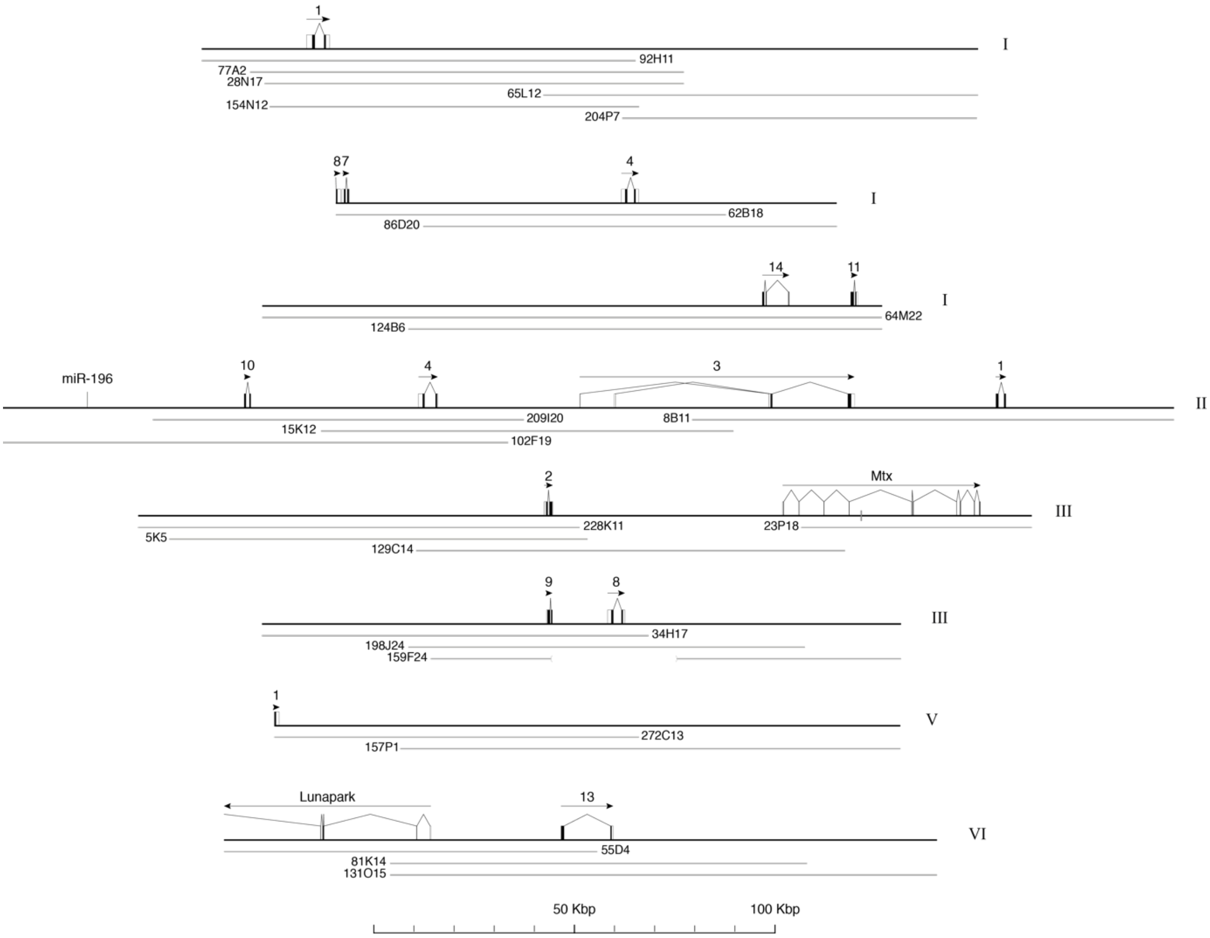
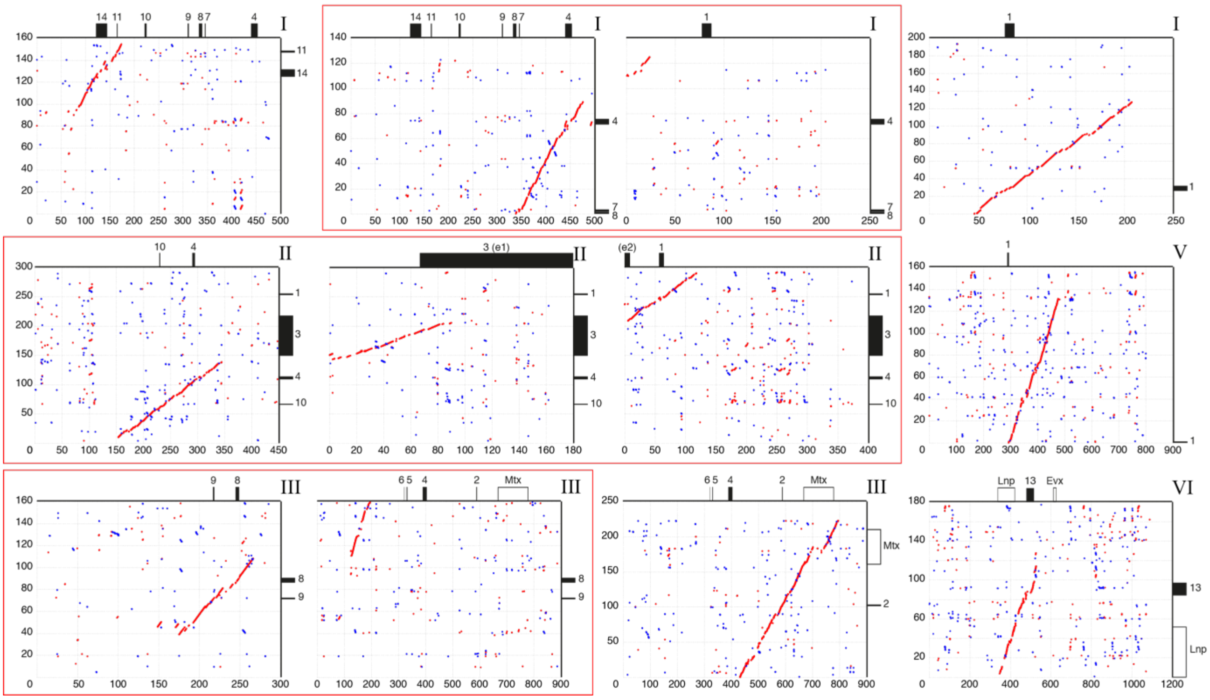
Hagfish and lamprey Hox genes reveal conservation of temporal colinearity in vertebrates

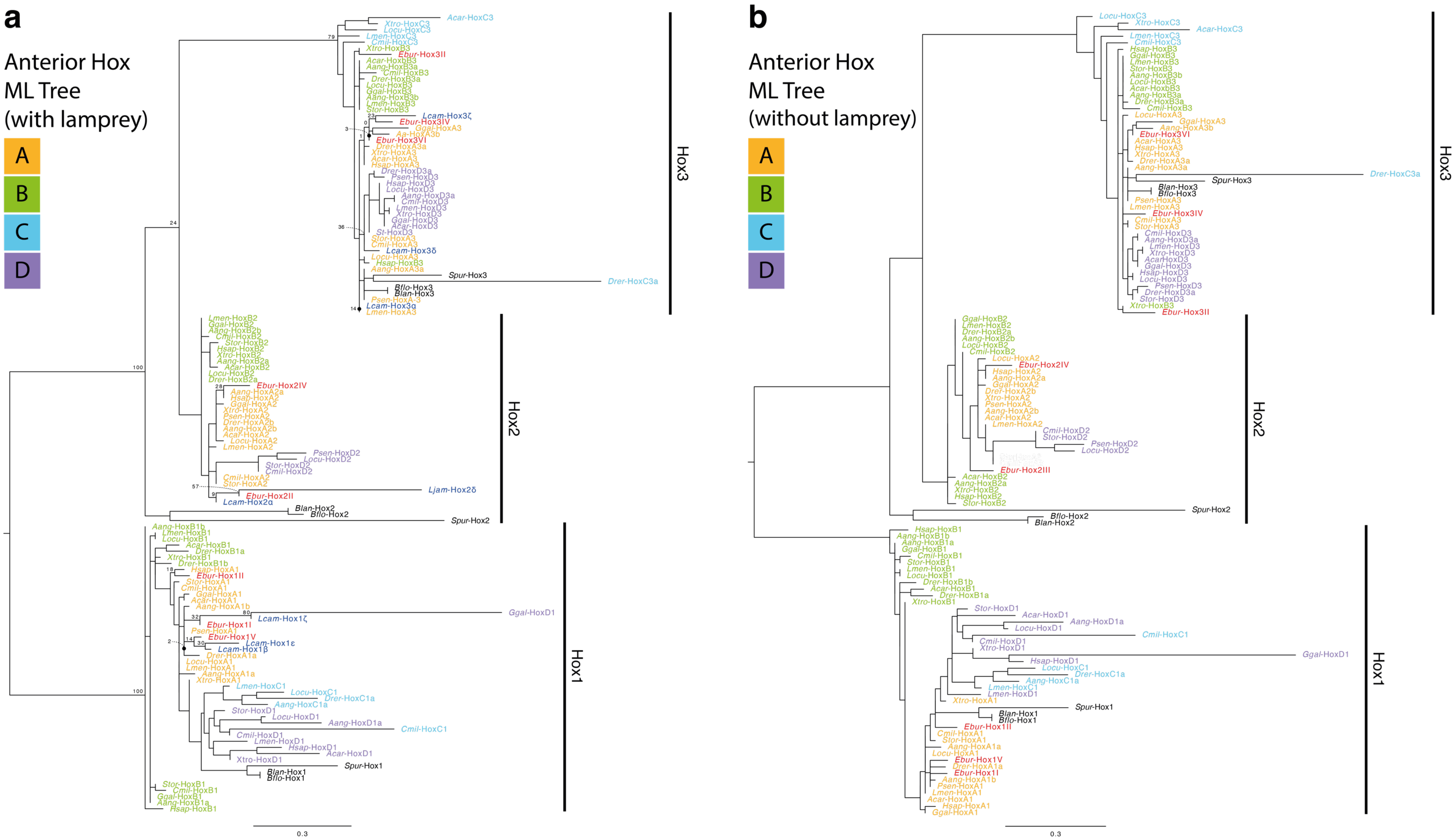
Juan Pascual-Anaya, Iori Sato, Fumiaki Sugahara, Shinnosuke Higuchi, Jordi Paps, Ren Yandong, Wataru Takagi, Adrián Ruiz-Villalba, Kinya G. Ota, Wen Wang, Shigeru Kuratani



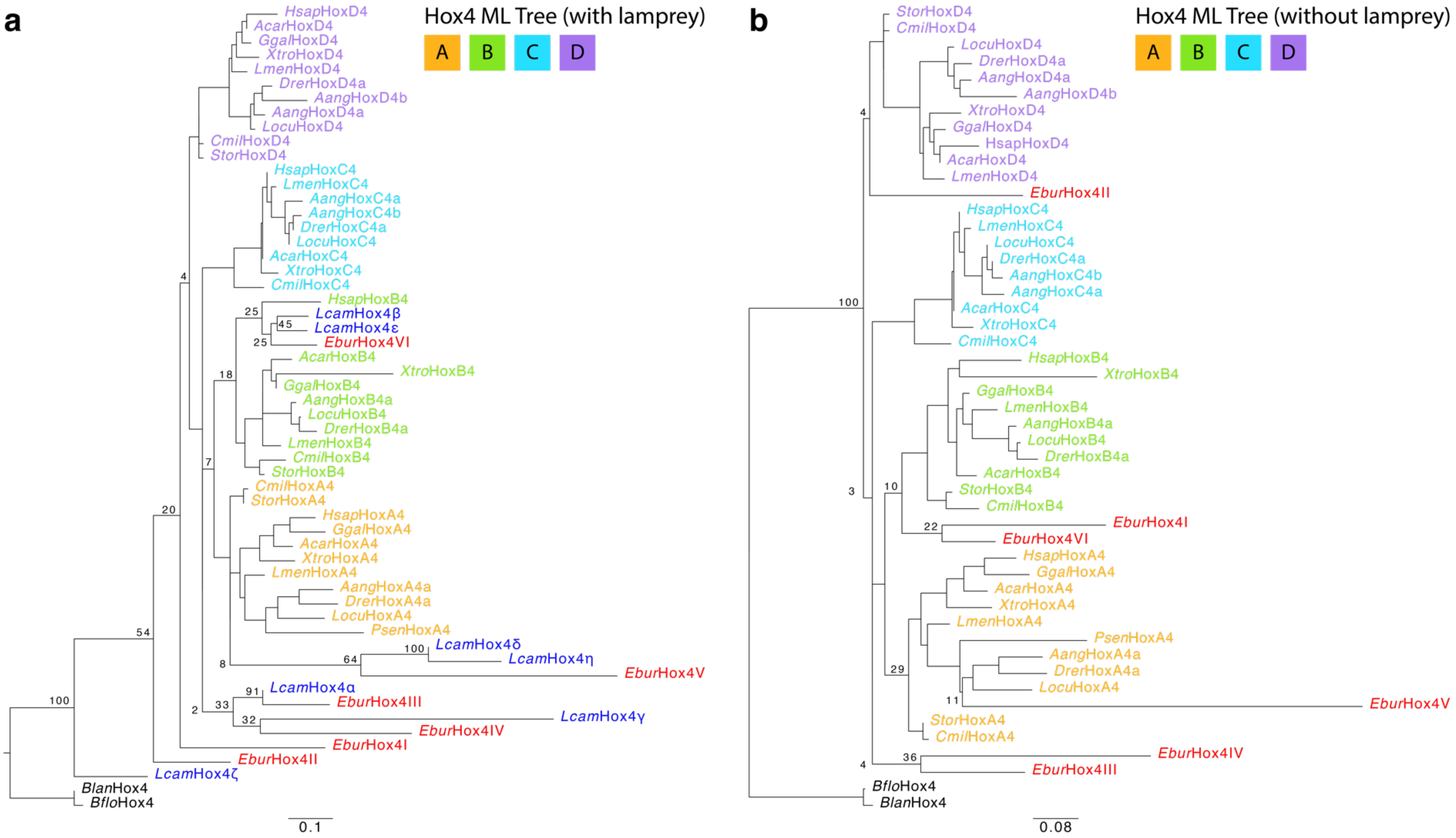
Supplementary Figure 1. BAC clones found in this study. The BAC clones sequenced are drawn to scale. Exons are indicated by boxes, of which filled and open ones indicate coding and unstranslated sequences, respectively. Introns are indicated by diagonal lines on top of adjacent exons. Arrows indicate the orientation of transcription. BAC 159F24 was found to contain a gap with respect to BACs 34H17 and 198J24. This was shown to be an artefact, problably introduced during BAC library construction, and not a genuine somatic genome rearrangement event (data now shown).



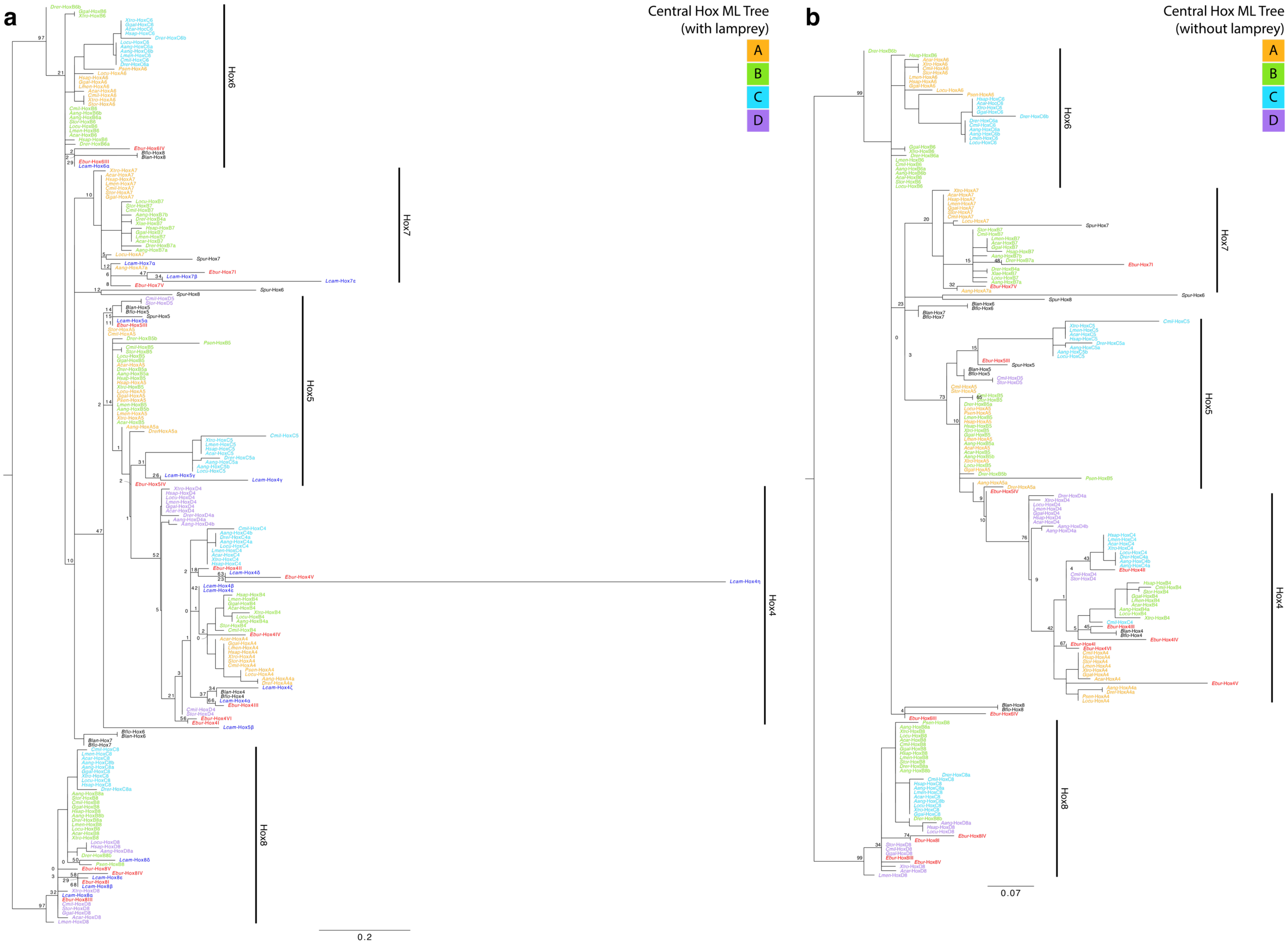
Supplementary Figure 2. Dotplots of BAC versus genome scaffold sequences. BAC clone sequences (y axis) were aligned against corresponding genomic regions (x axis) using MUMmer. Red and blue dots indicate alignments in the same or reverse orientation, respectively. Hox and non-Hox genes are indicated by filled and open boxes, respectively. Those scaffolds that can aligned with the same BAC (i.e., a single BAC clone aligning against contiguous scaffolds) are outlined by a red box. Axis indicate BAC and scaffold sizes in Kbp.



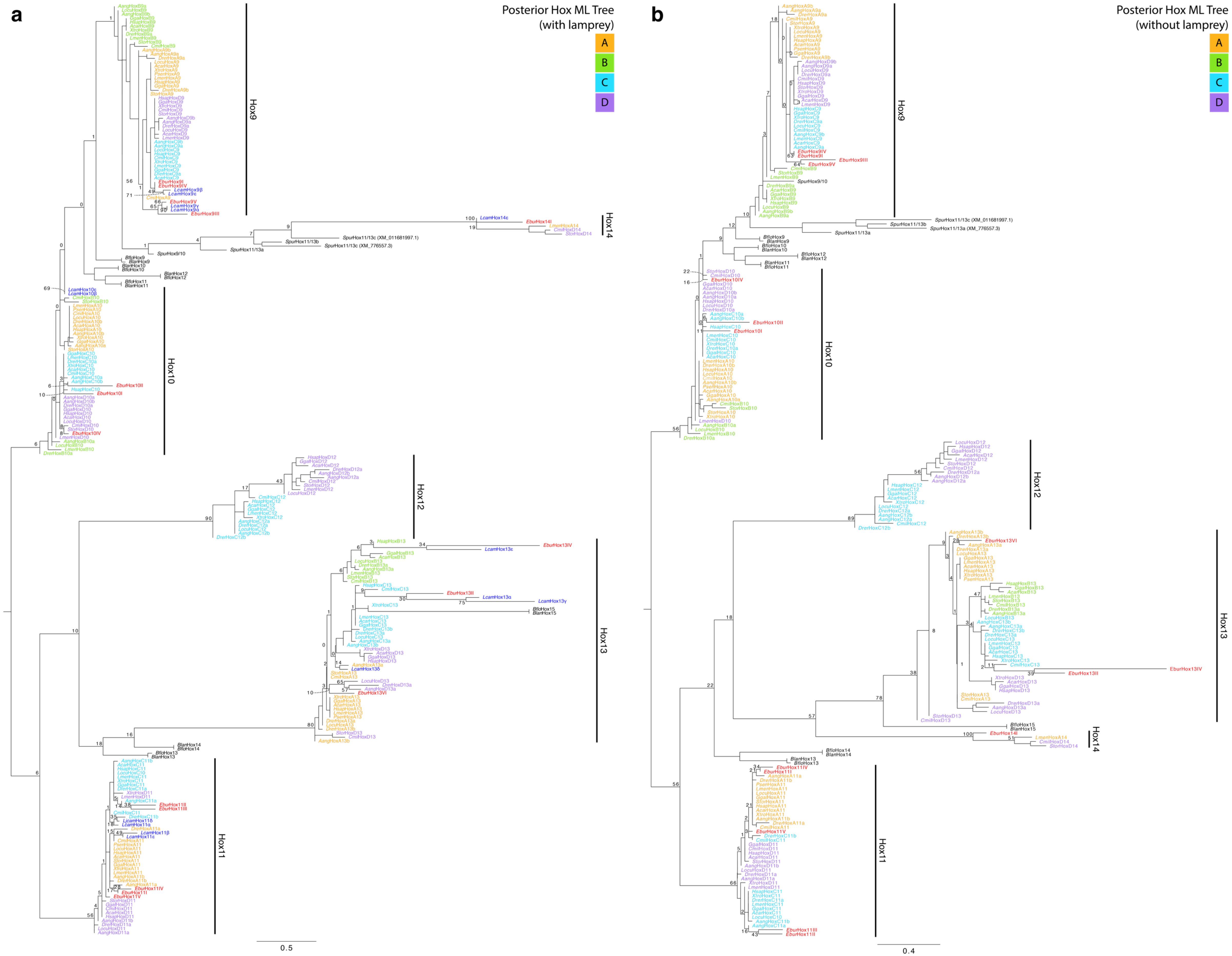
Supplementary Figure 3. RAxML trees (LG + Gamma + Invariants, 1000 bootstraps) of Hox genes from the anterior paralogy groups (Hox1-3). Hox genes from representative vertebrate species of all A (orange), B (green), C (light blue) and D (purple) Hox cluster types, including *S. torazame* sequences described in this study, were selected and analysed together with *E. burgeri* Hox sequences (in red), and including (**a**) or excluding (**b**) *L. camtschaticum* sequences (in blue). Hox1-3 gene sequences from amphioxus (*B. floridae* and *B. lanceolatum*) and sea urchin (*S. purpuratus*) were included in the analysis. *Hsap*, *Homo sapiens*; *Ggal*, *Gallus gallus*; *Acar*, *Anolis carolinensis*; *Xtro*, *Xenopus tropicalis* (*Xlae*, *Xenopus laevis*); *Lmen*, *Latimeria menadoensis*; *Drer*, *Danio rerio*; *Aang*, *Anguilla anguilla*; *Locu*, *Lepidosteus oculatus*; *Psen*, *Polypterus senegalensis*; *Cmil*, *Callorhinchus milii*; *Stor*, *Scyliorhinus torazame*; *Lcam*, *Lethenteron camtschaticum*; *Ebur*, *Eptatretus burgeri*; *Bflo*, *Branchiostoma floridae*; *Blan*, *Branchiostoma lanceaolatum*; *Spur*, *Strongylocentrotus purpuratus*.



Supplementary Figure 4. RAxML trees (LG + Gamma + Invariants, 1000 bootstraps) of Hox4 genes. Hox4 genes sequences from representative vertebrate species of all A (orange), B (green), C (light blue) and D (purple) Hox cluster types, including *S. torazame* sequences described in this study, were selected and analysed together with *E. burgeri* Hox4 sequences (in red), and including (**a**) or excluding (**b**) *L. camtschaticum* sequences (in blue). Hox4 gene sequences from amphioxus (*B. floridae* and *B. lanceolatum*) were included in the analysis. *Hsap*, *Homo sapiens*; *Ggal*, *Gallus gallus*; *Acar*, *Anolis carolinensis*; *Xtro*, *Xenopus tropicalis* (*Xlae*, *Xenopus laevis*); *Lmen*, *Latimeria menadoensis*; *Drer*, *Danio rerio*; *Aang*, *Anguilla anguilla*; *Locu*, *Lepidosteus oculatus*; *Psen*, *Polypterus senegalensis*; *Cmil*, *Callorhinchus milii*; *Stor*, *Scyliorhinus torazame*; *Lcam*, *Lethenteron camtschaticum*; *Ebur*, *Eptatretus burgeri*; *Bflo*, *Branchiostoma floridae*; *Blan*, *Branchiostoma lanceaolatum*.


a
Central Hox ML Tree
(with lamprey)
A
B
C
D
Hox6
Hox7
Hox5
Hox4
Hox8
0.2
b
Central Hox ML Tree
(without lamprey)
A
B
C
D
Hox6
Hox7
Hox5
Hox4
Hox8
0.07

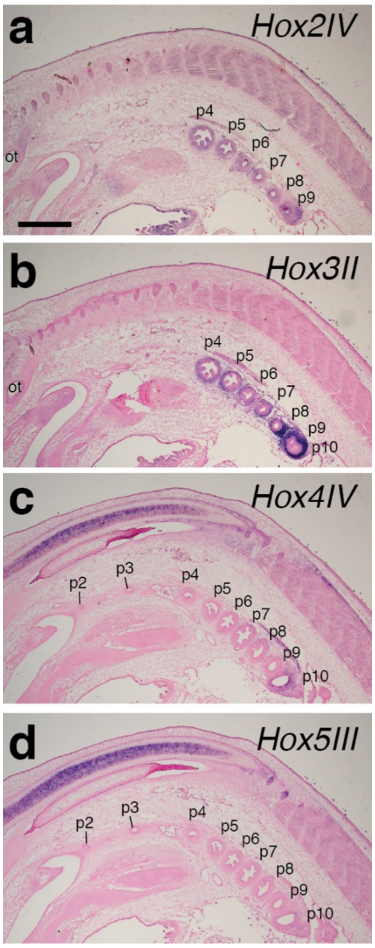
Supplementary Figure 5. RAxML trees (LG + Gamma + Invariants, 1000 bootstraps) of Hox genes from the central paralogy groups (Hox4-8). Hox genes from representative vertebrate species of all A (orange), B (green), C (light blue) and D (purple) Hox cluster types, including *S. torazame* sequences described in this study, were selected and analysed together with *E. burgeri* Hox sequences (in red), and including (**a**) or excluding (**b**) *L. camtschaticum* sequences (in blue). Hox4-8 gene sequences from amphioxus (*B. floridae* and *B. lanceolatum*) and sea urchin (*S. purpuratus*) were included in the analysis. *Hsap*, *Homo sapiens*; *Ggal*, *Gallus gallus*; *Acar*, *Anolis carolinensis*; *Xtro*, *Xenopus tropicalis* (*Xlae*, *Xenopus laevis*); *Lmen*, *Latimeria menadoensis*; *Drer*, *Danio rerio*; *Aang*, *Anguilla anguilla*; *Locu*, *Lepidosteus oculatus*; *Psen*, *Polypterus senegalensis*; *Cmil*, *Callorhinchus milii*; *Stor*, *Scyliorhinus torazame*; *Lcam*, *Lethenteron camtschaticum*; *Ebur*, *Eptatretus burgeri*; *Bflo*, *Branchiostoma floridae*; *Blan*, *Branchiostoma lanceaolatum*; *Spur*, *Strongylocentrotus purpuratus*.


a
Posterior Hox ML Tree
(with lamprey)
A
B
C
D
Hox9
Hox14
Hox10
Hox12
Hox13
Hox11
0.5
b
Posterior Hox ML Tree
(without lamprey)
A
B
C
D
Hox9
Hox10
Hox12
Hox13
Hox14
Hox11
0.4

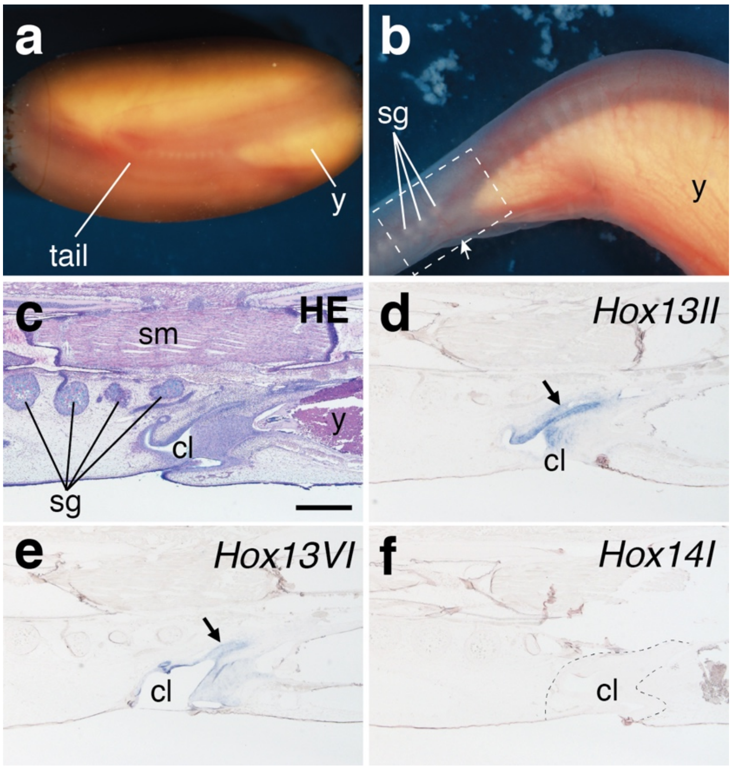
Supplementary Figure 6. RAxML trees (LG + Gamma + Invariants, 1000 bootstraps) of Hox genes from the posterior paralogy groups (Hox9-14). Hox genes from representative vertebrate species of all A (orange), B (green), C (light blue) and D (purple) Hox cluster types, including *S. torazame* sequences described in this study, were selected and analysed together with *E. burgeri* Hox sequences (in red), and including (**a**) or excluding (**b**) *L. camtschaticum* sequences (in blue). Hox9-15 gene sequences from amphioxus (*B. floridae* and *B. lanceolatum*) and Hox9/10 and Hox11/13a,b and c from sea urchin (*S. purpuratus*) were included in the analysis. *Hsap*, *Homo sapiens*; *Ggal*, *Gallus gallus*; *Acar*, *Anolis carolinensis*; *Xtro*, *Xenopus tropicalis* (*Xlae*, *Xenopus laevis*); *Lmen*, *Latimeria menadoensis*; *Drer*, *Danio rerio*; *Aang*, *Anguilla anguilla*; *Locu*, *Lepidosteus oculatus*; *Psen*, *Polypterus senegalensis*; *Cmil*, *Callorhinchus milii*; *Stor*, *Scyliorhinus torazame*; *Lcam*, *Lethenteron camtschaticum*; *Ebur*, *Eptatretus burgeri*; *Bflo*, *Branchiostoma floridae*; *Blan*, *Branchiostoma lanceaolatum*; *Spur*, *Strongylocentrotus purpuratus*


Hox1
Hox2
Hox3
Hox4
Hox5
Hox6/7
Hox8
Hox9
Hox10
Hox12
Hox11
Hox13
Hox14
2.0

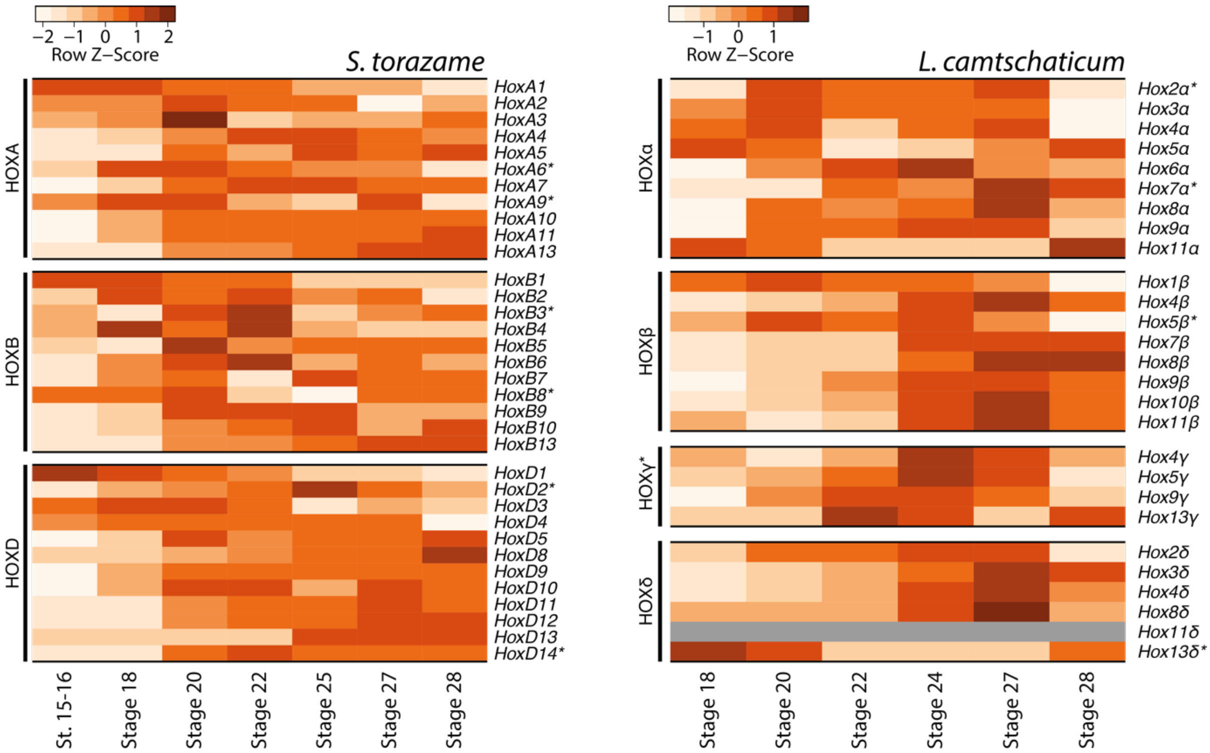
Supplementary Figure 7. RAxML trees (LG + Gamma + Invariants, 1000 bootstraps) of all Hox genes included in Supplementary Figs. 3-6. Hox genes from representative vertebrate species of all A , B, C and D Hox cluster types, including *S. torazame* sequences described in this study, were selected and analysed together with *E. burgeri* Hox sequences (in red) and *L. camtschaticum* sequences (in blue). Hox gene sequences from amphioxus (*B. floridae* and *B. lanceolatum*) and sea urchin (*S. purpuratus*) were included in the analysis. Note that no lamprey or hagfish Hox gene fall within the Hox12 group (open arrow). *Hsap*, *Homo sapiens*; *Ggal*, *Gallus gallus*; *Acar*, *Anolis carolinensis*; *Xtro*, *Xenopus tropicalis* (*Xlae*, *Xenopus laevis*); *Lmen*, *Latimeria menadoensis*; *Drer*, *Danio rerio*; *Aang*, *Anguilla anguilla*; *Locu*, *Lepidosteus oculatus*; *Psen*, *Polypterus senegalensis*; *Cmil*, *Callorhinchus milii*; *Stor*, *Scyliorhinus torazame*; *Lcam Lethenteron camtschaticum*; *Ebur*, *Eptatretus burgeri*; *Bflo*, *Branchiostoma floridae*; *Blan*, *Branchiostoma lanceaolatum*; *Spur*, *Strongylocentrotus purpuratus*



Supplementary Figure 8. Expression of *Hox2-5* paralogs in the pharynx. **a**-**d**, Colinear expression patterns of *Hox2IV* (**a**), *Hox3II* (**b**), *Hox4IV* (**c**) and *Hox5III* (**d**) in the pharyngeal pouches 4-10.



Supplementary Figure 9. Expression of *Hox13* paralogs in the cloaca. **a**, A hagfish egg containing a juvenile at stage 60. **b**, Juvenile extracted from the egg shown in **a**. The white arrowhead indicates the position of the cloaca. The white dashed square marks the area shown in panels **c**-**f**. (**c**) Sagittal section of an embryo at stage 60 stained with haematoxylin and eosin (HE) showing the histology of the cloacal region. **d**-**f**, Expression patterns of *Hox13II* (**d**), *Hox13VI* (**e**) and *Hox14I* (**f**) in a juvenile at stage 60. Hox13 genes, but not *Hox14I*, are expressed in the cloaca (black arrows). cl, cloaca; sg, slime glands; sm, skeletal muscle; y, yolk. Scale bar, 0.5 mm. In **b**-**f**, anterior is to the right, dorsal to the top. In **a**, the juvenile is coiled around itself, but dorsal is to the top, and the tail shows is posterior tip towards the left.



Supplementary Figure 10. Expression profiles of *Hox* genes per cluster during development of *S. torazame* and *L. camtschaticum*. Heatmaps of *Hox* genes expression in *S. torazame* (gnathostome) and *L. camtschaticum* 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 different clusters of both *S. torazame* and *L. camtschaticum*. Genes that escape collineatity, either by having a peak of expression earlier or later than expected are indicated by an asterisk. Lamprey HOX-γ cluster, comprised of only four Hox genes (largely degenerated), do not follow temporal collinearity. Lamprey *Hox11δ* is not found in our lamprey transcriptome assembly (grey row).

Supplementary Table 1. CEGMA and BUSCO completeness assessment results of *E. burgeri* transcriptome assemblies

Assembly	CEG-based CEGMA (248 proteins)				CVG-based CEGMA (233 proteins)				Vertebrata-based BUSCO (3023 proteins)				CVG-based BUSCO (233 proteins)			
	Complete		Partial		Complete		Partial		Complete		Fragmented		Complete		Fragmented	
	# Proteins	%	# Proteins	%	# Proteins	%	# Proteins	%	# Proteins	%	# Proteins	%	# Proteins	%	# Proteins	%
1	**246**	**99.19**	246	99.19	211	90.56	219	93.99	1537	50.84	1617	53.49	**215**	**92.27**	**221**	**94.85**
2	242	97.58	246	99.19	198	84.98	219	93.99	1532	50.68	1629	53.89	210	90.13	**221**	**94.85**
3	**246**	**99.19**	**247**	**99.60**	**214**	**91.85**	**220**	**94.42**	**1605**	**53.09**	**1674**	**55.38**	214	91.85	**221**	**94.85**
4	228	91.94	244	98.39	185	79.40	216	92.70	1400	46.31	1521	50.31	201	86.27	220	94.42

1, All-in-one, Trinity; 2, All-in-one, digital normalization within Trinity; 3, Integration of each embryo's Trinity-transcriptome with CD-HIT-EST; 4, SOAPdenovoTrans.

Note: Higher values in bold font.

Supplementary Table 2. CEGMA and BUSCO completeness assessment results of *L. camtschaticum* transcriptome assemblies

Assembly	CEG-based CEGMA (248 proteins)				CVG-based CEGMA (233 proteins)				Vertebrata-based BUSCO (3023 proteins)				CVG-based BUSCO (233 proteins)			
	Complete		Partial		Complete		Partial		Complete		Fragmented		Complete		Fragmented	
	# Proteins	%	# Proteins	%	# Proteins	%	# Proteins	%	# Proteins	%	# Proteins	%	# Proteins	%	# Proteins	%
1	234	94.35	243	97.98	186	79.83	216	92.70	1507	49.85	1677	55.47	214	91.85	**230**	**98.71**
2	**238**	**95.97**	243	97.98	184	78.97	218	93.56	1486	49.16	1658	54.85	212	90.99	229	98.28
3	231	93.15	244	98.39	189	81.12	219	93.99	1506	49.82	1690	55.90	214	91.85	228	97.85
4	235	94.76	**246**	**99.19**	**195**	**83.69**	**219**	**93.99**	**1590**	**52.60**	**1755**	**58.05**	**218**	**93.56**	**230**	**98.71**

1, Assembly of each embryo's data with Trinity (strand-specific information) and integration with CD-HIT-EST; 2, idem, but not taking into account strand-specific information; 3, All-in-one, Trinity, strand-specific; 4, Trinity+PASA

Note: Higher values in bold font.

Supplementary Table 3. CEGMA and BUSCO completeness assessment results of *S. torazame* transcriptome assemblies

Assembly	CEG-based CEGMA (248 proteins)				CVG-based CEGMA (233 proteins)				Vertebrata-based BUSCO (3023 proteins)				CVG-based BUSCO (233 proteins)			
	Complete		Partial		Complete		Partial		Complete		Fragmented		Complete		Fragmented	
	# Proteins	%	# Proteins	%	# Proteins	%	# Proteins	%	# Proteins	%	# Proteins	%	# Proteins	%	# Proteins	%
1	245	98.79	245	98.79	225	96.57	231	99.14	2586	85.54	**2679**	**88.62**	226	97.00	**231**	**99.14**
2	246	99.19	246	99.19	224	96.14	231	99.14	2569	84.98	2664	88.12	225	96.57	**231**	**99.14**
3	**247**	**99.60**	**247**	**99.60**	**227**	**97.42**	**232**	**99.57**	**2596**	**85.87**	2672	88.39	**227**	**97.42**	**231**	**99.14**

1, Assembly of each embryo's data with Trinity (strand-specific information) and integration with CD-HIT-EST; 2, idem, but not taking into account strand-specific information; 3, All-in-one, Trinity, strand-specific.

Note: Higher values in bold font.