

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

3 Juan Pascual-Anaya^{1*}, Iori Sato^{1†}, Fumiaki Sugahara^{1,2†}, Shinnosuke Higuchi¹, Jordi Paps³,
4 Ren Yandong^{4, 5}, Wataru Takagi¹, Adrián Ruiz-Villalba⁶, Kinya G. Ota⁷, Wen Wang⁴,
5 Shigeru Kuratani¹

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

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

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

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

11 ⁵ *University of Chinese Academy of Sciences, Beijing, 100049, China*

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

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

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

17

18 *Correspondence to:

19 Juan Pascual-Anaya

20 e-mail: jpascualanaya@gmail.com

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

23 Tel.: +81-78-306-3589

24 Fax: +81-78-306-3370

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

41

42 Hox genes are fundamental developmental genes with crucial roles for the early specification
43 of embryonic structures along the main anterior-posterior axis of bilaterian animals¹. Hox
44 genes are usually placed in the same genomic regions forming cluster(s). Hox clusters are
45 thought to be the result of several tandem duplication events of an ancestral proto-Hox gene²,
46 and while most invertebrates generally have a single Hox cluster, vertebrate genomes present
47 multiple clusters³. It is widely accepted that the genome of vertebrates has evolved through
48 two rounds (2R) of whole genome duplication (WGD) events (but see ref. 4 for an alternative
49 scenario), generating up to four paralogous loci for each single region of a pre-duplicative

50 genome⁵⁻⁸. Extant vertebrates are divided into two major groups: agnathans, represented by
51 the monophyletic group of cyclostomes (hagfish and lampreys); and gnathostomes,
52 encompassing all jawed-vertebrates in two major groups: cartilaginous fishes (e.g., sharks,
53 rays and chimaeras) and bony vertebrates (e.g., teleosts, coelacanth, amphibians, reptiles,
54 mammals). Tetrapod genomes, including mammals, contain four Hox clusters, named from
55 HoxA to HoxD, as the result of these 2R-WGD (Fig. 1a). Although the 2R-WGD events are
56 generally accepted, the timing of these events with respect to the divergence of cyclostomes
57 and gnathostomes is still a matter of intense debate^{4,9-11}. Despite extended research on
58 vertebrate genomes, this has mostly focused on representative species of gnathostomes, while
59 cyclostomes have remained poorly understood. A recent study of the genome of the Artic
60 lamprey, *Lethenteron camtschaticum* (or Japanese lamprey, *Lethenteron japonicum*),
61 suggested that lampreys had probably undergone a third round of WGD event (3R-WGD)¹⁰.
62 Whether this event is an independent, lineage-specific event remains a mystery, since the
63 Hox complement of the hagfish is unknown.

64 The position of Hox genes in the cluster determines their expression patterns. Spatial
65 colinearity refers to the property by which the anterior limit of expression of a given Hox
66 gene is generally more rostral than its upstream (more 5') counterpart. Spatial colinearity is
67 widely conserved among bilaterians studied so far, even in cases where the Hox cluster is
68 completely atomized¹². Temporal colinearity refers to the phenomenon describing the
69 temporal order of expression of Hox genes according to their position in the cluster, i.e.,
70 genes in the 3' part are expressed earlier, and was first described in the HoxD cluster of the
71 mouse^{13,14}. Indeed, this so-called whole-cluster temporal colinearity (WTC)¹⁵ phenomenon
72 had been described only in jawed vertebrates. The recent analysis of the scallop genome and
73 the reanalysis of Hox gene expression in a wide range of invertebrates has revealed that Hox
74 genes of these species follow what is called a subcluster-level temporal colinearity (STC),

75 i.e., that the cluster is divided into small, contiguous groups of Hox genes, each of these
76 displaying temporal colinearity¹⁵. This situation leads to the uncertainty of what was the
77 ancestral condition before deuterostomes and protostomes split. Moreover, temporal
78 colinearity has not been described in any cyclostome species so far. In *L. camtschaticum*,
79 Hox genes known to be in the same cluster were not expressed following temporal
80 colinearity¹⁶, and both the Hox gene repertoire and expression of the other major group of
81 cyclostomes, the hagfish, is mostly unknown^{3,17}. Therefore, the evolutionary origin of WTC
82 as observed in gnathostomes remains obscure.

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

92

93 **Results and discussion**

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

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

103 To determine the genomic organization of hagfish Hox genes we then screened a
104 BAC library built from blood genomic DNA. We found 25 BAC clones spanning only 15 out
105 of the 40 Hox genes (Supplementary Fig. 1, 2). Recently, it has been described that the
106 lamprey genome goes through somatic rearrangements, differentially eliminating stretches of
107 germ line-specific sequences, which might include protein-coding genes²⁰. Considering that
108 the hagfish, which is known to go through a chromosome elimination process in somatic
109 tissues during development²¹, might be losing Hox genes in somatic tissues, we decided to
110 generate a draft genome using genomic DNA obtained from the testis (germ line) of a single
111 individual. In our preliminary assembly, we found evidence for at least six Hox clusters
112 containing all 40 Hox genes found in the transcriptome and three microRNAs, together with
113 conserved syntenic non-Hox genes (Fig. 1b). The hagfish Hox repertoire and genomic
114 organization are overall very similar to the one described in the *L. camtschaticum* genome¹⁰
115 (number of genes —43 in the lamprey— and putative clusters —six in the lamprey—), raising
116 the possibility that the 3R-WGD event suggested to have occurred in the lamprey lineage¹⁰
117 took place before the split of lampreys and hagfish lineages. Surprisingly, we found a hagfish
118 Hox13 gene (*Hox13VI*) enclosed by two conserved syntenic genes: *Lunapark* (*Lnp*) and
119 *Even-skipped* (*Evx*). This suggests that a translocation event took place in the hagfish lineage,
120 likely together with a severe disintegration of a cluster involving large Hox gene losses.

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

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

131 The obscure orthology relationship between jawed and jawless vertebrate genes has
132 been broadly described for both Hox and non-Hox gene families²². It is unclear whether the
133 2R-WGD events that took place during early vertebrate evolution are shared or not among
134 cyclostomes and gnathostomes^{6,23,24}. The lack of one-to-one orthology relationships between
135 genes from both groups can be taken as evidence for independent WGD events. However,
136 despite their obscure phylogenetic relationship, Hox clusters of cyclostomes and
137 gnathostomes can still be the result of an ancestral 2R-WGD, if the duplicated regions
138 containing the Hox clusters had not completed rediploidization before the split of
139 cyclostomes and gnathostomes²⁵. Consequently, certain number of phylogenetic analyses
140 would support a shared WGD between cyclostomes and gnathostomes as it seems to be the
141 case⁹. These would correspond to those genes that had differentiated into different alleles
142 before the split of the two lineages.

143 Once confirmed the presence of clusters, we wondered whether the hagfish Hox genes
144 were expressed according to the spatial colinearity rule. Spatial colinearity has been observed
145 in the vast majority of bilaterians studied so far, included the lamprey^{16,26}, and with only few
146 exceptions²⁷. In both the lamprey and gnathostomes, nested expression of anterior Hox genes
147 is coupled to the morphological segmentation of the hindbrain into discrete rhombomeres,
148 and this is controlled by a highly conserved gene regulatory network, established at least in

149 the last common ancestor of vertebrates²⁸. The hagfish hindbrain is, as in the lamprey²⁹,
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 gnathostomes³⁰ (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 border³¹ (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 *Hox1III*, 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

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

176 The evolution of the expression domains of Hox3 genes in the hindbrain of different
177 vertebrates is particularly interesting. Considering the global expression pattern of Hox3
178 paralogs in each group, we observe that while in the lamprey (*Hox3 α*) and the shark
179 *Scyliorhinus canicula* (*Hoxb3*), Hox3 genes are expressed from r4 (refs. 26, 30, 32), in the
180 hagfish and osteichthyans Hox3 genes are expressed up to r5 (Fig. 3z). There are two
181 possible evolutionary explanations for this difference, both involving parallel evolutionary
182 events: either a caudal shift of Hox3 expression domains from r4 to r5 convergently
183 happened in both the lamprey and osteichthyan lineages, or, on the other hand, a rostral shift
184 from r5 to r4 occurred in the lamprey and chondrichthyans. In a different lamprey species,
185 *Petromyzon marinus*, the *Pm1Hox3* gene, orthologous of *L. camtschaticum* *Hox3 α* , was
186 found to be expressed from r5 like in mammals²⁸. This could favour the hypothesis of a
187 convergent expression shift in both the Arctic lamprey and the shark as lineage or species-
188 specific changes.

189 Following the spatial colinearity rule, most posterior PG Hox genes are expressed in
190 the most caudal regions of the embryo. One of the expression domains of Hox13 paralogue
191 genes are the most posterior parts of the hindgut. Concordantly, hagfish *Hox13II* and
192 *Hox13VI* were found around the cloacal region of a juvenile (stage 60; Supplementary Fig.
193 10), as in the lamprey and other vertebrates³³. Vertebrate Hox14 genes have also been
194 reported to be expressed in the most posterior parts of the hindgut of the lamprey and the
195 shark³³. However, we were not able to detect any signal for *Hox14I* transcripts in the cloaca
196 of the hagfish larva (Supplementary Fig.9).

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

199 segmentation²⁸ is well conserved in the hagfish. More important than the similarities,
200 elucidating what specific regulatory inputs account for lineage-specific differences in the
201 hindbrain Hox code, such as the striped expression of hagfish *Hox3II* and the different rostral
202 limits of expression of different Hox3 genes in different vertebrates, will be helpful to
203 determine how the hindbrain GRN diversified during vertebrate evolution, and what are the
204 functional and morphological implications of these differences.

205 To unravel the evolutionary origin of WTC in vertebrates, we further carried out a
206 comprehensive analysis of the developmental expression profile of Hox genes using embryos
207 from both jawed and jawless vertebrates. Together with the RNA-seq data generated for *E.*
208 *burgeri*, we sequenced RNA-seq libraries covering early to late developmental stages of the
209 lamprey *L. camtschaticum*³⁴ and the gnathostome catshark *Scyliorhinus torazame*³⁵ and
210 quantified the expression profiles of Hox genes. As expected, the expression profiles of *S.*
211 *torazame* Hox genes were consistent with temporal colinearity across all the clusters found in
212 other jawed vertebrates, showing a clear tendency of anterior Hox genes (Hox1-3) expressed
213 at earlier stages and posterior ones (Hox9-14) at later (Fig. 4; Supplementary Fig. 10).
214 Despite previous reports¹⁶, lamprey Hox genes (for which we found an unreported Hox1
215 gene, *Hox1ζ*) also followed the rule of temporal colinearity (Fig. 4; Supplementary Fig. 10).
216 Interestingly, the HOX- γ cluster has completely lost the temporal colinearity (Supplementary
217 Fig. 10). HOX- γ is one of the most degenerated clusters in the lamprey with only 4 Hox
218 genes¹⁰ (Fig. 1a), which might be a direct consequence of the lack of temporal colinearity. In
219 the hagfish, although obtaining a pool of embryos from a full developmental series is
220 unfeasible, a similar tendency was also observed: levels of posterior *Hox11.I*, *Hox11.V* and
221 all Hox13 genes are higher at stage 40-45 (comparable to lamprey stage 25-26) than at 28-30
222 (lamprey stage 22-23), while generally all anterior and central Hox genes levels are higher at
223 stage 28-30 than at later stages (Fig. 4).

224 The above observations imply that cyclostome Hox expression profiles, as in
225 gnathostomes, are consistent with the WTC rule, suggesting at least a vertebrate origin. In
226 order to determine whether WTC was present before the origin of vertebrates, we
227 investigated the Hox expression profiles of a chordate outgroup. Wang and colleagues¹⁵
228 described the tunicate *Ciona intestinalis* Hox gene expressions as according to the STC.
229 However, their statement was based on the reanalysis of data from whole mount *in situ*
230 hybridization³⁶, which is not a quantitative technique. Cephalochordates are the closest
231 lineage to vertebrates with an intact Hox cluster, and is thus very informative in this regard.
232 Expression profiles of Hox genes in the cephalochordate amphioxus *Branchiostoma*
233 *belcheri*³⁷ show that amphioxus *Hox1* to *Hox5* are expressed in an anterior Hox/early-
234 posterior/late manner. However, *Hox6*, *Hox10* and *Hox14* genes violate this pattern,
235 consistent with our previous report³⁸, and *Hox7-8*, *Hox11-13* and *Hox15* were not detected
236 during the stages assayed, contributing to the dismantling of the colinearity (Fig. 4). In most
237 invertebrate species where STC has been described, Hox1-2 or Hox1-3 was the most anterior
238 subgroup showing temporal colinearity¹⁵. The fact that in amphioxus Hox1-5 are expressed in
239 temporal order as a single group indicates that this expression pattern is reminiscent from a
240 genuine WTC, which was subsequently broken from Hox6 in the cephalochordate lineage
241 (Fig. 4). In addition, amphioxus Hox6-15 genes might still follow WTC at later stages than
242 the ones assayed here³⁷. The putative presence of WTC in both the cephalochordate and
243 vertebrate lineages implies that it was likely present in the last common ancestor of
244 vertebrates.

245 Taken together, our results depict a scenario in which chordate Hox genes are
246 expressed following WTC, and protostome Hox genes according to STC. This, importantly,
247 can offer a mechanistic answer to explain the radically different bauplans displayed by
248 chordates and protostome invertebrates. Deschamps and Duboule³⁹ have recently proposed

249 that temporal colinearity, as seen in mammals (WTC), is displayed only by animals that
250 follow a developmental strategy of anterior to posterior elongation, adding new regions to the
251 main body axis from a posterior growth zone. This temporal activation of Hox genes during
252 the posterior elongation, or Hox clock, translates during development into the spatial
253 colinearity observed along the main anterior-posterior body axis. The fact that the lamprey
254 and the hagfish also develop according to this posterior elongation –a developmental mode
255 thus very well conserved across vertebrates–, together with the presence of WTC and spatial
256 colinearity in the main axial structures of these animals (this study and refs. 16, 26), supports
257 Deschamps and Duboule’s hypothesis³⁹. This implies that this mechanism was present in the
258 last common ancestor of vertebrates, although some lineage-specific differences might have
259 occurred in the mechanism transmitting the Hox clock from the posterior progenitors into the
260 resulting axial structures (for instance, there are differences in the expression of Hox10 genes
261 between lampreys and amniotes in the tailbud and axial mesoderm, see ref. 16). Ultimately,
262 the question of whether the Hox cluster of the last common bilaterian ancestor was expressed
263 according to either whole-cluster or subcluster modes of temporal expression remains open.
264 A more detailed investigation of the temporal expression of Hox genes in non-chordate
265 deuterostome groups (namely, ambulacrarians –e.g., sea urchins, sea stars, acorn worms–)
266 will be thus needed to ultimately resolve this question^{40,41}.

267 It has been proposed that gnathostome Hox clusters are relatively compacted, or
268 ‘organized’, due to a consolidation process that was associated with the emergence of meta-
269 *cis* regulation of the cluster, and probably facilitated by the 2R-WGD events that occurred
270 during vertebrate evolution⁴². Hox clusters of cyclostomes are, on the other hand, more akin
271 to ‘disorganized’ types of clusters, like the one of amphioxus⁴² –because of their extremely
272 large sizes–, suggesting that this consolidation did not start in the last common ancestor of
273 vertebrates, but rather was a progressive gnathostome-specific process⁴² (Fig. 4). Further

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

284

285 **METHODS**

286 **Animal sampling, experiments and aquarium maintenance**

287 *E. burgeri* embryos (staged according to refs. 18 and 19) used in this study were obtained
288 from adult hagfish individuals captured in the Japan Sea off Shimane prefecture as previously
289 described⁴³, during August of a given year. Eggs were laid in a cage deposited in the natural
290 environment in the sea in October of the same year. Deposited eggs were then incubated in
291 laboratory aquariums with artificial sea water at 16 °C under controlled conditions, until
292 developing embryos are apparent around February or March of the following year. Hagfish
293 embryos used for RNA-seq were from adults captured in 2010, and were assayed in February
294 2011(total developing embryos 8 out of >150 eggs). Hagfish embryos used for *in situ*
295 hybridization were from adults captured in 2016 (stage 40), 2013 (stage 45) and 2014 (stage
296 60), and embryos fixed in March 2017, 2014 and 2015, respectively. Sections of stage 53
297 were from an embryo previously reported⁴⁴. Lamprey (*L. camtschaticum*) and cloudy
298 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 method⁴⁶ 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 MIRA⁴⁷ 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 v2.1.1⁴⁸ 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

324 further integrated with CD-HIT-EST⁴⁹ v4.6.4 with parameters ‘-c 0.98’. A fourth
325 assembly was done with SOAPdenovoTrans v1.03⁵⁰ using all reads simultaneously and
326 multiple k-mers (19, 21, 23, 25, 27, 29, 31: with ‘SOAPdenovo-Trans-31mer’
327 command; and 41, 51, 61, 71, 81, 91: with ‘SOAPdenovo-Trans-127mer’), with a final
328 integration with CD-HIT-EST. In the case of the lamprey and shark, reads were assembled
329 according to 3 different pipelines: (1) assembly with Trinity v2.1.1 of reads coming from each
330 pool of embryos independently, taking into account the strand-specific information (--
331 SS_lib_type RF), and integrated with CD-HIT-EST v4.6.4; (2) idem, but not taking into
332 account the strand-specific information; and (3) assembly of all reads together. In the case of
333 the lamprey, a fourth assembly strategy was carried out by means of integrating a genome-
334 guided assembly (option --genome_guided_bam of Trinity, mapping the RNA-seq reads
335 to *L. camtschaticum* 1.0 genome¹⁰ with the splice-aware mapper HISAT2⁵¹) and the above *de*
336 *novo* assembly #3, using the PASA v2.0.2⁵² pipeline
337 (http://pasapipeline.github.io/#A_ComprehensiveTranscriptome). Finally, completeness
338 assessments of all versions were done using CEGMA v2.5 and BUSCO v1.1b1 programs, as
339 previously described⁵³ (Supplementary Tables 1-3). The most complete versions of *E.*
340 *burgeri* and *L. camtschaticum* were selected for further analysis. In the case of *S. torazame*,
341 although strategy #3 gave as a result a more complete transcriptome in general, it contained
342 more fragmented Hox genes than version #1, and therefore we selected the latter. All *E.*
343 *burgeri*, *L. camtschaticum* and *S. torazame* RNA-seq data have been deposited in NCBI
344 GenBank under the BioProject number PRJNA371391. Amphioxus *B. belcheri* transcriptome
345 was assembled using previously published RNA-seq data, from the NCBI’s SRA database,
346 under BioProject numbers PRJNA310680³⁷ and PRJNA214454⁵⁴. *B. belcheri* RNA-seq reads
347 from the former BioProject were subjected to adaptor trimming with cutadapt v1.10⁵⁵. All *B.*
348 *belcheri* RNA-seq data was then assembled following the same above-mentioned pipeline for

349 the lamprey transcriptome (strategy #3), using the previously published *B. belcheri* genome⁵⁴
350 for the PASA pipeline.

351 All Trinity commands were executed using the `--group_pairs_distance 999`
352 parameter value⁵⁶.

353

354 **BAC library, clone screening and PacBio sequencing and assembly**

355 Blood was drawn from the caudal sub-cutaneous sinus of one adult specimen of *E. burgeri*
356 using a heparin-rinsed disposable syringe. The whole blood sample was immediately frozen
357 in liquid nitrogen, and used for DNA extraction. A BAC library consisting of 129,024 clones,
358 with an average insert size of 100 Kbp (~4.4X of the *E. burgeri* genome size), was
359 constructed using the pCCBAC1 vector⁵⁷ [CopyControlTM BAC Cloning Kit (*Hind*III)
360 (EPICENTRE)] and pooled into 96-well and 384-well plates according to the Matrix Pool
361 and Superpool Strategy⁵⁸ by Amplicon Express (Pullman, WA, USA). The BAC library was
362 screened for Hox-containing clones by means of PCR with specific primers⁵⁸. Positive BAC
363 clones were extracted with the QIAGEN Large-Construct kit, and sequenced in different
364 pools using SMRT technology on a PacBio *RS* platform using XL-C2 chemistry, or on a *RSII*
365 platform using P4-C2 chemistry. pCCBAC1 vector sequence were masked using a script
366 from ref. 59 with minor modifications. BAC clones were assembled using masked subreads
367 with MIRA⁴⁷ v4.9.5_2. The sequence of the BAC clones used in this study have been
368 deposited in GenBank (accession numbers MF182102-MF182109).

369

370 **Genome sequencing and assembly**

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

374 mate-pair (5-Kbp, 5~7-Kbp, 7~10-Kbp, 10~15-Kbp and 15~20-Kbp) libraries, generating
375 >300X coverage of the estimated 2.906 Gb-long genome of the hagfish. All short-read data
376 were corrected by SOAPec v2.01⁶⁰ using >40X data. Assembly of the hagfish genome was
377 performed with ABySS v1.9.0⁶¹ with a k-mer size of 79, followed by a scaffolding step with
378 SOAPdenovo v2.04-r241⁶⁰ software (parameter ‘-K 41 -d 1 -M 2 -F’). Gaps were
379 finally filled with GapCloser v1.12-r6⁶⁰. The resulting assembly (size, ~2.59 Gb; N50, ~439
380 Kbp) was used for the screening of Hox clusters. Hox-containing scaffolds were then aligned
381 against the BAC clones using MUMmer v3.23⁶² and visualized using mummerplot, bundled
382 within the same software. Sequences of Hox-containing scaffolds, as well as those of *E.*
383 *burgeri Lnp* and *Evx* (whose sequences are not complete in the genome) have been deposited
384 in GenBank under accession numbers MF398213-MF398235. A publication with more
385 detailed and in-depth analysis of the *E. burgeri* genome is now in preparation.

386

387 **Identification of Hox genes**

388 UniProt Knowledgebase (UniProtKB) database (<http://www.uniprot.org/>) was searched for
389 entries containing the term “Hox” and restricted to Eumetazoans (name:hox, taxonomy:6072;
390 UniProt release 2015_11). Resulting entries were downloaded and used as queries against the
391 transcriptome assembly and genome of the hagfish by means of TBLASTN (NCBI BLAST
392 v2.2.31+⁶³). The best BLAST hits were then used as queries against the whole UniProtKB
393 database using BLASTX. Those transcripts whose reciprocal best hit was a Hox gene were
394 kept and manually inspected for false positives. Lamprey Hox genes were downloaded from
395 NCBI GenBank¹⁰ and blasted against our lamprey transcriptome assembly to identify Hox
396 transcripts. *Hox4η*, *Hox7ε*, *Hox9ζ*, *Hox11δ*, *Hox13α*, *Hox13ε*, *Hox13ζ* and *Hox14ε* were not
397 found in our transcriptome assembly. We found an extra, unreported Hox1 paralogous gene,
398 which we named *Hox1ζ* (following the nomenclature from ref. 10). *Scyliorhinus canicula*

399 Hox genes³⁰ sequences were downloaded from NCBI GenBank and used as queries to
400 identify orthologous sequences in our *S. torazame* transcriptome by means of TBLASTN.
401 The *L. camtschaticum Hox1ζ* and *S. torazame* Hox gene sequences were deposited in
402 GenBank (accession numbers MF398236-MF398269).

403

404 **cDNA cloning and section *in situ* hybridization**

405 Selected Hox genes were cloned from cDNA prepared for a previous study⁴⁴ using specific
406 primers. *In situ* hybridization on paraffin wax-embedded sections of stage 45 and 60 hagfish
407 embryos was performed according to refs. 44, 45. Haematoxylin and eosin (H&E) staining on
408 paraffin sections of stage 60 embryo was carried out by standard protocol. H&E stained
409 sections were further stained with Alcian Blue⁶⁴.

410

411 **3D reconstruction of the hagfish embryos**

412 The 3D reconstruction images of hagfish embryos were reconstructed based on images taken
413 of 1 in every 10 histological sagittal sections at 6 μm, stained with standard haematoxylin and
414 eosin staining protocols for the stage 40 embryo, and 1 in 2 unstained sections at 8 μm for the
415 stage 45 embryo. Reconstructed images were acquired using Avizo software (Visualization
416 Sciences Group). Stage 53 reconstruction is from an embryo used previously¹⁹.

417

418 **Molecular phylogenetic analyses**

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

424 datasets based on different gene content were assembled: 1) Anterior genes (Hox1-3), 2) Hox
425 4 genes, 3) Central genes (Hox4-8), 4) Posterior genes (Hox9-14), and 5) all Hox genes
426 together. The datasets were aligned using MAFFT v7.123b⁶⁸ using the “*auto*” option, regions
427 of ambiguous alignment trimmed with Gblocks v0.91b⁶⁹ using the less stringent options.
428 Alignments were visually inspected with BioEdit v7.2.6⁷⁰. Phylogenetic trees were inferred
429 with RAxML v8.2.10⁷¹ using a random starting tree, the evolutionary model LG + Gamma +
430 Invariants with empirical base frequencies, and 1000 rapid bootstrap replicates. Trees were
431 edited using FigTree v1.4.3 (<http://tree.bio.ed.ac.uk/software/figtree/>).

432

433 **Expression profiling of Hox genes.**

434 RNA-seq reads from individual embryos were used to quantify transcripts of the selected
435 transcriptomes of *E. burgeri* and *S. torazame* using Perl scripts
436 ‘align_and_estimate_abundance.pl’ and ‘align_and_estimate_abundance.pl’, bundled with
437 Trinity v.2.1.1, and using RSEM v1.2.28⁷² as quantification method
438 (<https://github.com/trinityrnaseq/trinityrnaseq/wiki/Trinity-Transcript-Quantification>). Hox
439 transcripts of *S. torazame* were directly quantified using RSEM with data from each
440 embryonic stage. TPM values from either genes (for Hox genes represented with a single
441 transcript in the assemblies) or isoforms (for Hox genes represented with several transcripts)
442 were then selected and a heat map analyses of the log(TPM+0.1) were conducted in R using
443 heatmap.2 (gplots package⁷³) scaling by gene (row Z-score), and implemented in RStudio
444 v1.0.136⁷⁴ [with R v3.3.0 (2016-05-03)⁷⁵]. *B. belcheri* Hox transcripts were quantified using
445 previously published DGE-seq data³⁷ with DGE-EM v1.0.0⁷⁶ software, and FPKM values
446 were analysed as above.

447

448 **Data availability.**

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

453

454

455 **References**

- 456 1 Pearson, J. C., Lemons, D. & McGinnis, W. Modulating Hox gene functions during
457 animal body patterning. *Nat. Rev. Genet.* **6**, 893-904 (2005).
- 458 2 Garcia-Fernàndez, J. The genesis and evolution of homeobox gene clusters. *Nat.*
459 *Rev. Genet.* **6**, 881-892 (2005).
- 460 3 Pascual-Anaya, J., D'Aniello, S., Kuratani, S. & Garcia-Fernandez, J. Evolution of
461 Hox gene clusters in deuterostomes. *BMC Dev. Biol.* **13** (2013).
- 462 4 Smith, J. J. & Keinath, M. C. The sea lamprey meiotic map improves resolution of
463 ancient vertebrate genome duplications. *Genome Res* **25**, 1081-1090 (2015).
- 464 5 Dehal, P. & Boore, J. L. Two rounds of whole genome duplication in the ancestral
465 vertebrate. *PLoS Biol.* **3**, e314 (2005).
- 466 6 Putnam, N. H. *et al.* The amphioxus genome and the evolution of the chordate
467 karyotype. *Nature* **453**, 1064-1071 (2008).
- 468 7 Van de Peer, Y., Maere, S. & Meyer, A. The evolutionary significance of ancient
469 genome duplications. *Nat. Rev. Genet.* **10**, 725-732 (2009).
- 470 8 Van de Peer, Y., Maere, S. & Meyer, A. 2R or not 2R is not the question anymore.
471 *Nat. Rev. Genet.* **11**, 166 (2010).

- 472 9 Kuraku, S., Meyer, A. & Kuratani, S. Timing of genome duplications relative to the
473 origin of the vertebrates: did cyclostomes diverge before or after? *Mol. Biol. Evol.*
474 **26**, 47-59 (2009).
- 475 10 Mehta, T. K. *et al.* Evidence for at least six Hox clusters in the Japanese lamprey
476 (*Lethenteron japonicum*). *Proc Natl Acad Sci U S A* **110**, 16044-16049 (2013).
- 477 11 Smith, J. J. *et al.* Sequencing of the sea lamprey (*Petromyzon marinus*) genome
478 provides insights into vertebrate evolution. *Nat. Genet.* **45**, 415-421 (2013).
- 479 12 Seo, H. C. *et al.* Hox cluster disintegration with persistent anteroposterior order
480 of expression in *Oikopleura dioica*. *Nature* **431**, 67-71 (2004).
- 481 13 Dollé, P., Izpisúa-Belmonte, J. C., Falkenstein, H., Renucci, A. & Duboule, D.
482 Coordinate expression of the murine *Hox-5* complex homoeobox-containing
483 genes during limb pattern formation. *Nature* **342**, 767-772 (1989).
- 484 14 Izpisúa-Belmonte, J. C., Falkenstein, H., Dollé, P., Renucci, A. & Duboule, D. Murine
485 genes related to the *Drosophila AbdB* homeotic genes are sequentially expressed
486 during development of the posterior part of the body. *EMBO J.* **10**, 2279-2289
487 (1991).
- 488 15 Wang, S. *et al.* Scallop genome provides insights into evolution of bilaterian
489 karyotype and development. *Nat. Ecol. Evol.* **1**, 0120 (2017).
- 490 16 Takio, Y. *et al.* Hox gene expression patterns in *Lethenteron japonicum*
491 embryos—Insights into the evolution of the vertebrate Hox code. *Dev. Biol.* **308**,
492 606-620 (2007).

- 493 17 Stadler, P. F. *et al.* Evidence for independent *Hox* gene duplications in the hagfish
494 lineage: a PCR-based gene inventory of *Eptatretus stoutii*. *Mol. Phylogenet. Evol.*
495 **32**, 686-694 (2004).
- 496 18 Dean, B. On the Embryology of *Bdellostoma stouti*: A General Account of Myxinoid
497 Development from the Egg and Segmentation to Hatching. 220-276 (G. Fischer,
498 1899).
- 499 19 Oisi, Y., Ota, K. G., Kuraku, S., Fujimoto, S. & Kuratani, S. Craniofacial development
500 of hagfishes and the evolution of vertebrates. *Nature* **493**, 175-180 (2013).
- 501 20 Smith, J. J., Baker, C., Eichler, E. E. & Amemiya, C. T. Genetic consequences of
502 programmed genome rearrangement. *Curr. Biol.* **22**, 1524-1529 (2012).
- 503 21 Kohno, S.-i., Kubota, S. & Nakai, Y. in *The Biology of Hagfishes*. 81-100 (Springer
504 Netherlands, 1998).
- 505 22 Kuraku, S. Impact of asymmetric gene repertoire between cyclostomes and
506 gnathostomes. *Semin. Cell Dev. Biol.* **24**, 119-127 (2013).
- 507 23 Ohno, S. *Evolution by Gene Duplication*. (Springer-Verlag, 1970).
- 508 24 Kasahara, M. The 2R hypothesis: an update. *Curr. Opin. Immunol.* **19**, 547-552
509 (2007).
- 510 25 Martin, K. J. & Holland, P. W. Enigmatic orthology relationships between *Hox*
511 clusters of the African butterfly fish and other teleosts following ancient whole-
512 genome duplication. *Mol. Biol. Evol.* **31**, 2592-2611 (2014).
- 513 26 Takio, Y. *et al.* Evolutionary biology: lamprey *Hox* genes and the evolution of
514 jaws. *Nature* **429**, 1 p following 262 (2004).

- 515 27 Schiemann, S. M. *et al.* Clustered brachiopod Hox genes are not expressed
516 collinearly and are associated with lophotrochozoan novelties. *Proc. Natl. Acad.*
517 *Sci. U. S. A.* **114**, E1913-E1922 (2017).
- 518 28 Parker, H. J., Bronner, M. E. & Krumlauf, R. A *Hox* regulatory network of hindbrain
519 segmentation is conserved to the base of vertebrates. *Nature* **514**, 490-493
520 (2014).
- 521 29 Kuratani, S., Horigome, N., Ueki, T., Aizawa, S. & Hirano, S. Stereotyped axonal
522 bundle formation and neuromeric patterns in embryos of a cyclostome,
523 *Lampetra japonica*. *J. Comp. Neurol.* **391**, 99-114 (1998).
- 524 30 Oulion, S. *et al.* Evolution of repeated structures along the body axis of jawed
525 vertebrates, insights from the *Scyliorhinus canicula* Hox code. *Evol. Dev.* **13**, 247-
526 259 (2011).
- 527 31 Wingate, R. J. The rhombic lip and early cerebellar development. *Curr. Opin.*
528 *Neurobiol.* **11**, 82-88 (2001).
- 529 32 Murakami, Y. *et al.* Segmental development of reticulospinal and branchiomotor
530 neurons in lamprey: insights into the evolution of the vertebrate hindbrain.
531 *Development* **131**, 983-995 (2004).
- 532 33 Kuraku, S. *et al.* Noncanonical role of Hox14 revealed by its expression patterns
533 in lamprey and shark. *Proc. Natl. Acad. Sci. U. S. A.*, 0710947105 (2008).
- 534 34 Tahara, Y. Normal stages of development in the lamprey, *Lampetra reissneri*
535 (Dybowski). *Zool. Sci.* **5**, 109-118 (1988).

- 536 35 Ballard, W. W., Mellinger, J. & Lechenault, H. A series of normal stages for
537 development of *Scyliorhinus canicula*, the lesser spotted dogfish (*Chondrichthyes*:
538 *Scyliorhinidae*). *J. Exp. Zool.* **267**, 318-336 (1993).
- 539 36 Ikuta, T., Yoshida, N., Satoh, N. & Saiga, H. *Ciona intestinalis* Hox gene cluster: Its
540 dispersed structure and residual colinear expression in development. *Proc. Natl.*
541 *Acad. Sci. U. S. A.* **101**, 15118-15123 (2004).
- 542 37 Yang, K. Y. *et al.* Transcriptome analysis of different developmental stages of
543 amphioxus reveals dynamic changes of distinct classes of genes during
544 development. *Sci. Rep.* **6**, 23195 (2016).
- 545 38 Pascual-Anaya, J. *et al.* Broken colinearity of the amphioxus Hox cluster. *EvoDevo*
546 **3** (2012).
- 547 39 Deschamps, J. & Duboule, D. Embryonic timing, axial stem cells, chromatin
548 dynamics, and the Hox clock. *Genes Dev.* **31**, 1406-1416 (2017).
- 549 40 Aronowicz, J. & Lowe, C. J. Hox gene expression in the hemichordate *Saccoglossus*
550 *kowalevskii* and the evolution of deuterostome nervous systems. *Integr. Comp.*
551 *Biol.* **46**, 890-901 (2006).
- 552 41 Gonzalez, P., Uhlinger, K. R. & Lowe, C. J. The Adult Body Plan of Indirect
553 Developing Hemichordates Develops by Adding a Hox-Patterned Trunk to an
554 Anterior Larval Territory. *Curr. Biol.* **27**, 87-95 (2017).
- 555 42 Duboule, D. The rise and fall of Hox gene clusters. *Development* **134**, 2549-2560
556 (2007).
- 557 43 Oisi, Y., Kakitani, O., Kuratani, S. & Ota, K. G. in *In Situ Hybridization Methods* (ed
558 Giselbert Hauptmann) 249-262 (Springer New York, 2015).

559 44 Sugahara, F. *et al.* Evidence from cyclostomes for complex regionalization of the
560 ancestral vertebrate brain. *Nature* **531**, 97-100 (2016).

561 45 Sugahara, F., Murakami, Y. & Kuratani, S. in *In Situ Hybridization Methods* (ed
562 Giselbert Hauptmann) 263-278 (Springer New York, 2015).

563 46 Sultan, M. *et al.* A simple strand-specific RNA-Seq library preparation protocol
564 combining the Illumina TruSeq RNA and the dUTP methods. *Biochem. Biophys.*
565 *Res. Commun.* **422**, 643-646 (2012).

566 47 Chevreux, B., Wetter, T. & Suhai, S. in *Computer Science and Biology: Proceedings*
567 *of the German Conference on Bioinformatics (GCB)* 45-56 (1999).

568 48 Grabherr, M. G. *et al.* Full-length transcriptome assembly from RNA-Seq data
569 without a reference genome. *Nat. Biotechnol.* **29**, 644-652 (2011).

570 49 Li, W. & Godzik, A. Cd-hit: a fast program for clustering and comparing large sets
571 of protein or nucleotide sequences. *Bioinformatics* **22**, 1658-1659 (2006).

572 50 Xie, Y. *et al.* SOAPdenovo-Trans: de novo transcriptome assembly with short
573 RNA-Seq reads. *Bioinformatics* **30**, 1660-1666 (2014).

574 51 Kim, D., Langmead, B. & Salzberg, S. L. HISAT: a fast spliced aligner with low
575 memory requirements. *Nat. Methods* **12**, 357-360 (2015).

576 52 Haas, B. J. *et al.* Improving the *Arabidopsis* genome annotation using maximal
577 transcript alignment assemblies. *Nucleic Acids Res.* **31**, 5654-5666 (2003).

578 53 Hara, Y. *et al.* Optimizing and benchmarking de novo transcriptome sequencing:
579 from library preparation to assembly evaluation. *BMC Genomics* **16**, 977 (2015).

580 54 Huang, S. *et al.* Decelerated genome evolution in modern vertebrates revealed by
581 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 antero-posterior regionalization in avian embryos. *Mech. Dev.*
604 **120**, 227-240 (2003).

605 65 Amemiya, C. T. *et al.* The amphioxus *Hox* cluster: characterization, comparative
606 genomics, and evolution. *J. Exp. Zool. B Mol. Dev. Evol.* **310**, 465-477 (2008).

607 66 Holland, L. Z. *et al.* The amphioxus genome illuminates vertebrate origins and
608 cephalochordate biology. *Genome Res.* **18**, 1100-1111 (2008).

609 67 Pascual-Anaya, J., D'Aniello, S. & Garcia-Fernàndez, J. Unexpectedly large number
610 of conserved noncoding regions within the ancestral chordate *Hox* cluster. *Dev.*
611 *Genes Evol.* **218**, 591-597 (2008).

612 68 Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version
613 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772-780
614 (2013).

615 69 Castresana, J. Selection of conserved blocks from multiple alignments for their
616 use in phylogenetic analysis. *Mol. Biol. Evol.* **17**, 540-552 (2000).

617 70 Hall, T. A. BioEdit: a user-friendly biological sequence alignment editor and
618 analysis program for Windows 95/98/NT. *Nucleic Acids Symp. Ser.* **41**, 95-98
619 (1999).

620 71 Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-
621 analysis of large phylogenies. *Bioinformatics* **30**, 1312-1313 (2014).

622 72 Li, B. & Dewey, C. N. RSEM: accurate transcript quantification from RNA-Seq data
623 with or without a reference genome. *BMC Bioinformatics* **12**, 323 (2011).

624 73 gplots: Various R programming tools for plotting data. R package version 3.0.1
625 (2016).

626 74 RStudio: Integrated Development for R. (R. RStudio, Inc., Boston, MA, 2016).

627 75 R: A language and environment for statistical computing (Vienna, Austria, 2016).

628 76 Nicolae, M. & Măndoiu, I. in *Bioinformatics Research and Applications: 7th*
629 *International Symposium, ISBRA 2011, Changsha, China, May 27-29, 2011.*
630 *Proceedings* (eds Jianer Chen, Jianxin Wang, & Alexander Zelikovsky) 392-403
631 (Springer Berlin Heidelberg, 2011).

632

633 **Acknowledgments**

634 We thank Y. Oisi and S. Fujimoto for providing preliminary hagfish Hox sequences; J.M.
635 Martín-Durán, I. Maeso, M. Irimia and C. Böhmer for fruitful discussions; O. Kakitani for
636 hagfish sampling; K. Shirato for shark sampling; S. Shibuya and K. Yamamoto for
637 maintenance of aquarium tanks; S. Kuraku, K. Itomi, C. Tanegashima, K. Tatsumi and O.
638 Nishimura from the Phyloinformatics Unit, RIKEN CLST, for RNA-seq data production; J.
639 Huddleston and E. Eichler for providing the code to mask BAC vector sequences from
640 PacBio reads; I. Mandoiu for his help using DGE-EM software; and B. Chevreaux for his
641 help with the MIRA assembler. This work was supported by the Chinese Academy of
642 Sciences program XDB13000000 to W.W., and by a Grant-in-Aid for Scientific Research (A)
643 15H02416 (Japan Society for the Promotion of Science), a Grant-in-Aid for Scientific
644 Research on Innovative Areas (Research in a Proposed Research Area) 17H06384 (Ministry
645 of Education, Culture, Sports, Science and Technology of Japan), and a Naito Grant for the
646 Promotion of Focused Research (The Naito Foundation) to S.K.

647

648 **Author contributions**

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

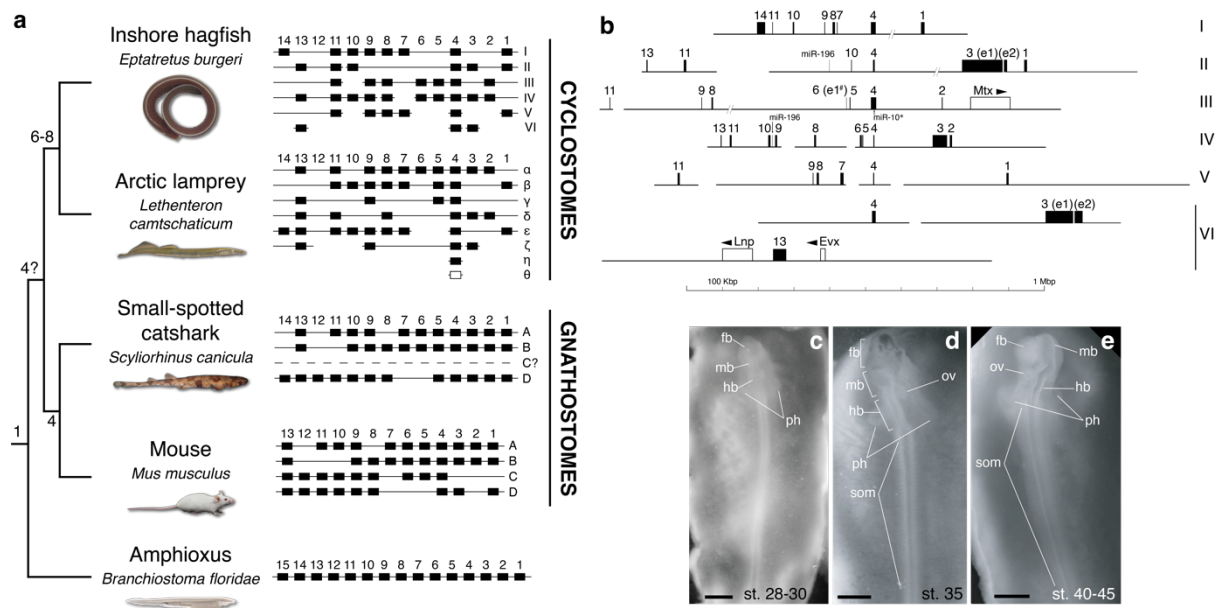
652 assembled the *E. burgeri* draft genome. J.P. performed the phylogenetic analyses. All authors
653 analysed and discussed the data. All authors approved the final version of the manuscript.

654

655 **Competing interests**

656 The authors declare no competing financial interests.

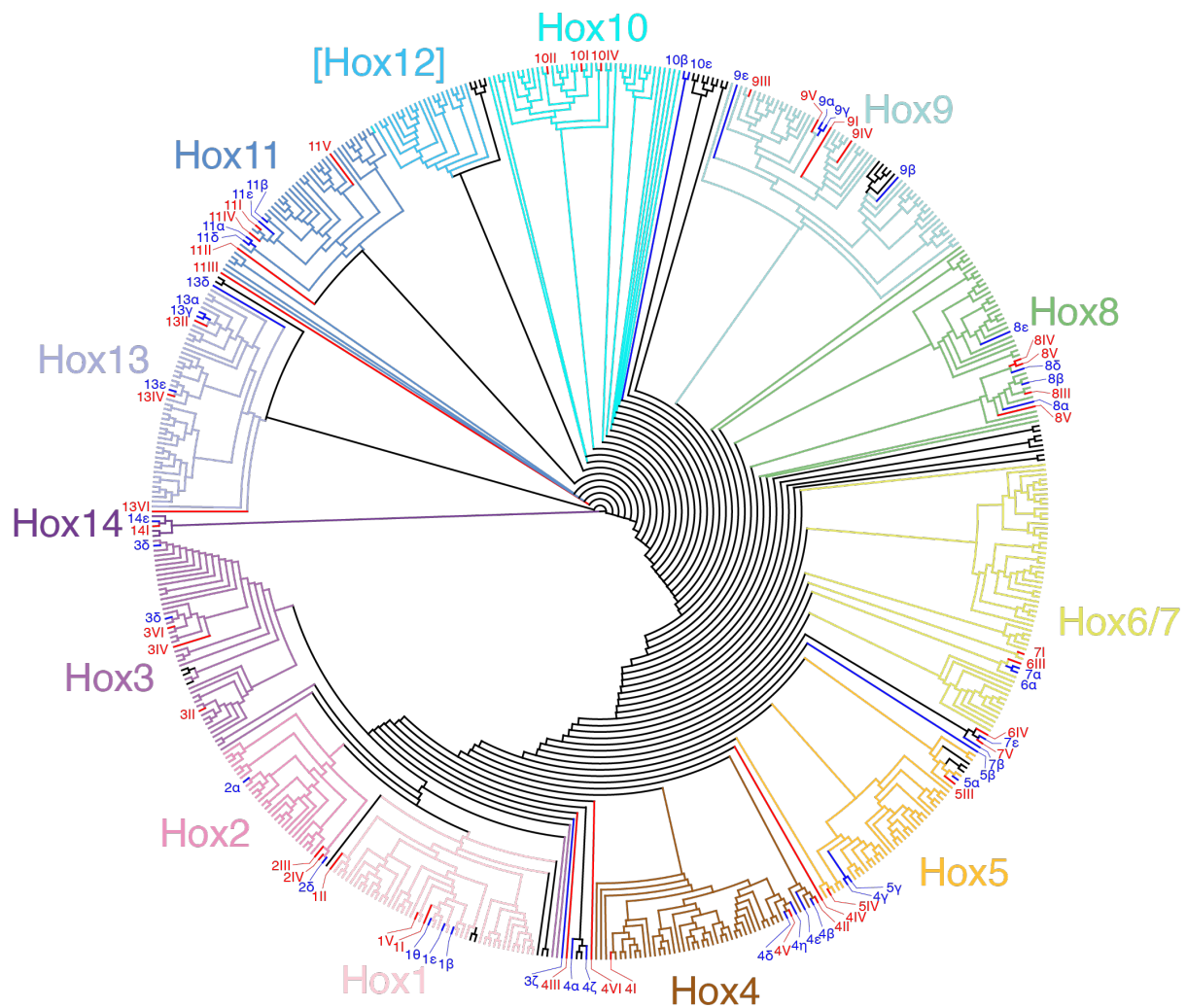
657



658

659 **Figure 1. Hox cluster evolution in chordates.** a, Phylogenetic tree of chordates, showing
 660 the two major groups of vertebrates –cyclostomes (hagfish and lamprey) and gnathostomes
 661 (jawed vertebrates, e.g., mouse and shark) – together with cephalochordates (amphioxus),
 662 displaying their known Hox repertoires. Numbers on the nodes indicate the putative number
 663 of Hox clusters in each last common ancestor. b, *E. burgeri* Hox genes and clusters found in
 664 this study, drawn to scale. All Hox genes are transcribed in the same orientation, from left to
 665 right. Orientation of transcription of non-Hox syntenic genes are indicated by arrowheads.
 666 Solid horizontal lines correspond to single scaffolds. Double diagonal lines separate two
 667 contiguous scaffolds, based on BAC sequences connecting them (Supplementary Fig. 2).
 668 *Hox3III* and *Hox3VI* genes have corresponding exons 1 and 2 in two different scaffolds, which
 669 have been put together based on both BAC and transcriptomics evidences. e1, exon 1; e2,
 670 exon 2. Asterisk over *miR-10III* indicates that this microRNA is within a 5'UTR intron of
 671 *Hox4III*; hash symbol, *Hox6III* exon2 is not assembled in place, but in a separate small contig
 672 containing its sequence. c-e, *E. burgeri* embryos used for the transcriptomics analysis, at
 673 Dean stages 28-30 (c), 35 (d) and 40-45 (e). fb, forebrain; hb, hindbrain; mb, midbrain; ov,
 674 otic vesicle; ph, pharynx; som, somites. Scale bars, 1 mm.

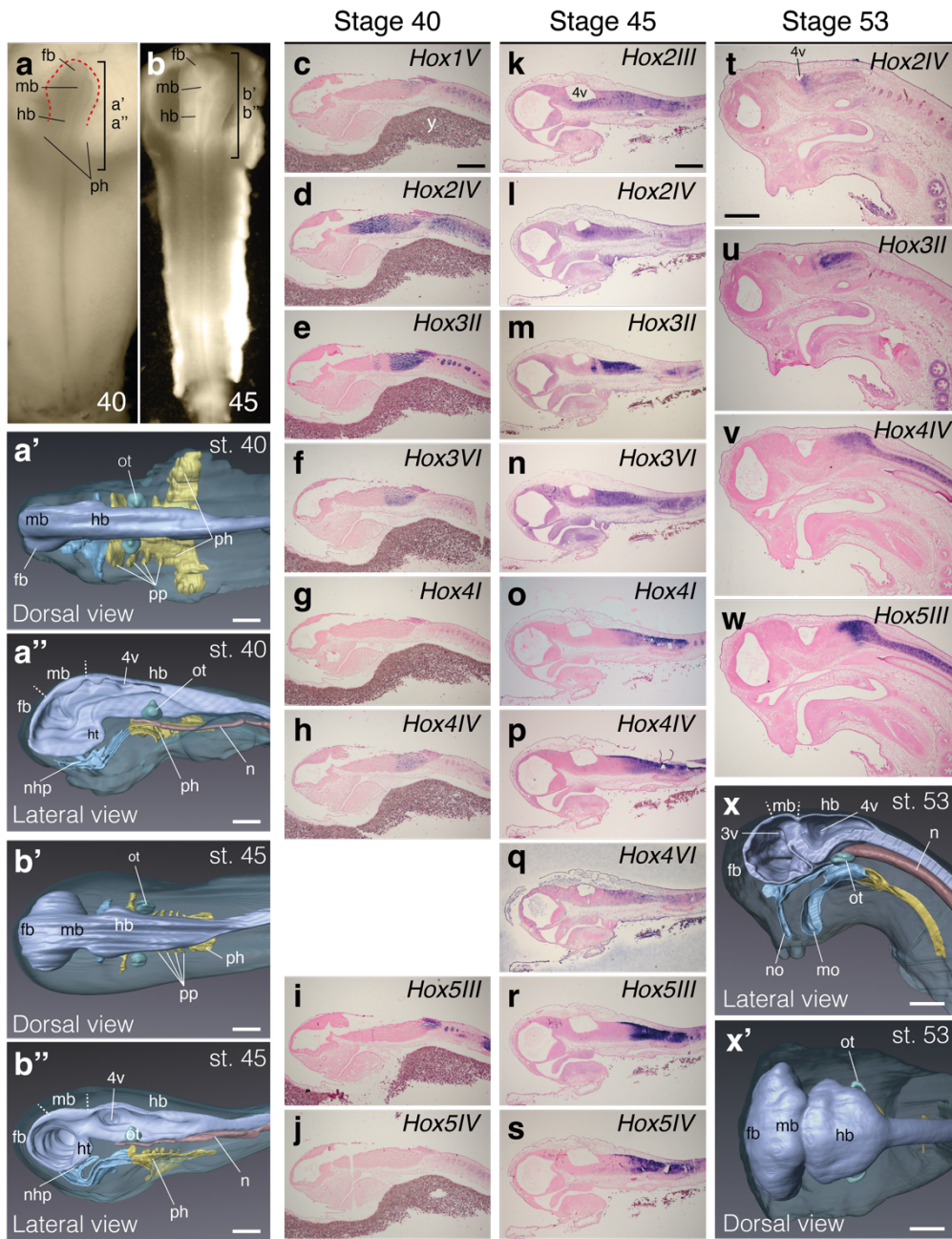
675



676

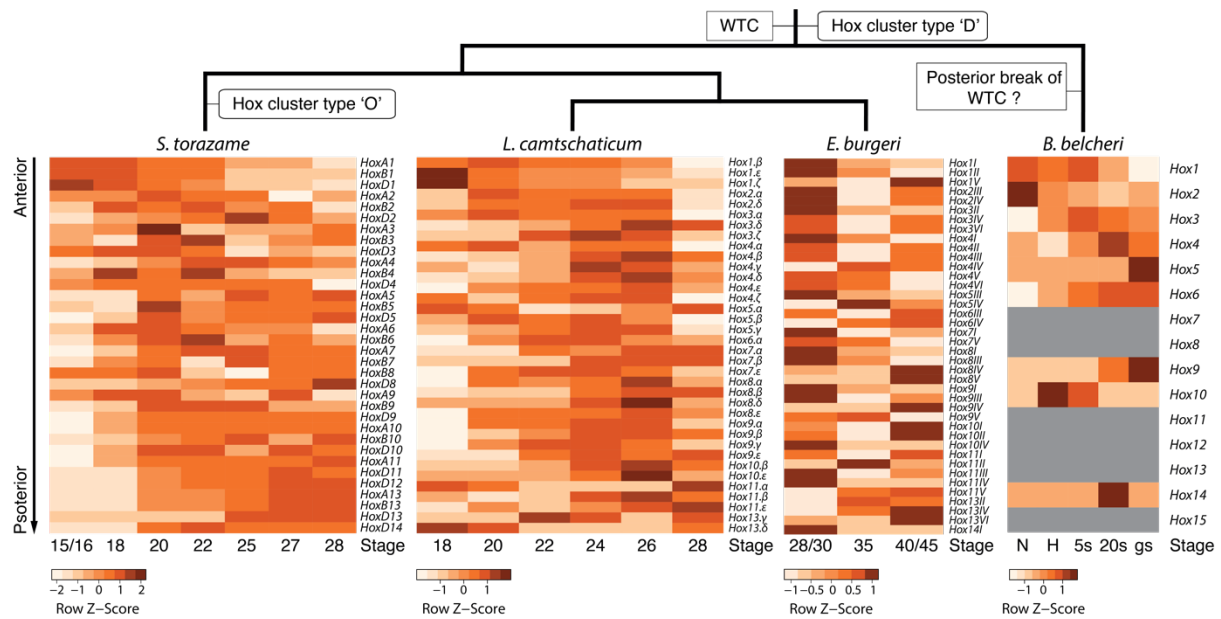
677 **Figure 2. Molecular phylogenetic tree of vertebrate Hox genes.** 1000-replicate Maximum
 678 Likelihood tree of representative Hox genes of all paralogy groups in vertebrates. The
 679 branches have been color-coded by paralogy group (Hox1-14). Red and blue branches denote
 680 *E. burgeri* and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

686



687

688 **Figure 3. Spatial colinearity of hagfish *Hox* genes in the hindbrain of *E. burgeri***
689 **embryos. a, b**, Embryos at stage Bashford Dean 40 (**a**) and 45 (**b**) used for *in situ*
690 hybridizations on sections. The inset square brackets mark the head regions, used for sagittal
691 sectioning. **a', a'', b', b''**, 3D Avizo reconstructions of the heads of the embryos shown in **a**
692 and **b**, respectively, showing the main internal anatomy of the brain and main head structures.
693 The central nervous systems are in purple; ectoderm is in light blue; endoderm is in yellow;
694 otic vesicle in green; and notochord is in light red. These embryos are the source of the
695 sections shown in **c-j** (stage 40) and **k-s** (stage 45). **c-w**, Spatial colinearity displayed by
696 expression patterns of *E. burgeri* *Hox1IV* (**c**), *Hox2III* (**k**), *Hox2IV* (**d, l, t**), *Hox3II* (**e, m, u**),
697 *Hox3VI* (**f, n**), *Hox4I* (**g, o**), *Hox4IV* (**h, p, v**), *Hox4VI* (**q**), *Hox5III* (**i, r, w**) and *Hox5IV* (**j, s**),
698 revealed by *in situ* hybridization on sections of an embryos shown in **a, b** and **x, x'**, 3D
699 Avizo reconstructions of the head of an embryo at stage Bashford Dean 53. The central
700 nervous systems are in purple; ectoderm is in light blue; endoderm is in yellow; otic vesicle
701 in green; and notochord is in light red. **y**, Expression patterns shown in (**c-j**) in the hindbrain
702 are aligned according to rhombomere (**r**) segmentation, showing nested expression patterns of
703 *Hox1-5* paralogs in the hagfish hindbrain. **z**, Schematic diagrams summarizing the expression
704 patterns of *Hox* genes in the hindbrain of the lamprey, the hagfish and mouse, with nested
705 anterior limits coinciding with rhombomere borders, and showing the overall conservation
706 among the groups. f, forebrain; h, hindbrain; ht, hypothalamus; m, midbrain; mo, mouth; n,
707 notochord; no, nasal opening; nhp, naso-hypophyseal plate; ot, otic vesicle; ph, pharynx; pp,
708 pharyngeal pouches; y, yolk; 3v, 4v, third and fourth ventricles, respectively. Scale bars, 0.5
709 mm. Asterisk indicates a different expression has been found in a separate species, *P.*
710 *marinus*, in which *Pm1Hox3* rostral limit is on the r4/r5 border²⁸.
711



712

713 **Figure 4. Developmental Hox genes expression profiling of Hox genes in chordates.** Heatmaps of
 714 Hox genes expression in *S. torazame* (gnathostome), *L. camtschaticum* and *E. burgeri*
 715 (agnathans), and *B. belcheri* (invertebrate chordate), coloured according to Z-score (standard
 716 deviations from mean expression level). Anterior Hox genes (top rows of heatmaps) tend to
 717 be expressed at higher levels at early stages of development than posterior genes (bottom
 718 rows of heatmaps) in both *S. torazame* and *L. camtschaticum*. On top, a phylogenetic tree
 719 with chordate relationships of the species studied here indicate the putative events that took
 720 place during evolution: in *B. belcheri*, temporal colinearity is appreciated between Hox1-5
 721 genes, indicating WTC was likely present in the last common ancestor of chordates, and a
 722 secondary escape of the posterior half of the cluster from it occurred independently in the
 723 amphioxus lineage. The large sizes of both amphioxus and agnathan Hox clusters implies that
 724 the common ancestor of vertebrates had a so-called ‘disorganized’ (D) cluster type, while the
 725 consolidation towards an ‘organized’ (O) type occurred in the gnathostome lineage⁴², after
 726 the split between jawed and jawless vertebrates. In *B. belcheri*, grey rows indicate genes with
 727 a FPKM value of 0 in all stages. N, amphioxus neurula stage; H, hatching stage; 5s, 5-somite
 728 stage; 20s, 20-somite stage; gs, 1- or 2-gill slit larvae.