Parallel evolution of amphioxus and vertebrate small-scale gene duplications

Marina Brasó-Vives[1,2], Ferdinand Marlétaz[3], Amina Echchiki[1,2], Federica Mantica[4], Rafael D. Acemel[5], José L. Gómez-Skarmeta[5,#], Lorlane L. Targa[6,7], Pierre Pontarotti[6,7,8], Juan J. Tena[5], Ignacio Maeso[5,9], Hector Escriva[10], Manuel Irimia[4,11,12], Marc Robinson-Rechavi[1,2*]

1. Department of Ecology and Evolution, University of Lausanne, Lausanne, Switzerland.
2. Swiss Institute of Bioinformatics (SIB), Lausanne, Switzerland.
3. Department of Genetics, Evolution and Environment (GEE), University College London, London, UK.
4. Centre for Genomic Regulation (CRG), Barcelona Institute of Science and Technology, Barcelona, Spain.
5. Andalusian Centre for Developmental Biology (CABD), CSIC-Pablo Olavide University, Sevilla, Spain.
6. IRD, APHM, MEPHI, Aix Marseille Université, Marseille, France.
7. IHU-Méditerranée Infection, Marseille, France.
8. CNRS, France.
9. Department of Genetics, Microbiology and Statistics, University of Barcelona, Barcelona, Spain.
10. Oceanological Observatory of Banyuls-sur-Mer (OOB), CNRS-Sorbonne University, Banyuls-sur-Mer, France.
11. Pompeu Fabra University (UPF), Barcelona, Spain.
12. ICREA, Barcelona, Spain.

* Corresponding author.

# Deceased.

Abstract

Background: Amphioxus are non-vertebrate chordates characterized by a slow morphological and molecular evolution. They share the basic chordate body-plan and genome organization with vertebrates but lack their "2R" whole-genome duplications and their developmental complexity. For these reasons, amphioxus are frequently used as an outgroup to study vertebrate genome evolution and Evo-Devo. Aside from whole-genome duplications, genes continuously duplicate on a smaller scale. Small-scale duplicated genes can be found in both amphioxus and vertebrate genomes, while only the vertebrate genome has duplicated genes product of their 2R whole-genome duplications. Here, we explore the history of small-scale gene duplications in the amphioxus lineage and compare it to small- and large-scale gene duplication history in vertebrates.

Results: We present a study of the European amphioxus (*Branchiostoma lanceolatum*) gene duplications thanks to a new, high-quality genome reference. We find that, despite its overall slow molecular evolution, the amphioxus lineage has had a history of small-scale duplications similar to the one observed in vertebrates. We find parallel gene duplication profiles between amphioxus and vertebrates, and conserved functional constraints in gene duplication. Moreover, amphioxus gene duplicates show levels of expression and patterns of functional specialization similar to the ones observed in vertebrate duplicated genes. We also find strong conservation of gene synteny between two distant amphioxus species, *B. lanceolatum* and *B. floridae*, with two major chromosomal rearrangements.

Conclusions: In contrast to their slower molecular and morphological evolution, amphioxus small-scale gene duplication history resembles that of the vertebrate lineage both in quantitative and in functional terms.

Keywords

Amphioxus
Branchiostoma lanceolatum
Gene duplication
Small-scale duplication
Genome assembly
Vertebrate
Ohnolog
Comparative genomics

Background

Amphioxus, the extant members of the subphylum Cephalochordata, are small marine non-vertebrate chordates key to the study of chordate and early vertebrate evolution [1–3]. Cephalochordates, together with tunicates, are the only two known extant non-vertebrate chordate lineages. In contrast to tunicates, which present very derived genomes and adult morphology, the amphioxus lineage has had slower molecular evolution and presents ancestral phenotypic characters both during development and in adult individuals [4–7]. For these reasons, amphioxus have been historically used as outgroup models to study vertebrate anatomy, embryonic development, and genomics [6]. Comparative genomics of amphioxus and other chordates has been key to understanding vertebrate genome and transcriptome origins [6–10] and continues to shed light on chordate ancestor genome structure [11].

Among other findings, the study of amphioxus genomes validated the "2R" hypothesis proposing two rounds of whole-genome duplication (WGD) in early vertebrate evolution [6,9,12]. Duplication, both whole-genome and small-scale, is an important contributor to genome evolution [13–17]. Vertebrate genomes notably include both small-scale gene duplicates and ohnologs derived from WGDs (2R and fish- or amphibian-specific). In the absence of WGDs in the amphioxus lineage, the study of gene duplications in both amphioxus and vertebrates is key to understanding the differences in gene duplication history between these lineages, and the role of duplication in both lineages' evolution. Notably, to understand the role of duplication in vertebrates, we need to characterize their role in the reference outgroup lineage.

Studying gene duplication in amphioxus is complicated by their very high heterozygosity. They present some of the highest measured heterozygosity rates in animals, exceeded by few others such as the purple sea urchin *Strongylocentrotus purpuratus* [6,18,19]. High heterozygosity can result in alternative haplotypes assembled as separate loci [20,21], and in difficulty distinguishing between alleles and recent paralogs. Duplicated sequences must neither be collapsed into a single region nor placed as alternative haplotypes [22,23]. Thus, assembling the genome with long-read and long-distance data, as well as using haplotype collapsing methods, is critical to the study of gene duplications in amphioxus [11].

The European amphioxus (*Branchiostoma lanceolatum*) is one of the best studied species of amphioxus. It has an ecological range expanding from the northeastern Atlantic Ocean to the Mediterranean Sea [1] and a diploid karyotype of 19 pairs of chromosomes [24]. For this study, we built the first long-read and proximity-ligation based high-quality genome assembly, and we improved gene annotation of this species. Thanks to the reliability provided by a good-quality genome reference and annotation, we present here an analysis of the genome evolution of the lineage leading to the European amphioxus with special focus on amphioxus gene duplication history compared to that of vertebrates. Surprisingly, we find that amphioxus has a similar small-scale gene duplication rate to vertebrates, with parallel functional patterns of duplication between the two lineages.

Results

High-quality *B. lanceolatum* genome assembly and annotation

We constructed BraLan3, a high-quality genome assembly combining PacBio sequencing reads with chromosome conformation Hi-C data of the European amphioxus, *Branchiostoma lanceolatum* (see Methods). We re-annotated and validated protein-coding genes in this new genome reference (see Methods). The high-quality of BraLan3 can be seen in basic genome assembly quality statistics (Figure 1, Table 1). It has an N50 of 23,752,511 bp for a total length of 474,791,770 bp. Moreover, 96.78% of the BraLan3 sequence is found within the 19 chromosomes of *B. lanceolatum*'s haploid karyotype [24]. The gene annotation of BraLan3 contains 27,102 genes of which 96.97% (26,282 genes) are in chromosomes and 97.66% (26,468 genes) are supported by strong evidence (see Methods). This new annotation has a BUSCO completeness score of 97.6%. These numbers represent a strong improvement in genome assembly and gene annotation quality for *B. lanceolatum* relative to the previous short-read based genome [8]. The quality of BraLan3 resembles that of vertebrate model species and is as good or better than those of other amphioxus species genomes in all statistics (Figure 1, Table 1).

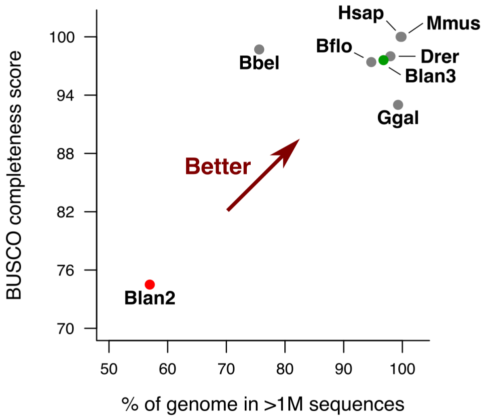

Figure1. BraLan3 genome assembly and annotation quality comparison. Percentage of each genome assembly sequence in chromosomal-sized sequences (>1M nucleotides) versus BUSCO completeness score of each genome annotation. Drer, Ggal, Mmus, Hsap, Bflo and Bbel correspond to the genome assemblies for zebrafish (GRCz11), chicken (GRCg6a), mouse (GRCm39), human (GRCh38), the American amphioxus (*B. floridae*) [9] and the Asian amphioxus (*B. belcheri*) [7], respectively. Blan2 corresponds to the previously available genome reference for the European amphioxus [8] and Blan3 corresponds to BraLan3, the genome reference for the European amphioxus presented in this study. BUSCO completeness score was performed with the metazoan gene universe in all cases. Table 1 contains this figure's numbers and other genome statistics.

	Total length (Mbp)	N50 (Mbp)	L50	# sequences >1Mbp	% length in >1Mbp	# gaps in >1Mbp	BUSCO completeness
BraLan3	474.79	23.75	8	19	96.78%	1523	97.60%
BraLan2	495.35	1.30	79	108	56.95%	54787	74.50%
B. floridae	513.46	25.44	9	20	94.70%	20731	97.40%
B. belcheri	426.12	2.33	52	121	75.60%	12649	98.70%
Zebrafish (GRCz11)	1,373.45	54.30	11	25	97.96%	17943	98.00%
Chicken (GRCg6a)	1,065.35	91.32	4	37	99.27%	866	93.00%
Human (GRCh38)	3,099.73	145.14	9	25	99.71%	997	100.00%
Mouse (GRCm39)	2,728.21	130.53	9	21	99.86%	268	100.00%

Table 1. BraLan3 assembly and annotation quality comparison. All statistics were calculated with the chromosome-level assembly when available or, alternatively, with the scaffold-level assembly; scaffold vs. chromosome-level especially impacts N50 and L50. BUSCO completeness score was performed with the metazoan gene universe in all cases. BraLan3 corresponds to the assembly presented in this work, BraLan2 refers to the previously available *B. lanceolatum* assembly [8].

As an example of the completeness of the assembly, we were able to identify for the first time in this species genome assembly three *ProtoRAG* transposon sequences. The recombination-activating genes (RAG1 and RAG2) in jawed vertebrates, essential for V(D)J recombination in immunoglobulin genes, have their origin in a transposon domestication in vertebrate evolution, a RAGB transposon. An active *ProtoRAG* transposon has been described in the *B. belcheri* genome containing both RAG1 and RAG2 genes flanked by terminal inverted repeats (TIR) [25]. We screened BraLan3 for the presence of *ProtoRAG* transposons and found two full *ProtoRAG* transposon copies in chromosomes 12 and 18 (with 98.8% sequence similarity) and an incomplete copy with a truncated RAG1 in chromosome 16 (Supplementary Note 1). These *B. lanceolatum ProtoRAG* transposons present the same structure as the *ProtoRAG* of *B. belcheri* without a PHD in RAG2, unlike in jawed vertebrates. In addition, we found 13 Miniature Inverted-repeat Transposable Elements (MITE) in 10 different BraLan3 chromosomes, suggesting the *ProtoRAG* transposon is active in this amphioxus genome. Together, these results show that this transposon has been active in amphioxus at least since the split between the *B. belcheri* lineage and the *B. lanceolatum* and *B. floridae* lineage.

Gene duplication profiles in amphioxus and vertebrates

In order to compare the gene duplication history of *B. lanceolatum* to that of vertebrates, we inferred ortholog groups (orthogroups) for chordates using the new BraLan3 gene annotation. We included in the analysis two other amphioxus species, the American amphioxus (*B. floridae*) and the Asian amphioxus (*B. belcheri*), and four vertebrate model species: zebrafish (*D. rerio*), chicken (*G. gallus*), mouse (*M. musculus*) and human (*H. sapiens*). We did not include ascidian genomes due to their highly derived status [2], which makes them less useful to study amphioxus-specific patterns. We only included protein-coding genes, where orthologs and paralogs can be more easily identified using sequence similarity (see Methods). Among all chordate orthogroups, we distinguished between single-copy orthogroups, lineage-specific duplicated orthogroups (either amphioxus- or vertebrate-specific), parallelly duplicated orthogroups, and lineage-specific orthogroups (Figure 2A). All orthogroups are derived from a single chordate ancestor gene (except for putative *de novo* gene births, see Figure 2A, Methods). This is, gene duplications predating chordates are classified in different orthogroups and, thus, not considered in our approach. For duplications in vertebrates, we distinguish between *small-scale duplicated genes* and *ohnologs*, the latter corresponding to genes duplicated during the 2R rounds of genome duplication at the origin of vertebrates. In order to retrieve patterns common to all vertebrates, teleost-specific ohnolog genes were not considered (see Methods for specifications on ohnolog definition). In the case of amphioxus, all gene duplications were considered small-scale duplications.

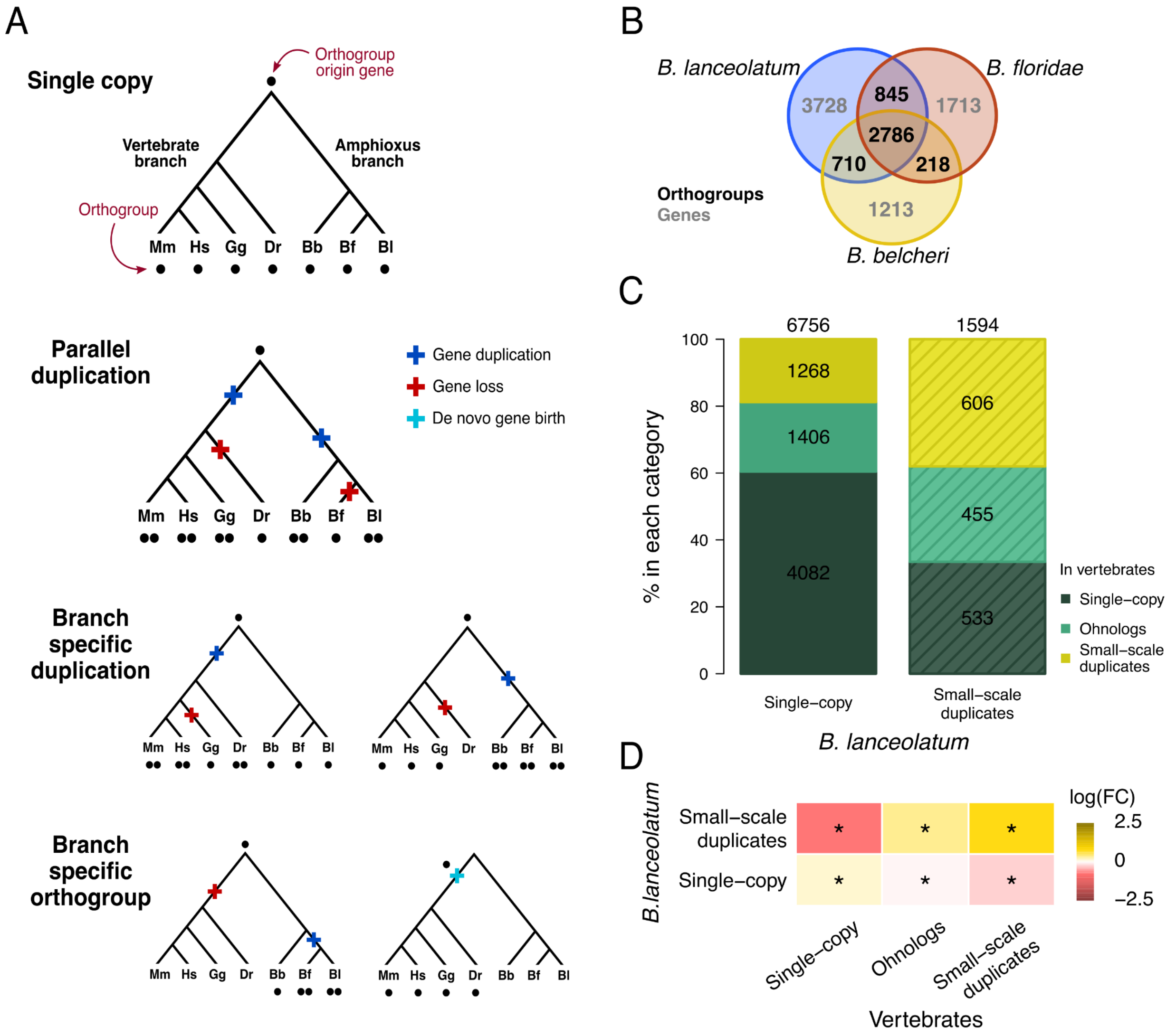

Figure 2. Gene duplication patterns in the amphioxus and vertebrate lineage. A. Schematic representation of the chordate gene orthogroups classification. Specific cases of possible orthogroup evolutionary histories are represented as examples. Mm, Hs, Gg, Dr, Bb, Bf and Bl correspond to mouse (*M. musculus*), human (*H. sapiens*), chicken (*G. gallus*), zebrafish (*D. rerio*), and the Asian (*B. belcheri*), American (*B. floridae*) and European amphioxus (*B. lanceolatum)* respectively. **B.** Venn diagram representing the number of amphioxus specific orthologous groups (black) and amphioxus species-specific genes (gray). **C.** Percentage of each vertebrate orthogroup type according to *B. lanceolatum* orthogroup type. **D.** Results of hypergeometric test for shared duplication profile between vertebrates and *B. lanceolatum*. log(FC) corresponds to the binary logarithm of the ratio between observed and expected values (fold change). P-values were corrected with Bonferroni multiple testing correction; all cases have p-value ≤ 0.0002 (*); corresponding values are in Table S1.

We detected 8705 orthogroups shared by amphioxus and vertebrates (65.6% of amphioxus groups and 68.4% of vertebrate groups). Thus, a large part of gene orthogroups are common and still detectable by sequence similarity, despite 500 million years of independent evolution between the two lineages. This is the set of shared orthogroups between amphioxus and vertebrates that we have considered, although others are likely to exist which are more divergent in sequence and thus undetected with the methods used [26]. We identified 4559 amphioxus lineage-specific orthogroups. These orthogroups are either lost in vertebrates, born *de novo* in the amphioxus lineage, or diverged enough to elude sequence similarity based orthology. The majority of these amphioxus-specific orthogroups are shared among the three amphioxus species included in the analysis (Figure 2B). *B. lanceolatum* shares slightly more orthogroups with *B. floridae* than it does with *B. belcheri*, consistent with the known Branchiostoma phylogeny in which *B. lanceolatum* and *B. floridae* are sister groups [27,28]. *B. floridae* and *B. belcheri* share less orthogroups than each of them does with *B. lanceolatum*, possibly due to differences in genome assembly and annotation, which supports the higher quality of BraLan3.

Around half of all genes are duplicated in both vertebrate and amphioxus genomes (Table 2). These results suggest that duplication contributes to the amphioxus gene repertoire in similar proportions to that of vertebrates. On the other hand, they do differ in the distribution of duplications among orthogroups. Slightly less than 20% of amphioxus orthogroups have duplications, while consistently more than 20% of orthogroups in vertebrates do. Zebrafish shows an even higher proportion of orthogroups with duplications, as expected because of the teleost WGD (Table 2). The average number of duplicated genes in duplicated orthogroups is consistently higher in amphioxus than in vertebrates. The examination of the chromosomal location of duplicated genes showed that tandemly duplicated genes are more abundant in *B. lanceolatum* than in vertebrates (Figure S2A). In conclusion, amphioxus and vertebrate genomes have similar proportions of duplicated genes but their distribution across orthologous groups is different, very likely because of differences in the two types of gene duplication mechanisms in the two lineages [6,29].

		Genes			Orthogroups			
Lineage	**Species**	**Total**	**Duplicated**	**%**	**Total**	**Duplicated**	**%**	**Mean # of genes in duplicated orthogroups**
Cephalochordates	*B. lanceolatum*	25946	11973	46.15	12691	2468	19.45	4.86
	B. floridae	20541	10192	49.62	10568	1934	18.30	5.27
	B. belcheri	18413	8469	45.99	10533	1805	17.14	4.69
Vertebrates	*D. rerio*	23912	13510	56.50	10587	3727	35.20	3.63
	G. gallus	16558	7416	44.79	10455	2399	22.95	3.09
	M. musculus	21691	10660	49.14	12236	2847	23.27	3.75
	H. sapiens	19715	9799	49.70	12205	2997	24.56	3.29

Table 2. Gene duplication prevalence in different species of amphioxus and vertebrates. Percentages are calculated from the total number of genes or the total number of orthogroups in each species, respectively.

Gene duplication patterns are conserved between amphioxus and vertebrates (Figure 2C, 2D, Table S1): genes which are duplicated in one lineage tend to be duplicated in the other lineage. As expected, this trend is especially strong for small-scale duplicated genes (1.69 times more orthologous duplicates than expected by chance), but surprisingly, also holds for vertebrate ohnologs (1.28 times more than expected). While there is also an enrichment in orthogroups which are single-copy in both lineages (1.09 times more than expected), it is less strong than the enrichment of parallel small-scale duplication. These results are robust to the inclusion of lineage-specific genes in the analysis (Figure S3, Table S2) and are consistent with recent observations in other amphioxus species [11]. Moreover, duplicated orthogroups with a low number of gene copies (<=2.5 mean number of genes) in amphioxus tend to have a low number of copies in vertebrates (Spearman's $\varrho = 0.28$; Table S3, contingency table chi-squared test p-value < 0.0001). This is true for both small-scale duplicates and ohnologs (Table S3).

Functional patterns of duplicate genes

The parallelism in duplication profiles between amphioxus and vertebrates extends to the level of functional categories. There is a strong positive relation in the proportion of duplications (including both small-scale and ohnologs) in different functional categories between the two lineages (Figures 3A and 3D). The categories with the lowest rates of duplication in both lineages correspond to basic functional categories, such as mitochondrial, transcriptional, translational or cell cycle-related biological processes or ribosome, DNA, RNA and nuclear-related molecular functions (Table S4). Conversely, many functional categories related to regulation, signaling, immune system or response have more than 80% or even 90% of genes duplicated in both lineages

(Table S4). These results suggest strong selection for preserving single-copy genes in some functions, and either a tolerance to duplication or selection favoring duplication in others. Interestingly, both vertebrate small-scale duplicates and ohnologs contribute to the general correlation of constraints on duplication in functional categories between the amphioxus and the vertebrate lineage (Figures 3B, 3C, 3E and 3F, Table S4). This means that the functional categories that retained ohnolog genes from the 2R WGDs in vertebrates have frequently duplicated through small-scale duplication in amphioxus. Together, these results highlight that, despite well established results on different constraints on paralog retention between small-scale duplicates and ohnologs [30–33], selection for preserving single-copy genes in certain basic functional categories drives a strong signal of parallelism in copy number between sub-phyla.

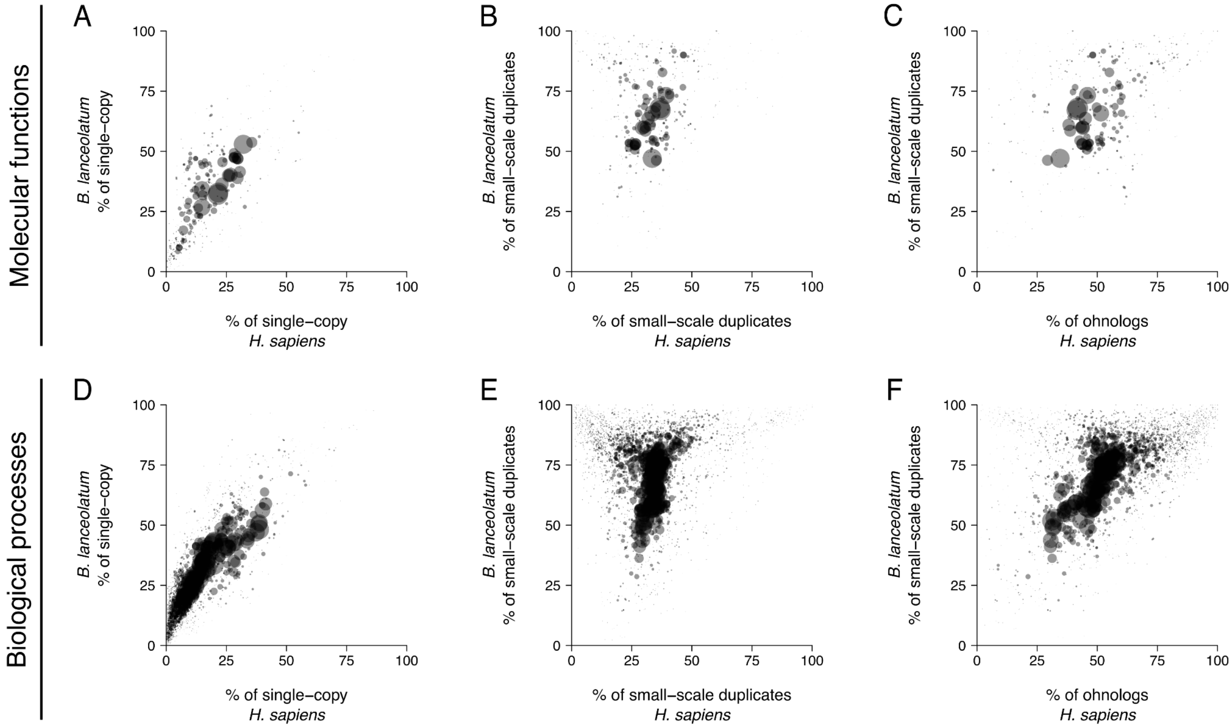

Figure 3. Functional categories' parallelism in duplications between amphioxus and vertebrates. *B. lanceolatum* genes orthologous to a human gene belonging to a given gene-ontology term (GO term) were considered to belong to this same GO term. Only GO terms with a minimum of 50 genes in both humans and *B. lanceolatum* were considered, point size is proportional to the number of human genes in each functional category. **A.** Percentage of single-copy genes for molecular function GO terms in *H. sapiens* compared to *B. lanceolatum*. **B.** Percentage of small-scale duplicated genes for molecular function GO terms in *H. sapiens* compared to *B. lanceolatum*. **C.** Percentage of ohnolog genes for molecular function GO terms in *H. sapiens* compared to the percentage of small-scale duplications in *B. lanceolatum*. **D-F.** Respectively similar to A-C for biological processes GO terms. No statistical test of correlation was done because different GO terms are not independent.

Expression and evolution of duplicate genes

To characterize *B. lanceolatum* duplicated genes independently of GO annotations in human, we re-analysed *B. lanceolatum* RNA-seq from Marlétaz et al. 2018 [8]. Amphioxus duplicated genes have lower levels of expression than single-copy genes. This is true both in adult tissues (Figure 4A) and during embryonic development (Figure 4B), and whatever the profile of the corresponding vertebrate orthologs: single-copy, ohnologs, small-scale duplicates, or missing (amphioxus-specific). The lower expression of amphioxus duplicated genes with respect to single-copy genes is preserved within functional categories (Figure S4). We also observe lower expression of amphioxus-specific genes even when they are single-copy (Figures 4A and 4B). Furthermore, both the amphioxus duplicated genes and the amphioxus-specific genes show higher levels of adult tissue-specificity and developmental stage-specificity than other genes (Figure 4C and 4D).

Together, these results show, on the one hand, that conserved genes between amphioxus and vertebrates that were not duplicated in the amphioxus lineage have a major role in the transcriptome of amphioxus. They show higher gene expression and less tissue and stage specificity, suggesting a role in the maintenance of basic cellular processes in amphioxus. On the other hand, the lower expression and higher specificity of amphioxus duplicates and of amphioxus-specific genes proposes secondary but more specialized roles of these genes in this species. These results are consistent with the analysis of functional categories, where duplicated genes are less present in core functional categories and less expressed whenever they are present, and coherent with previously described patterns of expression of duplicated and of young or rapidly-evolving genes [34–39].

In order to further explore the evolution of amphioxus duplicated genes, we compared expression in amphioxus and zebrafish single-copy or duplicated orthologs (Figures 4E, following the approach of Marlétaz et al. 2018 [8]). The distribution of the differences in observed gene expression patterns between amphioxus and zebrafish single-copy orthologs (blue histograms in Figure 4E) has a strong centrality and symmetry. This shows that the dominant pattern is conservation in gene expression conditions between amphioxus and zebrafish in the absence of duplication. On the contrary, genes specifically duplicated in either one or the other lineage show a loss of expression domains compared to their single-copy orthologs in the other lineage. This effect is seen as an asymmetry in the histograms (solid red histograms in Figure 4E) and happens in all cases: amphioxus-specific duplicated genes (Figure 4E left), vertebrate-specific small-scale duplicates (Figure 4E center) and vertebrate ohnologs (Figure 4E right). When unifying the expression profiles of duplicates in the same orthogroup, the skewness of the curve is reverted in both lineages and for both duplicate types (shaded red histograms in Figure 4E and Figure 4F). This is, when all duplicated genes within an orthogroup are accounted for together, they tend to have an expression profile similar to the one of their single-copy ortholog. We observe this pattern for small-scale duplicates in both lineages and for ohnologs, and when splitting amphioxus duplicated genes in either multichromosomal, monochromosomal distant and tandemly duplicated genes (Figure S2B). These results, previously reported specifically for vertebrate ohnologs [8], support a general trend of specialization in all types of duplicates in both lineages and are consistent with expectations for subfunctionalization and specialization.

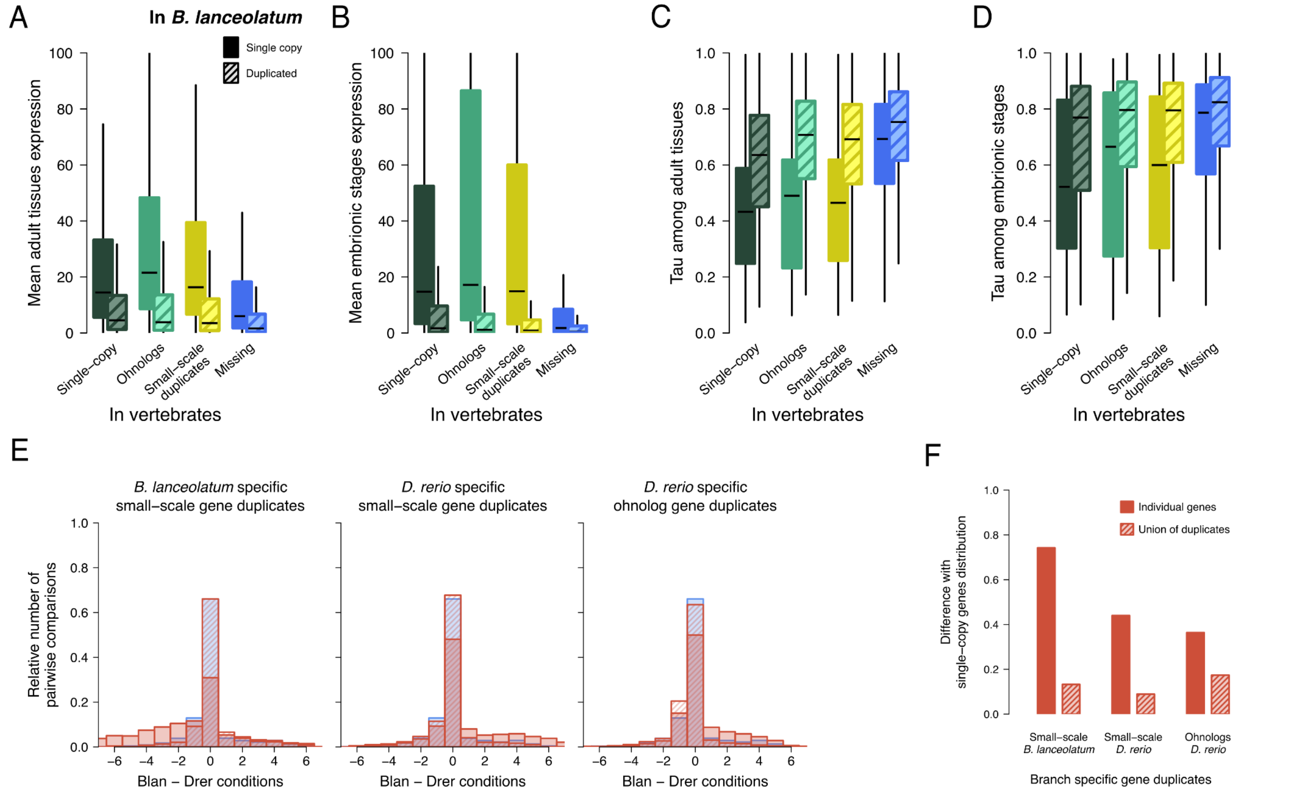

Figure 4. ***B. lanceolatum*** **duplicated gene expression evolution relative to vertebrates. A-B.** Mean gene expression (TPM) across adult amphioxus tissues (**A**) or embryonic development stages (**B**) for amphioxus single-copy (solid boxplots) or duplicated (shaded boxplots) genes, relative to their status in vertebrates. **C-D.** Tissue-specificity (**C**) or stage-specificity (**D**) for the same gene groups as A and B. Tau values range from 0 (ubiquitous expression) to 1 (specific expression). **E.** Distribution of differences in number of gene expression domains between amphioxus and zebrafish for one-to-one orthologs (blue), for branch specific duplicated genes (solid light red), and for the union of expression patterns of duplicated genes within each orthogroup (shaded red). Left, center and right correspond to *B. lanceolatum* specific small-scale duplicates, zebrafish specific small-scale duplicates and ohnologs respectively. **F.** Quantification of the difference between individual genes [union of duplicates] distribution and single-copy distribution of expression differences between *B. lanceolatum* and zebrafish (i.e., difference between solid light red [shaded red] and blue in E) for each one of the cases in E.

Amphioxus gene synteny

B. lanceolatum and *B. floridae* diverged between 20 and 150 million years ago (Mya) but share around 500 My of common evolution since their lineage split with that of other chordates [27,28]. The presence of chromosome scale genome references for these two species [9] allows comparison of gene order. While one-to-one orthologs show an overall conservation of synteny between the two species, there are also two important chromosomal rearrangements (Figure 5). Although both species share the same number of chromosomes (n = 19), we observe two large chromosomal fusions and splits between the two species. The largest chromosome of *B. lanceolatum* (chromosome 1) appears to be homologous to two smaller *B. floridae* chromosomes (chromosomes 9 and 11). Conversely, chromosome 3 of *B. floridae* appears to be homologous to *B. lanceolatum* chromosomes 15 and 19. In order to shed light on the ancestral state of these chromosomal fusions or splits, we compared the *B. lanceolatum* and *B. floridae* synteny to the fragmented *B. belcheri* genome reference at the gene level (Figure S5). We found no evidence for either chromosomal fusion in *B. belcheri*. This is, none of the *B. belcheri* scaffolds map to both chromosomes 9 and 11 of *B. floridae* or to both chromosomes 15 and 19 of *B. lanceolatum*. These results match recent evidence on *B. belcheri* and *B. floridae* karyotype reconstruction, pointing towards a chromosomal fusion in the *B. floridae* branch with respect to *B. belcheri* [11]. Here, we see that this chromosomal fusion is *B. floridae* specific (is not shared with *B. lanceolatum*) and that there is an additional chromosomal fusion in *B. lanceolatum*.

Gene collinearity of one-to-one orthologs between *B. lanceolatum* and *B. floridae* is generally conserved, with frequent medium scale intra-chromosomal rearrangements (inversions and translocations). This observation is true for the majority of chromosomes but fails for the smallest chromosomes (*B. lanceolatum* chromosomes 16, 17 and 18 that correspond to *B. floridae* chromosomes 17, 18 and 19 respectively) where we find a loss of gene collinearity and a dispersed distribution of one-to-one orthologs. Unlike one-to-one orthologs, amphioxus-specific gene duplicates show a scattered pattern of chromosome distribution between *B. lanceolatum* and *B. floridae* (Figure S6A), with local gene expansions. As expected, they show no evidence of any large-scale duplication event. Overall, although inter-chromosomal duplications have been frequent in the amphioxus lineage, small-scale duplications are found with more probability within the same chromosome.

Between *B. lanceolatum* and vertebrate (*G. gallus* in Figure S6B) one-to-one orthologs, we observe several conserved synteny blocks with a complete loss of gene collinearity. This is consistent with previously reported results for *B. floridae* [9,10], and suggests that in addition to chromosomal rearrangements, there were numerous middle and small-scale rearrangements affecting genes in both lineages, as commonly observed in vertebrates [40,41]. These have broken gene collinearity while preserving the context of conserved synteny blocks. In the case of duplicated genes, we find some synteny block conservation between chordate lineages with a general dispersal of duplicated genes across the genome. We also observe multiple local gene expansions, more frequent in the vertebrate lineage (Figure S6C).

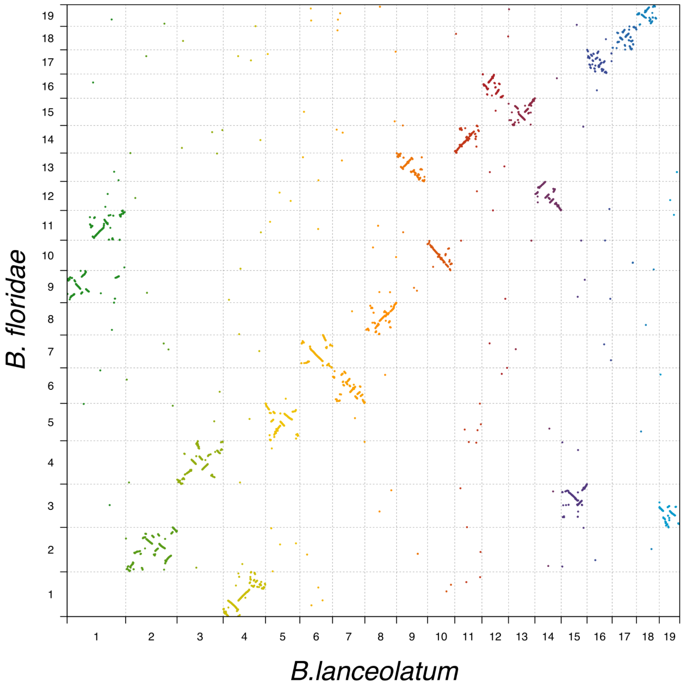

Figure 4. Single-copy genes synteny conservation between *B. lanceolatum* and *B. floridae*. Every gene is represented by its mid-point coordinates in the genome.

Discussion

Amphioxus are generally characterized by slow molecular evolution, notably slow protein sequence evolution, relative to the two other lineages of chordates, ascidians and vertebrates [6,7]. This provides a useful contrast to the vertebrate diversity of phenotypes and genomes, and has led to the use of amphioxus to understand many features of vertebrate evolution, serving as a proxy for the ancestral state. Given the role of gene duplication in evolutionary change, one could expect to find a low duplication rate and a minor role of duplicate genes in the amphioxus lineage. Yet, we and others [11] find a similar amount of small-scale gene duplication activity in amphioxus as in vertebrates, and we show a strong parallelism in terms of genes, biological functions, genome distribution, and gene expression profiles.

The extent of amphioxus-specific duplication has been difficult to study until recently. This is because the high heterozygosity of amphioxus genomes was a problem for the reliable assembly of duplicated genes [6,19]. In this study, we constructed both high-quality assembly and annotation for *B. lanceolatum*, allowing us to infer a reliable set of paralogs, and to study these duplications genome-wide. The usage of long-read and long-range scaffolding data as well as specifically designed methods for alternative haplotype filtering provided a high quality reference on top of which we cautiously annotated genes. As is usual when annotating the genome of a non-model organism, we filtered gene models by similarity to known genes or proteins. This can lead to false negatives, especially putative faster-evolving duplicate or lineage-specific genes. To minimize this issue while avoiding false positives, we also retained both gene predictions with a BLAST hit in other amphioxus annotations (recovering 143 genes) and gene predictions expressed above the background in at least three samples (recovering 986 genes). This has allowed us to recover a total of 11973 duplicated genes in *B. lanceolatum*, a similar proportion of the total gene set as in most vertebrates. With different methods, Huang et al. [11] recently reported high quality genomes for several other species of amphioxus and also found similar proportions of gene gains (including duplicated genes) in the vertebrate and amphioxus lineages.

The parallelism of duplication patterns between vertebrates and amphioxus is striking. The same functional categories are enriched or depleted. Notably, regulatory genes are over-represented in amphioxus duplicated genes, as they are in other lineages, including yeasts, plants and vertebrates [42,43]. This shows that amphioxus have continued to evolve complex regulatory processes or pathways, despite remaining apparently simpler than other chordates. This is reinforced by the parallelism of duplicate gene evolutionary patterns, with a similar dominance of expression specialization in amphioxus and vertebrates. Moreover, synteny analysis shows a dispersed distribution across the genome of small-scale duplicates in the amphioxus lineage, similar to the one known for vertebrates [44], pointing towards common gene duplication mechanisms and dynamics. Of note, Huang et al. [11] report an excess of segmental duplications in other amphioxus species.

While our new genome assembly confirms the overall conservative evolution of amphioxus synteny, we find chromosomal rearrangements between *B. lanceolatum* and *B. floridae* that diverged between 20 and 150 Mya [27,28]. Our observations corroborate the conclusions derived from the very recently reported high quality genomes of three other amphioxus [11] ancestral *Branchiostoma* karyotype had 20 pairs of chromosomes, and we report two independent chromosomal fusions in *B. lanceolatum* and *B. floridae*. These rearrangements, together with the duplication patterns, illustrate how amphioxus genomes continue to be shaped by large scale evolutionary events.

Conclusions

The amphioxus lineage has a history of small-scale gene duplications similar to the one observed in the vertebrate lineage, and there is a conservation of the constraints on gene duplication between these two lineages. Our results highlight the durability of the selection preventing duplication in certain gene families and allowing or favoring it in others. In the amphioxus and vertebrate lineages, around 500My of independent evolution and the large diversification of vertebrate's phenotypes have not erased most of these constraints on gene duplication despite the 2R WGDs in early vertebrate evolution.

Methods

Genome assembly

We generated 9.46M PacBio reads through 20 cells of PacBio RSII, with a N50 of 10998 and totalling 80.2Gb of raw data, which represents 146x coverage of the haploid *B. lanceolatum* genome. Canu (v1.9) was run on the PacBio data using the parameters correctedErrorRate=0.065, ovlMerDistinct=0.975 and batOptions="-dg 3 -db 3 -dr 1 -ca 500 -cp 50" to maximize haplotype separation [45]. With these settings, Canu yielded, a diploid assembly of 931.9Mb with a N50 of 1.02Mb that was subjected to polishing with the Arrow algorithm (v2.3.2) as implemented in the GenomicConsensus package (PacBio) after aligning back the raw PacBio reads with Pbmm2 (v1.1.0) relying on Minimap2 (v2.17). After polishing, we filtered haplotigs and heterozygous regions from the assembly with purge_dups (v1.0.0), relying on coverage depth and sequence reciprocal alignment [46]. We estimated coverage by aligning PacBio reads with Minimap2, and we used a coverage cutoff of 105 to distinguish between homozygous and heterozygous regions. We further scaffolded the resulting haploid assembly (size: 507.4Mb with a N50 of 1.48Mb) using chromatin conformation capture data (HiC data). HiC data was aligned and duplicated and spurious read pairs were filtered out using Juicer (v6be7c0f) [47] and BWA-MEM (v0.7.17). The subsequent read pairs were used as an input by 3D-DNA to perform contact-based scaffolding [48]. The resulting assembly was manually edited using Juicebox to correct problems and subjected to a final round of optimization with 3D-DNA. This assembly notably had 19 large scaffolds consistent with the chromosome number of *B. lanceolatum* [24], in addition to multiple smaller scaffolds. To reduce the computational load of downstream analyses, we removed small scaffolds that did not have multi-intronic gene models or that had a high percent (≥50%) of repeat content.

Two additional modifications were done to this intermediate assembly. First, the assembly of the Irx cluster on chromosome 7 was manually curated based on previous information [49]. In short, the ortholog of *IrxC* was already located on chr7, but *IrxA* and *IrxB* were each present in a different small scaffold, which were introduced in the correct location of chr7. Second, we detected various gene models that contained in-frame stop codons. Upon further inspection, these were found to be due to indels introduced by PacBio sequencing. To correct these errors, we extracted all unique exons obtained through an initial gene annotation with Mikado and StringTie+Transdecoder (see below) and blasted them against a combined Trinity assembly generated with Illumina RNA-seq data from multiple tissues and developmental stages [8]. The blast output was parsed to conservatively detect single indels within 10 nt alignment windows, supported by at least five different Trinity transcripts coming from at least two independent RNA-seq samples. These indels corresponded to 3672 insertions and 2690 deletions with respect to the Trinity assembly, mainly in UTR exons and/or lowly supported gene models, and were edited in the genome sequence to produce the final assembly.

Annotation

Stage- and organ-specific transcriptomes (RNA-seq) [8] were aligned using STAR and assembled as individual transcriptomes using StingTie [50]. The StingTie assemblies (GTF files) were merged using Taco [51]. The RNA-seq data was also assembled in bulk with Trinity as reported in Marletaz et al. 2018 [8] and aligned to the new assembly using Minimap2 with the 'splice:hq' parameter. These transcriptomes were leveraged using the Mikado tool [52]. We also generated spliced protein alignment using Exonerate from the annotation of the *B. floridae* genome [9] assuming at least 65% protein identity and a maximum intron size of 250kb. We converted the Mikado transcriptome assembly and the Exonerate alignment into hints for the Augustus gene prediction tool [53], as 'exon' and 'CDS' hints, respectively. Augustus was run using the previously defined model and the aforementioned hints while allowing hinted splices ('ATAC'). Augustus yielded 37,787 gene models, of which 7,101 overlapped with a repeat (all exons with at least 50% overlap with repeats) and were excluded. We constructed a repeat library using RepeatModeller (v2.0) and masked repeats using RepeatMasker (v4.0.7). We also generated a transcriptome dataset by aligning the assembled transcriptome using PASA (v2.3.2) and used it to add UTRs and isoforms to the Augustus gene models through two rounds of processing.

A series of additional corrections were then applied to this initial annotation. First, 5' and 3' UTRs were extended based on a GTF merge of the StingTie assemblies on which Transdecoder (PMID: 23845962) was run to identify open reading frames. 3' UTRs were extended up to 5 Kbp unless they overlap with downstream models in the same strand and provided these 3' UTRs had read support (otherwise, only 0.5 Kbp extensions were allowed). Upon extension, the end of the gene models were modified if there was no annotated STOP codon and an in-frame STOP codon was added with the UTR extension. Similarly, for gene models with no annotated start codon, 5' UTRs were extended when possible based on StingTie information, and a start codon were added when an upstream in-frame ATG was introduced with the extension. In addition, we noticed some Augustus gene models whose annotated start codon seemed upstream than expected based on a protein alignment with the human orthologs. To identify and correct these cases, when the initial M of the human ortholog aligned with an M in the amphioxus sequence, this M was selected as the new starting codon. Finally, we performed a search for potential chimeras and broken genes comparing Augustus (default) and StingTie+Transdecoder gene models, as described in [8]. These pairs of overlapping models were manually inspected and corrected by substituting the Augustus gene models by StingTie+Transdecoder ones when appropriate. Finally, gene models were named based on the following nomenclature, BLAGXXYZZZZZ, where XX corresponded to the chromosome (small scaffolds were indicated by 90), Y was 0 or 1 if the gene model was derived from Augustus or StingTie+Transdecoder, respectively, and ZZZZZ followed a consecutive order in the chromosomes.

To validate the new annotation, we used two parallel strategies; search for sequence similarity to a known gene product and search for evidence of expression. We ran BLASTp for all genes in our annotation against three protein sequence pools; UniProt database [54] and the current gene annotation for both *B. floridae* and *B. belcheri*. For all protein sequence pools and each gene independently, we reported a *strong* evidence of sequence similarity for genes with hits fulfilling query length / alignment length > 0.75, subject length / alignment length

> 0.75 and e-value < 10^{-8}; *weak* evidence for other genes with hits with e-value < 10^{-4}; and, *no evidence* for other genes. Gene expression above background noise in *B. lanceolatum* RNA-seq libraries as in [8] was reported as evidence of gene expression for every gene in the new annotation [55]. For every gene, we reported *strong* evidence of gene expression if it was expressed in more than 3 libraries; *weak* evidence if it was expressed in less than 3 libraries and *zero* evidence if expressed in none of the libraries. Of the total 27102 genes, 97.66% (26468) have strong evidence in at least one of the 4 strategies (see complete numbers on annotation validation in Table S5). All genes with reported strong evidence in at least one of the four approaches (3 sequence similarity searches and one gene expression validation) were used for subsequent analysis. This validation method combining sequence similarity to known proteins and gene expression, allows keeping gene models with low similarity, because they evolve fast or are short, while limiting false positives.

Genome assembly and annotation quality comparison

We compared the quality of BraLan3 with that of other relevant genome references; the previously available assembly for *B. lanceolatum* [8], and the available assemblies for *Branchiostoma belcheri* [7], *Branchiostoma floridae* [9], *Danio rerio* (GRCz11), *Gallus gallus* (GRCg6a), *Homo sapiens* (GRCh38) and *Mus musculus* (GRCm39). Published Branchiostoma genome references were downloaded from NCBI while vertebrate genome references were downloaded from ENSEMBL (release 103). All summary statistics were computed with the final genome assemblies (chromosome level instead of scaffold level if applicable). BUSCO (4.1.4) was run with the metazoan universe (metazoa_odb10) in all cases [56].

Ortholog and paralog analysis

Genome assemblies, gene annotations and protein sequences for *Branchiostoma belcheri* [7] and *Branchiostoma floridae* [9] were downloaded from NCBI genome browser and for *Danio rerio* (GRCz11), *Gallus gallus* (GRCg6a), *Homo sapiens* (GRCh38) and *Mus musculus* (GRCm39) were downloaded from ENSEMBL (release 103). Only BraLan3 genes with strong evidence in at least one validation strategy were used for this analysis (see annotation section). For all species, cDNA sequence of each gene was extracted from genome sequence according to annotated coordinates and was compared to the corresponding protein sequence. We filtered out genes with non-corresponding cDNA-protein sequence pairs (allowing for 10% of mismatches; see Table S6). The longest gene isoform (transcript) was used in all cases. Gene orthology analysis based on protein sequence similarity was performed with Broccoli (version 1.1) [57] using default parameters, DIAMOND (2.0.7.14) [58] and FastTree (2.1.10) [59]. Broccoli groups genes derived from one single gene in the last common ancestor of all considered species in a given orthogroup. By using this algorithm, we classified gene duplications preceding amphioxus and vertebrate lineages split in different orthogroups and, thus, they were not considered as duplicated in the subsequent analysis. This is, in the whole gene duplication analysis, only duplications posterior to the last common ancestor of amphioxus and vertebrates were considered.

We distinguished between single-copy, small-scale duplicated and ohnolog genes as follows. If a given species had more than one gene in a given Broccoli orthogroup, this orthogroup was considered duplicated (non-single-copy) in this species. If at least one amphioxus/vertebrate species presented a duplication in a given orthogroup, this orthogroup was considered as duplicated in the corresponding branch (amphioxus or vertebrate). Later on, in vertebrates, we differentiated small-scale duplications from ohnologs by classifying all orthogroups containing at least one ohnolog gene derived from the 2R WGDs as ohnolog orthogroups (list of 2R ohnologs from OHNOLOGS version 2.0 for the four vertebrate species [60]). In order to avoid classifying 3R teleost ohnologs (derived from the teleost-specific WGD) as small-scale duplications in vertebrates while focusing in 2R ohnologs, we considered as non-duplicated all orthogroups that, among vertebrates, were only duplicated in zebrafish and that were known to be retained in the 3R WGD (list of 3R ohnologs from OHNOLOGS version 2.0 for zebrafish [60]). Co-occurence of gene duplication status between amphioxus and vertebrate lineages in orthogroups was tested with a hypergeometric test and p-values were corrected with the Bonferroni correction.

Functional annotation

Human genes belonging to gene ontology (GO) molecular function and biological processes terms were retrieved by unifying iteratively all genes belonging to all child terms of a given GO term [61,62]. This analysis was restricted to human-*B. lanceolatum* orthologous genes, meaning that only *B. lanceolatum* genes with a human ortholog were considered and vice versa. *B. lanceolatum* genes orthologous to a human gene belonging to a given GO term were considered to belong to this same GO term. This same reasoning was applied back to human genes in order to avoid being more restrictive in considering a gene as belonging to a given GO term in human respect to *B. lanceolatum*. This is, if a human gene in a given orthogroup belonged to a given GO term, all the genes in the orthogroup were considered as belonging to this GO term. Only GO terms with a minimum of 50 genes in both humans and *B. lanceolatum* were considered.

Gene expression

Amphioxus RNA-Seq gene expression data [8] was pseudoaligned to BraLan3 gene annotation with Kallisto [63] using the *--single --rf-stranded -l 180 -s 20 --bias* parameters. Gene abundances were retrieved with tximport [64]. Zebrafish gene expression estimates in several tissues and developmental stages were retrieved from Bgee version 15 [55] using the R package BgeeDB [65]. Amphioxus average gene expression in embryonic stages, male gonads, muscle, neural tube, gut, grills and hepatic diverticulum were matched to zebrafish gene expression in embryo, testis, muscle tissue, brain, intestine, pharyngeal grill and liver respectively. Blastula, female gonads and epidermis were excluded from the analysis due to divergent patterns of gene expression in amphioxus. Gene expression specificity in amphioxus adult tissues and embryonic stages was estimated with the Tau index [66].

Declarations

Ethics approval and consent to participate

A ripe adult *B. lanceolatum* individual was collected in Argelès-sur-Mer (France) with a specific permission delivered by the Préfet des Pyrénées-Orientales. *B. lanceolatum* is a non-protected species.

Consent for publication

Not applicable.

Availability of data and materials

BraLan3 genome reference and annotation and the *B. lanceolatum* whole genome PacBio sequencing data and chromatin conformation capture data (HiC) we used to build it are available in the European Nucleotide Archive (ENA) under the accession number PRJEB49647 (as of submission, some data are still "pending" but should be available shortly). Code used for the analysis available in https://github.com/marinabraso/BraLan3.

Competing interests

The authors declare that they have no competing interests.

Funding

MRR acknowledges support from the Swiss National Science Foundation grant 31003A/173048, and from the Etat de Vaud. FMar acknowledges support from the Royal Society grant URF\R1\191161. MI acknowledges support from European Research Council ERC-StG-LS2-637591. HE acknowledges support from the Centre National de la Recherche Scientifique and Agence Nationale de la Recherche (ANR-16-CE12-0008-01 and ANR-19-CE13-0011). PP acknowledges support from the French Government under the «Investissements d'avenir» program managed by the Agence Nationale de la Recherche (Méditerranée Infection 10-IAHU-03). IM acknowledges support from the Spanish Ministry of Science and Innovation and the European Union (RYC-2016-20089 and PGC2018-099392-A-I00). JJT acknowledges support from the ERC (Grant Agreement No. 740041) and the Spanish Ministerio de Ciencia e Innovación (PID2019-103921GB-I00).

Authors' contributions

MBV performed the gene duplication analysis and results. MBV and MRR wrote the first draft of the manuscript; MBV, FMar, IM, HE, PP, MI and MRR wrote the final manuscript. HE provided the original samples. FMar, AE, FMan, MI, IM, RDA, JLGS and JJT performed the BraLan3 assembly and annotation. PP and LLT performed the RAG analysis. FMar contributed to the comparative synteny analysis. All authors contributed to result interpretation and discussion. All authors read and approved the final manuscript.

Acknowledgements

We thank Kamil S. Jaron for work on an earlier assembly, and discussion of those preliminary results. We acknowledge Diego A. Hartasánchez and all the members of MRR group for valuable discussions. Furthermore, we thank Consolée Aletti and the Genomic Technologies Facility at University of Lausanne for technical contributions.

Bibliography

1. Bertrand S, Escriva H. Evolutionary crossroads in developmental biology: amphioxus. Development. 2011;138:4819–30.
2. Holland LZ. Genomics, evolution and development of amphioxus and tunicates: The Goldilocks principle. J Exp Zoolog B Mol Dev Evol. 2015;324:342–52.
3. Escriva H. My Favorite Animal, Amphioxus: Unparalleled for Studying Early Vertebrate Evolution. BioEssays. 2018;40:1800130.
4. Delsuc F, Philippe H, Tsagkogeorga G, Simion P, Tilak M-K, Turon X, et al. A phylogenomic framework and timescale for comparative studies of tunicates. BMC Biol. 2018;16:39.
5. Benito-Gutiérrez È, Gattoni G, Stemmer M, Rohr SD, Schuhmacher LN, Tang J, et al. The dorsoanterior brain of adult amphioxus shares similarities in expression profile and neuronal composition with the vertebrate telencephalon. BMC Biol. 2021;19:110.
6. Putnam NH, Butts T, Ferrier DEK, Furlong RF, Hellsten U, Kawashima T, et al. The amphioxus genome and the evolution of the chordate karyotype. Nature. 2008;453:1064–71.
7. Huang S, Chen Z, Yan X, Yu T, Huang G, Yan Q, et al. Decelerated genome evolution in modern vertebrates revealed by analysis of multiple lancelet genomes. Nat Commun. 2014;5:5896.
8. Marlétaz F, Firbas PN, Maeso I, Tena JJ, Bogdanovic O, Perry M, et al. Amphioxus functional genomics and the origins of vertebrate gene regulation. Nature. 2018;564:64–70.
9. Simakov O, Marlétaz F, Yue J-X, O'Connell B, Jenkins J, Brandt A, et al. Deeply conserved synteny resolves early events in vertebrate evolution. Nat Ecol Evol. 2020;4:820–30.
10. Louis A, Roest Crollius H, Robinson-Rechavi M. How much does the amphioxus genome represent the ancestor of chordates? Brief Funct Genomics. 2012;11:89–95.
11. Huang Z, Xu L, Cai C, Zhou Y, Liu J, Zhu Z, et al. Three amphioxus reference genomes reveal gene and chromosome evolution of chordates. bioRxiv. 2022;doi:10.1101/2022.01.04.475009.
12. Furlong RF, Holland PWH. Were vertebrates octoploid? Philos Trans R Soc Lond B Biol Sci. 2002;357:531–44.
13. Conant GC, Wolfe KH. Turning a hobby into a job: How duplicated genes find new functions. Nat Rev Genet. 2008;9:938–50.
14. Ohno S. Evolution by gene duplication. Springer; 1970.
15. Zhang J. Evolution by gene duplication: an update. Trends Ecol Evol. 2003;18:292–8.
16. Magadum S, Banerjee U, Murugan P, Gangapur D, Ravikesavan R. Gene duplication as a major force in evolution. J Genet. 2013;92:155–61.
17. Kuzmin E, Taylor JS, Boone C. Retention of duplicated genes in evolution. Trends Genet. 2022;38:59–72.
18. Urchin Genome Sequencing Consortium. The Genome of the Sea Urchin Strongylocentrotus purpuratus. Science. 2006;314:941–52.
19. Bi C, Lu N, Han T, Huang Z, Chen J-Y, He C, et al. Whole-Genome Resequencing of Twenty *Branchiostoma belcheri* Individuals Provides a Brand-New Variant Dataset for *Branchiostoma*. BioMed Res Int. 2020;2020:1–15.
20. Huang S, Chen Z, Huang G, Yu T, Yang P, Li J, et al. HaploMerger: Reconstructing allelic relationships for polymorphic diploid genome assemblies. Genome Res. 2012;22:1581–8.
21. Huang S, Kang M, Xu A. HaploMerger2: rebuilding both haploid sub-assemblies from high-heterozygosity diploid genome assembly. Bioinformatics. 2017;33:2577–9.
22. Hartasánchez DA, Brasó-Vives M, Heredia-Genestar JM, Pybus M, Navarro A. Effect of Collapsed

Duplications on Diversity Estimates: What to Expect. Genome Biol Evol. 2018;10:2899–905.
23. Alkan C, Sajjadian S, Eichler EE. Limitations of next-generation genome sequence assembly. Nat Methods. 2011;8:61–5.
24. Colombera D. Male chromosomes in two populations of Branchiostoma lanceolatum. Experientia. 1974;30:353–5.
25. Huang S, Tao X, Yuan S, Zhang Y, Li P, Beilinson HA, et al. Discovery of an Active RAG Transposon Illuminates the Origins of V(D)J Recombination. Cell. 2016;166:102–14.
26. Herrera-Úbeda C, Marín-Barba M, Navas-Pérez E, Gravemeyer J, Albuixech-Crespo B, Wheeler GN, et al. Microsyntenic Clusters Reveal Conservation of lncRNAs in Chordates Despite Absence of Sequence Conservation. Biology. Multidisciplinary Digital Publishing Institute; 2019;8:61.
27. Subirana L, Farstey V, Bertrand S, Escriva H. Asymmetron lucayanum: How many species are valid? PLOS ONE. 2020;15:e0229119.
28. Igawa T, Nozawa M, Suzuki DG, Reimer JD, Morov AR, Wang Y, et al. Evolutionary history of the extant amphioxus lineage with shallow-branching diversification. Sci Rep. 2017;7:1157.
29. Jaillon O, Aury J-M, Brunet F, Petit J-L, Stange-Thomann N, Mauceli E, et al. Genome duplication in the teleost fish Tetraodon nigroviridis reveals the early vertebrate proto-karyotype. Nature. 2004;431:946–57.
30. Davis J, Petrov D. Do disparate mechanisms of duplication add similar genes to the genome? Trends Genet. 2005;21:548–51.
31. Brunet FG, Crollius HR, Paris M, Aury J-M, Gibert P, Jaillon O, et al. Gene Loss and Evolutionary Rates Following Whole-Genome Duplication in Teleost Fishes. Mol Biol Evol. 2006;23:1808–16.
32. Makino T, McLysaght A. Positionally biased gene loss after whole genome duplication: Evidence from human, yeast, and plant. Genome Res. 2012;22:2427–35.
33. Roux J, Liu J, Robinson-Rechavi M. Selective Constraints on Coding Sequences of Nervous System Genes Are a Major Determinant of Duplicate Gene Retention in Vertebrates. Mol Biol Evol. 2017;34:2773–91.
34. Brohard-Julien S, Frouin V, Meyer V, Chalabi S, Deleuze J-F, Le Floch E, et al. Region-specific expression of young small-scale duplications in the human central nervous system. BMC Ecol Evol. 2021;21:59.
35. Shew CJ, Carmona-Mora P, Soto DC, Mastoras M, Roberts E, Rosas J, et al. Diverse Molecular Mechanisms Contribute to Differential Expression of Human Duplicated Genes. Mol Biol Evol. 2021;38:3060–77.
36. Guschanski K, Warnefors M, Kaessmann H. The evolution of duplicate gene expression in mammalian organs. Genome Res. 2017;27:1461–74.
37. Jiang X, Assis R. Natural Selection Drives Rapid Functional Evolution of Young Drosophila Duplicate Genes. Mol Biol Evol. 2017;34:3089–98.
38. Huerta-Cepas J, Dopazo J, Huynen MA, Gabaldón T. Evidence for short-time divergence and long-time conservation of tissue-specific expression after gene duplication. Brief Bioinform. 2011;12:442–8.
39. O'Toole ÁN, Hurst LD, McLysaght A. Faster Evolving Primate Genes Are More Likely to Duplicate. Mol Biol Evol. 2018;35:107–18.
40. Robinson-Rechavi M, Boussau B, Laudet V. Phylogenetic Dating and Characterization of Gene Duplications in Vertebrates: The Cartilaginous Fish Reference. Mol Biol Evol. 2004;21:580–6.
41. Brasó-Vives M, Povolotskaya IS, Hartasánchez DA, Farré X, Fernandez-Callejo M, Raveendran M, et al. Copy number variants and fixed duplications among 198 rhesus macaques (Macaca mulatta). PLOS Genet. 2020;16:e1008742.
42. Cañestro C, Albalat R, Irimia M, Garcia-Fernàndez J. Impact of gene gains, losses and duplication modes on the origin and diversification of vertebrates. Semin Cell Dev Biol. 2013;24:83–94.
43. Maere S, Bodt SD, Raes J, Casneuf T, Montagu MV, Kuiper M, et al. Modeling gene and genome duplications in eukaryotes. Proc Natl Acad Sci. National Academy of Sciences; 2005;102:5454–9.

44. Kaessmann H. Origins, evolution, and phenotypic impact of new genes. Genome Res. 2010;20:1313–26.
45. Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. Canu: scalable and accurate long-read assembly via adaptive *k* -mer weighting and repeat separation. Genome Res. 2017;27:722–36.
46. Guan D, McCarthy SA, Wood J, Howe K, Wang Y, Durbin R. Identifying and removing haplotypic duplication in primary genome assemblies. Bioinformatics. 2020;36:2896–8.
47. Durand NC, Shamim MS, Machol I, Rao SSP, Huntley MH, Lander ES, et al. Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. Cell Syst. 2016;3:95–8.
48. Dudchenko O, Batra SS, Omer AD, Nyquist SK, Hoeger M, Durand NC, et al. De novo assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. Science. 2017;356:92–5.
49. Maeso I, Irimia M, Tena JJ, González-Pérez E, Tran D, Ravi V, et al. An ancient genomic regulatory block conserved across bilaterians and its dismantling in tetrapods by retrogene replacement. Genome Res. 2012;22:642–55.
50. Kovaka S, Zimin AV, Pertea GM, Razaghi R, Salzberg SL, Pertea M. Transcriptome assembly from long-read RNA-seq alignments with StringTie2. Genome Biol. 2019;20:278.
51. Niknafs YS, Pandian B, Iyer HK, Chinnaiyan AM, Iyer MK. TACO produces robust multisample transcriptome assemblies from RNA-seq. Nat Methods. 2017;14:68–70.
52. Venturini L, Caim S, Kaithakottil GG, Mapleson DL, Swarbreck D. Leveraging multiple transcriptome assembly methods for improved gene structure annotation. GigaScience. 2018;7:giy093.
53. Keller O, Kollmar M, Stanke M, Waack S. A novel hybrid gene prediction method employing protein multiple sequence alignments. Bioinformatics. 2011;27:757–63.
54. The UniProt Consortium, Bateman A, Martin M-J, Orchard S, Magrane M, Agivetova R, et al. UniProt: the universal protein knowledgebase in 2021. Nucleic Acids Res. 2021;49:D480–9.
55. Bastian FB, Roux J, Niknejad A, Comte A, Fonseca Costa SS, de Farias TM, et al. The Bgee suite: integrated curated expression atlas and comparative transcriptomics in animals. Nucleic Acids Res. 2021;49:D831–47.
56. Manni M, Berkeley MR, Seppey M, Simão FA, Zdobnov EM. BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes. Mol Biol Evol. 2021;38:4647–54.
57. Derelle R, Philippe H, Colbourne JK. Broccoli: Combining Phylogenetic and Network Analyses for Orthology Assignment. Mol Biol Evol. 2020;37:3389–96.
58. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. Nat Methods. 2015;12:59–60.
59. Price MN, Dehal PS, Arkin AP. FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments. PLoS ONE. 2010;5:10.
60. Singh PP, Isambert H. OHNOLOGS v2: a comprehensive resource for the genes retained from whole genome duplication in vertebrates. Nucleic Acids Res. 2019;gkz909.
61. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene Ontology: tool for the unification of biology. Nat Genet. 2000;25:25–9.
62. The Gene Ontology Consortium, Carbon S, Douglass E, Good BM, Unni DR, Harris NL, et al. The Gene Ontology resource: enriching a GOld mine. Nucleic Acids Res. 2021;49:D325–34.
63. Bray NL, Pimentel H, Melsted P, Pachter L. Near-optimal probabilistic RNA-seq quantification. Nat Biotechnol. 2016;34:525–7.
64. Soneson C, Love MI, Robinson MD. Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. F1000Research. 2015;4:1521.
65. Komljenovic A, Roux J, Robinson-Rechavi M, Bastian FB. BgeeDB, an R package for retrieval of curated expression datasets and for gene list expression localization enrichment tests. F1000Research. 2016;5:2748.

66. Yanai I, Benjamin H, Shmoish M, Chalifa-Caspi V, Shklar M, Ophir R, et al. Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. Bioinformatics. 2005;21:650–9.