PLOS | ONE

RESEARCH ARTICLE

Comparative transcriptomics of Entelegyne spiders (Araneae, Entelegynae), with emphasis on molecular evolution of orphan genes

David E. Carlson[1,2]*, **Marshal Hedin**[1]

1 Department of Biology, San Diego State University, San Diego, California, United States of America, **2** Department of Ecology & Evolution, Stony Brook University, Stony Brook, New York, United States of America

* david.carlson@stonybrook.edu

Abstract

Next-generation sequencing technology is rapidly transforming the landscape of evolutionary biology, and has become a cost-effective and efficient means of collecting exome information for non-model organisms. Due to their taxonomic diversity, production of interesting venom and silk proteins, and the relative scarcity of existing genomic resources, spiders in particular are excellent targets for next-generation sequencing (NGS) methods. In this study, the transcriptomes of six entelegyne spider species from three genera (*Cicurina travisae*, *C. vibora*, *Habronattus signatus*, *H. ustulatus*, *Nesticus bishopi*, and *N. cooperi*) were sequenced and *de novo* assembled. Each assembly was assessed for quality and completeness and functionally annotated using gene ontology information. Approximately 100 transcripts with evidence of homology to venom proteins were discovered. After identifying more than 3,000 putatively orthologous genes across all six taxa, we used comparative analyses to identify 24 instances of positively selected genes. In addition, between ~ 550 and 1,100 unique orphan genes were found in each genus. These unique, uncharacterized genes exhibited elevated rates of amino acid substitution, potentially consistent with lineage-specific adaptive evolution. The data generated for this study represent a valuable resource for future phylogenetic and molecular evolutionary research, and our results provide new insight into the forces driving genome evolution in taxa that span the root of entelegyne spider phylogeny.

OPEN ACCESS

Citation: Carlson DE, Hedin M (2017) Comparative transcriptomics of Entelegyne spiders (Araneae, Entelegynae), with emphasis on molecular evolution of orphan genes. PLoS ONE 12(4): e0174102. https://doi.org/10.1371/journal.pone.0174102

Editor: Matjaž Kuntner, Scientific Research Centre of the Slovenian Academy of Sciences and Art, SLOVENIA

Received: October 15, 2016

Accepted: March 4, 2017

Published: April 5, 2017

Data Availability Statement: All raw read fastq files are available from the NCBI Sequence Read Archive (accession numbers SRX761208, SRX652499, SRX652504, SRX763246, SRX761332, and SRX652508). In addition, the six transcriptome assemblies produced for this manuscript have been deposited to the Dryad Digital Repository under the following DOI: doi:10.5061/dryad.3pq3q. The files may be viewed at the following link: http://datadryad.org/review?doi=doi:10.5061/dryad.3pq3q.

Introduction

The rise of high throughput sequencing technologies (also known as next-generation sequencing, hereafter NGS) has created new research opportunities in many fields of biology, including evolutionary biology and systematics (e.g., [1,2]). For non-model organisms, shotgun sequencing of a transcriptome can be a useful and cost-effective means of gaining insight into genome-wide biological processes ([3]). One way that transcriptomic data have been used to

Funding: This study was funded by a grant to MH from the United States Fish and Wildlife Service (contract #F13PX00770). https://www.fws.gov/. This study was also funded by support to MH from the San Diego State University Grants Program (https://newscenter.sdsu.edu/researchaffairs/ugp.aspx). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

study molecular and organismal evolution is through comparative analyses of sequences from multiple taxa. Comparative transcriptomics has been used, for example, to detect positive selection in genomes ([4,5]), estimate transcriptome-wide rates of molecular evolution ([6]), and resolve difficult phylogenetic questions (e.g., [2,7–9]).

Another area of research that has greatly benefited from the availability of NGS data is the study of orphan genes. Orphan genes are genes that are detected only in a particular lineage and lack recognizable protein-coding homologs in other taxa ([10]). The relative ease with which genomic data can be collected for non-model organisms is expanding our ability to identify and analyze the evolutionary dynamics of orphan genes ([11–13]). These taxonomically restricted genes are often thought to be implicated in lineage-specific adaptation to unique environmental conditions ([10]) and are thus of major interest to researchers. While examples of likely adaptive orphan genes have been identified ([14–16]), generalizations regarding the origin and maintenance of orphan genes have heretofore been limited by the availability of genomic or transcriptomic data.

Entelegynae is a clade of spiders that have traditionally been united by a number of shared derived morphological features, such as highly modified and complex male pedipalps ([17]). With more than 38,000 described species ([18]), the entelegynes include many of the most species-rich spider families, such as wolf spiders (Lycosidae), jumping spiders (Salticidae), sheat weavers (Linyphiidae), and orb weavers (Araneidae). Substantial research effort has focused on elucidating entelegyne phylogeny (e.g., [9,19–22]). Two of the largest entelegyne clades include the "RTA clade," where males possess a palpal knob called the retrolateral tibial apophysis, and the Araneoidea, which consists of the ecribellate orb weavers and kin ([17,18]). Phylogenomic data estimate the age of Entelegynae as roughly 154–290 MA, the age of Araneoidea as 114–233 MA, and the age of the RTA clade as 83–201 MA (see Fig 1**; [9**]). Although spiders produce many unique silk and venom proteins, available genomic resources for spiders—and arachnids more broadly—remain limited and are not commensurate with the rich taxonomic diversity of this clade. While the first two spider genomes have been published ([23]) and spider-specific orthologous gene models are now available ([9]), large-scale, well-annotated genetic data remain unavailable for most spider taxa. As such, members of this understudied taxonomic group make excellent subjects for both molecular evolutionary research and increased DNA sequencing efforts.

In recent years, a number of studies utilizing NGS data have aimed to reveal information about the evolution of entelegyne spiders; however, much of this research focused on the evolution of a narrow set of proteins. For example, multiple transcriptomic studies characterizing spider venom [24–28]) and silk proteins ([29–32]) have been published. Conversely, studies that expand their scope to a broader array of genomic sequences have focused primarily on a fairly small phylogenetic range of taxa. Comparative transcriptomics have been used, for example, to investigate the evolution of sociality in three *Stegodyphus* species ([33]), and patterns of selection on ecomorphological diversification in Hawaiian *Tetragnatha* ([34,35]). Only very recently (e.g., [36]) have researchers begun using NGS data to examine patterns of genome-wide molecular evolution across divergent entelegyne taxa.

Six species representing three congeneric pairs of entelegyne genera were chosen for transcriptome assembly and analysis. These genera represent both the Araneoidea and RTA clades, and differ in ecology. *Nesticus* (Nesticidae) and *Cicurina* (Dictynidae) prefer sheltered microhabitats, such as in caves or under rocks in forests ([37,38]), while *Habronattus* jumping spiders (Salticidae) are active open-habitat visual hunters ([39,40]). *Cicurina* and *Habronattus* are both members of the RTA clade, while *Nesticus* is found within Araneoidea. The use of these relatively divergent taxa, which span the root of the entelegyne phylogeny, in conjunction with our congeneric pair sampling strategy enabled investigation of patterns of transcriptome

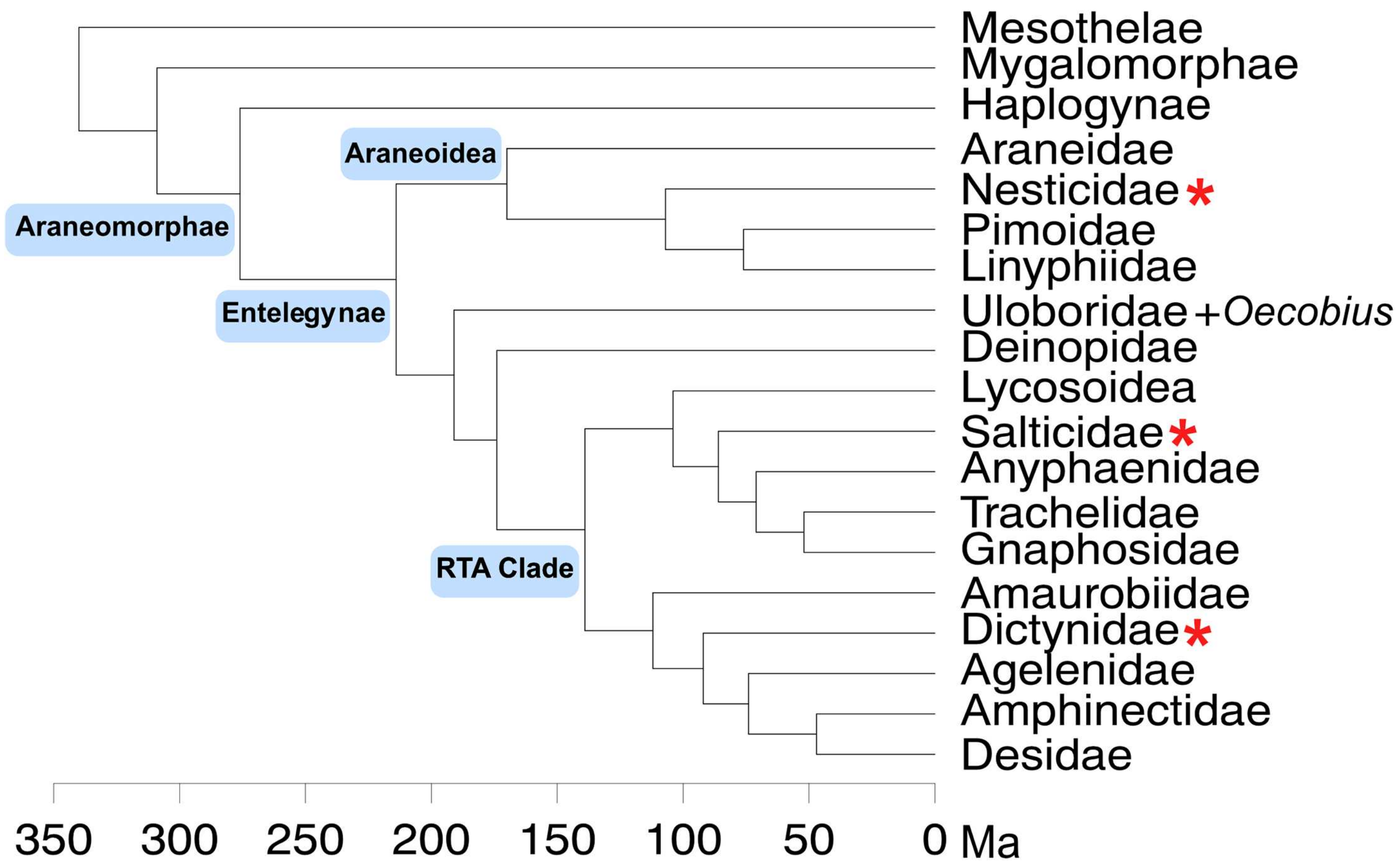

Fig 1. Time-calibrated backbone spider phylogeny. Modified from [9], this phylogeny shows the placement of major clades. Families containing taxa used in this study are highlighted with red asterisks.

https://doi.org/10.1371/journal.pone.0174102.g001

composition and rates of molecular evolution both within and among entelegyne spider lineages. After *de novo* assembling each transcriptome, assessing the quality and completeness of each assembly, and performing functional annotation, we estimated substitution rates among putatively-orthologous genes in order to detect signatures of positive selection. In addition, we used the presence of novel coding sequences shared among congeneric pairs to identify lineage-specific orphan genes as well as undescribed genes likely present in a variety of entelegyne species. Our results provide evidence for a relatively low rate of positive selection among more than 3,300 single-copy orthologs. In addition, we find that many of the unique, uncharacterized orphan genes in each transcriptome exhibit elevated rates of substitution likely due to a combination of relaxed functional constraint in some cases and positive selection in others. These results are consistent with the hypothesis that some orphan genes are involved in lineage-specific adaptation.

Materials and methods

RNA extraction and sequencing

RNA was extracted from field-collected specimens preserved in RNAlater using RNeasy minikits (Qiagen, Valencia, CA), and shipped to the Genomic Services Lab at HudsonAlpha (www.hudsonalpha.org) where non-normalized cDNA libraries were sequenced on an Illumina Hiseq 2000 platform. As detailed in [41], cave-dwelling *Cicurina* were collected under a permit from the US Fish & Wildlife Service, issued to Cyndee Watson. *Habronattus* and *Nesticus* samples were collected from roadside right-of-way habitats in California and North Carolina where specific permissions were not required. *Habronattus* and *Nesticus* specimens were sequenced using 50 bp paired-end reads, while *Cicurina* specimens were sequenced using 100 bp paired-end reads. S1 Table summarizes the specimens/tissues used in each RNA extraction and the number of raw Illumina reads generated for each species.

De novo transcriptome assembly

FastQC version 0.10.1 ([42]) was used to calculate read quantity, Phred scores, and to confirm quality of the raw reads prior to assembly. The perl script Trim Galore! (http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/) was used to remove low quality sequences and trim Illumina adaptors. Sequences less than 30 base pairs in length, as well as those consisting of greater than one percent ambiguity characters, were removed with PRINSEQ Lite ([43]). Raw reads passing quality control were assembled *de novo* into transcripts using Trinity ([44]) with the default k-mer size of 25 and minimum transcript length set to 200 bp. Briefly, Trinity uses overlapping k-mers to assemble contigs, which are grouped together into clusters likely stemming from alternatively spliced isoforms of the same gene or products of closely related paralogs. Then, De Bruijn graphs are built for each cluster and used to reconstruct individual transcripts for each isoform and paralogous gene ([44]).

DeconSeq ([45]) was used to remove likely contaminant sequences. Reads were mapped back to assembled transcripts using Bowtie ([46]) in RSEM ([47]) to estimate abundance levels. To remove likely assembly errors, low abundance transcript outliers (i.e., individual isoform transcripts whose expression levels represented less than one percent of the overall expression for that gene) were filtered out of the assembly.

Assembly summary statistics and completeness assessment

After quality filtering, Trinity and the python script sizecutter.py ([48]) were used to calculate a variety of assembly summary statistics. Read coverage depth was calculated with BEDtools using the genomeCoverageBed program ([49]). Read coverage is dependent on the level of expression for a given transcript, which varies greatly among sequences. Because of this, mean read coverage is expected to be inconsistent across different parts of the transcriptome ([44,50]). As such, we computed both the mean coverage and the proportion of each transcriptome with three different levels of read coverage.

The completeness and quality of each assembled transcriptome was evaluated using several procedures. The first measure was the Core Eukaryotic Genes Mapping Approach (CEGMA; [51,52]) which attempts to quantify the relative gene content of a particular set of sequences. This involves mapping assemblies to 248 core eukaryotic genes (CEGs) that have been found to be highly conserved and present in low copy number across six model eukaryotic organisms. To complement CEGMA analyses, we downloaded a recently published dataset consisting of hidden markov model (HMM) profiles from 4,934 spider-specific core ortholog groups ([9]). After translating putative open reading frames (ORFs) from each of the six transcriptomes into protein sequences using Trinity's TRANSDECODER utility, we searched the spider ortholog group set against our translated proteins. We then calculated the proportion of total HMMs that successfully matched at least one sequence in each of the assembled transcriptomes (at a max e-value threshold of 1 x 10^{-10}). Finally, we measured the percent of total gene length that was captured by each assembled transcript using the "ortholog hit ratio" (OHR; [53]). This metric compares the length of an assembled transcript to the full length of its top Blast hit. Under this method, an OHR close to zero indicates a poorly assembled transcript, an

OHR of one indicates that the full length of the gene was captured, while an OHR greater than one suggests the presence of one or more insertions in the transcript. In order to calculate OHR values, XML output files containing hits to the NCBI's non-redundant (hereafter nr) protein database (see below) for each species were processed with the ortholog_hit_ratio.py script ([54]), which divides the length of the hit in the query sequence by the total length of the subject sequence.

Functional annotation

To annotate assembled sequences, standalone Blastx searches of the nr protein database were performed using a maximum e-value threshold of 1 x10^{-5}, a high scoring segment-pair (HSP) length cutoff of greater than 33 and 20 maximum hits per sequence. Blastx results were then imported into Blast2GO (B2G; [55]), GO terms were mapped from Accession IDs, and sequences with at least one significant hit were annotated with GO terms using an e-value hit filter of 1x10^{-6}, an annotation cut off of 55 and a GO weight of 5. In addition, each transcriptome was annotated with GO terms using B2G's InterProScan ([56]) option, which assigns functional information to sequence data using protein domain and motif information. Each set of annotations was reduced to include a more manageable subset of higher level GO terms by mapping the annotations to a generic GO Slim ontology using the GO-Slim function.

Venom sequence analysis

To identify potential venom-associated proteins, we downloaded the Arachnoserver spider toxin database ([57]) and inferred homology between the sequences in this database and ORFs from each transcriptome. Blastx hits with > 50% sequence identity and query coverage as well as e-values less than 1 x10^{-15} were considered for further analysis. To this dataset we added successfully annotated sequences whose top blast hit to the nr protein database consisted of a toxin or venom protein. The translated polypeptide sequences corresponding to this list of transcripts were input into the Arachnoserver's SpiderP utility to look for evidence of signal peptides and propeptides. If one or both of these were found to be present in a sequence with homology to a known venom protein, this was interpreted as evidence that the transcript is a venom gene. Further functional investigation of these putative venom genes was accomplished using an enrichment analysis in B2G. GO terms present in successfully annotated venom genes were compared against the entire set of functionally annotated transcripts using Fisher's exact test to look for GO terms that are enriched in transcripts likely to function in venom production. P-values were corrected for multiple testing using Benjamini & Hochberg's false discovery rate (FDR) method ([58]).

Identification and alignment of orthologous genes

A set of translated ORFs from each assembly was input into InParanoid ([59,60]), which we used to infer orthologs among each pairwise combination of species. This method identifies orthologous coding sequence pairs based upon similarity scores from reciprocal protein Blast searches. Pairwise orthologs were then grouped into clusters containing one gene from each of the six species using QuickParanoid (available from http://pl.postech.ac.kr/QuickParanoid/). Occasionally, multiple isoforms from the same gene were present in a cluster. In such cases, all but one isoform was removed, leaving only one coding sequence per species in each cluster. Because InParanoid identifies orthologs as well as in-paralogs (i.e., genes that duplicated in one species following a speciation event; [60]), clusters containing more than one non-isoform sequence from any single species were systematically removed from the dataset and were not used in subsequent analyses.

The accuracy of downstream analyses can be negatively impacted by multiple sequence alignment (MSA) errors ([61–64]). Because of this, particular care was taken to minimize such errors. First, we used the phylogeny-aware aligner PRANK v.150803 ([65]) to produce a codon-based MSA for each set of orthologs. Next, each PRANK alignment was processed using GUIDANCE2 v.2.0.1 ([66]) with default settings to detect and remove unreliable alignment positions. Each alignment then received another round of filtering to remove gaps and poorly conserved position using Gblocks v.0.91b ([67]). As a final quality control measure, any MSAs that exhibited 60% pairwise identity, 40% identity across all sequences, or that were 100 bp in length were removed.

Tests for positive selection

To identify genes under positive selection, we estimated rates of substitution for each ortholog alignment using the codeml program in PAML v4.9a ([68]). Tests for positive selection were implemented using the branch-site model ([69,70]); this method allows ω—the ratio of the rate of non-synonymous substitutions per non-synonymous site (Ka) to the rate of synonymous substitutions per synonymous site (Ks)—to vary both among different lineages and at different sites in the sequence. A value of $\omega < 1$ is commonly accepted as evidence of purifying selection, while a Ka/Ks value greater than one may indicate positive selection [71]). A phylogeny representing the well-established relationships among the six species (e.g., [9]) was used in all selection analyses. The branch-site analyses required specification of a "foreground" branch in which ω is allowed to exceed 1, while ω is limited to a value ≤ 1 in the remaining "background" branches (Yang 2007). Nine branches, corresponding to each of the six species and each of the three genera, were chosen as foreground branches in successive analyses.

Each alternative model analysis was compared with a null model in which the foreground branch's ω value could not exceed one using a likelihood-ratio test (i.e., 2^*(Log Likelihood$_{\text{HypA}}$– Log Likelihood$_{\text{HypB}}$)). All analyses were replicated three times with different starting values to ensure that likelihood values reached global optima. Likelihood values were compared across replicate runs, and the run with the highest value was chosen for the LRT. X^2 test statistics resulting from LRTs were used to generate p-values, which were corrected for multiple comparisons using the FDR method ([58]) as implemented in the p.adjust function in R v3.0.2 ([72]).

For genes inferred to be under positive selection by the branch-site analysis, we used the Bayes Empirical Bayes (BEB; [69]) approach in codeml to identify the specific amino acids most likely under selection. This method accounts for uncertainty in the Maximum Likelihood estimates of parameters in order to calculate the posterior probability that each amino acid site in a sequence is included in the class of sites with $\omega > 1$. A particular amino acid was considered to have strong evidence for being under positive selection if the BEB posterior probability value was ≥ 0.95. To investigate possible functions of positively selected genes (PSGs), we retrieved GO terms for each of these genes and performed an enrichment analysis using B2G ([55]).

Orphan gene search

Orphan genes are putative protein-coding sequences that lack homologs in other taxa and may represent examples of lineage-specific adaptive evolution ([10]). Here, orphan genes were identified and validated by comparing potentially protein-coding transcripts that did not successfully hit to sequences in the nr database from one species with similar transcripts in its closely related congener. First, we used TRANSDECODER to find ORFs within transcripts that lacked a significant Blastx hit to the nr protein database. For genes with multiple isoforms, a single exemplar isoform (the longest) was selected for further analysis. The ORFs from each species were then searched using Tblastx against ORFs in their corresponding congeneric transcriptome that similarly lacked Blastx hits. Two transcripts were designated as putative orphan genes if they were reciprocal best hits of one another and the Tblastx results fulfilled the following criteria: a maximum e-value 1 x 10^{-10}, a minimum protein hit length of 67 amino acids (201 nucleotides) and minimum total HSP coverage of 40 percent of the query sequence length. The putatively novel genes discovered through this method were then searched against sequences from the two other genera to identify novel transcripts that are shared among a wider range of Entelegyne taxa. GO terms derived from Interproscan were retrieved for sequences identified as lineage-specific genes.

Because the MSA procedure applied to the orthologs is not applicable for pairwise alignments, we implemented an alternative method to align and edit congeneric orphan sequences. Each pair of orphan genes was translated into amino acids using T-Coffee v.11.00 ([73]), and the resulting translated sequences were aligned with M-Coffee ([74]) using the t_coffe_pair, mafft_pair, muscle_pair, and kalign_pair alignment algorithms. We then used T-Coffee's TCS procedure ([75]) to output a score ranging from 0 to 9 for each alignment position, depending on how consistently that position was aligned by the four different methods. In order to reduce the incidence of alignment errors, we filtered out alignment positions that received a TCS score lower than 8. Nucleotide MSAs were then generated from each filtered amino acid alignment using T-Coffee's seq_reformat program. As before, Gblocks v.0.91b ([67]) was used to remove gaps and poorly conserved positions from alignments. Only pairwise sequence alignments (PSA) $\geq$ 100 bp in length and with $\geq$ 60% identity were used in subsequent analyses.

If many putative orphan genes are actually examples of lineage specific adaptive evolution, we expect ω ratios for these sequences to be elevated compared to the genome-wide set of protein coding sequences for which there is no *a priori* expectation of positive selection. To test this, pairwise maximum likelihood estimates of Ka, Ks and ω were calculated for each set of congeneric orphan gene alignments using codeml (Runmode = -2). These evolutionary rate parameters were compared with similar estimates derived from each pairwise comparison of congeneric orthologs used in tests for positive selection. Significant differences in Ka, Ks and ω between these two datasets were assessed using nonparametric Mann-Whitney U tests implemented with the wilcox.test function in R [72]. We also clustered each set of ω estimates that ranged from 0.0 to 1.0 into three bins of equal size ($\omega \leq 0.33$, $0.33 < \omega \leq 0.67$, and $0.67 < \omega \leq 1.0$) and designated these as strong, moderate and weak purifying selection, respectively. Although these designations are arbitrary, the differences in the proportion of sequences that fall into each bin reflect variation in the degree to which orphan and ortholog genes are subject to functional constraint. Prior to analysis, unreliable estimates of ω due to extremely low (Ks = 0) or extremely high (Ks > 3) sequence divergence were removed from all datasets.

Results and discussion

Transcriptome assemblies

Read data for all six species have been submitted to the NCBI Sequence Read Archive (SRA Accession numbers SRX761208, SRX652499, SRX652504, SRX763246, SRX761332, and SRX652508). Assemblies have been deposited to the Dryad Digital Repository (available at http://dx.doi.org/10.5061/dryad.3pq3q). Assembly size ranged from ~41,000 transcripts in *N. cooperi* to more than 165,000 in *C. vibora* (Table 1). Isoform content appears to be proportionally much higher in *Cicurina* than in other assemblies. The *Cicurina* transcriptomes had the lowest contig N50 scores, while the highest values were associated with the *Nesticus* transcriptomes. The mean and median transcript lengths followed a similar pattern, with *Nesticus* having the highest, *Cicurina* having the lowest and *Habronattus* exhibiting intermediate results

Table 1. Transcriptome assembly summary statistics.

	C. travisae	*C. vibora*	*H. signatus*	*H. ustulatus*	*N. bishopi*	*N. cooperi*
Transcripts	144,351	165,397	55,081	57,456	52,484	41,169
Components ("genes")	122,700	146,297	53,337	55,642	51,216	35,817
N50	676	418	814	1002	1266	1161
Mean length	519	407	581	640	747	714
Median length	305	276	345	353	396	392
Maximum length	12,541	11,555	11,379	16,104	14,486	14,677
Nucleotides	74,927,623	67,360,479	31,976,435	36,757,749	39,233,450	29,393,279
% GC content	34.99	36.45	32.97	33.46	36.21	37.55
% of reads aligning to transcripts	83.14	79.96	86.01	90.26	89.4	84.91
Mean Read Coverage	22	18	50	52	41	22

https://doi.org/10.1371/journal.pone.0174102.t001

(see Table 1). Most reads were successfully mapped back onto the assembled transcripts, though there was variability in the proportions of reads aligned across species. This proportion is highest in *H. ustulatus* (~90%) and lowest for *C. vibora* (~80%). GC content was generally similar among the different taxa, ranging from 32.97% in *H. signatus* to 37.55% in *N. cooperi*. GC values are consistent with those of the velvet spider, *Stegodyphus mimosarum*, but lower than the 39.1% GC content seen in the tarantula *Acanthoscurria* ([23]).

We defined read coverage as the number of reads that successfully map to each transcriptomic base pair position. As seen in Table 1, the mean read coverage was highest in the *Habronattus* transcriptomes (~50x) and lowest in the *Cicurina* and *Nesticus cooperi* transcriptomes (18-22x). It should be noted that these values are all underestimates of the true average read coverage because the maximum coverage at any particular nucleotide position was capped at 200x in the Bedtools analysis. The proportion of each transcriptome with a given level of read coverage (summarized in S1 Fig) was relatively similar across *Habronattus* and *Nesticus* taxa. Approximately 90% of the nucleotides in these four transcriptomes have at least 5x coverage, while there is a broader range—36% to 57%—among these taxa for the proportion of total bases meeting the $\geq$ 20x coverage cutoff, although only in *N. cooperi* is this proportion lower than 50%. Coverage levels were markedly lower in *Cicurina*. Only about 66% of the bases found in the *C. travisae* transcriptome have coverage depths of 5x or greater, while the proportion of nucleotides with this level of coverage was slightly less than 55% in *C. vibora*. The percentage of nucleotides meeting or exceeding the 20x level of read coverage in *C. travisae* and *C. vibora* is approximately 26% and 20%, respectively. In the context of previously published spider RNA-Seq studies, the range of read coverage levels reported here is neither remarkably high nor unusually small (e.g., [33,76–78]).

Based on the metrics reported in Table 1, the assembly process resulted in assembled transcriptomes that are broadly similar across the *Habronattus* and *Nesticus* taxa—although there are a few notable exceptions seen in *N. cooperi*, such as fewer reads, transcripts and lower coverage. With many more transcripts that are often relatively short, however, the *Cicurina* assemblies are in some ways qualitatively different from the other four. Although the transcript quantity and length found in the *Cicurina* assemblies still fall within the range of values seen in other published RNA-Seq studies (e.g., [79,80]), the fact that these assemblies contain a much larger number of transcripts requires explanation. The *Cicurina* transcriptomes include a proportionally larger number of isoforms when compared to other taxa, which could be an indication of greater transcriptional complexity in these species. Alternatively, the large quantity of somewhat shorter transcripts might be an indication of sequence fragmentation. The *Cicurina* transcriptomes were sequenced using 100 bp reads, while the other transcriptomes were sequenced with 50 bp reads. However, it is unclear why longer reads would lead to a greater frequency of fragmented transcripts. A more important factor may be that the verified concentration of *Cicurina* RNA was much lower than in other taxa. In addition, biased base composition—such as AT-rich DNA sequences—is known to cause problems for NGS methods that rely on PCR amplification during library prep, resulting in low coverage and potential assembly difficulties (e.g., [81,82]). Even though all six transcriptomes are AT-rich, it may be the case that a combination of low RNA concentration and high AT content led to lower overall coverage levels for *Cicurina*, resulting in some low-coverage transcripts being assembled into multiple, shorter fragments.

Quality and completeness assessment

Results of the CEGMA analyses are summarized in Table 2. More than 90 percent of the CEGs are present at least in partial length for four of the six taxa, while the *C*. *vibora* and *N*. *cooperi* assemblies captured 87% and 88% of the CEGs, respectively. These results are broadly consistent with previous CEGMA analyses of spider transcriptomic data ([32,35]), and confirm that all six transcriptomes contain the majority of conserved eukaryotic protein-coding genes. As summarized in S3 Table, searching the spider-specific OGs against translated protein sequences resulted in patterns very similar to those seen in the CEGMA analysis. Species-specific differences aside, our results indicate that most genes in the spider ortholog group dataset are also present in our assemblies.

Results of the OHR analyses are shown in Fig 2 and S2 Table. In each species there is a bimodal distribution of OHR values with a peak near 0.0 and an additional peak at 1.0. Despite these similarities, however, the relative sizes of the two peaks and the overall distribution of OHR values differ dramatically in some cases. In particular, the histograms from the two *Cicurina* assemblies show peaks near 0.0 that are larger than the peaks near 1.0. Similarly, the mean and median OHR values are lowest in the *Cicurina* transcriptomes and highest in the *N*. *bishopi* and *H*. *signatus* transcriptomes. These results suggest that a smaller proportion of genes were assembled to full length in the *Cicurina* transcriptomes compared to the other taxa. Despite these larger proportional differences, the number of genes with an OHR greater than 0.5 and an OHR greater than 0.8 was relatively stable across taxa (see S2 Table). It should be noted that the OHR is a conservative approach that may underestimate the percentage of total gene length captured if the phylogenetic distance between the query sequence and its closest hit is great ([54]). Because spider-specific annotated sequences are still sparse in current databases, many transcripts hit to sequences found in distantly related organisms, possibly resulting in OHR values that are lower than they might otherwise be if compared with sequences from closer relatives. Nevertheless, these results suggest that while there are differences in the proportions of fully assembled genes across the six species, each transcriptome successfully captured at least several thousand genes that are approximately full length. Overall, multiple

Table 2. Results of CEGMA analyses.

Species	Full-length mapped CEGs (%)	Full-length + partially mapped CEGs (%)
C. travisae	217 (87.5)	227 (91.53)
C. vibora	198 (79.84)	216 (87.10)
H. signatus	222 (89.5)	233 (94.0)
H. ustulatus	238 (96.0)	243 (98.0)
N. bishopi	229 (92.34)	242 (97.58)
N. cooperi	203 (81.85)	219 (88.31)

https://doi.org/10.1371/journal.pone.0174102.t002

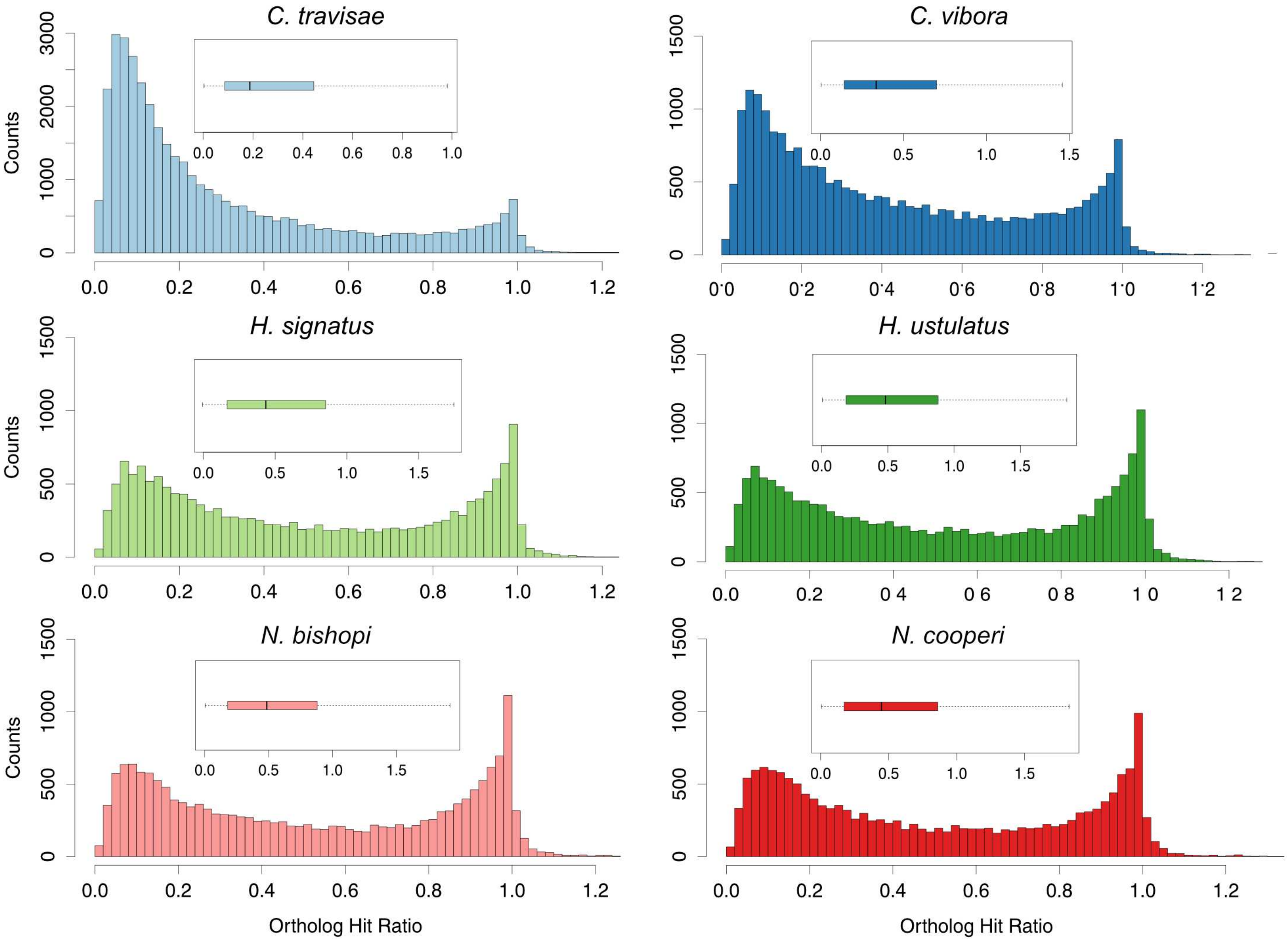

Fig 2. Distribution of OHR values. Histograms (main figure) and boxplots (inset) depict the distribution of OHR values in each species. Outliers have been removed from boxplots to better visualize results. Note the axis scale differences in the *C. travisae* plots.

https://doi.org/10.1371/journal.pone.0174102.g002

analyses point to the conclusion that the *H. ustulatus* and *N. bishopi* transcriptomes are the highest quality and most complete assemblies. Although the *Cicurina* assemblies show evidence of fragmentation, the quality and completeness assessment confirms the presence of many full-length protein-coding genes and supports the conclusion that the assembly process was generally successful.

Functional annotation

Putative homology between transcripts and sequences from the nr protein database was assigned using Blastx. The number of successful hits resulting from the standalone Blastx searches ranged from approximately 39,000 in *C. travisae* to about 16,600 in *H. signatus*, although these are likely overestimates of the number of genes discovered, since they include, in some cases, isoforms derived from the same coding sequence. The percentage of all transcripts with hits varied from only 14% for *C. vibora* to just over 40% for *N. cooperi* (S4 Table). There were no fungal taxa among the 100 most frequently hit species, suggesting that fungal contamination was likely not a significant problem in these assemblies.

Despite large variability in transcript quantity and number of Blastx hits, the number of transcripts that were successfully annotated with GO terms was consistent across the six transcriptomes. The highest number was seen in *C. travisae* (15,866 transcripts) and the lowest was in *H. signatus* (13,477 transcripts; S4 Table). Protein domain information provided by InterProscan resulted in the addition of 377–802 GO terms to sequences that failed to hit any proteins in the nr database. In the *Habronattus* and *Nesticus* transcriptomes, roughly 80% of sequences with Blastx hits were also successfully annotated with GO terms, while this proportion was substantially lower in the *Cicurina* transcriptomes. In particular, only about 40% of the *C. travisae* Blastx hits resulted in successful annotations. This suggests that many of these hits were not associated with GO terms and/or that many of the GO terms mapped to the sequences did not pass the annotation filter.

For all species, a majority of assembled transcripts could not be annotated with GO terms, suggesting that many of these transcripts may be sequence fragments, derived from non-coding RNAs, or are spider-specific genes that do not yet have homologs in the nr protein database. A relatively large proportion of sequences without homology to known proteins have been seen in other spider transcriptome studies (e.g., [32,76,78]). As well-characterized spider genomic data continue to be added to existing databases, there will presumably be a concomitant increase in the proportion of newly published spider sequences with homologous matches.

The total number of GO terms assigned to each transcriptome varied from about 64,000 in *C. travisae* to nearly 80,000 for *H. ustulatus*, although the number of distinct GO terms found in each transcriptome was in the low hundreds (see S4 Table). The fraction of successfully annotated transcripts with GO terms (Level 2) found in the three major categories—biological process, molecular function and cellular component—is depicted in Fig 3. Overall, the relative frequency of different GO terms is consistent across taxa, and with some exceptions, the relative abundance of many of the most common GO terms found (e.g., "cellular" and "metabolic" processes, "binding", "catalytic activity", etc.) is similar to findings from previous transcriptomic analyses of a variety of different taxa. These include studies of other spiders (e.g., [31,33,76]), other arachnids (e.g., [8,83,84]), other metazoans (e.g.,[53,85,86]), and even non-metazoans (e.g., [87,88]). One previously suggested explanation for these similarities (e.g., [7]) is that transcriptome assemblies from short-read RNA-Seq analyses are biased toward the capture of highly expressed, conserved housekeeping genes. Alternatively, homology-based methods may be biased toward annotating conserved housekeeping genes, particularly in species for which closely related sequences are largely unavailable in databases.

Venom genes

Blasting all six transcriptomes against both the nr protein and Arachnoserver spider toxins databases revealed evidence of homology to venom proteins in 146 transcripts. Of these, 101 transcripts were inferred to contain a signal peptide and/or propeptide sequence region. These 101 transcripts were considered to be putative venom protein-coding genes (S6 Table). Eighty-three of these were successfully functionally annotated, and an enrichment analysis identified five GO terms that are overrepresented (S5 Table). Based on the analysis, we found that venom transcripts are approximately 16 times more likely to be localized in the "extracellular region" (i.e., outside of the plasma membrane) than are non-venom transcripts. In fact, "extracellular region" was the only cellular component GO term found in any of the venom transcripts. Other studies that specifically targeted spider venom gland transcripts have also reported significant enrichment of the "extracellular region" term in venom genes (e.g., [28,77]). These venom-related transcripts are also ~ 4-8x more likely to be assigned functions related to

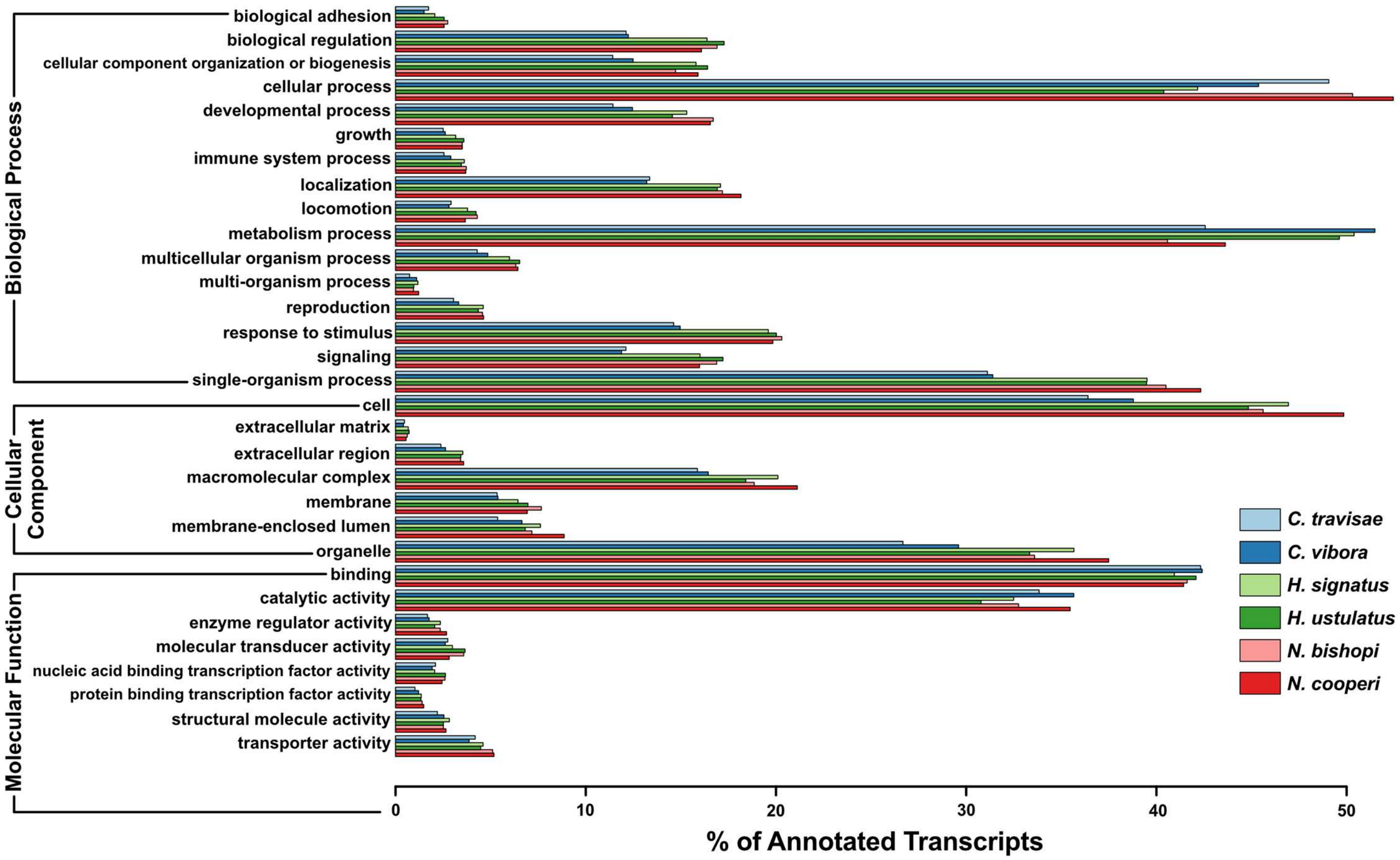

Fig 3. Results of GO analysis. Bar chart showing the proportion of annotated transcripts containing various GO terms (Level 2) across the 3 major categories.

https://doi.org/10.1371/journal.pone.0174102.g003

peptidase activity—which involves hydrolyzing the peptide bonds connecting amino acids in a polypeptide chain ([89])—cell adhesion, and enzyme regulation.

Some of the toxins most frequently inferred to be homologous to sequences in our transcriptomes include agatoxins, pisautoxins, and aranetoxins. Agatoxins were first isolated from the funnel weaver *Agelenopsis aperta* and are known to induce paralysis in prey items by blocking or otherwise interfering with the normal behavior of ion channels ([90]). First identified in the fishing spider *Dolomedes mizhoanus*, pisautoxins are proteins containing cystine knot motifs. It has been speculated that pisautoxins play a role in blocking calcium ion channels and/or inhibiting P2X3 receptors, but these functions have not been experimentally verified ([27]). Aranetoxins are a group of venoms that were first isolated from the Chinese orbweaver, *Araneus ventricosus*, whose specific molecular functions have yet to be elucidated ([91]). A fourth common venom protein is Techylectin-1-Phoneutria, which is a unique Ctenitoxin known only from the Brazilian armed spider *Phoneutria nigriventer* ([92]). This protein exhibits high sequence identity (> 50%) to the techylectin-5A protein found in the Japanese horseshoe crab, although the latter protein is not known to play a toxin-related role ([93]). Little is known about the biological activity of Techylectin-1-Phoneutria other than that it appears to be a carbohydrate-binding lectin with a fibrinogen C-terminal domain ([92]).

Based upon available data, it is not clear whether the frequencies of these particular venom families are primarily a reflection of the actual diversity of the toxins produced by *Cicurina*, *Habronattus* and *Nesticus* spiders or simply a product of biases in the taxonomic composition of the relatively small Arachnoserver database. While the homology-based inference of venom proteins presented here is suggestive, it would be unwise to draw definitive conclusions regarding these results until the functions of these putative venom genes can be validated by more targeted and perhaps experimental approaches.

Ortholog identification & tests for positive selection

The Transdecoder open reading frame prediction identified between 13,000 and 17,000 unique ORFs in each species. Using these coding sequences as input, the ortholog inference and clustering procedures implemented in InParanoid and QuickParanoid resulted in the discovery of 3,421 different clusters containing one gene from each of the six transcriptomes. For further analyses, each cluster was treated as a group of orthologous protein-coding genes. After implementing MSA and quality filtering procedures, 3,345 alignments remained, with a median gap-free length of ~851 bp (range of 105–7038 bp).

Using these 3,345 alignments, we tested for positive selection along each genus- and species-level branch of the canonical six species phylogeny using codeml's branch-site model. Replicate runs with different starting values for ω and κ (the transition-transversion ratio) generally resulted in very similar likelihood scores, but when the scores differed across replicates, the run with the higher score was used. After the FDR correction for multiple testing, LRTs between null and alternative models returned statistically significant (at a FDR of 0.1) signatures of positive selection along 24 branches from 22 ortholog alignments (Table 3). The BEB analysis found that 20 of these putative PSGs contained at least one codon belonging to the class of codons under positive selection at a posterior probability $\geq$ 0.95. On average, each PSG contained ~ four amino acids under positive selection. When these positively selected regions were present, they made up, on average, only 4.9% of the entire aligned sequence length. As expected, these results suggest that even in genes undergoing adaptive evolution, most amino acids are likely to be under strong purifying selection. The frequency of positive selection on each branch of the phylogeny is depicted in Fig 4A. Overall, instances of positive selection were relatively rare. Although PSGs are not heavily concentrated in any specific lineage, a slightly higher number are found on *Cicurina* branches. The *Cicurina* species, which are restricted to deep cave environments with high humidity and total darkness, also exhibit a higher degree of ecological specialization than the other taxa studied.

Of the 22 ortholog alignments with evidence for positive selection, all but one were successfully annotated with GO terms. Many of the most frequent terms under the biological process category relate to cellular, metabolic, single-organism cellular, developmental processes, and localization (Fig 4B); likewise, the common terms under the molecular function category include various types of binding, as well as "hydrolase", "oxidoreductase" and "transferase" activity. These latter three terms refer to the catalysis of various biochemical reactions, including hydrolysis, oxidation-reduction reactions, and the transfer of different chemical groups (e.g., methyl) from one compound to another ([89]). Owing perhaps to the small size of the selection dataset, Fisher's exact test did not find statistically significant evidence that GO terms were overrepresented in PSGs.

Although the homology-based gene identifications should be treated cautiously until more data are available, some tentative inferences can be made regarding the functions of genes under positive selection. One of the most striking patterns seen among the PSGs (listed in Table 3) is the presence of several proteins involved in muscle contraction, including myosin, actin and calcium ATPase. The codeml analysis detected selection on the myosin heavy chain on both the *Cicurina* and *Nesticus* branches. This protein is assembled into the backbone of

Table 3. Genes under positive selection as identified by codeml branch-site analysis.

Branch	Putative Protein ID	ω	FDR	Proportion under selection
Cicurina	Myosin heavy chain, muscle-like	17.8	4.48E-05	0.015
Nesticus		87.9	6.58E-02	0.009
Habronattus	Calcium-transporting ATPase sarcoplasmic/endoplasmic reticulum type	107.9	4.61E-03	0.014
Nesticus		314.8	4.36E-02	0.018
H. ustulatus	Alpha-actinin, sarcomeric	999	1.59E-06	0.008
C. travisae	Calcium homeostasis endoplasmic reticulum protein	114.5	1.03E-04	0.024
C. travisae	Peroxidasin	74.3	4.99E-03	0.036
N. bishopi	Hemocyanin G	999	4.53E-10	0.048
C. vibora	Homothorax 1	999	1.59E-06	0.018
Cicurina	Clotting factor B	267.6	4.52E-02	0.059
N. cooperi	S-adenosylmethionine synthase	999	1.68E-03	0.027
C. vibora	mRNA cap guanine-N7 methyltransferase	999	9.31E-02	0.012
H. signatus	Zinc finger protein 226	44.4	1.55E-02	0.028
C. vibora	tRNA (adenine(58)-N(1))-methyltransferase non-catalytic subunit TRM6-like	999	4.48E-05	0.017
Nesticus	Tropomyosin	152.9	4.36E-02	0.051
H. signatus	Ubiquitin-conjugating enzyme E2 R2	999	1.63E-02	0.033
H. ustulatus	Ras-related protein Rab-5C	696.5	4.48E-05	0.126
H. ustulatus	Lysozyme 1	999	7.80E-02	0.08
N. bishopi	L-azetidine-2-carboxylic acid acetyltransferase	999	4.52E-02	0.081
Cicurina	U24-ctenitoxin-Pn1a	26.7	8.75E-02	0.1
C. vibora	Mitochondrial fission factor	999	1.98E-02	0.011
C. vibora	Protein boule-like	999	4.52E-02	0.035
Nesticus	WD repeat-containing protein mio-B	227.5	4.52E-02	0.315
C. vibora	Eukaryotic translation initiation factor 4E-binding protein 1	999	4.93E-02	0.021

https://doi.org/10.1371/journal.pone.0174102.t003

muscle thick filaments and is responsible for controlling muscle contraction ([94]). In addition, we found evidence for selection on alpha-actin on the *H*. *ustulatus* branch. This protein is the major constituent of muscle thin filaments, which, combined with the myosin-dominated thick filaments, form sarcomeres ([95]). Another apparent muscle-related protein under selection in multiple lineages—in this case, *Habronattus* and *Nesticus*—is the sarco/endoplasimic form of calcium-transporting ATPase, which is responsible for removing calcium ions from the cytosol after muscle contraction ([96]). Tropomyosin, another protein involved in muscle contraction (although there are non-muscle forms as well; [97]), was found to be under selection in *N*. *bishopi*.

Other noteworthy PSGs include the appendage development gene *homothorax*-1 ([98]), the gene encoding respiratory protein hemocyanin G ([99]), and *Boule*, a gene known to be involved in metazoan gametogenesis ([100,101]). *Boule*, unlike many other reproductive proteins has previously been found to evolve primarily under purifying selection in other taxa ([102]), thus is it somewhat surprising to find it under positive selection in our data. We also found evidence in the *Cicurina* lineage of positive selection acting on a likely venom protein with high sequence similarity to the U24-ctenitoxin-Pn1a protein first sequenced from *Phoneutria nigriventer*. Although information is limited, this toxin is thought to be an inhibitor of cysteine proteinase activity ([92]).

In total this analysis identified evidence of adaptive molecular evolution in only 0.08% of the 29,835 (3,315 total ortholog alignments x nine lineages) branches examined. In comparison with several recent genome-wide scans for positive selection (e.g., [103–105]), our results suggest that a much smaller proportion of entelegyne spider genes have evolved under positive

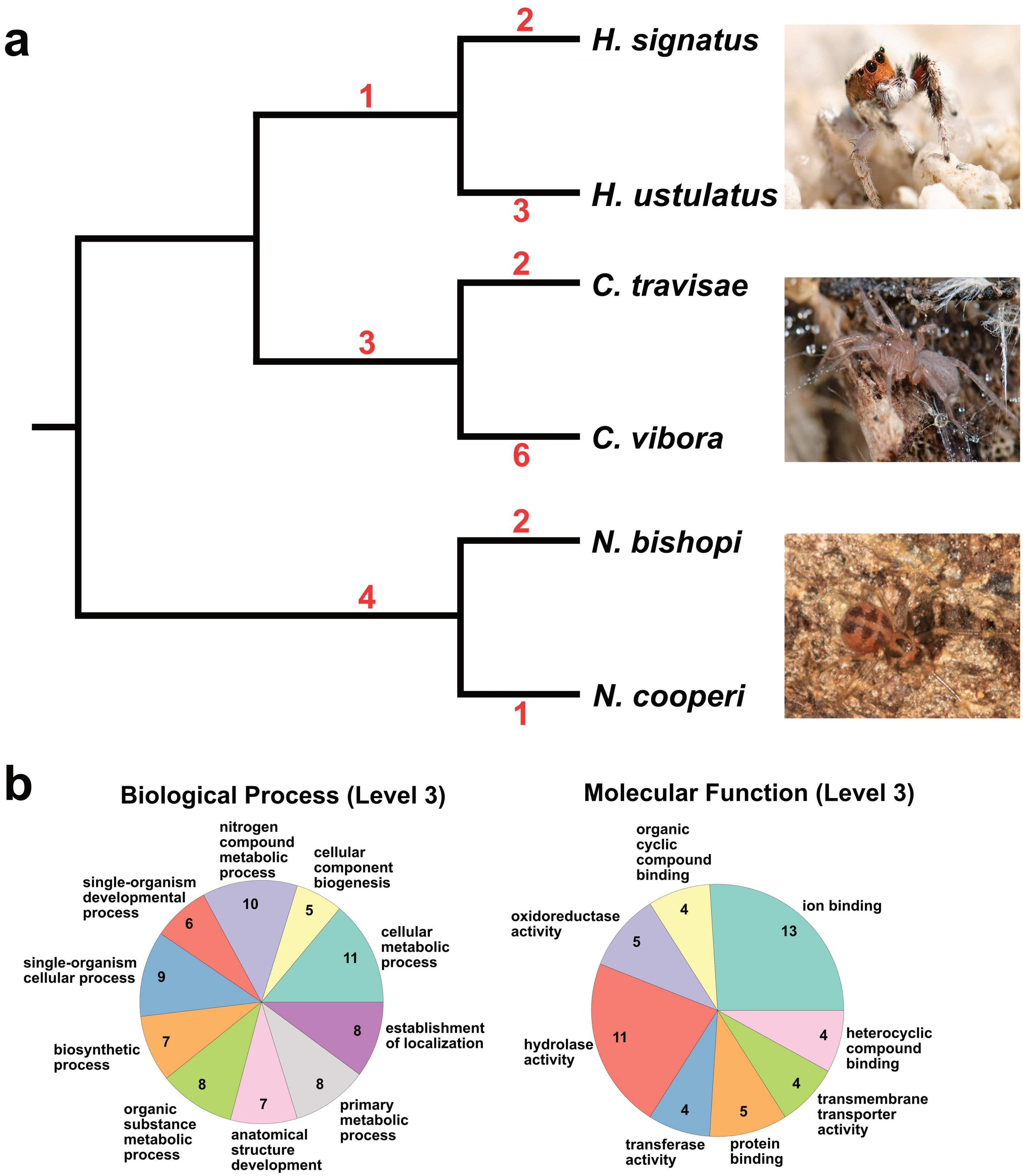

Fig 4. Tests for positive selection. (a) Phylogeny indicating the number of instances in which each branch was found to be under positive selection. *Nesticus* image used under a CC BY license with permission from Alan Cressler. (b) Most frequent GO terms found in PSGs.

https://doi.org/10.1371/journal.pone.0174102.g004

selection. The only other comparable study of spider genomics for a large number of loci found evidence of positive selection in a similarly low proportion of sequences ([34], but also see [35,36]). Although it is possible that our results may reflect the actual prevalence of positive selection in some entelegyne spider species, there are likely additional factors that contributed to the small number of inferred PSGs.

Our MSA strategy utilizing PRANK and GUIDANCE is a conservative approach that has been shown to reduce false negatives, albeit with a concomitant decrease in power to detect true positives ([106]). Preliminary analyses using different MSA procedures resulted in substantially larger numbers of inferred PSGs (data not shown), likely false positives resulting from alignment problems. Another factor that may have reduced the incidence of detected PSGs in our analysis is the sparse sampling of relatively divergent taxa. Previous work ([107]) has shown that the branch-site test has difficulty detecting positive selection along branches with saturated synonymous substitutions. Although observed divergence levels among congeneric species pairs was low (median Ks = 0.0105), the three genera are distant relatives ([9]), resulting in much higher divergences levels in interior branches (median Ks = 1.4036). Thus, it may be the case that our analysis missed some instances of adaptive protein evolution—especially those that occurred along longer interior branches—due to saturation. More broadly, other researchers (e.g. [108–110]) have found that power to detect positive selection on a particular phylogenetic branch increases with the addition of taxa. Given our limited sample size, it is also possible that the analysis lacked statistical power to detect positive selection events occurring on the phylogeny. Finally, another important factor likely driving down the incidence of observed adaptive evolution may have been our choice to examine only single-copy ortholog groups. Although other similar studies have found higher rates of positive selection among single-copy orthologs ([103–105]), it may be the case that these gene families are subject to lower rates of positive selection than multi-copy gene families. As more large-scale scans for positive selection are performed on spider genomic data, it will be important to see whether or not the pattern of low rates of positive selection in spider lineages seen here and in previous work is maintained.

Orphan genes

After removing isoforms and filtering out sequences without an ORF at least 200 bp in length, each transcriptome still included more than 2,000 potentially coding sequences without hits to the nr database. Using a paired congeneric sampling strategy, we verified the presence of lineage specific orphan genes under the assumption that a previously unknown sequence with coding potential is less likely to be the result of a sequencing error or assembly artifact if also present in a close congener. The results of this search are summarized with Venn diagrams in Fig 5. The *Cicurina* transcriptomes shared 547 putatively novel genes; the *Habronattus* transcriptomes shared 1,116 novel genes, and the *Nesticus* taxa shared 907 putative orphan genes. When comparing shared genes across different genera there was a steep drop-off in the number of undescribed genes found in common (Fig 5); in total, we found evidence for 18 novel genes shared among all taxa.

The fact that *Cicurina* and *Habronattus*, which are sister taxa in this analysis, share fewer taxonomically-restricted genes than do *Habronattus* and *Nesticus* suggests that variation in transcriptome completeness (e.g., due to variability in the number of samples, sex and tissue types used) likely influenced the number of shared genes discovered. However, the lack of orphan genes in one transcriptome that are present in another is, in some cases, likely a reflection of true absence. Given the finding that orphan genes comprise ~10–20% of each newly sequenced genome (e.g., [111]), it is not unreasonable to conclude that many of the orphans

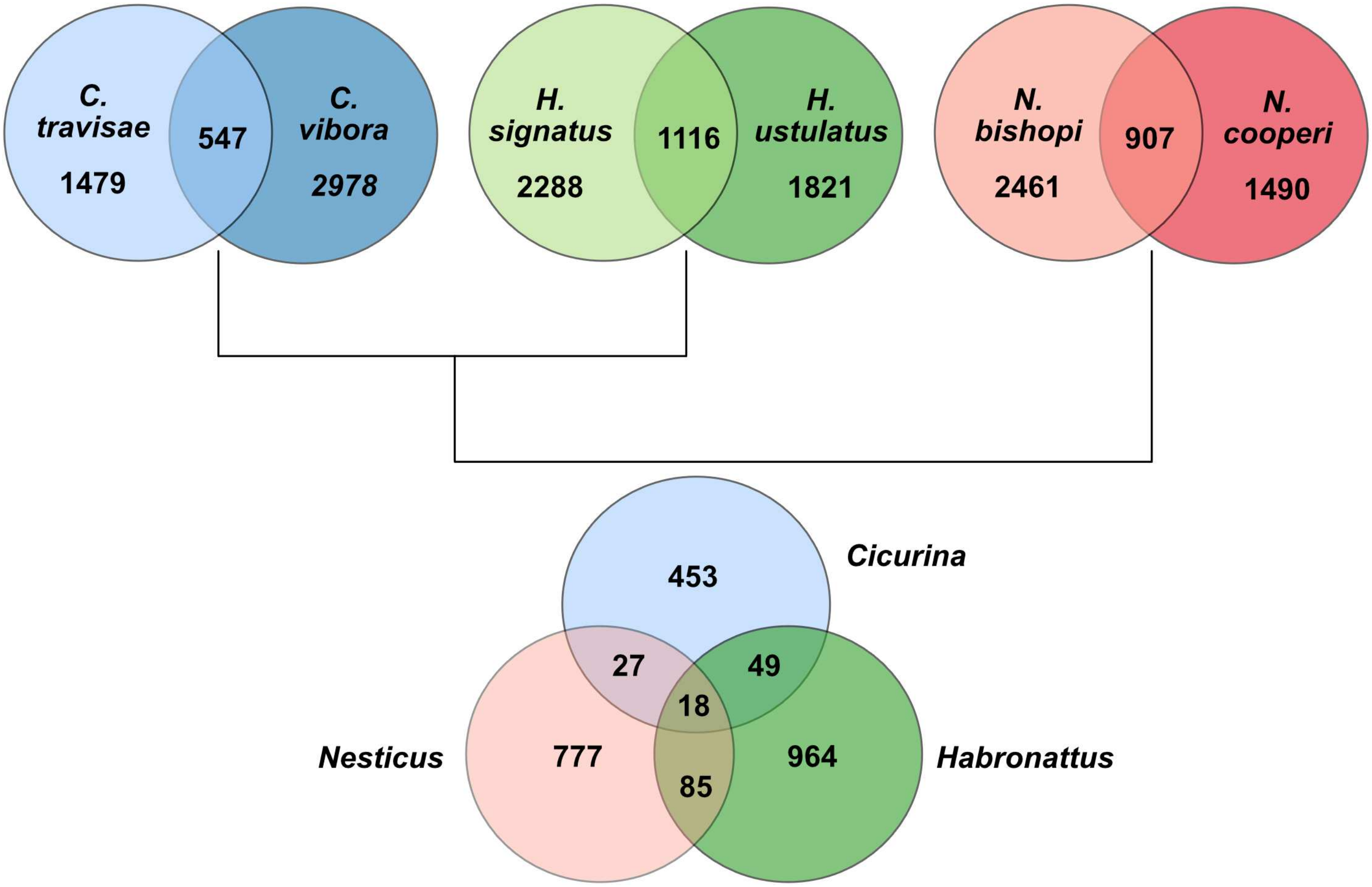

Fig 5. Venn diagrams showing results of the orphan gene search. Numbers within circle intersections represent genes shared among taxa. Numbers outside of intersections are genes exclusive to that taxon.

https://doi.org/10.1371/journal.pone.0174102.g005

found only in congeners are unique to those individual lineages. The taxa used in this analysis, while all entelegyne spiders, are nonetheless only distantly related. In fact, the divergence time between *Nesticus* and *Cicurina+Habronattus* (~154–290 MA; [9]) is roughly akin to the split between humans and monotremes. As such, it is likely that many of the genus-specific orphan genes detected in this analysis evolved *de novo* after the different lineages diverged from their respective common ancestors. Also, differences in the quantity of orphan genes seen among lineages might be influenced by lineage-specific losses of orphan genes (e.g., [112]).

Finally, another factor that undoubtedly contributed to the relatively small number of orphan genes shared among all transcriptomes were the specific quality control filters we chose. In preliminary analyses, using fewer and less stringent filters resulted in approximately 4-6x more taxonomically-restricted genes shared among genera as well as among all species. After exploring a number of different possible filtering parameters, the results reported here represent our best compromise between sensitivity and reliability. However, it is reasonable to presume that many true orphan genes were missed simply due to our choice of filters.

Because the novel genes discovered by this analysis lacked—by definition—hits to the nr database, the vast majority of them could not be functionally annotated. Nevertheless, each comparison between congeneric pairs revealed sequences to which GO terms could be applied using protein domain and motif information from InterProScan. Functional annotations were assigned to 56 *Cicurina*, 34 *Habronattus*, and 76 *Nesticus* orphans, respectively. Under the biological process category, the most frequent Level 2 terms were “cellular process”, “single-organism process” and “metabolic process”. An examination of more specific functions (GO Level 3 and above), suggests that many of these transcripts encode gene products with metabolic roles; 13 of the 19 most frequent terms involve metabolic processes (S2A Fig). Under the molecular function category, the two most common Level 2 terms are “binding” (85 transcripts) and “catalytic activity” (19 transcripts). At more specific GO levels, the vast majority of the most common terms relate to various forms of binding, including ion binding, protein and DNA binding (S2B Fig), suggesting that these are functions shared in common by many of the orphan genes. Overall, the GO terms assigned to orphan genes are qualitatively very similar to the most frequent GO terms found in all the annotated transcripts. Based on this, it seems likely that the functional annotations assigned by InterProScan were heavily biased toward highly conserved protein domains, making it difficult to infer what, if any, unique spider-specific roles these orphan genes perform.

After aligning and quality control filtering, we estimated substitution rates for 499 *Cicurina* orphan genes, 1104 *Habronattus* orphans, and 893 *Nesticus* orphans. Median aligned orphan lengths ranged from 408 bp for *Cicurina* to 510 bp for *Habronattus*, consistent with previous results showing that taxonomically restricted genes tend to be shorter than phylogenetically older sequences (e.g., [112,113]). The results of comparing Ka, Ks and ω between congeneric orphan genes and the congeneric orthologs are summarized in Fig 6 and Table 4. In nearly every case, estimates of all three evolutionary rate parameters showed a consistent and statistically significant trend toward larger values in the orphans than in the orthologs (Mann-Whitney U tests, $P < 2.2 \times 10^{-16}$ suggesting an increase in the overall rate of evolution in the orphan genes. The only exception is the comparison of Ks values in the *Nesticus* sequences, where orthologs have a slight and non-significantly higher rate of synonymous substitutions

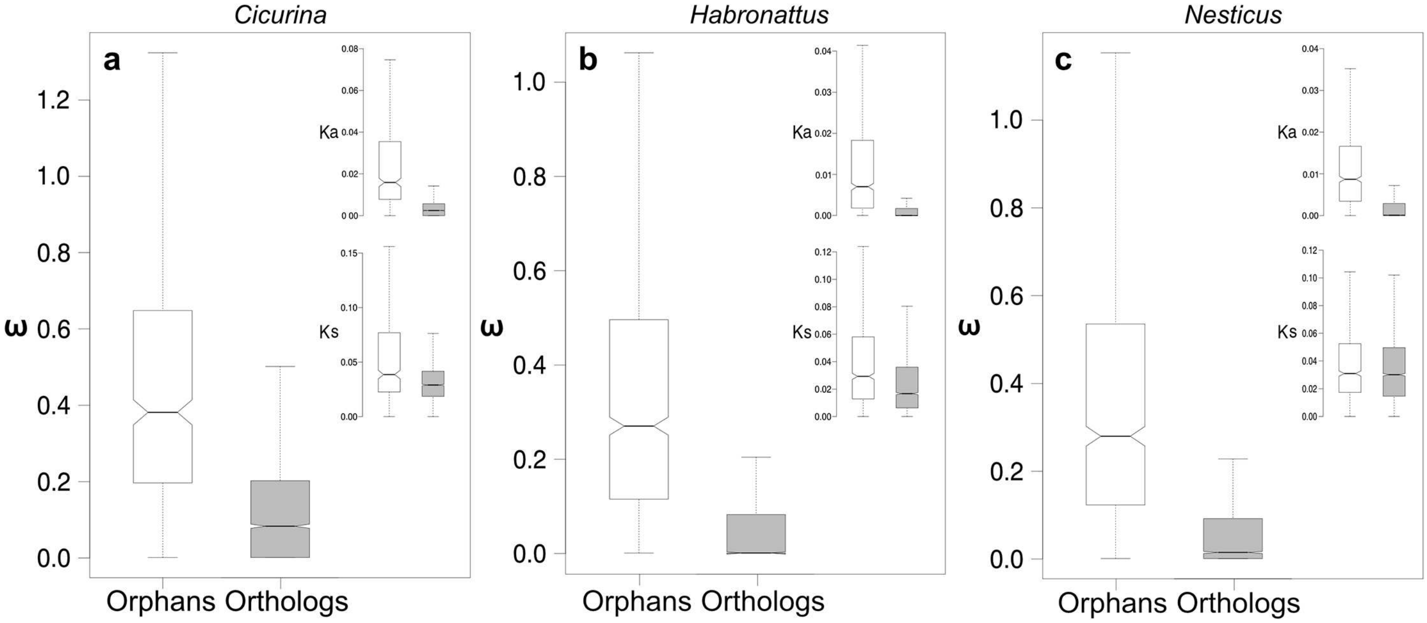

Fig 6. Boxplots of evolutionary rate comparisons between orthologs (grey) and orphan genes (white). (a) *Cicurina*, (b) *Habronattus* and (c) *Nesticus*. The main plot shows estimates of ω, while the inset shows Ka (top) and Ks (bottom). Outliers have been removed to better visualize the inter-quartile ranges. Mann-Whitney U tests indicate that all parameter estimates are significantly greater in orphan genes, with the exception of the *Nesticus* Ks comparison which shows no significant difference.

https://doi.org/10.1371/journal.pone.0174102.g006

Table 4. Comparison of substitution rates between congeneric orphan genes and orthologs.

	Cicurina		***Habronattus***		***Nesticus***	
	Orphans	**Orthologs**	**Orphans**	**Orthologs**	**Orphans**	**Orthologs**
Median ω	0.380	0.081	0.257	0.001	0.276	0.017
Median Ka	0.016	0.002	0.008	0	0.009	8E-04
Median Ks	0.039	0.029	0.033	0.021	0.032	0.032
ω ≤ 0.33 (%)	42.20	88.39	59.42	95.68	55.43	95.22
0.33 < ω ≤ 0.67 (%)	32.97	8.41	23.70	3.61	25.89	3.80
0.67 < ω ≤ 1.0 (%)	10.77	1.91	9.81	0.56	8.91	0.68
ω > 1.0 (%)	14.1	1.29	7.1	0.15	9.77	0.29

https://doi.org/10.1371/journal.pone.0174102.t004

(Mann-Whitney U test, P = 0.1389). Of particular importance, median estimates of *ω* ranged from ~ 4.7x to more than 200x larger in the orphan genes than in the orthologs. This high latter value is primarily a reflection of the especially small number of amino acid replacements in the *Habronattus* orthologs. In principle, this result could be explained by reduced synonymous substitution rates in the orphans. However, because Ks values were actually higher in most of the orphan genes, the concomitant increase seen in *ω* ratios is necessarily a result of excess nonsynonymous substitutions. It should be noted that we repeated this statistical analysis without first removing the most divergent sequences (Ks > 3), which did not change the overall outcomes reported here.

This discovery—elevated *ω* ratios driven by increased rates of nonsynonymous substitutions—is consistent with the hypothesis that orphan genes are involved in lineage-specific adaptive processes (e.g., [10,114,115]). However, positive selection is not the only potential explanation for this pattern. To some extent, our results might also be explained by a relaxation in purifying selection against the fixation of deleterious amino acids in orphan genes. For example, fewer orphans appear to be evolving under strong purifying selection as compared to the orthologs, and a larger fraction of the orphans show rates of substitution more consistent with weaker purifying selection or even neutral evolution (Table 4**)**. Given this, the elevated rates of nonsynonymous substitutions could suggest relaxed functional constraints or even incipient pseudogenization in some of the orphans. In particular, if single-copy orthologs tend to experience stronger than average purifying selection (e.g., [116]), this may have the effect of inflating the disparity in ω ratios seen between our ortholog and orphan gene datasets. However, most of the orphans appear to be evolving at rates primarily indicative of moderate-to-strong purifying selection (Table 4**)**, consistent with previous studies of orphan genes (e.g., [112]). This suggests that a decrease in the extent of purifying selection is unlikely to be the only factor driving the higher frequency of amino acid substitutions.

Importantly, the total proportion of comparisons in which we found pairwise congeneric estimates of *ω* > 1 is more than an order of magnitude greater in the orphans (9.54%) than in the orthologs (0.60%). This result strongly suggests an increased role for positive selection in orphan sequence evolution. Indeed, visual inspection of orphan alignments reveals many striking cases in which the number of amino acid substitutions greatly outnumber silent site substitutions in otherwise relatively conserved, high quality alignments. If relaxed selection alone were driving the observed increase in the rate of nonsynonymous substitutions, we would not expect Ka to exceed Ks in so many instances. Although these results are suggestive and in line with previous work finding evidence of adaptive evolution in orphans (e.g., [113,115,117]), we cannot definitively conclude that all orphan genes with *ω* > 1 have indeed undergone positive selection. This is because the tests for positive selection that we performed on the ortholog dataset are not applicable for pairwise sequence comparisons. This important caveat aside, our results point to a plausible conclusion that a substantial proportion of these lineage-specific spider genes have likely experienced bouts of adaptive protein evolution.

As genomic resources for non-model organisms have become more readily available, interest in the function, origin, and evolutionary dynamics of taxonomically restricted genes has increased (e.g., [11,111]). We believe that the orphan genes identified in this analysis represent fertile ground for future explorations aimed at understanding molecular evolution in entelegyne spiders. In particular, analysis of expression patterns in orphans should help substantiate the biochemical validity of these sequences and provide much needed information regarding their functional roles. In addition, we assert that the congeneric pair strategy implemented here is an effective means of identifying and validating new lineage-specific genes.

Supporting information

S1 Table. RNA extraction and sequencing information.
(XLSX)

S2 Table. Results of Ortholog Hit Ratio analyses.
(XLSX)

S3 Table. Matches to spider core ortholog dataset.
(XLSX)

S4 Table. Blastx and B2G functional annotation results.
(XLSX)

S5 Table. Results of enrichment analysis, showing GO terms overrepresented in venom-related transcripts.
(XLSX)

S6 Table. List of putative venom proteins found in the six transcriptomes.
(XLSX)

S1 Fig. Proportion of each transcriptome with $\geq$ 5x, $\geq$ 10x and $\geq$ 20x read coverage.
(TIFF)

S2 Fig. Distribution of the most common GO terms assigned to orphan genes in the (a) biological process and (b) molecular function categories. Only GO terms at levels 3 and above present in at least ten transcripts are shown.
(TIFF)

Acknowledgments

We wish to thank Shahan Derkarabetian, Keith Bradnam, and Jenny Felix who contributed to the laboratory work for this project and/or assisted perform analyses. Jim Starrett in particular helped with RNA extractions, and we thank Cheryl Hayashi for use of lab resources. Shahan Derkarabetian, Liz Waters, Rob Edwards, Joe Kelly, Laurel Yohe, Jesse Hollister, and two anonymous reviewers offered useful comments that improved the manuscript.

Author Contributions

Conceptualization: DEC MH.

Data curation: DEC MH.

Formal analysis: DEC.

Funding acquisition: MH.

Investigation: DEC.

Methodology: DEC MH.

Project administration: MH.

Resources: MH.

Supervision: MH.

Validation: DEC MH.

Visualization: DEC MH.

Writing – original draft: DEC MH.

Writing – review & editing: DEC MH.

References

1. Crawford NG, Faircloth BC, McCormack JE, Brumfield RT, Winker K, Glenn TC. More than 1000 ultraconserved elements provide evidence that turtles are the sister group of archosaurs. Biol Lett. 2012; 8: 783–786. https://doi.org/10.1098/rsbl.2012.0331 PMID: 22593086
2. Hejnol A, Obst M, Stamatakis A, Ott M, Rouse GW, Edgecombe GD, et al. Assessing the root of bilaterian animals with scalable phylogenomic methods. Proc Biol Sci. 2009; 276: 4261–4270. https://doi.org/10.1098/rspb.2009.0896 PMID: 19759036
3. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet. 2009; 10: 57–63. https://doi.org/10.1038/nrg2484 PMID: 19015660
4. Kunstner A, Wolf JB, Backstrom N, Whitney O, Balakrishnan CN, Day L, et al. Comparative genomics based on massive parallel transcriptome sequencing reveals patterns of substitution and selection across 10 bird species. Mol Ecol. 2010; 19: 266–276. https://doi.org/10.1111/j.1365-294X.2009.04487.x PMID: 20331785
5. Baldo L, Santos ME, Salzburger W. Comparative transcriptomics of Eastern African cichlid fishes shows signs of positive selection and a large contribution of untranslated regions to genetic diversity. Genome Biol Evol. 2011; 3: 443–455. https://doi.org/10.1093/gbe/evr047 PMID: 21617250
6. Elmer KR, Fan S, Gunter HM, Jones JC, Boekhoff S, Kuraku S, et al. Rapid evolution and selection inferred from the transcriptomes of sympatric crater lake cichlid fishes. Mol Ecol. 2010; 19: 197–211.
7. Hittinger CT, Johnston M, Tossberg JT, Rokas A. Leveraging skewed transcript abundance by RNA-Seq to increase the genomic depth of the tree of life. Proc Natl Acad Sci U S A. 2010; 107: 1476–1481. https://doi.org/10.1073/pnas.0910449107 PMID: 20080632
8. Hedin M, Starrett J, Akhter S, Schonhofer AL, Shultz JW. Phylogenomic resolution of paleozoic divergences in harvestmen (Arachnida, Opiliones) via analysis of next-generation transcriptome data. PLoS One. 2012; 7: e42888. https://doi.org/10.1371/journal.pone.0042888 PMID: 22936998
9. Garrison NL, Rodriguez J, Agnarsson I, Coddington JA, Griswold CE, Hamilton CA, et al. Spider phylogenomics: untangling the Spider Tree of Life. PeerJ. 2016; 4: e1719. https://doi.org/10.7717/peerj.1719 PMID: 26925338
10. Tautz D, Domazet-Loso T. The evolutionary origin of orphan genes. Nat Rev Genet. 2011; 12: 692–702. https://doi.org/10.1038/nrg3053 PMID: 21878963
11. Wissler L, Gadau J, Simola DF, Helmkampf M, Bornberg-Bauer E. Mechanisms and dynamics of orphan gene emergence in insect genomes. Genome Biol Evol. 2013; 5: 439–445. https://doi.org/10.1093/gbe/evt009 PMID: 23348040
12. Sun W, Zhao XW, Zhang Z. Identification and evolution of the orphan genes in the domestic silkworm, *Bombyx mori*. FEBS Lett. 2015; 589: 2731–2738. https://doi.org/10.1016/j.febslet.2015.08.008 PMID: 26296317
13. Prabh N, Rödelsperger C. Are orphan genes protein-coding, prediction artifacts, or non-coding RNAs? BMC Bioinformatics. 2016; 17: 226. https://doi.org/10.1186/s12859-016-1102-x PMID: 27245157
14. Khalturin K, Anton-Erxleben F, Sassmann S, Wittlieb J, Hemmrich G, Bosch TCG, et al. A Novel Gene Family Controls Species-Specific Morphological Traits in *Hydra*. PLoS Biol. 2008; 6: e278. https://doi.org/10.1371/journal.pbio.0060278 PMID: 19018660

15. Colbourne JK, Pfrender ME, Gilbert D, Thomas WK, Tucker A, Oakley TH, et al. The Ecoresponsive Genome of *Daphnia pulex*. Science. 2011; 331(6017): 555–61. https://doi.org/10.1126/science.1197761 PMID: 21292972

16. Carvunis A-R, Rolland T, Wapinski I, Calderwood MA, Yildirim MA, Simonis N, et al. Proto-genes and de novo gene birth. Nature. 2012; 487: 370–374. https://doi.org/10.1038/nature11184 PMID: 22722833

17. Coddington JA. Phylogeny and classification of spiders. In: Ubick D, Paquin P, Cushing PE, Roth V, editors. Spiders of North America: an identification manual. American Arachnological Society; 2005. pp. 18–24.

18. Griswold CE, Ramírez MJ, Coddington JA, Platnick NI. Atlas of phylogenetic data for entelegyne spiders (Araneae: Araneomorphae: Entelegynae), with comments on their phylogeny. Proc Calif Acad Sci. 2005; 56: Suppl II:1–324.

19. Griswold CE, Coddington J a, Platnick, Forster RR. Towards a phylogeny of entelegyne spiders (araneae, araneomorphae, entelegynae). J Arachnol. 1999; 27: 53–63.

20. Miller JA, Carmichael A, Ramirez MJ, Spagna JC, Haddad CR, Rezac M, et al. Phylogeny of entelegyne spiders: affinities of the family Penestomidae (NEW RANK), generic phylogeny of Eresidae, and asymmetric rates of change in spinning organ evolution (Araneae, Araneoidea, Entelegynae). Mol Phylogenet Evol. 2010; 55: 786–804. https://doi.org/10.1016/j.ympev.2010.02.021 PMID: 20206276

21. Agnarsson I, Gregoric M, Blackledge TA, Kuntner M. The phylogenetic placement of Psechridae within Entelegynae and the convergent origin of orb-like spider webs. J Zool Syst Evol Res. 2013; 51: 100–106.

22. Wheeler WC, Coddington JA, Crowley LM, Dimitrov D, Goloboff PA, Griswold CE, et al. The spider tree of life: phylogeny of Araneae based on target-gene analyses from an extensive taxon sampling. Cladistics. 2016;

23. Sanggaard KW, Bechsgaard JS, Fang X, Duan J, Dyrlund TF, Gupta V, et al. Spider genomes provide insight into composition and evolution of venom and silk. Nature. 2014; 5.

24. Chen J, Zhao L, Jiang L, Meng E, Zhang Y, Xiong X, et al. Transcriptome analysis revealed novel possible venom components and cellular processes of the tarantula *Chilobrachys jingzhao* venom gland. Toxicon. 2008; 52: 794–806. https://doi.org/10.1016/j.toxicon.2008.08.003 PMID: 18778726

25. Gremski LH, da Silveira RB, Chaim OM, Probst CM, Ferrer VP, Nowatzki J, et al. A novel expression profile of the *Loxosceles intermedia* spider venomous gland revealed by transcriptome analysis. Mol Biosyst. 2010; 6: 2403–2416. https://doi.org/10.1039/c004118a PMID: 20644878

26. Jiang L, Zhang D, Zhang Y, Peng L, Chen J, Liang S. Venomics of the spider *Ornithoctonus huwena* based on transcriptomic versus proteomic analysis. Comp Biochem Physiol Part D Genomics Proteomics. 2010; 5: 81–88. https://doi.org/10.1016/j.cbd.2010.01.001 PMID: 20403776

27. Jiang L, Liu C, Duan Z, Deng M, Tang X, Liang S. Transcriptome analysis of venom glands from a single fishing spider *Dolomedes mizhoanus*. Toxicon. 2013; 73: 23–32. https://doi.org/10.1016/j.toxicon.2013.07.005 PMID: 23851222

28. Haney RA, Ayoub NA, Clarke TH, Hayashi CY, Garb JE. Dramatic expansion of the black widow toxin arsenal uncovered by multi-tissue transcriptomics and venom proteomics. BMC Genomics. 2014; 15: 366. https://doi.org/10.1186/1471-2164-15-366 PMID: 24916504

29. Prosdocimi F, Bittencourt D, da Silva FR, Kirst M, Motta PC, Rech EL. Spinning gland transcriptomics from two main clades of spiders (order: Araneae)—insights on their molecular, anatomical and behavioral evolution. PLoS One. 2011; 6: e21634. https://doi.org/10.1371/journal.pone.0021634 PMID: 21738742

30. Collin MA, Clarke TH 3rd, Ayoub NA, Hayashi CY. Evidence from multiple species that spider silk glue somponent ASG2 is a spidroin. Sci Rep. 2016; 6: 21589. https://doi.org/10.1038/srep21589 PMID: 26875681

31. Zhao YJ, Zeng Y, Chen L, Dong Y, Wang W. Analysis of transcriptomes of three orb-web spider species reveals gene profiles involved in silk and toxin. Insect Sci. 2014; 21: 687–698. https://doi.org/10.1111/1744-7917.12068 PMID: 24167122

32. Clarke TH, Garb JE, Hayashi CY, Haney RA, Lancaster AK, Corbett S, et al. Multi-tissue transcriptomics of the black widow spider reveals expansions, co-options, and functional processes of the silk gland gene toolkit. BMC Genomics. 2014; 15: 365. https://doi.org/10.1186/1471-2164-15-365 PMID: 24916340

33. Mattila TM, Bechsgaard JS, Hansen TT, Schierup MH, Bilde T. Orthologous genes identified by transcriptome sequencing in the spider genus *Stegodyphus*. BMC Genomics. 2012; 13: 70. https://doi.org/10.1186/1471-2164-13-70 PMID: 22333217

34. Yim KM, Brewer MS, Miller CT, Gillespie RG. Comparative transcriptomics of maturity-associated color change in Hawaiian spiders. J Hered. 2014; 105: 771–781.

35. Brewer MS, Carter RA, Croucher PJ, Gillespie RG. Shifting habitats, morphology, and selective pressures: developmental polyphenism in an adaptive radiation of Hawaiian spiders. Evolution. 2015; 69: 162–178. https://doi.org/10.1111/evo.12563 PMID: 25403483

36. Clarke TH, Garb JE, Hayashi CY, Arensburger P, Ayoub NA. Spider Transcriptomes Identify Ancient Large-Scale Gene Duplication Event Potentially Important in Silk Gland Evolution. Genome Biol Evol. 2015; 7: 1856–70. https://doi.org/10.1093/gbe/evv110 PMID: 26058392

37. Hedin MC. Molecular phylogenetics at the population/species interface in cave spiders of the southern Appalachians (Araneae: Nesticidae: *Nesticus*). Mol Biol Evol. 1997; 14: 309–324. PMID: 9066798

38. Paquin P, Duperre N. A first step towards the revision of *Cicurina*: redescription of type specimens of 60 troglobitic species of the subgenus Cicurella (Araneae: Dictynidae), and a first visual assessment of their distribution. Zootaxa. 2009; 1–67.

39. Griswold CE. A revision of the jumping spider genus *Habronattus* F.O.P. Cambridge (Araneae; Salticidae), with phenetic and cladistic analyses. University of California Press, Berkeley; 1987.

40. Maddison W, Hedin M. Phylogeny of *Habronattus* jumping spiders (Araneae: Salticidae), with consideration of genital and courtship evolution. Syst Entomol. 2003; 28: 1–21.

41. Hedin M. High-stakes species delimitation in eyeless cave spiders (*Cicurina*, Dictynidae, Araneae) from central Texas. Mol Ecol. 2015; 24: 346–361. https://doi.org/10.1111/mec.13036 PMID: 25492722

42. Andrews S. FastQC: A quality control tool for high throughput sequence data. Available: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/. 2010;

43. Schmieder R, Edwards R. Quality control and preprocessing of metagenomic datasets. Bioinformatics. 2011; 27: 863–864. https://doi.org/10.1093/bioinformatics/btr026 PMID: 21278185

44. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol. 2011; 29: 644–U130. https://doi.org/10.1038/nbt.1883 PMID: 21572440

45. Schmieder R, Edwards R. Fast identification and removal of sequence contamination from genomic and metagenomic datasets. PLoS One. 2011; 6: e17288. https://doi.org/10.1371/journal.pone.0017288 PMID: 21408061

46. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 2009; 10: R25. https://doi.org/10.1186/gb-2009-10-3-r25 PMID: 19261174

47. Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics. 2011; 12: 323. https://doi.org/10.1186/1471-2105-12-323 PMID: 21816040

48. Francis WR, Christianson LM, Kiko R, Powers ML, Shaner NC, Haddock SH. A comparison across non-model animals suggests an optimal sequencing depth for de novo transcriptome assembly. BMC Genomics. 2013; 14: 167. https://doi.org/10.1186/1471-2164-14-167 PMID: 23496952

49. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010; 26: 841–842. https://doi.org/10.1093/bioinformatics/btq033 PMID: 20110278

50. Surget-Groba Y, Montoya-Burgos JI. Optimization of de novo transcriptome assembly from next-generation sequencing data. Genome Res. 2010; 20: 1432–1440. https://doi.org/10.1101/gr.103846.109 PMID: 20693479

51. Parra G, Bradnam K, Korf I. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. Bioinformatics. 2007; 23: 1061–1067. https://doi.org/10.1093/bioinformatics/btm071 PMID: 17332020

52. Parra G, Bradnam K, Ning Z, Keane T, Korf I. Assessing the gene space in draft genomes. Nucleic Acids Res. 2009; 37: 289–297. https://doi.org/10.1093/nar/gkn916 PMID: 19042974

53. O'Neil ST, Dzurisin JD, Carmichael RD, Lobo NF, Emrich SJ, Hellmann JJ. Population-level transcriptome sequencing of nonmodel organisms *Erynnis propertius* and *Papilio zelicaon*. BMC Genomics. 2010; 11: 310. https://doi.org/10.1186/1471-2164-11-310 PMID: 20478048

54. Ewen-Campen B, Shaner N, Panfilio KA, Suzuki Y, Roth S, Extavour CG. The maternal and early embryonic transcriptome of the milkweed bug *Oncopeltus fasciatus*. BMC Genomics. 2011; 12: 61. https://doi.org/10.1186/1471-2164-12-61 PMID: 21266083

55. Conesa A, Gotz S, Garcia-Gomez JM, Terol J, Talon M, Robles M. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics. 2005; 21: 3674–3676. https://doi.org/10.1093/bioinformatics/bti610 PMID: 16081474

56. Quevillon E, Silventoinen V, Pillai S, Harte N, Mulder N, Apweiler R, et al. InterProScan: protein domains identifier. Nucleic Acids Res. 2005; 33: W116–20. https://doi.org/10.1093/nar/gki442 PMID: 15980438

57. Herzig V, Wood DL, Newell F, Chaumeil PA, Kaas Q, Binford GJ, et al. ArachnoServer 2.0, an updated online resource for spider toxin sequences and structures. Nucleic Acids Res. 2011; 39: D653–7. https://doi.org/10.1093/nar/gkq1058 PMID: 21036864

58. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate—a Practical and Powerful Approach to Multiple Testing. J R Stat Soc Ser B-Methodological. 1995; 57: 289–300.

59. Remm M, Storm CE, Sonnhammer EL. Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. J Mol Biol. 2001; 314: 1041–1052. https://doi.org/10.1006/jmbi.2000.5197 PMID: 11743721

60. Ostlund G, Schmitt T, Forslund K, Kostler T, Messina DN, Roopra S, et al. InParanoid 7: new algorithms and tools for eukaryotic orthology analysis. Nucleic Acids Res. 2010; 38: D196–203. https://doi.org/10.1093/nar/gkp931 PMID: 19892828

61. Schneider A, Souvorov A, Sabath N, Landan G, Gonnet GH, Graur D. Estimates of positive Darwinian selection are inflated by errors in sequencing, annotation, and alignment. Genome Biol Evol. 2009; 1: 114–118. https://doi.org/10.1093/gbe/evp012 PMID: 20333182

62. Markova-Raina P, Petrov D. High sensitivity to aligner and high rate of false positives in the estimates of positive selection in the 12 *Drosophila* genomes. Genome Res. 2011; 21: 863–874. https://doi.org/10.1101/gr.115949.110 PMID: 21393387

63. Yang Z, dos Reis M. Statistical properties of the branch-site test of positive selection. Mol Biol Evol. 2011; 28: 1217–1228. https://doi.org/10.1093/molbev/msq303 PMID: 21087944

64. Jordan G, Goldman N. The effects of alignment error and alignment filtering on the sitewise detection of positive selection. Mol Biol Evol. 2012; 29: 1125–1139. https://doi.org/10.1093/molbev/msr272 PMID: 22049066

65. Löytynoja A, Goldman N. Phylogeny-aware gap placement prevents errors in sequence alignment and evolutionary analysis. Science. 2008; 320: 1632–1635. https://doi.org/10.1126/science.1158395 PMID: 18566285

66. Sela I, Ashkenazy H, Katoh K, Pupko T. GUIDANCE2: accurate detection of unreliable alignment regions accounting for the uncertainty of multiple parameters. Nucleic Acids Res. 2015; 43: W7–14. https://doi.org/10.1093/nar/gkv318 PMID: 25883146

67. Castresana J. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol Biol Evol. 2000; 17: 540–552. PMID: 10742046

68. Yang Z. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol. 2007; 24: 1586–1591. https://doi.org/10.1093/molbev/msm088 PMID: 17483113

69. Yang Z, Wong WS, Nielsen R. Bayes empirical bayes inference of amino acid sites under positive selection. Mol Biol Evol. 2005; 22: 1107–1118. https://doi.org/10.1093/molbev/msi097 PMID: 15689528

70. Zhang J, Nielsen R, Yang Z. Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. Mol Biol Evol. 2005; 22: 2472–2479. https://doi.org/10.1093/molbev/msi237 PMID: 16107592

71. Nielsen R. Molecular signatures of natural selection. Annu Rev Genet. 2005; 39: 197–218. https://doi.org/10.1146/annurev.genet.39.073003.112420 PMID: 16285858

72. Venables WN, Smith DM. the R Development Core Team.(2008) An Introduction to R. Netw Theory Limited, Bristol. 2010;

73. Notredame C, Higgins DG, Heringa J. T-Coffee: A novel method for fast and accurate multiple sequence alignment. J Mol Biol. 2000; 302: 205–217. https://doi.org/10.1006/jmbi.2000.4042 PMID: 10964570

74. Wallace IM, O'Sullivan O, Higgins DG, Notredame C. M-Coffee: combining multiple sequence alignment methods with T-Coffee. Nucleic Acids Res. 2006; 34: 1692–1699. https://doi.org/10.1093/nar/gkl091 PMID: 16556910

75. Chang JM, Di Tommaso P, Notredame C. TCS: a new multiple sequence alignment reliability measure to estimate alignment accuracy and improve phylogenetic tree reconstruction. Mol Biol Evol. 2014; 31: 1625–1637. https://doi.org/10.1093/molbev/msu117 PMID: 24694831

76. Croucher PJ, Brewer MS, Winchell CJ, Oxford GS, Gillespie RG. 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. 2013; 14: 862. https://doi.org/10.1186/1471-2164-14-862 PMID: 24314324

77. He Q, Duan Z, Yu Y, Liu Z, Liu Z, Liang S. The venom gland transcriptome of *Latrodectus tredecimguttatus* revealed by deep sequencing and cDNA library analysis. PLoS One. 2013; 8: e81357. https://doi.org/10.1371/journal.pone.0081357 PMID: 24312294

78. Posnien N, Zeng V, Schwager EE, Pechmann M, Hilbrant M, Keefe JD, et al. A comprehensive reference transcriptome resource for the common house spider *Parasteatoda tepidariorum*. PLoS One. 2014; 9: e104885. https://doi.org/10.1371/journal.pone.0104885 PMID: 25118601

79. Riesgo A, Andrade SC, Sharma PP, Novo M, Perez-Porro AR, Vahtera V, et al. Comparative description of ten transcriptomes of newly sequenced invertebrates and efficiency estimation of genomic sampling in non-model taxa. Front Zool. 2012; 9: 33. https://doi.org/10.1186/1742-9994-9-33 PMID: 23190771

80. Tao X, Gu YH, Wang HY, Zheng W, Li X, Zhao CW, et al. Digital gene expression analysis based on integrated de novo transcriptome assembly of sweet potato [*Ipomoea batatas* (L.) Lam]. PLoS One. 2012; 7: e36234. https://doi.org/10.1371/journal.pone.0036234 PMID: 22558397

81. Harismendy O, Ng PC, Strausberg RL, Wang X, Stockwell TB, Beeson KY, et al. Evaluation of next generation sequencing platforms for population targeted sequencing studies. Genome Biol. 2009; 10: R32. https://doi.org/10.1186/gb-2009-10-3-r32 PMID: 19327155

82. Oyola SO, Otto TD, Gu Y, Maslen G, Manske M, Campino S, et al. Optimizing Illumina next-generation sequencing library preparation for extremely AT-biased genomes. BMC Genomics. 2012; 13: 1. https://doi.org/10.1186/1471-2164-13-1 PMID: 22214261

83. Niu JZ, Dou W, Ding TB, Shen GM, Zhang K, Smagghe G, et al. Transcriptome analysis of the citrus red mite, *Panonychus citri*, and its gene expression by exposure to insecticide/acaricide. Insect Mol Biol. 2012; 21: 422–436. https://doi.org/10.1111/j.1365-2583.2012.01148.x PMID: 22676046

84. Rendon-Anaya M, Delaye L, Possani LD, Herrera-Estrella A. Global transcriptome analysis of the scorpion *Centruroides noxius*: new toxin families and evolutionary insights from an ancestral scorpion species. PLoS One. 2012; 7: e43331. https://doi.org/10.1371/journal.pone.0043331 PMID: 22912855

85. Van Belleghem SM, Roelofs D, Van Houdt J, Hendrickx F. De novo transcriptome assembly and SNP discovery in the wing polymorphic salt marsh beetle *Pogonus chalceus* (Coleoptera, Carabidae). PLoS One. 2012; 7: e42605. https://doi.org/10.1371/journal.pone.0042605 PMID: 22870338

86. Wang JT, Li JT, Zhang XF, Sun XW. Transcriptome analysis reveals the time of the fourth round of genome duplication in common carp (*Cyprinus carpio*). BMC Genomics. 2012; 13: 96. https://doi.org/10.1186/1471-2164-13-96 PMID: 22424280

87. Grant JR, Lahr DJG, Rey FE, Burleigh JG, Gordon JI, Knight R, et al. Gene discovery from a pilot study of the transcriptomes from three diverse microbial eukaryotes: *Corallomyxa tenera*, *Chilodonella uncinata*, and *Subulatomonas tetraspora*. Protist Genomics. 2012; 1: 3–18.

88. Santoferrara LF, Guida S, Zhang H, McManus GB. De novo transcriptomes of a mixotrophic and a heterotrophic ciliate from marine plankton. PLoS One. 2014; 9: e101418. https://doi.org/10.1371/journal.pone.0101418 PMID: 24983246

89. Binns D, Dimmer E, Huntley R, Barrell D, O'Donovan C, Apweiler R. QuickGO: a web-based tool for Gene Ontology searching. Bioinformatics. 2009; 25: 3045–3046. https://doi.org/10.1093/bioinformatics/btp536 PMID: 19744993

90. Adams ME. Agatoxins: ion channel specific toxins from the American funnel web spider, *Agelenopsis aperta*. Toxicon. 2004; 43: 509–525. https://doi.org/10.1016/j.toxicon.2004.02.004 PMID: 15066410

91. Duan Z, Cao R, Jiang L, Liang S. A combined de novo protein sequencing and cDNA library approach to the venomic analysis of Chinese spider *Araneus ventricosus*. J Proteomics. 2013; 78: 416–427. https://doi.org/10.1016/j.jprot.2012.10.011 PMID: 23088928

92. Magrane M, Consortium U. UniProt Knowledgebase: a hub of integrated protein data. Database. 2011; 2011.

93. Gokudan S, Muta T, Tsuda R, Koori K, Kawahara T, Seki N, et al. Horseshoe crab acetyl group-recognizing lectins involved in innate immunity are structurally related to fibrinogen. Proc Natl Acad Sci U S A. 1999; 96: 10086–10091. PMID: 10468566

94. Wells L, Edwards KA, Bernstein SI. Myosin heavy chain isoforms regulate muscle function but not myofibril assembly. EMBO J. 1996; 15: 4454–4459. PMID: 8887536

95. Hooper SL, Thuma JB. Invertebrate muscles: muscle specific genes and proteins. Physiol Rev. 2005; 85: 1001–1060. https://doi.org/10.1152/physrev.00019.2004 PMID: 15987801

96. Arruda AP, Nigro M, Oliveira GM, de Meis L. Thermogenic activity of Ca2+-ATPase from skeletal muscle heavy sarcoplasmic reticulum: the role of ryanodine Ca2+ channel. Biochim Biophys Acta. 2007; 1768: 1498–1505. https://doi.org/10.1016/j.bbamem.2007.03.016 PMID: 17466935

97. Reese G, Ayuso R, Lehrer SB. Tropomyosin: an invertebrate pan-allergen. Int Arch Allergy Immunol. 1999; 119: 247–258. PMID: 10474029

98. Prpic NM, Damen WG. Expression patterns of leg genes in the mouthparts of the spider *Cupiennius salei* (Chelicerata: Arachnida). Dev Genes Evol. 2004; 214: 296–302. https://doi.org/10.1007/s00427-004-0393-5 PMID: 15014991

99. Voit R, Feldmaier-Fuchs G, Schweikardt T, Decker H, Burmester T. Complete sequence of the 24-mer hemocyanin of the tarantula *Eurypelma californicum*. Structure and intramolecular evolution of the subunits. J Biol Chem. 2000; 275: 39339–39344. https://doi.org/10.1074/jbc.M005442200 PMID: 10961996

100. Eberhart CG, Maines JZ, Wasserman SA. Meiotic cell cycle requirement for a fly homologue of human Deleted in Azoospermia. Nature. 1996; 381: 783–785. https://doi.org/10.1038/381783a0 PMID: 8657280

101. Karashima T, Sugimoto A, Yamamoto M. Caenorhabditis elegans homologue of the human azoospermia factor DAZ is required for oogenesis but not for spermatogenesis. Development. 2000; 127: 1069–1079. PMID: 10662646

102. Shah C, Vangompel MJ, Naeem V, Chen Y, Lee T, Angeloni N, et al. Widespread presence of human BOULE homologs among animals and conservation of their ancient reproductive function. PLOS Genet. 2010; 6: e1001022. https://doi.org/10.1371/journal.pgen.1001022 PMID: 20657660

103. Nery MF, Gonzalez DJ, Opazo JC. How to Make a Dolphin: Molecular Signature of Positive Selection in Cetacean Genome. PLoS One. 2013; 8: e65491. https://doi.org/10.1371/journal.pone.0065491 PMID: 23840335

104. Roux J, Privman E, Moretti S, Daub JT, Robinson-Rechavi M, Keller L. Patterns of positive selection in seven ant genomes. Mol Biol Evol. 2014; 31: 1661–1685. https://doi.org/10.1093/molbev/msu141 PMID: 24782441

105. Silva DN, Duplessis S, Talhinhas P, Azinheira H, Paulo OS, Batista D. Genomic Patterns of Positive Selection at the Origin of Rust Fungi. PLoS One. 2015; 10: e0143959. https://doi.org/10.1371/journal.pone.0143959 PMID: 26632820

106. Privman E, Penn O, Pupko T. Improving the performance of positive selection inference by filtering unreliable alignment regions. Mol Biol Evol. 2012; 29: 1–5. https://doi.org/10.1093/molbev/msr177 PMID: 21772063

107. Gharib WH, Robinson-Rechavi M. The branch-site test of positive selection is surprisingly robust but lacks power under synonymous substitution saturation and variation in GC. Mol Biol Evol. 2013; 30: 1675–1686. https://doi.org/10.1093/molbev/mst062 PMID: 23558341

108. Kosiol C, Vinar T, da Fonseca RR, Hubisz MJ, Bustamante CD, Nielsen R, et al. Patterns of positive selection in six Mammalian genomes. PLOS Genet. 2008; 4: e1000144. https://doi.org/10.1371/journal.pgen.1000144 PMID: 18670650

109. Anisimova M, Bielawski JP, Yang Z. Accuracy and Power of the Likelihood Ratio Test in Detecting Adaptive Molecular Evolution. Mol Biol Evol. 2001; 18: 1585–1592. Available: http://mbe.oxfordjournals.org/cgi/content/long/18/8/1585 PMID: 11470850

110. Wong WSW, Yang Z, Goldman N, Nielsen R. Accuracy and Power of Statistical Methods for Detecting Adaptive Evolution in Protein Coding Sequences and for Identifying Positively Selected Sites. Genetics. 2004; 168: 1041–1051. https://doi.org/10.1534/genetics.104.031153 PMID: 15514074

111. Khalturin K, Hemmrich G, Fraune S, Augustin R, Bosch TC. More than just orphans: are taxonomically-restricted genes important in evolution? Trends Genet. 2009; 25: 404–413. https://doi.org/10.1016/j.tig.2009.07.006 PMID: 19716618

112. Palmieri N, Kosiol C, Schlotterer C. The life cycle of *Drosophila* orphan genes. Elife. 2014; 3: e01311. https://doi.org/10.7554/eLife.01311 PMID: 24554240

113. Domazet-Loso T, Tautz D. An evolutionary analysis of orphan genes in *Drosophila*. Genome Res. 2003; 13: 2213–2219. https://doi.org/10.1101/gr.1311003 PMID: 14525923

114. Chen S, Krinsky BH, Long M. New genes as drivers of phenotypic evolution. Nat Rev Genet. 2013; 14: 645–660. https://doi.org/10.1038/nrg3521 PMID: 23949544

115. Zhao L, Saelao P, Jones CD, Begun DJ. Origin and spread of de novo genes in Drosophila melanogaster populations. Science (80-). 2014; 343: 769–772. https://doi.org/10.1126/science.1248286 PMID: 24457212

116. Kondrashov F a, Rogozin IB, Wolf YI, Koonin E V. Selection in the evolution of gene duplications. Genome Biol 3 Res. 2002; 3: 1–9.

117. Cai JJ, Petrov DA. Relaxed purifying selection and possibly high rate of adaptation in primate lineage-specific genes. Genome Biol Evol. 2010; 2: 393–409. https://doi.org/10.1093/gbe/evq019 PMID: 20624743