Lampreys, the jawless vertebrates, contain only two ParaHox gene clusters

Huixian Zhang1,2,*, Vydiyanathan Raviv,1 Boon-Hui Tay,1 Sumanty Tohari,1 Nisha E. Pillai1, Aravind Prasad1, Qiang Lin1, Sydney Brenner1,2,*, and Byrrapa Venkatnesab,c,2

1Institute of Molecular and Cell Biology, Agency for Science, Technology and Research, Biopolis, Singapore 138673, Singapore; 2Chinese Academy of Sciences (CAS) Key Laboratory of Tropical Marine Bioresources and Ecology, South China Sea Institute of Oceanology, Chinese Academy of Sciences, Guangzhou 510301, China; and 3Department of Paediatrics, Yong Loo Lin School of Medicine, National University of Singapore, Singapore 119228, Singapore

Contributed by Sydney Brenner, July 6, 2017 (sent for review March 20, 2017; reviewed by José Luis Gómez-Skarmeta and Nipam H. Patel)

ParaHox genes (Gsx, Pdx, and Cdx) are an ancient family of developmentally closely related to the Hox genes. They play critical roles in the patterning of brain and gut. The basal chordate, amphioxus, contains a single ParaHox cluster comprising one member of each family, whereas nonteleost jawed vertebrates contain four ParaHox genomic loci with six or seven ParaHox genes. Teleosts, which have experienced an additional whole-genome duplication, contain six ParaHox genomic loci with six ParaHox genes. Jawless vertebrates, represented by lampreys and hagfish, are the most ancient group of vertebrates and are crucial for understanding the origin and evolution of vertebrate gene families. We have previously shown that lampreys contain six Hox gene loci. Here we report that lampreys contain only two ParaHox gene clusters (designated as α and β-clusters) bearing five ParaHox genes (Gsxα, Pdxα, Cdxα, Gsxβ, and Cdxβ). The order and orientation of the three genes in the α-cluster are identical to that of the single cluster in amphioxus. However, the orientation of Gsxβ in the β-cluster is inverted. Interestingly, Gsxα is expressed in the eye, unlike its homologs in jawed vertebrates, which are expressed mainly in the brain. The lamprey Pdxα is expressed in the pancreas similar to jawed vertebrate Pdx genes, indicating that the pancreatic expression of Pdx was acquired before the divergence of jawless and jawed vertebrate lineages. It is likely that the lamprey Pdxα plays a crucial role in pancreas specification and insulin production similar to the Pdx of jawed vertebrates.

Significance

Lampreys and hagfishes are the only living members of jawless vertebrates, the most ancient lineage of vertebrates, and are therefore a crucial group for understanding the evolution of vertebrates. ParaHox genes (Gsx, Pdx, and Cdx) are an important family of developmental genes that play critical roles in the patterning of brain, pancreas, and posterior gut of jawed vertebrates. Here we show that lampreys contain two ParaHox gene clusters compared with four ParaHox loci in most jawed vertebrates. The lamprey Gsx/β gene is expressed specifically in the eye, an unusual expression domain for Gsx genes. The pancreatic expression of the lamprey Pdx gene suggests the crucial role of Pdx in pancreas specification and insulin production evolved in the common ancestor of vertebrates.
and a Cdx gene in the pancreas. We identified five ParaHox genes organized in two clusters in the Japanese lamprey genome. Because we could not determine the orthology of these genes to ParaHox genes in gnathostomes (see following sections), we designated the Gsx, Pdx, and Cdx genes on scaffold_1054 as Gsxα, Pdxα, and Cdxα (α-cluster), and the Gsx and Cdx genes on scaffold_14 as Gsxβ and Cdxβ (β-cluster) (Fig. 2). In addition to genome assembly searches, we carried out degenerate PCR, using genomic DNA as a template to determine whether there were any additional ParaHox genes in Japanese lamprey (Materials and Methods). The degenerate PCRs were able to amplify fragments of Gsxα, Pdxα, Cdxα, and Cdxβ genes, but not Gsxβ, because of the divergence of its sequences corresponding to the forward primers and the presence of a lamprey-specific intron (2.9 kb) in the region targeted by PCR primers (Materials and Methods). Nevertheless, no new ParaHox genes were identified in the PCR products.

ParaHox Genes in the Sea Lamprey Genome. Searches for ParaHox genes in the genome assembly of the sea lamprey (www.ensembl.org/Petromyzon_marinus/Info/Index) identified Gsxβ and Cdxβ genes on scaffold_GL477479 (Fig. 2), and the first exon of Pdxα on scaffold_GL481998. The intergenic region of the sea lamprey Gsxβ and Cdxβ genes spanned ∼22 kb and contained three gaps (100 bp, 377 bp, and 100 bp). We filled these gaps by genomic PCR (GenBank accession numbers: KX400882–KX400884) and noted that the intergenic region does not contain a Pdx gene. It is therefore likely that the intergenic region of Japanese lamprey Gsxβ and Cdxβ also does not contain a Pdx gene.

ParaHox Genes in Lamprey RNA-seq Data. To determine whether there are any more ParaHox genes that are not represented in the genome assembly or in degenerate PCR products, we generated RNA-seq data for four tissues (brain, eye, intestine from Japanese lamprey pancreas from brook lamprey, Lampetra planeri) that were identified as the main tissues expressing ParaHox genes in lampreys by RT-PCR analysis (Expression Patterns of Lamprey ParaHox Genes). Searches of the assembled transcripts revealed transcripts for Gsxα in the brain, Gsxβ in the eye, Cdxα and Cdxβ in the intestine, and Pdxα in the pancreas. We also searched RNA-seq transcripts of sea lamprey embryos from stages 23–25 and 26–28 (18) and identified fragments of transcripts for Cdxβ, Gsxα, and Pdxα (SI Appendix, Sequence File). We generated full-length cDNA sequence for sea lamprey Gsxα, Pdxα, and Cdxβ by 5′/3′-RACE, using total RNA from embryo stages 26–28 as a template (Materials and Methods). Overall, our analysis of genome and transcriptome sequences indicate that lampreys contain only five ParaHox genes (Gsxα, Gsxβ, Pdxα, Cdxα, and Cdxβ) that are organized in two clusters (Fig. 2).

Phylogenetic Analysis and Synteny. To determine the orthology of lamprey and gnathostome ParaHox genes, we carried out phylogenetic analysis (Bayesian Inference) of ParaHox genes from lamprey, hagfish, and representative gnathostomes, with amphi- phians as the outgroup. However, phylogenetic analysis was not informative, as the lamprey and hagfish (cyclostomes) genes generally formed an exclusive cluster outside the gnathostome clades (SI Appendix, Fig. S1). This pattern of independent clustering of cyclostome sequences outside gnathostomes sequences suggests the duplication of cyclostome sequences occurred independent from the duplication event in gnathostomes. However, this clustering pattern is likely to be an artifact because of the high GC-content of the lamprey genome, which affects the codon use pattern and amino acid composition of lamprey protein-coding sequences. As a result, sequences of lamprey paralogues are more similar to each other than to their orthologs in gnathostomes. A similar pattern of exclusive clustering has been previously observed for several lamprey genes such as Hox,

Fig. 1. ParaHox gene loci in jawed vertebrates (gnathostomes). Organization of ParaHox gene loci in representative gnathostomes. Genes are depicted as block arrows, with the direction of arrows denoting the transcriptional orientation. The star represents the whole-genome duplication in the teleost lineage. GSX gene in human is known as GSX. Chromosomal/scaffold numbers of the ParaHox genes are shown at the right of each locus.

Results

ParaHox Genes in the Japanese Lamprey Genome. We searched the germline genome assembly of the Japanese lamprey (lampreygenome.imcb.a-star.edu.sg) for ParaHox genes, using human, coelacanth, and elephant shark ParaHox proteins as TBLASTN queries. On the basis of a combination of sequence analysis and RACE and/or reverse-transcription PCR (RT-PCR; SI Appendix, Materials and Methods), we identified five ParaHox genes organized in two clusters in the Japanese lamprey genome. Because we could not determine the orthology of these genes to ParaHox genes in gnathostomes (see following sections), we designated the Gsx, Pdx, and Cdx genes on scaffold_1054 as Gsxα, Pdxα, and Cdxα (α-cluster), and the Gsx and Cdx genes on scaffold_14 as Gsxβ and Cdxβ (β-cluster) (Fig. 2). In addition to genome assembly searches, we carried out degenerate PCR, using genomic DNA as a template to determine whether there were any additional ParaHox genes in Japanese lamprey (Materials and Methods). The degenerate PCRs were able to amplify fragments of Gsxα, Pdxα, Cdxα, and Cdxβ genes, but not Gsxβ, because of the divergence of its sequences corresponding to the forward primers and the presence of a lamprey-specific intron (2.9 kb) in the region targeted by PCR primers (Materials and Methods). Nevertheless, no new ParaHox genes were identified in the PCR products.

ParaHox Genes in the Sea Lamprey Genome. Searches for ParaHox genes in the genome assembly of the sea lamprey (www.ensembl.org/Petromyzon_marinus/Info/Index) identified Gsxβ and Cdxβ genes on scaffold_GL477479 (Fig. 2), and the first exon of Pdxα on scaffold_GL481998. The intergenic region of the sea lamprey Gsxβ and Cdxβ genes spanned ∼22 kb and contained three gaps (100 bp, 377 bp, and 100 bp). We filled these gaps by genomic PCR (GenBank accession numbers: KX400882–KX400884) and noted that the intergenic region does not contain a Pdx gene. It is therefore likely that the intergenic region of Japanese lamprey Gsxβ and Cdxβ also does not contain a Pdx gene.

ParaHox Genes in Lamprey RNA-seq Data. To determine whether there are any more ParaHox genes that are not represented in the genome assembly or in degenerate PCR products, we generated RNA-seq data for four tissues (brain, eye, intestine from Japanese lamprey pancreas from brook lamprey, Lampetra planeri) that were identified as the main tissues expressing ParaHox genes in lampreys by RT-PCR analysis (Expression Patterns of Lamprey ParaHox Genes). Searches of the assembled transcripts revealed transcripts for Gsxα in the brain, Gsxβ in the eye, Cdxα and Cdxβ in the intestine, and Pdxα in the pancreas. We also searched RNA-seq transcripts of sea lamprey embryos from stages 23–25 and 26–28 (18) and identified fragments of transcripts for Cdxβ, Gsxα, and Pdxα (SI Appendix, Sequence File). We generated full-length cDNA sequence for sea lamprey Gsxα, Pdxα, and Cdxβ by 5′/3′-RACE, using total RNA from embryo stages 26–28 as a template (Materials and Methods). Overall, our analysis of genome and transcriptome sequences indicate that lampreys contain only five ParaHox genes (Gsxα, Gsxβ, Pdxα, Cdxα, and Cdxβ) that are organized in two clusters (Fig. 2).

Phylogenetic Analysis and Synteny. To determine the orthology of lamprey and gnathostome ParaHox genes, we carried out phylogenetic analysis (Bayesian Inference) of ParaHox genes from lamprey, hagfish, and representative gnathostomes, with amphibians as the outgroup. However, phylogenetic analysis was not informative, as the lamprey and hagfish (cyclostomes) genes generally formed an exclusive cluster outside the gnathostome clades (SI Appendix, Fig. S1). This pattern of independent clustering of cyclostome sequences outside gnathostomes sequences suggests the duplication of cyclostome sequences occurred independent from the duplication event in gnathostomes. However, this clustering pattern is likely to be an artifact because of the high GC-content of the lamprey genome, which affects the codon use pattern and amino acid composition of lamprey protein-coding sequences. As a result, sequences of lamprey paralogues are more similar to each other than to their orthologs in gnathostomes. A similar pattern of exclusive clustering has been previously observed for several lamprey genes such as Hox,
KCNA, and the p53 family (19–21). Thus, phylogenetic analysis was not useful in assigning orthology between lamprey and gnathostome ParaHox genes. Nevertheless, phylogenetic analysis was able to infer the orthology between Japanese lamprey and sea lamprey sequences and between lamprey and hagfish sequences with high statistical support (posterior probability values, 0.96–1.0; SI Appendix, Fig. S1). Thus, we infer that the hagfish Gsx and Cdx genes (SI Appendix, Fig. S2) are orthologs of Japanese lamprey Gsxα and Cdxα, respectively.

Japanese lamprey Gsxβ and Cdxβ genes are located on a large scaffold (5.3 Mb) and are flanked by several genes. We compared the synteny of the flankng genes with those flanking various gnathostome ParaHox genes to see whether synteny provides any clue to the orthology of these lamprey ParaHox genes. However, the combination of paralogous genes flanking the lamprey Gsxβ and Cdxβ genes is not found in any gnathostome ParaHox locus. For example, homologs of lamprey Pan3, Mtif3, and Gtf3a found in this locus (scaffold_14) are present in the gnathostome Gsx1-Pdx1-Cdx2 locus (SI Appendix, Fig. S2), whereas homologs of lamprey Kar and Rasl1b, also found on this locus (scaffold_14), are present in the gnathostome (e.g., elephant shark) Gsx2-Pdx2 locus (SI Appendix, Fig. S2). This mixed pattern of syntenic genes in the lamprey suggests paralogous genes flanking the ParaHox genes were lost independently in the lamprey after it diverged from the gnathostome lineage. As such, synteny is also not useful in informing the orthology of lamprey and gnathostome ParaHox genes.

Genomic Organization of Lamprey ParaHox Genes. The Japanese lamprey Gsxα and Pdxα are on the same strand of DNA, whereas Cdxα is on the opposite strand, similar to the organization of Gsx, Pdx, and Cdx genes in amphioxus and gnathostomes. However, the Japanese lamprey Gsxβ gene is on the same strand as Cdxβ, which is an unusual arrangement. The inverted orientation of Gsxβ indicates this gene underwent a local inversion in the lamprey lineage. The protein encoded by the lamprey Gsxβ contains several amino acid substitutions in the homeodomain (14 of 60) that is nearly totally conserved in the Gsx proteins of all other chordates, including Gsxα of lamprey (Fig. 3). These substitutions might have altered the DNA binding specificity of the lamprey Gsxβ protein. The lamprey Gsxβ, Cdxβ, and Pdxα genes are also unusual, in that they each contain an extra intron compared with their homologs in gnathostomes (Fig. 3 and SI Appendix, Fig. S3).

Expression Patterns of Lamprey ParaHox Genes. We examined the expression pattern of the Japanese lamprey ParaHox genes in the brain, eye, and intestine by RT-PCR. Both Cdxα and Cdxβ genes were found to express in the intestine (Fig. 4), similar to Cdx1, Cdx2, and Cdx4 genes of mammals (22) and Cdx1α and Cdx1b of zebrafish (23, 24). The Japanese lamprey Gsxα showed expression in the brain (Fig. 4), similar to Gsx1 and Gsx2 genes of mouse (25, 26) and zebrafish (27, 28). However, the Japanese lamprey Gsxβ did not show expression in the brain, but instead was found to express in the eye, in addition to a low level of expression in the intestine (Fig. 4). None of the vertebrate ParaHox genes characterized so far have been shown to express in the eye. Thus, the eye expression of lamprey Gsx may seem to be an expression domain specific to the lamprey lineage. In gnathostomes, Pdx genes express specifically in the pancreatic tissue (4). Because we did not have access to the pancreatic tissue of the Japanese lamprey, we obtained cDNA from the pancreatic tissue of metamorphic juvenile sea lamprey and checked for the expression of Pdxα. The sea lamprey Pdxα was indeed found to express in this tissue (Fig. 4).

Conserved Noncoding Elements. Noncoding elements conserved in distant vertebrates have been shown to often function as enhancers (20, 29, 30). To determine whether the cyclostome ParaHox gene loci contain any evolutionarily conserved regulatory elements, we generated MLAGAN alignments of ParaHox loci from lamprey, hagfish, and selected gnathostomes (elephant shark, human, coelacanth, spotted gar, and zebrafish) and predicted conserved noncoding elements (CNEs) that are at least 70% identical across >100-bp windows (Materials and Methods). Because the orthologous relationship between the cyclostome and gnathostome ParaHox genes was not clear, we aligned the Japanese lamprey and hagfish Gsxα locus with the gnathostome Gsx1 locus, as well as Gsx2 locus and predicted CNEs. Both lamprey and hagfish Gsxα loci were found to share a CNE with the gnathostome Gsx1 locus (Fig. S4), but none with the gnathostome Gsx2 locus (SI Appendix, Fig. S4). The sharing of a CNE between the Japanese lamprey and hagfish Gsxα locus provides further support to our inference based on phylogenetic analysis that these loci are indeed orthologous. The CNE identified in the lamprey and hagfish Gsxα locus provides further support to our inference based on phylogenetic analysis that these loci are indeed orthologous. The CNE identified in the lamprey and hagfish Gsxα locus provides further support to our inference based on phylogenetic analysis that these loci are indeed orthologous.
Comparison of intron positions in the Gsx/Gsh genes of gnathostomes and jawless vertebrates. Multiple alignments of Gsx/Gsh proteins from representative gnathostomes (human, coelacanth, spotted gar, and elephant shark), jawless vertebrates (Japanese lamprey, sea lamprey, and inshore hagfish), and the cephalochordate amphioxus showing a selected region of the alignment for the Gsx/Gsh family. Intron positions are shown as arrowheads. The numbers next to the arrowheads indicate intron phases. Introns identified in the lampreys are shown as red arrows. The amino acid positions of the human sequences are shown. The boxed region denotes the homeodomain. The unique amino acid substitutions in the Hox domain of lamprey Gsxβ are shown in red font.

**Fig. 3.** Comparison of intron positions in the Gsx/Gsh genes of gnathostomes and jawless vertebrates. Multiple alignments of Gsx/Gsh proteins from representative gnathostomes (human, coelacanth, spotted gar, and elephant shark), jawless vertebrates (Japanese lamprey, sea lamprey, and inshore hagfish), and the cephalochordate amphioxus showing a selected region of the alignment for the Gsx/Gsh family. Intron positions are shown as arrowheads. The numbers next to the arrowheads indicate intron phases. Introns identified in the lampreys are shown as red arrows. The amino acid positions of the human sequences are shown. The boxed region denotes the homeodomain. The unique amino acid substitutions in the Hox domain of lamprey Gsxβ are shown in red font.

**Fig. 4.** Expression pattern of lamprey Parahox genes. RT-PCR analysis of lamprey Parahox genes. Liver cDNA was used as a negative control. β-actin was amplified as an internal control to assess the quality of the cDNA. Because we did not have access to pancreatic tissue from Japanese lamprey, we used cDNA from the pancreatic tissue of sea lamprey for RT-PCR of Pdxα (gel image on the Right).
domain in which Cdx genes are known to express at high levels. In the ParaHox clusters of gnathostomes and amphioxus, Gsx genes are always located at the 5′ end of the cluster and are expressed in the brain, whereas Cdx genes are found at the 3′ end of the cluster and are typically expressed in the hindgut. The unusual expression of the lamprey Gsxβ in the intestine could be a result of the inversion of this gene in the lamprey lineage, which might have brought its promoter in close proximity to the intestine enhancer of Cdxβ, resulting in the somewhat leaky expression of Gsxβ in the intestine.

Most gnathostomes contain four paralogous loci for several genes, including the Hox genes and ParaHox genes, whereas amphioxus, a basal chordate, contains a single locus for these genes. The quadruple loci in gnathostomes have been attributed to the two rounds of WGD during the early evolution of vertebrates (33). We have previously shown that the Japanese lamprey contains at least six Hox gene loci, in contrast to four Hox clusters in most gnathostomes (20), suggesting the Hox clusters may have experienced an additional round of duplication in the lamprey lineage (20). However, the Japanese lamprey contains only two ParaHox gene loci. This suggests the duplication event that gave rise to the six Hox loci did not include the ParaHox cluster, implying that the additional Hox loci are not the result of WGD, but are instead a result of segmental duplication or duplications, a scenario recently proposed based on the meiotic map of the sea lamprey genome (34). According to this study, the jawless vertebrate and gnathostome lineages experienced only one round of WGD before their separation, which was followed by several independent segmental duplications in the two lineages (34). However, an alternative possibility is that the lamprey ParaHox clusters duplicated along with the Hox clusters as part of the WGD events, but subsequently experienced secondary losses, resulting in the retention of only five ParaHox genes in two loci. This possibility is supported by the presence of some Japanese lamprey paralogues whose gnathostome homologs are linked to the additional ParaHox gene loci. For example, gnathostome Pdgfra and Kit genes are linked to the ParaHox “B-cluster,” whereas their paralogues, Pdgfrb and Csf1r, are linked to the ParaHox “C-cluster” (SI Appendix, Fig. S2). The Japanese lamprey genome contains homologs of both Kit and Csf1r (SI Appendix, Fig. S7A). Although the lamprey Kit is linked to the β-cluster (SI Appendix, Fig. S2), Csf1r is located on scaffold_154 that does not contain any ParaHox gene. The Japanese lamprey genome also contains two Pdgfr genes (SI Appendix, Fig. S7B), located on scaffold_1691 and scaffold_22, that do not harbor any ParaHox genes. The presence of Kit and Csf1r, and two Pdgfr paralogues in the lamprey genome, suggests the ancestor of Japanese lamprey contained more than two ParaHox clusters, and the ParaHox genes in the additional clusters were subsequently lost in the lamprey lineage. Thus, it seems likely that the ParaHox loci in the lamprey lineage underwent at least two rounds of duplication, along with the Hox clusters, as a result of WGD events, and subsequently only two ParaHox loci were retained in the lamprey lineage because of extensive secondary loss of ParaHox genes.

The homolog of the Pdx gene in invertebrates, known as Xlox, is expressed in the gut, whereas Pdx in gnathostomes is expressed in the pancreas, an organ specific to vertebrates. In mammals, Pdx1 plays a crucial role in the development of the pancreas,
differentiation of beta cells, and maintenance of their function (35), and is therefore vital for insulin production. Inactivation of Pdx1 in beta cells of mice is known to result in maturity-onset diabetes (36) whereas mutations in human PDX1 (also known as IPF1) result in pancreatic agenesis (37). Characterization of a ParaHox cluster in hagfish had shown that the Pdx gene in this locus has become a pseudogene (17). Our study shows that the ortholog of this Pdx gene in lampreys is intact and expresses in the pancreas. Thus, the Pdx gene in lampreys is likely to be functional. In contrast to hagfish, whose islet organ (endocrine pancreas) comprises scattered follicles, lampreys contain discrete islet organs (38). The presence of a functional Pdx gene in lampreys supports the notion that the development of its discrete islet organ, as well as insulin production, is mediated by Pdx similar to that in gnathostomes. In humans, PDX1 functions in pancreas as part of a network comprising insulin, glucagon, SLC2A2, FOXA2, HNF1A, NEUROD1, PRKACB, and PRKACG (STRING database, https://string-db.org/). The brook lamprey pancreas RNA-seq data contains transcripts (SI Appendix) for genes encoding homologs of all these proteins except PRKACG. Thus, although lampreys do not possess a distinct pancreatic gland similar to jawed vertebrates, most of the genes associated with the Pdx-network of jawed vertebrate pancreas are present and expressed in the islet organ of lampreys. A previous study had shown that cartilaginous fishes possess a three-hormone (insulin, glucagon, and somatostatin) pancreas and that the four-hormone (the three plus pancreatic polypeptide) pancreas found in coelacanths and tetrapods was a later innovation (39). The brook lamprey pancreas RNA-seq data includes transcripts (SI Appendix) for the three hormones present in the pancreas of cartilaginous fishes, indicating that the three hormone pancreas had evolved in the common ancestor of all vertebrates.

Materials and Methods

The Japanese lamprey and sea lamprey genome assemblies were searched for ParaHox genes by TBLASTN, using selected vertebrate ParaHox proteins as queries. RNA-seq reads for the Japanese lamprey brain, eye, and intestine and brook lamprey pancreas were generated on the Illumina platform and assembled using Trinity (40). To look for additional ParaHox genes in the genome of the lamprey, degenerate PCR was carried out using genomic DNA as a template. Details of the primers used and PCR conditions are given in the SI Appendix. Phylogenetic analysis of ParaHox protein sequences were performed using Bayesian inference as implemented in MrBayes (version 3.2.6; mbayes.sourceforge.net). CNEs in the ParaHox loci were predicted by aligning the sequences with MLAGAN and visualized using VISTA (genome. ncbi.gov/vista/index.shtml). The Japanese lamprey and zebrafish CNEs were assayed for enhancer function in transgenic zebrafish using GFP as a reporter. Further details of the materials and methods are given in the SI Appendix.

Acknowledgments. We thank Michael Schorpp and Thomas Boehn for providing pancreatic tissue of brook lamprey, and Jonathan M. Wilson for sea lamprey pancreas cDNA. This work was supported by the Biomedical Research Council of the Agency for Science, Technology and Research of Singapore.


