# A New Orthology Assessment Method for Phylogenomic Data: Unrooted Phylogenetic Orthology

Jesús A. Ballesteros*[,1] and Gustavo Hormiga[1]
[1]Department of Biological Sciences, The George Washington University
*Corresponding author: E-mail: jabac@gwu.edu.
Associate editor: Xun Gu

## Abstract

**Current sequencing technologies are making available unprecedented amounts of genetic data for a large variety of species including nonmodel organisms. Although many phylogenomic surveys spend considerable time finding orthologs from the wealth of sequence data, these results do not transcend the original study and after being processed for specific phylogenetic purposes these orthologs do not become stable orthology hypotheses. We describe a procedure to detect and document the phylogenetic distribution of orthologs allowing researchers to use this information to guide selection of *loci* best suited to test specific evolutionary questions. At the core of this pipeline is a new phylogenetic orthology method that is neither affected by the position of the root nor requires explicit assignment of outgroups. We discuss the properties of this new orthology assessment method and exemplify its utility for phylogenomics using a small insects dataset. In addition, we exemplify the pipeline to identify and document stable orthologs for the group of orb-weaving spiders (Araneoidea) using RNAseq data. The scripts used in this study, along with sample files and additional documentation, are available at https://github.com/ballesterus/UPhO.**

***Key words:*** **Markov cluster, protein homology, spiders, Araneae, transcriptomics, genomics.**

## Introduction

Since Fitch (1970) made the distinction of two types of gene homology, namely orthology and paralogy, orthology assessment has become a central problem for evolutionary and molecular biologists. For phylogenetics, a fundamental assumption is that the *loci* used to infer evolutionary relationships are orthologs and the violation of this assumption results in phylogenetic error (Delsuc et al. 2005; Kumar et al. 2012; Lemmon and Lemmon 2013).

Although orthology and paralogy are defined on evolutionary terms with explicit phylogenetic properties (Fitch 1970; Thornton and DeSalle 2000; Remm et al. 2001), the vast majority of methods evaluate orthology largely or solely on more or less sophisticated measurements of similarity (Kristensen et al. 2011). Validation and bench-marking studies comparing these methods have shown that in spite of the several methodological refinements, most approaches do not perform significantly better than the basic BLAST (Altschul et al. 1990) score-based reciprocal best hit (RBH) strategy (Chen et al. 2007; Altenhoff and Dessimoz 2009; Salichos and Rokas 2011). Setting aside the issue of assessing accuracy of orthology predictions, one common feature of all orthology assessment analyses is the computational burden imposed by exhaustive pairwise comparisons for distance-based methods, and that of tree inference for the phylogeny-based methods (Trachana et al. 2011).

There is a growing interest in phylogeny-based methods for assessing orthology (Gabaldón 2008) and new ones have been recently proposed (Thornton and DeSalle 2000; Zmasek and Eddy 2002; Chiu et al. 2006; van der Heijden et al. 2007; Hejnol et al. 2009; Marcet-Houben and Gabaldón 2011; Yu et al. 2011; Ramazzotti et al. 2012). Many of these methods have been specifically developed to work on incomplete and sparse genomic datasets, such as those derived from RNAseq experiments and draft genomes (Hejnol et al. 2009; Dunn et al. 2013; Kocot et al. 2013; Yang and Smith 2014). The basic idea in these tree-based methods is to use the phylogeny of gene families to infer events of duplication-speciation of the gene copies; branches derived from speciation events in any given gene family tree are retained to represent a group of orthologs. In addition, probabilistic methods for orthology evaluation that estimate parameters for gene duplication and extinction (e.g., Sennblad and Lagergren 2009; Ullah et al., 2015) will not be discussed in the context of this contribution because these methods depend on a fixed species phylogeny as input parameter.

One shortcoming common to all available phylogenetic orthology methods and programs is that these conduct the tree pruning procedure in a specific order, from the leaves to the root in Agalma (Hejnol et al. 2009; Dunn et al. 2013) and PhyloTreePruner (PTP, Kocot et al. 2013) or from the root to the leaves (Marcet-Houben and Gabaldón 2011); therefore, the position of the root has the potential to affect the outcome of orthology evaluation. The inclusion of distant outgroups has been proposed as a potential solution to improve the orthology assessment (Marcet-Houben and Gabaldón 2011; Yang and Smith 2014); however, it is quite







Article

possible that such outgroup is not represented in the sampled genes or in many cases, where the phylogeny of the group of interest is unknown, such a decision could bias downstream results (Thornton and DeSalle 2000).

## Instant Phylogenomics Versus Stable Orthology Hypotheses

The increasing use of high throughput sequencing technologies in diverse areas of biological research is accumulating large amounts of sequence data from a variety of organisms (Delsuc et al. 2005; Telford and Copley 2011). For some clades, where several reference genomes are available, it has been possible to produce "stable" orthology hypotheses for use in phylogenetics in the form of ultra conserved elements (UCE, McCormack et al. 2012; Jarvis et al. 2014) or target regions (Lemmon et al. 2012; Hedtke et al. 2013). By stable, we mean that can be identified, individually accessed and further documented. Many of these orthology hypotheses are stored and continuously curated in public databases (Afrasiabi et al. 2013; Douzery et al. 2014), and therefore readily available to the rest of the scientific community.

Perhaps not surprisingly, such representation of reference genomes is not available for the vast majority of lineages (Richards 2015). There are ongoing projects toward closing this gap in reference genomes in several groups of organisms, including invertebrates (e.g., Evans et al. 2013; Beach 2014); however, such task has a long way to go before achieving the taxonomic representation and reliability comparable to groups such as vertebrates and some microorganisms mainly of medical and economic importance.

In the meantime, projects using RNA sequencing (transcriptomics) have proved to be a feasible and cost-effective alternative to bring nonmodel organisms into the genomic revolution. This approach is already making significant contributions to our understanding of evolutionary patterns in several groups of invertebrates such as annelids (Novo et al. 2013), arachnids (Sharma et al. 2014, 2015), mollusks (Kocot et al. 2011; Smith et al. 2011), ostracods (Oakley et al. 2013), hexapods (Misof et al. 2014), tardigrades (Campbell et al. 2011), and centipedes (Fernández et al. 2014b), just to mention a few examples.

In general, most phylotranscriptomics analyses consist of three basic steps: (i) sequence data acquisition, sanitation, and trimming, (ii) homology and orthology assessment, and (iii) species tree inference (Delsuc et al. 2005; Dunn et al. 2013; Lemmon and Lemmon 2013; Yang and Smith 2014). At least one method (PHYLDOG, Boussau et al., 2013) bypasses the orthology assessment stage as a whole by simultaneously estimating the gene and species evolutionary histories. Although in all cases considerable time and computing effort are spent in identifying orthologs from the raw data, the products of such efforts are rarely carried over as future reference. In some cases orthologs are not provided as a final product at all or such information is difficult to extract from scattered individual sequences or datasets that have been manipulated for phylogenetic inference. In fact, since the main goal in these studies is to infer phylogenetic relationships of the taxa involved ("instant phylogenomics"); all other results, including the orthologs, are discarded as byproducts suitable to change according to taxonomic representation or parameter manipulation. Consequently, these orthologs lack utility beyond the phylogenomic result and do not translate into stable orthology hypotheses.

As is the case for many nonmodel organisms, spiders (and Arachnida in general) have a restricted set of molecular markers available for phylogenetic inference and this set is heavily dominated by mitochondrial and/or ribosomal genes. In the case of orb-weaving spiders these molecular markers have consistently failed to provide statistically robust phylogenetic hypotheses, especially for the deeper nodes, like those at the inter-familial level and higher (Blackledge et al. 2009; Dimitrov et al. 2012). Several authors have already pointed out the limitations of these markers in illuminating standing evolutionary questions for spider systematics (Lopardo et al. 2011; Agnarsson et al. 2013; Hormiga and Griswold 2014). Nevertheless, few efforts have been made for reverting this trend. Traditional development of markers is a costly, time-consuming endeavor, and current efforts have been unsuccessful or have failed to extend beyond the focal taxon used in the design of primers (Agnarsson 2010; Bidegaray-Batista and Arnedo 2011; Brewer et al. 2014). Recent phylogenomic investigations in spiders are improving our understanding of major phylogenetic and evolutionary patterns (Bond et al. 2014; Fernández et al. 2014a), yet these efforts have not produced accessible orthology hypotheses.

We introduce Unrooted Phylogenetic Orthology (UPhO), as an assessment method that is not affected by the position of the root in the input gene family trees. This new approach is first demonstrated on a trivial example using vertebrate hemoglobins where its general properties are discussed. Then its use in a phylogenomic pipeline is demonstrated and contrasted against alternative methods using a small (14 spp.) dataset comprising insects and outgroups for which well curated genomes exist (see supplementary table 1, Supplementary Material online). Finally, its utility on transcriptomic data is evaluated on a set of 27 species of spiders (see supplementary table 2, Supplementary Material online), focusing on orb-weavers, to demonstrate its ability to identify orthologs of phylogenetic relevance at specific hierarchical levels. This spider dataset was evaluated under a variety of homology (clustering) strategies; clusters composed of "single copy" (SC) genes were also identified and analyzed separately. The resulting list of orthologs, including untrimmed nucleotide (NT) and amino acid (AA) sequence alignments, gene trees and functional annotations, are provided as supplementary files, Supplementary Material online which we hope will translate in long-term orthology hypotheses.

In addition to the orthology assessment criterion, the pipeline herein exemplified differs from phylogenomic approaches in its goal to provide long-term orthology hypotheses and preserving information for the use of these orthologs beyond immediate phylogenetic inference.

## New Approaches

The general procedures for orthology assessment and screening of *loci*, beginning with a collection of reference genomes (gene models or predicted transcripts) or transcriptomes, are depicted in figure 1.

Because the initial motivation for developing this pipeline is the recognition of orthologs and the sequence variation within a reference set, it offers several differences from other phylogenomic pipelines (cf. Delsuc et al. 2005):

The initial clustering of sequences (primary homology) is performed using an "all versus all" BLAST search strategy followed by clustering based on a similarity threshold (*e* value) and minimum alignment length. The initial clustering is minimally processed by filtering out redundant sets and enforcing a minimum number of taxa per cluster. This original clustering gathers sequences that satisfy criteria for reasonably suspecting homology. This procedure also differs from RBH approaches in that it is not limited to one single most similar sequence in the reference; and thus, all sequences in the radius of similarity are considered.

On RNAseq datasets, protein isoforms are not *a priori* removed from the orthology assessment (see "Discussion" section).

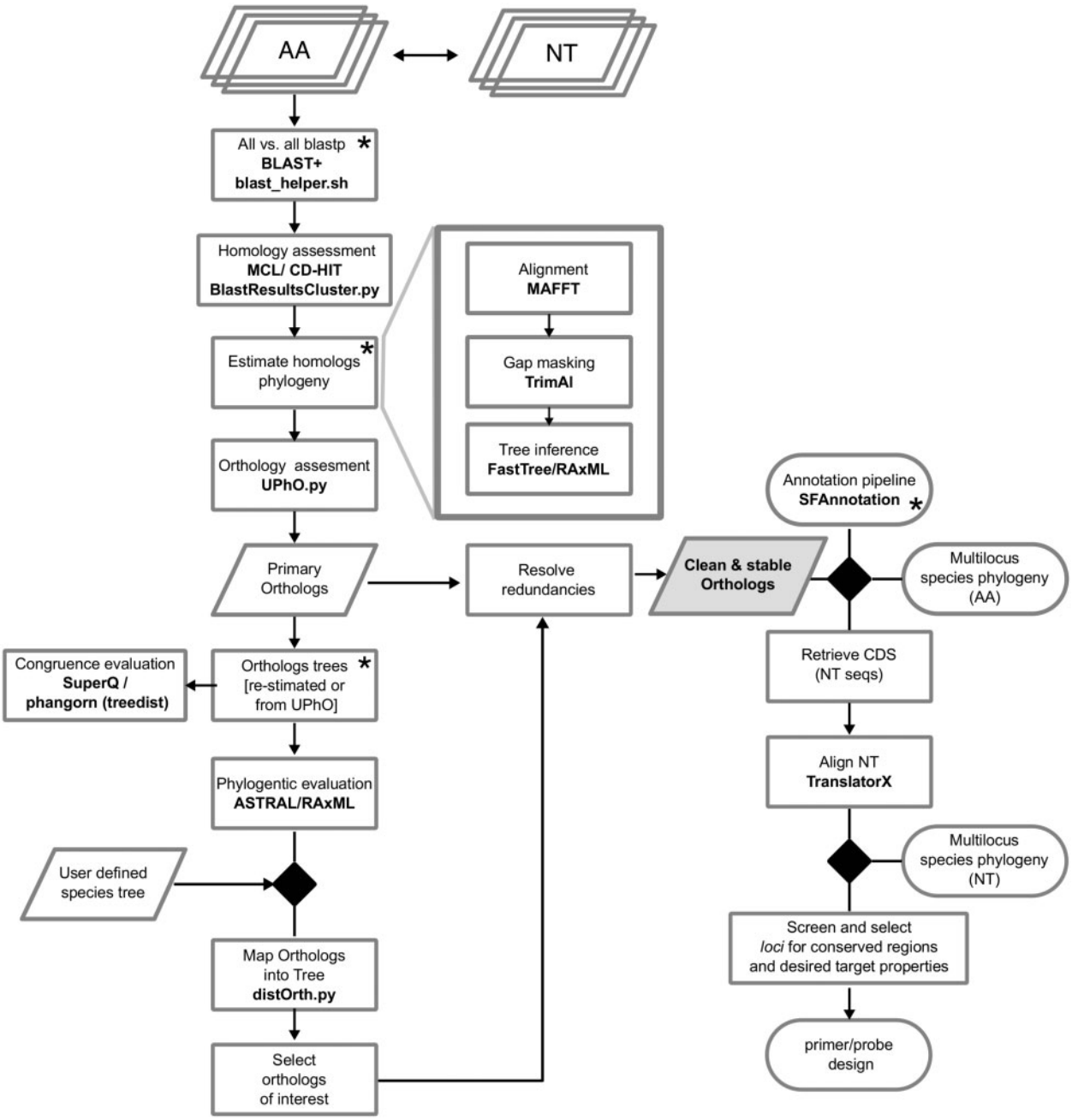


**Fig. 1.** General flowchart of the pipeline followed to identify and select stable orthology hypotheses. Computationally demanding tasks are identified with an *. The software used for each task is indicated in bold type. The list of clean orthogroups identified by this pipeline can be used to retrieve unprocessed sequences or alignments [either nucleotides (NT) or amino acids (AA)].

Emphasis is placed on the evolutionary properties of orthologs rather than on exhaustive pairwise comparisons based on similarity (distance).

Initial orthogroups may overlap in sequence composition. In the case of RNAseq datasets, this can be attributed to domain heterogeneity, resulting in partial homology, and alternative orthology assignment. For its direct use in phylogenetics these "primary" orthologs need to be inspected to avoid redundancy. Clustering methods that prevent or reduce redundancies can be preferred for phylogenetic purposes.

The distribution of orthologs on a reference tree is used to identify and select sets of orthologous genes with desired taxonomic representation.

## Results

### Unrooted Phylogentic Orthology: UPhO

Our UPhO test is implemented in a Python script (UPhO.py), using pure Python modules. This script (in the same manner as PTP) works independently of the rest of the pipeline. The algorithm to identify orthologs is based on the "maximum inclusive" criterion commonly used by several tree-based methods (Hejnol et al. 2009; Dunn et al. 2013; Kocot et al. 2013; Yang and Smith 2014). All of these methods rely on the "species-overlap" algorithm (Gabaldón 2008) and all require a rooted tree to define ancestor-descendant polarity. Our method applies the same species-overlap algorithm without the requirement of a rooted tree by applying orthology evaluation at the level of splits in the gene tree. In gene family trees, leaves represent individual genes associated to a source species, but a given species may be represented by more than one sequence.

#### *Orthologous Split*

For any split $S$ in a gene family tree $\tau$, let $|x|$ denote the number of unique species represented in $S$ and $|y|$ the cardinality of leave nodes (sequences) in $S$. A split is an orthologous split if satisfies $|x| = |y|$. Therefore, each species occurs only once in that split. All other splits not satisfying this condition are considered paralogous splits as these cases imply at least one duplication event.

#### *In-paralogous Split*

For any split in $\tau$, a split is an in-paralogous split if $|x| = 1$ and $|y| > 1$. Note that under this strictly phylogenetic definition, isoforms and alleles will behave as in-paralogous splits. This definition is similar to the ultra-paralogous definition in RIO (Zmasek and Eddy 2002), with the distinction of applying to unrooted networks and to a set of sequences instead of pairs.

Orthologous splits are equivalent to node-based orthologs when the root falls outside of the split in consideration; therefore, node-based orthologs are a special case of split-based orthologs.

UPhO takes as input one or many gene trees (Newick) and produces as output a text file (UPhO_orthologs.txt) with all the identifiers of the sequences found forming an "orthogroup" written in a single line. This list of orthologs can be used to fetch and write the sequences in the orthogroup to FASTA files. In addition, orthologous subtrees (derived from the input gene family trees) can be written to separate Newick files identifying the source gene family tree. The script allows the user to specify if in-paralogs (satisfying the above definition) are to be included in the set of orthologs. The user can also specify a minimum number of different taxa to include per orthogroup. As a default this threshold is set to five species and we advise against decreasing this value below four taxa. The script also allows the user to specify a "support" threshold (default to 0.0). When a different support value is provided, only orthogroups for which the incident branch shows a support value equal or greater than the specified threshold will be written to the output. Note that this criterion may reflect confidence in the monophyly of the orthogroup but does not mean that all nodes in that subtree are equally well supported (see "Discussion" section for a comparison to the support evaluation implemented in PTP).

### Vertebrate Hemoglobins

The gene tree used to explore the properties of UPhO is depicted in figure 2. The subtree on the branch leading to $\alpha$ hemoglobins (hba's) reflects the known species phylogeny; whereas that of $\beta$ hemoglobins (hbb's) shows *Gallus* sister to the *Danio* sequence.

When this gene tree is subject to orthology assessment using PTP it results in only one orthogroup with all hba's; even when rooted at the split between hba's and hbb's ("correct" rooting). The same tree but rooted on any edge within the hba's subtree would find hbb's as the single orthogroup.

UPhO recovers both hba's and hbb's as orthogroups regardless of the position of the root.

### Insects Proteomes

The search of homologs using Markov clustering as implemented in MCL (van Dongen 2000; Enright et al. 2002) with inflation parameter $i = 3$ resulted in 2,538 clusters. After sanitation 1,906 alignments were processed for gene tree estimations and orthology assessment with PTP and UPhO, retaining only orthogroups for which all 14 species were represented. Statistics of the MCL clusters using several inflation values are shown in the supplementary table 3, Supplementary Material online. Results of a partially overlapping dataset using OMA (Roth et al. 2008) are also presented (see details on the OMA analyses in supplementary material, Supplementary Material online).

UPhO consistently discovered more orthogroups than PTP and OMA (fig. 3) (see supplementary section, Supplementary Material online, on the OMA run). The same species phylogeny was consistently recovered by ASTRAL (supplementary fig. 1, Supplementary Material online) regardless of the orthology assessment method. This also includes the homolog gene trees with one representative sequence. Tree distances, using the symmetric difference metric ($d_s$, Bourque 1978; Robinson and Foulds 1979), were used to represent congruence of individual orthologous trees with the reference species trees. UPhO and PTP trees of orthologs showed similar mean and

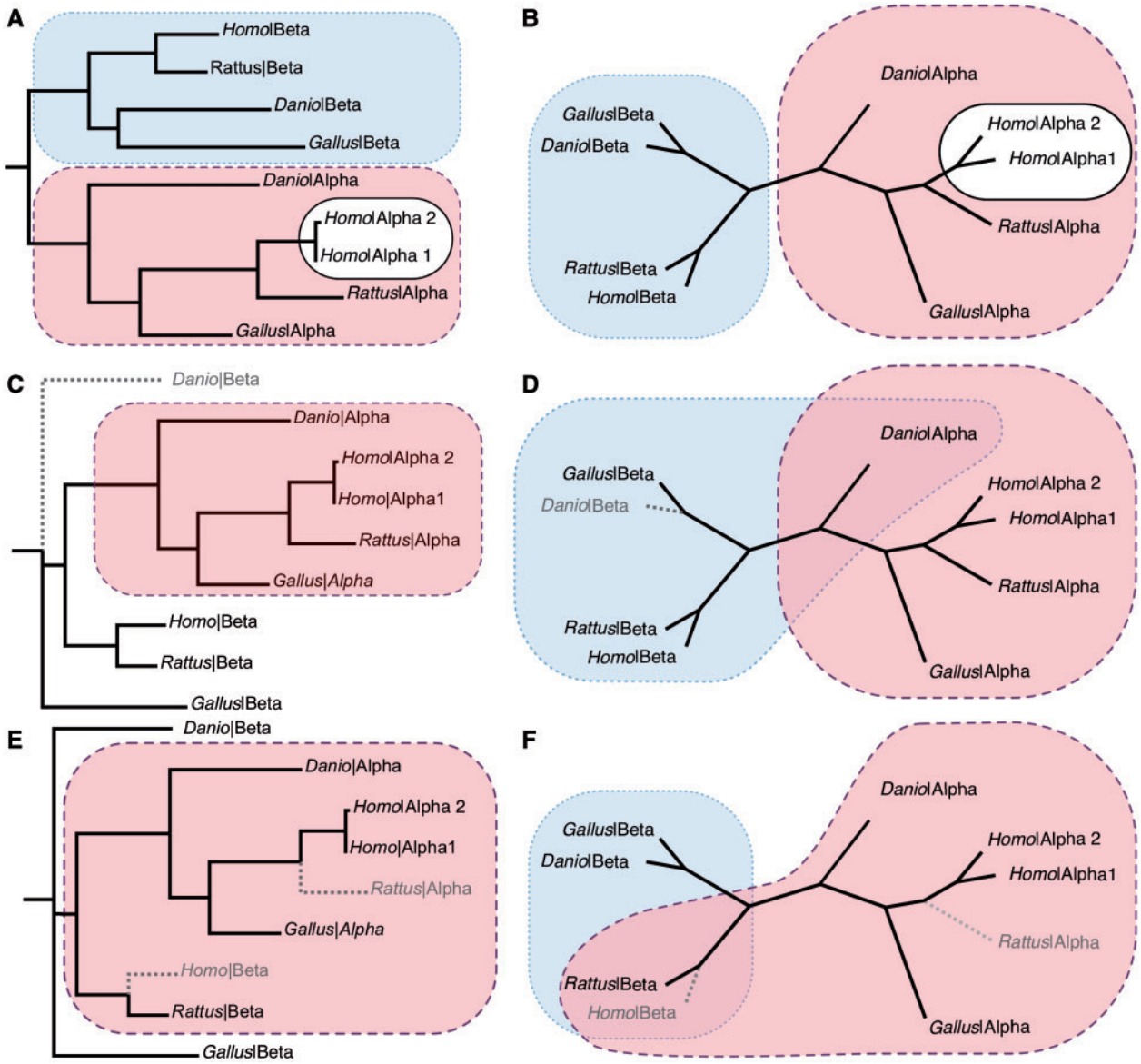


**Fig. 2.** Gene family example using vertebrate hemoglobins. (*A*) Gene family tree displayed with the "true rooting", both $\alpha$ and $\beta$ form monophyletic groups and would be detected by current phylogenetic orthology methods. Note the presence of "in-paralogs" (white circles). (*B*) The same tree as an unrooted network showing the grouping of our unrooted orthology method. (*C*) Arbitrary rooting of the tree would result in the recovery of only one orthologous group when using rooted criteria. (*D*) UPhO invariably recognizes the same two orthogroups; failure to include *Danio*|Beta causes *Danio*|Alpha to be included in both hba and hbb orthogroups. (*E*) and (*F*) Missing sequences could cause "erroneous" groupings regardless of the orthology assessment method.

distribution of $d_s$ distances. The mean $d_s$ of gene trees to the reference was shorter for OMA-derived gene trees.

## Orthologous *Loci* in Orb-Weaving Spiders

The relaxed BLAST clustering, using an $e = 10^{-5}$ cutoff, produced 12,122 clusters with at least five different species. Comparisons of the number of clusters obtained with various clustering strategies (BLAST, BLAST + MCL, and CD-HIT), before and after minimum taxon filtering, are shown in supplementary figure 2, Supplementary Material online. The distribution of number of sequences per cluster are shown in supplementary figure 3, Supplementary Material online and MCL clustering statistics, using various *i* values, are shown in supplementary table 4, Supplementary Material online.

Orthogroups and SC homologs in this dataset were required to include at least five species. In the case of the SC datasets, more strict clustering strategies resulted in more orthogroups (fig. 4*A*). The opposite pattern is found for the rest of the clusters when applying phylogenetic orthology inference. In this case, clustering refinement strategies, such as a more strict BLAST clustering ($e = 10^{-10}$) or the addition of Markov clustering strategy, resulted in fewer and more compact clusters (fig. 4*B*). Nevertheless, as shown in figure 5, different combinations of clustering strategies and orthology assessment variants, share most of their

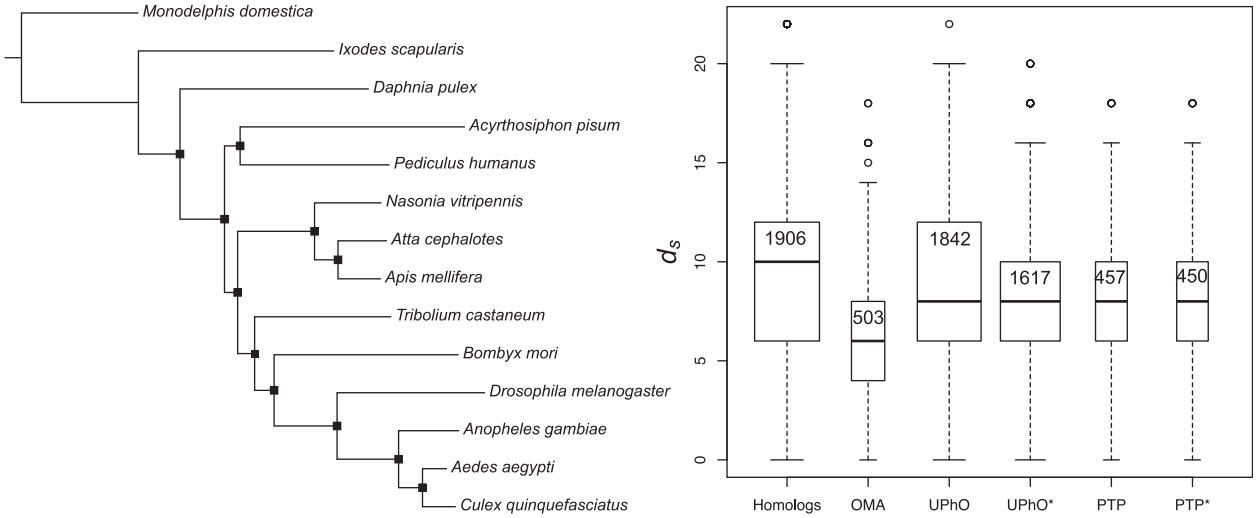


**Fig. 3.** ML tree of the concatenated matrix of 1,617 orthologs recovered with UPhO enforcing 95% support (left). Phylogenetic congruence of individual gene trees are presented as the distribution of topological distances ($d_s$) between this reference species tree and the subtrees of orthologs obtained by OMA, UPhO, and PTP. Results marked with * obtained with support threshold of 0.95, otherwise threshold set to 0 (right).

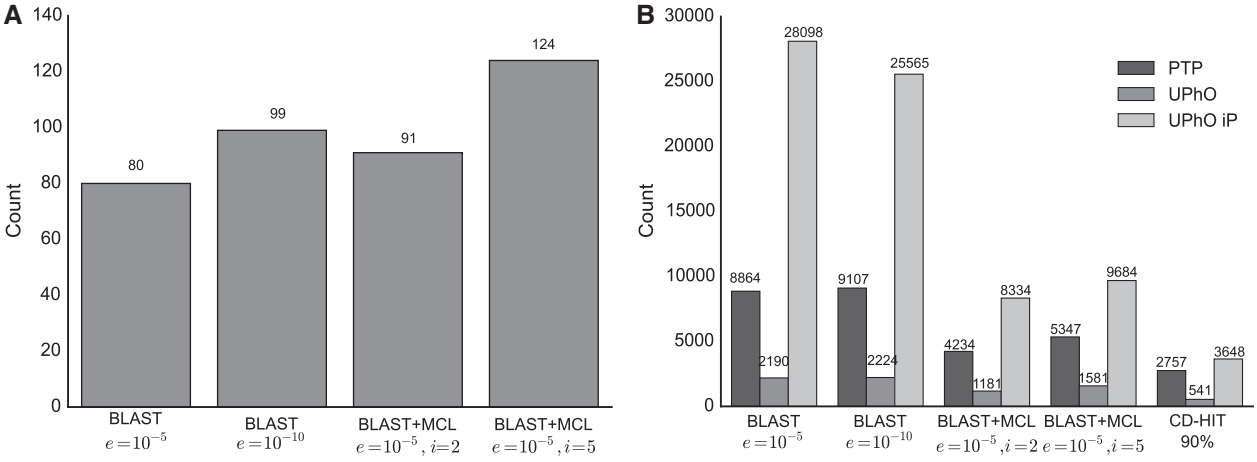


**Fig. 4.** (*A*) Number of SC genes found per clustering method (MCL = Markov clustering). (*B*) Number of orthologs recognized with the phylogenetic orthology criterion using PTP, and our unrooted phylogenetic method with in-paralogs (UPhO iP) and without in-paralogs (UPhO) starting from different homology clusters.

orthologous sequences with the ones discovered using the BLAST $e = 10^{-5}$ clustering strategy. For this reason, we selected the larger collection of orthologs in this last treatment for downstream analyses.

Our unrooted orthology criterion accepting in-paralogs discovered almost three times as many orthologs as PTP (fig. 4*B*). The implementation of Markov clustering reduced the amount of orthologs found with both PTP and UPhO. Runs of UPhO without in-paralogs, that is, discarding any splits with taxon redundancy (including paralogs, or true splicing variants, i.e., same *locus*) resulted in fewer orthologs than PTP even when more than one branch per tree could be pruned. This result suggests that most of the orthogroups recognized by PTP and UPhO with in-paralogs are composed by branches containing sequences from the same species, representing in-paralogs or splicing variants.

The phylogenetic structure for the SC datasets, as shown by their correspondent quartets networks by SuperQ (Grünewald et al. 2013) and the summary species trees, resulted in heterogeneous and conflicting phylogenetic signal. The monophyly of some well-established families was not resolved (e.g., Theridiidae). Tetragnathidae and Araneidae were monophyletic in all but one of the SC trees. These summarized annotated trees and networks of the SC datasets are presented in the supplementary figures 4–7, Supplementary Material online.

None of the SC datasets included all 27 species and from the distribution of orthologs on the topology is evident that some of the discrepancy can be attributed to missing taxa in the input gene family trees. In many cases, a particular species was represented in only one of the input trees. The corresponding networks were also consistently reticulated, indicating the lack of congruence between the input trees.

For the dataset tolerating redundancy in species, and thus subject to orthology evaluation, the test for phylogenetic congruence of the 28,098 primary orthogroups were

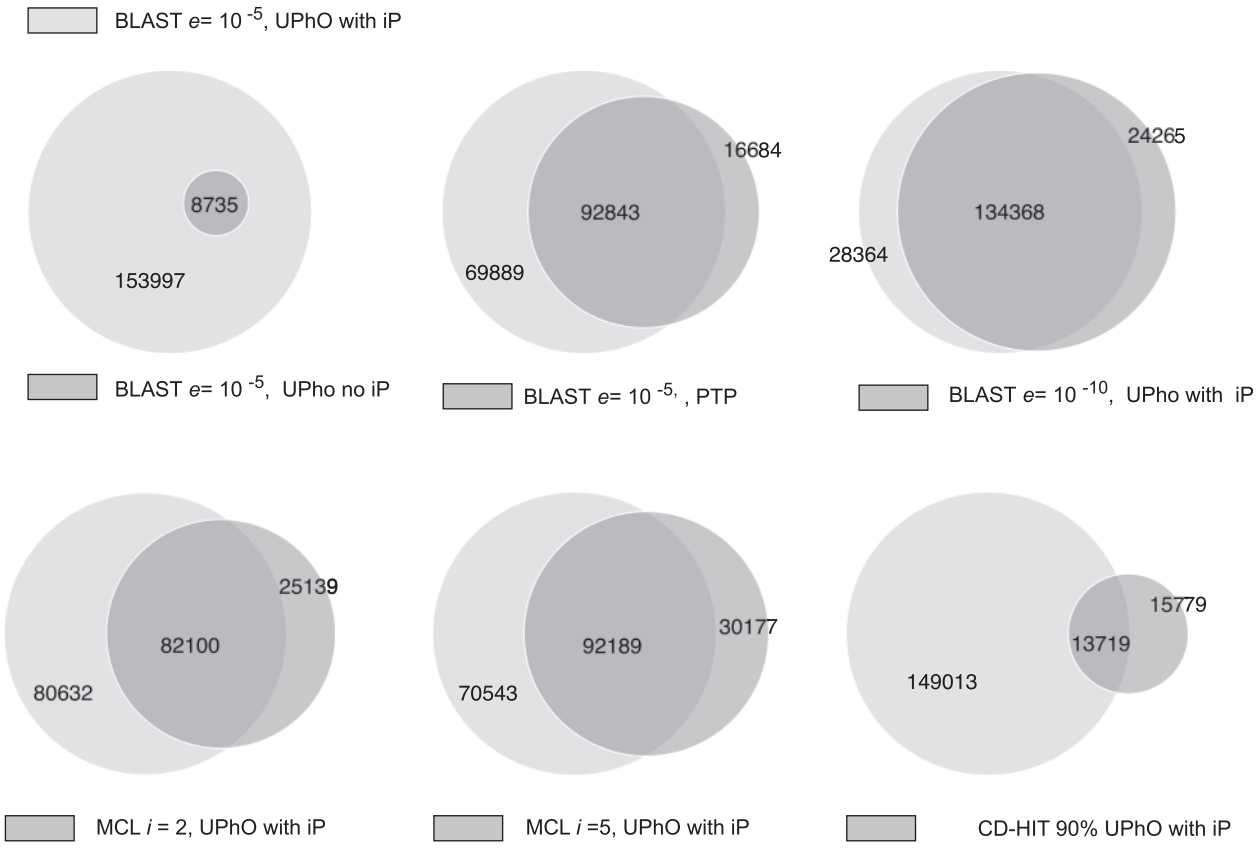


FIG. 5. Venn diagrams showing the overlap of sequence composition in the set of orthologs identified through different clustering methods and parameters. Sets of orthologs derived from different clustering strategies (dark gray) are contrasted with the orthologs found with UPhO with in-paralogs from the BLAST-based clusters using $e = 10^{-5}$ (light gray).

reanalyzed with RAxML resulting in 27,960 gene trees. The discrepancy between the number of orthogroups and the trees obtained is due to cases where the sequences in the orthogroups were identical or the sequence length was reduced to zero after re-alignment and trimming, which are not processed by RAxML. Such instances should be inspected in a case-by-case basis, but its effect is negligible for the search of targets or other analyses. In these cases, consensus sequences from these orthogroups result in either perfect consensus or complete discordance. Overall, the topology shown in this summary tree, figure 6 seems in agreement with current phylogenetic hypotheses and the quartets network (fig. 7) with a well-defined tree-like conformation indicates high congruence within the gene trees.

Using the distribution of orthologs on the summary tree, we retrieved the orthologs at the nodes representing the families Araneidae, Tetragnathidae, and the super family Araneoidea. The orthogroup composition between these sets is summarized in the Venn diagram shown in figure 8. These primary orthologs were processed to remove redundancies in the form of overlaps in sequences from the same cluster tree and removing subsets. This cleaning procedure was done independently to each node-based set of orthologs. For the family Tetragnathidae 5,235 cleaned orthologs were found, for Araneidae resulted in 5,882, and finally the non-redundant set of orthologs for Araneoidea resulted in 6,224 orthologs. The raw (untrimmed) nucleotide alignments of these sets and their functional annotation are provided in the supplementary material, Supplementary Material online. This information can be used, for example, in the design of probes or primers; a task that can be accomplished using available probe/primer design tools and tuned to experimental requirements.

## Discussion

### Unrooted Versus Rooted Orthology

The rationale behind UPhO follows previous phylogenetic orthology methods such as Agalma (Hejnol et al. 2009; Dunn et al. 2013); TreeKO (Marcet-Houben and Gabaldón 2011) and PTP (Kocot et al. 2013) and the "maximum inclusive" criterion described by Yang and Smith (2014). The main difference from available methods lies in that no assumption is made regarding the position of the root of input gene trees nor a reference species tree is required to draw these orthology hypotheses. Although carefully selected outgroups could improve the accuracy in orthology assessment, undisputed outgroups are not always available for some phylogenetic problems, especially for nonmodel organisms (Pearson et al. 2013). In addition, the fragmentary nature of most phylogenomic datasets inevitably results in sequence clusters for which the outgroup is not necessarily represented (Yang and Smith 2014), in which cases these gene family trees are either discarded or arbitrarily rooted.

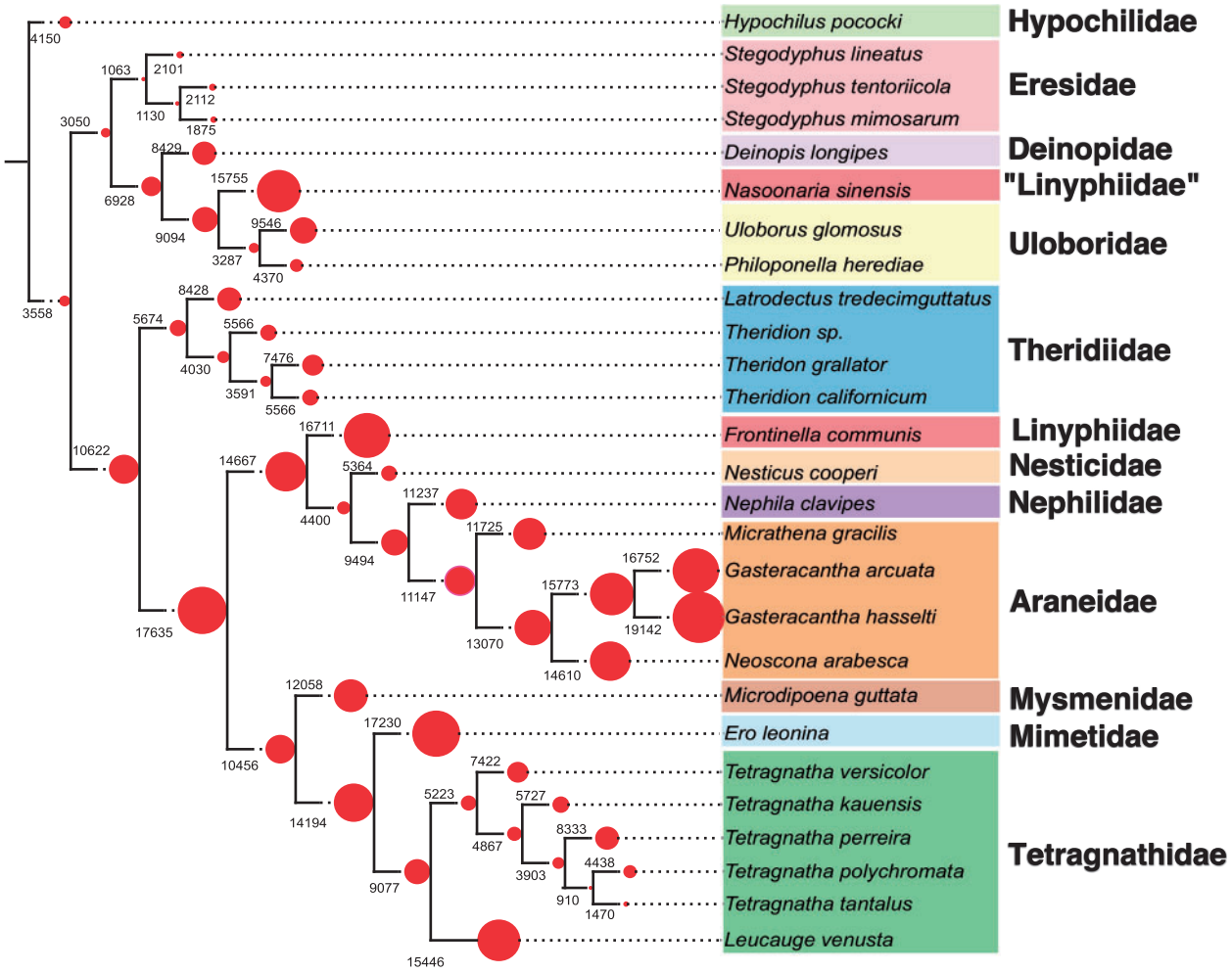


**Fig. 6.** Summary species tree derived from 27,960 individual gene trees with at least five species from the primary orthologs identified by UPhO with in-paralogs. The size of the node bubble is proportional to the number of orthologs mapped to that node. The species labeled *Nasoonaria sinensis* represents a case of misidentification and the specimen most likely belongs to the family Uloboridae.

## Phylogenetic Orthology Assessment on a Trivial Hemoglobin Dataset

A classic example for the concepts of orthology and paralogy is provided in the evolution of vertebrate hemoglobins. Consider the simplified tree of vertebrate hemoglobins in figure 2. This tree is derived from real sequences (see "Material and Methods" section) and represents a plausible gene-family tree subject to orthology assessment. In this example, cases of orthology, out-paralogy, and in-paralogy (hba-1 and hba-2) are represented. When "correctly" rooted, this gene tree implies an early duplication event in the common ancestor of the vertebrates originating two paralogous gene copies, with one branch leading to $\alpha$-hemoglobins (hba's) and the other to $\beta$-hemoglobins (hbb's). After this early duplication, cladogenesis (speciation) accounts for the subsequent divergences within hba's and hbb's; a more recent duplication event is represented in the hba branch leading to in-paralogous copies in *Homo*. In this example and for a reason we will not explore, the branch of hbb's does not reflect the "known" phylogeny of the vertebrate species involved. There are a variety of reasons, aside paralogy, that can cause such topological discordance, ranging from biological to methodological biases. If the tree is rooted on the "correct" edge any tree-based orthology tests should recover hba and hbb as mutually exclusive orthogroups. If the root is however placed at any other edge/node the orthology pruning would detect one sub-tree but not the other, even if we use one of the *Danio* sequences as outgroup. In fact the only way to objectively root this tree would require a more distant member of the globin protein family.

All phylogenetic orthology methods are susceptible to the confounding effects of stochastic extinction of gene copies and ours is no exception. Missing data and extinction of gene copies can also mislead correct orthology assessment. In our example, given a gene tree in which *Danio*|Beta is not represented, the unrooted orthology test will recognize the orthogroup of hba's and the orthogroup of hbb+*Danio*|Alpha, whereas for rooted methods the recovery of this pseudo-ortholog depends again on the position of the root, for example, if rooted on the edge leading to the leaf node *Gallus*|Alpha, it will result in the same mixture of hbb's with *Danio*|Alpha. This assignment error is particularly harmless, but other examples caused by incomplete sampling of gene copies would imply an erroneous phylogeny. For

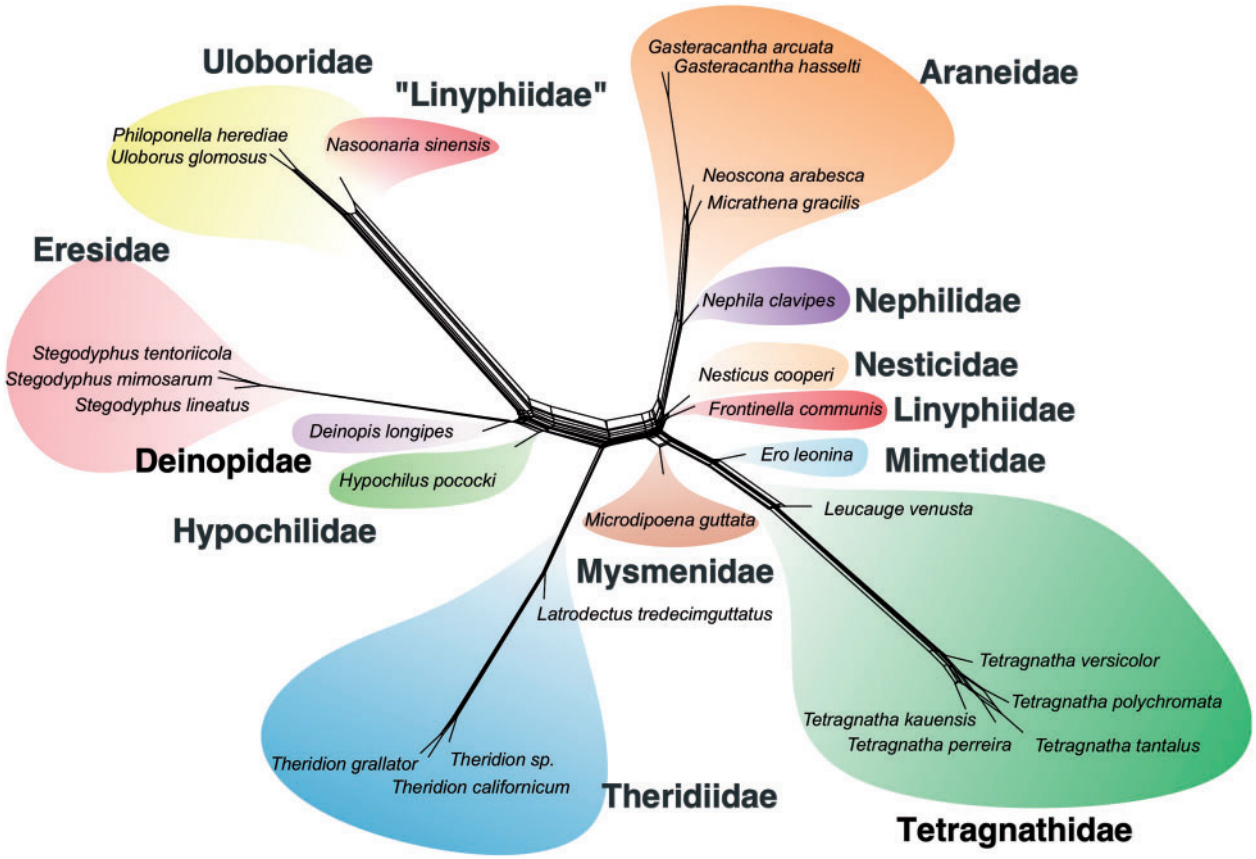


FIG. 7. Quartet networks from 27,960 gene trees with at least five species derived from the primary orthologs identified by UPhO with in-paralogs. The species labeled *N. sinensis* represents a case of misidentification and the specimen most likely belongs to the family Uloboridae.

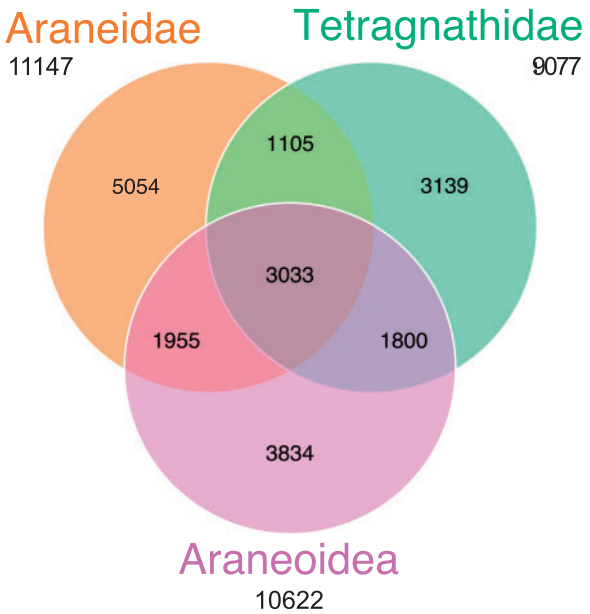


FIG. 8. Venn diagram showing the number of orthogroup shared between the node base composition of orthologs between the families Araneidae, Tetragnathidae, and the superfamily Araneoidea. This information can be used as an additional criterion for selecting targets.

example, if *Rattus*|Alpha and *Homo*|Beta are missing then the phylogenetic orthology criterion will group *Rattus*|Beta in the hba's orthogroup, and imply an erroneous species phylogeny. We are only able to recognize these groupings as erroneous because in this example we "know" the complete history of the gene duplication events and the relationships of the species involved. Solving this problem would require external evidence; luckily this evidence can also be found as congruence in the cumulative phylogenetic signal in other orthologs involving the same species. This simplified hemoglobin tree represents only one of the simplest cases of the entanglement of gene and species phylogenies.

## Insect Phylogeny from Proteomes

For the genomic dataset and starting from the same gene clusters, UPhO consistently recognized more orthologs than PTP and OMA. The reference phylogeny inferred from the maximum likelihood (ML) analyses of the concatenated dataset, shown in figure 3 is congruent with the one reported by Kocot et al. (2013). This tree is also in agreement with a recent phylogenomic survey of hexapods except for the position of the aphid, *Acyrthosiphon pisum*, shown to be sister to the rest of the included insect lineages instead of sister to *Pediculus humanus* (Misof et al. 2014). Notably, a fully congruent topology is resolved by ASTRAL regardless of the orthology assessment treatment (see supplemental fig. 1, Supplementary Material online). It is also worth noticing

the surprising level of phylogenetic congruence with the summary tree observed in the unevaluated (no orthology assessment performed) gene family trees. However, the leaves in these trees were not drawn randomly but are connected in the original gene family tree, from which visited paralogs were discarded. This result may explain why some orthology strategies such as RBH perform with reasonable accuracy in some datasets. Results from the OMA run produced individual gene trees in average more congruent than those found by UPhO and PTP. However, it must be noted that most of the sequences identified by OMA are represented in the UPhO orthologs and, many reasonably congruent orthogroups are not recovered by OMA (fig. 9). The increased congruence obtained by OMA is possibly due to limiting the inclusion of the highest weighted pair distances, resulting in minimally connected networks. On the other hand, UPhO includes all or most of these highly congruent orthogroups, and many others not recovered by OMA. If increased conservation is desired, these UPhO orthologs can be further refined using explicit criteria; for example, rates of evolution or phylogenetic informativeness.

## Orthologs from Spider Transcriptomes

In all clustering and orthology assessment treatments the family level taxa were found monophyletic with a sole exception: Linyphiidae, for which *Nasoonaria sinensis* consistently grouped with Uloboridae in all treatments. This odd grouping is clearly the result of specimen misidentification, given that the authors mention that this spider weaves "orbs above bushes" (Zhao et al. 2014). Linyphiid spiders do not build orbicular snares but sheet webs, with little to no resemblance to orb-webs. Uloboridae on the other hand, build typical orb-webs with radii and spirals.

The summary trees and networks derived from UPhO (figs. 6 and 7) and PTP (supplementary fig. 9, Supplementary Material online) were very similar, differing only on the position of Eresidae (*Stegodyphus* spp.), and both showing highly congruent (tree-like) quartet networks.

The phylogenetic structure derived from the set of strict orthologs (no in-paralogs) discovered with UPhO (supplementary fig. 9, Supplementary Material online) resulted in a similar topology than the preferred tree, except again for the placement of Eresidae; this time found within Araneoidea. Note however that the clade of eresids is represented by very few orthologs in all datasets, and that the network of quartets resulted in an overall reticulated pattern. For this particular type of data, the presence of isoforms would cause to reject several orthogroups by confounding these with in-paralogs. If strict orthogroups (without in-paralogs) are desired, splicing isoforms should be removed from the input data but the criterion for recognizing splicing variants should rely on explicit criteria additional to the assembler output.

## Sequence Clustering Matters

Initial homology assessment greatly influences the downstream analyses. The use of natural clustering methods such as MCL (van Dongen 2000) produces reasonable good initial homologies for phylogenetics. For documentation of orthology hypotheses, particularly from complex dataset such as from RNAseq this approach may result incomplete due to

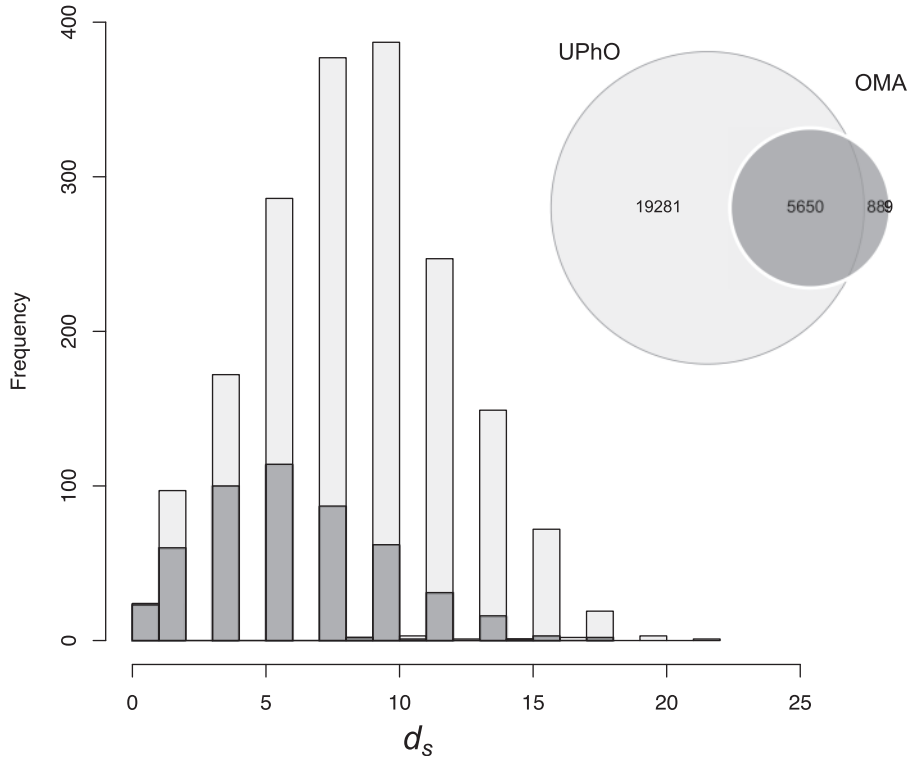


**Fig. 9.** Venn diagram showing the number of sequences shared by the orthogroups found between OMA (dark grey) and UPhO (light grey). The histogram shows the distribution of $d_s$ distances between these two sets of orthogroups.

the effects of ignoring distant and domain specific homologies (Enright et al. 2002). On the other hand investigating all neighboring sequences using a similarity radius from every query sequence may result in extreme redundancies for some datasets. For the insect dataset this exhaustive BLAST radius strategy produced too many clusters to be reasonably evaluated (>150,000 clusters). Alternatively, one could restrict the homology search to be centered on one or few reference species; for example, around a taxon of interest or preferring species for which good quality genomes exist. In cases of extreme redundancy, MCL is a much better alternative because this method automatically adapts to the relative connectivity of the dataset. However, this same adaptive connectivity may in some other cases fail to group distant homologs. Sequence clustering is a hard problem in bioinformatics and is very difficult to compare and evaluate different clustering sets. In the case of MCL, the exploration of different inflation parameter values (*i*) which affects the "granularity" or "tightness" of resulting clusters is strongly recommended. Low values of the inflation parameters tend to produce "coarse" clusters; that is to say larger, more loosely connected clusters. In opposition larger values of the *i* parameter produce more "fine grained" clusters, usually smaller and with tighter connectivity; refer to van Dongen (2000) for further details on MCL and its parameters. A conjecture that remains untested is that PTP, producing only one orthogroup per gene tree, would perform better when using more fine grained clusters (larger *i* value), whereas UPhO would be more resilient to the effects of coarse homologs (lower *i* value). In either case, initial homology should be subject to exploration and adjustment according to the properties of the data and the objectives of the study.

### Isoforms, Paralogs and Representative Sequences

The presence of splicing variants is an inherent complication when dealing with RNAseq data. A mature mRNA molecule can be composed of disjunct regions of the genome, separated by noncoding regions (introns); the combinatorial effects of this process greatly increases the diversity of transcripts (Matlin et al. 2005; Sammeth et al. 2008). When explicitly mentioned, phylogenetic studies based on transcriptomics data deal with splicing variants by removing all but one of the isoform from the assembly prior to orthology assessment (e.g., Hejnol et al. 2009; Fernández et al. 2014a; Sharma et al. 2014). Although this procedure can be justified as a mean to increase the occupancy in the final matrix, it must be noted that current *de novo* transcriptome assembly software such as Trinity (Haas et al. 2013) do not differentiate between paralogs, isoforms, and allelic variants during the assembly (Grabherr et al. 2011; Yang and Smith 2014). An alternative approach proposed by Bazinet et al. (2013) combines these variants into a consensus sequence incorporating ambiguity codes; thus avoiding errors associated with selecting an inappropriate representative but potentially degrading phylogenetic signal. Our pipeline retains the putative splicing variants and its homology is established at a later stage using explicit criteria. If these "isoforms" are found as forming a clade and thus retained in the orthogroup as in-paralogs the information derived from its similarities and differences is particularly relevant to the design of probes or primers. In other orthology pipelines this information can be irreparably lost in cases where supposed isoforms were originally removed. However, if one decides to use these orthogroups directly for phylogenetic inference, representative sequences can be easily selected, duplicates filtered, or consensus sequences produced (Bazinet et al. 2013); so that only one sequence per species is retained for phylogenetic analyses.

For designing targets and probes, the potential negative effects of retaining isoforms are negligible, as they should show identity in the shared domains. Allegedly, this identical blocks could inflate our confidence on the level of conservation of a particular region; on the other hand, being able to account for allelic variants, and in-paralogs allows us to take these variations into account when designing targets. On the other extreme, some real paralogs may exist lumped in the isoform category and should thus be subjected to orthology test. Teasing apart alleles from in-paralogs and isoforms could also be done by other means, as all these variants should show specific patterns of sequence variation; however, such *a priori* inspection is a time-consuming task that does not directly affect the orthology assessment and is therefore beyond the scope of this orthology pipeline. On the other hand, retaining isoform variants, acknowledges the possibility of independent protein domain homology and test the orthology assumption among all available sequence blocks (Thornton and DeSalle 2000; Gabaldón and Koonin 2013; Sonnhammer et al. 2014).

### UPhO Versus PTP

Another obstacle to a more widespread application of phylogeny-based methods has been the ties to the specific pipeline for which they were developed for, and thus are difficult to port (cf. Dunn et al. 2013; Yang and Smith 2014). An exception to this trend is found in PTP (Kocot et al. 2013), which is provided as a standalone program for the only purpose of pruning monophylectic subtrees from input gene trees and its corresponding FASTA file with the sequences. One important feature in PTP is its tolerance to the presence of in-paralogs (although such feature cannot be toggled off) and it also incorporates branch support based criteria by collapsing unsupported nodes before orthology evaluation. The output of PTP consists of "pruned" FASTA files and allows the user to define whether in-paralogs are included in the output FASTA, or only one of these should be retained. This method has been tested in some recent empirical studies (Bazinet et al. 2013; Andrade et al. 2014; Whelan et al. 2015).

One shortcoming of PTP is that it recovers only one ortholog (subtree) per gene family tree, the one with most leaves. In theory, one could apply iterative runs of PTP on the same input tree–FASTA pair, and recover more than one ortholog per input tree until no subtrees satisfying the conditions are found, provided that on each iteration the "orthogroup" found in the previous one is ignored. Our method (UPhO), on the other hand, returns all the branches from a given gene tree satisfying the orthology tests; these are then considered primary orthologs.

The rescue of in-paralogs in UPhO and PTP effectively increases the number of orthologs discovered. The terms out- and in-paralogs were introduced by Remm et al. (2001) to distinguish paralogous genes derived from one or more duplication events which occurred before the speciation event (out-paralogy) or after speciation (in-paralogy). The distinction is useful for phylogenetics because in an orthogroup with in-paralogous copies, any individual in-paralog implies the same species phylogeny. This on the other hand is not always the case for out-paralogs. It must be noted however, that our method is based on an operational definition of in-paralogy (see "Results" section) which differs from the one implemented by PTP. In UPhO in-paralogy is restricted only to duplication events confined at the level of terminal species, whereas in PTP adjacency of the nodes suffices (see example in supplementary fig. 10, Supplementary Material online). In PTP this operation allows for paraphyletic assemblies caused by a duplication event followed by speciation to fall within the in-paralog category. Although PTP approach to in-paralogy has no negative effects on phylogenetic inference (because the phylogenetic structure at the species level is not altered), we believe that such distinction is important for the recognition of stable orthology hypotheses where evidence of gene duplication may indicate the existence of paralogous copies. It must be noted also that isoforms and duplicated sequences would behave as in-paralogs using the definition outlined in the "Results" section. This behavior is in agreement with the phylogenetic properties of isoforms and not its definition.

### Branch Support in UPhO Versus PTP

Both PTP and UPhO allow the user to incorporate measurements of topological support from the input trees as an additional criterion during the orthology assessment. If a minimum support value is provided by the user, UPhO will return all orthogroups with equal or greater support value for the split defining that partitioning in the gene tree. This threshold is applied over all split evaluations including in-paralogy evaluations, therefore only "well" supported in-paralogs splits will be considered and evaluated as such. This approach is almost opposite to the support evaluation implemented in PTP, where support thresholds are used in collapsing low supported nodes and interpreting resulting polytomies as cases of in-paralogy. Therefore, while UPhO uses support information to reject unreliable groupings, PTP uses it to relax the evaluation. For the insect proteome dataset this effect might not be particularly problematic, as even homologs show reasonable phylogenetic congruence, but this distinction may be more important in other cases. Although a subtle difference, the support value is used in UPhO as a mean to increase our confidence in the orthologs found, whereas in PTP this criterion decreases such confidence. The argument provided by the authors of PTP is that such relaxation allows the rescue of sequences in cases where "a weakly supported tree topology incorrectly indicates paralogy"; however, it is clear that more orthologs are missed by ignoring other branches in the same input gene tree that do satisfy the phylogenetic orthology criterion.

The main difference found between PTP and UPhO lies in the amount of orthologs found (completeness). Such pattern is not particularly surprising as both methods rely on similar paradigms for orthology evaluation and thus no dramatic difference in qualities are to be expected.

Finally, the quality of the orthology hypotheses relies heavily on the quality of the gene trees, and in the cascade of procedures involved in the accuracy of the gene family phylogenies: alignment, masking, tree estimation, etc.

### Phylogenetic Structure and Congruence in Primary Orthologs

Our results suggest that the search of SC genes directly from the clustering produced poor results in the transcriptomic dataset. These groups with only one sequence per species were very rare in the dataset and most show heterogeneous taxon composition leading to indecisive (*sensu* Steel and Sanderson 2010) taxon sets for the 27 species tree. One of the potential explanations for the discrepancy observed in these genes lies in the fact that even though they are SC variants in the reference dataset, additional homologous copies could exist but were not detected. There are many factors that can affect our ability of detecting any particular gene copy derived from RNAseq experiments; for example, expression levels, tissue, or ontogenetic-specific expression, and, inaccuracies during assembly. Some of these factors apply as well to whole genome datasets, yet are much more accentuated on RNAseq. These caveats are not restricted to the SC dataset and should be considered as a potential problem to all orthology assessment analyses, this is: the inability to falsify the orthology assumption from a sampling of the universe of genes copies. This "missing" evidence refers as well to the homolog copies present in species not included in the analysis. Accuracy, measured as the resolution of well-established monophyletic groups, increases for pruned orthologs simply because more observations are taken into account, including the commonly discarded variants lumped in the "isoform" category during assembly. More evidence increases our confidence on our orthology hypotheses. Nevertheless, these original hypotheses should be subject to further testing.

## Materials and Methods

### Taxon Selection and Data Acquisition

Sequences from the vertebrate hemoglobin example were retrieved from NCBI entrez using "hemoglobin alpha/beta" as search terms. Accession numbers as follow: *Homo sapiens* (NG_000006.1, NG_000006.1, NG_000007.3), *Rattus norvegicus* (AC096051.7, X15009.1), *Gallus gallus* (AC172304.2, L17432.1), and *Danio rerio* (NM_131257.2, NM_131759.1).

The genomic dataset includes 14 proteomes obtained from the inparanoid8 database (Sonnhammer and Östlund 2014). Taxon sampling (see supplementary table 1, Supplementary Material online) follows closely the example used by Kocot et al. (2013) but unprocessed sequences were preferred.

The transcriptomic dataset focuses on the search for orthologs for the clade of ecribellate orb-weaving spiders

(Araneoidea) for which we included all currently available RNAseq experiments from the SRA (Mattila et al. 2012; Croucher et al. 2013; He et al. 2013; Bond et al. 2014; Fernandez et al. 2014a; Sharma et al. 2014; Yim et al. 2014; see supplementary table 2, Supplementary Material online). In addition, we included representatives of the cribellate orb-weavers (Deinopoidea), Eresidae, and one basal araneomorph (*Hypochilus*) as outgroups.

Assembly, BLAST searches, alignment, and tree inference tasks were performed in the high performance computing cluster of The George Washington University (Colonial One).

Custom scripts were written for all parts of the pipeline mainly using Python 2.7, some of which made use of the BioPython (Cock et al. 2009) and ETE2 (Huerta-Cepas et al. 2010) modules. These scripts are freely available from github at https://github.com/ballesterus/UPhO.git and https://github/ballesterus/PhyloUtensils.git under the GNU GPL software license http://www.gnu.org/licenses/gpl.html (last accessed April 13, 2016).

## Read Cleaning and Assembly

Quality of the reads was assessed before and after cleaning, using FastQC http://www.bioinformatics.babraham.ac.uk/projects/fastqc/ (last accessed April 13, 2016). The cleaning procedure consisted of clipping the starting 14 bases following Hansen et al. (2010), trimming the low-quality tails under a Phred score of 25 and discarding reads <30 bp. All trimming procedures were performed using the package "ShortRead" for the R environment (Morgan et al. 2009). Illumina reads were assembled *de novo* using Trinity (version r20140717, Haas et al. 2013). The reads were normalized before assembly to ameliorate the effects of over represented reads. Reads from Roche 454 sequencing were cleaned with the same quality and read length criteria, but were assembled using newbler (Margulies et al. 2006).

Whenever information on library preparation was available, it was used to modify pre-processing and assembly parameters. When the use of paired end strand-specific libraries was not clear, the reads were assembled using both flags, retaining assemblies with better scores based on N50 and the number of coding sequences. The search for open reading frames was performed using Transdecoder (Haas et al. 2013) retaining only coding regions >50 amino acids.

Assemblies from the same species but different runs (*Frontinella communis*, *Tetragnatha kauensis*, *Theridion grallator*, and *Theridon californicum*) were assembled independently. The resulting amino acid sequences were combined in nonredundant set with CD-HIT (Fu et al. 2012).

## Sequence Clustering Methods

We restricted all comparisons to protein coding sequences, and most of the analyses were done using amino acids sequences. By doing this, we intended to avoid some of the difficulties associated with the use of nucleotide sequences, including alignment accuracy and base composition heterogeneity (Breinholt and Kawahara 2013; Pearson 2013).

For each of the reference species FASTA file, sequence identifiers were renamed to include the name of each species using minreID.py. Protein coding sequences (AA sequences) were combined in one reference FASTA file. Using BLAST+ (version 2.29.29; Camacho et al. 2009) we created a local database to perform an all versus all search with blastp with a relaxed expectation value threshold of $e = 1 \times 10^{-3}$. The BLAST results were output as comma separated text file to facilitate parsing. This resulting output file contains all possible pairwise comparisons between all sequences included in our reference dataset. A helper shell script (blast_helper.sh) is provided to create local database of the reference sequences and performed the all versus all BLAST searches writing the results in csv output format.

This BLAST output file was used as the starting point to produce sequence clusters based on a additional *e* score thresholds of ($1 \times 10^{-5}$ and $1 \times 10^{-10}$), a minimum alignment length (50 sites), and minimum taxa representation.

For the insect proteome dataset, the minimum number was set to 14 species per cluster thus avoiding confounding effects of missing taxa during phylogenetic inference and allowing the use of pairwise tree distance ($d_s$) for evaluating phylogenetic congruence of the orthologs trees. For the spider transcriptome dataset, the minimum number of taxa was relaxed to five different species. The filtered sequences in each cluster were retrieved from the reference file and divided to produce one FASTA file per cluster (Get_fasta_from_Ref.py).

The initial BLAST $e = 1 \times 10^{-5}$ output file was further refined using Markov clustering algorithm as implemented in the program MCL (van Dongen 2000; Enright et al. 2002) following the procedures described in the documentation for clustering protein sequences from BLAST results and using various inflation parameters.

A special type of clusters are those composed exclusively of sequences of different species and thus putative "single copy genes" (SC dataset hereafter). The vast majority of the clusters, however, have some degree of taxon redundancy and thus these clusters of similar sequences are subject to orthology assessment. A comparison of SC and evaluated orthogroups is presented for the spider transcriptome dataset.

In addition, the initial reference sequences were clustered with CD-HIT (Fu et al. 2012) with a similarity threshold of 90%. The CD-HIT package is most commonly used in bioinformatic applications to eliminate or reduce redundant sequences from two or more sequence databases or files. However, it can also cluster sequences based on a similarity threshold (see supplementary documentation, Supplementary Material online for further details). This program also includes its own scripts to generate individual FASTA files from the clusters, but does not allow further filtering. The rationale for trying CD-HIT as clustering method was to test its performance in comparison to the most explicit and commonly used BLAST-based clustering. The most clear advantage of CD-HIT over our BLAST clustering implementation is its speed.

## Orthology Assessment

Orthology was assessed with the phylogenetic criterion (see "Discussion" section), as implemented in PTP (Kocot et al. 2013) and UPhO.

All cluster files were subject to orthology test (except for the SC datasets), which represent a trivial case for the phylogenetic orthology test. PTP requires the user to set as parameters the minimum number of taxa to consider an orthogroup, and a branch support value under which to collapse nodes. For the transcriptomic dataset this support value was set to 0 and thus nodes were not collapsed based on support. For the insect proteomes supports values of 0.0 and 0.95 were explored.

Graphs-based orthology evaluation using OMA was explored for the smaller insect dataset. The use of OMA for the larger spider transcriptome data resulted prohibitive due to constrains in time and availability of computing resources. The results from OMA are based on 13 species instead of 14 (see supplementary material, Supplementary Material online). Only orthogroups containing all 13 species were considered for the comparisons and the resulting FASTA files were processed for phylogenetic estimation with the same parameters used for obtaining gene family trees in the phylogenetic orthology analyses. The resulting gene trees were then used to evaluate phylogenetic congruence with PTP and UPhO, the only difference being that the missing tip was dropped for the comparisons with OMA gene trees.

## Redundancies

Depending on the original gene homologies (clustering), the primary orthologs derived from pruning input gene family trees may contain redundancies which potentially violate independence for phylogenetic analyses, that is, that a given sequence identifier is found in more than one orthogroup. These redundancies are due to two causes: (1) orthogroups derived from a different cluster/tree, where the orthology statement applies to parts of the sequence (domain orthologs) and (2) orthogroups derived from the same cluster/tree, which are caused by uncertainty on orthology assignment. For example, in figure 2D, the absence of Danio|Alpha causes Danio|Beta to be included in both hba and hbb orthogroups. The first type of redundancies should be tolerated, or contrasted with additional information, as they imply putative independent *loci*. In the case of subsets, only the largest orthogroup is retained. In the second case, redundancies were solved arbitrarily toward the largest orthogroup. These types of redundancies would require additional evidence to be properly tested. For direct phylogenetic inference, an additional cleaning procedure can be applied so that only mutually exclusive (no overlapping) orthogroups are used.

## Phylogenetic Inference, Trees and Networks

Sequence clusters were aligned using MAFFT automatic strategy (L-INSI for most cases, Katoh and Standley 2013). Gap-rich regions were masked using trimAl with "gappyout" flag (Capella-Gutiérrez et al. 2009). Phylogenetic trees for each cluster were computed using FastTree (Price et al. 2010) using the Jones Taylor Thorton (JTT) substitution matrix. FastTree was favored at this stage because of the larger number of alignments to process; the effects on the accuracy of these initial trees on the downstream analyses require further inspection.

An additional sanitation step was used on the trimmed alignments so that sequences composed by 50% or more ambiguous sites from the total alignment length were removed (Al2Phylo.py).

For the transcriptomic dataset, orthologs were aligned and trimmed in a similar fashion as the original clusters, but this time RAxML v.8.0 (Stamatakis 2014) was used to estimate trees. Tree searches were done using 100 rapid bootstraps with automatic protein model selection plus gamma (-m PROTGAMMAAUTO).

To facilitate and accelerate phylogenetic analyses, MAFFT, trimAl RAxML, and FastTree were wrapped along with gnu-parallel (Tange 2011) in a simple shell script (paMATRAX+.sh) that runs each step of the phylogenetic pipeline in as many parallel processes as possible.

Species trees were estimated from the collection of individual gene trees using ASTRAL (Mirarab et al. 2014). For the proteomic dataset, *Monodelphis domestica* was used to root the species tree; in the transcriptomic dataset this tree was rooted on *Hypochilus pococki* when present, or in its defect *Stegodyphus lineatus*.

For the insect proteome dataset, orthologs were concatenated in a supermatrix (geneStitcher.py). Individual alignments were inspected to guarantee only one sequence per species is included (regardless of the cause the duplication, Al2phylo.py) and analyzed without partitioning in RAxML. Since all orthologous trees and the supermatrix species tree have the same number of leaves, phylogenetic congruence of orthologs was measured using the $d_s$ metrics between individual trees of orthogroups (derived from UPhO) and the species tree (from supermatrix ML estimation). The R package "phangorn" (Schliep 2011) was used to compute these topological comparisons. In order to allow comparisons with PTP results, individual subtrees had to be edited by hand from the standard output of PTP since this program does not write trees to files.

For the transcriptomic dataset, phylogenetic congruence was assessed by computing quartet networks as implemented in SuperQ (Grünewald et al. 2013). Resulting networks were visualized in SplitsTree (Huson and Bryant 2006).

## Functional Annotation of Orthologous

To further document these orthology hypotheses, functional annotation was performed for the set of orthologs shared by the nodes representing the intersection of Araneoidea, Tetragnathidae, and Araneidae using the SFAnnotation pipeline (Yu and Kim 2014) with the "speedup" flag.

## Tree Annotation of Orthologs

With the purpose of representing the distribution of orthologs on the tree, a custom Python script (distOrth.py) was devised to "map" orthogroups onto trees in a similar fashion as in Laumer et al. (2015); but our implementation is not aiming to map decisive, *sensu* Steel and Sanderson (2010), taxon sets for specific nodes.

For each terminal node $t$, the composition of orthologs $G_t$ is given simply by counting all the orthogroups in

which the terminal is represented by at least one sequence $G_t = \{G_i, \ldots, G_n\}$ where $t \in G_{i,n}$.

The composition of orthologs of a particular split $S$ is defined as the collection of orthogroups represented in all terminals present in that split. $G_S = \cup G_t$ for every $t \in S$.

Finally, for internal nodes $n$, this set is calculated as the intersection of the orthogroups composition of all splits ($G_S$) incident to $n$, $G_n = \cap G_S$ for every $G_S$ where $n \notin S$.

This operation ensures that the orthologs mapped to any given internal node are represented by at least one species in each of the descendant branches and at least in one outgroup. The comparison of sequence composition of these sets can be used to explore sequence variation at specific hierarchical levels.

This operation assumes that (1) node composition assignment is independent of the root, and (2) missing a given ortholog on a leaf does not constitute evidence of its absence (it is therefore predictive). The goal of such annotation effort is to guide the selection of orthologs for future surveys.

It must be noted that the number of orthologs that map to a given node does not imply that the these orthologs support the grouping but only that contain sequences from species in the branches incident to that node.

Finally "untrimmed" nucleotide sequences of orthologs of interest are desirable for a variety of analyses. Nucleotide versions of the selected orthologs were retrieved from the reference NT file. This NT reference was produced by concatenating the cds files created along the amino acid ones by Transdecoder. The seamless transition from AA to NT is made possible because the sequence names in AA and NT version of the reference dataset are identical. These NT orthologs were then aligned using TranslatorX (Abascal et al. 2010) using the same AA alignment used in the orthology assessment thus preserving the reading frame.

## Concluding Remarks

We demonstrate the utility of UPhO in discovering orthologous groups from gene trees in the absence of a reference species tree. Orthologs detected by this new method can be used in phylogenetics and to outline stable orthology hypotheses. Nevertheless, the influence of initial homology assessment strategy (clustering) on downstream analyses must be noted; as it affects the estimation of gene trees and therefore orthology evaluation. This finding should raise a flag to question "automated" orthology methods that treat all datasets similarly regardless of the complexity of the problem. Initial data exploration should be carried out in order to select parameters and procedures better suited to the data and the research question. For many methods, such parameter exploration is discouraged by the computational burden of independent runs. In our pipeline, computational bottlenecks are reduced to initial BLAST searches (all vs. all) and gene family tree inference. But once these steps are completed these outputs can be processed under a variety of parameters and methods included but not limited to cluster refinement, filtering and the different options available for the orthology assessment.

The orthologs recognized can be directly used to infer the phylogeny of the reference organisms; however, it is not clear how our method compares with other phylogenomic approaches for which additional tests using empirical datasets would be required. Our method can rapidly populate genomic scale datasets for phylogenetic analyses by favoring the effort on the exploration of evolutionary relatedness of genes rather than the exhaustive pairwise comparisons of other approaches.

Tree-based orthology methods rely heavily on the quality of the input trees, which in turn depend on alignment accuracy, tree estimation parameters, and methods. Phylogenetic estimation and multiple sequence alignment algorithms are continuously being improved toward faster and more accurate estimates. Larger datasets, with more species and more complete genomic data (longer reads, higher coverage, etc.), will inevitably increase the amounts of evidence and our confidence in drawing stable orthology hypotheses for a variety of organisms.

Although the current trend in molecular phylogenetics seems to move toward big-data, not all phylogenetic problems require genomic scale datasets to reach robust solutions. The ability of target-specific orthologous *loci* with informativeness at a desired hierarchical level is in the long run more valuable than blindly sequencing random scattered regions of the genome.

The series of events leading to the observed genetic diversity in any collection of organisms are likely to be obscured by many biological factors such as gene duplication, deletions, mutation, recombination, etc., in addition to confounding factors in our methods (sampling error, fragmentary data, sequencing and assembly errors, etc.). The entangled evolutionary history of any collection of gene copies for a given group of organisms is unlikely to be grasped on by any single survey or method and thus populating and documenting orthology hypotheses for continuous testing is a central task in closing the gap of phylogenomic resources in non-model organisms.

## Note Added in Proof

Shortly after this paper was available on line, we were contacted by Casey Dunn, who pointed out to us that the algorithms and scripts used in Hejnol et al. (2009), Dunn (2013), and Yang and Smith (2014) (Agalma-like methods here after) are neither clade-based nor derived from previous clade-based methods (such as Zmazek and Eddy 2002; Chiu 2006; Gabaldon 2008). Instead, they are based on unrooted subtrees. Thus, Agalma-like methods are not sensitive to the position of the root in the input trees.

Although both methods are similar, UPhO and Agalma-like methods differ in the outcome of their orthology evaluation and further testing is required to compare their properties. The Agalma-like methods prune away only the largest subtree (or largest subtrees, if they have the same size), and then iterate this process on the remaining of the tree until it too has no more than one sequence per species. UPhO, on the other hand, prunes away all the subtrees satisfying the

minimum species threshold on a single pass, not only the largest subtrees.

## Supplementary Material

Supplementary figures S1–S10 and tables S1–S4 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

## Acknowledgments

Thanks to the GWU Computational Biology Institute and in particular to Adam Wong, who helped in the implementation of many of the packages and libraries used in this study. The GWU Colonial One computer cluster provided computing time for the analyses in this study. Lily Hughes, João Tonini and Andrew Thompson helped testing several versions of the scripts used in this study, providing substantial feedback. We thank the GWU Systematics Discussion Group and Rosa Fernández for comments on some aspects of the pipeline. Gonzalo Giribet, Kevin M. Kocot and Guillermo Ortí provided helpful comments, reviews and suggestions on an early versions of this manuscript. We are grateful to the editor and the two anonymous reviewers for their thoughtful comments and suggestions. This research was funded by National Science Foundation grants DEB1144492, DEB114417, DEB1457300, and DEB1457539 to G.H. and Gonzalo Giribet (H.U.). Additional support to J.A.B. was provided by GWU's Harlan Summer Fellowship and a Weintraub Graduate Fellowship.

## References

Abascal F, Zardoya R, Telford M. 2010. TranslatorX: multiple alignment of nucleotide sequences guided by amino acid translations. *Nucleic Acids Res.* 38(Suppl. 2):W7–W13.

Afrasiabi C, Samad B, Dineen D, Meacham C, Sjölander K. 2013. The PhyloFacts FAT-CAT web server: ortholog identification and function prediction using fast approximate tree classification. *Nucleic Acids Res.* 41:W242–W248.

Agnarsson I. 2010. The utility of ITS2 in spider phylogenetics: notes on prior work and an example from Anelosimus. *J Arachnol.* 38(2): 377–382.

Agnarsson I, Coddington J, Kuntner M. 2013. Systematics: progress in the study of spider diversity and evolution. In: Penny D, editor. Spider research in the 21st century: trends & perspectives. Castleton: Siri Scientific Press, p. 58–109.

Altenhoff A, Dessimoz C. 2009. Phylogenetic and functional assessment of orthologs inference projects and methods. *PLoS Comput Biol.* 5(1): e1000262.

Altschul S, Gish W, Miller W. 1990. Basic local alignment search tool. *J Mol Biol.* 215:403–410.

Andrade S, Montenegro H, Strand M, Schwartz M, Kajihara H, Norenburg J, Turbeville J, Sundberg P, Giribet G. 2014. A Transcriptomic approach to ribbon worm systematics (Nemertea): resolving the *Pilidiophora* problem. *Mol Biol Evol.* 31(12): 3206–3215.

Bazinet A, Cummings M, Mitter K, Mitter C. 2013. Can RNA-Seq resolve the rapid radiation of advanced moths and butterflies (Hexapoda: Lepidoptera: Apoditrysia)? An exploratory study. *PLoS One* 8(12): e82615.

Beach D. 2014 The global invertebrate genomics alliance (GIGA): developing community resources to study diverse invertebrate genomes. *J Hered.* 105(1): 1–18.

Bidegaray-Batista L, Arnedo M. 2011. Gone with the plate: the opening of the Western Mediterranean basin drove the diversification of ground-dweller spiders. *BMC Evol Biol.* 11(1): 317.

Blackledge T, Scharff N, Coddington J, Szüts T, Wenzel J, Hayashi C, Agnarsson I. 2009. Reconstructing web evolution and spider diversification in the molecular era. *Proc Natl Acad Sci U S A.* 106(13): 5229–5234.

Bond J, Garrison N, Hamilton C, Godwin R, Hedin M, Agnarsson I. 2014. Phylogenomics resolves a spider backbone phylogeny and rejects a prevailing paradigm for orb web evolution. *Curr Biol.* 24: 1765–1771.

Bourque M. 1978. Arbres de Steiner et réseaux dont certains sommets sont à localisation variable. [Ph. d. thesis]. [Canada]: Université de Montréal. Montréal.

Boussau B, Szöllosi G, Duret L, Gouy M, Tannier E, Daubin V. 2013. Genome-scale coestimation of species and gene trees. *Genome Res.* 23(2): 323–330.

Breinholt J, Kawahara A. 2013. Phylotranscriptomics: saturated third codon positions radically influence the estimation of trees based on next-gen data. *Genome Biol Evol.* 5(11): 2082–2092.

Brewer M, Carter R, Croucher P, Gillespie R. 2015. Shifting habitats, morphology, and selective pressures: developmental polyphenism in an adaptive radiation of Hawaiian spiders. *Evolution* 69(1): 162–178.

Brewer M, Cotoras D, Croucher PJ, Gillespie R. 2014. New sequencing technologies, the development of genomics tools, and their applications in evolutionary arachnology. *J Arachnol.* 42(1): 1–15.

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

Campbell L, Rota-Stabelli O, Edgecombe G, Marchioro T, Longhorn S, Telford M, Philippe H, Rebecchi L, Peterson K, Pisani D. 2011. MicroRNAs and phylogenomics resolve the relationships of Tardigrada and suggest that velvet worms are the sister group of Arthropoda. *Proc Natl Acad Sci U S A.* 108(38): 15920–15924.

Capella-Gutiérrez S, Silla-Martínez J, Gabaldón T. 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* 25(15): 1972–1973.

Chen A, Fand M, Vermunt J, Roos D. 2007. Assessing performance of orthology detection strategies applied to eukaryotic genomes. *PLoS One* 2(4): e383.

Chiu J, Lee E, Egan M, Sarkar I, Coruzzi G, DeSalle R. 2006. OrthologID: automation of genome-scale ortholog identification within a parsimony framework. *Bioinformatics* 22(6): 699–707.

Cock P, Antao T, Chang J, Chapman B, Cox C, Dalke A, Friedberg I, Hamelryck T, Kauff B, F and Wilczynski, et al. 2009. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* 25(11): 1422–1423.

Croucher P, Brewer M, Winchell C, Oxford G, Gillespie R. 2013. De novo characterization of the gene-rich transcriptomes of two color-polymorphic spiders, *Theridion grallator* and *T. californicum* (Araneae: Theridiidae), with special reference to pigment genes. *BMC Genomics* 14(1): 862.

Delsuc F, Brinkmann H, Philippe H. 2005. Phylogenomics and the reconstruction of the tree of life. *Nat Rev Genet.* 6(5): 361–375.

Dimitrov D, Lopardo L, Giribet G, Arnedo M, Alvarez-Padilla F, Hormiga G. 2012. Tangled in a sparse spider web: single origin of orb weavers and their spinning work unravelled by denser taxonomic sampling. *Proc R Soc Lond B Biol.* 279(1732): 1341–1350.

Douzery E, Scornavacca C, Romiguier J, Belkhir K, Galtier N, Delsuc F, Ranwez V. 2014. OrthoMaM v8: a database of orthologous exons and coding sequences for comparative genomics in mammals. *Mol Biol Evol.* 31(7): 1923–1928.

Dunn C, Howison M, Zapata F. 2013. Agalma: an automated de novo transcriptome assembly pipeline. *BMC Bioinformatics* 14: 330.

Enright A, Van Dongen S, Ouzounis C. 2002. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.* 30(7): 1575–1584.

Evans J, Brown S, Hackett K, Robinson G, Richards S, Lawson D, Elsik C, Coddington J, Edwards O, Emrich S, et al. 2013. The i5K initiative:

advancing arthropod genomics for knowledge, human health, agriculture, and the environment. *J Hered*. 104(5): 595–600.
Fernández R, Hormiga G, Giribet G. 2014a. Phylogenomic analysis of spiders reveals nonmonophyly of orb weavers. *Curr Biol*. 24: 1772–1777.
Fernández R, Laumer C, Vahtera V, Libro S, Kaluziak S, Sharma P, Pérez-Porro A, Edgecombe G, Giribet G. 2014b. Evaluating topological conflict in centipede phylogeny using transcriptomic data sets. *Mol Biol Evol*. 31(6): 1500–1513.
Fitch W. 1970. Distinguishing homologous from analogous proteins. *Syst Biol*. 19(2): 99–113.
Fu L, Niu B, Zhu Z, Wu S, Li W. 2012. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* 28(23): 3150–3152.
Gabaldón T. 2008. Large-scale assignment of orthology: back to phylogenetics? *Genome Biol*. 9(10): 235.
Gabaldón T, Koonin E. 2013. Functional and evolutionary implications of gene orthology. *Nat Rev Genet*. 14: 360–366.
Grabherr M, Haas B, Yassour M, Levin J, Thompson D, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, et al. 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol*. 29(7): 644–652.
Grünewald S, Spillner A, Bastkowski S, Bögershausen A Moulton V, 2013 SuperQ: computing supernetworks from quartets. *IEEE/ACM Trans Comput Biol Bioinform*. 10(1): 151–160.
Haas B, Papanicolaou A, Yassour M, Grabherr M, Blood P, Bowden J, Couger M, Eccles D, Li B, Lieber M, et al. 2013. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat Protoc*. 8(8): 1494–1512.
Hansen K, Brenner S, Dudoit S. 2010. Biases in Illumina transcriptome sequencing caused by random hexamer priming. *Nucleic Acids Res*. 38(12): e131.
He Q, Duan Z, Yu Y, Liu Z, Liu Z, Liang S. 2013. The venom gland transcriptome of *Latrodectus tredecimguttatus* revealed by deep sequencing and cDNA library analysis. *PLoS One* 8(11): e81357.
Hedtke S, Morgan M, Cannatella D, Hillis D. 2013. Targeted enrichment: maximizing orthologous gene comparisons across deep evolutionary time. *PLoS One* 8(7): e67908.
Hejnol A, Obst M, Stamatakis A, Ott M, Rouse G, Edgecombe G, Martinez P, Baguñà J, Bailly X, Jondelius U, et al. 2009. Assessing the root of bilaterian animals with scalable phylogenomic methods. *Proc R Soc Lond B Biol*. 276(1677): 4261–4270.
Hormiga G, Griswold C. 2014. Systematics, phylogeny, and evolution of orb-weaving Spiders. *Annu Rev Entomol*. 59: 487–512.
Huerta-Cepas J, Dopazo J, Gabaldón T. 2010. ETE: a Python environment for tree exploration. *BMC Bioinformatics* 11: 24.
Huson D, Bryant D. 2006. Application of phylogenetic networks in evolutionary studies. *Mol Biol Evol*. 23(2): 254–267.
Jarvis E, Mirarab S, Aberer A, Li B, Houde P, Li C, Ho S, Faircloth B, Nabholz B, Howard J, et al. 2014. Whole-genome analyses resolve early branches in the tree of life of modern birds. *Science* 346: 1320–1331.
Katoh K, Standley D. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol*. 30(4): 772–780.
Kocot K, Cannon J, Todt C, Citarella M, Kohn A, Meyer A, Santos S, Schander C, Moroz L, Lieb B, et al. 2011. Phylogenomics reveals deep molluscan relationships. *Nature* 477(7365): 452–456.
Kocot K, Citarella M, Moroz L, Halanych K. 2013. PhyloTreePruner: a phylogenetic tree-based approach for selection of orthologous sequences for phylogenomics. *Evol Bioinform*. 9: 429–435.
Kristensen D, Wolf Y, Mushegian A, Koonin E. 2011. Computational methods for gene orthology inference. *Brief Bioinform*. 12(5): 379–391.
Kumar S, Filipski A, Battistuzzi F, Pond S, Tamura K. 2012. Statistics and truth in phylogenomics. *Mol Biol Evol*. 29(2): 457–472.
Laumer C, Hejnol A, Giribet G. 2015. Nuclear genomic signals of the "microturbellarian" roots of platyhelminth evolutionary innovation. *eLife* 4: e05503.
Lemmon A, Emme S, Lemmon E. 2012. Anchored hybrid enrichment for massively high-throughput phylogenomics. *Syst Biol*. 61(5): 727–744.
Lemmon E, Lemmon A. 2013. High-throughput genomic data in systematics and phylogenetics. *Annu Rev Ecol Evol Syst*. 44(1): 99–121.
Lopardo L, Giribet G, Hormiga G. 2011. Morphology to the rescue: molecular data and the signal of morphological characters in combined phylogenetic analyses—a case study from mysmenid spiders (Araneae, Mysmenidae), with comments on the evolution of web architecture. *Cladistics* 27: 278–330.
Marcet-Houben M, Gabaldón T. 2011. TreeKO: a duplication-aware algorithm for the comparison of phylogenetic trees. *Nucleic Acids Res*. 39(10): e66.
Margulies M, Egholm M, Altman W, Attiya S, Bader J, Bemben L, Berka J, Braverman M, Chen Y, Chen Z, et al. 2006. Genome sequencing in open microfabricated high density picoliter reactors. *Nat Biotechnol*. 437(7057): 376–380.
Matlin A, Clark F, Smith C. 2005. Understanding alternative splicing: towards a cellular code. *Nat Rev Mol Cell Biol*. 6(5): 386–398.
Mattila T, Bechsgaard J, Hansen T, Schierup M, Bilde T. 2012. Orthologous genes identified by transcriptome sequencing in the spider genus *Stegodyphus*. *BMC Genomics* 13(1): 70.
McCormack J, Faircloth B, Crawford N, Gowaty P, Brumfield R, Glenn T. 2012. Ultraconserved elements are novel phylogenomic markers that resolve placental mammal phylogeny when combined with species-tree analysis. *Genome Res*. 22: 746–754.
Mirarab S, Reaz R, Bayzid MS, Zimmermann T, Swenson M, Warnow T. 2014. ASTRAL: genome-scale coalescent-based species tree estimation. *Bioinformatics* 30(17): i541–i548.
Misof B, Liu S, Meusemann K, Peters R, Donath A, Cristoph M, Flouri T, Beutel R, Niehuis O, Petersen M. 2014. Phylogenomics resolves the timing and pattern of insect evolution. *Science* 346(6210): 763–768.
Morgan M, Anders S, Lawrence M, Aboyoun P, Pagès H, Gentleman R. 2009. ShortRead: a bioconductor package for input, quality assessment and exploration of high-throughput sequence data. *Bioinformatics* 25(19): 2607–2608.
Novo M, Riesqo A, Fernández-Guerra A Giribet G, Ferna A, 2013. Pheromone evolution, reproductive genes, and comparative transcriptomics in mediterranean earthworms (Annelida, Oligochaeta, Hormogastridae). *Mol Biol Evol*. 30(7): 1614–1629.
Oakley T, Wolfe J, Lindgren A, Zaharoff A. 2013. Phylotranscriptomics to bring the understudied into the fold: monophyletic ostracoda, fossil placement, and pancrustacean phylogeny. *Mol Biol Evol*. 30(1): 215–233.
Pearson T, Hornstra H, Sahl J, Schaack S, Schupp J, Beckstrom-Sternberg S, O'Neill M, Priestley R, Champion M, Beckstrom-Sternberg J, et al. 2013. When outgroups fail; phylogenomics of rooting the emerging pathogen, *Coxiella burnetii*. *Syst Biol*. 62(5): 752–762.
Pearson W. 2013. An introduction to sequence similarity ("homology") searching. *Curr Protoc Bioinform*. 42(3.1): 1–8.
Price M, Dehal P, Arkin A. 2010. FastTree 2—approximately maximum-likelihood trees for large alignments. *PLoS One* 5(3): e9490.
Ramazzotti M, Berná L, Stefanini I, Cavalieri D. 2012. A computational pipeline to discover highly phylogenetically informative genes in sequenced genomes: application to *Saccharomyces cerevisiae* natural strains. *Nucleic Acids Res*. 40(9): 3834–3848.
Remm M, Storm C, Sonnhammer E. 2001. Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *J Mol Biol*. 314(5): 1041–1052.
Richards S. 2015. It's more than stamp collecting: how genome sequencing can unify biological research. *Trends Genet*. 31(7): 411–421.
Robinson D, Foulds L. 1979. Comparison of weighted labelled trees. In: Horadam A, Wallis W, editors. Combinatorial mathematics VI: proceedings of the sixth Australian conference on combinatorial mathematics, Armidale, Australia, August 1978. Berlin: Springer, p. 119–126.
Roth A, Gonnet G, Dessimoz C. 2008. Algorithm of OMA for large-scale orthology inference. *BMC Bioinformatics* 9: 518.
Salichos L, Rokas A. 2011. Evaluating ortholog prediction algorithms in a yeast model clade. *PLoS One* 6(4): e18755.
Sammeth M, Foissac S, Guigó R. 2008. A general definition and nomenclature for alternative splicing events. *PLoS Comput Biol*. 4(8): e1000147.

Schliep K. 2011. phangorn: phylogenetic analysis in R. *Bioinformatics* 27(4): 592–593.
Sennblad B, Lagergren J. 2009. Probabilistic orthology analysis. *Syst Biol*. 58(4): 411–424.
Sharma P, Fernández R, Esposito L, González-Santillán E, Monod L. 2015. Phylogenomic resolution of scorpions reveals multilevel discordance with morphological phylogenetic signal. *Proc R Soc Lond B Biol*. 282(1804): 20142953.
Sharma P, Kaluziak S, Pérez-Porro A, González V, Hormiga G, Wheeler W, Giribet G. 2014. Phylogenomic interrogation of Arachnida reveals systemic conflicts in phylogenetic signal. *Mol Biol Evol*. 31(11): 2963–2984.
Smith S, Wilson N, Goetz F, Feehery C, Andrade S, Rouse G, Giribet G, Dunn C. 2011. Resolving the evolutionary relationships of mollusks with phylogenomic tools. *Nature* 480(7377): 364–367.
Sonnhammer E, Gabaldón T, Sousa da Silva A, Martin M, Robinson-Rechavi M, Boeckmann B, Thomas P, Dessimoz C, the Quest for Orthologs consortium. 2014. Big data and other challenges in the quest for orthologs. *Bioinformatics* 30(21): 2993–2998.
Sonnhammer E, Östlund G. 2014. InParanoid 8: orthology analysis between 273 proteomes, mostly eukaryotic. *Nucleic Acids Res*. 43(D1): D234–D239.
Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30(9): 1312–1313.
Steel M, Sanderson M. 2010. Characterizing phylogenetically decisive taxon coverage. *Appl Math Lett*. 23(1): 82–86.
Tange O. 2011. GNU parallel—the command-line power tool. *login: The USENIX Magazine* 36(1): 42–47.
Telford M, Copley R. 2011. Improving animal phylogenies with genomic data. *Trends Genet* 27(5): 186–195.
Thornton J, DeSalle R. 2000. Gene family evolution and homology. *Annu Rev Genomics Hum Genet*. 1(48): 41–73.
Trachana K, Larsson T, Powell S, Chen W, Doerks T, Muller J, Bork P. 2011. Orthology prediction methods: a quality assessment using curated protein families. *Bioessays* 33(10): 769–780.
Ullah I, Sjöstrand J, Andersson P, Sennblad B, Lagergren J. 2015. Integrating sequence evolution into probabilistic orthology analysis. *Syst Biol*. 64(6): syv044.
van der Heijden R, Snel B, van Noort V, Huynen M. 2007. Orthology prediction at scalable resolution by phylogenetic tree analysis. *BMC Bioinformatics* 8: 83.
van Dongen S. 2000. Graphs clustering by flow simulation. [Ph. d. thesis]. [Utrecht]: University of Utrecht.
Whelan N, Kocot K, Moroz L, Halanych K. 2015. Error, signal, and the placement of Ctenophora sister to all other animals. *Proc Natl Acad Sci U S A*. 112(18): 5773–5778.
Yang Y, Smith S. 2014. Orthology inference in non-model organisms using transcriptomes and low-coverage genomes: improving accuracy and matrix occupancy for phylogenomics. *Mol Biol Evol*. 31(11): 3081–3092.
Yim K, Brewer M, Miller C, Gillespie R. 2014. Comparative transcriptomics of maturity-associated color change in Hawaiian spiders. *J Hered*. 105(S1): 771–781.
Yu C, Zavaljevski N, Desai V, Reifman J. 2011. QuartetS: a fast and accurate algorithm for large-scale orthology detection. *Nucleic Acids Res*. 39(13): e88.
Yu D, Kim B. 2014. SFannotation: a simple and fast protein function annotation system. *Genomics Inf*. 12(2): 76–78.
Zhao Y, Zeng Y, Chen L, Dong Y, Wang W. 2014. Analysis of transcriptomes of three orb-web spider species reveals gene profiles involved in silk and toxin. *Insect Sci*. 21(6): 687–698.
Zmasek C, Eddy S. 2002. RIO: analyzing proteomes by automated phylogenomics using resampled inference of orthologs. *BMC Bioinformatics* 3: 14.