Duplication and expression of Sox genes in spiders

Christian L. Bonatto Paese, Daniel J. Leite, Anna Schönauer, Alistair P. McGregor and Steven Russell

Abstract

Background: The Sox family of transcription factors is an important part of the genetic ‘toolbox’ of all metazoans examined to date and is known to play important developmental roles in vertebrates and insects. However, outside the commonly studied Drosophila model little is known about the repertoire of Sox family transcription factors 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 suggests 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 support the idea of conserved functions for some of these genes, including a potential role for SoxC and SoxD genes in CNS development and SoxF in limb development.

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, in particular, aspects of stem cell biology and nervous system development [3–5].

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 is represented by 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 with other groups specific to particular lineages [9]. While Sox-like sequences have been reported in the genome of the choanoflagellate Monosiga brevicollis, 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 [10–12].

While vertebrate Sox genes have been intensively studied due to their critical roles in development, with the exception of the fruit fly Drosophila melanogaster, they are less well characterised in invertebrates [3]. 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]. Dichaete (D) plays critical roles in early segmentation and nervous system development, while SoxNeuro (SoxN) is essential for CNS development, and where the expression of these two genes overlaps in the embryonic CNS they exhibit phenotypic redundancy [16–19].

The evolutionary conservation of Sox protein sequence and function has been shown in rescue or swap experiments, where mouse Sox2 rescues Dichaete null mutant phenotypes in the D. melanogaster embryo and Drosophila SoxN can replace Sox2 in mouse ES cells [20, 21]. Furthermore, a comparison of Dichaete 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 1000 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 but there is no known function for Sox21b [25, 26]. 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 similar role for Dichaete in the early 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 has been reported in the chelicerate P. tepidariorum [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, as well as retention of some paralogs generated from tandem duplications. Furthermore, while group B genes show highly conserved expression in the developing CNS, the expression of other spider Sox genes suggests they have evolved 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 15 and 14 sequences respectively. All but three of these contained the highly conserved RPMNAFMVW motif that is characteristic of Sox proteins and the three exceptions (ptSoxC-2, ptSoxB-like and ptSox21b-2) only show minor
conservative substitutions in this motif (see Fig. 3 for full alignments). 14 of the \textit{P. tepidariorum} sequences corresponded to annotated gene models. Moreover, two sequences were identical (\textit{ptSox21b–I}, aug3.24914.t1 and aug3.g24896.t1) and 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 (\textit{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 our subsequent analysis.

In the case of \textit{S. mimosarum} we identified 14 genomic regions, 11 of which correspond to annotated Sox genes. Reciprocal BLAST searches of \textit{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 \textit{D. melanogaster} capicua (\textit{cic}) and bobby sox (\textit{bbx}) genes in both spider species but here we do not consider these Sox-related genes further.

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 \textit{D. melanogaster} Sox genes and representatives of each subgroup from mouse (Additional file 1: Table S1). These analyses resulted in a clear classification of spider Sox genes into groups B–F as found in other invertebrate genomes (Fig. 1). Note that Group A only contains the \textit{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 \textit{D. melanogaster} yielded virtually identical results (Additional file 2: Figure S1). Following the recommended nomenclature for Sox genes [7], we have named the spider Sox genes as indicated in Additional file 1: Table S1. The naming of \textit{D. melanogaster} Sox genes is confusing with some carrying historic names based on their phenotype (\textit{Dichaete} and \textit{SoxN}), others named after cytological locations (\textit{Sox100B} and \textit{Sox102F}) and others with inappropriate numerical designations (\textit{D. melanogaster} Sox14 is a Group C gene while in vertebrates Sox14 is in Group B and \textit{D. melanogaster} Sox15 is in group F, while vertebrate Sox15 is in Group G). For these reasons we propose renaming the \textit{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 [45]. 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*, *Dichaete*, *Sox21a* and *Sox21b* (Additional file 1: Table S1).

In common with many other gene families in spiders [41], the Sox genes are mostly represented by two or more copies in each group (Fig. 2). In other arthropods examined to date, as well as the onychophoran *Euperipatoides kanangrensis* [46], there is usually only a single copy of each gene, although there is a recent report of two Group E genes in the millipede *G. marginata* [46]. In the case of spider Groups D and E, the duplications likely 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 (Fig. 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 (Additional file 2: Figure S1), *ptSoxC-1* groups with *smSoxC-1* and *ptSoxC-2* with *smSoxC-2*. *smSoxC-2* has undergone a local head-to-head duplication, with *smSoxC-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 Fig. 3). While there is an intron within the region encoding the DNA binding domains of 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.

While the Group B genes of insects and vertebrates show considerable sequence similarity in their DNA binding domains, they are clearly different in terms of their genomic organisation and functions. Vertebrate Group B genes are not linked in the genome and are subdivided into B1 (Sox1, 2 and 3) and B2 (Sox14 and 21). This classification manifests both at sequence and 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 [47–49]. 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 (*Dichaete*, *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 gene arrangement 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 Dichaete mainly acts as a transcriptional activator, a role inconsistent with that observed for vertebrate SoxB2 proteins [22, 50].

The phylogenies generated with the HMG domains from a range of species (Fig. 1; Additional file 2: Figure S1) or full-length proteins sequences from spiders and *D. melanogaster* (Additional file 3: Figure S2) 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 Dichaete-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 (Fig. 4), an organisation reminiscent of that found in insects. Indeed, the linkage of ptDichaete and ptSox21a-1 supports the idea that these genes were formed by a tandem duplication in the protostome/deuterostome ancestor [11, 15]. The separation of SoxN from the Dichaete/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 Dichaete (or Sox21a) occurred early in Sox evolution [11, 15] (Fig. 4).

Taken together, our analysis clearly shows that the spider genomes we examined have the full complement of Sox genes found in insects, have 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.
orientation (Additional file 4: Table 2 and Additional file 5: Table S3). These observations further evidence, in conjunction with phylogenetic relationships, that Group D genes were duplicated in the ancestor of both spiders.

The only tentative example of retained synteny within a species was in the SoxF group, where we found that the two SoxF genes of *P. tepidariorum* have an upstream flanking sequence with homology to a transposable element (TE) with matching transcriptional orientation. Interestingly, six of the thirteen *P. tepidariorum* Sox containing scaffolds also have TE-like sequences nearby (Fig. 4). Furthermore, of the nine *S. mimosarum* scaffolds that have flanking gene information, three have TEs flanking Sox genes (Additional file 5: Table S3). TEs
have previously been linked to the expansion of genes and their rearrangements [51, 52], 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), *ptSox21b-2* and *SoxB-like* on scaffold #642 (Fig. 4), as well as *smSoxC-2* and *smSoxC-3* that are adjacent on scaffold #4648 (Additional file 4: Table S2). 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 (Figs. 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, it’s sequence similarity to *ptSox21b-2* suggests that *SoxB-like* may have been pseudogenised (Fig. 4).

**Sox gene expression during *P. tepidariorum* embryogenesis**

We next studied the expression of Sox genes during embryogenesis in *P. tepidariorum* using in situ hybridisation. For the SoxB family genes *ptSox21a-1*, *ptSox21a-2*, *ptSox21b-2* and *Dichaete*, we did not detect any expression during embryogenesis. This might indicate that they are only expressed at very low levels, only 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 (Fig. 5a). This head-specific expression in *P. tepidariorum* is similar to early expression of *SoxN* observed in *D. melanogaster* [53] 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 (Fig. 5b). By mid stage 9, *ptSoxN* is strongly expressed in the head lobes and in the ventral nerve cord (Fig. 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 delaminate in clusters of cells rather than single neuroblast-like cells
found in dipterans and some hymenopterans [54]. 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 of *T. castaneum*, *E. kanangrensis* and *G. marginata* also shows that the SoxN orthologues in these species have widespread and early neuroectodermal expression [46]. Taken together with published SoxN expression, our results clearly support the view that throughout the Bacteria a SoxN class protein is a marker of the earliest stages of neural specification.

Another member of the B group, *ptSox21b-1*, shows expression in the nascent prosomal segments and in the posterior segment addition zone (SAZ) from stage 7 (Fig. 6a and b). At stage 8.2 expression is observed in the most anterior part of the germ band, which corresponds to the presumptive neuroectoderm in the future head and prosomal segments (Fig. 6c). At stages 9 and 10, strong expression is apparent throughout the ventral nerve cord, similar to *ptSoxN*. Comparing expression in the SAZ at different stages in these fixed preparations suggest that Sox21b-1 may be dynamic in this region (Fig. 6d and e).

In *T. castaneum*, Sox21b has similar expression to insect Dichaete genes, early in the SAZ and then in the developing CNS. In *E. kanangrensis* and *G. marginata*, there is no early Sox21b expression [46], however, in these species Dichaete is expressed during segmentation and then later in the CNS. This suggests that the role of Dichaete in *D. melanogaster* and *T. castaneum* segmentation [32] could extend to *E. kanangrensis* and *G. marginata*, whereas in spiders the closely related Sox21b-1 gene may play this role. The widespread expression of both SoxN and Sox21b-1 throughout the neuroectoderm strongly suggest that, as has been shown in vertebrates and flies, many cells in the developing CNS co-express two related SoxB genes. We confirmed their overlapping expression in the CNS, but not in the SAZ, with double in situ hybridisations, using *SoxN* and *Sox21b-1* probes (Additional file 6: Figure S3). While both genes clearly show extensive expression overlap throughout the developing CNS, we were interested to note that at the very lateral regions of the neuroectoderm, Sox21b-1 is uniquely expressed. This is similar to the situation in *Drosophila* where SoxN has a unique lateral expression domain [18, 19].

In the case of the Sox C genes, we did not detect any expression for *ptSoxC-2*. However, *ptSoxC-1* expression was found at mid-stage 6, in a pattern similar to that of

---

**Fig. 6** Expression of *ptSox21b-1*. Flat-mounted embryos at different stages of development after RNA in situ hybridization. **a** *ptSox21b-1* expression is detected from mid-stage 7 in the nascent segment (black arrowhead) and in the SAZ (white arrow). **b** At stage 8.1, expression in the SAZ appears to be dynamic (white arrow, c.f. Figure 6a), 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 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 appears to remain dynamic in the SAZ (black arrowhead). **e** At stage 10, expression is visible in the ventral nerve cord beneath the growing limb buds (black arrowhead) 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. Scale bars: 150 μm
ptSoxN in the most anterior part of the germ band in the presumptive neuroectoderm (Fig. 7a). By stage 8.2 expression is apparent in neuroectodermal progenitors along the germ band and at the anterior region of the SAZ (Fig. 7b), however by stage 9.1 (Fig. 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 (Fig. 7c).

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 vertebrate 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 [55]. 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 [46] 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 [23, 56].

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 (Fig. 8a). The *D. melanogaster* SoxD gene is also expressed at later stages of embryonic CNS development [57] 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 [46], as well as a known role for SoxD genes in aspects of vertebrate neurogenesis [55, 58],
again suggests conserved roles for SoxD during metazoan evolution.

*ptSoxE-1* is expressed in the developing limbs from stage 9 in small regions of the chelicerae, pedipalps and L1 buds, with broader expression in L2 and L3, and in two prominent foci in the L4 limbs, that correspond to the differentiating peripheral nervous system (PNS) (Fig. 8b). At the stages we examined we did not observed any expression of *ptSoxE-1* in opisthosomal segments 2 to 6 where the germline is believed to originate [59].

In *D. melanogaster*, the SoxE orthologue is associated with both endodermal and mesodermal differentiation, is expressed in the embryonic gut, malpighian tubules and gonad [60], 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 observed expression of SoxE genes in other invertebrates, associated with limb buds as we observed in the spider, but they also detected posterior expression associated with gonadogenesis [46]. These observations are particularly intriguing since the vertebrate Sox9 gene has a crucial function in testis development [61]. 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]. In vertebrates, Group E genes are required in neural crest cells that contribute to the PNS [3, 62, 63] 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 [64] where they differentiate from PNS progenitors.

Finally, the expression of ptSoxF-2 is only detected at stage 9, in single foci at the tips of the L1 segment limb buds (Fig. 8c). In D. melanogaster the SoxF gene is expressed in the embryonic PNS [57] 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. kanangensis and G. marginata, SoxF expression is also associated with the embryonic limb buds [46], 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 Dichaete/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 [46], 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 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 neo-functionalisation [65]. Clearly, 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/blast.php and http://metazoa.ensembl.org/Stegodyphus_mimosarum/Info/Index respectively. Gene models were retrieved from the P. tepidariorum Web Apollo genome annotations via https://apollo.nal.usda.gov/partep/jbrowse/ and from http://metazoa.ensembl.org/Stegodyphus_mimosarum/Info/Index. Sox gene sequences for other insects and vertebrates were retrieved from UniProt https://www.uniprot.org.

Multiple sequence alignments and phylogenetic analysis were performed with Clustal Omega [66] at http://www.ebi.ac.uk/Tools/msa/clustalo/ or with MUSCLE and PhyLM 3.0 [67, 68] at http://www.phylogeny.fr/index.cgi. Pairwise sequence alignments were performed with SIM [69] at http://web.expasy.org/sim/.

Synteny analysis of Sox genes in P. tepidariorum and S. mimosarum

The synteny of Sox genes was analysed to determine whether Sox genes were duplicated during the reported WGD [41].

For P. tepidariorum the AUGUSTUS gene models 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 [70].

For S. mimosarum the Sox gene models and their location in the genome were obtained from [44]. Similar to P. tepidariorum, the synteny of the five upstream and five downstream genes relative to each Sox gene were compared. Annotations of flanking genes was previously performed by Sanggaard et al [44].

Embryo collection and procedures

Embryos were collected from adult female spiders from the temperature controlled (25 °C) laboratory culture at
Oxford Brookes University. Embryos at stages 5 to 12 were fixed as described in [71] and staged according to [72].

**In situ hybridisation**

RNA in situ hybridisation was carried out as described in [71], with the following minor modifications: Proteinase K treatment and post-fixations steps in the original protocol were omitted, and prior to hybridization, the probes were heated to 80 °C for 5 min and immediately put on ice before adding to the pre-hybridization buffer. Fluorescent in situ hybridization was performed following [40]. Tyramide Signal Amplification (TSA) was performed with TSA kits from PerkinElmer (TSA Fluorescein and TSA Cyanine). Post hybridisation, nuclear staining was achieved by incubating embryos in 1 µg/ml 4–6-diamidino-2-phenylindol (DAPI) in PBS with 0.1% Tween-20 for 15 min. 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 microscope (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). Gene-specific cDNA fragments were amplified with Gene isolation and cloning [71], with the following minor modifications: Proteinase K treatment and post-fixations steps in the original protocol were omitted, and prior to hybridization, the probes were heated to 80 °C for 5 min and immediately put on ice before adding to the pre-hybridization buffer. Fluorescent in situ hybridisation was carried out as described in [71] and staged according to [72].

**Additional files**

- **Additional file 1**: Table S1. 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_JD is the gene or protein accession. Scaffold indicates chromsome or genomic scaffold location. Annotation is the designation from spider annotations. (XLSX 53 kb)
- **Additional file 2**: Figure S1. 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. (PNG 1363 kb)
- **Additional file 3**: Figure 2. Phylogeny of full-length Sox proteins from Drosophila and spiders. PhyLM tree of Sox genes from D. melanogaster (Dm), P. tepidariorum (Pt) and S. mimosarum (Sm) based on available full-length protein sequence (Additional file 1: Table S1). Branch support values from PhyML are indicated in red. (DOCX 15 kb)

**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 Accessed (Stowers Institute – USA) for pushing to keep the ball rolling.

**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.

**Availability of data and materials**

Gene models for P. tepidariorum and S. mimosarum were retrieved from https://www.hgsc.bcm.edu/arthropods/common-house-spidere-genome-project and from http://metazoa.ensembl.org/Stegodyphus_mimosarum/Info/Index. Sox gene sequences for animals were retrieved from UniProt. The annotated P. tepidariorum genome is available at https://iskrnal.usda.gov/JBrowse-partep and the assembly is deposited at NCBI: BioProject PRJNA167405 (Accession: AOMJ0000000).

**Authors' contributions**

Experiments were performed by CLBP, SR, AS and DJL. All authors contributed to data analysis and writing the manuscript. All authors read and approved the final manuscript.

**Ethics approval and consent to participate**

Not applicable.

**Consent for publication**

Not applicable.

**Competing interests**

The authors declare that they have no competing interests.

**Publisher’s Note**

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Received**: 30 May 2018 **Accepted**: 17 December 2018

**Published online**: 27 December 2018

**References**

1. Shimeld SM, Holland PW. Vertebrate innovations. Proc Natl Acad Sci U S A. 2000;97(9):4449–52.
2. 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):980–96.
3. Kamachi Y, Kondoh H. Sox proteins: regulators of cell fate specification and differentiation. Development. 2013;140(20):4129–44.
4. Sarkar A, Hocheldinger K. The sox family of transcription factors: versatile regulators of stem and progenitor cell fate. Cell Stem Cell. 2013;12(1):15–30.
5. Reiprich S, Wegner M. From CNS stem cells to neurons and glia. Sox for everyone. Cell Tissue Res. 2015;359(1):111–24.
53. Cremazy F, Berta P, Girard F. SoxNeuro, a new Drosophila Sox gene expressed in the developing central nervous system. Mech Dev. 2000;93:215–9.

54. Stollewerk A, Chipman AD. Neurogenesis in myriapods and chelicerates and its importance for understanding arthropod relationships. Integr Comp Biol. 2006;46(2):195–206.

55. 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;24:8834–46.

56. Bergland M, Ramskold D, Zawierut C, Klum S, Sandberg R, Muhr J. Sequentially acting Sox transcription factors in neural lineage development. Genes Dev. 2011;25:2453–64.

57. Cremazy F, Berta P, Girard F. Genome-wide analysis of Sox genes in Drosophila melanogaster. Mech Dev. 2001;109:371–5.

58. Lefebvre V. The SoxD transcription factors – Sox5, Sox6, and SoxE – are key cell fate modulators. Int J Biochem Cell Biol. 2010;42:429–32.

59. Schvager EE, Meng Y, Extavour CG. Vasa and piwi are required for mitotic integrity in early embryogenesis in the spider Parasteatoda tepidariorum. Dev Biol. 2015;402(2):276–90.

60. Loh SY, Russell S. A Drosophila group E Sox gene is dynamically expressed in the embryonic alimentary canal. Mech Dev. 2000;93:4.

61. Vidal VP, Charboissier M-C, deRooij DG, Schedl A. Sox9 induces testis development in XX transgenic mice. Nat Genet. 2001;28:216–7.

62. Bell DM, Leung KK, Wheatley SC, Ng LJ, Zhou S, Ling KW, Sham MH, Koopman P, Tarn PP, Cheah KS. SOX9 directly regulates the type-II collagen gene. Nat Genet. 1997;16(2):174–8.

63. Stolt CC, Wegner M. SoxE function in vertebrate nervous system development. Int J Biochem Cell Biol. 2010;42:437–40.

64. Stollewerk A, Weller M, Tautz D. Neurogenesis in the spider Cupiennius salei. Development. 2001;128(14):2673–88.

65. 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.

66. Stevers 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.

67. Edgar RC. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics. 2004;5:113.

68. 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):W465–9.

69. Huang X, Miller W. A time-efficient linear-space local similarity algorithm. Adv Appl Math. 1991;12:337–57.

70. Altschul SF, Wootton JC, Gertz EM, Agarwala R, Morgulis A, Schäffer AA, Yu YK. Protein database searches using compositionally adjusted substitution matrices. FEBS J. 2005;272:5101–9.

71. 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;130:1735–47.

72. Mittmann B, Wolff C. Embryonic development and staging of the cobweb spider Parasteatoda tepidariorum C. L. Koch, 1841 (syn. Achaearanea tepidariorum; Araneomorphae; Theridiidae). Dev Genes Evol. 2012;222(4):189–216.