Gene Novelties in Amphioxus Illuminate the Early Evolution of Deuterostomes

Authors: Qing Xiong[1,2]†, Kevin Yi Yang[1,2]†, Xi Zeng[3]†, Mingqiang Wang[1,4]†, Patrick Kwok-Shing Ng[1], Jun-Wei Zhou[1], Judy Kin-Wing Ng[1], Cherie Tsz-Yiu Law[1], Qiao Du[3], Kejin Xu[3], Laura Jane Falkenberg[5], Bingyu Mao[6], Jun-Yuan Chen[7], Stephen Kwok-Wing Tsui[1,2,8] *

Affiliations:

1 School of Biomedical Sciences, The Chinese University of Hong Kong, Hong Kong.
2 Hong Kong Bioinformatics Centre, The Chinese University of Hong Kong, Hong Kong.
3 Agricultural Bioinformatics Key Laboratory of Hubei Province, 3D Genomics Research Centre, College of Informatics, Huazhong Agricultural University, Wuhan, China.
4 Stanford Cardiovascular Institute, Stanford University School of Medicine, Stanford, CA, 94305, USA.
5 Simon FS Li Marine Science Laboratory, School of Life Sciences, The Chinese University of Hong Kong, Hong Kong.
6 State Key Laboratory of Genetic Resources and Evolution, Kunming Institute of Zoology, Chinese Academy of Sciences, 650223, Kunming, Yunnan, China.
7 State Key Laboratory of Palaeobiology and Stratigraphy, Nanjing Institute of Geology and Palaeontology, Chinese Academy of Sciences, Nanjing, China.
8 Centre for Microbial Genomics and Proteomics, The Chinese University of Hong Kong, Hong Kong.

* Correspondence to Professor Stephen Kwok-Wing Tsui, School of Biomedical Sciences, The Chinese University of Hong Kong, Shatin, N.T., Hong Kong. E-mail: kwtsui@cuhk.edu.hk; Tel: 852-3943 6381; Fax: 852-2603 5123.

† These authors contributed equally to this work.

Preprint Servers:

This manuscript has been deposited in *bioRxiv* (https://doi.org/10.1101/2022.05.18.492404).

Classification: BIOLOGICAL SCIENCES, Evolution

Keywords: gene novelty, genome evolution, amphioxus, chordate, deuterostome

Abstract

Amphioxus, as the best living proxy to the chordate ancestor, is considered an irreplaceable model organism for evolutionary studies of chordates and deuterostomes. In this study, a high-quality genome of the Beihai amphioxus, *Branchiostoma belcheri* beihai, was *de novo* assembled and annotated. Within four amphioxus genomes, a wide range of gene novelties were identified, revealing new genes that share unexpectedly high similarities with those from non-metazoan species. These gene innovation events have played roles in a range of amphioxus adaptations, including innate immunity responses, adaptation to anaerobic environments, and regulation of calcium balance. The gene novelties related to innate immunity, such as a group of lipoxygenases and a DEAD-box helicase, boosted amphioxus immune responses. The novel genes for alcohol dehydrogenase and ferredoxin could aid in the anaerobic tolerance of amphioxus. A proximally arrayed cluster of EF-hand calcium-binding protein genes were identified to resemble those of bacteria. The copy number of this gene cluster was linearly correlated to the sea salinity of the collection region, suggesting that it may enhance their survival at different calcium concentrations. Collectively, this comprehensive study on gene novelties of amphioxus reveals insights into the early genome evolution of chordates and deuterostomes and provides valuable resources for future research.

Significance statement

Amphioxus is a slow-evolving model organism for the study of chordate evolution, and while much research has focused on its ortholog evolution, relatively little attention has been paid to its gene innovations. In this study, we assembled and annotated a high-quality genome of the Beihai amphioxus and identified gene novelties across four amphioxus genomes. Many of these new genes resembled those from non-metazoan species, suggesting the possibility of horizontal gene transfer. These gene novelties have been found to enhance amphioxus adaptations, including innate immune responses and anaerobic tolerance. Of particular interest is a cluster of novel EF-hand calcium-binding protein genes located in proximity, which present similarity to bacterial genes. These genes may facilitate adaptation to different calcium concentrations in amphioxus.

Main text

Introduction

Amphioxus, a group of non-vertebrate chordates under Cephalochordata, diverged from the common ancestor of other two chordate groups, urochordates (tunicates) and vertebrates, around 550 million years ago (Ma) (1-3). Compared with tunicates, which were considered as a rapidly-evolving group (1), amphioxus shares a basic chordate body plan, undergoes slower molecular evolution and has fewer lineage-specific genomic changes (4, 5). Therefore, amphioxus is considered an irreplaceable model organism for evolutionary studies. With the advent of high-throughput sequencing, genome-based investigation has revolutionized the study of amphioxus evolution and greatly expanded our knowledge of the origin and evolution of vertebrates (5-12). Whole-genome comparison of the three chordate groups provided evidence supporting the 2R hypothesis, which posits that two rounds of whole-genome duplication occurred in the early vertebrate lineage (5, 6). The application of functional genomics to amphioxus has provided a wealth of information on the origins of the complex gene regulatory landscapes that are characteristic of vertebrates, shedding new light on the evolutionary history of this important group of animals (9).

In this study, we successfully constructed a high-quality genome of the Beihai amphioxus, *Branchiostoma* (*B.*) *belcheri* beihai, which was collected from the nearshore area of South China. Although amphioxus is widely regarded as the best extant proxy to the chordate ancestor and lacks the speciation and genome innovations of vertebrates (13-15), we performed a comparative genomic analysis of amphioxus to identify gene novelties and focused on genes that unexpectedly resemble those from non-metazoan species, including bacteria, fungi and plants. Our analysis revealed that

these gene novelties play crucial roles in a wide range of amphioxus adaptations, including innate immunity, anaerobic adaptation, and calcium balance. Overall, our comprehensive genomic analysis of the gene novelties in amphioxus provided insights into the genetic mechanisms that underlie the adaptations and diversification of chordates and deuterostomes.

Results and Discussion

Genome assembly and annotation

A high-quality genome of the Beihai amphioxus, *B. belcheri* beihai (16, 17), was *de novo* assembled, annotated, and compared with three other genome assemblies of the Xiamen amphioxus (*B. belcheri* xiamen), Florida amphioxus (*B. floridae*) and European amphioxus (*B. lanceolatum*) (Supplementary Note 1.1, Table S1, Fig. S1). For the genome assembly of *B. belcheri* beihai, the scaffold N50 length (4,185,906 bp) was second only to that of *B. floridae* (25,441,410 bp), the gap content was the lowest among the four amphioxi and the BUSCO completeness was one of the highest (97.2%) (Supplementary Note 1.2, Table S1). In the genome annotation of *B. belcheri* beihai, 44,745 protein-coding genes were annotated with 95.2% BUSCO completeness. The genome of *B. lanceolatum* was annotated with 38,190 protein-coding genes with 91.9% completeness (Supplementary Note 1.3, Table S1).

Phylogeny and comparative genomics

Along with other genome assemblies (Supplementary Note 2.1, Table S2), a phylogenetic analysis was performed, with the sponge *Amphimedon queenslandica* used as the outgroup. The deuterostome phylogeny of this organism was consistent with current opinions (18, 19) and supported the monophylies of both chordates and deuterostomes (Fig. 1). Additional deuterostome phylogenetic analyses further supported the monophylies and refuted confusing phylogenies (Supplementary Note 2.2, Fig. S2). Amphioxus was suggested as the outermost or first-evolving chordate group (Fig. 1 and S2A).

To gain insights into the genome evolution of amphioxus, we performed a comprehensive comparative genomics analysis of annotated animal genomes (Table S2). Based on the ultrametric time tree and phylogenomic orthology analysis (Supplementary Note 2.3), we observed significantly more assigned orthogroups in the four amphioxus genomes compared to other animal genomes, including over 2,000 amphioxus-specific orthogroups (Fig. S3). This observation piqued our interest in exploring the gene novelties that underlie the genome evolution of amphioxus.

Gene novelties

Within four annotated amphioxus genomes (Table S1), we identified a wide range of gene novelties conserved among them (Table 1), in which the involved genes resemble those from bacteria, fungi, plants or other eukaryotes. In addition to cautious filtering and rigorous validation to avoid false positivity in assembled sequences, the high conservation of those genes among four independently collected, sequenced, and assembled genomes of amphioxi (Table 1) precluded the occurrence of contamination. Previous studies have identified a number of gene novelties in stem chordates chordates (6) and deuterostomes (20). To avoid repeating these efforts and to focus on novel gene features in amphioxus, we excluded candidate novelties that are also conserved in jawed vertebrates, including the sialidase gene, which may have arisen in the deuterostome stem (Fig. S4, Supplementary Note 3.1). Also, we excluded novel genes that may originate from extreme sequence variation or new protein domain combinations, as these genes were less likely to be truly unique to amphioxus.

A gene novelty could generate many novel genes via gene duplications; for example, in *B. belcheri* beihai, 28 gene novelties generated 108 new genes (Table 1). According to their biological functions, these involved genes were classified into six categories, namely, innate immunity, anaerobic adaptation, calcium balance, amino acid metabolism, other metabolic genes, and others without explicit functional domains (Table 1). The expression level of these genes was analyzed in different tissues of *B. lanceolatum* because of its high-quality transcriptome data, and all the genes were expressed in at least one tissue (Fig. S5). Regardless of the inconsistent gene expression in tissue duplicates of muscle and male gonads, the most marked expression feature was that the genes coding for calcium-binding proteins were highly expressed in the guts and hepatic diverticula (Fig. S5) (Supplementary Note 3.2). Subsequently, we explored these novelty genes in three important functional groups, namely innate immunity, anaerobic adaptation, and calcium balance. More gene novelties were discussed in Supplementary Note 3.

Innate immunity

Innate immunity is the first-line nonspecific immune system against infections, and this immune system is observed in jawed and jawless vertebrates as well as invertebrates (21). Successful immunity evolution confers survival traits to a species. In the absence of an adaptative immune system in amphioxus (21), innate immunity is the most important weapon of amphioxus in response to various pathogens.

Lipoxygenases (LOXs) are important in arachidonic acid metabolism, participate in inflammatory lipid signaling and further regulate inflammatory and allergic responses (22, 23). A group of

duplicated LOX genes of amphioxus were identified similar to those of cyanobacteria or proteobacteria (Table 1). Notably, these LOX genes have unexpected homologs from only two living groups of jawless vertebrates (Cyclostomata) (24), the sea lamprey *Petromyzon marinus* and the hagfish *Eptatretus burger* (Fig. S6); however, such homologs are not observed in other metazoa. In the phylogenetic tree, these LOX genes of amphioxi are close to those of sea lamprey, cyanobacteria and proteobacteria (Fig. 2A). Three splice sites (locations of exon–intron boundaries) were conserved among the LOX genes of four amphioxi and the sea lamprey (Fig. S5). In the absence of genome annotation, two fragments of the LOX gene were recovered from the hagfish genome, and three other conserved splice sites were identified between those of the two jawless vertebrates (Fig. S6). Therefore, the LOX genes of four amphioxi and two jawless vertebrates originated from the same eukaryotic source.

The phylogenetic analysis revealed that these LOX genes were duplicated and diversified among the four amphioxi and *LOX_Bf_9* was an outlier gene (Fig. 2B). All the LOXs contain a C2-like domain (Fig. 2B). Except for the five LOXs of *B. floridae* and *B. lanceolatum*, the LOXs possessed one to five domains of the PYR (pyrabactin resistance)/PYL (PYR1-like)/RCAR (regulatory components of ABA receptor) family in the SRPBCC superfamily (SRPBCC domains, Fig. 2B) and these SRPBCC domains could be further divided into three groups (Fig. S7). In Fig. 2C, the gene synteny alignment revealed that many LOXs (12 of 27 genes) were proximally located (distance <20 Kb) and possibly generated by tandem gene duplications. *LOX_Bl_5* had an apparently wide expression in the gills, gut, and hepatic diverticula (Fig. S5).

These unusual LOX genes containing a C2-like domain were reported in amphioxi (22, 23) and first in jawless vertebrates, although the SRPBCC domains were not identified. Since SRPBCC superfamily proteins contain a deep hydrophobic ligand-binding pocket and can bind diverse ligands (25), we consider that these SRPBCC domains may also function as receptor domains in these LOX genes in response to various signals. A multifunctional LOX containing a SRPBCC domain was reported in the red alga *Pyropia haitanensis* and presented substrate flexibility, oxygenation position flexibility, and functional versatility (25). These LOX genes may enrich the inflammatory lipid signaling pathway of amphioxus.

DEAD-box helicases (DBHs) are known as sensors for nucleic acids and mediators of antiviral innate immunity (26). A DBH gene was identified as a single-copy gene that resembles those from bacteria, and it was conserved among four amphioxi (Table 1). In the phylogenetic tree (Fig. 3A), this DBH gene of amphioxi is close to those of the charophyte alga *Klebsormidium* (*K.*) *nitens*, soft coral *Dendronephthya gigantea*, hemichordate *Saccoglossus kowalevskii*, and a group of bacteria. Intriguingly, the DBH genes of amphioxi shared three conserved splice sites with those of charophyte alga, soft coral, and hemichordate (Fig. 3B and S7), which suggested that the DBH genes of four amphioxi, soft coral and hemichordate shared a common source with that of the charophyte alga. A homologous DBH gene was also identified in the genome of the other hemichordate, *Ptychodera flava*, in the absence of annotation. This DBH gene was widely expressed at low levels in the tissues of *B. lanceolatum* (Fig. S5). The single-copy DBH gene may serve as a novel sensor in the antiviral innate immunity of amphioxi. Another DBH gene was reported as a horizontally transferred gene in coral species (27).

In addition, another nine conserved splice sites in orange lines (Fig. 3B and S8) indicated close phylogenetic relationships among the DBHs of the four amphioxi, soft coral and hemichordate; another four conserved splice sites in blue lines (Fig. 3B and S8) suggested that the DBHs of the soft coral and hemichordate originated from the same ancestral gene, while two unusually conserved splice sites marked with black lines (S1 and S2, Fig. 3B and S8) could be explained by intron loss during gene evolution, which indicated S1 loss in ancestral amphioxus but S2 loss in both the ancestral amphioxus and hemichordate. Other gene novelties related to innate immunity can be found in Supplementary Note 3.3.

Anaerobic adaptation

Amphioxus lives in subtidal shallow water with extremely fluctuating dissolved oxygen level (28) and burrow in sandy substrate where is often rich in excrement and anaerobic (29). Amphioxus has no mature respiratory system and breathe mainly through at the surface over the coelomic cavity (30). Under anoxic conditions, the soft tissues of amphioxus (*B. lanceolatum*) could survive for as long as 124 days (31). Additionally, another amphioxus species, *Asymmetron inferum*, survives in anaerobic and sulfide-rich bottom environments (229 m depth) (32). Therefore, we suggest that amphioxi are tolerant to an anaerobic environment.

Alcohol dehydrogenases (ADHs) catalyze the oxidation of alcohols to aldehydes and ketones (33). In the phylogenetic tree (Fig. 3C), an ADH gene containing zinc-binding sites and belonging to the conserved medium-chain dehydrogenase/reductase (MDR) family was much closer to those of the

sponge *Amphimedon queenslandica*, the copepod *Eurytemora affinis*, and a bacterial group than to those of other metazoa. The splice sites of the ADH gene of amphioxi are highly conserved with those of the sponge but not the copepod (Fig. 3D and S11), which suggested that the ADH genes of amphioxi and *Amphimedon queenslandica* have the same eukaryotic origin. Among the 100 best hits of this ADH gene in the NCBI NR database, we noticed over ten hits from anaerobic bacteria and the ADH gene of *B. belcheri* beihai shared 64% identity with that of *Anaerolineae bacterium* (GenBank accession: MCD6286762.1).

Ferredoxin (FDX) contains an iron-sulfur cluster and is known as a mediator of electron transfer in many metabolic pathways (34). The FDX gene of amphioxus was identified to be highly conserved with the genes in certain plants, bacteria, or other eukaryotes (identities over 68%) but not with that of other metazoa (excluding Amphioxiformes, Table 1). This FDX gene was tandemly duplicated into two copies on opposite strands in *B. floridae*. All these FDX genes contain a Ubl_FUBI (Ubiquitin-like protein FUBI) domain (Fig. S12A), which may protect the FDX gene from ubiquitination and subsequent degradation (35).

Genes encoding FDXs, nitroreductases, NADH oxidase, and ADH3 have been suggested to be horizontally transferred from anaerobic prokaryotes to two intestinal protozoan parasites (36) and facilitate their survival under anaerobic conditions of intestinal lumen. Collectively, we consider that the ADH and FDX genes may play similar roles in amphioxus, thereby facilitating their survival in low-oxygen and even anoxic environments.

In addition, a group of atypical short-chain dehydrogenase/reductase subgroup 4 (SDR_a4) genes were identified mostly proximally arrayed (distance <20 Kb) and divided into two clusters in four amphioxi (Fig. S12B and C). The SDR_a4 gene was tandemly duplicated in all four amphioxi (Fig. S12B) and could be divided into two clusters in the phylogenetic tree (Fig. S12C). SDR_a4 genes function in redox sensor mechanisms and may also contribute to anaerobic adaptation (37, 38).

Therefore, we propose that amphioxus developed adaptation and tolerance to a low-oxygen environment with these gene novelties. In addition, as an ancient animal emerged around 550 Ma (39), amphioxus experienced three events with extreme drops in oxygen concentration (40). Since the gene novelties related to anaerobic adaptation are conserved among the four amphioxi in different oceans, they should have occurred before the geographical separation caused by the breakup of Pangaea (initiated approximately 200 Ma) (41). Therefore, we propose that the anaerobic adaptation-related gene novelties were possibly the result of the ancient extreme drops in oxygen concentration that occurred during 200–500 Ma, especially the two major drops that occurred in this period (40).

Calcium balance

Calcium balance and osmoregulation of fishes are important for their physiological functions (42). A voltage-dependent calcium channel (VDCC), EF-hand calcium-binding protein (EF-CaBP) group and calcium-binding protein (CML26) were identified as gene novelties (Table 1).

VDCC is a complex of several different subunits (43); however, whether this VDCC gene functions

independently as a calcium channel deserves further investigation. As shown in Fig. 3E, the VDCC genes of amphioxi, the hemichordate *Saccoglossus kowalevskii*, and three starfishes were located closer to those of other eukaryotes and plants, especially the VDCC genes of the choanoflagellate *Salpingoeca rosetta* (bootstrap value 53), whereas they were located far away from other metazoa, including many protostomes. The VDCC genes of amphioxi, hemichordate and three starfishes shared three conserved splice sites (Fig. 3F and S13), although no conserved splice sites are shared with choanoflagellate. This VDCC gene was proposed to be related to electrophysiological function and was widely expressed among tissues, but no apparent tissue-specific expression (Fig. S5).

EF-CaBPs may function in maintaining a high calcium concentration in the endoplasmic reticulum (44) and could be involved in bone formation (45). Of the 73 newly identified EF-CaBP genes in four amphioxi, the bit scores of the best hits to bacteria were significantly higher than those to other metazoa (Fig. 4A and S14).

This group of EF-CaBP genes was proximally arrayed in four amphioxi and presented high copy number variation. *B. belcheri* xiamen has 25 EF-CaBP genes, *B. belcheri* beihai has 19 genes, *B. floridae* has 17 genes, and *B. lanceolatum* has only 12 genes. Marine salinity is linearly correlated with the calcium ion concentration (46). Therefore, we further explored the relationship of the gene number variation of these proximally arrayed EF-CaBP genes and the sea surface salinity of the collection locations (Fig. 4B). The Mediterranean Sea has a much higher than average seawater salinity (approximately 35 parts per thousand) (47, 48), and the Atlantic Ocean is globally saltier than the Pacific Ocean (49). Because of the absence of high-quality salinity data for these nearshore

regions where the four amphioxi were obtained, we used the average sea surface salinity data in the analysis. The sea surface salinities of *B. belcheri* beihai, *B. belcheri* xiamen, *B. floridae* and *B. lanceolatum* were estimated as 27.65, 22.9, 34.7 and 37 practical salinity unit (PSU) respectively (50-53), and these values were linearly correlated with the gene number of the proximally arrayed EF-CaBP genes ($R^2 = 0.9041$, Fig. 4B).

In the phylogenetic tree (Fig. 4C), except for the outlier gene *Bf_10*, the EF-CaBP genes were classified into two large groups, Clan A and Clan B. Clan A was further classified into three clusters, Cluster 1, Cluster 2, and Cluster 3. In Cluster 1, no genes of *B. floridae* were identified while eight genes of two *B. belcheri* were assigned to four pairs, and they included four highly conserved genes, *Bbb_2*, *Bbb_3*, *Bbx_21* and *Bbx_22*. However, three and two tandemly arrayed genes in Cluster 2 were identified in *B. floridae* and *B. lanceolatum*, respectively, and variable duplications were observed in Cluster 3. The phylogenetic analysis indicated that at least three genes in Clan A appeared in the ancestral amphioxus species. For Clan B, five relatively unique genes of *B. floridae*, *Bf_1-5* in Cluster 4, were proximally arrayed, and many paired genes were identified in two *B. belcheri*, such as three pairs in Cluster 5. All 73 EF-CaBPs were proximally arrayed in four amphioxi (Fig. 4D) and all spaces between genes were below 20 Kb. Notably, four longer genes, *Bbb_6*, *Bbx_18*, *Bf_11* and *Bl_6*, contained more EF-hand motifs (Cluster 3, Fig. 4C).

Except for the large group of proximally arrayed EF-CaBPs above, the CML26 gene was a calcium-binding gene with a shorter length (Table 1). This CML26 gene was tandemly duplicated in *B. floridae* and *B. lanceolatum* (Table 1). Initially, we considered the CML26 gene (Table 1) to belong

to another event, but the conserved splice sites (Fig. S15) suggested that the CML26 gene and the large group of EF-CaBPs possibly originated from a common ancestor.

Along with the two CML26 genes, these EF-CaBP genes were apparently highly expressed in the gut and hepatic diverticulum, which are both considered digestive organs for calcium absorption (Fig. S5). Thus, the expansion of these EF-CaBP genes was proposed to occur in response to marine salinity and calcium concentration changes. The *EF-CaBP_Bl_11* and the two CML26 genes had the highest expression levels (Fig. S5).

Conclusions

As an important group of invertebrate chordates, amphioxi play a pivotal role in the evolutionary study of deuterostomes and chordates. In this study, the genome of the Beihai amphioxus, *B. belcheri* beihai, was *de novo* assembled and annotated (Table S1). Combined with the three amphioxus genomes (Table S1), we performed a phylogenetic analysis of the deuterostome genomes (Fig. 1 and S2), which supported the monophylies of deuterostomes and chordates.

We identified massive gene novelties conserved among the four amphioxus genomes (Table 1, Supplementary Note 3.1). Among these gene novelties, except for the TIR domains of amphioxi, which was mentioned in a previous report (54), all other genes were newly identified and reported. Certain involved genes had homologs from other deuterostome groups except for the jawed vertebrate group (Table 1 and S3). That is, the LOX gene had homologs from two jawless vertebrates; 8 genes had homologs from either tunicates or echinoderms or hemichordates; and only the VDCC gene had homologs from both echinoderms and hemichordates. The conservation of these genes can be well explained by frequent gene gain and loss in the early and rapid evolution of deuterostomes (Table S3) (19). For example, the LOX gene was suggested to be a novel gene that was introduced at the deuterostome stem but decayed in tunicates and jawed vertebrates. Those gene novelties could be well dated according to their conservation among different deuterostome groups, such as the three genes conserved with urochordates (SCP24, MCO and UF-TM3) were proposed as results of events occurred at the chordate stem, while their absence in vertebrates is due to gene decay.

The gene novelties were further comparatively analyzed with other groups like tunicates, thus providing additional insights into the early evolution of chordates and deuterostomes. A wide range of adaptation innovations were enabled by these genes, such as innate immunity-related events that contributed more complexity to the immune system in amphioxus, especially for amphioxus has no adaptative immune system.

Proposed mechanisms

Although it is difficult to rule out gene loss in other groups like protostomes and/or extreme divergence, those gene novelties, especially the genes with phylogenetic supports, can be explained as results of horizontal gene transfer (HGT) from cross-kingdom origins such as marine bacteria, considering many genes shared high similarity with those from unusual taxonomic groups. Despite unknown molecular mechanism, the HGT in amphioxus could be facilitated by external fertilization based on the weak-link model, which suggests that the unicellular fertilized egg cells of amphioxus are potential entry points for these foreign genes (55), especially when many potential HGT origin species are marine microbes, such as cyanobacteria.

The HGT of amphioxus could be explained by the 'you are what you eat hypothesis' (56), which means that the HGT genes originated from the food eaten by amphioxus. All modern deuterostomes were considered to have evolved from a filter-feeding ancestor (20). Along with other deuterostomes, such as tunicates, echinoderms, and hemichordates, amphioxus is known as a filter feeder (57, 58) and may obtain genes from marine microbes, including bacteria and algae, which are potential foods. For example, the charophyte alga *K. nitens* frequently appeared in the analysis of the HGT genes,

including the DBH, SCP24 and UF-TM1 genes (Fig. 3A and B, Fig. S15 and Table 1). The DBH gene was unusually identified in the charophyte alga *K. nitens* (Fig. 3A). The highly conserved splice sites among the DBH genes of the charophyte algae, soft coral, hemichordate, and amphioxi (Fig. 3B and S8) suggested that their DBH genes originated from a common ancient eukaryote, such as some marine algae. Although the charophyte alga *K. nitens* is considered a freshwater/terrestrial and nonmarine species (59), we proposed that the ancestor of *K. nitens* was possibly distributed in marine environments and served as food for these marine animals, especially because of the sister marine species *K. marinum* (no available genome sequences). Similarly, the SCP24 gene (Fig. S16) and UF-TM1 gene (Table 1) possibly originated from the marine ancestor of *K. nitens*.

New gene generation

The origin of new genes is a fundamental process of genome evolution (60). In this study, the comparative analysis of the gene novelties revealed insights into the new gene generation. Regardless of whether common gene duplication and diversification occurred, many genes of amphioxi underwent possible gene fusion or fission to generate new genes. Three novel models of new gene generation were summarized, including gene fusion (Fig. S25A), gene fission (Fig. S25B) and partial duplication or deletion models (Fig. S25C).

In the gene fusion model (Fig. S25A), we revealed that two distinct TIR domains fused with a wide range of endogenous genes (or domains) with immune functions (Fig. S10) and generated a series of new immune genes. In addition, the FDX gene fused with an endogenous Ubl_FUBI domain (Fig. S12A). Both gene fission and partial duplication or deletion were observed in the LOX family of

amphioxi (Fig. 2B). Since we found that the LOX genes of amphioxi were mostly similar to bacterial LOX genes containing up to three SRPBCC domains and three groups of SRPBCC domains could be clustered with three SRPBCC domains of the cyanobacterium *Coleofasciculus sp. FACHB-125* (GenBank accession: WP_190414990.1) (Fig. S7), we considered that the original LOX gene of amphioxi contains three SRPBCC domains. Then, they were split into LOX genes with only C2 domains and genes with only SRPBCC domain via gene fission (Fig. S25B) and generated a series of LOX genes with variable numbers of SRPBCC domains (Fig. 2B) via partial gene duplication or deletion (Fig. S25C).

Limitations of this study

During the identification of gene novelties, we ignored those candidates with homologous genes in jawed vertebrates. Although this avoided repeating efforts on the gene novelties of chordates (6) and deuterostomes (20), we may miss some important genes. Meanwhile, the filter standard of gene novelties may be too high to achieve highly comprehensive results, and many genes may be missed during filtration. Moreover, the lack of animal model experiments impeded our validation of the metabolic functions of some interesting genes.

Materials and Methods

Sample collection

The adults of the Beihai amphioxus, *B. belcheri*, were obtained from the sea near Dianbai District, Maoming City, Guangdong Province, China, cultured at 24–28 °C with air-pumped circulating artificial seawater in Beihai Marine Station of Nanjing University in Beihai City, Guangxi Province, China, and fed with seawater and sea alga (16, 17, 61).

Genomic DNA extraction and sequencing

Genomic DNAs were extracted using the Qiagen Blood & Cell Culture DNA Maxi Kit (Qiagen, Germany). Firstly, adult bodies of the Beihai amphioxus were washed twice with Phosphate-buffered saline (PBS, pH 7.4) and homogenized into powder form by mortar and pestle with liquid nitrogen for keeping low temperature. The homogenized samples were then incubated at 50 °C for proteolysis following the manufacturer's protocol. Genomic DNA was bound to the column, washed, and eluted by washing buffer and elution buffer respectively, and precipitated in 70% ethanol. Finally, air-dry pellets of genomic DNA were dissolved overnight in UltraPure™ DNase/RNase-free distilled water (Thermo Fisher Scientific, USA) overnight at room temperature. The integrity of genomic DNA was determined by electrophoresis in 0.5% agarose gel and analyzed by Agilent 2100 Bioanalyzer (Agilent Technologies, USA), and the quantity was detected by Nanodrop 2000 spectrophotometer (Thermo Fisher Scientific, USA) and Qubit Fluorometer (Thermo Fisher Scientific, USA). The genomic DNA of *B. belcheri* beihai was sequenced by Illumina HiSeq 2000 in 500-bp and 3,000-bp libraries to generate NGS reads and PacBio RS system for third-generation long reads.

De novo genome assembly and annotation

The initial genome assembly of *B. belcheri* beihai was constructed with PacBio long reads using Canu v1.8 (62). Then, scaffolding was performed by SSPACE Basic v2.0 (63) with paired-end Illumina short reads (64) and by SSPACE LongRead v1.1 (65) with PacBio long reads (66). The final sequence polishing was finished by Pilon v1.22 (67) with all Illumina short reads to generate the final genome assembly. The genome assembly was assessed by QUAST v5.0.2 (68) in continuity and the completeness of genome assembly was assessed by BUSCO v3.1.0 (69) with database metazoa_odb9.

With genome assembled, genome annotation was performed. Firstly, repeat masking was performed by *de novo* prediction with RepeatModeler v2.0.1 (70) and masking with RepeatMasker v4.0.8 (RepBase edition 20181026) (71). In the *de novo* prediction with RepeatModeler v2.0.1 (70), RECON v1.05 (72) and RepeatScout v1.0.6 (73) were used for the prediction of repeat family in the genome. Then, genome annotation was performed by Maker pipeline v2.31 (74). In the Maker pipeline, transcriptome assemblies and homologous proteins were used as evidence for alignment by Exonerate v2.4.0 (75), whilst gene prediction was completed by SNAP (lib v2017-03-01) (76), GeneMark v4.38 (77) and Augustus v3.3.1 (78). The quality of genome annotation was assessed by BUSCO v3.1.0 (69) with database metazoa_odb9.

Data collection and access

Besides of the *de novo* assembled and annotated genome of *B. belcheri* beihai, genome assemblies and annotations of other three amphioxi were downloaded from NCBI database (NCBI Assembly accession GCF_001625305.1 for *B. belcheri* xiamen, GCF_000003815.2 for *B. floridae* and GCA_900088365.1 for *B. lanceolatum*). The genome sequencing data, assembly, and annotation of *B. belcheri* beihai have been uploaded to NCBI database under the BioProject accession: PRJNA804338. The transcriptome data of *B. belcheri* beihai were from the BioProject accession: PRJNA310680. The genomes of hybrid amphioxi were not analyzed in this study, such as the haplotype genome of *B. japonicum* (NCBI Assembly accession GCA_013266295.2). Moreover, the recently available genome of *B. lanceolatum* (NCBI Assembly accession GCA_927797965.1) has a high assembly quality, but no transcriptome data could be found. Therefore, only the four amphioxus genomes were used in this study.

Phylogenetic analysis of deuterostomes

The genome sequences of deuterostome species were phylogenetically analyzed with those of other species. The overlapped single and complete BUSCO protein sequences were extracted by BUSCO v3.1.0 (69) with database metazoa_odb9, then aligned by MAFFT (79) and edited in Gblocks (80) with the option '-t=p' to generate sequence alignment of conserved amino-acid residues. Finally, the sequence alignment was used to construct the phylogenetic tree in maximum likelihood algorithm and 100 bootstrap replicates by RAxML v8.2.12 (81) with the options '-m PROTCATWAG -f a -# 100'. The phylogenetic tree was edited by the online tool Interactive Tree of Life (iTOL) (82).

The construct a ultrametric time tree, the alignment of codon sequences of conserved amino-acid residues were extracted and used for estimating divergence time by MCMCTree program in PAML v4.9j (83). The ultrametric time tree was final edited by the online tool Interactive Tree of Life (iTOL) (82).

Phylogenomic orthology analysis

Phylogenomic orthology and gene gain/loss analysis (84) was performed among protein sequences in 21 annotated genomes by OrthoFinder v2.5.4 (85, 86) and CAFÉ v4.2 (87) based on the ultrametric time tree. All the protein sequences were assigned into orthogroups according to the protein similarities on sequence level. These orthogroups were also considered as gene families. Specific orthogroups were summarized and further analyzed. Functional clusters of orthologous groups (COGs) of genes in specific orthogroups were performed by eggnog-mapper v2.1.5 (88, 89).

Identification of gene novelties

To identify gene novelties in amphioxus, we compared the best hits of all the annotated proteins of *B. belcheri* beihai to those from different taxonomic groups in UniRef50 database (UniProt, last updated on Dec. 15th, 2020) (90) as follows. Although amphioxus specific orthogroups have been extracted, we input all the annotated proteins of *B. belcheri* beihai, to avoid missing important gene novelties.

1) In UniRef50 database, collecting the proteins of different taxonomic groups according to the

taxon identifiers, namely non-deuterostome metazoa, bacteria, fungi, plants, and other eukaryotes (excluding fungi, plants, and metazoa), and building BLAST databases.

2) Searching the best hits of 44,745 annotated proteins of *B. belcheri* beihai in the five BLAST databases built in Step 1, using BLASTP v2.9.0 (91) with the options '-evalue 1e-6 -max_hsps 1 -max_target_seqs 1 -outfmt 6'.

3) The best hits in different taxonomic groups were compared and identified two groups of candidates of gene novelties. Firstly, when one protein had hit in the databases of bacteria, fungi, plants, or other eukaryotes, but not that of non-deuterostome metazoan, this gene was identified as group I candidates of gene novelties. Secondly, when the gene novelty index of a protein was calculated as dividing the bit score of the best hit in the databases of bacteria, fungi, plants, and other eukaryotes, by that of non-deuterostome metazoan; if the index ≥ 1.5, this gene was identified as group II candidates.

4) Collecting protein sequences of two groups of candidates, performing BLASTP online search in the NCBI non-redundant protein (NR) database (last updated on Nov. 1st, 2021), and then manually checking the top 100 hits. Except those from Amphioxiformes (including all amphioxi), if the best hits were from unusual taxonomic groups like bacteria, fungi, plants, and other eukaryotes (excluding fungi, plants, and metazoa), this gene was confirmed as an amphioxus specific gene; if some of the best hits were from other deuterostome groups except jawed vertebrates, this gene was considered as a novel gene conserved with other deuterostomes. Additionally, only the DBH and

ADH genes had homologs from the soft coral *Dendronephthya gigantea* and the sponge *Amphimedon queenslandica*, respectively (Table 1).

Therefore, the dominantly abundant and well annotated proteins of jawed vertebrates were used to filter out the candidate genes possibly introduced at stem chordates and deuterostomes, to avoid repeated study in the previous report on in stem chordates (6) and deuterostomes (20). The absence of the candidate genes in jawed vertebrates was double confirmed by searching using TBLASTN in the genomic DNA sequences and BLASTP in the annotated proteins at E-value cutoff of 1e-6. Those candidates supported by poor hits (e.g., low bit scores caused by repeated domains) were discarded either.

5) Once a candidate gene was confirmed, homologous genes of four amphioxi were searched by TBLASTN in the genomic DNA sequences and BLASTP in the annotated proteins at E-value cutoff of 1e-6. After again checking those homologous genes at Step 4 and filtering out false-positive genes; manual checking and curating based on transcriptome data, especially in splice sites (locations of exon–intron boundaries) using the program Integrative Genomics Viewer (IGV) v2.9.4 (92); finally, all homologous genes among four amphioxi (Table 1) were collected and further analyzed.

Most candidates were filtered out and the final 28 gene novelties were summarized in Table 1. Since those involved genes were conserved among up two deuterostome groups (except the VDCC gene among three groups, Table S3) and often had highly similar genes from unusual taxonomic groups (Table 1), we excluded the possibility of assembly contamination and false-positiveness as

implausible. For better reproducibility, the protein sequences coded by these representative genes can be found in Table S4. Some gene novelties were further validated by phylogenetic analysis.

Transcriptome and gene expression level

To explore the expression level of target genes, we collected the transcriptome data of different tissues of *B. lanceolatum* as follows. The SRA accessions of the transcriptome data in the NCBI database are epidermis: SRR6246024; cirri: SRR6246021; gill 1: SRR6246026; gill 2: SRR6246027; gut 1: SRR6246028; gut 2: SRR6246029; hepatic diverticulum 1: SRR6246030; hepatic diverticulum 2: SRR6246031; male gonad 1: SRR6246032; male gonad 2: SRR6246033; female gonad: SRR6246025; muscle 1: SRR6246034; muscle 2: SRR6246035; neural tube 1: SRR6246036; neural tube 2: SRR6246037. The transcript sequences of all the target genes and Glyceraldehyde 3-phosphate dehydrogenase (*GAPDH*) or elongation factor 1 alpha (*EF1A*) of *B. lanceolatum* were built as index and calculated in transcripts per million (TPM) values by the program Salmon v0.12.0 with the options '--type quasi -k 31'. The TPM values of all target genes of *B. lanceolatum* were divided by that of the reference genes, *GAPDH* or *EF1A (93)*, to generate normalized TPM values, and then the log2 (1+ normalized TPM) values were calculated and presented in the heatmaps generated by the program GraphPad Prism v9.0.0.

Comparative analysis of genes

All protein sequences of target genes were searched TBLASTN in the genomic DNA sequences and BLASTP in the annotated proteins of the four amphioxus genomes at E-value cutoff of 1e-6, to

acquire all the homologs. After manual check and validation based on transcriptome data, all protein sequences of target genes were collected in four amphioxi.

The sequence alignment of genes was performed by the online tool Clustal Omega (94). The gene synteny and splice sites (locations of exon–intron boundaries) of target genes were identified and checked in the program Integrative Genomics Viewer (IGV) v2.9.4 (92). The conserved sites were defined with at least one of two flanking amino acids identical among protein sequences. The domain structure of target genes was manually drawn based on the online analysis of BLASTP with the NCBI NR database, InterProScan (95) and Pfam search (96). Gene synteny alignment was manually drawn in PowerPoint 2021, in which the gene distance below 5 Kb defined as tandemly arrayed and below 20 Kb as proximally arrayed.

To further validate these gene novelty events, protein sequences of the target genes were phylogenetically analyzed with the best-hit protein sequences of other taxonomical groups, including bacteria, fungi, other eukaryotes (excluding metazoa, plants, and fungi) and other metazoa (excluding Amphioxiformes), in UniRef50 database and some sequences in the NCBI GenBank database. In addition, protein sequences of the multi-copy genes of amphioxi were phylogenetically analyzed to explore their duplication and divergence. All protein sequences were aligned and drawn into a phylogenetic tree, in which sequence alignment was performed by CLUSTAL W (97) and MUSCLE (98) in MEGA v10.2.2 (99), and all phylogenetic trees were constructed by MEGA v10.2.2 (99) with maximum likelihood (ML) algorithm in the JTT (Jones-Taylor-Thornton) model,

80% site coverage and 100 bootstrap replicates, and then edited by online tool Interactive Tree of Life (iTOL) (100). IQ-TREE 2 (101) was used with options "-st AA -m TEST -bb 1000 -alrt 1000" to test and validate the phylogenetic results. Phylogenetic noise like long branch attraction effect (102) was carefully excluded using the classic SAW method (103).

Data availability

All the data of *B. belcheri* beihai have been deposited under NCBI BioProject accession PRJNA804338.

Acknowledgments and funding sources

This work was supported by the General Research Fund from Research Grants Council (Reference numbers: 464710, 475113, 14119219, 14119420, 14175617) and Health and Medical Research Fund from Food and Health Bureau (Reference numbers: 06171016, 07181266) of Hong Kong.

Competing interests

We have no competing interest to disclose.

Supporting information

Supplementary Note 1–3

Tables S1–4

Figures S1–25

References

1. S. J. Bourlat *et al.*, Deuterostome phylogeny reveals monophyletic chordates and the new phylum Xenoturbellida. *Nature* **444**, 85-88 (2006).
2. F. Delsuc, H. Brinkmann, D. Chourrout, H. Philippe, Tunicates and not cephalochordates are the closest living relatives of vertebrates. *Nature* **439**, 965-968 (2006).
3. S. Bertrand, H. Escriva, Evolutionary crossroads in developmental biology: amphioxus. *Development* **138**, 4819-4830 (2011).
4. È. Benito-Gutiérrez *et al.*, The dorsoanterior brain of adult amphioxus shares similarities in expression profile and neuronal composition with the vertebrate telencephalon. *BMC Biol* **19**, 110 (2021).
5. N. H. Putnam *et al.*, The amphioxus genome and the evolution of the chordate karyotype. *Nature* **453**, 1064-1071 (2008).
6. L. Z. Holland *et al.*, The amphioxus genome illuminates vertebrate origins and cephalochordate biology. *Genome Res* **18**, 1100-1111 (2008).
7. S. Huang *et al.*, Decelerated genome evolution in modern vertebrates revealed by analysis of multiple lancelet genomes. *Nat Commun* **5**, 5896 (2014).
8. R. D. Acemel *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).
9. F. Marlétaz *et al.*, Amphioxus functional genomics and the origins of vertebrate gene regulation. *Nature* **564**, 64-70 (2018).
10. O. Simakov *et al.*, Deeply conserved synteny resolves early events in vertebrate evolution. *Nat Ecol Evol* **4**, 820-830 (2020).
11. M. Brasó-Vives *et al.*, Parallel evolution of amphioxus and vertebrate small-scale gene duplications. *Genome Biology* **23**, 243 (2022).
12. Z. Huang *et al.*, Three amphioxus reference genomes reveal gene and chromosome evolution of chordates. *Proceedings of the National Academy of Sciences* **120**, e2201504120 (2023).
13. K. S. Kassahn, V. T. Dang, S. J. Wilkins, A. C. Perkins, M. A. Ragan, Evolution of gene function and regulatory control after whole-genome duplication: comparative analyses in vertebrates. *Genome Res* **19**, 1404-1418 (2009).
14. A. Louis, H. Roest Crollius, M. Robinson-Rechavi, How much does the amphioxus genome represent the ancestor of chordates? *Brief Funct Genomics* **11**, 89-95 (2012).
15. P. Holland, The dawn of amphioxus molecular biology - a personal perspective. *Int J Dev Biol* **61**, 585-590 (2017).
16. K. Y. Yang *et al.*, Transcriptome analysis of different developmental stages of amphioxus reveals dynamic changes of distinct classes of genes during development. *Scientific Reports* **6**, 23195 (2016).
17. Q. Du *et al.*, Genomic Analysis of Amphioxus Reveals a Wide Range of Fragments Homologous to Viral Sequences. *Viruses* **15**, 909 (2023).
18. N. Satoh, D. Rokhsar, T. Nishikawa, Chordate evolution and the three-phylum system. *Proc Biol Sci* **281**, 20141729 (2014).
19. P. Kapli *et al.*, Lack of support for Deuterostomia prompts reinterpretation of the first Bilateria. *Science Advances* **7**, eabe2741 (2021).
20. O. Simakov *et al.*, Hemichordate genomes and deuterostome origins. *Nature* **527**, 459-465 (2015).
21. S. Bartl, M. Baish, I. L. Weissman, M. Diaz, Did the Molecules of Adaptive Immunity Evolve from the Innate Immune System? *Integrative and Comparative Biology* **43**, 338-346 (2003).

22. D. Yuan *et al.*, Ancestral genetic complexity of arachidonic acid metabolism in Metazoa. *Biochimica et Biophysica Acta (BBA) - Molecular and Cell Biology of Lipids* **1841**, 1272-1284 (2014).
23. D. Yuan, A. Xu, "Chapter 10 - Evolution of Inflammation-Related Lipid Metabolism Pathway and Inflammatory Lipid Signaling in Amphioxus" in Amphioxus Immunity, A. Xu, Ed. (Academic Press, 2016), https://doi.org/10.1016/B978-0-12-849903-0.00010-5, pp. 189-205.
24. P. Janvier, Modern look for ancient lamprey. *Nature* **443**, 921-923 (2006).
25. H.-m. Chen *et al.*, A multifunctional lipoxygenase from Pyropia haitanensis— The cloned and functioned complex eukaryotic algae oxylipin pathway enzyme. *Algal Research* **12**, 316-327 (2015).
26. F. Taschuk, S. Cherry, DEAD-Box Helicases: Sensors, Regulators, and Effectors for Antiviral Defense. *Viruses* **12** (2020).
27. D. Bhattacharya *et al.*, Comparative genomics explains the evolutionary success of reef-forming corals. *eLife* **5**, e13288 (2016).
28. C. D. Trowbridge *et al.*, Extreme oxygen dynamics in shallow water of a fully marine Irish sea lough. *Regional Studies in Marine Science* **11**, 9-16 (2017).
29. J. E. Carvalho, F. Lahaye, M. Schubert, Keeping amphioxus in the laboratory: an update on available husbandry methods. *Int J Dev Biol* **61**, 773-783 (2017).
30. A. Schmitz, M. Gemmel, S. F. Perry, Morphometric partitioning of respiratory surfaces in amphioxus (Branchiostoma lanceolatum Pallas). *Journal of Experimental Biology* **203**, 3381-3390 (2000).
31. D. E. Briggs, A. J. Kear, Decay of Branchiostoma: implications for soft-tissue preservation in conodonts and other primitive chordates. *Lethaia* **26**, 275-287 (1993).
32. T. Kon *et al.*, Phylogenetic position of a whale-fall lancelet (Cephalochordata) inferred from whole mitochondrial genome sequences. *BMC Evolutionary Biology* **7**, 127 (2007).
33. L. Di, A. Balesano, S. Jordan, S. M. Shi, The role of alcohol dehydrogenase in drug metabolism: beyond ethanol oxidation. *The AAPS Journal* **23**, 1-21 (2021).
34. J. Meyer, The evolution of ferredoxins. *Trends in Ecology & Evolution* **3**, 222-226 (1988).
35. T. G. Rossman, M. A. Visalli, E. V. Komissarova, fau and its ubiquitin-like domain (FUBI) transforms human osteogenic sarcoma (HOS) cells to anchorage-independence. *Oncogene* **22**, 1817-1821 (2003).
36. J. E. Nixon *et al.*, Evidence for lateral transfer of genes encoding ferredoxins, nitroreductases, NADH oxidase, and alcohol dehydrogenase 3 from anaerobic prokaryotes to Giardia lamblia and Entamoeba histolytica. *Eukaryot Cell* **1**, 181-190 (2002).
37. K. L. Kavanagh, H. Jörnvall, B. Persson, U. Oppermann, Medium- and short-chain dehydrogenase/reductase gene and protein families : the SDR superfamily: functional and structural diversity within a family of metabolic and regulatory enzymes. *Cell Mol Life Sci* **65**, 3895-3906 (2008).
38. J. Green, M. S. Paget, Bacterial redox sensors. *Nature Reviews Microbiology* **2**, 954-966 (2004).
39. S. Yuan, J. Ruan, S. Huang, S. Chen, A. Xu, Amphioxus as a model for investigating evolution of the vertebrate immune system. *Developmental & Comparative Immunology* **48**, 297-305 (2015).
40. R. A. Berner, J. M. VandenBrooks, P. D. Ward, Oxygen and Evolution. *Science* **316**, 557-558 (2007).
41. R. S. Dietz, J. C. Holden, The breakup of Pangaea. *Scientific American* **223**, 30-41 (1970).
42. P. F. Guerreiro, Juan, "Control of calcium balance in fish" in Fish Osmoregulation. (Science Publisher, USA, 2007), 10.1201/b10994-16.
43. A. C. Dolphin, A short history of voltage-gated calcium channels. *Br J Pharmacol* **147 Suppl 1**, S56-S62 (2006).
44. P. Camacho, L. John, Y. Li, R. M. Paredes, H. L. Roderick, "Calnexin and calreticulin, ER associated modulators of calcium transport in the ER" in Calreticulin. (Springer, 2003), pp. 126-132.

45. L. H. Allen, T. E. Hall, Calcium Metabolism, Intestinal Calcium-binding Protein, and Bone Growth of Rats Fed High Protein Diets. *The Journal of Nutrition* **108**, 967-972 (1978).
46. H. He, Y. Li, S. Wang, Q. Ma, Y. Pan, A High Precision Method for Calcium Determination in Seawater Using Ion Chromatography. *Frontiers in Marine Science* **7** (2020).
47. C. Millot, I. Taupier-Letage, "Circulation in the Mediterranean Sea" in The Mediterranean Sea, A. Saliot, Ed. (Springer Berlin Heidelberg, Berlin, Heidelberg, 2005), 10.1007/b107143, pp. 29-66.
48. F. J. Millero, R. Feistel, D. G. Wright, T. J. McDougall, The composition of Standard Seawater and the definition of the Reference-Composition Salinity Scale. *Deep Sea Research Part I: Oceanographic Research Papers* **55**, 50-72 (2008).
49. C. S. Jones, P. Cessi, Size Matters: Another Reason Why the Atlantic Is Saltier than the Pacific. *Journal of Physical Oceanography* **47**, 2843-2859 (2017).
50. Anonymous (2019) Tidal Flat Planning of Aquaculture Waters in Dianbai District, Maoming City. (The Government of Dianbai District, Maoming City, China).
51. Anonymous (2016) The marine environment of Xiamen City. (The Government of Xiamen City, China).
52. S. D. Meyers, A. J. Linville, M. E. Luther, Alteration of residual circulation due to large-scale infrastructure in a coastal plain estuary. *Estuaries and Coasts* **37**, 493-507 (2014).
53. A. Dańko, R. Schaible, M. J. Dańko, Salinity Effects on Survival and Reproduction of Hydrozoan Eleutheria dichotoma. *Estuaries and Coasts* **43**, 360-374 (2020).
54. S. Huang *et al.*, Genomic analysis of the immune gene repertoire of amphioxus reveals extraordinary innate complexity and diversity. *Genome Res* **18**, 1112-1126 (2008).
55. J. Huang, Horizontal gene transfer in eukaryotes: the weak-link model. *Bioessays* **35**, 868-875 (2013).
56. W. F. Doolittle, You are what you eat: a gene transfer ratchet could account for bacterial genes in eukaryotic nuclear genomes. *Trends in Genetics* **14**, 307-311 (1998).
57. H. U. Riisgård, I. Svane, Filter Feeding in Lancelets (Amphioxus), Branchiostoma lanceolatum. *Invertebrate Biology* **118**, 423-432 (1999).
58. K. J. Peterson, D. J. Eernisse, The phylogeny, evolutionary developmental biology, and paleobiology of the Deuterostomia: 25 years of new techniques, new discoveries, and new ideas. *Organisms Diversity & Evolution* **16**, 401-418 (2016).
59. K. Hori *et al.*, Klebsormidium flaccidum genome reveals primary factors for plant terrestrial adaptation. *Nature Communications* **5**, 3978 (2014).
60. M. Long, E. Betrán, K. Thornton, W. Wang, The origin of new genes: glimpses from the young and old. *Nature Reviews Genetics* **4**, 865-875 (2003).
61. X. Chen *et al.*, Identification and characterization of novel amphioxus microRNAs by Solexa sequencing. *Genome Biology* **10**, R78 (2009).
62. S. Koren *et al.*, Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res* **27**, 722-736 (2017).
63. M. Boetzer, C. V. Henkel, H. J. Jansen, D. Butler, W. Pirovano, Scaffolding pre-assembled contigs using SSPACE. *Bioinformatics* **27**, 578-579 (2011).
64. D. R. Bentley *et al.*, Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**, 53-59 (2008).
65. M. Boetzer, W. Pirovano, SSPACE-LongRead: scaffolding bacterial draft genomes using long read sequence information. *BMC Bioinformatics* **15**, 211 (2014).
66. J. Eid *et al.*, Real-time DNA sequencing from single polymerase molecules. *Science* **323**, 133-138 (2009).
67. B. J. Walker *et al.*, Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly

improvement. *PloS one* **9**, e112963 (2014).
68. A. Gurevich, V. Saveliev, N. Vyahhi, G. Tesler, QUAST: quality assessment tool for genome assemblies. *Bioinformatics* **29**, 1072-1075 (2013).
69. F. A. Simão, R. M. Waterhouse, P. Ioannidis, E. V. Kriventseva, E. M. Zdobnov, BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210-3212 (2015).
70. J. M. Flynn *et al.*, RepeatModeler2 for automated genomic discovery of transposable element families. *Proceedings of the National Academy of Sciences* **117**, 9451-9457 (2020).
71. M. Tarailo-Graovac, N. Chen, Using RepeatMasker to identify repetitive elements in genomic sequences. *Current protocols in bioinformatics* **Chapter 4**, Unit 4.10 (2009).
72. Z. Bao, S. R. Eddy, Automated de novo identification of repeat sequence families in sequenced genomes. *Genome Res* **12**, 1269-1276 (2002).
73. A. L. Price, N. C. Jones, P. A. Pevzner, De novo identification of repeat families in large genomes. *Bioinformatics* **21**, i351-i358 (2005).
74. B. L. Cantarel *et al.*, MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Res* **18**, 188-196 (2008).
75. G. S. C. Slater, E. Birney, Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* **6**, 31 (2005).
76. I. Korf, Gene finding in novel genomes. *BMC Bioinformatics* **5**, 59 (2004).
77. M. Borodovsky, A. Lomsadze, Eukaryotic gene prediction using GeneMark.hmm-E and GeneMark-ES. *Current protocols in bioinformatics* **Chapter 4**, Unit 4.6.1-10 (2011).
78. K. J. Hoff, M. Stanke, Predicting Genes in Single Genomes with AUGUSTUS. *Current protocols in bioinformatics* **65**, e57 (2019).
79. K. Katoh, K. Misawa, K. i. Kuma, T. Miyata, MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research* **30**, 3059-3066 (2002).
80. J. Castresana, Selection of Conserved Blocks from Multiple Alignments for Their Use in Phylogenetic Analysis. *Molecular Biology and Evolution* **17**, 540-552 (2000).
81. A. Stamatakis, RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312-1313 (2014).
82. I. Letunic, P. Bork, Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Research* **49**, W293-W296 (2021).
83. Z. Yang, PAML 4: Phylogenetic Analysis by Maximum Likelihood. *Molecular Biology and Evolution* **24**, 1586-1591 (2007).
84. M. W. Hahn, T. De Bie, J. E. Stajich, C. Nguyen, N. Cristianini, Estimating the tempo and mode of gene family evolution from comparative genomic data. *Genome Res* **15**, 1153-1160 (2005).
85. D. M. Emms, S. Kelly, OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol* **20**, 238 (2019).
86. D. M. Emms, S. Kelly, OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biol* **16**, 157 (2015).
87. T. De Bie, N. Cristianini, J. P. Demuth, M. W. Hahn, CAFE: a computational tool for the study of gene family evolution. *Bioinformatics* **22**, 1269-1271 (2006).
88. C. P. Cantalapiedra, A. Hernández-Plaza, I. Letunic, P. Bork, J. Huerta-Cepas, eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale. *Molecular Biology and Evolution* **38**, 5825-5829 (2021).
89. J. Huerta-Cepas *et al.*, eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology

resource based on 5090 organisms and 2502 viruses. *Nucleic acids research* **47**, D309-d314 (2019).

90. B. E. Suzek *et al.*, UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics* **31**, 926-932 (2015).

91. S. McGinnis, T. L. Madden, BLAST: at the core of a powerful and diverse set of sequence analysis tools. *Nucleic Acids Research* **32**, W20-W25 (2004).

92. H. Thorvaldsdóttir, J. T. Robinson, J. P. Mesirov, Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief Bioinform* **14**, 178-192 (2013).

93. Q.-L. Zhang *et al.*, Selection of reliable reference genes for normalization of quantitative RT-PCR from different developmental stages and tissues in amphioxus. *Scientific Reports* **6**, 37549 (2016).

94. F. Sievers, D. G. Higgins, Clustal Omega, accurate alignment of very large numbers of sequences. *Methods Mol Biol* **1079**, 105-116 (2014).

95. P. Jones *et al.*, InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236-1240 (2014).

96. J. Mistry *et al.*, Pfam: The protein families database in 2021. *Nucleic Acids Research* **49**, D412-D419 (2021).

97. J. D. Thompson, D. G. Higgins, T. J. Gibson, CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic acids research* **22**, 4673-4680 (1994).

98. R. C. Edgar, MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* **5**, 113 (2004).

99. S. Kumar, G. Stecher, M. Li, C. Knyaz, K. Tamura, MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. *Mol Biol Evol* **35**, 1547-1549 (2018).

100. I. Letunic, P. Bork, Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. *Bioinformatics* **23**, 127-128 (2006).

101. B. Q. Minh *et al.*, IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Molecular Biology and Evolution* **37**, 1530-1534 (2020).

102. J. P. Townsend, Z. Su, Y. I. Tekle, Phylogenetic Signal and Noise: Predicting the Power of a Data Set to Resolve Phylogeny. *Systematic Biology* **61**, 835-835 (2012).

103. M. E. Siddall, M. F. Whiting, Long-branch abstractions. *Cladistics* **15**, 9-24 (1999).

Table 1. A range of gene novelties identified in amphioxus

Functional category	Abbrev.	Representative gene ID①	Amino acid length	Best hit, metazoan②	Best hit③			Copy number④			
					E-value	Identity	Taxonomy	*Bbb*	*Bbx*	*Bf*	*Bl*
Innate immunity	LOX*	Bbb_152860.1⑤	696	3E-62	7E-168	39.6%	Bacteria, cyanobacteria/proteobacteria	6	5	9	7
	DBH*	Bbb_034850.1	1253	2E-141	0.0	39.1%	Bacteria, cyanobacteria/proteobacteria	1	1	1	1
	RBL	Bbb_161350.1	282	8E-17	4E-23	29.7%	Bacteria, Terrabacteria group	2	1	3	1
	TIR-a	Bbb_152643.1⑥	126	1E-4	4E-21	42.5%	Bacteria, *Solibacter usitatus*	4	5	2	2
	TIR-b	Bbb_258280.1⑥	95	N. A.	1E-10	43.3%	Bacteria, cyanobacteria	11	14	18	9
Anaerobic adaptation	ADH*	Bbb_340330.1	355	3E-78	1E-129	51.3%	Bacteria, Gammaproteobacteria	1	1	1	1
	FDX	Bbb_233460.1	171	1E-12	6E-46	74.5%	Plants, *Micractinium conductrix*	1	1	2	1
	SDR_a4	Bbb_177610.1	263	N. A.	5E-35	34.6%	Bacteria, *Methylomarinum vadi*	3	3	3	2
Calcium balance	VDCC*	Bbb_084320.1	524	3E-36	9E-83	37.2%	Other eukaryotes, *Salpingoeca rosetta*	1	1	1	1
	EF-CaBP	Bbb_077420.1	248	2E-6	5E-11	27.4%	Bacteria, Gammaproteobacteria	19	25	17	12
	CML26	Bbb_165620.1	153	4E-6	3E-11	36.2%	Bacteria, Gammaproteobacteria	1	1	2	2
Amino acid metabolism	SCP24*	Bbb_236620.1	474	1E-83	8E-136	46.6%	Plants, *Pycnococcus provasolii*	1	1	1	1
	GSA*	Bbb_006430.1	438	2E-32	2E-113	53.6%	Bacteria, Deltaproteobacteria	2	1	1	1
	GNAT	Bbb_266880.1	325	3E-26	7E-60	39.3%	Bacteria, Alphaproteobacteria	3	5	5	5
	DDC*	Bbb_001110.1	586	4E-49	2E-141	43.4%	Bacteria, *Bacteroidetes bacterium*	9	8	16	7
	DOD	Bbb_303310.1	173	3E-8	3E-36	42.5%	Other eukaryotes, *Tetrahymena*	2	1	1	1
	FR	Bbb_207740.1	241	9E-24	3E-72	54.0%	Other eukaryotes, Chlorarachniophyte	2	3	5	4
	DHFR	Bbb_115460.1	186	2E-29	1E-43	44.0%	Bacteria, *Ignavibacteriae bacterium*	1	1	1	1
Other metabolic genes	GH27⑦	Bbb_359430.1	419	8E-32	6e-82	40.1%	Bacteria, Terrabacteria group	1	1	1	1
	MCO*	Bbb_246040.1	640	4E-10	7E-67	30.7%	Plants, *Tetraselmis astigmatica*	5	3	8	4
	anmK	Bbb_029670.1	292	7E-29	2E-51	52.2%	Bacteria, cyanobacteria	1	1	1	1
Others	TPR	Bbb_379800.1	884	1E-86	0.0	37.6%	Other eukaryotes, Amoeba	1	1	1	1
	ApoLp	Bbb_173200.1	1205	N. A.	1E-158	36.8%	Bacteria, Gamma/Deltaproteobacteria	3	2	3	3
	HintN	Bbb_173190.1	789	2E-5	1E-118	36.3%	Bacteria, Gamma/Deltaproteobacteria	3	1	2	3
	CytP460	Bbb_256630.1	435	N. A.	2E-94	41.3%	Fungi, *Punctularia strigosozonata*	5	3	8	2
	UF-TM1	Bbb_158830.1	341	N. A.	6E-39	35.0%	Plants, *Klebsormidium nitens*	6	6	7	12
	UF-TM2*	Bbb_202120.1	1108	8E-7	4E-34	23.9%	Fungi, *Aspergillus thermomutatus*	1	1	1	1
	UF-TM3*	Bbb_320160.1	268	N. A.	1E-8	29.8%	Bacteria, *Streptomyces*	12	10	6	7

① To describe the gene features, a gene of *B. belcheri* beihai was selected as the representative gene. The protein sequences are listed in Table S4.

② The best hits of metazoan were of matches to genes of metazoa (excluding Amphioxiformes) in UniRef50 database. N.A., Not applicable, because no best hit has E-value smaller than the cutoff 0.001. In some cases, the gene has homologs in other metazoan species and marked in star (*); then these metazoan species were excluded.

③ The best hits in UniRef50 database were listed and displayed in detail. In many cases, the best matches were from a group of microbes, but not exactly a species, so the group name like cyanobacteria was listed.

④ Since many genes underwent gene duplications, copy number of four amphioxi were compared. Bbb, *B. belcheri* beihai; Bbx, *B. belcheri* xiamen; Bf, *B. floridae*; Bl, *B. lanceolatum*.

⑤ For the LOX gene Bbb_152860.1 contains many domains, only the partial 696-aa sequence containing its LOX domain was analyzed.

⑥ For the TIR-a and TIR-b, only their TIR domains were analyzed, which covered 35-160 of Bbb_152640.3 and 107-201 of Bbb_258280.1, respectively.

⑦ For the GH27 gene Bbb_359430.1, the best hit in NR database (6E-82 from Terrabacteria) has a much lower E-value than the best hit in UniRef50 database (6E-74 from plants), therefore we chose to present the best hit in NR database.

Abbreviation:

Innate immunity

LOX* lipoxygenase; homologs in the sea lamprey *Petromyzon marinus* and the hagfish *Eptatretus burger*;
DBH* DEAD/DEAH-box helicase; homologs in the soft coral *Dendronephthya gigantea* and hemichordates;
RBL ricin B-like lectin;
TIR-a Toll/interleukin-1 receptor domain a;
TIR-b Toll/interleukin-1 receptor domain b;

Anaerobic adaptation

ADH* alcohol dehydrogenase; homolog in the sponge *Amphimedon queenslandica*;
FDX ferredoxin;
SDR_a4 atypical SDR subgroup 4, containing a glycine-rich NAD(P)-binding motif;

(to be continued)

Calcium balance

VDCC *	voltage-dependent calcium channel (subunit alpha-2/delta-4-like); homologs in hemichordates and echinoderms;
EF-CaBP	EF-hand calcium-binding protein;
CML26	calmodulin-like protein 26;

Amino acid metabolism

SCP24 *	serine carboxypeptidase 24; homologs in tunicates;
GSA *	glutamate-1-semialdehyde aminotransferase; homologs in echinoderms;
GNAT	GNAT family N-acetyltransferase;
DDC *	DOPA decarboxylase; homologs in hemichordates;
DOD	DOPA 4,5-dioxygenase;
FR	folate receptor;
DHFR	dihydrofolate reductase;

Other metabolic genes

GH27	glycoside hydrolase family 27;
MCO *	multicopper oxidase; homologs in tunicates;
anmK	anhydro-N-acetylmuramic acid kinase;

Others

TPR	tetratricopeptide repeat protein;
ApoLp	apolipophorin III-like domain-containing protein;
HintN	Hedgehog/Intein domain-containing protein;
CytP460	cytochrome P460 family like protein;
UF-TM1	transmembrane protein 1 of unknown function;
UF-TM2 *	transmembrane protein 2 of unknown function; homologs in hemichordates;
UF-TM3 *	transmembrane protein 3 of unknown function; homologs in tunicates.

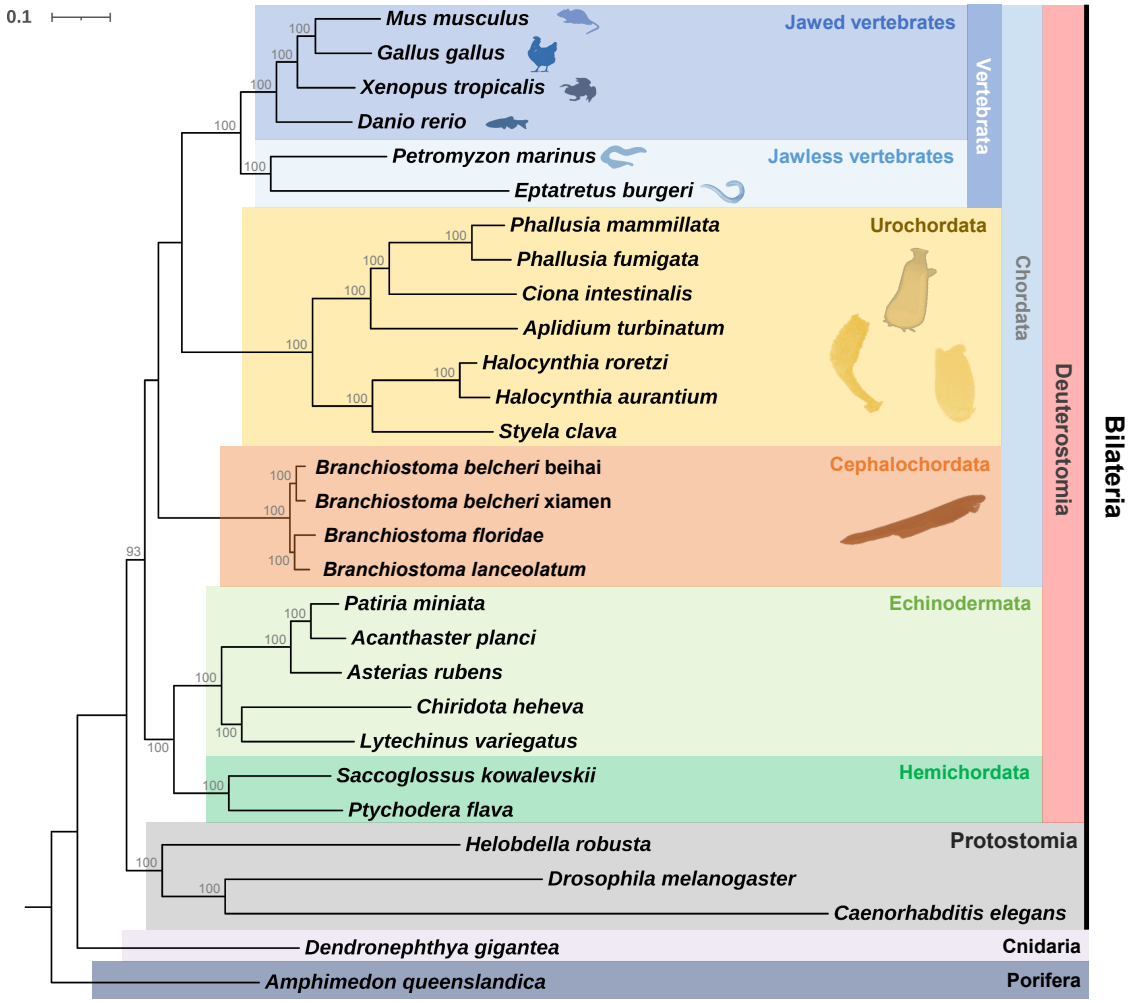

Fig. 1. Phylogenetic analysis of major deuterostome groups
The completeness of 29 genome assemblies (Table S2) was assessed by Benchmarking Universal Single-Copy Orthologs (BUSCO) using database metazoa_odb9. 112 single-copy BUSCOs overlapped by 29 genome assemblies were extracted in protein sequences and aligned by MAFFT for the phylogenetic analysis by RAxML with 100 bootstrap replicates (Supplementary Note 2.2).

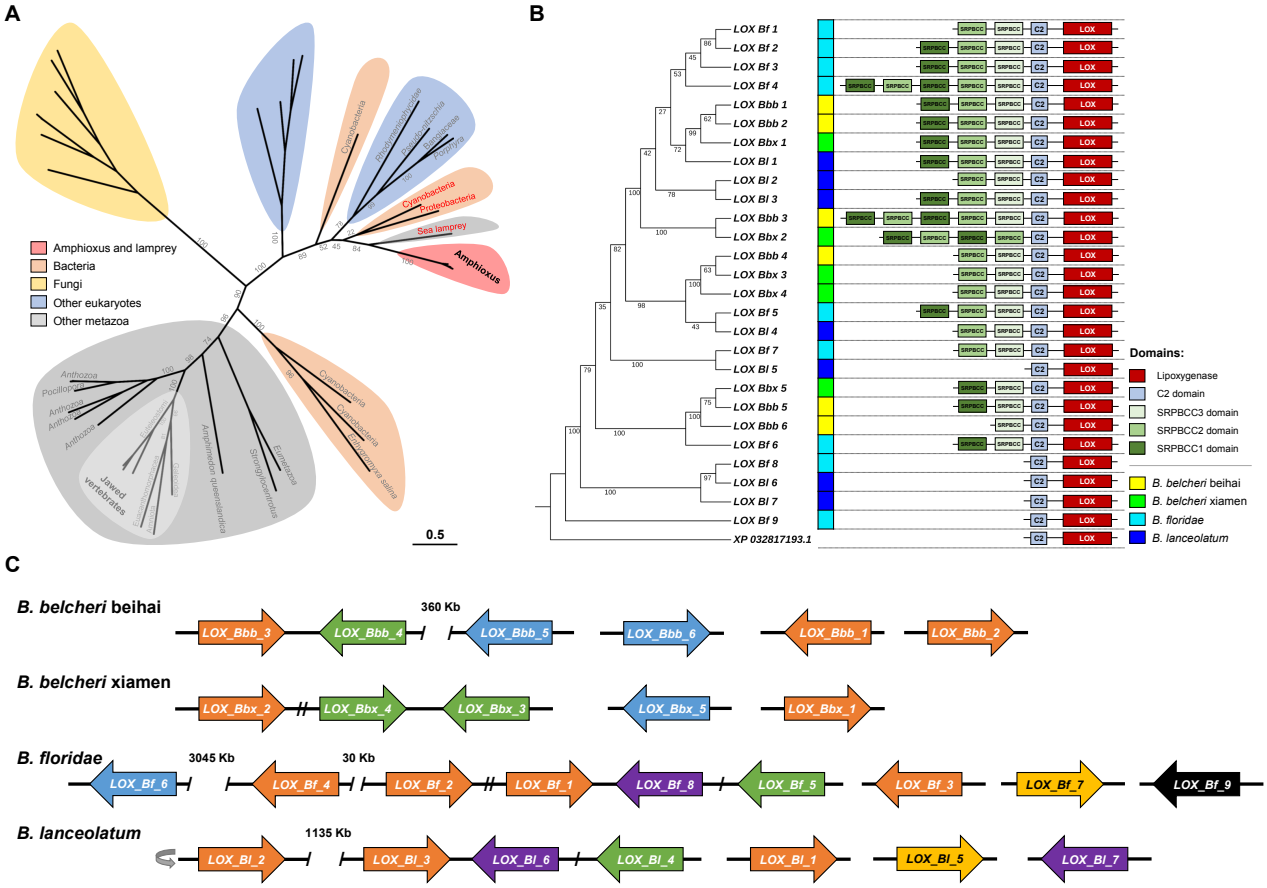

Fig. 2. A group of lipoxygenases (LOXs) in amphioxi
(A) The amphioxus LOX genes and their closest genes in the UniRef50 database from other taxonomic categories, including bacteria, fungi, other eukaryotes (excluding metazoa, plants, and fungi) and other metazoa (excluding Amphioxiformes) were collected for phylogenetic analysis. **(B)** Phylogenetic analysis of all LOX genes of four amphioxi and their domain architecture. The LOX gene of the sea lamprey *Petromyzon marinus* (GenBank accession: XP_032817193.1) was used as outgroup. C2, C2-like domain (universal calcium-binding domain); SRPBCC, PYR/PYL/RCAR family domain in SRPBCC superfamily. **(C)** Gene synteny alignment of the LOX genes of four amphioxi. The distance precision was set as 5 Kb, e.g., 30 Kb represents for 30-35 Kb. If not noted, a slash (/) between two genes represents 5 Kb, e.g., one slash represents 5-10 Kb. Except the outlier (*LOX_Bf_9*), LOXs located on clades supported by high bootstrap value (>95) in the phylogenetic tree were painted in the same color. The suffixes stand for four amphioxi, *Bbb* for *B. belcheri* beihai, *Bbx* for *B. beicheri* xiamen, *Bf* for *B. floridae* and *Bl* for *B. lanceolatum*. The grey turnover arrow means after reverse complement.

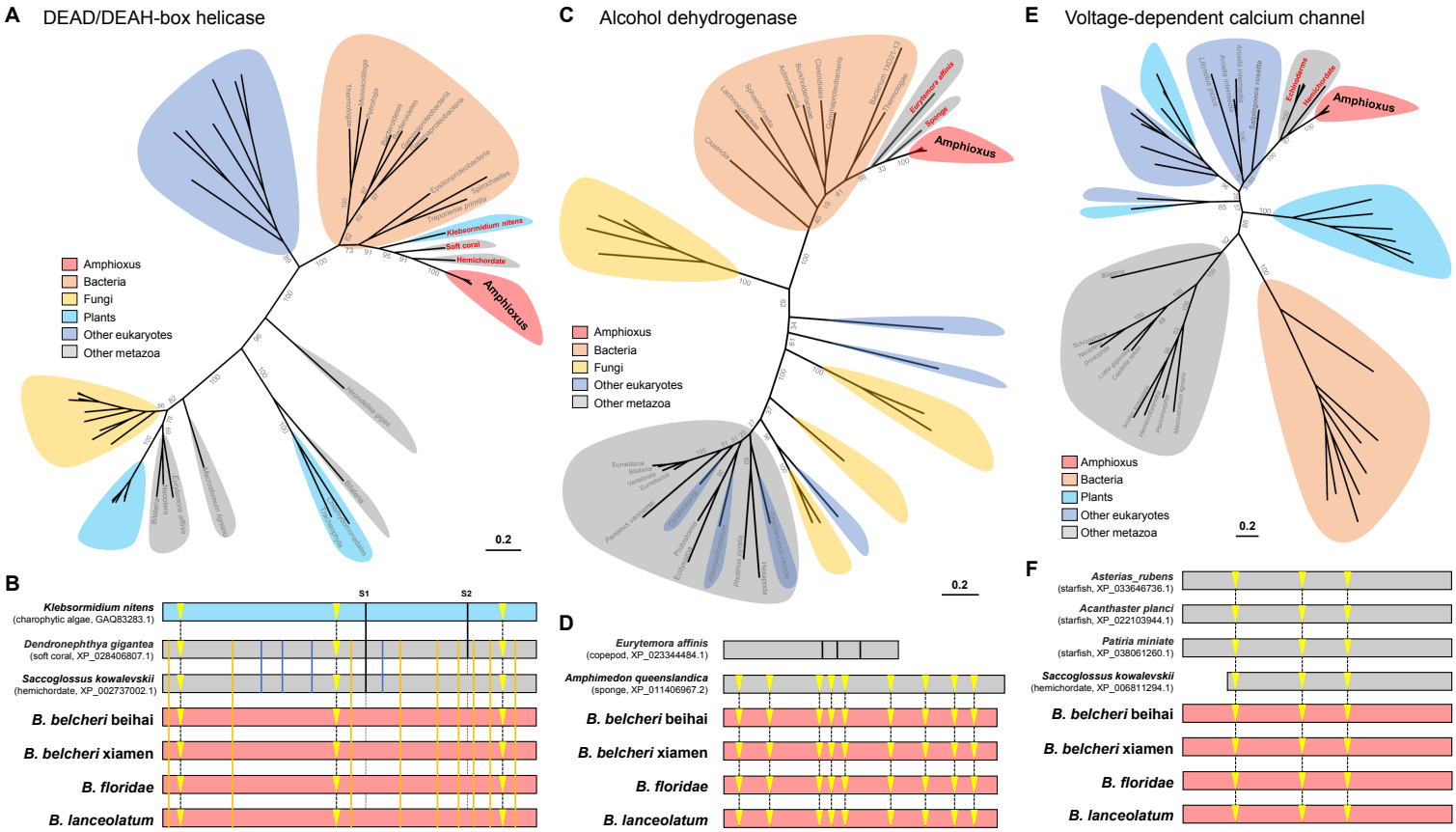

Fig. 3. Comparative analysis of a DEAD/DEAH-box helicase (DBH), an alcohol dehydrogenase (ADH) and a voltage-dependent calcium channel (VDCC)
(A) The DBHs of amphioxi and their closest genes in the UniRef50 database from other taxonomic categories, including bacteria, fungi, other eukaryotes (excluding metazoa, plants, and fungi) and other metazoa (excluding Amphioxiformes) were collected for phylogenetic analysis. **(B)** Alignment of splice sites of the DBH genes of amphioxi, the charophytic algae *K. nitens* (GenBank accession: GAQ83283.1), the soft coral *Dendronephthya gigantea* (GenBank accession: XP_028406807.1) and the hemichordate *Saccoglossus kowalevskii* (GenBank accession: XP_002737002.1). Conserved splice sites were marked by vertical lines (Fig. S8), and three conserved splice sites of all seven sequences were highlighted in yellow triangles. Two unusual splice sites were labeled as S1 and S2. **(C)** The ADHs of amphioxi and their closest genes in the UniRef50 database from other taxonomic categories were collected for phylogenetic analysis. **(D)** Alignment of splice sites of the ADH genes of amphioxi, the sponge *Amphimedon queenslandica* (GenBank accession: XP_011406967.2) and the copepod *Eurytemora affinis* (UniProt accession: UPI000C762174) (Fig. S11). **(E)** The VDCCs of amphioxi and their closest genes in the UniRef50 database from other taxonomic categories were collected for phylogenetic analysis. **(F)** Alignment of splice sites of the VDCC genes of amphioxi, the hemichordate *Saccoglossus kowalevskii* (GenBank accession: XP_006811294.1) and three starfishes (GenBank accessions: XP_033646736.1, XP_022103944.1 and XP_038061260.1) (Fig. S13).

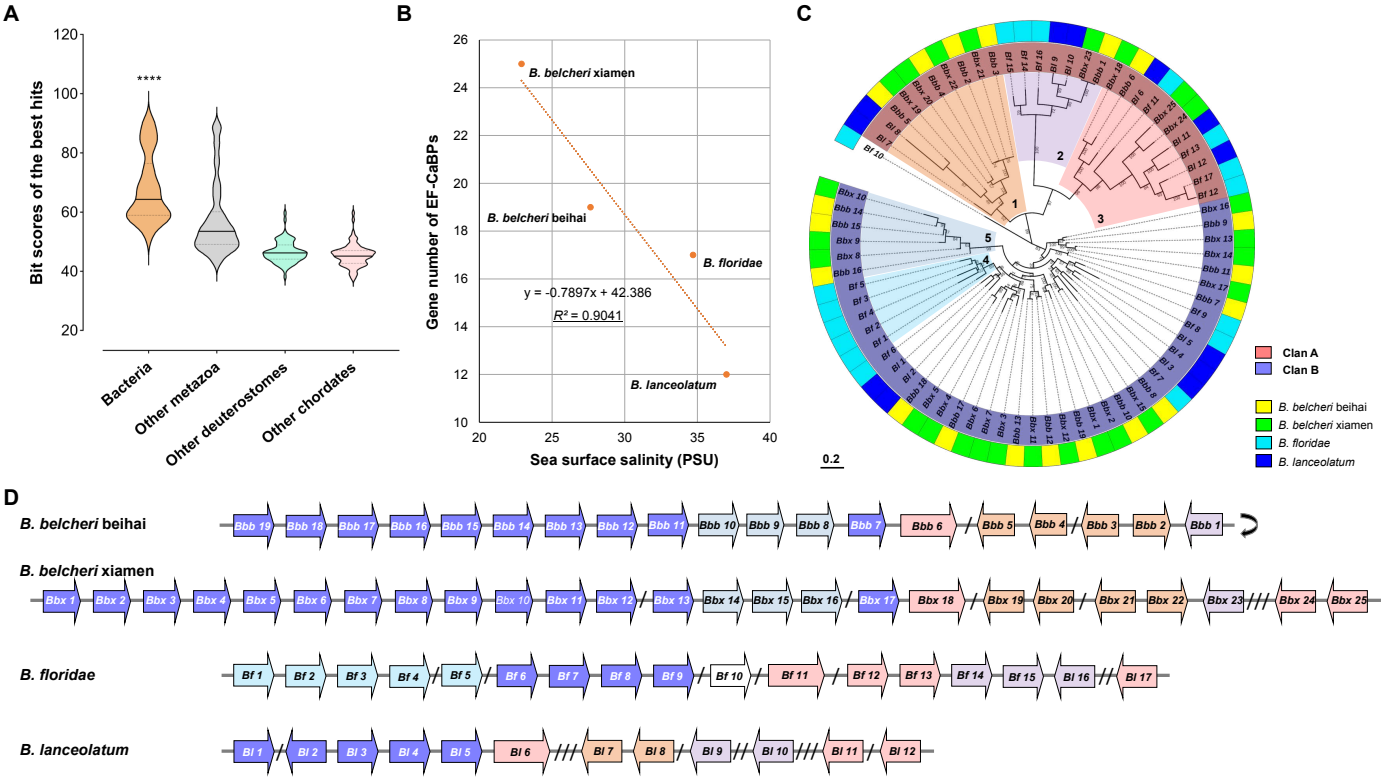

Fig. 4. A group of proximally arrayed EF-hand calcium-binding proteins (EF-CaBPs) **(A)** Of a group of EF-CaBP genes of amphioxus, bit scores of the best hits in the UniRef50 database to other metazoa, other deuterostomes and other chordates (excluding Amphioxiformes respectively) were compared with those to bacteria in violin plot with quartile lines. **** p <0.0001, when comparing the best hits to bacteria with other three groups, respectively. **(B)** Linear correlation of the gene number and the sea surface salinity of collection locations of four amphioxi. Linear regression was performed in $R^2 = 0.9041$. **(C)** Phylogenetic analysis of these EF-CaBPs of four amphioxi. Two large clans, Clan A and B, were highlighted and five specific clusters were marked. **(D)** Gene synteny alignment of the proximally arrayed EF-CaBP genes. A slash (/) represents 5 Kb, e.g., one slash represents 5-10 Kb, and three slashes represent 15-20 Kb, while none represents below 5 Kb. The grey turnover arrow means after reverse complement.