An examination of Chelicerate genomes reveals no evidence for a whole genome duplication among spiders and scorpions

Gregg W.C. Thomas[1], Michael T.W. McKibben[2], Matthew W. Hahn[3,4], Michael S. Barker[2]

[1]Informatics Group, Harvard University, Cambridge, MA, USA

[2]Department of Ecology & Evolutionary Biology, University of Arizona, Tucson, AZ, USA

[3]Department of Biology, Indiana University, Bloomington, IN, USA

[4]Department of Computer Science, Indiana University, Bloomington, IN, USA

Abstract

Whole genome duplications (WGDs) can be a key event in evolution, playing a role in both adaptation and speciation. While WGDs are common throughout the history of plants, only a few examples have been proposed in metazoans. Among these, recent proposals of WGD events in Chelicerates, the group of Arthropods that includes horseshoe crabs, ticks, scorpions, and spiders, include several rounds in the history of horseshoe crabs, with an additional WGD proposed in the ancestor of spiders and scorpions. However, many of these inferences are based on evidence from only a small portion of the genome (in particular, genes containing homeobox domains); therefore, genome-wide inferences with broader species sampling may give a clearer picture of WGDs in this clade. Here, we investigate signals of WGD in Chelicerates using whole genomes from 17 species. We employ multiple methods to look for these signals, including gene tree analysis of thousands of gene families, comparisons of synteny, and signals of divergence among within-species paralogs. We test several scenarios of WGD in Chelicerates using multiple species trees as a backbone for all hypotheses. While we do find support for at least one WGD in the ancestral horseshoe crab lineage, we find no evidence for a WGD in the history of spiders and scorpions using any genome-scale method. This study not only sheds light on genome evolution and phylogenetics within Chelicerates, but also demonstrates how a combination of comparative methods can be used to investigate signals of ancient WGDs.

Introduction

Whole genome duplications (WGDs) occur when an individual retains both sets of chromosomes from one or more parents. While such events are often highly deleterious, occasionally the combination of novel genetic material can provide advantages that allow the whole genome duplication to propagate, resulting in a polyploid species with more than 2n chromosomes in its genome. WGDs have been important evolutionary events, with some evidence pointing to an association between environmental stress and the success of polyploid species (Van de Peer, et al. 2021). WGDs are common in plants (Masterson 1994; Adams and Wendel 2005; Barker, et al. 2016; Initiative 2019), but there are also a smaller number of important genome duplications in the history of fungi (Wolfe and Shields 1997; Ma, et al. 2009) and vertebrates (Ohno 1970; Furlong and Holland 2002; McLysaght, et al. 2002).

Common processes in the evolution of polyploid species are diploidization, the reversion to disomic inheritance (Wolfe 2001), and fractionation,the loss of many of the excess genes and chromosomes that resulted from the WGD (Li, et al. 2021). The end result of these processes is a return of the gene-content of the polyploid species to a nearly diploid state, with most paralogous genes that resulted from the WGD being lost or unidentifiable as paralogs (Wolfe 2001). Nevertheless, even in paleopolyploid species that have had ancient WGDs and have undergone diploidization and fractionation, signatures of the WGD can remain in their genomes. For example, an excess of paralogs in the genome will have an origin that approximately coincides with the timing of the WGD. The timing of such events can be determined by multiple methods. One class of methods, generally referred to as gene tree-species tree reconciliation, uses gene tree topologies to map duplication events onto branches of the species tree (Pfeil, et al. 2005; Cannon, et al. 2015; Thomas, et al. 2017; Yan, et al. 2022). These topological methods can also potentially identify the

mode of polyploidy (Thomas, et al. 2017) and can more accurately identify independent WGDs when fractionation occurs during speciation (Redmond, et al. 2023). A second class of methods examines pairwise divergence between paralogs in the same species, with the expectation that a WGD event will lead to a peak of synonymous divergence (*K*s) between paralogs (Lynch and Conery 2000; Blanc and Wolfe 2004; Tiley, et al. 2018). Finally, there may also be syntenic evidence for the WGD in polyploids, where whole paralogous regions of the same genome (including both coding and non-coding sequence) trace their history to the WGD event (Tang, et al. 2008; Hao, et al. 2021).

Recently, WGDs have been proposed in the history of the Arthropod sub-phylum Chelicerata, which includes horseshoe crabs, sea spiders, mites, ticks, scorpions, and spiders. In horseshoe crabs, counts of gene duplications, paralog divergence estimates, and syntenic blocks all have been interpreted as a whole genome duplication (Nossa, et al. 2014; Shingate, Ravi, Prasad, Tay, Garg, et al. 2020). Examination of homeobox containing genes has also been used to suggest that there have been anywhere between one and three WGDs during the course of horseshoe crab evolution (Kenny, et al. 2016; Shingate, Ravi, Prasad, Tay, Garg, et al. 2020; Shingate, Ravi, Prasad, Tay and Venkatesh 2020). Similar approaches also form the basis for the claim that a WGD has occurred in the lineage ancestral to extant spiders and scorpions (Sharma, et al. 2014; Clarke, et al. 2015; Schwager, et al. 2017; Leite, et al. 2018; Fan, et al. 2021; Harper, et al. 2021; Aase-Remedios, et al. 2023). In both cases, the number of genes or genomes used for analysis has been limited. In addition, while the duplication of conserved gene clusters (i.e. the those containing homeobox sequences) may be indicative of a larger (perhaps whole genome) duplication event, it is too limited a dataset with which to confirm such an event (Noah, et al. 2020). As well as issues with the amount of data used for inferences, recent evidence supports an

alternate placement of horseshoe crabs in the chelicerate phylogeny. Traditionally, the aquatic horseshoe crabs have been thought to be sister to all arachnids (spiders, scorpions, mites, and ticks), which are mostly terrestrial (Weygoldt and Paulus 1979). However, the possibility of polyphyletic origins of arachnids has been considered (see Shultz 1990) and some molecular studies have supported a scenario of polyphyletic arachnids (Sharma, et al. 2014; Ballesteros and Sharma 2019; Noah, et al. 2020; Ontano, et al. 2021). Recently, Ballesteros, et al. (2022) presented strong evidence for horseshoe crabs being nested within arachnids, making arachnids polyphyletic. While the placement of horseshoe crabs tends to be highly dependent on the species sampling and alignment used (Ballesteros and Sharma 2019; Ontano, et al. 2021; Ballesteros, et al. 2022), this newly proposed species tree could substantially impact how WGDs are inferred within this group when phylogenetic methods are used (Noah, et al. 2020; McKibben, et al. 2024).

Here, we use whole-genome sequences from 17 chelicerate species, in combination with several different analytical methods, to look for ancient WGDs in this group. These methods include gene tree reconciliation, synonymous divergence between paralogs, and whole-genome analyses of synteny. Using multiple species trees as a backbone for analysis, we find no evidence for a WGD taking place in the history of spiders and scorpions. In contrast, our suite of methods all find some evidence for at least one WGD occurring during the evolution of horseshoe crabs, even in light of their possible new placement in the chelicerate phylogeny.

Methods

Data

To investigate the possible existence of whole genome duplication (WGD) events in chelicerates on a genome-wide scale, we took a multi-faceted approach. We initially downloaded 18 chelicerate

genomes with annotations available at the beginning of this project from various sources: NCBI's Assembly database (https://www.ncbi.nlm.nih.gov/assembly) Ensembl Metazoa (Yates, et al. 2022; release 51), the i5k database (Consortium 2013; Thomas, et al. 2020), and, for two samples, the data supplements of their genome publications (Fan, et al. 2021; Nong, et al. 2021). These genomes span the various taxonomic groups contained within the subphylum Chelicerata, including four species from the superorder Parasitiformes (mites and ticks), two species from the superorder Acariformes (mites), eight species from the order Araneae (spiders), one species from the order Scorpiones (scorpions), and four species from the order Xiphosura (horseshoe crabs) (Fig. 1). For this study, we treat Parasitiformes and Acariformes as orders. For phylogenetic analyses, we also include two insects (*Drosophila melanogaster* and *Bombyx mori*) as outgroups for tree rooting. See Supplemental Table S1 for full details of the samples and summaries of their assemblies and annotations.

We observed that annotations of one of the horseshoe crabs, *Tachypleus tridentatus*, contained 79,557 genes, more than twice as many as any other species in our sample, including the other horseshoe crabs. While on the surface this may indeed be indicative of a recent WGD in this species, we also note that the median gene length for this species is only 1,377 bp. While this is not the shortest gene length in our sample, it is considerably smaller than the rest of the horseshoe crabs, which all have a median gene length of over 8,500 bp (see Supplemental Table S1). Because this could be indicative of annotation error in this species and because we are interested in ancient rather than recent WGDs, we excluded this sample from our analyses. In total, our final dataset contained 17 chelicerate species and 2 outgroup insects for analyses that span almost 600 million years of genome evolution.

Gene tree reconciliation analysis

We extracted the coding sequence of the longest transcript from each gene in each of our 19 species and used FastOrtho (https://github.com/olsonanl/FastOrtho), which is a reimplementation of orthomcl (Li, et al. 2003), to cluster genes into gene families. Using an inflation value of 3, we inferred 49,561 gene families. We then extracted the sequences in each gene family and aligned each gene family with Guidance2 (Sela, et al. 2015) using MAFFT (Katoh and Standley 2013) as the underlying aligner, and removing any alignment columns with a score below 0.93. We also performed our own alignment filtering by removing columns in sliding windows of 3 codons that have 2 codons with 2 or more gaps in 50% of the sequences in that alignment. We also removed any sequences that were made up of greater than 20% gap characters and removed any alignments with sequences from fewer than 4 species or that were shorter than 33 codons after all filtering. See Supplementary Table S2 for alignment filtering details.

We translated the remaining 11,016 alignments from nucleotides to amino acids and inferred gene trees with IQ-TREE (Nguyen, et al. 2015) using ultrafast bootstrap (Hoang, et al. 2018); the gene trees were used to infer a species tree with ASTRAL-Multi (Rabiee, et al. 2019). For subsequent reconciliation analyses, we rooted our gene and species trees using the outgroup insects with Newick Utilities (nw_reroot; Junier and Zdobnov 2010). Gene trees that could not be rooted because there was no outgroup were excluded from reconciliation analyses. After rooting, we retained gene trees from 6,368 gene families. To further reduce possible gene tree inference error, we used bootstrap rearrangement implemented in Notung (Chen, et al. 2000) with a bootstrap threshold of 90. This method forces inferred duplications on branches in our gene trees with a bootstrap score less than this threshold to be resolved in such a way that minimizes the number of duplications and losses counted in the tree. We also ran our reconciliation analyses with a bootstrap threshold of 80 and with no bootstrap threshold.

We used these 6,368 rooted, bootstrap-resolved gene trees and a species tree as input to GRAMPA (Thomas, et al. 2017) to identify the placement of any WGDs in the chelicerate phylogeny. Briefly, GRAMPA performs least common ancestor (LCA) mapping from each gene tree to the species tree but allows for WGDs to be present in the species tree by representing them as multi-labeled trees (MUL-trees), in which one or more tip labels appear twice. By comparing LCA mapping scores between the input species tree and a set of MUL-trees defined by target lineages, GRAMPA can determine if a WGD has occurred on a hypothesized lineage, and the shape of the MUL-trees tested allows it to distinguish between allo- and auto-polyploidy. Importantly, tandem duplications do not affect GRAMPA's inferences since they will be spread across the branches in the input species tree, making this method suitable for detecting even ancient WGDs. For our runs, we set as target lineages for WGD identification those on which WGDs have previously been proposed: specifically, the branch leading to spiders and scorpions and the branch leading to horseshoe crabs. We also used multiple different species trees as input to GRAMPA to test the same scenarios. In addition to the species tree we inferred using ASTRAL (Fig. 1A), we tested for WGDs on two alternate species tree topologies. One alternate topology is based on a recently inferred phylogeny from Ballesteros, et al. (2022) in which horseshoe crabs group within arachnids (Fig. 1B). Because some molecular studies still propose a monophyletic Arachnida that does not include horseshoe crabs (Sharma, et al. 2014; Lozano-Fernandez, et al. 2019; Howard, et al. 2020), we also used a 'traditional' species tree topology, in which horseshoe crabs are sister to all arachnid species (Fig. 1C). For the 'traditional' tree, because of the unresolved placement of Acariformes and Parasitiformes (Sharma, et al. 2014; Ontano, et al. 2021), we simply use the topology recovered by Ballesteros, et al. (2022, their Figure 2A) and manually placed horseshoe crabs sister to arachnids.

Synteny analysis

We used multiple synteny-based methods to detect signatures of ancient WGDs across the 19 assemblies in our analyses. We estimated inter- and intraspecific synteny using MCScanX (Wang, et al. 2012) and the top five hits from an all-against-all BLAST (Camacho, et al. 2009). We used the default settings of MCScanX to detect and visualize collinear blocks. Given that ancient WGDs may be highly fractionated, we also relaxed the minimum block size from five to three genes and increased the maximum gaps allowed from 20 to 50 genes. These settings allow us to recover potentially highly fragmented blocks of synteny. In addition, we used *synmap.pl* from CoGe as an alternative method for syntenic block detection (Haug-Baltzell, et al. 2017). WGDs can also be detected using interspecific comparisons with an outgroup species that does not share the hypothesized WGD, which would be evident in the form of double conserved syntenic blocks. To capture this signal, we used the relaxed settings in MCScanX to compare *P. tepidariorum* to *T. urticae*.

Prior analyses also used SatsumaSynteny to recover gene clusters containing homeobox domains that were duplicated and resided in syntenic blocks with in *P. tepidariorum*. (Schwager, et al. 2017). To compare these analyses to our inferences of synteny, we use reciprocal best BLAST hits to find homologs of the homeobox clusters in the *P. tepidariorum* assembly. We then assessed whether these homeobox gene clusters reside in the intra- and interspecific syntenic blocks from our analyses, and we compared their gene classifications to those reported in Schwager, et al. (2017). Further, as MCScanX can mask tandem duplications when detecting collinearity, we manually compared the locations of homeobox containing gene clusters to those reported in Schwager, et al. (2017).

Synonymous divergence between paralogs (K_S*)*

To construct gene families and to estimate the age distribution of gene duplications we used the DupPipe pipeline (Barker, et al. 2008; Barker, et al. 2010). Briefly, DupPipe translates coding transcripts from nucleotide to peptide sequences and identifies reading frames by comparing Genewise (Birney, et al. 2004) alignments to the best-hit protein from a collection of proteins from the 19 sampled genomes. For all DupPipe runs, we used protein-guided DNA alignments to align our nucleic acid sequences while maintaining the reading frame. We estimated synonymous divergence (K_S) using PAML (Yang 2007) with the F3X4 model for each node in the gene-family phylogenies. We identified peaks of gene duplication as evidence for potential ancient WGDs in histograms of the age distribution of gene duplications (K_S plots). To infer ancient WGDs in the paralog age distributions we used a recently developed machine learning approach, SLEDGe (Sutherland, et al. 2024), to classify K_S plots with peaks consistent with an ancient WGD. Specifically, we applied the support vector machine classifier from SLEDGe on node K_S-values for species that had greater than 1,500 gene duplicates, subsampling down to 3,000 duplicates when more than 3,000 were present. For each K_S distribution that SLEDGe predicted as being indicative of a WGD, we also used mixture modeling and manual curation to identify significant peaks of gene duplication consistent with a WGD and to estimate their median paralog K_S values. We ran normalmixEM for a maximum of 400 iterations to fit the maximum number of *k*-components for each K_S distribution selected from a likelihood ratio test available in the boot.comp function from the mixtools R library (Benaglia, et al. 2009). Finally, to assess if WGD peaks in the paralog K_S distributions were shared between species, we used OrthoPipe from EvoPipes (Barker, et al. 2008; Barker, et al. 2010) to identify orthologs between species and PAML (Yang 2007) to estimate their K_S values using the same procedure and protein database as described for the DupPipe analyses. We then assessed species divergence by estimating the

median K_S of all orthologs with a K_S of 5 or lower for each species pair and compared to the median K_S of each WGD peak.

Results

Inference of the species tree

We used the genomes of 17 chelicerates and 2 insect outgroups to reconstruct the Chelicerata phylogeny, with an emphasis on Arachnids and horseshoe crabs. Using 11,016 gene trees we confirm the placement of Xiphosura (horseshoe crabs) as nested within Arachnids (Fig. 1A), in agreement with Ballesteros et al. (Fig 1B; Ballesteros, et al. 2022). However, our inferred tree differs from theirs in the placement of the superorders Acariformes and Parasitiformes. Our results show that Acariformes is sister to the spider, scorpion, and horseshoe crab clade, while Ballesteros et al. (2022) suggest that Parasitiformes is more closely related to them. However, the placement of these groups is also ambiguous in their analyses and has been contentious in previous studies (Sharma, et al. 2014; Ontano, et al. 2021).

Reconciliation analysis

We used the inferred species tree, as well as two other hypothesized sets of relationships, to test various hypotheses of WGD in the history of chelicerate evolution. Specifically, based on synteny and duplication of some gene families, multiple rounds of WGD have been proposed in horseshoe crabs (Nossa, et al. 2014; Kenny, et al. 2016; Shingate, Ravi, Prasad, Tay, Garg, et al. 2020; Shingate, Ravi, Prasad, Tay and Venkatesh 2020), and, based on the duplication of genes containing homeobox domains, one WGD has been proposed in the ancestor of spiders and scorpions (Schwager, et al. 2017). Using gene tree topologies from thousands of genes, GRAMPA (Thomas, et al. 2017) finds no evidence for a WGD in the history of spiders and scorpions using

either our inferred species tree, the one based on the Ballesteros et al. (2022) species tree, or the traditional species tree in which horseshoe crabs are sister to Arachnids (Figs. 1 and 2). In each case, we tested whether the species tree with a WGD proposed on any of the target lineages (H1 lineages in Fig. 1) better explains the duplication history of the genes in these genomes than a species tree with no proposed WGDs. However, in each case we find that the species tree without any proposed WGDs results in the lowest duplication and loss score (black shapes in Fig. 2). Our evidence is definitive for any WGD in the history of spiders and scorpions; however, we do see evidence for a large number of duplications on the branch leading to horseshoe crabs regardless of the species tree used (Fig. 1). We also find that the second- and third-lowest scoring scenarios when using our inferred species tree posit a WGD in horseshoe crabs (Fig. 2, Supplemental Table S3, Fig. S1). The horseshoe crab clade is also often inferred as being involved in a WGD in the next lowest scoring MUL-trees when using the other two species trees, but usually in more complicated scenarios (Figs. S1 and S2; Supplemental Tables S4 and S5). That is, while GRAMPA did not find a WGD in the history of horseshoe crabs as the single most parsimonious reconciliation, there are multiple pieces of evidence that point to one or more possibly occurring. Our results are consistent when using a lower bootstrap rearrangement threshold of 80 (Supplemental Table S6); with no bootstrap threshold, we infer allopolyploid scenarios that require unrealistic hybridizations (e.g. between horseshoe crabs and mites, leading to the rise of modern spiders and scorpions; Supplemental Table S7).

We also find that, when comparing reconciliation scores between species trees, our species tree and the Ballesteros et al. (2022) species tree both explain the history of gene duplication and loss better than the ‘traditional’ species tree in which horseshoe crabs are not nested within Arachnids (Fig. 2). This is further evidence in favor of the placement of this group within

Arachnida. While our species tree always better explains the data from rooted gene trees than Ballesteros et al. (2002), this should not be surprising since we inferred our tree from a superset of these data (both rooted and unrooted gene trees).

Synteny and K_S analyses

We next looked at other genome-wide signatures of WGDs among chelicerates. Specifically, we looked for intraspecific synteny blocks, which should be widespread in genomes that have undergone WGD, and distributions of synonymous divergence (K_S) of paralogs within each genome. If a WGD has occurred in the history of a genome, a secondary peak of K_S should be present in these distributions. Across both analyses, we again find no evidence for WGD in any spider or scorpion genomes but do find suggestive evidence for at least one occurring in the history of horseshoe crabs (Fig. 3). Only two species, *C. rotundicauda* and *T. gigas*, both horseshoe crabs, showed substantial amounts of intraspecific synteny. Both of these species, along with the other horseshoe crab, *L. polyphemus*, were also predicted by SLEDGe to have signatures of WGD in their K_S distributions (Fig. 3, Supplemental Table S8). Mixture models placed the median K_S of this duplication at ~0.85-1.35 (Fig. 3, Supplemental Table S8). The average ortholog divergence between the three horseshoe crabs was ~0.22, compared to the average divergence with *C. sculpturatus* at ~4.09, suggesting the WGD peak corresponds to the same branch identified with an excess number of gene duplications and losses in our gene tree topology reconciliation analysis above (Fig. 1, Fig. 3, Supplemental Table S9). In addition, one mite species, *Tetranychus urticae*, was predicted by SLEDGe to contain a WGD in its K_S distribution (Fig. 3). However, this species had few intraspecific syntenic blocks (Fig. 3; Supplemental Table S8) and no signal of excess duplication in the reconciliation analysis (Fig. 1). It is likely that the prediction made by SLEDGe in *T. urticae* is an artefact of assembly or annotation in this species.

Prior analyses by Schwager, et al. (2017) showed evidence that genes containing homeobox sequences were frequently duplicated in *P. tepidariorum*, a potential signature of WGD (see Discussion). Of the 145 homeobox gene clusters identified by Schwager, et al. (2017), we were able to detect the homologs of 105 in the *P. tepidariorum* assembly, 102 of which had 100% identity and coverage (Table S10). None of these homeobox genes were present in intraspecific syntenic blocs, regardless of method used (MCScanX defaults, MCScanX relaxed settings, snymap.pl). Rather, MCScanX labeled one homeobox homolog as a singleton, 76 as dispersed, 11 as proximal, and 21 as tandem duplicates (Table S10). Schwager, et al. (2017) reported similar results, however they also reported that a subset of these genes (namely *Lab, Pb, Hox3, Dfd, Scr, ftz, Antp, Ubx, adbA,* and *adbB*) were found in syntenic blocks detected by SatsumaSynteny, a different synteny program. Among these genes and their paralogs, we identified 13 in the *P. tepidariorum* assembly, 10 of which were annotated as tandem duplicates by MCScanX, a gene class masked during the collinearity detection process. To assess these homeobox genes in more detail, we manually compared their locations in Schwager, et al. (2017) to the *P. tepidariorum* assembly. Our results were similar to Schwager, et al. (2017), with *Scre, fts, Antp, Ubx, adbA,* and *adbB* found on the same scaffold; however, the remaining paralogs were located on five different scaffolds (Table S10). To further check if these genes are syntenic, and to better account for assembly quality, we also used relaxed settings in MCScanX to make interspecific syntenic inferences against *T. urticae* (see online data repository). Although we detected 248 collinear genes, none of the homeobox gene clusters were found in double conserved syntenic blocks (Table S10).

Discussion

Whole genome duplications (WGDs) can be a key event in the evolution of a species, possibly facilitating adaptation (Ohno 1970; Werth and Windham 1991; Adams and Wendel 2005; Crow and Wagner 2006). While prolonged processes of diploidization and fractionation can make more ancient WGDs harder to detect, multiple methods have been developed that have the potential to capture the signal of these events in extant genomes. Here, we used several of these methods to investigate the existence of ancient WGDs in the Chelicerates (Nossa, et al. 2014; Kenny, et al. 2016; Shingate, Ravi, Prasad, Tay, Garg, et al. 2020; Shingate, Ravi, Prasad, Tay and Venkatesh 2020). Several rounds of WGD have been proposed in the history of horseshoe crab evolution, and a single WGD has been proposed in the ancestor of spiders and scorpions (Sharma, et al. 2014; Clarke, et al. 2015; Schwager, et al. 2017; Leite, et al. 2018; Fan, et al. 2021; Harper, et al. 2021; Aase-Remedios, et al. 2023). The evidence for these events usually starts with the observation of the duplication of well-conserved gene family clusters, namely those containing homeobox domains. Further investigations of inter- and intraspecific synteny, gene tree topologies, and divergence have also been used previously, but until now have been limited to only a few genes or genomes.

Using 17 chelicerate whole genomes we find no evidence for a WGD in the history of spiders and scorpions. When reconciling gene tree topologies to a species tree that allows for the inference of WGDs, the best-scoring scenario is always the one without any WGDs, regardless of the input species tree topology used. For spiders and scorpions, we also see no excess intraspecific synteny or peaks in divergence of paralogs that would indicate a WGD. In contrast, all three methods find support for the widely recognized WGD in the history of horseshoe crabs.

It is possible that signatures of an ancient WGD in spiders and scorpions have been eroded by extensive fractionation and are additionally difficult to detect due to assembly quality. However,

a reexamination of data from previous papers finds that there was ambiguous support for a WGD within these as well. In a prior analysis, 10 Hox genes in a cluster were found to be duplicated, with a subset residing in syntenic blocks detected by SatsumaSynteny (Schwager, et al. 2017). Here, MCScanX and *synmap.pl* were not able to recover these synteny relationships, regardless of input parameters. Similarly, in our analyses none of the homeobox gene clusters were found in double conserved synteny with an outgroup. In addition to the Hox cluster, a number of other homeobox genes were found as duplicates by Schwager et al. (2017). MCScanX here labeled the majority of these homeobox genes as tandem duplicates, as in the original analyses. Leite et al. (2018) and Harper et al. (2021) similarly found many homeobox genes to be duplicates in spiders and scorpions, but no methods classified them as due to a WGD in those studies. Manual comparison of the relative locations of these genes in the annotation of *P. tepidariorum* here showed one cluster of the homeobox genes on a single scaffold, with the remaining paralogs scattered across five other scaffolds. These results may imply that the duplicated homeobox genes observed in some spiders and scorpions are the result of small-scale duplications. While homeobox gene clusters are thought to be relatively slowly evolving outside of WGDs, this is not always the case (Mulhair, et al. 2023; Mulhair and Holland 2024). Alternatively, collinear homeobox genes may be the only remaining signature of a shared WGD. However, in most cases duplicated homeobox genes are not taken alone as definitive evidence for a WGD (e.g. Amores, et al. 1998; Farhat, et al. 2023).

We do find evidence for WGDs during horseshoe crab evolution. While no MUL-trees are the single-most optimal solution in the gene tree analysis, we do find a burst of gene duplications on the branch leading to horseshoe crabs. This burst is observed regardless of the species tree considered (Fig. 1). Previously, anywhere from one to three WGDs have been proposed along the

horseshoe crab lineage. In fact, if multiple WGDs occurred, this may diminish the signal for any single proposed MUL-tree. Since our tests using GRAMPA are limited to a single MUL-tree, this may in turn hinder our ability to explicitly identify any single WGD as the most parsimonious scenario. In addition to the large number of duplications on the horseshoe crab lineage, we also observe notable intraspecific synteny and peaks in divergence of paralogs (Fig. 3).

In the course of our study of WGDs in Chelicerates, we also reconstructed a species tree for our 17 species (Fig. 1A). Using our whole genome data and including paralogs in our species tree inference (cf. Smith and Hahn 2021), we find that the horseshoe crabs (Xiphosura) are nested within Arachnids, though our species sampling prevents determining their placement with a higher resolution. This agrees with several recent molecular phylogenies of this group (Sharma, et al. 2014; Ballesteros and Sharma 2019; Noah, et al. 2020; Ontano, et al. 2021; Ballesteros, et al. 2022), and rejects a tree suggested by the biomes in which the organisms live, where the aquatic horseshoe crabs are nested within the mostly terrestrial arachnids (Fig. 1C). In this traditional monophyletic Arachnid tree, separate WGDs would need to be proposed for both spiders/scorpions and horseshoe crabs. However, the molecular trees allow the possibility that a single WGD took place in the ancestor of spiders, scorpions, and horseshoe crabs if they form a monophyletic group (Noah, et al. 2020). We also tested this scenario (Fig. 1A) and were able to rule out this possibility.

Our work shows that, even for ancient polyploids, whole genome comparative evidence can still find signals of WGDs. While the duplication of a single gene family can be a good initial clue that a WGD has occurred, as it was for vertebrates (Amores, et al. 1998), whole genome evidence is still needed for a more confident inference (Furlong and Holland 2002; McLysaght, et al. 2002; Hokamp, et al. 2003; Dehal and Boore 2005; Noah, et al. 2020). Our work shows that this is also the case for Chelicerates. In horseshoe crabs, duplications in homeobox containing gene

clusters coincide with synteny, peaks of synonymous divergence in intraspecific paralogs, and gene duplication reconciliation in the Chelicerate phylogeny. None of these additional pieces of evidence is present in the lineage leading to spiders and scorpions. Our work also adds to the growing body of evidence that horseshoe crabs are not sister to all arachnids as was traditionally thought, but rather are placed within the arachnid group.

Data availability

The genomes used in our analyses are available from their respective databases (see Supplemental Table S1). All other data generated for this project (gene alignments, gene trees, etc.) and scripts to parse and analyze it are available at https://github.com/gwct/spider-wgd.

Acknowledgements

We thank Zheng Li for helpful discussions on our analyses. Gene family analysis was performed on the FASRC Cannon cluster supported by the FAS Division of Science Research Computing Group at Harvard University. M.W.H. was supported by National Science Foundation grant DEB-1936187.

Figures

Figure 1

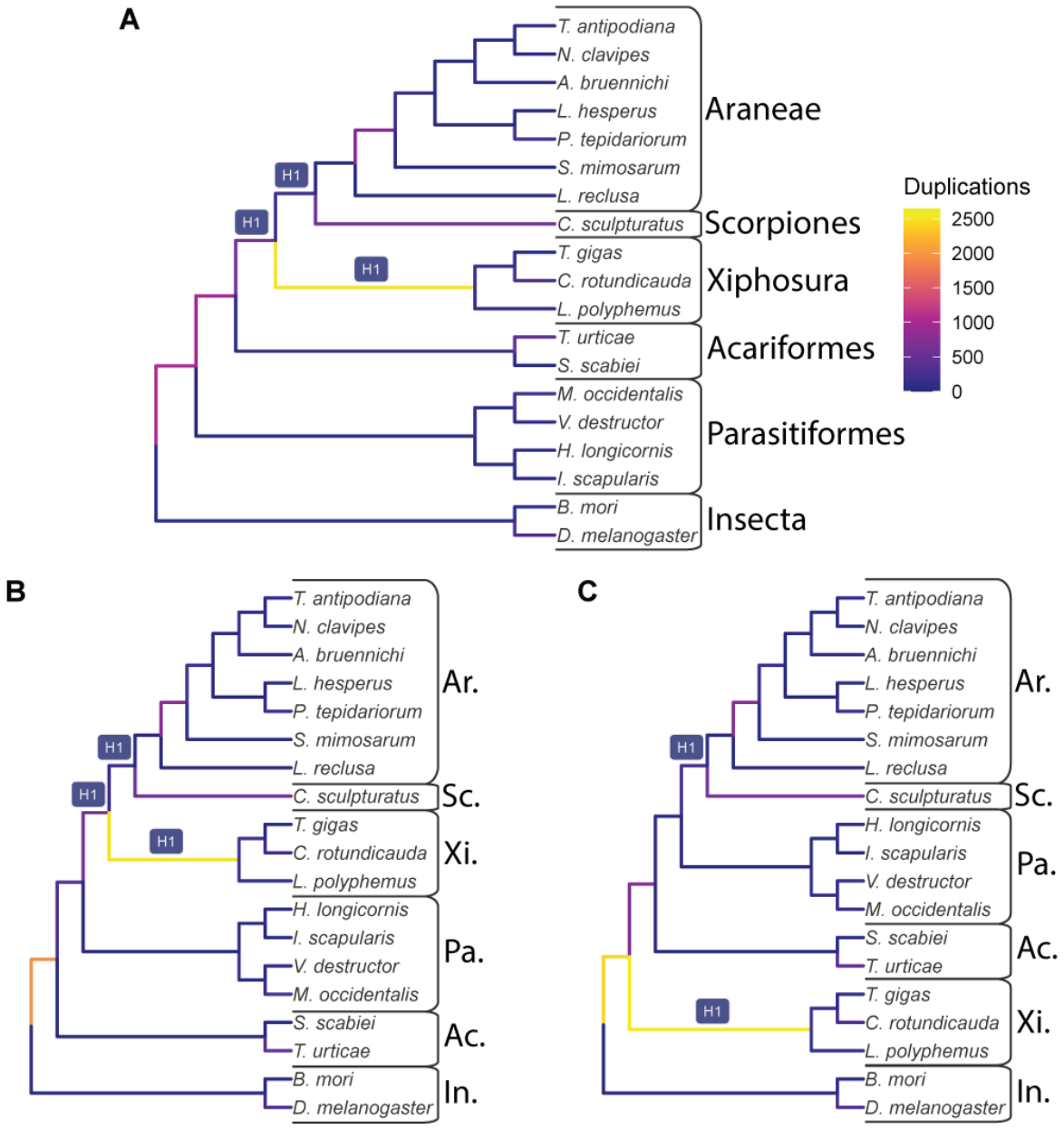


Figure 1: The input species trees used with GRAMPA, which are also the lowest scoring trees when considering possible WGDs at the branches labeled H1. Branches are shaded by the number of duplications that map to them. A) The species tree topology inferred in this study

from 11,016 gene families. B) The species tree inferred by Ballesteros, et al. (2022). C) A species tree that places horseshoe crabs (Xiphosura) sister to Arachnids. For all B and C, taxonomic groups are labeled as follows: Ar. = Araneae (spiders); Sc. = Scorpiones (scorpions); Xi. = Xiphosura (horseshoe crabs); Ac. = Acariformes (mites); Pa. = Parasitiformes (mites and ticks); In. = Insecta (insects).

Figure 2

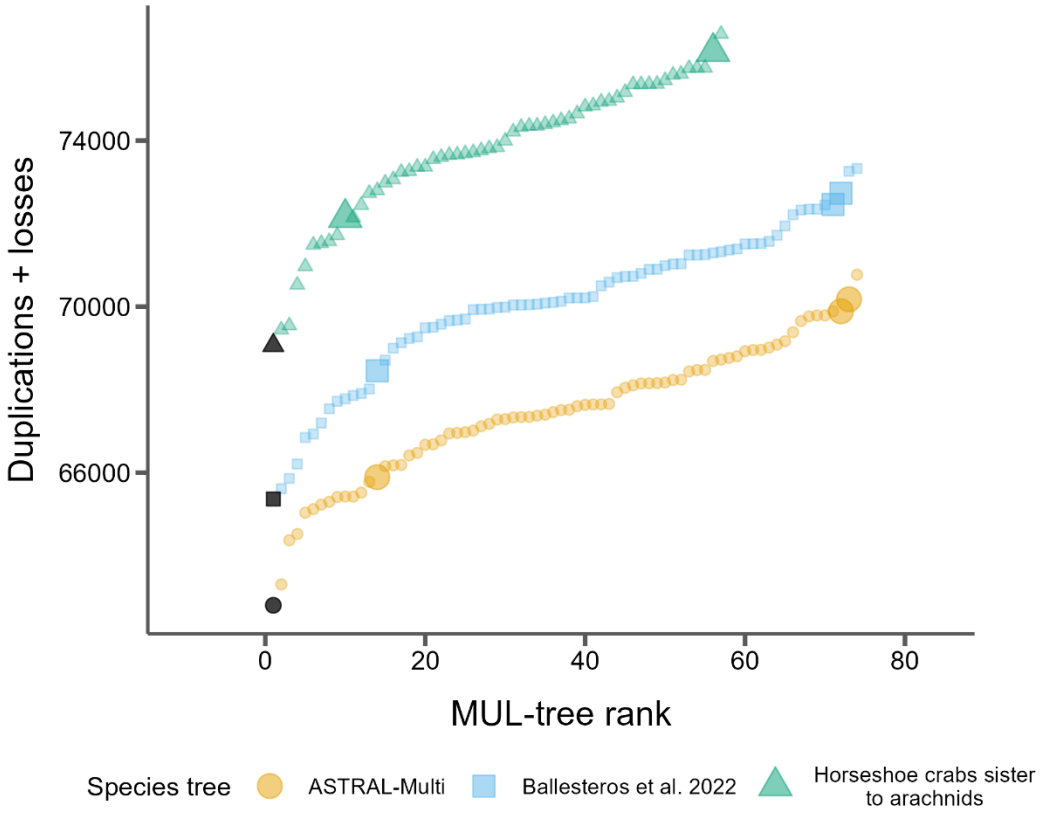
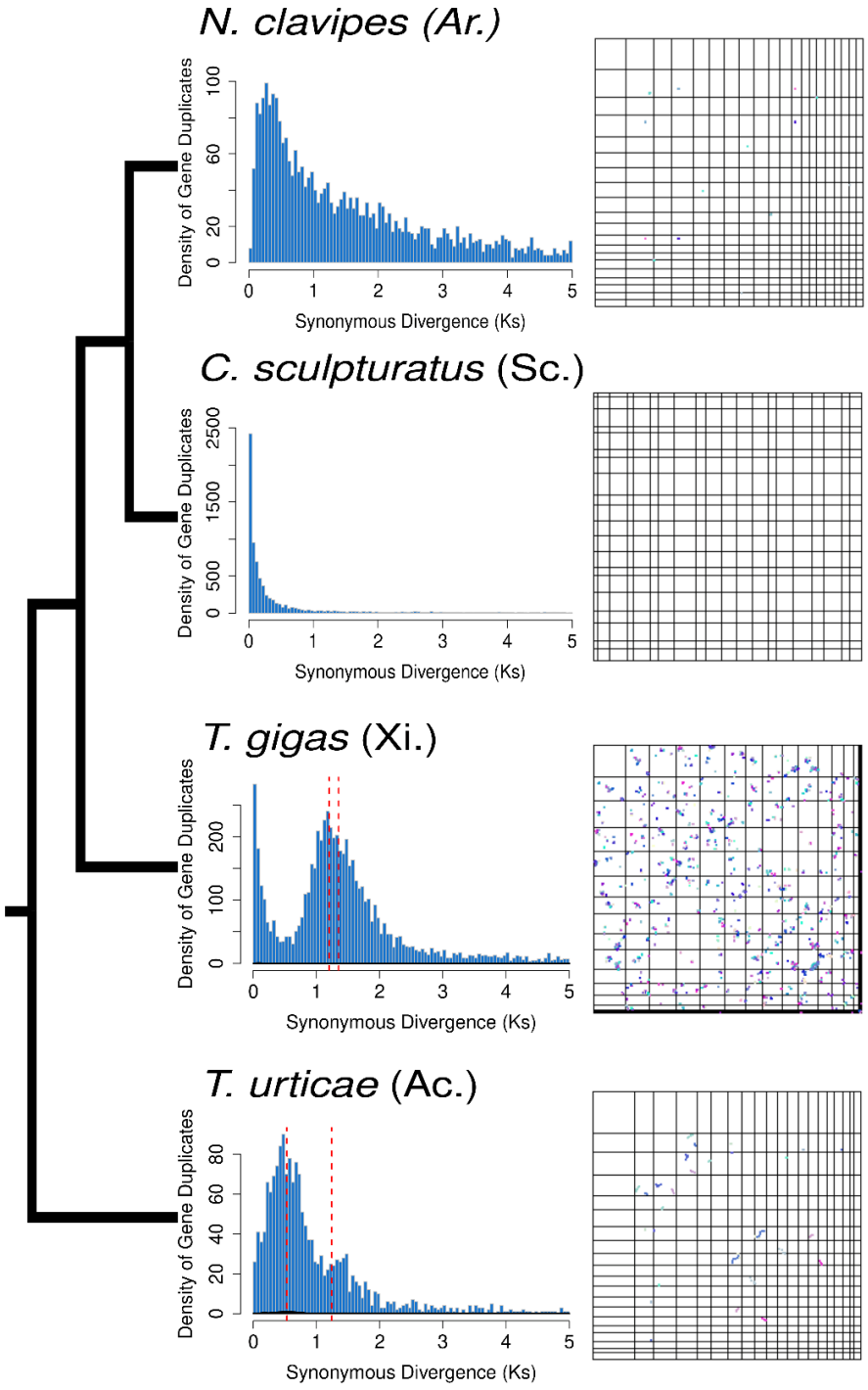


Figure 2: GRAMPA scores (duplications + losses) for every MUL-tree considered for each of the three species trees. Black points represent the input singly-labeled species tree with no WGD proposed. All other shaded points propose one WGD on one of the target H1 branches (see Fig. 1). Larger points indicate autopolyploidy scenarios and smaller dots indicate allopolyploidy scenarios.

405 *Figure 3*



406

407 **Figure 3**: Distributions of *K*s (left) and synteny (right) for select samples (See Figs. S5 and S6

408 for all samples) from Acariformes (Ac.), Xiphosura (Xi.), Araneae (Ar.) and Scorpiones (Sc.).

These samples all showed the highest levels of synteny among samples in each group. The species tree topology is shown on the far left. Red dotted lines indicate the median Ks of mixture models fit to distributions that were predicted by SLEDGe to be indicative of WGDs.

Supplemental Figure Legends

Figure S1

The lowest scoring MUL-trees from the GRAMPA analysis using our inferred species tree.

Figure S2

The lowest scoring MUL-trees from the GRAMPA analysis using the Ballesteros, et al. (2022) species tree.

Figure S3

The lowest scoring MUL-trees from the GRAMPA analysis using a traditional species tree with horseshoe crabs sister to arachnids.

Figure S4

Dot plots showing intra-species synteny for all species (19 panels, labeled with species name) with a max block size of 3.

Figure S5

Dot plots showing intra-species synteny for all species (19 panels, labeled with species name) with a max block size of 5.

Figure S6

Distributions of K_s between paralogs of all species (19 panels, labeled with species name). Dashed red lines indicate the median K_s of mixture models fit to each K_s distribution that was indicative of a WGD.

References

Aase-Remedios ME, Janssen R, Leite DJ, Sumner-Rooney L, McGregor AP. 2023. Evolution of the spider Homeobox gene repertoire by tandem and whole genome duplication. Molecular Biology and Evolution 40:msad239.

Adams KL, Wendel JF. 2005. Polyploidy and genome evolution in plants. Current Opinion in Plant Biology 8:135-141.

Amores A, Force A, Yan YL, Joly L, Amemiya C, Fritz A, Ho RK, Langeland J, Prince V, Wang YL, et al. 1998. Zebrafish hox clusters and vertebrate genome evolution. Science 282:1711-1714.

Ballesteros JA, Santibanez-Lopez CE, Baker CM, Benavides LR, Cunha TJ, Gainett G, Ontano AZ, Setton EVW, Arango CP, Gavish-Regev E, et al. 2022. Comprehensive species sampling and sophisticated algorithmic approaches refute the monophyly of Arachnida. Molecular Biology and Evolution 39:msac021.

Ballesteros JA, Sharma PP. 2019. A critical appraisal of the placement of Xiphosura (Chelicerata) with account of known sources of phylogenetic error. Systematic Biology 68:896-917.

Barker MS, Arrigo N, Baniaga AE, Li Z, Levin DA. 2016. On the relative abundance of autopolyploids and allopolyploids. New Phytologist 210:391-398.

Barker MS, Dlugosch KM, Dinh L, Challa RS, Kane NC, King MG, Rieseberg LH. 2010. EvoPipes.net: Bioinformatic tools for ecological and evolutionary genomics. Evolutionary Bioinformatics Online 6:143-149.

Barker MS, Kane NC, Matvienko M, Kozik A, Michelmore RW, Knapp SJ, Rieseberg LH. 2008. Multiple paleopolyploidizations during the evolution of the Compositae reveal parallel patterns of duplicate gene retention after millions of years. Molecular Biology and Evolution 25:2445-2455.

Benaglia T, Chauveau D, Hunter DR, Young DS. 2009. mixtools: An R package for analyzing mixture models. Journal of Statistical Software 32:1 - 29.

Birney E, Clamp M, Durbin R. 2004. GeneWise and Genomewise. Genome Research 14:988-995.

Blanc G, Wolfe KH. 2004. Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes. Plant Cell 16:1667-1678.

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

Cannon SB, McKain MR, Harkess A, Nelson MN, Dash S, Deyholos MK, Peng Y, Joyce B, Stewart CN, Jr., Rolf M, et al. 2015. Multiple polyploidy events in the early radiation of nodulating and nonnodulating legumes. Molecular Biology and Evolution 32:193-210.

Chen K, Durand D, Farach-Colton M. 2000. NOTUNG: a program for dating gene duplications and optimizing gene family trees. Journal of Computational Biology 7:429-447.

Clarke TH, Garb JE, Hayashi CY, Arensburger P, Ayoub NA. 2015. Spider transcriptomes identify ancient large-scale gene duplication event potentially important in silk gland evolution. Genome Biology and Evolution 7:1856-1870.

Consortium iK. 2013. The i5K Initiative: advancing arthropod genomics for knowledge, human health, agriculture, and the environment. Journal of Heredity 104:595-600.

Crow KD, Wagner GP. 2006. What is the role of genome duplication in the evolution of complexity and diversity? Molecular Biology and Evolution 23:887-892.

Dehal P, Boore JL. 2005. Two rounds of whole genome duplication in the ancestral vertebrate. PLoS Biology 3:e314.

Fan Z, Yuan T, Liu P, Wang LY, Jin JF, Zhang F, Zhang ZS. 2021. A chromosome-level genome of the spider *Trichonephila antipodiana* reveals the genetic basis of its polyphagy and evidence of an ancient whole-genome duplication event. Gigascience 10:1-15.

Farhat S, Modica MV, Puillandre N. 2023. Whole genome duplication and gene evolution in the hyperdiverse venomous gastropods. Molecular Biology and Evolution 40:msad171.

Furlong RF, Holland PW. 2002. Were vertebrates octoploid? Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences 357:531-544.

Hao Y, Mabry ME, Edger PP, Freeling M, Zheng C, Jin L, VanBuren R, Colle M, An H, Abrahams RS, et al. 2021. The contributions from the progenitor genomes of the mesopolyploid Brassiceae are evolutionarily distinct but functionally compatible. Genome Research 31:799-810.

Harper A, Baudouin Gonzalez L, Schonauer A, Janssen R, Seiter M, Holzem M, Arif S, McGregor AP, Sumner-Rooney L. 2021. Widespread retention of ohnologs in key developmental gene families following whole-genome duplication in arachnopulmonates. G3 11:jkab299.

509
510 Haug-Baltzell A, Stephens SA, Davey S, Scheidegger CE, Lyons E. 2017. SynMap2 and
511 SynMap3D: web-based whole-genome synteny browsers. Bioinformatics 33:2197-2198.
512
513 Hoang DT, Chernomor O, von Haeseler A, Minh BQ, Vinh LS. 2018. UFBoot2: Improving the
514 ultrafast bootstrap approximation. Molecular Biology and Evolution 35:518-522.
515
516 Hokamp K, McLysaght A, Wolfe KH. 2003. The 2R hypothesis and the human genome sequence.
517 Journal of Structural and Functional Genomics 3:95-110.
518
519 Howard RJ, Puttick MN, Edgecombe GD, Lozano-Fernandez J. 2020. Arachnid monophyly:
520 Morphological, palaeontological and molecular support for a single terrestrialization within
521 Chelicerata. Arthropod Struct Dev 59:100997.
522
523 Initiative OTPT. 2019. One thousand plant transcriptomes and the phylogenomics of green plants.
524 Nature 574:679-685.
525
526 Junier T, Zdobnov EM. 2010. The Newick utilities: high-throughput phylogenetic tree processing
527 in the UNIX shell. Bioinformatics 26:1669-1670.
528
529 Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7:
530 Improvements in performance and usability. Molecular Biology and Evolution 30:772-780.
531
532 Kenny NJ, Chan KW, Nong W, Qu Z, Maeso I, Yip HY, Chan TF, Kwan HS, Holland PWH, Chu
533 KH, et al. 2016. Ancestral whole-genome duplication in the marine chelicerate horseshoe crabs.
534 Heredity 119:190-199.
535
536 Leite DJ, Baudouin-Gonzalez L, Iwasaki-Yokozawa S, Lozano-Fernandez J, Turetzek N,
537 Akiyama-Oda Y, Prpic NM, Pisani D, Oda H, Sharma PP, et al. 2018. Homeobox gene duplication
538 and divergence in arachnids. Molecular Biology and Evolution 35:2240-2253.
539
540 Li L, Stoeckert CJ, Jr., Roos DS. 2003. OrthoMCL: Identification of ortholog groups for eukaryotic
541 genomes. Genome Research 13:2178-2189.
542
543 Li Z, McKibben MTW, Finch GS, Blischak PD, Sutherland BL, Barker MS. 2021. Patterns and
544 processes of diploidization in land plants. Annual Review of Plant Biology 72:387-410.
545
546 Lozano-Fernandez J, Tanner AR, Giacomelli M, Carton R, Vinther J, Edgecombe GD, Pisani D.
547 2019. Increasing species sampling in chelicerate genomic-scale datasets provides support for
548 monophyly of Acari and Arachnida. Nat Commun 10:2295.

549
550 Lynch M, Conery JS. 2000. The evolutionary fate and consequences of duplicate genes. Science
551 290:1151-1155.

552
553 Ma LJ, Ibrahim AS, Skory C, Grabherr MG, Burger G, Butler M, Elias M, Idnurm A, Lang BF,
554 Sone T, et al. 2009. Genomic analysis of the basal lineage fungus *Rhizopus oryzae* reveals a whole-
555 genome duplication. PLoS Genetics 5:e1000549.

556
557 Masterson J. 1994. Stomatal size in fossil plants: Evidence for polyploidy in majority of
558 angiosperms. Science 264:421-424.

559
560 McKibben MTW, Finch G, Barker MS. 2024. Species Tree Topology Impacts the Inference of
561 Ancient Whole-Genome Duplications Across the Angiosperm Phylogeny.
562 bioRxiv:2024.2001.2004.574202.

563
564 McLysaght A, Hokamp K, Wolfe KH. 2002. Extensive genomic duplication during early chordate
565 evolution. Nature Genetics 31:200-204.

566
567 Mulhair PO, Crowley L, Boyes DH, Harper A, Lewis OT, Consortium DToL, Holland PWH. 2023.
568 Diversity, duplication, and genomic organization of Homeobox genes in Lepidoptera. Genome
569 Research 33:32-44.

570
571 Mulhair PO, Holland PWH. 2024. Evolution of the insect Hox gene cluster: Comparative analysis
572 across 243 species. Seminars in Cell & Developmental Biology 152-153:4-15.

573
574 Nguyen LT, Schmidt HA, von Haeseler A, Minh BQ. 2015. IQ-TREE: A fast and effective
575 stochastic algorithm for estimating maximum-likelihood phylogenies. Molecular Biology and
576 Evolution 32:268-274.

577
578 Noah KE, Hao J, Li L, Sun X, Foley B, Yang Q, Xia X. 2020. Major Revisions in Arthropod
579 Phylogeny Through Improved Supermatrix, With Support for Two Possible Waves of Land
580 Invasion by Chelicerates. Evol Bioinform Online 16:1176934320903735.

581
582 Nong W, Qu Z, Li Y, Barton-Owen T, Wong AYP, Yip HY, Lee HT, Narayana S, Baril T, Swale T,
583 et al. 2021. Horseshoe crab genomes reveal the evolution of genes and microRNAs after three
584 rounds of whole genome duplication. Communications Biology 4:83.

585
586 Nossa CW, Havlak P, Yue JX, Lv J, Vincent KY, Brockmann HJ, Putnam NH. 2014. Joint assembly
587 and genetic mapping of the Atlantic horseshoe crab genome reveals ancient whole genome
588 duplication. Gigascience 3:9.

Ohno S. 1970. Evolution by Gene Duplication: Springer-Verlag.

Ontano AZ, Gainett G, Aharon S, Ballesteros JA, Benavides LR, Corbett KF, Gavish-Regev E, Harvey MS, Monsma S, Santibanez-Lopez CE, et al. 2021. Taxonomic sampling and rare genomic changes overcome long-branch attraction in the phylogenetic placement of pseudoscorpions. Molecular Biology and Evolution 38:2446-2467.

Pfeil BE, Schlueter JA, Shoemaker RC, Doyle JJ. 2005. Placing paleopolyploidy in relation to taxon divergence: A phylogenetic analysis in legumes using 39 gene families. Systematic Biology 54:441-454.

Rabiee M, Sayyari E, Mirarab S. 2019. Multi-allele species reconstruction using ASTRAL. Molecular Phylogenetics and Evolution 130:286-296.

Redmond AK, Casey D, Gundappa MK, Macqueen DJ, McLysaght A. 2023. Independent rediploidization masks shared whole genome duplication in the sturgeon-paddlefish ancestor. Nature Communications 14:2879.

Schwager EE, Sharma PP, Clarke T, Leite DJ, Wierschin T, Pechmann M, Akiyama-Oda Y, Esposito L, Bechsgaard J, Bilde T, et al. 2017. The house spider genome reveals an ancient whole-genome duplication during arachnid evolution. BMC Biology 15:62.

Sela I, Ashkenazy H, Katoh K, Pupko T. 2015. GUIDANCE2: accurate detection of unreliable alignment regions accounting for the uncertainty of multiple parameters. Nucleic Acids Research 43:W7-W14.

Sharma PP, Kaluziak ST, Perez-Porro AR, Gonzalez VL, Hormiga G, Wheeler WC, Giribet G. 2014. Phylogenomic interrogation of arachnida reveals systemic conflicts in phylogenetic signal. Molecular Biology and Evolution 31:2963-2984.

Shingate P, Ravi V, Prasad A, Tay BH, Garg KM, Chattopadhyay B, Yap LM, Rheindt FE, Venkatesh B. 2020. Chromosome-level assembly of the horseshoe crab genome provides insights into its genome evolution. Nature Communications 11:2322.

Shingate P, Ravi V, Prasad A, Tay BH, Venkatesh B. 2020. Chromosome-level genome assembly of the coastal horseshoe crab (*Tachypleus gigas*). Molecular Ecology Resources 20:1748-1760.

Shultz JW. 1990. Evolutionary morphology and phylogeny of Arachnida. Cladistics 6:1-38.

Smith ML, Hahn MW. 2021. New approaches for inferring phylogenies in the presence of paralogs. Trends in Genetics 37:156-169.

Sutherland BL, Tiley GP, Li Z, McKibben MT, Barker MS. 2024. SLEDGe: Inference of ancient whole genome duplications using machine learning. bioRxiv:2024.2001.2017.574559.

Tang H, Bowers JE, Wang X, Ming R, Alam M, Paterson AH. 2008. Synteny and collinearity in plant genomes. Science 320:486-488.

Thomas GWC, Ather SH, Hahn MW. 2017. Gene-tree reconciliation with MUL-trees to resolve polyploidy events. Systematic Biology 66:1007-1018.

Thomas GWC, Dohmen E, Hughes DST, Murali SC, Poelchau M, Glastad K, Anstead CA, Ayoub NA, Batterham P, Bellair M, et al. 2020. Gene content evolution in the arthropods. Genome Biology 21:15.

Tiley GP, Barker MS, Burleigh JG. 2018. Assessing the performance of *Ks* plots for detecting ancient whole genome duplications. Genome Biology and Evolution 10:2882-2898.

Van de Peer Y, Ashman TL, Soltis PS, Soltis DE. 2021. Polyploidy: An evolutionary and ecological force in stressful times. Plant Cell 33:11-26.

Wang Y, Tang H, Debarry JD, Tan X, Li J, Wang X, Lee TH, Jin H, Marler B, Guo H, et al. 2012. MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. Nucleic Acids Res 40:e49.

Werth CR, Windham MD. 1991. A model for divergent, allopatric speciation of polyploid pteridophytes resulting from silencing of duplicate-gene expression. The American Naturalist 137:515-526.

Weygoldt P, Paulus HF. 1979. Untersuchungen zur Morphologie, Taxonomie und Phylogenie der Chelicerata1 II. Cladogramme und die Entfaltung der Chelicerata. Journal of Zoological Systematics and Evolutionary Research 17:177-200.

Wolfe KH. 2001. Yesterday's polyploids and the mystery of diploidization. Nature Reviews Genetics 2:333-341.

Wolfe KH, Shields DC. 1997. Molecular evidence for an ancient duplication of the entire yeast genome. Nature 387:708-713.

669 Yan Z, Cao Z, Liu Y, Ogilvie HA, Nakhleh L. 2022. Maximum parsimony inference of
670 phylogenetic networks in the presence of polyploid complexes. Systematic Biology 71:706-720.

671
672 Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. Molecular Biology and
673 Evolution 24:1586-1591.

674
675 Yates AD, Allen J, Amode RM, Azov AG, Barba M, Becerra A, Bhai J, Campbell LI, Carbajo
676 Martinez M, Chakiachvili M, et al. 2022. Ensembl Genomes 2022: an expanding genome resource
677 for non-vertebrates. Nucleic Acids Res 50:D996-D1003.

678

679

680