Syst. Biol. 63(5):712–725, 2014

DOI:10.1093/sysbio/syu038
Advance Access publication June 14, 2014

A Protocol for Species Delineation of Public DNA Databases, Applied to the Insecta

DOUGLAS CHESTERS AND CHAO-DONG ZHU*

Key Laboratory of Zoological Systematics and Evolution (CAS), Institute of Zoology, Chinese Academy of Sciences, Beijing 100101, PR China
**Correspondence to be sent to: Key Laboratory of Zoological Systematics and Evolution, Institute of Zoology, Chinese Academy of Sciences, Beijing 100101, PR China; E-mail: zhucd@ioz.ac.cn.*

Received 16 August 2013; reviews returned 21 November 2013; accepted 10 May 2014
Associate Editor: Brian Wiegmann

Abstract.—Public DNA databases are composed of data from many different taxa, although the taxonomic annotation on sequences is not always complete, which impedes the utilization of mined data for species-level applications. There is much ongoing work on species identification and delineation based on the molecular data itself, although applying species clustering to whole databases requires consolidation of results from numerous undefined gene regions, and introduces significant obstacles in data organization and computational load. In the current paper, we demonstrate an approach for species delineation of a sequence database. All DNA sequences for the insects were obtained and processed. After filtration of duplicated data, delineation of the database into species or molecular operational taxonomic units (MOTUs) followed a three-step process in which (i) the genetic loci L are partitioned, (ii) the species S are delineated within each locus, then (iii) species units are matched across loci to form the matrix $L \times S$, a set of global (multilocus) species units. Partitioning the database into a set of homologous gene fragments was achieved by Markov clustering using edge weights calculated from the amount of overlap between pairs of sequences, then delineation of species units and assignment of species names were performed for the set of genes necessary to capture most of the species diversity. The complexity of computing pairwise similarities for species clustering was substantial at the cytochrome oxidase subunit I locus in particular, but made feasible through the development of software that performs pairwise alignments within the taxonomic framework, while accounting for the different ranks at which sequences are labeled with taxonomic information. Over 24 different homologs, the unidentified sequences numbered approximately 194,000, containing 41,525 species IDs (98.7% of all found in the insect database), and were grouped into 59,173 single-locus MOTUs by hierarchical clustering under parameters optimized independently for each locus. Species units from different loci were matched using a multipartite matching algorithm to form multilocus species units with minimal incongruence between loci. After matching, the insect database as represented by these 24 loci was found to be composed of 78,091 species units in total. 38,574 of these units contained only species labeled data, 34,891 contained only unlabeled data, leaving 4,626 units composed both of labeled and unlabeled sequences. In addition to giving estimates of species diversity of sequence repositories, the protocol developed here will facilitate species-level applications of modern-day sequence data sets. In particular, the $L \times S$ matrix represents a post-taxonomic framework that can be used for species-level organization of metagenomic data, and incorporation of these methods into phylogenetic pipelines will yield matrices more representative of species diversity. [Database partitioning; MOTU; multi-locus clustering; species delineation.]

Representing the sum of all sequence data made available to date, public DNA databases such as GenBank are a heterogeneous assemblage with no constraints in terms of representing biological diversity. Sequence data are submitted in which the overlap with other sequences may be high or low, and in which the evolutionary distances to other members of the database are similarly unconstrained. Determining the contents of a sequence database often relies on the annotation given to entries, although in practice this information may be incomplete, vague, or even incorrect (Vilgalys 2003; Nilsson et al. 2006; Wägele et al. 2009). Partial taxonomic labeling of sequences is particularly common in public databases. The increased rate of DNA sequence submission fueled by the continually reducing cost and the adoption of high-throughput technologies places greater demand on identification of samples by taxonomists. In some cases, this may compel improved taxonomic work, but identification to the species level for most insects can be time-consuming. The species is the fundamental biological unit and perhaps also the most valid (nonarbitrary) rank in taxonomy (Mayr 1982). Yet all too often this rank is omitted, with submission of insects sequences with incomplete species labels routine (Althoff 2008; Emery et al. 2009; Smith and Fisher 2009; Pinzon-Navarro et al. 2010; Burks et al. 2011; Pilgrim et al. 2011; Santos et al. 2011; Smith et al. 2012). Alphanumerical labels are often assigned to species fields of sequences in the absence of species-level identification, inconsistently referring either to the specimen or the putative species group (defined via sequence analysis) to which the specimen belongs. Thus, a large class of unidentified molecular data exists on GenBank (termed "dark taxa"; Page 2011), the utility of which would be substantially enhanced where the taxonomic limits and identities can be determined. Post submission classification of data using principles from DNA taxonomy and DNA barcoding is one means in which these sequences may be given taxonomic placement. Such molecular delineation and taxonomic assignment is routinely performed for single-locus data sets (Stackebrandt and Goebel 1994; Floyd et al. 2002; Hebert et al. 2003; Blaxter et al. 2005; O'Brien et al. 2005; Nilsson et al. 2009). For example, using a web service linked to a library of reference data, unidentified cytochrome oxidase subunit I (COI) query sequences differing from fully labeled references by less than approximately 1% can be assigned the species label of the latter (Ratnasingham and Hebert 2007). This process is straightforward in the case of COI barcode sequences since this particular gene region is standardized and widely utilized; an extensive

curated set of library COI sequences is available, with well-characterized evolutionary features (Savolainen et al. 2005; Ratnasingham and Hebert 2007). However, this implies that the assignment of species for any particular sequence is dependent on the representation of homologous gene fragments. The standardization in the case of COI is atypical, as databases are composed of numerous gene regions which overlap to varying degrees. Therefore, the taxonomic delineation of a database cannot be disentangled from the partitioning of the genetic loci on which delineation occurs.

Partitioning the contents of a database according to homology is commonly practiced in evolutionary studies, where two general approaches have emerged; (i) a search of the database using user-specified queries and (ii) grouping of database sequences according to internal criteria. In the former, the database search queries may be user specified (Dereeper et al. 2008; Smith et al. 2009) or from a standardized set of references (e.g., "left path" of the pipeline developed by Peters et al. 2011). The latter uses patterns in sequence similarity and overlap (Enright and Ouzounis 2000; Driskell et al. 2004; McMahon and Sanderson 2006; Sanderson et al. 2008; Thomson and Shaffer 2010) and has a history in the field of protein classification (Sonnhammer and Kahn 1994; Krause and Vingron 1998; Tian and Dickerman 2007; Ebersberger et al. 2009). This method typically consists of an all-against-all (pairwise) comparison of sequences, followed by clustering of overlapping pairs, where heuristics are adopted for larger databases (Tian and Dickerman 2007; Thomson and Shaffer 2010). An ideal partitioning resulting from this would be groups of sequences containing members that were mostly overlapping (i.e., having large regions of homology), and mostly nonoverlapping with members of other groups of sequences.

After the database has been grouped into homologs, the species memberships of sequences *within* each homolog would need to be determined. Species-level clustering of unidentified sequences is perhaps most commonly achieved with phenetic approaches (Hebert et al. 2003; Ratnasingham and Hebert 2013), consisting of computation of similarities (or distances) between sequence pairs followed by the grouping of highly similar sequences. The degree of similarity up to which sequences are grouped is known as the cutoff or threshold, and needs to be set to a level appropriate for the species level. In the case of DNA barcodes, the threshold is set at 97.8% similarity (Ratnasingham and Hebert 2013), although species clustering is not confined to organellar protein-coding genes, for example nuclear 28S rRNA sequences grouped where identical, have been found to correspond to presumed species groups in beetles (Monaghan et al. 2005). As the rate of evolution is known to vary across the genome (Roe and Sperling 2007), species clustering parameters require customization. An intuitive approach to deriving an optimal species-level cutoff is testing a range of reasonable values in reference data and adopting the one in which the resulting clusters most closely resemble established taxonomic species (Göker et al. 2009; Hibbett et al. 2011; Mende et al. 2013). The threshold maximizing congruence can be selected as the optimal, and applied to the unidentified sequence class. When applied to the unidentified sequences, the resulting clusters are a proxy for estimating species diversity.

As routine as sequence-based species clustering has become, there is little work on the practicality of consolidating clusters from multiple loci, where forming molecular operational taxonomic units (MOTUs) from a locus-partitioned database requires consolidating results among very many loci, in which incongruence is inevitable. In related settings, resolving conflict between genes usually involves extracting common signal, for example; selecting the predominant tree from single gene-tree distributions (Ané et al. 2007); inferring the species tree which minimizes both intraspecific structure and conflict between trees (O'Meara 2010); forming the most similar single-gene clusters by modification of their linkage parameters (Setaro et al. 2012); searching for a species tree that maximizes likelihood of the sequence data (Kubatko et al. 2009) or posterior probability over a distribution of trees (Liu and Pearl 2007; Heled and Drummond 2010) under models that incorporate multiple species and genes. In the context of single-gene MOTUs on a locus-partitioned database, consolidating results can be achieved in a graph framework, treating loci as partitions and species units as nodes. Species units may then be linked with minimal conflict using graph matching algorithms (Chesters and Vogler 2013). By maintaining loci as distinct but linked partitions, this allows formation of a two-dimensional matrix ($L \times S$) in which columns correspond to loci and rows to the delineated species units. This form of matrix is useful in many applications that use mined sequences, particularly the supermatrix approach to phylogenetic analysis, which is a widely adopted method used for building trees from public data (Driskell et al. 2004; McMahon and Sanderson 2006; Goloboff et al. 2009; Thomson and Shaffer 2010; Jones et al. 2011a; Peters et al. 2011).

In the current study, we perform a species delineation of the insect DNA database. The insects are selected as a case study for the protocol, being both hyper-diverse and well represented on sequence databases, while still posing significant challenges for taxonomy. In order to reconstruct a set of putative species units over the set of genetic data present, we perform homolog partitioning optimized for the purpose of species-level clustering. The loci are then consolidated into a single matrix, this gives an estimate of the species diversity and an analyzable species-level matrix in which the impediment of incomplete labeling has been addressed.

Materials and Methods

Protocol Overview

Figure 1 illustrates the three-step process whereby a sequence database of unknown composition (Fig. 1a)

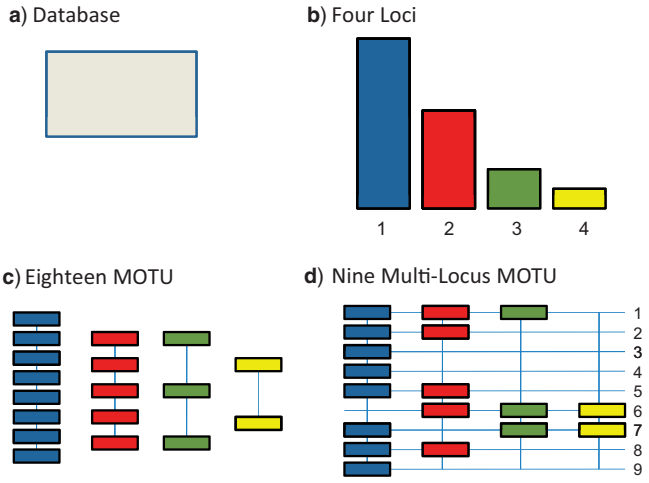

FIGURE 1. Illustration of database delineation. a) A database is delineated with no prior assumptions on composition, sampling patterns, or delineation parameters, thus the contents of the database are initially unknown. b) The database is partitioned into loci by identifying then obtaining the predominantly used fragments. Four loci are shown, with the size of each bar reflecting the amount of data present for the locus. Typically one predominant gene is found, as depicted by the first bar. c) Next each locus is delineated into species units, which consists of sequences being grouped according to species name, and sequences clustered according to similarity where unidentified. d) Finally multilocus species units are formed by linking together those from different loci, resulting in the delineation matrix $L \times S$ (where $L = 4$ and $S = 9$).

can be organized into an $L \times S$ matrix. The steps are the primary partitioning of the database into loci (Fig. 1b), followed by species clustering performed separately on each of the partitioned loci (Fig. 1c) and then integration of single-locus species units to create the final delineation matrix (Fig. 1d). Each row of the $L \times S$ matrix corresponds to a named species or global MOTU (see Table 1 for definitions) of unidentified sequences. The protocol is implemented as a series of steps amenable to automation, where the flowchart in Figure 2 gives the main steps required.

The Insect Database

The Invertebrate flat file release (as of March 2013) was downloaded from the GenBank ftp site (ftp://ftp.ncbi.nih.gov/genbank/) and the taxonomy database (taxdump.tar.gz) from ftp://ftp.ncbi.nih.gov/pub/taxonomy/. The GenBank flatfile database was parsed for specific fields (accession, NCBI taxon ID, gene name, and DNA sequence). The NCBI taxonomic hierarchy was used in assigning species labels (both Linnaean binomials where identified and alphanumerical labels where unidentified) to each sequence, via the NCBI taxon ID (the "taxon" type in the "db_xref" field). Where subspecies names were used as species IDs, we assigned the containing species name (scientific name), ignoring synonyms. All sequences with a taxon identifier downstream of the Insecta node (NCBI taxonomy ID: 50557) were selected from the invertebrate division, and then dereplicated using Usearch (Edgar 2010), in which sequences identical across their whole length (command line option "-derep_fullseq") were removed (except if the identical sequences were labeled as different species) to form a reduced redundancy database. This reduced redundancy insect database was used in all further analyses, and is made available on Dryad (http://dx.doi.org/10.5061/dryad.k7t50, filename insecta.fas, see Supplementary Material online, available from http://www.sysbio.oxfordjournals.org).

TABLE 1. Definitions

(Bi-)Partite matching	Linking members between two different groups
Cardinality	Count of the number of links, e.g., in a graph
Clustering optimization	Derivation of the best parameters for clustering, under some criteria
Global MOTU	Sequence groups broadly representing species after consolidating information from multiple loci, where any one global MOTU may have one or more locus represented
HA Rand index	Measure of the similarity between two alternative groupings
Hit span fraction	A measure of the proportion of the length of two sequences which overlap
Homology	Evolutionary relatedness, here assumed where sequences share similarity over much of their length
Markov clustering	Heuristic approach to grouping, here applied to measures of sequence overlap

Partitioning Gene Fragments

Gene fragments were first partitioned by clustering sequences according to degree of overlap. Overlap was determined according to local alignments between the complete database and a random subset thereof (Figs. 3 and 2(step 1)). A local Blast database was created from the file of insect sequences using makeblastdb (Camacho et al. 2009), then all sequence pairs between the database and random subset compared (Fig. 2(step 2)) using the BlastN algorithm (Blast+ v. 2.2.28) under the command line settings "-task blastn -dust no -strand both." The dust filter was deactivated (-dust no) to prevent hit fragmentation (Sanderson et al. 2008). There is little consensus on an appropriate *e*-value under such searches, with values spanning over 10 orders of magnitude (e.g., Yona et al. 2000; Sasson et al. 2002; Krause et al. 2005; McMahon and Sanderson 2006; Tian and Dickerman 2007); therefore, we performed a number of search replicates under a randomly selected *e*-value taking a value between 10 and $1e{-}10$. Where alignment is performed to determine homology and overlap only, the degree of sequence similarity is of less concern since both distantly and closely related homologs are sought. The length of the alignment is most relevant, for which we used the hit span fraction (Fig. 2(step 3))

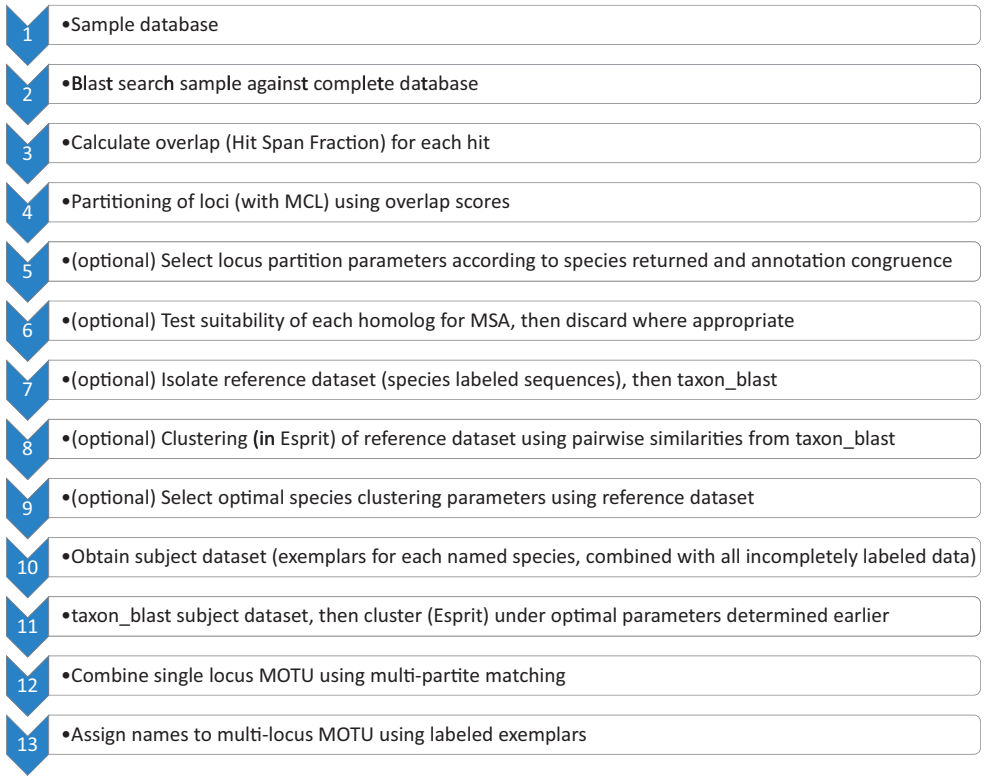

FIGURE 2. The sequence of steps for the protocol developed herein. Step numbers are referred to in the text. Note, not all steps are necessary for each iteration. In particular, locus parameter optimization (step 5), processing, and assessment for multiple sequence alignment (step 6), and species parameter optimizations (steps 7–9) may not be required. Abbreviations; MSA, multiple sequence alignment; MCL, Markov cluster process; MOTU, molecular operational taxonomic unit.

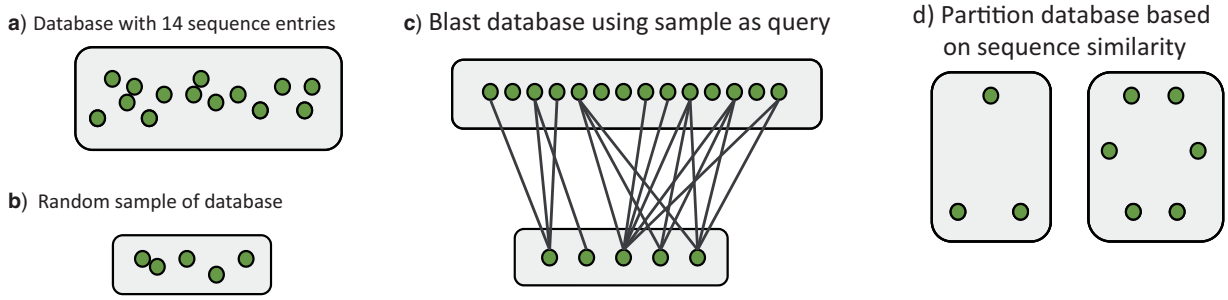

FIGURE 3. Illustrating an approach to partitioning a database by locus. This approach requires no prior assumptions on the composition of the database, being based only on the structure therein. a) the database; small circles denote individual members (DNA sequence entries). b) A random subset of the database is sampled. c) The sample (c lower) are used as alignment queries against the complete database (c upper), the links indicate sequence overlap. d) Partitioning according to similarity gives two loci, one containing three members and the other containing six.

according to McMahon and Sanderson (2006). These pairwise overlap values were then input into MCL (Markov CLuster process; van Dongen 2000) for locus-partitioning (Fig. 2(step 4)). Briefly, this program uses edge weights (here, hit span fraction) as transition probabilities. The probabilities are altered during Markov rounds whereby the stronger relationships become more robust and weak links broken, until a robust partitioning of the data is created. The resulting gene clusters are primarily influenced by the inflation parameter, with tight gene families created with larger values, and larger gene groupings where using smaller values (Krause et al. 2005). For each replicate, we randomly set the inflation as either 1.1, 1.4, 2, 4, 5, or 6. Fragment clustering was repeated where variously selected subsets of the database were used as queries, and with randomly selected values for the key parameters.

For the purpose of creating the $L \times S$ matrix, an optimal locus-partitioning of the database is that in which (i) as much as the species diversity as possible has been integrated and (ii) each gene partition contains

mostly overlapping sequences, with minimal overlap between sequences of different partitions. The former was characterized for each replicate according to the number of species IDs placed into the partitioning (i.e., hit during the Blast search), whereas for the latter we propose the congruence between the Markov clustered partitions and the gene names assigned by sequence submitters (Fig. 2(step 5)). Gene names were first standardized (e.g., the labels CO1, COI, COX1, COXI, and CYTOCHROME OXIDASE I were all given the same identifier) then congruence between gene names and Markov clusters measured. A standard score of congruence is the Rand index, and several derived measures (Rand 1971; Milligan and Cooper 1986; Warrens 2008). These give an indication of the similarity between clusters made under two different settings, based on pairs of individuals in the data set which are (i) clustered in one setting and clustered in the second, (ii) clustered in one setting and not clustered in the second, (iii) not clustered in the first setting but clustered in the second, (iv) not clustered in either setting, where in the current case the two settings are (a) the groupings based on sequence similarity (i.e., MOTU) and (b) the groupings based on gene labels (usually morphospecies). We use the Hubert and Arabie-adjusted Rand index (HA Rand index; Hubert and Arabie 1985), as calculated in the Clues R package (R Development Core Team 2008; Chang et al. 2010).

Grouping of sequences in this way is no guarantee that sets can be aligned globally, which would be required where matrices are used for phylogenetic inference. Therefore, homologs were further assessed for this purpose. For this step (Fig. 2(step 6)), sequence orientation (such that the equivalent strand is maintained throughout) is both the primary processing requirement and means of assessment. Peters et al. (2011) propose orientation in initial reference to user-supplied representatives for each locus. We here compare this approach to a fully automated one, in which a "most representative sequence" (MRS) is sought, then acting as a seed for sequence orientation of the remainder. For selection of the MRS, we use a method broadly similar to the routine used in BlastAlign (Belshaw and Katzourakis 2005), except that we select the sampled member with maximal aligned length when compared against others (as opposed to maximizing the presence of landmarks). Next sequences for each homolog were oriented generally following Peters et al. (2011). A Blast search was performed of the seed sequence against other members of the gene, then the strand of hits used to orient where necessary. Sequences discarded in the first round (due to lack of sequence similarity) were subject to a second Blast round, against four randomly selected sequences successfully aligned against the original seed. Finally, oriented sequences were aligned with Clustal Omega (Sievers et al. 2011) under default parameters, and assessed visually for suitability for multiple sequence alignment.

Defining Taxonomic Units

The unidentified sequences were delineated into MOTUs initially on a locus by locus basis. In addition to indicating the otherwise unknown species diversity in the unidentified sequences, this permitted the assignment of species names to unidentified sequences in many cases. Species clustering parameters were obtained by first grouping sequences that had been labeled to species level (reference data), then selecting parameters in which the congruence between molecular clustering and taxonomic species was greatest. Sequences of the reference data set were grouped according to percent identities as calculated after pairwise Blast alignments under the command line settings "-word_size 20 -perc_ident 95 -evalue 1e-10" (Fig. 2(step 7)). After genetic distances were obtained, agglomerative single linkage clustering was performed using the "hcluster" algorithm as implemented in Esprit (Sun et al. 2009) (Fig. 2(step 8)). Clusters were generated under thresholds from 100 to 95, in steps of 0.1. Finally, the congruence between clustering of the reference data and their taxonomic labels (species) was then scored. As before, we test congruence using the HA Rand index (Fig. 2(step 9)), where in this case a clustering parameter producing the maximal HA Rand index indicates the one most likely to return sequence groups corresponding to established species.

Sequences not identified to species level (the subject data) were then clustered under the parameters deemed optimal by the HA Rand index. Alignment between all pairs of the larger partitions (particularly, COI) is both unfeasible and unnecessary; therefore, we carry out alignments within taxonomic groups above the species level. We developed a software tool (taxon_blast.pl) for pairwise alignments within the taxonomic framework of both fully and partially identified sequences. By default the script uses the genus, family, and order levels found to be most relevant in the current study. The script first identifies all sequences that have genus-level labeling, then all-against-all alignments are performed within each of these genera. As shown in Figure 4, many sequences are present in which taxonomic labeling is only given for ranks above genus; for example, those in which the species and genus label are both absent, but family label is given. In these cases, two sets of alignments *within each given family* need to be carried out: (i) all-against-all alignment of the sequences lacking genus-/species-level annotation and (ii) all-against-all alignment between the sequences in (i) and the remaining sequences that *do* have genus/species taxonomic annotation. This procedure is repeated at the order level due primarily to the large number of BOLD data only labeled to that rank (Fig. 4). Calculation of sequence similarities in the subject data with taxon_blast.pl was followed by hierarchical clustering with Esprit (Fig. 2(step 11)), as described previously.

MOTUs were by necessity clustered separately for each gene fragment, thus deriving an integrated

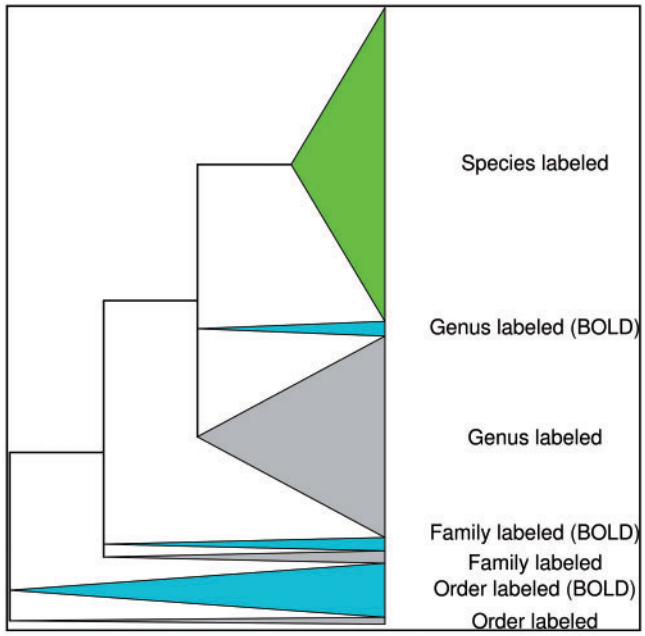

FIGURE 4. Tree structure depicting taxonomic rank and labeling for sequence data on GenBank. Splits correspond to labeling class (not Linnean groups). Labeling classes include taxonomic identifiers complete to the level of species, incomplete to the level of genus, family, and order level, and whether these are BOLD or non-BOLD derived. Bases are sized relative to the number of labels in the given class.

delineation of the database and a total estimate of species diversity requires the consolidation of results from the different loci (Fig. 1d). Where two MOTU from adjacent loci share a label (species or alphanumerical) they can be regarded as a single species unit, and their sequence data united as representing genomic data from that one species. This process relies on the matching of labels of which different conventions are used for unidentified sequences. Thus, cases are expected where MOTUs actually from the same species are not united due different labels. Creating global MOTUs by combining those separately delineated from different loci is not straightforward, since many of the latter are composed of multiple species IDs. In such cases, there is more than one way in which the MOTUs can be matched to those at adjacent loci, and thus a large space of possible configurations exists over the whole database. In order to generate the most reasonable global MOTUs, we performed maximal cardinality multipartite matching (Chesters and Vogler 2013) (Fig. 2(step 12)). The optimal set of matches is one in which the number of links between loci is maximized, in other words, the optimal $L \times S$ matrix is that in which the least number of single-locus MOTUs remain unlinked. The problem of maximal matching where more than two partitions (loci) is present, is NP-complete, thus a heuristic is used. The single hub heuristic splits the problem into a series of bipartite matchings, a graph problem which is solvable in polynomal time (Papadimitriou and Steiglitz 1982). For example, where combining MOTUs from three gene fragments, GeneA, GeneB, and GeneC; MOTUs from GeneA and GeneB are first matched by maximal cardinality bipartite matching to form the set of MOTUs GeneA–GeneB; then this MOTU set (GeneA–GeneB) is then matched by the bipartite algorithm to GeneC.

Assessment of the proposed protocol.—Species-level clustering of unlabeled data relies on parameter optimization using sequences with associated species labels (reference data); however, it is well known that mislabeling is prevalent in public databases, which is expected to impact the accuracy of clustering. In order to estimate the level of error in species clustering, a key set of 26 "model genera" is omitted from the reference data set during optimization of clustering parameters, and used later for measurement of clustering error. This is under the assumption that labeling accuracy would be greatest for these highly studied taxa. The 26 model genera used were: *Acyrthosiphon*, *Aedes*, *Anopheles*, *Apis*, *Bombyx*, *Culex*, *Dendroctonus*, *Drosophila*, *Galapaganus*, *Gryllus*, *Heliconius*, *Helicoverpa*, *Lucilia*, *Manduca*, *Melanoplus*, *Myrmica*, *Nasonia*, *Ostrinia*, *Papilio*, *Pediculus*, *Pheidole*, *Rhagoletis*, *Schistocerca*, *Spodoptera*, *Triatoma*, and *Tribolium*.

Next, we examined the sensitivity of species clustering to clade-specific parameter estimation (Huang et al. 2008; Meier et al. 2008). Since the rate in substitution may undergo clade-specific shifts, it might be assumed that clustering parameters are better assessed individually for groups. Parameter optimization and application of the optimal parameter to subject data are carried out as described earlier, but individually for each family.

Finally, we assessed an alternative gene partitioning approach for the purpose of species clustering. There have been several such approaches previously used, of which the one developed here requires minimal user decisions and input. Another tractable approach is simply placing sequence entries into separate files according to their gene labels. We assess this alternative approach using a modified version of "multiple_sequence_splitter" (Peters et al. 2011). The script retrieves the full GenBank entry for each sequence and splits the sequence by each "feature" (gene label), although we modify the original script in order to standardize gene names in a way similar to that described earlier (scoring genetic congruence in the section *Partitioning Gene Fragments*), which prevents the splitting of homologs which are alternatively named.

RESULTS

Insect Database

The invertebrate release was downloaded from GenBank, and all insect sequences selected. After dereplication, the insect database stood at 731,090 sequences. The database contained 382,363 sequences with a complete binomial species label, leaving 348,727 labeled with an alphanumerical identifier. In total, the database contained 43,465 binomials, and

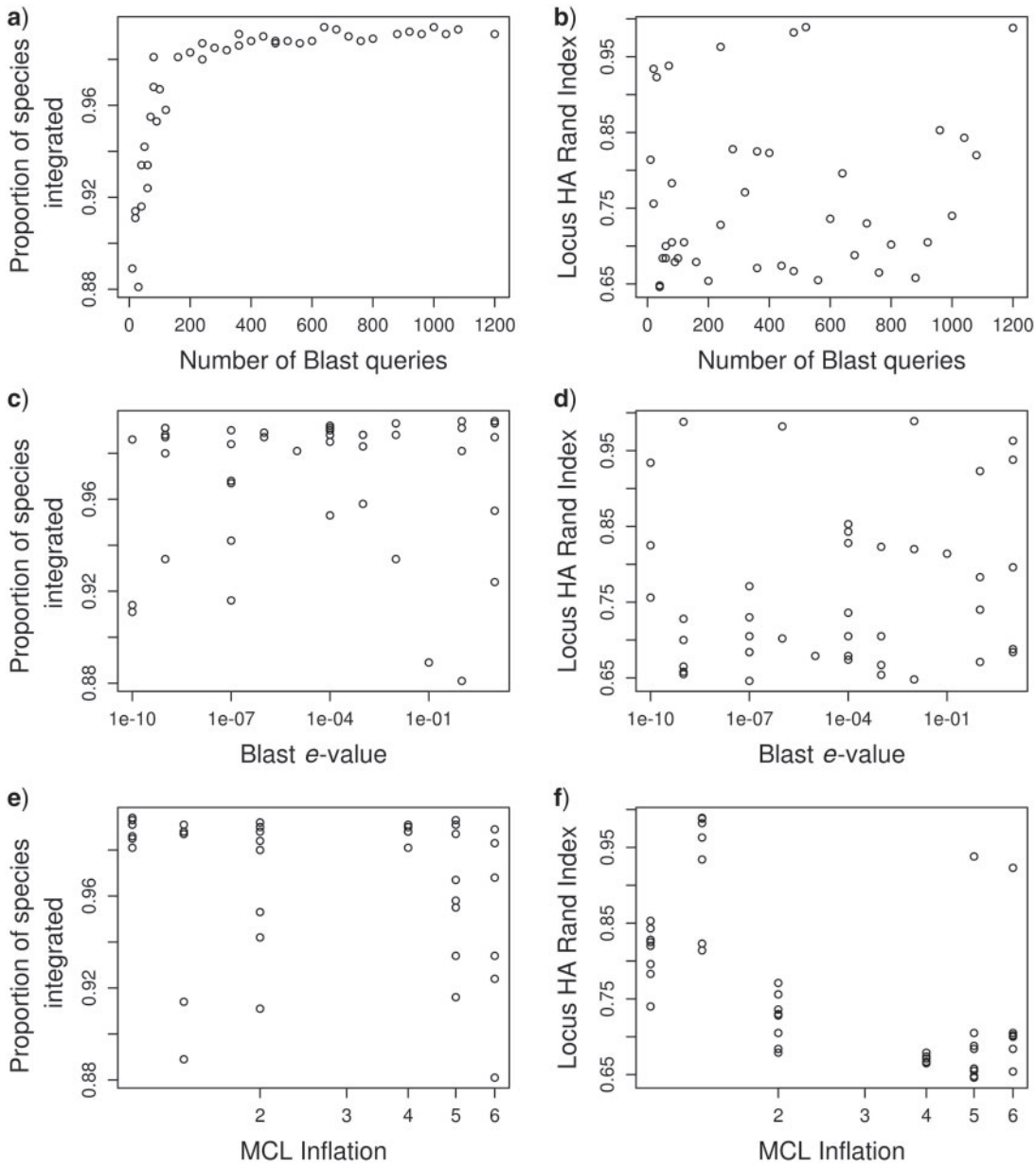

FIGURE 5. Results of locus-partitioning replicates under varying parameters. (a, c and e) give the proportion of species IDs in the insect database that are incorporated into the partition of homologs, whereas (b, d and f) give the locus HA Rand index, which is a score of congruence between the locus-partitioning and the gene names assigned to sequences. Three different parameters are assessed for their impact on these variables; (a and b) the number of sequences sampled as Blast queries for the complete insect database; (c and d) the *e*-value used for that Blast search, and (e and f) the MCL inflation used to group members into homologs.

alphanumerical species-level labels with taxonomic information to the level of genus (29,952), tribe (109), family (3557), or order (8449). The proportions of the main labeling classes are illustrated in Figure 4. The majority of entries labeled only to order level are BOLD submissions whereas most labeled up to genus level only were non-BOLD sequences. The former are largely interim sequence data releases from BOLD in which more complete taxonomic labeling is to be made available at a later stage, whereas it is unlikely that more complete taxonomic information will be given for most of the latter.

Fragment Clustering

The insect database was partitioned accorded to locus. Figure 5 gives the two optimality criteria (integration of species diversity and congruence between gene clusters and gene annotation) according to variation in key parameters. There was no noticeable impact of the choice of *e*-value (Fig. 5c,d) on the characteristics of the resulting clusters, whereas the proportion of species diversity hit in the Blast search was determined by the number of queries (Fig. 5a), and low Markov inflation values (1.1, 1.4) produced gene clusters which corresponded better

to the gene names assigned to the sequences (Fig. 5f). Capturing most of the species diversity of the database was achieved using a modest number of sampled queries. For example, a Blast search with a random sample of just 80 sequences and a relatively stringent *e*-value (1*e*−07) hit over 95% of all species IDs (82,748) including 95% of unidentified IDs (40,724/42,057), and just over half of the database in total (425,299/731,090). There were rapidly diminishing returns in terms of hitting more species by using a greater number of queries, for example, only an extra 208 species are found when doubling the number of queries from 600 to 1200. Further, although the database in total contained many tens of thousands of species, the majority of these could be found by consulting just a handful of the many gene partitions; the loci preferentially used in biodiversity studies. This is illustrated in Figure 6, which shows the example of the database partitioning under the settings: 520 queries, Blast *e*-value 1e−02, and MCL inflation 1.4. Twenty-four gene clusters are shown, ranked according to the number of sequences comprising each. As expected, most species IDs (63,933) are represented at the COI locus. The sampling for this gene is so marked that, whereas the 28S cluster contains 21,449 species IDs, only half of these (10,020) are not already present for COI. Figure 6b illustrates this curve. Adding more than 10 loci adds few additional species; for example, while the gene cluster composed primarily of the CAD gene contains 4225 species, only 132 of these are novel. The replicate described here was that with the greatest genetic HA Rand index (0.989), and so was selected for further species-level analyses. This replicate hits 68.3% of the database (499,471), but included 98.8% of the species labels contained in the database (84,480) and 99.0% of the unidentified labels (41,621). Statistics for the highest ranking 24 of the 162 gene clusters for this partitioning are given in Table 2.

An optional step (Fig. 2(step 6)) was implemented to assist where matrices are to be used for phylogenetics, in which genes globally unalignable are discarded. Each gene cluster was assessed for lack of similarity between its members. Peters et al. (2011) propose comparing members against a user-supplied representative, which we assess in addition to an automated algorithm using the MRS similar to that used in BlastAlign (Belshaw and Katzourakis 2005). Sequences were discarded where they lacked similarity to the reference. Where using a single reference, 34.6% of sequences were discarded where aligning the set against a representative sequence randomly selected from the model taxa, as opposed to 13.1% where aligning against the MRS. However, an iterative approach (members successfully oriented in an initial round are used to attempt aligning and orienting the discarded members) led to fewer sequences discarded both using the user model sequence or the MRS. Here, 11.5% of sequences are discarded where the initial reference is user supplied, with 13.1% when using the automated MRS. After orientation, the homologs were aligned and inspected manually. Based on the inspections, the orientation process was found to reflect

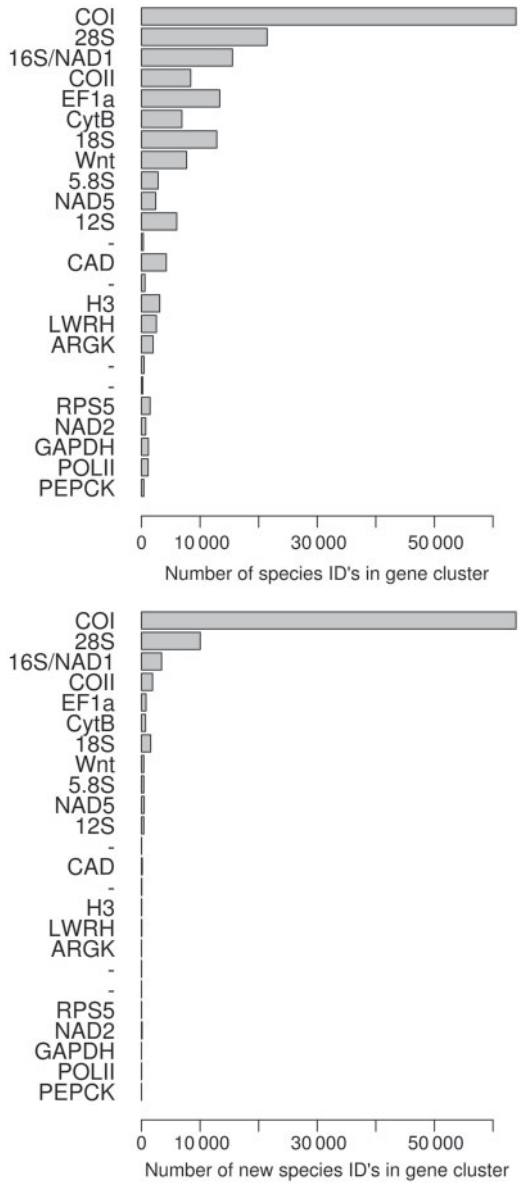

FIGURE 6. Counts of species IDs for the 24 homologs under analysis. Upper bar chart gives the number of species IDs, and lower gives the number of species IDs new to that locus. Predominant locus names are given, except where ambiguous (denoted "-").

the quality of the alignments. In homologs in which a large proportion of its members were discarded (lacking similarity to the reference), alignments of (only) the remaining sequences were of a poor quality. Based on inspection of alignments, were the matrix to be used for phylogenetics, four of the clusters (clusters 12, 14, 18, and 19) would certainly not be subject to further analysis. Notably, it was these clusters for which gene name assignment was ambiguous (containing a large number of sequences in which the gene name differed). Meanwhile, at the current scale, it was apparent that reliable global alignment of the two rRNA loci 28S and 12S would be challenging. Strategies for

TABLE 2. Counts for sequences and species for each gene partition

Locus	Sequences	Species IDs				
		Total in locus	Culminative over loci	Proportion of insect db	Proportion of this partitioning	New to locus
COI	272,167	63,933	63,933	0.747	0.756	63,933
28S	31,117	21,449	73,919	0.864	0.875	10,020
16S/NAD1	25,091	15,524	77,344	0.904	0.916	3425
COII	22,347	8394	79,220	0.926	0.938	1876
EF1a	21,761	13,352	79,987	0.935	0.947	767
Cytb	16,147	6885	80,671	0.943	0.955	684
18S	16,013	12,853	82,231	0.962	0.973	1560
Wnt	10,779	7697	82,618	0.966	0.978	387
5.8S	9826	2836	83,005	0.971	0.983	387
NAD5	9722	2423	83,446	0.976	0.988	441
12S	8591	6011	83,833	0.980	0.992	387
NA	8372	307	83,853	0.980	0.993	20
CAD	5391	4225	83,985	0.982	0.994	132
NA	4617	586	84,038	0.983	0.995	53
Histone3	3739	3101	84,072	0.983	0.995	34
LWRH	3248	2531	84,120	0.984	0.996	48
ARGK	2704	1977	84,128	0.984	0.996	8
NA	2545	437	84,144	0.984	0.996	16
NA	1740	215	84,148	0.984	0.996	4
RPS5	1686	1456	84,162	0.984	0.996	14
NAD2	1634	716	84,252	0.985	0.997	90
GAPDH	1551	1176	84,252	0.985	0.997	0
POLIII	1416	1125	84,270	0.985	0.998	18
PEPCK	1261	391	84,276	0.985	0.998	6

Note: Loci are ranked according to number of sequences. Predominant gene label is given in locus column where unambiguous, and NA otherwise.

multiple sequence alignment of problematic or saturated sequence sets have been proposed elsewhere (McMahon and Sanderson 2006; Smith et al. 2009; Thomson and Shaffer 2010).

Species Clustering

The diversity represented by the unidentified sequences was estimated by clustering into MOTUs, and species names were assigned where possible. The first step was clustering optimization for 24 loci; Table 3 ("Distance threshold" column) includes optimal thresholds for these fragments, with a noticeable trend for more permissive thresholds at mitochondrial markers compared with nuclear markers, and lower species-level congruence at the slower evolving genes (a correlation between the taxonomic HA Rand index and the optimal species threshold of $r = 0.4085$, $P = 0.066$, Pearson's product-moment correlation). These optimal thresholds were used to cluster the unidentified sequences, for delineation of MOTUs. Named species were also included in the clustering for assignment of species names to MOTUs. Calculation of sequence similarities between members of the COI locus was the rate-limiting step of the whole protocol, with that single locus requiring a running time roughly similar to all other loci combined. Computations were made feasible by performing all-against-all alignments within the taxonomic framework using taxon_blast.pl, which automated alignments within each of 5978 genera (the remaining 6442 genera only contained a single member), 146 families, and 16 orders. The Lepidoptera required most computation, with 21,344 sequences lacking any taxonomic labeling below order level requiring all-against-all alignment, followed by those 21,344 against the otherwise annotated sequences numbering 61,379. Still, taxon_blast.pl reduced the number of required alignments by an order of magnitude; 2.1 billion alignments were carried out for the COI locus, whereas 24.7 billion would have been required were each COI insect sequence aligned with each other. Each set of homologs was separately clustered using the corresponding optimal threshold, and species names were assigned to partially identified sequences where a MOTU contained unidentified sequences and no more than a single-named species. The species diversity of unidentified data potentially would range between two extreme scenarios; all sequences originating from a single species, or each unlabeled sequence from a different species. The results show a structure much closer to the latter; Table 3 gives the species clusterings for individual genes, where the putative species diversity represented by unidentified sequences was substantial. For example, COI was delineated into a total of 54,907 species units. This included 126,241 unidentified sequences clustered into an estimated 26,722 single-locus MOTUs, an average of 4.7 sequences per MOTU.

In order to derive global counts of the various species classes, the MOTUs were matched between

TABLE 3. Species clustering

Locus	Distance threshold	Species units	Labeled species	Unidentified members	MOTU with unidentified members	Species assignments	MOTU with >1 species
COI	0.029	54,907	31,056	126,241	26,722	15,297	1942
28S	0.004	20,928	13,080	11,910	8391	602	814
16S/NAD1	0.013	14,485	10,991	6988	3706	233	601
COII	0.024	7741	5757	6666	2084	167	355
EF1a	0.005	13,010	9875	5065	3297	186	415
Cytb	0.014	6779	4280	5133	2565	98	173
18S	0.000	12,302	8341	5343	4203	249	440
Wnt	0.007	7697	5804	2506	1962	67	212
5.8S	0.021	2465	2005	3242	522	102	174
NAD5	0.025	2356	1435	5858	929	8	85
12S	0.012	5842	4657	1829	1231	50	229
NA	0.025	1000	221	2430	778	0	8
CAD	0.003	4352	3034	1525	1337	24	40
NA	0.05	1048	437	1714	614	11	27
Histone3	0.018	2730	2155	1001	644	87	144
LWRH	0.007	2307	1691	1046	664	43	117
ARGK	0.004	1882	1146	1153	757	35	61
NA	0.05	480	340	662	144	23	23
NA	0.014	561	82	1511	478	0	3
RPS5	0.01	1454	1356	108	102	3	38
NAD2	0.015	674	269	636	410	2	21
GAPDH	0.01	1178	1084	121	95	0	21
POLIII	0.002	1201	791	466	418	8	25
PEPCK	0.008	588	171	841	416	0	10

Note: Predominant gene label is given in locus column where unambiguous, and NA otherwise.

loci as to minimize conflict between species labels. We derived a total of 78,091 species units in the Insecta, where 39,517 were global MOTUs which included unidentified sequences. Showing the results of a single genus, Table 4 illustrates the incongruity that is often encountered when global MOTUs are formed, some of which may be addressed by further parameterization. *Vespula* are represented on NCBI by data from a number of genes, and several species labels, of which four are partial identifications. Three of the sequences with (two different) unidentified labels are unambiguously placed in the matrix; the label TRU-2010 has been assigned to sequences from two different genes, both of which have been determined as unique species-level entities, and thus forming a single multilocus MOTU; and the sequence with the label BOLD:AAG7678 has clustered in a single MOTU with the named species *Vespula flavopilosa*. However, sequences labeled with CSM-2006 (which is this case refers to a voucher specimen) clustered with three named species over the different genes: *V. flavopilosa* for COI, *Vespula maculifrons* for 28S, and *Vespula pensylvanica* for 18S. The matrix in Table 4 is the maximal cardinality of a number of different configurations in which the MOTUs containing CSM-2006 could be matched. Users may consider the matching of alphanumerical IDs or voucher labels more valid links than species names, since this would preferentially concatenate different genes sequenced from the same individual, and reduce formation of chimeras. This can be achieved by application of a weighting regime, in which several options are made available (script multi_locus_MOTU.pl). The full delineation matrix (24 $L \times 78{,}091$ S) is made available, see Supplementary Material online (file name delineation_matrix).

Assessment of the database organization protocol.—The optimal parameters inferred earlier were applied to all species-labeled data for a set of 26 model genera in order to estimate the level of accuracy with which they formed a set of groups representative of species diversity (Supplementary Fig. 1a). There was a strong correspondence between inferred and actual number of species, particularly where thresholds are estimated on highly sampled genes. For example, the homolog with by far the greatest representation of named and unnamed data (COI) differs in the number of species by only 6%. However, 18S rRNA in particular appears less suited for the species clustering method implemented here, as it substantially underestimates the true number of species (estimating 24 when actually 71 were present in the model genera for 18S) despite the very stringent threshold which is inferred and used. 28S rRNA additionally may increase the error in species estimate. Although the deviation was much less marked for 28S, the dense sampling of this gene over the insects gives it greater influence on the final species count. Over all genes, the mean deviation from the presumed true number of species in the model genera was 27.4% (weighted according to the number of species per gene). In applications where the rows of the $L \times S$ matrix are required to reflect species diversity to a higher degree of accuracy, genes can be omitted and the matrix rebuilt under the more suitable loci. For example, the omission

TABLE 4. Section of the delineation matrix, where $L=6$ and $S=5$

COI	28S	16S/NAD1	CytB	18S	Wnt
CSM-2006 / BOLD:AAG7678 / *flavopilosa*	N	N	N	CSM-2006 / *pensylvanica*	N
DIMC068-09 / *intermedia*	N	N	*intermedia*	N	N
TRU-2010	N	N	TRU-2010	N	N
maculifrons	CSM-2006 / *maculifrons*	*maculifrons*	*maculifrons*	*maculifrons*	CSM-2006
N	*pensylvanica*	*pensylvanica*	N	N	N

Notes: The example contains a single genus *Vespula* (five additional species are not shown here). N indicates lack of sequence data for given homologs of a species unit, otherwise all species-level labels (Linnean where italicized, and alphanumerical otherwise) found in the species unit are given. Multiple individuals of a species unit are separated by '/'.

of 18S gives a matrix of 76,467 species units with 34,118 MOTU composed solely of unlabelled data, and an estimated accuracy of 21.7% (weighted mean deviation in the model genera), whereas omission of both 18S and 28S gives 71,116 units and an accuracy of 20.6%.

We next determined how inferred species diversity might be impacted by the range at which clustering parameters are optimized. For simplicity, the pipeline derives species clusters from a single threshold (for each locus) over all insects, although we would like to determine whether species counts have a tendency to deviate from these insect-wide values where thresholds are optimized at more local scales. Thus, optimal clustering parameters were inferred and applied individually for each family in the data set. A strong correspondence was found between the number of MOTUs where general thresholds or family-specific thresholds are used (Pearson's $R=0.854$, Supplementary Fig. 1b), particularly for the more species dense families, although there was a tendency to "overlump" where species clustering parameters are inferred for sparsely sampled families.

Finally, the species-level clustering procedures were performed on a set of primary homologs defined simply by the feature names on sequence entries (Peters et al. 2011). A high number of sequences lacking genetic annotation would be an issue for this, as these could not be assigned to any name-based homolog; however, these did not seem substantial (2589 out of 731,090). A comparative analysis was performed on the name partitioned data set, with the species units clustered on the MCL partitioned homologs. A general pattern differentiating these was a lesser number of sequences in a larger number of partitions where partitioning by gene name. The main genes defined (with number of sequences for name partitioned; MCL partitioned) were COI (157,089; 157,297), 28S (23,569; 24,990), COII (13,758; 12,423), CYTB (9280; 9413), EF1a (10,087; 14,940), 18S (11,779; 13,684), and Wingless (7671; 8310). Additionally, some of the cases where the MCL method produced a larger gene cluster were clearly the result of combining genes with overlapping sequence. These are usually adjacent on a chromosome and commonly sequenced in tandem. For example, where name annotation defined two separate partitions for 16S (14,751) and NAD1 (2514), the MCL partition grouped these together (17,979). A core set of 259,784 entries was identified which overlapped between the approximately 20 most species dense homologs of both methods. Summed over all loci, these core entries were clustered into 134,152 single-gene MOTUs in the MCL partitioned and 139,382 where name partitioned, whereas after multipartite matching the global species units numbered 73,253 (MCL) and 73,994 (name partitioned). These results indicate that while initial differences exist in the primary gene clustering, subsequent estimation of species diversity proceeds similarly.

DISCUSSION

We have developed a framework for species delineation of a database. Some of the steps of this protocol have been used previously for the purpose of organizing mined data for phylogenetic analysis (Driskell et al. 2004; McMahon and Sanderson 2006; Sanderson et al. 2008; Smith et al. 2009; Thomson and Shaffer 2010; Jones et al. 2011a; Peters et al. 2011; Hedtke et al. 2013). Many of these also require the partitioning of such data into $L \times S$ matrices, although none address the problem of defining the S-axis in the presence of unidentified data. Other pipelines are underway based on principles of DNA barcoding and DNA taxonomy, facilitating species-level clustering and annotation, but are designed for use with one (Wu et al. 2008; Jones et al. 2011b) or a small number (CBOL Plant Working Group 2009; Chesters and Vogler 2013) of loci specified *a priori*. There has been no progress in the combined delineation of both axes of the matrix, despite the interdependence of species clustering and fragment partitioning.

The delineation of species according to molecular divergence first requires the specification of the fragments upon which species clustering is performed. In principle, the automated partitioning of fragments allows the data set to "speak for itself" in terms of generating a data set maximally representing the species information content of the database. In practice, it is difficult to attain a fully objective delineation. For example, optimizing gene partitions according to their similarity to gene labeling reduces some arbitrary decisions and parameter selections by the user, but the partitioning is then inclined toward clustering sequences into functional gene units, which might not necessarily correspond to fragments commonly sequenced. This might explain the results here in which all sequences

of COI were placed into a single partition in spite of this gene being represented in the insects primarily by two nonoverlapping fragments (the five prime and three prime segments). Still, the partition optimizations are necessary only initially; the parameters suited for the formation of species dense gene sets as determined here can be applied in further studies.

The ability to delimit major divisions of a sequence repository could facilitate genetic-based approaches to quantifying species diversity. A long standing but fundamental question in biology is the total number of species on earth (Mora et al. 2011); public sequence databases have great potential to shed light on this question. There is an ongoing imperative to sample biodiversity at the molecular level, leading to a massive amount of data of a type which often retains intrinsic information on species boundaries (Barraclough et al. 2003; Acinas et al. 2004; Pons et al. 2006). Thus, information of global biodiversity exists on GenBank irrespective of the thoroughness of species labeling. Further, these data are also expected to contain information on hidden diversity. For example, molecular data suggest that species diversity of host-specific arthropods may be grossly underestimated (Smith et al. 2006; 2008; McBride et al. 2009), as has long been suspected (Erwin 1982). A delineation matrix generated from public sequence data represents a set of hypotheses that can be used for independent assessment of species inventories from traditional observational means, and the associated metadata (e.g., geographic origin, altitude, and interacting species) are invaluable in testing broad-scale hypotheses on patterns in biodiversity (e.g., Baselga et al. 2013).

In addition to giving species diversity estimates, it is evident that this pipeline could assist in the scaling up of supermatrix phylogenetics. For some time, sequences mined from public repositories constituted the bulk of genetic information used in phylogenetic analysis. Adoption of the principle herein yields supermatrices with some advantageous features. The implementation itself removes much of the guesswork and arbitrary decisions often required when selecting homologs, and the matrices formed are more representative both of genetic and species diversity. Supermatrices from which extraneous intraspecific data have been discarded retain phylogenetic information while being more streamlined and reducing computation time (Chesters and Vogler 2013). We demonstrate this can scale considerably, with the matrix formed here perhaps the largest produced (to date, the largest multilocus phylogenetic analysis was carried out by Goloboff et al. (2009), consisting of 73,060 taxa and 13 genes) although we do not address the downstream processes extensively covered elsewhere, such as matrix reduction and tree-searching itself (Sanderson and Driskell 2003). The ease with which the specific tools provided here can be applied in a given phylogenetic context is likely to be dependent on genetic and taxonomic structure. For example, any invertebrate studies selected are likely to show similar patterns in gene use and taxonomic sampling to that observed here, whereas application to plant divisions would require a new set of parameter optimizations for chloroplast markers. Further, the species delineations are dependent on the presence of reference (species labeled) data, which may be sparse for some taxa which have high diversity relative to their morphological distinctiveness.

The abilities in database organization afforded by this study have the potential to inform fields beyond phylogenetic analysis of mined data. Although we expect this pipeline initially to be run anew on a number of primary data sets, it may be valuable to establish a publically accessible database based on the $L \times S$ matrix for querying of new sequence data (manuscript in preparation). This can be viewed as an identification service similar to those which have long existed for single locus data (Maidak et al. 2001; Pruesse et al. 2007; Ratnasingham and Hebert 2007) but that includes the genomic dimension (L) where previously only the species dimension (S) was used. The need for the genomic dimension in species identification will become more pressing as current trends in monitoring of biodiversity are going beyond single-locus sequencing. The term genomic observatory has been coined referring to the need to characterize communities in the genomic era (Davies et al. 2012), and more generally, the integration of amplification free technologies in environmental sampling is generating extensive multigene data sets in which species diversity is unknown, and of which the sequence data are characteristically fragmentary (Taberlet et al. 2012; Zhao et al. 2013). Querying such data against an $L \times S$ matrix would allow them to be set into a species-level framework irrespective of the fragment or species amplified, where queries could then be (i) assigned to an existing row of the reference matrix that contains named species, (ii) assigned to a row only of unlabelled data, and (iii) unassignable to existing rows at the given thresholds. All of these possibilities are informative; (i) permits assignment of a species name to the query, (ii) would not return any species name although would return associated information such as geographic locations of putative conspecifics, and (iii) indicating novel species units. The matrix formed herein therefore represents an initial attempt at building a framework which represents both genetic and species diversity as currently accumulated, which can be utilized for quantifying both of these, in applications using DNA sequence data.

Software Availability

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

Supplementary Material

Supplementary material, including data files and/or online-only appendices, can be found in the Dryad data repository at http://dx.doi.org/10.5061/dryad.k7t50.

Funding

This work was supported mainly by the Knowledge Innovation Program of the Chinese Academy of Sciences [Grant No. KSXC2-EW-B-02]; the National Science Foundation, China [Grants Nos. 30870268, 31172048, and J1210002]; partially supported by the Public Welfare Project from the Ministry of Agriculture, China [Grant No. 201103024]; and the Program of Ministry of Science and Technology of the People's Republic of China [2012FY111100 to C.-D.Z.].

Acknowledgements

This article was immeasurably improved by the review process, for which we would like to thank editors Brian Wiegmann and Frank Anderson, peer reviewer Rod Page, and three anonymous reviewers. The authors would also like to thank Alfried Vogler and Arong Luo for useful comments on early drafts. The bootstrap of the locus-partitioning analysis was performed on a computer cluster at the Institute of Zoology, Beijing; the authors would like to thank Xian-Bing Li for assistance using this system.

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.

Althoff D.M. 2008. A test of host-associated differentiation across the 'parasite continuum' in the tri-trophic interaction among yuccas, bogus yucca moths, and parasitoids. Mol. Ecol. 17:3917–3927.

Ané C., Larget B., Baum D.A., Smith S.D., Rokas A. 2007. Bayesian estimation of concordance among gene trees. Mol. Biol. Evol. 24:412–426.

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

Baselga A., Fujisawa T., Crampton-Platt A., Bergsten J., Foster P.G., Monaghan M.T., Vogler A.P. 2013. Whole-community DNA barcoding reveals a spatio-temporal continuum of biodiversity at species and genetic levels. Nat. Commun. 4:1892.

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

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.

Burks R.A., Heraty J.M., Gebiola M., Hansson C. 2011. Combined molecular and morphological phylogeny of Eulophidae (Hymenoptera: Chalcidoidea), with focus on the subfamily Entedoninae. Cladistics 27:581–605.

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.

CBOL Plant Working Group. 2009. A DNA barcode for land plants. Proc. Natl Acad. Sci. U. S. A. 106:12794–12797.

Chang F., Qiu W., Zamar R.H., Lazarus R., Wang X. 2010. clues: an R package for nonparametric clustering based on local shrinking. J. Stat. Softw. 33:1–16.

Chesters D., Vogler A.P. 2013. Resolving ambiguity of species limits and concatenation in multi-locus sequence data for the construction of phylogenetic supermatrices. Syst. Biol. 62:456–466.

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.

Dereeper A., Guignon V., Blanc G., Audic S., Buffet S., Chevenet F., Dufayard J., Guindon S., Lefort V., Lescot M., Claverie J.M., Gascuel O. 2008. Phylogeny.fr: robust phylogenetic analysis for the non-specialist. Nucleic Acids Res. 36:W465–W469.

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.

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.

Emery V.J., Landry J., Eckert C.G. 2009. Combining DNA barcoding and morphological analysis to identify specialist floral parasites (Lepidoptera: Coleophoridae: Momphinae: Mompha). Mol. Ecol. Resour. 9:217–223.

Enright A.J., Ouzounis C.A. 2000. GeneRAGE: a robust algorithm for sequence clustering and domain detection. Bioinformatics 16:451–457.

Erwin T.L. 1982. Tropical forests: their richness in Coleoptera and other arthropod species. Coleopterists Bull. 36:74–75.

Floyd R., Abebe E., Papert A., Blaxter M. 2002. Molecular barcodes for soil nematode identification. Mol. Ecol. 11:839–850.

Göker M., Garcia-Blazquez G., Voglmayr H., Telleria M.T., Martin M.P. 2009. Molecular taxonomy of phytopathogenic fungi: a case study in Peronospora. PLoS One 4:e6319.

Goloboff P.A., Catalano S.A., Mirande J.M., Szumik C.A., Arias J.S., Kallersjo M., Farris J.S. 2009. Phylogenetic analysis of 73 060 taxa corroborates major eukaryotic groups. Cladistics 25:211–230.

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 Lond B 270:S596–S599.

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.

Heled J., Drummond A.J. 2010. Bayesian inference of species trees from multilocus data. Mol. Biol. Evol. 27:570–580.

Hibbett D.S., Ohman A., Glotzer D., Nuhn M., Kirk P., Nilsson R.H. 2011. Progress in molecular and morphological taxon discovery in fungi and options for formal classification of environmental sequences. Fungal Biol. Rev. 25:38–47.

Huang D., Meier R., Todd P.A., Chou L.M. 2008. Slow mitochondrial COI sequence evolution at the base of the metazoan tree and its implications for DNA barcoding. J. Mol. Evol. 66:167–174.

Hubert L., Arabie P. 1985. Comparing partitions. J. Classif. 2:193–218.

Jones M.O., Koutsovoulos G.D., Blaxter M.L. 2011a. iPhy: an integrated phylogenetic workbench for supermatrix analyses. BMC Bioinformatics 12:30.

Jones M.O., Ghoorah A., Blaxter M.L. 2011b. jMOTU and taxonerator: turning DNA barcode sequences into annotated operational taxonomic units. PLoS One 6:e19259.

Krause A., Vingron M. 1998. A set-theoretic approach to database searching and clustering. Bioinformatics 14:430–438.

Krause A., Stoye J., Vingron M. 2005. Large scale hierarchical clustering of protein sequences. BMC Bioinformatics 6:15.

Kubatko L.S., Carstens B.C., Knowles L.L. 2009. STEM: species tree estimation using maximum likelihood for gene trees under coalescence. Bioinformatics 25:971–973.

Liu L., Pearl D.K. 2007. Species trees from gene trees: reconstructing Bayesian posterior distributions of a species phylogeny using estimated gene tree distributions. Syst. Biol. 56:504–514.

Maidak B.L., Cole J.R., Lilburn T.G., Parker C.T. Jr., Saxman P.R., Farris R.J., Garrity G.M., Olsen G.J., Schmidt T.M., Tiedje J.M. 2001. The RDP-II (ribosomal database project). Nucleic Acids Res. 29:173–174.

Mayr E. 1982. The growth of biological thought. Cambridge (MA): Harvard University Press.

McBride C.S., Van Velzen R., Larsen T.B. 2009. Allopatric origin of cryptic butterfly species that were discovered feeding on distinct host plants in sympatry. Mol. Ecol. 18:3639–3651.

McMahon M.M., Sanderson M.J. 2006. Phylogenetic supermatrix analysis of GenBank sequences from 2228 Papilionoid legumes. Syst. Biol. 55:818–836.

Meier R., Zhang G., Ali F. 2008. The use of mean instead of smallest interspecific distances exaggerates the size of the "barcoding gap" and leads to misidentification. Syst. Biol. 57:809–813.
Mende D.R., Sunagawa S., Zeller G., Bork P. 2013. Accurate and universal delineation of prokaryotic species. Nat. Methods 10:881–884.
Milligan G.W., Cooper M.C. 1986. A study of the comparability of external criteria for hierarchical cluster analysis. Multivar. Behav. Res. 21:441–458.
Monaghan M.T., Balke M., Gregory T.R., Vogler A.P. 2005. DNA-based species delineation in tropical beetles using mitochondrial and nuclear markers. Phil. Trans. R. Soc. B 360:1925–1933.
Mora C., Tittensor D.P., Adl S., Simpson A.G.B., Worm B. 2011. How many species are there on earth and in the ocean? PLoS Biol. 9:e1001127.
Nilsson R.H., Ryberg M., Kristiansson E., Abarenkov K., Larsson K.H., Koljalg U. 2006. Taxonomic reliability of DNA sequences in public sequence databases: a fungal perspective. PLoS One 1:e59.
Nilsson R.H., Ryberg M., Abarenkov K., Sjokvist E., Kristiansson E. 2009. The ITS region as a target for characterization of fungal communities using emerging sequencing technologies. FEMS Microbiol. Lett. 296:97–101.
O'Brien H.E., Parrent J.L., Jackson J.A., Moncalvo J.M., Vilgalys R. 2005. Fungal community analysis by large-scale sequencing of environmental samples. Appl. Environ. Microbiol. 71:5544–5550.
O'Meara B.C. 2010. New heuristic methods for joint species delimitation and species tree inference. Syst. Biol. 59:59–73.
Page R.D.M. 2011. Dark taxa: GenBank in a post-taxonomic world. Available from: URL http://iphylo.blogspot.co.uk/2011/04/dark-taxa-genbank-in-post-taxonomic.html (last accessed June 25, 2014).
Papadimitriou C.H., Steiglitz K. 1982. Combinatorial optimization: algorithms and complexity. Englewood Cliffs (NJ): Prentice-Hall.
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.
Pilgrim E.M., Jackson S.A., Swenson S., Turcsanyi I., Friedman E., Weigt L., Bagley M.J. 2011. Incorporation of DNA barcoding into a large-scale biomonitoring program: opportunities and pitfalls. J. N. Am. Benthol. Soc. 30:217–231.
Pinzon-Navarro S., Barrios H., Murria C., Lyal C.H.C., Vogler A.P. 2010. DNA-based taxonomy of larval stages reveals huge unknown species diversity in neotropical seed weevils (genus Conotrachelus): relevance to evolutionary ecology. Mol. Phylogenet. Evol. 56:281–293.
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.
Pruesse E., Quast C., Knittel K., Fuchs B.M., Ludwig W., Peplies J., Glockner F.O. 2007. SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. Nucleic Acids Res. 35:7188–7196.
R Development Core Team. 2008. R: a language and environment for statistical computing [Computer software and manual]. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0. Available from: URL http://www.R-project.org (last accessed June 25, 2014).
Rand W.M. 1971. Objective criteria for the evaluation of clustering methods. J. Am. Stat. Assoc. 66:846–850.
Ratnasingham S., Hebert P.D.N. 2007. BOLD: the barcode of life data system (http://www.barcodinglife.org). Mol. Ecol. Notes 7:355–364.
Ratnasingham S., Hebert P.D.N. 2013. A DNA-based registry for all animal species: the barcode index number (BIN) system. PLoS One 8:e66213.
Roe A.D., Sperling F.A. 2007. Patterns of evolution of mitochondrial cytochrome c oxidase I and II DNA and implications for DNA barcoding. Mol. Phylogenet. Evol. 44:325–345.
Sanderson M.J., Driskell A.C. 2003. The challenge of constructing large phylogenetic trees. Trends Plant Sci. 8:374–379.
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.
Santos A.M.C., Besnard G., Quicke D.L.J. 2011. Applying DNA barcoding for the study of geographical variation in host–parasitoid interactions. Mol. Ecol. Resour. 11:46–59.
Sasson O., Linial N., Linial M. 2002. The metric space of proteins—comparative study of clustering algorithms. Bioinformatics 18:S14–S21.
Savolainen V., Cowan R.S., Vogler A.P., Roderick G.K., Lane R. 2005. Towards writing the encyclopedia of life: an introduction to DNA barcoding. Phil. Trans. R. Soc. Lond. B 360:1805–1811.
Setaro S.D., Garnica S., Herrera P.I., Suarez J.P., Göker M. 2012. A clustering optimization strategy to estimate species richness of Sebacinales in the tropical Andes based on molecular sequences from distinct DNA regions. Biodivers. Conserv. 21:2269–2285.
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.
Smith D., Janzen D., Hallwachs W., Smith A. 2012. Hyperparasitoid wasps (Hymenoptera, Trigonalidae) reared from dry forest and rain forest caterpillars of Area de Conservación Guanacaste, Costa Rica. J. Hymenopt. Res. 29:119–144.
Smith M.A., Fisher B.L. 2009. Invasions, DNA barcodes, and rapid biodiversity assessment using ants of Mauritius. Front. Zool. 6:31.
Smith M.A., Woodley N.E., Janzen D.H., Hallwachs W., Hebert P.D.N. 2006. DNA barcodes reveal cryptic host-specificity within the presumed polyphagous members of a genus of parasitoid flies (Diptera: Tachinidae). Proc. Natl Acad. Sci. U. S. A. 103:3657–3662.
Smith M.A., Rodriguez J.J., Whitfield J.B., Deans A.R., Janzen D.H., Hallwachs W., Hebert P.D. 2008. Extreme diversity of tropical parasitoid wasps exposed by iterative integration of natural history, DNA barcoding, morphology, and collections. Proc. Natl Acad. Sci. U. S. A. 105:12359–12364.
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.
Sonnhammer E.L.L., Kahn D. 1994. Modular arrangement of proteins as inferred from analysis of homology. Protein Sci. 3:482–492.
Stackebrandt E., Goebel B.M. 1994. Taxonomic note: a place for DNA–DNA reassociation and 16S rRNA sequence analysis in the present species definition in bacteriology. Int. J. Syst. Evol. Microbiol. 44:846–849.
Sun Y., Cai Y., Liu L., Yu F., Farrell M.L., McKendree W., Farmerie W. 2009. ESPRIT: estimating species richness using large collections of 16S rRNA shotgun sequences. Nucleic Acids Res. 37:e76.
Taberlet P., Coissac E., Pompanon F., Brochmann C., Willerslev E. 2012. Towards next-generation biodiversity assessment using DNA metabarcoding. Mol. Ecol. 21:2045–2050.
Thomson R.C., Shaffer H.B. 2010. Rapid progress on the vertebrate tree of life. BMC Biol. 8:19.
Tian Y., Dickerman A.W. 2007. GeneTrees: a phylogenomics resource for prokaryotes. Nucleic Acids Res. 35:D328–D331.
van Dongen S. 2000. Graph clustering by flow simulation [PhD thesis]. University of Utrecht. Available from: URL http://www.library.uu.nl/digiarchief/dip/diss/1895620/full.pdf (last accessed June 25, 2014).
Vilgalys R. 2003. Taxonomic misidentification in public DNA databases. New Phytol. 160:4–5.
Wägele J.W., Letsch H., Klussmann-Kolb A., Mayer C., Misof B., Wägele H. 2009. Phylogenetic support values are not necessarily informative: the case of the Serialia hypothesis (a mollusk phylogeny). Front. Zool. 6:12.
Warrens M.J. 2008. On the equivalence of Cohen's kappa and the Hubert–Arabie adjusted Rand index. J. Classif. 25:177–183.
Wu D., Hartman A., Ward N., Eisen J.A. 2008. An automated phylogenetic tree-based small subunit rRNA taxonomy and alignment pipeline (STAP). PLoS One 3:e2566.
Yona G., Linial N., Linial M. 2000. ProtoMap: automatic classification of protein sequences and hierarchy of protein families. Nucleic Acids Res. 28:49–55.
Zhao X., Li Y., Liu S., Yang Q., Su X., Zhou L., Tang M., Fu R., Li J., Huang Q. 2013. Ultra-deep sequencing enables high-fidelity recovery of biodiversity for bulk arthropod samples without PCR amplification. GigaScience 2:1–4.