

Duplication and divergence of Sox genes in spiders

Christian L. B. Paese¹, Daniel J. Leite¹, Anna Schoenauer¹, Alistair P. McGregor^{1*}, Steven Russell^{2*}

¹ Department of Biological and Medical Sciences, Oxford Brookes University, Gypsy Lane, Oxford, OX3 0BP, UK

² Department of Genetics, University of Cambridge, Downing Street, Cambridge, CB2 3EH, UK.

* Correspondence: s.russell@gen.cam.ac.uk and amcgregor@brookes.ac.uk

Abstract

Background

The Sox family of transcription factors are present and conserved in the genomes of all metazoans examined to date and are known to play important developmental roles in vertebrates and insects. However, outside the commonly studied *Drosophila* model little is known about the extent or conservation of the Sox family in other arthropod species. Here we characterise the Sox family in two chelicerate species, the spiders *Parasteatoda tepidariorum* and *Stegodyphus mimosarum*, which have experienced a whole genome duplication (WGD) in their evolutionary history.

Results

We find that virtually all of the duplicate Sox genes have been retained in these spiders after the WGD. Analysis of the expression of Sox genes in *P. tepidariorum* embryos indicates that it is likely that some of these genes have neofunctionalised after duplication. Our expression analysis also strengthens the view that an orthologue of vertebrate Group B1 genes, *SoxNeuro*, is implicated in the earliest events of CNS specification in both vertebrates and invertebrates. In addition, a gene in the *Dichaete/Sox21b* class is dynamically expressed in the spider segment addition zone, suggestive of an ancient regulatory mechanism controlling arthropod segmentation as recently suggested for flies and beetles. Together with the recent analysis of Sox gene expression in the embryos of other arthropods, our findings are also indicative of conserved functions for some of these genes, including a role for *SoxC* and *SoxD* genes in CNS development, *SoxF* in limb development and a tantalising suggestion that *SoxE* genes may be involved in gonadogenesis across the metazoa.

Conclusions

Our study provides a new chelicerate perspective to understanding the evolution and function of Sox genes and how the retention of duplicates of such important tool-box genes after WGD has contributed to different aspects of spider embryogenesis. Future characterisation of the function of these genes in spiders will help us to better understand the evolution of the regulation of important developmental processes in arthropods and other metazoans including neurogenesis and segmentation.

Keywords: Sox genes, *Parasteatoda tepidariorum*, *Stegodyphus mimosarum* spider, evolution, development

Introduction

The evolution of metazoan life forms was in part driven by the acquisition of novel families of transcription factors and signalling molecules that were subsequently expanded by gene duplications and evolved new functions [1, 2]. One such family, encoded by Sox genes, encompasses a set of conserved

metazoan specific transcriptional regulators that play critical roles in a range of important developmental processes [3], in particular, aspects of stem cell biology and nervous system development [4, 5].

The Sox family is defined by a set of genes containing an HMG class DNA binding domain sharing greater than 50%

sequence identity with that of SRY, the Y-linked sex determining factor in eutherian mammals [6]. In the chordates the family contains approximately 20 genes, which have been subdivided into eight groups (A-H) based mainly on homology within the DNA binding domain but also related group-specific domains outwith the HMG domain [7, 8]. In all metazoans examined to date representatives of the Sox family have been identified and these are largely restricted to Groups B to F [9] with other groups specific to particular lineages. While Sox-like sequences have been reported in the genome of the choanoflagellate *Monosiga brevicollis* [10] these are more closely related to the non-sequence specific HMG1/2 class of DNA binding domain and thus true Sox genes are restricted to metazoans [11, 12].

While vertebrate Sox genes have been intensively studied due to their critical roles in development [3], with the exception of the fruit fly *Drosophila melanogaster*, they are less well characterised in invertebrates. *D. melanogaster* contains eight Sox genes (four group B and one each in groups C to F), which is generally consistent across the insect genomes examined to date [9, 13, 14]. Of particular interest are the Group B genes of insects, which share a common genomic organisation that has been conserved across all insects examined to date, with three genes closely linked in a cluster [13-15]. Critical roles in early segmentation and nervous system development have been shown for *Dichaete* (*D*) [16, 17], and in CNS development for *SoxNeuro* (*SoxN*), where both these group B genes show partial redundancy [18, 19]. The evolutionary conservation of Sox protein function as well as sequence has been shown in rescue or swap experiments, where mouse Sox2 rescues *D* null mutant phenotypes in the *D. melanogaster* embryo [20] and *Drosophila* SoxN can replace Sox2 in mouse ES cells [21]. Furthermore, a comparison of *D* and SoxN genomic binding in the *D. melanogaster* embryo with Sox2 and Sox3 binding in mouse embryonic or neural stem cells indicates that these proteins share a common set of over 2000 core target genes [22-24]. These and other studies suggest that Sox proteins have ancient roles, particularly in the CNS, where their functions have been conserved from flies to mammals.

Of the other two *D. melanogaster* group B genes, *Sox21a* plays a repressive role in maintaining adult intestinal stem cell populations [25, 26] but there is no known function for *Sox21b*. The group C gene *Sox14* is involved in the response to the steroid hormone ecdysone and is necessary for

metamorphosis [27]; *Sox102F* (Group D) has a role in late neuronal differentiation [28]; *Sox100B* (Group E) is involved in male testis development [29] and *Sox15* (Group F) is involved in wing metamorphosis and adult sensory organ development [30, 31].

While functional studies are lacking in other insects, gene expression analysis in *Apis mellifera* and *Bombyx mori* indicates that aspects of Sox function are likely to be conserved across species [13, 14]. More recently, a conserved role for *D* in the early embryonic segmentation of both *Drosophila* and the flour beetle *Tribolium castaneum* suggests that aspects of regulatory function as well as genomic organisation may have been conserved across insects [32]. Outside the insects little is known, however genome sequence analysis and gene expression studies suggest key roles for Sox family members in stem cell and cell fate processes in Ctenophores [12] and Porifera [33], as well as neural progenitor development in Cnidarians [34] and a Dioplopod [35]. Taken together with the extensive work in vertebrate systems, it is clear that Sox genes play critical roles in many aspects of metazoan development, at least some of which appear to be deeply conserved.

Arthropods comprise approximately 80% of living animal species [36] exhibiting a huge range of biological and morphological diversity that is believed to have originated during the Cambrian Period over 500 million years ago [37]. While the analysis of traditional model arthropods such as *D. melanogaster* has taught us much about conserved developmental genes and processes, it is only more recently that genomic and other experimental approaches are beginning to shed light on the way genes and regulatory networks are deployed to generate the diversity of body plans found in other insects [38] and more widely in chelicerates and myriapods [39]. In terms of the Sox family, recent work indicates conserved Group B expression in the early neuroectoderm of the myriapod *Glomeris marginata* [35] and neuroectodermal expression of a Group B gene in the chelicerate *P. tepidariorum* has been reported [40].

Chelicerates in particular offer an interesting system for exploring the evolution and diversification of developmental genes since it has emerged that some arachnid lineages, including spiders and scorpions, have undergone a whole genome duplication (WGD) [41]. Interestingly, duplicated copies of many developmental genes, including Hox genes and other regulatory factors such as microRNAs, have been

retained in *P. tepidariorum* and other arachnids [41, 42]. Thus, chelicerate genomes provide an opportunity to explore issues of gene retention, loss or diversification [43].

Here we report an analysis of the Sox gene family in the spiders, *P. tepidariorum* and *S. mimosarum*, and show that most duplicate Sox genes have been retained in the genomes of these spiders after the WGD. Furthermore, while group B genes show highly conserved expression in the developing CNS, the expression of other spider Sox genes suggests they play important roles and potentially novel functions in other aspects of embryogenesis.

Results and Discussion

Characterisation of Sox genes in spiders

In order to characterise the Sox gene complement of spiders we conducted TBLASTN searches of the genomes of *P. tepidariorum* [41] and *S. mimosarum* [44] using the HMG domain of the mouse Sox2 protein, recovering 16 and 15 sequences respectively. All but three of these contained the highly conserved RPMNAFMVW motif that is characteristic of Sox proteins and these three (*ptSoxC-2*, *ptSoxB-like* and *ptSox21b-2*) only show minor conservative substitutions in this motif. 14 of the *P. tepidariorum* sequences corresponded with annotated gene models. Two sequences were identical (*ptSox21b-1*, aug3.24914.t1 and aug3.g24896.t1), since the latter maps to a genomic scaffold of only ~7 kb we presume this represents an assembly error and thus consider them as a single gene. One genomic scaffold encoding a Sox domain (*ptSoxB-like*, Scaffold3643:28071..28299) is in a region of poor sequence quality and we cannot be sure it represents a *bona fide* gene but have nevertheless included it in subsequent analysis. In the case of *S. mimosarum* we identified 15 genomic regions, 11 of which correspond to annotated genes. Reciprocal BLAST searches of *D. melanogaster* or vertebrate genes recovered Sox proteins as top scoring hits. In addition to these true Sox gene sequences, we also recovered sequences that correspond to the *D. melanogaster capicua* (*cic*) and *bobby sox* (*bbx*) genes but do not consider these Sox-related genes further here.

To classify the spider Sox proteins we generated MUSCLE sequence alignments and PhyML maximum likelihood phylogenies using the HMG domains recovered from the BLAST searches, along with those from the eight *D. melanogaster* Sox genes and representatives of each

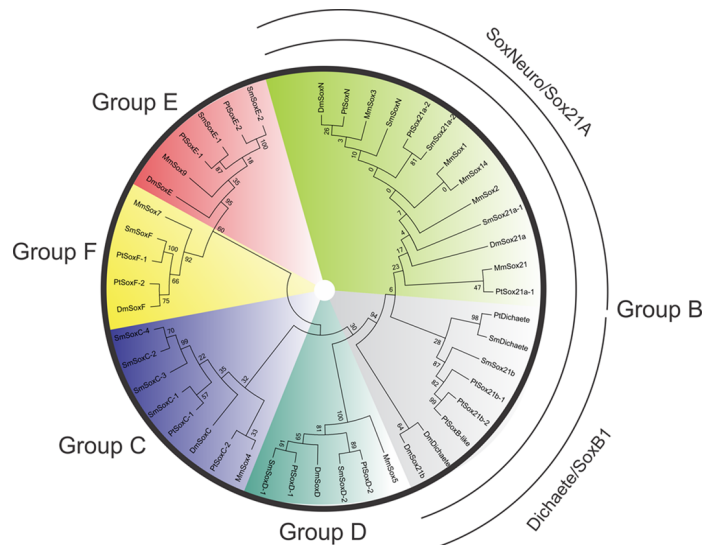


Figure 1. Phylogeny of Sox HMG domains in selected metazoans. Phylogenetic tree showing the relationship between *Mus musculus* (Mm), *D. melanogaster* (Dm), *P. tepidariorum* (Pt) and *S. mimosarum* (Sm) Sox genes based on HMG domain sequences. The grouped genes are divided into different colours as highlighted outside the circle.

subgroup from mouse (Supplementary Table 1). These analyses resulted in a clear classification into groups B-F as found in other invertebrate genomes (Figure 1). Note that Group A only contains the SRY gene specific to eutherian mammals and there are no Group G, H or I Sox genes found outside the vertebrates. Supporting this classification, phylogenetic trees constructed with the full length sequences of the predicted spider Sox proteins and those from *D. melanogaster* yielded virtually identical results (Supplementary Figure 1). Following the recommended nomenclature for Sox genes [7], we have named the spider Sox genes as indicated in Supplementary Table 1. The naming of *D. melanogaster* Sox genes is confusing with some carrying historic names based on phenotype (*Dichaete* and *SoxN*), others named after cytological locations (*Sox100B* and *Sox102F*) and others with inappropriate numerical designations (*D. melanogaster Sox14* is a Group C gene while the vertebrate Sox14s are in Group B and *D. melanogaster Sox15* is in group F, while vertebrate Sox15s are in Group G), thus we propose renaming the *D. melanogaster* group C-F genes according to the standard nomenclature used in the Sox field: these designations are already recognised as synonyms in FlyBase. With respect to the Group B genes, since the sequence and organisation of these appears to be invertebrate specific, we propose a nomenclature based on the current *D. melanogaster* gene names: *SoxN*, *D*, *Sox21a* and *Sox21b* (Supplementary Table 1).

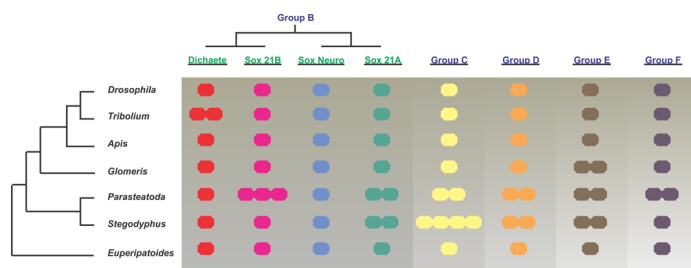


Figure 2. Repertoire of Sox genes in selected arthropods. Diagrammatic representation of the complement of Sox genes in insects (*Drosophila melanogaster*, *Tribolium castaneum* and *Apis mellifera*), the spiders (*Parasteatoda tepidariorum* and *Stegodyphus mimosarum*), the myriapod (*Glomeris marginata*) and an onychophoran (*Euperipatoides kanangrensis*). Each coloured circle represents a gene.

In common with many other gene families in spiders [41], the Sox genes are mostly represented by two or more copies in each group (Figure 2). In other arthropods examined to date, as well as the onychophoran *Euperipatoides kanangrensis* [45], there is usually only a single copy of each gene, although [45] recently reported two Group E genes in the millipede *G. marginata*. In the case of spider Groups D and E, the duplications clearly predate the divergence of the two spider species we analysed since the duplicates group together in the phylogenetic analysis and show extensive homology across the length of the coding sequence (Figure 1). With Group F, there is only one gene identified in *S. mimosarum* but two in *P. tepidariorum*. In the case of group C, there appears to have been additional duplication events in *S. mimosarum*. When we consider the full-length protein sequences (Supplementary Figure 1), *ptSoxC-1* groups with *smSoxC-1* and *ptSoxC-2* with *smSoxC-2*. *smSoxC-2* has undergone a local head-to-head duplication, with *ptSoxC-2* and *smSoxC-3* adjacent in the genome. *smSoxC-4* has no predicted gene model but the region of the genome encodes an uninterrupted HMG domain closely related to those of the *smSoxC-2* and *C-3* duplicates. Whether this is a *bona fide* gene remains to be determined.

In many organisms, some genes in Groups D, E and F contain an intron within the DNA binding domain sequence in a position that is highly conserved and specific for each group [7]: our analysis indicates that this is also the case for the spider genes in these three groups (see arrows in Figure 3). Interestingly, while there is an intron within the spider Group D genes, it has been lost in the *D. melanogaster* orthologue. Secondary intron loss is also observed in Group F, where mouse *Sox7* has no intron but the related *Sox17* and *Sox18*

genes do. The location of these HMG domain introns suggests they were present in the common ancestor of the vertebrates and the arthropods.

The Group B genes of insects and vertebrates are clearly different. Vertebrate Group B genes are subdivided into B1 (*Sox1*, 2 and 3) and B2 (*Sox14* and 21), a classification manifest both at the sequence and the functional levels, with Group B1 proteins acting as transcriptional activators particularly important for nervous system specification, while the Group B2 proteins act as transcriptional repressors [46-48]. In contrast, the organisation and functional classification of Group B genes in insects is subject to some debate. There is a clear orthologue of the Group B1 proteins, represented by *SoxN* in *D. melanogaster* and genes named *SoxB1* or *Sox2* in every invertebrate genome examined. The remaining three *D. melanogaster* Group B genes (*D*, *Sox21* and *Sox21b*) have been characterised as Group B2 based on sequence alignments with vertebrate proteins. In *D. melanogaster* these three genes are arranged in a cluster on Chromosome 3L, an organisation that is conserved across at least 300 MY of evolution, with a similar organisation found in flies, mosquitoes, wasps, bees and beetles [11, 13, 15]. While there is evidence that *Sox21a* has a repressive role consistent with the vertebrate B2 class [25, 26], considerable genomic evidence clearly shows *D* acts as a transcriptional activator, a role inconsistent with that observed for vertebrate *SoxB2* proteins [22, 49].

The phylogenies generated with the HMG domains from a range of species (Figure 1; Supplementary Figure 2) or full length proteins sequences from spiders and *D. melanogaster* (Supplementary Figure 1) support a classification of arthropod Group B genes where there is a single *SoxN* gene, one or more *Sox21a* genes and two or more *Dichaete-Sox21b* genes. In spiders we find strong support for a single *SoxN* gene, duplications of the *Sox21a* class and a single *D*-like gene in both species. In *P. tepidariorum* we find a duplication of the *Sox21b* genes and the possibility of a further tandem duplication of *ptSox21b-2* gene if the *ptSoxB-like* ORF is a genuine gene. *S. mimosarum*, in contrast has a single *Sox21b* class gene. Intriguingly, we find that two *P. tepidariorum* Group B genes (*ptDichaete* and *ptSox21a-1*) are located in the same genomic region, separated by over 200 kb of intervening DNA that is devoid of other predicted genes (Figure 4), an organisation reminiscent of that found in insects. Indeed, the linkage of *ptDichaete* and *ptSox21a-1* supports the

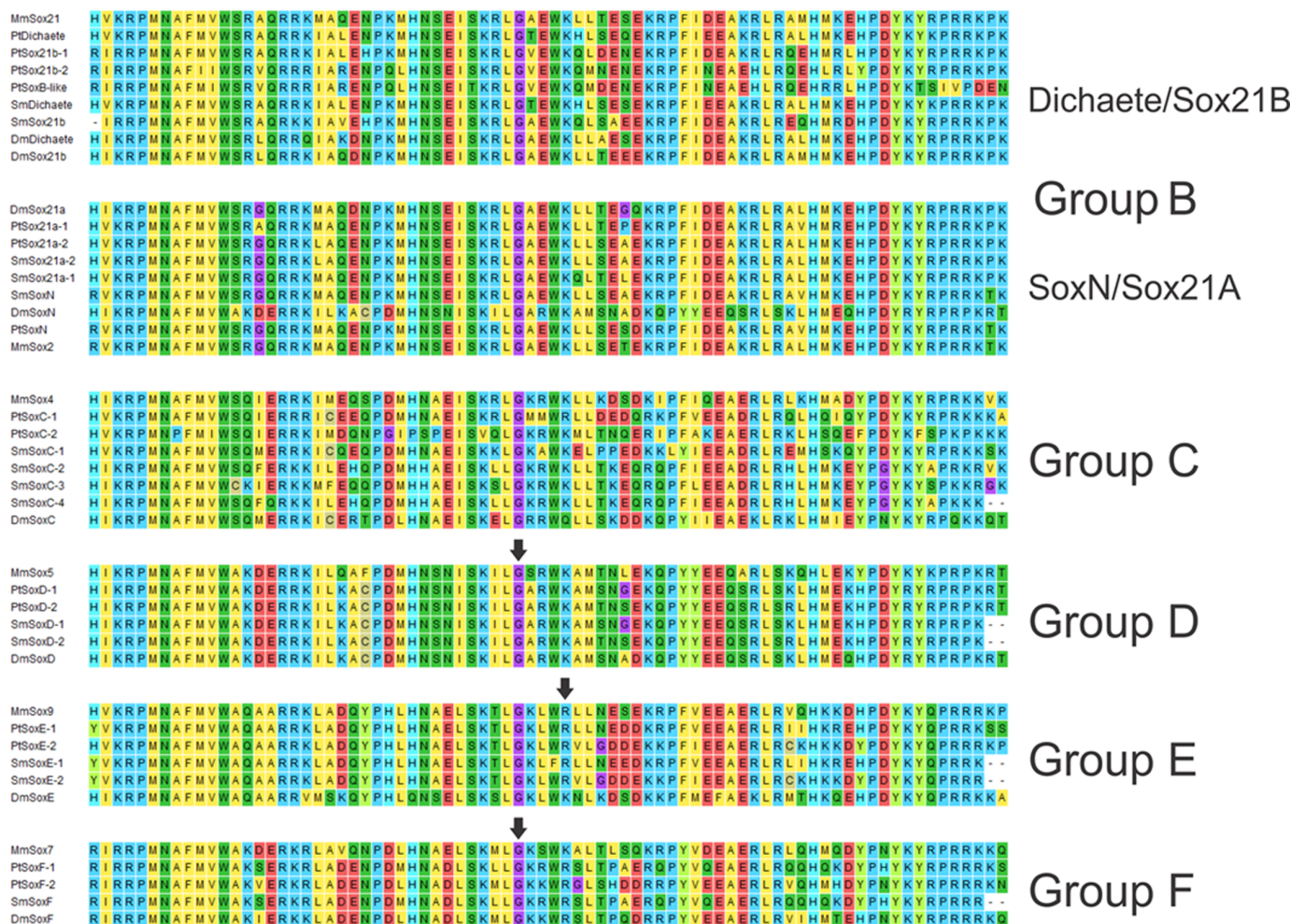


Figure 3. Multiple sequence alignment of HMG domain sequences from selected metazoans. *Mus musculus* (Mm), *D. melanogaster* (Dm), *P. tepidarius* (Pt) and *S. mimosarum* (Sm). Arrowheads indicate the locations of HMG domain introns and the bold underlined amino acids indicate the genes where the introns are present.

idea that these genes were formed by a tandem duplication in the protostome/deuterostome ancestor [11, 15]. The separation of *SoxN* from the *D/Sox21a-1* cluster in the spider suggests that either this fragmentation happened early in arthropod evolution [11] or that the duplication and separation of *SoxN* and *D* (or *Sox21a*) occurred early in Sox evolution [11, 15] (Figure 4).

Taken together, our analysis clearly shows that the spider genomes we examined have the full complement of Sox genes found in insects, mostly retained duplicates in Groups C, D, E and F after the WGD, and have a Group B organisation that more closely resembles insects than vertebrates.

Arrangement of *P. tepidarius* Sox genes after WGD

The phylogenetic relationships of Sox genes in *P. tepidarius* suggest that there are two paralogs of each Sox group except for *SoxN* and *D* (Figures 1 and 2). To investigate if all of these

duplicated Sox genes arose from the WGD event in the ancestor of these animals [41], the synteny of Sox genes was analysed in the *P. tepidarius* genome (Figure 4).

Most of the Sox genes in *P. tepidarius* were found dispersed in the genome on separate scaffolds consistent with the expectation that they arose via WGD. Analysis of the five upstream and five downstream genes flanking each Sox gene, however, revealed that dispersed duplicated Sox genes are generally not closely linked to other duplicated genes (Figure 4, Supplementary Table 2). While it is likely that this is a consequence of extensive loss of ohnologs and genomic rearrangements since the WGD 430 MYA, we cannot rule out that at least some of the duplicated Sox genes in this spider arose via tandem duplication followed by rearrangements after the WGD. The only tentative example of retained synteny was in the *SoxF* group where we found that the two *SoxF* genes of *P. tepidarius* have an upstream flanking sequence with

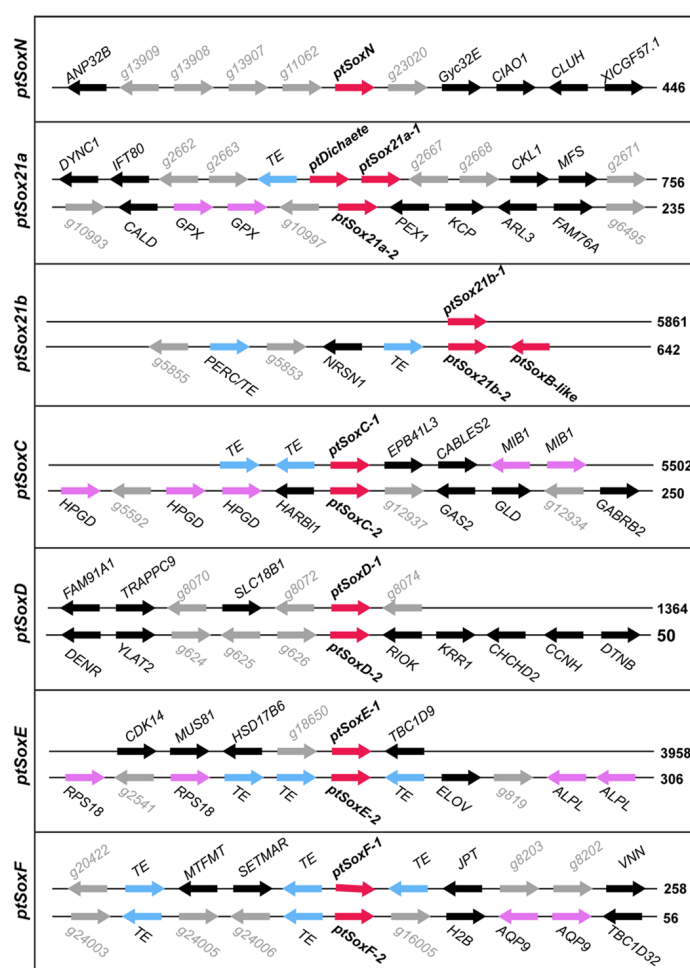


Figure 4. **Sox gene synteny in the *P. tepidariorum* genome.** The synteny of Sox genes (red) and flanking genes that have putative homology (black) compared between the Sox paralogs. Homology of flanking genes was also used to indicate tandem duplicates (pink), transposable elements (TEs) (blue). Genes that lack homology are shown in grey with their gene model IDs. Only the Sox genes were found in the same transcriptional orientation as upstream TEs. Of the thirteen Sox containing scaffolds, six scaffolds contained TEs that flank the Sox genes. Transcriptional direction is indicated by arrows. The DoveTail/HiRise scaffold ID numbers are given on the right.

homology to a transposable element (TE) with matching transcriptional orientation. Interestingly, six of the thirteen Sox containing scaffolds also have TE-like sequences nearby (Figure 4). TEs have previously been linked to the expansion of genes and their rearrangements [50, 51], however further analysis is needed to determine if TEs identified in this synteny analysis are involved in the evolution of Sox genes in spiders.

The exceptions to the dispersion of Sox genes in *P. tepidariorum* are *ptDichaete* and *ptSox21a-1* on scaffold #756 (as discussed above), as well as *ptSox21b-2* and *SoxB-like* on scaffold #642 (Figure 4). The sequences of the HMG domains of the clustered *ptSox21b-2* and *SoxB-like* genes grouped together with high bootstrap confidence indicative of a head-to-head tandem duplication (Figures 1 and 4). However, the HMG

domain of *SoxB-like* is split across two reading frames and although the sequence quality is poor in parts of this scaffold, its sequence similarity to *ptSox21b-2* suggests that *SoxB-like* may have been pseudogenised (Figure 4).

Sox Gene Expression during *P. tepidariorum* embryogenesis

We next studied the expression of Sox genes during embryogenesis in *P. tepidariorum*. For the Sox family genes *ptSox21a-1*, *ptSox21a-2*, *ptSox21b-2* and *D*, we did not detect any expression during embryogenesis. This might indicate that they are only expressed at very low levels or in a few cells or that these genes are used during post-embryonic development.

ptSoxN expression is visible from late stage 7 in the most anterior part of the germ band, a region corresponding to the presumptive neuroectoderm (Figure 5A). This head-specific expression in *P. tepidariorum* is similar to early expression of *SoxN* observed in *D. melanogaster* [52] and in *A. mellifera*, where *SoxB1* is expressed in the gastrulation fold and the anterior part of the presumptive neuroectoderm [13]. *ptSoxN* is subsequently expressed broadly in the developing head and follows neurogenesis in a progressive anterior-to-posterior pattern as new segments are added (Figure 5B). By mid stage 9, *ptSoxN* is strongly expressed in the head lobes and in the ventral nerve cord (Figure 5C), however, after this stage no further expression was detected. In both *D. melanogaster* and *A. mellifera*, *SoxN* expression is also observed throughout the neuroectoderm and becomes restricted to the neuroblasts [13, 18, 19].

In chelicerates, neurogenic progenitors were shown to delaminate in clusters of cells rather than single neuroblast-like cells found in dipterans and some hymenopterans [53]. However, even with these different modes of neurogenic differentiation, the expression of *SoxN* orthologues suggests this gene performs the same function. Indeed, the recent study by [45], of *T. castaneum*, *E. kanangrensis* and *G. marginata* also shows that the *SoxN* orthologues in these species have widespread and early neuroectodermal expression. Taken together these data clearly support the view that throughout the Bilateria a *SoxN* class protein is a marker of the earliest stages of neural specification.

Another member of the B group, *ptSox21b-1*, shows dynamic expression in the nascent prosomal segments and in the posterior segment addition zone (SAZ) from stage 7 (Figure 6A and B). At stage 8.2 expression is observed in the

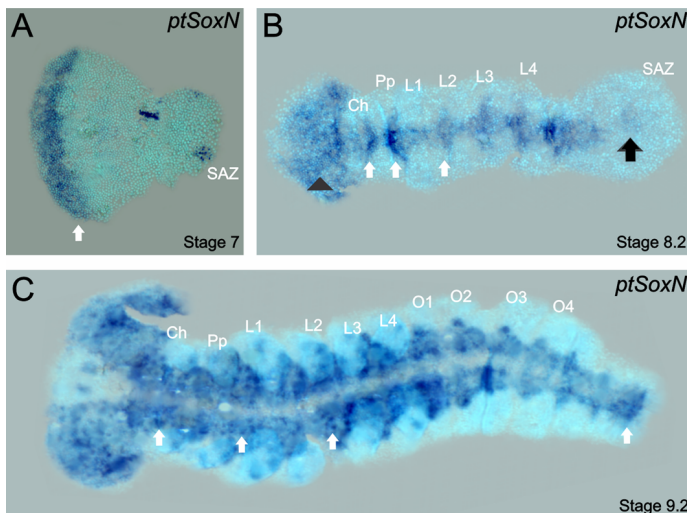


Figure 5. Expression of *ptSoxN* Flat-mount embryos at different stages of development after RNA *in situ* hybridization. A) *ptSoxN* expression is restricted to the presumptive neuroectoderm in the most anterior region of the germ band in stage 7 embryos (arrow). B) At stage 8.2, expression is in the most anterior part of the embryo (black arrowhead) and in the ventral nerve cord appearing sequentially from anterior to posterior, white arrows indicate expression in clusters that will subsequently broaden, expression in the posterior region adjacent to the SAZ is also observed (black arrowhead). C) At stage 9.2 expression is observed throughout the ventral nerve cord, with differentiating clusters indicated by arrows. Ch: chelicerae, L1 – L4: prosomal segments 1 to 4, O1 – O4: opisthosomal segments 1 to 4, Pp: pedipalps; SAZ: segment addition zone. Ventral views are shown for all embryos with the anterior to the left.

most anterior part of the germ band, which corresponds to the presumptive neuroectoderm in the future head and prosomal segments (Figure 6C). At stages 9 and 10, strong expression is apparent throughout the ventral nerve cord, similar to *ptSoxN*, and cyclical expression is also detected in the SAZ (Figure 6D and E).

In *T. castaneum*, *Sox21b* has similar expression to *D*, early in the SAZ and then in the developing CNS. In *E. kanangrensis* and *G. marginata*, there is no early *Sox21b* expression [45], however in these species, *D* is expressed during segmentation and then later in the CNS. This suggests that the role of *D* in segmentation in *D. melanogaster* and *T. castaneum* [32] could extend to *E. kanangrensis* and *G. marginata* but in spiders the closely related *Sox21b-1* gene may play this role.

For the Sox C group genes we did not detect any expression for *ptSoxC-2*. However, *ptSoxC-1* expression was detected at mid-stage 6, in a pattern similar to that of *ptSoxN* in the presumptive head and anterior segments (Figure 7A). By stage 8.2 expression is apparent in neuroectodermal progenitors along the germ band and at the anterior region of the SAZ (Figure 7B), however by stage 9.1 (Figure

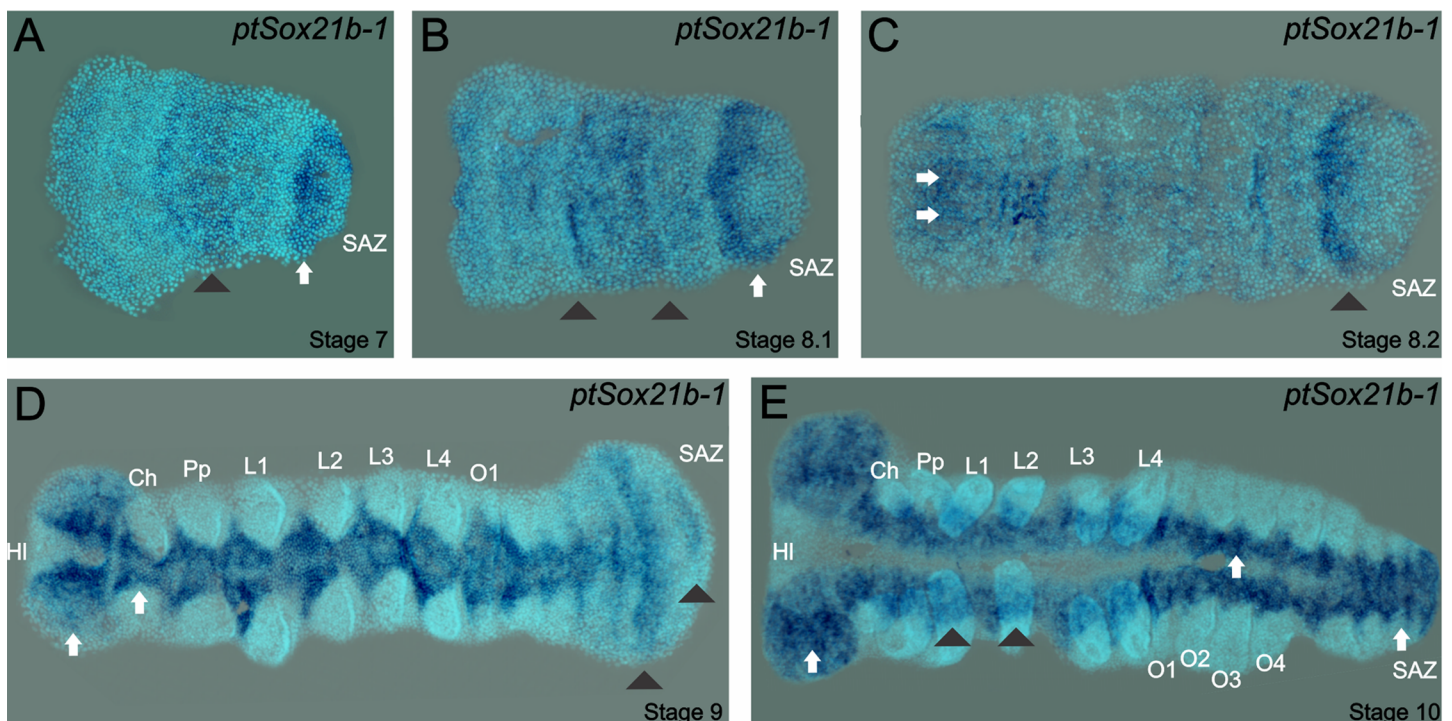


Figure 6. Expression of *ptSox21b-1* Flat-mount embryos at different stages of development after RNA *in situ* hybridization. A) *ptSox21b-1* expression is detected from mid-stage 7, where dynamic expression in the nascent segment (black arrowhead) and in the SAZ are indicated (white arrow). B) At stage 8.1, expression in the SAZ is dynamic (white arrow), and broadens in forming segments (black arrowheads). C) At stage 8.2, white arrows at the anterior indicate expression in the presumptive ventral nerve cord, with strong expression in the posterior SAZ still prominent (black arrowhead). D) At stage 9 strong expression in the entire anterior part of the ventral nerve cord is indicated by white arrows, expression is lower at the most posterior but remains dynamic in the SAZ (black arrowhead). E) At stage 10 expression is visible in the ventral nerve cord underneath the growing limb buds (black arrowheads) and becomes strong in the entire ventral nerve cord (white arrows). Ch: chelicerae, HL: head lobes, L1 – L4: prosomal segments 1 to 4, O1 – O4: opisthosomal segments 1 to 4, Pp: pedipalps; SAZ: segment addition zone. Ventral views are shown for all embryos with the anterior to the left.

7C) expression is lost from the SAZ. Interestingly, from stage 9.1, *ptSoxC-1* is expressed in the ventral nerve cord, from the head to the SAZ, however unlike the uniform expression of *ptSoxN*, *ptSoxC-1* is observed in clusters of cells, presumably undergoing neurogenic differentiation, progressively from the head through to opisthosomal segments as they differentiate in an anterior to posterior manner (Figure 7C).

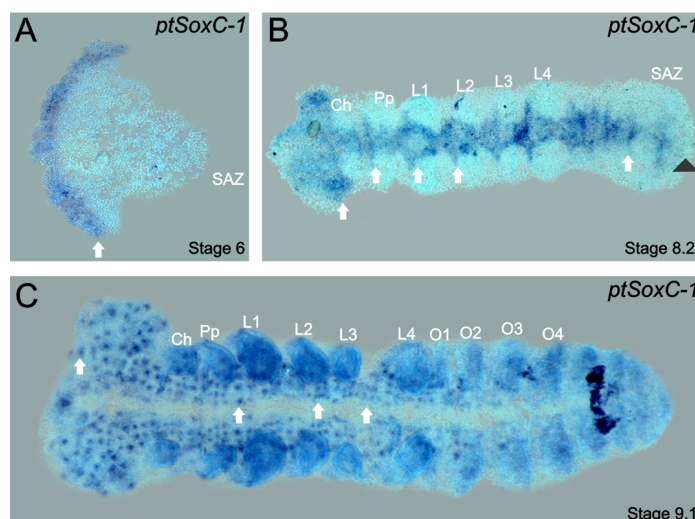


Figure 7. Expression of *ptSoxC-1*. Flat-mount embryos at different stages of development after RNA *in situ* hybridization. A) *ptSoxC-1* is strongly expressed in the presumptive neuroectoderm at stage 6 as indicated by the white arrow. B) At stage 8.2, strong expression is observed in the ventral nerve cord (white arrows) with the exception of the most posterior part of the SAZ (black arrowhead) C) At stage 9.1, expression is apparent in clusters of cells in the head and each anterior segment until the third opisthosomal segment (O3); white arrows indicate localized expression. The signal in the limb buds is background and staining at the most posterior part of the O5 segment is an artefact of incomplete chorion removal. Ch: chelicerae, L1 – L4: prosomal segments 1 to 4, O1 – O4: opisthosomal segments 1 to 4, Pp: pedipalps; SAZ: segment addition zone. Ventral views are shown for all embryos with the anterior to the left.

In *D. melanogaster* the single *SoxC* gene has been shown to play a role in the response to ecdysone at the onset of metamorphosis and has no known role in the embryonic CNS [27]. In contrast, the vertebrates *SoxC* genes (*Sox4*, 11 and 12) play critical roles in the differentiation of post-mitotic neurons, acting after the Group B genes, which specify neural progenitors [54]. In *A. mellifera*, late expression of the *SoxC* gene was observed in the embryonic cephalic lobes and in the mushroom bodies [13]. The expression of *SoxC* orthologues in the embryonic CNS of other invertebrates [45] suggests that this class of *Sox* gene may play a conserved role in aspects of neuronal differentiation, which has been lost in *D. melanogaster*. Interestingly, a comparison of target genes bound by *Sox11* in differentiating mouse neurons and *SoxN* in the *D. melanogaster* embryo shows a conserved set of neural

differentiation genes, suggesting that in *D. melanogaster* the role of *SoxC* in neuronogenesis has been taken over by *SoxN* [55].

We identified two genes in each of the *SoxD*, *E* and *F* families, however, we found no *in situ* evidence for expression of *SoxD-2*, *SoxE-2* or *SoxF-1* during the *P. tepidariorum* embryonic stages we examined. For *ptSoxD-2* we found no expression prior to stage 10, but we then observed expression in the ventral nerve cord from the head to the most posterior part of the opisthosoma (Figure 8A). The *D. melanogaster* *SoxD* gene is also expressed at later stages of embryonic CNS development [56] and has been shown to play roles in neurogenesis in the larval CNS [28]. While *SoxD* has been reported to be ubiquitously expressed in *A. mellifera* embryos, it is also expressed in the mushroom bodies of the adult brain [13]. Embryonic brain expression of *SoxD* orthologues in beetles, myriapods and velvet worms [45], as well as a known role for *SoxD* genes in aspects of vertebrate neurogenesis [54, 57] again suggests conserved roles for *SoxD* during metazoan evolution.

ptSoxE-1 is expressed in the developing limbs from stage 9 in small dots in the chelicerae, pedipalps and L1 buds, broader expression in L2 and L3, and in two dots in the L4 limb pairs, corresponding to the differentiating peripheral nervous system (PNS) (Figure 8B). We did not observe any expression of *ptSoxE-1* in opisthosomal segments 2 to 6 where the germline is believed to originate [58].

In *D. melanogaster* the *SoxE* orthologue is associated with both endodermal and mesodermal differentiation, is expressed in the embryonic gut, malpighian tubules and gonad [59], and has been shown to be required for testis differentiation during metamorphosis [29]. Both the *A. mellifera* *SoxE* genes are also expressed in the testis [13]. Janssen and colleagues [45], observed expression of *SoxE* genes in other invertebrates, associated with limb buds like in the spider, but they also detected posterior expression associated with gonadogenesis. These observations are particularly intriguing since the vertebrate *Sox9* gene has a crucial function in testis development [60]. Therefore, while we did not observe *SoxE* expression associated with early gonadogenesis it remains possible that the spider genes are used later in this process. We note that while the fly *SoxE* gene is expressed from the earliest stages of gonadogenesis, null mutant phenotypes are not apparent until the onset of metamorphosis [29].

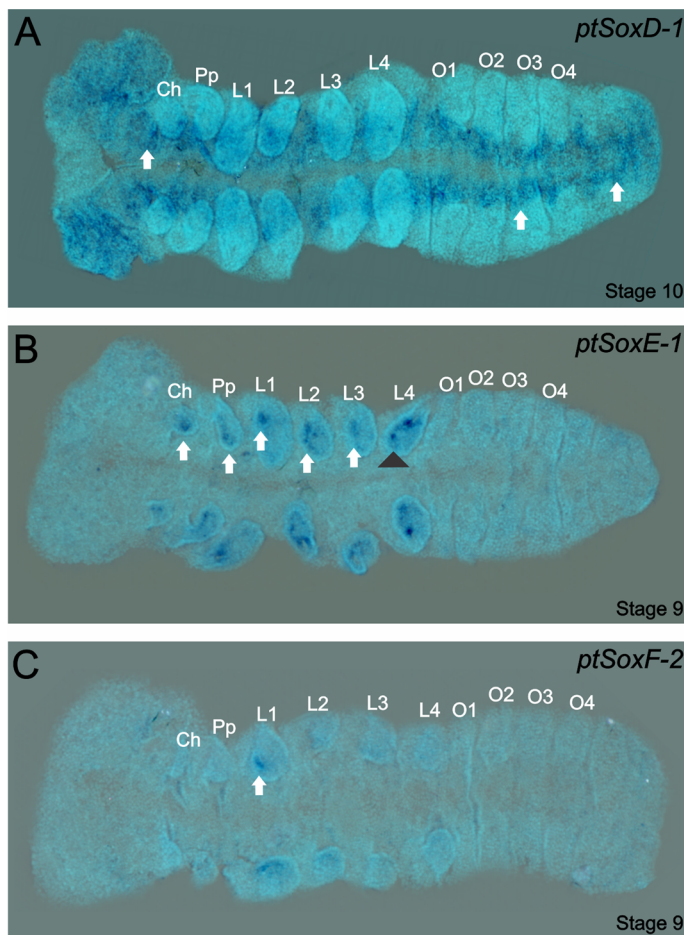


Figure 8. Expression of Sox D, E and F group orthologues. Flat-mount embryos at different stages of development after RNA *in situ* hybridization. A) *ptSoxD-1* expression is observed throughout the ventral nerve cord in stage 10 embryos as indicated by the arrows. B) *ptSoxE-1* expression at stage 9 is visible as single dots in the forming chelicerae, broader expression in the pedipalps and L1 to L3 (white arrows), and as two dots in the L4 limb bud as indicated by the black arrowhead. C) The expression of *ptSoxF-2* is only visible in the L1 limb buds forming at stage 9 (arrows). Ch: chelicerae, L1 – L4: prosomal segments 1 to 4, O1 – O4: opisthosomal segments 1 to 4, Pp: pedipalps; SAZ: segment addition zone. Ventral views are shown for all embryos with the anterior to the left.

In vertebrates, Group E genes are required in neural crest cells that contribute to the PNS [3, 61, 62] and we suggest the spider orthologue may have a similar function in the mechanoreceptors. These receptors are distributed all over the body, but the trichobothria only appear on the extremities of the limbs [63] where they differentiate from PNS progenitors.

Finally, the expression of *ptSoxF-2* is only detected at stage 9, in single dots at the tips of the L1 segment limb buds (Figure 8C). In *D. melanogaster* the *SoxF* gene is expressed in the embryonic PNS [56] and plays a role in the differentiation of sensory organ precursors [31], whereas in *A. mellifera* the *SoxF* orthologue is expressed ubiquitously throughout the embryo [13]. In *T. castaneum*, *E. kanangrensis* and *G. marginata* [45] *SoxF* expression is also associated with the

embryonic limbs, again suggesting that this was an ancestral function of this Sox family in the Euarthropoda.

Taken together, our study expands our understanding of a highly conserved family of transcriptional regulators that appear to have played prominent roles in metazoan evolution. Our analysis indicates that the classification of Sox genes in the invertebrates appears to be robust and that genes in all Groups have aspects of their expression patterns that suggest evolutionary conservation across the Bilateria. In particular, it is becoming increasingly clear that a *SoxN* orthologue (*SoxB1* in vertebrates) has a prominent role in the earliest aspects of CNS development. The finding that a *D/Sox21-b* class gene is implicated in the segmentation of both long and short germ band insects as well as the spider, and more widely in other arthropods [45], supports the view that formation of the segmented arthropod body plan is driven by an ancient mechanism [32], involving these Sox genes.

Conclusions

Our analysis provides insights into the fate of duplicate genes in organisms that have undergone WGD. We find that virtually all the duplicates have been retained in the spider genome but the expression analysis suggests that some have been possibly been subject to subfunctionalisation and/or neofunctionalisation. It is interesting to note that in teleost fish, which have also undergone WGD events, the pattern we observe for the Sox family in spiders is mirrored, with considerable gene retention and lineage-specific neofunctionalisation [64]. Indeed, future functional studies in *P. tepidariorum* will help to reveal the precise roles played by Sox genes during spider embryogenesis and how this relates to other metazoans.

Materials and Methods

Genome analysis

TBLASTN searches of the *P. tepidariorum* and *S. mimosarum* genomes were performed with the HMG domain of mouse Sox2 (UniProtKB - P48432) at <http://bioinf.uni-greifswald.de/blast/parasteatoda/bblast.php> and http://metazoa.ensembl.org/Stegodyphus_mimosarum/Info/Ind ex respectively. Gene models were retrieved from the *P. tepidariorum* Web Apollo genome annotations via <https://www.hgsc.bcm.edu/arthropods/common-house-spider-genome-project> and from http://metazoa.ensembl.org/Stegodyphus_mimosarum/Info/Ind

[ex](#). Sox gene sequences for other insects and vertebrates were retrieved from UniProt.

Multiple sequence alignments and phylogenetic analysis were performed with Clustal Omega [65] at <http://www.ebi.ac.uk/Tools/msa/clustalo/> or with MUSCLE and PhyLM 3.0 [66, 67] at <http://www.phylogeny.fr/index.cgi>. Pairwise sequence alignments were performed with SIM [68] at <http://web.expasy.org/sim/>.

Synteny analysis of Sox genes in *P. tepidariorum*

The synteny of Sox genes was analysed to determine whether Sox genes were duplicated in the proposed WGD [41]. AUGUSTUS gene models in *P. tepidariorum* are already mapped against the DoveTail/HiRise genome assembly [41] and using these data the locations of Sox genes along with five upstream and five downstream flanking genes were compared. Gene models were removed if they were partial, chimeric or artefacts of the AUGUSTUS annotation to the HiRise assembly. To infer putative homology of flanking genes, their protein sequences were compared with BLASTP to the NCBI non-redundant protein sequence database [69].

Embryo collection and procedures

Embryos were collected from adult female spiders from our temperature controlled (25°C) laboratory culture at Oxford Brookes University. Embryos at stages 5 to 12 were fixed as described in [70] and staged according to [71].

In situ hybridisation

RNA in situ hybridisation was carried out as in [70], with slight modifications. Proteinase K treatment and post-fixations steps were omitted, the probes were heated to 80°C for 5 minutes and immediately put on ice before adding to the pre-hybridization buffer. Nuclear staining was performed by incubation of embryos in 1 µg/ml 4-6-diamidino-2-phenylindol (DAPI) in PBS with 0.1% Tween-20 for 15 minutes. Embryos were mounted in glycerol on Poly-L-lysine (Sigma-Aldrich) coated coverslips, where the germband tissue attaches making it easier to remove the yolk before imaging. Images were taken with an AxioZoom V16 stereomicroscope (Zeiss) equipped with an Axiocam 506 mono and colour digital camera. Brightness and intensity of the pictures were adjusted in Corel PhotoPaint X5 (CorelDraw).

Gene isolation and cloning

Gene-specific cDNA fragments were amplified with primers designed with Primer Blast (<https://www.ncbi.nlm.nih.gov/tools/primer-blast/>) and PCR products cloned in the pCR4-TOPO vector (Invitrogen, Life Technologies). The primers to generate probe fragments for RNA *in situ* hybridization were designed to regions outside the consensus HMG domain to produce DNA fragments between 500-800 bp. The probes were *in vitro* transcribed as described in [70]. Primers and fragments size are described in Supplementary Table 3.

Supplementary Tables

Supplementary Table 1. HMG-domain and, where available full-length protein sequences from *D. melanogaster*, *P. tepidariorum*, *S. mimosarum* and *M. musculus*. Gene indicates the proposed names (or defined names for mouse). DB_Name indicates gene or gene model name from databases. DB_ID is the gene or protein accession. Scaffold indicates chromosome or genomic scaffold location. Annotation is the designation from spider annotations.

Supplementary Table 2. Gene and scaffold IDs of Sox and linked genes in the *P. tepidariorum* genome.

Supplementary Table 3. Genes, primers sequences and sizes for all the fragments used for in situ hybridisations.

End Matter

Availability of data and material

Gene models for *P. tepidariorum* and *S. mimosarum* were retrieved from <https://www.hgsc.bcm.edu/arthropods/common-house-spider-genome-project> and from http://metazoa.ensembl.org/Stegodyphus_mimosarum/Info/Ind [ex](#). Sox gene sequences for animals were retrieved from UniProt. The annotated *P. tepidariorum* genome is available at <https://i5k.nal.usda.gov/JBrowse-partep> and the assembly is deposited at NCBI: BioProject PRJNA167405 (Accession: AOMJ00000000).

Funding

This research was funded by a CNPq scholarship to CLBP (234586/2014-1), a grant from The Leverhulme Trust (RPG-2016-234) to APM and AS, and in part by a BBSRC grant (BB/N007069/1) to SR.

Authors' contributions

Experiments were performed by CLBP, SR, AS and DL. All authors contributed to data analysis and writing the manuscript.

Acknowledgements

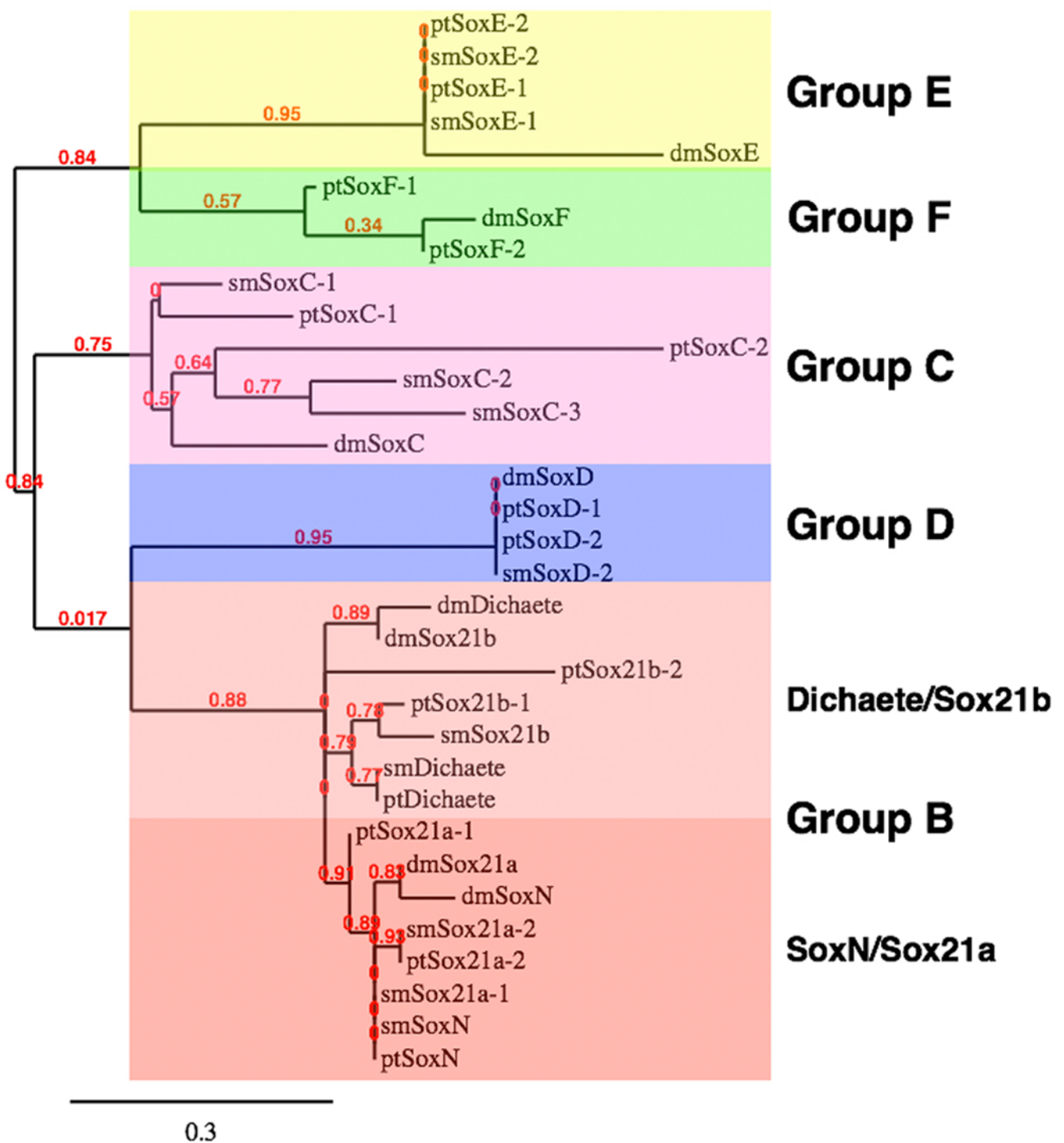
We thank Evelyn Schwager for assistance in identifying Sox genes in the *P. tepidariorum* genome. CLBP is immensely grateful for the help and discussion with the members of the Embryology course (Woods Hole – 2016), especially Joaquin Navajas Acedo (Stowers Institute – USA) for pushing to keep the ball rolling.

References

- Shimeld SM, Holland PW: **Vertebrate innovations**. *Proc Natl Acad Sci U S A* 2000, **97**(9):4449-4452.
- Larroux C, Luke GN, Koopman P, Rokhsar DS, Shimeld SM, Degnan BM: **Genesis and Expansion of Metazoan Transcription Factor Gene Classes**. *Mol Biol Evol* 2008, **25**(5):980-996.
- Kamachi Y, Kondoh H: **Sox proteins: regulators of cell fate specification and differentiation**. *Development* 2013, **140**(20):4129-4144.
- Sarkar A, Hochedlinger K: **The sox family of transcription factors: versatile regulators of stem and progenitor cell fate**. *Cell stem cell* 2013, **12**(1):15-30.
- Reiprich S, Wegner M: **From CNS stem cells to neurons and glia: Sox for everyone**. *Cell and tissue research* 2015, **359**(1):111-124.
- Sinclair A, Berta P, Palmer M, Hawkins J, Griffiths B, Smith M, Foster J, Frischauf A, Lovell-Badge R, Goodfellow P: **A gene from the human sex determining region encodes a protein with homology to a conserved DNA binding motif**. *Nature* 1990, **346**:240-244.
- Bowles J, Schepers G, Koopman P: **Phylogeny of the SOX family of developmental transcription factors based on sequence and structural indicators**. *Dev Biol* 2000, **227**:239-255.
- Heenan P, Zondag L, Wilson MJ: **Evolution of the Sox gene family within the chordate phylum**. *Gene* 2016, **575**(2 Pt 2):385-392.
- Phochanukul N, Russell S: **No backbone but lots of Sox: Invertebrate Sox genes**. *The international journal of biochemistry & cell biology* 2010, **42**(3):453-464.
- King N, Westbrook MJ, Young SL, Kuo A, Abedin M, Chapman J, Fairclough S, Hellsten U, Isogai Y, Letunic I et al: **The genome of the choanoflagellate *Monosiga brevicollis* and the origin of metazoans**. *Nature* 2008, **451**(7180):783-788.
- Zhong L, Wang D, Gan X, Yang T, He S: **Parallel Expansions of Sox Transcription Factor Group B Predating the Diversifications of the Arthropods and Jawed Vertebrates**. *PLoS ONE* 2011, **6**(1):e16570.
- Schnitzler CE, Simmons DK, Pang K, Martindale MQ, Baxevanis AD: **Expression of multiple Sox genes through embryonic development in the ctenophore *Mnemiopsis leidyi* is spatially restricted to zones of cell proliferation**. *Evodevo* 2014, **5**:15.
- Wilson MJ, Dearden PK: **Evolution of the insect Sox genes**. *BMC Evol Biol* 2008, **8**(1):120.
- Wei L, Cheng D, Li D, Meng M, Peng L, Tang L, Pan M, Xiang Z, Xia Q, Lu C: **Identification and characterization of Sox genes in the silkworm, *Bombyx mori***. *Mol Biol Rep* 2010:1-12.
- McKimmie C, Woerfel G, Russell S: **Conserved genomic organisation of group B Sox genes in insects**. *BMC Genetics* 2005, **6**:26.21-26.15.
- Russell SRH, Sanchez-Soriano N, Wright CR, Ashburner M: **The Dichaete gene of *Drosophila melanogaster* encodes a SOX-domain protein required for embryonic segmentation**. *Development* 1996, **122**:3669-3676.
- Nambu P, Nambu J: **The *Drosophila* fishhook gene encodes a HMG domain protein essential for segmentation and CNS development**. *Development* 1996, **122**:3467-3475.
- Buescher M, Hing FS, Chia W: **Formation of neuroblasts in the embryonic central nervous system of *Drosophila melanogaster* is controlled by SoxNeuro**. *Development* 2002:4193-4203.
- Overton P, Meadows L, Urban J, Russell S: **Evidence for differential and redundant function of the Sox genes Dichaete and SoxN during CNS development in *Drosophila***. *Development* 2002, **129**:4219-4228.
- Sanchez-Soriano N, Russell S: **The *Drosophila* Sox-domain protein Dichaete is required for the development of the central nervous system midline**. *Development* 1998, **125**:3989-3996.
- Niwa H, Nakamura A, Urata M, Shirae-Kurabayashi M, Kuraku S, Russell S, Ohtsuka S: **The evolutionally-conserved function of group B1 Sox family members confers the unique role of Sox2 in mouse ES cells**. In: *BMC Evol Biol*. BMC Evolutionary Biology; 2016: 1-12.
- Aleksic J, Ferrero E, Fischer B, Shen SP, Russell S: **The role of Dichaete in transcriptional regulation during *Drosophila* embryonic development**. *BMC Genomics* 2013, **14**:861.
- Ferrero E, Fischer B, Russell S: **SoxNeuro orchestrates central nervous system specification and differentiation in *Drosophila* and is only partially redundant with Dichaete**. *Genome Biol* 2014, **15**(5):R74.
- Carl S, Russell S: **Common binding by redundant group B Sox proteins is evolutionarily conserved in *Drosophila***. *BMC Genomics* 2015, **16**:292.
- Meng FW, Biteau B: **A Sox Transcription Factor Is a Critical Regulator of Adult Stem Cell Proliferation in the *Drosophila* Intestine**. *Cell Rep* 2015, **13**(5):906-914.
- Chen J, Xu N, Huang H, Cai T, Xi R: **A feedback amplification loop between stem cells and their progeny promotes tissue regeneration and tumorigenesis**. *eLife* 2016, **5**.
- Ritter AR, Beckstead RB: **Sox14 is required for transcriptional and developmental responses to 20-hydroxyecdysone at the onset of *drosophila* metamorphosis**. *Dev Dyn* 2010, **239**(10):2685-2694.
- Li A, Hooli B, Mullin K, Tate R, Bubnys A, Kirchner R, Chapman B, Hofmann O, Hide W, Tanzi RE: **Silencing of the *Drosophila* ortholog of Sox5 leads to abnormal neuronal development and behavioural impairment**. In: *Hum Mol Gen*. 2017: 1-36.
- Nanda S, Defalco T, Hui Yong Loh S, Phochanukul N, Camara N, Van Doren M, Russell S: **Sox100B, a *Drosophila* Group E Sox-domain Gene, Is Required for Somatic Testis Differentiation**. *Sex Dev* 2009, **3**(1):26-37.
- Dichtel-Danjoy M-L, Caldeira J, Casares F: **SoxF is part of a novel negative-feedback loop in the wingless pathway that controls proliferation in the *Drosophila* wing disc**. *Development* 2009, **136**(5):761-769.
- Miller SW, Avidor-Reiss T, Polyanovsky A, Posakony JW: **Complex interplay of three transcription factors in controlling the tormogen differentiation program of *Drosophila* mechanoreceptors**. *Dev Biol* 2009:1-14.
- Clark E, Peel A: **Evidence for the temporal regulation of insect segmentation by a conserved set of developmental transcription factors**. *bioRxiv* 2017, doi.org/10.1101/145151.
- Fortunato S, Adamski M, Bergum B, Guder C, Jordal S, Leininger S, Zwafink C, Rapp HT, Adamska M: **Genome-wide analysis of the sox family in the calcareous sponge *Sycon ciliatum*: multiple genes with unique expression patterns**. *Evodevo* 2012, **3**(1):14.

34. Richards GS, Rentzsch F: **Regulation of Nematostella neural progenitors by SoxB, Notch and bHLH genes.** *Development* 2015, **142**(19):3332-3342.
35. Pioro HL, Stollewerk A: **The expression pattern of genes involved in early neurogenesis suggests distinct and conserved functions in the diplopod Glomeris marginata.** *Dev Genes Evol* 2006, **216**(7-8):417-430.
36. Stork NE, McBroom J, Gely C, Hamilton AJ: **New approaches narrow global species estimates for beetles, insects, and terrestrial arthropods.** *Proc Natl Acad Sci U S A* 2015, **112**(24):7519-7523.
37. Valentine JW, Jablonski D, Erwin DH: **Fossils, molecules and embryos: new perspectives on the Cambrian explosion.** *Development* 1999, **126**(5):851-859.
38. Schmidt-Ott U, Lynch JA: **Emerging developmental genetic model systems in holometabolous insects.** *Curr Opin Genet Dev* 2016, **39**:116-128.
39. Leite DJ, McGregor AP: **Arthropod evolution and development: recent insights from chelicerates and myriapods.** *Curr Opin Genet Dev* 2016, **39**:93-100.
40. Akiyama-Oda Y, Oda H: **Multi-color FISH facilitates analysis of cell-type diversification and developmental gene regulation in the Parasteatoda spider embryo.** *Dev Growth Differ* 2016, **58**(2):215-224.
41. Schwager EE, Sharma PP, Clarke T, Leite DJ, Wierschin T, Pechmann M, Akiyama-Oda Y, Esposito L, Bechsgaard J, Bilde T et al: **The house spider genome reveals an ancient whole-genome duplication during arachnid evolution.** *BMC Biol* 2017, **15**(1):62.
42. Leite DJ, Ninova M, Hilbrant M, Arif S, Griffiths-Jones S, Ronshaugen M, McGregor AP: **Pervasive microRNA Duplication in Chelicerates: Insights from the Embryonic microRNA Repertoire of the Spider Parasteatoda tepidarium.** *Genome Biol Evol* 2016, **8**(7):2133-2144.
43. Hilbrant M, Damen WG, McGregor AP: **Evolutionary crossroads in developmental biology: the spider Parasteatoda tepidarium.** *Development* 2012, **139**(15):2655-2662.
44. Sanggaard KW, Bechsgaard JS, Fang X, Duan J, Dyrland TF, Gupta V, Jiang X, Cheng L, Fan D, Feng Y et al: **Spider genomes provide insight into composition and evolution of venom and silk.** *Nat Commun* 2014, **5**:3765.
45. Janssen R, Andersson E, S B, Fowler W, Höök L, Leyher J, Landström E, Mannelqvist A, Panara V, Smith K et al: **Embryonic expression patterns and phylogenetic analysis of panarthropod Sox genes: insight into nervous system development, segmentation and gonadogenesis.** *Submitted* 2017.
46. Pevny L, Placzek M: **Sox genes and neural progenitor identity.** *Curr Opin Neurobiol* 2005, **15**:7-13.
47. Uchikawa M, Kamachi Y, Kondoh H: **Two distinct subgroups of Group B Sox genes for transcriptional activators and repressors: their expression during embryonic organogenesis of the chicken.** *Mech Dev* 1999, **84**:103-120.
48. Popovic J, Stanisavljevic D, Schwirtlich M, Klajn A, Marjanovic J, Stevanovic M: **Expression analysis of SOX14 during retinoic acid induced neural differentiation of embryonic carcinoma cells and assessment of the effect of its ectopic expression on SOXB members in HeLa cells.** *PLoS One* 2014, **9**(3):e91852.
49. Shen SP, Aleksic J, Russell S: **Identifying targets of the Sox domain protein Dichaete in the Drosophila CNS via targeted expression of dominant negative proteins.** *BMC Dev Biol* 2013, **13**:1.
50. Faino L, Seidl MF, Shi-Kunne X, Pauper M, van den Berg GC, Wittenberg AH, Thomma BP: **Transposons passively and actively contribute to evolution of the two-speed genome of a fungal pathogen.** *Genome Res* 2016, **26**(8):1091-1100.
51. Levine MT, Vander Wende HM, Hsieh E, Baker EP, Malik HS: **Recurrent Gene Duplication Diversifies Genome Defense Repertoire in Drosophila.** *Mol Biol Evol* 2016, **33**(7):1641-1653.
52. Cremazy F, Berta P, Girard F: **SoxNeuro, a new Drosophila Sox gene expressed in the developing central nervous system.** *Mech Dev* 2000, **93**:215-219.
53. Stollewerk A, Chipman AD: **Neurogenesis in myriapods and chelicerates and its importance for understanding arthropod relationships.** *Integr Comp Biol* 2006, **46**(2):195-206.
54. Tanaka S, Kamachi Y, Tanouchi A, Hamada H, Jing N, Kondoh H: **Interplay of SOX and POU Factors in Regulation of the Nestin Gene in Neural Primordial Cells.** *Mol Cell Biol* 2004, **24**:8834-8846.
55. Bergsland M, Ramskold D, Zaouter C, Klum S, Sandberg R, Muhr J: **Sequentially acting Sox transcription factors in neural lineage development.** *Genes and Development* 2011, **25**:2453-2464.
56. Cremazy F, Berta P, Girard F: **Genome-wide analysis of Sox genes in Drosophila melanogaster.** *Mech Dev* 2001, **109**:371-375.
57. Lefebvre V: **The SoxD transcription factors – Sox5, Sox6, and Sox13 – are key cell fate modulators.** *Int J Biochem Cell Biol* 2010, **42**:429–432.
58. Schwager EE, Meng Y, Extavour CG: **vasa and piwi are required for mitotic integrity in early embryogenesis in the spider Parasteatoda tepidarium.** *Dev Biol* 2015, **402**(2):276-290.
59. Loh SHY, Russell S: **A Drosophila group E Sox gene is dynamically expressed in the embryonic alimentary canal.** *Mech Dev* 2000, **93**:4.
60. Vidal VPI, Charboissier M-C, deRooij DG, Schedl A: **Sox9 induces testis development in XX transgenic mice.** *Nat Genet* 2001, **28**:216-217.
61. Bell DM, Leung KK, Wheatley SC, Ng LJ, Zhou S, Ling KW, Sham MH, Koopman P, Tam PP, Cheah KS: **SOX9 directly regulates the type-II collagen gene.** *Nat Genet* 1997, **16**(2):174-178.
62. Stolt CC, Wegner M: **SoxE function in vertebrate nervous system development.** *Int J Biochem Cell Biol* 2010, **42**:437–440.
63. Stollewerk A, Weller M, Tautz D: **Neurogenesis in the spider Cupiennius salei.** *Development* 2001, **128**(14):2673-2688.
64. Voldoire E, Brunet F, Naville M, Volff JN, Galiana D: **Expansion by whole genome duplication and evolution of the sox gene family in teleost fish.** *PLoS One* 2017, **12**(7):e0180936.
65. Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, Lopez R, McWilliam H, Remmert M, Soding J et al: **Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega.** *Mol Syst Biol* 2011, **7**:539.
66. Edgar RC: **MUSCLE: a multiple sequence alignment method with reduced time and space complexity.** *BMC Bioinformatics* 2004, **5**:113.
67. Dereeper A, Guignon V, Blanc G, Audic S, Buffet S, Chevenet F, Dufayard JF, Guindon S, Lefort V, Lescot M et al: **Phylogeny.fr: robust phylogenetic analysis for the non-specialist.** *Nucleic Acids Res* 2008, **36**(Web Server issue):W465-469.
68. Huang X, Miller W: **A Time-Efficient Linear-Space Local Similarity Algorithm.** *Advances in Applied Mathematics* 1991, **12**:337-357.
69. Altschul S, JC W, EM G, R A, A M, AA S, Yu Y: **Protein database searches using compositionally adjusted substitution matrices.** *FEBS J* 2005, **272**:5101-5109.
70. Akiyama-Oda Y, Oda H: **Early patterning of the spider embryo: a cluster of mesenchymal cells at the cumulus produces Dpp signals received by germ disc epithelial cells.** *Development* 2003, **130**:1735-1747.
71. Mittmann B, Wolff C: **Embryonic development and staging of the cobweb spider Parasteatoda tepidarium C. L. Koch, 1841 (syn.: Achaearanea tepidarium; Araneomorphae; Theridiidae).** *Dev Genes Evol* 2012, **222**(4):189-216.

>



Supplementary Figure 1. Phylogeny of Group B Sox HMG domains

PhyLM tree and multiple sequence alignment of group B HMG domains from *Mus musculus* (Mm), *Drosophila melanogaster* (Dm), *Anopheles gambiae* (Ag), *Tribolium castaneum* (Tc), *Parasteatoda tepidariorum* (Pt) and *Stegodyphus mimosarum* (Sm). Branch support values from PhyML are indicated in red. Arrow indicates the conserved Isoleucine residue indicative of invertebrate Dichaete/Sox21b class genes [15].

