A single three-dimensional chromatin compartment in amphioxus indicates a stepwise evolution of vertebrate Hox bimodal regulation

Rafael D. Acemel[1, 5], Juan J. Tena[1, 5], Ibai Irastorza[1, 5], Ferdinand Marlétaz[2, 5], Carlos Gómez-Marín[1], Elisa de la Calle-Mustienes[1], Stéphanie Bertrand[3], Sergio G. Diaz[1], Daniel Aldea[3], Jean-Marc Aury[4], Sophie Mangenot[4], Peter W. H. Holland[2], Damien P. Devos[1, 6], Ignacio Maeso[1, 6], Hector Escrivá[3, 6], José Luis Gómez-Skarmeta[1, 6]

[1] Centro Andaluz de Biología del Desarrollo (CABD), Consejo Superior de Investigaciones Científicas/Universidad Pablo de Olavide, Sevilla, Spain.

[2] Department of Zoology, University of Oxford, Oxford, United Kingdom.

[3] UPMC Univ Paris 06, CNRS, UMR 7232, BIOM, Observatoire Océanologique de Banyuls sur Mer, Banyuls/Mer, France.

[4] Commissariat à l'Energie Atomique (CEA), Institut de Génomique (IG), Genoscope, Evry, France.

[5] These authors contributed equally to this work

[6] Correspondence to: jlgomska@upo.es, hescriva@obs-banyuls.fr, nacho.maeso@gmail.com, dpdev@upo.es

The HoxA and HoxD gene clusters of jawed vertebrates are organized into bipartite 3D chromatin structures that separate long-range regulatory inputs coming from the anterior and posterior Hox neighboring regions [1]. This architecture is instrumental in allowing vertebrate Hox genes to pattern disparate parts of the body, including limbs [2]. Almost nothing is known about how these 3D topologies originated. Here, we perform an extensive 4C-seq profiling of the Hox cluster in embryos of amphioxus, an invertebrate chordate. We find that, in contrast to vertebrates, the amphioxus Hox cluster is organized into a single chromatin interaction domain that includes long-range contacts mostly from the anterior side, bringing distant *cis*-regulatory elements into contact with Hox genes. We infer that the vertebrate Hox bipartite regulatory system is an evolutionary novelty built by combining ancient long-range regulatory contacts from DNA in the anterior Hox neighborhood with new regulatory inputs from the posterior side.

How the 3D organization of DNA in the nucleus impacts on regulation of gene expression is a topic of central importance in biology [3]. Despite recent progress in understanding chromatin organization, little is known about how such functional interactions evolve. Here, we study the evolutionary pathway leading to the bipartite 3D chromatin architecture regulating vertebrate Hox gene expression. In animals, chromatin is compartmentalized into Topological Associating Domains (TADs): megabase-scale chromatin regions within which DNA sequences preferentially interact with one another [4,5]. A paradigmatic example of how TADs organize gene regulatory information is presented by the vertebrate Hox clusters, which contain genes of pivotal importance for animal development [6]. Different chromosome conformation capture techniques have revealed that HoxA and HoxD genomic regions are each divided into two main adjacent TADs. These TADs compartmentalize long-range regulatory inputs coming from either side of the clusters into two major domains: enhancers distal to the 3' flank preferentially contact 'anterior' Hox genes whereas those beyond the 5' mostly interact with 'posterior' genes ([2,7-10], Fig. 1a). This bipartite regulatory topology provides gnathostomes with a versatile bimodal system allowing Hox genes to pattern multiple structures, including an ancestral role in antero-posterior axis patterning and novel roles in morphological innovations such as paired limbs [1].

To address whether TADs associated with HoxA and HoxD arose independently or have a shared ancestry dating to before the two vertebrate-specific genome duplications (2R WGDs, Supplementary Figure 1 [11]), we first studied synteny conservation around Hox clusters between and within species. In mouse, HoxA and HoxD neighboring regions are strikingly different, with many HoxA long-range *cis*-regulatory elements (CREs) embedded in introns of neighboring genes, but HoxD long-range CREs located in gene deserts (i.e. large intergenic regions devoid of coding genes). Data from divergent vertebrates, including elephant shark, show that this is a derived situation and all vertebrate Hox cluster neighborhoods were originally very similar. What is now a HoxD gene desert in mammals contained copies of HoxA neighboring genes [12], and gene-free regions surrounding the other two Hox clusters have also resulted from differential loss of neighboring genes coding exons [13] (Fig. 1b, Supplementary Figure 2, Supplementary Note). Thus, differences in the genomic organization of mammalian HoxA and HoxD regulation are derived, not ancestral. This implies that CREs currently engaged in Hox long-range bipartite contacts were primarily intronic and intergenic within a conserved array of neighboring protein-coding loci before Hox cluster duplications (Fig. 1b).

We investigated the ancestry of this arrangement examining the location of vertebrate Hox-neighboring genes in invertebrate genomes. We find that few of these homologues are closely linked to Hox clusters outside chordates, and that gene order and orientations are highly variable (e.g. vertebrate anterior-linked genes are frequently found at the posterior sides and the other way round Supplementary Figure 3). This shuffling of the Hox syntenic environment suggests that in the bilaterian ancestor, long-range Hox *cis*-regulatory interactions were either absent or not important enough to constraint microsynteny. In contrast, in amphioxus (a non-vertebrate chordate that retains many ancestral genomic and morphological features; [14-16]), synteny on the anterior side of the Hox cluster is strikingly conserved with vertebrates; gene order and orientations are almost identical to those inferred for the vertebrate ancestor (Fig. 1b). On the posterior side, most neighboring genes are different: only two immediately adjacent genes, *Evx* and *Lnp*, are conserved in position. The conservation of anterior flanking genes between vertebrates and amphioxus suggests that long-range regulatory interactions from the 3' side had become essential for Hox regulation at the base of the chordate lineage, imposing strong constraints to genomic rearrangements in this region. In the case of the posterior side, given the lack of synteny conservation in non-chordate animals, at present, we cannot discern whether

amphioxus or vertebrates have diverged the most from the syntenic organization of the chordate ancestor. Whatever the case, beyond *Evx* and *Lnp*, gene synteny has followed different evolutionary routes in two chordate groups, suggesting that, in stark contrast to anterior territories, the regulatory contribution of distant posterior regions was less important or even absent in their last common ancestor.

To evaluate this hypothesis experimentally, we compared Hox chromatin contacts between amphioxus and vertebrate embryos using circular chromosome conformation capture followed by high-throughput sequencing (4C-seq), a method that reveals distal chromatin contacts. Studies in mouse embryonic tissues and whole zebrafish embryos have demonstrated that 4C-seq efficiently resolves the organization of the HoxA and HoxD long-range contacts into two adjacent TADs [2,7,8,10,17]. We generated 4C-seq data for fourteen gene 'viewpoints' (8 Hox genes and 6 neighboring genes) in amphioxus embryos and compared these results with previously reported [8] and newly generated zebrafish data (4 Hox genes and 5 neighboring genes). In total, 73 4C-seq datasets were generated including replicates for all viewpoints and 3 amphioxus developmental stages (see Online Methods).

With these datasets, we first defined target interacting regions for each of the 4C-seq viewpoints (genomic regions showing a statistically significant read enrichment against a randomized background) and quantified the number of reads corresponding to each of these targets (Online Methods). These analyses revealed the characteristic bipartite distribution of anterior and posterior Hox genes long-range contacts previously reported in mouse and zebrafish [2,8,10,18] (Fig. 2a, Supplementary Figure 4). Zebrafish *hoxd4a* and *hoxd13a* show little contact overlap, with the majority of their interactions mapping towards opposing sides of the cluster (83.3% anterior and 76.6% posterior, respectively). In contrast, in amphioxus, Hox genes located at the edges of the cluster show the opposite trend: most *Hox2* and *Hox15* contacts converge towards the same direction, with their interacting regions located primarily within the Hox complex (75.2% and 74.2% respectively). In fact, regardless of their position within the cluster, anterior, central and posterior Hox genes exhibit 4C-seq profiles that overlap extensively, with no signs of a bipartite distribution (Fig. 2b, Supplementary Figure 5). Importantly, these Hox interaction profiles are developmentally stable, even though the number of active Hox genes in amphioxus changes dramatically from early gastrula to premouth embryo [19] (Supplementary Figure 6). This temporal stability is in line with previous findings in mouse and *Drosophila*,

where most long-range 3D chromatin interactions are organized similarly across tissues and developmental stages, with only some differences in the intensity of the contacts upon activation of different sets of distal enhancers [7,20,21]. However, despite this temporal uniformity, it is conceivable that in amphioxus TAD structures could be less similar across cell populations with different transcriptional activities than they are in vertebrates; thus, by using whole embryos we may be missing cell type-specific chromatin interactions.

We then correlated 4C-seq results with synteny data. Consistent with the high conservation of anterior neighboring genes, in the majority of amphioxus Hox viewpoints, a significant fraction of contacts map to the conserved anterior region (ranging from 14 to 24.8% for the promoters of *Hox2*, *Hox5*, *Hox6*, *Hox7* and *Hox9* Supplementary Table 1). Long-range interactions between Hox genes and anterior territories are even clearer when using 3' neighboring genes as viewpoints (Supplementary Figure 5, Supplementary Table 1). The amphioxus cluster contains 25.5% of *Hnrnpa* interactions, a similar fraction than its ortholog in zebrafish (33.4%), and in the case of amphioxus *Mtx2*, the percent of contacts corresponding to the Hox complex reaches 42.7%. In contrast, at the posterior side, we found striking differences between amphioxus and vertebrates. Hox genes contact posterior neighboring regions in both chordate lineages; however, the distribution of these 5' interactions is very different (Fig. 2a-b). In zebrafish, *hoxd13a* interactions enter into far distant 5' territories, well beyond the *evx2-lnpa* syntenic region, reaching vertebrate-specific posterior neighboring genes such as *atp5g3a* and *creb2*, consistent with previous reports on the location of zebrafish and mouse 5' long-range Hox enhancers [7,22,23]. In amphioxus, by contrast, the target interacting regions of the posterior-most Hox gene, *Hox15*, are circumscribed to the most proximal neighboring region, with no significant contacts crossing the *Lnp* promoter towards the amphioxus-specific territory. Thus, even within the only 5' region with synteny conservation, interaction profiles are different. In both cases, the Evx-Lnp contacts Hox genes, but while in amphioxus *Evxa* and *Lnp* show a clear interaction preference towards the Hox cluster side (66.1% and 73%, respectively), zebrafish *evx2* and *lnpa* preferentially contact vertebrate specific genomic regions (with only 26.8% and 20.7% of the contacts towards the cluster, respectively)(Supplementary Figures 4-5, Supplementary Table 1). Taken together, these results suggest that there is an inflexion point for long-range chromatin interactions around the *Evxa-Evxb-Lnp* region in amphioxus, with no significant Hox-contacts with 5' amphioxus-specific genes.

To better characterize vertebrate and amphioxus Hox topologies and identify interaction compartments, we generated virtual 3D chromatin architecture models using read counts of the 4C-seq signals as a proxy to distance from each viewpoint (Supplementary Figure 7 and Online Methods). As 4C-seq data correspond to pooled cells from whole embryos, our 3D models provide an average view of chromatin topologies, rather than the dynamic chromatin folding present in each individual cell. These integrative visualizations emphasize how strikingly different the vertebrate and amphioxus 3D chromatin architectures are (Fig. 3). In zebrafish, the HoxDa cluster sits between the two separate anterior and posterior chromatin domains; like a hinge on which the two sets of long-range regulatory inputs can swing. In contrast, the amphioxus Hox cluster appears as a large single chromatin domain that contains distant anterior neighboring genes but not posterior ones. To visualize boundaries between these chromatin domains, we developed a new approach to transform our 4C-seq-derived 3D modeling data into a heatmap of distances (analogous to those obtained by Hi-C, hereafter termed virtual Hi-C; see Online Methods and Supplementary Figures 8-10 for details on virtual Hi-C validations). As expected, zebrafish virtual Hi-C recovered the bipartite architecture that divides vertebrate HoxD clusters into the anterior and posterior TADs (Fig. 3b). In contrast, the amphioxus cluster is contained within a single TAD that includes the conserved anterior neighboring genes, but not the amphioxus-specific posterior genes (such as *Gpatch8*, which has its own interacting compartment) (Fig. 3d). Importantly, no boundaries bisect the cluster or separate Hox genes from anterior neighboring territories. In the case of *Lnp* and the amphioxus Evx genes the situation is less clear: although these loci seem to be part of their own small interaction domain, this region is not completely isolated from its two adjacent compartments (the one containing the Hox and the one including *Gpatch8*). This suggests the single Hox 3D chromatin domain present in amphioxus has a weaker contact border at its posterior side than at its anterior region, and that the *EvxA-EvxB-Lnp* territory can be considered as an extended boundary region (Fig. 3d).

To examine the functional significance of amphioxus Hox chromatin organization, we searched for putative enhancers active in 36 hpf amphioxus embryos (i.e. immediately preceding what can be regarded as a pharyngula stage in amphioxus, equivalent to the zebrafish phylotypic stage at 24 hpf) using ATAC-seq (Assay for Transposase-Accessible Chromatin using sequencing [24]; Fig. 4a). In agreement with the 3D chromatin topologies inferred from the virtual Hi-C results, the distribution of ATAC-seq peaks at either side of the amphioxus Hox gene

cluster revealed very different regulatory potentials for the two Hox neighboring regions (Supplementary Figure 11). While anterior territories are rich in putative distal enhancer regions, the posterior side contains comparatively much fewer ATAC-seq peaks. In fact, apart from the peaks tightly associated to the Evx genes or directly overlapping with transcriptional start sites and repetitive elements, we only found a single candidate enhancer region, within the intergenic region between *Evxb* and *Lnp*. We then tested four putative enhancer elements from within the TAD on the anterior side of the amphioxus cluster located at different distances from the closest Hox gene (elements 1655, 1739, 1784 and 1801, located 150kb, 66kb, 20kb and 3kb downstream of *Hox1*) and the aforementioned element identified at the posterior side (element 2473, 165 kb upstream of *Hox15*), by generating GFP-reporter zebrafish stable transgenes. All anterior enhancers promoted expression in the antero-posterior axis, consistent with expression patterns of amphioxus Hox genes but not with those of neighboring loci (Fig. 4b, Supplementary Figure 12 and [19]), suggesting that they are amphioxus Hox CREs. In contrast, the 2473 posterior element activates GFP expression in isolated neurons in the spinal cord, in a pattern reminiscent of the amphioxus *Evxa* gene (Fig. 4b, [25]) rather than a Hox gene. These experiments suggest that the 3D organization identified using 4C-seq and modeling brings long-range regulatory elements into proximity with amphioxus Hox genes mostly on the anterior side of the cluster (Fig. 4).

In summary, our results support a stepwise evolution of the bimodal regulatory machineries of HoxA and D clusters of jawed vertebrates (Fig. 4c). The relatively simple Hox cluster 3D topology of early bilaterian animals, where external long-range regulation was probably absent, changed profoundly in early chordate evolution, with newly incorporated distal regulatory inputs from anterior neighboring loci becoming a fundamental part of the Hox regulatory architecture. This unipolar topology was further developed in the vertebrate lineage. The acquisition of distal CREs interactions at the posterior side permitted the switch between two separate sets of long-range regulatory inputs, allowing an unprecedented plasticity in the developmental usage of the Hox patterning system in vertebrates.

Accession codes

Data sets presented in this study are available with accession GEO number: GSE68737.

Acknowledgments

We specially thank Dr. Juan Pascual-Anaya for helping with some figures and helpful discussions. We would also like to thank Dr. Fernando Casares, Dr. Isabel Almudí and Dr. Juan Ramón Martínez-Morales for fruitful discussions. Work was funded by grants from Ministerio de Economía y Competitividad (BFU2013-41322-P to JLG-S, Juan de la Cierva postdoctoral contract to IM); the Andalusian Government (BIO-396 to JLG-S; C2A (EE: 2013/2506) to DPD and II); European Research Council (ERC Grant Nº 268513) to PWH and FM; EMBO short fellowship to IM; Universidad Pablo de Olavide to JJT and Conicyt "Becas Chile" to DA.

Author Contributions

R.D.A. carried out the 4C-seq experiments with the help of C.G.-M. and S.G.D. J.J.T. performed the bioinformatic analysis of all the 4C-seq and ATAC-seq datasets. I.I. developed and applied the 3D modeling and virtual Hi-C procedures. F.M. generated the assembly and annotation of the Hox locus in the European amphioxus. J.M.A. and S.M. ensured sequencing project management at Génoscope. E.d.l.C.-M., R.D.A., J.J.T. and I.M. carried out the zebrafish reporter assays. S.B. and D.A. collected and processed the amphioxus embryonic material and performed in situ hybridizations. I.M. completed the amphioxus ATAC-seq experiments and the evolution of synteny analyses. J.L.G.-S., H.E., I.M. and D.D. conceived, designed and coordinated the project. J.L.G.-S. and I.M. wrote the manuscript with the help of P.W.H.H. All the authors revised and contributed to the final version of the text.

All experimental procedures using vertebrates were ethically approved by the Andalusian Government.

References

1. Lonfat, N. & Duboule, D. Structure, function and evolution of topologically associating domains (TADs) at HOX loci. *FEBS Lett* (2015).
2. Andrey, G. *et al.* A switch between topological domains underlies HoxD genes collinearity in mouse limbs. *Science* **340**, 1234167 (2013).
3. de Laat, W. & Duboule, D. Topology of mammalian developmental enhancers and their regulatory landscapes. *Nature* **502**, 499-506 (2013).
4. Gomez-Diaz, E. & Corces, V.G. Architectural proteins: regulators of 3D genome organization in cell fate. *Trends Cell Biol* **24**, 703-11 (2014).
5. Ciabrelli, F. & Cavalli, G. Chromatin-driven behavior of topologically associating domains. *J Mol Biol* **427**, 608-25 (2015).
6. Mallo, M. & Alonso, C.R. The regulation of Hox gene expression during animal development. *Development* **140**, 3951-63 (2013).

7. Montavon, T. *et al.* A regulatory archipelago controls Hox genes transcription in digits. *Cell* **147**, 1132-45 (2011).
8. Woltering, J.M., Noordermeer, D., Leleu, M. & Duboule, D. Conservation and divergence of regulatory strategies at Hox Loci and the origin of tetrapod digits. *PLoS Biol* **12**, e1001773 (2014).
9. Berlivet, S. *et al.* Clustering of tissue-specific sub-TADs accompanies the regulation of HoxA genes in developing limbs. *PLoS Genet* **9**, e1004018 (2013).
10. Lonfat, N., Montavon, T., Darbellay, F., Gitto, S. & Duboule, D. Convergent evolution of complex regulatory landscapes and pleiotropy at Hox loci. *Science* **346**, 1004-6 (2014).
11. Dehal, P. & Boore, J.L. Two rounds of whole genome duplication in the ancestral vertebrate. *PLoS Biol* **3**, e314 (2005).
12. Lehoczky, J.A., Williams, M.E. & Innis, J.W. Conserved expression domains for genes upstream and within the HoxA and HoxD clusters suggests a long-range enhancer existed before cluster duplication. *Evol Dev* **6**, 423-30 (2004).
13. Maeso, I. *et al.* An ancient genomic regulatory block conserved across bilaterians and its dismantling in tetrapods by retrogene replacement. *Genome Res* **22**, 642-55 (2012).
14. Paps, J., Holland, P.W. & Shimeld, S.M. A genome-wide view of transcription factor gene diversity in chordate evolution: less gene loss in amphioxus? *Brief Funct Genomics* **11**, 177-86 (2012).
15. Bertrand, S. & Escriva, H. Evolutionary crossroads in developmental biology: amphioxus. *Development* **138**, 4819-30 (2011).
16. Holland, L.Z. & Onai, T. Early development of cephalochordates (amphioxus). *Wiley Interdiscip Rev Dev Biol* **1**, 167-83 (2012).
17. Noordermeer, D. *et al.* Temporal dynamics and developmental memory of 3D chromatin architecture at Hox gene loci. *Elife* **3**, e02557 (2014).
18. Noordermeer, D. *et al.* The dynamic architecture of Hox gene clusters. *Science* **334**, 222-5 (2011).
19. Pascual-Anaya, J. *et al.* Broken colinearity of the amphioxus Hox cluster. *Evodevo* **3**, 28 (2012).
20. Dixon, J.R. *et al.* Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**, 376-80 (2012).
21. Ghavi-Helm, Y. *et al.* Enhancer loops appear stable during development and are associated with paused polymerase. *Nature* **512**, 96-100 (2014).
22. Gonzalez, F., Duboule, D. & Spitz, F. Transgenic analysis of Hoxd gene regulation during digit development. *Dev Biol* **306**, 847-59 (2007).
23. Gehrke, A.R. *et al.* Deep conservation of wrist and digit enhancers in fish. *Proc Natl Acad Sci U S A* **112**, 803-8 (2015).
24. Buenrostro, J.D., Giresi, P.G., Zaba, L.C., Chang, H.Y. & Greenleaf, W.J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat Methods* **10**, 1213-8 (2013).
25. Ferrier, D.E., Minguillon, C., Cebrian, C. & Garcia-Fernandez, J. Amphioxus Evx genes: implications for the evolution of the Midbrain-Hindbrain Boundary and the chordate tailbud. *Dev Biol* **237**, 270-81 (2001).

Figures

Fig. 1. Genomic organization of vertebrate and amphioxus Hox clusters. (**a**) Distribution of TADs (obtained from human cells Hi-C datasets [20]) and schematics of the chromatin architecture of HoxA and HoxD clusters, showing their similar 3D topology. Colored bars represent Hox (white), anterior (blue) and posterior (red) neighboring genes. (**b**) Microsynteny arrangements around the Hox clusters of gnathostomes, the pre-WGD vertebrate ancestor and amphioxus. Genes are represented by arrows according to the transcriptional orientation (white, Hox clusters; blue, anterior; red, posterior; grey, genes with non-conserved linkages), those outlined by dashed lines correspond to vertebrate paralogs that have been lost in at least one species. Question marks indicate genes whose status in the vertebrate ancestor could not be inferred. Slashes correspond to non-conserved amphioxus loci shown in Supplementary Figure 3.

Fig. 2. Comparative of the 4C-seq interaction profiles of zebrafish and amphioxus Hox clusters. Normalized 4C-seq profiles of several Hox genes promoters in zebrafish HoxDa locus (**a**) and amphioxus Hox locus (**b**). Spider plots show the statistically significant contacts to the left (blue arcs) and to the right (red arcs) of each viewpoint. Percentages of reads aligned to statistically significant targets at each side of viewpoints are indicated in blue (left contacts) or red (right contacts). Units in the y-axes correspond to normalized interacting counts. Green bars indicate the positions of the viewpoints.

Fig. 3. 3D chromatin architecture of amphioxus and zebrafish Hox clusters. 3D models of zebrafish HoxDa (**a**) and amphioxus Hox (**c**) regions. 4C-seq viewpoints are highlighted (blue, anterior genes; yellow, Hox genes; red, posterior genes). (**b**, **d**) Zebrafish and amphioxus virtual Hi-C consensus of all 3D model solutions. 4C-seq viewpoints are represented by circles with the same color scheme of the previous panels. Arrows point out the TAD border bisecting the zebrafish Hox cluster (b) and the absence of this border in the case of amphioxus (d).

Fig. 4. Regulatory compartments in the amphioxus Hox region and evolution of Hox 3D architecture. (**a**) Amphioxus virtual Hi-C heatmap and ATAC-seq profile at 36 hpf in the Hox region. Amphioxus ATAC-seq peaks tested in zebrafish are colored and highlighted by asterisks

(blue for those in the anterior region, red for the one at the posterior side of the cluster). (**b**) Lateral views of 24-hpf and 48-hpf (inset in 2473) embryos of zebrafish transgenic lines showing GFP expression driven by the amphioxus ATAC-seq peaks (1655, 1739, 1784, 1801 and 2473) highlighted in (a). Midbrain expression corresponds to the enhancer positive control included in the reporter constructs. Whole mount *in situ* hybridizations of *Hox1* and *Evxa* in 36-hpf amphioxus embryos are shown for comparisons. Anterior is to the left. (ey) eye; (hb) hindbrain; (mb) midbrain; (nc) neural crest cells; (ne) neurons; (no) notochord; (ot) otic vesicle, (op) olfactory placode, (sc) spinal chord. (**c**) 3D-architectures schematics showing an evolutionary scenario for the origin of the bimodal regulatory system of jawed vertebrates. The Hox-only chromatin domain of early bilaterians is first expanded by the anterior side in the chordate ancestor and by the posterior side at the origin of vertebrates, allowing the bipartition of the regulatory topologies of HoxA and HoxD clusters.

Online Methods

Genome sequencing and assembly

DNA was prepared from a single European amphioxus (*Branchiostoma lanceolatum*) mature male and sequenced using Illumina technology at Génoscope (Centre National de Séquençage, Evry, France). Briefly, two paired-end (180bp and 700bp) and 6 mate-pair libraries (3, 5, 8kb) were generated and sequenced at >200x total coverage.

Reads were quality-trimmed using sickle (v1.290, https://github.com/najoshi/sickle), error corrected using Musket (v1.0.6) [26] and overlapping libraries merged using Flash [27]. Assembly was carried out using SOAPdenovo (v2.04) [28] using a k-mer of 71 for contiging and of 35 for mapping and scaffolding. Gaps were subsequently filled using GapCloser (v1.2) [28] with an overlap parameters set to 31. The resulting assembly (N50: 649,215, size: 948.5Mb) contains allelic copies for most scaffolds (expected genome size ~500Mb) that we reconciled using the Haplomerger [29] pipeline relying on best-reciprocal lastz alignment after masking repeats using a custom library built with Repeatmodeler (http://www.repeatmasker.org). The Hox locus was extracted from the final assembly (N50: 1132648bp, size: 526.8Mb) and submitted together with the 4C-seq and ATAC-seq data (GSE68737).

Gene models were built using Evidence Modeler (EVM) [30] based on (i) *de novo* gene prediction obtained using Augustus [31] with a custom training based on CEGMA [32] report, (ii) split-aware alignment of human proteins using Exonerate [33] and transcriptome alignment. Models for known genes within the Hox region that were not present in these annotations were added manually. More details regarding genome *B. lanceolatum* assembly and annotation will be provided in a separate upcoming publication.

Synteny analyses and genome browsing

Hox neighboring genes were searched across the different studied species using tblastn and blastp. We compared the relative orientations and positions of these genes by browsing the genomes of the studied species through the NCBI (http://www.ncbi.nlm.nih.gov/blast/Blast.cgi), UCSC (http://genome.ucsc.edu/), and Ensembl Metazoa (http://metazoa.ensembl.org/info/website/species.html) webpages, using the following genome versions: Elephant shark (*Callorhinchus milii*) 6.1.3, *Lottia gigantea* v1.0, *Mus musculus* Build 38, *Saccoglossus kowalevskii* Build1.1, *Strigamia maritima* Smar1.0, *Trichoplax adhaerens* v1.0. In the case of the starfish *Acanthaster planci*, no gene annotation or genome browser were available for the published *A. planci* Hox genome scaffold (accession number DF933567.1) [34]. Therefore, we used tblastn to search for conserved neighboring genes and Genscan to predict genes *de novo*.

Mouse *Jazf2* pseudogenized exons were detected with VISTA [35] using elephant shark as a reference sequence, LAGAN as the alignment program and the following parameters: 100bp window and 65% of identity in 70bp.

Amphioxus procurement and culture

B. lanceolatum ripe adults were collected at the Racou beach in Argelès-sur-Mer (France). Gametes were collected by heat stimulation as previously described [36,37]. Fertilization was undertaken in Petri dishes filled with filtered sea water and embryos were cultured at 19°C.

Whole mount *in situ* hybridization

Partial cDNAs from *Gpatch8*, *Nfe2*, *Lnp*, *Slc20*, *Mtx2*, *Hnrnpa,* and *Cbx* of *B. lanceolatum* were amplified by RT-PCR and cloned in pGEM-T Easy vector. DIG-labeled RNA

probes were synthesized by *in vitro* translation after plasmid linearization using the appropriate enzymes. Fixation and whole-mount *in situ* hybridization (WISH) were performed as described in [38]. No expression could be detected using WISH for *Gpatch8*, *Lnp*, *Slc20* and *Mtx2*; the expression patterns for the rest of the genes are included in Supplementary Figure 12.

4C-seq

4C-seq assays were performed as previously reported [18,39-41]. For each zebrafish biological replicate, 500 embryos at 24hpf of the Tübingen strain were dechorionated using pronase and deyolked in 1 ml of Ginzburg Fish Ringers (55 mM NaCl, 1.8 mM KCl, 1.25 mM NaHCO3). Then, they were fixed in PBS1X-formaldehyde 2% for 15 minutes at room temperature. For amphioxus biological replicates, embryos (~8000 in the case of 8 hpf and ~4000 for the 15 hpf and 36 hpf stages) were concentrated by centrifugation at low speed in 2mL microtubes. They were fixed 15 minutes at room temperature in 1,5mL of MOPS buffer (0,1M MOPS pH 7,5, 2mM MgSO2, 1 mM EGTA, 0,5M NaCl) containing 1,85% formaldehyde. 155µL of 10% Glycine were added to both species samples to stop the fixation, followed by 5 washes with PBS (NaPBS in the case of amphioxus) at 4°C. Pellets were frozen in liquid nitrogen and kept at -80°C. Isolated cells were lysed (lysis buffer: 10 mM Tris-HCl pH 8, 10 mM NaCl, 0.3% IGEPAL CA-630 (Sigma-Aldrich, I8896), 1X protease inhibitor cocktail (cOmplete, Roche, 11697498001)) and the DNA digested with DpnII (New England Biolabs, R0543M) and Csp6I (Fermentas, Thermo Scientific, FD0214) as primary and secondary enzymes respectively. T4 DNA ligase (Promega, M1804) was used for both ligation steps. Specific primers were designed around the putative transcriptional start sites of the genes with Primer3 v. 0.4.0 [42]. Illumina adaptors were included in the primers sequence and 8 PCRs were performed with Expand Long Template PCR System (Roche, 11759060001) and pulled together. Two libraries coming from different biological replicates were generated for each 4C-seq experiment (i.e. for each viewpoint and for each developmental stage). These libraries were purified with a High Pure PCR Product Purification Kit (Roche, 11732668001), their concentrations measured using the Quanti-iTTM PicoGreen dsDNA Assay Kit (Invitrogen, P11496) and sent for deep sequencing. 4C-seq data were analyzed as previously described [17]. Briefly, raw sequencing data were demultiplexed and aligned using zebrafish July 2010 assembly (danRer7) and the *B. lanceolatum* reference genomes. Reads located in fragments flanked by two restriction sites of the same enzyme, in

fragments smaller than 40 bp, or within a window of 10 kb around the viewpoint (indicated by dashed lines in the different figures) were filtered out. Mapped reads were then converted to reads-per-first-enzyme-fragment-end units, and smoothed using a 30 fragments mean running window algorithm. 4C-seq data were normalized by total weight of reads within the window displayed in figures.

To calculate statistically significant contacting regions for each viewpoint, an average background level was estimated as previously described [43]. Briefly, fragments distribution in a window of 2 Mb around each viewpoint was randomized, excluding an internal window of 100 Kb around the viewpoint to avoid biases due to close contacts. Then, this randomized fragment distribution was smoothened as described above. This randomized profile was then used to calculate the p-value for each potential target in the observed 4C-seq distribution by means of Poisson probability function. Regions with p-values below 1E-5 were considered as statistically significant interacting targets.

To calculate the distribution of contacts at each side of the viewpoints we took into account only those reads overlapping the interacting targets, discarding also those mapped within the 100kb viewpoint window as previously reported [8]. The same approach was used to quantify the distribution of contacts in the three windows defined as follows: cluster (from the 5' UTR of the 5' most Hox genes (zebrafish *hoxd13a* and amphioxus *Hox15*) to the 3' UTR of the 3' most Hox genes (zebrafish *hoxd3a* and amphioxus *Hox1*)); anterior (downstream of zebrafish *hoxd3a* and amphioxus *Hox1*); and posterior (upstream of zebrafish *hoxd13a* and amphioxus *Hox15*).

3D Computational Modeling and virtual Hi-C

4C Data normalization

To equal the amount of reads in all experiments, we normalized the reads of the 4C-seq datasets. We then extracted the data relevant for the modeling by calculating the Z-score (see the sections below on Z-score thresholds optimization) of those reads as in [44].

Structure determination

The overall approach of the determination of the genomes structures was adapted from a previous work [44] with some variations, using the Integrative Modeling Platform (IMP) [45]. The procedure was divided in three stages:

1) Representation of the genome locus and translation of the data into spatial restraints. We represented the chromosomal fragment as a flexible string of beads where each bead corresponded to a number of consecutive fragments between 10 and 45, depending on the total size of the locus (Supplementary Figure 7c). The size of the beads representing those 20 fragments was proportional to the sum of the sizes of these fragments.

In order to impose connection between the beads, harmonic upper bound distance restraints were used between consecutive beads. This distance was the sum of the radii of both beads. Excluded volume restraints were imposed over all the beads so these would not overlap each other. The reach window of a viewpoint was defined as the area between the furthest upstream and downstream fragments with a Z-score above the upper Z-score (uZ) (Supplementary Figure 13). Harmonic distance restraints were applied between beads corresponding to the viewpoints and the rest of the beads, as long as these beads' Z-scores were above the uZ or below the lower Z-score (lZ). We used the absolute Z-score of the reads to give more weight to the most meaningful reads. Beads outside the reach window were restrained with harmonic lower bound distances, with a weight equal to the absolute Z-score. With the harmonic lower bound restraint we only impose the beads not to be closer than their computed distance (Supplementary Figure 7).

2) Optimization and sampling of the space of solutions. We combined a Monte Carlo exploration with a local optimization of conjugate gradients and simulated annealing. We started with an individual optimization of 5 steps of conjugate gradients from a entirely random configuration of beads followed by simulated annealing until the score difference between rounds was below 0.00001 or reached 0 (Supplementary Figure 7d). To sample the space of solutions exhaustively we computed 50.000 independent optimizations for each genome (Supplementary Figure 7e).

3) Analysis and assessment of the ensemble of models. We gathered the 200 models with the best score. Those solutions were then clustered according to their similarity measured by their Root Mean Square Deviation (RMSD). We used the Multiexperiment Viewer, MeV [46] with Hierarchical Clustering and K-Means clustering. All models grouped in two clusters that were the mirror image of each other (Supplementary Figure 14). The most representative models (i.e. the closest ones to the mean of all solutions within the most populated cluster) are displayed in

Fig. 3. Results were indistinguishable when we used the solutions of the other mirror image cluster.

Reconstruction of virtual Hi-C data

We used the models from the most populated cluster to generate the heat map plots that were equivalent to Hi-C data. First we superimposed all the models (Supplementary Figure 7f). To generate virtual Hi-C heatmap plots, we measured the distances between all beads in each model and calculated the mean of these distances (Supplementary Figure 7g).

Empirical calculation of the Maximum distance, the lZ and the uZ.

The calculation of these parameters was done as described previously [44] with little variations: The uZ varied between 0.2 and 1.4 in bins of 0.2. The lZ varied in bins of 0.2 between -1.4 and -0.2. The maximum distance varied from 3000 to 7000 in bins of 1000. Due to the heavy computational load, we did not consider thinner bins or higher or lower values.

For each set of parameters, we generated 500 models and calculated the mean distances between the viewpoints and the rest of the fragments and compared them to the distances that represented each set of 20 fragments of the normalized 4C data (Supplementary Figure 15 b, d, f).

The set of parameters that best fitted the 4C data were 0.2 for the uZ and -0.2 for the lZ in amphioxus, zebrafish and mouse. The best max distances were different for each species. To allow comparison, we needed to settle the same max distance for all three. Taking this into account, and for the sake of ease of visualization, we settled on the max distance of 7000, whose score was also amongst the best (Supplementary Figure 15 a, c, e).

Validation of the virtual Hi-C approach

To validate the virtual Hi-C method we followed two strategies:

1) Jackknife resampling. We tested the reproducibility and robustness virtual Hi-C results taking advantage of the extensive number of viewpoints available in our amphioxus and zebrafish Hox 4C-seq data. We performed additional modeling experiments by resampling our original datasets using different subsets of 4C data both in zebrafish and in amphioxus (Supplementary Table 2). We generated 500 models with the same parameters that we used for our initial modeling and reconstructed virtual Hi-C data for each subset. Subsequently, we calculated Spearman´s coefficients between the different subsets. This demonstrated that virtual

Hi-C results are very reproducible and robust to perturbations, with high correlations even when 60% of the viewpoints are eliminated (Supplementary Figure 10, Supplementary Table 2).

2) Modeling of other loci and shifted calculation of correlations. To validate the models and the virtual Hi-C derived from them, we generated models for diverse mouse genomic regions using previously published 4C-seq data (from the HoxD locus and two additional loci: *Wnt6-Ihh-Epha4-Pax3* and *Med13l-Tbx3-Tbx5-Rbm19* [17,47,48]). Using these models, we generated the virtual Hi-Cs and compared them with previously published experimental Hi-C data [20] (Supplementary Figures 8-9). These comparisons were performed shifting the window used for the modeling 25% of its size in each direction, in steps of 20Kb (see Supplementary Figure 8). For each comparison, Spearman's and Pearson's correlations were calculated. Due to the dominance of read counts corresponding to short distances, we calculated these correlations using bins separated by at least 240kb (HoxD and *Med13l-Tbx3-Tbx5-Rbm19*) or 480kb (*Wnt6-Ihh-Epha4-Pax3*), to account for the different size of these three loci (~2.12, ~2.48 and ~4.88Mb, respectively). In all cases, our 4C-seq-derived virtual Hi-C contact matrices accurately recapitulate the TAD organization and borders present in the experimental Hi-C maps, with Spearman's and Pearson's coefficients within the same range (from 0.63 to 0.88) of those typically obtained between different Hi-C experimental conditions (from 0.4 to 0.99 [20,49-51]) (Supplementary Figure 9, Supplementary Table 3).

ATAC-seq

ATAC-seq experiments in amphioxus embryos were performed as previously described [23,24]. Approximately 80000 cells (corresponding to thirteen 36 hpf embryos) were directly lysed in cold lysis buffer (10 µM Tris pH7.4, 10 µM NaCl, 3 µM MgCl2, 0.1% IGEPAL) after removing the seawater by centrifuging briefly. Then, the sample was incubated for 30 min at 37ºC with the TDE1 enzyme and purified with Qiagen Minelute kit. A PCR was performed with 13 cycles using Ad1F and Ad2.3R primers and KAPA HiFi hotstart enzyme (Kapa Biosystems). The resulting library was multiplexed and sequenced in a HiSeq 2000 lane. Reads were aligned using the mentioned *B. lanceolatum* assembly. Duplicated pairs or those ones separated by more than 2Kb were removed. The enzyme cleavage site was determined as the position -4 (minus strand) or +5 (plus strand) from each read start, and this position was extended 5 bp in both directions for signal visualization. For the zebrafish reporter assays of anterior elements, we

selected 4 regions including ATAC-seq peaks with no overlap with coding exons, transcriptional start sites and repetitive elements. We applied the same criteria to the posterior region, also excluding ATAC-seq peaks tightly associated with amphioxus Evx genes (i.e. those located in the Evx introns and within 5kb of the Evx transcribed regions). This rendered a single candidate element between *Evxb* and *Lnp* (see Supplementary Figure 11).

Transgenesis in zebrafish

Transgenesis assays were performed as previously reported [52]. Putative enhancers were amplified by PCR from amphioxus genomic DNA using the primers listed in Supplementary Table 4. The PCR fragments were subcloned in PCR8/GW/TOPO vector and, using Gateway technology (Life Technologies), were shuttled into an enhancer detection vector composed of a *gata2* minimal promoter, an enhanced GFP reporter gene, and a strong midbrain enhancer (z48) that works as an internal control for transgenesis in zebrafish [23]. Zebrafish transgenic embryos were generated using the Tol2 transposon/transposase method [53], with minor modifications. One-cell embryos were injected with a 2 nl volume containing 25 ng/µl of transposase mRNA, 20 ng/µl of purified constructs and 0.05% of phenol red. In order to ensure the reproducibility of the expression patterns observed in the reporter assays, three or more stable transgenic lines derived from different founders were generated for each construct.

Methods-only References

26. Liu, Y., Schroder, J. & Schmidt, B. Musket: a multistage k-mer spectrum-based error corrector for Illumina sequence data. *Bioinformatics* **29**, 308-15 (2013).
27. Magoc, T. & Salzberg, S.L. FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics* **27**, 2957-63 (2011).
28. Luo, R. *et al.* SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *Gigascience* **1**, 18 (2012).
29. Huang, S. *et al.* HaploMerger: reconstructing allelic relationships for polymorphic diploid genome assemblies. *Genome Res* **22**, 1581-8 (2012).
30. Haas, B.J. *et al.* Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biol* **9**, R7 (2008).
31. Stanke, M. *et al.* AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Res* **34**, W435-9 (2006).
32. Parra, G., Bradnam, K. & Korf, I. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics* **23**, 1061-7 (2007).
33. Slater, G.S. & Birney, E. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* **6**, 31 (2005).
34. Baughman, K.W. *et al.* Genomic organization of Hox and ParaHox clusters in the echinoderm, Acanthaster planci. *Genesis* **52**, 952-8 (2014).

35. Frazer, K.A., Pachter, L., Poliakov, A., Rubin, E.M. & Dubchak, I. VISTA: computational tools for comparative genomics. *Nucleic Acids Res* **32**, W273-9 (2004).
36. Fuentes, M. *et al.* Insights into spawning behavior and development of the European amphioxus (Branchiostoma lanceolatum). *J Exp Zool B Mol Dev Evol* **308**, 484-93 (2007).
37. Fuentes, M. *et al.* Preliminary observations on the spawning conditions of the European amphioxus (Branchiostoma lanceolatum) in captivity. *J Exp Zool B Mol Dev Evol* **302**, 384-91 (2004).
38. Somorjai, I., Bertrand, S., Camasses, A., Haguenauer, A. & Escriva, H. Evidence for stasis and not genetic piracy in developmental expression patterns of Branchiostoma lanceolatum and Branchiostoma floridae, two amphioxus species that have evolved independently over the course of 200 Myr. *Dev Genes Evol* **218**, 703-13 (2008).
39. Dekker, J., Rippe, K., Dekker, M. & Kleckner, N. Capturing chromosome conformation. *Science* **295**, 1306-11 (2002).
40. Hagege, H. *et al.* Quantitative analysis of chromosome conformation capture assays (3C-qPCR). *Nat Protoc* **2**, 1722-33 (2007).
41. Splinter, E., de Wit, E., van de Werken, H.J., Klous, P. & de Laat, W. Determining long-range chromatin interactions for selected genomic sites using 4C-seq technology: From fixation to computation. *Methods* (2012).
42. Rozen, S. & Skaletsky, H. Primer3 on the WWW for general users and for biologist programmers. *Methods Mol Biol* **132**, 365-86 (2000).
43. Smemo, S. *et al.* Obesity-associated variants within FTO form long-range functional connections with IRX3. *Nature* **507**, 371-5 (2014).
44. Bau, D. *et al.* The three-dimensional folding of the alpha-globin gene domain reveals formation of chromatin globules. *Nat Struct Mol Biol* **18**, 107-14 (2011).
45. Russel, D. *et al.* Putting the pieces together: integrative modeling platform software for structure determination of macromolecular assemblies. *PLoS Biol* **10**, e1001244 (2012).
46. Saeed, A.I. *et al.* TM4: a free, open-source system for microarray data management and analysis. *Biotechniques* **34**, 374-8 (2003).
47. Lupianez, D.G. *et al.* Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. *Cell* **161**, 1012-25 (2015).
48. van Weerd, J.H. *et al.* A large permissive regulatory domain exclusively controls Tbx3 expression in the cardiac conduction system. *Circ Res* **115**, 432-41 (2014).
49. Vietri Rudan, M. *et al.* Comparative Hi-C reveals that CTCF underlies evolution of chromosomal domain architecture. *Cell Rep* **10**, 1297-309 (2015).
50. Hou, C., Li, L., Qin, Z.S. & Corces, V.G. Gene density, transcription, and insulators contribute to the partition of the Drosophila genome into physical domains. *Mol Cell* **48**, 471-84 (2012).
51. Zhang, Y. *et al.* Spatial organization of the mouse genome and its role in recurrent chromosomal translocations. *Cell* **148**, 908-21 (2012).
52. Bessa, J. *et al.* Zebrafish enhancer detection (ZED) vector: A new tool to facilitate transgenesis and the functional analysis of *cis*-regulatory regions in zebrafish. *Developmental Dynamics* **238**, 2409-2417 (2009).
53. Kawakami, K. Transgenesis and gene trap methods in zebrafish by using the Tol2 transposable element. *Methods Cell Biol* **77**, 201-22 (2004).

Figure 1

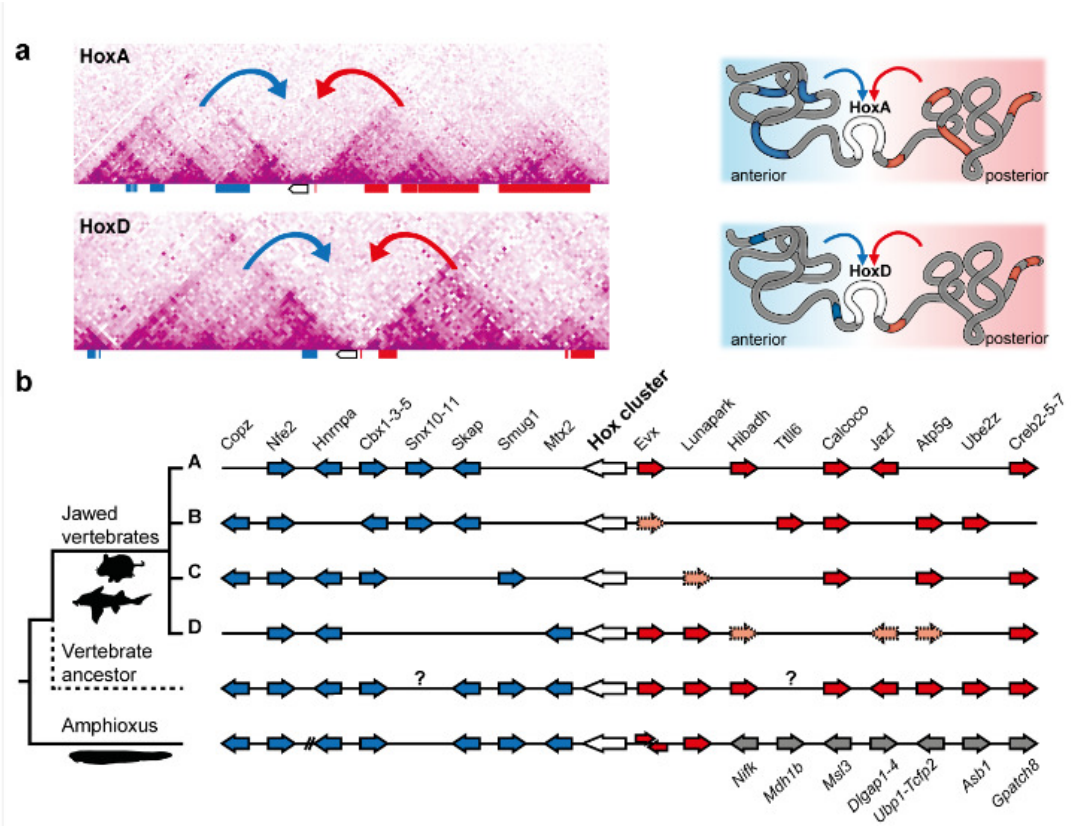

a
HoxA
HoxD
HoxA
anterior
posterior
HoxD
anterior
posterior
b
Copz
Nfe2
Hnrnpa
Cbx1-3-5
Snx10-11
Skap
Smug1
Mtx2
Hox cluster
Evx
Lunapark
Hibadh
Ttll6
Calcoco
Jazf
Atp5g
Ube2z
Creb2-5-7
A
B
C
D
Jawed vertebrates
Vertebrate ancestor
Amphioxus
?
?
#
Nifk
Mdh1b
Msl3
Dlgap1-4
Ubp1-Tcfp2
Asb1
Gpatch8

Figure 2

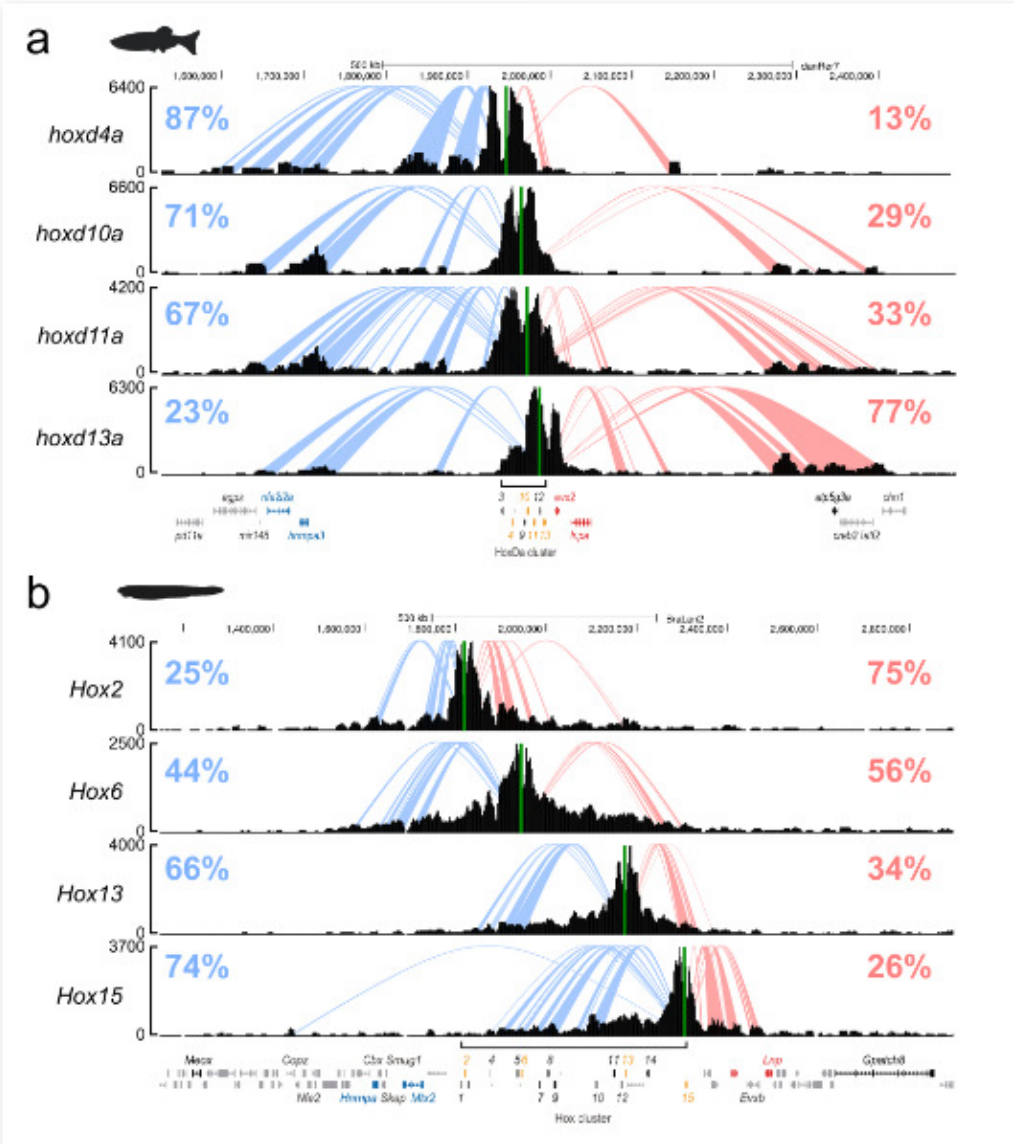

a
hoxd4a
87%
13%
hoxd10a
71%
29%
hoxd11a
67%
33%
hoxd13a
23%
77%
b
Hox2
25%
75%
Hox6
44%
56%
Hox13
66%
34%
Hox15
74%
26%
Hox cluster

Figure 3

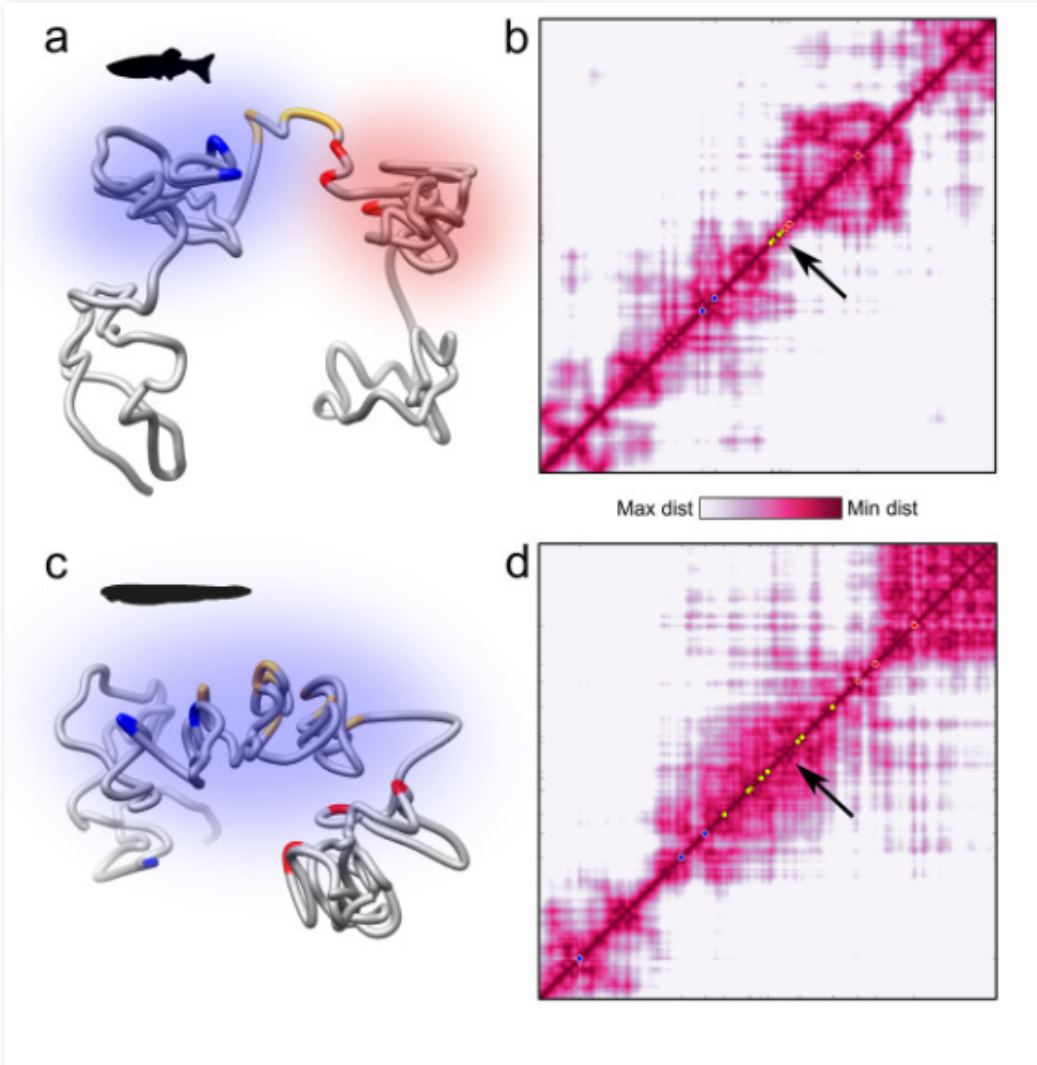
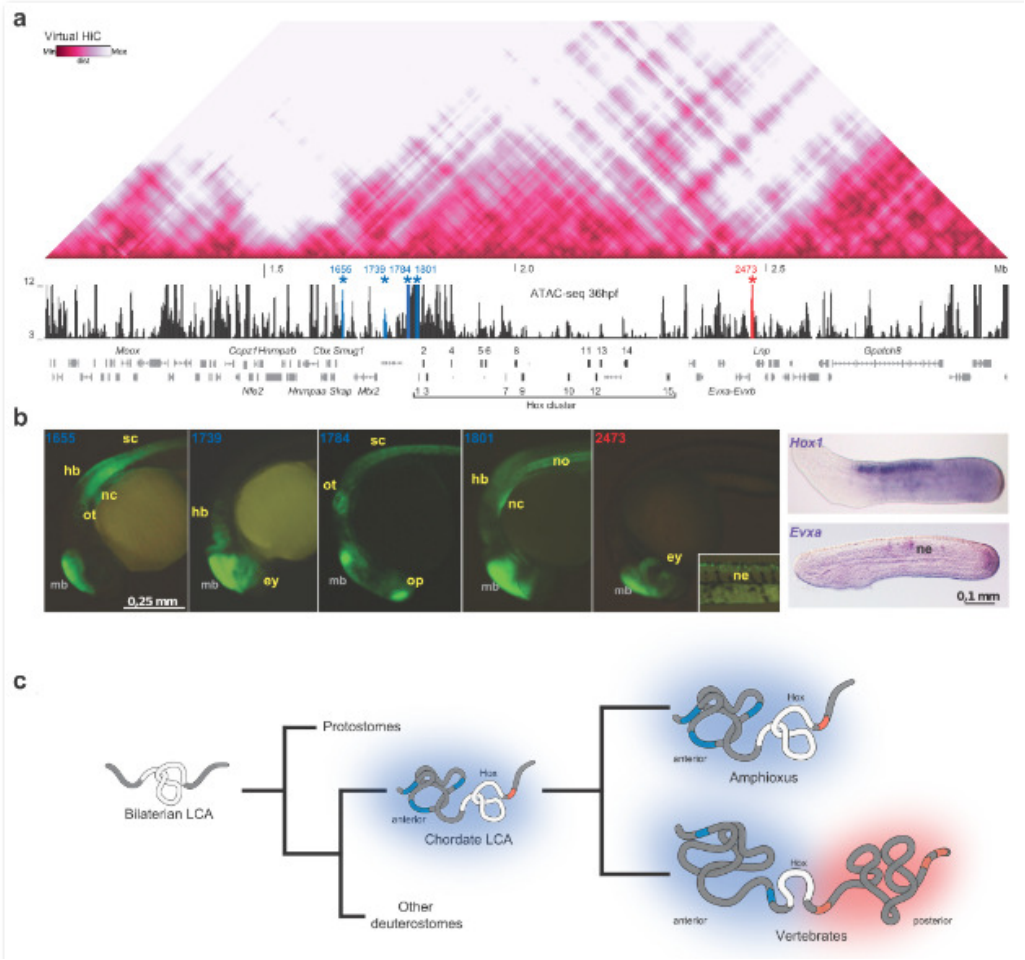


Figure 4


a
Virtual HiC
ATAC-seq 36hpf
Mb
Meox
Cdx2 Hmpaab
Cbx Smug1
Lsp
Gpatch8
Mx2
Hmpaa Skap
Mx2
Hox cluster
Evxa-Evxb
b
1655
sc
hb
nc
ot
mb
0,25 mm
1739
hb
mb
ey
1784
sc
ot
mb
op
1801
no
hb
nc
mb
2473
ey
ne
mb
Hox1
Evxa
ne
0,1 mm
c
Bilaterian LCA
Protostomes
Hox
anterior
Chordate LCA
Other
deuterostomes
Amphioxus
Vertebrates
posterior