The draft genome of horseshoe crab Tachypleus tridentatus reveals its evolutionary scenario and well-developed innate immunity

CURRENT STATUS: ACCEPTED

BMC Genomics BMC Series

Yan Zhou
Fudan University

Yuan Liang
Fudan University

Qing Yan
Chinese National Human Genome Center at Shanghai

Liang Zhang
Chinese National Human Genome Center at Shanghai

Dianbao Chen
Minjiang University

Lingwei Ruan
Ministry of Natural Resources of the People's Republic of China

Yuan Kong
Minjiang University

Hong Shi
Ministry of Natural Resources of the People's Republic of China

Mingliang Chen
Ministry of Natural Resources of the People's Republic of China

Jianming Chen

chenjianming@tio.org.cn*Corresponding Author*
ORCiD: https://orcid.org/0000-0003-0400-3092

DOI:
10.21203/rs.2.9427/v6

SUBJECT AREAS

Molecular Biology & Genetics

KEYWORDS

Tachypleus tridentatus, Genome, Evolution, Innate immunity, Coagulation

Abstract

Background: Horseshoe crabs are ancient marine arthropods with a long evolutionary history extending back approximately 450 million years, which may benefit from their innate immune systems. However, the genetic mechanisms underlying their abilities of distinguishing and defending against invading microbes are still unclear. Results: Here, we describe the 2.06 Gbp genome assembly of Tachypleus tridentatus with 24,222 predicted protein-coding genes. Comparative genomics shows that T. tridentatus and the Atlantic horseshoe crab Limulus polyphemus have the most orthologues shared among two species, including genes involved in the immune-related JAK-STAT signalling pathway. Divergence time dating results show that the last common ancestor of Asian horseshoe crabs (including T. tridentatus and C. rotundicauda) and L. polyphemus appeared approximately 130 Mya (121-141), and the split of the two Asian horseshoe crabs was dated to approximately 63 Mya (57-69). Hox gene analysis suggests two clusters in both horseshoe crab assemblies. Surprisingly, selective analysis of immune-related gene families revealed the high expansion of conserved pattern recognition receptors. Genes involved in the IMD and JAK-STAT signal transduction pathways also exhibited a certain degree of expansion in both genomes. Intact coagulation cascade-related genes were present in the T. tridentatus genome with a higher number of coagulation factor genes. Moreover, most reported antibacterial peptides have been identified in T. tridentatus with their potentially effective antimicrobial sites. Conclusions: The draft genome of T. tridentatus would would provide important evidence for further clarifying the taxonomy and evolutionary relationship of Chelicerata. The expansion of conserved immune signalling pathway genes, coagulation factors and intact antimicrobial peptides in T. tridentatus constitutes its robust and effective innate immunity for self-defence in marine environments with an enormous number of invading pathogens and may affect the quality of the adaptive properties with regard to complicated marine environments.

Background

Horseshoe crabs are marine arthropods, representing an ancient family with an evolutionary history record extending back approximately 450 million years (1). Based on their static morphology and

their position in the arthropod family tree, they have been therefore labelled “living fossils” for a long time (2). There are now few types of existing horseshoe crabs with narrow distribution.

Tachypleus tridentatus (Leach, 1819), an extant horseshoe crab species, is mainly distributed from coastal Southeast China to western Japan and in a few islands in Southeast Asia (3). Similar to other invertebrates, *T. tridentatus* relies entirely on its innate immune system, including haemolymph coagulation, phenoloxidase activation, cell agglutination, release of antibacterial substances, active oxygen formation and phagocytosis (4-8), which operates on pattern-recognition receptors (PRRs) upon the detection of pathogen-associated molecular patterns (PAMPs) present on surface of microbes, such as lipopolysaccharides, lipoproteins and mannans (9). Upon recognition, PRRs trigger diverse signal transduction pathways, including the Toll pathway, IMD pathway, JAK-STAT and JNK pathways, that can produce immune-related effectors (10). Previous studies have investigated important signalling pathways and gene families from other arthropods, such as insects, crustaceans and myriapods, revealing extensive conservation and functional diversity among innate immune components across arthropods (11, 12). Currently, the immune molecular mechanisms of how horseshoe crabs achieve distinguishing "self" and "non-self" antigenic epitopes, also known as pathogen-associated molecular patterns (PAMPs), has not yet been established.

The Atlantic horseshoe crab, *Limulus polyphemus* (Linnaeus 1758), is the most extensively investigated species of horseshoe crabs, occupying a large latitudinal range of coastal and estuarine habitats along the west Atlantic coast from Maine to Florida in eastern North America and along the eastern Gulf and around the Yucatán peninsula of Mexico (3, 13, 14). A previous research about the genome of *L. polyphemus* with a high assembly quality has published, focusing on the full repertoire of Limulus opsins, which could provide insight into the visual system of horseshoe crabs. (15). In order to obtain the genome characteristics not only of *T. tridentatus* but also of the xiphosuran lineage and try to reduce errors of only using a single draft-quality genome, the comparative genomic study of immune systems within *T. tridentatus* and *L. polyphemus* were included.

Here, we present an analysis of the *T. tridentatus* genome sequence together with comparative genomic and divergence time analyses on other available Chelicerata genomes to date, including the

previously released *L. polyphemus* assembly (15). Particular attention was paid to gene families related to assessing the genomic and phenotypic changes of horseshoe crabs, as well as exploring immune signalling pathways, antimicrobial peptides and coagulation factors that may contribute to their robust and effective innate immunity for self-defence in marine environments with enormous number of invading pathogens and may have important implications for the continuation of this species.

Results

General genome features

The genomic DNA isolated from *T. tridentatus* was sequenced to 124× coverage and assembled into a 2.06-Gb genome. The k-mer analysis yielded an estimated genome size of 2.22 Gb with a depth peak of 78×. The final draft assembly consists of 143,932 scaffolds with an N50 scaffold size of 165 kb, among which the longest scaffold size is 5.28 Mb and the shortest is 1 kb. The GC content of the genome is 32.03% (Table 1). A total of 24,222 protein-coding genes were conservatively predicted in the *T. tridentatus* genome in this study. The average exon and intron lengths predicted for the assembly are 333 bp and 3,792 bp, respectively. A total of 88.25% of the predicted genes were assigned and annotated by comparing to the NCBI non-redundant database, KEGG database (16) and InterPro database (17).

Repeat annotation

The screening of repeat contents from the RepeatMasker (18) analysis based on similarity alignments identified 20.29 Mb in *T. tridentatus*, representing 0.99% of the genome size. Most of the identified repeat sequences were simple repeats (0.77%). To estimate of repeat sequences which are more difficult to detect in the draft assembly, RepeatModeler (19) was used to predict potential existing but unidentified repeats. Based on this analysis, repeat elements totalled 34.83% in *T. tridentatus*, including a 13.26% proportion of transposable elements. Meanwhile, long interspersed elements (LINEs) composed the largest portion at 6.21%. LTR elements (1.72%) and DNA elements (5.33%) were also detected in the *T. tridentatus* genome. To determine the reliability of the repeat contents

screening by RepeatMasker and RepeatModeler, we also performed repeat analysis of the *L. polyphemus* genome for reference. Similar results were obtained with the identification of repeat sequences representing 1.11% and 34.24% in *L. polyphemus*, respectively. Given that RepeatMasker use similarity of known repeat sequences in the Repbase database to identify repeats in the input sequence, this suggests that the repeat sequences from horseshoe crabs have a great difference compared with existing homologous repeats.

Assembly assessment

The completeness of the *T. tridentatus* genome assembly was assessed using the transcriptome data of the embryonic sample at Stage 21 (the hatch-out stage) of *T. tridentatus* (20). It was found that 99.04% of the transcriptome contigs were aligned to the assembly scaffolds, with an e-value cut-off of 10^{-30}. To further confirm the completeness of the predicted genes, the commonly used genome assembly validation pipeline BUSCO (21) gene mapping method with 1,066 BUSCO Arthropoda gene sets were utilized. The predicted genes of *T. tridentatus* reveals 98.7% conserved proteins of homologous species with 1,052 BUSCOs (76.6% complete single-copy BUSCOs, 10.8% complete duplicated BUSCOs and 11.3% fragmented BUSCOs). Only 1.3% of the benchmarked universal single-copy orthologous groups of arthropod genes were missing in the assembly. This demonstrated that most of the evolutionarily conserved core genes were found in *T. tridentatus* genome, suggesting a remarkable completeness of genome assembly and predicted gene repertoire of *T. tridentatus*.

Phylogeny analysis and divergence time dating

Two *L. polyphemus* assemblies have been previously documented (15, 22), one of which was selected to perform comparative genomics according to a relatively higher assembly level. The OrthMCL (23) calculation resulted in a total of 12,116 orthologous groups in the genomes of *T. tridentatus* and *L. polyphemus*. Of these, 10,968 orthologues contained genes found in both horseshoe crab genomes, with 15,905 *T. tridentatus* and 20,390 *L. polyphemus* genes included; moreover, approximately 6,880 of the shared genes were single-copy. Functional enrichment analysis showed that these shared

genes were involved in several important pathways (p-value < 0.05), such as metabolic pathways (pyruvate, glycerolipid, amino sugar, nucleotide sugar and so on), ribosome biogenesis and DNA replication. The analysis also identified 1,418 protein-coding genes that were only present in *T. tridentatus*. In total, 1,956 genes were only specific to *L. polyphemus*. To place *T. tridentatus* with the most current understanding of the evolution of Chelicerata species, phylogenetic and comparative genomic analyses of *T. tridentatus* and 11 other Chelicerata as well as one Myriapoda outgroup were conducted. The phylogenetic tree was rooted using the centipede *S. maritima* as the outgroup (Figure 1a). Strong bootstrap support was obtained for spider, mite and tick clades, forming a monophyletic group. *T. tridentatus* and *L. polyphemus* were grouped together, forming the Xiphosura clade. The comparative genomic analysis of the 14 species revealed 14,479 orthologous groups containing genes in at least two different species, among which 1,993 shared groups were commonly distributed in all sampled species, with 111 single-copy orthologues (Figure 1b). The single-copy genes enriched for KEGG pathways such as ribosome, oxidative phosphorylation, proteasome, metabolic pathways, and carbon metabolism. Additionally, *T. tridentatus* and *L. polyphemus* had the most orthologues shared among these two species (2,720 (22.2%) and 2,648 (21.5%)). Pathway enrichment of these genes showed significant enrichment (p-value < 0.01) for neuroactive ligand-receptor interaction, FoxO signalling pathway and AGE-RAGE signalling pathway in diabetic complications. The latter two KEGG pathways include the important JAK-STAT signalling pathway genes related to innate immunity in arthropods. With respect to species-specific genes, 1,124 genes were unique to *T. tridentatus*. *C. sculpturatus* had the most (7,328) expanded species-unique genes, followed by 6,247 *N. clavipes*-specific gene families. In contrast, only 161 genes were unique to *T. mercedesae*. The numbers of species-specific genes in *T. tridentatus* and *L. polyphemus* were in between, with 1,124 and 857, respectively. Nevertheless, considering the fragmentation of the draft genome, there may exist more coding genes in the analysed genomes. The species-specific genes described here only refer to the results based on the draft genomes.

The divergence time estimate results for the 7 Chelicerata species showed that the last common ancestor of Asian horseshoe crabs (including *T. tridentatus*) and *L. polyphemus* was dated to 130 Mya

(121-141) and that the split of the Asian horseshoe crabs *T. tridentatus* and *C. rotundicauda* was dated to 63 Mya (57-69) (Figure 2), while the internal split of *T. tridentatus* from southern coastal China to the Korean Peninsula was dated to 12 Mya (11–14). Both the species tree and time tree suggested that horseshoe crabs are closely related to scorpions and that the split of scorpions from horseshoe crabs was dated to 440 Mya (412-468).

Two Hox gene clusters

Hox genes, which are a highly conserved subclass of homeobox super-class genes that have been extensively investigated, are usually distributed in clusters (24-27). Analysis of the Hox gene family showed that the *T. tridentatus* assembly contained 46 Hox genes, while 43 Hox genes were identified in *L. polyphemus* (Table S1). This is the most complete set of Hox genes we obtained based on homeobox domains from these two horseshoe crab assemblies. We found that most Hox genes had at least two representatives in both genomes, which was consistent with a previous whole-genome duplication study in horseshoe crabs (28).

We further examined the positions of the identified Hox genes in the two genomes and found two clusters of adjacently distributed Hox1 and Hox4 in the *T. tridentatus* assembly. In *L. polyphemus*, there was one Hox cluster of adjacent Hox1 and Hox4 genes and one additional Hox1, Hox2 and Hox3 cluster. Other clusters, such as adjacent Hox2 and Hox3 clusters and longer clusters of Hox4, Hox7, Ubx, AbdA and AbdB genes found in the two assemblies, could probably be connected to the two clusters mentioned above. Based on the Hox gene positions in the assemblies, our analysis is consistent with a previous study and suggests that there are possibly two Hox gene clusters present in horseshoe crabs if Hox genes are linearly arranged in clusters along the anterior-posterior axis similar to the ancestral arthropod *Drosophila* (29).

Expansion of crucial gene families of the innate immune signalling pathways in *T. tridentatus* and *L. polyphemus*

Immune-related genes can be broadly classified into pattern recognition receptors (PRRs), signaling

transduction pathways and effectors. We manually searched the *T. tridentatus* and *L. polyphemus* genomes and *T. tridentatus* transcriptome for homologues of essential immune-related genes. PRRs in *T. tridentatus* and *L. polyphemus* show large amounts of expansion, and key genes in the signal transduction pathways also exhibit a certain degree of expansion (Figure 3). We examined six PRR families in *T. tridentatus* and *L. polyphemus*, which included the peptidoglycan recognition proteins (PGRPs), thioester-containing proteins (TEPs), fibrinogen-related proteins (FREPs), down syndrome cell adhesion molecules (Dscams), galectins and C-type lectins (CTLs). The results revealed 42 FREPs and 117 Dscams in *T. tridentatus* that were extensively present in both horseshoe crab genomes with functional domains.

Recognition of PAMPs by PRRs triggers signal transduction pathways through transcriptional activation. All known gene family components that play important roles in innate immune signal transduction in arthropods (such as the Toll, IMD, JAK-STAT, and JNK pathways) (30-32) are present in the genomes of *T. tridentatus* and *L. polyphemus*. From our analyses, we show that IMD and JAK-STAT pathway genes in *T. tridentatus* and *L. polyphemus* exhibit a certain degree of expansion. The orthologue analysis for shared genes in horseshoe crabs with their close evolutionary related species found that horseshoe crabs have the most unique gene orthologues shared in only two species, including the abovementioned expanded gene families.

Regarding the IMD signalling pathway, imd and IKK exit as a single gene, and we discovered multiple copies of genes encoding death-related ced-3/Nedd2-like proteins (Dredds), MAPKKK transforming growth factor - b (TGFb) - activated kinase 1 (Tak1) and Relish proteins within *T. tridentatus* and *L. polyphemus*. For Dredds, the phylogeny tree shows one branch including 7 corresponding genes identified in the two horseshoe crabs and 1 gene in *C. sculpturatus*. Another branch encompasses 2 genes in *P. tepidariorum* (Figure 4a). The Dredds are required for Tak1 activation. For Tak1 homologue analysis, one branch consisting of two gene copies in *T. tridentatus* and *L. polyphemus* showed gene expansion (Figure 4b). Moreover, main components of the JAK-STAT signalling pathway, including the receptor Domeless and the Janus Kinase and STAT transcription factor, were identified in both *T. tridentatus* and *L. polyphemus*, indicating that the JAK-STAT pathway has remained intact in

horseshoe crabs. Two STAT homologue candidates were identified in the *T. tridentatus* genome with the typical functional domains, including a DNA binding domain and an SH2 domain which are conserved compared to those reported in insects and shrimps (33). Plausible homologs of major components of the JNK signalling were also identified in both *T. tridentatus* and *L. polyphemus*. Phylogenetic analysis of JNKs showed that there were three branches consisting of a pair of corresponding genes identified in the *T. tridentatus* and *L. polyphemus* genomes and one branch formed by a pair of genes in *C. sculpturatus* and *S. mimosarum* (Figure 4c).

Antimicrobial peptide diversity in *T. tridentatus*

A hallmark of the *T. tridentatus* host defence system is the production of antimicrobial peptides, which act as innate immune effectors (34). We searched the *T. tridentatus* genome for antimicrobial peptide genes and identified most of the antibacterial peptides that have been reported, including one anti-LPS, two tachyplesin and two big defensin peptides (Figure 3).

The anti-LPS gene found in the *T. tridentatus* genome contains an antimicrobial peptide (AMP) region between G23 to R83 with two conserved cysteine residues as well as a hydrophobic NH^2-terminal and cationic residues clustered in its disulphide loop, which are supposed to act as an affinity site in combination with LPS (35, 36). The tachyplesin family includes constitutively expressed cationic peptides comprised of 17-18 amino acids that strongly inhibit the growth of both Gram-negative and -positive bacteria, including pathogenic microorganisms from marine bivalves such as *Bonamia ostreae*, *Perkinsus marinus* and Vibrio P1, and can also have strong inhibitory effects on the growth of fungi (37, 38). In this study, we identified two tachyplesin precursors in *T. tridentatus*, each of which consists of 77 amino acids encompassing a putative signal peptide sequence, a mature tachyplesin peptide sequence, a C-terminal arginine followed by the amidation signal residues Gly-Lys-Arg and a 22-aa peptide in the C-terminal portion (38). In addition to this, two big defensin protein precursors are also present in the *T. tridentatus* genome, one of which is 118 amino acids in length and contains a hydrophobic N-terminal half and a cationic C-terminal half, which may be closely related to its biological activity for broad antimicrobial properties (39).

Intact coagulation cascades in *T. tridentatus*

Serine protease-dependent rapid coagulation in horseshoe crabs has been shown to play a key role upon the activation of immune pathways in response to pathogen detection (40). We found that *T. tridentatus* and *L. polyphemus* have all the coagulation-related genes while other related species lack a part of the coagulation pathway (Table 2), indicating a wider diversity of coagulation factors and a relatively intact coagulation cascade present in horseshoe crabs. Factor G, a heterodimer that is specifically activated by the fungal cell wall component 1,3-β-D-glucan, is a special serine protease precursor that provides another starting point for the clotting reaction (41, 42). We identified 4 factor G sequences in our *T. tridentatus* genome and transcriptome assembly, including genes encoding the alpha and beta subunits, respectively. However, we failed to identify any clotting factor G homologues in other Chelicerata species.

Discussion

A draft genome of *T. tridentatus* can provide the Chelicerata clade another high-quality publicly available sequence, and would provide an important source for eliminating the uncertainty associated with the evolution of Chelicerata. To date, two papers describing the *T. tridentatus* genome have been published, revealing 2.16 Gb and 1.94 Gb *T. tridentatus* genomes, providing valuable genomic and transcriptomic resources for future studies to exploit horseshoe crabs (43, 44). Using a parallel experiment, the assembly size in this study was between the two previous *T. tridentatus* assemblies. Besides, the number of protein-coding genes predicted in *T. tridentatus* genome was lower that that from the other two published *T. tridentatus* genomes (34,966 and 25,252) but higher than that from *L. polyphemus* (23,287) (18-20). Considering that previous phylogenetic studies only used transcriptomic data with multiple representations of one gene or obtained low bootstrap support for Arachnida, our phylogenetic tree using 111 single-copy orthologous groups of 13 Chelicerata species and 1 outgroup was unsuccessful to support Euchelicerata (Xiphosura plus Arachnida). Even so, the relatively wider species sampling range and more comprehensive information of this study would be helpful to explore the Chelicerata taxa. We further investigated the divergence time using

mitochondrial coding sequences from 7 Chelicerata species, and our analyses suggest that the diversification of the *Limulidae* and *T. tridentatus* lineages was congruent approximately 121-141 Mya, and the lineages of the two Asian horseshoe crabs *T. tridentatus* and *C. rotundicauda* was also congruent approximately 57-69 Mya. According to the continental drift theory, before the Triassic Period, virtually all continents were joined to form the supercontinent Pangea, with the breakup of Pangea commencing in the Triassic Period (45). Approximately 170–120 million years ago (MYA), Pangea broke up into the following two supercontinents: Laurasia and Gondwana (46). The subsequent lineage divergence within reptiles (47), amphibians (48, 49), mammals (50) and even plants (51) matches the separation and fragmentation of Laurasia and Gondwana. Laurasia fragmented during the mid-Mesozoic Era (52), but until late-Cretaceous Period, the Eurasian and North American plates were still joined together (53). The ancestor of horseshoe crabs (or their progenitor species) likely originated in the Mesozoic waters of Europe (54, 55). After the final breakup of the Eurasian and North American plates, the European land mass formed as the shallow seas disappeared and the ancestors of the horseshoe crab migrated. One group migrated to the west along the east coast of North America from Maine to south Florida and from the Gulf of Mexico to the Yucatan Peninsula and evolved into the Atlantic species *L. polyphemus*. The second group migrated to the east through the Tethys, is found along Asia from Japan to India, and evolved into *T. tridentatus*, *T. gigas*, and *C. rotundicauda*. There is evidence showing that the India–Asia collision was underway in northern Pakistan ca. 56–55 Mya (56). The diversity of the Asian horseshoe crabs *T. tridentatus* and *C. rotundicauda* may be related to the India–Asia collision.

Hox genes have been well studied to play essential roles in determining the anterior-posterior axis as well as organogenesis during embryonic development (57-61). Research focused on Hox genes suggests the duplication of clusters from one common ancestor (62). Examination of Hox cluster genes with detailed scaffold positions in the *T. tridentatus* and *L. polyphemus* genomes in this study revealed two Hox gene clusters, suggesting that one round of whole-genome duplication occurred in *T. tridentatus* and *L. polyphemus*. Considering that draft genome data was utilized, the identification and scaffold positions in the Hox genes analysis in both horseshoe crab genomes probably provide a

minimum estimate. More intact Hox clusters may be recovered with more complete assemblies; thus, we can assume that at least one round of whole-genome duplication occurred in both *T. tridentatus* and *L. polyphemus*.

Horseshoe crabs are an independent ancient group with distributions in Southeast Asia and North America. There must be inherent reasons for their survival against complicated marine environmental changes. Some clues could be found from the comparison of the previously determined *L. polyphemus* genomes and our presented *T. tridentatus* genome. Among them, the immune system undoubtedly provides an important guarantee. We searched the two horseshoe crab genomes to investigate the molecular basis of immune signalling pathways, which affect the specificity of foreign pathogen recognition and performance for releasing immune response substances. Genes that are not found in the *T. tridentatus* genome were supplemented by transcriptome data to restore as accurate of numbers as possible. Although the draft genome is incomplete, the gene count errors should be small enough for the predicted genes according to the BUSCO evaluation results of 98.7% conserved homologous proteins. From the results we have obtained, the two horseshoe crabs showed high sequence homology for most immune signaling-related gene families, which have been studied previously in other arthropods. Homologs of the FREP and Dscam families play an important role, and pattern recognition receptors with their corresponding functional domains were extensively present in the two horseshoe crab genomes. The Dscam family in arthropods has evolved to recognize a variety of pathogens, and this is supported by the abundant receptor diversity due to alternative splicing of hypervariable regions (63-66). The abundantly expressed PRRs in the *T. tridentatus* transcriptome suggest an effective ability to recognize a broad range of pathogens, which may be inducted to cope with a great diversity of invading microorganisms in the marine environment. Pathogen recognition would further lead to the activation of signal transduction and the amplification of immune responses, thus producing immune factors that are resistant to microbial activity. Several gene family members in the IMD and JAK-STAT signalling transduction pathway exhibit a certain degree of expansion in *T. tridentatus* and *L. polyphemus*. Research in Drosophila has shown that the IMD pathway preferentially recognizes the peptidoglycan molecules present on the surface of Gram-negative bacteria through

peptidoglycan recognition receptors (PGRPs), leading to the generation of specific AMPs. The JAK-STAT pathway in arthropods, which is analogous to a cytokine-signalling pathway in mammals, has been implicated as having defence abilities against viral, bacterial or protozoan pathogens (67-69). The expanded PRRs and signalling pathways might predict that the innate immunity of *T. tridentatus* has strong signal reception and lineage-specific signal transduction abilities. The interactions among the components in *T. tridentatus* should be paid more attention.

The generation of AMPs is an important aspect of immune responses in horseshoe crabs. The transcription of specific AMPs, which act as important innate immune effectors, are activated through the Toll, IMD, JAK-STAT or JNK pathway based on the recognition of bacteria, fungi, viruses or parasites. We identified most of the antimicrobial peptides isolated in previous studies in *T. tridentatus* assembly, including anti-LPS, tachyplesin, and big defensin peptides. However, in the *L. polyphemus* genome, we failed to find any of the tachyplesin family genes, probably because these antimicrobial peptides are usually shorter and have a high degree of species specificity. There is increasing evidence that AMPs in horseshoe crabs not only possess broad-spectrum antimicrobial capabilities but also have a strong resistibility to enveloped viruses, parasites and tumour cells (70-73). Moreover, antimicrobial peptide counterparts in horseshoe crab haemolymph have also been identified in other evolutionarily conserved animal population representatives (74-76). Thus, more work needs to be undertaken to study the specific function of the AMPs in *T. tridentatus*.

Serine protease-dependent haemolymph coagulation is a major component of the innate immune system in horseshoe crabs. Activation of the horseshoe crab coagulation cascade consists of four coagulation factors including factor C, factor B, factor G and proclotting enzyme (77, 78). Factor C and factor G are two serine protease zymogens that act as biosensors and can be activated by LPS or (1→3)-β-D-glucan, which are major components of the cell walls of Gram-negative bacteria and fungi, respectively. We found relatively high number of genes for all four coagulation factors in the *T. tridentatus* assembly, indicating a complete coagulation cascade in horseshoe crabs. Other Chelicerata species can lack factor G and coagulogen. This might be because of a limited annotation of the draft genome assemblies or there might be other special biosensors involved in their antifungal

recognition process. Although the more direct internal cause for the long-term evolutionary conservation and success of horseshoe crabs with abilities to maintain morphological stability in a fluctuating marine environment may be the combination of a wide feeding spectrum, substantial saline tolerance and insensitivity to temperature by horseshoe crabs (79, 80), there is a presumption that the haemolymph of horseshoe crabs may improve their adaptive strategies and increase their population survival rate (81, 82).

Conclusions

Our draft assembly of *T. tridentatus* was sequenced into a 2.06 Gb of genome consisting of 143,932 scaffolds with an N50 scaffold size of 165 kb. In total, 24,222 protein-coding genes were conservatively predicted as present in the *T. tridentatus* genome, revealing 98.7% completeness with 1,052 BUSCOs.

Further analysis of the *T. tridentatus* genome included phylogeny analysis and divergence time dating using newly published Chelicerata species, as well as selective analysis of Hox genes, innate immunity-related genes, coagulation factors and antimicrobial peptides. We found that the two horseshoe crabs *T. tridentatus* and *L. polyphemus* had the most orthologues shared among two species, and were enriched for the immune-related JAK-STAT signalling pathway. Furthermore, two clusters of Hox genes were predicted in both assemblies, suggesting that at least one round of whole-genome duplication occurred in both *T. tridentatus* and *L. polyphemus*. The innate immunity gene investigation showed that conserved pattern recognition receptor gene families and several signal transduction pathway genes involved in IMD and JAK-STAT exhibit a certain degree of expansion in both horseshoe crab genomes. Moreover, all of the antibacterial peptides reported in previous studies, as well as their effective antimicrobial sites, have also been identified in the *T. tridentatus* genome. Beyond that, according to our analysis, intact coagulation cascade-related genes were identified in the *T. tridentatus* genome, with high numbers of total gene counts. Apart from the abovementioned results, there may remain other aspects of horseshoe crabs that may be greatly conducive to their adaptive advantages in marine environments through their long evolutionary history that need to be studied further and established in future studies.

Methods

T. tridentatus specimens and DNA extraction

T. tridentatus is endangered according to the Red List (https://www.iucnredlist.org/species/21309/149768986). The use of the *T. tridentatus* species collected in this study was to separate 0.1g leg skeleton muscle tissue for DNA extraction. This study was reviewed and approved by the Animal Care and Use Committee of Minjiang University. All the required collection and permits had been obtained before any work was started. Only one *T. tridentatus* was purchased in the aquatic market in Xiamen, Fujian Province, China and was released after 0.1g leg muscle collection. Total genomic DNA was extracted using an E.Z.N.A.® Insect DNA Extraction Kit (Omega Bio-Tek, Inc., Norcross, GA, USA) at the State Key Laboratory of Genetic Engineering in Fudan University, Shanghai, China. The genomic research about *T. tridentatus* covers the evolution history of this endangered species, and may reveal genomic characteristics that limit their survival, which could provide a basis for the species protection.

Genome sequencing and assembly

The Illumina TruSeq Nano DNA Library Prep Kit (Illumina, San Diego, CA) was used to construct the 400-bp and 800-bp paired-end sequencing libraries. Long mate-pair libraries featuring 3-kb inserts were prepared using a CHGC kit (CHGC, China). The Nextera MP Sample Prep Kit (Illumina, San Diego, CA) was used to build long mate-pair libraries with 8- and 12-kb insert sizes. Whole-genome sequencing was performed on the Illumina HiSeq 2500 and HiSeq X Ten platforms generating 250-bp and 150-bp paired-end reads, respectively. Five libraries of nominal insert sizes, including 400 bp, 800 bp, 3- kb, 8- kb and 12- kb, were sequenced at expected genome coverages of 60×, 30×, 7×, 7× and 20×, respectively. The generated clean reads from the 400 bp, 800 bp and 3- kb insert size libraries were used for the estimation of the genome size via JELLYFISH v1.116 (83) with a k-mer length of 17. The genome size of *T. tridentatus* was estimated using online scripts (https://github.com/josephryan/estimate_genome_size.pl) from a k-mer distribution generated by jellyfish. Contigs were assembled using Velvet (84) by cutting the short reads into k-mers and

establishing the de Bruijn table to correct and complete the contigs, which were further scaffolded and gap-filled using SSPACE-STANDARD (85) with long mate-pair reads (8- and 12-kb).

Automated annotation

tridentatus genome assembly automatic gene prediction was established using AUGUSTUS (Version 3.3) (86). The predicted genes were annotated by comparing to the NCBI non-redundant database (NR), Kyoto Encyclopedia of Genes and Genomes (KEGG) database (16) and InterPro database (17) with an E-value threshold of 10^{-5}. The three annotation results were combined as the annotation of the predicted genes. The benchmarking sets of universal single-copy orthologues (BUSCOs) (21) were used to assess the completeness of the predicted genes with 1066 Arthropoda datasets. The repeat contents of *T. tridentatus* and *L. polyphemus* were first analysed using RepeatMasker (version open-4.0.5) (18) with merostomata as the query species and running with rmBLASTn (version 2.2.27+) (87) and RepBase (version 20140131) (88). RepeatModeler (version open-1.0.11) (19) was used to build the repeat database, which was further masked by RepeatMasker.

Transcriptome analysis

RNA-seq raw data obtained from the embryonic sample at Stage 21 (the hatch-out stage) of *T. tridentatus* were downloaded from the NCBI SRA database (accession number SRX330201) (20). De novo assembly of the transcriptome was performed with Trinity (89) with default parameters. Protein-coding sequences and the longest transcript ORFs were predicted via Transdecoder (89).

Orthology and phylogeny analysis

The published genome assembly, coding sequences and protein sequences for the Atlantic horseshoe crab *L. polyphemus* submitted by Washington University were downloaded from NCBI with RefSeq ID 2304488, accession GCF_000517525.1. The non-redundant protein sequences in *L. polyphemus* were selected by sorting the protein scaffold positions and filtering out overlapped proteins. Protein sequences of other 11 Chelicerata species, including three Araneae (*Nephila clavipes*, *Parasteatoda*

tepidariorum, and *Stegodyphus mimosarum)*, seven Acari (*Varroa jacobsoni*, *Tropilaelaps mercedesae*, *Sarcoptes scabiei*, *Metaseiulus occidentalis*, *Tetranychus urticae*, *Varroa destructor*, and *Ixodes scapularis*), and one scorpion (*Centruroides sculpturatus*) were downloaded from the GenBank database. Additional outgroup protein sequences from the centipede *Strigamia maritima* were downloaded from the UniProt database. OrthoMCL (23) was used to perform orthologous gene family clustering and provide calculation results for the number of genes in each orthologous group. Selected single copy orthologues were subjected to multiple sequence alignments using ClustalW v2.0.12 (90) and Maximum Likelihood phylogenetic trees were built using the PROTGAMMAWAG model with 1000 bootstraps implemented in RAxML v8.2.12 (91), with centipede as an out-group. The KOBAS 3.0 web server (92-94) was used for the functional gene set enrichment of shared and species-specific genes from *L. polyphemus* and *T. tridentatus* and the single-copy gene set for the 13 Chelicerata species and centipede with the KEGG database (16).

Divergence time estimates

Thirteen intact mitochondrial coding sequences of 7 species were downloaded from the NCBI gene database and used to build the time tree containing three horseshoe crabs (*Tachypleus tridentatus*, *Carcinoscorpius rotundicauda* and *Limulus Polyphemus*), one scorpion (*Mesobuthus gibbosus*) and three Araneae (*Argiope bruennichi*, *Wadicosa fidelis* and *Tetragnatha nitens*). The 13 mitochondrial coding sequences for *T. tridentatus* collected in Korea were used to query for identical genes in our *T. tridentatus* genome and transcriptome assembly using blastn (blast, Basic Local Alignment Search Tool). Alignment of sequences with identities of at least 97% was considered to be the counterparts of mitochondrial coding sequences in *T. tridentatus* identified in this study. Multiple sequence alignments of concatenated mitochondrial coding sequences of 7 species were processed using ClustalW v2.0.12 (90). BEAUti v2.5.1 (95-97) was used to generate the BEAST input XML files, after which the tree was dated using BEAST v.2.5.1. The Hasegawa-Kishino-Yano (HKY) substitution model with empirical frequencies and a strict clock model were used. Fossil information for all Chelicerata species were used to calibrate the tree with a normal distribution of 530 Mya and standard deviation

5 Mya (98). A chain length of 6,000,000 generations was run when sampling every 1,000 generations. Software Tracer was used to analyse the BEAST log file output. TreeAnnotator (99) and FigTree were used for tree production and tree visualization, respectively.

Identification of Hox genes, immune pathway genes and coagulation factors

Protein sequences of Hox genes, essential immune signalling-related genes and coagulation factors of species closely related to horseshoe crabs were downloaded from the NCBI protein database and used as query sequences. Blastp was then used with an e-value of 10^{-15} to search for homologues using the *T. tridentatus* and *L. polyphemus* genomes as databases. Tblastn was used with an e-value of 10^{-15} to search for corresponding transcripts in the *T. tridentatus* transcriptome. Transcripts found in the transcriptome were further compared to the genome of *T. tridentatus* by blastn with an e-value of 10^{-5} to complement identified genes and to reduce the data omission in the genome. Putative genes were selected based on positive scores and alignment length percentages defined by dividing the alignment length by query length, and then filtrated according to their annotation in the NCBI non-redundant database (NR) and InterPro database (17).

Dredd, Tak1 and JNK gene phylogenetic analysis

MEGA (version 7.0) (100) was used to construct two neighbour-joining trees and a maximum likelihood tree with 1,000 bootstraps using putative protein sequences as the default parameters. Domains from all the selected genes annotated as IPR001309, IPR001932 and IPR000719 in the InterPro database were used for multiple alignment. Dredd from *Stegodyphus mimosarum*, Tak1 from *Xenopus tropicalis* and JNK from *Parasteatoda tepidariorum* were chosen as outgroups for the above phylogenetic trees.

Identification of putative antimicrobial peptides

Because antimicrobial peptides (AMPs) are commonly short in length, display wide variety and have

large differences in their structures and functions within species, tblastn was used to identify potential AMPs using the *T. tridentatus* and *L. polyphemus* genome assemblies as databases. Identified putative AMPs were further used as query sequences to search for corresponding transcripts. The prediction of ORFs was performed using NCBI ORF finder (https://www.ncbi.nlm.nih.gov/orffinder/) and the annotation of those ORFs was performed by blastp with the NR database.

List Of Abbreviations

blast	Basic Local Alignment Search Tool
KEGG	Kyoto Encyclopedia of Genes and Genomes
PAMPs	Pathogen-Associated Molecular Patterns
LINEs	Long Interspersed Elements
LTR	Long Terminal Repeat
BUSCO	Benchmarking Universal Single-Copy Orthologs
FoxO	The Forkhead Box O
RAGE	Receptor for Advanced Glycation Endproducts
JAK	Janus kinase
STAT	Signal Transducer and Activator of Transcription Protein
Hox genes	Homeotic Genes
Ubx	Ultrabithorax
AbdA	Abdominal-A
AbdB	Abdominal-B
PRRs	Pattern Recognition Receptors
FREPs	Fibrinogen-Related Proteins
Dscams	Down Syndrome Cell Adhesion Molecules
IMD	Immune deficiency
JNK	c-Jun N-terminal Kinase
LPS	Lipopolysaccharide
AMP	Antimicrobial Peptide
ORF	Open Reading Frame

Declarations

Ethics approval and consent to participate

The experimental protocol in this study for *Tachypleus tridentatus* genome sequencing was established according to the Animal Care and Use Committee of Minjiang University and was approved by the Animal ethics committee, Institute of Oceanography, Minjiang University.

Consent for publication

Not applicable

Availability of data and material

All data are available from the following databases as described. The raw sequences for *T. tridentatus* have been deposited in the NCBI SRA: BioProject ID PRJNA510236, BioSample ID SAMN10600682, accession SRX6412182 - SRX6412188. The whole genome assembly has been deposited at GenBank under the BioProject ID PRJNA510236, accession QCWK00000000. The version described in this paper is version QCWK01000000. RNA-seq raw data obtained from the embryonic sample at Stage 21 (the

hatch-out stage) of *T. tridentatus* were downloaded from the NCBI SRA database (accession number SRX330201). The published genome assembly, coding sequences and protein sequences for the Atlantic horseshoe crab *L. polyphemus* submitted by Washington University were downloaded from NCBI with RefSeq ID 2304488, accession GCF_000517525.1. Protein sequences of three Araneae (*Nephila clavipes*, *Parasteatoda tepidariorum*, and *Stegodyphus mimosarum*), seven Acari (*Varroa jacobsoni*, *Tropilaelaps mercedesae*, *Sarcoptes scabiei*, *Metaseiulus occidentalis*, *Tetranychus urticae*, *Varroa destructor*, and *Ixodes scapularis*), and one scorpion (*Centruroides sculpturatus*) were downloaded from NCBI Genbank with following accession numbers: GCA_002102615.1, GCF_000365465.2, GCA_000611955.2, GCA_002532875.1, GCA_002081605.1, GCA_000828355.1, GCA_000255335.1, GCA_000239435.1, GCA_002443255.1, GCA_000208615.1 and GCA_000671375.2. 15,047 protein sequences from the centipede *Strigamia maritima* were downloaded from UniProt (https://www.uniprot.org/uniprot/?query=Strigamia+maritima&sort=score). Thirteen intact mitochondrial coding sequences of 7 species were downloaded from the NCBI gene database with following accession numbers: 7768785-7768797 for *Tachypleus tridentatus*, 14049643-14049655 for *Carcinoscorpius rotundicauda*, 803775-803787 for *Limulus Polyphemus*, 3183285-3183286, 3183289-3183291, 3183293, 3183295, 3183297-3183298, 3183300, 3183305, 3183313, 3183315 for *Mesobuthus gibbosus*, 19591985 - 19591997 for *Argiope bruennichi*, 22834025- 22834037 for *Wadicosa fidelis* and 26047144, 26047146-26047148, 26047151, 26047156, 26047158, 26047160, 26047167-26047170, 26047172 for *Tetragnatha nitens.*

Competing interests

The authors declare that they have no competing interests.

Funding

This work was supported in part by the National Key R&D Program of China (2017YFC1404504), the Major State Basic Research Development Program of China (973 Program) (2015CB755906), Project 2018N2001 from Department of Fujian Science and Technology and Program for Innovative Research Team in Science and Technology in Fujian Province University, and Beihai Pilot City Program for the National Innovative Development of the Marine Economy (BHSFS002). The funding bodies played no

role in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

Authors' contributions

YZ, MLC and JMC designed the study. MLC, DBC, LWR, YK, HS collected horseshoe crab samples. MLC, YL, QY, and LZ carried out the experiments. YZ, YL, QY analyzed the data. YZ, MLC, JMC, YL and YK wrote the manuscript. All authors read and approved the final manuscript.

Acknowledgements

Not applicable

References

1. Rudkin DM, Young GA, Nowlan GS. The oldest horseshoe crab: a new xiphosurid from Late Ordovician Konservat-Lagerstätten deposits, Manitoba, Canada. Palaeontology. 2008;51:1-9.
2. Fisher DC. The Xiphosurida: archetypes of bradytely? In: Eldredge N, Stanley SM editors. Living fossils. New York: Springer; 1984. p. 196-213.
3. Sekiguchi K. Biology of horseshoe crabs. Portland: International Specialized Book Service Incorporated; 1988.
4. Iwanaga S, Kawabata S-i, Muta T. New types of clotting factors and defense molecules found in horseshoe crab hemolymph: their structures and functions. The Journal of Biochemistry. 1998;123:1-15.
5. Tan NS, Ho B, Ding JL. High-affinity LPS binding domain (s) in recombinant factor C of a horseshoe crab neutralizes LPS-induced lethality. The FASEB Journal. 2000;14:859-70.
6. Muta T, Iwanaga S. The role of hemolymph coagulation in innate immunity. Current Opinion in Immunology. 1996;8:41-7.
7. Nellaiappan K, Sugumaran M. On the presence of prophenoloxidase in the hemolymph of the horseshoe crab, Limulus. Comparative Biochemistry and Physiology Part B:

Biochemistry and Molecular Biology. 1996;113:163-8.

8. Pieters J. Evasion of host cell defense mechanisms by pathogenic bacteria. Current Opinion in Immunology. 2001;13:37-44.

9. Akira S, Uematsu S, Takeuchi O. Pathogen recognition and innate immunity. Cell. 2006;124:783-801.

10. Lemaitre B, Hoffmann J. The host defense of Drosophila melanogaster. Annual Review of Immunology. 2007;25:697-743.

11. Iwanaga S, Lee B-L. Recent advances in the innate immunity of invertebrate animals. BMB Reports. 2005;38:128-50.

12. Chipman AD, Ferrier DE, Brena C, Qu J, Hughes DS, Schröder R, et al. The first myriapod genome sequence reveals conservative arthropod gene content and genome organisation in the centipede Strigamia maritima. PLoS Biology. 2014;12:e1002005.

13. Anderson LI, Shuster CN. Throughout geologic time: Where have they lived? Cambridge: Harvard University Press; 2003.

14. Faurby S, King TL, Obst M, Hallerman EM, Pertoldi C, Funch P. Population dynamics of American horseshoe crabs - historic climatic events and recent anthropogenic pressures. Molecular Ecology. 2010;19:3088-100.

15. Battelle B-A, Ryan JF, Kempler KE, Saraf SR, Marten CE, Warren WC, et al. Opsin repertoire and expression patterns in horseshoe crabs: evidence from the genome of Limulus polyphemus (Arthropoda: Chelicerata). Genome Biology and Evolution. 2016;8:1571-89.

16. KEGG database. https://www.kegg.jp/. Accessed 23 April 2018.

17. InterPro database. http://www.ebi.ac.uk/interpro/search/sequence/. Accessed 23 April 2018.

18. Tarailo-Graovac M, Chen N. Using RepeatMasker to identify repetitive elements in genomic sequences. Current Protocols in Bioinformatics. 2009;25:4.10. 1-4.. 4.
19. Smit AF, Hubley R. RepeatModeler Open-1.0. 2008.
20. Chen M, Wang C, Wang W, Ji G, Hu B, Du M, et al. De novo assembly and characterization of early embryonic transcriptome of the horseshoe crab Tachypleus tridentatus. PloS One. 2016;11:e0145825.
21. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. 2015;31:3210-2.
22. Simpson SD, Ramsdell JS, Watson III WH, Chabot CC. The draft genome and transcriptome of the Atlantic horseshoe crab, Limulus polyphemus. International Journal of Genomics. 2017;2017:1-14.
23. Li L, Stoeckert CJ, Roos DS. OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Research. 2003;13:2178-89.
24. Holland PW. Evolution of homeobox genes. Wiley Interdisciplinary Reviews: Developmental Biology. 2013;2:31-45.
25. Hui JH, McDougall C, Monteiro AS, Holland PW, Arendt D, Balavoine G, et al. Extensive chordate and annelid macrosynteny reveals ancestral homeobox gene organization. Molecular Biology and Evolution. 2011;29:157-65.
26. Gehring WJ, Affolter M, Burglin T. Homeodomain proteins. Annual Review of Biochemistry. 1994;63:487-526.
27. Ferrier D. When is a Hox gene not a Hox gene? The importance of gene nomenclature. Evolving Pathways: key themes in Evolutionary Developmental Biology, Cambridge University Press: Cambridge, UK. 2008:175-93.
28. Kenny NJ, Chan KW, Nong W, Qu Z, Maeso I, Yip HY, et al. Ancestral whole-genome

duplication in the marine chelicerate horseshoe crabs. Heredity. 2016;116:190.

29. Hughes CL, Kaufman TC. Hox genes and the evolution of the arthropod body plan. Evolution & Development. 2002;4:459-99.

30. Tanji T, Ip YT. Regulators of the Toll and Imd pathways in the Drosophila innate immune response. Trends in Immunology. 2005;26:193-8.

31. Boutros M, Agaisse H, Perrimon N. Sequential activation of signaling pathways during innate immune responses in Drosophila. Developmental Cell. 2002;3:711-22.

32. Li F, Xiang J. Signaling pathways regulating innate immune responses in shrimp. Fish & Shellfish Immunology. 2013;34:973-80.

33. Chen WY, Ho KC, Leu JH, Liu KF, Wang HC, Kou GH, et al. WSSV infection activates STAT in shrimp. Developmental & Comparative Immunology. 2008;32:1142-50.

34. Miyata T, Tokunaga F, Yoneya T, Yoshikawa K, Iwanaga S, Niwa M, et al. Antimicrobial peptides, isolated from horseshoe crab hemocytes, tachyplesin II, and polyphemusins I and II: chemical structures and biological activity. The Journal of Biochemistry. 1989;106:663-8.

35. Aketagawa J, Miyata T, Ohtsubo S, Nakamura T, Morita T, Hayashida H, et al. Primary structure of limulus anticoagulant anti-lipopolysaccharide factor. Journal of Biological Chemistry. 1986;261:7357-65.

36. Muta T, Miyata T, Tokunaga F, Nakamura T, Iwanaga S. Primary structure of anti-lipopolysaccharide factor from American horseshoe crab, Limulus polyphemus. The Journal of Biochemistry. 1987;101:1321-30.

37. Nakamura T, Furunaka H, Miyata T, Tokunaga F, Muta T, Iwanaga S, et al. Tachyplesin, a class of antimicrobial peptide from the hemocytes of the horseshoe crab (Tachypleus tridentatus). Isolation and chemical structure. Journal of Biological Chemistry. 1988;263:16709-13.

38. Shigenaga T, Muta T, Toh Y, Tokunaga F, Iwanaga S. Antimicrobial tachyplesin peptide precursor. cDNA cloning and cellular localization in the horseshoe crab (Tachypleus tridentatus). Journal of Biological Chemistry. 1990;265:21350-4.

39. Saito T, Kawabata S-i, Shigenaga T, Takayenoki Y, Cho J, Nakajima H, et al. A novel big defensin identified in horseshoe crab hemocytes: isolation, amino acid sequence, and antibacterial activity. The Journal of Biochemistry. 1995;117:1131-7.

40. Muta T, Iwanaga S. Clotting and immune defense in Limulidae. Invertebrate Immunology: Springer; 1996. p. 154-89.

41. Morita T, Tanaka S, Nakamura T, Iwanaga S. A new (1→ 3)-β-D-glucan-mediated coagulation pathway found in limulus amebocytes. FEBS Letters. 1981;129:318-21.

42. Muta T, Seki N, Takaki Y, Hashimoto R, Oda T, Iwanaga A, et al. Purified horseshoe crab factor G reconstruction and characterization of the (1,3)-β-D-glucan-sensitive serine protease cascade. Journal of Biological Chemistry. 1995;270:892-7.

43. Gong L, Fan G, Ren Y, Chen Y, Qiu Q, Liu L, et al. Chromosomal level reference genome of Tachypleus tridentatus provides insights into evolution and adaptation of horseshoe crabs. Molecular Ecology Resources. 2019;19:744-56.

44. Liao YY, Xu PW, Kwan KY, Ma ZY, Fang HY, Xu JY, et al. Draft genomic and transcriptome resources for marine chelicerate Tachypleus tridentatus. Scientific Data. 2019;6:190029.

45. Dietz RS, Holden JC. Reconstruction of Pangaea: breakup and dispersion of continents, Permian to present. Journal of Geophysical Research. 1970;75:4939-56.

46. Pyron RA. Biogeographic analysis reveals ancient continental vicariance and recent oceanic dispersal in amphibians. Systematic Biology. 2014;63:779-97.

47. Gamble T, Bauer AM, Greenbaum E, Jackman TR. Evidence for Gondwanan vicariance in an ancient clade of gecko lizards. Journal of Biogeography. 2008;35:88-104.

48. Roelants K, Bossuyt F. Archaeobatrachian paraphyly and Pangaean diversification of crown-group frogs. Systematic Biology. 2005;54:111-26.

49. San Mauro D, Vences M, Alcobendas M, Zardoya R, Meyer A. Initial diversification of living amphibians predated the breakup of Pangaea. The American Naturalist. 2005;165:590-9.

50. Springer MS, Murphy WJ, Eizirik E, O'Brien SJ. Placental mammal diversification and the Cretaceous–Tertiary boundary. Proceedings of the National Academy of Sciences. 2003;100:1056-61.

51. Mao K, Milne RI, Zhang L, Peng Y, Liu J, Thomas P, et al. Distribution of living Cupressaceae reflects the breakup of Pangea. Proceedings of the National Academy of Sciences. 2012;109:7793-8.

52. Gardner JD, Nydam RL. Mesozoic and Cenozoic lissamphibian and squamate assemblages of Laurasia—introduction to the special issue. Palaeobiodiversity and Palaeoenvironments. 2013;93:391-5.

53. Wicander R, Monroe JS. Historical geology : evolution of Earth and life through time. 4th ed. USA: Brooks/Cole; 2000.

54. Shuster JC. A pictorial review of the natural history and ecology of the horseshoe crab Limulus polyphemus, with reference to other Limulidae. Progress in Clinical and Biological Research. 1982;81:1-52.

55. Briggs DE, Moore RA, Shultz JW, Schweigert G. Mineralization of soft-part anatomy and invading microbes in the horseshoe crab Mesolimulus from the Upper Jurassic Lagerstätte of Nusplingen, Germany. Proceedings of the Royal Society of London B: Biological Sciences. 2005;272:627-32.

56. Ding L, Qasim M, Jadoon IA, Khan MA, Xu Q, Cai F, et al. The India–Asia collision in north Pakistan: Insight from the U–Pb detrital zircon provenance of Cenozoic foreland

basin. Earth and Planetary Science Letters. 2016;455:49-61.

57. Gehring W. Homeotic genes, the homeobox, and the spatial organization of the embryo. Harvey Lectures. 1985;81:153-72.

58. Gehring WJ. A history of the homeobox. Guidebook to the Homeobox Genes. 1994:1-10.

59. De Robertis EM. The homeobox in cell differentiation and evolution. Guidebook to the Homeobox Genes. 1994.

60. Holland P, Hogan B. Expression of homeo box genes during mouse development: a review. Genes & Development. 1988;2:773-82.

61. Denis D. Guidebook to the homeobox genes. Oxford: Oxford University Press; 1994.

62. Garcia-Fernàndez J. The genesis and evolution of homeobox gene clusters. Nature Reviews Genetics. 2005;6:881.

63. Schmucker D, Clemens JC, Shu H, Worby CA, Xiao J, Muda M, et al. Drosophila Dscam is an axon guidance receptor exhibiting extraordinary molecular diversity. Cell. 2000;101:671-84.

64. Graveley BR, Kaur A, Gunning D, Zipursky SL, Rowen L, Clemens JC. The organization and evolution of the dipteran and hymenopteran Down syndrome cell adhesion molecule (Dscam) genes. RNA. 2004;10:1499-506.

65. Brites D, McTaggart S, Morris K, Anderson J, Thomas K, Colson I, et al. The Dscam homologue of the crustacean Daphnia is diversified by alternative splicing like in insects. Molecular Biology and Evolution. 2008;25:1429-39.

66. Yue Y, Meng Y, Ma H, Hou S, Cao G, Hong W, et al. A large family of Dscam genes with tandemly arrayed 5′ cassettes in Chelicerata. Nature Communications. 2016;7:11252.

67. Souza-Neto JA, Sim S, Dimopoulos G. An evolutionary conserved function of the JAK-

STAT pathway in anti-dengue defense. Proceedings of the National Academy of Sciences. 2009;106:17841-6.

68. Gupta L, Molina-Cruz A, Kumar S, Rodrigues J, Dixit R, Zamora RE, et al. The STAT pathway mediates late-phase immunity against Plasmodium in the mosquito Anopheles gambiae. Cell Host & Microbe. 2009;5:498-507.

69. Liu L, Dai J, Zhao YO, Narasimhan S, Yang Y, Zhang L, et al. Ixodes scapularis JAK-STAT pathway regulates tick antimicrobial peptides, thereby controlling the agent of human granulocytic anaplasmosis. The Journal of Infectious Diseases. 2012;206:1233-41.

70. Morvan A, Iwanaga S, Comps M, Bachere E. In vitro activity of the Limulus antimicrobial peptide tachyplesin I on marine nivalve pathogens. Journal of Invertebrate Pathology. 1997;69:177-82.

71. Morimoto M, Mori H, Otake T, Ueba N, Kunita N, Niwa M, et al. Inhibitory effect of tachyplesin I on the proliferation of human immunodeficiency virus in vitro. Chemotherapy. 1991;37:206-11.

72. Murakami T, Niwa M, Tokunaga F, Miyata T, Iwanaga S. Direct virus inactivation of tachyplesin I and its isopeptides from horseshoe crab hemocytes. Chemotherapy. 1991;37:327-34.

73. Chen Y, Xu X, Hong S, Chen J, Liu N, Underhill CB, et al. RGD-Tachyplesin inhibits tumor growth. Cancer Research. 2001;61:2434-8.

74. Zasloff M. Magainins, a class of antimicrobial peptides from Xenopus skin: isolation, characterization of two active forms, and partial cDNA sequence of a precursor. Proceedings of the National Academy of Sciences. 1987;84:5449-53.

75. Moore KS, Wehrli S, Roder H, Rogers M, Forrest JN, McCrimmon D, et al. Squalamine: an aminosterol antibiotic from the shark. Proceedings of the National Academy of

Sciences. 1993;90:1354-8.

76. Merchant ME, Leger N, Jerkins E, Mills K, Pallansch MB, Paulman RL, et al. Broad spectrum antimicrobial activity of leukocyte extracts from the American alligator (Alligator mississippiensis). Veterinary Immunology and Immunopathology. 2006;110:221-8.

77. Kawabata S-i, Osaki T, Iwanaga S. Innate immunity in the horseshoe crab. New York: Humana Press; 2003.

78. Kawabata S. Clotting cascade and defense molecules found in the hemolymph of the horseshoe crab. New Directions in Invertebrate Immunology. 1996:255-83.

79. Ehlinger G, Tankersley R. Ecology of horseshoe crabs in microtidal lagoons. US: Springer; 2009.

80. Walls EA, Berkson J, Smith SA. The horseshoe crab, Limulus polyphemus: 200 million years of existence, 100 years of study. Reviews in Fisheries Science. 2002;10:39-73.

81. Mikkelsen T. The secret in the blue blood. Beijing: Science Press; 1988.

82. Leibovitz L, Lewbart G. Diseases and symbionts: vulnerability despite tough shells. The American Horseshoe Crab Harvard University Press, Cambridge. 2003:245-75.

83. Marçais G, Kingsford C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics. 2011;27:764-70.

84. Zerbino D, Birney E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Research. 2008:gr. 074492.107.

85. Boetzer M, Pirovano W. SSPACE-LongRead: scaffolding bacterial draft genomes using long read sequence information. BMC Bioinformatics. 2014;15:211.

86. Stanke M, Tzvetkova A, Morgenstern B. AUGUSTUS at EGASP: using EST, protein and genomic alignments for improved gene prediction in the human genome. Genome Biology. 2006;7:S11.

87. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. Journal of Molecular Biology. 1990;215:403-10.

88. Jurka J. Repbase update: a database and an electronic journal of repetitive elements. Trends in Genetics. 2000;16:418-20.

89. Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, et al. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nature Protocols. 2013;8:1494.

90. Larkin MA, Blackshields G, Brown N, Chenna R, McGettigan PA, McWilliam H, et al. Clustal W and Clustal X version 2.0. Bioinformatics. 2007;23:2947-8.

91. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 2014;30:1312-3.

92. Wu J, Mao X, Cai T, Luo J, Wei L. KOBAS server: a web-based platform for automated annotation and pathway identification. Nucleic Acids Research. 2006;34:W720-W4.

93. Xie C, Mao X, Huang J, Ding Y, Wu J, Dong S, et al. KOBAS 2.0: a web server for annotation and identification of enriched pathways and diseases. Nucleic Acids Research. 2011;39:W316-W22.

94. Ai C, Kong L. CGPS: A machine learning-based approach integrating multiple gene set analysis tools for better prioritization of biologically relevant pathways. Journal of Genetics and Genomics. 2018;45:489-504.

95. Bouckaert R, Heled J, Kühnert D, Vaughan T, Wu C-H, Xie D, et al. BEAST 2: a software platform for Bayesian evolutionary analysis. PLoS Computational Biology. 2014;10:e1003537.

96. Heled J, Drummond AJ. Calibrated tree priors for relaxed phylogenetics and divergence time estimation. Systematic Biology. 2011;61:138-49.

97. Hasegawa M, Kishino H, Yano T-a. Dating of the human-ape splitting by a molecular

clock of mitochondrial DNA. Journal of Molecular Evolution. 1985;22:160-74.

98. Tanaka G, Hou X, Ma X, Edgecombe GD, Strausfeld NJ. Chelicerate neural ground pattern in a Cambrian great appendage arthropod. Nature. 2013;502:364.

99. Helfrich P, Rieb E, Abrami G, Lücking A, Mehler A, editors. TreeAnnotator: versatile visual annotation of hierarchical text relations. Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018); 2018.

100. Kumar S, Stecher G, Tamura K. MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. Molecular Biology and Evolution. 2016;33:1870-4.

101. McTaggart SJ, Conlon C, Colbourne JK, Blaxter ML, Little TJ. The components of the Daphnia pulex immune system as revealed by complete genome sequencing. BMC Genomics. 2009;10:175.

Tables

Table 1 Summary of the *Tachypleus tridentatus* genome assembly and annotation statistic

Summary of the *Tachypleus tridentatus* genome assembly and annotation statistics.	
Tachypleus tridentatus assembly statistics	
Assembly size (Gb)	2.06
Number of scaffolds	143,932
N50 scaffold length (kb)	165
Largest scaffold (kb)	5,278
Shortest scaffold (kb)	1
GC content	32.03%
Average exon length (bp)	333
Average intron length (bp)	3,792
Tachypleus tridentatus assembly annotation statistics	
Total number of genes	24,222
% BUSCOs[a]	87.4 [10.8], 11.3, 1.3

[a] of 1,066 arthropod BUSCOs Complete [Duplicated], Fragmented, Missing, in the assembly

Table 2 Coagulation Cascade genes in 2 horseshoe crabs, 1 scorpione, and 3 spiders.

	Horseshoe Crabs		Scorpiones
Species	*T. tridentatus*	*L. Polyphemus*	*C. sculpturatus*
Factor C	2	4	5
Factor B	10	12	8
Factor G	4	1*	0
Proclotting Enzyme	9	11	7
Coagulogen	6	6	0
Total	30	33	20

* Identified in previous study

Additional File Legend

Table S1 Comparison of homeobox ANTP class genes between *T. tridentatus* and *L. polyphemus* genome.

Figures

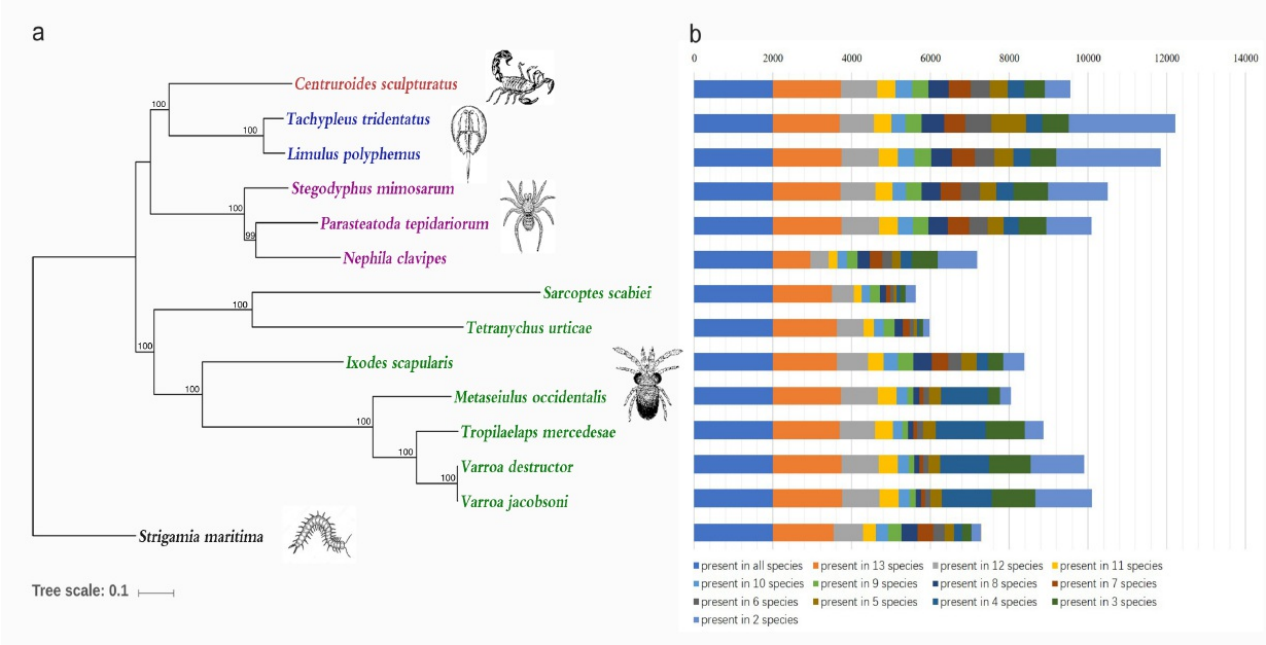


Figure 1

Comparative genomics. (a) Phylogenetic placement among T. tridentatus and other Chelicerata species. The phylogeny with 111 single-copy orthologous genes presented in all 14 species was built using RAxML. The tree was rooted with S. maritima. (b). Orthology comparsion among T. tridentatus and other Chelicerata species. There were 2,720 (22.2%) and 2,648 (21.5%) orthologs of T. tridentatus and L. polyphemus shared in only two species. C. sculpturatus had the most expanded species unique genes (7,328), followed with 6,247 N. clavipes specific genes. The number of species specific genes of T. tridentatus and L. polyphemus showed in the middle position with 1,124 and 857, respectively.

	T. tridentatus (genome)	*T. tridentatus* (transcriptom)	*L. polyphemus*	*S. maritima*	*D. melanogaster*
Pattern recognition receptors					
PGRP	2	2	2	16	13
TEP like	23	2	24	4	6
FREP like	42	32	46	13	13
Dscam like	117	57	118	1	4
Galectin	5	2	5	-	-
CTL	27	25	27	-	-
Signaling and Transduction					
Toll pathway					
Toll like	18	8	17	36	9
spz like	8	2	13	1	6
Myd88	1	1	1	1	1
tube	1	1	1	0	1
pelle	3	1	1	1	1
cactus	1	1	1	1	1
dorsal	1	1	2	1	1
TRAF2	8	8	6	-	-
IMD pathway					
imd	1	1	1	~1	1
Dredd	4	4	3	1	1
Tak1	3	3	3	1	1
Relish	6	6	4	2	1
IKK	1	1	1	-	-
Other					
domeless	5	4	5	1	1
Jak (hop)	1	0	1	1	1
Stat92E	2	2	2	1	1
JNK (bsk)	3	3	3	1	1
Hem	1	1	1	1	1
Effectors					
Anti-LPS factor	1	-	~1	-	-
Tachyplesin	2	-	-	-	-
Big defensin	2	-	-	-	-

Figure 2

Bayesian maximum-clade-credibility tree based on the concatenated mitochondrial coding

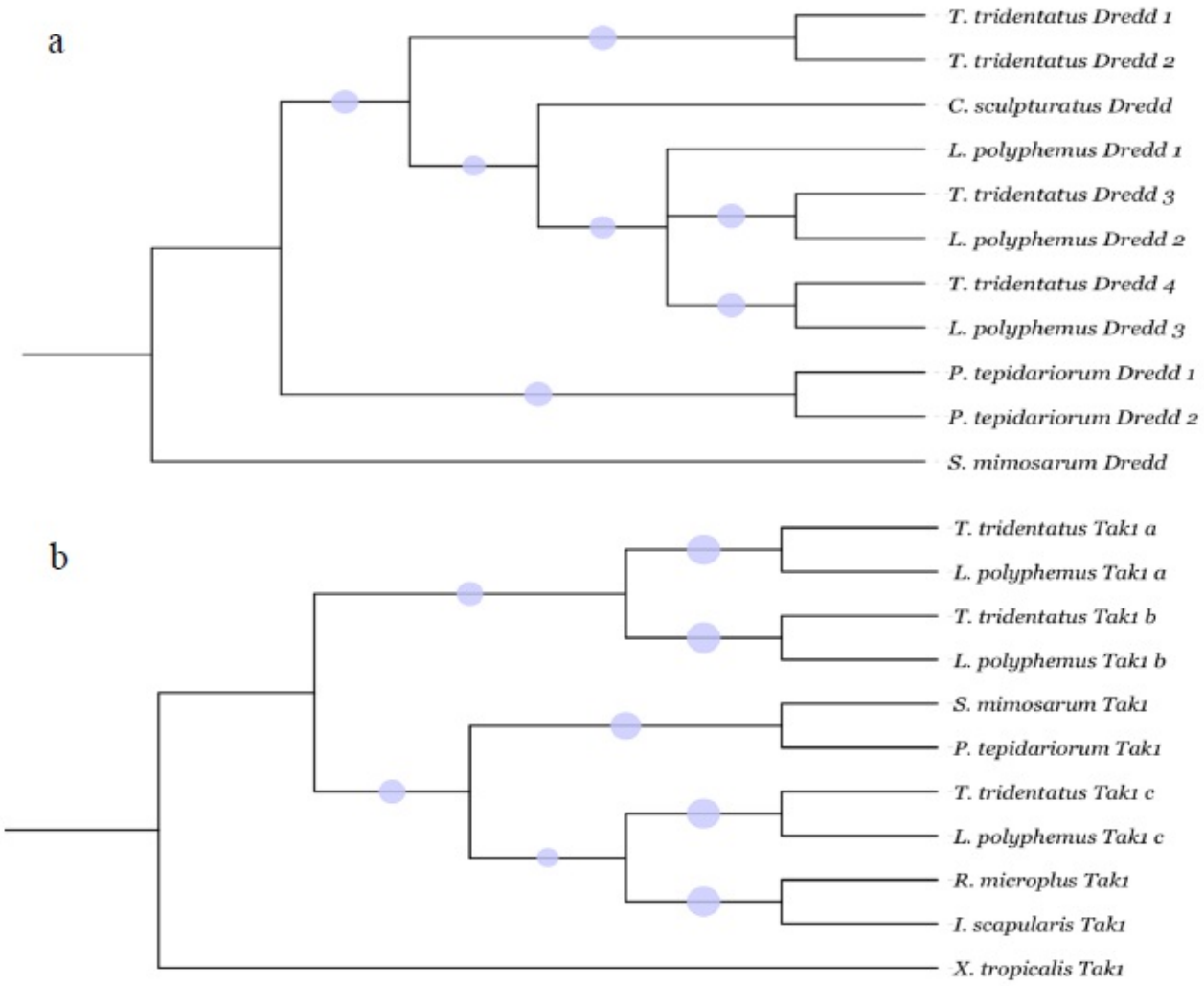
genes dataset in BEAST 2.5.1 with a strict clock, showing the estimated divergence time of Chelicerata species. Node shows the mean estimated divergence times in million years ago (MYA). Purple bars indicate 95% confidence levels. On the time axis, the green bar shows the divergence time for split of the scorpion from horseshoe crabs; the brown bar shows the inner split time of the three spiders; the blue bar shows the origin of the the last common ancestor of Asian horseshoe crabs (including T. tridentatus) and L. polyphemus; the red bar shows the inner split of C. rotundicauda and T. tridentatus.

	T. tridentatus (genome)	*T. tridentatus* (transcriptom)	*L. polyphemus*	*S. maritima*	*D. melanogaster*
Pattern recognition receptors					
PGRP	2	2	2	16	13
TEP like	23	2	24	4	6
FREP like	42	32	46	13	13
Dscam like	117	57	118	1	4
Galectin	5	2	5	-	-
CTL	27	25	27	-	-
Signaling and Transduction					
Toll pathway					
Toll like	18	8	17	36	9
spz like	8	2	13	1	6
Myd88	1	1	1	1	1
tube	1	1	1	0	1
pelle	3	1	1	1	1
cactus	1	1	1	1	1
dorsal	1	1	2	1	1
TRAF2	8	8	6	-	-
IMD pathway					
IMD	1	1	1	~1	1
Dredd	4	4	3	1	1
Tak1	3	3	3	1	1
Relish	6	6	4	2	1
IKK	1	1	1	-	-
Other					
domeless	5	4	5	1	1
JAK (hop)	1	0	1	1	1
Stat92E	2	2	2	1	1
JNK (bsk)	3	3	3	1	1

Hem	1	1	1	1	1
Effectors					
Anti-LPS factor	1	-	~1	-	-
Tachyplesin	2	-	-	-	-
Big defensin	2	-	-	-	-

Figure 3

Presence of immune related gene families in T. tridentatus and L. polyphemus. Counts of immune related genes are shown for T. tridentatus, L. polyphemus, S. maritima (12) and D. melanogaster (104). The gene number counts according to results of blastp based on NR annotation and InterPro from the genome of T. tridentatus and L. polyphemus and the transcriptome of T. tridentatus. Abbreviations: PGRP, peptidoglycan recognition protein; TEP, thioester-containing protein; FREP, fibrinogen-related protein; CTL, C-type lectin.



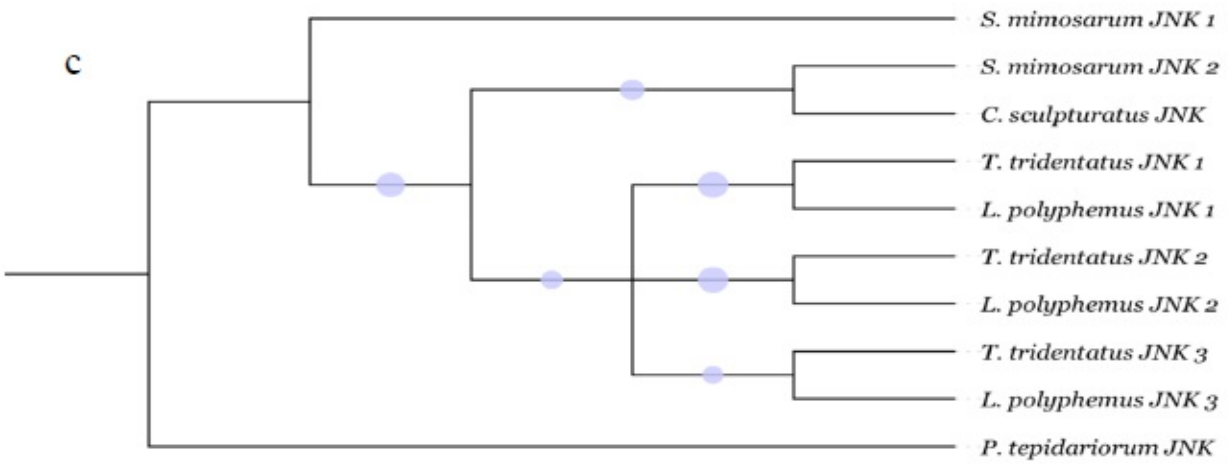


Figure 4

Phylogenetic analysis of immune signaling pathway related genes. (a) Phylogeny of Dredd genes involved in IMD signaling pathways among multiple Chelicerata species. (b) Phylogeny of Tak1 genes involved in IMD signaling pathways among multiple Chelicerata species. (c) Phylogeny of JNK genes involved in JNK signaling pathways among multiple Chelicerata species. The two Neighbor-Joining and one Maximum Likelihood trees were constructed using MEGA with 1000 bootstrap, and rooted with S. mimosarum for Dredds, X. tropicalis for Tak1 and P. tepidariorum for JNKs.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

Supplementary Table 1.doc