branch of the reference tree to which query is attached); and SAP was bootstrap-based probability of placement from taxonomically constrained NJ analysis. Optimization consisted of selection of parameters in which congruence in taxonomic groupings between references and queries (where of known taxa) is highest (e.g. Sauer and Hausdorf 2012; Mende et al. 2013). Congruence was calculated on species-level assignments only, the rationale being that speciation processes leaves signatures in DNA (Barraclough et al. 2003; Acinas et al. 2004; Pons et al. 2006), leading to broadly applicable thresholds for species membership, which have been well characterized in COI (Hebert et al. 2004). Benchmarking was conducted according to Chesters et al. (2015), using newly available COI's of known species (an alternative to the widely used "leave-one-out" query generation). Query Datasets (QD) 1-8 were mined from NCBI's daily updates (of the month November 2015, from ftp://ftp.ncbi.nih.gov/genbank/daily-nc/). Briefly, the daily files were downloaded, insect COI's extracted from the database as previously, and sequences (i) overlapping the references (e.g. updated entries) (ii) identical, (iii) not labeled to species level, removed. The taxonomic distribution of the data was then checked, and eight data sets selected for analyses (Carabidae, Staphylinidae, Cucujiformia, Chironomidae, Muscoidea, Oestroidea, Apoidea and Obtectomera).

A further set of query data sets were mined for reanalysis, and submitted for automated assignment to the insect tree of life; QD9 was 1050 Sanger sequenced insect COI barcodes (Porter et al. 2014; the same data set also analyzed by Gibson et al. 2014); QD10 was of 673 arthropod COI (~614 bp) from metabarcoding of material Malaise trapped in three locations in south China (Yu et al. 2012); and QD11 was composed of 7478 Hymenoptera barcode sequences (Stahlhut et al. 2013). Query to (fixed) reference alignments were conducted using Pynast or Mothur (align.seqs function), and followed by taxonomic profiling. For QD1-8 phylogeny-based taxonomic assignment was conducted using the four methods described above, whereas for QD9-12 RaxML EPA only was applied.

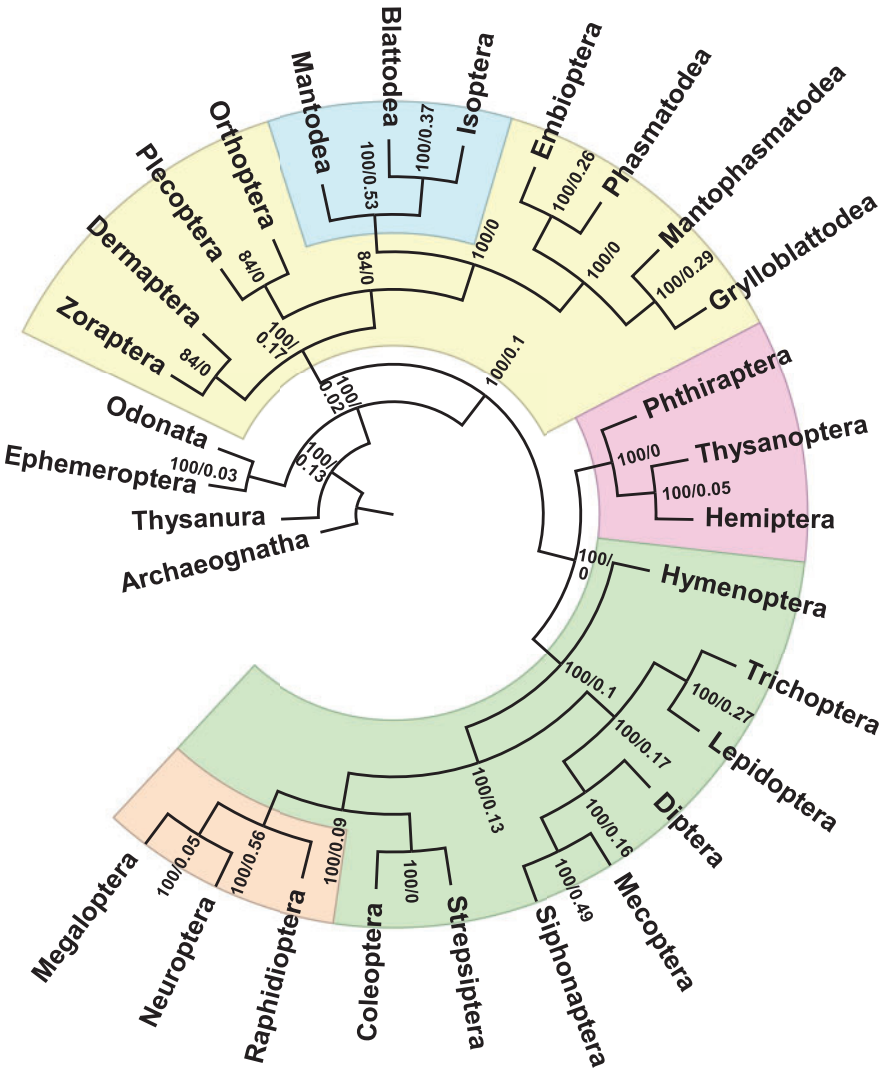

FIGURE 3. Backbone phylogeny of the insects, including all 29 orders (outgroups not shown). Generated from a matrix reduced of orthologs according to the incidence of compositional heterogeneity and branch-length variance, and information content (Nucl5). Node support is indicated by bootstrap percentage/Internode Certainty (note, IC of 0.1 corresponds to ratio of 7:3 in support of the inferred node, over the most prevalent conflicting node, whereas a value close to 0 implies a competing topology with similar support). Several key insect groupings are boxed.

RESULTS AND DISCUSSION

Impacts of Supermatrix Construction on Inference

The preferred backbone tree is given in Figure 3 (nucl5, an identical topology observed for nucl0, nucl1, nucl4 and nucl6), and species-level tree in Figure 4 (Nucl-Mt-Sp, integrating all partitions). Tree Unconstr-Sp (using only the species-rich partition) proved difficult to parse due to ubiquitous higher-rank polyphylys, and thus is not considered further. Tree Nucl-Sp (species-level tree constrained using only nuclear partition) is discussed for intra-ordinal relationships, and mitogenome Unconstr-Mt for basal intra-ordinal and sister order relationships (Talavera and Vila 2011). A general review of these and other trees is given in Supplementary Materials available on Dryad, otherwise, discussion of topological results is limited to certainty in the backbone trees; basal relationships are usually the foremost of candidates for scrutiny, those of broadest interest and often the most difficult to resolve (Talavera and Vila 2011).

Initial phylogenomic analysis was repeated on four supermatrices, each liberally filtered for different sources of bias. Conflicts were observed, particularly regarding positions of the two orders Plecoptera and Zoraptera in Polyneoptera (Supplementary Fig. 2 available on Dryad). Further, conflicting topologies all received high bootstrap support, which is consistent with the emerging evidence (Salichos and Rokas 2013; Whelan et al. 2015). This topological variation prompted further analysis of tree-space via gene-trees. Gene trees were first binned (Mirarab et al. 2014), for ease of interpretation generating the fewest number of bins by splitting gene-tree pairs only in fully supported (100% bootstrap support) conflicting topologies. There was no

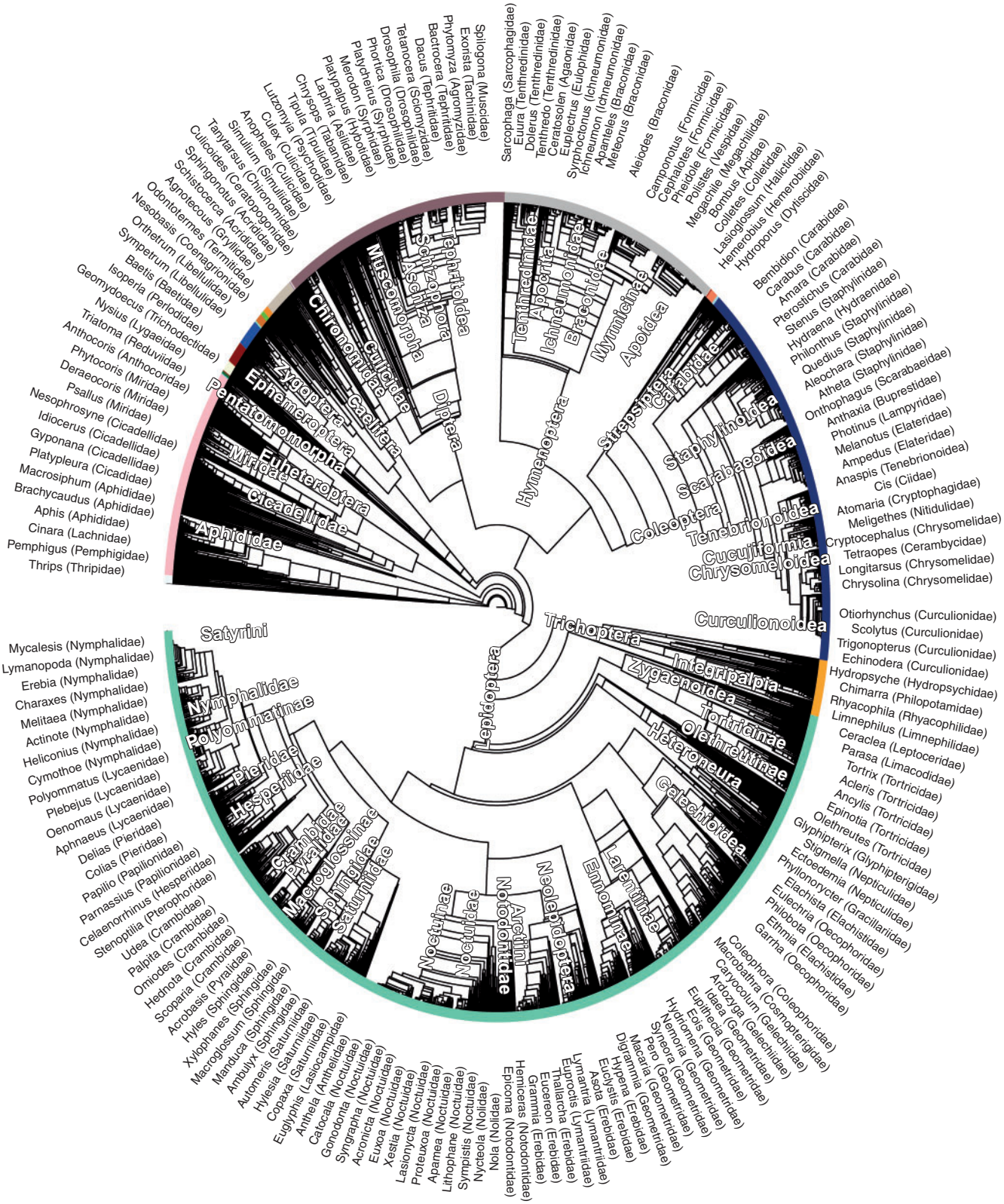

FIGURE 4. Insect phylogeny of 49358 species. Subset of tip and internal nodes are labeled. Tree circumference is variously shaded according to the 29 insect orders. High resolution version with more complete labeling is available in Supplementary Materials available on Dryad (species_level_tree_HR.png), as is complete Newick format tree (species_level_tree.nwk)

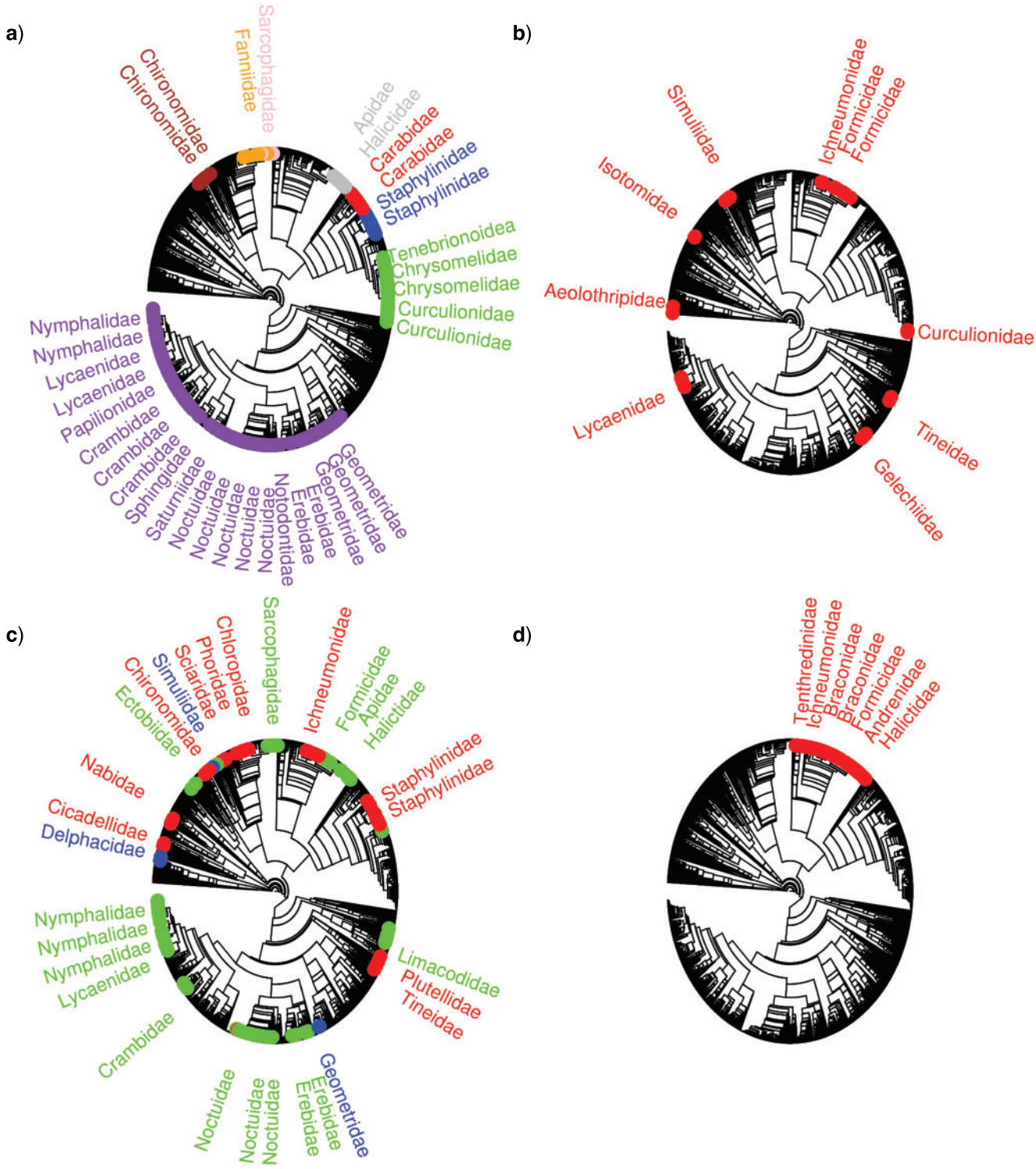

FIGURE 5. Results of evolutionary placement (Berger et al. 2011) of various query DNA barcode data sets to the insect tree of life. Colored and labeled tips indicate positions of significant query placement. a) QD1-8, all of known species, and each differently colored; b) QD9, 1050 insect DNA barcodes (Porter et al. 2014); c) QD10, 673 arthropod COI from Malaise trapping in three sites in south China, where green = Xishuangbanna, blue = Honghe, red = Kunming (Yu et al. 2012); d) QD11, 7478 Hymenoptera barcodes (Stahlhut et al. 2013).

clear evidence that topological variants described by the 17 bins thus formed were caused by the level of compositional heterogeneity, branch-length variance or information content per se (Supplementary Fig. 3 and Table 1 available on Dryad). Still, supermatrices were formed accounting for all measured bias, one filtered at the level of ortholog, and one at the level of ortholog-bin, each giving identical topologies (Fig. 3).

TABLE 2. Accuracy of phylogeny-based taxonomic profiling

Taxon	Number of queries	Proportion in references	RaxML	FastTree	SAP	BAGpipe
Carabidae	312	0.73	0.705	0.821	0.682	0.821
Staphylinidae	380	0.86	0.913	0.892	0.4263	0.874
Cucujiformia	878	0.74	0.897	0.885	0.4612	0.755
Chironomidae	718	0.87	0.811	0.801	0.448	0.763
Muscoidea	782	0.95	0.896	0.909	0.5179	0.898
Oestroidea	441	0.81	0.739	0.82	0.5532	0.800
Apoidea	73	0.85	0.767	0.821	0.808	0.767
Obtectomera	582	0.79	0.706	0.761	0.754	0.756

Notes: Column 1 names the eight test data set composed of species-labeled sequences; Column 2 gives the number of sequences in the test data set; Column 3 gives the proportion of sequences of each test query set which are from species also included in the reference tree. Columns 4–7 give the rate of correct species level assignments for each of four methods. SAP and BAGpipe implement both phylogenetic inference and taxonomic assignment, for RaxML and FastTree, taxonomies were parsed from placed phylogenies using bagpipe_phylo.

Although robust bootstrap support is observed in transcriptome trees, and some topological consistency, low internode certainty (second node values of Fig. 3) suggests some inferred bipartitions are not greatly supported over alternatives. Internode certainty well reflected topological variation observed in initial supermatrix filterings (nucl1-4; Supplementary Fig. 2 available on Dryad). Higher support was observed in the Endopterygota, a group comprising 85% of all insects. Within these, topologies were congruent on all matrix reductions herein, and with recent character rich morphological and nuclear studies (Beutel et al. 2010; Trautwein et al. 2012; Misof et al. 2014). Outside Endopterygota, Dictyoptera (Blattaria, Mantodea and Isoptera) was supported, as was a sister relationships of the latter two (a grouping termed Condylognatha). Monophyly of the Paraneoptera (Phthiraptera, Thysanoptera, and Hemiptera) was consistent across matrix treatments, whereas support was lower and some conflicts observed across treatments for several taxa in the Polyneoptera.

Benchmarking Phylogenetic Placement of Insect DNA Barcodes

The necessity for a comprehensive tree for taxonomic profiling is not implicit since local profiling methods exist that generate local trees de novo. However, phylogenetic placement of eight benchmark data sets (Table 2 and Fig. 5a) revealed species assignments were generally more accurate with methods conducted in the context of a more comprehensive reference phylogeny. Significant improvements were observed between assignment on the whole tree versus individually for each query (FastTree versus SAP, $P = 0.007813$, weighted Wilcoxon signed rank test). There were indicative (nonsignificant, a strict Bonferroni correction cutoff being 0.0083) improvements in accuracy where conducted on the whole tree as opposed to using grouped queries (FastTree versus Bagpipe, $P = 0.01563$) or using grouped queries as opposed to building small reference trees individually for each query (Bagpipe versus SAP, $P = 0.02344$). By extension, accuracy within an individual method might be improved by adjustment of parameters to use more comprehensive reference trees, as observed to a point (<1/32 of edges) with increased reference size (fraction of insertion edges) under the RaxML EPA heuristic (Berger et al. 2011). In the case of BAGpipe, a minor (nonsignificant, $P = 0.148$) increase in accuracy is observed using 54 query+reference groups (clustering at an average linkage of 85% similarity forming groups broadly coincident with genera) compared with 76 groups (clustering at 90%).

The reduced accuracy of local assignment methods sometimes observed may be partly an organizational effect. Under an empirical distribution of query and references, some queries are assigned many references, and others so few as to obstruct normal functioning of subsequent phylogenetic profiling steps (e.g. Papadopoulou et al. 2014). For example, despite a standardized effort in reference retrieval for SAP and Bagpipe, the former proved less successful in construction of appropriate reference sets. In the test data set Cucujiformia, for the 646 queries of species that were included in the references, Bagpipe returned the correct species name for over twice the number of queries (445 versus 201). The majority (87%) of cases in which SAP did not return the correct species were due to failure to capture conspecifics into the phylogenetic framework that placement occurs. These results indicate that organizational subdivision of de novo phylogenies may work to counteract gains in accuracy usually observed in phylogeny-based taxonomic profiling.

Taxonomic Profiling of Community Data on the Insect Tree of Life

ML based evolutionary placement was next applied for reanalysis of previously published insect DNA barcoding data sets (QD9-QD12; Fig. 5b–d). Note reference data used herein would not necessarily be equivalent to that used earlier publications (informal comparisons are made nonetheless). Statistically reliable placements were inferred based on Likelihood ratio (calculated from the local likelihood under placement to

the optimal branch versus alternatives) under thresholds determined empirically above. Porter et al. (2014) applied the prokaryote RDP Bayesian Classifier (Wang et al. 2007) to 1052 specimens (949 insects), and assigned taxa to the rank of order (Diptera >Hymenoptera >Coleoptera). The same set of specimens was analyzed by Gibson et al. (2014) with Megan-assigned taxonomies (Huson et al. 2011) using Blast top-hits, and a rate of family, genus assignment of 51%, 19% respectively (Diptera >Hymenoptera >Coleoptera >Lepidoptera). Submitted to phylogeny-based taxonomic assignment herein, the rate of family and genus assignments were 129 (Isotomidae >Formicidae >Tineidae), 85 (Cryptopygus >Dryadaula >Pseudomyrmex). Yu et al. (2012) generated 673 unique haplotypes from three Malaise samples. SAP (Munch et al. 2008) assignment gave rates of species, genus and family identification each ~36%, and where OTU counts were presented per order (Lepidoptera >Diptera >Hymenoptera >Coleoptera >non-Endopterygota). Phylogeny-based assignment herein gave counts of family, genus assignment of 130 (Erebidae >Ichneumonidae >Noctuidae >Nymphalidae = Pyralidae), 60 (including Condica, Zizina, Bradysia, Adoretus). Stahlhut et al. (2013) presented >7000 sub-arctic Hymenoptera sequences dominated by the parasitoids (Ichneumonoidea >Chalcidoidea >Diapriidae) and tabulated to family level. Here, after dereplication of 7478 sequences to 4057 unique haplotypes and EPA assignment to the insect phylogeny, significant placements were attained for just under half, with counts to family (Ichneumonidae >Braconidae >Tenthredinidae >Diapriidae) and genus (Atractodes >Dolichovespula >Bombus >Syrphophilus) of 2561 and 953, respectively.

Conclusions

Objective realization of the tree of life requires that some considerable methodological gaps be addressed. Genomic matrix processing strongly influences backbone topology, current approaches to matrix processing are largely ad hoc, and reliance on bootstrapping (in previous insect studies and more generally) may not have adequately described the level of node support. Furthermore, phyloinformatics protocols are poorly aligned to the structure of informative data. As presented herein, pipelines forked along gene- and species-rich partitions with subsequent integration enable upscaling, permitting generation of trees that are both well resolved at the deep level while still holding information on diversity through to the rank of species. The resulting draft insect tree is a resource for further studies on diversification and enables advanced descriptions of insect community data. Placement of thousands of barcodes to the tree herein is not only a demonstration of utility, but also that further increasing species-breadth through incorporation of upcoming barcode releases is a trivial addition to the computational framework.

In addition to species-breadth, inclusion of the genomic dimension in phylogenetic frameworks is required since there is a current trend in PCR-free genetic characterization of communities, although phylogeny-based placement of pooled metagenomes is a considerable bioinformatical challenge (Gómez-Rodríguez et al. 2015; Andújar et al. 2015). However, with an advance in the generation rate of reference genomes (Crampton-Platt et al. 2015), multi-gene organization of metagenomic data (Chesters et al. 2015), organization of informative partitions into a comprehensive phylogenetic framework (herein), and development of tree-based genomic placement, there is great potential for advancing the understanding of natural communities (Davies et al. 2012).

Supplementary Material

Data available from the Dryad Digital repository at http://dx.doi.org/10.5061/dryad.27114.

"Online Supplementary Document": Supplementary methods, results, discussion, figures and tables.

"constraints": readable set of relational constraints on species-level inference.

"species_level_tree_HR.png": High resolution version of tree depicted in Figure 4, also with more complete labeling.

"species_level_tree.nwk": complete species-level insect tree in Newick format, appropriate for input into high capacity tree viewing software and other analyses. Additionally available at TreeBase: http://purl.org/phylo/treebase/phylows/study/TB2:S20032.

Supermatrices in nexus format; "supermatrix_nuclear_genomic.nex," "supermatrix_mitogenomic.nex," "supermatrix_species_level.nex."

Acknowledgments

The author is indebted to Toby Hunt (Wellcome Trust Sanger Institute, Hinxton, UK) for giving an introduction to many of the key components of insect phyloinformatics. The author would also like to thank Chao-Dong Zhu and Qing-Yan Dai (Institute of Zoology, Beijing, China) for essential support during this work. Several editors and reviewers contributed to improving this work, of which the author would particularly like to thank Frank Anderson (Southern Illinois University, Carbondale, IL, USA) for extensive suggestions.

Funding

This work was supported by a grant (No 31471975) from the National Science Foundation of China, a grant (No 2015VBC058) from CAS President's International Fellowship Initiative (PIFI) for visiting scientists, and a grant (No 31550110209) from the National Natural Science Foundation of China's International (Regional) Cooperation and Exchange Program, each to the author.

Software Availability

A Linux implementation of the protocol described here is made freely available under the GNU general public license at https://sourceforge.net/projects/sophi/.

References

Acinas S.G., Klepac-Ceraj V., Hunt D.E., Pharino C., Ceraj I., Distel D.L., Polz M.F. 2004. Fine-scale phylogenetic architecture of a complex bacterial community. Nature 430:551–554.

Andújar C., Arribas P., Ruzicka F., Crampton-Platt A., Timmermans M.J., Vogler A.P. 2015. Phylogenetic community ecology of soil biodiversity using mitochondrial metagenomics. Mol. Ecol. 24: 3603–3617.

Barraclough T.G., Birky Jr. C.W., Burt A. 2003. Diversification in sexual and asexual organisms. Evolution 57:2166–2172.

Belshaw R., Katzourakis A. 2005. BlastAlign: a program that uses blast to align problematic nucleotide sequences. Bioinformatics 21:122–123.

Benson D.A., Karsch-Mizrachi I., Lipman D.J., Ostell J., Wheeler D.L. 2005. GenBank. Nucleic Acids Res. 33:D34–D38.

Berger S.A., Krompass D., Stamatakis A. 2011. Performance, accuracy, and web server for evolutionary placement of short sequence reads under maximum likelihood. Syst. Biol. 60:291–302.

Bergsten J. 2005. A review of long-branch attraction. Cladistics 21: 163–193.

Beutel R.G., Friedrich F., Hornschemeyer T., Pohl H., Hunefeld F., Beckmann F., Meier R., Misof B., Whiting M.F., Vilhelmsen L. 2010. Morphological and molecular evidence converge upon a robust phylogeny of the megadiverse Holometabola. Cladistics 27:341–355.

Bininda-Emonds O.R.P. 2004. The evolution of supertrees. Trends Ecol. Evol. 19:315–322.

Blaxter M., Mann J., Chapman T., Thomas F., Whitton C., Floyd R., Abebe E. 2005. Defining operational taxonomic units using DNA barcode data. Phil. Trans. R. Soc. B 360:1935–1943.

Bocak L., Barton C., Crampton-Platt A., Chesters D., Ahrens D., Vogler A.P. 2014. Building the Coleoptera tree-of-life for >8000 species: composition of public DNA data and fit with Linnaean classification. Syst. Entomol. 39:97–110.

Camacho C., Coulouris G., Avagyan V., Ma N., Papadopoulos J., Bealer K. Madden T.L. 2009. BLAST+: architecture and applications. BMC Bioinformatics 10:1–421.

Cameron S.L. 2014. Insect mitochondrial genomics: implications for evolution and phylogeny. Ann. Rev. Entomol. 59:95–117.

Caporaso J.G., Bittinger K., Bushman F.D., DeSantis T.Z., Andersen G.L., Knight R. 2010. PyNAST: a flexible tool for aligning sequences to a template alignment. Bioinformatics 26:266–267.

Castresana J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol. Biol. Evol. 17:540–552.

Chesters D., Zhu C.-D. 2014. A protocol for species delineation of public DNA databases, applied to the insecta. Syst. Biol. 63:712–725.

Chesters D., Zheng W.-M., Zhu C.-D. 2015. A DNA barcoding system integrating multigene sequence data. Methods Ecol. Evol. 6:930–937.

Clare E.L., Fraser E.E., Braid H.E., Fenton M.B., Hebert P.D.N. 2009. Species on the menu of a generalist predator, the eastern red bat (*Lasiurus borealis*): using a molecular approach to detect arthropod prey. Mol. Ecol. 18:2532–2542.

Crampton-Platt A., Timmermans M.J., Gimmel M.L., Kutty S.N., Cockerill T.D., Vun Khen C., Vogler A.P. 2015. Soup to tree: the phylogeny of beetles inferred by mitochondrial metagenomics of a Bornean rainforest sample. Mol. Biol. Evol. 32:2302–2316.

Dávalos L.M., Perkins S.L. 2008. Saturation and base composition bias explain phylogenomic conflict in Plasmodium. Genomics 91: 433–442.

Davies N., Meyer C., Gilbert J.A., Amaral-Zettler L., Deck J., Bicak M., Rocca-Serra P., Assunta-Sansone S., Willis K., Field D. 2012. A call for an international network of genomic observatories (GOs). GigaScience 1:5.

Driskell A.C., Ané C., Burleigh J.G., McMahon M.M., O'Meara B.C., Sanderson M.J. 2004. Prospects for building the tree of life from large sequence databases. Science 306:1172–1174.

Dunn C.W., Hejnol A., Matus D.Q., Pang K., Browne W.E., Smith S.A., Seaver E., Rouse G.W., Obst M., Edgecombe G.D., Sørensen M.V., Haddock S.H., Schmidt-Rhaesa A., Okusu A., Kristensen R.M., Wheeler W.C., Martindale M.Q., Giribet G. 2008. Broad phylogenomic sampling improves resolution of the animal tree of life. Nature 452:745–749.

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 R.C. 2010. Search and clustering orders of magnitude faster than BLAST. Bioinformatics 26:2460–2461.

Erickson D.L., Jones F.A., Swenson N.G., Pei N., Bourg N.A., Chen W., Davies S.J., Ge X.-j., Hao Z., Howe R.W., Huang C.-L., Larson A.J., Lum S.K.Y., Lutz J.A., Ma K., Meegaskumbura M., Mi X., Parker J.D., Fang-Sun I., Wright S.J., Wolf A.T., Ye W., Xing D., Zimmerman J.K. and Kress W.J. 2014. Comparative evolutionary diversity and phylogenetic structure across multiple forest dynamics plots: a mega-phylogeny approach. Front. Genet. 5:358.

Filipski A., Tamura K., Billing-Ross P., Murillo O., Kumar S. 2015. Phylogenetic placement of metagenomic reads using the minimum evolution principle. BMC Genomics 16 (Suppl 1):S13.

Foster P.G. 2004. Modeling compositional heterogeneity. Syst. Biol. 53:485–495.

Garrison N.L., Rodriguez J., Agnarsson I., Coddington J.A., Griswold C.E., Hamilton C.A., Hedin M., Kocot K.M., Ledford J.M., Bond J.E. 2016. Spider phylogenomics: untangling the Spider Tree of Life. PeerJ. 4:e1719.

Gibson J., Shokralla S., Porter T.M., King I., van Konynenburg S., Janzen D.H., Hallwachs W., Hajibabaei M. 2014. Simultaneous assessment of the macrobiome and microbiome in a bulk sample of tropical arthropods through DNA metasystematics. Proc. Natl Acad. Sci. USA 111:8007–8012.

Göker M., García-Blázquez G., Voglmayr H., Tellería M.T., Martín M.P. 2009. Molecular taxonomy of phytopathogenic fungi: a case study in Peronospora. PLoS One 4:e6319.

Gómez-Rodríguez C., Crampton-Platt A., Timmermans M.J.T.N., Baselga A., Vogler A.P. 2015. Validating the power of mitochondrial metagenomics for community ecology and phylogenetics of complex assemblages. Methods Ecol. Evol. 6:883–894

Grimaldi D., Engel M.S. 2005. The evolution of the insects. New York: Cambridge University Press.

Hajibabaei M., Shokralla S., Zhou X., Singer G.A.C, Baird D.J. 2011. Environmental barcoding: a next-generation sequencing approach for biomonitoring applications using river benthos. PLoS One 6:e17497.

Haran J., Timmermans M.J.T.N., Vogler A.P. 2013. Mitogenome sequences stabilize the phylogenetics of weevils (Curculionoidea) and establish the monophyly of larval ectophagy. Mol. Phylogenet. Evol. 67:156–166.

Hasegawa M., Hashimoto T. 1993. Ribosomal RNA trees misleading? Nature 361:23.

Hebert P.D.N., Ratnasingham S., DeWaard J.R. 2003. Barcoding animal life: cytochrome c oxidase subunit 1 divergences among closely related species. Proc. R. Soc. B. Biol. Sci. 270:S596–S599.

Hebert P.D., Stoeckle M.Y., Zemlak T.S., Francis C.M. 2004. Identification of birds through DNA barcodes. PLoS Biol. 2:e312.

Hedtke S.M., Patiny S., Danforth B.N. 2013. The bee tree of life: a supermatrix approach to Apoid phylogeny and biogeography. BMC Evol. Biol. 13:138.

Heraty J.M., Ronquist F., Carpenter J.M., Hawks D., Schulmeister S., Dowling A.P., Murray D., Munro J., Wheeler W.C., Schiff N., Sharkey M. 2011. Evolution of the hymenopteran megaradiation. Mol. Phylogenet. Evol. 60:73–88.

Huemer P., Mutanen M., Sefc K.M., Hebert P.D.N. 2014. Testing DNA barcode performance in 1000 species of European Lepidoptera: large geographic distances have small genetic impacts. PLoS One 9:e115774.

Huson D.H., Mitra S., Ruscheweyh H.J., Weber N., Schuster S.C. 2011. Integrative analysis of environmental sequences using MEGAN4.

Genome Res. 21:1552–1560.
Hinchliff C.E., Smith S.A., Allman J.F., Burleigh J.G., Chaudhary R., Coghill L.M., Crandall K.A., Deng J., Drew B.T., Gazis R., Gude K., Hibbett D.S., Katz L.A., Laughinghouse H.D., McTavish E.J., Midford P.E., Owen C.L., Ree R.H., Rees J.A., Soltis D.E., Williams T., Cranstonk K.A. 2015. Synthesis of phylogeny and taxonomy into a comprehensive tree of life. Proc. Natl. Acad. Sci. USA 112:12764–12769.
Ives A.R., Helmus M.R. 2010. Phylogenetic metrics of community similarity. Am. Nat. 176:E128–E142.
Jeffroy O., Brinkmann H., Delsuc F., Philippe H. 2006. Phylogenomics: the beginning of incongruence? Trends Genet. 22:225–231.
Jetz W., Thomas G.H., Joy J.B., Hartmann K., Mooers A.O. 2012. The global diversity of birds in space and time. Nature 491(7424): 444–448.
Ji Y., Ashton L., Pedley S.M., Edwards D.P., Tang Y., Nakamura A., Kitching R., Dolman P.M., Woodcock P., Edwards F.A., Larsen T.H., Hsu W.W., Benedick S., Hamer K.C., Wilcove D.S., Bruce C., Wang X., Levi T., Lott M., Emerson B.C., Yu D.W. 2013. Reliable, verifiable and efficient monitoring of biodiversity via metabarcoding. Ecol. Lett. 16:1245–1257.
Johnson B.R., Borowiec M.L., Chiu J.C., Lee E.K., Atallah J., Ward P.S. 2013. Phylogenomics resolves evolutionary relationships among ants, bees, and wasps. Curr. Biol. 23:2058–2062.
Katoh K., Misawa K., Kuma K., Miyata T. 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res. 30:3059–3066.
Kobert K., Salichos L., Rokas A., Stamatakis A. 2016. Computing the internode certainty and related measures from partial gene trees. Mol. Biol. Evol. 33:1606–1617.
Kocot K.M., Citarella M.R., Moroz L.L., Halanych K.M. 2013. PhyloTreePruner: a phylogenetic tree-based approach for selection of orthologous sequences for phylogenomics. Evol. Bioinformatics 9:429–435.
Kress W.J., Erickson D.L., Swenson N.G., Thompson J., Uriarte M., Zimmerman J.K. 2010. Advances in the use of DNA barcodes to build a community phylogeny for tropical trees in a Puerto Rican forest dynamics plot. PLoS One 5:e15409.
Kueck P., Meusemann K., Dambach J., Thormann B., von Reumont B.M., Waegele J.W. 2010. Parametric and non-parametric masking of randomness in sequence alignments can be improved and leads to better resolved trees. Front. Zool. 7:1–12.
Lartillot N., Philippe H. 2004. Bayesian phylogenetic software based on mixture models. Mol. Biol. Evol. 21:1095–1109.
Lemmon A.R., Brown J.M., Stanger-Hall K., Lemmon E.M. 2009. The effect of ambiguous data on phylogenetic estimates obtained by Maximum Likelihood and Bayesian Inference. Syst. Biol. 58:130–145.
Letsch H.O., Meusemann K., Wipfler B., Schütte K., Beutel R., Misof B. 2012. Insect phylogenomics: results, problems and the impact of matrix composition. Proc. Biol. Sci. 279:3282–3290.
Maddison D.R., Schulz K.-S. editors. 2007. The tree of life web project. Available from: http://tolweb.org
Matsen F.A., Gallagher A., McCoy C.O. 2013. Minimizing the average distance to a closest leaf in a phylogenetic tree. Syst. Biol. 62:824–836.
Matsen F.A., Kodner R.B., Armbrust E.V. 2010. pplacer: linear time maximum-likelihood and Bayesian phylogenetic placement of sequences onto a fixed reference tree. BMC Bioinformatics 11:538.
McMahon M.M., Deepak A., Fernández-Baca D., Boss D., Sanderson M.J. 2015. STBase: one million species trees for comparative biology. PLoS One 10:e0117987.
McMahon M.M., Sanderson M.J. 2006. Phylogenetic supermatrix analysis of GenBank sequences from 2228 papilionoid legumes. Syst. Biol. 55:818–836.
Mende D.R., Sunagawa S., Zeller G., Bork P. 2013. Accurate and universal delineation of prokaryotic species. Nat. Methods 10: 881–884.
Meusemann K., von Reumont B.M., 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 J.W., Misof B. 2010. A phylogenomic approach to resolve the arthropod tree of life. Mol. Biol. Evol. 27:2451–2464.
Meyer B., Meusemann K., Misof B. 2011. MARE: MAtrix REduction - A tool to select optimized data subsets from supermatrices for phylogenetic inference. Version 0.1.2-rc. Zentrum für molekulare Biodiversitätsforschung (zmb) am ZFMK, Adenauerallee 160, 53113 Bonn, Germany.
Mirarab S., Bayzid M.S., Boussau B., Warnow T. 2014. Statistical binning enables an accurate coalescent-based estimation of the avian tree. Science 346:1250463.
Misof B., Liu S., Meusemann K., Peters R.S., Donath A., Mayer C., Frandsen P.B., Ware J., Flouri T., Beutel R.G., Niehuis O., Petersen M., Izquierdo-Carrasco F., Wappler T., Rust J., Aberer A.J., Aspöck U., Aspöck H., Bartel D., Blanke A., Berger S., Böhm A., Buckley T.R., Calcott B., Chen J., Friedrich F., Fukui M., Fujita M., Greve C., Grobe P., Gu S., Huang Y., Jermiin L.S., Kawahara A.Y., Krogmann L., Kubiak M., Lanfear R., Letsch H., Li Y., Li Z., Li J., Lu H., Machida R., Mashimo Y., Kapli P., McKenna D.D., Meng G., Nakagaki Y., Navarrete-Heredia J.L., Ott M., Ou Y., Pass G., Podsiadlowski L., Pohl H., von Reumont B.M., Schütte K., Sekiya K., Shimizu S., Slipinski A., Stamatakis A., Song W., Su X., Szucsich N.U., Tan M., Tan X., Tang M., Tang J., Timelthaler G., Tomizuka S., Trautwein M., Tong X., Uchifune T., Walzl M.G., Wiegmann B.M., Wilbrandt J., Wipfler B., Wong T.K., Wu Q., Wu G., Xie Y., Yang S., Yang Q., Yeates D.K., 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 K.M., Zhou X. 2014. Phylogenomics resolves the timing and pattern of insect evolution. Science 346:763–767.
Munch K., Boomsma W., Huelsenbeck J.P., Willerslev E., Nielsen R. 2008. Statistical assignment of DNA sequences using Bayesian phylogenetics. Syst. Biol. 57:750–757.
Nesnidal M.P., Helmkampf M., Bruchhaus I., Hausdorf B. 2010. Compositional heterogeneity and phylogenomic inference of metazoan relationships. Mol. Biol. Evol. 27:2095–2104.
Papadopoulou A., Chesters D., Coronado I., De la Cadena G., Cardoso A., Reyes J.C., Maes J.M., Rueda R.M., Gómez-Zurita J. 2014 Automated DNA-based plant identification for large-scale biodiversity assessment. Mol. Ecol. Res. 15:136–152.
Pereira H.M., Ferrier S., Walters M., Geller G.N., Jongman R.H., Scholes R.J., Bruford M.W., Brummitt N., Butchart S.H., Cardoso A.C., Coops N.C., Dulloo E., Faith D.P., Freyhof J., Gregory R.D., Heip C., Höft R., Hurtt G., Jetz W., Karp D.S., McGeoch M.A., Obura D., Onoda Y., Pettorelli N., Reyers B., Sayre R., Scharlemann J.P., Stuart S.N., Turak E., Walpole M., Wegmann M. 2013. Essential biodiversity variables. Science 339: 277–278
Peters R.S., Meyer B., Krogmann L., Borner J., Meusemann K., Schutte K., Niehuis O., Misof B. 2011. The taming of an impossible child - a standardized all-in approach to the phylogeny of Hymenoptera using public database sequences. BMC Biol. 9:55.
Philippe H., Brinkmann H., Lavrov D.V., Littlewood D.T.J., Manuel M., Wörheide G., Baurain D. 2011. Resolving difficult phylogenetic questions: why more sequences are not enough. PLoS Biol 9:e1000602.
Philippe H., Snell E.A., Bapteste E., Lopez P., Holland P.W.H. 2004. Phylogenomics of eukaryotes: impact of missing data on large alignments. Mol. Biol. Evol. 21:1740–1752.
Pickett S.B., Bergey C.M., Di Fiore A. 2012. A metagenomic study of primate insect diet diversity. Am. J. Primatol. 74:622–631.
Pons J., Barraclough T.G., Gomez-Zurita J., Cardoso A., Duran D.P., Hazell S., Kamoun S., Sumlin W.D., Vogler A.P. 2006. Sequence-based species delimitation for the DNA taxonomy of undescribed insects. Syst. Biol. 55:595–609.
Porter T.M., Gibson J.F., Shokralla S., Baird D.J., Golding G.B., Hajibabaei M. 2014. Rapid and accurate taxonomic classification of insect (class Insecta) cytochrome c oxidase subunit 1 (COI) DNA barcode sequences using a naïve Bayesian classifier. Mol. Ecol. Res. 14:929–942.
Price M.N., Dehal P.S., Arkin A.P. 2009. FastTree: computing large minimum-evolution trees with profiles instead of a distance matrix. Mol. Biol. Evol. 26:1641–1650.
Pruesse E., Peplies J., Glöckner F.O. 2012. SINA: accurate high-throughput multiple sequence alignment of ribosomal RNA genes. Bioinformatics 28:1823–1829.

Rainford J.L., Hofreiter M., Nicholson D.B., Mayhew P.J. 2014. Phylogenetic distribution of extant richness suggests metamorphosis is a key innovation driving diversification in insects. PLoS One 9:e109085.
Regier J.C., Mitter C., Zwick A., Bazinet A.L., Cummings M.P., Kawahara A.Y., Sohn J.-C., Zwickl D.J., Cho S., Davis D.R., Baixeras J., Brown J., Parr C., Weller S., Lees D.C., Mitter K.T. 2013. A large-scale, higher-level, molecular phylogenetic study of the insect order Lepidoptera (moths and butterflies). PLoS One 8:e58568.
Rokas A., Williams B.L., King N., Carroll S.B. 2003. Genome-scale approaches to resolving incongruence in molecular phylogenies. Nature 425:798–804.
Sanderson M.J., Boss D., Chen D., Cranston K.A., Wehe A. 2008. The PhyLoTA browser: processing GenBank for molecular phylogenetics research. Syst. Biol. 57:335–346.
Sanderson M.J., Driskell A.C., Ree R.H., Eulenstein O., Langley S. 2003. Obtaining maximal concatenated phylogenetic data sets from large sequence databases. Mol. Biol. Evol. 20:1036–1042.
Salichos L., Rokas A. 2013. Inferring ancient divergences requires genes with strong phylogenetic signals. Nature 497: 327–331
Sauer J., Hausdorf B. 2012. A comparison of DNA-based methods for delimiting species in a Cretan land snail radiation reveals shortcomings of exclusively molecular taxonomy. Cladistics 28: 300–316.
Savard J., Tautz D., Richards S., Weinstock G.M., Gibbs R.A., Werren J.H., Tettelin H., Lercher M.J. 2006. Phylogenomic analysis reveals bees and wasps (Hymenoptera) at the base of the radiation of Holometabolous insects. Genome Res. 16:1334–1338.
Schloss P.D., Westcott S.L., Ryabin T., Hall J.R., Hartmann M., Hollister E.B., Lesniewski R.A., Oakley B.B., Parks D.H., Robinson C.J., Sahl J.W., Stres B., Thallinger G.G., Van Horn D.J., Weber C.F. 2009. Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. Appl. Environ. Microbiol. 75:7537–7541.
Shokralla S., Gibson J.F., Nikbakht H., Janzen D.H., Hallwachs W., Hajibabaei M. 2014. Next-generation DNA barcoding: using next-generation sequencing to enhance and accelerate DNA barcode capture from single specimens. Mol. Ecol. Res. 14:892–901
Sievers F., Wilm A., Dineen D., Gibson T.J., Karplus K., Li W., Lopez R., McWilliam H., Remmert M., Soding J., Thompson J.D., Higgins D.G. 2011. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol. Syst. Biol. 7:539.
Simmons M.P. 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.
Simon S., Strauss S., von Haeseler A., Hadrys H. 2009. A phylogenomic approach to resolve the basal pterygote divergence. Mol. Biol. Evol. 26:2719–2730.
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 S.A., Beaulieu J.M., Donoghue M.J. 2009. Mega-phylogeny approach for comparative biology: an alternative to supertree and supermatrix approaches. BMC Evol. Biol. 9:37.
Smith S.A., Brown J.W., Hinchliff C.E. 2013. Analyzing and synthesizing phylogenies using tree alignment graphs. PLOS Comput. Biol. 9:e1003223.
Smith M.A., Eveleigh E.S., McCann K.S., Merilo M.T., McCarthy P.C., Van Rooyen K.I. 2011. Barcoding a quantified food web: crypsis, concepts, ecology and hypotheses. PLoS One 6:e14424.
Stahlhut J.K., Fernández-Triana J., Adamowicz S.J., Buck M., Goulet H., Hebert P.D., Huber J.T., Merilo M.T., Sheffield C.S., Woodcock T., Smith M.A. 2013. DNA barcoding reveals diversity of Hymenoptera and the dominance of parasitoids in a sub-arctic environment. BMC Ecol. 13:2.
Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30:1312–1313.
Sukumaran J., Holder M.T. 2010. DendroPy: a Python library for phylogenetic computing. Bioinformatics 26:1569–1571.
Taberlet P., Coissac E., Pompanon F., Brochmann C., Willerslev E. 2012. Towards next-generation biodiversity assessment using DNA metabarcoding. Mol. Ecol. 21:2045–2050.
Talavera G., Castresana J. 2007. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. Syst. Biol. 56:564–577.
Talavera G., Vila R. 2011. What is the phylogenetic signal limit from mitogenomes? The reconciliation between mitochondrial and nuclear data in the Insecta class phylogeny. BMC Evol. Biol. 11:315.
Tang M., Tan M., Meng G., Yang S., Su X., Liu S., Song W., Li Y., Wu Q., Zhang A., Zhou X. 2014. Multiplex sequencing of pooled mitochondrial genomes - a crucial step toward biodiversity analysis using mito-metagenomics. Nucleic Acids Res. 42:e166.
Timmermans M.J.T.N., Dodsworth S., Culverwell C.L., Bocak L., Ahrens D., Littlewood D.T.J., Pons J., Vogler A.P. 2010. Why barcode? High-throughput multiplex sequencing of mitochondrial genomes for molecular systematics. Nucleic Acids Res. 38:e197.
Timmermans M.J.T.N., Lees D.C., Simonsen T.J. 2014. Towards a mitogenomic phylogeny of Lepidoptera. Mol. Phylogenet. Evol. 79:169–178.
Trautwein M.D., Wiegmann B.M., Beutel R., Kjer K.M., Yeates D.K. 2012. Advances in insect phylogeny at the dawn of the postgenomic era. Ann. Rev. Entomol. 57:449–468.
van Dongen S. 2000. Graph clustering by flow simulation [PhD thesis]. University of Utrecht.
Wang Q., Garrity G.M., Tiedje J.M. Cole J.R. 2007. Naïve Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. Appl. Environ. Microbiol. 73:5261–5267
Wiens J.J., Lapoint R.T., Whiteman N.K. 2015. Herbivory increases diversification across insect clades. Nat. Commun. 6:8370.
Whelan N.V., Kocot K.M., Moroz L.L., Halanych K.M. 2015. Error, signal, and the placement of Ctenophora sister to all other animals. Proc. Natl Acad. Sci. USA 112:5773–5778
Yu D.W., Ji Y., Emerson B.C., Wang X., Ye C., Yang C., Ding Z. 2012. Biodiversity soup: metabarcoding of arthropods for rapid biodiversity assessment and biomonitoring. Methods Ecol. Evol. 3:613–623.
Zmasek C.M., Eddy S.R. 2001. ATV: display and manipulation of annotated phylogenetic trees. Bioinformatics 17:383–384.
Zwick A., Regier J.C., Mitter C., Cummings M.P. 2011. Increased gene sampling yields robust support for higher-level clades within Bombycoidea (Lepidoptera). Syst. Entomol. 36:31–43.