Amphioxus functional genomics reveals the evolution of vertebrate regulatory traits

Ferdinand Marletaz [1,39,40], Panos N. Firbas [2,40], Ignacio Maeso [2,40], Juan J. Tena [2,40], Ozren Bogdanovic [3,4,5,40], Malcolm Perry [6,7,40], Chris D.R. Wyatt [8,9], Elisa de la Calle-Mustienes [2], Stephanie Bertrand [10], Demian Burguera [8,9,11], Rafael D. Acemel [2], Simon J. van Heeringen [12], Silvia Naranjo [2], Carlos Herrera-Ubeda [11], Ksenia Skvortsova [3], Sandra Jimenez-Gancedo [2], Daniel Aldea [2], Yamile Marquez [8,9], Lorena Buono [2], Iryna Kozmikova [13], Jon Permanyer [8,9], Alexandra Louis [14,15,16], Beatriz Albuixech-Crespo [11], Yann Le Petillon [10], Anthony Leon Florian [10], Lucie Subirana [10], Paul Edward Duckett [3], Ensieh Farahani [2], Jean Marc Aury [17], Sophie Mangenot [17], Patrick Wincker [18], Ricard Albalat [11], Èlia Benito-Gutiérrez [19], Cristian Cañestro [11], Filipe Castro [20], Salvatore D'Aniello [21], David E.K. Ferrier [22], Shengfeng Huang [23], Vincent Laudet [10], Gabriel A.B. Marais [24], Pierre Pontarotti [25], Michael Schubert [26], Hervé Seitz [27], Ildiko Somorjai [28], Tokiharu Takahashi [29], Olivier Mirabeau [30], Anlong Xu [23,31], Jr-Kai Yu [32], Piero Carninci [33,34], Juan Ramon Martinez-Morales [2], Hugues Roest Crollius [14,15,16], Zbynek Kozmik [13], Matt Weirauch [35,36], Jordi Garcia-Fernàndez [11], Ryan Lister [5,37], Boris Lenhard [6,7,38], Peter W.H. Holland [1], Hector Escriva [10,41], Jose Luis Gómez-Skarmeta [2,41], Manuel Irimia [8,9,41]

[1] Department of Zoology, University of Oxford, Oxford, UK.

[2] Centro Andaluz de Biología del Desarrollo (CABD), CSIC-Universidad Pablo de Olavide-Junta de Andalucía, Seville, Spain.

[3] Genomics and Epigenetics Division, Garvan Institute of Medical Research, Sydney, New South Wales, Australia.

[4] St Vincent's Clinical School, Faculty of Medicine, University of New South Wales, Sydney, New South Wales, Australia.

[5] Australian Research Council Centre of Excellence in Plant Energy Biology, School of Molecular Sciences, The University of Western Australia, 35 Stirling Hwy, Crawley, WA 6009, Australia.

[6] Institute of Clinical Sciences, Faculty of Medicine, Imperial College London, London W12 0NN, UK.

[7] Computational Regulatory Genomics, MRC London Institute of Medical Sciences, London W12 0NN, UK.

[8] Centre for Genomic Regulation (CRG), Barcelona Institute of Science and Technology (BIST), Dr Aiguader 88, Barcelona 08003, Spain.
[9] Universitat Pompeu Fabra (UPF), Barcelona, Spain.
[10] Sorbonne Universités, UPMC Univ Paris 06, CNRS, Biologie Intégrative des Organismes Marins (BIOM), Observatoire Océanologique, F-66650, Banyuls/Mer, France.
...
[40] Co-first authors
[41] Corresponding authors: HE, hescriva@obs-banyuls.fr; JLGS, jlgomska@upo.es; MI, mirimia@gmail.com
[Complete list of affiliations after Acknowledgements]

Running Title: Functional evolutionary genomics of amphioxus

Key words: whole genome duplication, regulatory genomics, transcriptomics, DDC model, complexity, hourglass model, chordate evolution

ABSTRACT

All chordates share a fundamental bodyplan that was greatly elaborated in vertebrates. Vertebrates also evolved highly distinctive genomes, sculpted by two whole genome duplications (WGD). To investigate the evolution of genome regulation in chordates, we characterized promoters, DNA methylation, chromatin accessibility, histone modifications and transcriptomes in multiple tissues and throughout development of the cephalochordate amphioxus. These data revealed an intermediate stage in the evolution of differentially methylated regulatory elements, and high conservation of gene expression and its underlying *cis*-regulatory logic between amphioxus and vertebrates, maximally at a developmental phylotypic period. We also unraveled the principal route of regulatory evolution following WGD: over 80% of broadly expressed gene families with multiple paralogs in vertebrates have members that restricted their ancestral expression, undergoing specialization rather than subfunctionalization. Counter-intuitively, vertebrate genes that underwent expression restriction increased the complexity of their regulatory landscapes. Altogether, these data pave the way for a better understanding of the regulatory principles underlying key vertebrate innovations.

INTRODUCTION

All vertebrates share multiple morphological and genomic novelties [1,2]. The most prominent genomic difference from non-vertebrate chordates is the reshaping of the gene complement that followed two rounds of whole genome duplication (WGD or 2R), that likely occurred at the base of the vertebrate lineage [3,4]. These large-scale mutational events are hypothesized to have facilitated the evolution of vertebrate morphological innovations, at least in part through the preferential retention of 'developmental' gene families and transcription factors (TF) after duplication [4,5]. However, duplicate genes and their associated regulatory elements were initially identical, and could not drive innovation without regulatory and/or protein-coding changes.

To date, the impact of vertebrate WGDs on gene regulation remains obscure, both concerning the fates of duplicate genes and the acquisition of the unique genomic traits that are characteristic of vertebrate genomes. These traits include numerous features often associated with gene regulation, such as unusually large intergenic and intronic regions [6,7], a distinct set

of highly conserved non-coding regions (HCNRs) [5,8], and high global 5-methylcytosine (5mC) content and 5mC-dependent regulation of embryonic transcriptional enhancers [9].

To investigate the evolution and origins of vertebrate gene regulation, appropriate species must be used for comparison. Previous studies have largely focused on phylogenetic ranges that are either too short (e.g. human vs. mouse) or too long (e.g. human vs. fly, human vs. nematode), resulting in limited insights into the origins of vertebrate genome regulation. In the first case, comparisons among closely related species (e.g. mammals [10-16] or *Drosophila* species [17-20]), for which orthology of non-coding regions can be readily determined from genome alignments, have allowed fine-grained analyses of TF binding evolution. These studies have revealed high rates of repurposing and TF binding turnover [11,12,15,21], despite high conservation of tissue-dependent expression [22,23]. However, turnover rates and evolutionary conservation vary considerably across tissues/cell types [15,24], developmental stages [24] and types and location of regulatory elements (i.e. promoters versus enhancers, proximal versus distal) [11-13,19,24]. Moreover, it is unclear if similar evolutionary trends are present outside mammals and flies and at larger evolutionary distances. In the second case, three-way comparisons of human, fly and nematode by the modENCODE consortium revealed no detectable conservation at the *cis*-regulatory level [25] and very little conservation of gene expression [26]. Moreover, a wealth of studies has demonstrated that the genomes of fruitflies and nematodes are highly derived [27-29]. Thus, comprehensive functional genomic data for a slow-evolving, closely related outgroup is still missing for a proper investigation of the origins and evolution of the vertebrate regulatory genome. Furthermore, only by comparison between vertebrates and a closely related outgroup will it be possible to elucidate the impact of WGDs on gene regulation.

Unlike flies, nematodes and most non-vertebrates, amphioxus belongs to the chordate phylum. Therefore, although it lacks specializations and innovations of vertebrates, it shares with them a basic body plan, including a dorsal neural tube with an anterior brain, segmented somites, notochord, and a ventral gut with a hepatic diverticulum homologous to the vertebrate liver [1] (see Supplementary Information for further details). For these reasons, amphioxus has been widely used as a reference outgroup to infer ancestral versus novel features during vertebrate evolution. Here, to investigate how the unique functional genome architecture of vertebrates has evolved, we undertook a comprehensive study of the transcriptome and regulatory genome of amphioxus.

RESULTS

Functional genome annotation of amphioxus

We generated a deep resource of genomic, epigenomic and transcriptomic data for the Mediterranean amphioxus (*Branchiostoma lanceolatum*) (Fig. 1a and Supplementary Datasets 1-5). We characterized chromatin accessibility using Assay for Transposase-Accessible Chromatin sequencing (ATAC-seq), genome-wide 5mC patterns using MethylC-seq, and transcription start sites using Cap Analysis of Gene Expression followed by high throughput sequencing (CAGE-seq) for multiple amphioxus developmental stages and adult tissues (Fig. 1a). We also used ChIP-seq to determine the locations of specific histone modifications associated with distinct functional chromatin states for three developmental stages, focusing on H3K4me3 (active promoters), H3K27ac (active promoters and enhancers), and H3K27me3 (polycomb-repressed regulatory regions, typically associated with developmentally regulated genes). All these datasets were complemented with deep coverage RNA sequencing (RNA-seq) for 16 developmental stages and 9 adult tissues (Fig. 1a), comprising a total of 52 individual samples. The ATAC-seq, MethylC-seq, CAGE-seq, ChIP-seq and RNA-seq datasets were mapped to a *de novo* sequenced and assembled *B. lanceolatum* genome with 150x coverage, a total size of 495.4 Mbp, a scaffold N50 of 1.29 Mbp and 4% of gaps (Extended Data Fig. 1a-c, Supplementary Tables 1 and 2; Supplemental Information). To facilitate access by the research community, we integrated these resources into a UCSC hub (Fig. 1b; http://amphiencode.github.io), together with an intra-cephalochordate sequence conservation (phastCons) track derived from a 3-species amphioxus genome alignment, an annotation of transposable elements and other repeated sequences (Extended Data Fig. 1d), which represent 32% of the genome and are mostly unmethylated (Extended Data Fig. 1e,f), and a catalog of high confidence long non-coding RNAs (Extended Data Fig. 1g, Supplementary Dataset 6). To enable broader evolutionary comparisons, we reconstructed a set of orthologous gene families for multiple vertebrate and non-vertebrate species (Supplementary Dataset 7), and generated several equivalent datasets for zebrafish and medaka (Extended Data Fig. 2a), complemented with published data for other vertebrate species (Supplementary Datasets 2 and 3). In addition, local and genome-wide orthology and paralogy syntenic links between amphioxus and ten other genomes can be visualised on a dedicated Genomicus server [30] (Extended Data Fig. 1h).

A comprehensive functional annotation of the *B. lanceolatum* genome identified 88,391 putative DNA *cis*-regulatory elements (PREs) defined using ATAC-seq, and 20,569 orthology-supported protein-coding genes (Supplemental Information). We divided these PREs into promoters (from -1,000 bp to +500 bp from a transcription start site [TSS, which were highly supported by CAGE-seq data, Extended Data Fig. 2b]), and gene body, proximal and distal PREs (inside gene models or within or further than 5 kbp away from TSSs, respectively) (Fig. 1c). Equivalent analyses using zebrafish data yielded 256,018 potential regulatory regions, with a significantly higher proportion of these being distal PREs (Fig. 1c; $p < 2.2 \times 10^{-16}$, Fisher's exact test of distal vs. non-distal PREs). A significantly larger overall PRE-TSS distance distribution was observed for all vertebrates compared to amphioxus (Fig. 1d), even when correcting by the average intergenic distance of each species (Extended Data Fig. 2c) ($p<2.2\times10^{-16}$ for all vertebrate-vs-amphioxus comparisons, one-sided Mann Whitney tests). Amphioxus PREs showed a similar enrichment for enhancer-associated chromatin marks to previously validated enhancers (Extended Data Fig. 2d and Supplementary Table 3). Consistently, 10/11 tested PREs drove specific (6) or general (4) GFP expression in transgenic zebrafish assays (Fig. 1e and Extended Data Fig. 2e), and 3/3 specific expression in transgenic amphioxus assays (Extended Data Fig. 2f). Moreover, 32/36 (89%) of previously reported amphioxus enhancers overlapped PREs defined by our ATAC-seq data (Supplementary Table 3). Therefore, a significant fraction of PREs likely act as developmentally regulated transcriptional enhancers. To further investigate the nature of our PRE catalog, we analyzed their H3K27 acetylation dynamics during development using ChIP-seq data. Between 36.3 and 43.9% of PREs were active at a given developmental stage based on H3K27ac levels (Extended Data Fig. 2g), and a substantial fraction of these became active/inactive during the time course (Extended Data Fig. 2h). Interestingly, dynamic PREs were enriched for distinct TF binding motifs, including pluripotency- and differentiation-associated TFs in early and late active PREs, respectively (Extended Data Fig. 2i). In contrast, PREs that were inactive across the time course were enriched for CTCF motifs, among others, suggesting PREs also encompass other types of regulatory elements.

WGD-assisted disentanglement of bidirectional promoters in vertebrates

Dedicated analysis of CAGE-seq data defined core promoters at a single nucleotide resolution, revealing that amphioxus promoters display a mixture of pan-metazoan, pan-vertebrate and unique features. At the sequence level, ubiquitous promoters had a unique architecture, not described in any other organism, characterized by a WW (where W is A or

T) dinucleotide enrichment preceding the TSS followed by two stretches of SS (C or G) dinucleotide-enriched regions, and with a nucleosome positioned in between (around position +90 from the TSS) (Extended Data Fig. 3a,d). Moreover, the dominant TSS was usually asymmetrically located in the 5'-most region of the promoter (Extended Data Fig. 3a). On the other hand, as in other metazoans, ubiquitous promoters were often broad (i.e. with multiple TSS positions within a stretch of sequence) (Extended Data Fig. 3a), and, as in vertebrates, were enriched in YY1 motifs (particularly the narrowest ones) (Extended Data Fig. 3e,f). In contrast, embryo- and tissue-specific promoters did not share these features (Extended Data Fig. 3b,c), and had similar architectures at the sequence level to those described for other metazoans [31-33].

We also identified a large fraction of bidirectional promoters: 3,950 out of 15,884 (25%) pairs of promoters associated with two neighboring protein-coding genes fell within 1 kbp of each other and were in opposite orientations (Extended Data Fig. 4a and Supplementary Information). Pairs of bidirectional promoters displayed a marked inter-promoter distance periodicity (Extended Data Fig. 4b), with a period consistent with the spacing of zero, one or two nucleosomes, based on NucleoATAC signal [34] (Extended Data Fig. 4c). Bidirectional promoters were most common among ubiquitous promoters (Extended Data Fig. 4a), and the associated genes were significantly enriched in housekeeping functions (Extended Data Fig. 4d). Notably, the fraction of CAGE-seq-supported bidirectional promoters decreased progressively from amphioxus to mouse (1,752/13,654, 12.83%; $p < 2.2 \times 10^{-16}$, Fisher's exact test) and to zebrafish (1,098/14,014, 7.84%; $p < 2.2 \times 10^{-16}$, Fisher's exact test), suggesting that differential paralog elimination after each round of WGD (two in tetrapods, three in teleosts) resulted in a disentangling of a large fraction of the bidirectional promoters present in the chordate ancestor. To test this possibility, we identified 372 protein-coding gene pairs with high-confidence orthologs in all three species that were likely arranged as bidirectional promoters in the last common chordate ancestor (Extended Data Fig. 4e, Supplementary Materials). As expected, most of these ancestral bidirectional promoters were lost in vertebrates, particularly in stem vertebrates (54.5%), with only very few amphioxus-specific losses (5.3%). This disentanglement, however, was not accompanied in vertebrates by a general increase in the fraction of bidirectional promoters with antisense non-coding transcription (Extended Data Fig. 4f).

Developmental demethylation of PREs precedes the origin of vertebrates

To understand the evolution of the DNA methylation-mediated gene regulatory processes that are characteristic of vertebrates, we next investigated the pattern and dynamics of 5mC during amphioxus embryonic development and adulthood. Similar to other non-vertebrates [35-37], the majority of the amphioxus genome exhibited very low levels of CpG methylation (Fig. 2a; only ~30% of CpG sites were methylated at any stage, compared to ~93% in zebrafish), with nearly all the 5mC occurring in gene bodies, where the proportion of methylated CpGs correlated positively with gene expression levels (Fig. 2b), but negatively with H3K27me3, H3K4me3 and CpG density (Extended Data Fig. 5a,b).

Despite the overall low levels of 5mC in amphioxus, there was notable similarity with vertebrates in terms of developmental dynamics. As in zebrafish and frogs [9], global levels of 5mC displayed a slight developmental decrease (Extended Data Fig. 5c). Accordingly, 77% (22,333/28,972) of differentially methylated regions (DMRs; Supplementary Dataset 8) showed decreased 5mC in adult hepatic diverticulum compared to 36 hours post-fertilization (hpf) embryos (liver hypo-DMRs; Extended Data Fig. 5d-f). This is also consistent with the onset of expression of the single amphioxus *Tet* ortholog, involved in 5mC demethylation, at mid-late neurula stages (Extended Data Fig. 5g). To assess whether some of these DMRs may have regulatory potential, we used ATAC-seq data to identify hepatic-specific PREs that are not active during development. Clustering these PREs based on 5mC content revealed two distinct subsets, one with and one without hepatic-specific hypomethylation (Fig. 2c and Supplementary Table 4); a similar clustering pattern was not observed for embryo-specific PREs (Extended Data Fig. 6a). Both groups of hepatic-specific PREs were enriched for binding sites of liver-specific TFs, such as *Hnf4a* [38], as well as broadly expressed TFs like *Foxa* [39] (Extended Data Fig. 6b). Interestingly, differentially methylated PREs (cluster 1) also displayed significant hypomethylation in the three other studied adult tissues (gut, muscle, notochord), as assessed by reduced representation bisulfite sequencing (RRBS), suggestive of an organism-wide demethylation event at these PREs in adults (Fig. 2d). Notably, the broadly expressed *Foxa* has previously been shown to act as a pioneer factor, participating in 5mC removal at regulatory regions in various mammalian cell types via diverse mechanisms [40,41].

PREs from both clusters were preferentially associated with genes with metabolic functions (Extended Data Fig. 6c). However, whereas the PREs with no differential 5mC (cluster 2) were associated with hepatic-specific genes, the PREs with specific hepatic hypomethylation (cluster 1) were primarily associated with genes that displayed widespread expression (Fig. 2e

and Extended Data Fig. 6d) with low intra-tissue variation (Extended Data Fig. 6e). Moreover, they were mainly located within gene bodies, unlike other ATAC-seq peaks (Fig. 2f). Altogether, these data suggest that demethylation of these PREs may contribute to their identification as adult-specific transcriptional *cis*-regulatory elements within continuously hypermethylated gene-body contexts, characteristic of non-vertebrate species. Interestingly, we found differentially methylated PREs in orthologous zebrafish introns for 14 of the differentially methylated PREs identified in amphioxus, indicating potential evolutionary conservation. Strikingly, these 14 genes included four components of the Hippo pathway, including orthologs of the transcriptional effectors *Yap* and *Tead* (Fig. 2g and Extended Data Fig. 6f-i; Supplementary Table 5). Additional cases included genes that harbored PREs that are likely to regulate neighboring liver-specific genes ("bystander" genes; Supplementary Information). In summary, although amphioxus largely shows a non-vertebrate-like global methylation pattern, it presents the first detected occurrence in a non-vertebrate species of vertebrate-like PREs that are regulated in tight association with differential 5mC during development.

A phylotypic period in chordate embryogenesis

To investigate the evolutionary conservation of chordate development at the molecular level, we next conducted transcriptome comparisons across species and stages. Previous comparative analyses among vertebrate transcriptomes [42,43] showed a developmental period of maximal similarity in gene expression, coinciding with the so-called vertebrate phylotypic period, in agreement with the hourglass model [44,45]. However, similar comparisons with tunicates and amphioxus have thus far not resolved a phylotypic period shared across all chordates [43]. To improve the resolution of the transcriptomic comparisons, we increased the number and sequencing depth of sampled amphioxus developmental stages, and tested whether a period of maximal gene expression similarity exists between amphioxus and vertebrates. Pairwise comparisons of stage-specific RNA-seq data from developmental time courses of amphioxus against zebrafish, medaka, frog and chicken (Supplementary Table 6) revealed a consistent period of higher similarity between amphioxus and all vertebrate species (Fig. 3a,b and Extended Data Fig. 7), corresponding to the 4-7 somite neurula (18-21 hpf). In vertebrates, the stages with the highest similarity to this amphioxus stage occurred slightly earlier than those reported as the vertebrate phylotypic period [42,43]. Interestingly, the amphioxus phylotypic period matches the onset of expression of *Tet* demethylase, as reported also for vertebrates [9]. We also made pairwise comparisons between relative TF motif

enrichment in dynamic ATAC-seq peaks active at each stage (Supplementary Information). This was also consistent with the hourglass model, showing that the two most similar stages in terms of the motif content of their active PREs were those preceding the phylotypic period identified above (Fig. 3c; early neurula [15hpf] amphioxus and 80% epiboly zebrafish). In contrast, comparisons within amphioxus species showed that the sequence conservation for the same ATAC-seq-defined genomic regions was slightly higher at pre-mouth larva (36hpf), after the putative chordate phylotypic period (Fig. 3d).

To further examine the extent of gene expression conservation between developmental stages of amphioxus and vertebrates, we next selected eight equivalent embryonic stages in amphioxus and zebrafish based on landmarks such as fertilization, gastrulation and organogenesis (Supplementary Table 7). We used Mfuzz [46] clustering to identify sets of genes with similar temporal expression profiles across these comparable time courses in each species. Subsequent pairwise comparisons between amphioxus and zebrafish revealed pairs of clusters with highly significant overlap of orthologous genes (Extended Data Fig. 8a). In most cases, these profiles had very similar temporal dynamics with respect to the equivalent developmental landmarks, despite markedly distinct cell type compositions and differentiation dynamics of the embryos. Interestingly, clusters with significant ortholog overlap that showed similar temporal dynamics in both species were highly enriched for "intracellular" gene functions and components, such as nucleic acid binding and nucleus, whereas the heterochronic clusters were enriched for membrane-related and extracellular functions and components (Extended Data Fig. 8b and Supplementary Dataset 9; Supplementary Methods). When looking specifically at developmental pathways, genes from the Hedgehog and Hippo pathway more often fell into homochronic cluster pairs with significant ortholog overlap (Extended Data Fig. 8c). Altogether, these comparative transcriptomic analyses of chordate development reveal extensive conservation of temporal gene expression between amphioxus and vertebrate development, with a putative chordate phylotypic period displaying the highest resemblance.

Regulatory conservation underlying the adult chordate body plan

To have a first general assessment of the extent of conservation or divergence in gene expression among chordates at adult stages, we used Neighborhood Analysis of Conserved Co-expression (NACC)[47], a method developed to compare heterogeneous, non-matched sample sets across species. Using human as reference, we found significant conservation of

vertebrate and amphioxus orthologous gene expression patterns compared to gene sets with randomized orthology relationships (Fig. 4a). To investigate this conservation in more detail, we next compared gene co-expression profiles in a wide range of adult tissues and a subset of embryonic stages from amphioxus and zebrafish (Supplementary Table 8). We used Weighted Gene Co-expression Network Analysis (WGCNA)[48] to identify sets of genes (modules) with co-regulated expression in each species (Supplementary File 1). Pairwise comparison of these modules between the two species revealed multiple pairs with highly significant levels of ortholog overlap (Fig. 4b). These included modules with conserved tissue-specific expression that were enriched for coherent GO categories (Supplementary File 1). For example, one pair (① in Fig. 4b,c) comprises genes with high expression in organs with ciliated cells (e.g. spermatozoa, gill bars, etc.), reflected by enrichment in cilium, microtubule and cell projection GO terms in both species (Fig. 4d). Similarly, we observed conserved neural (② in Fig. 4b,c,e), as well as muscle, gut, hepatic, skin and metabolism-related modules (Supplementary File 1).

To assess whether this transcriptomic conservation was mirrored at the *cis*-regulatory level, we performed similar pairwise comparisons between modules based on relative TF binding site motif enrichment. For this, we scanned for TF motifs in ATAC-seq peaks located in the proximal regions (-5/+1 kbp around each gene's primary TSS), and calculated their relative enrichment within each WGCNA module (Supplementary Information). We found a significant positive correlation between relative motif enrichment scores for a large fraction of the inter-specific pairs of modules with high transcriptomic similarity (Fig. 4c). In such cases, the most enriched TF motifs within each cluster were highly consistent between amphioxus and zebrafish. These included TFs with well-known roles in tissue-specific development and differentiation (e.g. *Rfx* for cilia, *Hox* for brain, *Hnf1a* for liver and gut, *Ghrl* for skin, *Mef2* for muscle, and *Elf1* and *Spic* for immune function) (Fig. 4f, Supplementary File 1, and Supplementary Dataset 10). Altogether, our results show a high level of transcriptomic and *cis*-regulatory conservation underlying basic cellular processes and differentiated tissues in adult amphioxus and vertebrates.

Higher regulatory information in vertebrate genomes

Amphioxus represents an ideal outgroup for investigating the impact of WGDs on gene regulatory complexity in vertebrates. We thus first asked whether the number of putative regulatory regions is higher per gene in vertebrates than in amphioxus (Supplementary

Information). When comparing two developmental stages just prior to the proposed chordate phylotypic period, we observed significantly more ATAC-seq peaks (i.e. PREs) per gene regulatory landscape (as defined by GREAT [49]) in zebrafish than in amphioxus (Fig. 5a; $p < 2.2 \times 10^{-16}$, Wilcoxon Sum Rank test). This difference is particularly evident for gene families that have retained multiple copies after genome duplication (1:2, 1:3 or 1:4 gene number ratio), but is also clearly detectable for 1:1 orthologs (Fig. 5b). Moreover, in gene families with two, three or four ohnologs (paralogs derived from the WGDs), the number of ATAC-seq peaks is very uneven between ohnologs: the paralog with the lowest number of associated elements generally has a comparable number to the amphioxus ortholog, but dramatic regulatory expansions were observed for some ohnologs (Fig. 5c). The same patterns were detected for all amphioxus and zebrafish developmental stages, and also observed for medaka and mouse genomes (Extended Data Fig. 9a). These differences were robust to depth of sequencing, and were still significant after downsampling vertebrate ATAC-seq reads down to 15-20% of the effective coverage in amphioxus (Fig. 5d, Extended Data Fig. 9b; Supplementary Information). We also detected a higher number of peaks associated with regulatory (*trans-dev*) genes compared to housekeeping genes in all species (Extended Data Fig. 9c), consistent with their higher frequency of retention in multiple copies after WGD [4] (Fig. 5b). Furthermore, using circular chromosome conformation capture followed by sequencing (4C-seq) to experimentally determine the span of the regulatory landscapes for 58 genes from eleven *trans-dev* gene families in amphioxus, zebrafish and mouse we also found a significantly higher numbers of ATAC-seq peaks in both vertebrate species than in amphioxus (Extended Data Fig. 9d, Supplementary Table 9).

As expected [7], the higher number of PREs in zebrafish was associated with larger intergenic regions in this species (Extended Data Fig. 9e). However, the differences in PRE complements were not attributable only to a large increase in genome size in vertebrates, as subsets of amphioxus and zebrafish genes with matched length distributions of GREAT or intergenic regions displayed a higher number of ATAC-seq peaks in the latter species (Extended Data Fig. 9f,g; $p < 2.2 \times 10^{-16}$, Wilcoxon rank sum test). Further investigation of matched distributions show that these differences are particularly strong in genes with large regulatory landscapes (>50 kbp), which have a much higher number of ATAC-seq peaks in zebrafish than in amphioxus (Fig. 5e). Thus, although the number of ATAC-seq peaks and GREAT region size are positively correlated in all species, larger regions in amphioxus do not scale at the same rate as in vertebrates (Fig. 5f,g), consistent with the lower proportion of

distal PREs identified in this species (Fig. 1c,d). In summary, these analyses reveal that there was a large increase in the number of regulatory regions in the evolution of vertebrates (and/or a decrease in amphioxus), particularly of distal regulatory elements, and that this trend is enhanced for specific retained gene copies after the WGDs, pointing at unequal rates of regulatory evolution for different ohnologs.

Increased regulatory complexity in functionally specialized ohnologs

The Duplication-Degeneration-Complementation (DDC) model hypothesizes that retention of duplicate genes could be driven by reciprocal loss of regulatory elements and restriction of paralogs to distinct subsets of the ancestral expression domain [50]. Although intuitively attractive, this hypothesis has been difficult to test for the vertebrate WGDs due to lack of transcriptomic and regulatory data from appropriate outgroups. DDC predicts that individual gene duplicates would each have more restricted expression than an unduplicated outgroup, but their summation would not. To investigate this, we focused on seven homologous tissues and two equivalent developmental stages in amphioxus, zebrafish, frog and mouse (Supplementary Table 10), and binarized the expression (on or off) of each gene in each sample (i.e. expression domain) based on fixed cut-offs (Fig. 6a and Supplementary Information). When comparing genes that returned to single copy status after vertebrate WGDs (2,478 gene families), we detected no difference between amphioxus and vertebrates in expression bias, as measured by subtracting the number of positive expression domains in amphioxus from that of vertebrates (Fig. 6a,b and Extended Data Fig. 10a,b). In contrast, when vertebrate genes from families with multiple ohnologs were compared to their single amphioxus ortholog, the distributions were strongly skewed, with many vertebrate genes displaying far more restricted expression domains (Fig. 6b and Extended Data Fig. 10a,b). Remarkably, the symmetrical pattern was fully recovered when the expression of all vertebrate members was combined or the raw expression values summed for each member within a paralogy group (Fig. 6a,b and Extended Data Fig. 10a,b). Consistent with these results, comparison of *Tau* values (an alternative measure of gene expression restriction [51]) between amphioxus and vertebrate orthologs showed a marked bias towards higher *Tau* values in vertebrates (Extended Data Fig. 10c-e). These analyses indicate that, when multiple genes are retained after WGD, many gene copies restrict their expression domains, but the joint expression of all members is similar to that of the single amphioxus ortholog, and likely of the ancestral gene.

Although the above findings are consistent with the DDC model, they are also compatible with an alternative model in which a subset of duplicate genes becomes more 'specialized' in expression pattern while one or more paralogs retain the ancestral broader expression. To distinguish between these alternatives, we analyzed a subset of multi-gene families in which both the single amphioxus ortholog and the union of the vertebrate ohnologs were expressed across all nine compared samples (Supplementary Information). We then identified (Fig. 6c): (i) gene families in which all vertebrate paralogs were expressed in all domains ('redundancy'), (ii) gene families in which none of the vertebrate members had expression across all domains ('subfunctionalization')[50], and (iii) gene families in which one or more vertebrate ohnologs were expressed in all domains, but at least one ohnolog was not ('specialization'). We obtained very similar results for the three studied vertebrate species (Fig. 6d): between 80 and 88% of gene families fell into categories (ii) or (iii), with loss of ancestral expression domains in at least one member. Moreover, we found specialization (iii) to be consistently more frequent than subfunctionalization (ii) as a fate for vertebrate ohnologs with broad ancestral expression.

Interestingly, ohnologs that have experienced strong specialization (defined as having two or fewer remaining expression domains) have more often retained expression in neural tissues, followed by testis, particularly in mammals (Fig. 6e and Extended Data Fig. 10f,g; representative examples are illustrated by *in situ* hybridization in Extended Data Fig. 10h,i and Supplementary Information). Furthermore, they showed the fastest rates of sequence evolution and the highest dN/dS ratio between human and mouse, whereas genes from redundant families and those ohnologs from specialized families that retain the full ancestral expression displayed the lowest levels of sequence divergence (Fig. 6f and Extended Data Fig. 10j-l). This is consistent with strongly specialized ohnologs having their coding sequence optimized to perform their function in a specific tissue, and also with the evolution of novel functions (neofunctionalization). Strikingly, we found that ohnologs from specialized families that have lost expression domains showed significantly more associated regulatory elements than those with the full ancestral expression (Fig. 6g). In fact, we observed a strong positive relationship between the number of ancestral expression domains lost and the number of putative regulatory elements associated with specialized ohnologs (Extended Data Fig. 10m). This implies that specialization of gene expression after WGD does not occur primarily through loss of ancestral tissue-specific regulatory elements, but rather by complex

remodeling of regulatory landscapes involving recruitment of novel tissue-specific regulatory elements.

DISCUSSION

By surveying functional genomic elements in the cephalochordate amphioxus and deploying novel comparative approaches, we have deepened our understanding of the origin and evolution of chordate genome regulation and function. First, we identified PREs in amphioxus whose activation is tightly associated with differential DNA demethylation in adult tissues, a mechanism previously thought to be vertebrate-specific. Previously published non-vertebrate methylomes either compared various species at a single developmental time point or within adult tissues [35,52-54], or interrogated developmental 5mC dynamics without exploring active chromatin marks or chromatin accessibility [55]. Therefore, it is possible that this phenomenon may not be limited to amphioxus, and that it may be subsequently found in other non-vertebrate species. In amphioxus, such elements usually fall within gene bodies of widely expressed genes, suggesting that gene regulation by demethylation could have originated as a mechanism to allow better definition of enhancers in a hypermethylated intragenic context. Assuming this as the ancestral scenario, this mechanism would have been co-opted into new genomic contexts (i.e. demethylation of distal intergenic enhancers, which are much more numerous in vertebrates than in amphioxus) later in the evolution of vertebrate genomes, which are characterized by their pervasive, genome-wide hypermethylation. Second, we found remarkable conservation at the transcriptomic level in embryonic development between amphioxus and vertebrates, and in adult tissue identity, which is mirrored at the *cis*-regulatory level by putative binding sites for multiple tissue-specific TFs. These results provide a regulatory framework accounting for overall body plan conservation amongst chordates, showing that organs such as brain and liver, which have very different rates of TF turnover and repurposing in vertebrates [15,24,56], can nevertheless both show deep regulatory conservation across chordates. Third, we identified a consistently higher number of open chromatin regions per gene in vertebrates than in amphioxus. This pattern is observed at a genome-wide level, but is particularly evident for distal PREs and in retained ohnologs after WGD, which are enriched for regulatory genes with large regulatory landscapes. This would be consistent with the regulatory priming hypothesis [57], in which the potential to accumulate distal regulatory elements in Topological Associating Domains (TADs) could only be fully realized after the release of ancestral regulatory constraints by the retention of multiple ohnologs. Finally, we detected a large degree of expression specialization for retained

ohnologs, with the vast majority of multi-gene families with broad ancestral expression having at least one member that restricted its breadth of expression. Through this mechanism, vertebrates have increased their repertoire of tightly regulated genes, potentially contributing to tissue- and organ-specific evolution. Remarkably, gene expression specialization was accompanied by faster protein-coding sequence evolutionary rates, and an increase in the number of regulatory elements rather than a decrease. Taken together, these observations indicate that the two rounds of WGD not only caused an expansion and diversification of gene repertoires in vertebrates, but also allowed functional and expression specialization of the extra copies by increasing the complexity of their gene regulatory landscapes. We suggest that these changes to the gene regulatory landscapes underpinned the evolution of morphological specializations in vertebrates.

REFERENCES

1 Bertrand, S. & Escriva, H. Evolutionary crossroads in developmental biology: amphioxus. *Development* **138**, 4819-4830 (2011).

2 Janvier, P. Facts and Fancies about Early Fossil Chordates and Vertebrates. *Nature* **520**, 483-489 (2015).

3 Dehal, P. & Boore, J. L. Two rounds of whole genome duplication in the ancestral vertebrate. *PLoS Biol* **3**, e314 (2005).

4 Putnam, N. *et al.* The amphioxus genome and the evolution of the chordate karyotype. *Nature* **453**, 1064-1071 (2008).

5 Holland, L. Z. *et al.* The amphioxus genome illuminates vertebrate origins and cephalochordate biology. *Genome Res* **18**, 1100-1111 (2008).

6 Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860 - 921 (2001).

7 Nelson, C., Hersh, B. & Carroll, S. The regulatory content of intergenic DNA shapes genome architecture. *Genome Biology* **5**, R25 (2004).

8 Vavouri, T. & Lehner, B. Conserved noncoding elements and the evolution of animal body plans. *Bioessays* **31**, 727-735 (2009).

9 Bogdanović, O. *et al.* Active DNA demethylation at enhancers during the vertebrate phylotypic period. *Nat Genet* **48**, 417-426 (2016).

10 Berthelot, C., Villar, D., Horvath, J. E., Odom, D. T. & Flicek, P. Complexity and conservation of regulatory landscapes underlie evolutionary resilience of mammalian gene expression. *Nat Ecol Evol* **2**, 152-163 (2018).

11 Cotney, J. *et al.* The evolution of lineage-specific regulatory activities in the human embryonic limb. *Cell* **154**, 185-196 (2013).

12 Reilly, S. K. *et al.* Evolutionary genomics. Evolutionary changes in promoter and enhancer activity during human corticogenesis. *Science* **347**, 1155-1159 (2015).

13 Villar, D. *et al.* Enhancer evolution across 20 mammalian species. *Cell* **160**, 554-566 (2015).

14 Stergachis, A. B. *et al.* Conservation of trans-acting circuitry during mammalian regulatory evolution. *Nature* **515**, 365-370 (2014).

15 Vierstra, J. *et al.* Mouse regulatory DNA landscapes reveal global principles of cis-regulatory evolution. *Science* **346**, 1007-1012 (2014).

16 Zhou, X. *et al.* Epigenetic modifications are associated with inter-species gene expression variation in primates. *Genome Biol* **15**, 547 (2014).
17 Bradley, R. K. *et al.* Binding site turnover produces pervasive quantitative changes in transcription factor binding between closely related Drosophila species. *PLoS Biol* **8**, e1000343 (2010).
18 He, Q. *et al.* High conservation of transcription factor binding and evidence for combinatorial regulation across six Drosophila species. *Nat Genet* **43**, 414-420 (2011).
19 Paris, M. *et al.* Extensive divergence of transcription factor binding in Drosophila embryos with highly conserved gene expression. *PLoS Genet* **9**, e1003748 (2013).
20 Khoueiry, P. *et al.* Uncoupling evolutionary changes in DNA sequence, transcription factor occupancy and enhancer activity. *Elife* **6**, pii: e28440 (2017).
21 Odom, D. T. *et al.* Tissue-specific transcriptional regulation has diverged significantly between human and mouse. *Nat Genet* **39**, 730-732 (2007).
22 Chan, E. T. *et al.* Conservation of core gene expression in vertebrate tissues. *J Biol* **8**, 33 (2009).
23 Brawand, D. *et al.* The evolution of gene expression levels in mammalian organs. *Nature* **478**, 343-348 (2011).
24 Nord, A. S. *et al.* Rapid and pervasive changes in genome-wide enhancer usage during mammalian development. *Cell* **155**, 1521-1531 (2013).
25 Boyle, A. P. *et al.* Comparative analysis of regulatory information and circuits across distant species. *Nature* **512**, 453-456 (2014).
26 Gerstein, M. B. *et al.* Comparative analysis of the transcriptome across distant species. *Nature* **512**, 445-448 (2014).
27 Hendrich, B. & Tweedie, S. The methyl-CpG binding domain and the evolving role of DNA methylation in animals. *Trends Genet* **19**, 269-277 (2003).
28 Irimia, M. *et al.* Extensive conservation of ancient microsynteny across metazoans due to cis-regulatory constraints. *Genome Res* **22**, 2356-2367 (2012).
29 Simakov, O. *et al.* Insights into bilaterian evolution from three spiralian genomes. *Nature* **493**, 526-531 (2013).
30 Muffato, M., Louis, A., Poisnel, C. E. & Crollius, H. R. Genomicus: a database and a browser to study gene synteny in modern and ancestral genomes. *Bioinformatics* **26**, 1119-1121 (2010).
31 Celniker, S. E. *et al.* Unlocking the secrets of the genome. *Nature* **459**, 927-930 (2009).
32 Nepal, C. *et al.* Dynamic regulation of the transcription initiation landscape at single nucleotide resolution during vertebrate embryogenesis. *Genome Res* **23**, 1938-1950 (2013).
33 Forrest, A. R. *et al.* A promoter-level mammalian expression atlas. *Nature* **507**, 462-470 (2014).
34 Schep, A. N. *et al.* Structured nucleosome fingerprints enable high-resolution mapping of chromatin architecture within regulatory regions. *Genome Res* **25**, 1757-1770 (2015).
35 Wang, X. *et al.* Genome-wide and single-base resolution DNA methylomes of the Pacific oyster Crassostrea gigas provide insight into the evolution of invertebrate CpG methylation. *BMC Genomics* **15**, 1119 (2014).
36 Albalat, R., Martí-Solans, J. & Cañestro, C. DNA methylation in amphioxus: from ancestral functions to new roles in vertebrates. *Brief Funct Genomics* **11**, 142-155 (2012).
37 Huang, S. *et al.* Decelerated genome evolution in modern vertebrates revealed by analysis of multiple lancelet genomes. *Nat Commun* **5**, 5896 (2014).

584 38 Odom, D. T. *et al.* Control of pancreas and liver gene expression by HNF transcription
585 factors. *Science* **303**, 1378-1381 (2004).
586 39 Aldea, D., Leon, A., Bertrand, E. & Escriva, H. Expression of Fox genes in the
587 cephalochordate Branchiostoma lanceolatum. *Front Ecol Evol* **3**, 80 (2015).
588 40 Zhang, Y. *et al.* Nucleation of DNA repair factors by FOXA1 links DNA
589 demethylation to transcriptional pioneering. *Nat Genet* **48**, 1003-1013 (2016).
590 41 Yang, Y. A. *et al.* FOXA1 potentiates lineage-specific enhancer activation through
591 modulating TET1 expression and function. *Nucleic Acids Res* **44**, 8153-8164 (2016).
592 42 Irie, N. & Kuratani, S. Comparative transcriptome analysis reveals vertebrate
593 phylotypic period during organogenesis. *Nat Commun* **2**, 248 (2011).
594 43 Hu, H. *et al.* Constrained vertebrate evolution by pleiotropic genes. *Nat Ecol Evol*,
595 doi: 10.1038/s41559-41017-40318-41550 (2017).
596 44 Yanai, I. Development and Evolution through the Lens of Global Gene Regulation.
597 *Trends Genet*, pii: S0168-9525(0117)30171-30173 (2017).
598 45 Duboule, D. Temporal colinearity and the phylotypic progression: a basis for the
599 stability of a vertebrate Bauplan and the evolution of morphologies through
600 heterochrony. *Dev Suppl*, 135-142 (1994).
601 46 Kumar, L. & Futschik, M. E. Mfuzz: A software package for soft clustering of
602 microarray data. *Bioinformatics* **2**, 5-7 (2007).
603 47 Yue, F. *et al.* A comparative encyclopedia of DNA elements in the mouse genome.
604 *Nature* **515**, 355-364 (2014).
605 48 Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation
606 network analysis. *BMC Bioinformatics* **9**, 559 (2008).
607 49 McLean, C. Y. *et al.* GREAT improves functional interpretation of cis-regulatory
608 regions. *Nat Biotechnol* **28**, 495-501 (2010).
609 50 Force, A. *et al.* Preservation of duplicate genes by complementary, degenerative
610 mutations. *Genetics* **151**, 1531-1545 (1999).
611 51 Yanai, I. *et al.* Genome-wide midrange transcription profiles reveal expression level
612 relationships in human tissue specification. *Bioinformatics* **21**, 650-659 (2005).
613 52 Zemach, A., McDaniel, I. E., Silva, P. & Zilberman, D. Genome-wide evolutionary
614 analysis of eukaryotic DNA methylation. *Science* **328**, 916-919 (2010).
615 53 Feng, S. *et al.* Conservation and divergence of methylation patterning in plants and
616 animals. *Proc Natl Acad Sci USA* **107**, 8689-8694 (2010).
617 54 Schwaiger, M. *et al.* Evolutionary conservation of the eumetazoan gene regulatory
618 landscape. *Genome Res* **24**, 639-650 (2014).
619 55 Riviere, G. *et al.* Dynamics of DNA methylomes underlie oyster development. *PLoS*
620 *Genet* **13**, e1006807 (2017).
621 56 Maeso, I. & Tena, J. J. Favorable genomic environments for cis-regulatory evolution:
622 A novel theoretical framework. *Semin Cell Dev Biol* **57**, 2-10 (2016).
623 57 Darbellay, F. & Duboule, D. Topological Domains, Metagenes, and the Emergence of
624 Pleiotropic Regulations at Hox Loci. *Curr Top Dev Biol* **116**, 299-314 (2016).
625
626
627 **ACKNOWLEDGMENTS**
628 This research was funded primarily by the European Research Council (ERC) under the
629 European Union's Horizon 2020 research and innovation program (grant agreement No's
630 ERC-AdG-LS8-740041 to JLGS, ERC-StG-LS2-637591 to MI), the Spanish Ministerio de

631 Economía y Competitividad (BFU2016-74961-P and BFU2014-55738-REDT to JLGS, RYC-
632 2016-20089 to IM, and BFU2014-55076-P to MI), the 'Centro de Excelencia Severo Ochoa
633 2013-2017'(SEV-2012-0208 to the CRG), the 'Unidad de Excelencia María de Maetzu 2017-
634 2021'(MDM-2016-0687 to the Department of Gene regulation and morphogenesis of CABD),
635 and the Andalusian Government (BIO-396 to JLGS). In addition, support was provided by the
636 ERC under the Seventh Framework Program FP7 (FP7/2007-2013 ERC grant 268513 to
637 PWHH), the People Program (Marie Curie Actions) of the European Union's Seventh
638 Framework Program FP7 under REA grant agreement number 607142 (DevCom) to JLGS,
639 Australian Research Council Discovery Early Career Researcher Award (DECRA;
640 DE140101962 to OB), Royal Society International Exchanges grant to PWHH, the Spanish
641 Ministerio de Economía y Competitividad (BFU2014-58449-JIN to JJT, BFU2014-58908P to
642 JGF, BFU2016-80601-P to CC, BIO2015-67358-C2-1-P to RA), ICREA - Generalitat de
643 Catalunya (Academia Prize to JGF), the CNRS and the ANR (ANR16-CE12-0008-01 to HE)
644 and the Institut Universitaire de France to SB. We further acknowledge the support of the
645 CERCA Programme / Generalitat de Catalunya. NM held a Marie Sklodowska-Curie Grant
646 (658521), DB an APIF fellowship from University of Barcelona, CDRW a La Caixa PhD
647 fellowship, YM an EMBO Long Term postdoctoral fellowship (ALTF 1505-2015). We thank
648 the CRG Genomics Unit for their help with high-throughput sequencing.
649

650 **COMPLETE LIST OF AFFILIATIONS**

651 [1] Department of Zoology, University of Oxford, Oxford, UK.
652 [2] Centro Andaluz de Biología del Desarrollo (CABD), CSIC-Universidad Pablo de Olavide-
653 Junta de Andalucía, Seville, Spain.
654 [3] Genomics and Epigenetics Division, Garvan Institute of Medical Research, Sydney, New
655 South Wales, Australia.
656 [4] St Vincent's Clinical School, Faculty of Medicine, University of New South Wales, Sydney,
657 New South Wales, Australia.
658 [5] Australian Research Council Centre of Excellence in Plant Energy Biology, School of
659 Molecular Sciences, The University of Western Australia, 35 Stirling Hwy, Crawley, WA
660 6009, Australia.
661 [6] Institute of Clinical Sciences, Faculty of Medicine, Imperial College London, London W12
662 0NN, UK.
663 [7] Computational Regulatory Genomics, MRC London Institute of Medical Sciences, London
664 W12 0NN, UK.

[8] Centre for Genomic Regulation (CRG), Barcelona Institute of Science and Technology (BIST), Dr Aiguader 88, Barcelona 08003, Spain.

[9] Universitat Pompeu Fabra (UPF), Barcelona, Spain.

[10] Sorbonne Universités, UPMC Univ Paris 06, CNRS, Biologie Intégrative des Organismes Marins (BIOM), Observatoire Océanologique, F-66650, Banyuls/Mer, France.

[11] Department of Genetics, School of Biology, and Institut de Biomedicina (IBUB), University of Barcelona, Diagonal 643, Barcelona 08028, Spain.

[12] Radboud University, Faculty of Science, Department of Molecular Developmental Biology, Radboud Institute for Molecular Life Sciences.

[13] Institute of Molecular Genetics of the Czech Academy of Sciences, Videnska 1083, Praha 4, Czech Republic.

[14] Ecole Normale Supérieure, Institut de Biologie de l'ENS, IBENS, Paris, F-75005 France.

[15] Inserm, U1024, Paris, F-75005 France.

[16] CNRS, UMR 8197, Paris, F-75005 France.

[17] Genoscope, Commissariat à l'Energie Atomique et aux Energies Alternatives (CEA), Institut de Biologie François-Jacob, Evry, France.

[18] Genoscope, Commissariat à l'Energie Atomique et aux Energies Alternatives (CEA), Institut de Biologie François-Jacob, CNRS UMR 8030, Université d'Evry, France.

[19] Department of Zoology, University of Cambridge, Downing Street, CB2 3EJ Cambridge, United Kingdom .

[20] Interdisciplinary Centre of Marine and Environmental Research (CIIMAR/CIMAR) and Faculty of Sciences (FCUP), Department of Biology, University of Porto (U.Porto), Porto, Portugal.

[21] Biology and Evolution of Marine Organisms, Stazione Zoologica Anton Dohrn Napoli, Villa Comunale 1, 80121 Napoli, Italy.

[22] The Scottish Oceans Institute, Gatty Marine Laboratory, University of St Andrews, East Sands, St Andrews, Fife, KY16 8LB, UK.

[23] State Key Laboratory of Biocontrol, School of Life Sciences, Sun Yat-sen University, Guangzhou, 510275, People's Republic of China.

[24] Laboratoire de Biométrie et Biologie Evolutive (UMR 5558), CNRS/Université Lyon 1, Villeurbanne, France.

[25] EBM Aix Marseille Univ, CNRS, Centrale Marseille, I2M, Marseille, France.

[26] Sorbonne Universités, UPMC Université Paris 06, CNRS, Laboratoire de Biologie du Développement de Villefranche-sur-Mer, Observatoire Océanologique de Villefranche-sur-Mer, 181 Chemin du Lazaret, 06230 Villefranche-sur-Mer, France.

[27] Institut de Génétique Humaine, UMR 9002 CNRS-Université de Montpellier, 141 rue de la Cardonille, 34396 Montpellier CEDEX 5, France.

[28] School of Biology, University of St Andrews, Biomedical Sciences Research Complex, North Haugh, St Andrews, KY16 9ST, UK.

[29] School of Medical Sciences, Faculty of Biology, Medicine and Health, University of Manchester, Oxford Road, Manchester M13 9PT, UK.

[30] INSERM Unit 830 'Genetics and Biology of Cancers', Institut Curie Research Center, Paris, France.

[31] School of Life Sciences, Beijing University of Chinese Medicine, Dong San Huang Road, Chao-yang District, Beijing, 100029, People's Republic of China.

[32] Institute of Cellular and Organismic Biology, Academia Sinica, Taipei, Taiwan.

[33] RIKEN Center for Life Science Technologies (Division of Genomic Technologies), 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama, 230-0045 Japan.

[34] RIKEN Omics Science Center (OSC), 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama 230-0045, Japan.

[35] Center for Autoimmune Genomics and Etiology, Divisions of Biomedical Informatics and Developmental Biology, Cincinnati Children's Hospital Medical Center, Cincinnati, OH, USA.

[36] Department of Pediatrics, University of Cincinnati College of Medicine, Cincinnati, OH, USA.

[37] Harry Perkins Institute of Medical Research, 6 Verdun St, Nedlands, WA 6009, Australia.

[38] Sars International Centre for Marine Molecular Biology, University of Bergen, N-5008 Bergen, Norway.

[39] Current address: Molecular Genetics Unit, Okinawa Institute of Science and Technology, 1919-1 Tancha, Onna-son, Kunigami-gun, Okinawa 904-0495, Japan

[40] Co-first authors

[41] Corresponding authors: HE, hescriva@obs-banyuls.fr; JLGS, jlgomska@upo.es; MI, mirimia@gmail.com

MAIN FIGURES

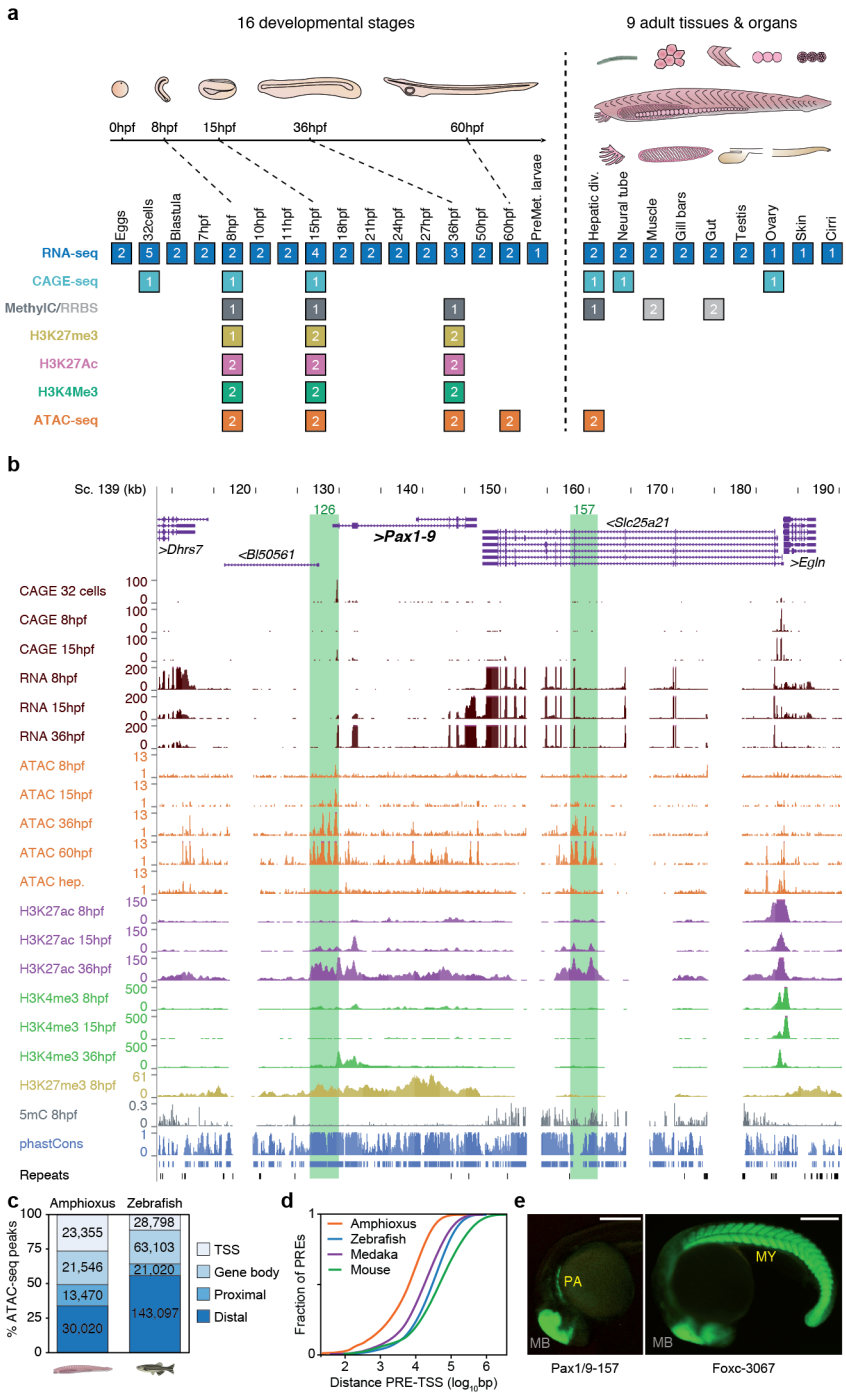

Figure 1 | Amphioxus datasets, genome browser and enhancer reporter assays

a, Summary of the 94 amphioxus samples generated in this study, comprising eight functional genomic datasets (RNA-seq, CAGE-seq, MethylC-seq or RRBS, H3K27me3, H3K27ac, H3K4me3, ATAC-seq) and 25 biological conditions (16 developmental stages and 9 adult tissues and organs). The number of replicates is indicated for each sample type. **b**, Amphioxus genome browser excerpt showing a selection of available tracks, including gene annotation,

sequence conservation (phastCons), repeats and several epigenomic and transcriptomic datasets. Green rectangle shadings highlight the ATAC-seq peaks driving the tissue-specific GFP expression shown in (e) (Pax1/9-157) and Extended Data Fig. 2f (Pax1/9-126). **c**, Percentage of all amphioxus and zebrafish ATAC-seq peaks (i.e. PREs) according to their genomic location. Regions were characterized as transcription start sites ('TSS') if located within 1 kbp upstream and 0.5 kbp downstream of a first annotated nucleotide of a transcript, 'gene body' if located within a orthology-supported gene, 'proximal' within 5 kbp upstream of (but not overlapping with) a TSS, and 'distal' if they do not belong to the aforementioned categories. **d**, Cumulative distributions of the distance between each PRE and the closest TSS in each species. **e**, Lateral views of embryos from stable transgenic zebrafish lines showing GFP expression driven by the DNA sequences underlying amphioxus ATAC-seq peaks associated with *Pax1/9* (highlighted in [b]; 26hpf) and *Foxc* (21hpf) genes. Midbrain expression corresponds to the positive control enhancer included in the reporter constructs. Anterior is to the left and dorsal to the top. EN, endoderm; MB, midbrain; MY, myotomes; PA, pharyngeal arches. Scale bar corresponds to 250 μm.

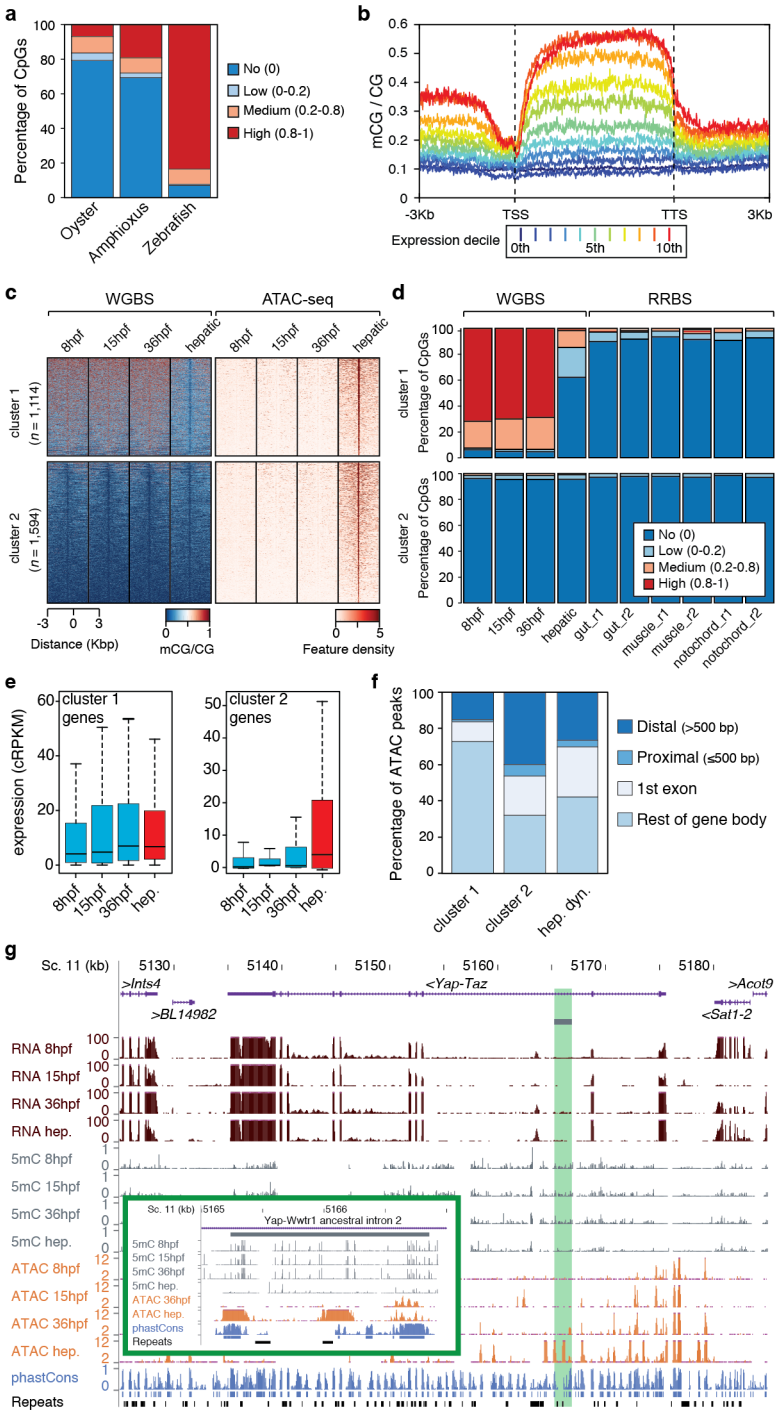

Figure 2 | 5mC patterns and dynamics in the amphioxus genome

a, Percentage of methylated CpG dinucleotides in oyster (mantle tissue), amphioxus (8hpf), and zebrafish (1K cell stage) samples. **b,** 5mC levels across gene bodies from different expression deciles (0^{th} - not expressed, 10^{th} - highest expression). TTS, transcription termination site. **c,** K-means clustering (n=2) of 5mC signal over hepatic-specific open chromatin regions (ATAC-seq peaks). **d,** Percentage of methylated CpG dinucleotides as assessed by whole genome bisulfite sequencing (WGBS) and RRBS in embryos and adult

tissue, in differentially accessible regions from (c) (cluster 1 - developmentally demethylated, cluster 2 - constitutively hypomethylated). **e,** Distribution of expression levels for genes associated with ATAC-seq peaks displaying distinct 5mC patterns in (c). **f,** Genomic distribution of regions with distinct 5mC patterns from (c). "Hep. dyn" correspond to dynamic ATAC-seq peaks that are active in hepatic diverticulum. **g**, Example of a potentially conserved (zebrafish - amphioxus) DMR associated with *Yap1*, a major TF of the Hippo pathway. Inset corresponds to the region highlighted in green. The two ohnologous genomic regions in zebrafish are shown in Extended Data Fig. 6f-i.

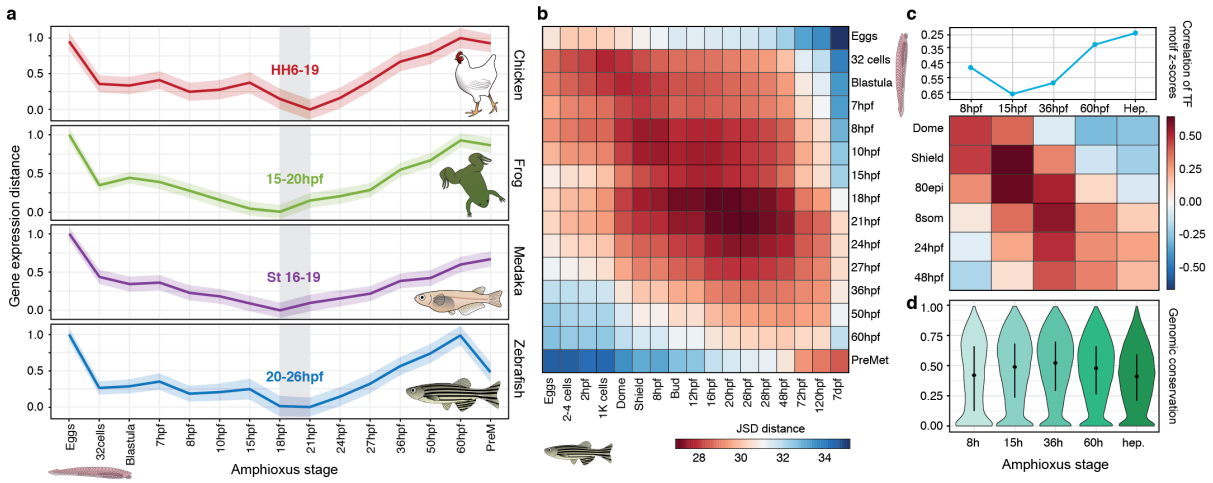

Figure 3 | A phylotypic period in chordate embryogenesis

a, Stages of minimal transcriptomic divergence (Jensen-Shannon Distance, JSD) to each amphioxus stage in four vertebrate species. The grey box outlines the 'phylotypic' period of minimal divergence, with the corresponding vertebrate periods indicated (the range given by the two closest stages). Dispersions correspond to the standard deviation computed over 100 bootstrap resamplings of the ortholog set. **b,** Heatmap of pairwise transcriptomic distances (Jensen-Shannon metrics) between amphioxus and zebrafish stages. As in (a), smaller distance (red) indicates higher similarity. **c,** Zebrafish and amphioxus pairwise correlation of relative TF motif enrichment z-scores in ATAC-seq peaks active at different developmental stages. Top panel shows the maximal correlation for five amphioxus stages to the zebrafish stages. Bottom panel shows a heatmap representation of all pairwise correlations between the two species. **d,** Sequence conservation levels in active chromatin regions at successive stages of embryonic development visualized as distribution of average phastCons scores derived from the whole genome alignment of three cephalochordate species (*B. lanceolatum*, *B. floridae*, and *B. belcheri*).

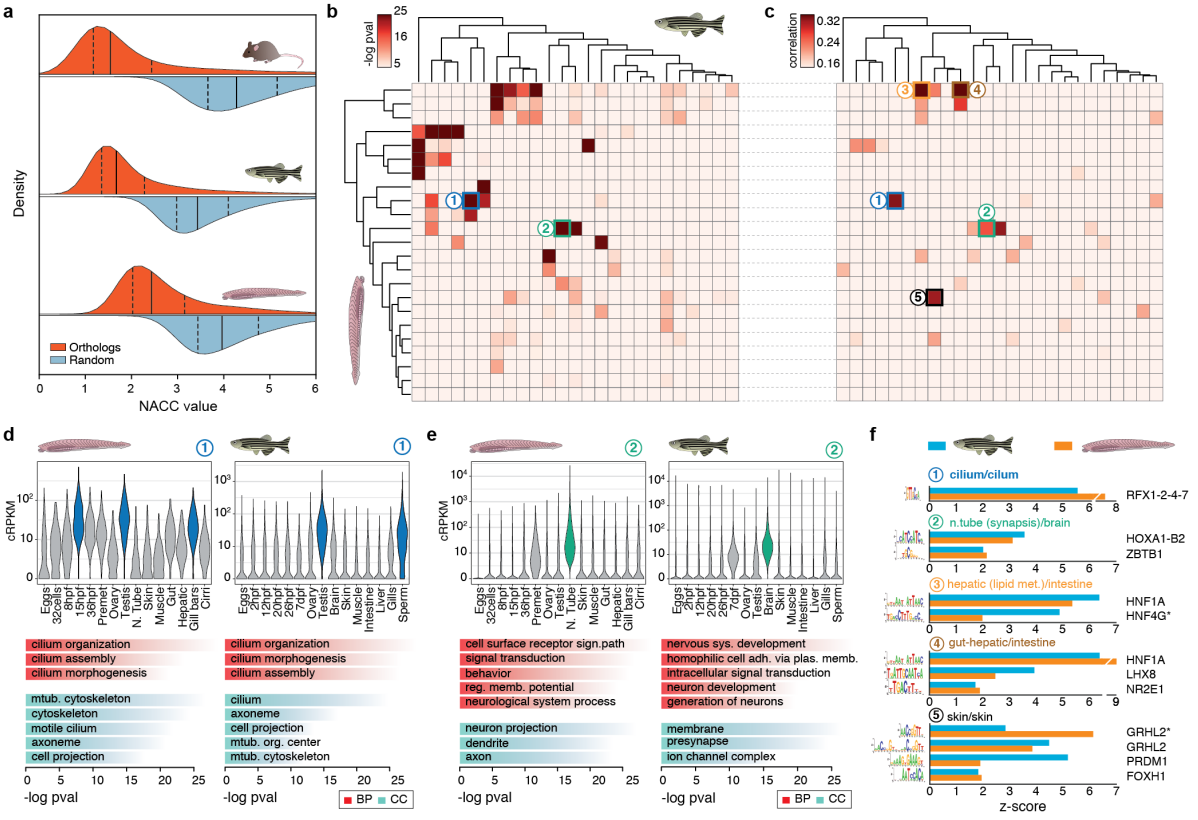

Figure 4 | Transcriptomic and *cis*-regulatory conservation of adult chordate tissues

a, Distributions of NACC values for orthologous genes (in red) or random orthology assignments (blue) in three chordate species (mouse, zebrafish and amphioxus) against human. Lower NACC values imply higher conservation of relative expression. **b**, Heatmap showing the level of statistical significance of orthologous gene overlap between WGCNA modules in the two species as derived from hypergeometric tests. **c**, Heatmap of all pairwise correlations between the modules of the two species, based on the relative TF motif z-scores for each module. Modules are ordered according to the clustering in (b). **d**,**e**, Distribution of expression values using the cRPKM (corrected for mappability Reads Per Kbp and Million mapped reads) metric for all genes within a given module across each sample (top) and enriched GO terms within each module (bottom) for two pairs of modules (① and ② in [b,c]) that are highly conserved both at the gene and *cis*-regulatory levels. BP, Biological Process; CC, Cellular Component. **f**, Examples of TF binding site motifs with high z-scores from highly correlated pairs of modules between zebrafish and amphioxus.

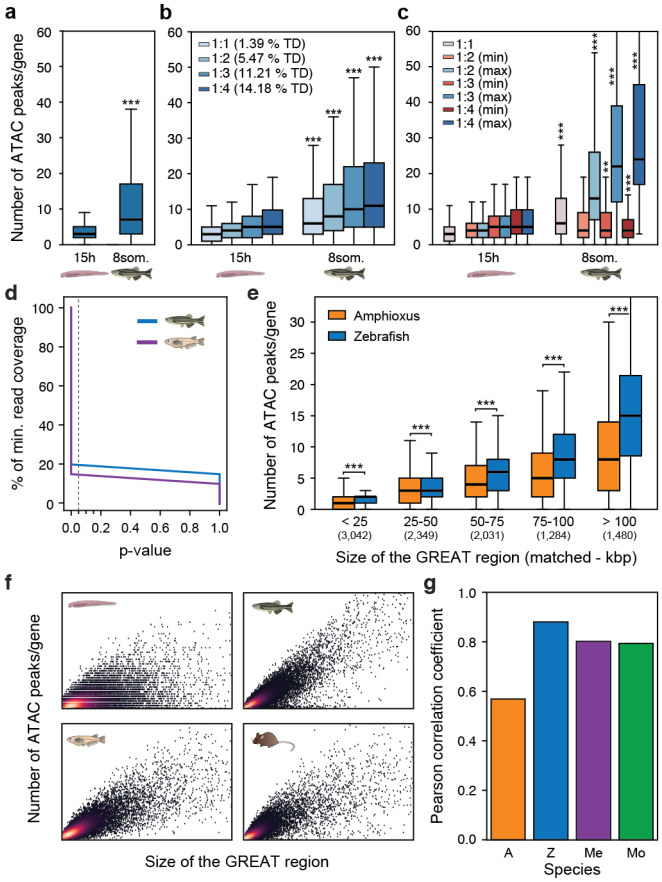

Figure 5 | Higher regulatory complexity in vertebrate regulatory landscapes

a, Distribution of the number of ATAC-seq peaks within each gene's regulatory landscape (as estimated by GREAT [49]; Supplementary Information) at pre-phylotypic developmental stages (15 hpf and 8 somites). **b**, As in (a), but with orthologous gene families separated according to the number of retained copies per family in vertebrates (from 1 to 4, using mouse as a reference). The percentage of developmental regulatory genes (*trans-dev*, TD) in each category is indicated. **c**, As in (b), but in the cases of gene families with more than one ohnolog, only the genes with the lowest ('min', in red) and the highest ('max', in blue) number of ATAC-seq peaks are plotted for each gene family. **d,** P-values of a Mann-Whitney U test against the amphioxus peak number distribution using 100% of the minimum read coverage for different levels of downsampling of the zebrafish and medaka samples. **e,** Distributions of the number of ATAC-seq peaks per gene among subsets of amphioxus and zebrafish genes matched by GREAT region size (+/- 500 bp) and binned by size as indicated. The number of genes in each group is indicated in the x-axis. **f,** Density scatterplot of the number of ATAC-seq peaks (y-axis) versus the size of the GREAT region (x-axis) for each gene and species. **g,** Pearson correlation coefficients for the values showed in (f). P-values in

a-c and **e**, correspond to one-sided Mann Whitney tests of the zebrafish distribution versus the equivalent amphioxus one.

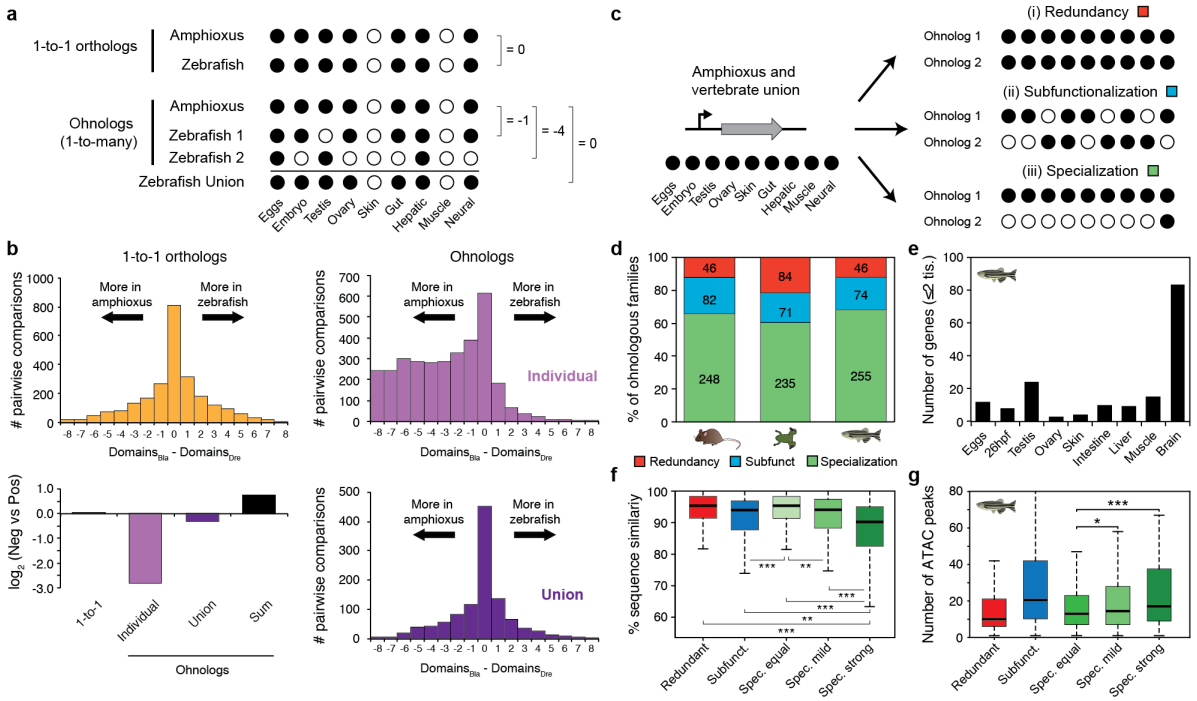

Figure 6 | Expression specialization is the main fate after whole genome duplication

a, Schematic summary of the analysis shown in (b). Expression is binarized (on or off) for each amphioxus and vertebrate gene across the nine comparable samples, based on an arbitrary expression cut-off (normalized cRPKM>5). For each vertebrate gene, the number of positive expression domains is subtracted from the number of domains in which the single amphioxus ortholog is expressed. Black/White circles represent on/off expression, respectively. **b,** Distribution of the difference in positive domains between zebrafish and amphioxus for 1-to-1 orthologs (yellow), individual ohnologs (lilac) and the union of all vertebrate ohnologs in a family (purple). Bottom left: $\log_2$ of the ratio between zebrafish genes with negative score (more domains in amphioxus) and with positive score (more domains in zebrafish) for each category. "Sum" (black), binarization of family expression is performed after summing the raw expression values for all ohnologs. **c,** Schematic summary of the analyses shown in (d), representing the three possible fates after WGD: Redundancy, all ohnologs are expressed in all domains; Subfunctionalization, none of the ohnologs are expressed in all domains; Specialization, at least one of the ohnologs in expressed in all domains, but at least one is not. **d,** Distribution of fates after WGD for families of ohnologs inferred to be ancestrally expressed in all nine studied domains for each vertebrate species. **e,** Number of ohnologs with strong specialization (two or fewer remaining expression domains) in zebrafish expressed in each domain. **f,** Distribution of the percentage of nucleotide

sequence similarity between human and mouse for different classes of ohnologs based on their fate after WGD. Ohnologs from specialized families are divided into “Spec. equal” (maintaining all expression domains), “Spec. mild” (which have lost expression domains, but maintained more than two), “Spec. strong” (with two or fewer remaining expression domains). **g,** Distribution of the number of ATAC-seq peaks within GREAT regions for zebrafish ohnologs for each category. P-values in **f** and **g** correspond to Wilcoxon Sum Rank tests between the indicated groups. * 0.5 < P-value ≤ 0.01, ** 0.01 < P-value ≤ 0.001, *** P-value < 0.001.