# Title

Alphabetical by last name author list:

Michael Barker

Matthew Hahn

Michael McKibben

Gregg Thomas

Li Zheng

# Abstract

# Introduction

Whole genome duplications occur when an individual retains both sets of chromosomes from each parent. This can be highly detrimental to the survival of the individual, but occasionally the influx of novel genetic material can provide adaptive advantages that allow the whole genome duplication to propagate, resulting in a polyploid species with more than 2N chromosomes in its genome. Polyploid species can arise in one of two ways, with both parental genomes being from the same species, called autopolyploidy, or with a hybridization event where the parental genomes come from different species, called allopolyploidy. But regardless of the origin of the polyploid, whole genome duplication can be a key evolutionary event in the history of a species, and there is some evidence pointing to an association between environmental stress and the success of polyploid species. While whole genome duplications are thought to be common in plants, there is also evidence pointing to important genome duplications in the history of fungi and vertebrates.

A common process in the evolution of polyploid species is diploidization, which is the loss of many of the excess genes and chromosomes that resulted from the whole genome duplication. The end result of diploidization is a return of the ploidy level of the species to diploid, with 2N chromosomes. Nevertheless, even in paleopolyploid species that have had ancient WGDs and undergone diploidization, signatures of the WGD can remain in their genomes. For example, an excess of paralogs in the genome may have occurred at a certain time. This means that when gene tree topologies of the gene families that include these paralogs are inferred along with homologous genes from other species, the paralogs will often be inferred on the same branch when mapped to the species tree. This also may lead to spikes in synonymous divergence between paralogs around the timing of the WGD event. There may also be syntenic evidence for the WGD in paleopolyploids, where regions of the extant genome that can trace their history to different parental sub-genomes align to each other.

Recently, whole genome duplications 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 and syntenic blocks suggest whole genome duplication has occurred during their evolution. Examination of the *hox* gene cluster is also used to suggest that there have been anywhere between 1 to 3 WGDs during the course of horseshoe crab evolution. Similar approaches also form the basis for the claim that a WGD has occurred in the lineage ancestral to extant spiders and scorpions. In both cases, the number of genes or genomes has been limited. Additionally, recent evidence supports an alternate placement of horseshoe crabs in the chelicerate phylogeny. Traditionally, horseshoe crabs have been thought to be sister to all arachnids (spiders, scorpions, mites, and ticks). However, Ballesteros et al. (2022) present strong evidence for horseshoe crabs being nested within arachnids, sister to spiders and scorpions. This newly proposed species tree could substantially impact how WGDs are inferred within this group.

Here we use whole genome sequences from 19 chelicerates species along with a myriad of methods to detect ancient WGDs, including the clustering of paralogs in gene tree topologies, synonymous divergence of paralogs, and intraspecific synteny. We test multiple species tree hypotheses and 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 new placement in the chelicerate phylogeny.

# Methods

To investigate the possible existences of whole genome duplication events in chelicerates on a genome-wide scale we took a multi-faceted approach to analyze gene family evolution, divergence, and synteny of 19 available genomes in this group. We downloaded these genomes from various sources (NCBI, Ensembl, i5k, data supplements) with annotations at the beginning of this project. 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). For phylogenetic analyses, we also include two insects (*Drosophila melanogaster* and *Bombyx mori*) as outgroups for tree rooting. See Supplemental Table SX for full details of the samples and summaries of their assemblies and annotations. We observed that one of the horseshoe crab annotations, *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 whole genome duplication in this species, we also note that the median gene length for this species is only 1,377bp which, while not the shortest in our sample, is considerably smaller than the rest of the horseshoe crabs, which all have a median gene length of over 8,500bp. Because this could be indicative of annotation error in this species and because we are interested in more ancient whole genome duplications, we decided to exclude this sample from our analyses. This results in a dataset of 19 chelicerate species and 2 outgroup insects for analysis that spans almost 600 million years of genome evolution.

## Gene family analysis

We identified the coding sequence of the longest transcript from each gene in each of our 21 species and used FastOrtho, which is a reimplementation of orthomcl, to cluster genes into gene families. Using an Inflation value of 3, we inferred 49,561 gene families (probably a Supplementary Figure). We then extracted the sequences in each gene family, correcting for inconsistencies resulting from the data originating from various sources and aligned each gene family with Guidance2 using MAFFT as the underlying aligner, and removing any alignment columns with a score below 0.93. We then 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 SX for full alignment filtering stats. We translated the remaining 11,016 alignments from nucleotides to amino acids and inferred gene trees with IQ-Tree and a species tree using all gene families with ASTRAL-Multi. We rooted our gene and species trees using the outgroup insects with Newick Utilities (nw\_reroot). However, our outgroups were not present in every gene family, in which case the gene tree could not be rooted and was excluded from subsequent analyses. After rooting, we retained gene trees from 6,368 gene families. Then, to reduce possible gene tree inference error, we used bootstrap rearrangement implemented in Notung 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 then used these 6,368 gene trees and a species tree as input to GRAMPA 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 reticulations to be present in the species tree by representing them as multi-labeled trees (MUL-trees), in which one or more tip label appears twice. By comparing LCA mapping scores between the input species tree and a set of MUL-trees defined by target lineages, it can determine if a whole genome duplication has occurred based on an excess of duplications being inferred on a single lineage. For our runs, we set as target lineages for WGD identification those on which WGDs have previously been proposed, specifically the lineages leading to spiders and scorpions and the lineage leading to horseshoe crabs. We also used multiple species trees as input to GRAMPA and tested the same scenarios. First, a recently inferred phylogeny from Ballesteros et al (2021) in which they find that horseshoe crabs group within arachnids, specifically sister to spiders and scorpions, and a ‘traditional’ species tree topology, in which horseshoe crabs are sister to all arachnid species. For the ‘traditional’ tree, because of the unresolved placement of Acariformes and Parasitiformes, we simply use the topology recovered by Ballesteros et al. (2022) and manually move horseshoe crabs to be sister to arachnids.

## Synteny analysis

We used syntenic estimates to test for paleopolyploid ancestry in 16 of the 21 Chelicerate species that had annotated reference genomes. Specifically, we used MCScanX, which uses BLASTP and a novel chain score, with default settings to detect and visualize intraspecific syntenic blocks [(Wang et al. 2012)](https://paperpile.com/c/ufcF67/KKG9).

## Synonymous divergence

To supplement our syntenic inferences, we used DupPipe to calculate the Ks for paralogs in each genome [(Barker et al. 2010)](https://paperpile.com/c/ufcF67/aizL). We then visualized the distributions of Ks values with matplotlib available in python3 [(Hunter 2007)](https://paperpile.com/c/ufcF67/QdPr) and visually assessed for signatures of paleopolyploid ancestry in the form of peaks in the distributions.

Previous work in the Chelicerata used Hox gene duplications as evidence of shared paleopolyploid ancestry [(Schwager et al. 2017)](https://paperpile.com/c/ufcF67/SeQw). Stuff about how the hox genes were found. We then used the duplicate gene classifier available through MCScanX to identify the mode of duplication that formed each Hox gene paralog.

# Results

We sought to assess the occurrence of whole genome duplications in Chelicerates, the Arthropod sub-phylum consisting of horseshoe crabs, mites, ticks, spiders, and scorpions using whole genome data across the group and in light of new molecular evidence that Xiphosura (horseshoe crabs) are nested within Arachnids rather than sister to them. We first used the genomes of 17 chelicerates and 2 insect outgroups to reconstruct the Chelicerate phylogeny, with an emphasis on Arachnids and horseshoe crabs. Using 6,368 genes we confirm the placement of Xiphosura as nested within Arachnids (Fig. 1A), in agreement with Ballesteros et al (2022) (Fig 1B). 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. suggest that Parasitiformes is more closely related to them. The placement of these groups has been contentious in previous studies as well …

We used these species trees as the basis to test various hypotheses of whole genome duplication (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, and based on the duplication of a singly gene family, *Hox*, a single WGD has been proposed in the ancestor of spiders and scorpions. We find that, when using gene tree topologies to count duplications and losses in thousands of genes, there is no evidence for a WGD in the history of spiders and scorpions using our inferred species tree, the Ballesteros et al (2021) species tree, or the traditional species tree in which horseshoe crabs are sister to Arachnids (Fig. 1). In each case, we tested whether a tree with a reticulation at any of the proposed H1 nodes better explains the duplication history of the genes in these genomes than a bifurcating species tree, which would be indicative of WGD, and in each case we find the bifurcating species tree results in the lowest duplication and loss score, indicating no WGD has occurred. This evidence is definitive for any WGD in the history of spiders and scorpions, however we do see evidence for large scale duplications on the branch leading to horseshoe crabs in each species tree (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, some supp fig of the trees?). That is, while our method of counting duplications and losses did not explicitly show a WGD in the history of horseshoe crabs, there are multiple pieces of evidence that point to one or more occurring. In fact, if multiple WGDs occurred, this may diminish the signal for any single proposed reticulation. Since our tests are limited to a single reticulation, this may in turn hinder our ability to explicitly identify any single WGD as the most parsimonious scenario.

We also find that, when comparing duplication and loss scores between species trees, our species tree and the Ballesteros et al. (2021) species tree both explain the history of gene evolution 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 as sister to spiders and scorpions. And while our species tree always better explains the data, this is unsurprising since we inferred our tree from these data.

We next looked at other genome-wide signatures of WGDs on a subset of species. Specifically, we looked for intraspecific syntenic blocks, which should be widespread in genomes that have undergone WGD, and distributions of synonymous divergence (Ks) of paralogs within each genome. If a WGD has occurred in the history of a genome, a secondary peak of Ks should be present in these distributions. With both of these analyses we again find no evidence for WGD in any spider or scorpion genomes and suggestive evidence for at least one occurring in the history of horseshoe crabs. Only two species, *C. rotundicauda* and *T. gigas*, both horseshoe crabs, showed substantial amounts of intraspecific synteny. Both of these, along with the other horseshoe crab, *L. polyphemus*, also have distinct peaks in their Ks distributions, indicating the possibility of WGD occurring in their ancestor. This is the same branch identified with an excess number of gene duplications and losses in our gene tree topology analysis above (Fig. 1)

# Discussion

# Data availability

# Acknowledgements

# References