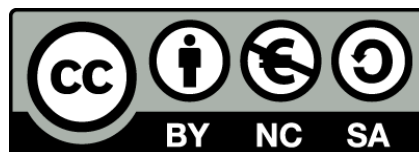




UNIVERSITAT DE
BARCELONA

Conservation of different mechanisms of Hox cluster regulation within chordates

Carlos Herrera Úbeda

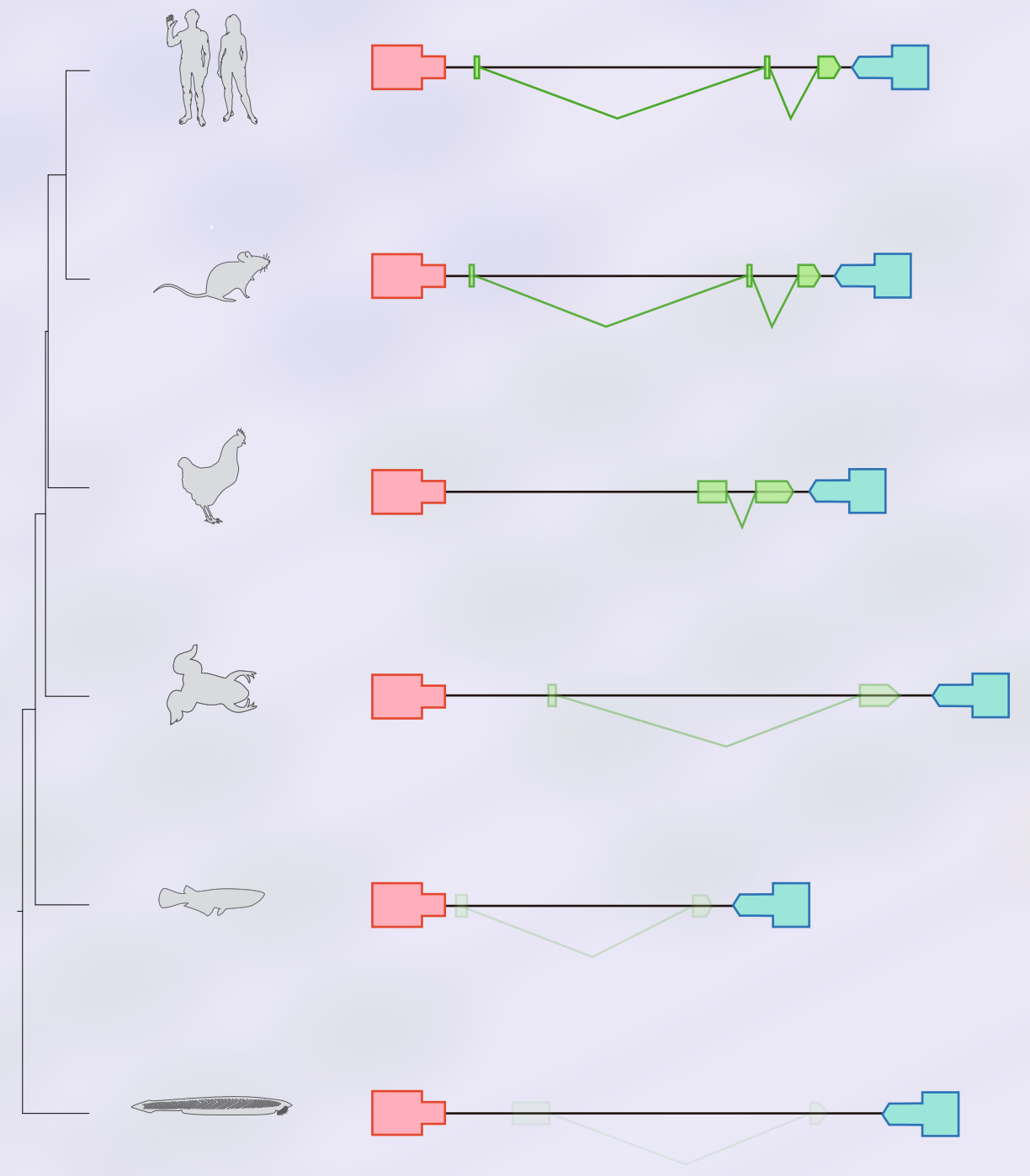


Aquesta tesi doctoral està subjecta a la llicència **Reconeixement- NoComercial – Compartir Igual 4.0. Espanya de Creative Commons.**

Esta tesis doctoral está sujeta a la licencia **Reconocimiento - NoComercial – Compartir Igual 4.0. España de Creative Commons.**

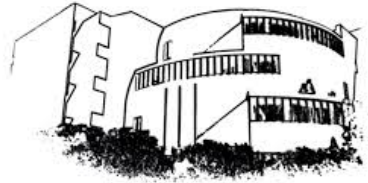
This doctoral thesis is licensed under the **Creative Commons Attribution-NonCommercial-ShareAlike 4.0. Spain License.**

CONSERVATION OF DIFFERENT MECHANISMS OF HOX CLUSTER REGULATION WITHIN CHORDATES





UNIVERSITAT DE
BARCELONA



Facultat de Biologia

Departamento de Genética, Microbiología y Estadística
Programa de Doctorado de Genética
2019

Tesis Doctoral presentada por

Carlos Herrera Úbeda

bajo el título

**Conservation of different mechanisms of Hox cluster regulation
within chordates**

para optar al título de doctor por la Universitat de Barcelona

Firmado, el doctorando

Firmado, el Director, responsable asimismo de la tutela de la tesis

Jordi Garcia Fernàndez

Barcelona, 27 de Septiembre de 2019

We are LEGIØN
David Haller

Summary	1
Acknowledgements	3
Introduction	9
1. Evolution and development	9
1.1. Dawn of the Evo-Devo	9
1.2. Changes vs Conservation	9
1.3. Amphioxus	10
<u>Overview</u>	10
<u>Development</u>	12
<u>Functional genomics</u>	13
1.4. <i>Xenopus tropicalis</i>	14
1.5. Hox Cluster	15
1.6. ParaHox Cluster	17
2. Long non-coding RNAs	19
2.1. Overview of lncRNAs	19
<u>History</u>	19
<u>Origin and characteristics of lncRNAs</u>	21
<u>Evolution of lncRNAs</u>	22
<u>Classification</u>	23
<u>Challenges</u>	24
2.2. <i>Hotairm1 (HOX antisense intergenic RNA myeloid 1)</i>	25
<u>Overview</u>	25
<u>Mechanism of <i>Hotairm1</i></u>	26
Objectives	33
Results	37
Article RI	39
Article RII	67
Article RIII	83
Discussion	137
1. Expanding the tools for understanding Evolution	137
2. Conserved lincRNAs within Chordata	138
2.1. <i>lncRNA fraction in amphioxus</i>	138
2.2. <i>Novelty of the LincOFinder pipeline</i>	138

2.3. <i>Origins of Hotairm1</i>	140
2.4. <i>Conservation on the expression between frog and lancelet Hotairm1</i>	141
2.5. <i>Conservation of the mechanism of function between frog and human Hotairm1</i>	143
2.6. <i>Conservation of Hotairm1 across Chordata</i>	144
3. Conserved regulation of Hox Cluster in Chordata	146
3.1. <i>Analysis of Pdx and Cdx knockouts during development</i>	146
3.2. <i>Gut-enrichment in both datasets</i>	147
3.3. <i>Conservation of the anterior and middle Hox cluster regulation via Cdx</i>	148
Conclusions	153
Annexes	157
Article AI	157
Article AII	173

Summary

In this thesis we have covered the importance of finding underlying conservation events to better understand the regulatory mechanisms of important development orchestrators like the Hox cluster. As an example of these non-evident conservation, we have shown two cases, as described below.

The first case studied, after developing a software able to detect homologous long noncoding RNAs by means of microsynteny analyses, is the conservation of *Hotairm1* in Chordata. For assessing the homology of this lincRNA, first we had to identify the lincRNA fraction within the *B. lanceolatum* transcriptome. With a reliable lincRNA dataset, we used our pipeline, LincOFinder, to identify orthologs between human and amphioxus through microsynteny. After the identification of *Hotairm1* as one of the lincRNAs with conserved microsynteny, we used *Xenopus* as a proxy to analyse the homologies in the expression and the function. We had to proceed this way due to the difficulties associated with the inhibition of genes in *B. lanceolatum*, and the unavailability of expression patterns for *Hotairm1* in the bibliography. After we successfully characterised *Hotairm1* expression in amphioxus and *Xenopus*, we injected morpholino oligonucleotides to target and inhibit the splicing of *Hotairm1* to promote an isoform imbalance. Through the phenotype obtained and the performing of qPCRs, we were able to deduct the mechanism of *Hotairm1* and successfully relate this mechanism with the one described in human cells. With all the data obtained we were able to strongly suggest that the amphioxus *Hotairm1* is homologous to the *Xenopus* and human *Hotairm1*, thus being conserved in most of the lineages within chordates.

The second case studied was the conservation of the regulation of the Hox cluster mediated by *Cdx*. When analysing the *B. floridae* knockouts of *Cdx* and *Pdx* obtained using the TALEN technique, we found a severe phenotype of the developing larvae in *Cdx*^{-/-} and a mild phenotype in *Pdx*^{-/-}. The *Cdx*^{-/-} phenotype consisted in the disruption of posterior gut development, as well as an underdevelopment of the postanal tail, coupled with a non-opening anus. When looking at changes in the expression of the Hox cluster in this *Cdx*^{-/-} embryos, we found collinear misregulation of the expressed Hox genes, with the most anterior Hox cluster genes upregulated, and the most posterior ones downregulated. This is very similar to findings seen in triple morpholino knockdowns of the *Cdx* genes in *Xenopus*, indicating that in both, *Xenopus* and amphioxus, *Cdx* is regulating the Hox cluster through a homologous mechanism.

Acknowledgements

Bueno, llego el mejor momento de escribir la tesis, el de dar las gracias. Siempre me gusta leer los agradecimientos de las tesis que ojeo, porque es como mirar una foto hecha de palabras del momento en el que se escribió. Así que espero sacar a todo el que pueda en mi foto, y si alguien que esperaba salir no se encuentra, será porque se me ha quedado el cerebro frito. Voy a seguir un orden tan organizado como yo, así que será totalmente aleatorio.

Eso sí, empezare por mi familia. En general ya sabéis que os quiero, y que sin vosotros no habría podido hacer la tesis blibliu, pero hay cosas que son dignas de reseñar. Como por ejemplo tus visitas brother, que aun siendo yo el peor organizador de planes del mundo sigues haciendo de tanto en tanto y que me encantan. También las tuyas sister (y Edu) que, aunque menos habituales, siempre acaban incluyendo un room escape, o un laser tag, o un apartamento con extrañas habitaciones en Badalona... lo típico. También me gustaría agradecer aquí las discusiones (de lo que sea) que tengo contigo mamá, que me obligan a estar siempre rápido de mente, y que cuando me aburra en el larguísimo viaje de 15 minutos de vuelta a casa, estés en el teléfono. A ti papá te debo mi pasión por los videojuegos, por los ordenadores y por la tecnología en general, pero mas que por todo eso (que no es poco) quiero darte las gracias por enseñarme que prácticamente todo tiene arreglo si le dedicas el tiempo y la energía suficiente y que desmontar algo es la mejor manera de aprender como funciona. Lo dicho, os quiero, incluida a ti Blanca (no prima, tu vas luego) que aunque aun no has nacido y tardaras bastante en poder leer esto quiero que sepas que tu tío pensaba mucho en ti durante su tesis. Blanca (ahora si primi), gracias por tu sonrisa y por encontrarte conmigo en Mallorca; Pablo, gracias por tu humor negro, siempre consigues hacerme reír. Al resto de mi familia, sois demasiados para ponerlos a todos, pero sabéis que os quiero y que me encanta veros cuando voy de visita. Pondría mas cosas, pero no es plan de hacer unos agradecimientos mas largos que la Introducción no?

Ahora toca el turno a la gente del lab. Bea, tu primero. Gracias por haberme enseñado desde que entre en el lab a hacer las cosas bien. No solo en ciencia, sino en la vida en general. Mother of volumes, sexadora de amphioxus, la que no deja que el laboratorio arda, gracias por todo. Kike, gracias por... esto... tu... un día me hiciste un café! XD obviamente es broma gil, gracias por todas las discusiones que hemos tenido, sobre moral, economía, política y sobretodo ciencia y por las veces que me pones contra las cuerdas (dialécticamente, aquel pulso lo gane yo). Solo por haber aguantado todas las canciones que he hecho con la palabra "Kika", te has ganado un rincón de cielo. Ari ari adnaaaaa!! Te toca. El lab esh un lugar mash aburrido deshde que te fuishite... Graciash por las cancionesh de frozen, las nochesesh eshcribiendo el master y no she, por ti, en general ashin a lo loco. Aina, para tu els agraiments seran en català. Ets la persona que mes m'ha fet voler parlar-lo bé. Gracies per el "sexí com tu", per acollir-me en Cambridge i per la connexion mental absoluta. Et trobo a faltar moltíssim i a més ja ningú li demana a Kika si es felíç. Coral, obviamente gracias por peluquerias Coral, y bueno por Coral Industries en general. Y por los masajes! El próximo me lo das en Ámsterdam no? Vero, ahora que me acerco al final de la tesis puedo decir que esto si que es la gloria. Miquel, a ti gracias por toda la ciencia y las conversaciones del brujo (y sus cartas). Eudi, shaval, para ti me faltan palabras. Gracias por confiar en mi para tus proyectos, ya sean cienteficos o científicos. Esto se acaba, y no se me ocurre mejor persona junto a quien acabarlo. Mi compañero de las antípodas, let's finish this! Al resto de gente del lab, ya sean del pasado (Maria, Demi, Nidia, Roser, Sara, Nieves, Sisco, Jose, Sheila, Jordi, Elena italiana, etc) o del presente (la otra Maria, Dani y Susana) gracias por haber compartido esto conmigo y no haberme estrangulado en el proceso. Kike no era el único que tenia que aguantar las Kika melodías. O frozen. O el hielo seco. El chiste de las patatas, etc, etc, ya me entendéis. Hacer la tesis aquí ha sido genial en gran parte por todos vosotros, así que lo dicho, gracias.

De los otros labs, las moscas, esos bichos tan majos, gracias a José por seguirme el ritmo silbando entre otras cosas, y a Qi por dejarse adoptar. La china pesará como un gato pequeño, pero tiene la mala leche de un oso rabioso, o eso dicen, porque conmigo eres puro amor. Al resto de moscas, no he coincidido tanto con vosotros, pero siempre que he necesitado algo me habéis ayudado así que gracias también. Marta, la reina de las colaboraciones. Gracias por tu dominio sobre los Xenopus, y por creer en mi lo suficiente como para ponerte a inyectar un morpholino de un lcnRNA. Podrian haber salido muchas cosas mal, pero decidiste intentarlo

Acknowledgements

anyway y estoy encantado de compartir parte de esta tesis contigo. Y a Nuria, porque tu para mi simbolizas todo lo que tiene que ser el PhDDay. También quiero agradecer a Raquel por aquellas practicas en las que los alumnos dejaban de existir y por tu dedicación hacia el resto de doctorandos.

Ahora tocan los esclavitos. En total he tenido cinco, y de todos ellos he aprendido algo. Primero Lorena, gracias, porque tu me enseñaste a enseñar bien, y por como me pedias que te explicase el porqué de las cosas que hacías. A Celia, gracias por estar conmigo en uno de los momentos clave, el clonaje de los primeros lncRNAs conservados. El único que te salio no era un lncRNA al final, pero lo sigo usando como control así que gracias. To Jan, thanks for helping me to design the LincOFinder pipeline. I'll learn Python someday, I promise. And thanks also for staying overtime on your lab to help me with the paper. Michal, gracias por esforzarte tanto en todo lo que haces, y por ser la otra única persona del lab que conoce al brujo. Y por último Claudia. La verdad es que conocerte no entraba en mis planes, pero no me puedo alegrar mas de que decidieras unirme a la expedición amphioxil. Quizás los experimentos no nos han salido todo lo bien que nos gustaría, pero has estado a mi lado todo el tiempo que ha durado la escritura de la tesis (distrayéndome sobre todo), y por eso te doy las gracias. Bueno por eso y por los macarrones con salchichas.

Y sigo con los jefes! A Jordi, obviously, por darme esta oportunidad. La verdad es que después de lo mal que lo hice en una de tus clases de master, no se que es lo que viste en mi, pero te agradezco todo lo que has hecho por mi, que es bastante. Desde llevarme a un congreso internacional recién empezado el PhD, hasta confiar en que tarde o temprano me acabaría saliendo la ish de Hotairm1. A Manu, por toda la guidance que me has dado y tus respuestas quasi-instantaneas. No hay duda que no hayas resuelto y tu paciencia conmigo ha sido infinita. Gracias de verdad. And finally Peter, thank you for making me feel like at home when I was in Oxford. I have really enjoyed my time there, and that is because of your kindness and joy. Oh! and also thanks to Roddi, Emily, Amy, Tom, Serena, Sonia, Vas, Gui, Matthew, etc. You are great guys. I'd like to say more, but I'm close to the 2 pages now. I hope to see all of you soon!

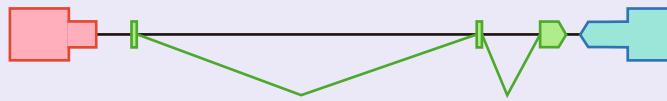
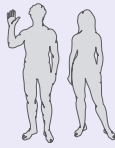
Bueno, ahora los de Tudela (y uno de Pamplona). Me queda poco espacio, así que hare un agradecimiento general a la cuadrilla, gracias por estar ahí. Da igual el momento, la distancia, lo que sea, siempre estais ahí para lo que necesite, y espero estarlo yo también siempre. A veces podemos discutir por alguna cosa que otra, pero si discutimos es porque nos importamos los unos a los otros, y quiero que sepáis que os llevo siempre conmigo dondequiera que este. Aquí aprovecho para incluirte, Mikel, porque eres como de la cuadrilla para mi.

Y creo que solo me quedan dos personas, que ahora mismo son las dos mas importantes en esta etapa. Primero Gorka, gracias por haber compartido todo este tiempo conmigo. 5 años viviendo juntos se dice rápido, pero es mas de lo que duro la carrera. La verdad es que es difícil expresar todo lo que te debo, o mejor dicho lo que Nala y yo te debemos, sin llenar 3 hojas mas. Gracias por todo lo que hemos compartido, por todas las partidas al Division, por los juegos abandonados a medias, por las series, películas, animes. Gracias por escuchar mis sueños enrevesados, y por ver capítulos repetidos de los simpsons. Por el phenomena y los regalos de cumpleaños. Por la decoración cutre (pero nuestra) del piso en Navidad y por los miercoles al Balmes. Por el verano de cine a la fresca, y portaventura. Por Legion, Leftovers, Always sunny, etc. Voy a parar ya porque podría seguir y no es plan, pero espero que pilles la idea.

Y por ultimo, y ahora si que si acabo de verdad, Elena. Que puedo decirte que no te haya dicho ya? Eres lo mejor que me llevo de esta tesis. Casi casi el motivo por el que decidí, pasara lo que pasara, quedarme en este laboratorio. Sabes que me conquistaste desde que dibujaste aquella Xbox en el DrawSomething aunque ya me hubiera fijado en ti cuando me diste clase. Me da pena dejar de vivir con Gorka, pero me muero de ganas de empezar a vivir contigo. Gracias por hacerme querer ser la mejor versión de mi mismo, y por enseñarme que aunque a veces los planes no salgan bien, casi siempre se pueden rescatar. Gracias por haber estado a mi lado todo este tiempo y por servirme de inspiración para darlo todo en cada cosa que hago. 88. Te quiero.

PS: A Nala y Ponyo, miau miau. <3

Introduction



Introduction

1. Evolution and development

1.1. Dawn of the Evo-Devo

Through the 20th century the evolutionary trend was the synthetic theory (Mayr and Provine 1980). This theory, also referred as “modern synthesis”, put Mendelian genetics, ideas of natural selection, population genetics, and macro and micro evolution events together. Yet it didn’t consider any of the discoveries and breakthroughs that took place in developmental biology research, as far as to consider an organism a direct reflection of its coding genes (Mayr and Provine 1980). However, in the last third of the 20th century, a new approach arose to include developmental biology in the framework of evolution (Müller 2007). Instead of focussing only in the changes within coding genes in order to explain different phenotypes, this new approach, named Evo-Devo, explains them through the variation in regulation of genes during development, where a slight change in spatiotemporal expression of one gene can have ripples that affect the adult animal in ways that the synthetic theory could not fully explain.

From this point of view, although mutations within the coding sequence of a gene are still relevant and can impact in several ways the phenotype of an organism (Cheng *et al.* 2009), mutations in regulatory regions such as enhancers, or transcription factor binding sites that have a role during development gain new relevance. Moreover, they can be identified as responsible for some of the evolutionary novelties like the powered fly present in bats (Sears *et al.* 2006), or the repetition-based body plan of the snakes (Woltering 2012). In addition, this view opened the door for using additional features to determine the evolutionary history of an organism, beyond just the sequence analysis.

1.2. Changes vs Conservation

When studying evolution, by the own definition of the word, one would look for changes, comparing two species and stating the observable differences. Although while this is true and the main aim of Evo-Devo, it is not the only item of its research. Another one that is of interest here is the conservation, as in the light of conservation, we are able to determine what has really changed and vice versa.

One of the striking discoveries in Evo-devo, thanks to molecular genetics, was the concept of developmental gene toolkit. As we have stated, the different phenotypic

traits between organisms are mostly due, not to mutations in coding regions, but to changes in regulation during development. This regulation is done by a deeply conserved set of genes in most of metazoans, known as the developmental gene toolkit (Gerhart and Kirschner 2007; Newman 2006). Most of those genes code for transcription factors, cell receptors, ligands, and morphogens and generally, a mutation in these genes turns out to be deleterious, or to produce a heavily impaired organism. Changes in their promoters or on their binding site, however, can result in an altered spatiotemporal expression that gives rise to a modified trait, often without a dramatic change in the viability of the organism (Sears *et al.* 2006). However, a change so disruptive as to produce an evolutionary novelty although it could still be viable, would wreak havoc in some of the essential regulatory networks of development, unless a concomitant gene duplication took place. Hence, the easiest way to overcome the selective pressure and develop new or modified traits is through conservation, duplication, and change (Ohno 1970). This concept is the core of the Duplication-Degeneration-Complementation (DDC) (Force *et al.* 1999) and the Duplication-Degeneration-Innovation (DDI) (Jimenez-Delgado *et al.* 2009) models, to explain the linkage of gene duplication to changes in gene regulation.

Some genetic networks are used in such a myriad of processes and are so fine-tuned that a strong selective pressure is placed upon them. Still, changes in the regulation and in the sequence can take place, and as long as the structure of the network is not heavily affected, they will not produce a dramatic effect. Of course, there are exceptions, either due to a better gene network overtaking the role of the ancient one, or due to the survival by “chance” or by absence of selective pressure, of an organism after a dramatic genetic event. Nevertheless, in most of the cases, understanding the process of generation and the selective pressures of these deep conservations would be helpful to gain better insights in the mechanisms of evolution.

1.3. *Amphioxus*

Overview

Since its scientific discovering in the 18th century (Holland and Holland 2017) the cephalochordate amphioxus has been an intriguing animal. First considered to be a mollusc (Pallas 1794) and later the simplest kind of vertebrates (Yarrell 1836), the most

Introduction

accurate description in those early years came from Goodsir (Goodsir 1844) who proposed this animal to be somehow in between invertebrates and vertebrates.

It has been thoroughly and continuously studied except for a half-century gap in the 20th century (Holland and Holland 2017), and recently has found a new spotlight thanks to the advances in high throughput sequencing and analysis techniques. Nowadays, we can place it pretty accurately within the phylogenetic tree of metazoans, being in a basal position inside the phylum Chordata. This comes along with the fact of possessing a pre-duplicative genome, which retains most of the ancestral chordate genome characteristics before the two round of full genome duplication that occurred at the origins and early evolution of vertebrates (2R) (Dehal and Boore 2005) (Figure I1). These perks make amphioxus to be currently and thoroughly studied in development research (Garcia-Fernández and Benito-Gutiérrez 2009; Escriva 2018), and to be considered a great model organism to study the transition from invertebrates to vertebrates in the frame of Evo-Devo.

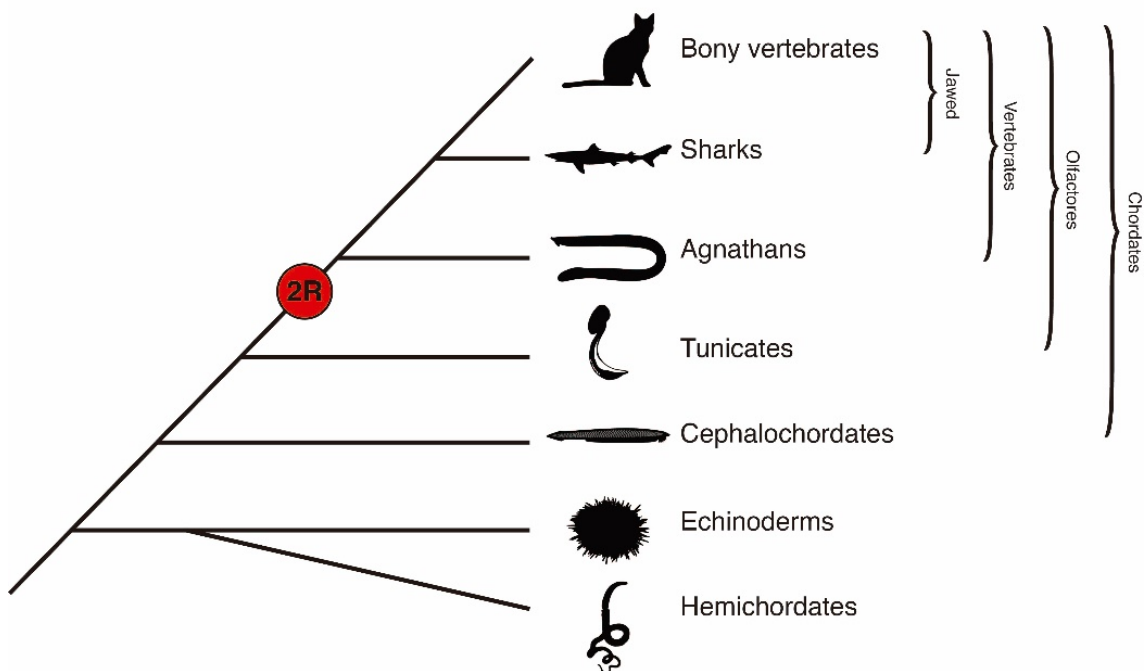


Figure I1. Simplified phylogenetic tree showing the position of cephalochordates in the context of the 2R. Adapted from Garcia-Fernandez and Benito-Gutierrez 2009

Development

Amphioxus development has been intensively studied classically (Bertrand *et al.* 2017; Conklin 1932; Mansfield *et al.* 2015). After the fertilization, which takes place externally in the sea, the chorion of the newly form zygote expands (Figure I2 A), and cells start to rapidly and synchronously divide until the embryo is composed of 128 cells, when a spherical and hollow blastula is formed. When gastrulation begins, the vegetal pole flattens and starts invaginating, leaving at the end of gastrulation two cell layers (Figure I2 D). These surround the archenteron which will form the future digestive system, open to the exterior through the blastopore. The mesoderm and endoderm are within the internal layer, and the external layer contains the ectoderm with ciliated cells that allows the embryo to move inside the chorion.

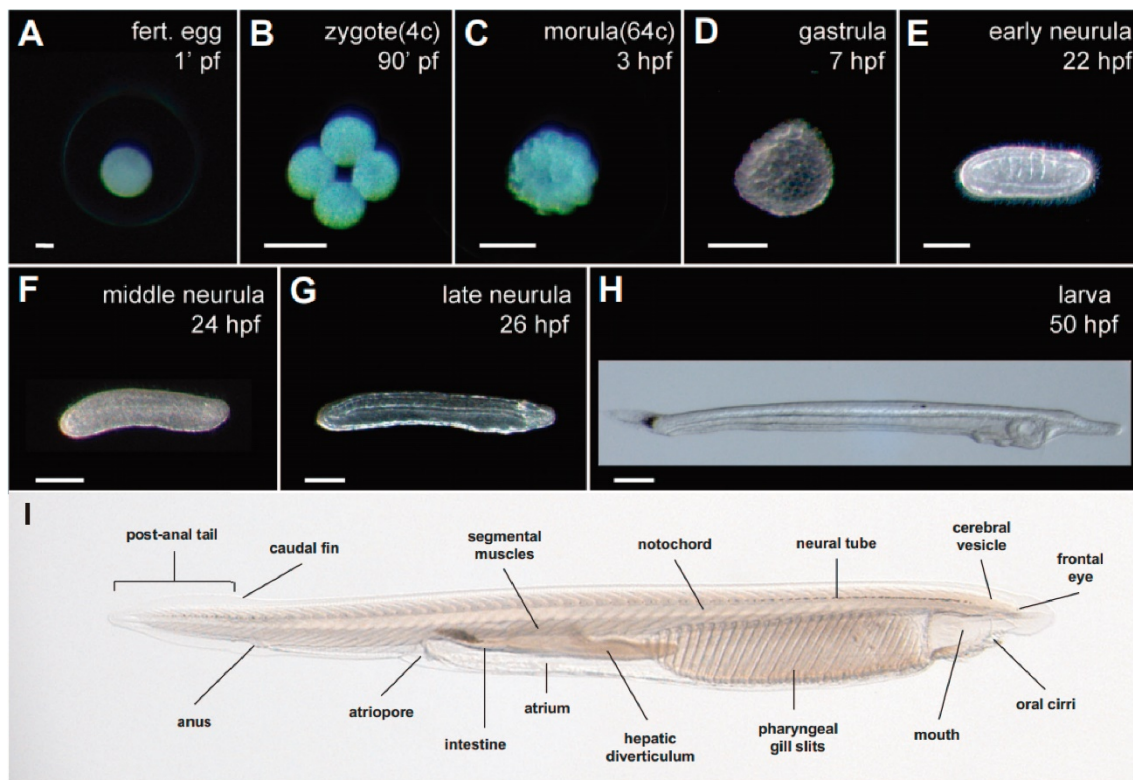


Figure I2. Developmental stages of *Branchiostoma lanceolatum* embryos grown at 19°C in Barcelona. Anterior to the right. (A) Fertilized egg; (B) zygote at four cell stage; (C) morula at 64 cells stage; (D) gastrula; (E) early neurula stage; (F) middle neurula stage; (G) late neurula stage; (H) larva stage with one gill slit; (I) Adult animal. Adapted from Garcia-Fernandez *et al.* 2009.

When gastrulation ends, the neurulation process initiates with the embryo breaking the chorion and swimming freely (Figure I2 E). The V-shaped dorsal neuroectoderm elongates in postero-anterior direction as the developing neural tube progressively closes thanks to the upwards extension of the neural labia in the same direction. This

Introduction

process leaves a small opening in the anterior end of the neural tube known as neuropore. The internal cell layers rearrange to give rise to the notochord and the somites. The first pairs of somites will be formed from the invagination of the archenteron while the rest of the somites will come from the tail bud. The ventral epithelium of the archenteron will fold over itself to give rise to the digestive tube.

The stage after the neurula is the larva (Figure 12 H). In this stage, the neural tube, the somites and the notochord keep developing through the antero-posterior axis. The most notable changes of this stage are the mouth development located in the left side of the larva, the formation of the cerebral vesicle, the anus opening, and the formation of the gill slits. Before the transition to juvenile and final adult stages, a metamorphosis takes place where most of the asymmetries of the larva are addressed, leaving an almost symmetric amphioxus juvenile.

Functional genomics

Despite being able to study some of the functions of genes in amphioxus using drugs dissolved in the growing medium of the developing amphioxus (Escriva *et al.* 2002), direct injections in the oocyte to produce knock-downs, or transient expression or even a knock-out animal remained elusive until the last decade. Alternatives as to feed them microRNAs or dissolve morpholinos in the medium were tried without success. But after the first successful injections of fluorophores in the unfertilized egg (Garcia-Fernandez *et al.* 2009), successive milestones were reached towards experimental Evo-devo in amphioxus. These milestones include the injection of mRNAs that were properly translate into proteins (Hirsinger *et al.* 2015), the injection of reporter constructs to properly study the role of regulatory regions (Liu *et al.* 2013) and finally amphioxus directed mutagenesis via TALEN method (G. Li *et al.* 2014). This, put together with the ability to close the reproductive cycle of amphioxus in the lab and obtaining mature fertile adults from in-vitro fertilized embryos, allowed researchers to produce stable heterozygotic knockout lines to study the effects in development of the absence of key genes.

In this work, from the several species of amphioxus, we will focus on two, *Branchiostoma lanceolatum* and *Branchiostoma floridae*. Their speciation took place approximately 190 million years ago (Cañestro *et al.* 2002) and their main morphological differences are the slightly bigger size and faster development of *B. floridae*.

1.4. *Xenopus tropicalis*

X. tropicalis is an African clawed frog used frequently to study developmental biology. The endurance of the eggs that allows to perform dissections and injections and the huge numbers in which they are produced make this animal a great model for studying gene function (Harland and Grainger 2011). The development of *Xenopus* (Figure 13) is normalised in the Nieuwkoop and Faber atlas (Nieuwkoop and Faber 1994). After fertilization takes place in the animal pole, a series of rapid synchronous cell divisions ends up forming the blastula. After the mid-blastula transition, the embryo stops relying on the maternal mRNA and starts expressing its genes in what is known as the embryonic genome activation. Furthermore, cell division is desynchronized (Schmitt *et al.* 2014), and this results in the gastrulation of the embryo, with the invagination of the dorsal cells and epiboly of the cells from the animal pole. This will establish the ectoderm, mesoderm and endoderm (Gilbert and Barresi 2017) that will give rise to the different tissues of the developing embryo. The next stage is the neurulation, where the neural plate located in the dorsal part of the embryo folds forming the neural tube. In addition, the cells on the edge of the neural plate will become the neural crest. Finally, the rest of the organs will be developed until the tadpole is fully formed.

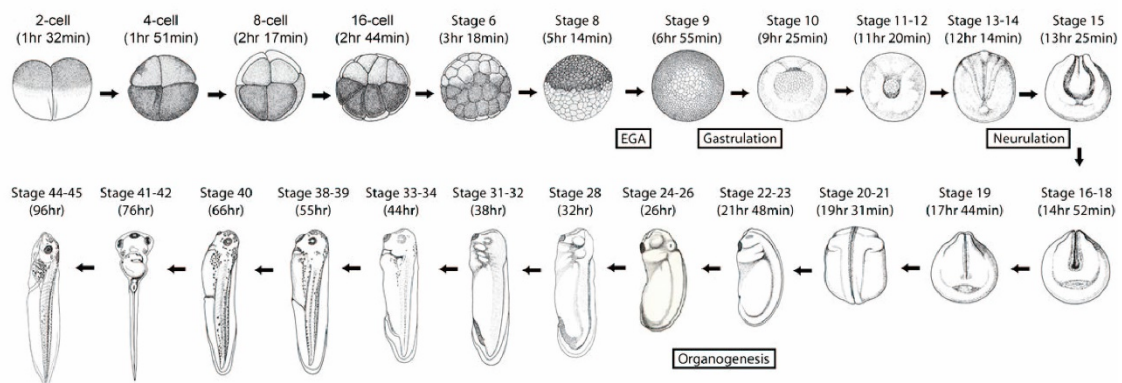


Figure 13. Developmental stages of *Xenopus tropicalis* grown at 23°C; Times expressed in hours post fertilization; (EGA) Embryonic Genome Activation. Adapted from Tan *et al.* 2013.

In this study we will use one of the most established methods of knocking-down genes in *Xenopus*, the injection of morpholino oligonucleotides (Eisen and Smith 2008). These oligonucleotides are designed to bind specifically to a target. In most of the cases, the morpholinos are designed to bind to a mRNA preventing it from being used in most of the biological processes, like translation into protein, or preventing it from being spliced if the morpholino targets the splice site of the pre-mRNA.

Introduction

1.5. Hox Cluster

One of the keys to understand Evo-devo is in the homeotic genes, called this way due to the displacement of body parts (homeosis) that the mutations in these genes caused in *Drosophila*. They are master regulators of the development in metazoans, although their origin is set sometime during eukaryote evolution, due to the absence of any related gene in prokaryotes (Bürglin 2001). Probably, the most important set of homeotic genes during development is a set of paralogs, in most of the cases presented as a cluster, known as the Hox cluster.

The Hox cluster is formed by a set of genes that have in common a Homeobox domain. They were discovered in *Drosophila* by Lewis (Lewis 1993) who called them *Antennapedia* genes (as it changed an antenna for a leg). In *D. melanogaster* there are 8 Hox genes separated in two cluster (ANTP and UBX) and their discovery opened a whole new research field, due mainly to their role in the antero-posterior axis determination during development.

The Hox genes are highly conserved in most of the bilaterians, and they can be related to similar functions in organisms as different as *D. melanogaster* and *H. sapiens* (Hueber *et al.* 2010). But what set these genes apart are the spatial and temporal collinearity present in the cluster. The first members or anterior Hox genes are expressed earlier during development and in a more anterior region than their counterparts at the end of the cluster (Figure I4).

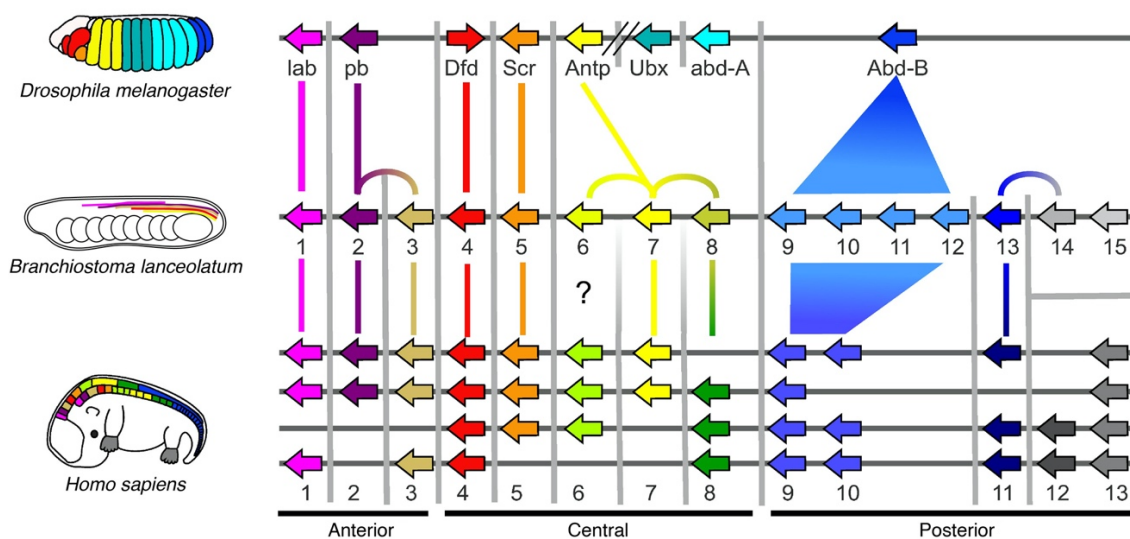


Figure I4. Scheme of the Hox cluster spatial collinearity in fly, amphioxus and human with their corresponding orthologies. Adapted from Garcia-Fernandez *et al.* 2009, and Hueber *et al.* 2010

The regulation of the hox cluster is a tight process where a group of different elements play a role, like Retinoic Acid (RA) (Balmer and Blomhoff 2002). An excess of RA produces a posteriorisation of the developing embryo due to a variation in the expression limits of Hox genes. This is because RA response elements located in the regulatory regions of Hox genes are activated by RA, thus being ectopically expressed in the case of an excess of RA (Balmer and Blomhoff 2002). Long noncoding RNAs (lncRNAs), like *Hotair* (Rinn *et al.* 2007), that closes the chromatin of the HoxD cluster in mammals through the polycomb repressing complex II (PRC2), also play an important role in the regulation of this cluster. In addition, the non-coding elements surrounding the sequences of Hox genes are key to understand their regulation and evolution, as the changes in these sequences are the ones that end up building the differences of expression between the different animals. Finally, other homeobox genes outside the Hox cluster can have a role in its regulation, as is the case of the Parahox cluster gene, *Cdx*. As reported by Marletaz *et al.* (Marlétaz *et al.* 2015) the knock-down of the three *Cdx* paralog genes in frog produces a collinear downregulation, with the anterior Hox genes upregulated, the middle ones mildly downregulated, and the posterior ones heavily downregulated (Figure I5). This may mean that *Cdx* interacts as a sliding scale, either by remodelling the chromatin, or by direct control of the transcription.

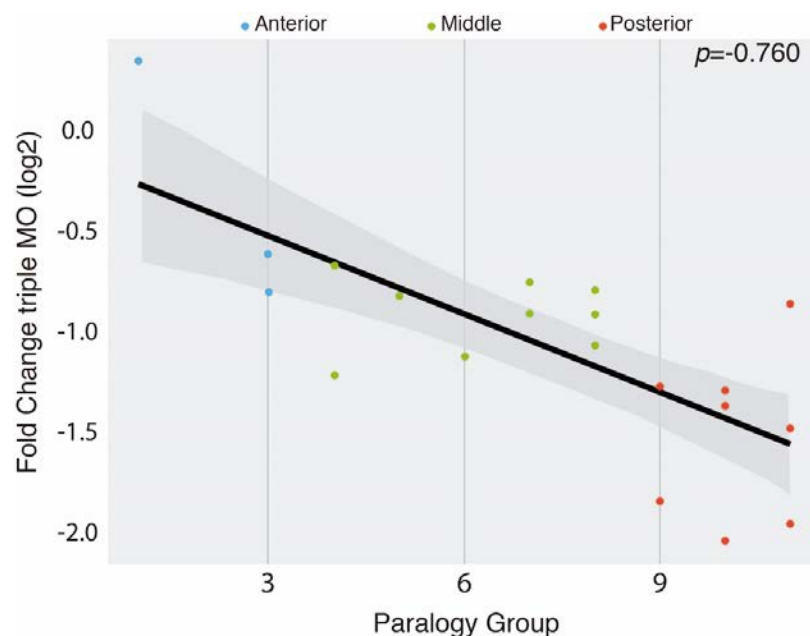


Figure I5. Fold-change in Hox gene expression caused by triple MO injection plotted against paralogy group assignment; each data point represents one Hox gene. Only genes with statistically significant change in expression are included. Colours denote anterior (blue), middle (green), and posterior (red) paralogy group assignments, assigning group 3 to anterior. Adapted from Marletaz *et al.* 2015.

Introduction

1.6. ParaHox Cluster

The ParaHox cluster was described by Brooke, Garcia-Fernández and Holland (1998) in *B. floridae* more than twenty years ago. They found the presence of three homeobox genes, *Cdx*, *Xlox* (*Pdx*) and *Gsx* sharing a region of 32Kb in the same chromosome, hence forming a cluster. The ParaHox cluster evolutionary origin is tightly linked to the Hox cluster, as there was an ancestral duplication forming two twin clusters, one of them giving rise to the ParaHox cluster and the other to the Hox cluster (Garcia-Fernández 2005) (Figure 6). The spatial collinearity was already defined in these ancestral twin clusters, as there is also a collinearity in the expression of ParaHox cluster that can be related with the well established collinearity existing in the Hox cluster.

In jawed vertebrates, after the 2R event, the most common arrangement of these genes is three *Cdx* paralogs, two *Gsx* and one *Pdx* (Figure I6). They are distributed in four chromosomes with one of them containing the full three-gene cluster, and the rest of the genes disaggregated. Their functions, as seen by disruption of the different genes, are related to neural and gut development. *Gsx* genes are expressed in the developing brain and have a role in differentiating the lateral ganglionic eminence neuronal subtypes (P. W. H. Holland 2013; Pei *et al.* 2011); *Pdx* is expressed in the midgut where it participates in the development of the pancreas and the development of the proximal duodenum (Hideaki Kaneto *et al.* 2007; A. M. Holland *et al.* 2013) whereas in the adult it is key to maintain the function of insulin-secreting beta cells; the three members of the *Cdx* gene family in jawed vertebrates have very similar roles, with an expression located in the posterior part of the embryo, and a role in the posterior patterning of the embryo, including the posterior gut (Marlétaz *et al.* 2015).

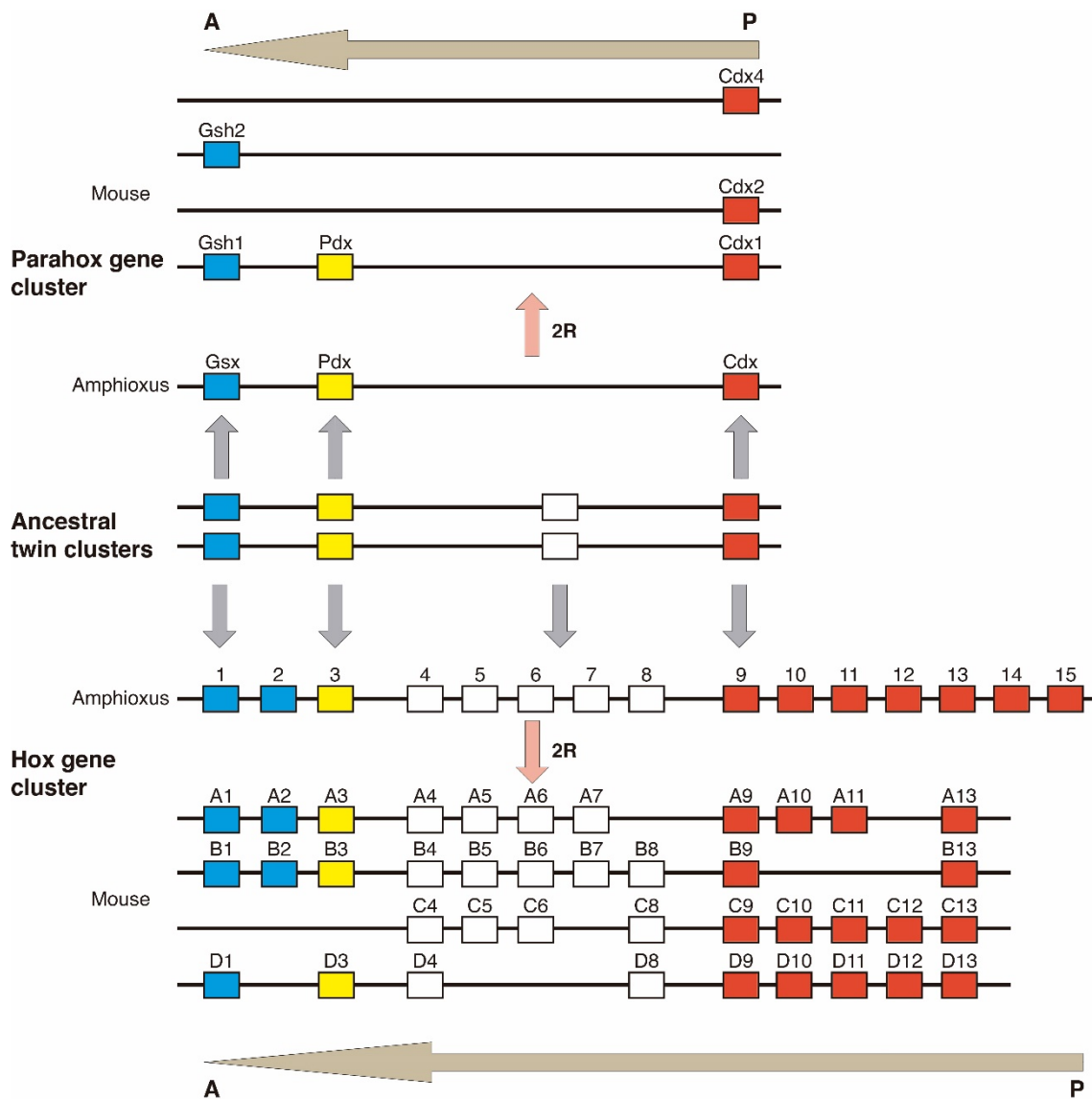


Figure I6. Scheme of the Hox and ParaHox cluster evolution through the 2R from ancestral twin clusters. Horizontal arrows denote polarity of spatial collinearity (A, anterior; P, posterior). Adapted from Brooke, Garcia and Holland 1998

Introduction

2. Long non-coding RNAs

2.1. Overview of lncRNAs

History

From the central dogma of molecular biology established in 1958 (Crick 1970), our view on the role that the RNA-based molecules have in the regulation of cell processes has changed significantly (Figure 17). Thanks to the advances in molecular biology, non-coding RNAs are not referred as “junk” DNA anymore, and their importance is properly considered.

lncRNAs are characterised by being more than 200nt in length and having a normal transcription process like a coding gene. They have the same features of coding mRNAs: (i) transcribed by RNA polymerase II, (ii) spliced and (iii) 3' polyadenylated. They just lack a coding open reading frame, and therefore they don't produce a protein (Jarroux *et al.* 2017). The first eukaryotic lncRNA discovered was the imprinted gene *H19* (Cai and Cullen 2007), expressed only maternally. Although it was evident that *H19* had a role during development, the molecular function remained a mystery until *Xist*, another lncRNA involved in dosage compensation, was functionally characterized (Cai and Cullen 2007; Kallen *et al.* 2013). The functional characterization of *Xist* is a milestone for lncRNAs research on its own. Until then, the lncRNAs (when considered as gene regulators) were assumed to act only upon their close genetic neighbourhood. But *Xist* is capable of imprinting the whole X chromosome in mammals (Rastan 1994) thus has a key function for the proper functioning of the organisms where it is present. These two examples demonstrated the versatility of a single lncRNA in the regulation of complex regulatory circuits and stimulated the community to put more effort in the characterization of lncRNAs.

One of the fruits of these efforts was the discovery of *Hotair*. Rinn *et al.* (2007) found a lncRNA located in the HoxC cluster whose regulation affected in trans the Hox genes of a different cluster in a different chromosome, the HoxD cluster. They found that *Hotair* coupled with the Polycomb Repressive Complex 2 (PRC2) silenced gene expression of the HoxD cluster, being the first identified trans-acting lncRNA.

Introduction

Origin and characteristics of lncRNAs

There are several ways lncRNAs can be established in the genome, like for example, by the process of pseudogenization. The accumulation of several mutations that break translation and the posterior co-option of the still transcribed pseudogene, can produce lncRNAs that are not apparently homologues to their original gene anymore. *Xist* is a good example of this process as it is derived from an ancestral *Lnx3* gene in eutherians. (Duret *et al.* 2006). Another event that can give rise to a new lncRNA gene is the co-option of RNA-derived transposable elements. These transposable elements can provide lncRNAs with transcription start sites (TSS), splicing sites, etc, and can also provide functional secondary structures (like protein binding sites) (Kapusta *et al.* 2013). Finally, *de novo* lncRNAs, although uncommon, can be formed from acquisition of transcriptional regulatory elements in a previously non-transcribed intergenic region, like in the case of *Poldi* (Heinen *et al.* 2009).

We have mentioned that new lncRNAs may be co-opted into new functions, like imprinting the X chromosome in the case of the *Lnx3*-derived lncRNA, *Xist*. But how do they act without being translated into a protein? The answer to this key question in most of the cases is through their secondary structure. lncRNAs (as mRNAs do (Hall *et al.* 1982)) fold themselves into a secondary structure that allows them to interact by themselves with several other elements within the cell. They can have several functions thanks to this folding, like acting as a scaffold for protein complexes (Fang and Fullwood 2016; X. Li *et al.* 2014) or binding to chromatin remodelling elements to direct them to a specific region, like in the case of *Hotair*. This makes lncRNAs really versatile regulators, as they can even have several functions, like the previously mentioned *H19*, a good example of a multitasking lncRNA. In addition, their prompt readiness (some lncRNAs can start having function right after being transcribed (Cloutier *et al.* 2016)) and lower stability than the mRNA, makes them sharp regulators of the processes where they have a role. This could explain the overall lower transcription levels of lncRNAs when compared with coding mRNAs, as they can be transcribed, act, and be degraded in a very short time window, showing just a peak of expression through a timeline, and in some cases keeping their overall expression level low.

Evolution of lncRNAs

If we base our study of lncRNA evolution in the study of the nucleotide sequence, we will find that in most of the cases they are not conserved, except in some closely related species. Due to their intrinsic characteristics, lncRNAs do not seem to have a selective pressure placed upon the nucleotide sequence at the same level that coding genes have. This does not mean, however, that they are not conserved through evolution. For example, *Mhrt* is a well-known lncRNA characterised in humans and mice. Although its sequence conservation between these two species is around thirty percent, they are known to be homologs with the same function in both species (Han *et al.* 2014). When comparing intergenic lncRNAs (lincRNAs), the conservation is somewhat higher, but still almost absent when comparing slightly distant species like human and zebrafish. Ulitsky *et al.* (2011) showed how *Megamind*, a lincRNA with less than 100nt of conserved region between human and zebrafish had an orthologous function. They were able to knockdown the expression of this lincRNA in zebrafish and then rescue the WT phenotype injecting the human *Megamind*. These findings probed that the conservation of lncRNAs is “multidimensional”, as defined by Diederichs (2014).

We may assign up to four dimensions of lncRNA conservation (Figure I8). The first one is the sequence conservation, as in the coding genes. Although lower, it still can be used to find homologous lncRNA in closely related species. The second dimension is the conservation of their secondary structure. As we have point previously, lncRNAs can produce secondary structures that are linked to their function, making conservation of the secondary structure something to be expected. Unfortunately, *in silico* secondary structure predictions are not fully reliable yet for comparison analyses in most of the cases. Besides, there are studies indicating that there could be no strong secondary structure conservation at all, leaving this dimension at the same level of the sequence level (Rivas *et al.* 2017). The third dimension is referred as the conservation of function. The problem is based in how to properly state if two lncRNA with the same function are homologous or the result of an evolutive convergence event. If we found a lncRNA with the same function present in two distant species without sequence or structure conservation, we can still find evidence of an evolutionary common origin in the fourth dimension, the conservation of locus synteny. This fourth dimension can be further stretched if we look at microsynteny conservation (Irimia *et al.* 2012) as we have done

Introduction

in this study. The conservation of just few genes loci in distant species is a good indicator of selective pressure that can maintain the function of a lncRNA across great evolutionary distances.

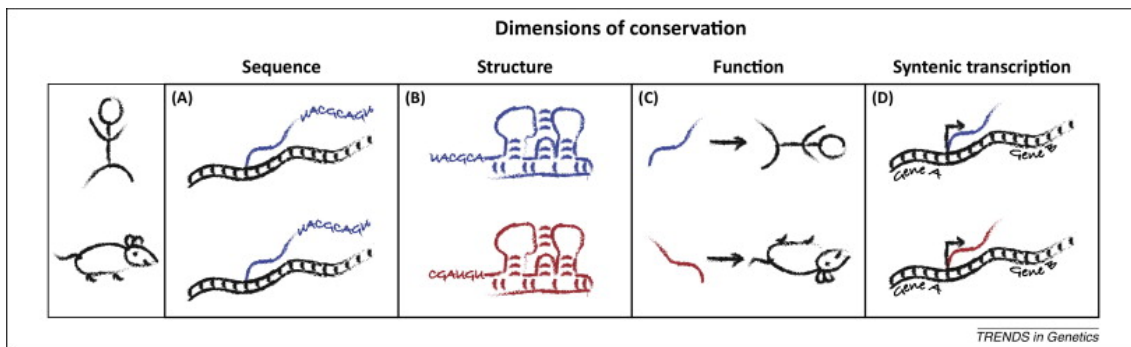


Figure 18. Graphic representation of the four dimensions of lncRNA conservation from Diederichs 2014

Classification

Although there are several ways to classify a lncRNA (Jarroux *et al.* 2017), in this study we will focus on two of them. The first one regards their location within the genome relative to their coding neighbours (Figure 19). Under this classification, lncRNAs can be: (i) intergenic, also referred as lincRNAs, when they don't overlap any other coding gene; (ii) intronic, when they are located within the intron (or introns) of a coding gene in the same strand; (iii) antisense, located in the same locus as a coding gene but in the opposite strand; (iv) bidirectional, similar to the intergenic ones, but in the opposite direction of a coding gene, where they may partially overlap with the 5' region of said coding gene, and sharing a bidirectional promoter; (v) overlapping, these lncRNAs span across one or several coding genes.



Figure 19. Schematic representation of the different lncRNA classes (red) respect their coding neighbours (black and grey).

The second classification highlighted in this study is the one referring to function. There are several functions that can be addressed to lncRNAs, but the main ones would be:

(i) Precursor lncRNAs; this kind of lncRNAs produce shorter regulatory RNAs like miRNAs or siRNAs. One example is the previously mentioned *H19* which also has a function in imprinting (Cai and Cullen 2007).

(ii) lncRNAs sponges; in this case these molecules will be the target of miRNAs acting as a decoy, sequestering them and modulating the silencing response of said miRNAs. They have a huge importance in plant regulation (Zhou *et al.* 2014).

(iii) Guide lncRNAs; as the name says, they can recruit protein complexes and guide them to specific loci. The mechanism of targeting involves in some cases a triple helix structure between the DNA and the lncRNA, but no RNA/DNA sequence complementarity is usually found.

(iv) lncRNA scaffolds; they have a role in the assembly of ribonucleoprotein complexes. These complexes can behave differently when their assembly is mediated by a lncRNA, thus providing another level of regulation for an already existing complex. *Hotair* and its assembly of the PRC2 is a good example of these lncRNAs. In addition, a subtype of this kind of lncRNAs, referred as architectural lncRNAs, is essential in the assembly of some complexes.

Challenges

So far, we have slightly mentioned some of the problems when working with lncRNAs, like their lower expression levels compared with mRNAs or they short expression window, but the challenges do not end there. When trying to characterize the function of a lncRNA by loss-of-function analysis we cannot just disrupt the open reading frame like in coding genes. Morpholino oligonucleotides, for example, can be used effectively, as long as the lncRNA is spliced (Ulitsky *et al.* 2011), by targeting one of the splice sites to impede the splicing and therefore the proper folding of the lncRNA, but this limits the possible MOs to the number of splice sites in the gene. Interference RNAs can be used to degrade the lncRNA (Leone and Santoro 2016), but they are limited to cytoplasmatic lncRNAs, and some of them act in the nucleus. Deletions and inversions of the whole locus can effectively halt the lncRNA transcription, but this could also affect to the regulatory sequences present in that locus for neighbouring genes.

Their proper classification is also a challenge by itself. There are several ways to classify them as we have seen, and some more that can be considered. Although lncRNAs are characterized by not having a normal open reading frame, some of them can have small ORFs up to 100 codons. These small ORFs (smORFs) can be transcribed and some of them even translated into small functional peptides (Couso and Patraquim 2017). This

Introduction

could mean a lncRNA could be true non-coding, smORF producing lncRNA or dual, having functions associated as a lncRNA and also as a smORF producer.

Finally, evolutionary speaking, they have a massive genomic generation and decay as exposed by Neme and Tautz (2016). They found that when analysing the genus *Mus*, if they collapse all the polyadenylated transcripts found in the nine species analysed in an ancestral-like genome, the complete genome would be undergoing transcription. This means that although there are some lncRNAs that are conserved through evolution, as we have seen and will expand in this study, in most of the cases they will be almost species specific.

2.2. *Hotairm1 (HOX antisense intergenic RNA myeloid 1)*

Overview

In this study, we will focus in one of the well-known conserved lncRNAs, *Hotairm1*. This gene was first identified being expressed in myeloid lineage cells by Zhang *et al.* (2009). When analysing the intergenic regions of the Hox cluster, they found that this lncRNA had transcription levels associated with retinoic acid-mediated myeloid differentiation, hence its name. They also stated that the knockdown with shRNA attenuated the transcription of the posterior HoxA cluster genes.

Its role during the myeloid differentiation was clear, but it was also found playing a role in the differentiation of neurons. Although the basal expression levels are low, its expression is dynamically regulated during neuronal differentiation, showing a sharp increase in early differentiating neurons (Lin *et al.* 2011). That means that its function as a regulator of the Hox cluster seen in myeloid differentiation is probably not restricted to that process, and that it may be regulating the Hox cluster in other processes like neurogenesis (Gavalas *et al.* 2003).

In the next years, thanks to its relatively high (for a lncRNA) sequence conservation, *Hotairm1* was traced back to the origin of therian mammals (Yu *et al.* 2012) after being discovered in marsupials. Later, it was found in several avian and reptile transcriptomes, making its origin now being at the origin of the amniotes (Gardner *et al.* 2015). But although it was known to be conserved, and to be misregulated in several types of cancer (Esfandi *et al.* 2019; Q. Li *et al.* 2018; Ren *et al.* 2019; Song *et al.* 2019) its

mechanism of function remained a mystery until the findings of Wang and Dostle (2017).

Mechanism of *Hotairm1*

Using NT2-D1 cell lines, a kind of neuroectodermal lineage precursor that differentiates under the exposure to RA with an induction of Hox cluster genes, Wang & Dostle were able to elucidate the mechanism of action of *Hotairm1*. First, they found that there were actually two isoforms at play, one spliced and one unspliced. After knocking down *Hotairm1* with a shRNA and two siRNA and analysing the expression levels of the two isoforms and the HoxA cluster, they deduced that the spliced isoform must be repressing the middle HoxA genes, while the unspliced isoform must be promoting the expression of the anterior HoxA genes. They also found that *Hotairm1* was acting through remodelling of the chromatin, coupling with the PRC2 in its spliced form and with UTX/MLL complex in its unspliced form (Figure I10). Interestingly however, the function and expression of *Hotairm1* was, to our knowledge, not studied in any whole animal system, neither in any embryo, up to this thesis.

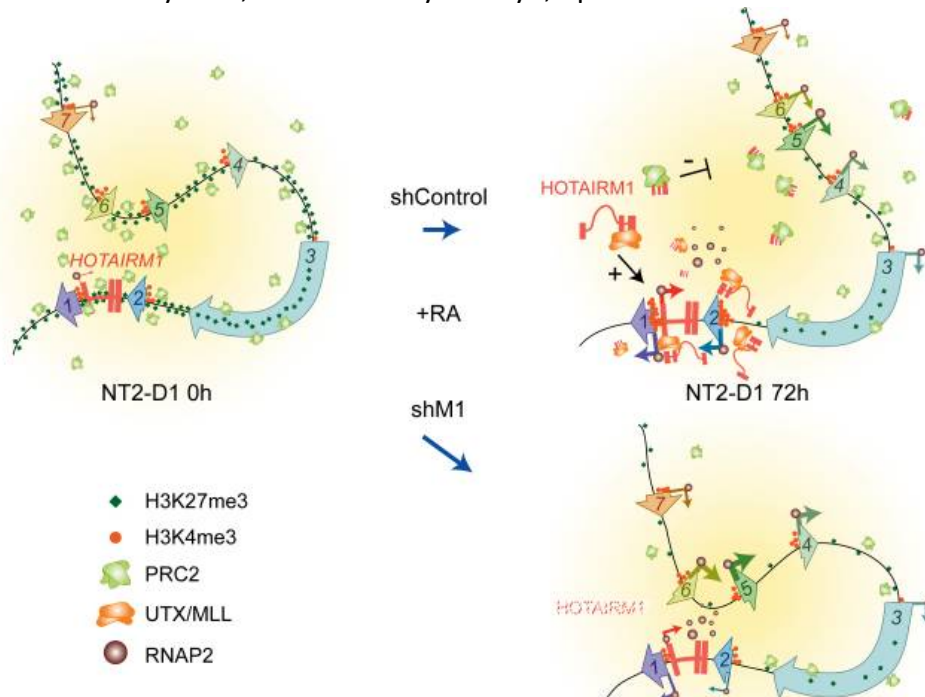


Figure I10. Diagram illustrating how HOTAIRM1 regulates the expression of HOXA genes in NT2-D1 cells. The model suggests how HOTAIRM1 contributes to the collinear activation of proximal HOXA genes through modulation of both spatial chromatin organization and the distribution of histone-modifying complexes, from Wang and Dostle 2017.

Introduction

Bibliography of Introduction

- Balmer, J. E., & Blomhoff, R. (2002). Gene expression regulation by retinoic acid. *Journal of Lipid Research*, 43(11), 1773–1808. doi:10.1194/jlr.R100015-JLR200
- Bertrand, S., Le Petillon, Y., Somorjai, I. M. L., & Escriva, H. (2017). Developmental cell-cell communication pathways in the cephalochordate amphioxus: actors and functions. *The International Journal of Developmental Biology*, 61(10-11-12), 697–722. doi:10.1387/ijdb.170202sb
- Brooke, N. M., Garcia-Fernández, J., & Holland, P. W. H. (1998). The ParaHox gene cluster is an evolutionary sister of the Hox gene cluster. *Nature*, 392(6679), 920–922. doi:10.1038/31933
- Bürglin, T. (2001). Homeobox. In *Encyclopedia of Genetics* (pp. 958–962). Academic Press. doi:10.1006/RWGN.2001.0625
- Cai, X., & Cullen, B. R. (2007). The imprinted H19 noncoding RNA is a primary microRNA precursor. *RNA*, 13(3), 313–316. doi:10.1261/rna.351707
- Cañestro, C., Albalat, R., Hjelmqvist, L., Godoy, L., Jörnvall, H., & González-Duarte, R. (2002). Ascidian and Amphioxus Adh Genes Correlate Functional and Molecular Features of the ADH Family Expansion During Vertebrate Evolution. *Journal of Molecular Evolution*, 54(1), 81–89. doi:10.1007/s00239-001-0020-2
- Cheng, F., Chen, W., Richards, E., Deng, L., & Zeng, C. (2009). SNP@Evolution: a hierarchical database of positive selection on the human genome. *BMC Evolutionary Biology*, 9(1), 221. doi:10.1186/1471-2148-9-221
- Cloutier, S. C., Wang, S., Ma, W. K., Al Husini, N., Dhoondia, Z., Ansari, A., et al. (2016). Regulated Formation of lncRNA-DNA Hybrids Enables Faster Transcriptional Induction and Environmental Adaptation. *Molecular Cell*, 61(3), 393–404. doi:10.1016/j.molcel.2015.12.024
- Conklin, E. G. (1932). The embryology of amphioxus. *Journal of Morphology*, 54(1), 69–151. doi:10.1002/jmor.1050540103
- Couso, J.-P., & Patraquim, P. (2017). Classification and function of small open reading frames. *Nature Reviews Molecular Cell Biology*, 18(9), 575–589. doi:10.1038/nrm.2017.58
- Crick, F. (1970). Central Dogma of Molecular Biology. *Nature*, 227(5258), 561–563. doi:10.1038/227561a0
- Dehal, P., & Boore, J. L. (2005). Two Rounds of Whole Genome Duplication in the Ancestral Vertebrate. *PLoS Biology*, 3(10), e314. doi:10.1371/journal.pbio.0030314
- Diederichs, S. (2014). The four dimensions of noncoding RNA conservation. *Trends in Genetics*, 30(4), 121–123. doi:10.1016/j.tig.2014.01.004
- Duret, L., Chureau, C., Samain, S., Weissenbach, J., & Avner, P. (2006). The Xist RNA gene evolved in eutherians by pseudogenization of a protein-coding gene. *Science (New York, N.Y.)*, 312(5780), 1653–5. doi:10.1126/science.1126316
- Eisen, J. S., & Smith, J. C. (2008). Controlling morpholino experiments: don't stop making antisense. *Development (Cambridge, England)*, 135(10), 1735–43. doi:10.1242/dev.001115
- Escriva, H. (2018). My Favorite Animal, Amphioxus: Unparalleled for Studying Early Vertebrate Evolution. *BioEssays*, 40(12), 1800130. doi:10.1002/bies.201800130
- Escriva, H., Holland, N. D., Gronemeyer, H., Laudet, V., & Holland, L. Z. (2002). The retinoic acid signaling pathway regulates anterior/posterior patterning in the nerve cord and pharynx of amphioxus, a chordate lacking neural crest. *Development (Cambridge, England)*, 129(12), 2905–16. <http://www.ncbi.nlm.nih.gov/pubmed/12050138>. Accessed 15 September 2019
- Esfandi, F., Taheri, M., Omrani, M. D., Shadmehr, M. B., Arsang-Jang, S., Shams, R., & Ghafouri-Fard, S. (2019). Expression of long non-coding RNAs (lncRNAs) has been dysregulated in non-small cell lung cancer tissues.

- BMC Cancer*, 19(1), 222. doi:10.1186/s12885-019-5435-5
- Fang, Y., & Fullwood, M. J. (2016). Roles, Functions, and Mechanisms of Long Non-coding RNAs in Cancer. *Genomics, Proteomics & Bioinformatics*, 14(1), 42–54. doi:10.1016/J.GPB.2015.09.006
- Force, A., Lynch, M., Pickett, F. B., Amores, A., Yan, Y. L., & Postlethwait, J. (1999). Preservation of duplicate genes by complementary, degenerative mutations. *Genetics*, 151(4), 1531.
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1460548/. Accessed 15 September 2019
- Garcia-Fernández, J. (2005). The genesis and evolution of homeobox gene clusters. *Nature Reviews Genetics*, 6(12), 881–892. doi:10.1038/nrg1723
- Garcia-Fernández, J., & Benito-Gutiérrez, È. (2009). It's a long way from amphioxus: descendants of the earliest chordate. *BioEssays*, 31(6), 665–675. doi:10.1002/bies.200800110
- Garcia-Fernandez, J., Jimenez-Delgado, S., Pascual-Anaya, J., Maeso, I., Irimia, M., Minguillon, C., et al. (2009). From the American to the European amphioxus: towards experimental Evo-Devo at the origin of chordates. *The International Journal of Developmental Biology*, 53(8–9–10), 1359–1366. doi:10.1387/ijdb.072436jg
- Gardner, P. P., Fasold, M., Burge, S. W., Ninova, M., Hertel, J., Kehr, S., et al. (2015). Conservation and Losses of Non-Coding RNAs in Avian Genomes. *PLOS ONE*, 10(3), e0121797. doi:10.1371/journal.pone.0121797
- Gavalas, A., Ruhrberg, C., Livet, J., Henderson, C. E., & Krumlauf, R. (2003). Neuronal defects in the hindbrain of Hoxa1, Hoxb1 and Hoxb2 mutants reflect regulatory interactions among these Hox genes. *Development*, 130(23), 5663–5679. doi:10.1242/dev.00802
- Gerhart, J., & Kirschner, M. (2007). The theory of facilitated variation. *Proceedings of the National Academy of Sciences of the United States of America*, 104 Suppl 1(Suppl 1), 8582–9. doi:10.1073/pnas.0701035104
- Gilbert, S. F., & Barresi, M. J. F. (2017). *Developmental Biology*. Oxford University Press.
https://books.google.es/books?id=Iq3dtAEACAAJ
- Goodsir, J. (1844). XV. *On the Anatomy of Amphioxus lanceolatus*; Lancelet, Yarrell. *Transactions of the Royal Society of Edinburgh*, 15(1), 247–263. doi:10.1017/S0080456800029938
- Hall, M. N., Gabay, J., Débarbouillé, M., & Schwartz, M. (1982). A role for mRNA secondary structure in the control of translation initiation. *Nature*, 295(5850), 616–618. doi:10.1038/295616a0
- Han, P., Li, W., Lin, C.-H., Yang, J., Shang, C., Nurnberg, S. T., et al. (2014). A long noncoding RNA protects the heart from pathological hypertrophy. *Nature*, 514(7520), 102–106. doi:10.1038/nature13596
- Harland, R. M., & Grainger, R. M. (2011). Xenopus research: metamorphosed by genetics and genomics. *Trends in Genetics*, 27(12), 507–515. doi:10.1016/j.tig.2011.08.003
- Heinen, T. J. A. J., Staubach, F., Häming, D., & Tautz, D. (2009). Emergence of a new gene from an intergenic region. *Current biology : CB*, 19(18), 1527–31. doi:10.1016/j.cub.2009.07.049
- Hideaki Kaneto, H., Takeshi Miyatsuka, T., Toshihiko Shiraiwa, T., Kaoru Yamamoto, K., Ken Kato, K., Yoshio Fujitani, Y., & Taka-aki Matsuoka, T. (2007). Crucial Role of PDX-1 in Pancreas Development, β -Cell Differentiation, and Induction of Surrogate β -Cells. *Current Medicinal Chemistry*, 14(16), 1745–1752.
doi:10.2174/092986707781058887
- Hirsinger, E., Carvalho, J. E., Chevalier, C., Lutfalla, G., Nicolas, J.-F., Peyri ras, N., & Schubert, M. (2015). Expression of Fluorescent Proteins in Branchiostoma lanceolatum by mRNA Injection into Unfertilized Oocytes. *Journal of Visualized Experiments : JoVE*, (95). doi:10.3791/52042
- Holland, A. M., Garcia, S., Naselli, G., MacDonald, R. J., & Harrison, L. C. (2013). The Parahox gene Pdx1 is required to maintain positional identity in the adult foregut. *The International Journal of Developmental Biology*, 57(5), 391–398. doi:10.1387/ijdb.120048ah

Introduction

- Holland, N. D., & Holland, L. Z. (2017). The ups and downs of amphioxus biology: a history. *The International Journal of Developmental Biology*, 61(10-11-12), 575–583. doi:10.1387/ijdb.160395LH
- Holland, P. W. H. (2013). Evolution of homeobox genes. *Wiley Interdisciplinary Reviews: Developmental Biology*, 2(1), 31–45. doi:10.1002/wdev.78
- Hueber, S. D., Weiller, G. F., Djordjevic, M. A., & Frickey, T. (2010). Improving Hox Protein Classification across the Major Model Organisms. *PLoS ONE*, 5(5), e10820. doi:10.1371/journal.pone.0010820
- Irimia, M., Tena, J. J., Alexis, M. S., Fernandez-Minan, A., Maeso, I., Bogdanovic, O., et al. (2012). Extensive conservation of ancient microsynteny across metazoans due to cis-regulatory constraints. *Genome Research*, 22(12), 2356–2367. doi:10.1101/gr.139725.112
- Jarroux, J., Morillon, A., & Pinskaya, M. (2017). Chapter 1: History, Discovery, and Classification of lncRNAs. In M.R.S. Rao (Ed.), *Long Non Coding RNA Biology* (pp. 1–46). Singapore: Springer Nature. doi:10.1007/978-981-10-5203-3_1
- Jimenez-Delgado, S., Pascual-Anaya, J., & Garcia-Fernandez, J. (2009). Implications of duplicated cis-regulatory elements in the evolution of metazoans: the DDI model or how simplicity begets novelty. *Briefings in functional genomics & proteomics*, 8(4), 266–275. doi:10.1093/bfpg/elp029
- Kallen, A. N., Zhou, X.-B., Xu, J., Qiao, C., Ma, J., Yan, L., et al. (2013). The Imprinted H19 lncRNA Antagonizes Let-7 MicroRNAs. *Molecular Cell*, 52(1), 101–112. doi:10.1016/j.molcel.2013.08.027
- Kapusta, A., Kronenberg, Z., Lynch, V. J., Zhuo, X., Ramsay, L., Bourque, G., et al. (2013). Transposable Elements Are Major Contributors to the Origin, Diversification, and Regulation of Vertebrate Long Noncoding RNAs. *PLoS Genetics*, 9(4), e1003470. doi:10.1371/journal.pgen.1003470
- Leone, S., & Santoro, R. (2016). Challenges in the analysis of long noncoding RNA functionality. *FEBS Letters*, 590(15), 2342–2353. doi:10.1002/1873-3468.12308
- Lewis, E. B. (1993). Clusters of master control genes regulate the development of higher organisms. *Current Science*, 64(9), 640–649. <http://www.jstor.org/stable/24095142>
- Li, G., Feng, J., Lei, Y., Wang, J., Wang, H., Shang, L.-K., et al. (2014). Mutagenesis at Specific Genomic Loci of Amphioxus *Branchiostoma belcheri* Using TALEN Method HHS Public Access. *J Genet Genomics*, 41(4), 215–219. doi:10.1016/j.jgg.2014.02.003
- Li, Q., Dong, C., Cui, J., Wang, Y., & Hong, X. (2018). Over-expressed lncRNA HOTAIRM1 promotes tumor growth and invasion through up-regulating HOXA1 and sequestering G9a/EZH2/Dnmts away from the HOXA1 gene in glioblastoma multiforme. *Journal of Experimental & Clinical Cancer Research*, 37(1), 265. doi:10.1186/s13046-018-0941-x
- Li, X., Wu, Z., Fu, X., & Han, W. (2014). lncRNAs: Insights into their function and mechanics in underlying disorders. *Mutation Research/Reviews in Mutation Research*, 762, 1–21. doi:10.1016/J.MRREV.2014.04.002
- Lin, M., Pedrosa, E., Shah, A., Hrabovsky, A., Maqbool, S., Zheng, D., & Lachman, H. M. (2011). RNA-Seq of Human Neurons Derived from iPS Cells Reveals Candidate Long Non-Coding RNAs Involved in Neurogenesis and Neuropsychiatric Disorders. *PLoS ONE*, 6(9), e23356. doi:10.1371/journal.pone.0023356
- Liu, X., Li, G., Feng, J., Yang, X., & Wang, Y.-Q. (2013). An efficient microinjection method for unfertilized eggs of Asian amphioxus *Branchiostoma belcheri*. *Development Genes and Evolution*, 223(4), 269–278. doi:10.1007/s00427-013-0441-0
- Mansfield, J. H., Haller, E., Holland, N. D., & Brent, A. E. (2015). Development of somites and their derivatives in amphioxus, and implications for the evolution of vertebrate somites. *EvoDevo*, 6(1), 21. doi:10.1186/s13227-015-0007-5

- Marlétaz, F., Maeso, I., Faas, L., Isaacs, H. V., & Holland, P. W. H. (2015). Cdx ParaHox genes acquired distinct developmental roles after gene duplication in vertebrate evolution. *BMC Biology*, *13*(1), 56. doi:10.1186/s12915-015-0165-x
- Mayr, E., & Provine, W. B. (1980). *The Evolutionary synthesis : perspectives on the unification of biology*. Harvard University Press.
- Müller, G. B. (2007). Evo–devo: extending the evolutionary synthesis. *Nature Reviews Genetics*, *8*(12), 943–949. doi:10.1038/nrg2219
- Neme, R., & Tautz, D. (2016). Fast turnover of genome transcription across evolutionary time exposes entire non-coding DNA to de novo gene emergence. *eLife*, *5*. doi:10.7554/eLife.09977
- Newman, S. A. (2006). The Developmental Genetic Toolkit and the Molecular Homology—Analogy Paradox. *Biological Theory*, *1*(1), 12–16. doi:10.1162/biot.2006.1.1.12
- Nieuwkoop, P., & Faber, J. (1994). *Normal table of Xenopus laevis (Daudin)*. New York: Garland Publishing. <https://books.google.es/books?id=a06nHgAACAAJ>
- Ohno, S. (1970). *Evolution by Gene Duplication*. Berlin, Heidelberg: Springer Berlin Heidelberg. doi:10.1007/978-3-642-86659-3
- Pallas, P. (1794). *Spicilegia zoologica: Tomus I. Continens quasrupedium, avium, amphobiorum ... - P.S. Pallas - Google Libros*. Berlin: Lange.
- Pei, Z., Wang, B., Chen, G., Nagao, M., Nakafuku, M., & Campbell, K. (2011). Homeobox genes Gsx1 and Gsx2 differentially regulate telencephalic progenitor maturation. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(4), 1675–80. doi:10.1073/pnas.1008824108
- Rastan, S. (1994). X chromosome inactivation and the Xist gene. *Current opinion in genetics & development*, *4*(2), 292–7. <http://www.ncbi.nlm.nih.gov/pubmed/8032207>. Accessed 15 September 2019
- Ren, T., Hou, J., Liu, C., Shan, F., Xiong, X., Qin, A., et al. (2019). The long non-coding RNA HOTAIRM1 suppresses cell progression via sponging endogenous miR-17-5p/ B-cell translocation gene 3 (BTG3) axis in 5-fluorouracil resistant colorectal cancer cells. *Biomedicine & Pharmacotherapy*, *117*, 109171. doi:10.1016/j.biopha.2019.109171
- Rinn, J. L., Kertesz, M., Wang, J. K., Squazzo, S. L., Xu, X., Bruggmann, S. A., et al. (2007). Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. *Cell*, *129*(7), 1311–23. doi:10.1016/j.cell.2007.05.022
- Rivas, E., Clements, J., & Eddy, S. R. (2017). A statistical test for conserved RNA structure shows lack of evidence for structure in lncRNAs. *Nature methods*, *14*(1), 45–48. doi:10.1038/nmeth.4066
- Schmitt, S. M., Gull, M., & Brändli, A. W. (2014). Engineering Xenopus embryos for phenotypic drug discovery screening. *Advanced Drug Delivery Reviews*, *69–70*, 225–246. doi:10.1016/j.addr.2014.02.004
- Sears, K. E., Behringer, R. R., Rasweiler, J. J., & Niswander, L. A. (2006). Development of bat flight: Morphologic and molecular evolution of bat wing digits. *Proceedings of the National Academy of Sciences*, *103*(17), 6581–6586. doi:10.1073/pnas.0509716103
- Song, L., Zhang, S., Duan, C., Ma, S., Hussain, S., Wei, L., & Chu, M. (2019). Genome-wide identification of lncRNAs as novel prognosis biomarkers of glioma. *Journal of Cellular Biochemistry*, *jcb.29259*. doi:10.1002/jcb.29259
- Tan, M. H., Au, K. F., Yablonoitch, A. L., Wills, A. E., Chuang, J., Baker, J. C., et al. (2013). RNA sequencing reveals a diverse and dynamic repertoire of the *Xenopus tropicalis* transcriptome over development. *Genome research*, *23*(1), 201–16. doi:10.1101/gr.141424.112
- Uliitsky, I., Shkumatava, A., Jan, C. H., Sive, H., & Bartel, D. P. (2011). Conserved function of lincRNAs in vertebrate

Introduction

- embryonic development despite rapid sequence evolution. *Cell*, 147(7), 1537–50.
doi:10.1016/j.cell.2011.11.055
- Wang, X. Q. D., & Dostie, J. (2017). Reciprocal regulation of chromatin state and architecture by HOTAIRM1 contributes to temporal collinear HOXA gene activation. *Nucleic acids research*, 45(3), 1091–1104.
doi:10.1093/nar/gkw966
- Woltering, J. M. (2012). From lizard to snake; behind the evolution of an extreme body plan. *Current genomics*, 13(4), 289–99. doi:10.2174/138920212800793302
- Yarrell, W. (1836). *A History of British Fishes*. J. Van Voorst. <https://books.google.es/books?id=FtsHAQAIAAJ>
- Yu, H., Lindsay, J., Feng, Z.-P., Frankenberg, S., Hu, Y., Carone, D., et al. (2012). Evolution of coding and non-coding genes in HOX clusters of a marsupial. *BMC genomics*, 13, 251. doi:10.1186/1471-2164-13-251
- Zhang, X., Lian, Z., Padden, C., Gerstein, M. B., Rozowsky, J., Snyder, M., et al. (2009). A myelopoiesis-associated regulatory intergenic noncoding RNA transcript within the human HOXA cluster. *Blood*, 113(11), 2526–34.
doi:10.1182/blood-2008-06-162164
- Zhou, X., Gao, Q., Wang, J., Zhang, X., Liu, K., & Duan, Z. (2014). Linc-RNA-RoR acts as a “sponge” against mediation of the differentiation of endometrial cancer stem cells by microRNA-145. *Gynecologic Oncology*, 133(2), 333–339. doi:10.1016/j.ygyno.2014.02.033

Objectives

The finding of conserved gene regulatory elements is at the very core of the Evo-devo discipline. Besides, the regulation of the Hox cluster is key to understand how the different organisms evolved, hence finding underlying conservations can help to unravel the detailed mechanisms of its regulation. As lncRNAs like *Hotairm1* and ParaHox cluster genes are known regulators of Hox genes, the objectives of this thesis are:

- The proper identification of the lncRNA fraction within the *Branchiostoma lanceolatum* transcriptome.
- The establishment of a new method to find underlying lincRNA orthologies using microsynteny.
- The characterisation of *Hotairm1* expression and function in chordates.
- The analysis of the ParaHox cluster genes *Cdx* and *Pdx* knockouts in *Branchiostoma floridae*.
- The analyses of the ParaHox gene *Cdx* in the regulation of the Hox cluster during development.

Results



Results

Indexes of quality of the articles included in the Thesis

Impact factor and ranking position in the research area

Article RI

Amphioxus functional genomics and the origins of vertebrate gene regulation

Ferdinand Marlétaz, Panos N. Firbas, ignacio Maeso*, Juan J. tena, Ozren Bogdanovic, Malcolm Perry, christopher D. r. Wyatt, elisa de la calle-Mustienes, Stephanie Bertrand, Demian Burguera, rafael D. Acemel, Simon J. van Heeringen, Silvia Naranjo, carlos Herrera-Ubeda, Ksenia Skvortsova, Sandra Jimenez-Gancedo, Daniel Aldea, Yamile Marquez, lorena Buono, iryna Kozmikova, Jon Permanyer, Alexandra louis, Beatriz Albuixech-crespo, Yann le Petillon, Anthony leon, Lucie Subirana, Piotr J. Balwierz, Paul edward Duckett, ensieh Farahani, Jean-Marc Aury, Sophie Mangenot, Patrick Wincker, ricard Albalat, Èlia Benito-Gutiérrez, cristian cañestro, Filipe castro, Salvatore D'Aniello, David e. K. Ferrier, Shengfeng Huang, Vincent laudet, Gabriel A. B. Marais, Pierre Pontarotti, Michael Schubert, Hervé Seitz, ildiko Somorjai, tokiharu takahashi, Olivier Mirabeau, Anlong Xu, Jr-Kai Yu, Piero carninci, Juan ramon Martinez-Morales, Hugues roest crollius, Zbynek Kozmik, Matthew t. Weirauch, Jordi Garcia-Fernàndez, ryan lister, Boris lenhard, Peter W. H. Holland, Hector escriva*, Jose luis Gómez-Skarmeta* & Manuel irimia,*

Journal: *Nature*

Impact factor= 41.57

Area/Ranking: Multidisciplinary 1/120

Article RII

Microsyntenic Clusters Reveal Conservation of lncRNAs in Chordates Despite Absence of Sequence Conservation

Carlos Herrera-Úbeda, Marta Marín-Barba, Enrique Navas-Pérez, Jan Gravemeyer, Beatriz Albuixech-Crespo, Grant N. Wheeler and Jordi Garcia-Fernàndez*

Journal: *Biology (MPDI)*

Impact factor= 4.42

Area/Ranking: Agricultural and Biological Sciences 22/272
Genetics and Molecular Biology 38/242

Article RIII

Mutation of amphioxus Pdx and Cdx demonstrates conserved roles for ParaHox genes in gut, anus and tail patterning

Yanhong Zhong*, Carlos Herrera-Úbeda*, Jordi Garcia-Fernàndez, Guang Li* and Peter WH Holland*

To be submitted to first decile or 1st quartile journals

Articles in the Annex section

Article AI

Characterization of the TLR Family in Branchiostoma lanceolatum and Discovery of a Novel TLR22-Like Involved in dsRNA Recognition in Amphioxus

Jie Ji, David Ramos-Vicente, Enrique Navas-Pérez, Carlos Herrera-Úbeda, José Miguel Lizcano, Jordi Garcia-Fernàndez, Hector Escrivà, Alex Bayés and Nerea Roher

Journal: *Frontiers in immunology*

Impact factor= 4,72

Area/Ranking: Immunology 45/216
Immunology and allergy 32/203

Article AII

Origin and evolution of the chordate central nervous system: insights from amphioxus genoarchitecture.

Beatriz Albuixech-Crespo, Carlos Herrera-Úbeda, Gemma Marfany, Manuel Irimia and Jordi Garcia-Fernàndez

Journal: *Int J Dev Biol*

Impact factor= 2,11

Area/Ranking: Embriology 10/19
Developmental biology 57/86

Signed: Jordi Garcia Fernàndez, PhD supervisor

Results

Article RI:

Abstract

Vertebrates have greatly elaborated the basic chordate body plan and evolved highly distinctive genomes that have been sculpted by two whole-genome duplications. Here we sequence the genome of the Mediterranean amphioxus (*Branchiostoma lanceolatum*) and characterize DNA methylation, chromatin accessibility, histone modifications and transcriptomes across multiple developmental stages and adult tissues to investigate the evolution of the regulation of the chordate genome. Comparisons with vertebrates identify an intermediate stage in the evolution of differentially methylated enhancers, and a high conservation of gene expression and its *cis*-regulatory logic between amphioxus and vertebrates that occurs maximally at an earlier mid-embryonic phylotypic period. We analyse regulatory evolution after whole-genome duplications, and find that—in vertebrates—over 80% of broadly expressed gene families with multiple paralogues derived from whole-genome duplications have members that restricted their ancestral expression, and underwent specialization rather than subfunctionalization. Counter-intuitively, paralogues that restricted their expression increased the complexity of their regulatory landscapes. These data pave the way for a better understanding of the regulatory principles that underlie key vertebrate innovations.

Article RI:

Amphioxus functional genomics and the origins of vertebrate gene regulation

Ferdinand Marlétaz, Panos N. Firbas, ignacio Maeso,* , Juan J. Tena, Ozren Bogdanovic, Malcolm Perry, christopher D. r. Wyatt, elisa de la calle-Mustienes, Stephanie Bertrand, Demian Burguera, rafael D. Acemel, Simon J. van Heeringen, Silvia Naranjo, carlos Herrera-Ubeda, Ksenia Skvortsova, Sandra Jimenez-Gancedo, Daniel Aldea, Yamile Marquez, lorena Buono, iryna Kozmikova, Jon Permanyer, Alexandra louis, Beatriz Albuixech-crespo, Yann le Petillon, Anthony leon, lucie Subirana, Piotr J. Balwierz, Paul edward Duckett, ensieh Farahani, Jean-Marc Aury, Sophie Mangenot, Patrick Wincker, ricard Albalat, Èlia Benito-Gutiérrez, cristian cañestro, Filipe castro, Salvatore D’Aniello, David e. K. Ferrier, Shengfeng Huang, Vincent laudet, Gabriel A. B. Marais, Pierre Pontarotti, Michael Schubert, Hervé Seitz, ildiko Somorjai, tokiharu takahashi, Olivier Mirabeau, Anlong Xu, Jr-Kai Yu, Piero carninci, Juan ramon Martinez-Morales, Hugues roest crolius, Zbynek Kozmik, Matthew t. Weirauch, Jordi Garcia-Fernàndez, ryan lister, Boris lenhard, Peter W. H. Holland, Hector escriba*, Jose luis Gómez-Skarmeta* & Manuel irimia,*

Marlétaz, F.; Firbas, P.N.; Maeso, I.; Tena, J.J.; Bogdanovic, O.; Perry, M.; Wyatt, C.D.R.; de la Calle-Mustienes, E.; Bertrand, S.; Burguera, D.; et al. Amphioxus functional genomics and the origins of vertebrate gene regulation. *Nature* **2018**, *564*, 64–70.

The PhD student is an author of this major multicentre article, on the functional analyses of the amphioxus genome. In particular, Carlos Herrera was the solely responsible of characterizing the long noncoding RNA complement on the crude genome data, for which he had to develop a complex pipeline, and had to classify and identify all the thousands of lncRNAs found. He also performed blast and comparison analyses of the lncRNAs found against current databases available. To my knowledge, the article has not been included in any other PhD thesis.

Signed: Jordi Garcia Fernàndez, PhD supervisor

Amphioxus functional genomics and the origins of vertebrate gene regulation

Ferdinand Marlétaz^{1,2,41}, Panos N. Firas^{3,41}, Ignacio Maeso^{3,41*}, Juan J. Tena^{3,41}, Ozren Bogdanovic^{4,5,6,41}, Malcolm Perry^{7,8,41}, Christopher D. R. Wyatt^{9,10}, Elisa de la Calle-Mustienes³, Stephanie Bertrand¹¹, Demian Burguera^{9,12}, Rafael D. Acemel³, Simon J. van Heeringen¹³, Silvia Naranjo³, Carlos Herrera-Ubeda¹², Ksenia Skvortsova⁴, Sandra Jimenez-Gancedo³, Daniel Aldea¹¹, Yamile Marquez⁹, Lorena Buono³, Iryna Kozmikova¹⁴, Jon Permanyer⁹, Alexandra Louis^{15,16,17}, Beatriz Albuixech-Crespo¹², Yann Le Petillon¹¹, Anthony Leon¹¹, Lucie Subirana¹¹, Piotr J. Balwierz^{7,8}, Paul Edward Duckett⁴, Ensieh Farahani³, Jean-Marc Aury¹⁸, Sophie Mangenot¹⁸, Patrick Wincker¹⁹, Ricard Albalat²⁰, Èlia Benito-Gutiérrez²¹, Cristian Cañestro²⁰, Filipe Castro²², Salvatore D'Aniello²³, David E. K. Ferrier²⁴, Shengfeng Huang²⁵, Vincent Laudet¹¹, Gabriel A. B. Marais²⁶, Pierre Pontarotti²⁷, Michael Schubert²⁸, Hervé Seitz²⁹, Ildiko Somorjai³⁰, Tokiharu Takahashi³¹, Olivier Mirabeau³², Anlong Xu^{25,33}, Jr-Kai Yu³⁴, Piero Carninci^{35,36}, Juan Ramon Martinez-Morales³, Hugues Roest Crollius^{15,16,17}, Zbynek Kozmik¹⁴, Matthew T. Weirauch^{37,38}, Jordi Garcia-Fernández¹², Ryan Lister^{6,39}, Boris Lenhard^{7,8,40}, Peter W. H. Holland¹, Hector Escriva^{11*}, Jose Luis Gómez-Skarmeta^{3*} & Manuel Irimia^{9,10*}

Vertebrates have greatly elaborated the basic chordate body plan and evolved highly distinctive genomes that have been sculpted by two whole-genome duplications. Here we sequence the genome of the Mediterranean amphioxus (*Branchiostoma lanceolatum*) and characterize DNA methylation, chromatin accessibility, histone modifications and transcriptomes across multiple developmental stages and adult tissues to investigate the evolution of the regulation of the chordate genome. Comparisons with vertebrates identify an intermediate stage in the evolution of differentially methylated enhancers, and a high conservation of gene expression and its cis-regulatory logic between amphioxus and vertebrates that occurs maximally at an earlier mid-embryonic phylotypic period. We analyse regulatory evolution after whole-genome duplications, and find that—in vertebrates—over 80% of broadly expressed gene families with multiple paralogues derived from whole-genome duplications have members that restricted their ancestral expression, and underwent specialization rather than subfunctionalization. Counter-intuitively, paralogues that restricted their expression increased the complexity of their regulatory landscapes. These data pave the way for a better understanding of the regulatory principles that underlie key vertebrate innovations.

All vertebrates share multiple morphological and genomic novelties¹. The most prominent genomic difference between vertebrates and non-vertebrate chordates is the reshaping of the gene complement that followed the two rounds of whole genome duplication (WGD)—the 2R hypothesis—that occurred at the base of the vertebrate lineage^{2,3}. These large-scale mutational events are hypothesized to have

facilitated the evolution of vertebrate morphological innovations, at least in part through the preferential retention of 'developmental' gene families and transcription factors after duplication^{3,4}. However, duplicate genes and their associated regulatory elements were initially identical and could not drive innovation without regulatory and/or protein-coding changes.

¹Department of Zoology, University of Oxford, Oxford, UK. ²Molecular Genetics Unit, Okinawa Institute of Science and Technology Graduate University, Onna-son, Japan. ³Centro Andaluz de Biología del Desarrollo (CABD), CSIC-Universidad Pablo de Olavide-Junta de Andalucía, Seville, Spain. ⁴Genomics and Epigenetics Division, Garvan Institute of Medical Research, Sydney, New South Wales, Australia. ⁵St Vincent's Clinical School, Faculty of Medicine, University of New South Wales, Sydney, New South Wales, Australia. ⁶Australian Research Council Centre of Excellence in Plant Energy Biology, School of Molecular Sciences, The University of Western Australia, Crawley, Western Australia, Australia. ⁷Institute of Clinical Sciences, Faculty of Medicine, Imperial College London, London, UK. ⁸Computational Regulatory Genomics, MRC London Institute of Medical Sciences, London, UK. ⁹Centre for Genomic Regulation (CRG), Barcelona Institute of Science and Technology (BIST), Barcelona, Spain. ¹⁰Universitat Pompeu Fabra (UPF), Barcelona, Spain. ¹¹Biologie Intégrative des Organismes Marins, BIOM, Observatoire Océanologique, CNRS and Sorbonne Université, Banyuls sur Mer, France. ¹²Department of Genetics, Microbiology and Statistics, Faculty of Biology, and Institut de Biomedicina (IBUB), University of Barcelona, Barcelona, Spain. ¹³Department of Molecular Developmental Biology, Faculty of Science, Radboud Institute for Molecular Life Sciences, Radboud University, Nijmegen, The Netherlands. ¹⁴Institute of Molecular Genetics of the Czech Academy of Sciences, Prague, Czech Republic. ¹⁵Institut de Biologie de l'ENS, IBENS, Ecole Normale Supérieure, Paris, France. ¹⁶Inserm, U1024, Paris, France. ¹⁷CNRS, UMR 8197, Paris, France. ¹⁸Genoscope, Institut de biologie François-Jacob, Commissariat à l'Energie Atomique (CEA), Université Paris-Saclay, Evry, France. ¹⁹Génomique Métabolique, Genoscope, Institut de biologie François Jacob, Commissariat à l'Energie Atomique (CEA), CNRS, Université Evry, Université Paris-Saclay, Evry, France. ²⁰Department of Genetics, Microbiology and Statistics, Faculty of Biology and Institut de Recerca de la Biodiversitat (IRBio), University of Barcelona, Barcelona, Spain. ²¹Department of Zoology, University of Cambridge, Cambridge, UK. ²²Interdisciplinary Centre of Marine and Environmental Research (CIIMAR/CIMAR) and Faculty of Sciences (FCUP), Department of Biology, University of Porto, Porto, Portugal. ²³Biology and Evolution of Marine Organisms, Stazione Zoologica Anton Dohrn Napoli, Naples, Italy. ²⁴The Scottish Oceans Institute, Gatty Marine Laboratory, University of St Andrews, St Andrews, UK. ²⁵State Key Laboratory of Biocontrol, School of Life Sciences, Sun Yat-sen University, Guangzhou, China. ²⁶Laboratoire de Biométrie et Biologie Evolutive (UMR 5558), CNRS and Université Lyon 1, Villeurbanne, France. ²⁷IRD, APHM, Microbe, Evolution, PHYLOGÉNIE, Infection, IHU Méditerranée Infection and CNRS, Aix Marseille University, Marseille, France. ²⁸Sorbonne Université, CNRS, Laboratoire de Biologie du Développement de Villefranche-sur-Mer, Institut de la Mer de Villefranche-sur-Mer, Villefranche-sur-Mer, France. ²⁹UMR 9002 CNRS, Institut de Génétique Humaine, Université de Montpellier, Montpellier, France. ³⁰Biomedical Sciences Research Complex, School of Biology, University of St Andrews, St Andrews, UK. ³¹School of Medical Sciences, Faculty of Biology, Medicine and Health, University of Manchester, Manchester, UK. ³²INSERM U830, Équipe Labellisée LNCC, SIREDO Oncology Centre, Institut Curie, PSL Research University, Paris, France. ³³School of Life Sciences, Beijing University of Chinese Medicine, Beijing, China. ³⁴Institute of Cellular and Organismic Biology, Academia Sinica, Taipei, Taiwan. ³⁵RIKEN Center for Life Science Technologies (Division of Genomic Technologies) (CLST DGT), Yokohama, Japan. ³⁶Laboratory for Transcriptome Technology, RIKEN Center for Integrative Medical Sciences, Yokohama, Japan. ³⁷Center for Autoimmune Genomics and Etiology, Divisions of Biomedical Informatics and Developmental Biology, Cincinnati Children's Hospital Medical Center, Cincinnati, OH, USA. ³⁸Department of Pediatrics, University of Cincinnati College of Medicine, Cincinnati, OH, USA. ³⁹Harry Perkins Institute of Medical Research, Nedlands, Western Australia, Australia. ⁴⁰Sars International Centre for Marine Molecular Biology, University of Bergen, Bergen, Norway. ⁴¹These authors contributed equally: Ferdinand Marlétaz, Panos N. Firas, Ignacio Maeso, Juan J. Tena, Ozren Bogdanovic, Malcolm Perry. *e-mail: nacho.maeso@gmail.com; hescriva@obs-banyuls.fr; jlgomska@upo.es; mirimia@gmail.com

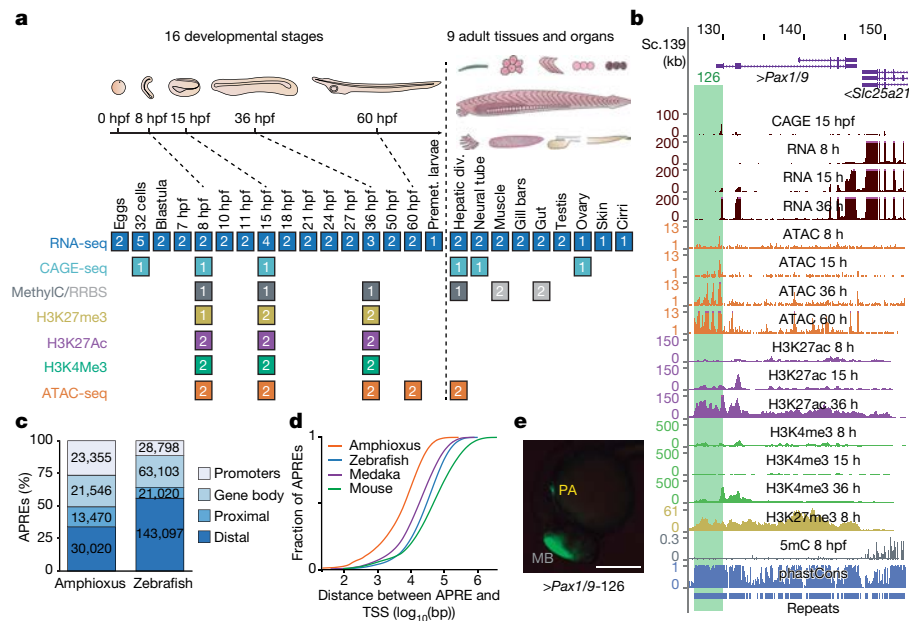


Fig. 1 | Functional genome annotation of amphioxus. **a**, Summary of the 94 amphioxus samples generated in this study, comprising eight functional-genomic datasets. The number of biological replicates is indicated for each sample type. div., diverticulum; MethyC/RRBS, methylC sequencing and reduced representation bisulfite sequencing; Premet., premetamorphic. **b**, Genome browser excerpt showing a selection of available tracks, including gene annotation, sequence conservation (using phastCons), repeats and several epigenomic and transcriptomic datasets. Green rectangle highlights the APRE tested in **e**. **c**, Numbers and proportions of amphioxus and zebrafish APREs according to their

genomic location. Promoters, within 1-kbp upstream and 0.5-kbp downstream of an annotated TSS; gene body, within an orthology-supported gene; proximal, within 5-kbp upstream of (but not overlapping with) a TSS; distal, not in the aforementioned categories. **d**, Cumulative distributions of the distance between each APRE and the closest annotated TSS in each species. **e**, Lateral view of a representative transgenic zebrafish 26-hpf embryo showing GFP expression driven by an amphioxus APRE associated with *Pax1/9* (*Pax1/9-126*, highlighted in **b**) in pharyngeal arches (PA; $n = 4/4$). Positive-control enhancer was expressed in the midbrain (MB). Scale bar, 250 μ m.

To date, the effect of vertebrate WGDs on gene regulation have remained poorly understood—both in terms of the fates of duplicate genes and the acquisition of the unique genomic traits that are characteristic of vertebrates. These traits include numerous features that are often associated with gene regulation, such as unusually large intergenic and intronic regions^{5,6}, high global 5-methylcytosine (5mC) content and 5mC-dependent regulation of embryonic transcriptional enhancers⁷. To investigate these traits, appropriate species must be used for comparisons. Previous studies have largely focused on phylogenetic distances that are either too short (such as human versus mouse) or too long (such as human versus fly or nematode), resulting in limited insights. In the first case, comparisons among closely related species (for example, between mammals^{8–11})—for which the orthology of non-coding regions can be readily determined from genomic alignments—have allowed fine-grained analyses of the evolution of transcription-factor binding. In the second case, three-way comparisons of human, fly and nematode by the modENCODE consortium revealed no detectable conservation at the *cis*-regulatory level¹² and very little conservation of gene expression¹³. Moreover, the genomes of flies and nematodes are highly derived^{14–16}. Thus, we lack comprehensive functional genomic data from a slow-evolving, closely related outgroup that would enable an in-depth investigation of the origins of the vertebrate regulatory genome and of the effect of WGDs on gene regulation.

Unlike flies, nematodes and most non-vertebrates, amphioxus belongs to the chordate phylum. Therefore, although amphioxus lacks the specializations and innovations of vertebrates, it shares with them a basic body plan and has multiple organs and structures homologous to those of vertebrates¹. For these reasons, amphioxus has widely been used as a reference outgroup to infer ancestral versus novel features during vertebrate evolution. Here, we undertook a comprehensive study of the transcriptome and regulatory genome of amphioxus to investigate how the unique functional genome architecture of vertebrates evolved.

Functional genome annotation of amphioxus

We generated an exhaustive resource of genomic, epigenomic and transcriptomic data for the Mediterranean amphioxus (*B. lanceolatum*), comprising a total of 52 sample types (Fig. 1a and Supplementary Data 2, datasets 1–5). These datasets were mapped to a *B. lanceolatum* genome that was sequenced and assembled de novo, with 150 \times coverage, a total size of 495.4 Mbp, a scaffold N50 of 1.29 Mbp and 4% gaps (Extended Data Fig. 1a–c). To facilitate access by the research community, we integrated these resources into a UCSC Genome Browser track hub (Fig. 1b; available at <http://amphiencode.github.io/Data/>), together with an intra-cephalochordate sequence conservation track and a comprehensive annotation of repetitive elements (Extended Data Fig. 1d–f) and long non-coding RNAs (Extended Data Fig. 1g and Supplementary Data 2, dataset 6). To enable broader evolutionary comparisons, we reconstructed orthologous gene families for multiple vertebrate and non-vertebrate species (Supplementary Data 2, dataset 7), generated several equivalent datasets for zebrafish and medaka (Extended Data Fig. 2a), and built a dedicated server for synteny comparisons (Extended Data Fig. 1h).

A comprehensive functional annotation of the *B. lanceolatum* genome identified 88,391 putative *cis*-regulatory elements of DNA as defined by assay for transposase-accessible chromatin using sequencing (ATAC-seq) (these elements are hereafter referred to as APREs), as well as 20,569 protein-coding genes supported by orthology. We divided the APREs into promoters—around transcription start sites (TSSs), which were highly supported by cap analysis gene-expression sequencing (CAGE-seq) data, Extended Data Fig. 2b—and gene-body, proximal and distal APREs (Fig. 1c). Equivalent analyses using zebrafish data yielded 256,018 potential regulatory regions, with a significantly higher proportion of these being distal APREs (Fig. 1c; $P < 2.2 \times 10^{-16}$, one-sided Fisher's exact test). A significantly larger global TSS distance in APREs was observed for all vertebrates compared to amphioxus (Fig. 1d), even after correcting for differences in average intergenic

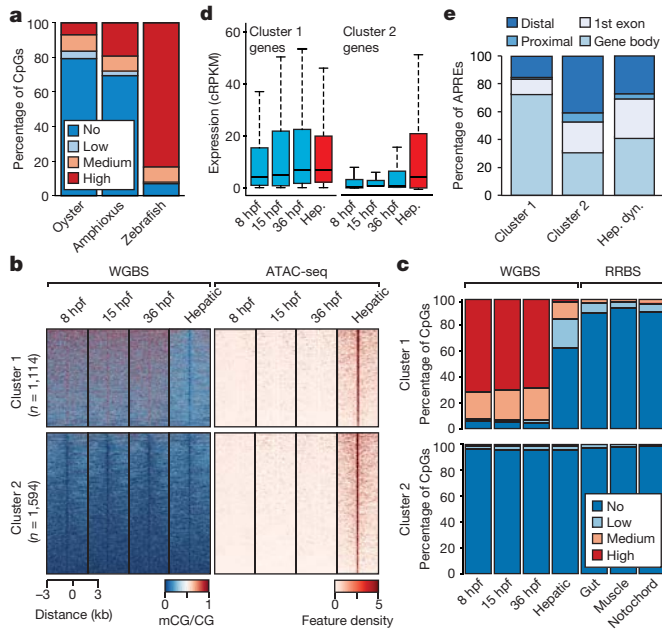


Fig. 2 | 5mC patterns and dynamics in the amphioxus genome. **a**, Percentage of methylated CpG dinucleotides in oyster (mantle, $n = 14,779,123$), amphioxus (8 hpf, $n = 19,657,388$) and zebrafish (1,000-cell stage, $n = 38,989,847$) samples. Low, >0–20%; medium, 20–80%; high, >80%. **b**, k -means clustering ($n = 2$) of 5mC signal over hepatic-specific APREs. **c**, Percentage of methylated CpG dinucleotides as assessed by whole-genome bisulfite sequencing (WGBS) and reduced representation bisulfite sequencing (RRBS) in embryos and adult tissues in APREs from **b**. **d**, Distribution of expression levels for genes associated with APREs displaying distinct 5mC patterns in **b**. Cluster 1: 1,114 genes; cluster 2: 1,594 genes. cRPKM, corrected (per mappability) reads per kb of mappable positions and million reads. Hep, hepatic diverticulum. **e**, Genomic distribution of regions with distinct 5mC patterns from **b**. Hep. dyn., dynamic APREs active in the hepatic diverticulum.

length among species (Extended Data Fig. 2c; $P < 2.2 \times 10^{-16}$ for all vertebrate-versus-amphioxus comparisons, one-sided Mann–Whitney tests). Amphioxus APREs showed enrichment for enhancer-associated

chromatin marks (Extended Data Fig. 2d), which were highly dynamic during embryo development (Extended Data Fig. 2e–g), and consistently drove GFP expression in zebrafish or amphioxus transgenic assays (93% (14/15), Fig. 1e and Extended Data Fig. 2h, i). Moreover, 89% (32/36) of previously reported amphioxus enhancers overlapped APREs defined by our data. Therefore, a large fraction of APREs probably act as developmentally regulated transcriptional enhancers.

Disentangling vertebrate bidirectional promoters

Analyses of core promoters, defined by CAGE-seq, at single-nucleotide resolution revealed that amphioxus promoters display a mixture of pan-metazoan, pan-vertebrate and unique features (Extended Data Fig. 3 and Supplementary Information). These analyses also identified that 25% (3,950/15,884) of neighbouring protein-coding genes were arranged in bidirectional promoters. Bidirectional promoters were most common among ubiquitous promoters (Extended Data Fig. 4a), displayed a marked periodicity in the distance between promoters (Extended Data Fig. 4b, c) and were associated with genes that were significantly enriched in housekeeping functions (Extended Data Fig. 4d). Notably, the fraction of bidirectional promoters defined by CAGE-seq decreased progressively from amphioxus to mouse (12.83% (1,752/13,654)) and to zebrafish (7.84% (1,098/14,014)), which suggests a disentanglement of ancestral bidirectional promoters after each round of WGD (two in tetrapods and three in teleosts). Consistently, the majority of a set of 372 putatively ancestral, bidirectional promoters were lost in vertebrates—particularly in stem vertebrates (54.5%)—with only very few amphioxus-specific losses (5.3%) (Extended Data Fig. 4e, f).

Developmental DNA demethylation of APREs

Similar to other non-vertebrates^{17–19}, the amphioxus genome exhibited very low levels of CpG methylation (Fig. 2a); nearly all of the 5mC occurred in gene bodies, in which the proportion of methylated CpGs correlated positively with gene-expression levels but negatively with the density of H3K27me3 and H3K4me3 histone marks and CpG dinucleotides (Extended Data Fig. 5a–c). However, as in zebrafish and frogs⁷, global levels of 5mC displayed a decrease during development (Extended Data Fig. 5d–g), coinciding with the onset of expression of the amphioxus orthologue of TET demethylase (Extended Data Fig. 5h).

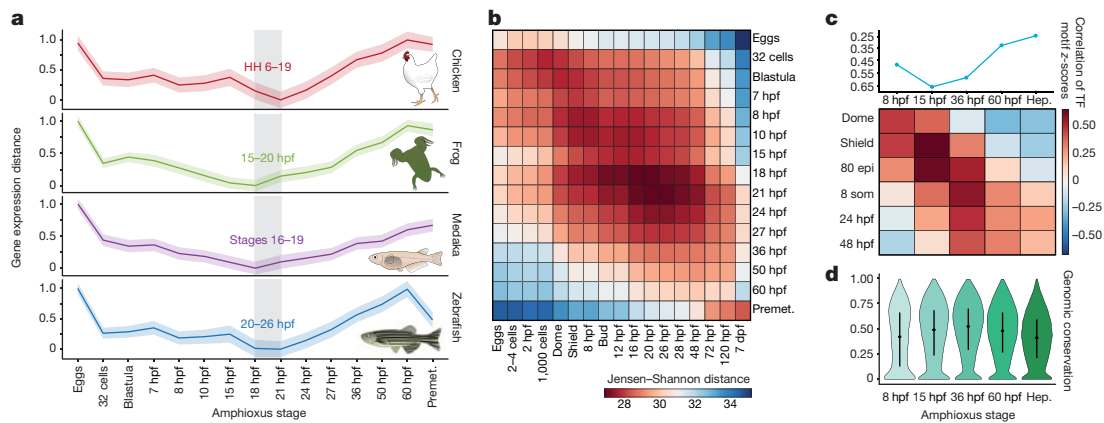


Fig. 3 | The hourglass model and chordate embryogenesis. **a**, Stages of minimal transcriptomic divergence (using the Jensen–Shannon distance metric) from four vertebrate species to each amphioxus stage. The grey box outlines the period of minimal divergence, with the corresponding vertebrate periods indicated (the range is given by the two less divergent stages). Dispersions correspond to the standard deviation computed on 100 bootstrap re-samplings of the orthologue sets (amphioxus–chicken: 5,720; amphioxus–zebrafish: 5,673; amphioxus–frog: 5,883; and amphioxus–medaka: 5,288). HH, Hamburger–Hamilton stage. **b**, Heat map of pairwise transcriptomic Jensen–Shannon distances between amphioxus (vertical) and zebrafish (horizontal) stages. A smaller distance (red) indicates higher similarity. dpf, days post-fertilization. **c**, Zebrafish and amphioxus pairwise Pearson correlation of relative enrichment z -scores for transcription-factor (TF) motifs in dynamic APREs, active at different developmental stages. Top, maximal correlation for each amphioxus stage against the zebrafish stages. Bottom, heat map with all pairwise correlations. 80 epi, 80% epiboly stage; 8 som, 8-somite stage. **d**, Sequence conservation levels within the cephalochordates of active APREs at each developmental stage, visualized as the distribution of average phastCons scores. The number of APREs at 8 hpf = 5,282; at 15 hpf = 17,387; at 36 hpf = 21,089; at 60 hpf = 22,674; and in hepatic diverticulum (hep) = 16,551. Dots correspond to the mean values and lines represent the interquartile range.

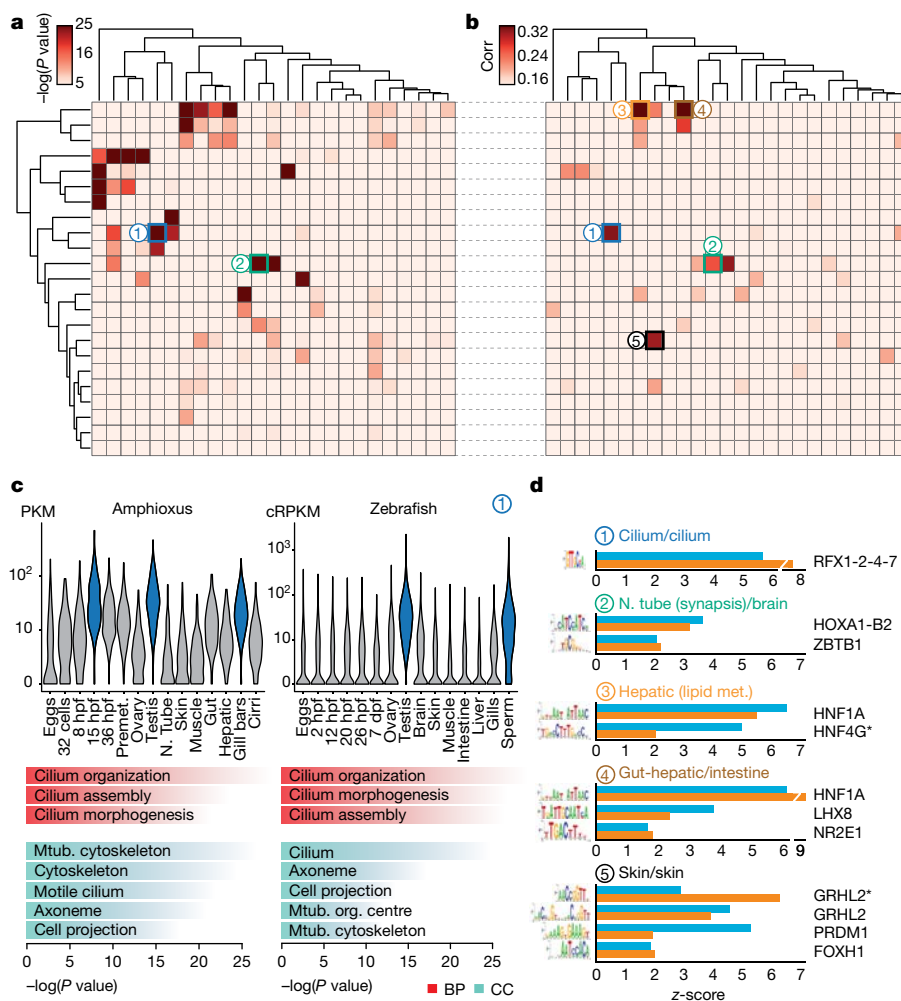


Fig. 4 | Transcriptomic and *cis*-regulatory conservation of adult chordate tissues. **a**, Heat map showing the level of raw statistical significance of orthologous gene overlap between modules produced by weighted correlation network analysis (WGCNA), from amphioxus (vertical) and zebrafish (horizontal) as derived from upper-tail hypergeometric tests. **b**, Heat map of all pairwise Pearson correlations (corr) between the modules of the two species, based on the relative z-scores of transcription-factor motifs for each module (242 super-families of motifs). Modules are clustered as in **a**. **c**, Distribution of expression values (cRPKMs) for all genes within the cilium modules across each sample (top), and enriched Gene Ontology terms within each module (bottom) for a pair of modules (labelled '1' in **b**; 1,681 and 261 genes in zebrafish and amphioxus, respectively). BP, biological process; CC, cellular component. *P* values correspond to uncorrected two-sided Fisher's exact tests as provided by topGO. Mtub., microtubule; N. tube, neural tube; org., organizing. **d**, Transcription-factor binding-site motifs with high z-scores from highly correlated pairs of modules between zebrafish (blue) and amphioxus (orange). Numbers correspond to those circles in **b**. RFX1-2-4-7 denotes RFX1, RFX2, RFX4 and RFX7; HOXA1-B2 denotes HOXA1 and HOXB2; asterisk denotes an alternative motif.

To assess whether these 5mC dynamics may have regulatory potential, we identified adult hepatic diverticulum-specific APREs that are inactive during development. Unlike embryo-specific APREs (Extended Data Fig. 6a), the clustering of these adult APREs on the basis of 5mC content revealed two distinct subsets, one with hepatic-specific and one with constitutive hypomethylation (Fig. 2b). Differentially methylated APREs (cluster 1) also displayed robust hypomethylation in other adult tissues (Fig. 2c), which suggests that demethylation at these APREs occurs organism-wide. Both groups of hepatic-specific APREs were enriched for binding sites of liver-specific transcription factors—such as Hnf4a—as well as broadly expressed transcription factors such as Foxa (Extended Data Fig. 6b), which is a pioneer factor that participates in 5mC removal at regulatory regions in mammals²⁰.

APREs from both clusters were preferentially associated with genes with metabolic functions (Extended Data Fig. 6c). However, only APREs with hepatic-specific hypomethylation (cluster 1) were primarily associated with genes that displayed steady widespread expression (Fig. 2d and Extended Data Fig. 6d, e); these APREs were mainly located within gene bodies (Fig. 2e). These data suggest that demethylation of these APREs may contribute to their identification as adult-specific, transcriptional *cis*-regulatory elements within continuously hypermethylated gene-body contexts, which is characteristic of non-vertebrate species. Fourteen zebrafish gene families contained differentially methylated APREs in introns that are orthologous to those identified in amphioxus—amongst these are four genes that encode components of the Hippo pathway, including the transcriptional effectors Yap (*yap1* and *wvtr1*) and Tead (*tead1a* and *tead3a*) (Extended Data Fig. 6f, g).

The hourglass model and chordate embryogenesis

Previous comparative analyses among vertebrate transcriptomes^{21,22} showed a developmental period of maximal similarity in gene expression that coincides with the so-called phylotypic period, consistent with the hourglass model²³. However, similar comparisons with tunicates and amphioxus have thus far not resolved a phylotypic period shared across all chordates²². Pairwise comparisons of stage-specific RNA sequencing (RNA-seq) data from developmental time courses of amphioxus against zebrafish, medaka, frog (*Xenopus tropicalis*) and chicken revealed a consistent period of highest similarity (Fig. 3a, b and Extended Data Fig. 7) that occurred slightly earlier than those reported for vertebrates; in amphioxus, this corresponds to the neurula at the 4–7-somite stage (18–21 hours post fertilization (hpf)). At the regulatory level, pairwise comparisons between the relative enrichment of transcription-factor motifs in sets of dynamic APREs that were active at each stage were also consistent with an earlier hourglass model²⁴ (Fig. 3c). By contrast, at a shorter timescale, comparisons between different species of amphioxus showed that the sequence conservation for the same APREs was higher after the putative chordate phylotypic period (Fig. 3d).

Regulatory conservation shapes chordate body plan

Additional comparisons of embryo transcriptomes and neighbour hood analysis of conserved co-expression²⁵ showed a high conservation of developmental and global expression patterns and of gene functions between amphioxus and vertebrates (Extended Data Fig. 8 and Supplementary Information). Further pairwise comparison of co-regulated gene modules across tissues between amphioxus and zebrafish revealed multiple pairs with highly significant levels of orthologue overlap (Fig. 4a). These included modules with conserved tissue-specific

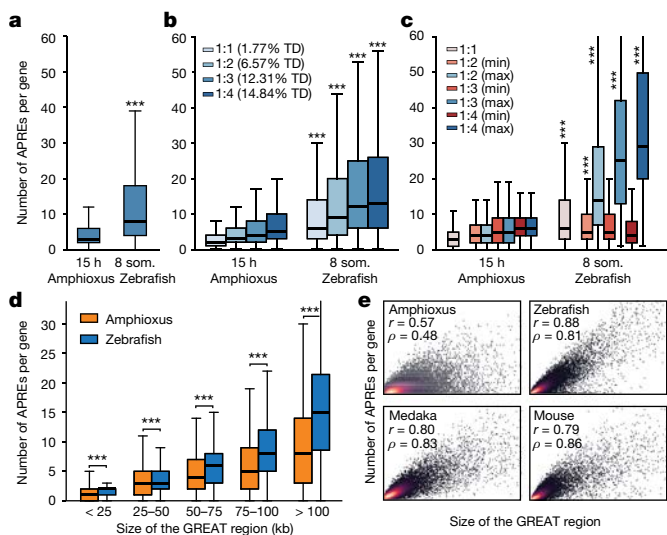


Fig. 5 | Higher regulatory complexity in vertebrate regulatory landscapes. **a**, Distribution of the number of APREs within the regulatory landscape of each gene (as estimated by GREAT), at comparable pre-phylogenetic developmental stages (15 hpf for amphioxus and 8 somites for zebrafish). $n = 6,047$ and $9,239$ genes for amphioxus and zebrafish, respectively. **b**, As in **a**, but with gene families separated according to the number of retained ohnologues per family in vertebrates (from 1 to 4, using mouse as a reference). The percentage of developmental regulatory genes (TD, trans-dev) in each category is indicated. **c**, As in **b**, but only the genes with the lowest ('min', in red) and the highest ('max', in blue) number of APREs are plotted for each ohnologue family. **d**, Distributions of the number of APREs per gene among subsets of amphioxus and zebrafish genes matched by GREAT-region size (± 500 bp) and binned by size as indicated. **e**, Density scatter plot of the number of APREs (y axis) versus the size of the GREAT region (x axis) per gene and species. Pearson (r) and Spearman (ρ) correlation coefficients are indicated. Sample sizes: amphioxus, 20,053; zebrafish, 20,569; medaka, 15,978; mouse, 18,838. **a–d**, $*** P < 0.001$; one-sided Mann–Whitney tests of the zebrafish distribution versus the equivalent amphioxus distribution. Exact P values and sample sizes are provided in Supplementary Data 2, dataset 8.

expression that were enriched for coherent Gene Ontology categories, including genes with high expression in organs with ciliated cells (for example, spermatozoa and gill bars) (labelled '1' in Fig. 4a–c) as well as neural, muscle, gut, liver, skin and metabolism-related modules (Supplementary Data 1). We also found a significant positive correlation between relative motif-enrichment scores for many pairs of modules (Fig. 4b); the most-enriched transcription-factor motifs within each cluster were highly consistent between amphioxus and zebrafish (Fig. 4d).

Higher regulatory information in vertebrate genomes

To investigate the effect of WGDs on the evolution of vertebrate gene regulation, we first asked whether the number of putative regulatory regions per gene is higher in vertebrates than in amphioxus. We observed significantly more APREs in the regulatory landscape of each gene (as defined by the 'Genomic Regions Enrichment of Annotations Tool' (GREAT)²⁶) in zebrafish than in amphioxus (Fig. 5a). This difference is particularly evident for gene families that have retained multiple copies after WGD (known as ohnologues; Fig. 5b), for which the number of APREs is very uneven between copies, with marked regulatory expansions observed for some ohnologues (Fig. 5c). The same patterns were detected for all developmental stages of amphioxus and zebrafish, as well as for medaka and mouse genomes, and were highly robust to down-sampling of ATAC-seq coverage in vertebrates (Extended Data Fig. 9a–c). We also detected a higher number of peaks associated with regulatory genes ('trans-dev' genes that are involved in the regulation of embryonic development) compared to housekeeping genes in all species (Extended Data Fig. 9d), consistent

with the higher frequency of retention of trans-dev genes in multiple copies after WGD³ (Fig. 5b). Comparison of regulatory landscapes—determined experimentally using circular chromosome conformation capture followed by sequencing (4C-seq)—for 58 genes from 11 trans-dev gene families in amphioxus, zebrafish and mouse showed similar results (Extended Data Fig. 9e).

As expected, the higher number of APREs in zebrafish was associated with larger intergenic regions in this species (Extended Data Fig. 9f). However, the differences in APRE complements were not attributable only to an increase in genome size in vertebrates, as subsets of amphioxus and zebrafish genes with matched distributions of GREAT or intergenic-region lengths also displayed a higher number of APREs in zebrafish (Extended Data Fig. 9g, h). Further investigation of matched distributions showed that these differences were particularly great in genes with large regulatory landscapes (>50 kb) (Fig. 5d). Thus, larger regions in amphioxus did not scale at the same rate as in vertebrates in terms of regulatory complexity (Fig. 5e), which is consistent with the overall lower proportion of distal APREs identified in this species (Fig. 1c, d). In summary, these analyses reveal a large increase in the number of regulatory regions during vertebrate evolution (and/or a decrease in these regions in amphioxus)—particularly of distal regulatory elements—and that this trend is enhanced for specific gene copies retained after the WGDs, pointing to unequal rates of regulatory evolution for different ohnologues.

More-complex regulation in specialized ohnologues

The duplication–degeneration–complementation (DDC) model hypothesizes that the retention of duplicate genes could be driven by reciprocal loss of regulatory elements and restriction of paralogues to distinct subsets of the ancestral expression pattern²⁷. In particular, the DDC model predicts that individual paralogues would each have more restricted expression than an unduplicated outgroup, but that their summation would not. To test this, we binarized the expression ('on' or 'off') of each gene in nine homologous expression domains in amphioxus, zebrafish, frog and mouse (Fig. 6a). When comparing genes that returned to single-copy status after WGDs, we detected no expression bias between amphioxus and vertebrates (Fig. 6a, b and Extended Data Fig. 10a, b). By contrast, when vertebrate ohnologues were compared to their single amphioxus orthologues, the distributions were strongly skewed and many vertebrate genes displayed far more restricted expression domains (Fig. 6b and Extended Data Fig. 10a, b; similar results were obtained by comparing τ values²⁸, Extended Data Fig. 10c–e). The symmetrical pattern was fully recovered when the expression of all vertebrate members was combined, or when the raw expression values were summed for each member within a paralogy group (Fig. 6a, b and Extended Data Fig. 10a, b).

Although the above findings are consistent with the DDC model, they are also compatible with an alternative model in which a subset of duplicate genes becomes more 'specialized' in expression pattern while one or more paralogues retain the broader ancestral expression²⁹. To distinguish between these alternatives, we analysed a subset of multi-gene families in which both the single amphioxus orthologue and the union of the vertebrate ohnologues—and thus probably the ancestral gene—were expressed across all nine samples that we compared. We then identified (i) gene families in which all vertebrate paralogues were expressed in all domains (termed 'redundancy'), (ii) gene families in which none of the vertebrate members had expression across all domains (termed 'subfunctionalization')²⁷ and (iii) gene families in which one or more vertebrate ohnologues were expressed in all domains, but at least one ohnologue was not (termed 'specialization') (Fig. 6c). We obtained very similar results for the three vertebrate species we studied (Fig. 6d): between 80 and 88% of gene families were subfunctionalized or specialized, which implies that ancestral expression domains have been lost in at least one member. Moreover, specialization was consistently more frequent than subfunctionalization as a fate for ohnologues with broad ancestral expression.

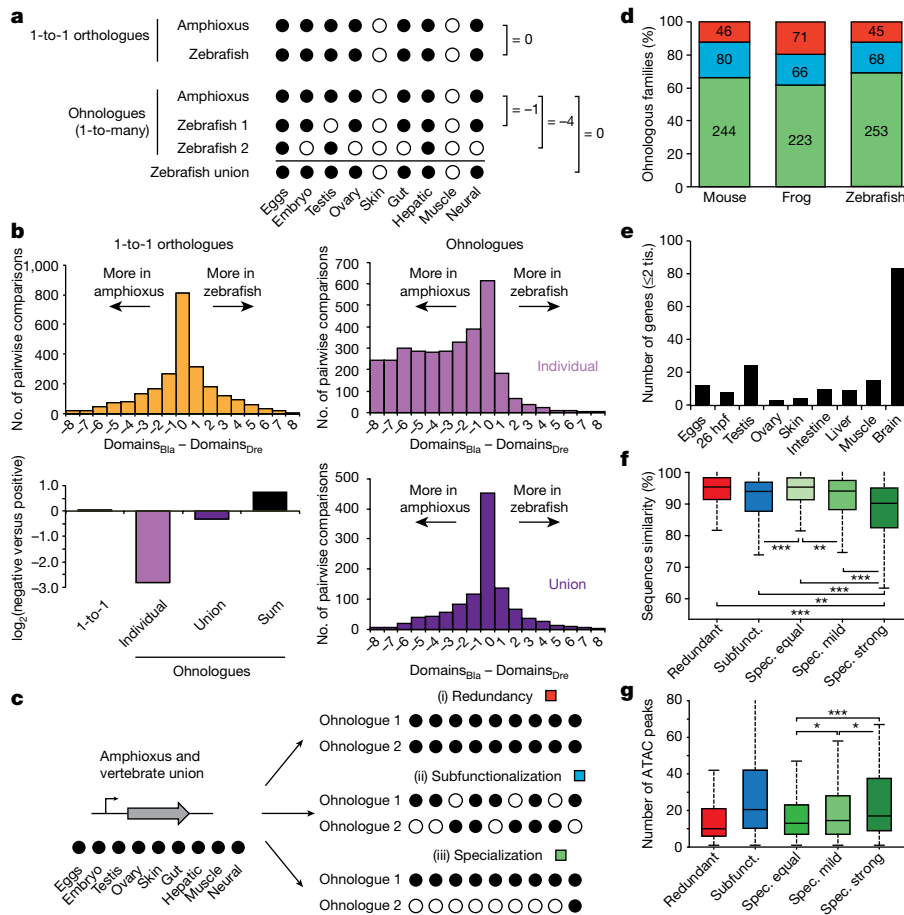


Fig. 6 | Expression specialization is the main fate after WGD.

a, Schematic of the analysis shown in **b**. Expression is binarized for each gene across the nine homologous samples ('on', black dots; normalized cRPKM > 5). **b**, Distribution of the difference in positive domains between zebrafish (domains_{Dre}) and amphioxus (domains_{Gla}) for 1-to-1 orthologues (2,478 gene pairs; yellow), individual ohnologues (3,427 gene pairs in 1,135 families; lilac) and the union of all vertebrate ohnologues in a family (purple). Bottom left, log₂ of the ratio between zebrafish genes with negative and positive score for each category. 'Sum' (black), binarization of family expression after summing the raw expression values for all ohnologues. **c**, Schematic of the analyses shown in **d**, representing the three possible fates after WGD. **d**, Distribution of fates after WGD for families of ohnologues. **e**, Number of ohnologues with strong

specialization in zebrafish expressed in each domain. Tis., tissue. **f**, Distribution of the percentage of nucleotide sequence similarity between human and mouse by family type. Ohnologues from specialized families are divided into 'spec. equal' (which maintain all expression domains), 'spec. mild' (which have lost some but maintained more than two expression domains) and 'spec. strong' (≤2 expression domains remain). Subfunct., subfunctionalized. **g**, Distribution of the number of APREs within GREAT regions for zebrafish ohnologues for each category. Only statistical comparisons within specialized families are shown. *P* values in **f** and **g** correspond to two- and one-sided Wilcoxon sum-rank tests between the indicated groups, respectively. *0.05 > *P* value ≥ 0.01, **0.01 > *P* value ≥ 0.001, ****P* value < 0.001. Exact *P* values and sample sizes are provided in Supplementary Data 2, dataset 8.

Ohnologues that have experienced strong specialization (≤2 remaining expression domains) retained expression more often in neural tissues (Fig. 6e and Extended Data Fig. 10f–i) and were generally not expressed in additional vertebrate-specific tissues (Supplementary Information). Furthermore, they showed the fastest rates of sequence evolution (Fig. 6f and Extended Data Fig. 10j–l), consistent with an optimization of their coding sequence to perform their function in a specific tissue and/or with the evolution of novel functions (neofunctionalization). Ohnologues from specialized families that have lost expression domains showed significantly more associated APREs than ohnologues with the full ancestral expression (Fig. 6g). We observed a strong positive relationship between the number of ancestral expression domains lost and the number of APREs associated with specialized ohnologues (Extended Data Fig. 10m). This implies that the specialization of gene expression after WGD does not occur primarily through loss of ancestral tissue-specific enhancers, but rather by a complex remodelling of regulatory landscapes that involves recruitment of novel, tissue-specific regulatory elements.

Discussion

By applying functional genomics approaches to the cephalochordate amphioxus, we have deepened our understanding of the origin and

evolution of chordate genomes. We identified APREs in amphioxus, the activation of which is tightly associated with differential DNA demethylation in adult tissues—a mechanism previously thought to be specific to vertebrates. Additional cases may be subsequently found in other non-vertebrate species when similar multi-omics datasets are analysed. In amphioxus, APREs of this type usually fall within gene bodies of widely expressed genes, which suggests that gene regulation by demethylation could have originated as a mechanism to allow better definition of enhancers in a hyper-methylated intragenic context. If so, this mechanism could have been co-opted into new genomic contexts—that is, distal intergenic enhancers—later in the evolution of vertebrate genomes, which are characterized by their pervasive, genome-wide hypermethylation.

We also found a consistently higher number of open chromatin regions per gene in vertebrates than in amphioxus. This pattern is observed at a genome-wide level, but is particularly evident for distal APREs and in gene families that retain multiple ohnologues after WGD; these families are enriched for regulatory genes with large regulatory landscapes. Finally, we detected a large degree of specialization in expression for retained ohnologues, with the vast majority of multi-gene families with broad ancestral expression having at least one member

that restricted its expression breadth. Through this mechanism, vertebrates have increased their repertoire of tightly regulated genes, which has potentially contributed to tissue-specific evolution. Gene-expression specialization was accompanied by faster evolution of protein-coding sequences, and by an increase—rather than a decrease—in the number of regulatory elements. Taken together, these observations indicate that the two rounds of WGD not only caused an expansion and diversification of gene repertoires in vertebrates, but also allowed functional and expression specialization of the extra copies by increasing the complexity of their gene regulatory landscapes. We suggest that these changes to the gene regulatory landscapes underpinned the evolution of morphological specializations in vertebrates.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at <https://doi.org/10.1038/s41586-018-0734-6>.

Received: 22 November 2017; Accepted: 18 October 2018;

Published online 21 November 2018.

- Bertrand, S. & Escriva, H. Evolutionary crossroads in developmental biology: amphioxus. *Development* **138**, 4819–4830 (2011).
- Dehal, P. & Boore, J. L. Two rounds of whole genome duplication in the ancestral vertebrate. *PLoS Biol.* **3**, e314 (2005).
- Putnam, N. H. et al. The amphioxus genome and the evolution of the chordate karyotype. *Nature* **453**, 1064–1071 (2008).
- Holland, L. Z. et al. The amphioxus genome illuminates vertebrate origins and cephalochordate biology. *Genome Res.* **18**, 1100–1111 (2008).
- International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
- Nelson, C. E., Hersh, B. M. & Carroll, S. B. The regulatory content of intergenic DNA shapes genome architecture. *Genome Biol.* **5**, R25 (2004).
- Bogdanović, O. et al. Active DNA demethylation at enhancers during the vertebrate phylogenetic period. *Nat. Genet.* **48**, 417–426 (2016).
- Berthelot, C., Villar, D., Horvath, J. E., Odom, D. T. & Flicek, P. Complexity and conservation of regulatory landscapes underlie evolutionary resilience of mammalian gene expression. *Nat. Ecol. Evol.* **2**, 152–163 (2018).
- Reilly, S. K. et al. Evolutionary changes in promoter and enhancer activity during human corticogenesis. *Science* **347**, 1155–1159 (2015).
- Villar, D. et al. Enhancer evolution across 20 mammalian species. *Cell* **160**, 554–566 (2015).
- Vierstra, J. et al. Mouse regulatory DNA landscapes reveal global principles of cis-regulatory evolution. *Science* **346**, 1007–1012 (2014).
- Boyle, A. P. et al. Comparative analysis of regulatory information and circuits across distant species. *Nature* **512**, 453–456 (2014).
- Gerstein, M. B. et al. Comparative analysis of the transcriptome across distant species. *Nature* **512**, 445–448 (2014).
- Hendrich, B. & Tweedie, S. The methyl-CpG binding domain and the evolving role of DNA methylation in animals. *Trends Genet.* **19**, 269–277 (2003).
- Irimia, M. et al. Extensive conservation of ancient microsynteny across metazoans due to cis-regulatory constraints. *Genome Res.* **22**, 2356–2367 (2012).
- Simakov, O. et al. Insights into bilaterian evolution from three spiralian genomes. *Nature* **493**, 526–531 (2013).
- Wang, X. et al. Genome-wide and single-base resolution DNA methylomes of the Pacific oyster *Crassostrea gigas* provide insight into the evolution of invertebrate CpG methylation. *BMC Genomics* **15**, 1119 (2014).
- Albalat, R., Martí-Solans, J. & Cañestro, C. DNA methylation in amphioxus: from ancestral functions to new roles in vertebrates. *Brief. Funct. Genomics* **11**, 142–155 (2012).
- Huang, S. et al. Decelerated genome evolution in modern vertebrates revealed by analysis of multiple lancelet genomes. *Nat. Commun.* **5**, 5896 (2014).
- Zhang, Y. et al. Nucleation of DNA repair factors by FOXA1 links DNA demethylation to transcriptional pioneering. *Nat. Genet.* **48**, 1003–1013 (2016).
- Irie, N. & Kuratani, S. Comparative transcriptome analysis reveals vertebrate phylogenetic period during organogenesis. *Nat. Commun.* **2**, 248 (2011).
- Hu, H. et al. Constrained vertebrate evolution by pleiotropic genes. *Nat. Ecol. Evol.* **1**, 1722–1730 (2017).
- Duboule, D. Temporal colinearity and the phylogenetic progression: a basis for the stability of a vertebrate Bauplan and the evolution of morphologies through heterochrony. *Development* **1994 Suppl.**, 135–142 (1994).
- Bogdanović, O. et al. Dynamics of enhancer chromatin signatures mark the transition from pluripotency to cell specification during embryogenesis. *Genome Res.* **22**, 2043–2053 (2012).

- Yue, F. et al. A comparative encyclopedia of DNA elements in the mouse genome. *Nature* **515**, 355–364 (2014).
- McLean, C. Y. et al. GREAT improves functional interpretation of cis-regulatory regions. *Nat. Biotechnol.* **28**, 495–501 (2010).
- Force, A. et al. Preservation of duplicate genes by complementary, degenerative mutations. *Genetics* **151**, 1531–1545 (1999).
- Yanai, I. et al. Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. *Bioinformatics* **21**, 650–659 (2005).
- Sandve, S. R., Rohlf, R. V. & Hvidsten, T. R. Subfunctionalization versus neofunctionalization after whole-genome duplication. *Nat. Genet.* **50**, 908–909 (2018).

Acknowledgements This research was funded primarily by the European Research Council (ERC) under the European Union's Horizon 2020 and Seventh Framework Program FP7 research and innovation programs (ERC-AdG-LS8-740041 to J.L.G.-S., ERC-StG-LS2-637591 to M.I., a Marie Skłodowska-Curie Grant (658521) to I.M. and a FP7/2007-2013-ERC-268513 to P.W.H.H.), the Spanish Ministerio de Economía y Competitividad (BFU2016-74961-P to J.L.G.-S., RYC-2016-20089 to I.M., BFU2014-55076-P and BFU2017-89201-P to M.I. and BFU2014-55738-REDT to J.L.G.-S., M.I. and J.R.M.-M.), the 'Centro de Excelencia Severo Ochoa 2013-2017' (SEV-2012-0208), the 'Unidad de Excelencia María de Maetzu 2017-2021' (MDM-2016-0687), the People Program (Marie Curie Actions) of the European Union's Seventh Framework Program FP7 under REA grant agreement number 607142 (DevCom) to J.L.G.-S., and the CNRS and the ANR (ANR16-CE12-0008-01) to H.E. O.B. was supported by an Australian Research Council Discovery Early Career Researcher Award (DECRA; DE140101962). We acknowledge the support of the CERCA Programme/Generalitat de Catalunya and of the Spanish Ministry of Economy, Industry and Competitiveness (MEIC) to the EMBL partnership. Additional sources of funding for all authors are listed in Supplementary Information.

Reviewer information *Nature* thanks D. Duboule and P. Flicek for their contribution to the peer review of this work.

Author contributions F.M., P.N.F., I.M., J.J.T., O.B., M.P., B.L., P.W.H.H., H.E., J.L.G.-S. and M.I. contributed to concept and study design. F.M., P.N.F., I.M., J.J.T., O.B., M.P., C.D.R.W., R.D.A., S.J.v.H., C.H.-U., K.S., Y.M., A. Louis, P.J.B., P.E.D., M.T.W., J.G.-F., R.L., B.L., P.W.H.H., J.L.G.-S. and M.I. performed computational analyses and data interpretation. O.B., E.d.I.C.-M., S.B., D.B., R.D.A., S.N., S.J.-G., D.A., L.B., J.P., B.A.-C., Y.L.P., A. Leon, L.S., E.F., P.C., J.R.M.-M., R.L., B.L., H.E., J.L.G.-S. and M.I. obtained biological material and generated next-generation sequencing data. I.M., J.J.T., E.d.I.C.-M., I.K., R.D.A., Z.K. and J.L.G.-S. performed transgenic assays. J.-M.A., S.M. and P.W. sequenced the genome. R.A., E.B.-G., C.C., F.C., S.D., D.E.K.F., S.H., V.L., G.A.B.M., P.P., M.S., H.S., I.S., T.T., O.M., A.X. and J.-K.Y. contributed to genome sequencing and gene family curation. I.M., H.E., J.L.G.-S. and M.I. coordinated the project. F.M., I.M., P.W.H.H. and M.I. wrote the main text, with input from all authors. Detailed contributions are listed in Supplementary Information. Animal illustrations by J.J.T., released under a Creative Commons Attribution (CC-BY) Licence.

Competing interests The authors declare no competing interests.

Additional information

Extended data is available for this paper at <https://doi.org/10.1038/s41586-018-0734-6>.

Supplementary information is available for this paper at <https://doi.org/10.1038/s41586-018-0734-6>.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

Correspondence and requests for materials should be addressed to I.M., H.E., J.L.G. or M.I.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or

format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

METHODS

No statistical methods were used to predetermine sample size. The experiments were not randomized and investigators were not blinded to allocation during experiments and outcome assessment.

Animal husbandry and embryo staging. Amphioxus gametes were obtained by heat stimulation as previously described^{30,31}. Embryos were obtained by in vitro fertilization in filtered seawater and cultured at 19 °C. Staging was done based on previous publications^{32,33}; correspondence between developmental stages and hpf are provided in Supplementary Table 1. All protocols used for vertebrate species (zebrafish and medaka) have been approved by the Institutional Animal Care and Use Ethic Committee (PRBB–IACUEC, for CRG) or the Ethics Committee of the Andalusian Government (license numbers 450-1839 and 182-41106, for CABD-CSIC), and implemented according to national and European regulations. All experiments were carried out in accordance with the principles of the 3Rs (replacement, reduction and refinement).

Genome sequencing and assembly. Genomic DNA was extracted from a single *B. lanceolatum* adult male collected in Argeles-sur-Mer, France. The genome was sequenced using a combination of Illumina libraries from a range of inserts at Genoscope (897 million reads in total, with a paired-end coverage of 150×; Supplementary Table 2). A diploid assembly was generated using SOAPdenovo assembler³⁴ using a *k*-mer of 71. After gap closing, haplotypes were reconciled with Haplomerger³⁵.

Genome annotation. We generated deep coverage RNA-seq for 16 developmental stages and 9 adult tissues (4.16 billion reads in total). The bulk of strand-specific transcriptomic data was assembled de novo with Trinity³⁶, aligned and assembled into loci with the PASA pipeline³⁷. De novo gene models were built using Augustus³⁸ and subsequently refined with EVM³⁹ using PASA assemblies and aligned proteins from other species. In parallel, all strand-specific RNA-seq reads were mapped to the genome using Tophat⁴⁰, assembled using Cufflinks⁴¹ and open reading frames were predicted using Trans-decoder⁴². Models obtained using both these approaches were reconciled yielding a total 218,070 transcripts from 90,927 unified loci, of which 20,569 were protein-coding and had homologues in at least one of the other studied species (see ‘Comparative genomics’). Gene Ontology (GO) terms were assigned to amphioxus proteins based on their PFAM and Interpro domains, as well as blastp hits against human proteins (1×10^{-6}).

Repeats were annotated and filtered with RepeatMasker using a custom library generated with RepeatModeller. Long non-coding RNAs were identified by filtering all transcripts for protein-coding potential using CPAT⁴³ trained with zebrafish transcripts, and further discarding those that had a positive hit in a HMM search against the NR and PFAM databases (Extended Data Fig. 1g).

Comparative genomics. We used OMA⁴⁴ to reconstruct gene families and infer homology relationships based on well-established phylogenetic relationships between species⁴⁵, and further merged families sharing Ensembl paralogues with ‘Euteleostomi’ or ‘Vertebrata’ ancestry. To define the set of high-confidence orthologue families (Supplementary Data 2, dataset 9), we retained families with two to four copies in three out of five vertebrates (excluding teleosts) and subjected them to phylogenetic reconciliation.

To assess genome sequence conservation, reciprocal whole-genome alignments of *Branchiostoma floridae*, *Branchiostoma belcheri* and *B. lanceolatum* were performed using LASTZ and processed with phastCons⁴⁶ to produce conservation scores. The distribution of phastCons scores in APREs was determined using ‘dynamic’ ATAC-seq peaks that showed no temporal discontinuity in activity.

Comparative transcriptomics. To investigate the evolutionary conservation of chordate development at the molecular level, newly generated data from zebrafish, medaka and amphioxus, as well as available data from the SRA (frog and chicken), were compared (Supplementary Data 2, dataset 3 and Supplementary Table 3). Gene expression was estimated with Kallisto⁴⁷ using Ensembl transcriptome annotations (Supplementary Table 4), and summing up transcripts per million (TPMs) from all transcript isoforms to obtain one individual gene-expression estimate per sample. We used single-copy orthologues to pair genes and used the Jensen-Shannon distance metrics after quantile normalization of TPMs to score distance between pairs of transcriptomes:

$$JSD_s = \sqrt{\frac{1}{2} \sum_{g=0}^{n_{og}} p_g \times \log \left(\frac{p_g}{\frac{1}{2}(p_g + q_g)} \right) + \frac{1}{2} \sum_{g=0}^{n_{og}} q_g \times \log \left(\frac{q_g}{\frac{1}{2}(p_g + q_g)} \right)}$$

Statistical robustness towards gene sampling was assessed by calculating transcriptomic distances based on 100 bootstrap replicates and estimating the standard deviation over these replicates.

To obtain groups of genes with similar dynamics of expression during development, genes were clustered based on their cRPKM⁴⁸ using the Mfuzz package⁴⁹. For this purpose, eight comparable stages were selected in amphioxus and zebrafish on the basis of conserved developmental landmarks such as fertilization,

gastrulation and organogenesis (Supplementary Table 5). The statistical significance of the orthologous gene overlap between pairs of clusters was assessed using upper-tail hypergeometric tests.

Modules of co-expressed genes across stages and adult tissues were inferred using WGCNA⁵⁰ with default parameters in amphioxus (17 samples) and zebrafish (27 samples) (Supplementary Table 6). The statistical significance of the orthologous gene overlap between pairs of clusters was assessed using upper-tail hypergeometric tests. The numbers of transcription-factor binding-site motifs detected in APREs in the basal regions of genes from any given cluster were standardized using *z*-scores.

To have a general assessment of the extent of conservation or divergence in gene expression among chordates at adult stages, we used neighbourhood analysis of conserved co-expression (NACC)²⁵, a method developed to compare heterogeneous, non-matched sample sets across species. NACC relies on comparisons of average distances between pairs of orthologous (genes A and B), the 20 genes with the closest transcriptomic distance (\bar{A} and \bar{B}) and their reciprocal orthologues in the other species ($\bar{A}B$ and $\bar{B}A$), and is calculated as follows:

$$NACC = \frac{1}{2} [(\bar{A}B - \bar{A}) + (\bar{B}A - \bar{B})]$$

NACC calculations were performed for each family that contained a single amphioxus member and up to eight members in zebrafish and were also performed with randomized orthology relationships as a control.

Regulatory profiling. ATAC-seq. For amphioxus, medaka and zebrafish, ATAC-seq was performed in two biological replicates by directly transferring embryos in the lysis buffer, following the original protocol^{51,52}. ATAC-seq libraries were sequenced to produce an average of 66, 83 and 78 million reads for amphioxus, zebrafish and medaka, respectively. Reads were mapped with Bowtie2 and nucleosome-free pairs (insert < 120 bp) retained for peak-calling using MACS2⁵³, and the irreducible discovery rate was used to assess replicability. Nucleosome positioning was calculated from aligned ATAC-seq data using NucleoATAC⁵⁴. **Chromatin immunoprecipitation with sequencing (ChIP-seq).** Embryos of undetermined gender were fixed in 2% formaldehyde and ChIP was performed as previously described for other species⁵⁵. Chromatin was sonicated and incubated with the corresponding antibody (H3K4me3: ab8580, H3K27ac: ab4729 and HeK27me3: ab6002, from Abcam). An average of 30 million reads per library was generated. Reads were mapped with Bowtie2 and peaks called with MACS2⁵³, assuming default parameters.

4C-seq. Embryos of undetermined gender were fixed in 2% formaldehyde and chromatin was digested with DpnII and Csp6. Specific primers targeted the TSSs of the studied genes and included Illumina adapters. An average 5 million reads were generated for each of the two biological replicates. After mapping, reads were normalized per digestion fragment cut and interactions were identified using peakC⁵⁶ with low-coverage regions excluded.

MethylC-seq and RRBS. Genomic DNA was extracted as previously described⁵⁷, sonicated, purified and end-repaired. Bisulfite conversion was performed with the MethylCode Bisulfite Conversion Kit (Thermo Fisher Scientific). After Illumina library construction, an average of 73 million reads per sample were sequenced. RRBS libraries were prepared similarly to those for MethylC-seq, but with restriction digestion with MspI instead of sonication and PCR amplification. An average of 46 million reads per sample was generated. Reads were mapped to an in silico, bisulfite-converted *B. lanceolatum* reference genome^{7,58}. Differentially methylated regions in the CpG context were identified as previously described⁷. Differential transcription-factor motif enrichment was obtained with DiffBind from Bioconductor.

CAGE-seq. Libraries were constructed using the non-amplifying non-tagging Illumina CAGE protocol⁵⁹. Mouse CAGE-seq data were obtained from FANTOM5⁶⁰. Reads were aligned using Bowtie. Nearby individual CAGE TSSs were combined using the distance-based clustering method in CAGER⁶¹ to produce tag clusters, which summarize expression at individual promoters. Tag clusters were clustered across samples to produce comparable promoter regions, referred to as ‘consensus clusters’. The consensus clusters were then grouped by expression patterns using a self-organizing map⁶². We investigated the relative presence and enrichment of the following features: TATA box, YY1 motif, GC and AT content, SS and WW dinucleotides, first exons and nucleosome positioning signal. Heat maps were plotted for visualization by scanning either for exact dinucleotide matches or for position weight matrix matches at 80% of the maximum score. Position weight matrices for TATA and YY1 were taken from the JASPAR vertebrate collection.

Cis-regulatory comparisons. Depending on the analysis, an APRE was associated with a specific gene if it was located within: (i) the ‘basal’ region of the gene (–5 kb to +1 kb of the TSS; for comparisons of enriched motif composition) or (ii) the GREAT region of the gene (up to ±1 Mb of the TSS unless another basal region was found; for comparing the number of APREs per gene)²⁶. Stratification of gene

sets by GREAT or intergenic-region size between amphioxus and zebrafish was done using the function stratify from the matt suite⁶³, with a range of ± 500 bp.

The DNA-binding specificity of each transcription factor was predicted on the basis of the binding domain similarity to other transcription-factor family members, as previously performed⁶⁴. Transcription-factor motifs from CIS-BP version 1.02⁶⁴ were downloaded and clustered using GimmeMotifs⁶⁵ ($P \leq 0.0001$). Two hundred and forty-two clusters of motifs were assigned to one or more orthologous groups in both amphioxus and zebrafish and used for all analyses (Supplementary Data 2, dataset 10). These motifs were detected in APREs using the tools gimme threshold and gimme scan from GimmeMotifs⁶⁵.

Effect of WGDs on gene expression. Gene expression was binarized (1 if the normalized cRPKM > 5 , and 0 otherwise) across nine comparable samples in amphioxus and three vertebrate species (mouse, frog and zebrafish) (Supplementary Table 7). Then, for each amphioxus gene and vertebrate orthologue, the expression bias was measured by subtracting the number of positive-expression domains in amphioxus from that of vertebrates (Fig. 6a). The amphioxus gene-expression pattern was also compared to the union of the orthologues, as well as the pattern after binarizing the expression for the sum of cRPKM values of all family members. The analysis was restricted to families with a single member in amphioxus

Next, we selected those orthologue families for which the ancestral expression included the nine studied domains, as inferred from having expression in the single amphioxus orthologue and in the union of the family. For each gene family, we then defined (Fig. 6c): (i) redundancy (all vertebrate paralogues were expressed in all domains), (ii) subfunctionalization (none of the vertebrate members had expression across all domains²⁷), and (iii) specialization (one or more vertebrate orthologues were expressed in all domains, but at least one orthologue was not). Members of the later type were subdivided into 'strong' and 'mild' specialization if they retained ≤ 2 or more expression domains. We examined the transcript sequence similarity as well as the dN/dS between human and mouse (retrieved from Biomart), and the number of APREs associated with genes from different categories. Finally, we computed the τ tissue-specificity index as previously described²⁸, to assess more broadly the tissue specificity of orthologues.

Transgenic assays in zebrafish and amphioxus. Enhancer reporter assays in zebrafish embryos were performed as previously described⁶⁶. Selected peaks were first amplified, cloned into a PCR8/GW/TOPO vector and transferred into a detection vector (including a *gata2* minimal promoter, a GFP reporter gene and a strong midbrain enhancer (z48) as an internal control)⁶⁷. Transgenic embryos were generated using the Tol2 transposon and transposase method⁶⁸. Three or more independent stable transgenic lines were generated for each construct as reported in Supplementary Table 8. For amphioxus reporter assays, selected peaks were amplified and transferred into a detection vector (including the *Branchiostoma* minimal actin promoter, a GFP reporter gene and piggyBac terminal repeats). Transgenic embryos were generated by the piggyBac transposase method.

In situ hybridization. Gene fragments that were synthetically designed or amplified by PCR from cDNA were sub-cloned into pBluescript II SK and used as templates for probe synthesis using the DIG labelling kit (Roche) and T3 RNA polymerase. Embryos at different developmental stages were fixed in PFA 4% dissolved in MOPS-EGTA buffer and in situ hybridization carried out as previously described⁶⁹, using BCIP/NBT as a chromogenic substrate.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Code availability. Custom code is available at <https://gitlab.com/groups/FunctionalAmphioxus>.

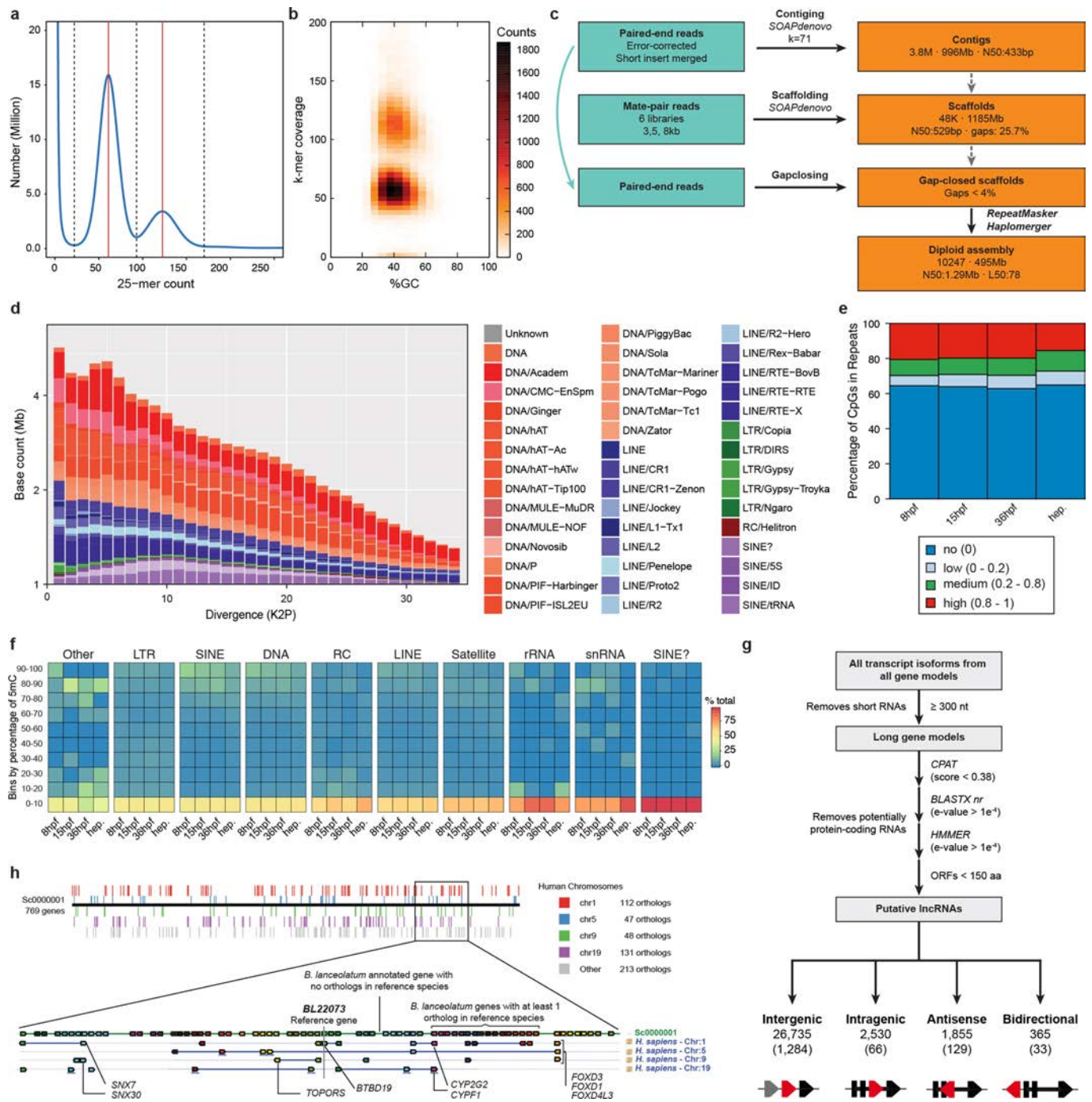
Data availability

Next-generation sequencing data have been deposited in Gene Expression Omnibus (GEO) under the following accession numbers: GSE106372 (ChIP-seq), GSE106428 (ATAC-seq), GSE106429 (CAGE-seq), GSE106430 (RNA-seq), GSE102144 (MethylC-seq and RRBS) and GSE115945 (4C-seq). Raw genome sequencing data and the genome assembly have been submitted to European Nucleotide Archive (ENA) under the accession number PRJEB13665. UCSC hub and annotation files are available at <http://amphiencode.github.io/>.

30. 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.* **302B**, 384–391 (2004).
31. Fuentes, M. et al. Insights into spawning behavior and development of the European amphioxus (*Branchiostoma lanceolatum*). *J. Exp. Zool. B Mol. Dev. Evol.* **308B**, 484–493 (2007).
32. Hirakow, R. & Kajita, N. Electron microscopic study of the development of amphioxus, *Branchiostoma belcheri tsingtauense*: the gastrula. *J. Morphol.* **207**, 37–52 (1991).
33. Hirakow, R. & Kajita, N. Electron microscopic study of the development of amphioxus, *Branchiostoma belcheri tsingtauense*: the neurula and larva. *Kaibogaku Zasshi* **69**, 1–13 (1994).

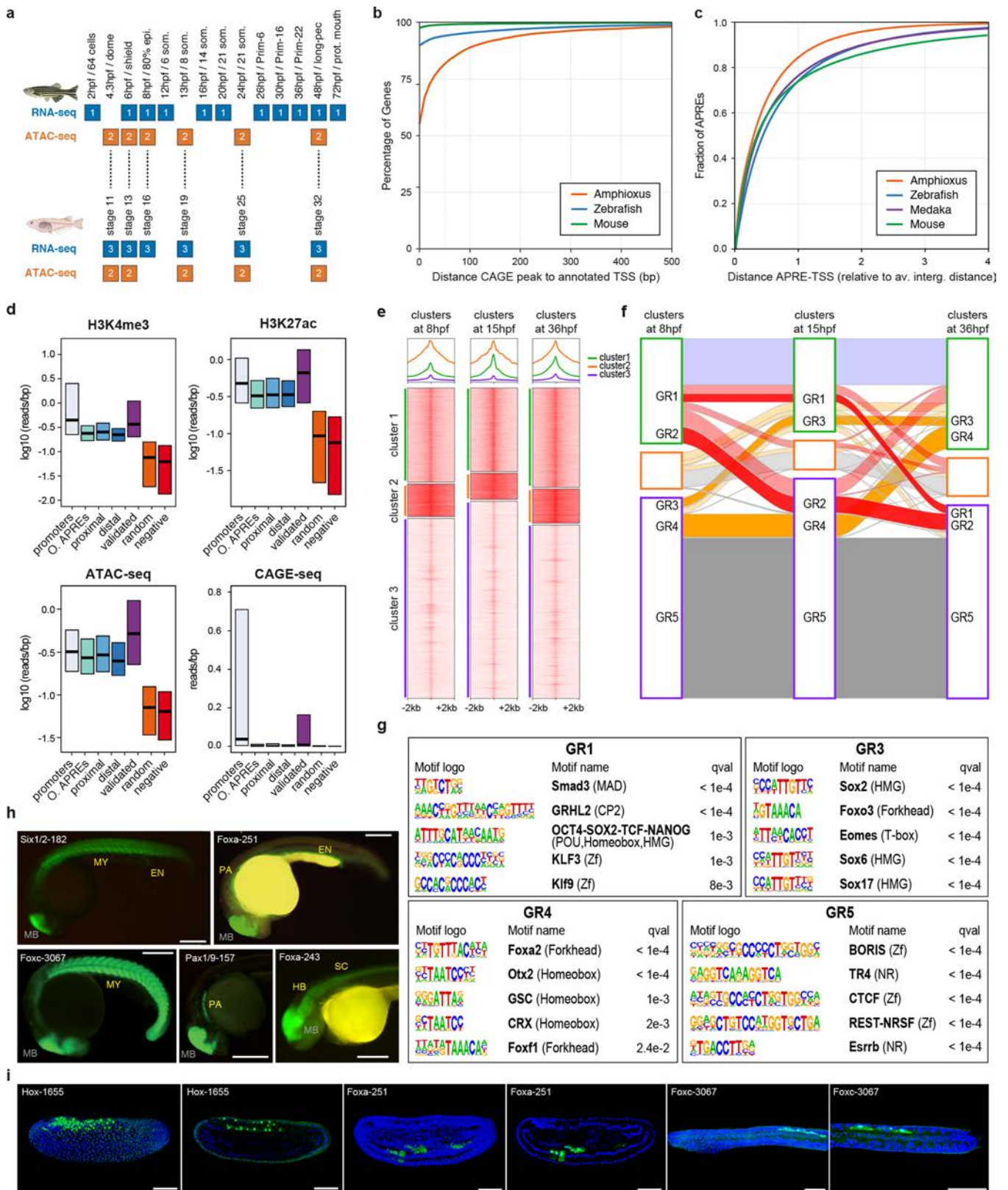
34. Luo, R. et al. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *Gigascience* **1**, 18 (2012).
35. Huang, S. et al. HaploMerger: reconstructing allelic relationships for polymorphic diploid genome assemblies. *Genome Res.* **22**, 1581–1588 (2012).
36. Grabherr, M. G. et al. Full-length transcriptome assembly from RNA-seq data without a reference genome. *Nat. Biotechnol.* **29**, 644–652 (2011).
37. Haas, B. J. et al. Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res.* **31**, 5654–5666 (2003).
38. Keller, O., Kollmar, M., Stanke, M. & Waack, S. A novel hybrid gene prediction method employing protein multiple sequence alignments. *Bioinformatics* **27**, 757–763 (2011).
39. Haas, B. J. et al. Automated eukaryotic gene structure annotation using EvidenceModeler and the program to assemble spliced alignments. *Genome Biol.* **9**, R7 (2008).
40. Kim, D. et al. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.* **14**, R36 (2013).
41. Trapnell, C. et al. Transcript assembly and quantification by RNA-seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* **28**, 511–515 (2010).
42. Haas, B. J. et al. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat. Protocols* **8**, 1494–1512 (2013).
43. Wang, L. et al. CPAT: Coding-Potential Assessment Tool using an alignment-free logistic regression model. *Nucleic Acids Res.* **41**, e74 (2013).
44. Roth, A. C., Gonnet, G. H. & Dessimoz, C. Algorithm of OMA for large-scale orthology inference. *BMC Bioinformatics* **9**, 518 (2008).
45. Altenhoff, A. M., Gil, M., Gonnet, G. H. & Dessimoz, C. Inferring hierarchical orthologous groups from orthologous gene pairs. *PLoS ONE* **8**, e53786 (2013).
46. Siepel, A. et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* **15**, 1034–1050 (2005).
47. Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.* **34**, 525–527 (2016).
48. Labbé, R. M. et al. A comparative transcriptomic analysis reveals conserved features of stem cell pluripotency in planarians and mammals. *Stem Cells* **30**, 1734–1745 (2012).
49. Kumar, L. & Futschik, M. E. Mfuzz: a software package for soft clustering of microarray data. *Bioinformatics* **2**, 5–7 (2007).
50. Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* **9**, 559 (2008).
51. 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**, 1213–1218 (2013).
52. Fernández-Miñán, A., Bessa, J., Tena, J. J. & Gómez-Skarmeta, J. L. Assay for transposase-accessible chromatin and circularized chromosome conformation capture, two methods to explore the regulatory landscapes of genes in zebrafish. *Methods Cell Biol.* **135**, 413–430 (2016).
53. Zhang, Y. et al. Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* **9**, R137 (2008).
54. Schep, A. N. et al. Structured nucleosome fingerprints enable high-resolution mapping of chromatin architecture within regulatory regions. *Genome Res.* **25**, 1757–1770 (2015).
55. Bogdanović, O., Fernández-Miñán, A., Tena, J. J., de la Calle-Mustienes, E. & Gómez-Skarmeta, J. L. The developmental epigenomics toolbox: ChIP-seq and MethylCap-seq profiling of early zebrafish embryos. *Methods* **62**, 207–215 (2013).
56. Geeven, G., Teunissen, H., de Laat, W. & de Wit, E. peakC: a flexible, non-parametric peak calling package for 4C and Capture-C data. *Nucleic Acids Res.* **46**, e91 (2018).
57. Bogdanović, O. & Veenstra, G. J. Affinity-based enrichment strategies to assay methyl-CpG binding activity and DNA methylation in early *Xenopus* embryos. *BMC Res. Notes* **4**, 300 (2011).
58. Lister, R. et al. Hotspots of aberrant epigenomic reprogramming in human induced pluripotent stem cells. *Nature* **471**, 68–73 (2011).
59. Murata, M. et al. Detecting expressed genes using CAGE. *Methods Mol. Biol.* **1164**, 67–85 (2014).
60. The FANTOM Consortium and the RIKEN PMI and CLST (DGT). A promoter-level mammalian expression atlas. *Nature* **507**, 462–470 (2014).
61. Haberer, V., Forrest, A. R., Hayashizaki, Y., Carninci, P. & Lenhard, B. CAGER: precise TSS data retrieval and high-resolution promoterome mining for integrative analyses. *Nucleic Acids Res.* **43**, e51 (2015).
62. Wehrens, R. & Buydens, L. M. C. Self- and super-organising maps in R: the kohonen package. *J. Stat. Softw.* **21**, 1–19 (2007).
63. Gohr, A. & Irimia, M. Matt: Unix tools for alternative splicing analysis. *Bioinformatics* (2018).
64. Weirauch, M. T. et al. Determination and inference of eukaryotic transcription factor sequence specificity. *Cell* **158**, 1431–1443 (2014).
65. van Heeringen, S. J. & Veenstra, G. J. GimmeMotifs: a de novo motif prediction pipeline for ChIP-sequencing experiments. *Bioinformatics* **27**, 270–271 (2011).
66. 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. *Dev. Dyn.* **238**, 2409–2417 (2009).
67. Gehrke, A. R. et al. Deep conservation of wrist and digit enhancers in fish. *Proc. Natl Acad. Sci. USA* **112**, 803–808 (2015).

68. Kawakami, K. Transgenesis and gene trap methods in zebrafish by using the *Tol2* transposable element. *Methods Cell Biol.* **77**, 201–222 (2004).
69. Somorjai, I., Bertrand, S., Camasses, A., Haguenaier, 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**, 703–713 (2008).
70. Tena, J. J. et al. Comparative epigenomics in distantly related teleost species identifies conserved *cis*-regulatory nodes active during the vertebrate phylotypic period. *Genome Res.* **24**, 1075–1085 (2014).
71. Acemel, R. D. et al. A single three-dimensional chromatin compartment in amphioxus indicates a stepwise evolution of vertebrate Hox bimodal regulation. *Nat. Genet.* **48**, 336–341 (2016).



Extended Data Fig. 1 | Summary of genomic assembly and repeat annotation. **a**, Spectrum of 25-mers in Illumina sequencing data that shows the bimodal distribution that is characteristic of highly polymorphic species. **b**, Heat map showing k -mer decomposition (y axis) across GC content (x axis). Both peaks show comparable GC content, which is consistent with them representing haploid versus diploid k -mers. **c**, Flow chart of the steps followed to obtain the *B. lanceolatum* assembly. **d**, Repeat landscape and its evolutionary history, shown by the proportion of repetitive elements with a given divergence (K2P) to their consensus in the repeat library (repeatScout). **e**, Percentage of methylated CpG dinucleotides within repetitive elements, at three developmental stages and in the adult hepatic diverticulum. **f**, Distribution of average levels of 5mC of different repeat families. Colour key indicates the percentage of repeats in each family with corresponding levels of average methylation. **g**, Computational pipeline to identify long non-coding RNAs (lncRNAs). Categories: antisense, lncRNA overlaps with a protein-coding gene in the reverse strand; intragenic, lncRNA overlaps with a protein-coding gene in the same strand; bidirectional, within 1 kbp of a TSS of a protein-coding gene in the antisense strand, probably a product of a

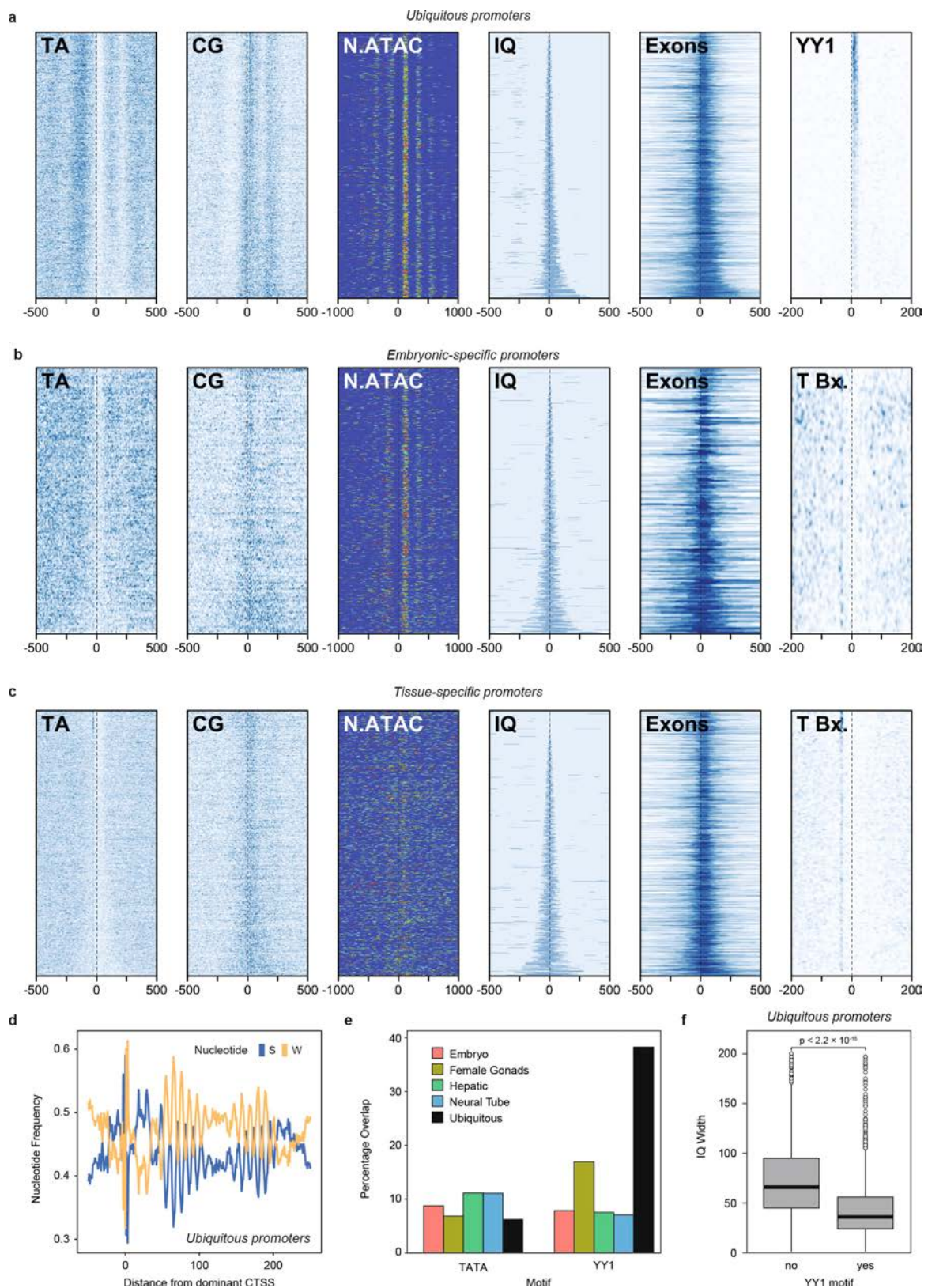
bidirectional promoter; intergenic, lncRNA does not overlap with any protein-coding gene. The total number in each category is indicated, with the number of those that are multi-exonic in parentheses. **h**, Quadruple conserved syntenic scaffold between amphioxus and human. Top, amphioxus scaffold Sc0000001 aligned against the four human chromosomes with which it shares the highest number of orthologues (chr1, chr5, chr9 and chr19). In this scaffold, 277 out of 551 genes have clear orthologues in human, and 203 of these have orthologues on at least one of the four mentioned chromosomes. The black horizontal line represents the amphioxus scaffold Sc0000001 aligned against the four human chromosomes with which it shares the highest number of orthologues (chr1, chr5, chr9 and chr19). In this scaffold, 277 out of 551 genes have clear orthologues in human, and 203 of these have orthologues on at least one of the four mentioned chromosomes. The black horizontal line represents the amphioxus scaffold Sc0000001: 7,736,434–8,850,041. On the top line, each amphioxus gene with at least one orthologue in the nine reference species is represented with an oriented coloured box. Human genes located in the four orthologous chromosomes are aligned underneath, in boxes of colours that correspond to those of their amphioxus pro-orthologues. The Genomicus server dedicated to amphioxus can be accessed at <http://genomicus.biologie.ens.fr/genomicus-amphioxus>.



Extended Data Fig. 2 | See next page for caption.

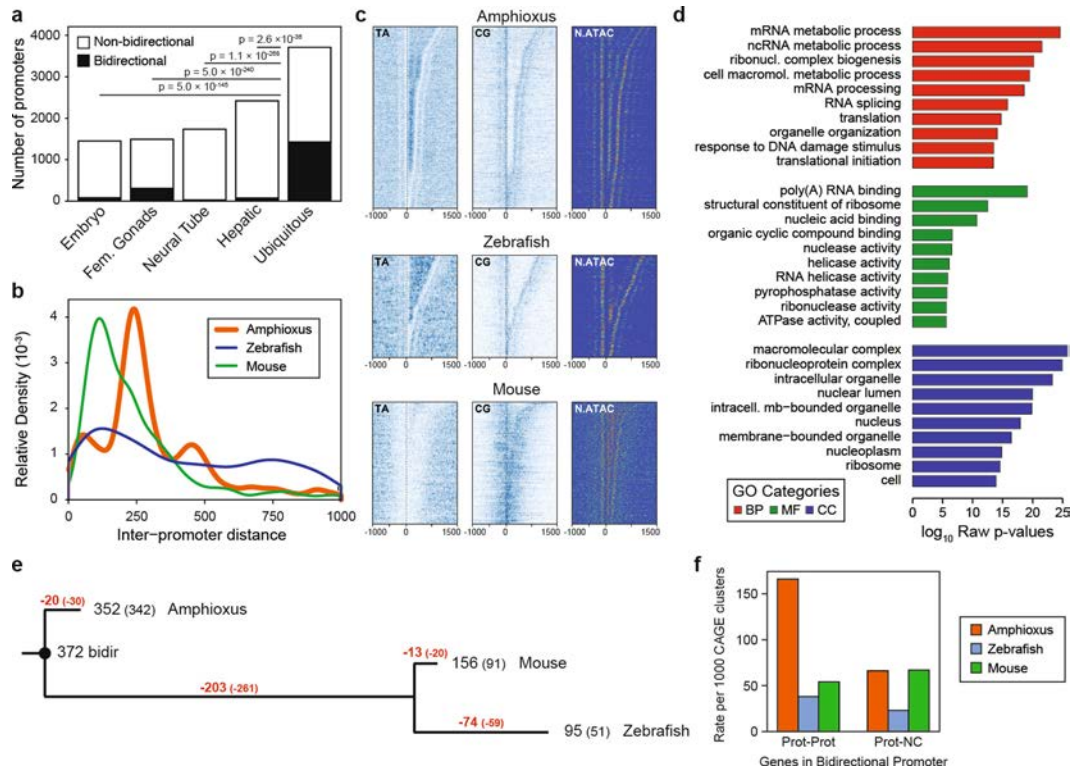
Extended Data Fig. 2 | Dynamics of chromatin marks on APREs and reporter assays. **a**, Summary of the zebrafish and medaka RNA-seq and ATAC-seq datasets generated for this study. Dashed lines indicate equivalent developmental stages in the two species, based on a previous study⁷⁰. The number of biological replicates is indicated for each experiment. Zebrafish 24-hpf ATAC-seq data are from a previous study⁶⁷. **b**, Cumulative distribution of the distance between CAGE-seq peaks and the closest annotated TSSs for genes with expression cRPKM > 5 in any of the samples covered by CAGE-seq (see Fig. 1a). Only CAGE-seq peaks within 1 kbp of an annotated TSS were tested (amphioxus: 10,435 peaks; zebrafish, 23,326 peaks; and mouse, 23,443 peaks). **c**, Cumulative distribution of distances between each APRE and the closest annotated TSS normalized by the average intergenic distance of the species (amphioxus, 83,471; zebrafish, 252,774; medaka, 174,139; and mouse, 216,857 APREs, as per Fig. 1c). **d**, Signal distribution of different marks within functional-genomic regions in amphioxus. \log_{10} of read counts of H3K4me3, H3K27ac and ATAC-seq, and raw read counts of CAGE-seq in promoters of homology-supported, protein-coding genes ($n = 26,501$), other APREs ('O. APREs', all APREs that do not overlap a TSS from any gene model; $n = 48,341$), proximal APREs ($n = 24,622$), distal APREs ($n = 11,881$), previously validated enhancers ($n = 43$; Supplementary Table 9), random regions ($n = 88,413$) and negative regions (excluding ATAC-seq peaks, $n = 88,413$). For region designation, see Fig. 1c. For clarity, whiskers and outliers are not displayed. **e**, *k*-means clustering of APREs based on H3K27ac signal in three developmental stages. Cluster 1 and 3 APREs were considered as active and inactive, respectively. Average H3K27ac profiles are represented in the top panels. The number of APREs per cluster and stage are provided in Supplementary Data 2, dataset 8. **f**, Alluvial plot that shows the dynamics of each APRE among

the clusters described in **e**. APREs that remained active (cluster 1 in all stages) along the three developmental stages are represented in blue, constitutively inactive APREs (cluster 3 in all stages) in dark grey and dynamic APREs in red or orange (if inactivated or activated, respectively, during development). Five groups of APREs of special interest are highlighted with stronger colours and named GR1–GR5. **g**, Representative enriched DNA motifs found in each of the groups described in **f**. GR1 APREs were enriched in early motifs (for example, Smad3 and Oct4, Sox2 and Nanog); GR3 APREs in motifs of transcription factors involved in the generation of the three germ layers (for example, Foxo3, Sox6 and Sox17); GR4 APREs in tissue-specific transcription factors (for example, Foxa2, Otx2 and Crx); and GR5 APREs in CTCF and CTCF-like (BORIS) motifs. *q* values as provided by Homer. **h**, Lateral views of embryos from stable transgenic zebrafish lines at 24 hpf (except for Foxa-243, at 48 hpf) showing GFP expression driven by the amphioxus APREs listed in Supplementary Table 8 and highlighted in Supplementary Fig. 1. The number of independent founders with the same expression were as follows: Six1/2-182 (5/5), Foxa-243 (3/3), Foxa-251 (4/4), FoxC-3067 (6/6) and Pax1/9-157 (3/3). Midbrain expression corresponds to the positive-control enhancer included in the reporter constructs. EN, endoderm; HB, hindbrain; MY, myotomes; PA, pharyngeal arch; SC, spinal cord. Scale bar, 250 μ m. **i**, Lateral views of transient transgenic amphioxus embryos, showing GFP expression driven by the APREs highlighted in Supplementary Fig. 1a, b (Foxa-251 ($n = 46$ out of 52) and Foxc-3067 ($n = 27$ out of 35), respectively) and in a previous study⁷¹ (Hox-1655, $n = 72$ out of 80). For each element, left panels correspond to 3D rendering from sub-stacks and right panels to *z*-stack sagittal sections. Scale bar, 50 μ m. Anterior is to the left and dorsal to the top.



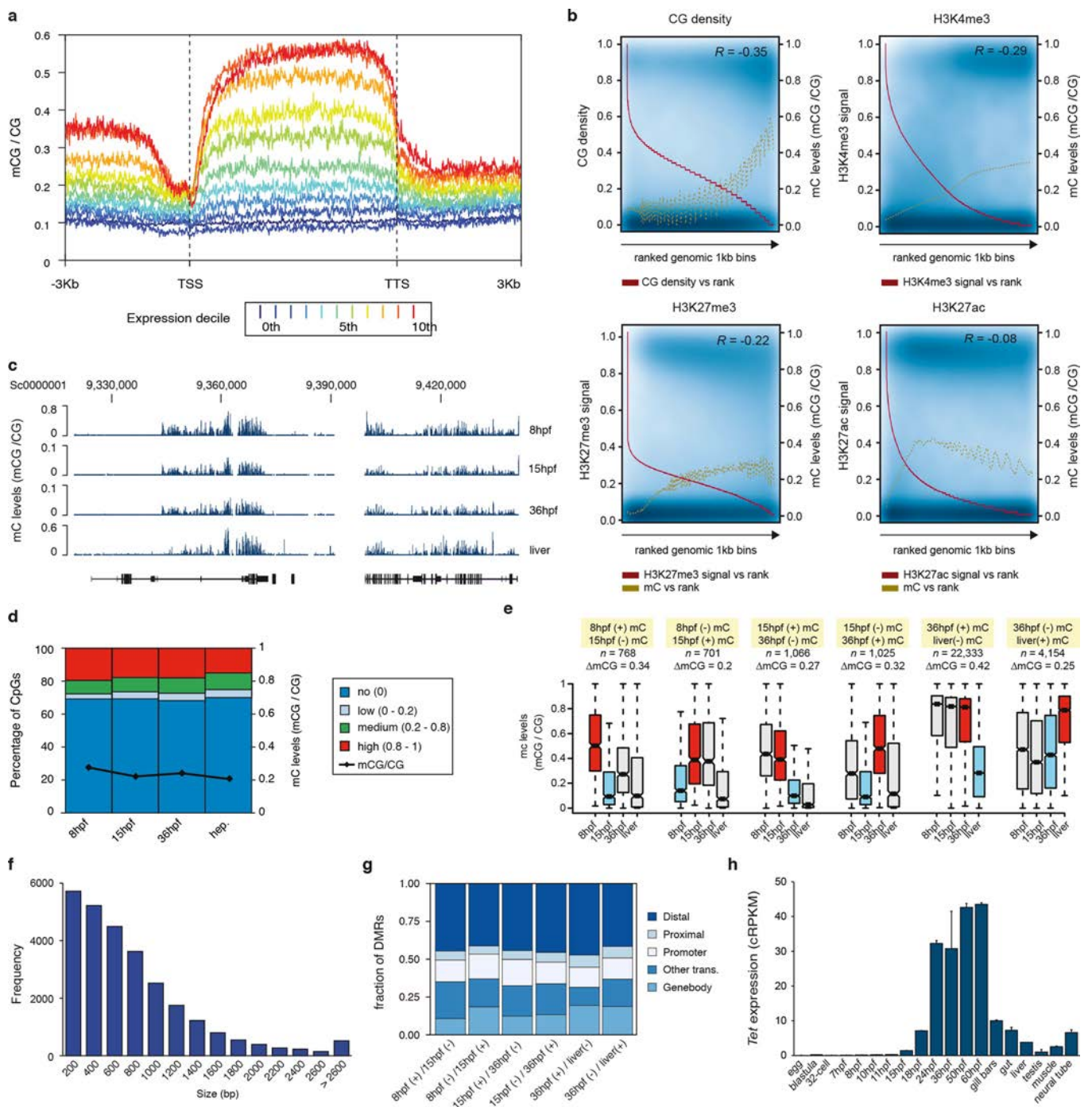
Extended Data Fig. 3 | Features of amphioxus promoters derived from CAGE-seq. **a–c.** Heat maps showing AT and CG signal, nucleosome positioning (derived from the NucleoATAC signal), promoter width (interquartile (IQ) range), first exon length and YY1 (**a**) or TATA box (**b**, **c**) motifs around ubiquitous (**a**, $n = 3,710$), embryonic-specific (**b**, $n = 1,451$) and tissue-specific (**c**, $n = 4,154$) promoters, sorted by promoter width. Position 0 corresponds to the main TSS. **d.** Ubiquitous promoters show strong evidence for a nucleosome positioned downstream

of the CAGE TSS, as judged from the 12-bp periodicity of W and S nucleotide density. **e.** Per cent of promoters of each category that have associated TATA box or YY1 motifs. Number of promoters: embryo, 1,451; female gonads, 1,494; hepatic, 2,420; neural tube, 1,734; and ubiquitous, 3,710. **f.** IQ width distribution of ubiquitous promoters ($n = 3,710$) with and without an associated YY1 motif. P value corresponds to two-sided Wilcoxon sum-rank tests.



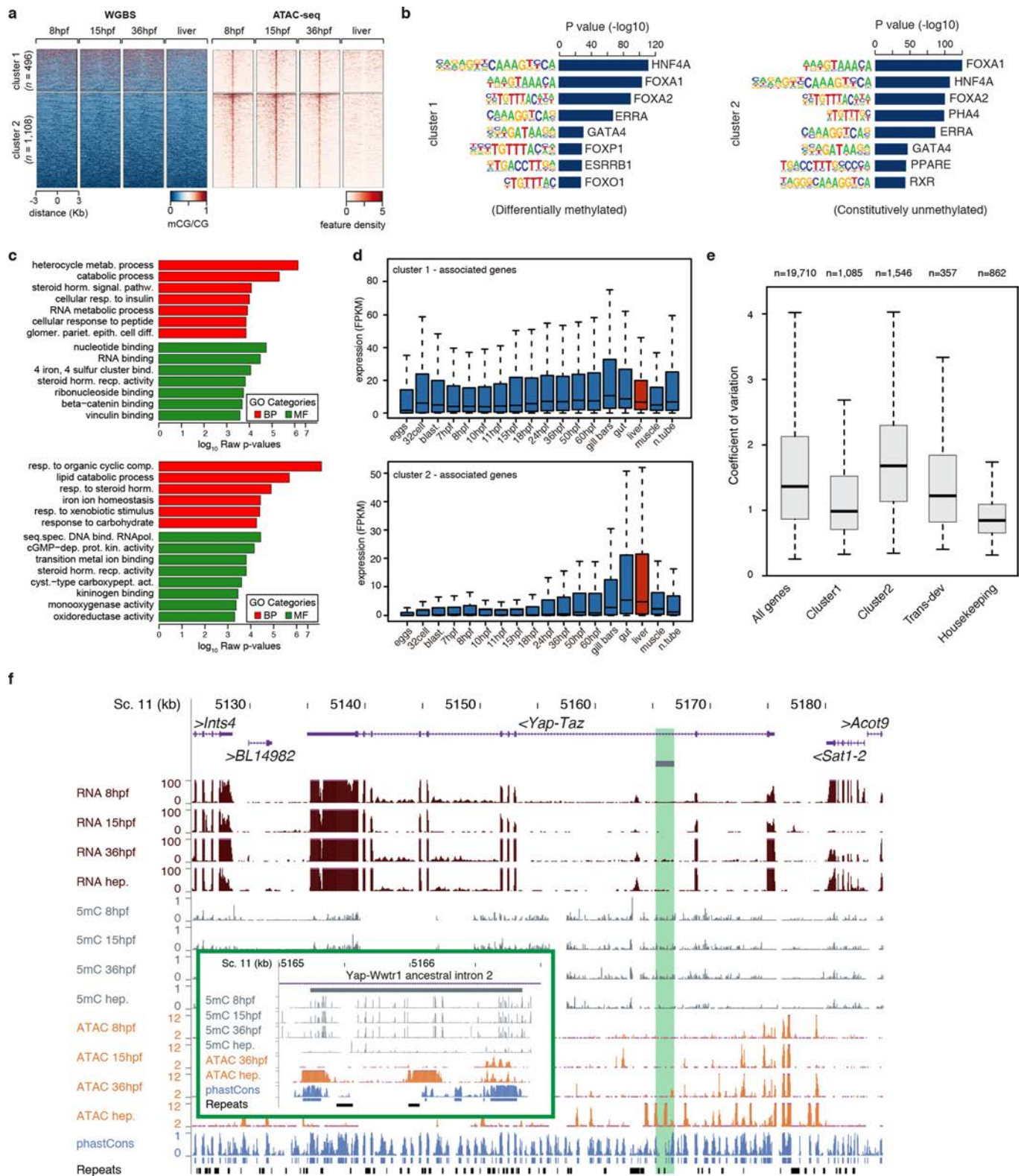
Extended Data Fig. 4 | Characteristics and evolution of bidirectional promoters. **a**, Number of bidirectional and non-bidirectional promoters identified for each regulatory category. *P* values correspond to two-sided Fisher's exact tests against ubiquitous promoters. **b**, Distribution of distance between bidirectional promoters in each species (amphioxus, 1,975; zebrafish, 549; and mouse, 876 pairs of promoters). The distance between amphioxus peaks closely corresponds to integral nucleosome spacing. **c**, Heat maps of TA, CG and nucleosome occupancy (derived from the NucleoATAC signal) around bidirectional promoter pairs in amphioxus ($n = 1,975$), mouse ($n = 876$) and zebrafish ($n = 549$), arranged by the distance between the two CAGE TSSs. In amphioxus, both TA and NucleoATAC signals indicate regions in which 0, 1 or 2 nucleosomes separate promoters. **d**, Enriched GO terms for genes associated with bidirectional promoters in amphioxus. Uncorrected *P* values correspond to two-sided Fisher's exact tests as provided by topGO.

e, Inferred evolutionary dynamics of 372 putatively ancestral bidirectional promoters among chordate groups. Red, number of inferred losses and disentanglements; black, number of detected bidirectional promoters by CAGE-seq (in brackets) or microsynteny (neighbouring genes in a 5' to 5' orientation) for each species. In parentheses, number of lost and disentangled (red) or retained (black) bidirectional promoters when considering only the cases supported by CAGE-seq. **f**, In vertebrates, disentanglement was not accompanied by a general increase in the fraction of bidirectional promoters with antisense non-coding transcription, as shown by the relative number of CAGE clusters identified as bidirectional promoters that are composed of two protein-coding genes ('Prot-Prot') or of one protein-coding and one non-coding or non-annotated locus ('Prot-NC'). The total number of uniquely annotated, protein-coding-associated CAGE promoters was amphioxus, 11,789; mouse, 13,654; and zebrafish, 14,014.



Extended Data Fig. 5 | 5mC dynamics in amphioxus. **a**, 5mC levels across gene bodies ($n = 20,569$) from different expression deciles (0th, not expressed; 10th, highest expression). TTS, transcription termination site. **b**, Scatter plots of levels of 5mC and CpG density, H3K4me3, H3K27me3 and H3K27ac in 1-kbp genomic bins sorted on the basis of feature rank. The red line tracks anti-correlation between feature density and rank number (a low rank number implies high feature density). The golden line represents a smoothing spline of 5mC signal versus feature rank number. Pearson correlation coefficients (R) are displayed in the top right corner of each panel. **c**, UCSC browser excerpt of 5mC patterns for selected regions. **d**, Percentage of methylated CpG dinucleotides in 8-hpf ($n = 19,657,388$), 15-hpf ($n = 21,247,615$), 36-hpf ($n = 21,702,000$) and hepatic (adult, $n = 19,240,245$) amphioxus samples. Black line indicates the fraction between methylated and non-methylated CpGs at each stage. **e**, Box

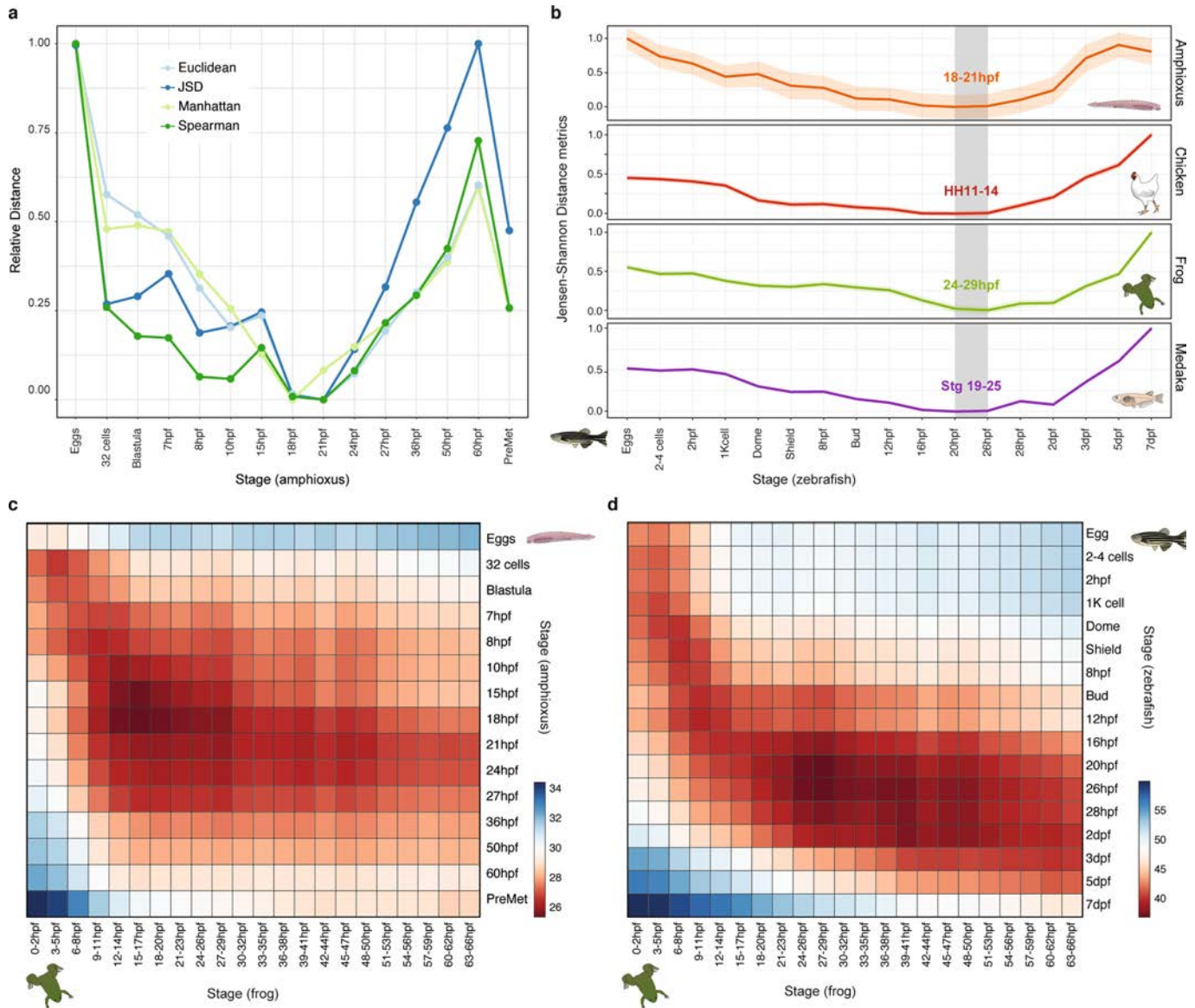
plots of average 5mC levels in different types of differentially methylated regions (DMRs) at each stage. ΔmCG denotes the change in the fraction of methylated CpGs between the two stages used for identification of DMRs (red (hyper) and blue (hypo) boxes). The number of DMRs were as follows: 8 hpf(+)-15 hpf(-), 768; 8 hpf(-)-15 hpf(+), 701; 15 hpf(+)-36 hpf(-), 1,066; 15 hpf(-)-36 hpf(+), 1,025; 36 hpf(+)-liver(-), 22,333; and 36 hpf(-)-liver(+), 4,154. The coordinates for all DMRs are provided in Supplementary Data 2, dataset 11. **f**, Distribution of DMR sizes (in bp). **g**, Genomic distribution of DMRs identified for each sample. 'Other trans,' DMRs that overlap with gene models that were not defined as being supported by orthology. **h**, Expression (cRPKM) of the amphioxus *Tet* orthologue in embryos and adult tissues. Error bars represent standard error of the mean (the number of replicates for each RNA-seq dataset is provided in Fig. 1a).



Extended Data Fig. 6 | See next page for caption.

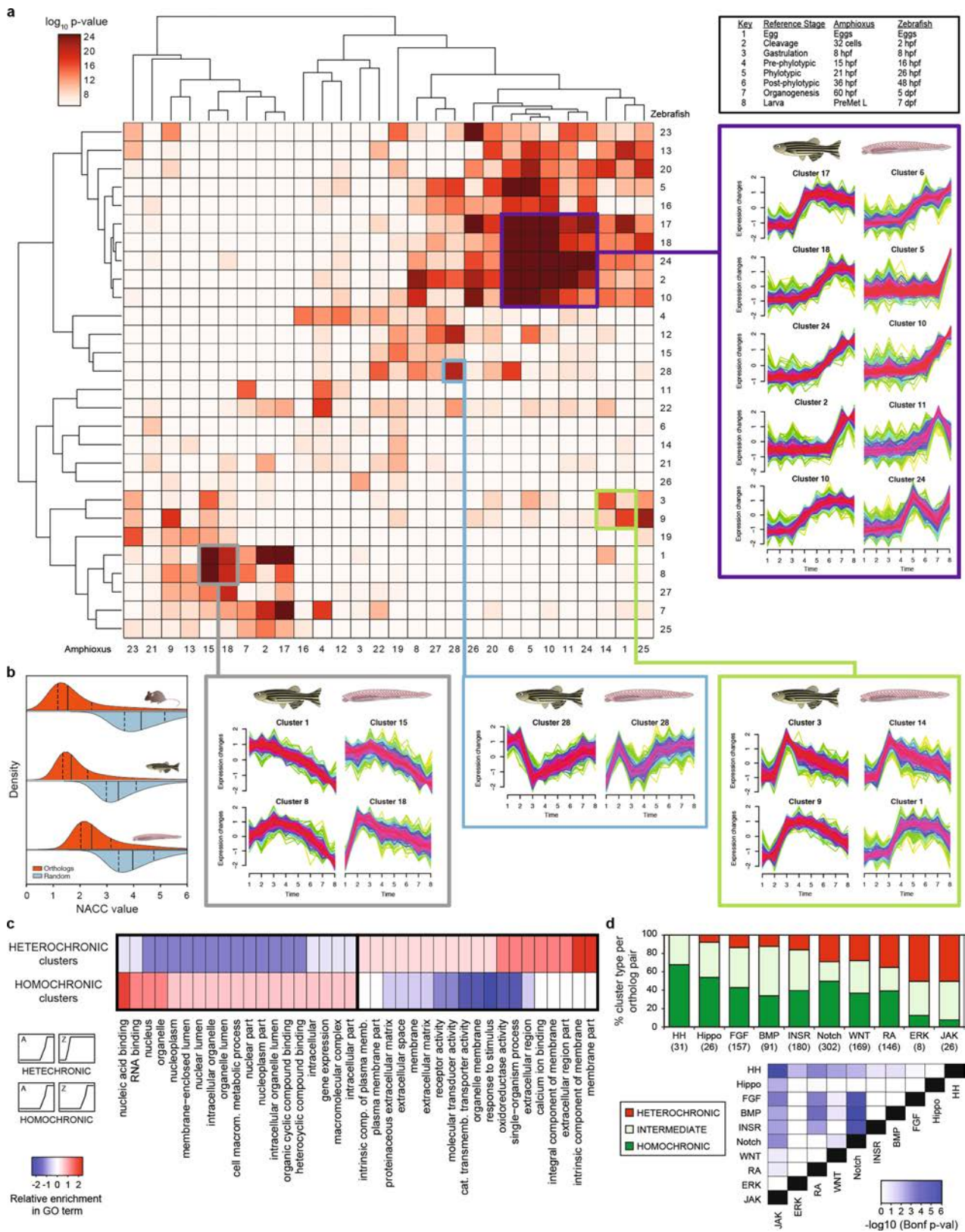
Extended Data Fig. 6 | Developmental 5mC dynamics at APREs in amphioxus. **a**, *k*-means clustering ($n = 2$) of 5mC signal over embryo-specific open-chromatin regions (that is, APREs), assessed by ATAC-seq (Supplementary Table 10). **b**, The most significantly enriched transcription-factor binding-site motifs in APREs that display different developmental 5mC patterns in Fig. 2b. Uncorrected *P* values as provided by MEME. All plotted motifs had Benjamini-corrected *q* values of 0. **c**, GO enrichment for genes associated with cluster 1 (top) or cluster 2 (bottom) APREs from Fig. 2b. Uncorrected *P* values correspond to two-sided Fisher's exact tests as calculated by topGO. **d**, Distribution of expression values (cRPKM) across all samples for genes associated with cluster 1 (top, $n = 1,114$) or cluster 2 (bottom, $n = 1,594$) APREs from

Fig. 2b. **e**, Distribution of the coefficients of variation for genes associated with cluster 1 or cluster 2 APREs from Fig. 2b, as well as all ($n = 19,710$), trans-dev ($n = 357$) and house-keeping ($n = 862$) amphioxus genes. **f**, Example of a potentially conserved (zebrafish to amphioxus) DMR associated with *yap1*, a major transcription factor of the Hippo pathway. The inset corresponds to the region highlighted in green. The two orthologous genomic regions in zebrafish are shown in Supplementary Fig. 2. Additional cases included genes that contained APREs that are likely to regulate neighbouring liver-specific genes ('bystander' genes) (Supplementary Table 11). The number of replicates for each experiment displayed in each track is provided in Fig. 1a.



Extended Data Fig. 7 | Periods of maximal transcriptomic similarity across chordate development. **a**, Stages of minimal transcriptomic distance obtained in the comparison between amphioxus and zebrafish for four alternative distance methods (Euclidean, Manhattan and Jensen–Shannon distances, and Spearman correlation). Values are normalized to minimal (0) and maximal (1) for each metric. **b**, Stages of minimal transcriptomic divergence shown as the smallest Jensen–Shannon distance between zebrafish stages and four chordate species. The shaded area surrounding the line that connects the stages is the standard deviation, derived from 100 bootstrap replicates of the orthologous gene set.

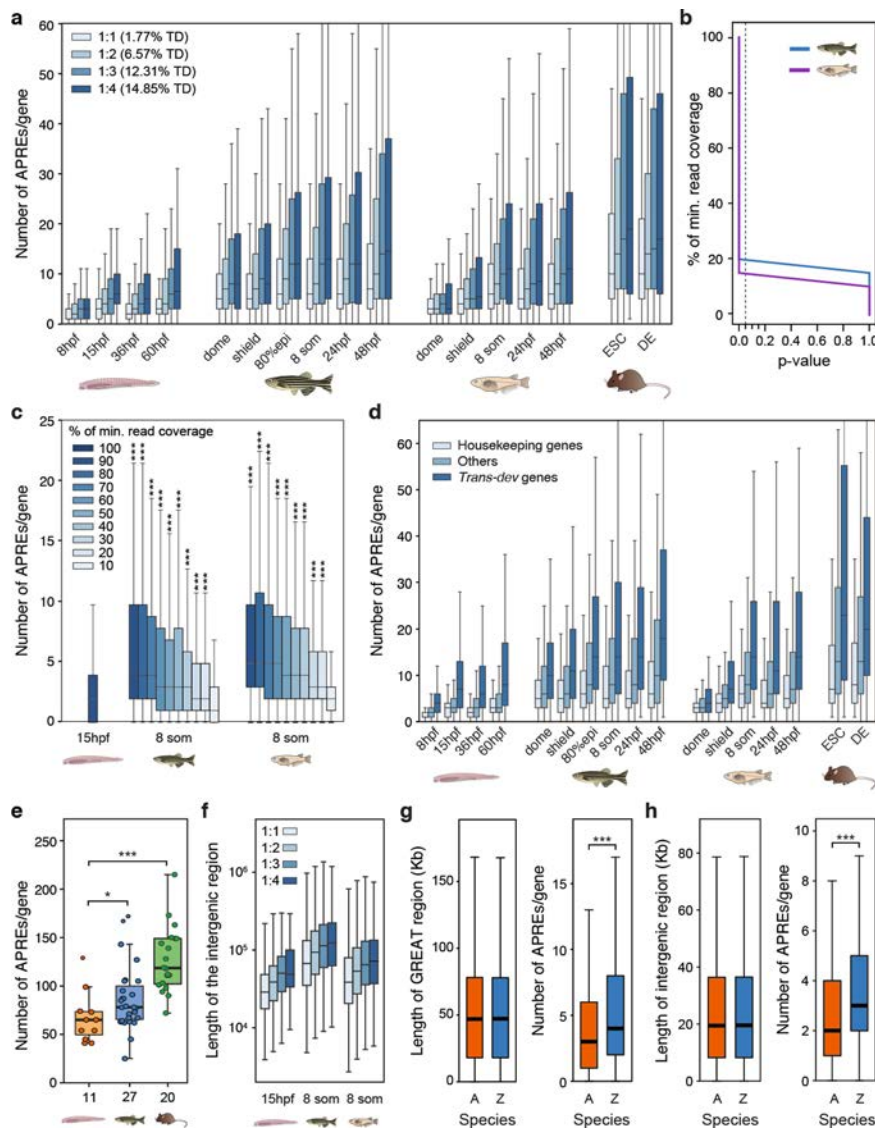
The grey box outlines the ‘phylogenic’ period of minimal divergence; the corresponding periods are indicated for each species as the range provided by the two closest stages. **c**, **d**, Heat maps of pairwise transcriptomic distances (Jensen–Shannon distance metric) between pairs of chordate species, amphioxus and frog (**c**), and zebrafish and frog (**d**). In both heat maps, the smallest distance (red) indicates maximal similarity of the transcriptome. The periods of minimal divergence of the transcriptome are earlier for the amphioxus–frog comparison than for the zebrafish–frog comparison.



Extended Data Fig. 8 | See next page for caption.

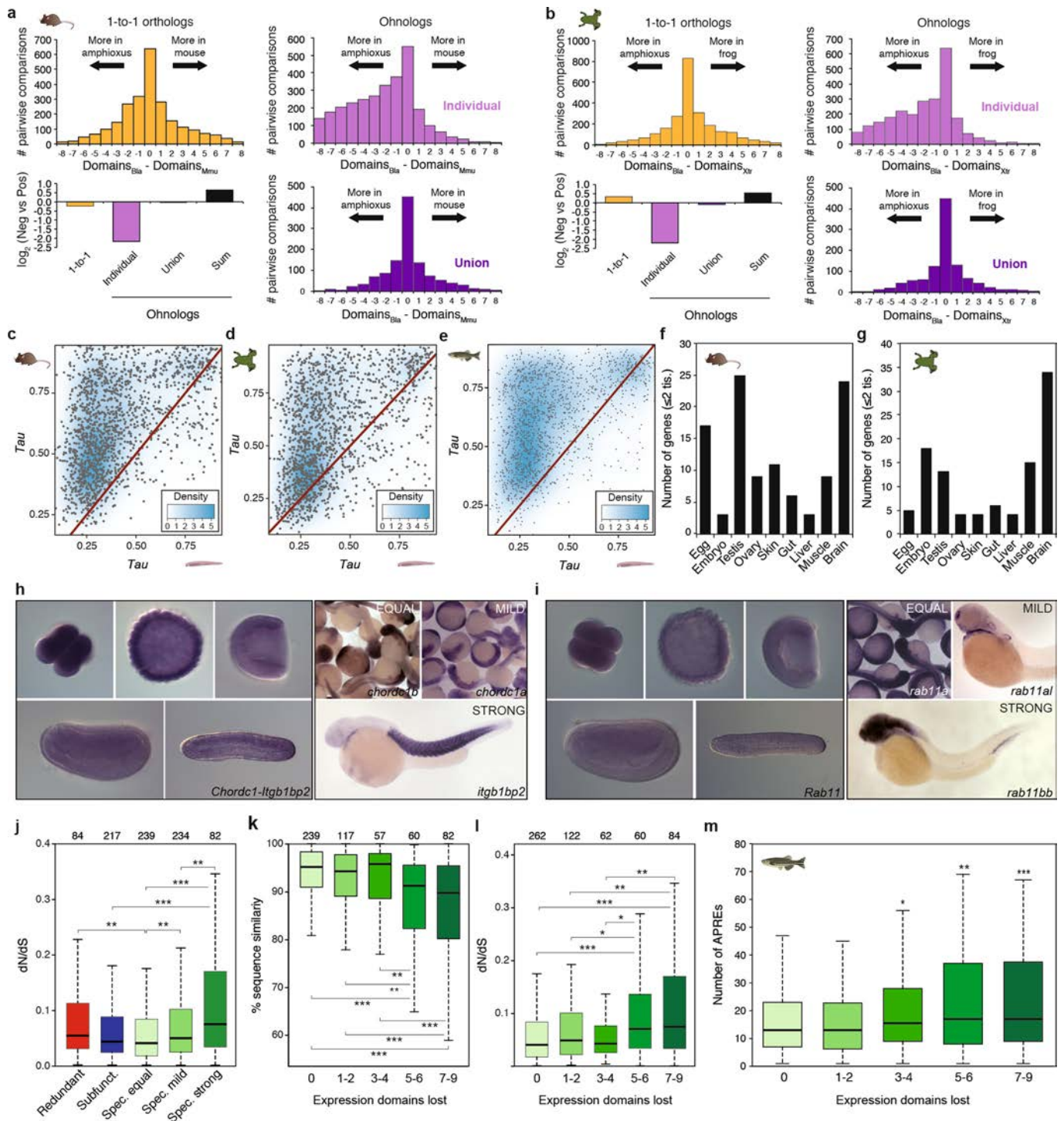
Extended Data Fig. 8 | Comparison of temporal gene expression profiles in amphioxus and zebrafish. **a**, Heat map showing the significance of orthologous gene overlap between Mfuzz clusters across eight matched developmental stages in amphioxus and zebrafish as derived from an upper-tail hypergeometric test. Some clusters with highly significant overlap are highlighted, and their corresponding temporal expression profiles are shown. The profiles of all clusters for the two species are included in Supplementary Figs. 3, 4. Exact P values and sample sizes are provided in Supplementary Data 2, dataset 8. **b**, Distributions of NACC values for orthologous genes (in red) or random orthology assignments (blue) for each species against human. Lower NACC values imply higher conservation of relative expression. Solid lines show the median, and the dashed lines mark the interquartile range. The number of orthologue pairs were as follows: mouse, 15,109; zebrafish, 16,480;

and amphioxus, 8,633. **c**, Differentially enriched GO terms among pairs of zebrafish and amphioxus Mfuzz clusters with significant orthologue overlap ($P < 10^{-10}$ upper-tail hypergeometric test) with homochronic (48 pairs) and heterochronic (35 pairs) patterns. The GO enrichment of a group was calculated as the number of cluster pairs with significant enrichment for that given term (Supplementary Data 2, dataset 12). **d**, Top, per cent of zebrafish genes from each developmental pathway we studied, based on the temporal similarity of their corresponding Mfuzz cluster (homochronic, heterochronic or intermediate). Only genes belonging to clusters with significant orthologue overlap were analysed; the number of genes is provided in parenthesis below the pathway name. Bottom, pairwise comparisons between developmental pathway distributions. P values correspond to Bonferroni-corrected, two-sided, three-way Fisher's exact tests.



Extended Data Fig. 9 | Higher regulatory content in vertebrate genomes. **a**, Distribution of the number of APREs per the regulatory landscape of a gene (as determined by GREAT²⁶), at different developmental stages or cell lines of four chordate species (amphioxus, zebrafish, medaka and mouse). Orthologous gene families are split according to the number of orthologues that are retained per family (from 1 to 4, using mouse as a reference species for the orthologue counts). The percentage of developmental regulatory genes (trans-dev, TD) in each category is indicated. **b**, *P* values of one-sided Mann–Whitney *U* tests against the amphioxus peak-number distribution using 100% of the minimum read coverage for different levels of down-sampling of the zebrafish and medaka samples. **c**, Distribution of the number of APREs in the GREAT region of the gene, called after down-sampling the reads of the two vertebrate samples to different fractions of the sample with the minimum effective coverage in our study (~21 reads per kbp for the 36-hpf sample in amphioxus). Asterisks correspond to the significance of the *P* values of Mann–Whitney *U* tests against the amphioxus peak-number distribution using 100% of the minimum-read coverage. The number of genes per box was as follows: amphioxus, 20,569; zebrafish,

20,053; and medaka, 15,978. **d**, As in **a**, but with gene families separated according to functional categories (housekeeping, trans-dev and others). **e**, Number of APREs per regulatory landscape determined using 4C-seq, for 58 members of 11 trans-dev families. The number of genes probed in each species is indicated on the *x* axis. **f**, Distribution of the length of the intergenic regions from the genes plotted in **a** for the indicated stages. **g**, Distributions of GREAT-region sizes (left) and number of APREs per gene (right) for a subset of 10,186 pairs of genes with matched GREAT-region size distributions (± 500 bp) in amphioxus and zebrafish. **h**, Distributions of intergenic-region sizes (left) and number of APREs per gene (right) for a subset of 13,941 pairs of genes with matched intergenic-region size distributions (± 500 bp) in amphioxus and zebrafish. *P* values correspond to Mann–Whitney *U* tests: * $0.05 > P$ value ≥ 0.01 , ** $0.01 > P$ value ≥ 0.001 , ****P* value < 0.001 . In **a** and **d**, all comparisons between each distribution of a vertebrate species and the equivalent distribution in amphioxus produced significant *P* values (*P* value < 0.001); for simplicity, in these panels asterisks are not shown. Exact *P* values and sample sizes are provided in Supplementary Data 2, dataset 8.



Extended Data Fig. 10 | See next page for caption.

Extended Data Fig. 10 | Regulatory evolution after vertebrate WGD.

a, b, For each mouse (**a**) or frog (**b**) gene, the number of positive-expression domains across nine equivalent samples is subtracted from the number of domains in which the single amphioxus orthologue is expressed. The distribution of the difference in domains between the amphioxus and the vertebrate species is plotted for 1-to-1 orthologues (2,450 and 2,484 gene pairs for mouse and frog, respectively; yellow), individual ohnologues (3,011 and 2,637 gene pairs in 1,212 and 1,094 families for mouse and frog, respectively; lilac) and the union of all vertebrate ohnologues in a family (purple). Bottom left, \log_2 of the ratio between the sum of all mouse (**a**) or frog (**b**) genes with negative versus positive score for each orthology group. 'Sum' (black), binarization of family expression is performed after summing the raw expression values for all ohnologues. **c–e**, Density scattered plot of the τ values for pairs of mouse (**c**, $n = 1,502$), frog (**d**, $n = 1,495$) and zebrafish (**e**, $n = 1,498$) and amphioxus orthologues from multi-gene families in vertebrates. **f, g**, Number of ohnologues with strong specialization (≤ 2 remaining expression domains) in mouse (**f**) or frog (**g**) expressed in each tissue or

developmental stage. **h, i**, Representative in situ hybridization assays in zebrafish embryos for different members of specialized families (right) and for the single amphioxus orthologue (left) (Chordc1 and Itgb1bp2 (**h**) and Rab11 (**i**)). Zebrafish image data for this paper were retrieved from the Zebrafish Information Network (ZFIN), University of Oregon, Eugene, OR 97403-5274; (<http://zfin.org/>, accessed May 2018) and are used with the permission of B. Thisse. Amphioxus in situ hybridization was performed once using 10 embryos per probe, all of which showed the same expression pattern. **j**, Distribution of the dN/dS ratio between human and mouse for different classes of ohnologues based on their fate after WGD. **k, l**, Distribution of the percentage of nucleotide sequence similarity (**k**) or dN/dS ratio (**l**) between human and mouse for ohnologues grouped by the number of expression domains lost. **m**, Distribution of the number of APREs within GREAT regions for zebrafish ohnologues grouped by the number of expression domains lost. *P* values in **j–m** correspond to Wilcoxon sum-rank tests. * $0.5 > P$ value ≥ 0.01 ; ** $0.01 > P$ value ≥ 0.001 ; *** P value < 0.001 .

Results

Article RII:

Abstract

Homologous long non-coding RNAs (lncRNAs) are elusive to identify by sequence similarity due to their fast-evolutionary rate. Here we develop LincOFinder, a pipeline that finds conserved intergenic lncRNAs (lincRNAs) between distant related species by means of microsynteny analyses. Using this tool, we have identified 16 bona fide homologous lincRNAs between the amphioxus and human genomes. We characterized and compared in amphioxus and *Xenopus* the expression domain of one of them, *Hotairm1*, located in the anterior part of the Hox cluster. In addition, we analyzed the function of this lincRNA in *Xenopus*, showing that its disruption produces a severe headless phenotype, most probably by interfering with the regulation of the Hox cluster. Our results strongly suggest that this lincRNA has probably been regulating the Hox cluster since the early origin of chordates. Our work pioneers the use of syntenic searches to identify non-coding genes over long evolutionary distances and helps to further understand lncRNA evolution.

Article RII:

Microsyntenic Clusters Reveal Conservation of lncRNAs in Chordates Despite Absence of Sequence Conservation

Carlos Herrera-Úbeda, Marta Marín-Barba, Enrique Navas-Pérez, Jan Gravemeyer, Beatriz Albuixech-Crespo, Grant N. Wheeler and Jordi Garcia-Fernàndez*




Herrera-Úbeda, C.; Marín-Barba, M.; Navas-Pérez, E.; Gravemeyer, J.; Albuixech-Crespo, B.; Wheeler, G.N.; Garcia-Fernàndez, J. Microsyntenic Clusters Reveal Conservation of lncRNAs in Chordates Despite Absence of Sequence Conservation. *Biology (Basel)*. **2019**, *8*, 61.

The PhD candidate is the first author, and most of the work was planned and performed by him. This included the design of the synteny program LincOFinder, the identification of homologous lincRNAs between amphioxus and vertebrates, and the in-depth characterization of *Hotairm1*. The latest included the analyses of morpholino-treated *Xenopus* embryos, by in situ hybridization, and qPCR analyses. To my knowledge, the article has not been included in any other PhD thesis.

Signed: Jordi Garcia Fernàndez, PhD supervisor

Article

Microsyntenic Clusters Reveal Conservation of lncRNAs in Chordates Despite Absence of Sequence Conservation

Carlos Herrera-Úbeda ¹, Marta Marín-Barba ², Enrique Navas-Pérez ¹, Jan Gravemeyer ³, Beatriz Albuixech-Crespo ¹, Grant N. Wheeler ² and Jordi Garcia-Fernández ^{1,*}

¹ Department of Genetics, Microbiology and Statistics, Faculty of Biology, University of Barcelona, 08028 Barcelona, Spain

² School of Biological Sciences, University of East Anglia, Norwich Research Park, Norwich NR4 7TU, UK

³ German Cancer Research Center, 69120 Heidelberg, Germany

* Correspondence: jordigarcia@ub.edu

Received: 23 July 2019; Accepted: 21 August 2019; Published: 24 August 2019



Abstract: Homologous long non-coding RNAs (lncRNAs) are elusive to identify by sequence similarity due to their fast-evolutionary rate. Here we develop LincOFinder, a pipeline that finds conserved intergenic lncRNAs (lincRNAs) between distant related species by means of microsynteny analyses. Using this tool, we have identified 16 bona fide homologous lincRNAs between the amphioxus and human genomes. We characterized and compared in amphioxus and *Xenopus* the expression domain of one of them, *Hotairm1*, located in the anterior part of the Hox cluster. In addition, we analyzed the function of this lincRNA in *Xenopus*, showing that its disruption produces a severe headless phenotype, most probably by interfering with the regulation of the Hox cluster. Our results strongly suggest that this lincRNA has probably been regulating the Hox cluster since the early origin of chordates. Our work pioneers the use of syntenic searches to identify non-coding genes over long evolutionary distances and helps to further understand lncRNA evolution.

Keywords: lncRNAs; genome_evolution; synteny; amphioxus

1. Introduction

Identifying and understanding the factors that underlie the evolution of morphological complexity is one of the central issues in the field of evolutionary developmental biology (or evo-devo). From the initial claims that gene duplication and neofunctionalization were at the core of phenotypic change [1], the current view also takes into account the fine-tuning of gene regulation [2] and increasing the proteome and interactome complexity through additional processes. In this regard, molecular mechanisms such as alternative splicing or RNA-editing, and the RNA world, with molecules like small miRNA or long non coding RNAs (lncRNAs), allow deeper and multifaceted levels of gene regulation [3]. lncRNA-mediated regulation stands out as a quick and efficient mechanism to modulate gene expression, as these molecules are function-ready almost immediately after or even during transcription and can be rapidly degraded by the cellular machinery [3–5]. These characteristics make lncRNAs sharp regulators of the myriad biological processes in which they are involved, such as chromatin remodeling, protein scaffolding or gene expression regulation through direct binding to genomic enhancers [6,7].

The study of lncRNAs from an evolutionary perspective has been hindered by their lack of strong primary sequence conservation [8,9], their apparent lack of secondary structure conservation [10], and their massive genomic generation and decay rate [11]. The cephalochordate *Branchiostoma lanceolatum*

represents the earliest branching chordate lineage and holds a genome that seems to have retained many of the features of the ancestral pre-duplicative vertebrate [12,13]. Searches of lncRNAs conserved between amphioxus and vertebrates based on sequence similarity have been unsuccessful [14,15], probably due to the long evolutionary distance that separates these lineages [16]. Recently, however, a strategy to identify conserved intergenic lncRNAs (lincRNAs) by means of syntenic analyses has been successfully attempted, but limited to closely related species [17,18].

Here, we have developed a pipeline called LincOFinder that finds conserved clusters of microsynteny between two distant organisms surrounding an intergenic lncRNA. Furthermore, we use this tool to study the conservation and evolution of the lincRNA repertoire in the chordate lineage, finding up to 16 lincRNAs putatively conserved between amphioxus and human. Finally, we further study the case of *Hotairm1*, assessing its developmental expression in amphioxus and *Xenopus* and showing that its inhibition during *X. tropicalis* development produces a severe headless phenotype, probably by disrupting the chromatin dynamics of the anterior Hox cluster. Overall, our work pioneers the use of syntenic searches to identify non-coding genes over long evolutionary distances and helps to further understand lincRNA evolution in the frame of the invertebrate-vertebrate transition.

2. Materials and Methods

2.1. Amphioxus and Human Coding and lincRNA Datasets

We used the intergenic and bidirectional fractions from the lncRNAs dataset provided by Marlétaz et al. [15] to obtain an amphioxus lincRNA fraction (1318 genes), and their protein-coding genes supported by orthology as the coding fraction (10,832 genes). The human coding genes were obtained from the Ensembl annotation of the Ch38.96 genome assembly [19]. Finally, the orthologous gene families described in Marlétaz et al. [15] were used to assess the amphioxus/human gene orthologies.

2.2. LincOFinder (lincRNA Orthology Finder)

LincOFinder (<https://github.com/cherrera1990/LincOFinder>) is a program designed to identify shared microsyntenic clusters surrounding lincRNAs between two species. These are named the “reference” (*Ref*) and the “interrogated” species (*Int*). In the first, nonautomated step, the genes of both species (only the coding genes for *Int*) need to be arranged according to their position within the corresponding chromosome or scaffold, then a virtual coordinate according to their position is established (e.g., the first gene in chromosome A will be chrA-1, the second chrA-2, etc.). Furthermore, the orthologies between the genes of both species are established, using sets of known orthologous families or with the help of programs like Orthofinder [20]. Once the data is properly formatted as indicated in the ReadMe.md of LincOFinder, each annotated lincRNA from *Ref* is used as a reference point. The three upstream and three downstream genes neighboring the lincRNA are selected, and the orthology coordinates from *Int* are parsed into a distance matrix (Figure 1). The reasoning behind selecting only the three upstream and three downstream genes is to try to be astringent enough to comply with the definition of microsynteny [21] but at the same time allowing insertions and deletions up to a certain degree and the discovering, in case they exist, of larger clusters. Only genes present in the same chromosome are taken into account for distance assessment, and comparisons between paralogs of the same *Ref* gene are avoided. Then, a UPGMA hierarchical clustering algorithm [22] is used to create viable distance clusters (the ones that comply with the previously stated restrictions), and the cluster with the minimum distance between two neighboring genes is selected, thus identifying possible microsynteny clusters. If several possible clusters are formed, then they are displayed separately. These microsynteny clusters should be further filtered by selecting only those that harbor adjacent genes. Finally, candidates are manually curated by looking for the presence of lincRNAs in the microsyntenic region of *Int* (the algorithm is blind to *Int* ncRNAs due to the possibility of missing unannotated syntenic lincRNAs that could be, for example, present in the form of ESTs). This step can be done using a genome browser such as UCSC [23] (Figure 1).

possibility of missing unannotated syntenic lincRNAs that could be, for example, present in the form of ESTs). This step can be done using a genome browser such as UCSC [23] (Figure 1).

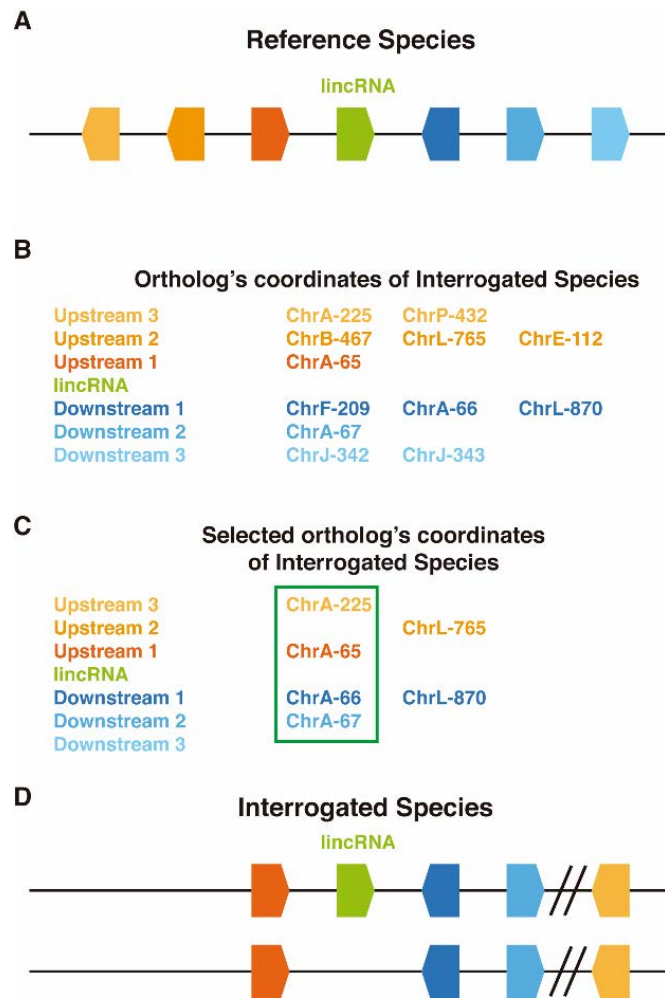


Figure 1. Diagram of LincFinder mechanism. (A) Representation of the Reference species region where a lincRNA is present. (B) Formatted table of orthologs and virtual coordinates from the three upstream and downstream regions generated to the algorithm. (C) Selection of the best clusters according to the minimum distance between regions. (D) Representation of a conserved orthologous cluster in the *Int* species, where the presence of a lincRNA is manually confirmed (a) or discarded (b) (w).

2.3.2.3. *Xenopus* Embryos and MO Injections

All experiments were performed in compliance with the relevant laws and institutional guidelines of the University of the East Anglia. The research has been approved by the local ethical review committee according to UK Home Office under Project License PPL 70/8876. *X. tropicalis* females were primed 24 h before the eggs were required and induced 5 h prior the experiment both with Chorulon (human Chorionic Gonadotropin). *X. tropicalis* males were also primed with Chorulon. Eggs were naturally obtained and fertilized with a sperm solution in Leibovitz's L-15 Medium supplemented with 10% calf serum and left at room temperature for 5 min. After that, embryos were immersed in 0.05× MMR (Mare's Modified Ringer's) (50 mM NaCl, 1 mM KCl, 0.8 mM MgCl₂, 1 mM CaCl₂, 2.5 mM HEPES, pH 7.5) for 20 min at room temperature, then washed in 20× cysteine, pH = 8 for 7 min and rinsed several times with 0.05× MMR. Embryos were incubated in a BSA (bovine serum albumin) coated Petri dish in a 0.05× MMR at 26 °C.

Morpholino oligos (MO) were designed and provided by Gene Tools. Morpholino sequences are: Standard control CCTCTTACCTCAGTTACAATTTATA, Hota1 AATGACTATTTGCTTCCTTACCGCT, Microinjection control CCTCTTACCTCAGTTACAATTTATA, Hota1 AATGACTATTTGCTTCCTTACCGCT. Microinjections were carried out in 3% Ficoll PM400 (Sigma St. Louis, MO, USA) at 2-cell stage in both cells, and then embryos were incubated at 26 °C. Once *Xenopus* embryos reached the

appropriate stage, they were snap frozen in liquid nitrogen for RNA extraction or fixed for whole mount in situ hybridization using MEMFA (3.7% formaldehyde, 1× MEM salts), then washed in PBST (PBS, 0.1% Tween), dehydrated in a serial dilution of ethanol and kept at 4 °C or −20 °C.

2.4. *Xenopus* RNA Extraction, cDNA Synthesis, PCR and Quantitative PCR

RNA was extracted using High Pure RNA isolation kit (Roche Basel, Switzerland) and 1 µg of RNA was taken to synthesize cDNA using Maxima First Strand cDNA synthesis kit (ThermoFisher Waltham, MA, USA). RT-PCR was performed using SYBR Green detection method. Primers were designed using Primer3plus; *gapdh* was used as a control housekeeper gene. Primer sequences are indicated in 5′–3′ direction.

hoxa1—F: AGAAGTTTGCCGGTCTCCTT, R: AAGCCATATCCCCAGCTTT

hoxa4—F: CAGTATCCACCCCGAAAAGA, R: GGGTTCCTCCACTGTAAT

hoxa5—F: GTCAGTGCAACCCCAAATCT, R: TTCCTTCTGGCCCTCCTAT

hoxa6—F: GGAAGTACAGCAGCCCTGTC, R: GTAGGTCTGCCTCCCTCTCC

hoxa7—F: GACTCCCATTTCGCATCTA, R: GGTAACGGGTGTAGGTCTGG

gapdh—F: ACTACCGTCCATGCCTTCAC, R: TCAGGGATGACTTCCCAAC

For RT-PCR, ~100 ng of cDNA was used for amplification and the total PCR product was loaded to a 10 mg/mL agarose gel.

P300—F: GATTGCTACACCACCTTCTC, R: CCATGGGAGTCTTGACAATC

hotairm1—F1: CACAGTGCAGATGTCAGTGC, F2: CTACGGAGAGATACTTGACAC, R1: ATGCA CCGTGTGATCAGTCG, R2: AAGCAATAACCGAGGCCTCT

2.5. *Xenopus* Whole-Mount In Situ Hybridization

In situ hybridization was carried out as previously described [24]. Probes were synthesized for *X. tropicalis hoxa1* and *hotairm1* using the following primers (sequences 5′–3′) *hoxa1*F: GATCGTTTTGTGGTTCGGACG, *hoxa1*R: GCAGCAATTTCTACCCTGCG, *hotairm1*F: CTACGGAGA GATACTTGACAC, *hotairm1*R: AAGCAATAACCGAGGCCTCT.

Otx2 and engrailed vectors for probe synthesis were kindly provided by Professor N. Papalopulu (Manchester).

2.6. *Amphioxus* Embryo Collection and Whole-Mount In Situ Hybridization

Ripe adult amphioxus specimens were collected in Banyuls-sur-mer, France. Spawning was induced as previously described [25] in a dry laboratory. After in vitro fertilization, embryos were cultured at 18 °C until they reach the desired stage and fixed with 4% PFA in MOPS buffer overnight at 4 °C.

The hybridization chain reaction (HCR) [26] in situ v3.0 kit by Molecular Instruments (Los Angeles, CA, USA) was used following the protocol provided by the manufacturer for zebrafish embryos, with some adjustments to the probe and hairpins concentration (2pMol and 18pMol respectively) and using nests with a 0.4 µm mesh. The sequence provided for the probe synthesis for *Hotairm1* was the following (5′–3′) AAGGAGAGACGAAAGTCACCGGGACAAACCGGAGGATGTCTCGG AGGACCCTACCACCGCTCCCGCCTGTGCTCTACAGGTCACCAGGTGGGGATAGCACAAACATG TCCTCTTAGACATCTCTACTACACGCAGCTTGCTACCTGAAAGTTATCATATCTAGAATGTATAT CTGCTTCAGTGTAAGCAACG.

3. Results and Discussion

3.1. Conserved Microsynteny Clusters

In order to identify conserved lincRNAs across chordates, we developed a pipeline called LincOFinder and used previously described *Branchiostoma lanceolatum-Homo sapiens* orthology families to detect conserved microsynteny clusters around specific amphioxus lincRNAs [15,23] (see Methods).

Here we present the most reliable set of homologous lincRNAs that we were able to produce. Although in our pipeline three upstream and three downstream genes are considered, the output must be trimmed to extract the bona fide orthologous lincRNAs. In this case, we decided to restrict the distance between coding genes to one, and to consider only the clusters formed by one upstream gene, the lincRNA and one downstream gene. From the 32 clusters, only the 16 presented in Table 1 were considered to have a bona fide orthologous lincRNA. We also analyzed under these restrictions the clusters formed by two upstream genes and the lincRNA and by the lincRNA and two downstream genes (Table S1, Table S2). The rate of orthologous lincRNA finding was around 45% in the aforementioned analyses and the whole raw output is available in the supplementary info (File S1). Using this approach, we were able to obtain a list of 16 lincRNAs putatively conserved between human and amphioxus (Table 1). To our knowledge, this list represents the best set of highly curated lincRNAs with the deepest evolutionary conservation reported to date [27]. The main advantage of LincOFinder over other methods based on lincRNA sequence conservation is that it relies on microsyntenic conservation and a proper establishment of interspecific orthology relationships, which are more evolutionary constrained than the highly mutable nucleotide sequences of lincRNAs. In conclusion, LincOFinder can help to uncover conserved lincRNAs over deep evolutionary distances, in any species for which proper gene annotation data is available.

Table 1. Putatively conserved lincRNAs between *Homo sapiens* and *Branchiostoma lanceolatum*. Analysis of the genes surrounding the lincRNA focusing on the three core genes (upstream1, lincRNA and downstream1). Some lincRNAs can be ascribed to more than one hypothetically conserved microsyntenic cluster.

Orthologous lincRNAs ¹	State of the Cluster in Human ²	Human Orthologous lincRNA ³
BL20528 Sc0000000 28 +	* Conserved microsynteny	ENST00000623777.1_1
BL38782 Sc0000000 30 +	* Conserved microsynteny	HOTAIRM1
BL90848 Sc0000001 150 –	Correct order but strands inverted. In addition, there are two lincRNAs surrounding the cluster	AL354977.2
BL79733 Sc0000007 52 +	One gene with the strand inverted	BANCR
BL84418 Sc0000009 86 +	One gene with the strand inverted	TCONS_12_00008513
BL91140 Sc0000010 144 –	* Conserved microsynteny	TCONS_00027655
BL91143 Sc0000010 145 –	Couple of lincRNAs in amphioxus, synteny conserved in the coding genes, but not in the lincRNA	TCONS_00006308
BL53024 Sc0000015 55 +	* Conserved microsynteny	RP11-181G12.4
BL82992 Sc0000016 15 –	One gene with the strand inverted	TCONS_00027115
BL78145 Sc0000039 54 +	One gene with the strand inverted	TCONS_00000550
BL55463 Sc0000050 45 +	* Conserved microsynteny	TCONS_00024711
BL68900 Sc0000072 2 +	* Conserved microsynteny	TCONS_00011710
BL54861 Sc0000089 4 –	Synteny conserved in the coding genes	AI219887
BL41904 Sc0000219 3 –	Problematic region with several lincRNAs and massive distances	AC109136.1
BL72725 Sc0000229 14 +	Problematic region with several lincRNAs and massive distances	AC124852.1
BL59605 Sc0000234 6 –	* Conserved microsynteny	TCONS_00007813
BL38170 Sc0000240 5 +	* Conserved microsynteny	BC043517
	* Conserved microsynteny	LINC00114
	* Conserved microsynteny	TCONS_00011870
	* Conserved microsynteny	LOC100132215

¹ GeneID, Scaffold, virtual coordinates and strand separated by “|”. ² Description of the synteny of the cluster status in human. ³ ID of the putative human orthologous lincRNA. * indicates a perfect match in strand and order of the three core genes.

3.2. Conservation of HOTAIRM1 across Chordates

3.2. Conservation of HOTAIRM1 across Chordates

HOTAIRM1 was selected for further study because its conservation across several vertebrate lineages has been previously underscored [28,29] and its mechanism of action has been thoroughly studied [30]. HOTAIRM1 was identified for the first time in myelopoietic human cells [31], during a screen for transcriptionally active intergenic elements within the HoxA cluster. In amphibians, it is situated in the Hox cluster between *Hox1* and *Hox2* and between *Hoxa1* and *Hoxa2* in vertebrates (Figure 2). According to our microsynteny analysis, *Hotair1* is conserved in most of the chordate species analyzed, with the notable exception of zebrafish (Figure 2). Nonetheless, *Hotair1* appears to be present in the outgroup chordates like spotted gar (data not shown), *A. mitchelli* and *A. baileyi* (data not shown). Given that the Hox cluster is disintegrated, as well as in teleosts due to the presence of *Hox1* in this chordate subphylum [92] and due to the loss of *Hoxa1* in teleosts due to the presence of *Hoxa1* in this chordate subphylum [92], we were unable to confirm the presence of *Hotair1* in the genome of the ray-finned fish *Parachanna mitchelli* and *A. baileyi* possibly because of the microsynteny in this region due to a lineage-specific loss or alternatively because the lncRNA annotation was deficient [33].

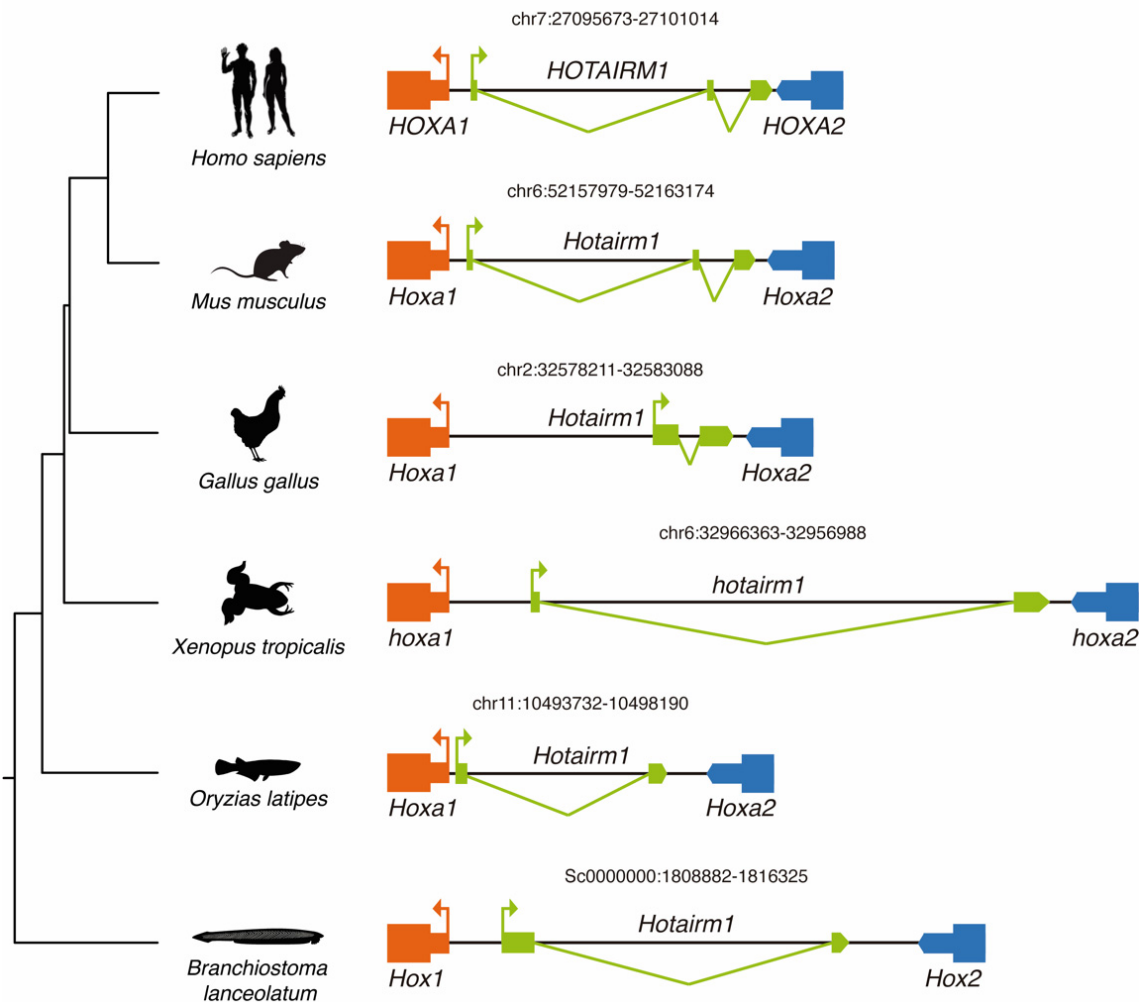


Figure 2. Schematic representation of the genomic locus of *HotairM1* across several chordate species. Genome or scaffold position is indicated above each *HotairM1* locus.

Our results add up to previous studies that have described the presence of *HotairM1* in mammals [28], birds and reptiles [29], strongly suggesting that *HotairM1* was retained within the HoxA cluster after the vertebrate-specific rounds of genome duplication [34] and that its origin predates, at least, the appearance of extant chordate lineages more than 500 million years ago.

3.3. HOTAIRM1 Expression Patterns in *Amphioxus* and *Xenopus*
 3.3. HOTAIRM1 Expression Patterns in *Amphioxus* and *Xenopus*

The expression domain of *Hotairm1* is mostly unknown, although it has been observed to be significantly increased or decreased in several types of cancer [35–37]. Furthermore, its expression is dynamically regulated during neuronal differentiation, showing a sharp increase in early differentiating neurons [38]. According to available RNA-seq data [15], the expression of *Hotairm1* during *B. lanceolatum* development peaks at 27 h post fertilization (hpf) (File S1). To investigate *Hotairm1* expression during amphioxus development, we performed fluorescent in situ hybridization in embryos from 18 hpf to 48 hpf. At 18 hpf we couldn't detect any signal, while at 21 hpf the expression of *Hotairm1* appeared in scattered cells in the presomitic mesoderm and in the neural plate partly overlapping *Hox1* expression domain (data not shown) [39]. At 30 hpf, 36 hpf and 48 hpf *Hotairm1* expression is restricted to the neural tube from the 5th somite towards the anterior developing neural tube, probably reaching the developing neural tube, probably reaching the Di-Mesencephalic primordium (DiMes) [39,40] (Figure 3A). Relevantly, the expression domain of *Hotairm1* in this developmental stage overlaps with *Hox1* which is also expressed in the developing neural tube and localized in the hindbrain neural tube and localized in the hindbrain (Figure 3C,C') [41,42].

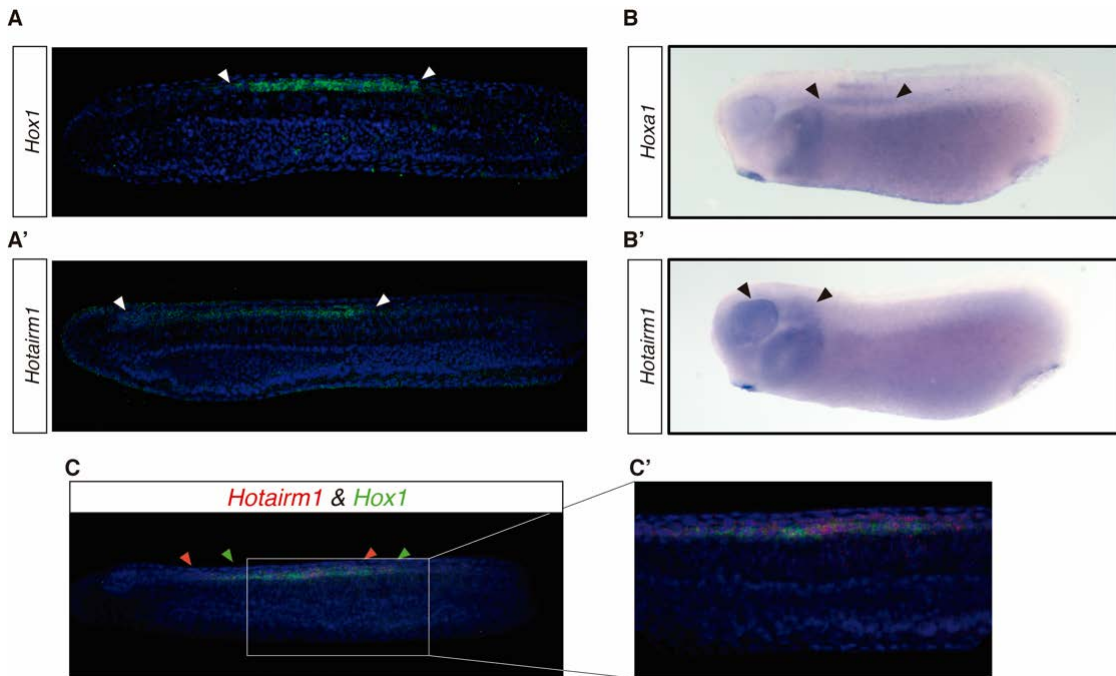


Figure 3. In situ hybridization (*ish*) in *B. lanceolatum* and *X. tropicalis*. Anterior to the left, dorsal is up. (A,A') Fluorescent ISH in *B. lanceolatum* whole mounts for (A) *Hox1* and (A') *Hotairm1* in 30 hpf embryos. White arrows mark the anterior and posterior limits of the expression domain. (B,B') Colorimetric whole mount *ish* in *X. tropicalis* tadpoles for (B) *hoxa1* and (B') *hotairm1*. Black arrows mark the anterior and posterior limits of the expression domain. (C,C') Fluorescent double ISH in *B. lanceolatum* whole mounts for (C) *Hox1* and (C') *Hotairm1* in a 36 hpf embryo and (C) the detailed zone where *Hotairm1* peaks its expression. Green arrows mark the anterior posterior limits of expression and red arrows mark the ones of *Hotairm1*.

In the vertebrate *X. tropicalis*, the expression of *hotairm1* was detected in the midbrain, hindbrain and the pharyngeal arch (Figure 3B'). At the same time, *hoxa1* in this species is expressed in the pharyngeal arch and the anterior developing neural tube (Figure 3B) [43].

These results suggest that the expression domain of *hotairm1* is conserved between *B. lanceolatum* and *X. tropicalis*, being expressed in the anterior half of the developing neural tube and partly overlapping with *Hox1* and *hoxa1* expression, respectively (Figure 3).

3.4. HOTAIRM1 Function and Expression Conservation

3.4. HOTAIRM1 Function and Expression Conservation

Remarkably, HOTAIRM1 has been described to act as a regulator of the chromatin state within the nucleus in HOTAIRM1-deficient cells [30]. In fact, the regulation of HOTAIRM1 isoform state within the nucleus is conserved in mice [30]. Wang & Dostie [30] regulation of the HOTAIRM1 isoform state in the presence of a polycomb repressor complex (PRC2) in flies is due to the fact that the spliceosome binds to the polycomb repressor complex (PRC2) and changes the chromatin state of the HOTAIRM1 gene (HOTAIRM1-HOXA5) to repress the expression of the proximal Hox genes (HOXA4 and HOXA3) and promote the expression of the proximal Hox genes (HOXA4 and HOXA3) expression balance of its isoforms during the functional development. We tried to alter the expression balance of its isoforms during *Xenopus laevis* development by splice order inhibition, this was used as a morpholino targeting the splice junction, thus forcing an isoform switch towards the unspliced state (Figure 4).

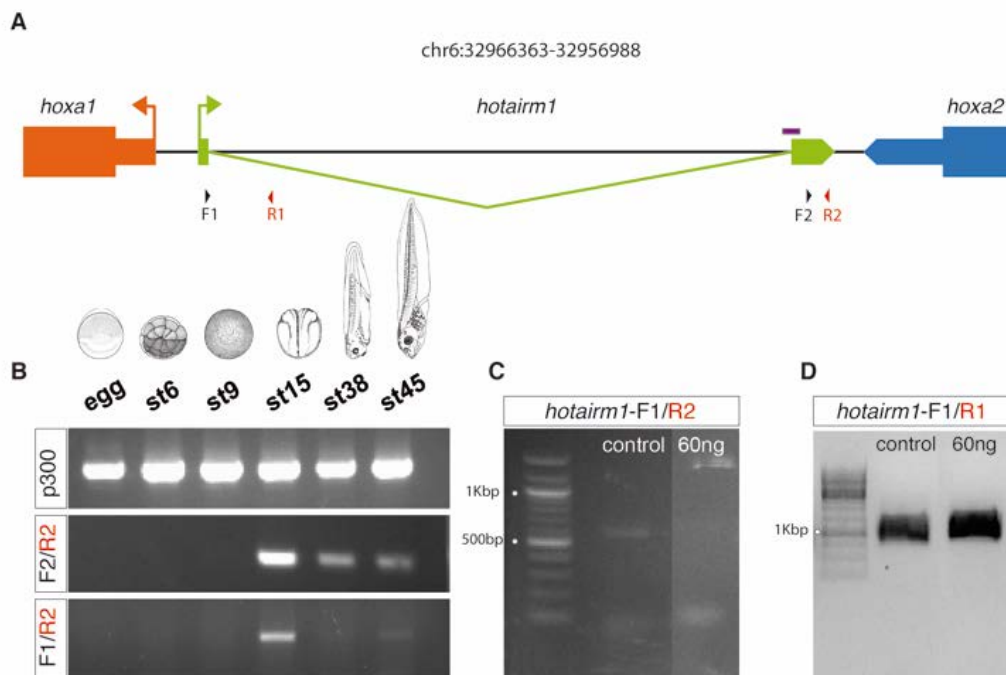


Figure 4. Isoform switch of *hotairm1* expression towards the unspliced state using a morpholino. (A) Genomic map of the *hotairm1* gene (black and red arrows) and the morpholino targeting the splice junction (green arrow) and the unspliced isoform (black arrow) and the spliced isoform (red arrow) of *hotairm1* in *Xenopus laevis* (p and f) are indicated. (B) Expression of *hotairm1* isoforms (p and f) in the embryonic stages (egg, st6, st9, st15, st38, st45). (C) Inhibition of the MO-related isoform in *Xenopus laevis* embryos. (D) Assessment of the presence of the unspliced isoform of *hotairm1* in MO treated embryos as well as in the control embryos at st18.

Strikingly, the morpholino treatment resulted in a headless tadpole-stage embryo (Figure 5A). This phenotype is characterized by the decrease of expression of brain markers such as *otx2* (forebrain-midbrain boundary marker) and *engrailed* (midbrain-hindbrain boundary marker) (Figure 5B,C). These results suggest that alterations in the balance of *Hotairm1* isoforms produce a severe disruption in the development of the anterior part of the central neural system (Figure 5).

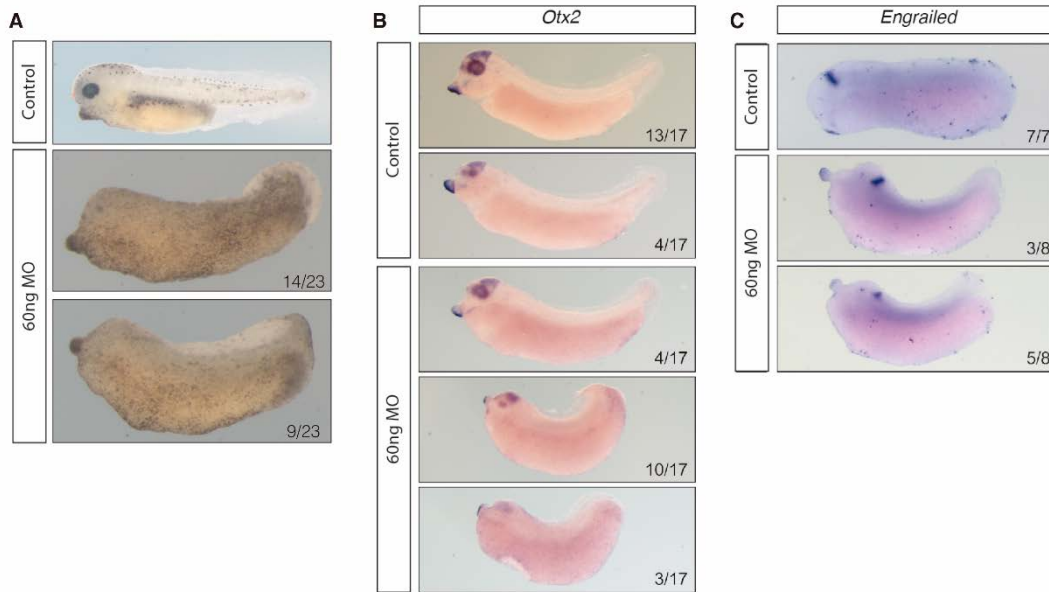


Figure 5. MO treated embryos and in situ hybridization in MO treated embryos. Anterior to the left, dorsal is up. (A) Control *X. tropicalis* MO treated embryos with normal development. 60ng Notaim1-MO treated embryos with a posteriorization of the anterior part of the embryo. (B) Whole mount colorimetric *ish* of *Otx2* in *X. tropicalis* stage 26 control embryos and MO treated embryos showing the reduced expression domain of *Otx2* in MO treated embryos. (C) Whole mount colorimetric *ish* of *engrailed* in *X. tropicalis* stage 26 control embryos and MO treated embryos showing a clear reduction in the expression in the MO treated embryos.

Finally, in order to check whether the expression of HoxA genes was altered in MO treated *Xenopus* embryos, we performed Real Time quantitative PCR at stage 18, when neurulation is taking place. Our results show a significant upregulation of medial Hox genes *hoxa5* and *hoxa6*, and a downregulation of *hoxa4*, compared with control embryos. Remarkably, no significant change in the expression of the proximal Hox gene, *hoxa1*, was observed. These results suggest that *HOXA* function is partially conserved between *Xenopus* and human (Figure 6).

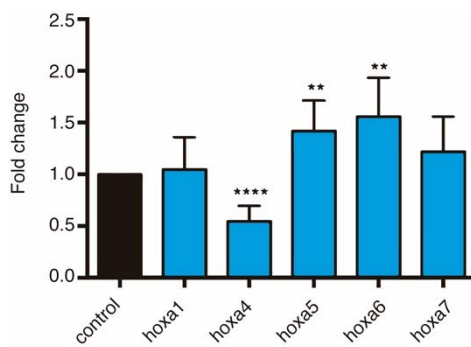


Figure 6. qPCR between 60ng MO treated embryos and control samples at stage 18. * shows statistically significance compared with control samples (Student's t-test, $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$, **** $p < 0.0001$). *Gmptn* was used as a reference gene.

4. Conclusions

We have developed a novel pipeline called lincOFinder that establishes bona fide microsyntenic clusters to detect deeply conserved lincRNAs. Applying this tool to investigate the invertebrate-vertebrate transition, we have managed to identify 18 lincRNA putatively conserved between amphioxus and humans. To our knowledge, this represents the first successful identification of homologous lincRNAs over very long evolutionary distances. We show that one of these conserved

homologous lincRNAs over very long evolutionary distances. We show that one of these conserved lincRNAs, *Hotairm1*, is expressed along the anterior half of the neural tube during amphioxus and *Xenopus* development. The injection of MO targeting the 3' splice junction triggers an imbalance between the spliced and unspliced form resulting in the disruption of the proximal and medial *hoxa* genes. This change in *hoxa* expression produces in a tadpole a patterning defect in the anterior neural system leading to a headless phenotype. However, further work needs to be done to elucidate the molecular mechanism underlying this severe phenotype. This nonetheless, is a reliable indicative that this lincRNA is at least to some degree conserved in amphioxus, *Xenopus* and human, allowing us to infer that it is conserved in the phylum Chordata and that regulation of the Hox cluster by lincRNAs may be traced back to the origin of chordates.

Supplementary Materials: The following are available online at <http://www.mdpi.com/2079-7737/8/3/61/s1>, Table S1: Putatively conserved lincRNAs between *Homo sapiens* and *Branchiostoma lanceolatum*. Analysis of the genes surrounding the lincRNA focusing on three genes (upstream2, upstream1 and lincRNA); Table S2: Putatively conserved lincRNAs between *Homo sapiens* and *Branchiostoma lanceolatum*. Analysis of the genes surrounding the lincRNA focusing on three genes (lincRNA, downstream1 and downstream2); File S1: Homology_Expression_RawOut.xlsx.

Author Contributions: Conceptualization, C.H.-U., E.N.-P. and M.M.-B.; methodology, C.H.-U. and M.M.-B.; software, C.H.-U. and J.G.; validation, C.H.-U., and M.M.-B.; formal analysis, C.H.-U. and J.G.; investigation, C.H.-U., B.A.-C. and M.M.-B.; resources, J.G.-F., G.N.W.; data curation, C.H.-U. and J.G.; writing—original draft preparation, C.H.-U., M.M.-B. and E.N.-P.; writing—review and editing, C.H.-U., M.M.-B., J.G.-F., G.N.W. and E.N.-P.; visualization, C.H.-U. and M.M.-B.; supervision, J.G.-F. and G.N.W.; project administration, J.G.-F.; funding acquisition, J.G.-F. and G.N.W.

Funding: This research was funded by grant BFU2017-86152-P (Ministerio de Ciencia, Innovación y Universidades, Spanish Government) to J.G.-F., C.H.-U. holds a predoctoral FPI contract (Ministerio de Ciencia, Innovación y Universidades, Spanish Government) and People Program (Marie Curie Actions) of the European Union's Seventh Framework Program FP7 under REA Grant agreement number 607142 (DevCom) to GW.

Acknowledgments: We wish to thank Mael Irimia for help, advice and support, and Tom Lewin for the modifications to the HCR protocol for *ish*.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

- Holland, P.W.H.; Garcia-Fernández, J.; Williams, N.A.; Sidow, A. Gene duplications and the origins of vertebrate development. *Development* **1994**, *1994*, 125–133.
- Schmitz, J.F.; Zimmer, F.; Bornberg-Bauer, E. Mechanisms of transcription factor evolution in Metazoa. *Nucleic Acids Res.* **2016**, *44*, 6287–6297. [[CrossRef](#)] [[PubMed](#)]
- Morris, K.V.; Mattick, J.S. The rise of regulatory RNA. *Nat. Rev. Genet.* **2014**, *15*, 423–437. [[CrossRef](#)] [[PubMed](#)]
- Zampetaki, A.; Albrecht, A.; Steinhofel, K. Long Non-coding RNA Structure and Function: Is There a Link? *Front. Physiol.* **2018**, *9*, 1201. [[CrossRef](#)] [[PubMed](#)]
- Wan, Y.; Kertesz, M.; Spitale, R.C.; Segal, E.; Chang, H.Y. Understanding the transcriptome through RNA structure. *Nat. Rev. Genet.* **2011**, *12*, 641–655. [[CrossRef](#)] [[PubMed](#)]
- Ponting, C.P.; Oliver, P.L.; Reik, W. Evolution and functions of long noncoding RNAs. *Cell* **2009**, *136*, 629–641. [[CrossRef](#)]
- Fico, A.; Fiorenzano, A.; Pascale, E.; Patriarca, E.J.; Minchiotti, G. Long non-coding RNA in stem cell pluripotency and lineage commitment: functions and evolutionary conservation. *Cell. Mol. Life Sci.* **2019**, *76*, 1459–1471. [[CrossRef](#)]
- Diederichs, S. The four dimensions of noncoding RNA conservation. *Trends Genet.* **2014**, *30*, 121–123. [[CrossRef](#)]
- Jathar, S.; Kumar, V.; Srivastava, J.; Tripathi, V. Technological developments in lincRNA biology. In *Advances in Experimental Medicine and Biology*; Springer: Singapore, 2017; Vol. 1008, pp. 283–323.
- Rivas, E.; Clements, J.; Eddy, S.R. A statistical test for conserved RNA structure shows lack of evidence for structure in lincRNAs. *Nat. Methods* **2017**, *14*, 45–48. [[CrossRef](#)]

11. Neme, R.; Tautz, D. Fast turnover of genome transcription across evolutionary time exposes entire non-coding DNA to de novo gene emergence. *Elife* **2016**, *5*. [[CrossRef](#)]
12. Garcia-Fernandez, J.; Benito-Gutierrez, E. It's a long way from amphioxus: descendants of the earliest chordate. *Bioessays* **2009**, *31*, 665–675. [[CrossRef](#)]
13. Paps, J.; Holland, P.W.H.; Shimeld, S.M. A genome-wide view of transcription factor gene diversity in chordate evolution: less gene loss in amphioxus? *Brief. Funct. Genomics* **2012**, *11*, 177–186. [[CrossRef](#)]
14. Putnam, N.H.; Butts, T.; Ferrier, D.E.K.; Furlong, R.F.; Hellsten, U.; Kawashima, T.; Robinson-Rechavi, M.; Shoguchi, E.; Terry, A.; Yu, J.-K.; et al. The amphioxus genome and the evolution of the chordate karyotype. *Nature* **2008**, *453*, 1064–1071. [[CrossRef](#)]
15. Marlétaz, F.; Firbas, P.N.; Maeso, I.; Tena, J.J.; Bogdanovic, O.; Perry, M.; Wyatt, C.D.R.; de la Calle-Mustienes, E.; Bertrand, S.; Burguera, D.; et al. Amphioxus functional genomics and the origins of vertebrate gene regulation. *Nature* **2018**, *564*, 64–70. [[CrossRef](#)]
16. Bertrand, S.; Escriva, H.; Williams, N.A.; Holland, N.D.; Holland, L.Z. Evolutionary crossroads in developmental biology: amphioxus. *Development* **2011**, *138*, 4819–4830. [[CrossRef](#)]
17. Pegueroles, C.; Iraola-Guzmán, S.; Chorostecki, U.; Ksiezopolska, E.; Saus, E.; Gabaldón, T. Transcriptomic analyses reveal groups of co-expressed, syntenic lncRNAs in four species of the genus *Caenorhabditis*. *RNA Biol.* **2019**, *16*, 320–329. [[CrossRef](#)]
18. Bush, S.J.; Muriuki, C.; McCulloch, M.E.B.; Farquhar, I.L.; Clark, E.L.; Hume, D.A. Cross-species inference of long non-coding RNAs greatly expands the ruminant transcriptome. *Genet. Sel. Evol.* **2018**, *50*, 20. [[CrossRef](#)]
19. Zerbino, D.R.; Achuthan, P.; Akanni, W.; Amode, M.R.; Barrell, D.; Bhai, J.; Billis, K.; Cummins, C.; Gall, A.; Girón, C.G.; et al. Ensembl 2018. *Nucleic Acids Res.* **2018**, *46*, D754–D761. [[CrossRef](#)]
20. Emms, D.M.; Kelly, S. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biol.* **2015**, *16*, 157. [[CrossRef](#)]
21. Bogdanovic, O.; Alexis, M.S.; Tena, J.J.; Maeso, I.; Fernandez-Minan, A.; Fraser, H.B.; Gomez-Skarmeta, J.L.; Roy, S.W.; Irimia, M.; de la Calle-Mustienes, E. Extensive conservation of ancient microsynteny across metazoans due to cis-regulatory constraints. *Genome Res.* **2012**, *22*, 2356–2367.
22. Sokal, R.R. A statistical method for evaluating systematic relationships. *Univ. Kansas, Sci. Bull.* **1958**, *38*, 1409–1438.
23. Kent, W.J.; Sugnet, C.W.; Furey, T.S.; Roskin, K.M.; Pringle, T.H.; Zahler, A.M.; Haussler, D. The human genome browser at UCSC. *Genome Res.* **2002**, *12*, 996–1006. [[CrossRef](#)]
24. Monsoro-Burq, A.H. A Rapid Protocol for Whole-Mount In Situ Hybridization on *Xenopus* Embryos. *Cold Spring Harb. Protoc.* **2007**, 2007. [[CrossRef](#)]
25. Fuentes, M.; Benito, E.; Bertrand, S.; Paris, M.; Mignardot, A.; Godoy, L.; Jimenez-Delgado, S.; Oliveri, D.; Candiani, S.; Hirsinger, E.; et al. Insights into spawning behavior and development of the european amphioxus (*Branchiostoma lanceolatum*). *J. Exp. Zool. Part B Mol. Dev. Evol.* **2007**, *308B*, 484–493. [[CrossRef](#)]
26. Choi, H.M.T.; Schwarzkopf, M.; Fornace, M.E.; Acharya, A.; Artavanis, G.; Stegmaier, J.; Cunha, A.; Pierce, N.A. Third-generation in situ hybridization chain reaction: multiplexed, quantitative, sensitive, versatile, robust. *Development* **2018**, *145*, dev165753. [[CrossRef](#)]
27. Ulitsky, I. Evolution to the rescue: using comparative genomics to understand long non-coding RNAs. *Nat. Rev. Genet.* **2016**, *17*, 601–614. [[CrossRef](#)]
28. Yu, H.; Lindsay, J.; Feng, Z.-P.; Frankenberg, S.; Hu, Y.; Carone, D.; Shaw, G.; Pask, A.J.; O'Neill, R.; Papenfuss, A.T.; et al. Evolution of coding and non-coding genes in HOX clusters of a marsupial. *BMC Genomics* **2012**, *13*, 251. [[CrossRef](#)]
29. Gardner, P.P.; Fasold, M.; Burge, S.W.; Ninova, M.; Hertel, J.; Kehr, S.; Steeves, T.E.; Griffiths-Jones, S.; Stadler, P.F. Conservation and Losses of Non-Coding RNAs in Avian Genomes. *PLoS One* **2015**, *10*, e0121797. [[CrossRef](#)]
30. Wang, X.Q.D.; Dostie, J. Reciprocal regulation of chromatin state and architecture by HOTAIRM1 contributes to temporal collinear HOXA gene activation. *Nucleic Acids Res.* **2017**, *45*, 1091–1104. [[CrossRef](#)]
31. Zhang, X.; Lian, Z.; Padden, C.; Gerstein, M.B.; Rozowsky, J.; Snyder, M.; Gingeras, T.R.; Kapranov, P.; Weissman, S.M.; Newburger, P.E. A myelopoiesis-associated regulatory intergenic noncoding RNA transcript within the human HOXA cluster. *Blood* **2009**, *113*, 2526–2534. [[CrossRef](#)]

32. Sekigami, Y.; Kobayashi, T.; Omi, A.; Nishitsuji, K.; Ikuta, T.; Fujiyama, A.; Satoh, N.; Saiga, H. Hox gene cluster of the ascidian, *Halocynthia roretzi*, reveals multiple ancient steps of cluster disintegration during ascidian evolution. *Zool. Lett.* **2017**, *3*, 17. [[CrossRef](#)]
33. Pascual-Anaya, J.; Sato, I.; Sugahara, F.; Higuchi, S.; Paps, J.; Ren, Y.; Takagi, W.; Ruiz-Villalba, A.; Ota, K.G.; Wang, W.; et al. Hagfish and lamprey Hox genes reveal conservation of temporal colinearity in vertebrates. *Nat. Ecol. Evol.* **2018**, *2*, 859–866. [[CrossRef](#)]
34. Dehal, P.; Boore, J.L. Two Rounds of Whole Genome Duplication in the Ancestral Vertebrate. *PLoS Biol.* **2005**, *3*, e314. [[CrossRef](#)]
35. Esfandi, F.; Taheri, M.; Omrani, M.D.; Shadmehr, M.B.; Arsang-Jang, S.; Shams, R.; Ghafouri-Fard, S. Expression of long non-coding RNAs (lncRNAs) has been dysregulated in non-small cell lung cancer tissues. *BMC Cancer* **2019**, *19*, 222. [[CrossRef](#)]
36. Li, Q.; Dong, C.; Cui, J.; Wang, Y.; Hong, X. Over-expressed lncRNA HOTAIRM1 promotes tumor growth and invasion through up-regulating HOXA1 and sequestering G9a/EZH2/Dnmts away from the HOXA1 gene in glioblastoma multiforme. *J. Exp. Clin. Cancer Res.* **2018**, *37*, 265. [[CrossRef](#)]
37. Song, L.; Zhang, S.; Duan, C.; Ma, S.; Hussain, S.; Wei, L.; Chu, M. Genome-wide identification of lncRNAs as novel prognosis biomarkers of glioma. *J. Cell. Biochem.* **2019**. [[CrossRef](#)]
38. Lin, M.; Pedrosa, E.; Shah, A.; Hrabovsky, A.; Maqbool, S.; Zheng, D.; Lachman, H.M. RNA-Seq of Human Neurons Derived from iPS Cells Reveals Candidate Long Non-Coding RNAs Involved in Neurogenesis and Neuropsychiatric Disorders. *PLoS One* **2011**, *6*, e23356. [[CrossRef](#)]
39. Albuixech-Crespo, B.; López-Blanch, L.; Burguera, D.; Maeso, I.; Sánchez-Arrones, L.; Moreno-Bravo, J.A.; Somorjai, I.; Pascual-Anaya, J.; Puellas, E.; Bovolenta, P.; et al. Molecular regionalization of the developing amphioxus neural tube challenges major partitions of the vertebrate brain. *PLoS Biol.* **2017**, *15*, e2001573. [[CrossRef](#)]
40. Albuixech-Crespo, B.; Herrera-Úbeda, C.; Marfany, G.; Irimia, M.; Garcia-Fernández, J. Origin and evolution of the chordate central nervous system: insights from amphioxus genoarchitecture. *Int. J. Dev. Biol.* **2017**, *61*, 655–664. [[CrossRef](#)]
41. Schubert, M.; Holland, N.D.; Laudet, V.; Holland, L.Z. A retinoic acid-Hox hierarchy controls both anterior/posterior patterning and neuronal specification in the developing central nervous system of the cephalochordate amphioxus. *Dev. Biol.* **2006**, *296*, 190–202. [[CrossRef](#)]
42. Zieger, E.; Candiani, S.; Garbarino, G.; Croce, J.C.; Schubert, M. Roles of Retinoic Acid Signaling in Shaping the Neuronal Architecture of the Developing Amphioxus Nervous System. *Mol. Neurobiol.* **2018**, *55*, 5210–5229. [[CrossRef](#)]
43. McNulty, C.L.; Peres, J.N.; Bardine, N.; van den Akker, W.M.R.; Durston, A.J. Knockdown of the complete Hox paralogous group 1 leads to dramatic hindbrain and neural crest defects. *Development* **2005**, *132*, 2861–2871. [[CrossRef](#)]



Results

Article RIII:

Abstract

The homeobox genes *Pdx* and *Cdx* are widespread in the animal kingdom but functional data are available for few lineages. Current data suggest ancient roles in patterning the bilaterian through-gut, although additional roles such as promoting posterior extension of the body and patterning neural and mesodermal tissues are seen in some taxa. Here we use TALENs to engineer frameshift mutations in the *Pdx* and *Cdx* genes of amphioxus *Branchiostoma floridae*. Homozygous *Pdx* mutants have a defect in cell fate specification in amphioxus endoderm, manifest as loss of a midgut GFP-expressing region. The anus fails to open in homozygous *Cdx* mutants, which also have defects in posterior body extension and epidermal tail fin development. The body axis and tailfin phenotypes are a consequence of increased retinoic acid signalling, likely mediated through up-regulation of a *Cyp26* gene. Transcriptome analysis reveals extensive gene expression changes, with a disproportionate effect on gut-enriched genes, and a colinear-like effect of *Cdx* on Hox genes. These data are consistent with *Pdx* and *Cdx* having conserved roles in gut patterning and *Cdx* having an ancient role in promoting body axis extension.

Article RIII:

Mutation of amphioxus Pdx and Cdx demonstrates conserved roles for ParaHox genes in gut, anus and tail patterning

Yanhong Zhong[†], Carlos Herrera-Úbeda[†], Jordi Garcia-Fernàndez, Guang Li^{*} and Peter WH Holland^{*}

The PhD candidate is a first joint author, of an article that will be submitted soon to high impact journals. He was in charge of analysing the transcriptome of knockout amphioxus for the ParaHox genes *Pdx1* and *Cdx*. The KOs were generated in China, and all the informatic analyses and discussion on the results, elaborated by the PhD candidate. To my knowledge, the article has not been included in any other PhD thesis.

Signed: Jordi Garcia Fernàndez, PhD supervisor

Mutation of amphioxus *Pdx* and *Cdx* demonstrates conserved roles for ParaHox genes in gut, anus and tail patterning

Yanhong Zhong¹⁺, Carlos Herrera^{2,3+}, Jordi Garcia- Fernández³,
Guang Li^{1*}, Peter WH Holland^{2*}

1. State Key Laboratory of the Cellular Stress Biology,
School of Life Sciences, Xiamen University, Xiamen, China
2. Department of Zoology, University of Oxford, Oxford OX1 3SZ
3. Department of Genetics, Microbiology & Statistics,
University of Barcelona, 08028 Barcelona, Spain

+ Equal contribution

* Corresponding authors

Abstract

The homeobox genes *Pdx* and *Cdx* are widespread in the animal kingdom but functional data are available for few lineages. Current data suggest ancient roles in patterning the bilaterian gut, with additional roles in some taxa such as promoting axis extension and patterning neural and mesodermal tissues. Here we use TALENs to engineer frameshift mutations in the *Pdx* and *Cdx* genes of amphioxus *Branchiostoma floridae*. Homozygous *Pdx* mutants have a defect in cell fate specification in amphioxus endoderm, manifest as loss of a midgut GFP-expressing region. The anus fails to open in homozygous *Cdx* mutants, which also have defects in posterior body extension and epidermal tail fin development. The body axis and tailfin phenotypes are a consequence of increased retinoic acid signalling, likely mediated through up-regulation of a *Cyp26* gene. Transcriptome analysis reveals extensive gene expression changes with a disproportionate effect on gut-enriched genes and a colinear-like effect of *Cdx* on Hox genes. These data are consistent with *Pdx* and *Cdx* having conserved roles in gut patterning and *Cdx* having an ancient role in promoting body axis extension.

Introduction

The discovery that three non-Hox homeobox genes, *Gsx*, *Pdx* (*Xlox*) and *Cdx*, are organized in a small gene cluster raised the possibility that these 'ParaHox' genes might act in a coordinated way, in a comparable way to Hox gene clusters. The clustering was first described in a cephalochordate or amphioxus, *Branchiostoma floridae* (Brooke et al. 1998), and was later shown in human, *Xenopus*, *Polypterus*, *Amia* and several other vertebrates, an echinoderm and a hemichordate (Ferrier et al. 2005; Mulley et al. 2006; Illes et al. 2009; Annunziata et al. 2013; Ikuta et al. 2013). The clustering of ParaHox genes is likely to be the ancestral condition for bilaterians, although the cluster has been broken in many taxa through genomic rearrangements, individual gene losses or duplications followed by losses (Brooke et al. 1998; Mulley et al. 2006; Garstang and Ferrier 2013).

The two ParaHox genes with clearest similarity in expression are *Pdx* and *Cdx*. The *Pdx* gene, also called *Pdx1*, *Xlox*, *Lox*, *lpf1*, *ldx1* or *Stf1*, has been studied extensively in vertebrates and is expressed in endoderm of the developing gut, with sharp anterior and posterior limits, and later in pancreas and where duodenum meets stomach (Wright et al. 1998; Offield et al. 1996; Holland et al. 2013). In adult mammals, Pdx protein is a transcriptional activator of *insulin* and other genes in β cells of the endocrine pancreas (Wang et al. 2018). In amphioxus embryos, *Pdx* is also expressed in a sharp domain of midgut endoderm and in two neural cells as described below (Brooke et al. 1998). The amphioxus *Cdx* gene is also expressed in the gut, more posteriorly where the anus will break through and, in other posterior tissues at early developmental stages (Brooke et al. 1998). Similarly, vertebrate *Cdx* genes are expressed strongly in caudal regions, especially posterior gut, although this is complicated by the presence of multiple paralogues with subtly different patterns plus additional sites, such as *Cdx2* in trophoderm of mouse blastocyst (Beck et al. 1995; van den Akker et al. 2002; Illes et al. 2009; Marletaz et al. 2015). *Gsx* is rather different and expressed in neural tissue in amphioxus rather than gut; similarly, two mouse *Gsx* genes are expressed in brain (Brooke et al. 1998).

Expression in regions of the gut is clearest commonality between *Pdx* and *Cdx* genes, and described in many taxa in addition to chordates including echinoderms, annelids and molluscs. The consistency, and restriction to anteroposterior domains, led to the hypothesis that ancestrally these genes were components of a system for patterning of the bilaterian gut (Brooke et al. 1998; Holland 2001; Garcia-Fernàndez 2005; Cole et al. 2009; Samadi and Steiner 2010). The 'through-gut', with distinct mouth, digestive regions and anus allowing a unidirectional flow of ingested food, is a key character of bilaterian animals. Although the anus has been secondarily lost in some taxa, such as platyhelminths, it most likely dates to the base of the bilaterian clade. Intimately associated with a bilaterally symmetrical body with active directed locomotion, the evolution of a through-gut may have facilitated the evolution of predation and active burrowing, key drivers of animal diversification in the Cambrian (Brooke et al. 1998; Holland 2015; but see Hejnol et al. 2015). Experimental data illuminating the function of *Pdx* and *Cdx* genes in a wide range of bilaterians would help testing of this hypothesis. Specifically, we need to know if these two homeobox genes specify region-specific cell fates, instructing cells to differentiate along a path appropriate to their head to tail position.

In mice, deletion of the *Pdx* gene results in lack of pancreas (Jonsson et al. 1994; Offield et al. 1996); similarly, in humans pancreatic agenesis has been reported in patients with mutations in both *Pdx*

alleles (Stoffers et al. 1997; Schwitzgebel et al. 2003). Further insight comes from conditional deletion of the *Pdx* gene in adult mice, which caused homeotic-like transformation of endoderm cells in the *Pdx*-expressing region (Holland et al. 2013). Similarly, ectopic expression of *Pdx1* in chick embryo gut caused endodermal cells to change molecular identity and behaviour (Grapin-Botton et al. 2001). Together these findings indicate a role for *Pdx* in specification of region-specific endodermal cell fate in vertebrates. A similar specification role is also likely in sea urchin larvae, where morpholino knock-down of *Pdx* activity caused the sphincter between midgut and hindgut not to form (Cole et al. 2009; Annunziata and Arnone 2014). No insights come from *Drosophila* or nematodes since they have lost the *Pdx* gene in evolution, and functional data are difficult to obtain in other invertebrate taxa.

Gene loss is less of a problem for the *Cdx* gene and functional data are available for vertebrates, sea urchin, *Drosophila* and nematode. For example, heterozygous mutation in the mouse *Cdx2* gene caused homeotic-like transformation of posterior gut cells into a more anterior phenotype (Beck et al. 1999) and inhibiting *Cdx* function in sea urchin development allowed posterior gut cells to express a more anterior marker (Cole et al. 2009). Mutational studies show that in *Drosophila*, *Cdx* (*cad*) is necessary for invagination of the hindgut, and in *Caenorhabditis elegans* *Cdx* (*pal-1*) has roles in development of the rectum (Wu and Lengyel 1998; Edgar et al. 2001). Although these similarities suggest that a role in posterior gut or anus formation is widespread and may date to the base of Bilateria, it is important to test this in additional taxa.

In at least some taxa, posterior *Cdx* expression is associated with an important additional role: promotion of axis extension or tail growth. Animals in which a tail extension role has been demonstrated include vertebrates such as mice, *Xenopus* and zebrafish (van den Akker et al. 2002; Chawengsaksophak et al. 2004; Shimizu et al. 2005; Faas and Isaacs 2009), an ascidian (Katsuyama et al. 1999), and several short-germ arthropods: crustacean *Artemia*, beetle *Tribolium* and cricket *Gryllus* (Copf et al. 2004; Shinmyo et al. 2005). There is no similar function in *Drosophila*, but this could be a secondary condition because *Cdx* was recruited into a specialized network of interactions deployed to pattern the rapidly developing long-germ band embryo (Charité et al. 1998). It is also not clear if the tail extension role of *Cdx* is homologous between phyla. To emphasize that the body patterning and tail extension roles may be distinct, vertebrate *Cdx* genes have been described as having ‘biphasic’ properties: an early phase associated with specification of cell fates, and a later phase controlling tail growth (van den Akker et al. 2002). We were interested to resolve whether posterior expression of *Cdx* in amphioxus development is connected with cell fate specification in posterior gut and anus, or axial extension, or both.

Pdx and *Cdx* genes can also have roles in tissues outside the gut and tail. In mouse, mutations of *Cdx1* or *Cdx2* can cause homeotic transformations of mesodermally-derived vertebrae, at least in part by subtly altering Hox gene expression along the body axis (Subramanian et al., 1995, van den Akker et al., 2002). Interfering with *Cdx* function in *Xenopus* causes disruption to neural patterning and again Hox gene expression is affected (Isaacs et al. 1998; Faas and Isaacs 2009). In *Drosophila*, *Cdx* has been described as a posterior homeotic gene specifying cell fate, not just of hindgut but also of the anal plates situated on the exterior of the most posterior segment (Moreno and Morata 1999). Similarly, many cell types are affected by mutation of *Cdx* (*pal-1*) in the nematode

Caenorhabditis elegans, including a specific fate change in posterior epidermal rays effected through regulation of a Hox gene (Hunter et al. 1999). Turning to amphioxus, the *Cdx* gene is expressed in posterior neural tube and mesoderm, as well as gut, and *Pdx* is expressed in two putative receptor cells in the developing neural tube (Brooke et al. 1998; Osborne et al. 2009). These two cells lie adjacent to a single large *Mitf*-expressing pigment cell at the level of the fifth somite pair (Brooke et al. 1998; Yu et al. 2008) and are the first to develop in an extensive series of photosensory Organs of Hesse (dorsal ocelli). This particular Hesse organ differs from others in having two receptor cells instead of one (Lacalli and Stach 2016).

The recent development of mutagenesis methods applicable to amphioxus opens up new opportunities for testing gene functions in an animal occupying a pivotal phylogenetic position (Li et al. 2014, 2017). Here we use Transcription Activator-Like Effector Nucleases (TALENs) to generate stable germline mutations in the amphioxus *Pdx* or *Cdx* genes. Analysis of homozygous mutants reveals cell fate changes in specific regions of the gut. We also find that mutation of amphioxus *Cdx* also affects posterior epidermal cell fate and blocks posterior growth through a similar molecular pathway to vertebrates. These findings are consistent with an ancient bilaterian role of *Pdx* and *Cdx* in gut patterning and a role for *Cdx* in axis extension that dates at least to the base of Chordata.

Results

Generation of stable *Pdx* and *Cdx* mutant lines in amphioxus

We designed constructs encoding TALEN pairs targeted to the first coding exon of amphioxus *Pdx* and *Cdx* (Figure 1; Figures S1 to S6). In each case, TALEN pairs spanned a restriction endonuclease site to facilitate mutation detection. For each gene, two TALEN mRNAs were co-injected into unfertilized eggs of *B. floridae* and, after fertilization, successful mutagenesis was detected by PCR on pools of neurula stage embryos (Table S1, Figure S7). Remaining embryos were reared to maturity to generate mosaic founder F0 animals; to identify which animals carried germline mutations, founders were spawned and mutations typed by PCR on sperm or pools of neurulae generated by outcrossing. Embryos from F0 x wild type crosses were reared to maturity to generate F1 heterozygous mutants. Since each F1 animal may carry a slightly different mutation in the target gene, lines were expanded by crossing with wild type animals, rearing embryos to maturity and intercrossing offspring. Since we have no linked markers, inheritance of mutations was followed by PCR on amphioxus tail clips in live adults and PCR after in situ hybridization on individual embryos.

We generated amphioxus lines with 4 bp, 11 bp and 13 bp deletions (4 Δ , 11 Δ , 13 Δ) in the *Pdx* gene (Figure 1). Each *Pdx* deletion causes a frameshift and is predicted to give a protein comprising the first 31-33 amino acids of wild type protein followed by peptide sequence from a different reading frame with no strong similarity to any known protein (Figure S8). We also generated a *Cdx* mutant line with a 7 bp deletion (7 Δ) predicted to give a short peptide comprising the first 5 amino acids of *Cdx* plus 4 additional amino acids before a stop codon (Figure 1; Figure S9). None of the predicted products contains a homeodomain.

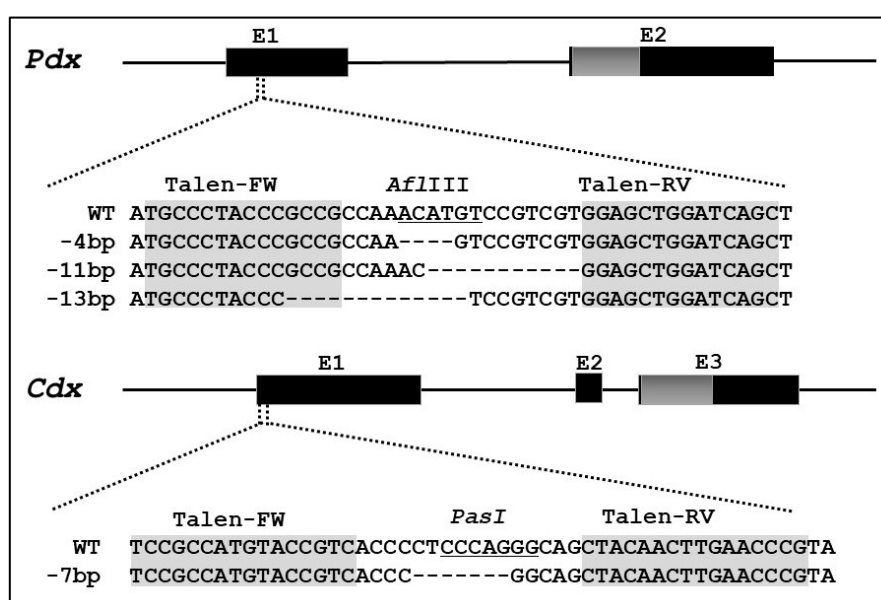


Figure 1: Gene structure, sequences targeted by TALENs and mutations generated. E1, E2, E3, coding regions of exons; grey, homeobox; WT, wild type sequence.

***Pdx* is necessary for correct spatial differentiation of midgut**

Embryos derived from crosses between heterozygous *Pdx* 4Δ mutants, and crosses between mutant lines, generated pools of morphologically identical embryos. Genotyping individual embryos revealed that homozygous *Pdx* mutants were identical to wild type and heterozygous embryos in external appearance. Since *Pdx* is strongly expressed in gut and the first two Hesse organ receptor cells in 13 h neurulae, we asked whether marker genes for these tissues were affected in *Pdx* 4Δ mutants. In situ hybridisation revealed no difference in expression of *melanopsin* (*Mop*), a gene expressed in the Hesse organ receptors (Figure 2A,B), or *Mitf*, a gene expressed in the adjacent pigment cell (Figure 2C,D). We also detected no difference in expression pattern in *Pdx* itself, *Cdx* or *Ilp1* in the endoderm of neurulae (Figure 2E-J).

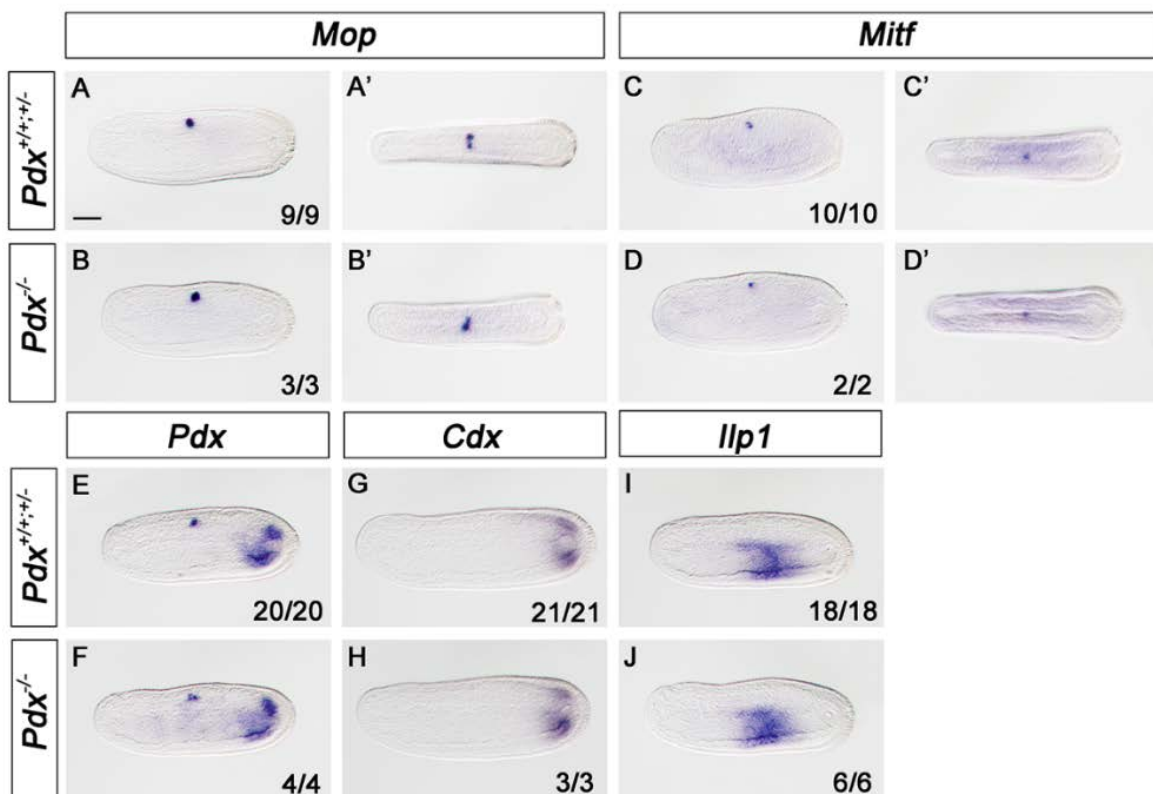


Figure 2: In situ hybridisation to wild type and *Pdx* 4Δ mutant 13 h neurulae stage amphioxus embryos for (A,B) *melanopsin* (*Mop*), (C,D) *Mitf*, (E,F) *Pdx*, (G,H) *Cdx* and (I,J) *Ilp1*. Embryos were genotyped by PCR after hybridisation.

As development proceeded to the larval stage, a morphological phenotype was visible in mutants when viewed under fluorescent illumination. In wild type and heterozygous larvae, endogenous green fluorescence was detected in cirri around the mouth and in a small clearly delineated patch in the endoderm of the midgut region. In contrast, in homozygous 4Δ mutant embryos, the principal fluorescence is in the buccal cirri. The difference in green fluorescence is first detectable at 6 d of development, and persists to at least 66 days of development (Figure 3, Figure S10). The same phenotype was observed in 4Δ/11Δ and 11Δ/13Δ compound heterozygotes (Figure S11).

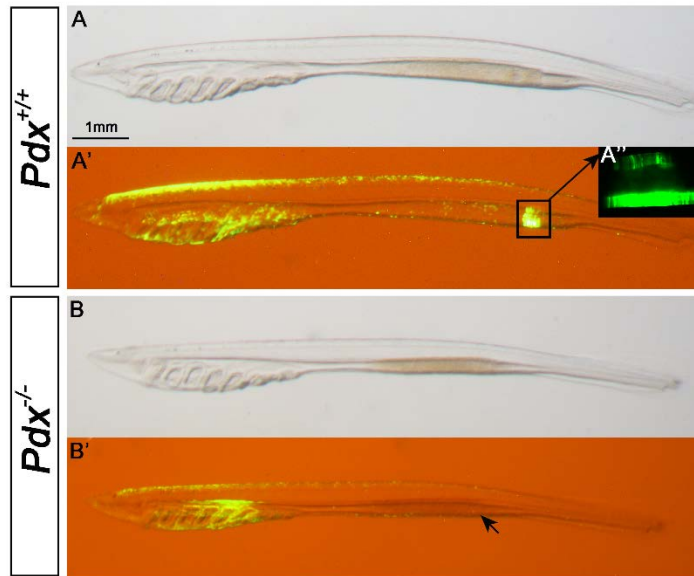


Figure 3. Green fluorescence in 7-gill (16 day) amphioxus larvae showing endodermal patch in wild type (A,A') but not homozygous *Pdx* 4 Δ mutants (B,B'). Inset A'' shows confocal imaging of sibling larva showing fluorescence located in endoderm cells.

***Cdx* is necessary for formation of the anus and posterior axial extension**

Homozygous *Cdx*^{-/-} mutants are identical to wild type amphioxus in external appearance through gastrula and neurula stages, and until the mouth opens around 24 hours post-fertilization (Figure 4; Figure S12). A clear morphological difference begins to manifest between 2 and 3 days of development as marked axial truncation. In normal 2-day larvae, the gut narrows in width very sharply at a position 80% of the distance from mouth to anus and this discontinuity provides a useful positional marker in the gut. However, the distance from the gut narrowing to the posterior end of the gut is shorter in mutant larvae, with all posterior tissues reduced, indicating that the entire body axis is truncated caudally (Figure 4A,B). As development proceeds, the posterior of the animal continues to extend in wild type larvae, but this axial extension is halted in mutants. By 6 days of development, mutant larvae are approximately half the length of wild type larvae (Figure S13). The length difference is predominantly the result of the cessation of tail extension, plus curvature of the body.

A second morphological difference becomes evident between 30 to 40 hours of development at the posterior end of the gut. In homozygous mutant animals, the posterior end of the gut remains closed by the epithelial cell layer of the endoderm and the anal opening does not form; in wild type animals, the anus perforates neatly to generate the through-gut in preparation for feeding to commence (Figure 4C,D). When cultured algal cells are provided to 2-day and 3-day larvae, mutant larvae take up food material and in some cases the closed anal region ruptures leaving a ragged terminus to the gut.

In addition, the tail fin is markedly smaller in mutants compared to wild type animals, and develops from a smaller zone of the posterior ectoderm (Figure 4A-D). Development of the amphioxus tail fin is driven by extension of specialised epidermal cells containing long intracellular ciliary rootlets, of

which a major component is the coiled-coil protein Rootletin or Crocc (Flood 1975, Koop et al 2011, Mansfield and Holland 2013). In situ hybridisation to homozygous *Cdx* mutant larvae reveals the amphioxus *Rootletin* gene is expressed by fewer cells than in control sibling larvae (Figure 4E-H). This alteration of *Rootletin* gene expression is consistent with a smaller number of epidermal cells being specified to differentiate into tail fin in *Cdx* mutant animals. The posteriorly expressed *Cyp26-3* gene is also greatly reduced in expression in *Cdx* mutant embryos (Figure 4I-L).

Thus, *Cdx* is necessary for correct fate specification of posterior ectoderm cells, for formation of the anus and for tail extension.

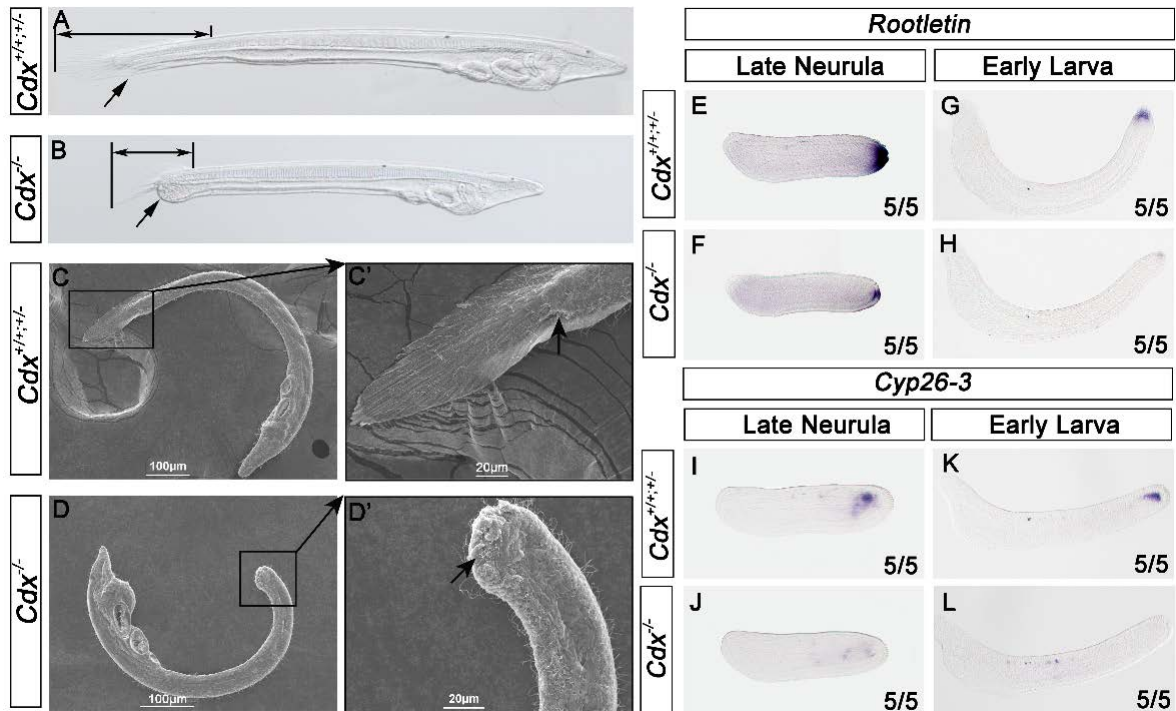


Figure 4. *Cdx* mutant phenotype. (A) Wild-type and (B) homozygous mutant 40 h larvae showing anus or lack of anus (arrow) and distance from gut restriction to tip of tail (double-headed arrow). Scanning electron micrographs showing (C) and (C') anus and tail fin in 40 h wild type larva and (D) and (D') closed anus and reduced tail fin in mutant larvae. In situ hybridisation to *Rootletin* RNA in (E) 16 h wild type embryo, (F) 16 h homozygous mutant embryo, (G) 21 h wild type embryo, (H) 21 h homozygous mutant embryo. In situ hybridisation to *Cyp26-3* RNA in (I) 16 h wild type embryo, (J) 16 h homozygous mutant embryo, (K) 21 h wild type embryo, (L) 21 h homozygous mutant embryo.

***Cdx* acts via the retinoic acid pathway in amphioxus posterior development**

The imperforate anus, tail truncation and tail fin phenotype observed in *Cdx* mutant larvae are similar to a suite of structural changes caused by exogenous treatment of larvae with retinoic acid, RA (Koop et al 2011). We therefore asked whether the phenotype we observe in *Cdx* mutant larvae is caused by disruption to RA signalling, specifically an increase in RA in the posterior region of mutant animals. Comparison between the two amphioxus studies should be made with caution,

however, as there are differences in experimental design. In our study *Cdx* gene function has been absent throughout development (after initiation of zygotic gene expression), while the exogenous RA study involved treatment of late larvae in which anus and tail fin had already formed.

If mutation of *Cdx* causes an increase in RA, we reasoned that dampening RA signalling may rescue the mutant phenotype. We treated amphioxus embryos derived from *Cdx* heterozygote crosses with an inverse agonist of RA signalling BMS493 at 5 h post-fertilisation (Figure 5). Consistent with the prediction, homozygous mutant larvae treated with BMS493 had enlarged tail fins compared to untreated mutants with the fin developing from a greater number of epidermal cells, although rescue was not complete (Figure 5b,d). Posterior growth of the body increased marginally, but the anus did not open. Treating wild type or heterozygous embryos with BMS493 gave the opposite phenotype to *Cdx* mutation (Figure 5a,c): an enlarged larval tail fin derived from a more extensive region of posterior ectoderm (consistent with the results of Carvalho et al. 2017). These phenotypes were presaged by subtle changes to expression of the *Rootletin* gene at earlier developmental stages, as detected by in situ hybridisation to 18 h embryos (Figure 5e-j). BMS493-treated *Cdx* mutant embryos had an enlarged patch of *Rootletin* gene expression, in some cases similar to wild type untreated embryos. BMS493-treated wild type and heterozygote embryos had an even larger *Rootletin* expression domain. Therefore, we infer that the role of *Cdx* in tail fin formation and tail extension acts, at least in part, through the RA pathway. An opposite interaction, RA acting on *Cdx*, has been shown previously (Osborne et al. 2009).

Further insight was obtained by examining a key player in the RA pathway, a cytochrome P450 family 26 (*Cyp26*) gene encoding an enzyme that degrades and clears excess RA (Thatcher & Isoherranen 2009). Amphioxus has three closely related tandemly-arranged *Cyp26* genes, derived from cephalochordate-specific tandem duplication (Albalat et al. 2001, Carvalho et al. 2017). In contrast to *Rootletin* gene expression, which responds in opposite directions to *Cdx* mutation (down) and BMS493 treatment (up), we find expression of amphioxus *Cyp26-3* is affected similarly by the two conditions. In homozygote *Cdx* mutants, posterior *Cyp26-3* expression is down-regulated but not abolished (Figure 5k,l); treatment with BMS493 also down-regulates *Cyp26-3* expression, but to a more extreme degree (Figure 5m,n).

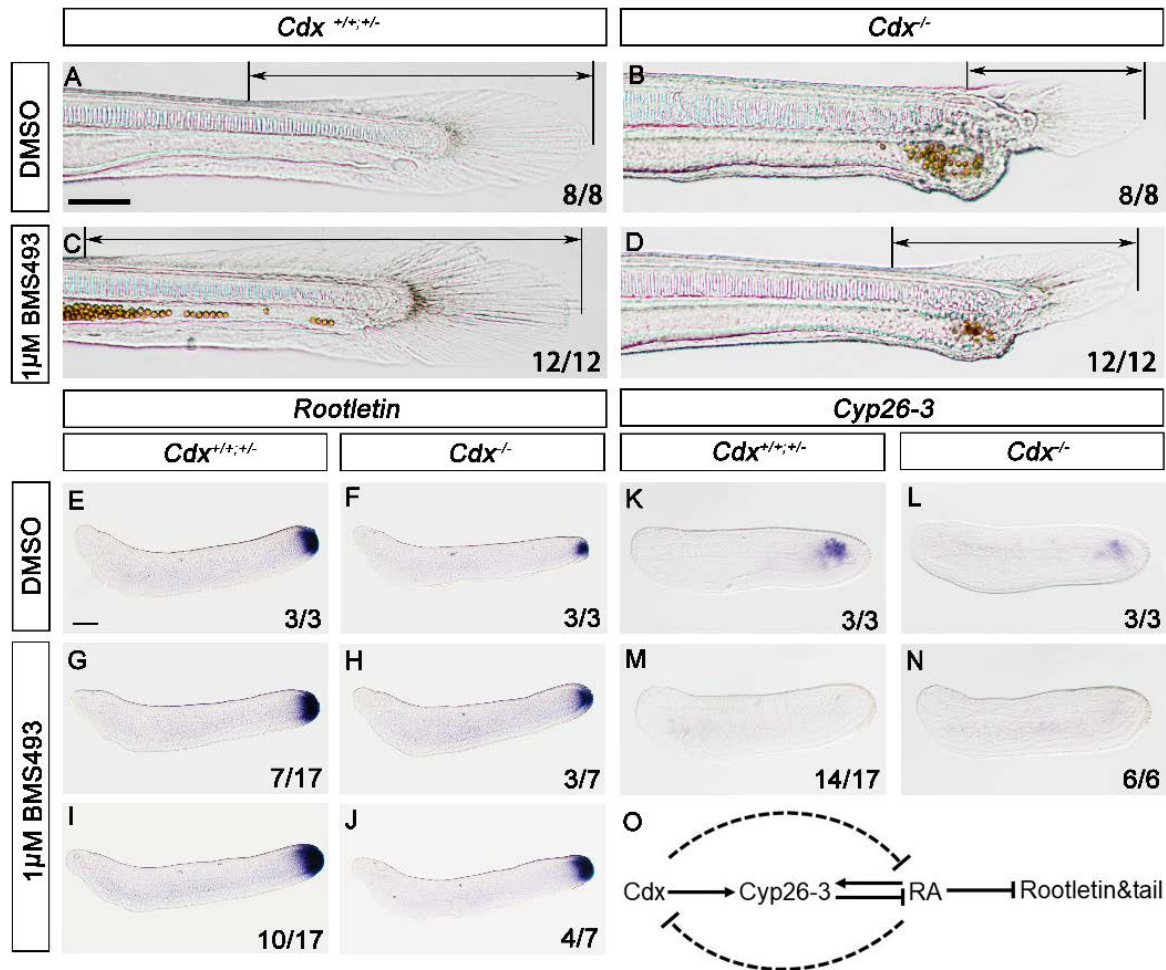


Figure 5: Effect of inhibiting the RA pathway interacts with *Cdx* function. The large tail fin evident in wild type and *Cdx*^{+/-} heterozygous ('normal') amphioxus larvae (A, 30 h) is severely truncated in *Cdx*^{-/-} homozygous mutants and develops from fewer epidermal cells (B). Inhibition of the RA pathway using BMS493 enlarges the tail fin of normal larvae (C) and partially rescues the truncation in *Cdx*^{-/-} mutants (D). Tail fin development is presaged by expression of the *Rootletin* gene in 18 h embryos, visualised by in situ hybridisation (E), which is also down-regulated in *Cdx*^{-/-} mutants. Inhibition of the RA pathway results on up-regulation (partial rescue) of *Rootletin* RNA to varying degrees in normal (G, I) and mutant (H, J) embryos. In contrast, posterior expression of *Cyp26-3* in 16 h embryos (K) is down-regulated in *Cdx*^{-/-} mutants (L) and by inhibition of RA action (M, N). These data are consistent with a model (O) involving inhibition of RA signalling by *Cdx* and a negative feedback loop; interactions in dotted lines not deduced from current work (RA inhibition of *Cdx* from Osborne et al. 2009).

To investigate if *Cyp26* genes could be direct transcriptional targets of amphioxus *Cdx*, as in mouse (Savory et al. 2009), we injected unfertilized amphioxus eggs with a construct in which a luciferase reporter gene is under the control of 3.1 kb DNA sequence 5' of the *B. floridae* *Cyp26a-3* gene, and assayed luciferase activity at the late neurula stage. The DNA sequence includes several putative *Cdx* binding sites (Figure S14) of which two have a close match to the consensus of Amin et al. (2016); mutation of either site singly, or both together, decreases luciferase activity consistent with *Cdx*-binding positively regulating *Cyp26-3* expression (Figure 6; Supplementary Information: Tables S3, S4).

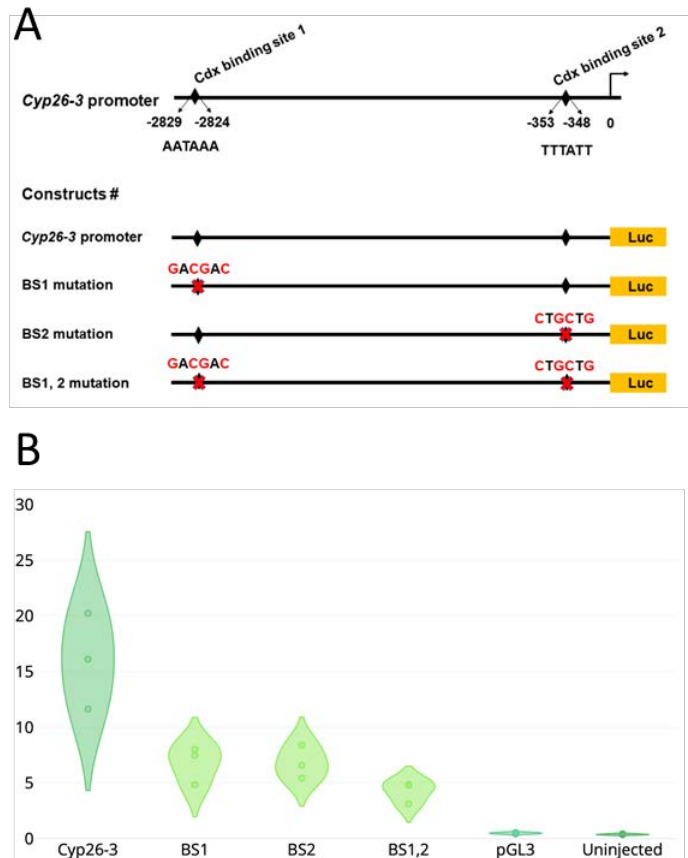


Figure 6: Reporter gene analysis of *Cyp26-3*. (A) Region 5' of *Cyp26-3* gene showing two putative Cdx binding sites. Numbers are distance before start codon. Four report gene constructs were tested in amphioxus embryos with 0, 1 or 2 Cdx sites mutated. (B) Relative luciferase expression for each construct and uninjected control.

Taken together, these results are consistent with a model in which a major role of *Cdx* expression is to ensure RA activity is kept at a low level in the posterior part of the amphioxus embryo. Disruption of *Cdx* leads to excess RA and morphological disruption. We suggest this action of *Cdx* occurs, at least in part, via positive regulation of *Cyp26-3* explaining why *Cyp26-3* expression drops in *Cdx* mutants (Figure 5o). Inhibition of RA action using BMS493 dampens the *Cdx* mutant phenotype, implying the phenotype is mediated by the RA pathway. BMS493 also down-regulates *Cyp26-3* expression consistent with a negative feedback loop between RA and *Cyp26* activity as previously shown in zebrafish (Dobbs-McAuliffe et al. 2004, Emoto et al. 2005, D'Aniello et al 2013).

Effects of *Cdx* and *Pdx* mutation on amphioxus transcriptome

To examine the downstream molecular consequences of *Pdx* and *Cdx* mutation in more detail, we sequenced replicate transcriptomes of mutant embryos or larvae soon after morphological phenotypes are clearly evident: 6 days for *Pdx*^{-/-}, 34 h and 42 h for *Cdx*^{-/-} (42 h analysed here, combined analysis in Supplementary Data). After mapping reads to an amphioxus supertranscriptome assembly, read counts were calculated to estimate gene expression levels.

Comparisons were made to transcriptome data from pools of mixed wild type and heterozygous siblings from the same crosses to identify genes that are differentially up- or down-regulated after mutation; these will include direct and indirect targets of *Cdx* and *Pdx* (Figure 7a,b; Supplementary Data).

We first asked which of the multiple amphioxus GFP-encoding genes (Bomati et al. 2009; Li et al. 2009) are affected in *Pdx* mutants. In *Pdx*^{-/-} mutants, we found 11 supercontigs encoding GFP were significantly down-regulated (1.8 to 8.3 fold). Ten derive from four highly similar, tandemly duplicated amphioxus *GFP* genes that cannot be distinguished using short read sequence data: *GFP-8*, *GFP-10*, *GFP-12* and *GFP-13* (nomenclature of Li et al. 2009; Table S5). We conclude that one or more of these closely related *GFP* genes were strongly down-regulated when regionalisation of the gut was disrupted by mutation of *Pdx* (Figure 7c).

Second, we used transcriptome data to ask if signalling pathways were disrupted by ParaHox gene mutation. In *Pdx*^{-/-} mutants, we detect significant changes to expression of genes encoding putative components of the insulin-signalling pathway, but not to the gut-expressed insulin-like peptide (*Ilp1*) gene itself (Figure 7c). We detect ~1.6 fold down-regulation of two insulin-like growth factor-binding protein-like (*IGFBP*) contigs and 1.4 fold up-regulation of an insulin-like peptide receptor (*ILPR*) contig (Figure 7c; Table S6). In *Cdx*^{-/-} mutants, we found down-regulation of eight supercontigs from the intracellular lipid-binding protein (*iLBP*) gene family, which in vertebrates includes genes encoding CRABP (cellular retinoic acid-binding protein), CRBP (cellular retinol-binding proteins) and FABP (Fatty acid-binding proteins). In amphioxus, these genes have undergone extensive independent duplication (Holland et al. 2008, Albalat et al. 2009). The contigs affected by *Cdx* mutation correspond to amphioxus *iLBP4* (4.2 to 13.5 fold down-regulation) and *iLBP6* (1.6 fold down-regulation; Fig 7d and Table S7). The biochemical activities of these genes are unclear; it is possible that one or both encode proteins that bind RA, inhibiting the RA pathway (Albalat et al. 2009). In contrast, we detect no significant change to expression levels of *FGF* or *Wnt* genes (Supplementary Data). Downregulation of *iLBP* genes in *Cdx* mutants suggests a second possible mechanism through which *Cdx* suppresses RA activity in the posterior of amphioxus: through positive regulation of iLBPs.

Third, we examined the effect of amphioxus *Cdx* mutation on Hox genes, because in vertebrates the role of *Cdx* genes in body axis elongation is mediated, at least in part, through activation of central (and some posterior) Hox genes (Pownall et al. 1996, Isaacs et al. 1998, van den Akker et al. 2002, Young et al. 2009). There is also evidence that vertebrate *Cdx* genes have repressive effects on the most anterior Hox genes, giving a colinear-like response across Hox clusters: in *Xenopus tropicalis*, perturbation of *Cdx* gene activity caused up-regulation of anterior Hox genes (paralogy groups 1 and 2) and down-regulation of Hox genes from paralogy group 5 to 10/11 (Marlétaz et al. 2015). In amphioxus, Hox genes differ greatly in expression intensity in normal embryos (Table S7, Figures S15 and S16; Marlétaz et al. 2018). Despite differences in absolute expression, we detect a colinear-like response of Hox genes to amphioxus *Cdx* gene activity. In amphioxus *Cdx*^{-/-} mutants, *Hox-1* expression is up-regulated (1.43x), *Hox2*, *Hox3* and *Hox4* are unaffected, *Hox5* and *Hox6* are mildly down-regulated, and *Hox7* is strongly down-regulated, although not each change is significant when considered in isolation (Figure 7e; Table S7 and Figure S17). This colinear-like response is consistent with amphioxus *Cdx* in normal development activating central and posterior Hox genes and repressing anterior Hox genes as part of an axis patterning system.

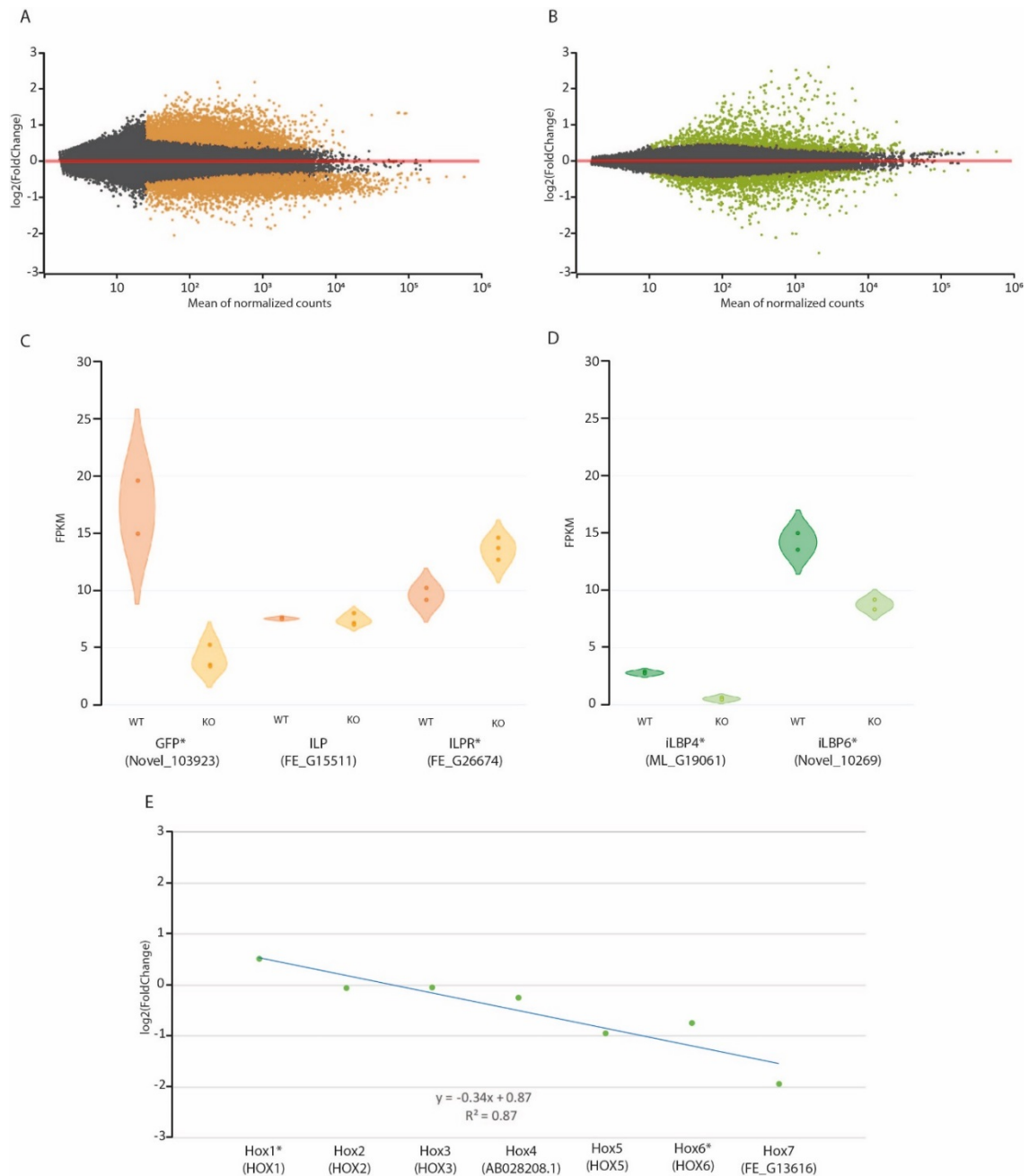


Figure 7: Transcriptome analysis of *Pdx* and *Cdx* mutant amphioxus embryos and larvae. (A) Fold change in expression (log₂ scale) between wild type and *Pdx* mutant larvae in relation to mean expression level (read counts). Coloured dots are supercontigs meeting significance criteria. (B) Fold change in expression (log₂ scale) between wild type and *Cdx* mutant embryos in relation to mean expression level (read counts). Coloured dots are supercontigs meeting significance criteria. (C) Expression level (FPKM) compared between wild type and *Pdx* mutant larvae (KO) for selected *GFP*, *ILP1* and *ILPR* supercontigs, showing down-regulation of *GFP* and up-regulation of *ILPR* in mutants. (D) Expression level compared between wild type and *Cdx* mutant embryos (KO) for selected *iLBP* supercontigs, showing down-regulation in mutants. (E) Colinear-like response in expression level fold change of Hox genes to *Cdx* mutation.

Pdx and *Cdx* mutation affects gut-associated gene expression

To investigate further the association between *Pdx* and *Cdx* genes and development of the gut, we tested whether gene sets affected by mutation include a high proportion of ‘gut-enriched’ genes. Using transcriptome data published for adult tissues of *B. lanceolatum* (Marlétaz et al. 2018), we defined genes as gut-enriched if expression level in gut was at least twice the expression level in seven out of eight other tissues (neural tube, muscle, gill bars, hepatic diverticulum, testis, ovary, skin, cirri; Supplementary Data). The 2083 gut-enriched genes were represented by 4705 contigs in our study; of these, 482 were differentially expressed in *Pdx*^{-/-} mutants and 218 were differentially regulated in *Cdx*^{-/-} mutants (Supplementary Data). This equates to 8.3% of the *Pdx* differentially expressed contigs and 15.3% of the *Cdx* differentially expressed contigs (Figure 8a).

To test if this represents enrichment, we ran 1000 simulations of each sampling (5831 or 1428 differentially expressed contigs) and assessed overlap with a dataset of 4705 contigs (the size of the gut-enriched dataset) chosen randomly from the supertranscriptome. Mean overlaps were 3.3%, with the experimental data being a highly significant outlier in each case (arrows in Figure 8b,c). Hence, mutation of *Pdx* or *Cdx* has a disproportionate and significant effect on expression of gut-enriched genes.

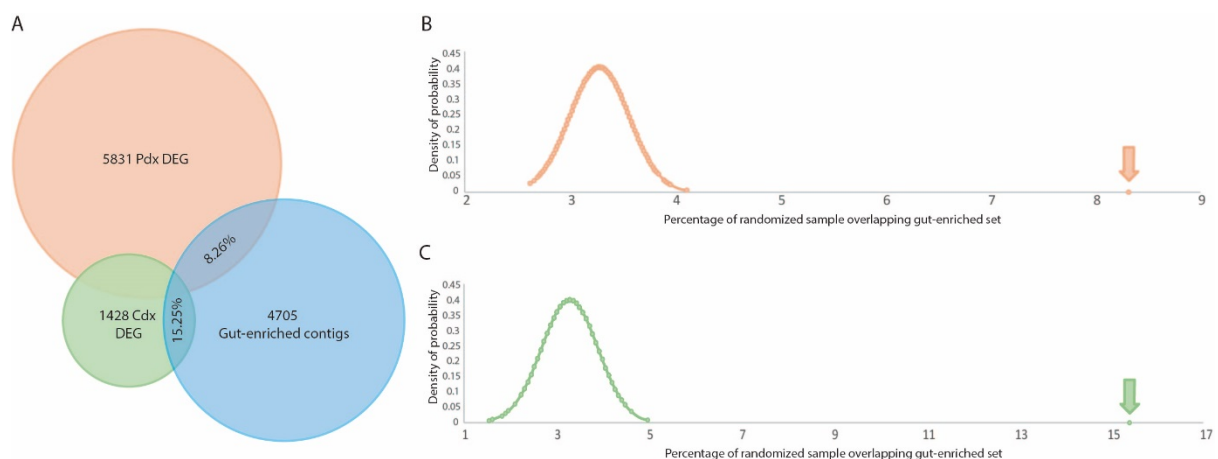


Figure 8. Mutation of amphioxus *Pdx* or *Cdx* has a disproportionate effect on gut-enriched genes. (A) The supercontigs that are expressed differentially between wild type and *Pdx* mutant larvae (orange), or between wild type and *Cdx* mutant embryos (green), overlap with the set of supercontigs classified as gut-enriched (blue). (B) The degree of overlap (orange and green arrows) is outside the range of overlap generated by random sampling (orange and green curves).

Discussion

The recent development of technologies for generating targeted mutations has great potential in comparative and evolutionary developmental biology. The most widely used technology is based CRISPR/cas, but this has not yet been applied successfully to cephalochordates. In contrast, TALENs have been used to generate inherited mutations in several genes in *B. floridae* (Li et al. 2014, 2017). The method is efficient and reproducible, although not straightforward because of the practical difficulties of rearing amphioxus from egg to adult, long generation times (4 to 6 months) and the lack of linked markers meaning that alleles must be tracked by PCR on sperm, embryos or tail clips. Here we describe successful introduction of germline mutations into two ParaHox genes, *Pdx* and *Cdx*, each predicted to generate frameshifts and production of ‘nonsense’ peptides lacking a homeodomain. Each produces a clear phenotype that can be related to known gene expression patterns. We cannot be certain that all are complete null mutations or if alternative start codon usage, perhaps at low frequency, generates a hypomorphic allele. This is a possibility for *Pdx* in particular since the mutant phenotype is subtle and not evident at embryonic stages. We argue that these mutations allow robust conclusions to be drawn about important roles of amphioxus *Pdx* and *Cdx* genes, although we may not uncover every role. Importantly, the choice of amphioxus widens the phylogenetic spread of taxa in which the function of ParaHox genes has been studied, permitting new evolutionary insights.

The widespread expression of *Pdx* and *Cdx* genes in the gut of bilaterian animals, localised to defined anteroposterior domains, is consistent with an ancient role for these genes in specifying correct region-specific development of the gut. We asked if mutation of amphioxus *Pdx* or *Cdx* causes cell fate changes in specific regions of the gut, notably the midgut for *Pdx* and anus for *Cdx*. One of the earliest regional markers in the endoderm of amphioxus is a gene encoding an insulin-like peptide, *Ilp1*, expressed in a broad domain spanning from the pharyngeal region back to the midgut from gastrula to neurula stages, and later in larvae especially in the pharyngeal region (Holland et al. 1997; Lecroisey et al. 2015). Amphioxus *Ilp1* is a pro-orthologue of vertebrate *insulin* and *IGF* genes (Chan et al. 1990; Holland et al. 1997; Lecroisey et al. 2015), and in vertebrates the *insulin* gene is a direct transcriptional target of *Pdx*. It seemed reasonable, therefore, to ask if *Ilp1* expression was altered in *Pdx*^{-/-} mutant amphioxus embryos. This is also possible because amphioxus *Pdx* and *ILP* genes are both expressed in midgut endoderm at neurula stages. We do note, however, that the expression domains are not precisely coincident, with the domain of *Ilp1* gene expression larger in anteroposterior extent and extending more rostrally. This differs from the situation in vertebrates, where early *Pdx* expression is necessary for formation of the pancreas and anterior duodenum, and the *Pdx*-positive domain includes the organ that will later express *insulin* (Jonsson et al. 1994; Offield et al. 1996). We found that mutation of amphioxus *Pdx* did not abolish *Ilp1* expression in neurulae, nor did it significantly alter *Ilp1* transcript abundance in larvae (other components of *insulin* signaling were affected in transcriptome analysis). It is possible there was undetected transformation of one set of *Ilp*-positive cells into other *Ilp*-positive cells in *Pdx* mutant embryos, or some other subtle change, but we could not assess this. We also stress that even if *Pdx* does not specify the *Ilp1* domain in embryos, this does not exclude the possibility of *Pdx* being a transcriptional regulator of the *Ilp1* gene at later stages or in adults, comparable to *Pdx* regulation of *insulin* in the adult vertebrate pancreas (Ohneda et al. 2000).

A later regional marker evident in the developing amphioxus endoderm is a neatly demarcated stripe of green fluorescence, first evident clearly in the midgut of 6-day larvae and persisting until at least 46 days of development. Confocal imaging confirmed this is not fluorescence from ingested algae as it is present inside endoderm cells. Previous studies involving UV-excitation of amphioxus larvae and adults have described green fluorescence primarily in ovaries and buccal cirri (Deheyn et al. 2007; Li et al. 2009), and this is attributed to the presence of multiple *GFP* genes (Bomati et al. 2009; Li et al. 2009). However, to our knowledge, the fluorescent stripe in larval gut has not been described previously, and it proved to be a fortuitous marker of endoderm regionalization. The most striking phenotype we observed in *Pdx*^{-/-} mutant amphioxus was absence of the GFP stripe. Transcriptome analysis reveals the GFP responsible for the stripe is encoded by one or more of an array of very similar *GFP* genes (*GFP-8*, *GFP-10*, *GFP-12*, *GFP-13*). We do not suggest these genes are necessarily under direct transcriptional control of *Pdx* protein; instead, we interpret loss of the GFP stripe, and lower abundance of *GFP* transcripts, as the consequence of incorrect cell fate specification in a localized region of endoderm. Since overall shape and size of the gut seems normal, we suggest these cells have been transformed to another endodermal cell type. The similar fluorescence in buccal cirri in mutants and wild type larvae indicates the phenotype is not a general failure of *GFP* gene expression.

A gut phenotype is also evident in amphioxus *Cdx* mutants, specifically the failure of the anus to open. Opening of the anus in amphioxus development must involve rearrangement of cell junctions within epithelial cell layers, such that the epithelial cells at the extreme posterior of the gut lessen their connections with some of their neighbours and form new junctions with epithelial cells of the epidermis. In this way, two nested epithelial cell layers (gut and epidermis) become contiguous, with the edges of the anus opening marking the boundary where the two layers fused. Mathematically, this marks a topological transition from a sphere to a torus (Jockusch and Dress 2003). Biologically, the propensity to break and reform epithelial cell junctions must be a property specific to the extreme posterior cells of the gut (and/or epidermis). Hence, loss of this property in *Cdx*^{-/-} mutant larvae likely reflects a change in cell fate specification, with terminal gut cells losing a region-specific character.

An independent assessment of the roles of *Pdx* and *Cdx* in gut development was made through transcriptome analysis. A category of genes we define as 'gut-enriched' were disproportionately represented among the (up and down) differentially-regulated genes in *Pdx*^{-/-} and *Cdx*^{-/-} mutant animals. The gut-enriched dataset was defined by examining adult transcriptomes, and includes genes encoding digestive enzymes and other proteins associated with gut functions. Although only 8.3% and 15.3% of *Pdx* and *Cdx* differentially expressed contigs fall in this category, the number of gut-associated genes affected will be higher than this since our definition of 'gut-enriched' is relatively strict. We therefore interpret the enrichment as further indication that correct development of the gut, with functional regionalisation, is severely perturbed in *Pdx*^{-/-} and *Cdx*^{-/-} mutants.

Thus, alterations to morphology, changes to positional markers and transcriptome analysis all support the hypothesis that *Pdx* and *Cdx* genes are essential for regional cell fate specification in the endoderm of the developing amphioxus gut. Combining with expression and functional data from other taxa, we argue that specification of correct cell fates in the middle and posterior of the gut has

been a function of *Pdx* and *Cdx* genes since the origin of Bilateria. This conclusion is consistent with the proposal that ParaHox genes played a role in the evolution of a through-gut, a key innovation of the bilaterians and a possible contributor to sedimentary mixing and animal diversification in the Cambrian (Brooke et al. 1998; Holland 2015).

We also asked if amphioxus *Pdx* and *Cdx* genes have roles in development of tissues outside the gut, because *Pdx* is expressed in the first Hesse organ to form in the neural tube and *Cdx* is expressed in all germ layers at the posterior, including epidermis and neurectoderm (Brooke et al. 1998). Hesse organ receptors are primary rhabdomic photoreceptors and express the amphioxus *melanopsin* (*Mop*) gene encoding the microvillar light transducing protein (Koyanagi et al. 2005, Pantzartzi et al. 2017). We find that mutation of the *Pdx* gene did not remove *Mop* expression from these cells, as assayed by in situ hybridisation, and the associated *Mitf*-positive pigment cell also formed. In summary, we do not detect cell fate changes in the amphioxus neural tube in *Pdx* mutant embryos and larvae. We cannot conclude there is no function in Hesse organ receptors and additional mutants would need to be analysed to test this. Similarly, our analyses of amphioxus *Cdx* mutants did not reveal clear homeotic or cell fate transformations in neural tube (or indeed mesoderm), although these would be hard to detect since there is little morphological distinction along the nerve cord or between somites for much of the body. It is possible there are subtle changes, especially since we detect shifts to Hox gene expression levels in a comparable manner to those observed in *Xenopus* embryos with disrupted *Cdx* function. In both *Xenopus* (Marlétaz et al. 2015) and amphioxus (this study), disruption of *Cdx* function has a colinear-like effect on Hox genes: the most 'anterior' or 3' Hox genes (paralogy group 1) respond in the opposite direction to middle or posterior Hox genes (paralogy group 5 to 7 in amphioxus, 5 to 11 in *Xenopus*). We do detect a change in cell fate in posterior epidermis in *Cdx* mutants, manifest as a smaller region of epidermis expressing the *Rootletin* gene. The protein product of this gene drives cell shape changes underpinning tail fin development (Flood 1975; Koop et al 2011; Mansfield and Holland 2013); hence, this change in epidermal cell fate has a direct effect on larval morphology.

The other major developmental process in which *Cdx* genes are implicated in a wide range of taxa is posterior growth of the body or axial elongation. This role is well characterised in vertebrates and is evident in some arthropods, although homology of the process between distant taxa is not proven. A possible role for amphioxus *Cdx* in tail extension cannot be deduced from expression pattern alone, because posterior expression could simply reflect the anus and tail fin roles discussed above. We found that mutation of the amphioxus *Cdx* gene dramatically disrupts growth of the body axis. The cessation of posterior growth we observe has some similarity in appearance to the consequence of disrupting *Cdx* function in vertebrate embryos, but is possibly less severe. However, we do detect some mechanistic similarities. The most important effectors through which *Cdx* genes control tail extension in vertebrates seem to be central Hox genes and the retinoic acid (RA), FGF and Wnt signaling pathways (Young et al. 2003; Savory et al. 2009; Amin et al. 2016). In amphioxus *Cdx* mutants, we did not detect effects on FGF or Wnt signaling components, but we do find clear effects on Hox gene expression and RA signaling. The role of vertebrate Hox genes in axial extension is complex. Although mouse Hox gene mutants do not generally have a tail phenotype (apart from *Hoxb13*; Economides et al. 2003), ectopic expression of central genes, such as *Hoxa5* or *Hoxb8*, can partially rescue the tail truncation phenotype of *Cdx* mutants (Young et al. 2009). An opposite effect was found for *Hoxa13*, *Hoxb13* and *Hoxc13* suggesting a role for the most posterior Hox genes in tail

growth termination (Young et al. 2009; Aires et al. 2019). Furthermore, there is evidence that *Cdx2* is a direct activator of anterior and central Hox genes and these likely stimulate axial growth (Amin et al. 2016). In amphioxus, we do not have definitive evidence for Hox gene involvement in tail development, but the activation of middle Hox genes by *Cdx* is consistent with this possibility and a striking similarity to vertebrates.

There are similarities between amphioxus and vertebrate *Cdx* genes with respect to RA signaling in the tail. A role for RA (or rather lack of RA) in vertebrate tail extension is well known (Ruiz i Altaba and Jessell 1991; Herrmann 1995; Padmanabhan 1998). Similarly, addition of ectopic RA causes axis truncation in amphioxus larvae implying low posterior RA levels are necessary for posterior growth (Koop et al. 2011). Here we show that *Cdx* is upstream of RA activity in amphioxus, as it is in vertebrates, as indicated by the finding that inhibition of RA action partially rescues the axial truncation phenotype, and the reduced tail fin, of *Cdx* mutants. There are several potential molecular mechanisms by which *Cdx* genes could suppress RA action. Young et al. (2009) found that mutation of mouse *Cdx2* and *Cdx4* genes causes down-regulation of the gene encoding an RA-clearing enzyme *Cyp26A1*. Savory et al. (2009) found this is a direct transcriptional effect and impacts RA signaling. Similarly, we find down-regulation of expression of a *Cyp26* gene in amphioxus *Cdx* mutants, and show this is also a direct transcriptional effect. The inference is that in both mice and amphioxus, *Cdx* genes suppress RA signaling in the tail through the same mechanism, activation of *Cyp26* expression, permitting axis extension. There may be a second route by which amphioxus *Cdx* dampens posterior RA signaling, through positive regulation of genes encoding putative RA-binding proteins (*iLBP4* and *iLBP6*), although currently it is unclear if these proteins bind RA or another molecule. The similarities lead us to conclude that the role of *Cdx* in axial extension is homologous between amphioxus and vertebrates, and therefore dates back at least to the base of Chordata. The function is possibly older, dating to the origin of Bilateria, although this conclusion is more tentative since there is less evidence of mechanistic similarity in arthropods.

Conclusions

- We have generated stable lines carrying disabling frameshift mutations in the *Pdx* and *Cdx* homeobox genes of amphioxus *Branchiostoma floridae*.
 - Homozygous *Pdx* and *Cdx* mutants have defects in developing midgut and posterior gut respectively. These defects are interpreted as changes to region-specific cell fate and are evident in transcriptomes as quantitative changes to expression of gut-enriched genes. *Cdx* mutation also affects Hox gene expression and patterning of posterior epidermis.
 - These data are consistent with the hypothesis that *Pdx* and *Cdx* have conserved roles in gut patterning. The emergence of these roles may have been instrumental in the evolution of a through-gut facilitating efficient feeding and burrowing, contributing to the evolutionary radiation of bilaterians.
 - Homozygous *Cdx* mutants have severely truncated posterior growth, with this axial elongation defect probably caused by down-regulation of *Cyp26* leading to increased retinoic acid signalling. Mechanistic similarity to vertebrates indicates a role for the *Cdx* in axis extension dates at least the base of chordates and possibly earlier
-

Acknowledgements

We thank Enrico D'Aniello, Salvatore D'Aniello, Jr-Kai Yu, José Luis Gómez-Skarmeta, Sebastian Shimeld, Rodrigo Pracana, Ricard Albalat and Thurston Lacalli for helpful advice and discussion. This research was funded by the Elizabeth Hannah Jenkinson fund, QR Global Challenges Research Fund to the University of Oxford, National Natural Science Foundation of China (No. 31872186), Fundamental Research Funds for the Central Universities, China (No. 20720160056), Ministry of Science, Innovation and Universities, Spanish Government (BFU2017-86152-P) and C.H.-U. holds a predoctoral FPI contract (Ministry of Science, Innovation and Universities, Spanish Government).

Methods

Amphioxus culture and targeted mutation

Amphioxus (*Branchiostoma floridae*) were obtained from a stock maintained by Jr-Kai Yu originating from Tampa, Florida. Cultures were maintained in Xiamen University under previously described conditions (Li et al. 2012). Gametes were obtained using thermal-shock (20°C to 26°C) following Li et al. (2013); fertilization and culturing of embryos at 26 °C was carried out as described by Liu et al. (2013). The TALEN method was used to generate *Pdx* and *Cdx* mutants with TALEN pairs designed to target coding sequence (Figures S1-S6). TALEN construct assembly, mRNA synthesis and mutation efficacy assays were conducted following the methods described in Li et al. (2014). Mosaic founder animals were spawned to generate F1 heterozygotes, using PCR and sequencing to detect, characterize and follow mutant alleles (Li et al., 2017; Supplementary Information, Table S1). Homozygous mutants were generated by crossing heterozygous animals.

Whole-mount in situ hybridization (WMISH)

Coding sequences of *Cdx*, *Pdx*, *Rootletin*, *Ilp1* and *Cyp26* genes were amplified from cDNA libraries from amphioxus embryos using primers listed in Supplementary Information. PCR products were cloned into the pGEM-T Easy vector (Promega, USA) and confirmed by DNA sequencing, linearized, cleaned using phenol-chloroform, and used for templated synthesis of digoxigenin-labelled antisense RNA probes using T7 or SP6 RNA polymerase (Promega, USA). Embryos and larvae were fixed in 4% paraformaldehyde in MOPS buffer at 4°C for 24 hours and stored in 70% ethanol at -20°C. Hybridisation and detection was performed as previously described (Yu et al., 2009). To ascertain the genotype of embryos after WMISH, an extra 30 min wash in 500 ml filtered sea water or PBS was performed to remove fixative before processing for hybridisation.

Scanning electron microscopy (SEM) and GFP detection

SEM used an adaptation of the methods of Inoué and Osatake (1988). In brief, embryos were fixed in 2.5% glutaraldehyde in PBS at 4°C overnight, washed 3 x 10 min in 0.1M PBS (pH7.4), and transferred to 100% ethanol through a graded ethanol series. Specimens were then washed 5 x 10 min in 100% tertiary butyl alcohol and stored at 4°C overnight. Samples were dried in a vacuum freeze dryer, placed on conductive tape, sprayed with platinum and observed under JSM-6390 scanning electron microscope. For GFP detection, amphioxus larvae were mounted in 1% methylcellulose in sea

water and photographed under an SZX10 fluorescent stereoscope (Olympus, Japan) or an LSM 780 confocal microscope (Zeiss, Germany).

BMS493 treatment

The retinoic acid antagonist BMS493 (Sigma Aldrich) was dissolved in DMSO as a 1mM stock solution. Stock solution was diluted in filtered seawater to 1 μ M and applied to amphioxus embryos from 5 hours post-fertilization onwards (early gastrula stage). Control embryos were treated with filtered seawater containing an equal amount of DMSO. Most embryos were washed and fixed at 16 or 18 hours post-fertilization for in situ hybridization; others were continuously cultured in BMS493 until 30 hours post-fertilization for morphological observation.

Reporter gene assays

A 3.1 kb region 5' of *Cyp26-3* was cloned into pGL3 (Promega) and modified by PCR to generate three mutant versions with altered Cdx-binding sites (Supplementary Information). Solutions containing 3 ng/ μ L Renilla luciferase vector pRL-TK (Promega), 20% glycerol, 5 mg/ml Texas Red Dextran, with or without 30 ng/ μ L of one of the *Cyp26-3* luciferase constructs, were microinjected into unfertilized *B. floridae* eggs as previously described (Liu et al. 2013). For each experiment, ~60 embryos were collected and assayed at 16 hours post fertilization; uninjected embryos from the same batch were used as negative control. Levels of luciferase and Renilla were detected with the Dual Luciferase Kit (Promega) using a GloMax luminometer with an integration of 10 seconds; the level of luciferase was normalized to Renilla activity for each experiment. All experiments were repeated three times.

Embryo genotyping

Genotyping of live embryos was performed as described in Li G, et al. (2017), and genotyping of embryos fixed with 4% PFA-MOPS-EGTA or analyzed by whole-mount in situ hybridization was conducted with the same protocol except that an extra 30-min wash in 500 ml filtered sea water or PBS was added before lysis, to remove the fixative.

RNA sequencing

Embryos and larvae obtained from crosses between *Pdx*^{+/-} 4 Δ x *Pdx*^{+/-} 11 Δ and *Cdx*^{+/-} 7 Δ x *Cdx*^{+/-} 7 Δ were sorted by visual phenotype at 6 days post-fertilization for *Pdx*, and 34 h or 42 h post-fertilization for *Cdx*. Samples for RNA analysis were pools of ~20 embryos or larvae. Each cross gave two matched sibling pools: 'mutant' individuals with a morphological phenotype (inferred *Pdx* 4 Δ /11 Δ and *Cdx* 7 Δ /7 Δ) and 'control' individuals without clear phenotype (mixture of heterozygotes and wild type). RNA sequencing was performed by BGI (Shenzhen, China) on the BGISEQ platform, >60 million 100 nt paired end reads per sample, using mRNA enrichment, random priming reverse transcription and low PCR cycle number. Reads were mapped using STAR v2.7.0 (Dobin et al. 2013) using --twopassmode Basic and --outSJfilterOverhangMin 12 12 12 12 options (Davidson et al. 2017) to a *B. floridae* supertranscriptome of 215,495 contigs assembled de novo from independent RNA-seq reads (45.2 to 52.4 million mapped reads per sample); genes may be represented by more than one contig. Hox genes were assembled manually to remove and correct an artefactual fusion between 6 distinct Hox genes. Since supertranscripts can join exons not normally spliced together, STAR aligner settings were adjusted to permit mapping near non-canonical splice junctions (STAR --

quantMode GeneCounts --twopassMode Basic --outSJfilterOverhangMin 12 12 12 12). STAR mapped 91% and 90% of reads from the *Pdx* and *Cdx* experiments. Reads were quantified using featureCounts (Liao et al. 2014) in the Subread package v1.6.3 allowing multiple mapping, thereby permitting analysis of duplicate genes (featureCounts -O -M -p -B -fraction).

Differential Gene Expression analysis

Analysis of differentially expressed genes (DEGs) used DESeq2 v3.8 (Love et al. 2014) including Principal Component Analysis to test for outlier samples and batch effects. In the *Cdx* experiment, two control and two mutant 42 h samples grouped distinctly from one control and one mutant 34 h sample; two DEG analyses were therefore performed, 42h alone and 34h/42h combined (Supplementary Information). In the *Pdx* experiment, one control sample (WT2) grouped aberrantly and was excluded from DEG analysis which used 3 mutant and 2 control samples (Supplementary Information). For a contig to be considered differentially expressed, we required expression change of $>0.5 \log_2$ (Fold Change) and adjusted p-value <0.05 , plus absolute expression level of >2 fpkm in at least one condition (Marlétaz et al. 2015). To assess accuracy of embryo sorting, raw reads matching mutant or wild type allele sequences were counted. Embryo pools classed as *Cdx*^{-/-} had 4.3 to 5.3% of *Cdx* 'wild type' reads, suggesting that 1 or 2 embryos in each mutant pool of 20 were heterozygous; the control pool had 38 to 70% wild type reads, consistent with a mix of heterozygous and genetic wild type embryos. The same method could not be applied accurately to *Pdx* mutants because control pools had predominantly wild type reads (91 to 100%), implying down-regulation of mutant allele expression in heterozygotes. This obviates applicability of read counts for assessing heterozygote number. Gut-enriched genes were identified from published *B. laneceolatum* data (Marlétaz et al. 2018; NCBI GEO GSE106430) as genes with higher mean expression level in gut than any other adult tissue (eggs and embryos excluded) and at least double expression level in gut than in 7 out of 8 other adult tissues; the 2083 genes were matched to contigs in the current study using blastn with an e-value cut off of 1^{e-70} giving 4705 gut-enriched supercontigs.

References

- Aires R, de Lemos L, Nóvoa A, Jurberg AD, Mascrez B, Duboule D, Mallo M (2019) Tail bud progenitor activity relies on a network comprising Gdf11, Lin28, and Hox13 genes. *Developmental Cell* 48: 383-395
- Albalat R, Brunet F, Laudet V, Schubert M (2011) Evolution of retinoid and steroid signaling: vertebrate diversification from an amphioxus perspective. *Genome Biol Evol* 3: 985–1005
- Amin S, Neijts R, Simmini S, van Rooijen C, Tan SC, Kester L, van Oudenaarden A, Creyghton MP, Deschamps J (2016) Cdx and T Brachyury co-activate growth signaling in the embryonic axial progenitor niche. *Cell Reports* 17: 3165-3177
- Annunziata R, Martinez P, Arnone MI (2013) Intact cluster and chordate-like expression of ParaHox genes in a sea star. *BMC Biol.* 11: 68
- Annunziata R, Arnone MI (2014) A dynamic regulatory network explains ParaHox gene control of gut patterning in the sea urchin. *Development* 141: 2462-2472
- Beck F, Erler T, Russell A, James R (1995). Expression of Cdx-2 in the mouse embryo and placenta: possible role in patterning of the extra-embryonic membranes. *Devel. Dyn.* 204: 219 -227
- Beck F, Chawengsaksophak K, Waring P, Playford RJ, Furness JB (1999) Reprogramming of intestinal differentiation and intercalary regeneration in *Cdx2* mutant mice. *Proc Natl Acad Sci USA* 96: 7318-7323
- Bomati EK, Manning G, Deheyn DD (2009) Amphioxus encodes the largest known family of green fluorescent proteins, which have diversified into distinct functional classes. *BMC Evol Biol* 9: 77
- Brooke NM, Garcia-Fernández J, Holland PWH (1998) The ParaHox gene cluster is an evolutionary sister of the Hox gene cluster. *Nature* 392: 920–922
- Carvalho JE, Theodosiou M, Chen J, Chevret P, Alvarez S, De Lera AR, Laudet V, Croce JC, Schubert M (2017) Lineage-specific duplication of amphioxus retinoic acid degrading enzymes (CYP26) resulted in sub-functionalization of patterning and homeostatic roles. *BMC Evol Biol* 17: 24
- Chan SJ, Cao QP, Steiner DF (1990) Evolution of the insulin superfamily: cloning of a hybrid insulin/insulin-like growth factor cDNA from amphioxus. *Proc Natl Acad Sci USA* 87: 9319-9323
- Charité J, de Graaff W, Consten D, Reijnen MJ, Korving J, Deschamps J (1998) Transducing positional information to the Hox genes: critical interaction of cdx gene products with position-sensitive regulatory elements. *Development* 125: 4349-4358
- Chawengsaksophak K, de Graaff W, Rossant J, Deschamps J, Beck F (2004) *Cdx2* is essential for axial elongation in mouse development. *Proc Natl Acad Sci USA* 101: 7641-7645
- Cole AG, Rizzo F, Martinez P, Fernandez-Serra M, Arnone MI (2009) Two ParaHox genes, *SpLox* and *SpCdx*, interact to partition the posterior endoderm in the formation of a functional gut. *Development* 136: 541-549
- Copf T, Schröder R, Averof M (2004) Ancestral role of caudal genes in axis elongation and segmentation. *Proc Natl Acad Sci USA* 101: 17711-17715

- D'Aniello E, Rydeen AB, Anderson JL, Mandal A, Waxman JS (2013) Depletion of retinoic acid receptors initiates a novel positive feedback mechanism that promotes teratogenic increases in retinoic acid. *PLoS Genetics* 9: e1003689.
- Davidson NM, Hawkins AD, Oshlack A (2017) SuperTranscripts: a data driven reference for analysis and visualisation of transcriptomes. *Genome Biol* 18: 148
- Deheyn DD, Kubokawa K, McCarthy JK, Murakami A, Porrachia M, Rouse GW, Holland ND (2007) Endogenous green fluorescent protein (GFP) in amphioxus. *Biol Bulletin* 213: 95-100
- Dobbs-McAuliffe B, Zhao Q, Linney E (2004) Feedback mechanisms regulate retinoic acid production and degradation in the zebrafish embryo. *Mech Devel* 121: 339–350
- Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR (2013) STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 29: 15-21
- Edgar LG, Carr S, Wang H, Wood WB (2001) Zygotic expression of the caudal homolog pal-1 is required for posterior patterning in *Caenorhabditis elegans* embryogenesis. *Devel Biol* 229: 71-88
- Economides KD, Zeltser L, Capecchi MR (2003) *Hoxb13* mutations cause overgrowth of caudal spinal cord and tail vertebrae. *Devel Biol* 256: 317-330.
- Emoto Y, Wada H, Okamoto H, Kudo A, Imai Y (2005) Retinoic acid-metabolizing enzyme Cyp26a1 is essential for determining territories of hindbrain and spinal cord in zebrafish. *Devel Biology* 278: 415–427
- Faas I, Isaacs HV (2009) Overlapping functions of *Cdx1*, *Cdx2*, and *Cdx4* in the development of the amphibian *Xenopus tropicalis*. *Devel Dynamics* 238: 835-852
- Ferrier DEK, Dewar K, Cook A, Chang JL, Hill-Force A, Amemiya C (2005) The chordate ParaHox cluster. *Current Biology* 15: R820-822
- Flood PR (1975) Fine structure of the notochord of amphioxus. *Symp Zoolog Soc Lond* 36: 81-104
- Fröbuis AC, Seaver EC (2006) ParaHox gene expression in the polychaete annelid *Capitella* sp. I. *Devel Genes Evol* 216: 81-88
- Garcia-Fernández J (2005) The genesis and evolution of homeobox gene clusters. *Nature Rev Gen* 6: 881-892
- Garstang M, Ferrier DEK (2013) Time is of the essence for ParaHox gene clustering. *BMC Biology* 11: 72
- Grapin-Botton A, Majithia AR, Melton DA (2001) Key events of pancreas formation are triggered in gut endoderm by ectopic expression of pancreatic regulatory genes. *Genes Devel* 15: 444-454
- Hejnal A, Martín-Durán JM (2015) Getting to the bottom of anal evolution. *Zoologischer Anzeiger* 256: 61-74
- Herrmann K (1995) Teratogenic effects of retinoic acid and related substances on the early development of the zebrafish (*Brachydanio rerio*) as assessed by a novel scoring system. *Toxicology in Vitro* 9: 267-283
- Holland AM, Garcia S, Naselli G, MacDonald RJ, Harrison LC (2013) The Parahox gene *Pdx1* is required to maintain positional identity in the adult foregut. *Int J Devel Biol* 57: 391-398

- Holland LZ et al. (2008) The amphioxus genome illuminates vertebrate origins and cephalochordate biology. *Genome Res* 18: 1100-1111
- Holland ND, Holland, LZ (1993) Embryos and larvae of invertebrate deuterostomes. In: Essential Developmental Biology: a Practical Approach (Stern CD, Holland PWH, eds.) IRL Press, Oxford. pp 21-32
- Holland PWH (2015) Did homeobox gene duplications contribute to the Cambrian explosion? *Zoological Lett* 1: 1
- Holland PWH, Patton SJ, Brooke NM, Garcia- Fernandez J (1997) Genetic patterning of the ectoderm and endoderm in amphioxus: from homeobox genes to hormones. In: Advances in Comparative Endocrinology, Proceedings of the 13th International Congress on Comparative Endocrinology (Kawashima S, Kikuyama S, eds.), pp. 247–252
- Houle M, Sylvestre JR, Lohnes D (2003) Retinoic acid regulates a subset of *Cdx1* function in vivo. *Development* 130: 6555-6567
- Hunter CP, Harris JM, Maloof JN, Kenyon C (1999) Hox gene expression in a single *Caenorhabditis elegans* cell is regulated by a *caudal* homolog and intercellular signals that inhibit wnt signaling. *Development* 126: 805-814
- Ikuta T, Chen Y-C, Anunziata R, Ting H-C, Tung CH, Koyanagi R, Tagawa K, Humphreys T, Fujiyama A, Saiga H, Satoh N, Yu J-K, Arnone MI, Su Y-H (2013) Identification of an intact ParaHox cluster with temporal colinearity but residual spatial colinearity in the hemichordate *Ptychodera flava*. *BMC Evol Biol*. 13: 129
- Illes JC, Winterbottom E, Isaacs HV (2009) Cloning and expression analysis of the anterior ParaHox genes *Gsh1* and *Gsh2* from *Xenopus tropicalis*. *Devel Dynamics* 238: 194-203
- Inoué T, Osatake H (1988) A new drying method of biological specimens for scanning electron microscopy: the t-butyl alcohol freeze-drying method. *Arch Histol Cytol* 51: 53-59
- Isaacs HV, Pownall ME, Slack JM (1998) Regulation of Hox gene expression and posterior development by the *Xenopus caudal* homologue *Xcad3*. *EMBO Journal* 17: 3413–3427
- Iulianella A, Beckett B, Petkovich M, Lohnes D (1999) A molecular basis for retinoic acid-induced axial truncation. *Devel Biology* 205: 33-48
- Jockusch H, Dress A (2003) From sphere to torus: a topological view of the metazoan body plan. *Bulletin Math Biol* 65: 57-65
- Jonsson J, Carlsson L, Edlund T, Edlund H (1994) Insulin-promoter-factor 1 is required for pancreas development in mice. *Nature* 371: 606-609
- Katsuyama Y, Sato Y, Wada S, Saiga H (1999) Ascidian tail formation requires *caudal* function. *Devel Biol* 213: 257-268
- Koyanagi M, Kubokawa K, Tsukamoto H, Shichida Y, Terakita A (2005) Cephalochordate melanopsin: evolutionary linkage between invertebrate visual cells and vertebrate photosensitive retinal ganglion cells. *Current Biol* 15: 1065-1069

- Koop D, Holland LZ, Setiamarga D, Schubert M, Holland ND (2011) Tail regression induced by elevated retinoic acid signaling in amphioxus larvae occurs by tissue remodeling, not cell death. *Evolution Devel* 13: 427-435.
- Lacalli T, Stach T (2016) Acrania (Cephalochordata). In: Structure and Evolution of Invertebrate Nervous Systems (Schmidt-Rhaesa A, Harzsch S, Purschke G eds.) Oxford University Press, Oxford. pp 719-728.
- Lecroisey C, Le Pétillon Y, Escriva H, Lammert E, Laudet V (2015) Identification, evolution and expression of an insulin-like peptide in the cephalochordate *Branchiostoma lanceolatum*. *PLoS ONE* 10: e0119461
- Li G, Zhang QJ, Zhong J, Wang YQ (2009) Evolutionary and functional diversity of green fluorescent proteins in cephalochordates. *Gene* 446: 41-49
- Li G, Yang X, Shu Z, Chen X, Wang Y (2012) Consecutive spawnings of Chinese amphioxus, *Branchiostoma belcheri*, in captivity. *PLoS ONE* 7: e50838
- Li G, Shu Z, Wang Y (2013) Year-round reproduction and induced spawning of Chinese amphioxus, *Branchiostoma belcheri*, in laboratory. *PLoS ONE* 8: e75461
- Li G, Feng J, Lei Y, Wang J, Wang H, Shang L-K, Liu D, Zhao H, Zhu Y, Wang Y (2014) Mutagenesis at specific genomic loci of amphioxus *Branchiostoma belcheri* using TALEN method. *J Genet Genomics* 41: 215-219
- Li G, Liu X, Xing C, Zhang H, Shimeld SM, Wang Y (2017) Cerberus-Nodal-Lefty-Pitx signaling cascade controls left-right asymmetry in amphioxus. *PNAS* 14(14):3684-3689
- Liu X, Li G, Feng J, Yang X, & Wang YQ (2013) An efficient microinjection method for unfertilized eggs of Asian amphioxus *Branchiostoma belcheri*. *Dev Genes Evol* 223(4):269-278.
- Liao Y, Smyth GK, Shi W (2013) featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* 30: 923-930
- Lohnes D (2003) The *Cdx1* homeodomain protein: an integrator of posterior signaling in the mouse. *BioEssays* 25: 971-980
- Love MI, Huber W, Anders S (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* 15: 550
- Marlétaz F, Maeso I, Faas L, Isaacs HV, Holland PWH (2015) Cdx ParaHox genes acquired distinct developmental roles after gene duplication in vertebrate evolution. *BMC Biology* 13: 56
- Marlétaz F et al. (2018) Amphioxus functional genomics and the origins of vertebrate gene regulation. *Nature* 564: 64-70
- Mansfield JH, Holland ND (2015) Amphioxus tails: source and fate of larval fin rays and the metamorphic transition from an ectodermal to a predominantly mesodermal tail. *Acta Zoologica* 96: 117-125
- Moreno E, Morata G (1999) *Caudal* is the Hox gene that specifies the most posterior *Drosophila* segment. *Nature* 400: 873-877
- Mulley JF, Chiu C-H, Holland PWH (2006) Breakup of a homeobox cluster after genome duplication in teleosts. *PNAS* 103: 10369-10372

- Ohneda K, Ee H, German M (2000) Regulation of *insulin* gene transcription. *Semin Cell Devel Biol* 11: 227-233
- Offield MF, Jetton TL, Labosky PA, Ray M, Stein RW, Magnuson MA, Hogan BLM, Wright CVE (1996) *PDX-1* is required for pancreatic outgrowth and differentiation of the rostral duodenum. *Development* 122: 983-995
- Osborne PW, Benoit G, Laudet V, Schubert M, Ferrier DEK (2009) Differential regulation of ParaHox genes by retinoic acid in the invertebrate chordate amphioxus (*Branchiostoma floridae*). *Devel Biology* 327: 252-262
- Padmanabhan R (1998) Retinoic acid-induced caudal regression syndrome in the mouse fetus. *Reprod. Toxicol.* 12: 139-151
- Pantartzzi CN, Pergner J, Kozmikova I, Kozmik Z (2017) The opsin repertoire of the European lancelet: a window into light detection in a basal chordate. *Int J Devel Biology* 61: 763-772
- Pownall ME, Tucker AS, Slack JM, Isaacs HV (1996) eFGF, *Xcad3* and Hox genes form a molecular pathway that establishes the anteroposterior axis in *Xenopus*. *Development* 122: 3881-3892
- Ruiz i Altaba A, Jessell T (1991) Retinoic acid modifies mesodermal patterning in early *Xenopus* embryos. *Genes & Devel* 5: 175-187
- Thatcher JE, Isoherranen N (2009) The role of CYP26 enzymes in retinoic acid clearance. *Expert Opinion on Drug Metab & Tox* 5: 875-886
- Samadi L, Steiner G (2010) Conservation of ParaHox genes' function in patterning of the digestive tract of the marine gastropod *Gibbula varia*. *BMC Devel Biol* 10: 74
- Savory JG, Bouchard N, Pierre V, Rijli FM, De Repentigny Y, Kothary R, Lohnes D (2005) *Cdx2* regulation of posterior development through non-Hox targets. *Development* 136: 4099-4110
- Schwitzgebel VM, Mamin A, Brun T, Ritz-Laser B, Zaiko M, Maret A, Jornayvaz FR, Theintz GE, Michielin O, Melloul D, Philippe J (2003) Agenesis of human pancreas due to decreased half-life of insulin promoter factor 1. *J Clin Endocrinology & Metab* 88: 4398-4406
- Shimizu T, Bae YK, Muraoka O, Hibi M (2005) Interaction of Wnt and *caudal*-related genes in zebrafish posterior body formation. *Devel Biol* 279: 125-141
- Shinmyo Y, Mito T, Matsushita T, Sarashina I, Miyawaki K, Ohuchi H, Noji S (2005) *caudal* is required for gnathal and thoracic patterning and for posterior elongation in the intermediate-germband cricket *Gryllus bimaculatus*. *Mech Devel* 122: 231-239
- Stoffers DA, Zinkin NT, Stanojevic V, Clarke WL, Habener JF (1997) Pancreatic agenesis attributable to a single nucleotide deletion in the human IPF1 gene coding sequence. *Nature Genet* 15: 106-110
- Subramanian V, Meyer BI, Gruss P (1995) Disruption of the murine homeobox gene *Cdx1* affects axial skeletal identities by altering the mesodermal expression domains of Hox genes. *Cell* 83: 641-653.
- van den Akker E, Forlani S, Chawengsaksophak K, de Graaff W, Beck F, Meyer BI, Deschamps J (2002) *Cdx1* and *Cdx2* have overlapping functions in anteroposterior patterning and posterior axis elongation. *Development* 129: 2181-2193
- Wang X et al. (2018) Genome-wide analysis of PDX1 target genes in human pancreatic progenitors. *Molecular Metab* 9: 57-68

Willey A (1894) *Amphioxus and the Ancestry of the Vertebrates*. Macmillan, London

Wollesen T, Monje SV, McDougall C, Degnan BM, Wanninger A (2015) The ParaHox gene *Gsx* patterns the apical organ and central nervous system but not the foregut in scaphopod and cephalopod mollusks. *EvoDevo* 6: 41

Wu LH, Lengyel JA (1998) Role of *caudal* in hindgut specification and gastrulation suggests homology between *Drosophila* amnioproctodeal invagination and vertebrate blastopore. *Development* 125: 2433-2442

Young T, Rowland JE, van de Ven C, Bialecka M, Novoa A, Carapuco M, van Nes J, de Graaff W, Duluc I, Freund JN, Beck F (2009) *Cdx* and *Hox* genes differentially regulate posterior axial growth in mammalian embryos. *Developmental Cell* 17: 516-526

Yu JK, Meulemans D, McKeown SJ, Bronner-Fraser M (2008) Insights from the amphioxus genome on the origin of vertebrate neural crest. *Genome Res* 18: 1127-1132

Yu JK, Holland LZ (2009) Amphioxus whole-mount in situ hybridization. *Cold Spring Harb Protoc* 4: 1-6 doi:10.1101/pdb.prot5286

SUPPLEMENTARY INFORMATION

**Mutation of amphioxus *Pdx* and *Cdx* demonstrates
conserved roles for ParaHox genes in gut, anus and tail patterning**

SECTION 1: GENE STRUCTURES OF AMPHIOXUS *PDX* AND *CDX* GENES

SECTION 2: GENERATION OF MUTATIONS IN AMPHIOXUS *PDX* AND *CDX* GENES

SECTION 3: CLONING GENES FOR IN SITU HYBRIDISATION PROBES

SECTION 4: MORPHOLOGY OF *PDX* MUTANT AMPHIOXUS

SECTION 5: MORPHOLOGY OF *CDX* MUTANT AMPHIOXUS

SECTION 6: REPORTER GENE ANALYSIS OF *CYP26-3*

SECTION 7: DIFFERENTIAL EXPRESSION OF CANDIDATE TARGET GENES

SECTION 1: GENE STRUCTURES OF AMPHIOXUS *PDX* AND *CDX* GENES

(a) Verification of *Pdx* gene structure

The *Branchiostoma floridae Pdx (Xlox)* gene is comprised of two coding exons, confirmed here by aligning an assembled *Pdx* transcript sequence to the previously reported sequence of PAC clone 33B4 (GenBank AC129948.3; Ferrier et al. 2005); Figure S1.



Figure S1: *B. floridae Pdx* open reading frame and gene structure.

(b) Verification of *Cdx* gene structure

The *B. floridae Cdx* gene is generally depicted as comprised of two coding exons. However, we assembled a *B. floridae Cdx* transcript that aligned to three regions of the sequenced genomic PAC clone 36D2 (NCBI GenBank AC129947.4; Ferrier et al. 2005); Figure S2. We propose the small second

exon is used differentially and note it is also present in predicted isoforms X1 and X2 (GenBank XP_019625318.1 and XP_019625319.1), but not X3 (XP_019625325.1), of *B. belcheri* Cdx.

Complete predicted *B. floridae* Cdx open reading frame (homeodomain in bold)

MYRHPSQGSYNLNPYNYATAHPAYPAEYGYQVPPAVNAGENLQQTAAAAAWQSAAAFGSHGAGQRPEEWD
GRGYNCTAGTGLTAGPTGSCTAFPGMDYPVPVGAIQANSPAVSGVTTNSTNSQRPQH SRNPYDWMRKS NYS
TSPPPVL SVRGMPPQGRKDGGRCEILGPDGKTRT**KDKYRVVYSDHQRL E L E K E F Y S N K Y I T I K R K V Q L A N E**
L G L S E R Q V K I W F Q N R R A K Q R K M A K R K E L Q H P G G Q G G S D D G G G V M G E V S T L T V G P P P H Q L T L N P S G V A A S T L
SNPALPPSSSPLMTSAMTHAVTL PSCVPSS

Cdx open reading frame

ATGTACCGTCACCCCTCCCAGGGCAGCTACAACCTGAACCCGTACAACCTACGCCACGGCGCACCCCTGCCTA
CCCCGCGGAGTACGGACAGTACCAGGTCCCGCCTGCCGTCAACGCCGGCGAGAACCTACAGCAGACGGCCG
CCGCCCGCGTGGCAGTCCGCCGAGCCTTCGGCTCGCACGGGGCCGGACAGAGGCCAGAGGAATGGGAC
GGTCGCGGGTACAACCTGCACGGCGGGGACCGGGCTGACCGCCGGCCCGACCGGGTCTGTACAGCCTTCCC
CGGGATGGACTACCTGTCCCCGTGGTGCATCCAGGCCAACAGCCCTGCCGTGTGGGAGTGACGACCA
ACTCTACCAACAGTCAGAGACCACAGCACAGCAGAAATCCGTACGACTGGATGAGGAAAAGCAACTACTCC
ACAAGTCCTCCCCAGTGTGTCCGTGCGAGGCATGCCGCCGAGGGCAGAAAGGATGGCGGCAGATGTGA
GATTCTAGGCCCTGATGGTAAGACGAGGACGAAGGATAAGTACCGGGTGGTTTATTCCGACCATCAGCGCC
TGGAGCTGGAGAAGGAGTTCTACTCCAACAAGTACATCACCATCAAGAGGAAGGTTTCAGCTGGCGAACGAA
CTGGGCTGTGCGAGCGCCAGGTCAAGATCTGGTTCCAGAACAGGCGCGCCAAGCAGCGCAAGATGGCCAA
GCGGAAGGAGCTGCAGCATCCGGGCGGGCAGGGCGGGAGTGACGATGGGGGAGGGGTGATGGGAGAGGTGT
CCACACTCACGGTAGGCCCCCCACCCACAGCTCACCC TAAACCCAGCGGCGTGGCGGCCCTCCACCCTC
AGCAACCCCGCTCTCCCCCGTCTCCTCCCTCTCATGACCAGCGCCATGACGCATGCAGTGACGTTGCC
GTCGTGTGTTCTTCTCCTCGTGA

Exon structure in BAC clone 36D2

42769bp–42329bp

MYRHPSQGSYNLNPYNYATAHPAYPAEYGYQVPPAVNAGENLQQTAAAAAWQSAAAFGSHGAGQRPEEWD
GRGYNCTAGTGLTAGPTGSCTAFPGMDYPVPVGAIQANSPAVSGVTTNSTNSQRPQH SRNPYDWMRKS NYS
TSPPP

32571bp–32467bp

VLSVRGMPPQGRKDGGRCEILGPD

31997bp–31569bp

GKTRT**KDKYRVVYSDHQRL E L E K E F Y S N K Y I T I K R K V Q L A N E L G L S E R Q V K I W F Q N R R A K Q R K M A K R K E L Q**
HPGGQGGSD DGGGVMGEVSTLTVGPPPHQLTLNPSGVAASTLSNPALPPSSSPLMTSAMTHAVTL PSCVPSS
S

Figure S2: *B. floridae* Cdx open reading frame and gene structure.

SECTION 2: GENERATION OF MUTATIONS IN AMPHIOXUS *PDX* AND *CDX* GENES

TALEN sequences used to target exon 1 of *B. floridae Pdx* and *Cdx* genes. For mutagenesis using TALENS, two in vitro transcribed RNAs are injected for each gene; each mRNA includes a region of Repeat Variable Di-residues (RVDs) encoding a sequence specific DNA-binding peptide, coupled to the catalytic domain of *FokI* nuclease. If two RVDs flank a site of interest, dimerization activates *FokI* nuclease activity, introducing DNA breaks leading to deletion mutations. Figures S3 to S6 give the sequences of the four in vitro transcribed RNAs used, from the T3 RNA polymerase binding site to the restriction enzyme site used for plasmid linearization before in vitro transcription.

Pdx forward TALEN

```
AATTAACCCTCACTAAAGGGAAGCTTGCTTGTCTTTTTCGAGAAGCTCAGAATAAACGCTCAACTTTGGCAGATCTAACTCGAGAAA
GATATTGTATATATCGTAACAATAGGAGGTTCAACAATGGCTTCCCTCCCAAGAAAAAGAGAAAGGTTAGTTGGAAGGACGCAA
GTGGTTGGTCTAGAGTGGATCTACGCACGCTCGGCTACAGTCAGCAGCAGCAAGAGAAGATCAAACCGAAGGTGCGTTCGACAGTGGC
GCAGCACCACGAGGCACTGGTGGGCCATGGGTTTACACACGCGCACATCGTTGCGCTCAGCCAACACCCGGCAGCGTTAGGGACCGTC
GCTGTACGTATCAGCACATAATCACGGCGTTGCCAGAGGGACACACGAAGACATCGTTGGCGTCGGCAACAGTGGTCCGGCGCAC
GCGCCTGGAGGCTTGTCTACGGATGCGGGGGAGTTGAGAGGTCGCGCTTACAGTTGGACACAGGCCAACTTGTGAAGATTGCAAA
ACGTGGCGGCGTGACCGCAATGGAGGCAAGTGCATGCATCGCGCAATGCACTGACGGGTGCCCCCTGAACCTGACCCCGGACCAAGTG
GTGGTATCGCCAGCAACGTTGGCGCAAGCAAGCGCTCGAAACGTTGACGCGGCTTTCGCGGTGCTGTGCCAGGACCATGGCCTGA
CCCCGGACCAAGTGGTGGCTATCGCCAGCAACAATGGCGGCAAGCAAGCGCTCGAAACGTTGACGCGGCTGTTGCCGGTGTGTGCCA
GGACCATGGCCTGACTCCGGACCAAGTGGTGGCTATCGCCAGCCACGATGGCGGCAAGCAAGCGCTCGAAACGTTGACGCGGCTGTTG
CCGGTGTGTGCCAGGACCATGGCCTGACTCCGGACCAAGTGGTGGCTATCGCCAGCCACGATGGCGGCAAGCAAGCGCTCGAAACG
TGCAGCGGCTGTTGCCGGTGTGTGCCAGGACCATGGCCTGACTCCGGACCAAGTGGTGGCTATCGCCAGCCACGATGGCGGCAAGCA
AGCGCTCGAAACGTTGACGCGGCTGTTGCCGGTGTGTGCCAGGACCATGGCCTGACCCCGGACCAAGTGGTGGCTATCGCCAGCAAC
GGTGGCGGCAAGCAAGCGCTCGAAACGTTGACGCGGCTGTTGCCGGTGTGTGCCAGGACCATGGCCTGACCCCGGACCAAGTGGTGG
CTATCGCCAGCAACATGGCGGCAAGCAAGCGCTCGAAACGTTGACGCGGCTGTTGCCGGTGTGTGCCAGGACCATGGCCTGACTCC
GGACCAAGTGGTGGCTATCGCCAGCCACGATGGCGGCAAGCAAGCGCTCGAAACGTTGACGCGGCTGTTGCCGGTGTGTGCCAGGAC
CATGGCCTGACTCCGGACCAAGTGGTGGCTATCGCCAGCCACGATGGCGGCAAGCAAGCGCTCGAAACGTTGACGCGGCTGTTGCCGG
TGCTGTGCCAGGACCATGGCCTGACTCCGGACCAAGTGGTGGCTATCGCCAGCCACGATGGCGGCAAGCAAGCGCTCGAAACGTTGCA
GCGGCTGTTGCCGGTGTGTGCCAGGACCATGGCCTGACCCCGGACCAAGTGGTGGCTATCGCCAGCAACAATGGCGGCAAGCAAGCG
CTCGAAACGTTGACGCGGCTGTTGCCGGTGTGTGCCAGGACCATGGCCTGACTCCGGACCAAGTGGTGGCTATCGCCAGCCACGATG
CGCGCAAGCAAGCGCTCGAAACGTTGACGCGGCTGTTGCCGGTGTGTGCCAGGACCATGGCCTGACTCCGGACCAAGTGGTGGCTAT
CGCCAGCCACGATGGCGGCAAGCAAGCGCTCGAAACGTTGACGCGGCTGTTGCCGGTGTGTGCCAGGACCATGGCCTGACCCCGGAC
CAAGTGGTGGCTATCGCCAGCAACAATGGCGGCAAGCAAGCGCTCGAAACGATTGTGGCCAGCTGAGCCGGCCTGATCCGGCGTTGG
CCGGTTGACCAACGACCACTCGTCCCTTGGCCTGCCTCGCGGACGTCCTGCCATGGATGCAGTGAAAAAGGGATTGCCGACCGC
GCCGGAATTGATCAGAAGAGTCAATCGCCGATTGGCGAACGCACGTCCTCATCGCGTTGCCCTAGATCCCAGCTAGTGAAATCTGAA
TTGGAAGAGAAGAAATCTGAACTTAGACATAAATGAAATATGTGCCACATGAATATATTGAATTGATTGAAATCGCAAGAAATCAA
CTCAGGATAGAATCTTGAATGAAGGTGATGGAGTCTTTATGAAGGTTTATGGTTATCGTGGTAAACATTTGGGTGGATCAAGGAA
ACCAGACGGAGCAATTTACTGTGCGATCCTCATTGATTACGGTGTGATCGTTGATACTAAGGCATATTCAGGAGGTTATAATCTT
CCAATTGGTCAAGCAGATGAAATGCAAGATATGTGCAAGAGAATCAAACAAGAAACAAGCATATCAACCCATATGAATGGTGGAAAG
TCTATCCATCTTCAGTAACAGAAATTAAGTTCTTGTGTGAGTGGTCATTTCAAAGGAAACTACAAGCTCAGCTTACAAGATTGAA
TCATATCACTAATTGTAATGGAGCTGTTCTTAGTGTAGAAGAGCTTTTGTGGTGGGAGAAATGATTAAAGCTGGTACATTGACACTT
GAGGAAGTGAGAAGGAAATTTAATAACGGTGAATAAACTTTTAAATAGGCTAGTACTGACTAGGATCTGGTTACCCATAAACAGCC
TCAAGAACACCCGAATGGAGTCTCTAAGCTACATAATACCAACTTACACTTACAAAATGTTGTCCCCAAAATGTAGCCATTTCGTATC
TGCTCCTAATAAAAAGAAAGTTTCTTACATTTCAAAAAAATAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
CCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCCGCAATGCCTGCAGTGCAGTACAGGATCCCCGGGTACCGAGCTC
```

Figure S3: *Pdx* forward TALEN. Red highlight = T3 promoter; blue text = 5' and 3' untranslated regions; red text = open reading frame for TALEN peptide sequence; red underlined text = sequence encoding RVDs; green highlight = *SacI* linearization site.

Pdx reverse TALEN

AATTAACCCCTCACTAAAGGG AAGCTTGCCTTGTCTTTTTCGAGAAGCTCAGAATAAACGCTCAACTTTGGCAGATCTAACTCGAGAAA
GATATTGTATATATATCGTAACAATAGGAGGTTCAACAATGGCTTCCTCCCTCCAAAGAAAAGAGAAAGGTTAGTTGGAAGGACGCAA
GTGGTTGGTCTAGAGTGGATCTACGCACGCTCGGCTACAGTCAGCAGCAGCAAGAGAAGATCAAACCGAAGGTGCGTTTCGACAGTGGC
GCAGCACCACGAGGCACTGGTGGCCATGGGTTTACACACGCGCACATCGTTGCGCTCAGCCAACACCCGGCAGCGTTAGGGACCGTC
GCTGTCACGTATCAGCACATAATCACGGCGTTGCCAGAGGCGACACACGAAGACATCGTTGGCGTCGGCAAACAGTGGTCCGGCGCAC
GCGCCCTGGAGGCTTGTCTACGGATGCGGGGGAGTTGAGAGTCCGCGTTACAGTTGGACACAGGCCAACTTGTGAAGATTGCAAA
ACGTGGCGGCGTGACCGCAATGGAGGCAGTGCATGCATCGCGCAATGCACTGACGGGTGCCCCCTGAACCTGACCCCGGACCAAGTG
GTGGCTATCGCCAGCAACAATGGCGGCAAGCAAGCGCTCGAAACGGTGCAGCGGCTGTTGCCGGTGTGTGCCAGGACCATGGCCTGA
CTCCGGACCAAGTGGTGGCTATCGCCAGCCACGATGGCGGCAAGCAAGCGCTCGAAACGGTGCAGCGGCTGTTGCCGGTGTGTGCCA
GGACCATGGCTGACCCCGGACCAAGTGGTGGCTATCGCCAGCAACGGTGGCGGCAAGCAAGCGCTCGAAACGGTGCAGCGGCTGTTG
CCGGTGTGTGCCAGGACCATGGCCTGACCCCGGACCAAGTGGTGGCTATCGCCAGCAACAATGGCGGCAAGCAAGCGCTCGAAACGG
TGCAGCGGCTGTTGCCGGTGTGTGCCAGGACCATGGCCTGACCCCGGACCAAGTGGTGGCTATCGCCAGCAACAATGGCGGCAAGCA
AGCGCTCGAAACGGTGCAGCGGCTGTTGCCGGTGTGTGCCAGGACCATGGCCTGACCCCGGACCAAGTGGTGGCTATCGCCAGCAAC
GGTGGCGGCAAGCAAGCGCTCGAAACGGTGCAGCGGCTGTTGCCGGTGTGTGCCAGGACCATGGCCTGACTCCGGACCAAGTGGTGG
CTATCGCCAGCCACGATGGCGGCAAGCAAGCGCTCGAAACGGTGCAGCGGCTGTTGCCGGTGTGTGCCAGGACCATGGCCTGACTCC
GGACCAAGTGGTGGCTATCGCCAGCCACGATGGCGGCAAGCAAGCGCTCGAAACGGTGCAGCGGCTGTTGCCGGTGTGTGCCAGGAC
CATGGCCTGACCCCGGACCAAGTGGTGGCTATCGCCAGCAACAATGGCGGCAAGCAAGCGCTCGAAACGGTGCAGCGGCTGTTGCCGG
TGCTGTGCCAGGACCATGGCCTGACCCCGGACCAAGTGGTGGCTATCGCCAGCAACAATGGCGGCAAGCAAGCGCTCGAAACGGTGC
GCGGCTGTTGCCGGTGTGTGCCAGGACCATGGCCTGACCCCGGACCAAGTGGTGGCTATCGCCAGCCACGATGGCGGCAAGCAAGCG
CTCGAAACGGTGCAGCGGCTGTTGCCGGTGTGTGCCAGGACCATGGCCTGACCCCGGACCAAGTGGTGGCTATCGCCAGCAACAATGG
GCGGCAAGCAAGCGCTCGAAACGGTGCAGCGGCTGTTGCCGGTGTGTGCCAGGACCATGGCCTGACTCCGGACCAAGTGGTGGCTAT
CGCCAGCCACGATGGCGGCAAGCAAGCGCTCGAAACGGTGCAGCGGCTGTTGCCGGTGTGTGCCAGGACCATGGCCTGACTCCGGAC
CAAGTGGTGGCTATCGCCAGCCACGATGGCGGCAAGCAAGCGCTCGAAACGGTGCAGCGGCTGTTGCCGGTGTGTGCCAGGACCATG
GCCTGACCCCGGACCAAGTGGTGGCTATCGCCAGCAACAATGGCGGCAAGCAAGCGCTCGAAACGGTGCAGCGGCTGTTGCCGGTGT
GTGCCAGGACCATGGCCTGACCCCGGACCAAGTGGTGGCTATCGCCAGCCACGATGGCGGCAAGCAAGCGCTCGAAACGGTGCAGCG
CAGCTGAGCCGCGCTGATCCGGCGTTGGCCGCGTTGACCAACGACCACTCGTCGCCTTGGCCTGCCTCGGGCGGACGTCCTGCCATGG
ATGCAGTGAAAAGGGATTGCCGCACGCGCCGGAATTGATCAGAAGAGTCAATCGCCGATTGGCGAACGCACGTCCCATCGGTTGC
CTCTAGATCCAGCTAGTGAATCTGAATTGGAAGAGAAGAAATCTGAACCTAGACATAAATGAAATATGTGCCACATGAATATAT
GAATTGATTGAAATCGCAAGAAATCAACTCAGGATAGAATCCTTGAAATGAAGGTGATGGAGTTCCTTATGAAGGTTTATGGTTATC
GTGGTAAACATTTGGGTGGATCAAGGAAACCAGACGGAGCAATTTACTGTGCGATCTCCTATTGATTACGGTGTGATCGTTGATAC
TAAGGCATATTCAGGAGTTATAATCTTCCAATTGGTCAAGCAGATGAAATGCAAAGATATGTCGAAGAGAATCAAACAAGAAACAAG
CATATCAACCCTAATGAATGGTGGAAAGTCTATCCATCTTCAGTAACAGAATTTAAGTCTTGTGTTGTGAGTGGTCATTTCAAAGGAA
ACTACAAGCTCAGCTTACAAGATTGAATCATATCACTAATGTAATGGAGCTGTTCTTAGTGTAGAAGAGCTTTTGATTGGTGGAGA
AATGATTAAGCTGGTACATTGACTTGGAGGAGTGAGAAGGAAATTTAATAACGGTGAAGATAAATTTTAAATAGGCTAGTACTGA
CTAGGATCTGGTTACCACTAAACCAGCCTCAAGAACACCCGAATGGAGTCTCTAAGCTACATAATACCAACTTACACTTACAAAATGT
TGTCACCCAAAATGTAGCCATTCTGATCTCCTAATAAAAAAGAAAGTTTCTTACATTCTAAAAAAGAAAAAAGAAAAAAGAAAAA
AAAACCCCCCCCCCCCCCCCCCCCCCCCCGCATGCCTGCAGGTCGACTAGGATCCCCGGTACC GAGCTC

Figure S4: *Pdx* reverse TALEN. Red highlight = T3 promoter; blue text = 5' and 3' untranslated regions; red text = open reading frame for TALEN peptide sequence; red underlined text = sequence encoding RVDs; green highlight = *SacI* linearization site.

Cdx forward TALEN

AATTAACCCTCACTAAAGGGAAGCTTGCTTGTCTTTTTCGAGAAGCTCAGAATAAACGCTCAACTTTGGCAGATCTAACTCGAGAAA
GATATTGTATATATCGTAAACAATAGGAGGTTCAACAATGGCTTCCCTCCCTCCAAAAGAAAAGAGAAAGGTTAGTTGGAAGGACGCCAA
GTGGTTGGTCTAGAGTGGATCTACGCACGCTCGGCTACAGTCAGCAGCAGCAAGAGAAGATCAAACCGAAGGTGCGTTTCGACAGTGGC
GCAGACCACGAGGCACTGGTGGCCATGGGTTTACACACGCGCACATCGTTGCGCTCAGCCAACACCCGGCAGCGTTAGGGACCGTC
GCTGTCACGTATCAGCACATAATCACGGCGTTGCCAGAGGCGACACAGAAAGACATCGTTGGCGTCGGCAAACAGTGGTCCGGCGCAC
GGCCCTGGAGGCCCTTGCTCACGGATGGGGGGAGTTGAGAGGTCGCCGTTACAGTTGGACACAGGCCAACTTGTGAAGATTGCAAA
ACGTGGCGCGTGACCGCAATGGAGGCAGTGCATGCATCGCGCAATGCACGTGACGGGTGCCCCCTGAACCTGACCCCGGACCAAGTG
GTGGCTATCGCCAGCCACGATGGCGGCAAGCAAGCGCTCGAAAACGGTGCAGCGGCTGTTGCCGGTGTGTGCCAGGACCATGGCCTGA
CTCCGGACCAAGTGGTGGCTATCGCCAGCCACGATGGCGGCAAGCAAGCGCTCGAAAACGGTGCAGCGGCTGTTGCCGGTGTGTGCCA
GGACCATGGCTGACCCCGGACCAAGTGGTGGCTATCGCCAGCAACAATGGCGGCAAGCAAGCGCTCGAAAACGGTGCAGCGGCTGTTG
CCGGTGTGTGCCAGGACCATGGCCTGACTCCGGACCAAGTGGTGGCTATCGCCAGCCACGATGGCGGCAAGCAAGCGCTCGAAAACGG
TGCAGCGGCTGTTGCCGGTGTGTGCCAGGACCATGGCCTGACTCCGGACCAAGTGGTGGCTATCGCCAGCCACGATGGCGGCAAGCA
AGCGCTCGAAAACGGTGCAGCGGCTGTTGCCGGTGTGTGCCAGGACCATGGCCTGACCCCGGACCAAGTGGTGGCTATCGCCAGCAAC
ATTGGCGGCAAGCAAGCGCTCGAAAACGGTGCAGCGGCTGTTGCCGGTGTGTGCCAGGACCATGGCCTGACCCCGGACCAAGTGGTGG
CTATCGCCAGCAACGGTGGCGGCAAGCAAGCGCTCGAAAACGGTGCAGCGGCTGTTGCCGGTGTGTGCCAGGACCATGGCCTGACCCCG
GGACCAAGTGGTGGCTATCGCCAGCAACAATGGCGGCAAGCAAGCGCTCGAAAACGGTGCAGCGGCTGTTGCCGGTGTGTGCCAGGAC
CATGGCCTGACCCCGGACCAAGTGGTGGCTATCGCCAGCAACGGTGGCGGCAAGCAAGCGCTCGAAAACGGTGCAGCGGCTGTTGCCGG
TGCTGTGCCAGGACCATGGCCTGACCCCGGACCAAGTGGTGGCTATCGCCAGCAACAATTGGCGGCAAGCAAGCGCTCGAAAACGGTGC
GCGGCTGTTGCCGGTGTGTGCCAGGACCATGGCCTGACCCCGGACCAAGTGGTGGCTATCGCCAGCCACGATGGCGGCAAGCAAGCG
CTCGAAAACGGTGCAGCGGCTGTTGCCGGTGTGTGCCAGGACCATGGCCTGACTCCGGACCAAGTGGTGGCTATCGCCAGCCACGATG
GCGGCAAGCAAGCGCTCGAAAACGGTGCAGCGGCTGTTGCCGGTGTGTGCCAGGACCATGGCCTGACCCCGGACCAAGTGGTGGCTAT
CGCCAGCAACAATGGCGGCAAGCAAGCGCTCGAAAACGGTGCAGCGGCTGTTGCCGGTGTGTGCCAGGACCATGGCCTGACCCCGGAC
CAAGTGGTGGCTATCGCCAGCAACGGTGGCGGCAAGCAAGCGCTCGAAAACGGTGCAGCGGCTGTTGCCGGTGTGTGCCAGGACCATG
GCCTGACTCCGGACCAAGTGGTGGCTATCGCCAGCCACGATGGCGGCAAGCAAGCGCTCGAAAACGGTGCAGCGGCTGTTGCCGGTGT
GTGCCAGGACCATGGCCTGACCCCGGACCAAGTGGTGGCTATCGCCAGCAACAATTGGCGGCAAGCAAGCGCTCGAAAACGATTTGGCC
CAGCTGAGCCGGCCTGATCCGGCCTGGCCGCGTTGACCAACGACACCTCGTCGCCTTGGCCTGCCTCGGGGACGTCCTGCCATGG
ATGCAGTGAAAAGGGATTGCCGCACGCGCCGGAATTGATCAGAAGAGTCAATCGCCGATTGGCGAACGCACGTCCCATCGCGTTGC
CTCTAGATCCAGCTAGTGAATCTGAATTGGAAGAGAAGAAATCTGAACCTAGACATAAAATGAAATATGTGCCACATGAATATATT
GAATTGATTGAAATCGCAAGAAATCAACTCAGGATAGAATCCTTGAAATGAAGGTGATGGAGTTCTTTATGAAGGTTTATGGTTATC
GTGGTAAACATTTGGGTGGATCAAGGAAACCAGACGGAGCAATTTACTGTCCGATCTCCTATTGATTACGGTGTGATCGTTGATAC
TAAGGCATATTCAGGAGGTTATAATCTTCCAATTGGTCAAGCAGATGAAATGCAAAAGATATGTCGAAGAGAATCAAACAAGAAACAAG
CATATCAACCCTAATGAATGGTGGAAAGTCTATCCATCTTCAGTAACAGAATTTAAGTTCTTGTGTTGTGAGTGGTCATTTCAAAGGAA
ACTACAAGCTCAGCTTACAAGATTGAATCATATCACTAATTTGTAATGGAGCTGTTCTTAGTGTAGAAGAGCTTTTATTGGTGGAGA
AATGATTAAGCTGGTACATTGACACTTGAGGAAGTGAGAAGGAAATTTAATAACGGTGAATAAACTTTTTAATAGGCTAGTGACTGA
CTAGGATCTGGTTACCACTAAACCAGCCTCAAGAACACCCGAATGGAGTCTCTAAGCTACATAATACCAACTTACACTTACAAAATGT
TGTCACCAAAATGTAGCCATTTCGTATCTGCTCCTAATAAAAAAGAAAGTTTCTTACATTCTAAAAAATAAAAAAAAAAAAAAAAAA
AAAACCCCCCCCCCCCCCCCCCCCCCCCCGCATGCCTGCAAGTTCGACTAGGATCCCCGGTACC**GAGCTC**

Figure S5: Cdx forward TALEN. Red highlight = T3 promoter; blue text = 5' and 3' untranslated regions; red text = open reading frame for TALEN peptide sequence; red underlined text = sequence encoding RVDs; green highlight = *SacI* linearization site.

Cdx reverse TALEN

AATTAACCCCTCACTAAAGGAAGCTTGCTTGTCTTTTTCGAGAAGCTCAGAATAAACGCTCAACTTTGGCAGATCTAACTCGAGAAA
GATATTGTATATATCGTAACAATAGGAGGTTCAACAATGGCTTCCCTCCCTCCAAAGAAAAAGAGAAAGGTTAGTTGGAAGGACGCAA
GTGGTTGGTCTAGAGTGGATCTACGCACGCTCGGCTACAGTCAGCAGCAGCAAGAGAAAGATCAAACCGAAGGTGCGTTTCGACAGTGGC
GCAGCACCACGAGGCACTGGTGGGCCATGGGTTTACACACGCGCACATCGTTGCGCTCAGCCAACACCCGGCAGCGTTAGGGACCGTC
GCTGTACGATATCAGCACATAATCACGGCTTGCCAGAGGCGACACAGAAGACATCGTTGGCGTCGGCAAACAGTGGTCCGGCGCAC
GCGCCCTGGAGGCTTGCTCACGGATGCGGGGAGTTGAGAGTCCGCGTTACAGTTGGACACAGGCCAACTTGTGAAGATTGCAAA
ACGTGGCGGCTGACCGCAATGGAGGCGATGCATGCATCGCGCAATGCACTGACGGGTGCCCCCTGAACCTGACCCCGGACCAAGTG
GTGGCTATCGCCAGCAACATTTGGCGCAAGCAAGCGCTCGAAACGGTGCAGCGGCTGTTGCCGGTGTGTGCCAGGACCATGGCCTGA
CTCCGGACCAAGTGGTGGCTATCGCCAGCCACGATGGCGGCAAGCAAGCGCTCGAAACGGTGCAGCGGCTGTTGCCGGTGTGTGCCA
GGACCATGGCTGACCCCGGACCAAGTGGTGGCTATCGCCAGCAACAATGGCGGCAAGCAAGCGCTCGAAACGGTGCAGCGGCTGTTG
CCGGTGTGTGCCAGGACCATGGCTGACCCCGGACCAAGTGGTGGCTATCGCCAGCAACAATGGCGGCAAGCAAGCGCTCGAAACGG
TGCAGCGGCTGTTGCCGGTGTGTGCCAGGACCATGGCCTGACCCCGGACCAAGTGGTGGCTATCGCCAGCAACAATGGCGGCAAGCA
AGCGCTCGAAACGGTGCAGCGGCTGTTGCCGGTGTGTGCCAGGACCATGGCCTGACCCCGGACCAAGTGGTGGCTATCGCCAGCAAC
GGTGGCGGCAAGCAAGCGCTCGAAACGGTGCAGCGGCTGTTGCCGGTGTGTGCCAGGACCATGGCCTGACCCCGGACCAAGTGGTGG
CTATCGCCAGCAACGGTGGCGGCAAGCAAGCGCTCGAAACGGTGCAGCGGCTGTTGCCGGTGTGTGCCAGGACCATGGCCTGACTCC
GGACCAAGTGGTGGCTATCGCCAGCCACGATGGCGGCAAGCAAGCGCTCGAAACGGTGCAGCGGCTGTTGCCGGTGTGTGCCAGGAC
CATGGCCTGACCCCGGACCAAGTGGTGGCTATCGCCAGCAACAATGGCGGCAAGCAAGCGCTCGAAACGGTGCAGCGGCTGTTGCCGG
TGCTGTGCCAGGACCATGGCCTGACCCCGGACCAAGTGGTGGCTATCGCCAGCAACAATGGCGGCAAGCAAGCGCTCGAAACGGTGC
GCGGCTGTTGCCGGTGTGTGCCAGGACCATGGCCTGACCCCGGACCAAGTGGTGGCTATCGCCAGCAACAATGGCGGCAAGCAAGCG
CTCGAAACGGTGCAGCGGCTGTTGCCGGTGTGTGCCAGGACCATGGCCTGACCCCGGACCAAGTGGTGGCTATCGCCAGCAACGGTG
GCGGCAAGCAAGCGCTCGAAACGGTGCAGCGGCTGTTGCCGGTGTGTGCCAGGACCATGGCCTGACCCCGGACCAAGTGGTGGCTAT
CGCCAGCAACGGTGGCGGCAAGCAAGCGCTCGAAACGGTGCAGCGGCTGTTGCCGGTGTGTGCCAGGACCATGGCCTGACCCCGGAC
CAAGTGGTGGCTATCGCCAGCAACAATGGCGGCAAGCAAGCGCTCGAAACGGTGCAGCGGCTGTTGCCGGTGTGTGCCAGGACCATG
GCCTGACCCCGGACCAAGTGGTGGCTATCGCCAGCAACGGTGGCGGCAAGCAAGCGCTCGAAACGGTGCAGCGGCTGTTGCCGGTGT
GTGCCAGGACCATGGCCTGACCCCGGACCAAGTGGTGGCTATCGCCAGCAACAATGGCGGCAAGCAAGCGCTCGAAACGGTGGCC
CAGCTGAGCGGCGCTGATCGGCGGCTGGCCGCGTTGACCAACGACACCTCGTCGCTTGGCTGCCTCGGCGGACGTCCTGCCATGG
ATGCAGTGAAGGATTCGCGCAGCGCGGAATTGATCAGAAGAGTCAATCGCCGTATTGGCGAACGCACGTCCCATCGCGTTGC
CTCTAGATCCAGCTAGTGAATCTGAATTGGAAGAGAAGAAATCTGAACCTAGACATAAATGAAATATGTGCCACATGAATATATT
GAATTGATTGAAATCGCAAGAAATCAACTCAGGATAGAATCCTTGAAATGAAGGTGATGGAGTCTTTATGAAGGTTTATGGTTATC
GTGGTAAACATTTGGTGGATCAAGGAAACCAGACGGAGCAATTTATACTGTCGGATCTCCTATTGATTACGGTGTGATCGTTGATAC
TAAGGCATATTCAGGAGTTATAATCTTCCAATTTGGTCAAGCAGATGAAATGCAAAGATATGTGCAAGAGAATCAAACAAGAAACAAG
CATAACAACCTAATGAATGGTGGAAAGTCTATCCATCTTCAGTAACAGAATTTAAGTTCTTGTGTTGTGAGTGGTCATTTCAAAGGAA
ACTACAAGCTCAGCTTACAAGATTGAATCATATCACTAATTGTAATGGAGCTGTTCTTAGTGTAGAAGAGCTTTTGATTGGTGGAGA
AATGATTAAAGCTGTTACATTGACACTTGAGGAAGTGAGAAGGAAATTTAATAACGGTGAAGATAAACTTTTAAATAGGCTAGTACTGA
CTAGGATCTGGTTACCACTAAACCAGCCTCAAGAACACCCGAATGGAGTCTCTAAGCTACATAATACCAACTTACACTTACAAAATGT
TGTCCCCAAAATGTAGCCATTCTGATCTGCTCCTAATAAAAAGAAAGTTTCTTCACATTCAAAAAAGAAAAAAGAAAAAAGAAAA
AAAACCCCCCCCCCCCCCCCCCCCCCCCCGCATGCTGCAGTTCGACTAGGATCCCCGGTACCGAGCTC

Figure S6: Cdx reverse TALEN. Red highlight = T3 promoter; blue text = 5' and 3' untranslated regions; red text = open reading frame for TALEN peptide sequence; red underlined text = sequence encoding RVDs; green highlight = *SacI* linearization site.

Each TALEN pair is designed to flank a restriction endonuclease recognition site enabling mutation detection by digestion of a PCR amplified product (*AflIII* for *Pdx*, *PasI* for *Cdx*). Table 1 gives the mutation detection primers used.

Genes	Primer sequences (5'→3')
<i>Pdx</i> mutation detection	Forward: TTTCAAACGATACCGGACAAAC Reverse: CCACTGAGACTTCCAGGCGT
<i>Cdx</i> mutation detection	Forward: TACTGGTTTGTACGGCGAG Reverse: CTGGGGGAGGACTTGTGGAGTA

Table S1: Mutation detection primers

B. floridae eggs were injected with pairs of TALEN RNAs, fertilized and embryos reared to neurula stage, using the methods described in Li et al. (2014). PCR followed by restriction digestion revealed that both TALEN pairs introduced deletion mutations, with a higher frequency detected by the *Cdx* TALEN pair (Figure S7).

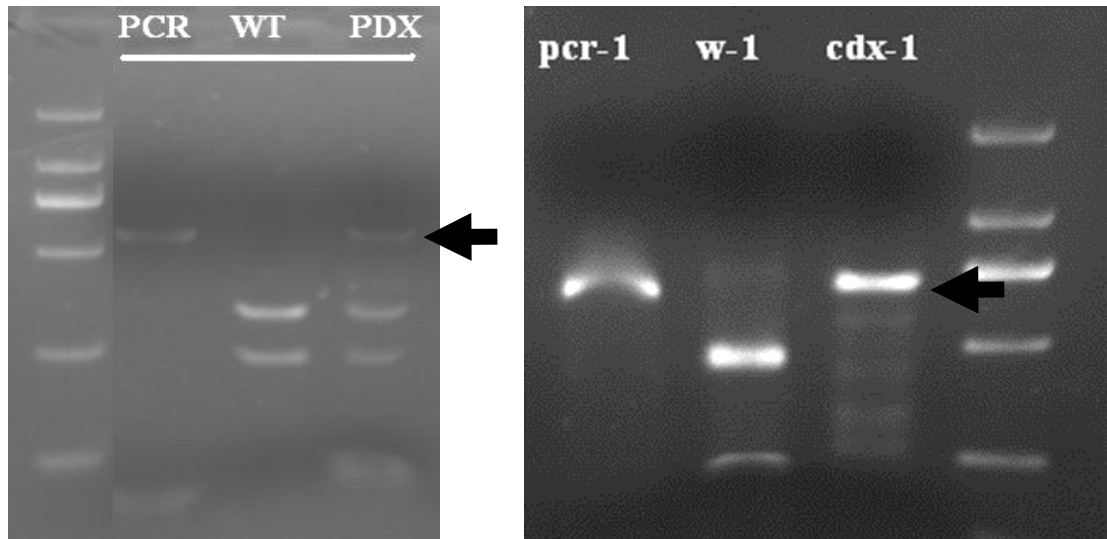


Figure S7: PCR analysis of embryos reared from injected eggs: Left hand gel: *Pdx* PCR products showing faint band of uncut amplification product (deletion mutants, arrow). Right hand gel: *Cdx* PCR products showing uncut amplification product (arrow).

The same PCR primers were used to track inheritance of the mutant alleles in adult amphioxus tail clips, in sperm or pools of embryos, and in single embryos after in situ hybridization, using methods described in Li et al (2017). Sequencing of PCR products was also used to verify exact nature of the mutations.

The TALEN-generated *Pdx* mutants used in this study have deletions of 4 bp, 11 bp and 13 bp (Figure 1, main text; Figure S8). The predicted mutant peptides are shown beneath the DNA sequences; no predicted product contains a homeodomain.

➤ 4 bp *Pdx* deleted region (red) and predicted protein product

ATGATCCCGGCGTCGTACCAGCAGCACCAGGCCCGGTCGTCTGCCTGTACGCCAACACCCAACAGCCGCAGC
ACGCCATGCCCTACCCGCCGCAA**ACAT**GTCCGTCGTGGAGCTGGATCAGCTGGACGCAGAACTTCCGGGCGG
CGGCATGCCGGGGCCCGGCCCATGGCGTCCCTCGGGACCAGGCCCGACCCAGCCCGTCCACCACGCCGGCCCG
CCCCGGCCCCgCAGTCCAGcTGTGCCGTCAACAGGAACGAGAACCCTGCCGTTCCCCTGGATGAAGACCACCA
AGTCGCAcGCTCACGCCCTGGAAGTCTCAGTGGCCAGGTGCGTCCTTcGCTGTTGAGGATGAGAACAAGAGAAC
GCGCACAGCCTACACCCGtGGCCAGCTCCTGGAGCTGGAGAAGGAGTTTCACTTCAACAAGTACATTTCCCGG
CCGCGCAGGATAGAGCTAGCCGCCATGCTCAACCTCACAGAGAGACACATCAAAATCTGGTTCCAGAACCGCC
GCATGAAGTGGAAAAAGGAGCAGGCaAAGCGGCGGCCGCTGCCCGAGTCTGCCCTCCAGCACGACCCCGGGGG
CAGCgGcGGGgCCGGCACC CGCGGGGGGGCGCCGAGTCCACGGGGACCAGCGGCACCGACCCCGAGACTTCA
CCGGTCAGAGAGCCGGTCTCGACGCCCTCCCGCCTCcACGTCTTTACCGGTGTCTCCACCTGTGAACCTCAGGTG
TACAGGGGACCTCAGCAGCTTCCACACGGGCGGGGTTACCGTTCCCCCGTGCACCAAACACTGCCTCATAG
CGTTACCGGACCGACAGAGCCCACTCCAACGGGAAAACCTCTCACAGAGCCTGGCCTTTTTCACGCTCCTGA

MIPASYQQHQARSSCLYANTQQPQHAMPYPPPSPSWSWISWTQNFRAAACRGPAPWRPRDQARPSPTTPARP
RPRSPAVPSTGTRTCRSPG*

➤ 11 bp *Pdx* deleted region (red) and predicted protein product

ATGATCCCGGCGTCGTACCAGCAGCACCAGGCCCGGTCGTCTGCCTGTACGCCAACACCCAACAGCCGCAGC
ACGCCATGCCCTACCCGCCGCAA**ATGTCGGTCTG**GGAGCTGGATCAGCTGGACGCAGAACTTCCGGGCGG
CGGCATGCCGGGGCCCGGCCCATGGCGTCCCTCGGGACCAGGCCCGACCCAGCCCGTCCACCACGCCGGCCCG
CCCCGGCCCCgCAGTCCAGcTGTGCCGTCAACAGGAACGAGAACCCTGCCGTTCCCCTGGATGAAGACCACCA
AGTCGCAcGCTCACGCCCTGGAAGTCTCAGTGGCCAGGTGCGTCCTTcGCTGTTGAGGATGAGAACAAGAGAAC
GCGCACAGCCTACACCCGtGGCCAGCTCCTGGAGCTGGAGAAGGAGTTTCACTTCAACAAGTACATTTCCCGG
CCGCGCAGGATAGAGCTAGCCGCCATGCTCAACCTCACAGAGAGACACATCAAAATCTGGTTCCAGAACCGCC
GCATGAAGTGGAAAAAGGAGCAGGCaAAGCGGCGGCCGCTGCCCGAGTCTGCCCTCCAGCACGACCCCGGGGG
CAGCgGcGGGgCCGGCACC CGCGGGGGGGCGCCGAGTCCACGGGGACCAGCGGCACCGACCCCGAGACTTCA
CCGGTCAGAGAGCCGGTCTCGACGCCCTCCCGCCTCcACGTCTTTACCGGTGTCTCCACCTGTGAACCTCAGGTG
TACAGGGGACCTCAGCAGCTTCCACACGGGCGGGGTTACCGTTCCCCCGTGCACCAAACACTGCCTCATAG
CGTTACCGGACCGACAGAGCCCACTCCAACGGGAAAACCTCTCACAGAGCCTGGCCTTTTTCACGCTCCTGA

MIPASYQQHQARSSCLYANTQQPQHAMPYPPPNGAGSAGRRTSGRRHAGARPHGVLGTRPDPARPPRRPAPGP
AVQLCRQQEREPAVPLDEDHQVARSRLEVSVARCVLRC*

➤ 13 bp *Pdx* deleted region (red) and predicted protein product

ATGATCCCGGCGTCGTACCAGCAGCACCAGGCCCGGTCGTCTGCCTGTACGCCAACACCCAACAGCCGCAGC
ACGCCATGCCCTACCC**GCCGCCAAACATG**TCCGTCGTGGAGCTGGATCAGCTGGACGCAGAACTTCCGGGCGG
CGGCATGCCGGGGCCCGGCCCATGGCGTCCCTCGGGACCAGGCCCGACCCAGCCCGTCCACCACGCCGGCCCG
CCCCGGCCCCgCAGTCCAGcTGTGCCGTCAACAGGAACGAGAACCCTGCCGTTCCCCTGGATGAAGACCACCA
AGTCGCAcGCTCACGCCCTGGAAGTCTCAGTGGCCAGGTGCGTCCTTcGCTGTTGAGGATGAGAACAAGAGAAC
GCGCACAGCCTACACCCGtGGCCAGCTCCTGGAGCTGGAGAAGGAGTTTCACTTCAACAAGTACATTTCCCGG
CCGCGCAGGATAGAGCTAGCCGCCATGCTCAACCTCACAGAGAGACACATCAAAATCTGGTTCCAGAACCGCC
GCATGAAGTGGAAAAAGGAGCAGGCaAAGCGGCGGCCGCTGCCCGAGTCTGCCCTCCAGCACGACCCCGGGGG
CAGCgGcGGGgCCGGCACC CGCGGGGGGGCGCCGAGTCCACGGGGACCAGCGGCACCGACCCCGAGACTTCA
CCGGTCAGAGAGCCGGTCTCGACGCCCTCCCGCCTCcACGTCTTTACCGGTGTCTCCACCTGTGAACCTCAGGTG
TACAGGGGACCTCAGCAGCTTCCACACGGGCGGGGTTACCGTTCCCCCGTGCACCAAACACTGCCTCATAG
CGTTACCGGACCGACAGAGCCCACTCCAACGGGAAAACCTCTCACAGAGCCTGGCCTTTTTCACGCTCCTGA

MIPASYQQHQARSSCLYANTQQPQHAMPYPPSWSWISWTQNFRAAACRGPAPWRPRDQARPSPTTPARPRR
SPAVPSTGTRTCRSPG*

Figure S8: Sites of 4 bp, 11 bp and 13 bp deletions in *Pdx* gene (red) and predicted protein products (out of frame amino acids underlined)

The TALEN-generated *Cdx* mutant used in this study has a deletion of 7 bp (Figure 1, main text; Figure S9). The predicted mutant peptide retains the first 5 amino acids before a frameshift, then 4 additional residues (underlined) before a premature stop codon.

7 bp *Cdx* deleted region (red) and predicted protein product

ATGTACCGTCACCC**CTCCCAG**GGCAGCTACAACCTTGAACCCGTACAACCTACGCCACGGCGCACCCCTGCC
TACCCCGCGGAGTACGGACAGTACCAGGTCCCGCCTGCCGTCAACGCCGGCGGAGAACCCTACAGCAGACG
GCCGCCGCCCGCGTGGCAGTCCGCGCAGCCTTCGGCTCGCACGGGGCCGGACAGAGGCCAGAGGAA
TGGGACGGTCGCGGTACAACCTGCACGGCGGGGACCGGGCTGACCGCCGGCCCGACCGGGTCTGTACA
GCCTTCCCCGGGATGGACTACCCCTGTCCCCGTCCGTGCCATCCAGGCCAACAGCCCTGCCGTGTCCGGA
GTGACGACCAACTCTACCAACAGTCAGAGACCACAGCACAGCAGAAATCCGTACGACTGGATGAGGAAA
AGCAACTACTCCACAAGTCTCCCCAGTGTCTCCGTGCGAGGCATGCCGCCGAGGGCAGAAAGGAT
GGCGGCAGATGTGAGATTCTAGGCCCTGATGGTAAGACGAGGACGAAGGATAAGTACCGGGTGGTTTAT
TCCGACCATCAGCGCCTGGAGCTGGAGAAGGAGTTCTACTCCAACAAGTACATCACCATCAAGAGGAAG
GTTTCAGCTGGCGAACGAACCTGGGCCTGTCCGAGCGCCAGGTCAAGATCTGGTTCCAGAACAGGCGCGCC
AAGCAGCGCAAGATGGCCAAGCGGAAGGAGCTGCAGCATCCGGGCGGGCAGGGCGGGAGTGACGATGGG
GGAGGGGTGATGGGAGAGGTGTCCACACTCACGGTAGGCCCCCCACCCCACCAGCTCACCTAAACCCC
AGCGGCGTGGCGGCCTCCACCCTCAGCAACCCCGCTCTCCCCCGTCTCTCCCTCTCATGACCAGC
GCCATGACGCATGCAGTGACGTTGCCGTCCGTGTGTTTCCTTCCTCGTGA

MYRHPPAATT*

Figure S9: Site of 7 bp deletion in *Cdx* gene (red) and predicted protein product (out of frame amino acids underlined)

SECTION 3: CLONING GENES FOR IN SITU HYBRIDISATION PROBES

Genes	Primer sequences (5'→3')
<i>Pdx</i>	Forward: GGTACCTACCCACGAGAAGGGTACGA Reverse: GAATTCGGAGAGCCGTTGTTGACGTA
<i>Cdx</i>	Forward: ATGTACCGTCACCCCTCCCAGGG Reverse: TCACGAGGAAGGAACACACGACG
<i>Ilp1</i>	Forward: GGTACCCAGGCATGAATCTATCCAGCG Reverse: GAATTCGGAAACTGCCTCCTAGACGTT
<i>Brachyury (Bra2)</i>	Forward: AGACCAGCGTCAACAACGAGATG Reverse: AACAACTGGAGCCCYATGAC
<i>Mop</i>	Forward: CTCGAGATGACTGAGCTGCCATCGTT Reverse: GATATCAGTTTGGATTCCGCCAGTCT
<i>Mitf</i>	Forward: ATGCAAGACGAGTCAGGTGTTG Reverse: TCATTGGAGCTGCAGGAGATCA
<i>Cyp26-3</i>	Forward: AAGACTCTCTCGTCAGTCGG Reverse: TGAAGGACAGCACGTCATCC
<i>Rootletin</i>	Forward: GAAGCGTGACCCGAGTACA Reverse: TTAGCCTCGGAAAGGGCTTG

Table S2: Primers used to clone genes for riboprobe synthesis. Some primer sequences include restriction endonuclease sites.

SECTION 4: MORPHOLOGY OF *PDX* MUTANT AMPHIOXUS

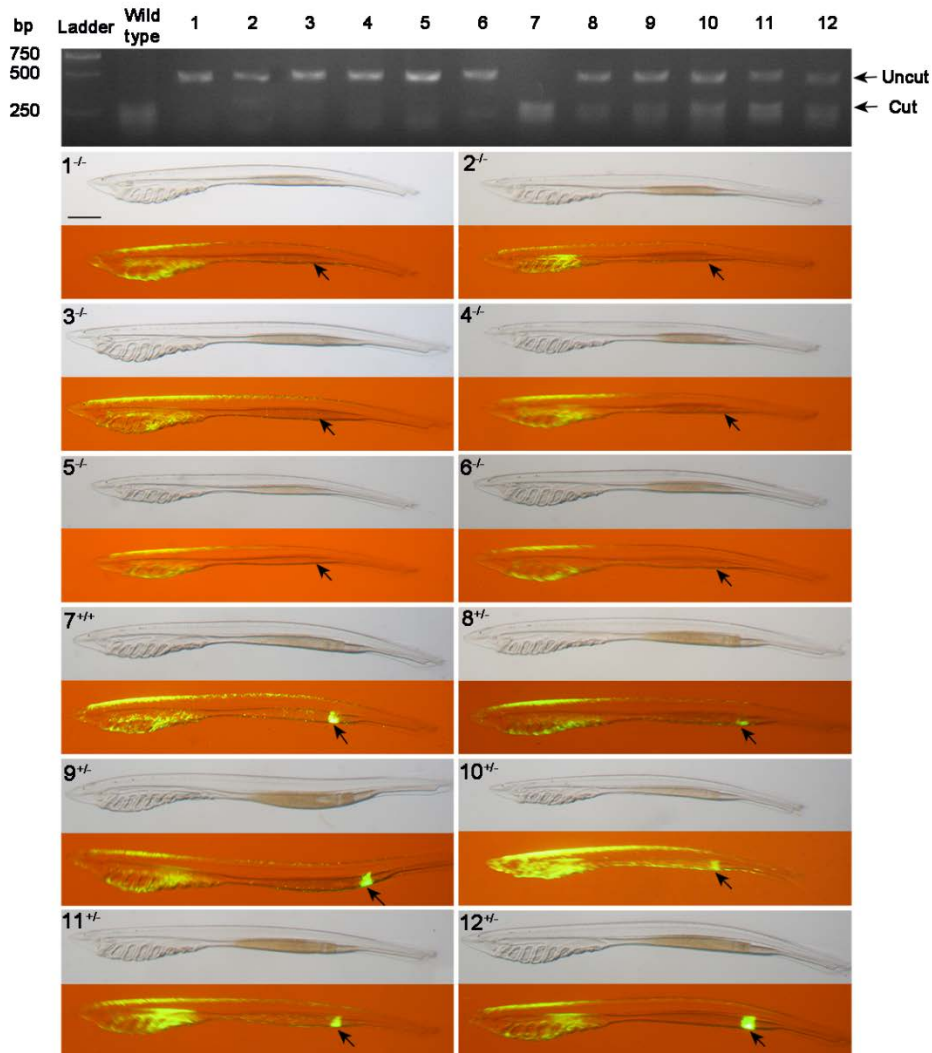
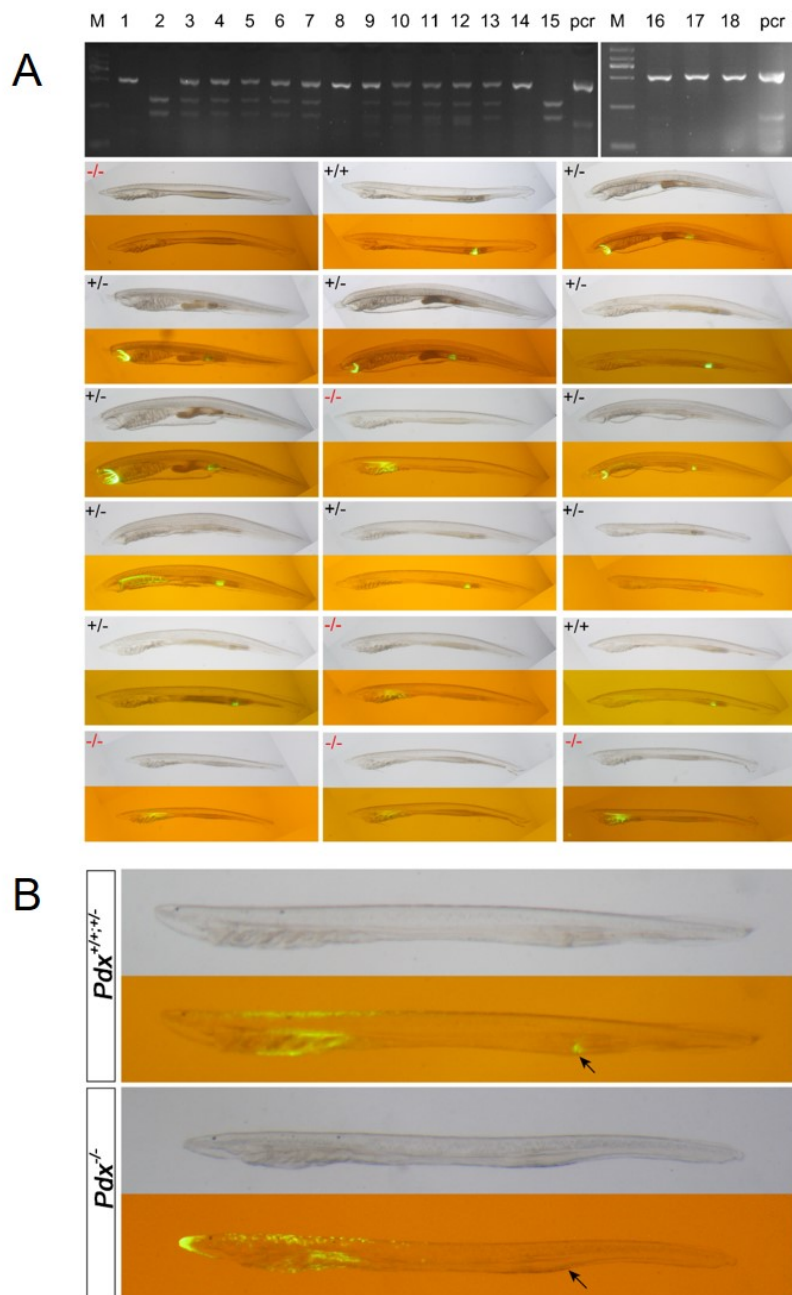


Figure S10: Amphioxus larvae at 16 d post-fertilisation (7-gill stage) showing endogenous green fluorescence in buccal cirri (all genotypes) and gut region (arrow; not seen in $-/-$ mutants). These larvae were generated by a cross between heterozygous animals with 4 bp *Pdx* deletion; larvae labelled $-/-$ are *Pdx* 4Δ homozygotes. Top panels show digestion products from PCR across the mutated region: mutation removes the restriction site, so the higher band indicates presence of the mutated allele, the lower band the wild type allele.

Figure S11: Fluorescence in *Pdx* compound heterozygous amphioxus larvae and siblings.



(A) Larvae at 46 days development showing endogenous green fluorescence in buccal cirri (all genotypes) and gut region (arrow; not in $-/-$ mutants). Larvae from a cross between female *Pdx* 11 Δ heterozygote and male *Pdx* 13 Δ heterozygote; larvae labelled $-/-$ are *Pdx* 11 Δ /13 Δ compound heterozygotes. Top panels show digestion products from PCR across the mutated region: mutation removes the restriction site, so the higher band indicates presence of the mutated allele, the lower band the wild type allele.

(B) Larvae at 6 days development (5 gill slits) showing endogenous green fluorescence in buccal cirri (all genotypes) and gut region (arrow; not in $-/-$ mutant). Larvae from a cross *Pdx* 4 Δ heterozygote and *Pdx* 11 Δ heterozygote; larva labelled $-/-$ is *Pdx* 4 Δ /11 Δ compound heterozygote.

SECTION 5: MORPHOLOGY OF *CDX* MUTANT AMPHIOXUS

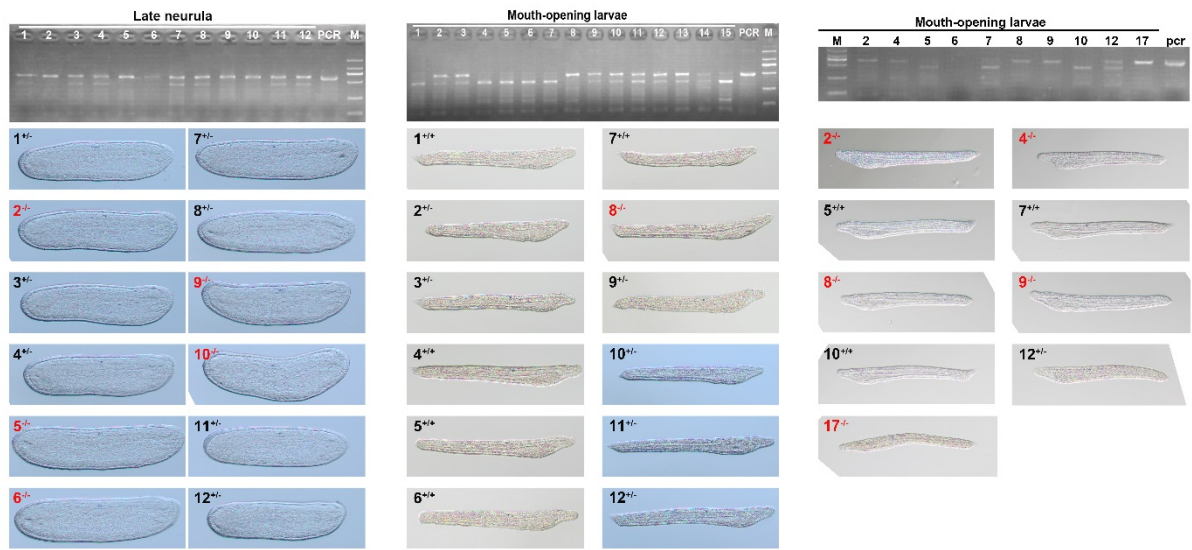


Figure S12. Morphological similarity of wild type and mutant *Cdx* embryos up to mouth opening. Top panels show digestion products from PCR across the mutated region: mutation removes the restriction site, so the higher band indicates presence of the mutated allele, the lower band the wild type allele.

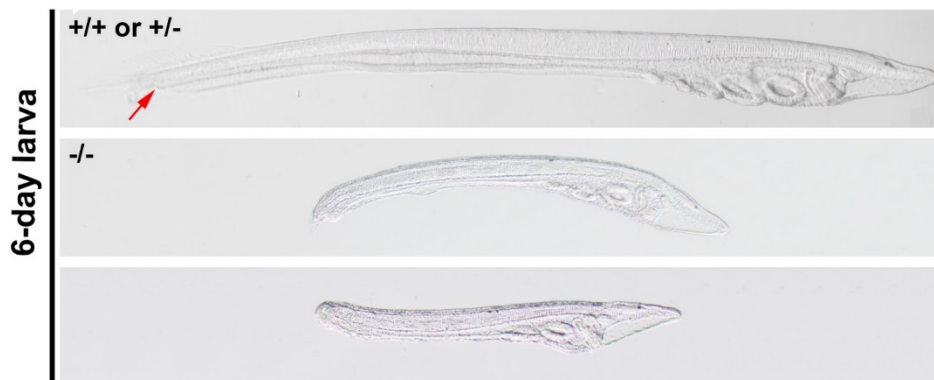


Figure S13. Extreme body truncation after 6 days of development in *Cdx* homozygous mutants.

SECTION 6: REPORTER GENE ANALYSIS OF CYP26-3

We identify five putative Cdx binding sites (TTTATT/AATAAA) in ~3.1 kb upstream of the ATG of *Cyp26a-3* (blue highlights in Figure S14). Four are upstream of the deduced transcriptional start site (black text) and one in the 5' untranslated region (blue text). Two were chosen for mutagenesis (bold and underlined); these have a purine residue following TTTATT and best match mouse (WTTWAYKRNH) (Amin et al. 2016) and human (NTTTATDRBHB) Cdx binding sites (<http://jaspar.binf.ku.dk/>)

```
BfCyp26a-3 promoter
TGGTGAGATTCGTGCGTACC GTCAAAGGACTTCTGTGATATCTATTTCTAACAAGGATACATCGCTCTGTTCAGAAATAGGATCTTGGCAA
TGGCAATAAGTACCAAAAGTATATATTATACGTTATGCAATACATTACAGTTGATACGACTAAGACGTTATTTCCAAGATTTTGACGACAGTG
GGTGGTAGACTTTAGGAAGACACTGGTGTCTATAGTCTCCAGGTAGACGTGTGATGGATGAAAACAGAACAGTCTGGCCAAATAGGGACC
AATAAA TTGTCATGGGATTTCCAGTAAGTCTTTTGCAGTCTGTTGAGCTATCTATACGCCGGCCGGTCAGTTAACTCCTACTCCTAGGCCAG
TTGATAAATACTCTTACTACCACTTAACTAGGCT TTTATT TCAAGCTTATCCAAAGTCCCCTGCCCTGTCACTATTCACTcagtagtctggt
agagattcTAAcAtatttaatttttgtgtgtctCACATGGTGTTTAACATACACTGTTTCAGCATACTCAATCAAACCACGACCTGACG
TCCAGTCACTTCTTATGATACATTACAGCTAtatgcccacaaaaaatcaagaccatagcatgtccagggtcaaaagatacaaaaagtcggtc
tgctgcagtaccaggtcacaaccagggtgccccaaatcaccctgaccttctgtctttgcaatacctaccacatatcaaatatcattataa
ccCAACCAGagttatggtgaccacaaaatccggaaaacacaaacagacacacagcatagacacacacagaaaacacaaaaactatgcctcc
atthttcatggaggtaacaatctGAAGAGACCTCAAAGCTGCAGACGGCAAGTGGCCATGTTACTATGATTGTTAGGTGCAGCTCAAACAGC
CCAGAtgttgacctgtgacctttacTAATTAGCTGTGAGACCGTGTGAGAAAGTGGCTGCAACCGTGTCCAAGAAGTCTACGGATATTG
AGTCAGAGGAAAAAGACTTTGTACAAGTGTGAAGAACTTCTCTATTTTATCATGACATATCTGCTGAGCAGTGGGTAGTCGACATCCC
AGTTGTCATAGCAAGAGGACAGGCTcagatccccccccccccccatacgaTACTGAAGCTGTTCTATATCATACATATCATGCAATTATG
TCTATTGATGATCTAGAAAAGCTCAAAGTCTAGCACGACAAAAGCGGACAGAGTTTGTAAACAACCTTCCGTCTCACTAACTTCGCTGTCT
CCGCACAACCTTAACTGTTATCTGTTTGAAGAAGGCATAGGATTCATTTTATAGAGTGAACAAAGATATCGACAGCGTTTCCCTCAAACCT
GTACCTATAAAAAGAGAACGTACAGACAGGGGTTACTTCCGGTTAAGGccctcttttgataagtgtgTAAAGTGGCGTGGCACTTGAGGA
GGGTTATCAGAAATTGTCAGTAGGCTCGAGGTCAATAACCCCAACCACTTTCACACGAGGACTCTCTCAGGAGATGTCTATTCCGATCTCTG
CAAATAGAATAGACAGGAGAACATTGCTCAGCGGTTACGGAACCGCAGTTTCTGCTTGTATGTCTATGTATGATCCGCGGCGCTGAGTGCCT
TCAGCTGCATATTAACGCCACTGCAATgagtgtgaccttgcattcggttgGTGGATCCCCTGGCCAAGCCATGCGCTATGTCAAAGAGACT
TTAGATATGGTACTttctgctttctgctgcttagagaaagcagcatgtgagCAAGAATTTGAGAAGGAgcatggtgtagtgaacacacaccact
accagtggaactagccccctgctgtaatgattacaaaagtgtgtggtggccggggtacagaaaaggagatgggtgcccgtcctatgcccactc
ggtacGGGAAGGACTTCAACTTACCTTAAGTAATGCCGTGTGGCTGATTAGGGTATCTGGTAACTTTTCGaaaaaaGAGAGTAAAGAGAGGA
AAATTTATTTCCCTCATAAAACCCCAATTAGCATTCCAGCGAAGGTTACACGGCAAGCCATATTTATACAAAAGAACGCCAATCATGTTAAT
ACCAGTATTACATTGATTTTTCATCTTCTGAAGACCTTTGATATACCAATAGTACAATAAGCCGTAAAGTTCGAACGTTCCGAACACCA
GGTACCCTAGCAGAAAAGAGATGAACTTTTCAGTGTATCAAACATTGTGTTTTCATGCAATGACTTTCagataatthttttcaagac
gTTAGGTATAccacgggggggggggggggttctgttGTGGTCCACAGAGAGACAGACGTGCGAGTTAAAAACTACGCGCACTACTGTCTAC
TGACCCCGACAGAAACCTTCAGTATACCCCGACCTACGTGGTGGGTAGCGGAGTGTGACAGAGACCCGGGTTTGACCCCTTGCACCTCAGC
AACACATCCAATG TTTATT TGCTTCTCTCGTATTCTGGGAATCTAGCAAAGATGACTAGTCGTGACAGCGCCCTGGCAGATAATTGACAGG
CCACTACAATTAAGTCCGTGTGTGAGCGATGGGCCGGGAGGTCAGGCAGGGCGTGGCTTGTCTTGTCTGTCTGAGTGGGAGGTTTGCAC
TCAGGAGCCAGCAAATAACGTAACATGTCCGTTCACCTCCACTGTCAATATTCAATTAGTCCCTCGGCGCGGCGATTGCACCTAT TTTATTAG
CCGCGTCCAATTGGACCTCTGTGAGCCAATCGGCGTGGTCAATTACAGGCGGTTCTTCAGACGACAAGGGCGGTGGCCCTGTGGGAGGTACG
ACTGTGACGTAATCGGGCGGCCCGTGTGTCGATCCAACGAGCTGAACCTCTGAATGAACCACTTATTAATGACATTGCGTGTCTGACG
GGGG AATAAA TACGAGTGTAGTGGTTCTCGAGCAGCCATTCTTGCCAGCCTCTGTTCTACACGAGAGGAGACTCTCTGCTCCGCCCAA
CCCCGGACCAGGCTCCCGTCGTATCACAGCTTGTAAACCAGCCACAAGA CTCTCTCGTGGTGGCAAAATG
```

Figure S14. DNA sequence upstream of start codon of *B. floridae Cyp26a-3* gene. Red ATG = start codon; blue text = 5' untranslated region; blue highlight = putative Cdx binding site; underlined blue highlight = binding sites chosen for mutagenesis; yellow highlight = cloning primers.

The 3.1 kb region was cloned into pGL3 basic vector (Promega) between *SacI* and *NheI* sites, and the resultant construct ('Cyp26-3 promoter') was used to make mutant constructs with one or two putative Cdx binding sites mutated ('BS1 mutation', 'BS2 mutation' and 'BS1,2 mutation') using a PCR method (Table S3). Controls were empty pGL3 vector and not injected samples.

Genes	Primer sequences (5'→3')
<i>Cloning primers</i>	Forward: TGGTGAGATTCGTGCGTACC Reverse: CTCTCTCGTCGGTCGGCAA
<i>Binding site 1 mutagenesis primers</i>	Forward: TCTGGCCAAATAGGGACCGAcgAcTTGTCATGGGATTCCAG Reverse: CTGGAAATCCCATGACAAGTcgTcGGTCCCTATTTGGCCAGA
<i>Binding site 2 mutagenesis primers</i>	Forward: CGCGGGCATTGCACTTATcTgcTgAGCCGCGGTCCAATTGGA Reverse: TCCAATTGGACCGCGGCTcAgcAgATAAGTGCAATGCCCGCG

Table S3: Primers used for cloning and mutagenesis of *Cyp26-3* promoter sequence

Injection solutions were prepared containing 3 ng/μL Renilla luciferase vector pRL-TK (Promega), 20% glycerol, 5 mg/ml Texas Red dextran, with or without 30ng/μL each of above luciferase constructs. Microinjection into unfertilized amphioxus eggs was conducted as previously described (Liu et al. 2013). For each experiment, ~60 embryos were collected at 16 hours post fertilization. Wild type embryos from the same batch were also collected and used as a negative control ('WT'). Levels of luciferase and Renilla were detected with the Dual Luciferase Kit (Promega Co.) using a GloMax luminometer with an integration of 10 seconds. The level of luciferase activity was normalized to the level of Renilla activity for each experiment. All experiments were repeated three times (Table S4).

Construct	<i>Cyp26-3</i> promoter	BS1 mutation	BS2 mutation	BS1,2 mutation	pGL3	WT
RFV1	16.0989	8.0092	8.3821	4.8000	0.5515	0.4338
RFV2	20.2335	4.8345	6.5707	3.1083	0.4145	0.3469
RFV3	11.6292	7.4662	5.4355	4.8504	0.4836	0.3582

Table S4: Raw values of relative luciferase for each construct and replicates RFV1 to RFV3.

SECTION 7: DIFFERENTIAL EXPRESSION OF CANDIDATE TARGET GENES

DNA sequences of supercontigs referred to below are in Supplementary Data.

(a) Analysis of GFP genes affected by mutation of amphioxus *Pdx*

In *Pdx* 4Δ/11Δ mutants, we found down-regulation of reads mapping to 11 contigs from the GFP gene family (GE_G14886, FE_G16436, NOVEL_103923, NOVEL_102722, NOVEL_50813, ML_G28518, FE_G16645, FE_G16414, ML_G19054, ML_G18969, NOVEL_74108). To analyse whether these are variants of the same amphioxus gene, or multiple genes, blast align (nucleotide vs nucleotide) (<https://blast.ncbi.nlm.nih.gov/Blast.cgi>) was used to search a database of DNA sequences of the open reading frames of the 13 amphioxus GFP open reading frames described by Li et al. (2009).

Results of blast align, compared to expression fold change data from DEG analysis:

Contig	Best hit	Weaker hits	Log 2 fold change	adjp
>FE_G16436 (short)	GFP-8	GFP-11	-1.59	3.05E-10
>NOVEL_103923	GFP-13	GFP-10,8,12	-2.08	4.00E-20
>NOVEL_102722	GFP-13	GFP-10,8,12	-2.21	1.21E-20
>GE_G14886 (short)	GFP-1	GFP-2	-0.64	0.009
>NOVEL_50813	GFP-10	GFP-8, 13, 12	-0.68	0.002
>ML_G28518	GFP-10	GFP-8,13,12	-0.68	0.004
>FE_G16645	GFP-8	GFP-10,13,12	-0.95	7.23E-07
>FE_G16414	GFP-10	GFP-12,9,1	-1.06	0.04
>ML_G19054	GFP-10	GFP-8,13,12	-1.26	2.59E-09
>ML_G18969	GFP-13	GFP-8,10,12	-2.05	5.78E-05
>NOVEL_74108	GFP-13	GFP-10,8,12	-3.06	1.30E-14

Table S5: GFP contigs affected in *Pdx* mutant transcriptomes

The most likely identities for these contigs are *GFP-8*, *GFP-10* and/or *GFP-13*. The single match to *GFP-1* is weaker and derived from a short contig so is less reliable. It should not be concluded that all three best hit genes (*GFP-8*, *GFP-10* and *GFP-13*) are changing in expression level, however, because four amphioxus genes - *GFP-8*, *GFP-10*, *GFP-12* and *GFP-13* - have highly similar nucleotide sequences and are the product of recent tandem gene duplication (Li et al. 2009). Short read sequence data, as generated in this study, cannot be unambiguously assigned to a particular gene and reads will be split by multimapping between them. We conclude that one, or more, of the genes *GFP-8*, *GFP-10*, *GFP-12* and *GFP-13* has been down-regulated in expression following *Pdx* mutation.

(b) Analysis of insulin signalling pathway genes affected by mutation of amphioxus *Pdx*

- i. Signalling peptides. We found three contigs with partial sequence similarity to insulin-like peptide (*ILP*) genes (FE_G15481, GE_G16157 and FE_G15511); these are expressed ~2 to 11 fpm. NCBI blastn analysis vs nr database (31/7/19) suggests the first two represent genes related to *ILP*; the third (FE_G15511) is true *ILP*. None of these contigs showed significant expression level changes in *Pdx* mutants.
- ii. Binding proteins. Supercontig M_G27744 encompassed three related contigs (below), each with stretches of 100% identity to a gene annotated as insulin-like growth factor binding protein 7 (*IGFBP7*) in *B. belcheri* (XM_019790737, LOC109486832) and its homologue in *B. floridae* (XM_002608818.1). These contigs showed clear down-regulation in *Pdx* mutants. A second supercontig FE_G33063 also matched this gene and showed similar down-regulation.

Contig	Transcript	Best hit	Log 2 fold change	adjp
ML_G27744	MLTU49440	IGFBP7	-0.67	0.002
	MLTU49442	IGFBP7	-0.73	0.0008
	MLTU49437	IGFBP7	-0.70	0.004
FE_G33063		IGFBP7	-0.62	0.0015

Table S6: IGFBP contigs affected in *Pdx* mutant transcriptomes

However, there are additional contigs with sequences matches to IGFBP genes (for example, FE_G30801, GE_G21098, ML_G26668) or IGFBP acid labile subunits (GE_G18919, ML_G15423) that do not show differential regulation

- iii. Receptors. Contig FE_G26674 encoding insulin-like peptide receptor (100% blastx hit to *B. floridae* ILP receptor BRAFLDRAFT_128184 XP_002585764.1) showed up-regulation in *Pdx* mutants (log 2 fold change 0.49, adjp = 0.009). The *B. lanceolatum* orthologue of this gene (second blastx hit), has been shown to bind peptides of insulin and an ILP analogue when expressed in cell culture, supporting its designation as an insulin-like peptide receptor (Pashmforoush et al. 1996).

(c) Analysis of iLBP genes affected by mutation of amphioxus *Cdx*

In *Cdx*^{-/-} mutants, we found 1.6- to 13.5-fold down-regulation of reads mapping to eight contigs from the intracellular lipid-binding protein (iLBP) gene superfamily which in vertebrates include CRABP (Cellular retinoic acid-binding protein), CRBP (cellular retinol-binding proteins) and FABP (Fatty acid-binding proteins). These contigs are FE_G16051, GE_G22703, ML_G19061, NOVEL_10269, NOVEL_50856, NOVEL_65559, NOVEL_74162, NOVEL_82533. To analyse whether these are variants of the same amphioxus gene, or multiple genes, BLASTX was used vs NCBI nr (30-31/7/2019). This revealed that seven contigs represent the gene *iLBP4* and one represents *iLBP6*.

Contig	Best hit	Log 2 fold change	adjp
FE_G16051	iLBP-4	-2.196	1.54E-22
GE_G22703	iLBP-4	-2.06	4.20E-15
ML_G19061	iLBP-4	-2.40	2.92E-09
NOVEL_50856	iLBP-4	-2.11	5.05E-14
NOVEL_65559	iLBP-4	-2.42	9.08E-61
NOVEL_74162	iLBP-4	-2.67	1.06E-14
NOVEL_82533	iLBP-4	-3.77	0.015
NOVEL_10269	iLBP-6	-0.699	0.002

Table S7: Intracellular lipid-binding protein contigs affected in *Cdx* mutant transcriptomes

The first seven contigs match *iLBP-4* of Albalat et al (2011), NCBI XP_002607338. The same gene was named *CRABP* by Jackman et al. (2004), NCBI AAQ72814.1, but does not group more closely with vertebrate *CRABP* in phylogenetic trees when a diversity of amphioxus *iLBP* genes is included (see Supplementary Figure S6 of Albalat et al 2011). The last contig matches *iLBP-6* of Albalat et al (2011), NCBI XP_002607336.

(d) Hox gene expression in wild type and *Cdx* mutants

In previous work from our laboratory and that of H.V. Isaacs (University of York, UK), we showed that disruption on *Cdx* function in *Xenopus tropicalis* has a colinear-like effect on Hox gene expression (Marlétaz et al. 2015). In addition to the expected down-regulation of posterior Hox genes after *Cdx* disruption, we detected higher expression of anterior Hox genes (consistent with *Cdx* genes activating posterior Hox genes and repressing anterior Hox genes in normal

development). We wished to test if an analogous colinear-like relationship to *Cdx* occurred in amphioxus.

First, we note that absolute levels of expression differ greatly between Hox genes in wild-type embryos, as estimated from transcriptome read mapping (calculated here from the *Cdx* experiment 42h samples). Only *Hox1*, *Hox3*, *Hox4* and *Hox6* have expression levels above 10 fpkm; *Hox2*, *Hox5* and *Hox7* have fewer reads counts and the more ‘posterior’ paralogy group genes barely any (Table S7; Figure S15).

Hox gene	Contig	Mean fpkm wild type	Mean fpkm <i>Cdx</i> mutant	adjp	Log2FoldChange
<i>Hox1</i>	HOX1	16.51668	23.66045	0.000546	0.518436
<i>Hox2</i>	HOX2	1.85504	1.780453	0.999998	-0.05927
<i>Hox3</i>	HOX3	22.38144	21.47403	0.999998	-0.05965
<i>Hox4</i>	AB028208.1	28.09742	23.44197	0.192511	-0.26125
<i>Hox5</i>	HOX5	1.842797	0.953663	0.672645	-0.9493
<i>Hox6</i>	HOX6	24.91482	14.74147	6.16E-05	-0.75688
<i>Hox7</i>	FE_G13616	0.241352	0.063064	0.073839	-1.93772
<i>Hox9</i>	FE_G13248	0.059181	0.057971	NA	-0.02971
<i>Hox15</i>	FE_G13062	0.043841	0.152279	0.266805	1.795219

Table S7: Hox gene expression changes in *Cdx* mutant transcriptomes

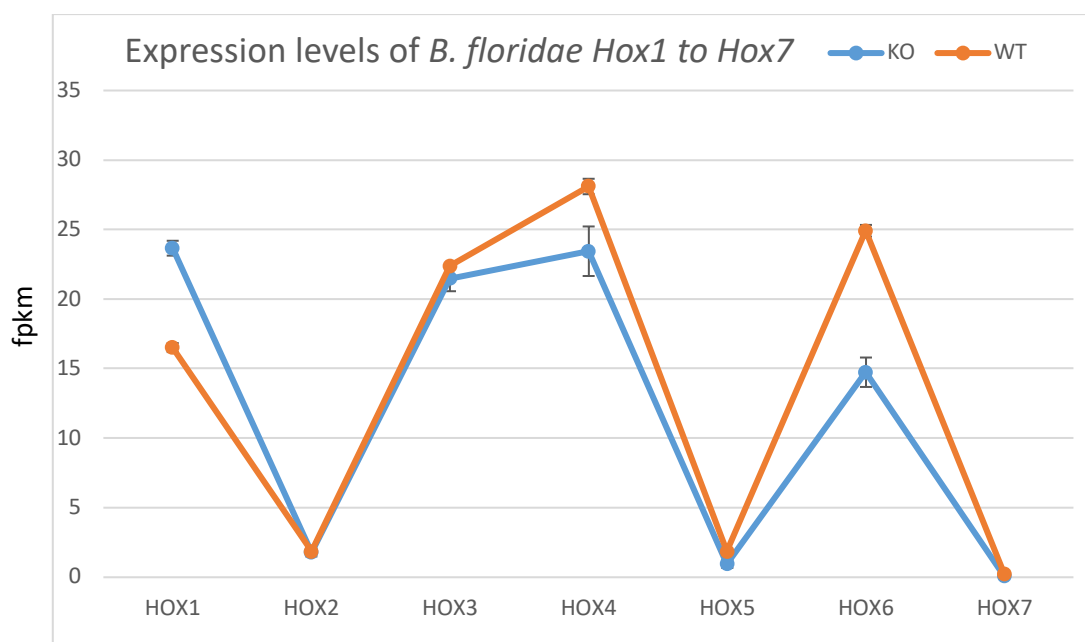


Figure S15: Expression levels (FPKM) of Hox genes in *B. floridae* determined from RNAseq data generated in the present study. Blue line, homozygous mutant; orange line, wild type and heterozygotes.

This finding is consistent with in situ hybridisation experiments to wild type *B. floridae* embryos, which detected strong signals for *Hox1*, *Hox3* and *Hox4*, but very weak expression of *Hox2* (Wada et al. 1999), although contrary to the stronger *Hox2* pattern reported by Schubert et al (2006). The read counts reported here for *B. floridae* are also consistent qualitatively with *B. lanceolatum*

transcriptome data (from a slightly later developmental stage) reported by Marlétaz et al. (2018), extracted and plotted below, apart from higher expression of *Hox1* in *B. lanceolatum* (Figure S16).

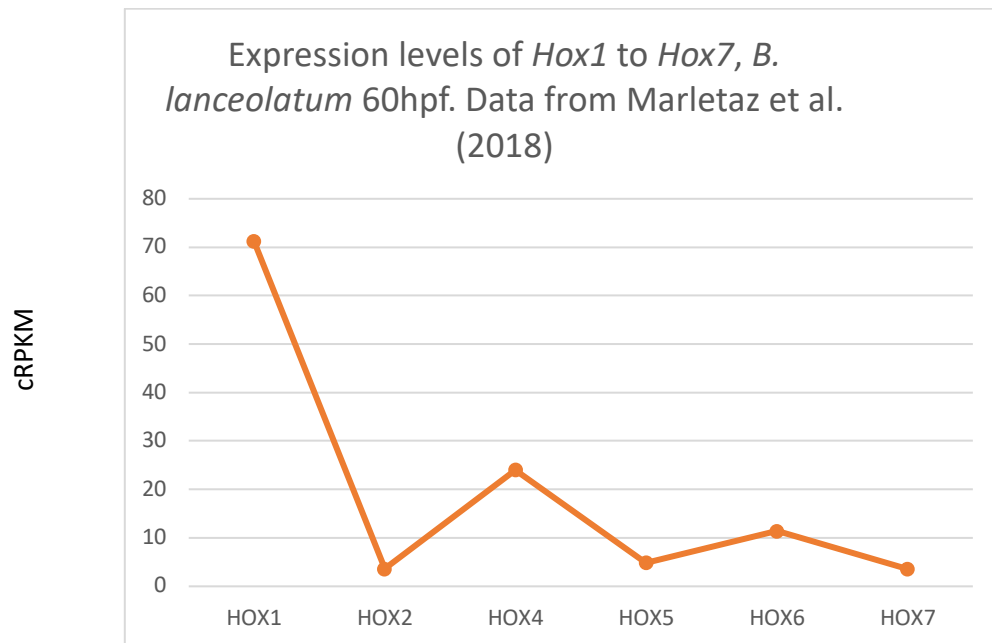


Figure S16: Expression levels (cRPKM) of Hox genes in *B. lanceolatum* determined from RNAseq data of Marlétaz et al. (2018)

Second, we detect a colinear-like response of Hox genes to mutation of *Cdx*, with paralogy group 1 gene expression higher in mutants, *Hox2*, *Hox3* and *Hox4* unaffected, *Hox5* and *Hox6* mildly down-regulated, and *Hox7* strongly down-regulated ((Table S7; Figure S15). Only the *Hox1* and *Hox6* expression changes are significant when each genes is considered in isolation; considering genes as a cluster and plotting mean changes collectively reveals a significant negative slope to the response (Figure S17).

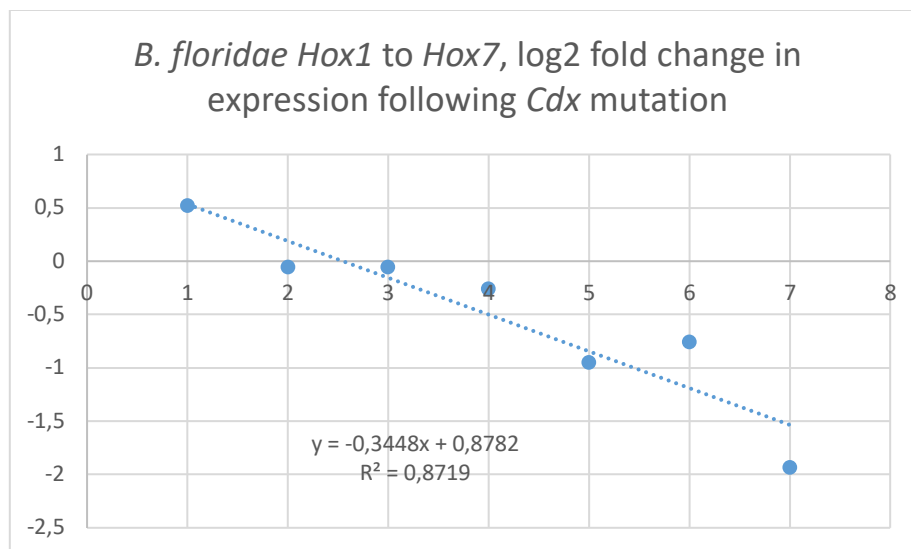
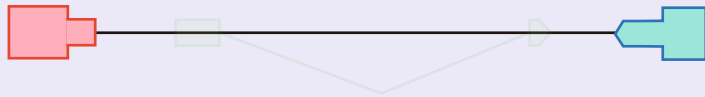


Figure S17: Expression level change (fold change log 2 scale) in Hox gene expression in *Cdx* homozygous mutants compared to wild type and heterozygotes.

Supplementary References

- Albalat R, Brunet F, Laudet V, Schubert M (2011) Evolution of retinoid and steroid signaling: vertebrate diversification from an amphioxus perspective. *Genome Biol Evol* 3: 985-1005
- Amin S, Neijts R, Simmini S, van Rooijen C, Tan SC, Kester L, van Oudenaarden A, Creyghton MP, Deschamps J (2016) Cdx and T Brachyury co-activate growth signaling in the embryonic axial progenitor niche. *Cell Reports* 17: 3165-3177
- Ferrier DEK, Dewar K, Cook A, Chang JL, Hill-Force A, Amemiya C (2005) The chordate ParaHox cluster. *Current Biology* 15: R820-822
- Jackman WR, Mougey JM, Panopoulou GD, Kimmel CB (2004) *crabp* and *maf* highlight the novelty of the amphioxus club-shaped gland. *Acta Zoologica* 85: 91-99
- Li G, Zhang QJ, Zhong J, Wang YQ (2009) Evolutionary and functional diversity of green fluorescent proteins in cephalochordates. *Gene* 446: 41-49
- Li G, Feng J, Lei Y, Wang J, Wang H, Shang L.-K., Liu D, Zhao H, Zhu Y, Wang Y (2014) Mutagenesis at specific genomic loci of amphioxus *Branchiostoma belcheri* using TALEN method. *J. Genet. Genomics* 41: 215-219.
- Li G, Liu X, Xing C, Zhang H, Shimeld SM, Wang Y (2017) Cerberus-Nodal-Lefty-Pitx signaling cascade controls left-right asymmetry in amphioxus. *PNAS* 14: 3684-3689
- Liu X, Li G, Feng J, Yang X, Wang Y-Q (2013) An efficient microinjection method for unfertilized eggs of Asian amphioxus *Branchiostoma belcheri*. *Dev Genes Evol.* 223: 269–278
- Marlétaz F, Maeso I, Faas L, Isaacs HV, Holland PWH (2015) Cdx ParaHox genes acquired distinct developmental roles after gene duplication in vertebrate evolution. *BMC Biology* 13: 56
doi.org/10.1186/s12915-015-0165-x
- Pashmforoush M, Chan SJ, Steiner DF (1996) Structure and expression of the insulin-like peptide receptor from amphioxus. *Molecular Endocrinology* 10: 857-866
- Schubert M, Holland ND, Laudet V, Holland LZ (2006) A retinoic acid-Hox hierarchy controls both anterior/posterior patterning and neuronal specification in the developing central nervous system of the cephalochordate amphioxus. *Dev Biol.* 296: [doi.org/190-202](https://doi.org/10.1016/j.ydbio.2006.04.457). 10.1016/j.ydbio.2006.04.457
- Wada H, Garcia-Fernández J, Holland PWH (1999) Colinear and segmental expression of amphioxus Hox genes. *Devel Biology* 213: 131-141

Discussion



Discussion

1. Expanding the tools for understanding Evolution

Before the dawn of Evo-Devo (Müller 2007), the genomic regulatory mechanisms that took place during development were not incorporated into the study of evolution. Once their role during development could be related to the phenotype of the adult organism, they became a great tool to elucidate the underlying mechanisms of evolution. In order to understand not only their role during evolution, but how these mechanisms work, we should identify their phylogenetic origin. This will provide not only the most ancestral-like version of the process studied, which may be more suitable to work with, but also will inform of the best fitted model organisms to use.

However, pinpointing the origin of a mechanism may be difficult in some cases, and this is particularly true when, for example, long noncoding RNAs are involved. LncRNAs are particularly difficult to analyse from an evolutive point of view in most of the cases. The main reasons are that they have a very low rate of sequence conservation (Diederichs 2014; Jathar et al. 2017) and that they tend to be subjected to huge turnover (Neme and Tautz 2016). Nonetheless, when some degree of conservation is found in something so variable as the lncRNAs, like the one presented in this work, it implies that a strong selective pressure is at play, hence studying it may well be giving interesting insights into the roads of regulatory evolution

Microsynteny is another example of the kind of conservation in elements otherwise seemingly volatile. Irimia *et al.* (2012) found syntenic conservation of gene pairs over large evolutionary distances and demonstrated that the linkage was due to regulatory constraints. Before those findings, the assumption that gene order was not under selective pressure kept us out from taking advantage of this feature to broaden our understanding of evolution.

These are just some of several examples (with more to come for sure) where finding underlying conservations helped to better identify the possible origins and reconstruct the formation and mechanics of genomic regulation processes that take place during development and shape the different organisms through evolution. In this work, one of these was analysed in further detail over long evolutionary distances on the road to vertebrates, including us humans.

2. Conserved lincRNAs within Chordata

2.1. *lincRNA fraction in amphioxus*

The identification of bona fide lincRNAs is, in most cases, a process of successive filtering that separates the coding from the non-coding fraction within the transcriptome (Pauli *et al.* 2012). However, these filters must be applied properly in order to be restrictive enough to discard all the probably-coding transcriptomic artefacts while identifying most of the true lincRNAs. In our case, more than 1500 loci were classified into one of the four different lincRNA classes assigned by our pipeline. This is around 1,5% of all the loci found in *B. lanceolatum*, a small percentage at first sight. But these 1500 lincRNAs were obtained after a final multiexonic filtering. If we take all the monoexonic plus multiexonic transcripts into account, more than 31000 loci are classified as lincRNAs, around 34% of the loci in *B. lanceolatum*. Nonetheless, we decided to restrict the analysis to the spliced lincRNAs to avoid any potential artefacts related with the RNA-seq and transcriptome assembly protocols, even if it implied losing such a huge number of candidates.

When looking at the tissue specificity of this lincRNA dataset (considering tissue specific any gene expressed more than twice in one tissue respect the others (Zhu *et al.* 2016) and having more than 5 cRPKM (corrected Reads per Kilobase per Million reads) in said tissue), we found that, as expected, gonads are the most enriched tissue in lincRNAs, although the neural tissue ranks fourth (Figure D1). This is not what we would had expected in terms of tissue specificity, as normally gonads and neural tissue are the most enriched tissues in lincRNAs (Necsulea *et al.* 2014). This enrichment may be caused by an overrepresentation in the RNA-seq data of gills, cirri and epidermal tissues, a cephalochordate-specific expansion of the lincRNAs in these tissues, or alternatively because additional levels of neural gene regulation are present in vertebrates but not in amphioxus, thanks to the higher presence of lincRNAs in the neural derivatives in vertebrates (Clark and Blackshaw 2017). The last option would be similar to the findings reported by Irimia *et al.* (M. Irimia *et al.* 2011; Manuel Irimia *et al.* 2009) and Burguera *et al.* (2017; Burguera Hernández 2017), where alternative splicing in neural tissues was shown to be dramatically increased at the origins and during the evolution of vertebrates, hence increasing the level of regulatory complexity present in this tissue.

Discussion

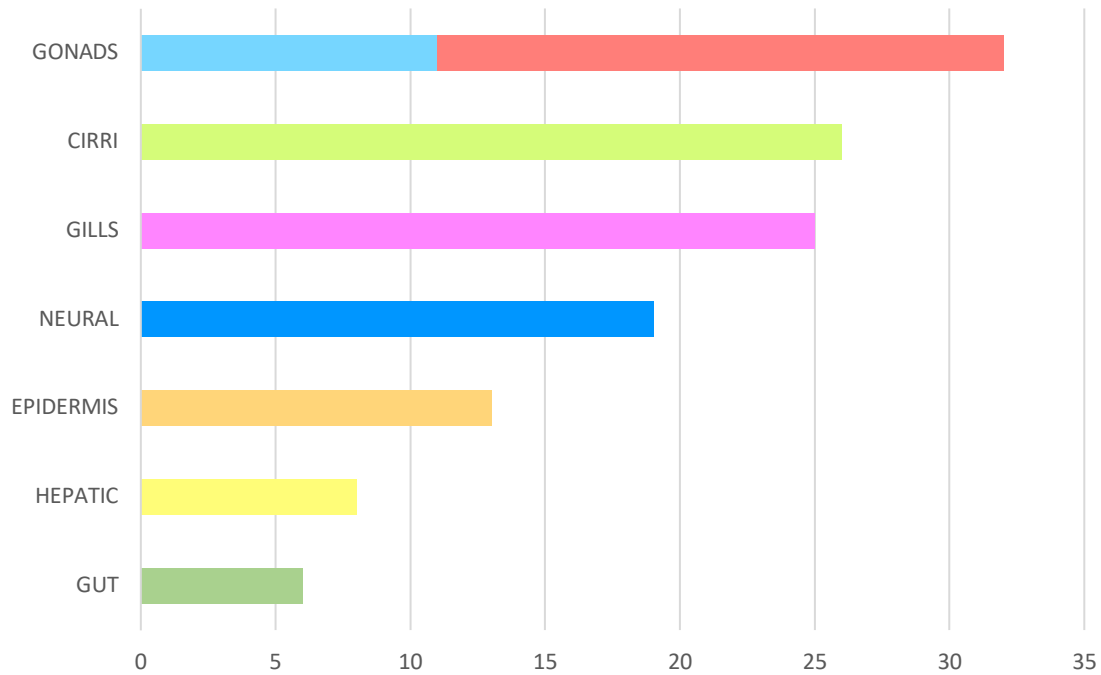


Figure D1. Number of tissue-specific multiexonic lincRNAs in *B. lanceolatum* in each of the 9 tissues analysed. Male (red) and female (light blue) gonads merged.

2.2. Novelty of the LincOFinder pipeline

Trying to assess the interspecific conservation of lincRNAs is problematic due to the inherent properties of these molecules (Diederichs 2014). Even between mouse and human, when comparing certain homologous lincRNAs, the conservation ratio is even lower than the mean ratio present in intronic regions (Han *et al.* 2014; Sorek and Ast 2003). Thus, if orthology relationships are being interrogated, taking into account the conservation of the nucleotide sequence is only useful if working with very closely related species.

LincRNAs apparently obtain their function through folding themselves into a secondary structure (Zampetaki *et al.* 2018). Although this would in theory allow them to maintain functional homology without conserving the nucleotide sequence, recent studies found that even the secondary structure seems not to be deeply conserved in orthologous lincRNAs (Rivas *et al.* 2017). In addition, carrying out secondary structure homology analyses, requires huge horsepower of computation hardware not easily available.

All these problems seemed to be solved by looking for conserved lincRNAs through synteny, performed by using BlastZ genomic alignments to detect conserved regions flanking a lincRNA in two different species (Ounzain *et al.* 2015). This solution has a main

handicap: the longer the evolutionary distance between the two species, the more difficult it becomes to find potential syntenic regions and hence orthologous lincRNAs. That is why we tried to spin the last concept a little bit to create a tool that can be used for virtually any two species with good quality genomic and transcriptomic data.

The main advantage of LincOFinder over other methods based on sequence conservation is that it relies on microsyntenic conservation and a proper establishment of interspecific orthology relationships, which are more evolutionary constrained than the highly mutable nucleotide sequences of lincRNAs, but at the same time more tolerant than genomic BlastZ alignments. This is the reason why lincOFinder can help to uncover conserved lincRNAs over deep evolutionary distances, in any species for which proper gene annotation data is available.

From our analyses, we identified, for the first time to our knowledge, *bona fide* homologous lincRNA between invertebrates (amphioxus) and vertebrates, which means 450 Myears of evolutionary distances

2.3. Origins of *Hotairm1*

From the list of putatively orthologous lincRNAs obtained by LincOFinder in *B. lanceolatum*, two of them were located on the Hox cluster and one, *Hotairm1*, was already annotated in some vertebrates. The bibliography allowed us to confirm that after its first identification in human cells, it was later found in several vertebrate lineages such as mammals (Yu *et al.* 2012), birds and reptiles (Gardner *et al.* 2015). These characteristics made us focus on the lincRNA *Hotairm1* and try to assess the extent of its conservation.

In most of the vertebrates analysed there were traces of *Hotairm1* presence, like an EST in *Xenopus*, reads mapping the genomic area between *Hoxaa1* and *Hoxaa2* in medaka, or a fragment of a conserved RNA motive in spotted gar. In the particular case of zebrafish, no *Hotairm1* evidence was found. This could be due to a lineage-specific loss as it is present in another teleost, i.e. medaka, and a secondary effect of the microsynteny breakage of the *Hoxa1* and *Hoxa2* paralogy groups that took place in zebrafish (Kurosawa *et al.* 2006). In the case of the cyclostomes *E. burger* and *P. marinus*, we could not find either any trace of *Hotairm1* within their genomes. As in zebrafish, the microsynteny is lost in this region, but in these organisms, we cannot

Discussion

confirm if the missing lincRNA could not be detected due to a lineage-specific loss or alternatively due to an incomplete lincRNA annotation (Pascual-Anaya *et al.* 2018). Finally, the Hox cluster disintegration that took place in tunicates (Sekigami *et al.* 2017) made the search for a microsyntenic lincRNA unavailable. Furthermore, due to the absence of an antisense transcript 5' of *Ciona intestinalis Hoxa1*, we could not confirm the presence or absence of *Hotairm1* in this chordate subphylum.

All the data obtained strongly suggests that *Hotairm1* was retained within the HoxA cluster after the two vertebrate-specific rounds of genome duplication (Dehal and Boore 2005; Pascual-Anaya *et al.* 2018), as in every species where *Hotairm1* was found, it was located within this cluster. We can also conclude that it was present at least in the last ancestor of chordates, with an origin that predates, at least, the appearance of extant chordate lineages more than 500 million years ago. We also checked in other invertebrate species to try to pinpoint the deepest origin of this lincRNA, but so far, we only found its presence in chordates, hence the current working hypothesis is that it originated earlier in, or close to, the phylum Chordata.

This would make *Hotairm1* the most deeply conserved *bona fide* lincRNA identified to date (along with the other lincRNAs identified in this work). And although *Hotairm1* is a well-known (at least in mammals and birds) conserved lincRNA, its finding among the orthologous lincRNAs set, is a sound indicative that LincOFinder works as intended.

2.4. Conservation on the expression between frog and lancelet *Hotairm1*

Though the function of *Hotairm1* has been thoroughly studied in several cancers (Esfandi *et al.* 2019; Li *et al.* 2018; Ren *et al.* 2019), all of the previous studies used human cell cultures as a model. This meant that there were not expression domains of *Hotairm1* in the bibliography to compare with the expression obtained in amphioxus. The areas where *Hotairm1* was detected in amphioxus (Figure RII-3) were concordant with the available functional data of *Hotairm1* in vertebrates, which claims that this lincRNA is expressed during development in a short time window coincident with neural development, and that it plays a role in the regulation of *Hoxa1* (*Hox1* in amphioxus) (Zhang *et al.* 2009; Lin *et al.* 2011; Wang and Dostie 2017). However, these promising equivalences were not strong enough to make a claim such as that this lincRNA is conserved between amphioxus and humans, as not developmental expression of

Hotairm1 was analysed by in situ hybridisation in any vertebrate. Therefore, we decided to extend our analysis to another vertebrate, where *hotairm1* was not even annotated, the frog *Xenopus tropicalis*.

As can be seen in Figure RII-3, the expression domain of *Hotairm1* in amphioxus and *Xenopus* resemble each other, and both are concordant with the aforementioned characteristics of this lincRNA found in the bibliography. In the embryos of amphioxus and *Xenopus* we detected *Hotairm1* in the anterior half of the developing neural tube. In addition, both (*Xenopus*'s and amphioxus's) could be interacting with *Hox1* and *hoxa1* expression domains, respectively, judging by the partial overlap of the expression domains. Furthermore, as seen in Figure D2, in both cases there is a peak of expression at neurulation that decays shortly after.

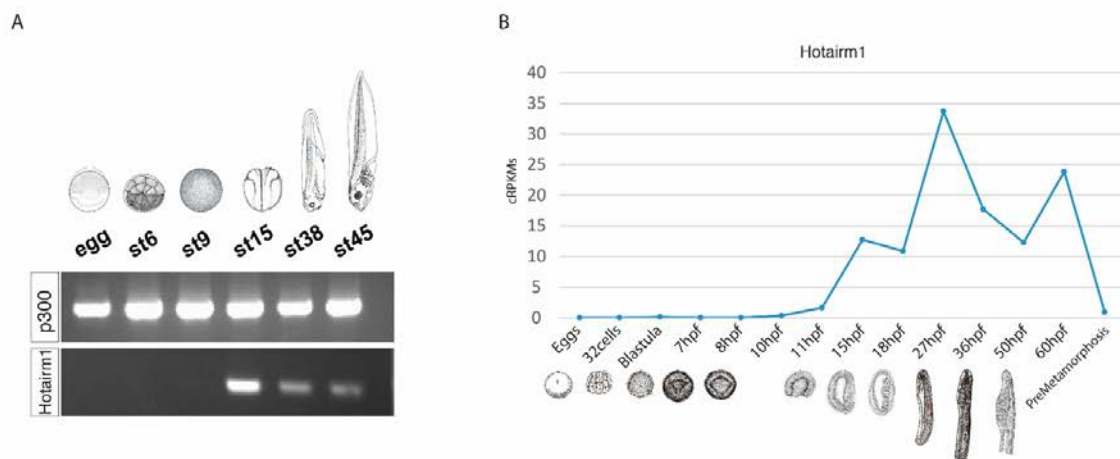


Figure D2. Expression profiles of *Hotairm1* in *Xenopus* and *B. lanceolatum*. **(A)** Expression levels of *hotairm1* determined by PCR in *Xenopus* through six timepoints with *p300* (a known gene expressed through development) as a positive control. **(B)** Expression levels of *Hotairm1* determined by RNA-seq in amphioxus through 14 timepoints. In both cases the peak of expression takes place in the Neurula state. Amphioxus glyphs adapted from Conklin 1932.

In humans, *Hotairm1* has at least two functional isoforms, one spliced and one unspliced (Wang and Dostie 2017). We found this extra unspliced isoform in *Xenopus* (Figure RII-4) but we were unable to amplify it by PCR in amphioxus. When looking at the RNA-seq data, there were several reads that could support the existence of this unspliced isoform in *Branchiostoma lanceolatum*, but they had a low depth. As for now, we still cannot firmly confirm nor reject the existence of this isoform in amphioxus. We should take into account, however, how low the overall levels of expression of *Hotairm1* are in amphioxus to consider if this low depth of supporting reads could be an indicator that the unspliced isoform exists, at very low levels, in amphioxus.

Discussion

Indeed, the resulting data from the expression analysis of *Hotairm1* in *Xenopus* and amphioxus, together with the current bibliography about this lincRNA, strongly supports the hypothesis that *Hotairm1* expression is conserved between *Xenopus tropicalis* and *Branchiostoma lanceolatum*.

2.5. Conservation of the mechanism of function between frog and human *Hotairm1*

Due to the technical limitations for generating *B. lanceolatum* knockouts, we decided to use the frog *Xenopus* for inferring the function of the lincRNA *Hotairm1* during development. Despite the usual criticism towards working with morpholinos due to the problems associated to this tool, it is an effective and well established mechanism of knocking-down genes in *Xenopus* (Eisen and Smith 2008; Heasman *et al.* 2000). Besides, our results indicate that what was mainly affected in our experiments was, as intended, the splicing of *hotairm1*.

As our data indicates in Figure RII-4, there is a clear switch in isoforms during *Xenopus* embryogenesis. Current bibliography suggests that, at least in cultured cells, *Hotairm1* influences the remodelling of chromatin in the anterior and middle part of the HoxA cluster (Wang and Dostie 2017), with the spliced isoform repressing middle HoxA genes and the unspliced isoform promoting anterior HoxA genes. According to that model of *Hotairm1*, we expected an upregulation of anterior HoxA genes, as more unspliced *hotairm1* is being produced. At the same time, we should observe an upregulation of middle HoxA genes, as the spliced isoform is no longer there to close the chromatin via the chromatin remodeller PRC2.

But then again, our observations in *Xenopus* slightly differ to our expectations. First, the anterior HoxA genes are not affected when morpholino-driven knockouts were produced (Figure RII-6). This could be explained by a chromatin accessibility level already at peak in the anterior HoxA genes under control conditions. In this case, increasing the amount of unspliced isoform would not increase the chromatin accessibility, thus we would not detect an upregulation of anterior Hox genes. In fact, a similar result was obtained by Wang & Dostie (2017) when inhibiting most of the spliced isoform. The unspliced isoform is also inhibited to a certain degree, but it is at an enough level to maintain the anterior HoxA cluster chromatin in an open state, making the expression of anterior HoxA genes invariant. Second, while most of the middle HoxA genes are

upregulated, *hoxa4* is downregulated. Several explanations could fit here, one being that in *Xenopus*, *hotairm1* acts as a transcription factor for *hoxa4*, or that the overexpression of *hoxa5* is coupled with an expansion of its expression domain that results in a downregulation of *hoxa4*. Unfortunately, these hypotheses could not be tested in the time being.

Slight differences apart, the *hoxa4* downregulation is coherent with the phenotype obtained as it presents an expansion of the anterior neural tube, prior to the hindbrain, in the developing embryo, which causes the headless phenotype. This could be the result of a cascade event, where the reduction of *hotairm1* spliced isoform expression domains generates a reduction of *hoxa4*, *otx2* (forebrain-midbrain boundary) (Gat-Yablonski 2011; Hidalgo-Sánchez *et al.* 2005), and *engrailed* (midbrain-hindbrain boundary) (Hidalgo-Sánchez *et al.* 2005) leading possibly to the anterior expansion of the anterior neural tube (posterior to the hindbrain) (Figure D3). Consistently, all these results make us consider that the mechanism of action is, at least in some levels, conserved between *Homo sapiens* and *Xenopus tropicalis*.

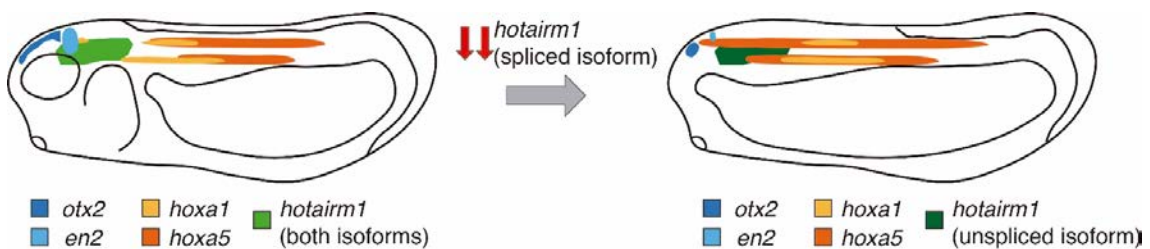


Figure D3. Schematic hypothetical representation of expression domains of *otx2*, *engrailed2*, *hoxa1*, *hoxa5* and *hotairm1* in *Xenopus* before and after the knock down of the spliced *hotairm1*.

2.6. Conservation of *Hotairm1* across Chordata

When we look at the expression of *Hotairm1*, there is a clear resemblance in both, the expression domain and the relationship with *Hox(a)1* between amphioxus and *Xenopus*. In addition, when we analyse the knocked-down embryos for the spliced isoform in *Xenopus* we obtained very similar results to the ones previously obtained by Wang & Dostie (2017) in human cells. *Xenopus* and mammalian development, including human development, is largely conserved, especially with regards to the neural tube closure and expansion. Thus, we can fairly conclude that *Hotairm1* regulation of the Hox cluster is conserved within gnathostomes, and probably vertebrates or even chordates. This

Discussion

backs up the hypothesis that *Hotairm1* appearance happened around the first chordate lineages and was maintained across evolution possibly due to its strong relationship with *Hox(a)1*, one of its known targets in humans (Li *et al.* 2018; Wang and Dostie 2017; Zhang *et al.* 2009). Even if we lack the supporting evidence of its presence within cyclostomes and tunicates, those cases could still be explained. As we stated, in cyclostomes because of an incomplete annotation due to the narrow time-window in which this lincRNA is expressed. In *C. intestinalis*, with the changes in collinearity expression suffered after the disintegration of the Hox cluster (Ikuta *et al.* 2004), *Hotairm1* could have lost its selective pressure, yielding to the high rates of turnover characteristics of lincRNAs (Neme and Tautz 2016) eventually being lost.

Summarizing, taking into account the fact that the amphioxus genome represents a good proxy to the ancestral chordate genome, we are prone to propose that this lincRNA is conserved in place, expression and function, at least at some degree, in the phylum Chordata, making *Hotairm1* the first identified and characterised case of a bona fide lincRNA with this level of evolutionary depth conservation.

3. Conserved regulation of Hox Cluster in Chordata

3.1. Analysis of *Pdx* and *Cdx* knockouts during development

When knocking-out key transcription factors during development, as the ParaHox cluster genes (Brooke *et al.* 1998), we should expect a huge number of genes (or contigs in our case) showing misregulation, whether being up or down expressed. This is because of a cascade effect where ParaHox genes regulate several genes, and these genes act upon others, etc... This is why we choose to be highly restrictive and applied several filters to identify significant differently expressed contigs (DEG), rather than obtaining a very long list of slightly distinctly expressed contigs in the KO vs WT embryos.

The analysis of the *Pdx*^{-/-} embryos is quite puzzling. If we look just at the developing embryos, we could think that we are looking at control animals. There is not an obvious phenotype, despite the genotyping telling us that those animals are in fact *Pdx*^{-/-}. Only one small detail can be used to distinguish KO from control animals, the absence of a small midgut-restricted area where an endogenous GFP is missing. This particular place matches with the place of expression of *Pdx*, that is probably acting as a midgut regulator (Brooke *et al.* 1998).

Excitingly however, when we look at RNA-seq data, *Pdx*^{-/-} animals tell a very different story, with around 6000 significant differentially expressed contigs. These numbers are, however, not translated into a quite strong phenotype. This apparent contradiction may be explained by four different hypothesis or a combination of them: (i) the targets of *Pdx* somehow are balancing each other and making the embryo to develop properly; (ii) the *Pdx* absence impact is going to be seen at long term, once the organism is fully developed and could be related with the misregulation of insulin-like proteins which are expressed just in the region of the gut affected in the *Pdx*^{-/-} organisms (P. W. H. Holland *et al.* 1997; Lecroisey *et al.* 2015); (iii) there are more subtle changes in the phenotype that the ones we identify, as minor defects in the midgut formation; (iv) most of the genes labelled *significant* may be borderline significant and the actual number of real significant DEGs is lower. With no doubt, it is difficult to pinpoint what is really happening, but we keep investigating to assess the precise extent of the *Pdx* knock-out. Just to note, mice knockouts for *Pdx1* are birth properly, the main and nearly one defect being the lack of pancreas, an endoderm derivative, hence no pancreas hormones are

Discussion

produced after birth, producing the metabolic death of the animal (A. M. Holland *et al.* 2013).

Cdx^{-/-} analysis, on the other hand, seems to be quite the opposite. The developing embryos start developing fairly normal, but at 48hpf, when the larvae start feeding and the anus should open, it remains closed. This ends with the *Cdx*^{-/-} larvae dying at ~7dpf. In addition, there is a disruption of the tail fin and the posterior gut. These characteristics are consistent with the bibliography about *Cdx* and its role as a regulator of posterior gut formation (Marlétaz *et al.* 2015). Curiously, this strong phenotype is caused by a fourth of the *Pdx*^{-/-} significant DEGs. But as we can see in the MA plots (Figure RIII-7), one main difference between the *Pdx*^{-/-} and *Cdx*^{-/-} significant DEGs is, that in the case of the later, the number of significant DEGs with a log2FoldChange greater than ±2 is quite higher, and these figures could explain the presence of such a strong phenotype with a lower amount of misregulated targets than *Pdx*^{-/-}. Nevertheless, another reason that could explain the strong phenotype is the relation of *Cdx* with two other key players in the embryo development, the Retinoic Acid (Shimizu *et al.* 2006) and the Hox cluster (Charite *et al.* 1998; Marlétaz *et al.* 2015).

3.2. Gut-enrichment in both datasets

Pdx and *Cdx* are well known for their role in the developing gut (Beck and Stringer 2010; Hideaki Kaneto *et al.* 2007), hence we wanted to assess how many gut-specific genes were being affected by each knock-out. To do so, we first had to determine what was the basal prevalence of any random set of genes in our DEG datasets. Therefore, we generated a randomized dataset, the size of the gut-specific sample, and then compare the percentages of inclusion of this randomized sample within both DEG datasets. For being statistically significant we made 1000 iterations of the randomized dataset and plotted them against the density of probability, alongside the percentage of inclusion of gut-specific genes in our DEG datasets.

These proved our DEG datasets to be significantly enriched in gut-specific genes but still the percentages of inclusion seem to be low for what we would expect in genes so clearly related with gut development. Then, we considered that a great fraction of these genes will not be tissue specific, either because they are not expressed in the adult animal, (where the tissue samples for determining the tissue specificity were collected),

or because they are housekeeping genes, or simply they are expressed in three or more different tissues. If we reduce the *Pdx*^{-/-} and the *Cdx*^{-/-} samples to their tissue specific fractions, these percentages bump up to 35% and 18% respectively. This would be closer to what we would expect. If we ran again the randomization analysis, but this time using only the tissue specific dataset as the selection tool, the enrichment in gut-specific genes is still significant. This proves that both datasets are reliably significantly enriched in gut-specific genes.

3.3. Conservation of the anterior and middle Hox cluster regulation via *Cdx*

Cdx has been known to be a Hox cluster regulator since very early after its discovering (Brooke *et al.* 1998; Charite *et al.* 1998). Unfortunately, how *Cdx* regulates the Hox cluster genes is still unclear. In a recent work in a *Cdx* absence scenario, Marletaz *et al.* (2015) showed a colinear regression between the changes in expression of Hox cluster genes and the position of these genes inside the cluster, with the anterior Hox genes being overexpressed, and the posterior Hox genes strongly downexpressed.

The results of these experiments carried out in *Xenopus tropicalis* using a quadruple morpholino were in fact replicated in *Branchiostoma floridae* *Cdx*^{-/-} embryos. This clearly points to a conservation of the regulation of the Hox cluster by *Cdx* in amphioxus and *Xenopus*, with *Cdx*-Hox gene interaction being a sliding scale. The conserved mechanism of action may be a chromatin remodelling process mediated by *Cdx* or, alternatively, by directly promoting or repressing Hox genes expression by binding to regulatory regions.

Despite amphioxus having only one *Cdx* gene, in comparison with the three that are present in vertebrates, the partial redundancy observed within the three vertebrate *Cdx* could explain this kind of conservation in function. As the three *Cdx* genes are expressed in similar regions of the developing embryo (Marlétaz *et al.* 2015) perhaps the three of them are needed to properly promote the expression of the posterior Hox genes and repress the anterior ones, and the differences rely in which of the four Hox cluster they mainly act. This would explain how this mechanism of Hox cluster regulation is conserved in amphioxus, as there is only one *Cdx* gene for one Hox cluster, making then amphioxus a simple model to look at this interesting and unsolved cross talk between Hox and ParaHox clusters.

Discussion

Bibliography of Discussion

- Beck, F., & Stringer, E. J. (2010). The role of Cdx genes in the gut and in axial development. *Biochemical Society Transactions*, 38(2), 353–357. doi:10.1042/BST0380353
- Brooke, N. M., Garcia-Fernández, J., & Holland, P. W. H. (1998). The ParaHox gene cluster is an evolutionary sister of the Hox gene cluster. *Nature*, 392(6679), 920–922. doi:10.1038/31933
- Burguera, D., Marquez, Y., Racioppi, C., Permanyer, J., Torres-Méndez, A., Esposito, R., et al. (2017). Evolutionary recruitment of flexible ESRP-dependent splicing programs into diverse embryonic morphogenetic processes. *Nature Communications*, 8(1), 1799. doi:10.1038/s41467-017-01961-y
- Burguera Hernández, D. (2017). Evolutionary recruitment and assembly of embryonic alternative splicing programs: insights from the Deuterostomia lineage. *TDX (Tesis Doctorals en Xarxa)*.
<http://www.tesisenred.net/handle/10803/456241>. Accessed 13 September 2019
- Charite, J., de Graaff, W., Consten, D., Reijnen, M. J., Korving, J., & Deschamps, J. (1998). Transducing positional information to the Hox genes: critical interaction of cdx gene products with position-sensitive regulatory elements. *Development*, 125(22).
- Clark, B. S., & Blackshaw, S. (2017). Understanding the Role of lncRNAs in Nervous System Development. *Advances in experimental medicine and biology*, 1008, 253–282. doi:10.1007/978-981-10-5203-3_9
- Conklin, E. G. (1932). The embryology of amphioxus. *Journal of Morphology*, 54(1), 69–151. doi:10.1002/jmor.1050540103
- Dehal, P., & Boore, J. L. (2005). Two Rounds of Whole Genome Duplication in the Ancestral Vertebrate. *PLoS Biology*, 3(10), e314. doi:10.1371/journal.pbio.0030314
- Diederichs, S. (2014). The four dimensions of noncoding RNA conservation. *Trends in Genetics*, 30(4), 121–123. doi:10.1016/j.tig.2014.01.004
- Eisen, J. S., & Smith, J. C. (2008). Controlling morpholino experiments: don't stop making antisense. *Development (Cambridge, England)*, 135(10), 1735–43. doi:10.1242/dev.001115
- Esfandi, F., Taheri, M., Omrani, M. D., Shadmehr, M. B., Arsang-Jang, S., Shams, R., & Ghafouri-Fard, S. (2019). Expression of long non-coding RNAs (lncRNAs) has been dysregulated in non-small cell lung cancer tissues. *BMC Cancer*, 19(1), 222. doi:10.1186/s12885-019-5435-5
- Gardner, P. P., Fasold, M., Burge, S. W., Ninova, M., Hertel, J., Kehr, S., et al. (2015). Conservation and Losses of Non-Coding RNAs in Avian Genomes. *PLOS ONE*, 10(3), e0121797. doi:10.1371/journal.pone.0121797
- Gat-Yablonski, G. (2011). Brain development is a multi-level regulated process--the case of the OTX2 gene. *Pediatric endocrinology reviews : PER*, 9(1), 422–30. <http://www.ncbi.nlm.nih.gov/pubmed/22783640>. Accessed 13 September 2019
- Han, P., Li, W., Lin, C.-H., Yang, J., Shang, C., Nurnberg, S. T., et al. (2014). A long noncoding RNA protects the heart from pathological hypertrophy. *Nature*, 514(7520), 102–106. doi:10.1038/nature13596
- Heasman, J., Kofron, M., & Wylie, C. (2000). β Catenin Signaling Activity Dissected in the Early Xenopus Embryo: A Novel Antisense Approach. *Developmental Biology*, 222(1), 124–134. doi:10.1006/dbio.2000.9720
- Hidalgo-Sánchez, M., Millet, S., Bloch-Gallego, E., & Alvarado-Mallart, R.-M. (2005). Specification of the meso-isthmo-cerebellar region: The Otx2/Gbx2 boundary. *Brain Research Reviews*, 49(2), 134–149. doi:10.1016/j.brainresrev.2005.01.010
- Hideaki Kaneto, H., Takeshi Miyatsuka, T., Toshihiko Shiraiwa, T., Kaoru Yamamoto, K., Ken Kato, K., Yoshio Fujitani, Y., & Taka-aki Matsuoka, T. (2007). Crucial Role of PDX-1 in Pancreas Development, β -Cell Differentiation, and Induction of Surrogate β -Cells. *Current Medicinal Chemistry*, 14(16), 1745–1752.

- doi:10.2174/092986707781058887
- Holland, A. M., Garcia, S., Naselli, G., MacDonald, R. J., & Harrison, L. C. (2013). The Parahox gene Pdx1 is required to maintain positional identity in the adult foregut. *The International Journal of Developmental Biology*, *57*(5), 391–398. doi:10.1387/ijdb.120048ah
- Holland, P. W. H., Patton, S. J., Brooke, N. M., & Garcia-Fernandez, J. (1997). Genetic patterning of ectoderm and endoderm in amphioxus: from homeobox genes to hormones. In *Advances in Comparative Endocrinology: Proceedings of the 13th International Congress on Comparative Endocrinology* (pp. 247–252).
- Ikuta, T., Yoshida, N., Satoh, N., & Saiga, H. (2004). *Ciona intestinalis* Hox gene cluster: Its dispersed structure and residual colinear expression in development. *Proceedings of the National Academy of Sciences of the United States of America*, *101*(42), 15118–23. doi:10.1073/pnas.0401389101
- Irimia, M., Denuc, A., Burguera, D., Somorjai, I., Martin-Duran, J. M., Genikhovich, G., et al. (2011). Stepwise assembly of the Nova-regulated alternative splicing network in the vertebrate brain. *Proceedings of the National Academy of Sciences*, *108*(13), 5319–5324. doi:10.1073/pnas.1012333108
- Irimia, M., Tena, J. J., Alexis, M. S., Fernandez-Minan, A., Maeso, I., Bogdanovic, O., et al. (2012). Extensive conservation of ancient microsynteny across metazoans due to cis-regulatory constraints. *Genome Research*, *22*(12), 2356–2367. doi:10.1101/gr.139725.112
- Irimia, Manuel, Rukov, J. L., Roy, S. W., Vinther, J., & Garcia-Fernandez, J. (2009). Quantitative regulation of alternative splicing in evolution and development. *BioEssays*, *31*(1), 40–50. doi:10.1002/bies.080092
- Jathar, S., Kumar, V., Srivastava, J., & Tripathi, V. (2017). Technological developments in lncRNA biology. In *Advances in Experimental Medicine and Biology* (Vol. 1008, pp. 283–323). doi:10.1007/978-981-10-5203-3_10
- Kurosawa, G., Takamatsu, N., Takahashi, M., Sumitomo, M., Sanaka, E., Yamada, K., et al. (2006). Organization and structure of hox gene loci in medaka genome and comparison with those of pufferfish and zebrafish genomes. *Gene*, *370*, 75–82. doi:10.1016/j.gene.2005.11.015
- Lecroisey, C., Le Pétilion, Y., Escriva, H., Lammert, E., & Laudet, V. (2015). Identification, Evolution and Expression of an Insulin-Like Peptide in the Cephalochordate *Branchiostoma lanceolatum*. *PLOS ONE*, *10*(3), e0119461. doi:10.1371/journal.pone.0119461
- Li, Q., Dong, C., Cui, J., Wang, Y., & Hong, X. (2018). Over-expressed lncRNA HOTAIRM1 promotes tumor growth and invasion through up-regulating HOXA1 and sequestering G9a/EZH2/Dnmts away from the HOXA1 gene in glioblastoma multiforme. *Journal of Experimental & Clinical Cancer Research*, *37*(1), 265. doi:10.1186/s13046-018-0941-x
- Lin, M., Pedrosa, E., Shah, A., Hrabovsky, A., Maqbool, S., Zheng, D., & Lachman, H. M. (2011). RNA-Seq of Human Neurons Derived from iPS Cells Reveals Candidate Long Non-Coding RNAs Involved in Neurogenesis and Neuropsychiatric Disorders. *PLoS ONE*, *6*(9), e23356. doi:10.1371/journal.pone.0023356
- Marlétaz, F., Maeso, I., Faas, L., Isaacs, H. V., & Holland, P. W. H. (2015). Cdx ParaHox genes acquired distinct developmental roles after gene duplication in vertebrate evolution. *BMC Biology*, *13*(1), 56. doi:10.1186/s12915-015-0165-x
- Müller, G. B. (2007). Evo–devo: extending the evolutionary synthesis. *Nature Reviews Genetics*, *8*(12), 943–949. doi:10.1038/nrg2219
- Necsulea, A., Soumillon, M., Warnefors, M., Liechti, A., Daish, T., Zeller, U., et al. (2014). The evolution of lncRNA repertoires and expression patterns in tetrapods. *Nature*, *505*(7485), 635–640. doi:10.1038/nature12943
- Neme, R., & Tautz, D. (2016). Fast turnover of genome transcription across evolutionary time exposes entire non-coding DNA to de novo gene emergence. *eLife*, *5*. doi:10.7554/eLife.09977

Discussion

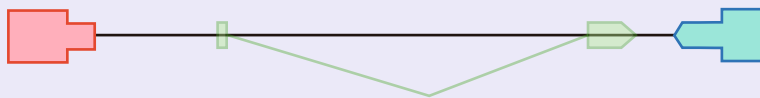
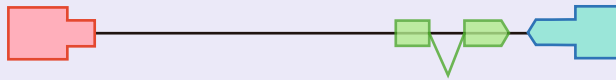
- Ounzain, S., Burdet, F., Ibberson, M., & Pedrazzini, T. (2015). Discovery and functional characterization of cardiovascular long noncoding RNAs. *Journal of Molecular and Cellular Cardiology*, *89*, 17–26. doi:10.1016/j.YJMCC.2015.09.013
- Pascual-Anaya, J., Sato, I., Sugahara, F., Higuchi, S., Paps, J., Ren, Y., et al. (2018). Hagfish and lamprey Hox genes reveal conservation of temporal colinearity in vertebrates. *Nature Ecology & Evolution*, *2*(5), 859–866. doi:10.1038/s41559-018-0526-2
- Pauli, A., Valen, E., Lin, M. F., Garber, M., Vastenhouw, N. L., Levin, J. Z., et al. (2012). Systematic identification of long noncoding RNAs expressed during zebrafish embryogenesis. *Genome Research*, *22*(3), 577–591. doi:10.1101/gr.133009.111
- Ren, T., Hou, J., Liu, C., Shan, F., Xiong, X., Qin, A., et al. (2019). The long non-coding RNA HOTAIRM1 suppresses cell progression via sponging endogenous miR-17-5p/ B-cell translocation gene 3 (BTG3) axis in 5-fluorouracil resistant colorectal cancer cells. *Biomedicine & Pharmacotherapy*, *117*, 109171. doi:10.1016/j.biopha.2019.109171
- Rivas, E., Clements, J., & Eddy, S. R. (2017). A statistical test for conserved RNA structure shows lack of evidence for structure in lncRNAs. *Nature methods*, *14*(1), 45–48. doi:10.1038/nmeth.4066
- Sekigami, Y., Kobayashi, T., Omi, A., Nishitsuji, K., Ikuta, T., Fujiyama, A., et al. (2017). Hox gene cluster of the ascidian, *Halocynthia roretzi*, reveals multiple ancient steps of cluster disintegration during ascidian evolution. *Zoological Letters*, *3*(1), 17. doi:10.1186/s40851-017-0078-3
- Shimizu, T., Bae, Y.-K., & Hibi, M. (2006). Cdx-Hox code controls competence for responding to Fgfs and retinoic acid in zebrafish neural tissue. *Development*, *133*(23), 4709–4719. doi:10.1242/dev.02660
- Sorek, R., & Ast, G. (2003). Intronic sequences flanking alternatively spliced exons are conserved between human and mouse. *Genome research*, *13*(7), 1631–7. doi:10.1101/gr.1208803
- Wang, X. Q. D., & Dostie, J. (2017). Reciprocal regulation of chromatin state and architecture by HOTAIRM1 contributes to temporal collinear HOXA gene activation. *Nucleic acids research*, *45*(3), 1091–1104. doi:10.1093/nar/gkw966
- Yu, H., Lindsay, J., Feng, Z.-P., Frankenberg, S., Hu, Y., Carone, D., et al. (2012). Evolution of coding and non-coding genes in HOX clusters of a marsupial. *BMC genomics*, *13*, 251. doi:10.1186/1471-2164-13-251
- Zampetaki, A., Albrecht, A., & Steinhofel, K. (2018). Long Non-coding RNA Structure and Function: Is There a Link? *Frontiers in physiology*, *9*, 1201. doi:10.3389/fphys.2018.01201
- Zhang, X., Lian, Z., Padden, C., Gerstein, M. B., Rozowsky, J., Snyder, M., et al. (2009). A myelopoiesis-associated regulatory intergenic noncoding RNA transcript within the human HOXA cluster. *Blood*, *113*(11), 2526–34. doi:10.1182/blood-2008-06-162164
- Zhu, J., Chen, G., Zhu, S., Li, S., Wen, Z., Bin Li, B., et al. (2016). Identification of Tissue-Specific Protein-Coding and Noncoding Transcripts across 14 Human Tissues Using RNA-seq. *Scientific Reports*, *6*(1), 28400. doi:10.1038/srep28400

Conclusions

The conclusions of this thesis are:

1. The lncRNA fraction in *B. lanceolatum* is composed of a total of 31.485 lncRNAs with 1.512 of them being multiexonic.
2. In vertebrates there are additional levels of regulation by lncRNAs in the neural tissue compared to amphioxus.
3. None of the lncRNAs identified had primary sequence conservation.
4. LincOFinder can be used to effectively find orthologous lincRNAs between distant species using microsynteny. In the case of amphioxus versus humans, we identified 16 cases, one of them was *Hotairm1*
5. *Hotairm1* was retained in the HoxA cluster after the 2R WGD.
6. *Hotairm1* expression domain is conserved between *Xenopus* and amphioxus.
7. *Hotairm1* mechanism of function is conserved between *Xenopus* and human.
8. *Hotairm1* is conserved in Chordata, being the first *bona fide* lincRNA identified with this degree of conservation.
9. In *B. floridae*, *Cdx* and *Pdx* knockouts have identifiable phenotypes and clear significant differentially regulated contigs.
10. *Cdx*^{-/-} and *Pdx*^{-/-} significant differentially expressed genes are significantly enriched in gut-specific contigs.
11. In *B. floridae*, the differences in expression of the Hox cluster in *Cdx*^{-/-} are resembling the ones seen in triple *Cdx* knockdowns in *Xenopus*.
12. The regulation of Hox cluster by *Cdx* is conserved in Chordata.

Annexes





Characterization of the TLR Family in *Branchiostoma lanceolatum* and Discovery of a Novel TLR22-Like Involved in dsRNA Recognition in Amphioxus

Jie Ji¹, David Ramos-Vicente^{1,2}, Enrique Navas-Pérez³, Carlos Herrera-Úbeda³, José Miguel Lizcano⁴, Jordi Garcia-Fernández³, Hector Escrivà⁵, Àlex Bayés^{1,2} and Nerea Roher^{1*}

¹ Department of Cell Biology, Animal Physiology and Immunology, Institute of Biotechnology and Biomedicine (IBB), Universitat Autònoma de Barcelona, Bellaterra, Spain, ² Molecular Physiology of the Synapse Laboratory, Biomedical Research Institute Sant Pau (IIB Sant Pau), Barcelona, Spain, ³ Department of Genetics, School of Biology and Institute of Biomedicine (IBUB), University of Barcelona, Barcelona, Spain, ⁴ Department of Biochemistry and Molecular Biology, Institute of Neurosciences, Universitat Autònoma de Barcelona, Bellaterra, Spain, ⁵ CNRS, Biologie Intégrative des Organismes Marins, BIOM, Sorbonne Université, Banyuls-sur-Mer, France

OPEN ACCESS

Edited by:

L. Courtney Smith,
George Washington University,
United States

Reviewed by:

Katherine Buckley,
Carnegie Mellon University,
United States
Loriano Ballarin,
Università degli Studi di Padova, Italy

*Correspondence:

Nerea Roher
nerea.roher@uab.cat

Specialty section:

This article was submitted to
Comparative Immunology,
a section of the journal
Frontiers in Immunology

Received: 11 July 2018

Accepted: 12 October 2018

Published: 02 November 2018

Citation:

Ji J, Ramos-Vicente D, Navas-Pérez E, Herrera-Úbeda C, Lizcano JM, Garcia-Fernández J, Escrivà H, Bayés À and Roher N (2018) Characterization of the TLR Family in *Branchiostoma lanceolatum* and Discovery of a Novel TLR22-Like Involved in dsRNA Recognition in Amphioxus. *Front. Immunol.* 9:2525. doi: 10.3389/fimmu.2018.02525

Toll-like receptors (TLRs) are important for raising innate immune responses in both invertebrates and vertebrates. Amphioxus belongs to an ancient chordate lineage which shares key features with vertebrates. The genomic research on TLR genes in *Branchiostoma floridae* and *Branchiostoma belcheri* reveals the expansion of TLRs in amphioxus. However, the repertoire of TLRs in *Branchiostoma lanceolatum* has not been studied and the functionality of amphioxus TLRs has not been reported. We have identified from transcriptomic data 30 new putative TLRs in *B. lanceolatum* and all of them are transcribed in adult amphioxus. Phylogenetic analysis showed that the repertoire of TLRs consists of both non-vertebrate and vertebrate-like TLRs. It also indicated a lineage-specific expansion in orthologous clusters of the vertebrate TLR11 family. We did not detect any representatives of the vertebrate TLR1, TLR3, TLR4, TLR5 and TLR7 families. To gain insight into these TLRs, we studied in depth a particular TLR highly similar to a *B. belcheri* gene annotated as bbtTLR1. The phylogenetic analysis of this novel BITLR showed that it clusters with the vertebrate TLR11 family and it might be more related to TLR13 subfamily according to similar domain architecture. Transient and stable expression in HEK293 cells showed that the BITLR localizes on the plasma membrane, but it did not respond to the most common mammalian TLR ligands. However, when the ectodomain of BITLR is fused to the TIR domain of human TLR2, the chimeric protein could indeed induce NF- κ B transactivation in response to the viral ligand Poly I:C, also indicating that in amphioxus, specific accessory proteins are needed for downstream activation. Based on the phylogenetic, subcellular localization and functional analysis, we propose that the novel BITLR might be classified as an antiviral receptor sharing at least partly the functions performed by vertebrate TLR22. TLR22 is thought to be viral

teleost-specific TLR but here we demonstrate that teleosts and amphioxus TLR22-like probably shared a common ancestor. Additional functional studies with other lancelet TLR genes will enrich our understanding of the immune response in amphioxus and will provide a unique perspective on the evolution of the immune system.

Keywords: toll-like receptor, TLR, evolution, amphioxus, Poly I:C, TLR22

INTRODUCTION

There are two types of immunity in vertebrates. One is the innate immunity, which is genetically programmed to detect invariant features of invading microbes. The other is the adaptive immunity, which employs antigen receptors that are not encoded in the germ line but are generated *de novo* (1). The innate immune system is the first line of defense against infectious diseases (2). Immediately after infection, the innate response is activated to combat pathogens and synthesize inflammatory mediators and cytokines (3). However, the primary challenge of the innate immune system is how to discriminate a countless number of pathogens using a restricted number of receptors (2). As a response, a variety of receptors can recognize conserved motifs on pathogens (4). These conserved motifs are known as Pathogen-Associated Molecular Patterns (PAMPs) (5) and their recognition partners, are called Pattern Recognition Receptors (PRRs) (6).

Toll-like receptors (TLRs), among the most extensively studied PRRs, are type-I transmembrane proteins consisting of an ectodomain, a transmembrane (TM) domain and an intracellular Toll/interleukin-1 receptor (TIR) domain (7). The ectodomain, which functions as a PAMPs recognition domain, is arranged in tandem leucine-rich repeat (LRR), from one to many depending on the receptor type. The LRR contains a segment of 11 conserved residues with the consensus sequence LxxLxLxxNxL, where x can be any amino acid, L is a hydrophobic residue (leucine, valine, isoleucine, or phenylalanine) and N can be asparagine or cysteine (8). The TIR domain is present in the cytosol and is required for downstream signal transduction (9). Upon PAMP recognition, TLRs recruit TIR-domain containing adaptor proteins such as MyD88, TRIF, TIRAP/MAL, or TRAM, which initiate signal transduction pathways that culminate in the activation of NF- κ B, IRFs, or MAP kinases regulating the expression of cytokines, chemokines, or type I interferons (IFN), which finally protect the host against infections (10).

TLRs are expressed in innate immune cells such as dendritic cells and macrophages as well as non-immune cells like fibroblast and epithelial cells (10). TLRs are largely divided into two subfamilies based on their subcellular localizations: cell surface or intracellular. Ten and twelve functional TLRs have been identified in humans and mice, respectively. Human TLR1, TLR2, TLR4, TLR5 and TLR6 are expressed on the cell surface and recognize mainly microbial membrane components such as lipids, lipoproteins and proteins. Human TLR3, TLR7, TLR8, TLR9 and murine TLR11, TLR12, TLR13, which are expressed in intracellular vesicles such as those in the endoplasmic reticulum (ER), endosomes, lysosomes and endolysosomes, and recognize

nucleic acids (9, 11–13). Recently, the sequencing of the genome in five bony fish species has allowed the discovery of at least 16 TLR types in teleosts (14).

There are two structural types of TLRs according to the TLR ectodomain structure: sccTLRs and mccTLRs. The sccTLRs are characterized by the presence of a single cysteine cluster on the C-terminal end of LRRs (a CF motif), which is juxtaposed to the plasma membrane. Most TLRs found in deuterostomes have this domain organization. The mccTLRs are characterized by an ectodomain with two or more CF motifs and another cysteine cluster on the N-terminal side of the LRRs (NF motif). They are systematically found in protostomes but have also been identified in the invertebrate deuterostome *S. purpuratus* and the cnidarian *N. vectensis* (15). Both sccTLR and mccTLR share a common TLR structure: LRR+TM+TIR. According to the ectodomain architecture and phylogenetic criteria, vertebrate TLRs can be classified into six families: 1, 3, 4, 5, 7 and 11. TLR1 family includes TLR1/2/6/10/14/18/24/25 as well as TLR27; TLR3, 4 and 5 families only include TLR3, 4 and 5 itself; TLR7 family includes TLR7/8/9; TLR11 family includes two subfamilies: 11 (TLR11/12/16/19/20/26) and 13 (TLR13/21/22/23) (16, 17).

A variety of TLRs are capable of recognizing viruses. Among human TLRs, the envelope proteins from viruses are mainly recognized by TLR2 and TLR6. Viral nucleic acids are recognized by TLR3 (ssRNA or dsDNA), TLR7 (ssRNA), TLR8 (ssRNA), and TLR9 (dsDNA or CpG motifs) (18). In teleosts, it has been reported that Poly I:C could be recognized by different TLRs. Teleost TLR13 was firstly reported in Miiuy croaker (*Miichthys miiuy*) which showed cytoplasmic localization in HeLa cells. It could respond to both *Vibrio anguillarum* and Poly I:C injection *in vivo* and Poly I:C stimulation in leukocytes (19). In fugu (*Takifugu rubripes*), TLR3 localizes in the endoplasmic reticulum and recognizes relatively short dsRNA, whereas TLR22 recognizes long dsRNA on the cell surface (20). Grass carp (*Ctenopharyngodon idella*) TLR22 is expressed in many tissues and is highly abundant in the gills. Infection of grass carp with grass carp reovirus (GCRV), a dsRNA virus, induces a rapid up-regulation of TLR22 gene expression in the spleen (21). Japanese flounder (*Paralichthys olivaceus*) TLR22 is mainly expressed in peripheral blood leukocytes (PBL) and could be induced by both peptidoglycan and Poly I:C (22), whereas TLR3 gene expression in PBLs increased upon stimulation with Poly I:C and CpG ODN 1668 (23). Both TLR3 and TLR22 gene transcription had also been studied in large yellow croaker. Basal gene transcription was high in several immune organs and could be up-regulated after injection of Poly I:C in the anterior kidney (TLR22), spleen (TLR3 and 22), liver (TLR3) and blood (TLR3) (23). In the common carp (*Cyprinus carpio* L.), TLR22 was transcribed in

almost all the tissues. When fish was challenged with Poly I:C or *Aeromonas hydrophila*, the transcription of this TLR was up-regulated in a variety of tissues (24). Overall, TLRs with immune function have been found from cnidarians to mammals which imply a conserved evolution. TLR3 is found both in mammals and teleost whereas TLR22 is present in many fish species and *Xenopus*, but absent from birds and other terrestrial animals (25). The origin of the TLRs involved in dsRNA virus recognition is still under study. The current hypothesis is that specific fish TLR duplication results from the fish specific Whole Genome Duplication (WGD) (26–28), but here we show that, in amphioxus, exists an ortholog of the TLR11 subfamily possessing TLR22 functional similarities, pointing out that a TLR22-like function was present in the ancestor of chordates.

Amphioxus belongs to an ancient chordate lineage which shares key anatomical and developmental features with vertebrates and tunicates (also known as urochordates) (29). All chordates have a similarly organized genome though amphioxus has relatively little duplication (30). Thus amphioxus, with its phylogenetic position diverging at the base of chordates and its genomic simplicity, is a good non-vertebrate model to understand the evolution of vertebrates (31). *Branchiostoma lanceolatum* (Mediterranean amphioxus) has been extensively studied together with other amphioxus species such as *Branchiostoma belcheri* (Asian amphioxus), *Branchiostoma japonicum* (Asian amphioxus) and *Branchiostoma floridae* (Florida amphioxus) (32). To date, genomic data have revealed that *B. floridae* has 48 TLRs (33). However, only one full-length TLR, annotated as bbtTLR1, was functionally characterized in *B. belcheri tsingtauense* until now. The experimental data supports the immunological function of this TLR that together with MyD88 is involved in the activation of NF- κ B signaling pathway (34). Further studies of TLRs in amphioxus are required to better understand the ancestors and functional evolution of vertebrate TLRs.

In this study, we investigated the total number of TLR genes in *B. lanceolatum* and studied their phylogenetic and evolutionary relationships with vertebrate and invertebrate TLRs. We also examined the total number of TLR genes in *B. floridae* and *B. belcheri* according to our definition of a true TLR. We studied the basal gene expression of all the TLRs in adult amphioxus (*B. lanceolatum*). Moreover, we cloned the full length of a novel TLR in *B. lanceolatum* and we further investigated its subcellular localization and PAMP binding specificity using NF- κ B luciferase assay in a mammalian expression system. Exhaustive phylogenetic analysis combined with functional data has allowed us to explore the evolution and function of this novel TLR compared with vertebrate TLRs.

MATERIALS AND METHODS

Sequence Analysis: Phylogeny and Bioinformatics

To characterize the TLR repertoire of *B. lanceolatum*, we performed a search using the BbtTLR1 sequence (GenBank: DQ400125.2) and an unpublished transcriptome of

B. lanceolatum derived from several adult tissues and embryonic stages. The transcriptome data were obtained from an exhaustive collection of 52 RNA-Seq datasets using the Illumina technology. From 15 embryonic stages, one pre-metamorphosis stage and 9 adult organs, a total of 4.2 billion Illumina reads with a volume of 871 Gbp were obtained. These embryonic stages are eggs, 32 cells, blastula, 7, 8, 10, 11, 15, 18, 21, 24, 27, 36, 50, and 60 hpf. The adult tissues are neural tube, gut, hepatic tissue, gills, epidermis, muscle, female and male gonads, and cirri. For the transcriptome assembly, TopHat2 was used mapping each strand-specific RNA-seq sample against the recently assembled *B. lanceolatum* genome. Gene models were built using Cufflinks and each annotation merged using Cuffmerge to produce a single collection of transcripts. The transcriptome was translated into predicted proteins using the TransDecoder suite v3.0.1. From the PFAM database v30.0, we downloaded the hidden Markov models profile collection (Pfam-A.hmm.gz) and extracted the two profiles for the protein domains that we were looking for, the TIR and the LRR domains. HMMER 3.1b was then used with the hmmsearch mode to identify the predicted proteins with these domains. Finally, a manually curated annotation was performed. Specific primers for each *B. lanceolatum* TLR were designed using NCBI primer designing tool (<https://www.ncbi.nlm.nih.gov/tools/primer-blast/>) and are shown in **Supplementary Table 1**.

To study the phylogenetic relationship of *B. lanceolatum* and vertebrate TLRs, we performed the maximum-likelihood analysis. *Drosophila melanogaster* Toll sequences and vertebrate TLR protein sequences were obtained from the National Center for Biotechnology Information (NCBI, <https://www.ncbi.nlm.nih.gov/>) and UniProt (<http://www.uniprot.org/>) (**Supplementary Table 2**). TLR sequences of *Lytechinus variegatus* (35) and *Saccoglossus kowalevskii* (36) were obtained from online repositories and a search similar to the one carried on in *B. lanceolatum* (**Supplementary Data 1**). In all the phylogenetic analysis, we only included the sequences that have a complete TIR domain. For full-length protein, sequences were aligned with MAFFT (37) choosing L-INS-i method which optimizes alignments for sequences containing hypervariable regions flanked by one alignable domain. For TIR domain, sequences were aligned with MAFFT choosing G-INS-i method which allows to align the entire region with a global conservation. The alignment was trimmed using TrimAL (38) with “Automated 1” mode. The phylogenetic reconstruction was done using IQ-TREE (39) and its built-in ModelFinder software (40). Branch support was calculated running 1,000 replicates of the SH-like approximate likelihood ratio test (SH-aLRT) (41) and ultrafast bootstrap (42).

The TLR sequences of *B. floridae* and *B. belcheri* were obtained from the databases of JGI (<http://genome.jgi.doe.gov/Brafl1/Brafl1.home.html>) and LanceletDB (<http://genome.bucm.edu.cn/lancelet/index.php>), respectively. The open reading frame was identified through sequence translation with ExPASy software (<http://web.expasy.org/translate/>). Transmembrane regions were predicted using TMHMM server v2.0 (<http://www.cbs.dtu.dk/services/TMHMM/>). The number of LRR domains was

predicted using LRRfinder software (<http://www.lrrfinder.com/lrrfinder.php>). Full-length protein domain was predicted by the Simple Modular Architecture Research Tool (SMART) (<http://smart.embl-heidelberg.de/>). The single cysteine cluster TLRs (sccTLRs) and multiple cysteine cluster TLRs (mccTLRs) were characterized according to Leulier and Lemaitre (15). The first annotated sequence was selected according to the blastp software in NCBI. The molecular weight of BITLR was calculated with ProtParam (<http://web.expasy.org/protparam/>). The sequence of BITLR was examined for the presence of a signal peptide using SignalP (<http://www.cbs.dtu.dk/services/SignalP/>). N-linked glycosylation site was predicted with NetNGly 1.0 server (<http://www.cbs.dtu.dk/services/NetNGly/>). Multiple sequence alignment of BITLR and fish TLR22 was performed by Clustal Omega (<https://www.ebi.ac.uk/Tools/msa/clustalo/>).

The phylogenetic analysis of three *Branchiostoma* species (*B. floridae*, *B. belcheri* and *B. lanceolatum*) TLRs was performed using TIR domain sequences. The TIR domain sequences of vertebrates, *S. kowalevskii* and *D. melanogaster* were included. All the TIR domain sequences were identified from the full-length protein using SMART software. Prior to the analysis, sequences were aligned with MAFFT choosing G-INS-i method. The alignment was trimmed using TrimAL with “Automated 1” mode. The phylogenetic analysis was done using IQ-TREE and its built-in ModelFinder software. Branch support was calculated running 1,000 replicates of the SH-like approximate likelihood ratio test and ultrafast bootstrap.

The phylogenetic analysis of BITLR and BbtTLR1 was performed with the full-length protein using IQ-TREE software. The *D. melanogaster* Toll and the vertebrate TLR sequences were included in the analysis. The sequences were aligned with MAFFT choosing L-INS-i method. The alignment was trimmed using TrimAL with “Automated 1” mode. In the analysis, branch support was calculated running 1,000 replicates of the SH-like approximate likelihood ratio test and ultrafast bootstrap.

Animals

Branchiostoma lanceolatum adults were collected in the bay of Argelès-sur-Mer, France (latitude 42° 32' 53" N and longitude 3° 03' 27" E) with a specific permission delivered by the Prefect of Region Provence Alpes Côte d'Azur. *B. lanceolatum* is not a protected species. Amphioxus were kept in the laboratory in 60-l glass tanks with ~50-l seawater and 5 cm height of sand on the bottom. Water temperature was maintained around 17°C and the salinity ranged between 40 and 45 PSU. The photoperiod was set to 14 h light/10 h dark. The animals were not fed with extra food during the experiment.

RNA Isolation, cDNA Synthesis and RT-PCR

Total RNA was extracted from the whole animal using TRI reagent (Sigma-Aldrich) according to the manufacturer's protocol. The homogenization was performed with a Polytron homogenizer (Kinematica). The quality of the RNA was assessed with a Bioanalyzer (Agilent Technologies) and the concentration was measured with a Nanodrop (Thermo scientific). The RNA

was purified using an RNeasy micro kit (Qiagen) and DNase treated according to manufacturer's instructions and stored at -80°C. The first-strand cDNA was synthesized with SuperScript III first-strand synthesis system (Thermo Fisher Scientific). RT-PCR reactions were performed with primers specific for each TLR under following conditions: initial denaturation at 94°C for 5 min, followed by 35 cycles of denaturation at 94°C for 45 s, annealing at 60°C for 45 s, and extension at 72°C for 50 s, and a final extension at 72°C, 7 min. Glyceraldehyde-3-phosphate dehydrogenase (GAPDH) was used as a reference gene. PCR products were separated in 1% agarose gel electrophoresis and stained with GelGreen Nucleic Acid Gel Stain (Biotium). Agarose gel imaging was performed with a GelDoc XR system (Bio-Rad). Six of the PCR products were purified and sequenced.

Full-Length cDNA Cloning of BLTLR

A DNA BLAST search of NCBI database was conducted using BbtTLR1 sequence from *B. belcheri* (GenBank: DQ400125.2). We obtained a sequence (GenBank: AF391294.1) from *B. floridae* showing 82% identity. In addition, a DNA BLAST search using bbtTLR1 was performed in the genome scaffold of *B. lanceolatum* and we identified a short sequence (ContigAmph29716) showing 83% identity. The forward primer (Table 1) was designed based on the conserved region between bbtTLR1 *B. belcheri* and *B. floridae* sequence. The reverse primer (Table 1) was designed based on the ContigAmph29716 sequence. We cloned a fragment of around 2,000 bp by PCR using the cDNA prepared from the whole animal. The 5'-end was obtained by 5' RACE (Invitrogen) using gene specific primers (Table 1). A fragment of ~600 bp was obtained. The 3'-end was obtained by 3' RACE (Invitrogen) using gene specific primer (Table 1). A fragment of ~1,000 bp was obtained. Finally, a PCR amplification was carried out to obtain the full-length sequence with Expand high fidelity PCR system (Roche) using the full-length primers (Table 1) designed in the non-coding regions from both 5' to 3'-ends. All the fragments were separated by electrophoresis and cloned into the pGEM-T Easy Vector (Promega). Sequencing was carried out using T7 and SP6 primers (Servei de Genòmica i Bioinformàtica, IBB-UAB).

TABLE 1 | Primers used for cloning and RT-qPCR.

Category	Primer	Sequence (5'-3')	Product size (bp)
Fragment	Forward	GGGACGATCCAGTCACGCTG	2,190
	Reverse	GACACCAACGGCTGCGCAG	
5'RACE	Reverse1	GAGTGAAGAACAGTGA	684
	Reverse2	GTCATTCCTCCAAGGTTCAAAGAAGTC	
3'RACE	Forward	CGAAGACAGGCGATGGGT	1,119
Full-length	Forward	AGAGAGAGAAAAGTCCAGGCC	3,077
	Reverse	TTTCTGTCTCGACGGTCCCTT	
RT-qPCR	Forward	TCACACGCTTTCTACGGCTT	122
	Reverse	AGGCTTAGGTTCCAGTACGGT	
GAPDH	Forward	CCCCACTGGCCAAAGGTATCA	154
	Reverse	GCTGGGATGATATTCTGGTGGGC	

LPS and Poly I:C Treatment *in vivo*

Adult amphioxys were treated with either 10 µg/ml bacterial lipopolysaccharide (LPS) from *Escherichia coli* O111:B4 strain (Sigma-Aldrich) or 10 µg/ml Poly I:C, a synthetic analog of dsRNA viruses (Invivogen) by bath immersion. The stocks of LPS and Poly I:C solution were prepared in PBS (Sigma-Aldrich) and diluted to the indicated working concentrations with sterile seawater. Seawater sterilization was performed with 0.22 µm sterile filter. PBS prepared in seawater (1% v/v) was used as a control. Three, 6, 12 and 24 h after immersion, 3 animals from each group were sampled separately. The animals were frozen in liquid nitrogen immediately and stored in -80°C until use. Total RNA was prepared from the whole animal and the first-strand cDNA was synthesized for RT-qPCR analysis.

RT-qPCR Analysis

RT-qPCR was carried out to analyze the relative transcription level of BITLR after LPS and Poly I:C treatments. The analysis was performed in the CFX384 Touch Real-Time PCR Detection System (Bio-Rad) using the iTaq universal SYBR green supermix kit (Bio-Rad) following the manufacturer's protocol. The RT-qPCR primers (Table 1) were designed to detect the transcription level of BITLR. GAPDH gene was used as a reference gene. 10^{-1} and 10^{-2} -fold cDNA dilutions were used for BITLR and GAPDH gene expression analysis, respectively. Each PCR mixture consisted of 5 µl of SYBR green supermix, 0.5 µM of primers, 2.5 µl of diluted cDNA, and 1.5 µl sigma water in a final volume of 10 µl. All samples were run in triplicate using the following steps: initial denaturation at 95°C for 3 min, 39 cycles of 95°C for 10 s and 60°C for 30 s, and finally, 95°C for 10 s, increase every 0.5°C for 5 s from 65 to 95°C . The relative transcription levels were calculated using the $2^{-\Delta\Delta\text{Ct}}$ method (43). All the data were analyzed using GraphPad software and significant differences were analyzed by one-way analysis of variance (ANOVA) using the value of ΔCt (normalize each technical repeat's gene-specific Ct value by subtracting from it the reference gene Ct value) (44).

Plasmids

To study the subcellular localization of BITLR in HEK293 cell, the coding sequence was cloned into pIRES2-EGFP vector (BD Biosciences Clontech, 6029-1) with two HA-tags (YPYDVPDYA) at 3' end (named BITLRHA) using XhoI and EcoRI as restriction sites. For testing the specific ligand binding of BITLR, the ectodomain and transmembrane domain (amino acids 1-774) of BITLR fused with human TLR2 cytoplasmic region (amino acids 611-784; NCBI: NP_001305716.1) was cloned into pIRES2-EGFP vector (named chimeric BITLR) between SacII and EcoRI restriction sites. The eukaryotic expression vector pIRES2-EGFP was purchased from BD Biosciences. The NF- κ B-dependent luciferase reporter vector (pNF κ B) and the Renilla luciferase vector (pRenilla) were provided by Dr. José Miguel Lizcano. All the plasmids were confirmed by sequencing and agarose gel electrophoresis digested with the corresponding restriction enzymes. All the plasmids were purified at large scale using NucleoBond Maxi endotoxin-free plasmid isolation kit (Fisher Scientific) and stored at -20°C until use.

Cell Culture, Transient Transfection and Stable Cell Lines

HEK293 cells were grown in complete medium: DMEM (Life Technologies, 31885) supplemented with 10% (v/v) FBS (Gibco) and 1% (v/v) penicillin and streptomycin (Gibco) at 37°C and 5% CO_2 . Plasmids were transiently transfected in HEK293 cells using linear polyethylenimine (PEI, CliniScience) at a ratio of 3:1 (µg PEI: µg plasmid). HEK293 cell lines stably expressing BITLRHA and chimeric BITLR were generated by Geneticin selection (Invitrogen, G418). In brief, 24 h after transient transfection, the culture medium was substituted with selective medium containing 1 mg/ml G418. Selective medium was refreshed every 2–3 days until the G418-resistant foci could be identified and all non-transfected cells (control) were dead (around 2 weeks). The colonies were picked and expanded in selective culture medium containing 1 mg/ml G418 for the following 2 weeks. Then, HEK293 stable cell lines were isolated via GFP-positive cell sorting (FACSJazz) in order to enrich the stable cell line. Finally, the HEK293 stable cell lines were cultured in DMEM complete medium at 37°C and 5% CO_2 .

Flow Cytometry

To assess the transient transfection efficiency of plasmid BITLRHA in HEK293 cells, flow cytometry was performed using a FACS Canto (Becton Dickinson, USA). In brief, HEK293 cells were seeded on 6-wells plate (Thermo Scientific) at 50% density. The cells were transfected with empty vector (pIRES2-EGFP) and BITLRHA plasmid using PEI as described above. Non-transfected cells were used as negative control. Cells were detached using TrypLE (Gibco) and re-suspended in PBS for cytometry analysis at 24, 48 and 72 h after transfection. The cytometer was set to detect the GFP signal and a total 10,000 events were recorded. The raw data were analyzed with Flowing software (Finland) and GraphPad software. Flow cytometry was also used to assess the percentage of transfected cells when setting up the stable cell lines BITLRHA and chimeric BITLR.

Western Blot Analysis

HEK293 cells were transiently transfected with empty vector (pIRES2-EGFP) and BITLRHA plasmid as described above. Cells were lysed in 200 µl cell lysis buffer (250 mM saccharose, 150 mM Tris, 5 mM EDTA, 125 mM DTT, 5% SDS, 2.5% bromophenol blue and 7.5% β -mercaptoethanol in water) and detached on ice using a cell scraper (BD Falcon) at 24, 48 and 72 h after transfection. The lysed cells were subjected to sonication for 10 s and centrifugation. After heating at 100°C for 5 min, the cell extracts were loaded into 10% SDS-PAGE and then transferred to PVDF membranes (EMD Millipore) using a Mini-protean Tetra (Bio-Rad). After 1 h blocking in 5% (w/v) BSA (Sigma-Aldrich) in TBST (50 mM Tris, 150 mM NaCl and 0.1% Tween 20), membranes were incubated with 1 µg/µl mouse anti-HA primary antibody (Covance, MMS-101P) overnight at 4°C , followed by incubation with a secondary HRP-conjugated antibody for 1 h at room temperature (RT). Proteins were visualized with a GelDoc system (Bio-Rad) by adding the SuperSignal West Pico chemiluminescent substrate (Thermo Fisher Scientific).

Immunofluorescence and Confocal Microscopy

HEK293 cells were seeded (50% density) on 24 × 24 mm cover glasses (Labbox) coated with Poly-D-lysine hydrobromide (Sigma-Aldrich). The BITLRHA plasmid was transiently transfected as described above. Cells were washed 3 times with DMEM at 48 h after transfection. For non-permeabilization, cells were blocked with 2% BSA in DMEM for 10 min at 37°C, and then incubated with mouse anti-HA primary antibody (1/500 diluted in DMEM) for 1 h at 37°C. Cells were washed 3 times with DMEM and fixed with 4% paraformaldehyde (PFA, Sigma-Aldrich) for 15 min at RT. After PBS washing, for transient transfection, fixed cells were incubated with anti-mouse Alexa Fluor 555 secondary antibody (Invitrogen) at 1:1,000 dilution for 2 h at RT; for stable transfection, cells were incubated with 5 µg/ml wheat germ agglutinin (WGA) conjugated with Alexa Fluor 647 for 10 min at RT before applying the secondary antibody at 1:1,000 dilution for 2 h at RT. For permeabilization, cells were washed with DMEM for 3 times and fixed with 4% PFA for 15 min at RT. After 3 washes with PBS, for transient transfection, cells were permeabilized with 0.2% Triton X-100 (Sigma-Aldrich) for 15 min at RT; for stable transfection, cells were incubated with 5 µg/ml WGA for 10 min at RT and then permeabilized with 0.1% Tween (Sigma-Aldrich) for 10 min at RT or the freeze and thaw method according to Mardones and González (45). After that, cells were blocked with 2% BSA in PBS for 1 h at RT, incubated with mouse anti-HA primary antibody (1/1,000 dilution) overnight at 4°C, followed by incubation with secondary anti-mouse AlexaFluor 555 antibody (Invitrogen) at 1:1,000 dilution for 2 h at RT. For both methods, cover glasses with cells were placed on SuperFrost Plus slides (Thermo scientific) covered with Fluoroshield with DAPI mounting medium (Sigma-Aldrich). Confocal imaging was performed using a Leica SP5 confocal microscope with a 63 × oil objective. The images were analyzed with Fiji software (46).

Ligand Stimulation and NF-κB Luciferase Reporter Assay

Human TLR1-9 agonist kit (tlrl-kit1hw) and murine TLR13 agonist (tlrl-orn19) were purchased from Invivogen. HEK293 stable cell lines were used to minimize the deviation among different experiments. The stable cell lines were transfected with 0.5 µg/ml pNFκB and 0.05 µg/ml pRenilla (0.5 ml per well) using PEI. Renilla was used as internal control to normalize the differences in the reporter due to different transfection efficiencies. Twenty-four hours after transfection, cells were treated with indicated concentrations of ligands (Supplementary Table 3) for 16 h. As a positive control, 20 ng/ml human TNFα (Sigma-Aldrich) was used. The experiment was performed in triplicate. Luciferase activity assay was performed with the Dual-luciferase reporter assay system (Promega) using the Victor3 (PerkinElmer) according to the manufacturer's instructions. Briefly, after removing the growth medium from the well, cells were washed with PBS (2X). One hundred µl of passive lysis buffer (PLB) were added to each well. Then, the NF-κB-dependent firefly luciferase reporter was measured by adding

100 µl of luciferase assay reagent II (LAR II). After quantifying the firefly luminescence, the reaction was quenched. The Renilla luciferase reaction was initiated by adding 100 µl Stop & Glo Reagent to the same well and the Renilla luminescent signal was detected. The luciferase activity was expressed as the ratio of NF-κB-dependent firefly luciferase activity divided by Renilla luciferase activity.

RESULTS

The TLR Family in *B. lanceolatum*

A search for TIR and LRR domains was performed and proteins with both domains were selected as candidates. Then, these candidates were manually curated and a list of putative TLRs was obtained (Supplementary Data 2). Despite there are TLR-related molecules lacking extracellular LRR domains reported in some species of *Hydra* and coral (15), we only considered those sequences with at least one LRR domain, one TM domain and one TIR domain to obtain our final list of true TLR candidates. Using this rule, we obtained 30 TLRs. In order to understand the evolution of TLR of *B. lanceolatum*, we performed a phylogenetic analysis with representative vertebrate and invertebrate TLR sequences. Other authors had used either the full-length protein or the TIR domain to study the TLR evolution (32, 46–48). Therefore, we used full-length protein to perform the phylogenetic analysis when the sequences were complete, or TIR domain when there were incomplete or truncated sequences. The phylogenetic analysis of *B. floridae*, *B. belcheri* and *B. lanceolatum* using TIR domain sequences showed that there are two major clusters of TLRs (mccTLRs and sccTLRs) in *Branchiostoma*. However, we obtained a single clade with almost all the *Branchiostoma* sequences, clustered with vertebrate TLR3, 5 and 7 families (Supplementary Figure 1). This approach did not allow the identification of inter-taxa relationships between vertebrate and *Branchiostoma* TLR families. Roach et al. predicted that a strong selective pressure for specific PAMPs recognition maintains a largely unchanged repertoire of TLR recognition in vertebrates (16). Thus, we did phylogenetic analysis using the highly refined full-length TLR sequences of *B. lanceolatum* to better understand the evolutionary relationships with vertebrate TLRs. The phylogenetic analysis showed that the vertebrate TLRs were grouped into six clusters (TLR1, TLR3, TLR4, TLR5, TLR7 and TLR11 families) with high branch support within their own clusters confirming the reliability of the tree (Figure 1 and Supplementary Figure 2). Twenty *B. lanceolatum* sequences formed a strongly supported clade distinct from the mccTLR sequences and grouped with the TLR11 family. One TLR (Bl19922) is not clustered with any TLRs, probably because it is an N-terminal truncated sequence. Moreover, six *B. lanceolatum* TLRs, which were identified as mccTLR (invertebrate type) were clustered separately from the main vertebrate branch (Figure 1 and Supplementary Table 4).

The transcription of the 30 TLRs in *B. lanceolatum* was confirmed by RT-PCR analysis in adult animals. Each primer pair was designed based on the nucleotide sequences reconstructed from transcript sequences of *B. lanceolatum*. We found gene

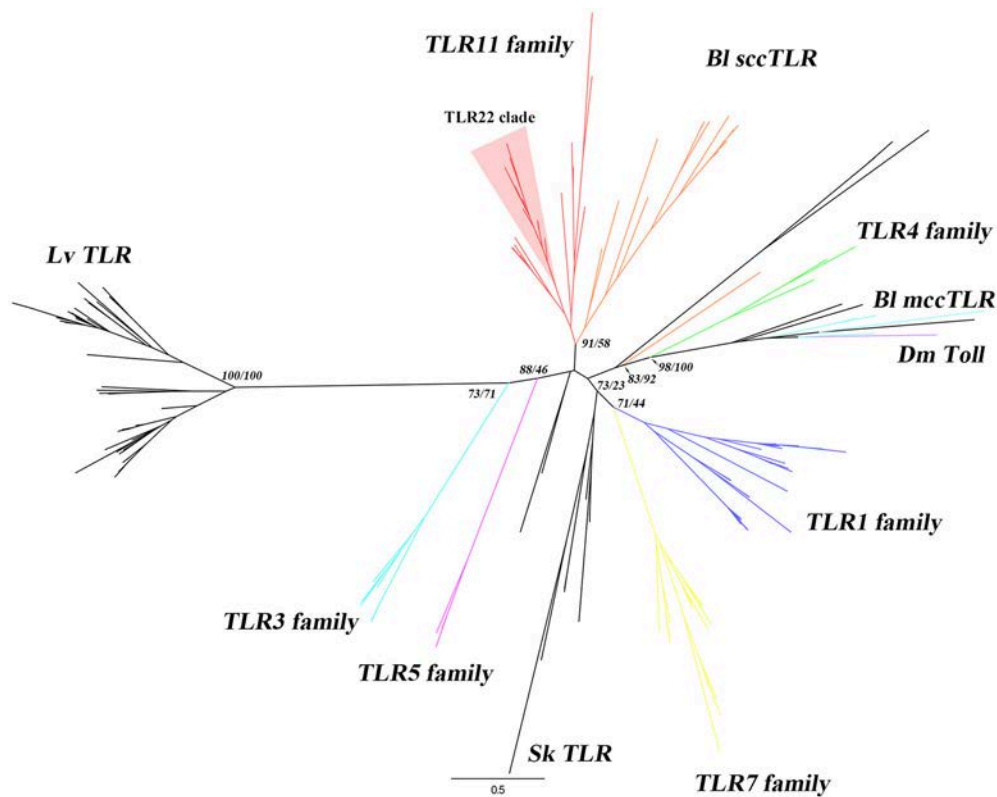


FIGURE 1 | Phylogenetic analysis of *B. lanceolatum* TLRs. The phylogenetic tree was constructed using maximum-likelihood (IQ-TREE) with the full-length protein sequences. TLR sequences of *B. lanceolatum*, *S. kowalevskii*, *L. variegatus*, representative vertebrates and *D. melanogaster* Toll were used. Three TLR sequences (BI10262, BI22164 and BI08928c) with incomplete TIR domain were removed from the analysis. Sequences were aligned with MAFFT choosing L-INS-i method and the alignments were trimmed using TrimAL with “Automated 1” mode. The best evolutionary model was established by ModelFinder according to BIC. The branch labels (numbers) are SH-aLRT support (%)/ultrafast bootstrap support (%) at the tree nodes. The tree was generated in FigTree. Dm Toll, Bl mccTLRs, Bl sccTLRs, Bl TLR, Lv TLR and 6 vertebrate TLR families (highlighted in different colors) are shown. TLR22 clade is shown with a red background. The detailed tree with all node supports can be found in **Supplementary Figure 2**.

transcription in basal conditions for all the 30 TLRs. The TLRs with gene ID of BITLR, BI48785, BI18798b, BI08928b and BI30396 showed a weak transcription while others were strongly expressed (**Figure 2**). Five of the genes were sequenced using specific primers confirming the identity of these genes (data not shown).

To better understand the *Branchiostoma* TLR evolution, we compared the domain structure of *B. lanceolatum*, *B. belcheri* and *B. floridae*. Therefore, we identified a total number of 30 TLRs in *B. lanceolatum*, 22 TLRs in *B. floridae* and 37 TLRs in *B. belcheri* (**Supplementary Tables 4–6**) according to the common TLR pattern. We also discriminated sccTLR and mccTLR in these three species according to the domain structure and phylogenetic analysis (**Supplementary Figure 1**). There are 3 mccTLRs in *B. floridae*, 5 mccTLRs in *B. belcheri* and 6 mccTLRs in *B. lanceolatum*. In addition, the mccTLRs found in the three *Branchiostoma* species consistently blast with invertebrate type TLRs (**Supplementary Tables 4–6**). We also studied the number of LRR from each TLR using LRRfinder software. The results showed that the LRR number of TLRs in the three species ranges from 1 to 25.

Identification and Characterization of a Novel BITLR

We focused on the amphioxus TLR11 family described in section The TLR Family in *B. lanceolatum* and specifically in a *B. lanceolatum* TLR sequence (BITLR) because it was highly similar to the published bbtTLR1 (GenBank: DQ400125.2). This *B. belcheri* gene was annotated as TLR1 based on phylogenetic and functional data (34). Nonetheless, our phylogenetic analysis pointed out that BITLR was a clear TLR11 family member. TLR11 family includes several teleost specific members (e.g., TLR19 or TLR22) that are not present in mammalian genomes and it is of great interest to know whether they are present in a more basal organism. To begin, we cloned the full-length of this novel BITLR (GenBank: MG437061) and its 5' and 3'-UTRs were obtained based on three orthologous found in the *Branchiostoma* genus. The length of the novel BITLR cDNA is 3,772 bp, containing a 227 bp long 5'UTR, a 2,913 bp ORF (which encodes a putative 970 amino acid-long protein), and a 616 bp long 3'UTR with a putative polyadenylation signal (AATAAA) 17 nucleotides upstream of the poly(A) tail (**Supplementary Figure 3**). SMART domain analysis predicted

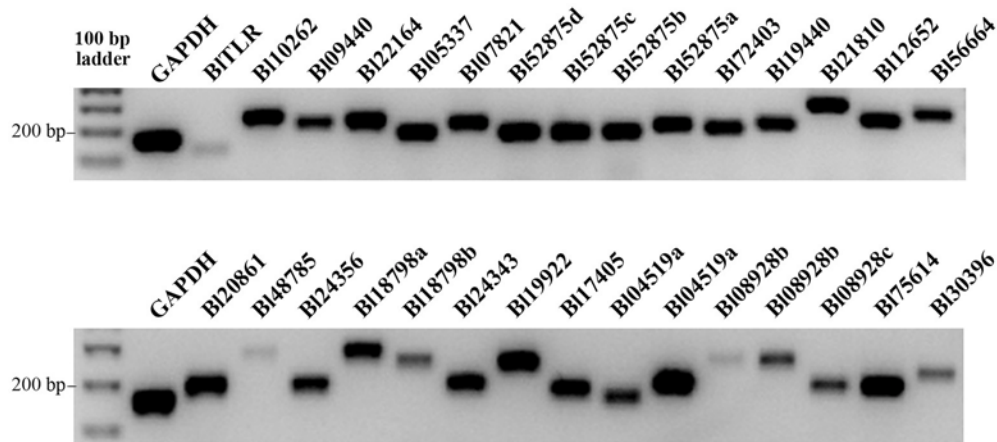


FIGURE 2 | Expression of TLR genes in *B. lanceolatum*. The cDNA used in all amplifications was prepared from whole animals. RT-PCR reactions were accomplished using equal number of cycles, the PCR products were loaded equally on two 1% agarose gels and GAPDH was used as a reference gene. Images were taken with the same exposure time using a Geldoc.

that the BITLR protein has the following domains: a C-terminal TIR domain (from residue 800 to 947), a transmembrane (TM) domain (from residue 752 to 774), a N-terminal signal peptide (first 27 residues), 21 tandem extracellular leucine-rich repeats (LRRs), a leucine rich repeat C-terminal domain (LRRCT) and a LRR N-terminal domain (LRRNT). The domain diagram of BITLR was made with IBS software (49) and shown in **Supplementary Figure 4**. The LRRs are flanked by one LRRCT and one LRRNT domain. The BITLR has only one LRRCT like most of the TLRs found in deuterostomes (scTLRs). The highly conserved consensus sequence (LxxLxLxxNxL) of each LRR was identified with the LRRfinder (**Supplementary Figures 3, 4**). Ten potential N-linked glycosylation sites were predicted by NetNGly 1.0: N¹⁰¹-N¹¹⁴-N¹⁵⁴-N¹⁶³-N²⁷⁶-N³⁷⁵-N³⁹³-N⁵²²-N⁵⁷³-N⁶³². The deduced molecular weight of BITLR protein is 111.3 kDa and the full-length protein showed 78.8% identity with the bbtTLR1 of *B. belcheri*. Three conserved boxes were identified in TIR domain of BITLR (**Supplementary Figure 3**). Box 1 and 2 are involved in binding downstream signaling molecules while box 3 is involved in the localization of the receptor through interactions with cytoskeletal elements (50). Importantly, a key residue in box 2 (Proline 681 in human TLR2 sequence) involved in MyD88 signaling was substituted by Ala in the BITLR sequence (51).

Expression Analysis of BITLR After LPS and Poly I:C Treatment

We performed RT-qPCR to investigate the expression profile of the BITLR in response to PAMP administration. This approach is often used to identify which family a putative TLR belongs to. Two representative PAMPs of bacterial and viral infection (LPS and Poly I:C, respectively) were used to challenge amphioxus *in vivo*. Amphioxus were immersed in 10 μ g/ml LPS or 10 μ g/ml Poly I:C to mimic the natural infection route. The gene transcription of BITLR was analyzed by RT-qPCR in a time course at 3, 6, 12 and 24 h post-immersion (**Figure 3**). However, no significant differences in gene expression were observed in any of the LPS

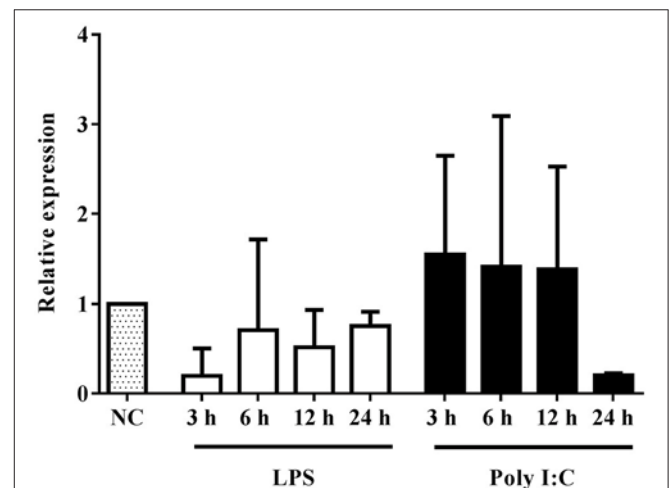


FIGURE 3 | Expression of BITLR gene after LPS or Poly I:C treatment. Animals previously immersed in 10 μ g/ml LPS or 10 μ g/ml Poly I:C, were collected at 3, 6, 12 and 24 h. The untreated animals were used as a control and assigned a value of 1 in the histogram. GAPDH was used as a reference gene. The bars indicate mean expression of 3 individual animals \pm S.D. Significant differences of mean values were analyzed according to one-way ANOVA followed by Tukey's test.

or Poly I:C-treated groups, indicating that 10 μ g/ml LPS or Poly I:C administered by immersion within this time frame could not significantly induce up- or down-regulation of the BITLR gene in adult amphioxus.

Subcellular Localization of BITLR in HEK293 Cells

We used HEK293 cells because these cells could be efficiently transfected and they have been extensively used for the study of TLR subcellular localization. Cells were transiently transfected with empty vector and the vector expressing the full-length BITLR. Flow cytometry analysis showed that cells

were successfully transfected at 24, 48 and 72 h and the transfection efficiency at 48 and 72 h (both around 60%) was higher than at 24 h (around 30%) post-transfection (**Figure 4A**). Western blot analysis confirmed that the BITLR protein was properly expressed in HEK293 cells, and it was not degraded by intracellular proteases. The BITLR protein was detected at 24, 48 and 72 h post-transfection (**Figure 4B**). The transcription levels were much higher at 48 and 72 h than at 24 h which agrees with the cytometry results. The molecular weight of BITLR protein was around 135 kDa which is slightly bigger than the theoretical one (111.27 kDa). This may be due to post-translational modifications such as glycosylation, phosphorylation, ubiquitination, ubiquitin-like modifications or S-nitrosylation among others.

To explore the subcellular localization of BITLR, we overexpressed the HA-tagged BITLR in HEK293 cells and we visualized the localization using immunofluorescence and confocal microscopy. We did not observe the HA-tagged BITLR in both transient and stable transfected cells when the cells were not permeabilized (**Figures 5B,D**). Non-transfected cells were used as a control (**Figure 5A**). This result indicates that, first, BITLR might be an intracellular protein; second, BITLR might localize on the plasma membrane but could not be detected in non-permeabilized cells due to the HA-tag location at the C-terminal. To further understand the localization of BITLR, we performed the assay with a plasma membrane marker (WGA) and different permeabilization methods. Interestingly, when the cells were permeabilized using different permeabilization methods (from weak to strong), we found that BITLR was mainly localized on the plasma membrane in both transient and stable transfected cells (**Figures 5C,E,F**).

BITLR Could Respond to Poly I:C in HEK293 Cells

Mammalian TLRs can transactivate the transcription factor NF- κ B in response to ligand binding. Usually, each TLR has a restricted PAMPs preference (**Supplementary Table 3**) and the

NF- κ B reporter assay allows functional discrimination between TLRs. To shed light on the role of novel BITLR in PAMPs recognition, a HEK293 cell line stably expressing BITLR was generated. However, the BITLR stable cells could not activate the NF- κ B promoter stimulated by any of the tested PAMPs (data not shown). To further study the receptor activity, we design a chimeric receptor fusing the ectodomain of BITLR with the TIR domain of human TLR2 and we generated a stable cell line. This approach has been used before to ensure a correct downstream signaling avoiding the differences in the set of adaptors and accessory proteins between vertebrates and non-vertebrates (34, 52). The chimeric BITLR stable cells responded to Poly I:C (LMW and HMW) which usually binds to TLR3 or TLR22. Conversely, other ligands, including Pam2CSK4 for TLR1/2, HKLM for TLR2, LPS for TLR4, flagellin for TLR5, FSL-1 for TLR2/6, imiquimod for TLR7, ssRNA for TLR8, ODN2006 for TLR9, ORN Sa19 for TLR13 (mouse) failed to induce NF- κ B transactivation (**Figure 6**). Human recombinant TNF α was used as a positive control since it is a well-known NF- κ B activator. In order to confirm that the up-regulation of luciferase activity is due to the Poly I:C recognition by the chimeric BITLR but not by endogenous TLRs, we performed the luciferase assay using chimeric BITLR stable cells and HEK293 cells without chimeric BITLR. The NF- κ B luciferase activity was up-regulated in chimeric BITLR stable cells with respect to HEK293 cells treated with Poly I:C (LMW and HMW; **Figure 6**).

Our results showed that the novel BITLR localized at the plasma membrane and responded to Poly I:C. These characteristics are only compatible with TLR22, thus we postulated that the novel receptor is a TLR22-like receptor. The alignment of BITLR with 12 teleost TLR22 sequences showed that BITLR had 27.8–30.8% of identity with fish TLR22 (**Supplementary Table 7**).

To further explore the phylogenetic relationship of BITLR, BbtTLR1 and vertebrate TLRs, phylogenetic trees were constructed based on full-length protein sequence using maximum-likelihood analysis (**Supplementary Figure 5**). As

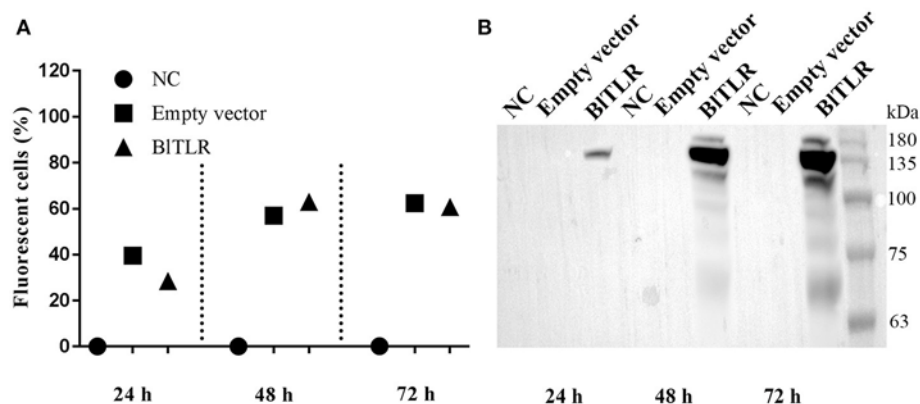


FIGURE 4 | BITLR expression in HEK293 cells. **(A)** HEK293 cells transfected with an empty vector (pIRES2-EGFP) and a vector expressing BITLR were analyzed at 24, 48 and 72 h post-transfection by flow cytometry. Non-transfected cells (NC) were used as a control. Transfection efficiency was evaluated as the percentage of GFP positive cells. **(B)** Non-transfected cells, cells transfected with the empty vector and the vector expressing BITLR with HA tag were analyzed at 24, 48 and 72 h post-transfection by western blot. Protein molecular weight standards (Nitorlab) are shown on the right side.

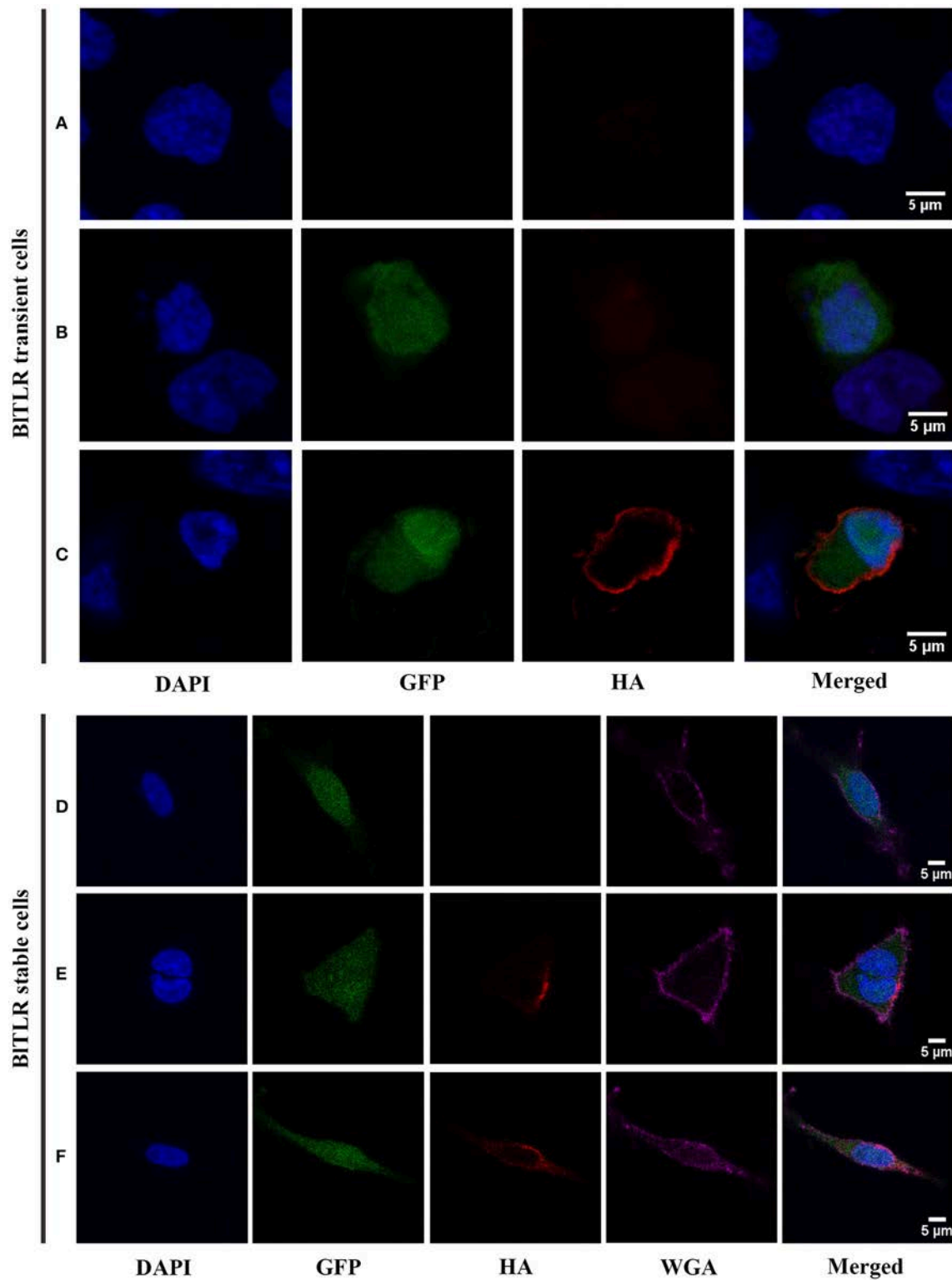


FIGURE 5 | Subcellular localization of BITLR in HEK293 cells. Confocal images showing HEK293 cells transiently transfected (**A–C**) or stably transfected (**D–F**) with BITLR. (**A**) Not transfected cells; (**B**) Cells transfected with BITLR and non-permeabilized; (**C**) Cells transfected with BITLR and permeabilized with 0.2% Triton X-100. (**D**) BITLR stable cells not permeabilized; (**E**) BITLR stable cells permeabilized using freeze and thaw protocol; (**F**) BITLR stable cells permeabilized with 0.1% Tween-20. Nuclei are stained with DAPI (in blue). Transfected cells were GFP labeled (in green). BITLR was detected with anti-HA antibody and AF555-conjugated anti-mouse IgG (in red). Plasma membrane was stained with WGA AF647 conjugated (in purple). Figures were analyzed with Fiji software.

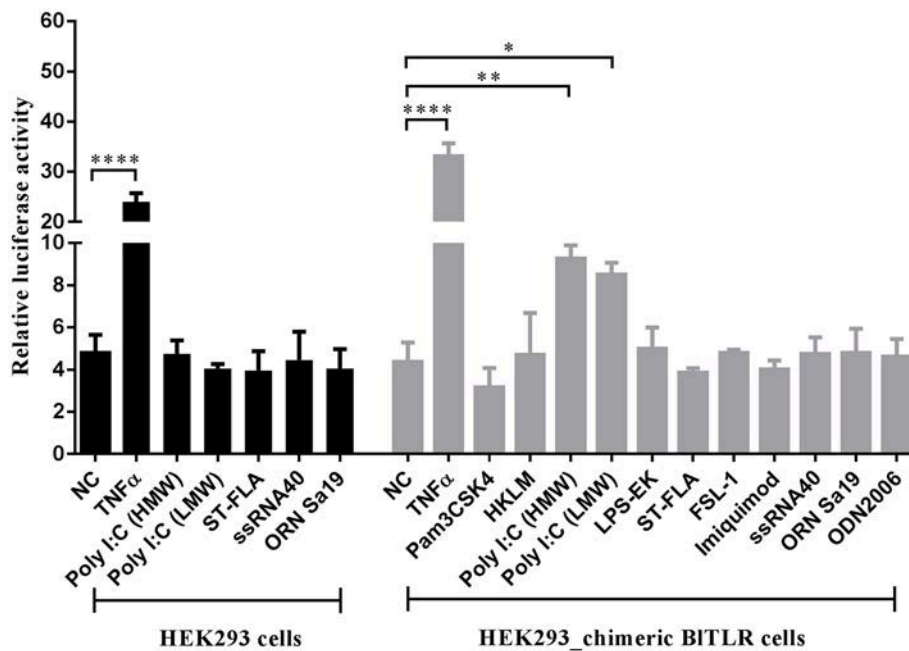


FIGURE 6 | HEK293 cells expressing chimeric BITLR induce the activation of NF- κ B in response to Poly I:C. HEK293 chimeric BITLR stable cells were treated with 11 different ligands (gray columns). Non-transfected HEK293 cells were treated with five potential ligands (black columns). Non-treated cells (NC) and cells treated with human TNF α (20 ng/ml) were used as negative and positive controls, respectively. The luciferase activity was expressed as the ratio of NF- κ B-dependent firefly luciferase activity divided by Renilla luciferase activity. Bars represented mean \pm S.D. Significant differences of mean values were analyzed according to one-way ANOVA followed by Tukey's test. * P < 0.05; ** P < 0.01; **** P < 0.0001.

expected, *D. melanogaster* Tolls clustered independently and all the vertebrate TLRs clustered into six clades. Furthermore, BITLR clustered in the same TLR11 family clade together with BbtTLR1. The SH-aLRT support (%) and ultrafast bootstrap support (%) are 94.6 and 94 (Supplementary Figure 5). This result indicates that BITLR is very likely to be a member of TLR11 family and could be identified as TLR11, 12, 13, 19, 20, 21, 22, or 23. Overall, all the results strongly support the identification of the novel receptor that carries the TLR22 function (BITLR22-like).

DISCUSSION

TLRs play crucial roles in the innate immune system by recognizing PAMPs from pathogens in vertebrates. In addition, TLRs have multiple functions ranging from developmental signaling to cell adhesion in protostomes (48). The study of TLRs may help to understand the role of TLR-mediated responses which could increase our range of strategies to treat infectious diseases and manipulate immune responses by drug intervention (53). From the evolutionary point of view, TLRs are conserved across invertebrates to vertebrates and absent from non-animal phyla (plants and fungi). However, there are vast structural and functional divergences in TLRs between invertebrates and vertebrates (15). In vertebrates, humans and mice have 10 and 12 TLRs, respectively and at least 16 TLRs have been identified in teleost; in urochordates, *Ciona intestinalis* has only two TLRs (54) whereas *Ciona savignyi* has between

8 and 20 (16); but in cephalochordates, *B. floridae* has an expansion of 48 TLRs according to Huang et al. (33). This expansion of TLRs in invertebrate deuterostomes remains to be understood by a comprehensive and thorough study of TLRs evolution. Amphioxus is a good model to study the invertebrate-chordate to vertebrate transition and the evolution of vertebrates. Therefore, studying TLR functions in such organism could improve our understanding of the ancestral innate immune system of vertebrates.

In this study, we identified 30 TLRs in *B. lanceolatum*, 22 TLRs in *B. floridae* and 37 TLRs in *B. belcheri* according to the basic TLR structure: "LRR+TM+TIR." Differences in the total number of *B. floridae* TLRs between Huang et al. and our data probably reflects discrepancies in the consensus of what is the basic structure of TLRs. Our stringent rule includes only those putative receptors with a TIR domain, a transmembrane domain and at least one LRR domain, known as true TLRs (15). Our available transcriptomic data maybe do not include all the possible TLRs. Probably the total number of TLRs in the 3 species of lancelet should be similar. Among them, we identified 6 mccTLRs in *B. lanceolatum*, 3 mccTLRs in *B. floridae* and 5 mccTLRs in *B. belcheri*. This finding is different from the observation by Huang et al. concerning amphioxus TLR family: it has a high rate of domain combination acquisition and therefore a high number of TLRs (prediction of 36 sccTLRs and 12 mccTLRs) (33). Importantly, Bányai and Patthy provided evidence to dispute that the rate of protein innovation is exceptionally high in lancelets. They surmised these high rates are

likely due to gene prediction errors (55). This might be the reason why there are less TLRs found in our study than the genomic prediction. Interestingly, if we remove 3 mccTLR sequences in *B. floridae* from our list, the total number of TLRs would be the same as reported by Tassia et al. which identified 19 TLRs (56). Moreover, the RT-PCR analysis showed that all the 30 TLRs of *B. lanceolatum* were truly expressed in adult animals. Our work shows that amphioxus and vertebrates share a conserved TLR framework in terms of protein structure. On the other hand, amphioxus TLRs maintain some features of invertebrates, such as the mccTLRs which are mainly found in protostomes (15). The function of remaining TLRs in PAMPs recognition remains unclear and needs further investigation.

We cloned the full-length sequence of BITLR22-like from Mediterranean amphioxus (*B. lanceolatum*). The full-length protein showed the highest identity (78.8%) with bbtTLR1 of *B. belcheri* that was annotated by the authors as a TLR1 based on the expression analysis after PAMPs injection *in vivo* (34) but the authors did not study the subcellular localization or the direct ligand specificity. The domain analysis of BITLR22-like protein sequence showed that it has a complete vertebrate-like ectodomain including a LRRCT, 21 LRRs and a LRRNT. The ectodomain forms a horseshoe structure to bind the specific PAMPs including the LRRCT that is responsible for dimerization which is necessary for complete ligand binding (57–59). The full-length protein sequences of BITLR22 are highly similar to the TLR22 of many fish species, suggesting that they may have similar ligand recognition, intracellular signal transduction pathway mechanisms and localization.

In mammals, TLRs can be divided into two main groups according to localization: on the cell surface or in intracellular compartments (60). Among human TLRs, the ones located at the plasma membrane (TLR1, 2, 4, 5 and 6) recognize microbial pathogenic components of the cell wall, while the others (TLR3, 7, 8 and 9) located intracellularly in endosomes or lysosome recognizing nucleic acids (4). However, the above ligand recognition pattern in non-mammalian organisms may be not always as in mammals. For instance, mouse TLR13 recognizes a conserved 23S ribosomal RNA (rRNA) from bacteria in the endolysosomal compartment (11). In teleost, TLR13 of *M. croaker* could respond to Poly I:C both *in vivo* and *in vitro* and is localized in the cytoplasm of HeLa cells (19). Fugu TLR22 recognizes long-sized dsRNA on the cell surface whereas TLR3 resides in the endoplasmic reticulum and recognizes relatively short-sized dsRNA (20). TLR22 of grass carp (*C. idella*) recognizes Poly I:C stimulation in CIK (*C. idella* kidney) cell line and is localized on the cell membrane (21). In our study, immunofluorescence and confocal microscopy showed that BITLR22-like is mainly localized on the plasma membrane.

In mammals, TLRs can recognize specific PAMPs with high levels of sensitivity (61). To test BITLR22-like ligand specificity, we performed different assays with commercially available mammalian TLRs ligands, using NF- κ B activity as a reporter. We could not observe significant differences of NF- κ B activation in HEK293 cells expressing BITLR22-like. There are different possible explanations but apart from problems with protein expression levels, intracellular degradation or incorrect

trafficking, the two most likely reasons could be: (1) BITLR22-like could not directly recognize PAMPs and the recognition process might require the assistance of other proteins that are specific for amphioxus and are not present in a mammalian system. For instance, *D. melanogaster* Tolls do not bind any PAMPs directly (62) and mammalian TLR4 cannot recognize LPS without the assistance of MD2 and CD14 (63) or; (2) BITLR22-like has a TIR domain that interacts with a species specific adaptor protein not present in mammalian cells. This hypothesis could be supported by the fact that P681 (human TLR2), extremely important to activate MyD88 signaling pathways in mammals (51), was not present in BITLR22-like neither in BbtTLR1 (**Supplementary Figure 3**). Thus, we could hypothesize that the absence of this Pro in the TLR22 sequence (Ala in BITLR22-like) explains why the TIR domain of BITLR22-like cannot activate MyD88 dependent signaling pathway in HEK293 unless we combine the ectodomain of TLR22 with the human TLR2 TIR domain. To test this hypothesis, we designed a chimeric protein containing the ectodomain and transmembrane domain of BITLR22-like fused to the human TLR2 TIR domain and we tested whether it could respond to ligand stimulation when stably transfected in HEK293 cells. Indeed, the cells expressing chimeric BITLR22-like activated significantly the NF- κ B reporter in response to both LMW and HMW Poly I:C. The magnitude of the stimulation is similar to other published data. For instance, Ji et al. characterized the activation of IFN and NF- κ B pathways by a teleost TLR19, and they found similar fold changes (around 2-fold change) as in our data (2.12 ± 0.1 -fold change Poly I:C HMW and 1.95 ± 0.09 -fold change Poly I:C LMW) (64). Other authors also have obtained similar fold-changes in the NF- κ B reporter assay (65, 66). On the other hand, Voogdt et al. showed an extremely high activation of the NF- κ B signaling pathway after flagellin stimulation but the main difference with our approach is that they used cells stably expressing NF- κ B reporter (67). Poly I:C is a specific ligand of vertebrate TLR3 including many fish species (20, 23, 65, 68), of *M. croaker* TLR13 (19) and of different fish TLR22 (20–22, 24, 69).

The phylogenetic analysis of BITLR22-like protein sequence and representative vertebrate TLR protein sequences revealed that BITLR22-like clusters with the vertebrate TLR11 family. Interestingly, the phylogenetic analysis of *B. floridae* TIR domain and vertebrate TLRs has indicated that 33 variable-type TLRs show a paraphyletic relationship with the vertebrate TLR11 lineage (33). The TLR11 family is represented in humans only by a pseudogene and the major divisions of the TLR11 family are clearly very ancient (16). Moreover, BITLR22-like has a single domain structure of the ectodomain which should be classified into TLR13 subfamily (**Supplementary Table 8**) according to the ectodomain architecture analysis of vertebrate TLRs (17). Taken together with its plasma membrane localization and functional analysis, we could further confirm the annotation of this TLR as an ortholog of vertebrate TLR11s, carrying a TLR22-like function and probably share a common ancestor with the fish specific TLR22. Overall, we provide evidence suggesting that TLR22 function may be an ancient and evolutionarily conserved antiviral response which emerged in Chordates.

DATA AVAILABILITY STATEMENTS

The datasets for this manuscript are not publicly available yet because the paper in which all these data are presented is in the final review process. Once the paper is published all the transcriptomic and genome data will be freely available, but until then requests to access the datasets should be directed to Dr. Jordi Garcia-Fernández (jordigarcia@ub.edu) and Dr. Hector Escrivà (hescrivà@obs-banyuls.fr).

ETHICS STATEMENT

All experimental procedures were approved by the Human and Animal Experimentation Ethics Committee of the Universitat Autònoma de Barcelona and were done in strict accordance with the recommendations of the European Directive (2010/63/EU) on the protection of animals used for scientific purposes.

AUTHOR CONTRIBUTIONS

JJ performed all the experiments and phylogenetic analysis. DR-V and AB helped to design the phylogenetic analysis. EN-P, CH-Ú and JG-F provided the TLR gene candidates from the transcriptome of *B. lanceolatum* and assisted with the phylogenetic analysis. JL provided the reporter assay plasmids and helped in the reporter assay design. HE provided the amphioxus and partial sequence of BITLR from the genome scaffold of *B. lanceolatum*. NR and JJ designed the experiments. JJ and NR prepared all the tables and figures and wrote the manuscript. All authors contributed to the correction of the final manuscript.

ACKNOWLEDGMENTS

This work was supported by grants from the Spanish Ministry of Science, European Commission and AGAUR to NR (AGL2015-65129-R MINECO/FEDER and 2014SGR-345 AGAUR). JJ was supported by a Ph.D. fellowship from the China Scholarship Council (201306300075), NR was supported by the Ramón y Cajal program (RYC-2010-06210, 2010, MINECO). AB and DR-V thanks to the grants BFU2012-34398, BFU2015-69717-P (MINECO/FEDER), SAF2014-52624-REDT, Ramón y Cajal Fellowship RYC-2011-08391, Career Integration Grant 304111 (EU), CERCA Programme/Generalitat de Catalunya and AGAUR 2014-SGR-297 and 2017-SGR-1776. HE was supported by the ANR (ANR16-CE12-0008-01). We thank N. Barba (Servei de Microscòpia) and Dr. M. Costa (Servei de Cultius Cel·lulars, Producció d'Anticossos i Citometria) from Universitat Autònoma de Barcelona for their technical assistance.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fimmu.2018.02525/full#supplementary-material>

Supplementary Figure 1 | Phylogenetic analysis of *B. floridae*, *B. belcheri* and *B. lanceolatum* TLRs. The phylogenetic tree was constructed by

maximum-likelihood method (IQ-TREE) using TIR domain sequences of *B. floridae*, *B. belcheri*, *B. lanceolatum*, *S. kowalevskii* and representative vertebrate TLRs. *D. melanogaster* Toll sequence was used as an outgroup to root the tree. Sequences were aligned by MAFFT with G-INS-i method and the best evolutionary model was established by ModelFinder according to BIC. Branch support was calculated running 1,000 replicates of the SH-aLRT and ultrafast bootstrap and they are represented as percentages at the tree nodes. The tree was generated in FigTree. Outgroup, mccTLRs and six vertebrate TLR families (by colors) were shown in the figure. BITLR is indicated by a red arrow.

Supplementary Figure 2 | Complete phylogenetic analysis of *B. lanceolatum* TLRs. The phylogenetic tree was constructed by IQ-TREE using full-length protein sequences. This tree is a more detailed version of the tree shown in **Figure 1**. All the values of SH-aLRT support and ultrafast bootstrap support are shown at the tree nodes. Outgroup, mccTLRs and 6 vertebrate TLR families (highlighted in different colors) are shown. The red arrow indicates BITLR. Additional information about the sequences can be found in **Supplementary Table 2, Supplementary Datas 1, 2**.

Supplementary Figure 3 | Nucleotide and deduced amino acid sequences of BITLR. Predicted transcription start site (TSS) is marked with a curved arrow. TATA box is boxed with a rectangle. The putative STAT5 and APIB transcription factor binding sites have a thick underline. The start codon (ATG), the stop codon (TAA) and the polyadenylation signal sequence (AATAAA) are in bold. The predicted signal peptide and the transmembrane region are underlined. The potential N-linked glycosylation sites are underlined and in bold. LRRCT domain predicted by LRRfinder is double underlined. The TIR domain predicted by SMART is underlined and highlighted in gray. The consensus sequence of LRR domain predicted by LRRfinder is highlighted in gray. The three consensus sequences of Toll/interleukin-1 receptor homology domain were boxed and underlined in gray: box 1 (FDAFISY), box 2 (GYKLC—RD—PG) and box3 (a conserved W surrounded by basic residues).

Supplementary Figure 4 | Predicted domain architecture of BITLR protein. The domain structure was predicted using the SMART program. Signal peptide (SP), leucine-rich repeat N-terminal domain (LRRNT), leucine-rich repeat (LRR), leucine rich repeat C-terminal domain (LRRCT), Transmembrane domain (TM) and Toll/interleukin-1 receptor (TIR) domain are indicated in figure. Figure was prepared with IBS software.

Supplementary Figure 5 | Phylogenetic analysis of BITLR. The phylogenetic tree was constructed by maximum-likelihood method (IQ-TREE) using full-length protein sequences. BITLR, BbtTLR1 and representative vertebrate TLR sequences were used in the analysis. *D. melanogaster* Toll was used as an outgroup to root the tree. Sequences were aligned with MAFFT choosing L-INS-i method and the alignments were trimmed using TrimAl with "Automated 1" mode. The best evolutionary model was established by ModelFinder according to BIC. One-thousand replicates of the SH-aLRT support and ultrafast bootstrap support are represented as percentages at the tree nodes. The tree was generated in FigTree. Outgroup and six vertebrate TLR families (by colors) are shown in figure. BITLR is indicated by a red arrow.

Supplementary Table 1 | Primers used for RT-PCR analysis.

Supplementary Table 2 | Vertebrate and invertebrate protein sequences used in the phylogenetic analysis.

Supplementary Table 3 | TLR ligands used in this study.

Supplementary Table 4 | TLRs in *B. lanceolatum*.

Supplementary Table 5 | TLRs in *B. floridae*.

Supplementary Table 6 | TLRs in *B. belcheri*.

Supplementary Table 7 | Protein sequence identity of BITLR and fish TLR22.

Supplementary Table 8 | Ectodomain architecture of vertebrate TLRs and BITLR.

Supplementary Data 1 | TLR sequences of *L. variegatus* and *S. kowalevskii* used in the phylogenetic analysis. The TIR domain of each TLR is highlighted in yellow.

Supplementary Data 2 | Identified DNA and putative protein sequences of TLRs in *B. lanceolatum*. The TIR domain of each TLR is highlighted in yellow.

REFERENCES

- Iwasaki A, Medzhitov R. Regulation of adaptive immunity by the innate immune system. *Science* (2010) 327:291–5. doi: 10.1126/science.1183021
- Aderem A, Ulevitch RJ. Toll-like receptors in the induction of the innate immune response. *Nature* (2000) 406:782–7. doi: 10.1038/35021228
- Aderem A, Underhill DM. Mechanisms of phagocytosis in macrophages. *Annu Rev Immunol.* (1999) 17:593–623. doi: 10.1146/annurev.immunol.17.1.593
- Akira S, Uematsu S, Takeuchi O. Pathogen recognition and innate immunity. *Cell* (2006) 124:783–801. doi: 10.1016/j.cell.2006.02.015
- Janeway CA. Approaching the asymptote? Evolution and revolution in immunology. *Cold Spring Harb Symp Quant Biol.* (1989) 54:1–13. doi: 10.1101/SQB.1989.054.01.003
- Janeway CA, Medzhitov R. Innate immune recognition. *Annu Rev Immunol.* (2002) 20:197–216. doi: 10.1146/annurev.immunol.20.083001.084359
- Takeda K, Kaisho T, Akira S. Toll-like receptors. *Annu Rev Immunol.* (2003) 21:335–376. doi: 10.1146/annurev.immunol.21.120601.141126
- Kobe B, Kajava AV. The leucine-rich repeat as a protein recognition motif. *Curr Opin Struct Biol.* (2001) 11:725–32. doi: 10.1016/S0959-440X(01)00266-4
- Kawai T, Akira S. The role of pattern-recognition receptors in innate immunity: update on Toll-like receptors. *Nat Immunol.* (2010) 11:373–84. doi: 10.1038/ni.1863
- Kawasaki T, Kawai T. Toll-like receptor signaling pathways. *Front Immunol.* (2014) 5:461. doi: 10.3389/fimmu.2014.00461
- Oldenburg M, Kruger A, Ferstl R, Kaufmann A, Nees G, Sigmund A, et al. TLR13 recognizes bacterial 23S rRNA devoid of erythromycin resistance-forming modification. *Science* (2012) 337:1111–5. doi: 10.1126/science.1220363
- Brinkmann MM, Spooner E, Hoebe K, Beutler B, Ploegh HL, Kim YM. The interaction between the ER membrane protein UNC93B and TLR3, 7, and 9 is crucial for TLR signaling. *J Cell Biol.* (2007) 177:265–75. doi: 10.1083/jcb.200612056
- Mathur R, Oh H, Zhang D, Park SG, Seo J, Koblansky A, et al. A mouse model of salmonella typhi infection. *Cell* (2012) 151:590–602. doi: 10.1016/j.cell.2012.08.042
- Temperley ND, Berlin S, Paton IR, Griñ n DK, Burt DW. Evolution of the chicken Toll-like receptor gene family: a story of gene gain and gene loss. *BMC Genomics* (2008) 9:62. doi: 10.1186/1471-2164-9-62
- Leulier F, Lemaitre B. Toll-like receptors — taking an evolutionary approach. *Nat Rev Genet.* (2008) 9:165–78. doi: 10.1038/nrg2303
- Roach JC, Glusman G, Rowen L, Kaur A, Purcell MK, Smith KD, et al. The evolution of vertebrate Toll-like receptors. *Proc Natl Acad Sci USA.* (2005) 102:9577–82. doi: 10.1073/pnas.0502272102
- Wang J, Zhang Z, Liu J, Zhao J, Yin D. Ectodomain architecture affects sequence and functional evolution of vertebrate toll-like receptors. *Sci Rep.* (2016) 6:26705. doi: 10.1038/srep26705
- Kumar H, Kawai T, Akira S. Toll-like receptors and innate immunity. *Biochem Biophys Res Commun.* (2009) 388:621–5. doi: 10.1016/j.bbrc.2009.08.062
- Wang Y, Bi X, Chu Q, Xu T. Discovery of toll-like receptor 13 exists in the teleost fish: *Miuy croaker* (Perciformes, Sciaenidae). *Dev Comp Immunol.* (2016) 61:25–33. doi: 10.1016/j.dci.2016.03.005
- Matsuo A, Oshiumi H, Tsujita T, Mitani H, Kasai H, Yoshimizu M, et al. Teleost TLR22 recognizes RNA duplex to induce IFN and protect cells from Birnaviruses. *J Immunol.* (2008) 181:3474–85. doi: 10.4049/jimmunol.181.5.3474
- Su J, Heng J, Huang T, Peng L, Yang C, Li Q. Identification, mRNA expression and genomic structure of TLR22 and its association with GCRV susceptibility/resistance in grass carp (*Ctenopharyngodon idella*). *Dev Comp Immunol.* (2012) 36:450–62. doi: 10.1016/j.dci.2011.08.015
- Hirono I, Takami M, Miyata M, Miyazaki T, Han HJ, Takano T, et al. Characterization of gene structure and expression of two toll-like receptors from Japanese flounder, *Paralichthys olivaceus*. *Immunogenetics* (2004) 56:38–46. doi: 10.1007/s00251-004-0657-2
- Hwang SD, Ohtani M, Hikima J, Jung TS, Kondo H, Hirono I, et al. Molecular cloning and characterization of Toll-like receptor 3 in Japanese flounder, *Paralichthys olivaceus*. *Dev Comp Immunol.* (2012) 37:87–96. doi: 10.1016/j.dci.2011.12.004
- Li H, Yang G, Ma F, Li T, Yang H, Rombout JHWM, et al. Molecular characterization of a fish-specific toll-like receptor 22 (TLR22) gene from common carp (*Cyprinus carpio* L.): evolutionary relationship and induced expression upon immune stimulants. *Fish Shellfish Immunol.* (2017) 63:74–86. doi: 10.1016/j.fsi.2017.02.009
- Ishii A, Kawasaki M, Matsumoto M, Tochinali S, Seya T. Phylogenetic and expression analysis of amphibian *Xenopus* Toll-like receptors. *Immunogenetics* (2007) 59:281–93. doi: 10.1007/s00251-007-0193-y
- Jatillon O, Aury JM, Brunet F, Petit JL, Stange-Thomann N, Maucell E, et al. Genome duplication in the teleost fish *Tetraodon nigroviridis* reveals the early vertebrate proto-karyotype. *Nature* (2004) 431:946–57. doi: 10.1038/nature03025
- Pietretti D, Wiegertjes GF. Ligand specificities of Toll-like receptors in fish: Indications from infection studies. *Dev Comp Immunol.* (2014) 43:205–22. doi: 10.1016/j.dci.2013.08.010
- Kasahara M, Naruse K, Sasaki S, Nakatani Y, Qu W, Ahsan B, et al. The medaka draft genome and insights into vertebrate genome evolution. *Nature* (2007) 447:714–9. doi: 10.1038/nature05846
- Putnam NH, Butts T, Ferrier DEK, Furlong RF, Hellsten U, Kawashima T, et al. The amphioxus genome and the evolution of the chordate karyotype. *Nature* (2008) 453:1064–71. doi: 10.1038/nature06967
- Schubert M, Escriva H, Xavier-Neto J, Laudet V. Amphioxus and tunicates as evolutionary model systems. *Trends Ecol Evol.* (2006) 21:269–77. doi: 10.1016/j.minpro.2005.11.002
- Theodosiou M, Colin A, Schulz J, Laudet V, Peyrieras N, Nicolas JE, et al. Amphioxus spawning behavior in an artificial seawater facility. *J Exp Zool Part B Mol Dev Evol.* (2011) 316 B:263–75. doi: 10.1002/jez.b.21397
- Yu JKS, Holland LZ. Cephalochordates (Amphioxus or Lancelets): a model for understanding the evolution of chordate characters. *Cold Spring Harb Protoc.* (2009) 2009.pdb.emo130. doi: 10.1101/pdb.emo130
- Huang S, Yuan S, Guo L, Yu YY, Li J, Wu T, et al. Genomic analysis of the immune gene repertoire of amphioxus reveals extraordinary innate complexity and diversity. *Genome Res.* (2008) 18:1112–6. doi: 10.1101/gr.069674.107
- Yuan S, Huang S, Zhang W, Wu T, Dong M, Yu Y, et al. An amphioxus TLR with dynamic embryonic expression pattern responses to pathogens and activates NF- κ B pathway via MyD88. *Mol Immunol.* (2009) 46:2348–356. doi: 10.1016/j.molimm.2009.03.022
- Cameron RA, Samanta M, Yuan A, He D, Davidson E. SpBase: The sea urchin genome database and web site. *Nucleic Acids Res.* (2009) 37:D750–4. doi: 10.1093/nar/gkn887
- Simakov O, Kawashima T, Marlétaz F, Jenkins J, Koyanagi R, Mitros T, et al. Hemichordate genomes and deuterostome origins. *Nature* (2015) 527:459–65. doi: 10.1038/nature16150
- Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol Biol Evol.* (2013) 30:772–80. doi: 10.1093/molbev/mst010
- Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. trimAl: A tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* (2009) 25:1972–73. doi: 10.1093/bioinformatics/btp348
- Nguyen LT, Schmidt HA, Von Haeseler A, Minh BQ. IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol Biol Evol.* (2015) 32:268–74. doi: 10.1093/molbev/msu300
- Kalyaanamoorthy S, Minh BQ, Wong TKF, Von Haeseler A, Jermini LS. ModelFinder: Fast model selection for accurate phylogenetic estimates. *Nat Methods* (2017) 14:587–9. doi: 10.1038/nmeth.4285
- Guindon S, Dufayard JE, Lefort V, Anisimova M, Hordijk W, Gascuel O. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst Biol.* (2010) 59:307–21. doi: 10.1093/sysbio/syq010
- Hoang DT, Chernomor O, von Haeseler A, Minh BQ, Vinh LS. UFBoot2: improving the ultrafast bootstrap approximation. *Mol Biol Evol.* (2018) 35:518–22. doi: 10.1093/molbev/msx281
- Livak KJ, Schmittgen TD. Analysis of relative gene expression data using real-time quantitative PCR and the 2(- $\Delta\Delta C(T)$) method. *Methods* (2001) 25:402–8. doi: 10.1006/meth.2001.1262
- Schmittgen TD, Livak KJ. Analyzing real-time PCR data by the comparative CT method. *Nat Protoc.* (2008) 3:1101–8. doi: 10.1038/nprot.2008.73

45. Mardones G, González A. Selective plasma membrane permeabilization by freeze-thawing and immunofluorescence epitope access to determine the topology of intracellular membrane proteins. *J Immunol Methods* (2003) 275:169–77. doi: 10.1016/S0022-1759(03)00015-2
46. Schindelin J, Arganda-Carreras I, Frise E, Kaynig V, Longair M, Pietzsch T, et al. Fiji: an open-source platform for biological-image analysis. *Nat Methods* (2012) 9:676–82. doi: 10.1038/nmeth.2019
47. Brennan JJ, Gilmore TD. Evolutionary origins of toll-like receptor signaling. *Mol Biol Evol.* (2018) 35:1576–87. doi: 10.1093/molbev/msy050
48. Buckley KM, Rast JP. Dynamic evolution of toll-like receptor multigene families in echinoderms. *Front Immunol.* (2012) 3:136. doi: 10.3389/fimmu.2012.00136
49. Liu W, Xie Y, Ma J, Luo X, Nie P, Zuo Z, et al. IBS: an illustrator for the presentation and visualization of biological sequences: Fig. 1. *Bioinformatics* (2015) 31:3359–61. doi: 10.1093/bioinformatics/btv362
50. Subramaniam S, Stansberg C, Cunningham C. The interleukin 1 receptor family. *Dev Comp Immunol.* (2004) 28:415–28. doi: 10.1016/j.dci.2003.09.016
51. Brown V, Brown RA, Ozinsky A, Hesselberth JR, Fields S. Binding specificity of toll-like receptor cytoplasmic domains. *Eur J Immunol.* (2006) 36:742–53. doi: 10.1002/eji.200535158
52. Wang M, Wang L, Guo Y, Sun R, Yue F, Yi Q, Song L. et al. The broad pattern recognition spectrum of the Toll-like receptor in mollusk Zhikong scallop *Chlamys farreri*. *Dev Comp Immunol.* (2015) 52:192–201. doi: 10.1016/j.dci.2015.05.011
53. O'Neill LAJ, Hennessy EJ, Parker AE. Targeting Toll-like receptors: emerging therapeutics? *Nat Rev Drug Discov.* (2010) 9:293–307. doi: 10.1038/nrd3203
54. Sasaki N, Ogasawara M, Sekiguchi T, Kusumoto S, Satake H. Toll-like receptors of the Ascidian *Ciona intestinalis* prototypes with hybrid functionalities of vertebrate Toll-like receptors. *J Biol Chem.* (2009) 284:27336–43. doi: 10.1074/jbc.M109.032433
55. Bányai L, Patthy L. Putative extremely high rate of proteome innovation in lancelets might be explained by high rate of gene prediction errors. *Sci Rep.* (2016) 6:30700. doi: 10.1038/srep30700
56. Tassia MG, Whelan NV, Halanych KM. Toll-like receptor pathway evolution in deuterostomes. *Proc Natl Acad Sci U S A.* (2017) 114:7055–60. doi: 10.1073/pnas.1617722114
57. Wang Y, Liu L, Davies DR, Segal DM. Dimerization of Toll-like receptor 3 (TLR3) is required for ligand binding. *J Biol Chem.* (2010) 285:36836–41. doi: 10.1074/jbc.M110.167973
58. Botos I, Segal DM, Davies DR. The structural biology of Toll-like receptors. *Structure* (2011) 19:447–59. doi: 10.1016/j.str.2011.02.004
59. Leonard JN, Ghirlando R, Askins J, Bell JK, Margulies DH, Davies DR, et al. The TLR3 signaling complex forms by cooperative receptor dimerization. *Proc Natl Acad Sci USA.* (2008) 105:258–63. doi: 10.1073/pnas.0710779105
60. Takeda K, Akira S. Toll-like receptors in innate immunity. *Int Immunol.* (2005) 17:1–14. doi: 10.1093/intimm/dxh186
61. Akira S, Takeda K, Kiyoshi T. Toll-like receptor signalling. *Nat Rev Immunol.* (2004) 4:499–511. doi: 10.1038/nri1391
62. Hoffmann JA. The immune response of *Drosophila*. *Nature* (2003) 426:33–8. doi: 10.1038/nature02021
63. Shimazu R, Akashi S, Ogata H, Nagai Y, Fukudome K, Miyake K, et al. MD-2, a Molecule that confers lipopolysaccharide responsiveness on Toll-like receptor 4. *J Exp Med.* (1999) 189:1777–82. doi: 10.1084/jem.189.11.1777
64. Ji J, Rao Y, Wan Q, Liao Z, Su J. Teleost-specific TLR19 localizes to endosome, recognizes dsRNA, recruits TRIF, triggers both IFN and NF- κ B pathways, and protects cells from grass carp reovirus infection. *J Immunol.* (2018) 200:573–85. doi: 10.4049/jimmunol.1701149
65. Alexopoulou L, Czopik Holt A, Medzhitov R, Flavell RA. Recognition of double-stranded RNA and activation of NF-kappa B by Toll-like receptor 3. *Nature* (2001) 413:732–8. doi: 10.1038/35099560
66. Sepulcre MP, Alcaraz-Perez F, Lopez-Munoz A, Roca FJ, Meseguer J, Cayuela ML, Mulero V, et al. Evolution of lipopolysaccharide (LPS) recognition and signaling: fish TLR4 does not recognize LPS and negatively regulates NF- κ B activation. *J Immunol.* (2009) 182:1836–45. doi: 10.4049/jimmunol.0801755
67. Voogdt CGP, Wagenaar JA, van Putten JPM. Duplicated TLR5 of zebrafish functions as a heterodimeric receptor. *Proc Natl Acad Sci USA.* (2018) 115:E3221–9. doi: 10.1073/pnas.1719245115
68. Rodriguez MF, Wiens GD, Purcell MK, Palti Y. Characterization of Toll-like receptor 3 gene in rainbow trout (*Oncorhynchus mykiss*). *Immunogenetics* (2005) 57:510–9. doi: 10.1007/s00251-005-0013-1
69. Ding X, Lu DQ, Hou QH, Li SS, Liu XC, Zhang Y, et al. Orange-spotted grouper (*Epinephelus coioides*) toll-like receptor 22: molecular characterization, expression pattern and pertinent signaling pathways. *Fish Shellfish Immunol.* (2012) 33:494–503. doi: 10.1016/j.fsi.2012.05.034

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2018 Ji, Ramos-Vicente, Navas-Pérez, Herrera-Úbeda, Lizcano, García-Fernández, Escrivà, Bayés and Roher. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Origin and evolution of the chordate central nervous system: insights from amphioxus genoarchitecture

BEATRIZ ALBUIXECH-CRESPO¹, CARLOS HERRERA-ÚBEDA¹, GEMMA MARFANY¹, MANUEL IRIMIA^{*,2,3}
and JORDI GARCIA-FERNÁNDEZ^{*,1}

¹Dept. Genetics, Microbiology and Statistics and Institute of Biomedicine (IBUB), Faculty of Biology, University of Barcelona, ²EMBL/CRG Systems Biology Research Unit, Centre for Genomic Regulation (CRG), The Barcelona Institute for Science and Technology, Barcelona, Spain and ³Universitat Pompeu Fabra (UPF), Barcelona, Spain

ABSTRACT The vertebrate brain is arguably the most complex anatomical and functional structure in nature. During embryonic development, the central nervous system (CNS) undergoes a series of morphogenetic processes that eventually obscure the major axes of the early neural plate to our perception. Notwithstanding this complexity, the “genoarchitecture” of the developing neural tube brings into light homologous regions between brains of different vertebrate species, acting as a molecular barcode of each particular domain. Those homologous regions and their topological interrelations constitute the ancestral, deeply conserved, *bauplan* of the vertebrate brain. Remarkably, although simpler, the cephalochordate amphioxus shares multiple features of this *bauplan*, serving as a privileged reference point to understand the origins of the vertebrate brain. Here, we review the development of the chordate CNS in view of the latest morphological and genoarchitectonic data from amphioxus. This comparison reveals that the amphioxus CNS is far from simple and provides unique insights into the structure of the vertebrate CNS and its evolutionary origins. In particular, we summarize recent research in amphioxus and vertebrates that has challenged views on the major partitions of the vertebrate brain, proposing a novel organization of the chordate CNS *bauplan* that better reflects developmental and evolutionary data.

KEY WORDS: *bauplan*, brain partition, DiMes, vertebrate CNS, cephalochordate

Introduction

Homology is a fundamental concept in biology and one of the key pillars of comparative embryology. Therefore, it is not surprising that since its early years, Evo-Devo researchers have discussed whether anatomy or gene expression are better indicators of “strict” homology (“*the same organ in different animals under every variety of form and function*”, Owen 1843, pp.379). These controversies started soon after developmental gene expression data from single genes were first compared between two species (e.g., Abouihel *et al.*, 1997, Wray and Abouihel 1998), and still persist in the post-genomic era (Tschopp & Tabin, 2016), when entire transcriptomes can be compared.


It is not surprising, then, that searching for homologies, divergences, and innovations in the most complex and fascinating structure in nature – the central nervous system (CNS) of vertebrates – has

been, and surely will be, the subject of study from very different perspectives by many researchers. Currently, there are large amounts of gene expression data that can be used as topological markers in the developing vertebrate CNS. As all vertebrate neural tubes are homologous and develop under similar principles, it is generally assumed that gene expression patterns, which provide positional information and give insights into the specification and maintenance of brain derivatives, are, overall, good markers of homology (Morona *et al.*, 2011, 2016). However, comparisons between vertebrate and non-vertebrate CNS are more challenging, as similarity of (relative) gene expression patterns are evaluated between highly diverse, and often quite dissimilar, anatomical structures, without a clear

Abbreviations used in this paper: AP, anteroposterior; ARCH, archencephalic protagma, CNS, central nervous system; DEU, deuteroencephalic protagma; DI, diencephalon; DiMes, DiMesenphalic primordium; MB, midbrain.

***Address correspondence to:** Manuel Irimia. Centre for Genomic Regulation (CRG), Barcelona Institute of Science and Technology (BIST), Barcelona, Spain.

E-mail: mirimia@gmail.com  <http://orcid.org/0000-0002-2179-2567> or

Jordi Garcia-Fernández. Dep. Genetics, Microbiology and Statistics and Institute of Biomedicine (IBUB), Faculty of Biology, University of Barcelona. Av. Diagonal 643, 08028 Barcelona, Spain. E-mail: jordigarcia@ub.edu - <http://www.ub.edu/genetica/evo-devoen/garciare.htm>  <http://orcid.org/0000-0001-5677-5970>

Submitted: 10 October, 2017; Accepted: 30 October, 2017.

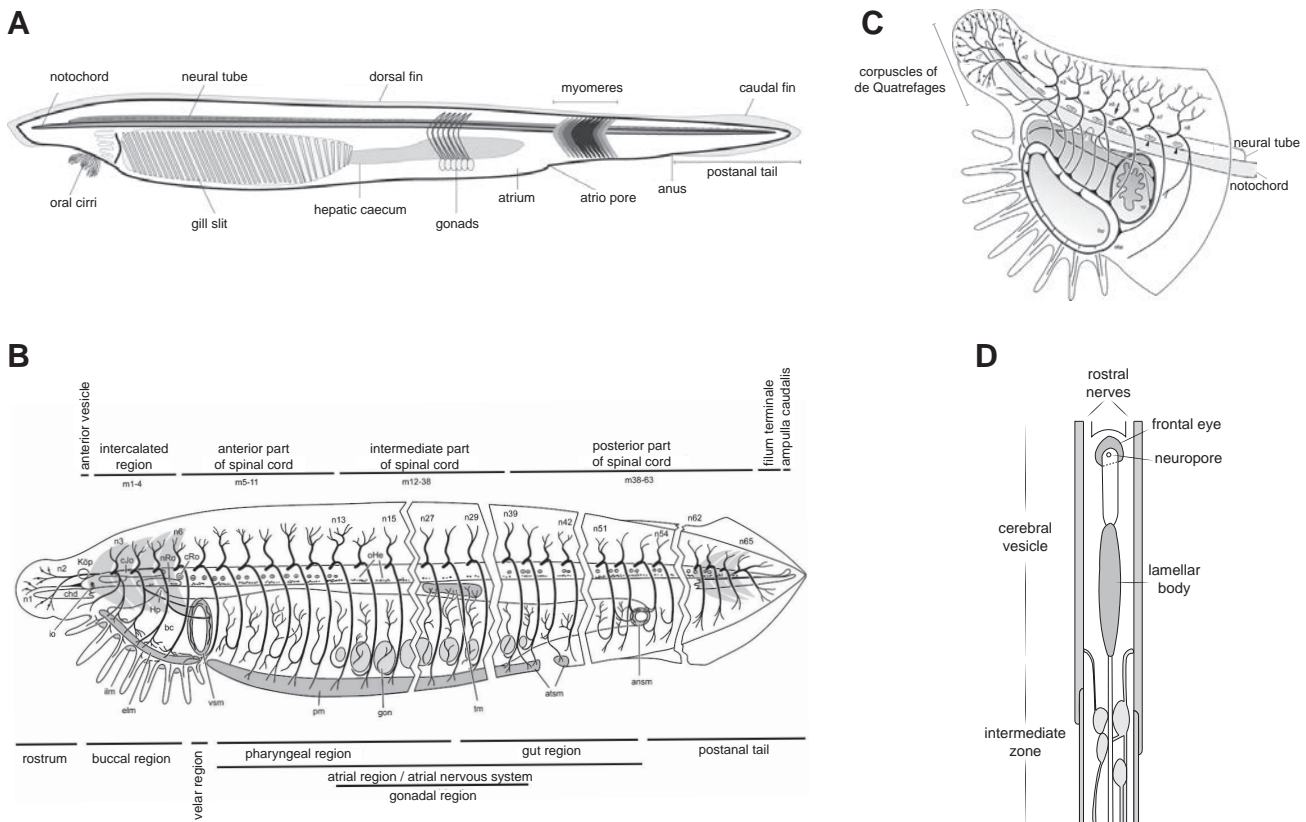


Fig. 1. Amphioxus anatomy and neural system. (A) Schematic representation of an adult amphioxus, indicating major anatomical structures. (B) Innervation of the adult amphioxus, dorsal nerves 1, 2, 3... (n1, n2, n3...), infundibular organ (io), notochord (chd), Kölliker's pit (Köp), Joseph cells (cJo), nucleus of Rohde (nRo), Rohde cell (cRo), Hatschek's pit (Hp), inner labial muscle (ilm), external labial muscle (elm), buccal cavity (bc), velar sphincter muscle (vsm), pterygeal muscle (pm), gonad (gon), organ of Hesse (oHe), trapezius muscle of atriocoelomic funnel (tm), atriopore sphincter muscle (atsm), anal sphincter muscle (ansm). (C) Detail of innervation in the most anterior tip, velar plexus (vp), inner buccal plexus (ibp), outer buccal plexus (obp), first cell of Rohde (arrow), Hatschek's pit (white arrowhead), anastomosis between nerves 1 and 2 (open arrowhead), two of the neuromuscular contact zones (black arrowheads). (D) Schematic representation of the neural tube (dorsal view, anterior is up). (B,C) Reproduced with slight modifications from Witch and Lacalli (2005, © 2008 Canadian Science Publishing or its licensors, reproduced with permission). (D) Adapted from Lacalli (1994). Anterior is to the right and dorsal is up in (A,B,C).

homologous topological reference. Hence, the study of amphioxus, which has a centralized CNS that develops from a neural plate that is unambiguously homologous to that of vertebrates, has been particularly useful to shed light into how the CNS of vertebrates has originated and evolved.

Here, after an introduction to the key phylogenetic position of the amphioxus in the path to vertebrates, we review classic and recent studies on vertebrates and cephalochordates to provide a comparative view of the development and morphology of their CNS. We begin by setting the conceptual framework of our comparisons: comparative neuroanatomy and the prosomeric model. We will then review recent results suggesting a novel ontogenetic organization of vertebrate brain regions as inferred from comparative studies with amphioxus, and, finally, we will discuss the impact of this organization on our understanding of key adult brain structures in cephalochordates.

The amphioxus: key to understanding chordate and vertebrate transitions

For centuries, amphioxus has attracted the attention of zoologists,

embryologists and, since 1980's, also of molecular biologists. First described by Peter Simon Pallas as a molluscan slug in 1774 (Pallas, 1774), later classified as a close relative of agnathan vertebrates, hagfish and lampreys (Costa, 1834), its phylogenetically privileged position is now thought to be that of being the most basal-branching extant chordate (Delsuc *et al.*, 2006). In this Special Issue of the *Int. J. Dev. Biol.*, readers can learn about the fascinating quest of several research groups to collect, get to reproducing in captivity, and work with amphioxus embryos, first using large amounts of radioactive ^{32}P probes, then performing many PCRs, and from struggling for genome sequencing funding to, of late, trying all available and possible gene modification protocols (see Holland, this issue).

Cephalochordates share the chordate phylum with the Olfactores (Urochordata and Vertebrata) (Delsuc *et al.*, 2006), with an estimated time of divergence from the last common ancestor at around 550 million years ago (Blair and Hedges, 2005; Delsuc *et al.*, 2008). The main characteristic of chordates is the presence of a notochord, which extends along much of the body; in the particular case of cephalochordates, the notochord seems to extend beyond the most anterior tip of the neural tube (Fig. 1A), although this anterior-most part has been suggested to be homologous to the

prechordal plate (see below). The very name of amphioxus is in fact an allegory of its shape (amphis = both, oxy = sharp). Unlike vertebrates, the amphioxus notochord is maintained throughout life and is the only skeletal structure present in the adult body. The hollow neural tube, which is dorsal to the notochord, is another of the morphological characteristics shared with vertebrates, in addition to the gill slits in the pharynx, the myomeres (or segmental muscle “bricks”), and the post-anal tail (García-Fernández and Benito-Gutiérrez, 2009)

Not only has its prototypical body plan (bauplan) and phylogenetic position enticed evolutionary biologists, but also its genome is highly remarkable. One of the first genes cloned in amphioxus (Holland *et al.*, 1992) supported the notion that the amphioxus had a prototypical genome with respect to the vertebrate – namely human – genome. In 1994, after the cloning of the single amphioxus Hox gene cluster (García-Fernández and Holland, 1994), Peter Holland together with one of the co-authors here (back when they were both very young) and other colleagues, proposed that the origin of vertebrates was linked to the double duplication of an ancestral genome that was very similar to the present amphioxus genome, the so-called 2R hypothesis (Holland *et al.*, 1994). Susumu Ohno already noticed the potential of gene duplication in evolution years before, based on earlier measures of genome size, and even proposed that the genome was duplicated somewhere in the path to vertebrates (Ohno, 1970). But it was not until 2005, with the publication of the first urochordate genome (Dehal and Boore, 2005), and definitely in 2008, with the publication of the amphioxus genome sequence draft (Putman *et al.*, 2008), that this hypothesis was corroborated: about 95% of the human genome could be traced back to a double polyploidisation of an amphioxus-like genome. In these 14 years, the idea that those new genes were instrumental to the morphological innovations of vertebrates shifted subtly from emphasizing additional protein-coding sequences to the relevance of duplication and innovation of regulatory sequences, through the Duplication, Degeneration, and Complementation model (the DDC model, Force *et al.*, 1999), or the Duplication, Degeneration, and Innovation model of regulatory DNA (the DDI model, Jiménez-Delgado *et al.*, 2009). More recently, the link between genome duplication and epigenetic changes (Acemel *et al.*, 2016) also supports the view that the 2R events at the origin and early evolution of vertebrates were instrumental to the path that led to us, humans.

In this review, we will show how amphioxus has also illuminated the origin and evolution of the vertebrate CNS, arguably the most paramount and defining organ of our clade.

Conceptual framework: the prosomeric model

A major breakthrough in our understanding of the bauplan of vertebrate CNS came with the publication of the prosomeric model by Puelles and Rubenstein (Puelles and Rubenstein, 1993; Puelles, 1995; Puelles,

2010; Puelles and Rubenstein, 2015). This model proposes that the vertebrate CNS is composed of basic units of neural development, the neuromeres, whose organizational principles are identifiable in the longitudinal and transversal axes across all studied vertebrate species (Fig. 2 B-B’). The prosomeric model facilitates comparisons across vertebrate species, since neuromeres refer to the ‘intrinsic’ axes, as established in the early neural plate, instead of the ‘extrinsic’ axes that had been used in other neuroanatomic approximations (Kuhlenbeck, 1967; Puelles and Rubenstein, 1993), and which are greatly affected by the complex morphogenetic events undergone by the neural tube upon closure (Fig. 2).

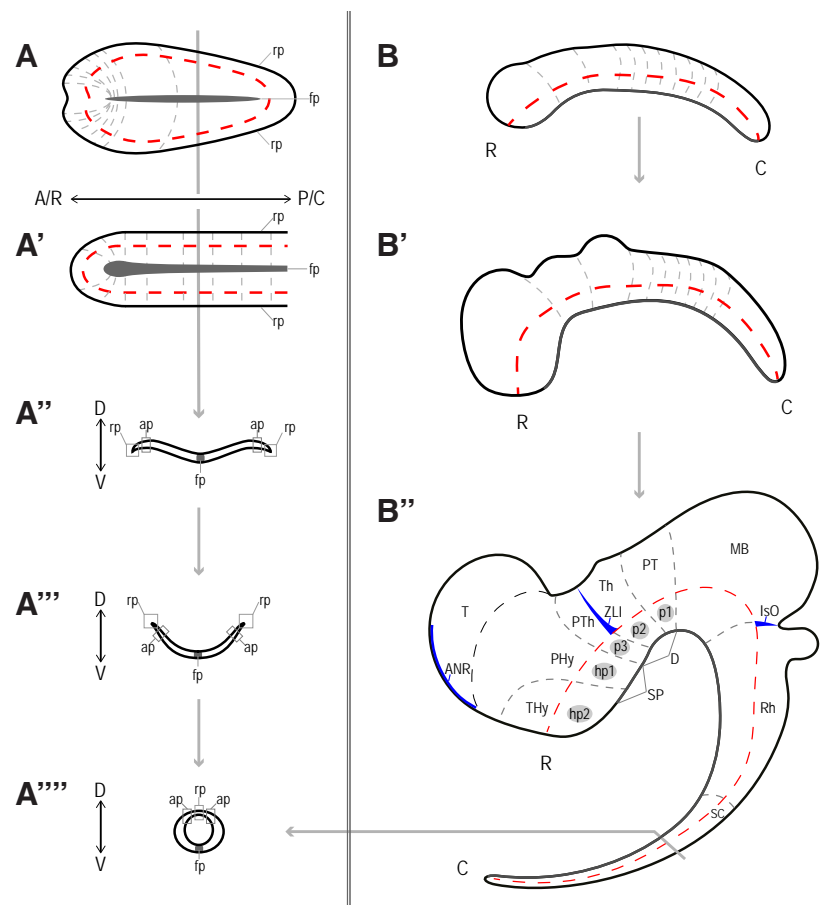


Fig. 2. Chordate neurulation and vertebrate central nervous system development.

The vertebrate (A) and amphioxus (A') neural plate (dorsal view) fold (A'', A''') to form the neural tube (A''''), similarly, generating a hollow nerve tube that closes dorsally by the roofplate (rp). (B) Schematic lateral views of bending of the vertebrate neural tube as a result of differential growth, resulting in an arching of its internal axis (red dotted line). Some neuromeres are indicated as the regionalization of the neural tube proceeds (grey dotted lines). The forebrain is subdivided into secondary prosencephalon (SP) with peduncular region and terminal prosomeres (hp1 and hp2, respectively), which include the hypothalamic region, telencephalon (T), and the optic vesicles, and Diencephalon (D), with prosomeres 1, 2, and 3 (p1–p3), which are represented by the pretectum (PT), thalamus (Th), and prethalamus (PTh), respectively. More caudally, midbrain (MB), rhombencephalon (Rh), and spinal cord (SC) regions are identified (for simplicity, their respective neuromeric components are not depicted). The position of the secondary organizers zona limitans intrathalamica (ZLI) and the isthmus organizer (IsO) are represented by blue spikes. Lateral view. Dark grey represents floorplate (fp) in all schemes; terminal hypothalamic prosomere (Thy), peduncular hypothalamic prosomere (PHY). A/R, anterior/rostral; P/C, posterior/caudal; D, dorsal; V, ventral. Adapted from Albuixech-Crespo *et al.*, (2017).

These primary axes are easy to identify in the early neural plate (Fig. 2A). The anteroposterior (AP) axis overlaps the longitudinal axis, whereas the mediolateral axis will correspond to the dorsoventral (DV) axis after neural tube closure, in which the neural plate midline and border will become the neural tube floor and roof, respectively. Interestingly, this layout results in an anatomical singularity in the neural plate, since the presumptive basal (ventral) plate, alar (dorsal) plate and roof run concentrically across the floor in the anterior and posterior tips of the neural plate (Fig. 2A, B-B"). As development progresses, the AP axis bends due to the differential growth of different territories of the neural tube (Fig. 2 B-B"). Despite this bending, each neuromere is established according to its position with respect to the AP axis (Fig. 2B). Moreover, the DV axis maintains its orthogonal relationship with the AP axis. Thus, the prosomeric model organically defines the basic developmental units of a tissue whose morphology is highly variable and plastic, assisting comparisons neuroanatomies of equivalent bauplans (Puelles and Rubenstein, 1993; Puelles, 1995; Puelles and Rubenstein, 2015).

Moreover, an additional crucial aid for comparative neuroanatomy is provided by gene expression patterns. During the development of the CNS, different populations of neural progenitors are established based on their molecular codes. The pool of transcriptionally active genes is different for each particular group of cells, and the dynamics of these specification processes eventually result into positional identities in the neuroepithelium, what is known as neural patterning. This relationship between specific spatiotemporal gene expression patterns and the positional identity of the neural progenitors is the basis for the genoarchitecture. This concept has been exploited by Evo-Devo approaches, facilitating comparisons of developing vertebrate brains and allowing direct homology assignments (Ferran *et al.*, 2007; Ferran *et al.*, 2009; Medina *et al.*, 2011; Puelles and Ferran, 2012). Similar approaches have also been undertaken to compare domains between more distantly related species. In particular, similarity of relative gene expression patterns along the AP axis have received much attention, reaching the conclusion that a subset of genes have likely maintained their relative AP positions since the last common ancestor of Bilateria (Lowe *et al.*, 2003; Castro *et al.*, 2006; Hirth, 2010; Irimia *et al.*, 2010). However, as mentioned above, these inferences are often obscured by unavoidable problems encountered when comparing highly divergent CNSs.

Secondary anteroposterior organizers: major players in vertebrate brain development and evolution

The vertebrate CNS becomes regionalized along its AP axis very early in development. At the neural plate stage, expression patterns of genes such as *Otx2*, *Gbx2*, *Fz3*, *Irx* or *Pax6* start establishing molecular subdivisions (Hidalgo-Sánchez *et al.*, 2005; Rodríguez-Seguel *et al.*, 2009). As development proceeds, the neural tube closes, bends, and the classical general AP subdivisions become morphologically apparent: the prosencephalon (which will be further subdivided into secondary prosencephalon and diencephalon), the midbrain, the rhombencephalon and the spinal cord. All these large AP domains will progressively regionalize and subdivide into neuromeres. For example, the rhombencephalon will metamere into 11 rhombomeres (the neuromeres of the rhombencephalon), and the diencephalon into 3 prosomeres, with distinctive alar

components: the prethalamus (p3), the thalamus (p2) and the pretegmentum (p1), in a rostrocaudal order (Puelles and Rubenstein, 1993; Puelles, 1995) (Fig. 2).

Many of these subdivisions are orchestrated by the action of the secondary organizers (blue dashes in Fig. 2B"). An organizer is a cellular domain that releases particular regulatory morphogens that act upon its neighboring tissues. Three main secondary AP organizers have been described for vertebrates, the anterior neural ridge (ANR), located in the most rostral part of the neural tube; the *zona limitans intrathalamica* (ZLI), located between the prethalamus and the thalamus, and the isthmic organizer (IsO), which develops in the boundary between the midbrain and the rhombencephalon. Each of these organizers releases a specific combination of morphogens that influences the development of the surrounding structures. In particular, the ANR expresses *Fgf8*, which is essential for the correct regionalization of the prosencephalon and necessary for the correct formation of the telencephalon, a dorsal outgrowth of the secondary prosencephalon (Kiecker and Lumsden, 2012; Vieira *et al.*, 2010). The reference morphogen in the ZLI is *Shh*, which is crucial for the differential specification of the thalamus (Crossley *et al.*, 1996; Chi *et al.*, 2003; Vieira *et al.*, 2005; Hirata, 2006; Vue *et al.*, 2009). Regarding the IsO, *Fgf8* acts again as morphogen, but in this case together with *Wnt1*. This latter region can be identified from very early stages of development by the apposition of the expression patterns of *Otx2* and *Gbx2*, which act antagonistically and are crucial to define the position and function of the IsO (Hidalgo-Sánchez *et al.*, 2005; Kiecker and Lumsden, 2012).

Interestingly, the organizers and their associated genetic networks have been conserved since the origin of vertebrates (Osorio *et al.*, 2005). However, the extent of evolutionary conservation is under debate when larger phylogenetic distances are considered, namely between vertebrates and other chordates or other deuterostomes (e.g. hemichordates). Several articles in recent years have discussed the presence/absence and homology/convergence of secondary organizers and their potential functionality as bona fide organizers (see Holland *et al.*, 1997; Kozmik *et al.*, 1999; Shimeld, 1999; Lowe *et al.*, 2003; Takahashi and Holland, 2004; Lowe *et al.*, 2006; Denes *et al.*, 2007; Hirth, 2010; Steinmetz *et al.*, 2011; Pani *et al.*, 2012; Arendt *et al.*, 2015; Yao *et al.*, 2016, for engaging discussions and controversies). As expected, one of the clades at the center of these controversies is the cephalochordate amphioxus.

The amphioxus CNS: challenging major subdivisions of the vertebrate brain

The amphioxus CNS is composed of a neural tube placed dorsally to the notochord (Fig. 1). Rostrally, the notochord continues in appearance beyond the extension of the neural tube, although the anterior-most tip of the notochord has been proposed to be a homolog of the vertebrate prechordal plate (Albuixech-Crespo *et al.*, 2017). A subtle widening in the anterior part of the neural tube, which corresponds to the cerebral vesicle, is barely noticeable in the adult, although this swelling is much more obvious at larval stages. Beyond this, there are neither anatomically recognizable neuromeres nor other morphological landmarks in the amphioxus neural tube to easily identify a segmental organization besides the visible dorsal nerves arranged in series (Fig. 1). However, examination at the histological and cellular level reveals visible

histoarchitectonic differences in the AP and DV axes in larvae as well as in adults, which reflect differences or subdivisions between regions along the neural tube (Wicht and Lacalli, 2005) (Figs. 1 and 2; for detailed descriptions and cyto- and histo-architecture of the amphioxus nervous system, see Lacalli (this issue, and references therein)).

These differential histoarchitectonic arrangements imply that there must be a subjacent molecular regionalization of the amphioxus CNS. In fact, many genes involved in neural development in vertebrates are also expressed in cephalochordate neural development. Literature that describes, compares and establishes correspondences between expression patterns of genes involved

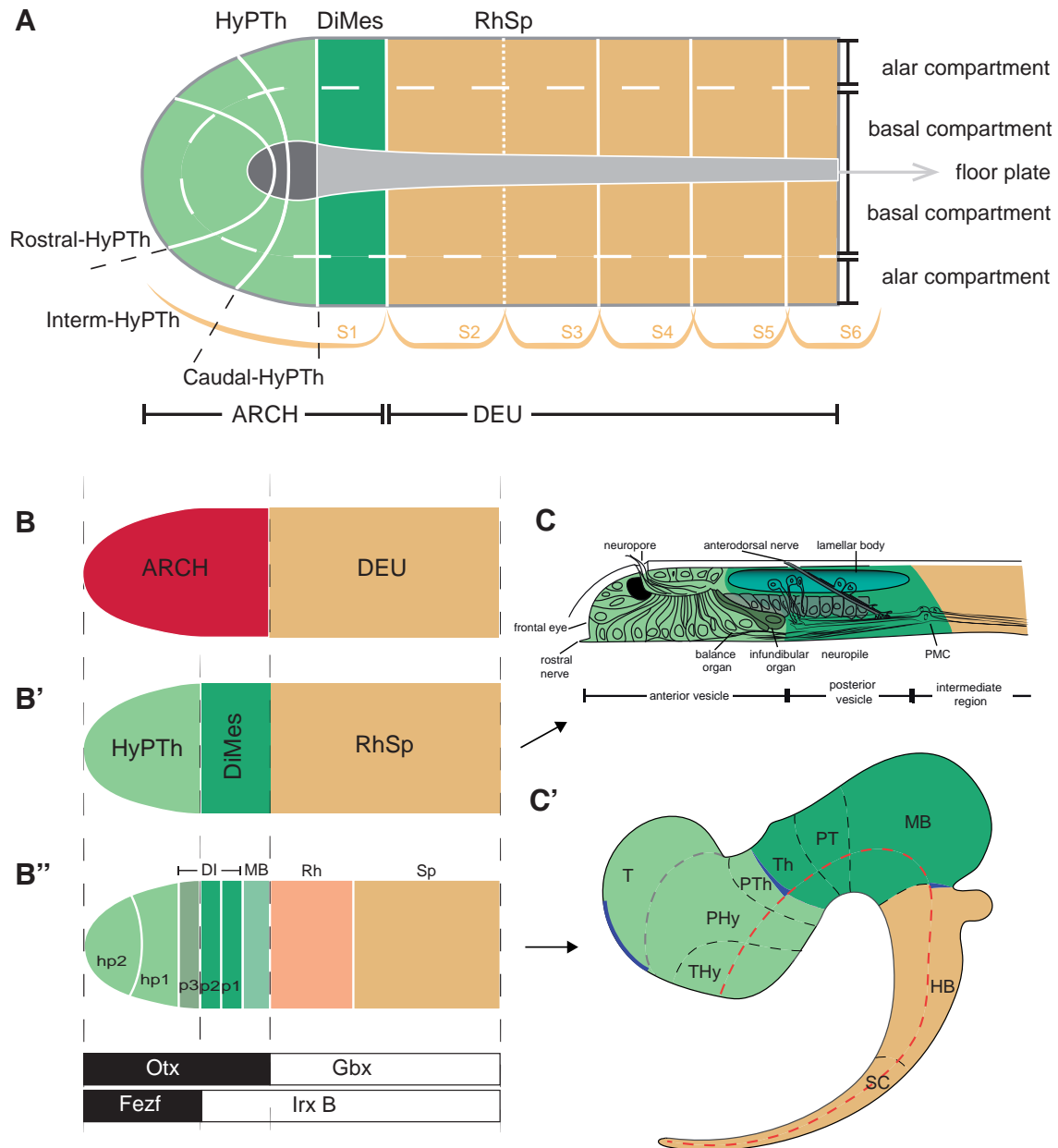


Fig. 3. Ontogenetic correspondences and proposed homologies between cephalochordates and vertebrates CNS. (A) Summary of all identified anteroposterior and dorsoventral partitions of the neural plate of amphioxus, dorsal-view scheme, anterior is to the left. Hypothalamo-prethalamic primordium (HyPTh), DiMesencephalic primordium (DiMes), Rhombencephalo-Spinal primordium (RhSp). S1-S6 indicates the relative position of the six most anterior somites. (B-B'') Topological comparison of major molecular subdivisions between an ancestor with a bipartite brain (B), cephalochordates (B') and vertebrates (B''). Archencephalic prototagma (ARCH), deuteroencephalic prototagma (DEU). The amphioxus DiMes seems to be homologous to the thalamus (p2), pretectecum (p1) and midbrain (MB), whereas the diencephalon (DI) is neither an ontogenetic nor an evolutionary unit, as the prethalamus (p3) is derived from a vertebrate-specific subdivision of the HyPTh. Gene expression domains of relevant genes (Otx, Gbx, Fezf and IrxB) that mark the three main subdivisions of the chordate brain are indicated below. (C-C') Correspondence of the three main subdivisions (colour coded) of the neural plate to the larval stages of amphioxus (C) and a representative neural tube of vertebrates (C'). Blue spikes indicate the secondary organizers IsO (left) and ZLI (right) in C'. Primary motor center (PMC), hindbrain (HB). Adapted from Albuixech-Crespo et al., (2017).

in vertebrate and amphioxus neural regionalization is abundant. Very recently, some of us with our collaborators (Albuixech-Crespo *et al.*, 2017) mapped the expression of 48 genes whose orthologs in vertebrates have a well-established morphological interpretation within the prosomeric model. These mappings were done at a single developmental stage of amphioxus (7-somite neurula stage), allowing homochronic comparisons of gene expression patterns. With these data, the authors proposed a genoarchitectonic model for amphioxus consistent with the prosomeric model (Fig. 3). As previously suggested (Castro *et al.*, 2006), this model proposes that the incipient amphioxus neural tube is molecularly divided into a rostral archencephalic (ARCH) domain and a caudal deuterocephalic (DEU) domain from very early stages, similarly to vertebrates. This division is highlighted by the abutting expression domains of *Otx* and *Gbx*. Furthermore, the amphioxus ARCH has two major subdivisions, which we termed Hypothalamo-Prethalamic primordium (HyPTh; further subdivided into Rostral-HyPTh, Interm-HyPTh and Caudal-HyPTh) and Di-Mesencephalic primordium (DiMes) (Fig. 3B), whose boundary is defined by the abutting expression of *Fez* on the most caudal HyPTh domain (Caudal-HyPTh) with that of the *lrx* genes in the DiMes. In contrast, and as mentioned above, the vertebrate ARCH is traditionally divided into three main regions, the secondary prosencehalon (hypothalamus plus telencephalon), the diencephalon (pretectum, thalamus and prethalamus) and the midbrain. Remarkably, considering its topological position relative to several markers as *Fez* and *lrx*, the Caudal-HyPTh may be homologous to the vertebrate prethalamus. In addition, the small amphioxus DiMes, consisting of two rows of cells along the AP axis at the mid-neurula stage, seems to be homologous to the thalamus, pretectum and midbrain all together. Therefore, the last two observations, together with multiple lines of experimental embryological evidence (Gardner and Barald, 1991; Martinez *et al.*, 1991; Bally-Cuif *et al.*, 1992; Bloch-Gallego *et al.*, 1996; Vieira *et al.*, 2005; Vue *et al.*, 2009; Hirata, 2006), profoundly challenge textbook subdivisions of the vertebrate brain: (i) the diencephalon loses its coherence as a single entity, since it has neither evolutionary nor developmental support, and (ii) the vertebrate thalamus,

pretectum and midbrain should be considered an evolutionary elaboration of an ancestral, perhaps amphioxus-like, DiMes region.

The hypothesis that the diencephalon is neither an evolutionary nor an ontogenetic primordial subdivision of the vertebrate brain was further supported by knock-out and knock-down experiments in mouse and zebrafish (Albuixech-Crespo *et al.*, 2017). Patterning of the DiMes-like territory in vertebrates occurs under the control of the secondary organizers ZLI and IsO, which, despite ongoing debate, seem to be absent in amphioxus (Shimeld, 1999; Pani *et al.*, 2012). Interfering with the vertebrate ZLI turns the thalamic region into a pretectum-like territory, ablating the IsO abolishes the midbrain and expands the pretectum territory, and a quadruple morpholino in zebrafish that alters both ZLI and IsO transforms those three DiMes regions of vertebrates (thalamus, pretectum and midbrain) into a more molecularly homogeneous structure that resembles the amphioxus DiMes region (Fig. 4). Therefore, these results suggest a close relationship between the gain or loss of secondary organizers during evolution and the origin or loss of specific partitions in the brain among the major chordate groups.

Amphioxus genoarchitectonic model provides novel insights into the ontogeny of adult structures and their homology with vertebrates

Our genoarchitectonic model of the developing amphioxus neural tube also sheds light into the ontogenetic origins of several larval and adult brain structures, and their possible evolutionary relationships with vertebrate derivatives. Ontogenetic assignments between adult, larval and embryonic structures and domains have many limitations in amphioxus. First, the lack of cell tracing data on the formation of neural derivatives hampers direct extrapolations of adult neural populations from early stages of development. Second, there is no information about the possible migration processes that the progenitor populations of the terminal derivatives of the adult neural tube experience. Third, there are no robust continuous reference landmarks. Two main references have been used in the literature to infer correspondences among

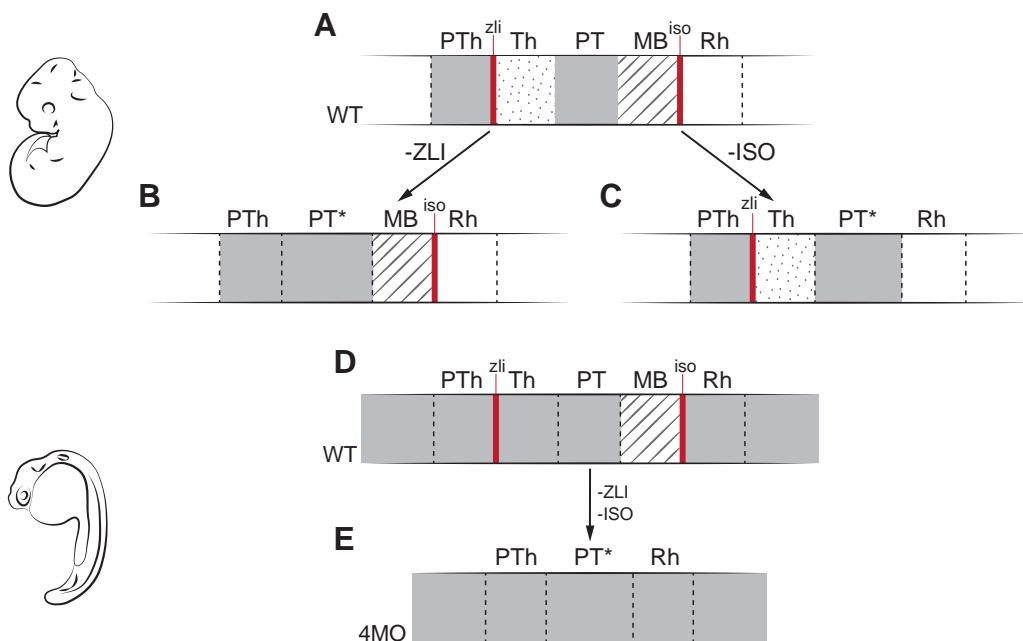


Fig. 4. Experimental disruption of vertebrate secondary organizers recapitulates the amphioxus brain. (A) Schematic representation of major neural compartments (PTh, Th, PT, MB, and Rh) around DiMes, delimited by the secondary organizers (ZLI and IsO: red lines) in wild type (WT) vertebrates; solid gray indicates Pax6 expression. (B) Abrogation of the ZLI turns the thalamic region into a pretectum-like territory (PT*). (C) Genetic depletion of the IsO abolishes the midbrain and expands the pretectum territory (PT*). (D) Wild type and (E) quadruple morpholino of *otx1a*, *otx2*, *eng2a*, and *eng2b* in zebrafish eliminates ZLI and IsO, and thus the whole DiMes loses its subdivisions and mimics the uniform amphioxus DiMes. Adapted from Albuixech-Crespo *et al.*, (2017).

neural structures. One is the position of these structures respect to the myomeres or somites in adults or embryos, respectively. Unfortunately, however, it has not yet been proven whether the position and the size of myomeres are fixed or permanent in relation to the neural tube, hampering ontogenetic assignments. The other one is innervation, which presents concordances with the myomeric asymmetry endorsed by the myoseptal position of the neural roots in the neural tube of the adult amphioxus. However, these are difficult to extrapolate to early embryos.

Given these caveats, extrapolations on the correspondence of the neural derivatives in the adult amphioxus with respect the progenitor territories in the neural plate stage during development must come from meticulous correlations that include information for both its topological position and the combination of gene expression patterns during neural ontogeny, always keeping in mind that gene expression patterns could be very dynamic and that they do not provide deep lineage information if considered independently. Keeping these limitations in mind, we revise here previous literature and propose some ontogenetic correspondences for important terminal derivatives of the anterior CNS of amphioxus, and discuss their implications to assign vertebrate homologies

Frontal eye and visual system

One of the most visible structures of the amphioxus neural tube is the frontal eye, which can be observed from early larval stages. The topological position of the frontal eye suggests that it arises from the most anterior part of the neural plate (Lacalli, 1996), the acroterminal domain (Fig. 3), as it occurs in vertebrates (McMahon and Bradley, 1990; Puelles *et al.*, 2012; Puelles and Rubenstein, 2015; Albuixech-Crespo *et al.*, 2017). The genoarchitecture of this area in the neural plate suggests its involvement in the development of this photoreceptive structure. Several genes expressed very early in amphioxus development are involved in eye structures in a wide number of bilaterians, such as *Rx*, *Six3/6* and *Lhx2/9* (Zuber *et al.*, 2003; Kamijyo *et al.*, 2015; Albuixech-Crespo *et al.*, 2017), even though the development of the amphioxus frontal eye becomes apparent only later in development by the emergence of the cells with dendrites and cilia that protrude from the neuropore and by the presence of the pigment cup. Remarkably, at larval stages, when the frontal eye is evident, the resulting cellular types that form it maintain a characteristic genoarchitecture. Both in vertebrates and cephalochordates *Gi* and *c-Opsin* are expressed in photoreceptors, whereas *Mitf* and *Pax2/5/8* appear in pigment cells (Vopalensky *et al.*, 2012). In vertebrates, two eye regions develop as a result of the splitting of the morphogenetic eye field by the action of signals coming from the prechordal mesoderm (Zuber *et al.*, 2003), while in amphioxus it stays as a single medial structure.

The amphioxus frontal eye innervates the neuropile through serotonergic neurons (Vopalensky *et al.*, 2012). This neuropile develops immediately caudal to the infundibular organ, which marks the transition between the anterior vesicle and the posterior region of the cerebral vesicle. It projects to the primary motor center (PMC), located more caudally, in the region immediately posterior to the limit between the 1st and the 2nd myomeres (Fig. 3C). Part of this region expresses *Pax4/6* in larval stages (Suzuki *et al.*, 2014) and, topologically, it seems to correspond to the *Pax4/6*-positive DiMes region described in our genoarchitectonic model (Albuixech-Crespo *et al.*, 2017). Interestingly, the frontal eye of amphioxus as well as the paired eyes of the lamprey larva project to a photoreceptor

visual center that is *Otx*- and *Pax4/6*-positive, located in the most posterior region of the cerebral vesicle in amphioxus larvae, and in the caudal prosencephalon in lampreys (Suzuki *et al.*, 2014) (Fig. 2A). In the latter study, Suzuki and colleagues suggested that the mesencephalic optic projections (which are retino-tectal, *Pax6*-negative and *Otx2*-positive) are secondarily developed, specific to vertebrates, and implicated in vision image formation. In consequence, the retino-pretecal primary projections, which appear in earlier stages of lamprey larvae, would be more similar to the ancestral state in vertebrates projecting to a visual *Pax6*-positive center, as it occurs in amphioxus. Lacalli (1996) identified the visual center of amphioxus tracking the projections from the frontal eye through electronic microscopy, and named it *tectum*, as a reference of homology to the midbrain in vertebrates. However, other authors concluded that cephalochordates lack a proper homolog to the tectal region, based on the absence of expression of *Dmbx* in the amphioxus neural tube (Takahashi and Holland, 2004), which is expressed in the midbrain of vertebrates. Our genoarchitectonic model of the neural plate of amphioxus described above sheds light into the evolutionary common origin for the midbrain, thalamus and pretectum of vertebrates from a DiMes-like territory that was likely ancestrally involved in visual processing. Based on these data, it is tempting to speculate that a DiMes/pretectum-like function was ancestral, and that the origin of the midbrain as an independent unit could have been linked to the origin of the vision with image formation present in vertebrates, but not in amphioxus.

Balance organ and circadian clock

The topology and morphology of cells associated with the circadian clock and the balance organ in the adult amphioxus, which are embedded in the ventral commissure and associated to the lamellar body (a photoreceptor organ) (Figs. 1D and 3C), as well as immunoreactive to GABA, led some authors to propose its homology to the mammalian suprachiasmatic nucleus (Anadón *et al.*, 1998; Castro and Becerra, 2015). The suprachiasmatic nucleus in vertebrates is a derivative of the alar part of the most rostral hypothalamic territory (Puelles and Rubenstein, 2015), which is *Nkx2.2*-positive. Also, in vertebrates, *Lhx1* (also known as *Lim1*) is involved in the establishment and maintenance of this circadian rhythm in the suprachiasmatic nucleus (Hatori *et al.*, 2014). In amphioxus, the genoarchitecture of the most rostral part of the neural plate is coincident, being *Nkx2.2*- and *Lhx1/5*-positive, further supporting the similarity between the two structures. However, the fact that vertebrate suprachiasmatic nucleus arises from the alar part of the hypothalamus and the amphioxus balance organ is ventral in origin casts doubts on the homology between the two derivatives.

Infundibular organ

As mentioned above, the infundibular organ is located in the most anterior part of the floor plate of the neural tube (Fig. 3C) and secretes Reissner's fiber to the central canal of the neural tube (Olsson *et al.*, 1994; López-Avalos *et al.*, 1997). Its functions have not been characterized in amphioxus, although it seems to be involved in axon guidance and establishment of commissures (Lehmann and Naumann, 2005). Both the amphioxus infundibular organ and the vertebrate subcommissural and flexural organs (which produce Reissner's fiber in vertebrates) are formed near developing commissures (López-Avalos *et al.*, 1997). The subcommissural organ develops in the pretectum (p1) and is the

primary and permanent producer of Reissner's fiber in vertebrates. Based on these observations, some authors have proposed its correspondence with the amphioxus infundibular organ (Olsson, 1955; Lichtenfeld *et al.*, 1999). In contrast, the flexural organ is located in the most rostral floor plate of vertebrates. Therefore, given the topological position of the amphioxus infundibular organ, a probable derivative of the HyPTh, it seems more likely that, if any, the vertebrate flexural organ is a homolog of the amphioxus infundibular organ, even though the secreting action of the flexural organ is transitory (Lichtenfeld *et al.*, 1999). However, it is also possible these chordate organs carry on analogous functions but are not homologous structures.

Lamellar body

The lamellar body (Fig. 3C) has often been proposed as a homolog of the pineal gland of vertebrates (Lacalli, 1996). Both develop from a *Pax4/6*-positive territory and are (at least ancestrally) photoreceptive structures. However, the amphioxus lamellar body seems to develop from the DiMes area, which does not include in its genoarchitecture the marker *Rx*, essential for the development of the pineal gland in vertebrates (Ruiz and Anadón, 1991; Vopalensky *et al.*, 2012; Rath *et al.*, 2013). Moreover, the lamellar body is not immunoreactive to melatonin and does not produce the enzymes necessary for its synthesis (Vernadakis *et al.*, 1998; Falcón *et al.*, 2014). In addition, all the genes related to circadian rhythms in amphioxus are expressed in the balance organ and not in the lamellar body (Schomerus *et al.*, 2008). Therefore, despite its similar topology, the proposed homology between the lamellar body and the pineal gland of vertebrates is not fully conclusive.

Concluding remarks

Although morphologically much simpler than vertebrates, a detailed genoarchitecture analysis of the amphioxus CNS has revealed a high level of complexity. Furthermore, it has shown many similarities to vertebrates, providing support to a common bauplan being already present in the last common ancestor of chordates. Many major AP and DV subdivisions have direct counterparts between amphioxus and vertebrates, even though they are often further subdivided in vertebrates. An example of this is the DiMes homologous region, which does not seem to be molecularly regionalized in the developing amphioxus neural tube, but is subdivided into thalamus, pretectum and midbrain in vertebrates. These comparative analyses also help to reveal cryptic ontogenetic correspondences and homologies between cephalochordate and vertebrate CNS terminal derivatives, thus shedding light into the origin and evolution of the complex vertebrate CNS and the specialized features of the amphioxus CNS. One surprising conclusion of these analyses is that the originally simple and uniform DiMes of amphioxus seems to give rise to a relatively large number of functional adult and larval structures (Fig. 3C), indicating a hidden and unexplored complexity-generation potential for this region during late amphioxus CNS development.

Acknowledgements

Authors' research is funded by: the Spanish Ministry of Economy and Competitiveness; European FEDER funds (grants number BFU2014-57516-P to JGF, SAF2016-80937-R to GM and BFU2014-55076-P to MI); the European Union Horizon 2020 research and innovation program (grant agreement No ERC-StG-LS2-637591 to MI); and by the 'Centro

de Excelencia Severo Ochoa 2013-2017', SEV-2012-0208 to CRG. We would also like to acknowledge the support of the CERCA Programme / Generalitat de Catalunya. CH-B holds an FPI fellowship. The authors wish to thank José Luis Ferrán and Ignacio Maeso for comments and fruitful discussions.

References

- ABOUHEIF E, AKAM M, DICKINSON W, HOLLAND PWH, MEYER A, PATEL N, RAFF R, LOUISE R, WRAY G (1997). Homology and developmental genes. *Trends Genet* 13: 432–433.
- ACAMPORA D, AVANTAGGIATO V, TUORTO F, SIMEONE A (1997). Genetic control of brain morphogenesis through *Otx* gene dosage requirement. *Development* 124: 3639–3650.
- ACEMEL R, TENA J, IRASTORZA-AZCARATE I, MARLETAZ F, GOMEZ-MARIN C, DE LA CALLE-MUSTIENES E, BERTRAND S, DIAZ S, ALDEA D, AURY JM, MANGENOT S, HOLLAND PWH, DEVOS D, MAESO I, ESCRIVA H, GOMEZ-SKARMETA JL (2016). A single three-dimensional chromatin compartment in amphioxus indicates a stepwise evolution of vertebrate Hox bimodal regulation. *Nat Genet* 48: 336–341.
- ALBUIXECH-CRESPO B, LÓPEZ-BLANCH L, BURGUERAD, MAESO I, SÁNCHEZ-ARRONES L, MORENO-BRAVO JA, SOMORJAI I, PASCUAL-ANAYA J, PUELLES E, BOVOLENTA P, GARCIA-FERNÁNDEZ J, PUELLES L, IRIMIA M, FERRAN JL (2017). Molecular regionalization of the developing amphioxus neural tube challenges major partitions of the vertebrate brain. *PLoS Biol* 15: e2001573.
- ANADÓN R, ADRIO F, RODRÍGUEZ-MOLDES I (1998). Distribution of GABA immunoreactivity in the central and peripheral nervous system of amphioxus (branchiostoma lanceolatum pallas). *J Comp Neurol* 401: 293–307.
- ARENDD D, TOSCHES MA, MARLOW H (2015). From nerve net to nerve ring, nerve cord and brain - evolution of the nervous system. *Nat Rev Neurosci* 17: 61–72.
- BALLY-CUIF L, ALVARADO-MALLART RM, DARNELL DK, WASSEF M (1992). Relationship between *Wnt-1* and *En-2* expression domains during early development of normal and ectopic met-mesencephalon. *Development* 115:999–1009.
- BLAIR JE, HEDGES SB (2005). Molecular phylogeny and divergence times of deuterostome animals. *Mol Biol Evol* 22: 2275–2284.
- BLOCH-GALLEGO E, MILLET S, ALVARADO-MALLART RM (1996). Further observations on the susceptibility of diencephalic prosomeres to *En-2* induction and on the resulting histogenetic capabilities. *Mech Dev* 58:51–63.
- BONE Q (1960). The central nervous system in amphioxus. *J Comp Neurol* 115: 27–64.
- CASTRO A, BECERRA M (2015). Neuronal Organization of the Brain in the Adult Amphioxus (*Branchiostoma lanceolatum*): A Study With Acetylated Tubulin Immunohistochemistry. *J Comp Neurol* 523: 2211–2232.
- CASTRO LFC, RASMUSSEN SLK, HOLLAND PWH, HOLLAND ND, HOLLAND LZ (2006). A *Gbx* homeobox gene in amphioxus: Insights into ancestry of the ANTP class and evolution of the midbrain/hindbrain boundary. *Dev Biol* 295: 40–51.
- CHI CL, MARTINEZ S, WURST W, MARTIN GR (2003). The isthmus organizer signal *FGF8* is required for cell survival in the prospective midbrain and cerebellum. *Development* 130: 2633–2644.
- COSTA O (1834). Cenni zoologici ossia descrizione sommaria delle specie nuove di animali scoperti in diverse contrade del regno nell' anno 1834. *Annuario Zoologico* 1834.
- CROSSLEY PH, MARTINEZ S, MARTIN GR (1996). Midbrain development induced by *FGF8* in the chick embryo. *Nature* 380: 66–68.
- DEHAL P, BOORE JL (2005). Two Rounds of Whole Genome Duplication in the Ancestral Vertebrate. *PLoS Biol* 3: e314.
- DELSUC F, BRINKMANN H, CHOURROUT D, PHILIPPE H (2006). Tunicates and not cephalochordates are the closest living relatives of vertebrates. *Nature* 439: 965–968.
- DELSUC F, TSAGKOGEOGA G, LARTILLOT N, PHILIPPE H (2008). Additional molecular support for the new chordate phylogeny. *Genesis* 46: 592–604.
- DENES AS, JÉKELY G, STEINMETZ PRH, RAIBLE F, SNYMAN H, PRUD'HOMME B, FERRIER DEK, BALAVOINE G, ARENDT D, JEKELY G, STEINMETZ PRH, RAIBLE F, SNYMAN H, PRUD'HOMME B, FERRIER DEK, BALAVOINE G, ARENDT D (2007). Molecular Architecture of Annelid Nerve Cord Supports Common Origin of Nervous System Centralization in Bilateria. *Cell* 129: 277–288.

- FALCÓN J, COON SL, BESSEAU L, CAZAMÉA-CATALAN D, FUENTÈS M, MAGNANOU E, PAULIN C-H, BOEUF G, SAUZET S, JØRGENSEN EH, MAZAN S, WOLF YI, KOONIN E V, STEINBACH PJ, HYODO S, KLEIN DC (2014). Drastic neofunctionalization associated with evolution of the timezyme AANAT 500 Mya. *Proc Natl Acad Sci USA* 111: 314–319.
- FERRAN JL, DE OLIVEIRA ED, MERCHÁN P, SANDOVAL JE, SÁNCHEZ-ARRONES L, MARTÍNEZ-DE-LA-TORRE M, PUELLES L (2009). Genoarchitectonic profile of developing nuclear groups in the chicken pretectum. *J Comp Neurol* 517: 405–451.
- FERRAN JL, SÁNCHEZ-ARRONES L, BARDET SM, SANDOVAL JE, MARTÍNEZ-DE-LA-TORRE M, PUELLES L (2008). Early pretectal gene expression pattern shows a conserved anteroposterior tripartition in mouse and chicken. *Brain Res Bull* 75: 295–298.
- FERRAN JL, SANCHEZ-ARRONES L, SANDOVAL JE, PUELLES L (2007). A model of early molecular regionalization in the chicken embryonic pretectum. *J Comp Neurol* 505: 379–403.
- FORCE A, LYNCH M, PICKETT FB, AMORES A, YAN Y, POSTLETHWAIT J (1999). *Genetics* 151: 1531–1545.
- GARCIA-FERNANDEZ J, BENITO-GUTIERREZ E (2009). It's a long way from amphioxus: descendants of the earliest chordate. *Bioessays* 31: 665–675.
- GARCIA-FERNANDEZ J, HOLLAND PWH (1994). Archetypal organization of the amphioxus Hox gene cluster. *Nature* 370: 563–566.
- GARDA AL, ECHEVARRÍA D, MARTÍNEZ S (2001). Neuroepithelial co-expression of Gbx2 and Otx2 precedes Fgf8 expression in the isthmus organizer. *Mech Dev* 101: 111–118.
- GARDNER CA, BARALD KF (1991). The cellular environment controls the expression of engrailed-like protein in the cranial neuroepithelium of quail-chick chimeric embryos. *Development* 113: 1037–1048.
- HATORI M, GILL S, MURE LS, GOULDING M, O'LEARY DD, PANDA S (2014). Lhx1 maintains synchrony among circadian oscillator neurons of the SCN. *eLife* 3: e03357.
- HIDALGO-SÁNCHEZ M, MILLET S, BLOCH-GALLEGO E, ALVARADO-MALLART RM (2005). Specification of the meso-isthmo-cerebellar region: The Otx2/Gbx2 boundary. *Brain Res Rev* 49: 134–149.
- HIRATA T (2006). Zinc-finger genes Fez and Fez-like function in the establishment of diencephalon subdivisions. *Development* 133: 3993–4004.
- HIRTH F (2010). On the origin and evolution of the tripartite brain. *Brain Behav Evol* 76: 3–10.
- HOLLAND LZ, HOLLAND ND (2007). A revised fate map for amphioxus and the evolution of axial patterning in chordates. *Integr Comp Biol* 47: 360–372.
- HOLLAND LZ, KENE M, WILLIAMS N a, HOLLAND ND (1997). Sequence and embryonic expression of the amphioxus engrailed gene (AmphiEn): the metameric pattern of transcription resembles that of its segment-polarity homolog in *Drosophila*. *Development* 124: 1723–1732.
- HOLLAND ND, PANGANIBAN G, HENYEE EL, HOLLAND LZ (1996). Sequence and developmental expression of AmphiDII, an amphioxus Distal-less gene transcribed in the ectoderm, epidermis and nervous system: insights into evolution of cranial forebrain and neural crest. *Development* 122: 2911–2920.
- HOLLAND PWH, HOLLAND LZ, WILLIAMS NA, HOLLAND ND (1992). An amphioxus homeobox gene: sequence conservation, spatial expression during development and insights into vertebrate evolution. *Development* 116: 653–661.
- HOLLAND PW, GARCIA-FERNÁNDEZ J, WILLIAMS N, SIDOW A (1994). Gene duplications and the origins of vertebrate development. *Dev Suppl* 1994: 125–133.
- IRIMIA M, PIÑEIRO C, MAESO I, GÓMEZ-SKARMETA JL, CASARES F, GARCIA-FERNÁNDEZ J, PIÑEIRO C, MAESO I, GÓMEZ-SKARMETA JL, CASARES F, GARCIA-FERNANDEZ J, PIÑEIRO C, MAESO I, GÓMEZ-SKARMETA JL, CASARES F, GARCIA-FERNÁNDEZ J (2010). Conserved developmental expression of Fezf1 in chordates and *Drosophila* and the origin of the Zona Limitans Intrathalamica (ZLI) brain organizer. *Evodevo* 1: 7.
- JIMENEZ-DELGADOS, PASCUAL-ANAYA J, GARCIA-FERNÁNDEZ J (2009). Implications of duplicated cis-regulatory elements in the evolution of metazoans: the DDI model or how simplicity begets novelty. *Brief Funct Genomic Proteomic* 8: 266–275.
- KAMIJYO A, YURA K, OGURA A (2015). Distinct evolutionary rate in the eye field transcription factors found by estimation of ancestral protein structure. *Gene* 555: 73–79.
- KIECKER C, LUMSDEN A (2012). The Role of Organizers in Patterning the Nervous System. *Annu Rev Neurosci* 35: 347–367.
- KOZMIK Z, HOLLAND ND, KALOUSOVA A, PACES J, SCHUBERT M, HOLLAND LZ (1999). Characterization of an amphioxus paired box gene, AmphiPax2/5/8: developmental expression patterns in optic support cells, nephridium, thyroid-like structures and pharyngeal gill slits, but not in the midbrain-hindbrain boundary region. *Development* 126: 1295–1304.
- KUHLENBECK H (1967). The Central Nervous System of Vertebrates. *Yale J Biol Med* 40: 170.
- LACALLI TC (1996). Frontal Eye Circuitry, Rostral Sensory Pathways and Brain Organization in Amphioxus Larvae: Evidence from 3D Reconstructions. *Philos Trans R Soc B Biol Sci* 351: 243–263.
- LACALLI TC (2002). Sensory pathways in amphioxus larvae I. Constituent fibres of the rostral and anterodorsal nerves, their targets and evolutionary significance. *Acta Zool* 83: 149–166.
- LACALLI TC, HOLLAND ND, WEST JE (1994). Landmarks in the anterior central nervous system of amphioxus larvae. *Phil Trans R Soc Lond* 344: 165–185.
- LEHMANN C, NAUMANN WW (2005). Axon pathfinding and the floor plate factor Reissner's substance in wildtype, cyclops and one-eyed pinhead mutants of *Danio rerio*. *Dev Brain Res* 154: 1–14.
- LICHTENFELD J, VIEHWEG J, SCHÜTZENMEISTER J, NAUMANN WW (1999). Reissner's substance expressed as a transient pattern in vertebrate floor plate. *Anat Embryol (Berl)* 200: 161–174.
- LÓPEZ-AVALOS MD, CIFUENTES M, GRONDONA JM, MIRANDA E, PÉREZ J, FERNÁNDEZ-LLEBREZ P (1997). Rostral floor plate (flexural organ) secretes glycoproteins immunologically similar to subcommissural organ glycoproteins in dogfish (*Scyliorhinus canicula*) embryos. *Brain Res Dev Brain Res* 102: 69–75.
- LOWE CJ, TERASAKI M, WU M, FREEMAN JR. RM, RUNFT L, KWAN K, HAIGO S, ARONOWICZ J, LANDER E, GRUBER C, et al., (2006). Dorsoventral patterning in hemichordates: Insights into early chordate evolution. *PLoS Biol* 4: 1603–1619.
- LOWE CJ, WU M, SALIC A, EVANS L, LANDER E, STANGE-THOMANN N, GRUBER CE, GERHART J, KIRSCHNER M (2003). Anteroposterior patterning in hemichordates and the origins of the chordate nervous system. *Cell* 113: 853–865.
- MARTINEZ S, WASSEF M, ALVARADO-MALLART RM (1991). Induction of a mesencephalic phenotype in the 2-day-old chick prosencephalon is preceded by the early expression of the homeobox gene en. *Neuron* 6: 971–981.
- MCMAHON AP, BRADLEY A (1990). The Wnt-1 (int-1) proto-oncogene is required for development of a large region of the mouse brain. *Cell* 62: 1073–1085.
- MEDINA L, BUPESH M, ABELLÁN A (2011). Contribution of genoarchitecture to understanding forebrain evolution and development, with particular emphasis on the amygdala. *Brain Behav Evol* 78: 216–236.
- MERCHÁN P, BARDET SM, PUELLES L, FERRAN JL (2011). Comparison of Pretectal Genoarchitectonic Pattern between Quail and Chicken Embryos. *Front Neuroanat* 5: 23.
- MORONA R, FERRAN JL, PUELLES L, GONZALEZ A (2017). Gene expression analysis of developing cell groups in the pretectal region of *Xenopus laevis*. *J Comp Neurol* 519: 1–38.
- MORONA R, FERRAN JL, PUELLES L, GONZÁLEZ A (2011). Embryonic genoarchitecture of the pretectum in *Xenopus laevis*: A conserved pattern in tetrapods. *J Comp Neurol* 519: 1024–1050.
- OHNO S (1970). Evolution by gene duplication. Springer-Verlag, New York.
- OLSSON R (1955). Structure and Development of Reissner's Fibre in the caudal end of amphioxus and some lower vertebrates. *Acta Zool* 36: 167–198.
- OLSSON R, YULIS R, RODRIGUEZ EM (1994). Cell & Tissue The infundibular organ of the lancelet (*Branchiostoma lanceolatum*, Acrania): an immunocytochemical study. *Cell Tissue Res* 277: 107–114.
- OSORIO J, MAZAN S, RETAUX S (2005). Organisation of the lamprey (*Lampetra fluviatilis*) embryonic brain: insights from LIM-homeodomain, Pax and hedgehog genes. *Dev Biol* 288: 100–112.
- OWEN R (1848). On the archetype and homologies of the vertebrate skeleton. (Ed. John van Voorst).
- PALLAS P (1774). Spicilegia zoologica: quibus novae imprimis et obscurae animalium species iconibus, descriptionibus atque commentariis illustrantur. *Berolini*. Fasciculus 10: 1-411.
- PANI AM, MULLARKEY EE, ARONOWICZ J, ASSIMACOPOULOS S, GROVE EA, LOWE CJ (2012). Ancient deuterostome origins of vertebrate brain signalling centres. *Nature* 483: 289–294.

- PUELLES L (1995). A segmental morphological paradigm for understanding vertebrate forebrains. *Brain Behav Evol* 46: 319–337.
- PUELLES L (2010). Forebrain Development: Prosomere Model. *Encycl Neurosci*: 315–319.
- PUELLES L, FERRAN JL (2012). Concept of neural genoarchitecture and its genomic fundament. *Front Neuroanat* 6: 47.
- PUELLES L, MARTINEZ-DE-LA-TORRE M, BARDET SM, RUBENSTEIN JLR (2012). The mouse nervous system. In *Hypothalamus* (eds Watson C, Paxinos G, Puelles L). Academic Press/Elsevier, London, San Diego, pp. 221–312.
- PUELLES L, RUBENSTEIN JL (1993). Expression patterns of homeobox and other putative regulatory genes in the embryonic mouse forebrain suggest a neuromeric organization. *Trends Neurosci* 16: 472–479.
- PUELLES L, RUBENSTEIN JLR (2015). A new scenario of hypothalamic organization: rationale of new hypotheses introduced in the updated prosomeric model. *Front Neuroanat* 9: 27.
- POTNAM N, BUTTS T, FERRIER D, FURLONG R, HELLSTEN U, KAWASHIMA T, ROBINSON-RECHAVI M, SHOGUCHI E, TERRY A, YU JR-KAI, BENITO-GUTIERREZ E, DUBCHAK I, GARCIA-FERNANDEZ J, GIBSON-BROWN J, GRIGORIEV I, HORTON A, DE JONG P, JURKA J, KAPITONOV V, KOHARA Y, KUROKI Y, LINDQUIST E, LUCAS S, OSOEGAWA K, PENNACCHIO L, SALAMOV A, SATOU Y, SAUKA-SPENGLER T, SCHMUTZ J, SHIN-I T, TOYODA A, BRONNER-FRASER M, FUJIYAMA A, HOLLAND LZ, HOLLAND PWH, SATOH N, ROKHSAR D (2008). The amphioxus genome and the evolution of the chordate karyotype. *Nature* 453: 1064–1071.
- RATH MF, ROHDE K, KLEIN DC, MØLLER M (2013). Homeobox genes in the rodent pineal gland: Roles in development and phenotype maintenance. *Neurochem Res* 38: 1100–1112.
- RODRÍGUEZ-SEGUEL E, ALARCÓN P, GÓMEZ-SKARMETA JL (2009). The *Xenopus* *lrx* genes are essential for neural patterning and define the border between prethalamus and thalamus through mutual antagonism with the anterior repressors *Fezf* and *Arx*. *Dev Biol* 329: 258–268.
- RUIZ S, ANADÓN R (1991). The fine structure of lamellate cells on the brain of amphioxus (*Branchiostoma lanceolatum*, Cephalochordata). *Cell Tissue Res*: 597–600.
- SÁNCHEZ-ARRONES L, FERRÁN JL, RODRÍGUEZ-GALLARDO L, PUELLES L (2009). Incipient forebrain boundaries traced by differential gene expression and fate mapping in the chick neural plate. *Dev Biol* 335: 43–65.
- SCHOMERUS C, KORF H-W, LAEDTKE E, MORET F, ZHANG Q, WICHT H (2008). Nocturnal behavior and rhythmic period gene expression in a lancelet, *Branchiostoma lanceolatum*. *J Biol Rhythms* 23: 170–181.
- SHIMELD SM (1999). The evolution of the hedgehog gene family in chordates: insights from amphioxus hedgehog. *Dev Genes Evol* 209: 40–47.
- STEINMETZ PRH, KOSTYUCHENKO RP, FISCHER A, ARENDT D (2011). The segmental pattern of *otx*, *gbx*, and *Hox* genes in the annelid *Platynereis dumerilii*. *Evol Dev* 13: 72–79.
- SUZUKI DG, MURAKAMI Y, ESCRIVAH, WADAH (2014). A comparative examination of neural circuit and brain patterning between the lamprey and amphioxus reveals the evolutionary origin of the vertebrate visual center. *J Comp Neurol* 0: 1–11.
- TAKAHASHI T, HOLLAND PWH (2004). Amphioxus and ascidian *Dmbx* homeobox genes give clues to the vertebrate origins of midbrain development. *Development* 131: 3285–3294.
- TSCHOPP P, TABIN C. (2017). Deep homology in the age of next-generation sequencing. *Philos Trans R Soc Lond B Biol Sci* 372: 20150475.
- VERNADAKIS A J, BEMIS WE, BITTMAN EL (1998). Localization and partial characterization of melatonin receptors in amphioxus, hagfish, lamprey, and skate. *Gen Comp Endocrinol* 110: 67–78.
- VIEIRA C, GARDAAL, SHIMAMURAK, MARTINEZ S (2005). Thalamic development induced by *Shh* in the chick embryo. *Dev Biol* 284: 351–363.
- VIEIRAC, POMBEROA, GARCÍA-LOPEZ R, GIMENOL, ECHEVARRIAD, MARTÍNEZ S (2010). Molecular mechanisms controlling brain development: An overview of neuroepithelial secondary organizers. *Int J Dev Biol* 54: 7–20.
- VOPALENSKY P, PERGNER J, LIEGERTOVAM, BENITO-GUTIERREZ E, ARENDT D, KOZMIK Z (2012). Molecular analysis of the amphioxus frontal eye unravels the evolutionary origin of the retina and pigment cells of the vertebrate eye. *Proc Natl Acad Sci USA* 109: 15383–15388.
- VUE TY, BLUSKE K, ALISHAHI A, YANG LL, KOYANO-NAKAGAWA N, NOVITCH B, et al., (2009). Sonic hedgehog signaling controls thalamic progenitor identity and nuclei specification in mice. *J Neurosci* 29: 4484–4497.
- WICHT H, LACALLI TC (2005). The nervous system of amphioxus: structure, development, and evolutionary significance. *Can J Zool* 83: 122–150.
- WICHT H, LAEDTKE E, KORF HW, SCHOMERUS C (2010). Spatial and temporal expression patterns of *Bmal* delineate a circadian clock in the nervous system of *Branchiostoma lanceolatum*. *J Comp Neurol* 518: 1837–1846.
- WRAY G, ABOUHEIF E (1998). When is homology not homology? *Curr Opin Genet Dev* 8: 675–680.
- YAO Y, MINOR PJ, ZHAO Y-T, JEONG Y, PANI AM, KING AN, SYMMONS O, GAN L, CARDOSO W V, SPITZ F, LOWE CJ, EPSTEIN DJ (2016). Cis-regulatory architecture of a brain signaling center predates the origin of chordates. *Nat Genet*. 48: 575–580.
- ZUBER ME, GESTRI G, VICZIAN AS, BARSACCHI G, HARRIS W a (2003). Specification of the vertebrate eye by a network of eye field transcription factors. *Development* 130: 5155–5167.

