Hagfish and lamprey Hox genes reveal conservation of temporal colinearity in vertebrates

Juan Pascual-Anaya[1]*, Iori Sato[1†], Fumiaki Sugahara[1,2†], Shinnosuke Higuchi[1], Jordi Paps[3], Ren Yandong[4, 5], Wataru Takagi[1], Adrián Ruiz-Villalba[6], Kinya G. Ota[7], Wen Wang[4], Shigeru Kuratani[1]

[1] *Evolutionary Morphology Laboratory, RIKEN, Kobe 650-0047, Japan*

[2] *Division of Biology, Hyogo College of Medicine, Nishinomiya 663-8501, Japan*

[3] *School of Biological Sciences, University of Essex, Colchester CO4 3SQ, UK.*

[4] *State Key Laboratory of Genetic Resources and Evolution, Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming 650223, China.*

[5] *University of Chinese Academy of Sciences, Beijing, 100049, China*

[6] *Cell Therapy Program, Foundation of Applied Medical Research (FIMA), University of Navarra, Pamplona, Spain.*

[7] *Laboratory of Aquatic Zoology, Marine Research Station, Institute of Cellular and Organismic Biology, Academia Sinica, Yilan 26242, Taiwan*

[†]Iori Sato and Fumiaki Sugahara contributed equally to this work.

*Correspondence to:

Juan Pascual-Anaya

e-mail: jpascualanaya@gmail.com

Evolutionary Morphology Laboratory, RIKEN, 2-2-3 Minatojima-minami, Chuo-ku, Kobe, Hyogo 650-0047, Japan

Tel.: +81-78-306-3589

Fax: +81-78-306-3370

Hox genes exert fundamental roles for proper regional specification along the main rostro-caudal axis of animal embryos. Hox genes are generally expressed in restricted spatial domains according to their position in the cluster (spatial colinearity), a feature that is conserved across bilaterians. In jawed vertebrates (gnathostomes), the position in the cluster also determines the onset of expression of Hox genes, a feature known as whole-cluster temporal colinearity (WTC), while in invertebrates this phenomenon is displayed as a subcluster-level temporal colinearity (STC). However, little is known about the expression profile of Hox genes in jawless vertebrates (cyclostomes), and therefore the evolutionary origin of WTC, as seen in gnathostomes, remains a mystery. Here we show that Hox genes in cyclostomes are expressed according to WTC during development. We have investigated the Hox repertoire and Hox gene expression profiles in three different species —a hagfish, a lamprey and a shark— encompassing the two major groups of vertebrates and found that these are expressed following a whole-cluster, temporally staggered pattern, indicating that WTC has been conserved during the last 500 million years despite drastically different genome evolution and morphological outputs between jawless and jawed vertebrates.

Hox genes are fundamental developmental genes with crucial roles for the early specification of embryonic structures along the main anterior-posterior axis of bilaterian animals[1]. Hox genes are usually placed in the same genomic regions forming cluster(s). Hox clusters are thought to be the result of several tandem duplication events of an ancestral proto-Hox gene[2], and while most invertebrates generally have a single Hox cluster, vertebrate genomes present multiple clusters[3]. It is widely accepted that the genome of vertebrates has evolved through two rounds (2R) of whole genome duplication (WGD) events (but see ref. 4 for an alternative scenario), generating up to four paralogous loci for each single region of a pre-duplicative

genome[5-8]. Extant vertebrates are divided into two major groups: agnathans, represented by the monophyletic group of cyclostomes (hagfish and lampreys); and gnathostomes, encompassing all jawed-vertebrates in two major groups: cartilaginous fishes (e.g., sharks, rays and chimaeras) and bony vertebrates (e.g., teleosts, coelacanth, amphibians, reptiles, mammals). Tetrapod genomes, including mammals, contain four Hox clusters, named from HoxA to HoxD, as the result of these 2R-WGD (Fig. 1a). Although the 2R-WGD events are generally accepted, the timing of these events with respect to the divergence of cyclostomes and gnathostomes is still a matter of intense debate[4,9-11]. Despite extended research on vertebrate genomes, this has mostly focused on representative species of gnathostomes, while cyclostomes have remained poorly understood. A recent study of the genome of the Artic lamprey, *Lethenteron camtschaticum* (or Japanese lamprey, *Lethenteron. japonicum*), suggested that lampreys had probably undergone a third round of WGD event (3R-WGD)[10]. Whether this event is an independent, lineage-specific event remains a mystery, since the Hox complement of the hagfish is unknown.

The position of Hox genes in the cluster determines their expression patterns. Spatial colinearity refers to the property by which the anterior limit of expression of a given Hox gene is generally more rostral than its upstream (more 5') counterpart. Spatial colinearity is widely conserved among bilaterians studied so far, even in cases where the Hox cluster is completely atomized[12]. Temporal colinearity refers to the phenomenon describing the temporal order of expression of Hox genes according to their position in the cluster, i.e., genes in the 3' part are expressed earlier, and was first described in the HoxD cluster of the mouse[13,14]. Indeed, this so-called whole-cluster temporal colinearity (WTC)[15] phenomenon had been described only in jawed vertebrates. The recent analysis of the scallop genome and the reanalysis of Hox gene expression in a wide range of invertebrates has revealed that Hox genes of these species follow what is called a subcluster-level temporal colinearity (STC),

i.e., that the cluster is divided into small, contiguous groups of Hox genes, each of these displaying temporal colinearity[15]. This situation leads to the uncertainty of what was the ancestral condition before deuterostomes and protostomes split. Moreover, temporal colinearity has not been described in any cyclostome species so far. In *L. camtschaticum*, Hox genes known to be in the same cluster were not expressed following temporal colinearity[16], and both the Hox gene repertoire and expression of the other major group of cyclostomes, the hagfish, is mostly unknown[3,17]. Therefore, the evolutionary origin of WTC as observed in gnathostomes remains obscure.

Here, we provide a comprehensive analysis of different transcriptomics and genomics resources for the Japanese inshore hagfish, *Eptatretus burgeri*. The hagfish Hox repertoire consists of at least 40 Hox genes, including six Hox4 genes that might suggest the presence of at least 6 Hox clusters, suggesting that the 3R-WGD described for the lamprey could be shared in cyclostomes. Finally, we have comprehensively compared the developmental expression levels of Hox genes during development of four different chordate species, including the hagfish and the lamprey, and conclude that temporal colinearity likely originated in the last common ancestor of chordates, and it was certainly well established at least in the last common ancestor of extant vertebrates.

Results and discussion

To gain insights into vertebrate Hox evolution (Fig. 1a), especially with regards to the evolution of temporal colinearity, we decided to comprehensively analyse the Hox repertoire and expression of Hox genes during development of both the lamprey and the hagfish. First, we screened both the developmental transcriptome and the genome of *E. burgeri*. For the developmental transcriptome, RNA-seq data was generated from three different whole

hagfish embryos at Bashford Dean stages 28/30, 35 and 40/45 (refs. 18 and 19; Fig. 1c-e) and from the head region of a hatched juvenile. In total, we found 40 bona fide Hox genes in the developmental transcriptome of the hagfish, including the 5' and 3' untranslated regions for most cases (Fig. 1b).

To determine the genomic organization of hagfish Hox genes we then screened a BAC library built from blood genomic DNA. We found 25 BAC clones spanning only 15 out of the 40 Hox genes (Supplementary Fig. 1, 2). Recently, it has been described that the lamprey genome goes through somatic rearrangements, differentially eliminating stretches of germ line-specific sequences, which might include protein-coding genes[20]. Considering that the hagfish, which is known to go through a chromosome elimination process in somatic tissues during development[21], might be losing Hox genes in somatic tissues, we decided to generate a draft genome using genomic DNA obtained from the testis (germ line) of a single individual. In our preliminary assembly, we found evidence for at least six Hox clusters containing all 40 Hox genes found in the transcriptome and three microRNAs, together with conserved syntenic non-Hox genes (Fig. 1b). The hagfish Hox repertoire and genomic organization are overall very similar to the one described in the *L. camtschaticum* genome[10] (number of genes —43 in the lamprey— and putative clusters —six in the lamprey—), raising the possibility that the 3R-WGD event suggested to have occurred in the lamprey lineage[10] took place before the split of lampreys and hagfish lineages. Surprisingly, we found a hagfish Hox13 gene (*Hox13VI*) enclosed by two conserved syntenic genes: *Lunapark* (*Lnp*) and *Even-skipped* (*Evx*). This suggests that a translocation event took place in the hagfish lineage, likely together with a severe disintegration of a cluster involving large Hox gene losses.

Phylogenetic analysis and best BLAST hits show that the hagfish genome contains representative Hox genes of most of the vertebrate paralogy groups (PG) between PG1 and PG14 (Fig. 2; Supplementary Figs. 3-7). Interestingly, the hagfish genome does not contain

any member of the PG12 (Fig. 1b, Supplementary Figs. 6, 7), a feature shared with the lamprey[3,10,11] (Fig. 1a). Phylogenetic analysis of the posterior Hox genes suggests that a shared cyclostome loss of the PG12 is the most plausible scenario. We were, however, unable to clarify one-to-one orthology relationships between gnathostome HoxA-D paralogs and lamprey and hagfish Hox genes. Therefore, we named the hagfish Hox genes with a different nomenclature from the one used for the lamprey and gnathostomes counterparts, using roman numbers: I-VI (Fig. 1b).

The obscure orthology relationship between jawed and jawless vertebrate genes has been broadly described for both Hox and non-Hox gene families[22]. It is unclear whether the 2R-WGD events that took place during early vertebrate evolution are shared or not among cyclostomes and gnathostomes[6,23,24]. The lack of one-to-one orthology relationships between genes from both groups can be taken as evidence for independent WGD events. However, despite their obscure phylogenetic relationship, Hox clusters of cyclostomes and gnathostomes can still be the result of an ancestral 2R-WGD, if the duplicated regions containing the Hox clusters had not completed rediploidization before the split of cyclostomes and gnathostomes[25]. Consequently, certain number of phylogenetic analyses would support a shared WGD between cyclostomes and gnathostomes as it seems to be the case[9]. These would correspond to those genes that had differentiated into different alleles before the split of the two lineages.

Once confirmed the presence of clusters, we wondered whether the hagfish Hox genes were expressed according to the spatial colinearity rule. Spatial colinearity has been observed in the vast majority of bilaterians studied so far, included the lamprey[16,26], and with only few exceptions[27]. In both the lamprey and gnathostomes, nested expression of anterior Hox genes is coupled to the morphological segmentation of the hindbrain into discrete rhombomeres, and this is controlled by a highly conserved gene regulatory network, established at least in

149 the last common ancestor of vertebrates[28]. The hagfish hindbrain is, as in the lamprey[29],
150 transiently segmented into rhombomeres during stage 45 (ref. 19). We investigated the
151 expression pattern of *E. burgeri* anterior Hox genes in three different developmental stages,
152 from mid-pharyngula (stage 40 and 45) to late-pharyngula (stage 53; Fig. 3), with especial
153 focus on their putative expression pattern in the hindbrain. We found that several hagfish
154 Hox1-5 genes were expressed with staggered anterior boundaries in the hindbrain, an
155 expression pattern reminiscent of that of the lamprey[26,28] and gnathostomes[30] (Fig. 3y, z). We
156 also found Hox2-5 genes expressed colinearly in the pharynx of a juvenile at stage 53
157 (Supplementary Fig. 8). In the hindbrain, the most rostral expression domain detected was
158 that of *Hox2IV*, at the border between rhombomeres 1 and 2 (r1/2), from stage 40 (Fig3d, l,
159 t). *Hox2III* signal is not revealed until stage 45, and is similar to that of *Hox2IV*, with its
160 rostral limit apparently around the lateral edge of the diamond shape of the 4th ventricle,
161 which in gnathostomes marks the r1/2 border[31] (Fig. 3k). The expression of Hox2 genes from
162 r2 rearwards is conserved in all vertebrates (Fig. 3z). In gnathostomes and the lamprey, r4 is
163 characterized by a strong expression of Hox1. We were able to find only a very faint
164 expression of only one of the Hox1 genes in the hagfish, *Hox1V*, not in r4 but probably
165 within r7 with an unclear rostral limit (Fig. 3c). We were not able to find any expression for
166 *Hox1I* and *Hox1II*, which could still be expressed in r4 at different stages. Hagfish *Hox3VI*
167 was expressed up to r6 (Fig. 3f, n), while, strikingly, *Hox3II* was found to be expressed in
168 two domains: r5, and from r7 onwards, i.e., with r6 being *Hox3II*-negative (Fig. 3e, m, u). We
169 also found that *Hox4IV* is expressed, as other vertebrate Hox4 genes, from r7 (Fig. 3h, p, v).
170 *Hox4I* is expressed later in development, at stage 45, with a very similar pattern to that of
171 *Hox4IV*, but slightly posteriorly (Fig. 3o). We also found a very weak signal for *Hox4VI* at
172 stage 45 (Fig. 3q). *Hox5III* is expressed the most posteriorly, apparently from the most
173 anterior part of the spinal cord at stage 40, its rostral limit shifting anteriorly into the

hindbrain by stage 45 and 53 (Fig. 3i, r, w), when transcripts of *Hox5IV* are also detected (Fig. 3s).

The evolution of the expression domains of Hox3 genes in the hindbrain of different vertebrates is particularly interesting. Considering the global expression pattern of Hox3 paralogs in each group, we observe that while in the lamprey (*Hox3α*) and the shark *Scyliorhinus canicula* (*Hoxb3*), Hox3 genes are expressed from r4 (refs. 26, 30, 32), in the hagfish and osteichthyans Hox3 genes are expressed up to r5 (Fig. 3z). There are two possible evolutionary explanations for this difference, both involving parallel evolutionary events: either a caudal shift of Hox3 expression domains from r4 to r5 convergently happened in both the lamprey and osteichthyan lineages, or, on the other hand, a rostral shift from r5 to r4 occurred in the lamprey and chondrichthyans. In a different lamprey species, *Petromyzon marinus*, the *Pm1Hox3* gene, orthologous of *L. camtschaticum Hox3α,* was found to be expressed from r5 like in mammals[28]. This could favour the hypothesis of a convergent expression shift in both the Arctic lamprey and the shark as lineage or species-specific changes.

Following the spatial colinearity rule, most posterior PG Hox genes are expressed in the most caudal regions of the embryo. One of the expression domains of Hox13 paralogue genes are the most posterior parts of the hindgut. Concordantly, hagfish *Hox13II* and *Hox13VI* were found around the cloacal region of a juvenile (stage 60; Supplementary Fig. 10), as in the lamprey and other vertebrates[33]. Vertebrate Hox14 genes have also been reported to be expressed in the most posterior parts of the hindgut of the lamprey and the shark[33]. However, we were not able to detect any signal for *Hox14I* transcripts in the cloaca of the hagfish larva (Supplementary Fig.9).

Overall, vertebrate Hox code is generally conserved in hagfish, particularly in the case of the hindbrain, suggesting that the GRN underlying vertebrate hindbrain

segmentation[28] is well conserved in the hagfish. More important than the similarities, elucidating what specific regulatory inputs account for lineage-specific differences in the hindbrain Hox code, such as the striped expression of hagfish *Hox3II* and the different rostral limits of expression of different Hox3 genes in different vertebrates, will be helpful to determine how the hindbrain GRN diversified during vertebrate evolution, and what are the functional and morphological implications of these differences.

To unravel the evolutionary origin of WTC in vertebrates, we further carried out a comprehensive analysis of the developmental expression profile of Hox genes using embryos from both jawed and jawless vertebrates. Together with the RNA-seq data generated for *E. burgeri*, we sequenced RNA-seq libraries covering early to late developmental stages of the lamprey *L. camtschaticum*[34] and the gnathostome catshark *Scyliorhinus torazame*[35] and quantified the expression profiles of Hox genes. As expected, the expression profiles of *S. torazame* Hox genes were consistent with temporal colinearity across all the clusters found in other jawed vertebrates, showing a clear tendency of anterior Hox genes (Hox1-3) expressed at earlier stages and posterior ones (Hox9-14) at later (Fig. 4; Supplementary Fig. 10). Despite previous reports[16], lamprey Hox genes (for which we found an unreported Hox1 gene, *Hox1ζ*) also followed the rule of temporal colinearity (Fig. 4; Supplementary Fig. 10). Interestingly, the HOX-γ cluster has completely lost the temporal colinearity (Supplementary Fig. 10). HOX-γ is one of the most degenerated clusters in the lamprey with only 4 Hox genes[10] (Fig. 1a), which might be a direct consequence of the lack of temporal colinearity. In the hagfish, although obtaining a pool of embryos from a full developmental series is unfeasible, a similar tendency was also observed: levels of posterior *Hox11.I, Hox11.V* and all Hox13 genes are higher at stage 40-45 (comparable to lamprey stage 25-26) than at 28-30 (lamprey stage 22-23), while generally all anterior and central Hox genes levels are higher at stage 28-30 than at later stages (Fig. 4).

The above observations imply that cyclostome Hox expression profiles, as in gnathostomes, are consistent with the WTC rule, suggesting at least a vertebrate origin. In order to determine whether WTC was present before the origin of vertebrates, we investigated the Hox expression profiles of a chordate outgroup. Wang and colleagues[15] described the tunicate *Ciona intestinalis* Hox gene expressions as according to the STC. However, their statement was based on the reanalysis of data from whole mount *in situ* hybridization[36], which is not a quantitative technique. Cephalochordates are the closest lineage to vertebrates with an intact Hox cluster, and is thus very informative in this regard. Expression profiles of Hox genes in the cephalochordate amphioxus *Branchiostoma belcheri*[37] show that amphioxus *Hox1* to *Hox5* are expressed in an anterior Hox/early-posterior/late manner. However, *Hox6*, *Hox10* and *Hox14* genes violate this pattern, consistent with our previous report[38], and *Hox7-8*, *Hox11-13* and *Hox15* were not detected during the stages assayed, contributing to the dismantling of the colinearity (Fig. 4). In most invertebrate species where STC has been described, Hox1-2 or Hox1-3 was the most anterior subgroup showing temporal colinearity[15]. The fact that in amphioxus Hox1-5 are expressed in temporal order as a single group indicates that this expression pattern is reminiscent from a genuine WTC, which was subsequently broken from Hox6 in the cephalochordate lineage (Fig. 4). In addition, amphioxus Hox6-15 genes might still follow WTC at later stages than the ones assayed here[37]. The putative presence of WTC in both the cephalochordate and vertebrate lineages implies that it was likely present in the last common ancestor of vertebrates.

Taken together, our results depict a scenario in which chordate Hox genes are expressed following WTC, and protostome Hox genes according to STC. This, importantly, can offer a mechanistic answer to explain the radically different bauplans displayed by chordates and protostome invertebrates. Deschamps and Duboule[39] have recently proposed

that temporal colinearity, as seen in mammals (WTC), is displayed only by animals that follow a developmental strategy of anterior to posterior elongation, adding new regions to the main body axis from a posterior growth zone. This temporal activation of Hox genes during the posterior elongation, or Hox clock, translates during development into the spatial colinearity observed along the main anterior-posterior body axis. The fact that the lamprey and the hagfish also develop according to this posterior elongation –a developmental mode thus very well conserved across vertebrates–, together with the presence of WTC and spatial colinearity in the main axial structures of these animals (this study and refs. 16, 26), supports Deschamps and Duboule's hypothesis[39].This implies that this mechanism was present in the last common ancestor of vertebrates, although some lineage-specific differences might have occurred in the mechanism transmitting the Hox clock from the posterior progenitors into the resulting axial structures (for instance, there are differences in the expression of Hox10 genes between lampreys and amniotes in the tailbud and axial mesoderm, see ref. 16). Ultimately, the question of whether the Hox cluster of the last common bilaterian ancestor was expressed according to either whole-cluster or subcluster modes of temporal expression remains open. A more detailed investigation of the temporal expression of Hox genes in non-chordate deuterostome groups (namely, ambulacrarians —e.g., sea urchins, sea stars, acorn worms—) will be thus needed to ultimately resolve this question[40,41].

It has been proposed that gnathostome Hox clusters are relatively compacted, or 'organized', due to a consolidation process that was associated with the emergence of meta-*cis* regulation of the cluster, and probably facilitated by the 2R-WGD events that occurred during vertebrate evolution[42]. Hox clusters of cyclostomes are, on the other hand, more akin to 'disorganized' types of clusters, like the one of amphioxus[42] —because of their extremely large sizes—, suggesting that this consolidation did not start in the last common ancestor of vertebrates, but rather was a progressive gnathostome-specific process[42] (Fig. 4). Further

functional analyses of the regulatory mechanisms of cyclostomes' Hox clusters, with special focus on determining the presence or absence of global regulatory elements outside the clusters, will be needed to clarify whether the consolidation process was indeed a consequence of the acquisition of a global regulatory mode for the cluster, or if, on the other hand, this meta-*cis* regulation was already present in the last common ancestor of vertebrates, before the consolidation process started. Moreover, the timing of the vertebrate 2R-WGD, i.e., whether or not these events are shared between gnathostomes and cyclostomes, is one of the most important questions that remain open about the origin of the vertebrate genome architecture, and solving it will be also helpful to decipher whether the vertebrate genome duplications facilitated the consolidation process.

METHODS

Animal sampling, experiments and aquarium maintenance

E. burgeri embryos (staged according to refs. 18 and 19) used in this study were obtained from adult hagfish individuals captured in the Japan Sea off Shimane prefecture as previously described[43], during August of a given year. Eggs were laid in a cage deposited in the natural environment in the sea in October of the same year. Deposited eggs were then incubated in laboratory aquariums with artificial sea water at 16 °C under controlled conditions, until developing embryos are apparent around February or March of the following year. Hagfish embryos used for RNA-seq were from adults captured in 2010, and were assayed in February 2011(total developing embryos 8 out of >150 eggs). Hagfish embryos used for *in situ* hybridization were from adults captured in 2016 (stage 40), 2013 (stage 45) and 2014 (stage 60), and embryos fixed in March 2017, 2014 and 2015, respectively. Sections of stage 53 were from an embryo previously reported[44]. Lamprey (*L. camtschaticum*) and cloudy catshark (*S. torazame*) embryos were obtained as previously described in refs. 45 and 44,

299 respectively. Lamprey and catshark embryos were staged according to refs. 34 and 35,
300 respectively. The sampling and experiments were conducted according to the institutional
301 and national guidelines for animal ethics, approved by the RIKEN Animal Experiments
302 Committee (approval ID: H14-25-24).
303
304 **RNA-seq data and transcriptome assemblies**
305 Total RNA samples from three whole embryos of *E. burgeri* (Fig. 1c-e) and the head region
306 of a hatched juvenile were used to prepare RNA-seq libraries and sequenced individually on
307 different HiSeq and MiSeq platforms (one embryo at stage 28/30: Illumina TruSeq RNA
308 Sample Prep Kit, non-strand-specific library, sequenced with a HiSeq1000 platform; one
309 embryo at stage 35 and one at 40/45: one strand-specific library each using TruSeq RNA
310 Sample Prep Kit modified with the dUTP method[46] and sequenced in a HiSeq2000, and a
311 further non-strand-specific library with Illumina TruSeq RNA Library Prep Kit and
312 sequenced in a MiSeq platform for the former; one juvenile's head: TruSeq RNA Sample
313 Prep Kit, non-strand-specific library, HiSeq1500). Total RNA samples from separate pools of
314 embryos of *L. camtschaticum* at stages 15/16, 18, 20, 22, 24, 26 and 28 (20-30 embryos per
315 stage), and of *S. torazame* at stages 15/16 (6 embryos), 18 (9), 20 (10), 22 (9), 25 (5), 27 (5)
316 and 28 (2) were used to prepare strand-specific libraries (Illumina TruSeq Stranded RNA Lib
317 Prep Kit). Lamprey and shark libraries were sequenced on a HiSeq1500 platform. Reads
318 coming from mitochondrial DNA were filtered out using mirabait (bundled with MIRA).
319 Then, reads were preprocessed with MIRA[47] v.4.9.5_2, using the option '`parameters =`
320 `-GE:ppo=yes`' in the manifest file. In the case of the hagfish, the resulting reads were then
321 assembled with Trinity v2.1.1[48] following 3 different strategies: (1) assembly of all reads
322 together; (2) idem, but including a digital normalization step within Trinity (`--`
323 `normalize_reads`), and (3), assembly of RNA-seq data from each embryo separately and

further integrated with CD-HIT-EST[49] v4.6.4 with parameters '`-c 0.98`'. A fourth assembly was done with SOAPdenovoTrans v1.03[50] using all reads simultaneously and multiple k-mers (19, 21, 23, 25, 27, 29, 31: with '`SOAPdenovo-Trans-31mer`' command; and 41, 51, 61, 71, 81, 91: with '`SOAPdenovo-Trans-127mer`'), with a final integration with CD-HIT-EST. In the case of the lamprey and shark, reads were assembled according to 3 different pipelines: (1) assembly with Trinity v2.1.1of reads coming from each pool of embryos independently, taking into account the strand-specific information (`--SS_lib_type RF`), and integrated with CD-HIT-EST v4.6.4; (2) idem, but not taking into account the strand-specific information; and (3) assembly of all reads together. In the case of the lamprey, a fourth assembly strategy was carried out by means of integrating a genome-guided assembly (option `--genome_guided_bam` of Trinity, mapping the RNA-seq reads to *L. camtschaticum* 1.0 genome[10] with the splice-aware mapper HISAT2[51]) and the above *de novo* assembly #3, using the PASA v2.0.2[52] pipeline (http://pasapipeline.github.io/#A_ComprehensiveTranscriptome). Finally, completeness assessments of all versions were done using CEGMA v2.5 and BUSCO v1.1b1 programs, as previously described[53] (Supplementary Tables 1-3). The most complete versions of *E. burgeri* and *L. camtschaticum* were selected for further analysis. In the case of *S. torazame*, although strategy #3 gave as a result a more complete transcriptome in general, it contained more fragmented Hox genes that version #1, and therefore we selected the latter. All *E. burgeri*, *L. camtschaticum* and *S. torazame* RNA-seq data have been deposited in NCBI GenBank under the BioProject number PRJNA371391. Amphioxus *B. belcheri* transcriptome was assembled using previously published RNA-seq data, from the NCBI's SRA database, under BioProject numbers PRJNA310680[37] and PRJNA214454[54]. *B. belcheri* RNA-seq reads from the former BioProject were subjected to adaptor trimming with cutadapt v1.10[55]. All *B. belcheri* RNA-seq data was then assembled following the same above-mentioned pipeline for

the lamprey transcriptome (strategy #3), using the previously published *B. belcheri* genome[54] for the PASA pipeline.

All Trinity commands were executed using the `--group_pairs_distance 999` parameter value[56].

BAC library, clone screening and PacBio sequencing and assembly

Blood was drawn from the caudal sub-cutaneous sinus of one adult specimen of *E. burgeri* using a heparin-rinsed disposable syringe. The whole blood sample was immediately frozen in liquid nitrogen, and used for DNA extraction. A BAC library consisting of 129,024 clones, with an average insert size of 100 Kbp (~4.4X of the *E. burgeri* genome size), was constructed using the pCCBAC1 vector[57] [CopyControl™ BAC Cloning Kit (*Hind*III) (EPICENTRE)] and pooled into 96-well and 384-well plates according to the Matrix Pool and Superpool Strategy[58] by Amplicon Express (Pullman, WA, USA). The BAC library was screened for Hox-containing clones by means of PCR with specifi primers[58]. Positive BAC clones were extracted with the QIAGEN Large-Construct kit, and sequenced in different pools using SMRT technology on a PacBio *RS* platform using XL-C2 chemistry, or on a *RS*II platform using P4-C2 chemistry. pCCBAC1 vector sequence were masked using a script from ref. 59 with minor modifications. BAC clones were assembled using masked subreads with MIRA[47] v4.9.5_2. The sequence of the BAC clones used in this study have been deposited in GenBank (accession numbers MF182102-MF182109).

Genome sequencing and assembly

Germ line DNA for whole genome shotgun (WGS) sequencing, derived from the testis of a single male hagfish, *E. burgeri*, was sequenced on an Illumina HiSeq 2500™ platform. In total, we sequenced five pair-end (174-bp, 234-bp, 242-bp, 279-bp and 612-bp) and five

mate-pair (5-Kbp, 5~7-Kbp, 7~10-Kbp, 10~15-Kbp and 15~20-Kbp) libraries, generating >300X coverage of the estimated 2.906 Gb-long genome of the hagfish. All short-read data were corrected by SOAPec v2.01[60] using >40X data. Assembly of the hagfish genome was performed with ABySS v1.9.0[61] with a k-mer size of 79, followed by a scaffolding step with SOAPdenovo v2.04-r241[60] software (parameter '`-K 41 -d 1 -M 2 -F`'). Gaps were finally filled with GapCloser v1.12-r6[60]. The resulting assembly (size, ~2.59 Gb; N50, ~439 Kbp) was used for the screening of Hox clusters. Hox-containing scaffolds were then aligned against the BAC clones using MUMmer v3.23[62] and visualized using mummerplot, bundled within the same software. Sequences of Hox-containing scaffolds, as well as those of *E. burgeri Lnp* and *Evx* (whose sequences are not complete in the genome) have been deposited in GenBank under accession numbers MF398213-MF398235. A publication with more detailed and in-depth analysis of the *E. burgeri* genome is now in preparation.

Identification of Hox genes

UniProt Knowledgebase (UniProtKB) database (http://www.uniprot.org/) was searched for entries containing the term "Hox" and restricted to Eumetazoans (name:hox, taxonomy:6072; UniProt release 2015_11). Resulting entries were downloaded and used as queries against the transcriptome assembly and genome of the hagfish by means of TBLASTN (NCBI BLAST v2.2.31+[63]). The best BLAST hits were then used as queries against the whole UniProtKB database using BLASTX. Those transcripts whose reciprocal best hit was a Hox gene were kept and manually inspected for false positives. Lamprey Hox genes were downloaded from NCBI GenBank[10] and blasted against our lamprey transcriptome assembly to identify Hox transcripts. *Hox4η*, *Hox7ε*, *Hox9ζ*, *Hox11δ*, *Hox13α*, *Hox13ε*, *Hox13ζ* and *Hox14ε* were not found in our transcritptome assembly. We found an extra, unreported Hox1 paralogous gene, which we named *Hox1ζ* (following the nomenclature from ref. 10). *Scyliorhinus canicula*

Hox genes[30] sequences were downloaded from NCBI GenBank and used as queries to identify orthologous sequences in our *S. torazame* transcriptome by means of TBLASTN. The *L. camtschaticum Hox1ζ* and *S. torazame* Hox gene sequences were deposited in GenBank (accession numbers MF398236-MF398269).

cDNA cloning and section *in situ* hybridization

Selected Hox genes were cloned from cDNA prepared for a previous study[44] using specific primers. *In situ* hybridization on paraffin wax-embedded sections of stage 45 and 60 hagfish embryos was performed according to refs. 44, 45. Haematoxylin and eosin (H&E) staining on paraffin sections of stage 60 embryo was carried out by standard protocol. H&E stained sections were further stained with Alcian Blue[64].

3D reconstruction of the hagfish embryos

The 3D reconstruction images of hagfish embryos were reconstructed based on images taken of 1 in every 10 histological sagittal sections at 6 μm, stained with standard haematoxylin and eosin staining protocols for the stage 40 embryo, and 1 in 2 unstained sections at 8 μm for the stage 45 embryo. Reconstructed images were acquired using Avizo software (Visualization Sciences Group). Stage 53 reconstruction is from an embryo used previously[19].

Molecular phylogenetic analyses

The Hox genes nucleotide sequences for different chordates and outgroups were mined from NCBI GenBank, Ensembl (www.ensembl.org), EchinoBase (http://www.echinobase.org/Echinobase/), or, in some instances, manually annotated (see Supplementary Table 4 for accession numbers of genes used in the analyses). Hox genes sequences of amphioxus species *B. lanceolatum* and *B. floridae* are from refs. 65-67. Five

datasets based on different gene content were assembled: 1) Anterior genes (Hox1-3), 2) Hox 4 genes, 3) Central genes (Hox4-8), 4) Posterior genes (Hox9-14), and 5) all Hox genes together. The datasets were aligned using MAFFT v7.123b[68] using the "*auto*" option, regions of ambiguous alignment trimmed with Gblocks v0.91b[69] using the less stringent options. Alignments were visually inspected with BioEdit v7.2.6[70]. Phylogenetic trees were inferred with RAxML v8.2.10[71] using a random starting tree, the evolutionary model LG + Gamma + Invariants with empirical base frequencies, and 1000 rapid bootstrap replicates. Trees were edited using FigTree v1.4.3 (http://tree.bio.ed.ac.uk/software/figtree/).

Expression profiling of Hox genes.

RNA-seq reads from individual embryos were used to quantify transcripts of the selected transcriptomes of *E. burgeri* and *S. torazame* using Perl scripts 'align_and_estimate_abundance.pl' and 'align_and_estimate_abundance.pl', bundled with Trinity v.2.1.1, and using RSEM v1.2.28[72] as quantification method (https://github.com/trinityrnaseq/trinityrnaseq/wiki/Trinity-Transcript-Quantification). Hox transcripts of *S. torazame* were directly quantified using RSEM with data from each embryonic stage. TPM values from either genes (for Hox genes represented with a single transcript in the assemblies) or isoforms (for Hox genes represented with several transcripts) were then selected and a heat map analyses of the log(TPM+0.1) were conducted in R using heatmap.2 (gplots package[73]) scaling by gene (row Z-score), and implemented in RStudio v1.0.136[74] [with R v3.3.0 (2016-05-03)[75]]. *B. belcheri* Hox transcripts were quantified using previously published DGE-seq data[37] with DGE-EM v1.0.0[76] software, and FPKM values were analysed as above.

Data availability.

RNA-seq data generated in this study have been deposited in SRA, under the BioProject number PRJNA371391. Sequences generated and analysed in this study have been deposited in NCBI GenBank under accession numbers MF182102-MF182109 and MF398213-MF398269.

References

1 Pearson, J. C., Lemons, D. & McGinnis, W. Modulating Hox gene functions during animal body patterning. *Nat. Rev. Genet.* **6**, 893-904 (2005).

2 Garcia-Fernàndez, J. The genesis and evolution of homeobox gene clusters. *Nat. Rev. Genet.* **6**, 881-892 (2005).

3 Pascual-Anaya, J., D'Aniello, S., Kuratani, S. & Garcia-Fernandez, J. Evolution of Hox gene clusters in deuterostomes. *BMC Dev. Biol.* **13** (2013).

4 Smith, J. J. & Keinath, M. C. The sea lamprey meiotic map improves resolution of ancient vertebrate genome duplications. *Genome Res* **25**, 1081-1090 (2015).

5 Dehal, P. & Boore, J. L. Two rounds of whole genome duplication in the ancestral vertebrate. *PLoS Biol.* **3**, e314 (2005).

6 Putnam, N. H. *et al.* The amphioxus genome and the evolution of the chordate karyotype. *Nature* **453**, 1064-1071 (2008).

7 Van de Peer, Y., Maere, S. & Meyer, A. The evolutionary significance of ancient genome duplications. *Nat. Rev. Genet.* **10**, 725-732 (2009).

8 Van de Peer, Y., Maere, S. & Meyer, A. 2R or not 2R is not the question anymore. *Nat. Rev. Genet.* **11**, 166 (2010).

9 Kuraku, S., Meyer, A. & Kuratani, S. Timing of genome duplications relative to the origin of the vertebrates: did cyclostomes diverge before or after? *Mol. Biol. Evol.* **26**, 47-59 (2009).

10 Mehta, T. K. *et al.* Evidence for at least six Hox clusters in the Japanese lamprey (Lethenteron japonicum). *Proc Natl Acad Sci U S A* **110**, 16044-16049 (2013).

11 Smith, J. J. *et al.* Sequencing of the sea lamprey (*Petromyzon marinus*) genome provides insights into vertebrate evolution. *Nat. Genet.* **45**, 415-421 (2013).

12 Seo, H. C. *et al. Hox* cluster disintegration with persistent anteroposterior order of expression in *Oikopleura dioica*. *Nature* **431**, 67-71 (2004).

13 Dollé, P., Izpisúa-Belmonte, J. C., Falkenstein, H., Renucci, A. & Duboule, D. Coordinate expression of the murine *Hox-5* complex homoeobox-containing genes during limb pattern formation. *Nature* **342**, 767-772 (1989).

14 Izpisúa-Belmonte, J. C., Falkenstein, H., Dollé, P., Renucci, A. & Duboule, D. Murine genes related to the *Drosophila AbdB* homeotic genes are sequentially expressed during development of the posterior part of the body. *EMBO J.* **10**, 2279-2289 (1991).

15 Wang, S. *et al.* Scallop genome provides insights into evolution of bilaterian karyotype and development. *Nat. Ecol. Evol.* **1**, 0120 (2017).

16 Takio, Y. *et al. Hox* gene expression patterns in *Lethenteron japonicum* embryos—Insights into the evolution of the vertebrate Hox code. *Dev. Biol.* **308**, 606-620 (2007).

17 Stadler, P. F. *et al.* Evidence for independent *Hox* gene duplications in the hagfish lineage: a PCR-based gene inventory of *Eptatretus stoutii*. *Mol. Phylogenet. Evol.* **32**, 686-694 (2004).

18 Dean, B. On the Embryology of *Bdellostoma stouti*: A General Account of Myxinoid Development from the Egg and Segmentation to Hatching. 220-276 (G. Fischer, 1899).

19 Oisi, Y., Ota, K. G., Kuraku, S., Fujimoto, S. & Kuratani, S. Craniofacial development of hagfishes and the evolution of vertebrates. *Nature* **493**, 175-180 (2013).

20 Smith, J. J., Baker, C., Eichler, E. E. & Amemiya, C. T. Genetic consequences of programmed genome rearrangement. *Curr. Biol.* **22**, 1524-1529 (2012).

21 Kohno, S.-i., Kubota, S. & Nakai, Y. in *The Biology of Hagfishes.* 81-100 (Springer Netherlands, 1998).

22 Kuraku, S. Impact of asymmetric gene repertoire between cyclostomes and gnathostomes. *Semin. Cell Dev. Biol.* **24**, 119-127 (2013).

23 Ohno, S. *Evolution by Gene Duplication*. (Springer-Verlag, 1970).

24 Kasahara, M. The 2R hypothesis: an update. *Curr. Opin. Immunol.* **19**, 547-552 (2007).

25 Martin, K. J. & Holland, P. W. Enigmatic orthology relationships between *Hox* clusters of the African butterfly fish and other teleosts following ancient whole-genome duplication. *Mol. Biol. Evol.* **31**, 2592-2611 (2014).

26 Takio, Y. *et al.* Evolutionary biology: lamprey *Hox* genes and the evolution of jaws. *Nature* **429**, 1 p following 262 (2004).

27 Schiemann, S. M. *et al.* Clustered brachiopod Hox genes are not expressed collinearly and are associated with lophotrochozoan novelties. *Proc. Natl. Acad. Sci. U. S. A.* **114**, E1913-E1922 (2017).

28 Parker, H. J., Bronner, M. E. & Krumlauf, R. A *Hox* regulatory network of hindbrain segmentation is conserved to the base of vertebrates. *Nature* **514**, 490-493 (2014).

29 Kuratani, S., Horigome, N., Ueki, T., Aizawa, S. & Hirano, S. Stereotyped axonal bundle formation and neuromeric patterns in embryos of a cyclostome, *Lampetra japonica*. *J. Comp. Neurol.* **391**, 99-114 (1998).

30 Oulion, S. *et al.* Evolution of repeated structures along the body axis of jawed vertebrates, insights from the *Scyliorhinus canicula* Hox code. *Evol. Dev.* **13**, 247-259 (2011).

31 Wingate, R. J. The rhombic lip and early cerebellar development. *Curr. Opin. Neurobiol.* **11**, 82-88 (2001).

32 Murakami, Y. *et al.* Segmental development of reticulospinal and branchiomotor neurons in lamprey: insights into the evolution of the vertebrate hindbrain. *Development* **131**, 983-995 (2004).

33 Kuraku, S. *et al.* Noncanonical role of Hox14 revealed by its expression patterns in lamprey and shark. *Proc. Natl. Acad. Sci. U. S. A.*, 0710947105 (2008).

34 Tahara, Y. Normal stages of development in the lamprey, *Lampetra reissneri* (Dybowski). *Zool. Sci.* **5**, 109-118 (1988).

35 Ballard, W. W., Mellinger, J. & Lechenault, H. A series of normal stages for development of *Scyliorhinus canicula*, the lesser spotted dogfish (*Chondrichthyes: Scyliorhinidae*). *J. Exp. Zool.* **267**, 318-336 (1993).

36 Ikuta, T., Yoshida, N., Satoh, N. & Saiga, H. *Ciona intestinalis* Hox gene cluster: Its dispersed structure and residual colinear expression in development. *Proc. Natl. Acad. Sci. U. S. A.* **101**, 15118-15123 (2004).

37 Yang, K. Y. *et al.* Transcriptome analysis of different developmental stages of amphioxus reveals dynamic changes of distinct classes of genes during development. *Sci. Rep.* **6**, 23195 (2016).

38 Pascual-Anaya, J. *et al.* Broken colinearity of the amphioxus Hox cluster. *EvoDevo* **3** (2012).

39 Deschamps, J. & Duboule, D. Embryonic timing, axial stem cells, chromatin dynamics, and the Hox clock. *Genes Dev.* **31**, 1406-1416 (2017).

40 Aronowicz, J. & Lowe, C. J. Hox gene expression in the hemichordate *Saccoglossus kowalevskii* and the evolution of deuterostome nervous systems. *Integr. Comp. Biol.* **46**, 890-901 (2006).

41 Gonzalez, P., Uhlinger, K. R. & Lowe, C. J. The Adult Body Plan of Indirect Developing Hemichordates Develops by Adding a Hox-Patterned Trunk to an Anterior Larval Territory. *Curr. Biol.* **27**, 87-95 (2017).

42 Duboule, D. The rise and fall of Hox gene clusters. *Development* **134**, 2549-2560 (2007).

43 Oisi, Y., Kakitani, O., Kuratani, S. & Ota, K. G. in *In Situ Hybridization Methods* (ed Giselbert Hauptmann) 249-262 (Springer New York, 2015).

44 Sugahara, F. *et al.* Evidence from cyclostomes for complex regionalization of the ancestral vertebrate brain. *Nature* **531**, 97-100 (2016).

45 Sugahara, F., Murakami, Y. & Kuratani, S. in *In Situ Hybridization Methods* (ed Giselbert Hauptmann) 263-278 (Springer New York, 2015).

46 Sultan, M. *et al.* A simple strand-specific RNA-Seq library preparation protocol combining the Illumina TruSeq RNA and the dUTP methods. *Biochem. Biophys. Res. Commun.* **422**, 643-646 (2012).

47 Chevreux, B., Wetter, T. & Suhai, S. in *Computer Science and Biology: Proceedings of the German Conference on Bioinformatics (GCB)* 45-56 (1999).

48 Grabherr, M. G. *et al.* Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* **29**, 644-652 (2011).

49 Li, W. & Godzik, A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**, 1658-1659 (2006).

50 Xie, Y. *et al.* SOAPdenovo-Trans: de novo transcriptome assembly with short RNA-Seq reads. *Bioinformatics* **30**, 1660-1666 (2014).

51 Kim, D., Langmead, B. & Salzberg, S. L. HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods* **12**, 357-360 (2015).

52 Haas, B. J. *et al.* Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res.* **31**, 5654-5666 (2003).

53 Hara, Y. *et al.* Optimizing and benchmarking de novo transcriptome sequencing: from library preparation to assembly evaluation. *BMC Genomics* **16**, 977 (2015).

54 Huang, S. *et al.* Decelerated genome evolution in modern vertebrates revealed by analysis of multiple lancelet genomes. *Nat. Commun.* **5**, 5896 (2014).

582 55 Martin, M. Cutadapt removes adapter sequences from high-throughput
583 sequencing reads. *EMBnet.journal* **17**, 10-12 (2011).

584 56 Macmanes, M. D. On the optimal trimming of high-throughput mRNA sequence
585 data. *Front. Genet.* **5**, 13 (2014).

586 57 Wild, J., Hradecna, Z. & Szybalski, W. Conditionally amplifiable BACs: switching
587 from single-copy to high-copy vectors and genomic clones. *Genome Res.* **12**,
588 1434-1444 (2002).

589 58 Bouzidi, M. F. *et al.* A sunflower BAC library suitable for PCR screening and
590 physical mapping of targeted genomic regions. *Theor. Appl. Genet.* **113**, 81-89
591 (2006).

592 59 Huddleston, J. *et al.* Reconstructing complex regions of genomes using long-read
593 sequencing technology. *Genome Res.* **24**, 688-696 (2014).

594 60 Luo, R. *et al.* SOAPdenovo2: an empirically improved memory-efficient short-
595 read de novo assembler. *GigaScience* **1**, 18 (2012).

596 61 Simpson, J. T. *et al.* ABySS: a parallel assembler for short read sequence data.
597 *Genome Res.* **19**, 1117-1123 (2009).

598 62 Kurtz, S. *et al.* Versatile and open software for comparing large genomes. *Genome*
599 *Biol.* **5**, R12 (2004).

600 63 Camacho, C. *et al.* BLAST+: architecture and applications. *BMC Bioinformatics* **10**,
601 421 (2009).

602 64 Nowicki, J. L., Takimoto, R. & Burke, A. C. The lateral somitic frontier: dorso-
603 ventral aspects of anterio-posterior regionalization in avian embryos. *Mech. Dev.*
604 **120**, 227-240 (2003).

65 Amemiya, C. T. *et al.* The amphioxus *Hox* cluster: characterization, comparative genomics, and evolution. *J. Exp. Zool. B Mol. Dev. Evol.* **310**, 465-477 (2008).

66 Holland, L. Z. *et al.* The amphioxus genome illuminates vertebrate origins and cephalochordate biology. *Genome Res.* **18**, 1100-1111 (2008).

67 Pascual-Anaya, J., D'Aniello, S. & Garcia-Fernàndez, J. Unexpectedly large number of conserved noncoding regions within the ancestral chordate *Hox* cluster. *Dev. Genes Evol.* **218**, 591-597 (2008).

68 Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772-780 (2013).

69 Castresana, J. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol. Biol. Evol.* **17**, 540-552 (2000).

70 Hall, T. A. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucleic Acids Symp. Ser.* **41**, 95-98 (1999).

71 Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312-1313 (2014).

72 Li, B. & Dewey, C. N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* **12**, 323 (2011).

73 gplots: Various R programming tools for plotting data. R package version 3.0.1 (2016).

74 RStudio: Integrated Development for R. (R. RStudio, Inc., Boston, MA, 2016).

75 R: A language and environment for statistical computing (Vienna, Austria, 2016).

76 Nicolae, M. & Măndoiu, I. in *Bioinformatics Research and Applications: 7th International Symposium, ISBRA 2011, Changsha, China, May 27-29, 2011. Proceedings* (eds Jianer Chen, Jianxin Wang, & Alexander Zelikovsky) 392-403 (Springer Berlin Heidelberg, 2011).

Acknowledgments

We thank Y. Oisi and S. Fujimoto for providing preliminary hagfish Hox sequences; J.M. Martín-Durán, I. Maeso, M. Irimia and C. Böhmer for fruitful discussions; O. Kakitani for hagfish sampling; K. Shirato for shark sampling; S. Shibuya and K. Yamamoto for maintenance of aquarium tanks; S. Kuraku, K. Itomi, C. Tanegashima, K. Tatsumi and O. Nishimura from the Phyloinformatics Unit, RIKEN CLST, for RNA-seq data production; J. Huddleston and E. Eichler for providing the code to mask BAC vector sequences from PacBio reads; I. Mandoiu for his help using DGE-EM software; and B. Chevreaux for his help with the MIRA assembler. This work was supported by the Chinese Academy of Sciences program XDB13000000 to W.W., and by a Grant-in-Aid for Scientific Research (A) 15H02416 (Japan Society for the Promotion of Science), a Grant-in-Aid for Scientific Research on Innovative Areas (Research in a Proposed Research Area) 17H06384 (Ministry of Education, Culture, Sports, Science and Technology of Japan), and a Naito Grant for the Promotion of Focused Research (The Naito Foundation) to S.K.

Author contributions

J.P.-A. conceived the project, designed the experiments, and wrote the paper. J.P.-A, F.S., S.H. and W.T. obtained the hagfish embryos. J.P.-A, I.S., F.S., S.H., W.T. and A.R.-V performed experiments. K.O. built the BAC library. R.Y. and W.W. sequenced and

assembled the *E. burgeri* draft genome. J.P. performed the phylogenetic analyses. All authors analysed and discussed the data. All authors approved the final version of the manuscript.

Competing interests

The authors declare no competing financial interests.

Figure legends

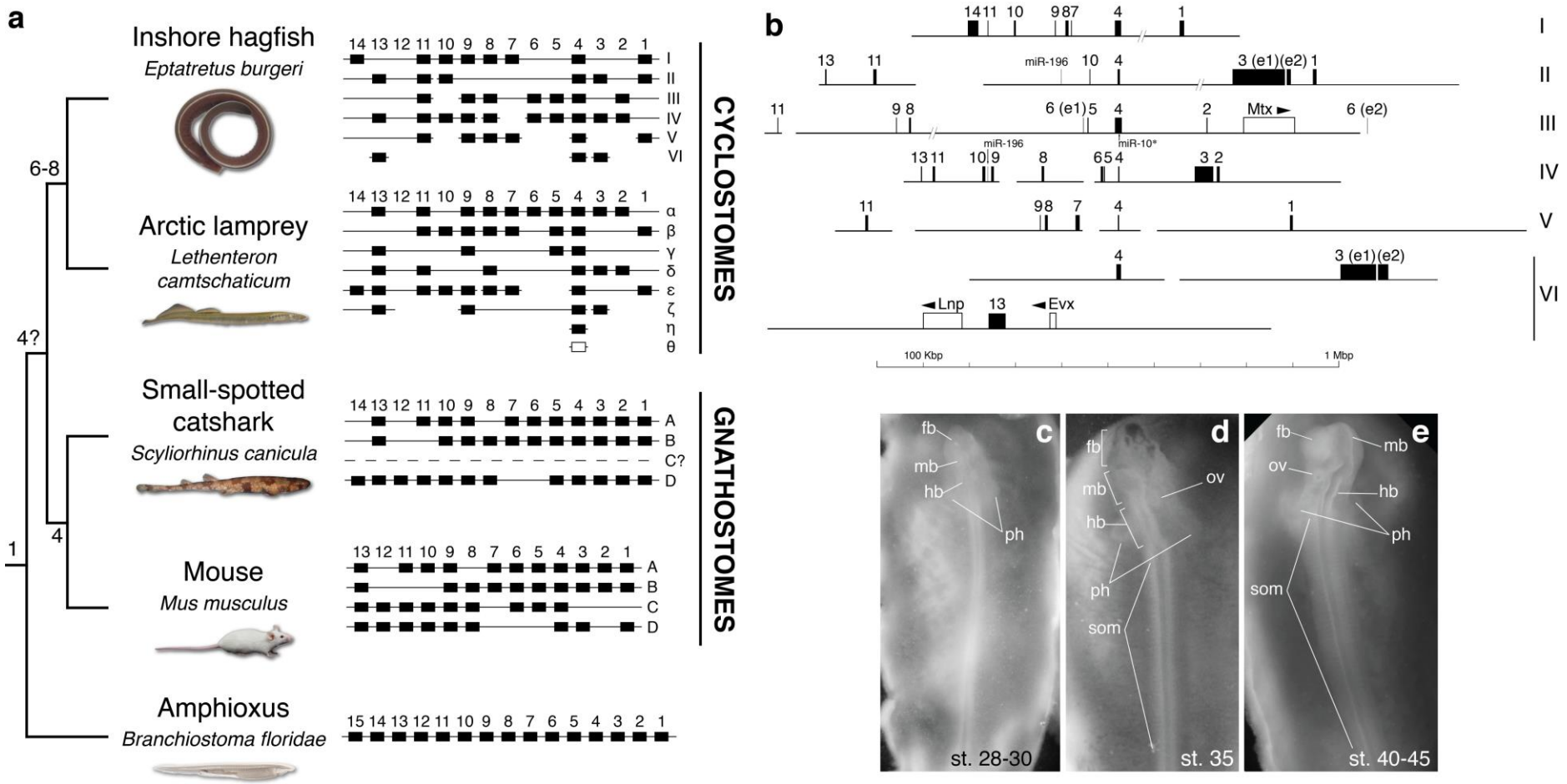
Figure 1. Hox cluster evolution in chordates. a, Phylogenetic tree of chordates, showing the two major groups of vertebrates –cyclostomes (hagfish and lamprey) and gnathostomes (jawed vertebrates, e.g., mouse and shark) – together with cephalochordates (amphioxus), displaying their known Hox repertoires. Numbers on the nodes indicate the putative number of Hox clusters in each last common ancestor. **b**, *E. burgeri* Hox genes and clusters found in this study, drawn to scale. All Hox genes are transcribed in the same orientation, from left to right. Orientation of transcription of non-Hox syntenic genes are indicated by arrowheads. Solid horizontal lines correspond to single scaffolds. Double diagonal lines separate two contiguous scaffolds, based on BAC sequences connecting them (Supplementary Fig. 2). *Hox3II* and *Hox3VI* genes have corresponding exons 1 and 2 in two different scaffolds, which have been put together based on both BAC and transcriptomics evidences. e1, exon 1; e2, exon 2. Asterisk over *miR-10III* indicates that this microRNA is within a 5'UTR intron of *Hox4III*; hash symbol, *Hox6III* exon2 is not assembled in place, but in a separate small contig containing its sequence. **c-e**, *E. burgeri* embryos used for the transcriptomics analysis, at Dean stages 28-30 (**c**), 35 (**d**) and 40-45 (**e**). fb, forebrain; hb, hindbrain; mb, midbrain; ov, otic vesicle; ph, pharynx; som, somites. Scale bars, 1 mm.

Figure 2. Molecular phylogenetic tree of vertebrate Hox genes. 1000-replicate Maximum Likelihood tree of representative Hox genes of all paralogy groups in vertebrates. The branches have been color-coded by paralogy group (Hox1-14). Red and blue branches denote *E. burgeri* and *L. camtschaticum* Hox genes, respectively. Black branches correspond to invertebrate Hox counterparts (amphioxus – *Branchiostoma floridae* and *Branchiostoma lanceolatum* – and sea urchin – *Strongylocentrotus purpuratus* –). Note that no hagfish or lamprey sequence have been found within the Hox12 group (denoted with square brackets). The same tree, with bootstrap values and branch tip names can be found in Supplementary Fig. 7.

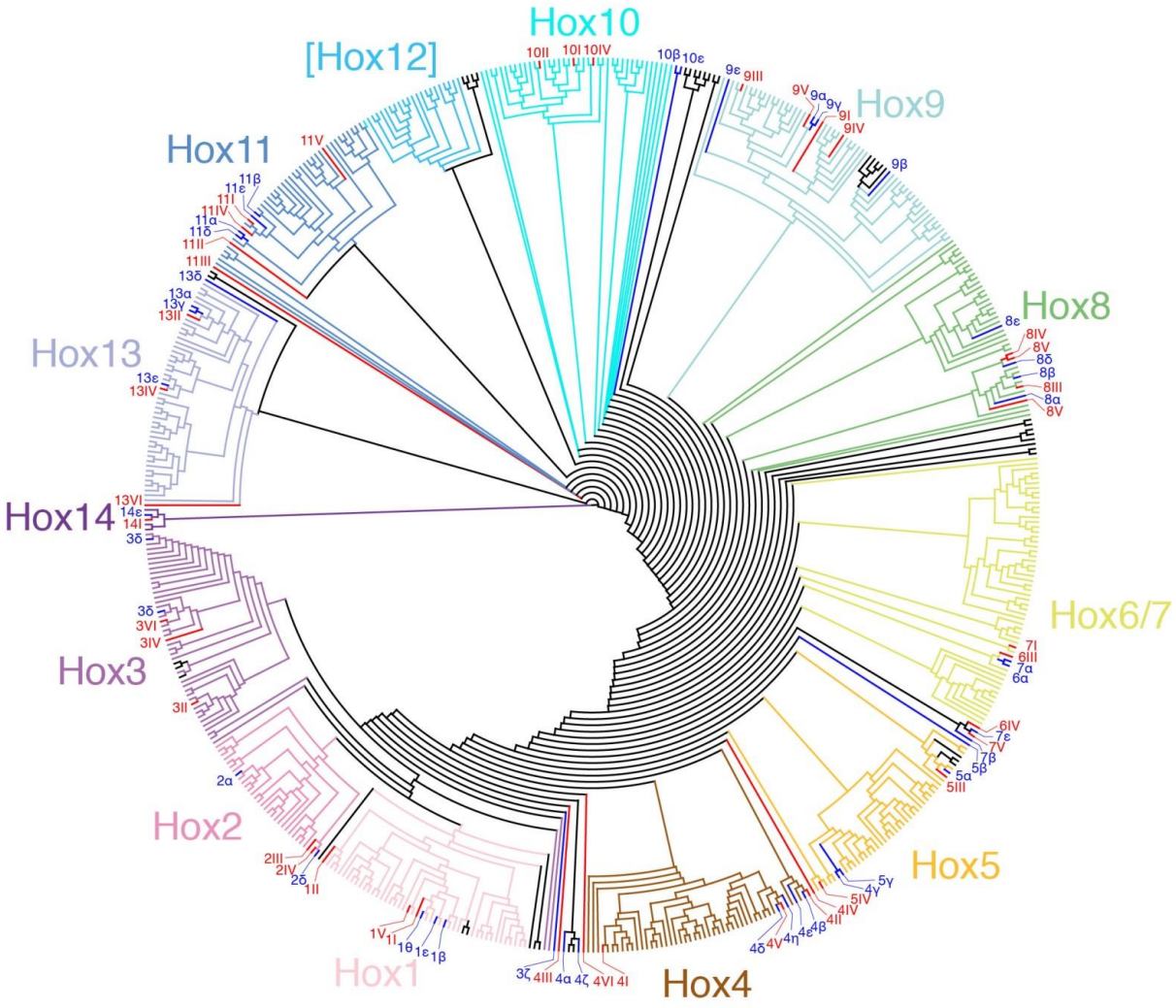
Figure 3. **Spatial colinearity of hagfish *Hox* genes in the hindbrain of *E. burgeri* embryos. a, b**, Embryos at stage Bashford Dean 40 (**a**) and 45 (**b**) used for in situ hybridizations on sections. The inset square brackets mark the head regions, used for sagittal sectioning. **a'**, **a''**, **b'**, **b''**, 3D Avizo reconstructions of the heads of the embryos shown in a and b, respectively, showing the main internal anatomy of the brain and main head structures. The central nervous systems are in purple; ectoderm is in light blue; endoderm is in yellow; otic vesicle in green; and notochord is in light red. These embryos are the source of the sections shown in **c**-**j** (stage 40) and **k**-**s** (stage 45). **c**-**w**, Spatial colinearity displayed by expression patterns of *E. burgeri Hox1IV* (**c**), *Hox2III* (**k**), *Hox2IV* (**d**, **l**, **t**), *Hox3II* (**e**, **m**, **u**), *Hox3VI* (**f**, **n**), *Hox4I* (**g**, **o**), *Hox4IV* (**h**, **p**, **v**), *Hox4VI* (**q**), *Hox5III* (**i**, **r**, **w**) and *Hox5IV* (**j**, **s**), revealed by *in situ* hybridization on sections of an embryos shown in **a**, **b** and **x**. **x**, **x'**, 3D Avizo reconstructions of the head of an embryo at stage Bashford Dean 53. The central nervous systems are in purple; ectoderm is in light blue; endoderm is in yellow; otic vesicle in green; and notochord is in light red. **y**, Expression patterns shown in (**c**-**j**) in the hindbrain are aligned according to rhombomere (r) segmentation, showing nested expression patterns of

Hox1-5 paralogs in the hagfish hindbrain. **z**, Schematic diagrams summarizing the expression patterns of Hox genes in the hindbrain of the lamprey, the hagfish and mouse, with nested anterior limits coinciding with rhombomere borders, and showing the overall conservation among the groups. f, forebrain; h, hindbrain; ht, hypothalamus; m, midbrain; mo, mouth; n, notochord; no, nasal opening; nhp, naso-hypophyseal plate; ot, otic vesicle; ph, pharynx; pp, pharyngeal pouches; y, yolk; 3v, 4v, third and fourth ventricles, respectively. Scale bars, 0.5 mm. Asterisk indicates a different expression has been found in a separate species, *P. marinus*, in which *Pm1Hox3* rostral limit is on the r4/r5 border[28].

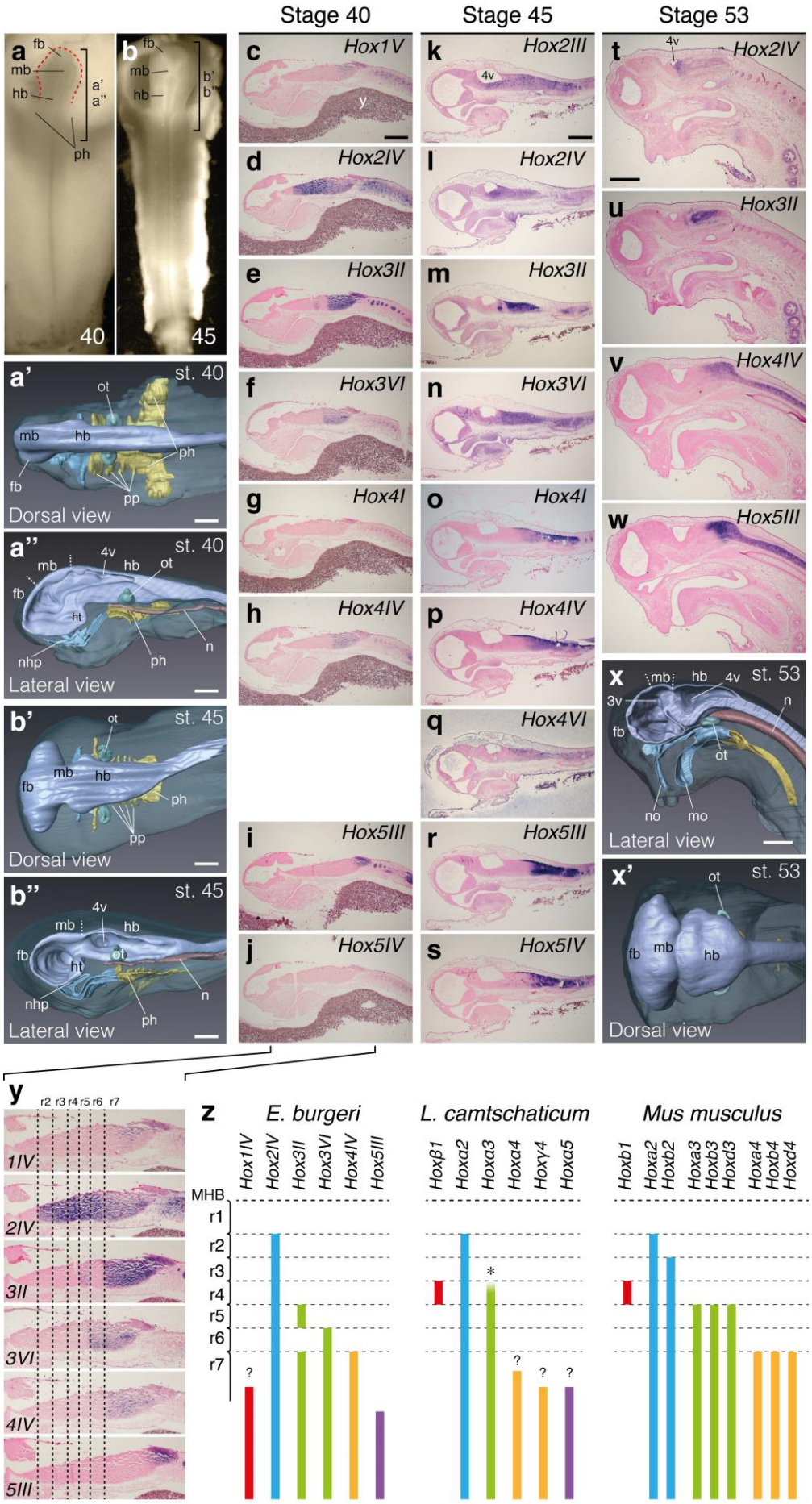
Figure 4. Developmental expression profiling of Hox genes in chordates. Heatmaps of Hox genes expression in *S. torazame* (gnathostome), *L. camtschaticum* and *E. burgeri* (agnathans), and *B. belcheri* (invertebrate chordate), coloured according to Z-score (standard deviations from mean expression level). Anterior Hox genes (top rows of heatmaps) tend to be expressed at higher levels at early stages of development than posterior genes (bottom rows of heatmaps) in both *S. torazame* and *L. camtschaticum*. On top, a phylogenetic tree with chordate relationships of the species studied here indicate the putative events that took place during evolution: in *B. belcheri*, temporal colinearity is appreciated between Hox1-5 genes, indicating WTC was likely present in the last common ancestor of chordates, and a secondary escape of the posterior half of the cluster from it occurred independently in the amphioxus lineage. The large sizes of both amphioxus and agnathan Hox clusters implies that the common ancestor of vertebrates had a so-called 'disorganized' (D) cluster type, while the consolidation towards an 'organized' (O) type occurred in the gnathostome lineage[42], after the split between jawed and jawless vertebrates. In *B. belcheri*, grey rows indicate genes with a FPKM value of 0 in all stages. N, amphioxus neurula stage; H, hatching stage; 5s, 5-somite stage; 20s, 20-somite stage; gs, 1- or 2-gill slit larvae.


a
Inshore hagfish
Eptatretus burgeri
Arctic lamprey
Lethenteron camtschaticum
Small-spotted catshark
Scyliorhinus canicula
Mouse
Mus musculus
Amphioxus
Branchiostoma floridae
6-8
4?
4
1
CYCLOSTOMES
GNATHOSTOMES
b
100 Kbp
1 Mbp
miR-196
miR-10*
Lnp
Evx
Mtx
c
d
e
fb
mb
hb
ph
ov
som
st. 28-30
st. 35
st. 40-45

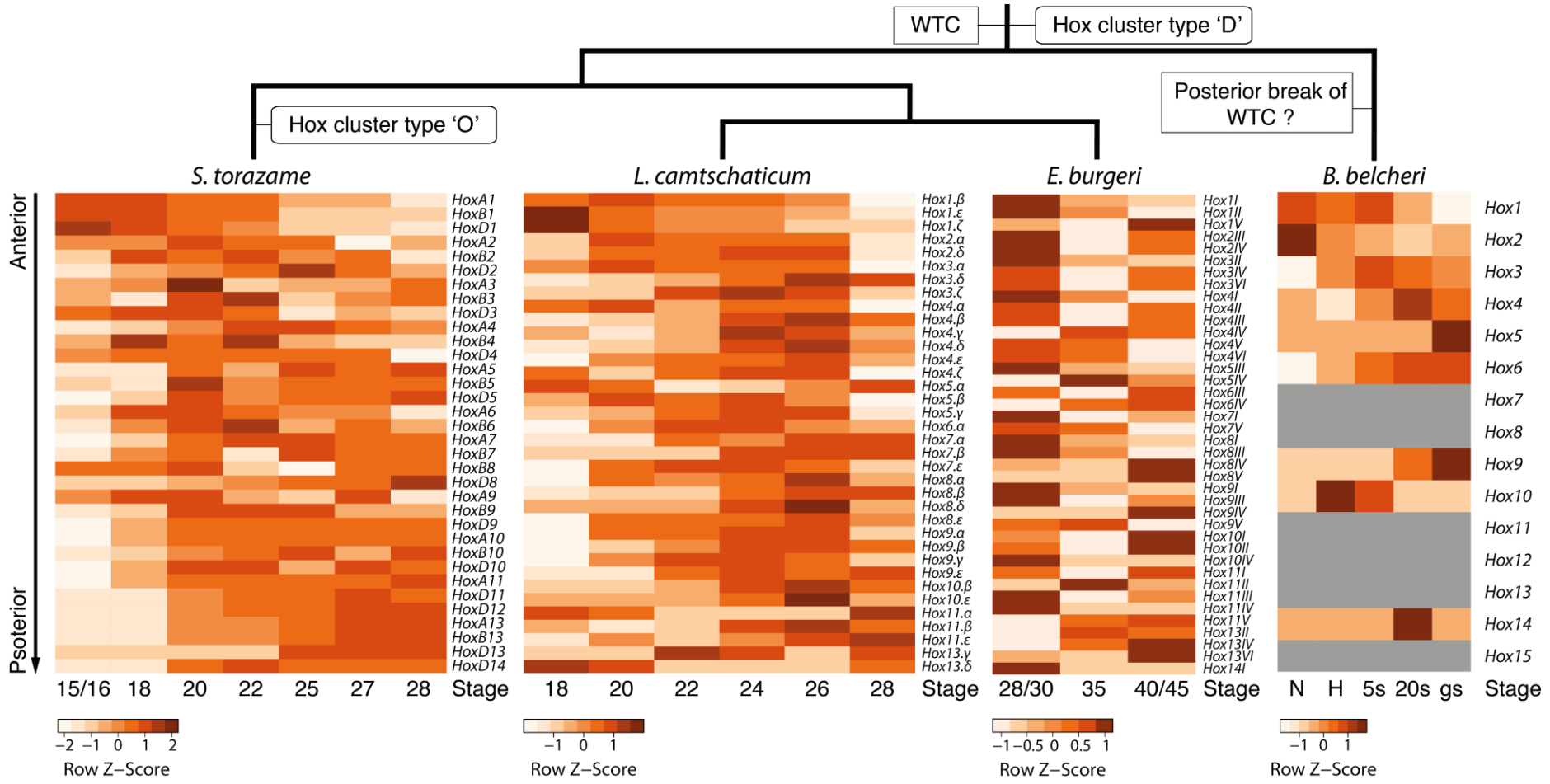
726


Hox10
[Hox12]
Hox11
Hox9
Hox8
Hox13
Hox14
Hox6/7
Hox3
Hox2
Hox5
Hox1
Hox4

727


Stage 40
Stage 45
Stage 53
a
b
fb
mb
hb
a'
a''
b'
b''
ph
40
45
a'
st. 40
ot
mb
hb
ph
fb
pp
Dorsal view
a''
st. 40
4v
ot
ht
n
nhp
Lateral view
b'
st. 45
b''
st. 45
c
Hox1V
y
d
Hox2IV
e
Hox3II
f
Hox3VI
g
Hox4I
h
Hox4IV
i
Hox5III
j
Hox5IV
k
Hox2III
l
Hox2IV
m
Hox3II
n
Hox3VI
o
Hox4I
p
Hox4IV
q
Hox4VI
r
Hox5III
s
Hox5IV
t
Hox2IV
u
Hox3II
v
Hox4IV
w
Hox5III
x
st. 53
3v
no
mo
x'
st. 53
y
r2 r3 r4 r5 r6 r7
1IV
2IV
3II
3VI
4IV
5III
z
E. burgeri
L. camtschaticum
Mus musculus
Hox1IV
Hox2IV
Hox3II
Hox3VI
Hox4IV
Hox5III
Hoxβ1
Hoxα2
Hoxα3
Hoxα4
Hoxγ4
Hoxα5
Hoxb1
Hoxa2
Hoxb2
Hoxa3
Hoxb3
Hoxd3
Hoxa4
Hoxb4
Hoxd4
MHB
r1
r2
r3
r4
r5
r6
r7

728


WTC
Hox cluster type 'D'
Hox cluster type 'O'
Posterior break of WTC ?
S. torazame
L. camtschaticum
E. burgeri
B. belcheri
Anterior
Psoterior
HoxA1 HoxB1 HoxD1 HoxA2 HoxB2 HoxD2 HoxA3 HoxB3 HoxD3 HoxA4 HoxB4 HoxD4 HoxA5 HoxB5 HoxD5 HoxA6 HoxB6 HoxA7 HoxB7 HoxB8 HoxD8 HoxA9 HoxB9 HoxD9 HoxA10 HoxB10 HoxD10 HoxA11 HoxD11 HoxD12 HoxA13 HoxB13 HoxD13 HoxD14
15/16 18 20 22 25 27 28 Stage
−2 −1 0 1 2
Row Z−Score
Hox1.β Hox1.ε Hox1.ζ Hox2.α Hox2.δ Hox3.α Hox3.δ Hox3.ζ Hox4.α Hox4.β Hox4.γ Hox4.δ Hox4.ε Hox4.ζ Hox5.α Hox5.β Hox5.γ Hox6.α Hox7.α Hox7.β Hox7.ε Hox8.α Hox8.β Hox8.δ Hox8.ε Hox9.α Hox9.β Hox9.γ Hox9.ε Hox10.β Hox10.ε Hox11.α Hox11.β Hox11.ε Hox13.γ Hox13.δ
18 20 22 24 26 28 Stage
−1 0 1
Row Z−Score
Hox1I Hox1II Hox1V Hox2III Hox2IV Hox3II Hox3IV Hox3VI Hox4I Hox4II Hox4III Hox4IV Hox4V Hox4VI Hox5III Hox5IV Hox6III Hox6IV Hox7I Hox7V Hox8I Hox8III Hox8IV Hox8V Hox9I Hox9III Hox9IV Hox9V Hox10I Hox10II Hox10IV Hox11I Hox11II Hox11III Hox11IV Hox11V Hox13II Hox13IV Hox13VI Hox14I
28/30 35 40/45 Stage
−1 −0.5 0 0.5 1
Row Z−Score
Hox1 Hox2 Hox3 Hox4 Hox5 Hox6 Hox7 Hox8 Hox9 Hox10 Hox11 Hox12 Hox13 Hox14 Hox15
N H 5s 20s gs Stage
−1 0 1
Row Z−Score

729