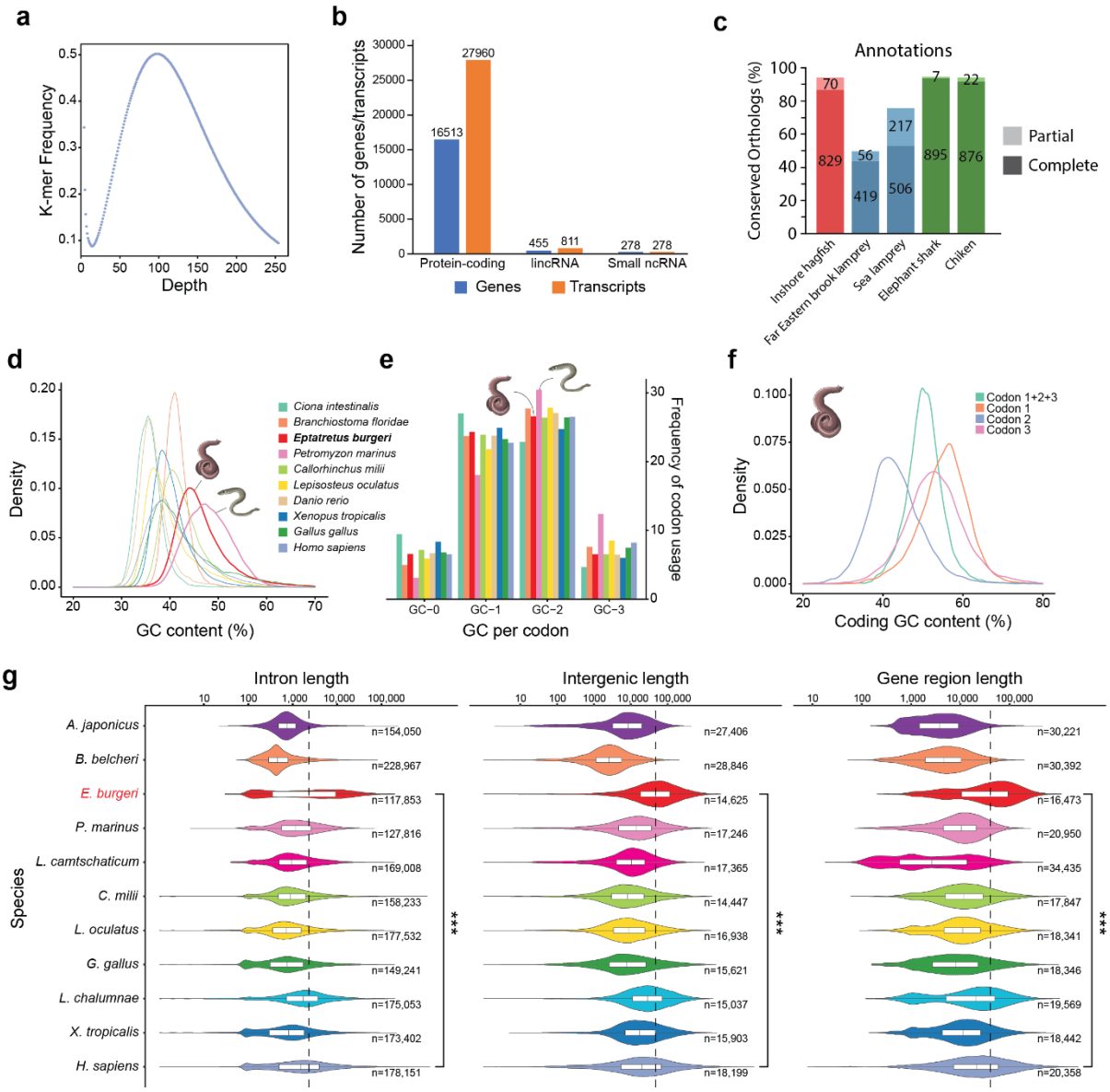
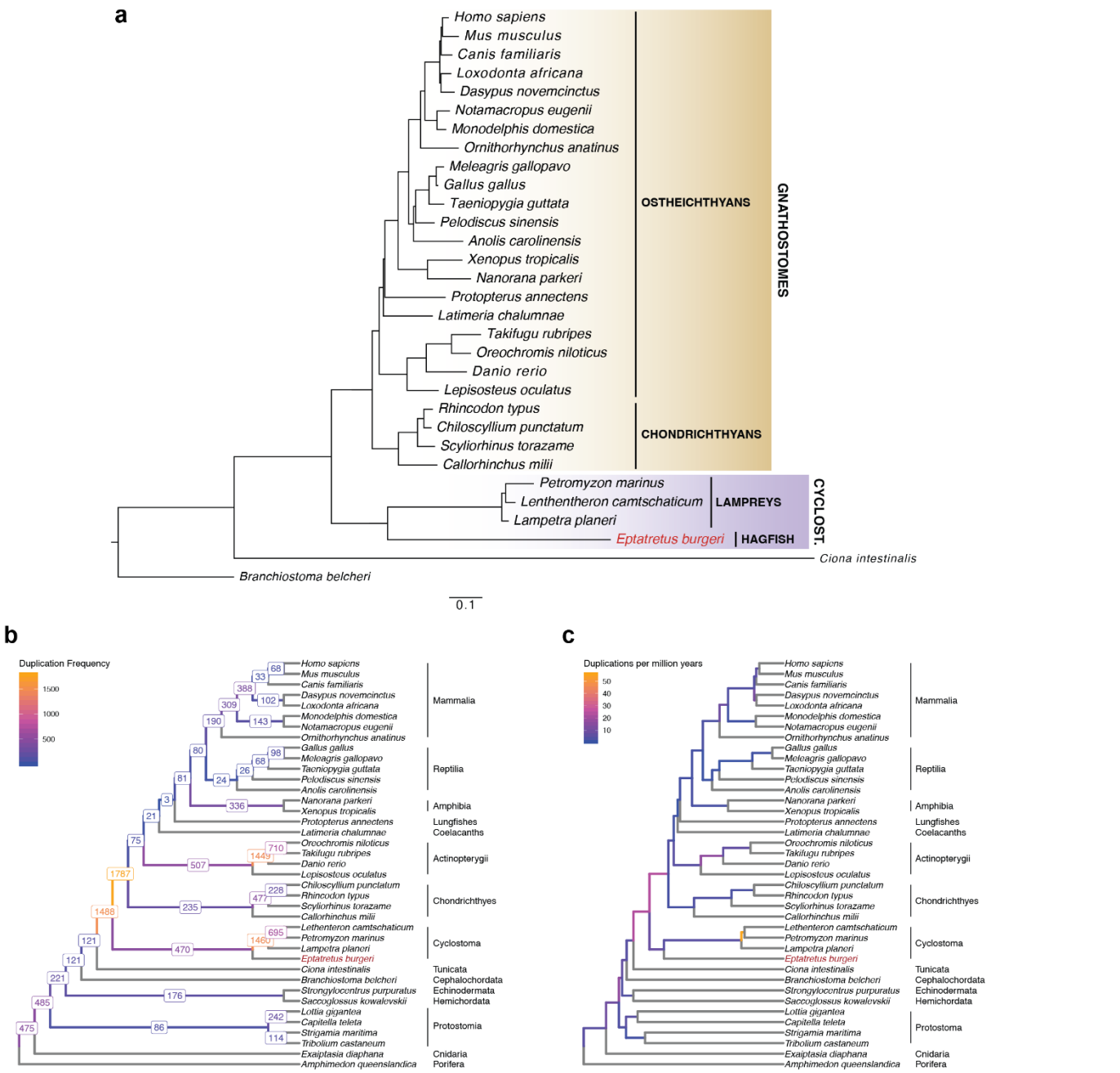
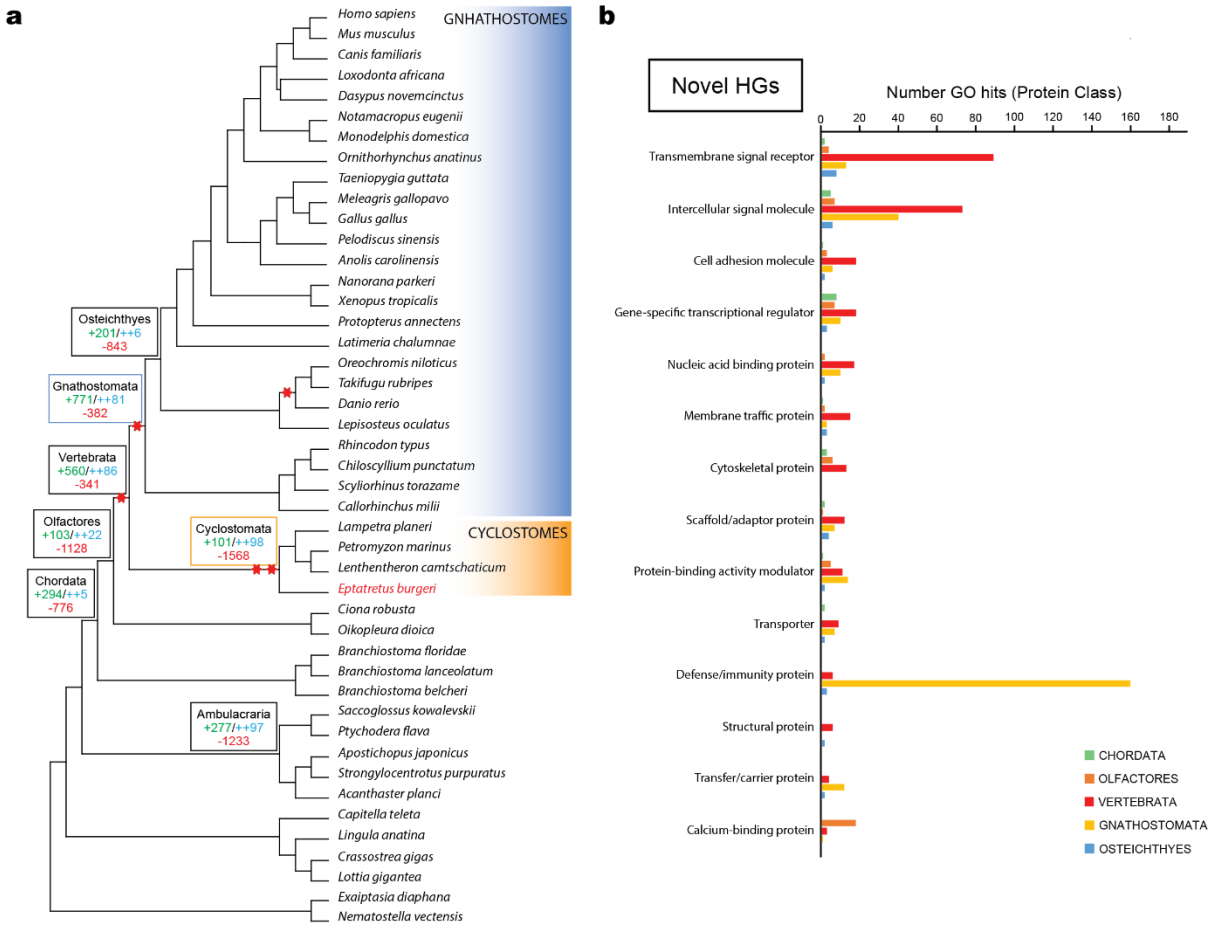


Extended Data Figure 1. Features of the hagfish genome. **a**, 17-mer distribution for inshore hagfish genome size estimation using all raw reads from short insert-size libraries. **b**, Counts for major classes of genes and transcripts from Ensembl annotation. **c**, Completeness assessment of the annotation of the inshore hagfish *E. burgeri* genome (red), three lamprey species (blue) and two jawed vertebrates (green). Numbers of conserved metazoan orthologs (metazoa_odb10 dataset, n =954) are indicated for each case. **d**, GC-content distribution of the hagfish genome and other chordates calculated from 10-kb non-overlapping windows. **e**, All codon type frequency in given chordate genomes according to GC-content. GC-0/1/2/3 indicates the number of G or C bases in a codon. **f**, Distribution of GC content at each codon position or at all codon positions (Codon1+2+3). For each protein coding gene, we only kept the longest coding sequence. For each coding sequence, we calculated the GC content at

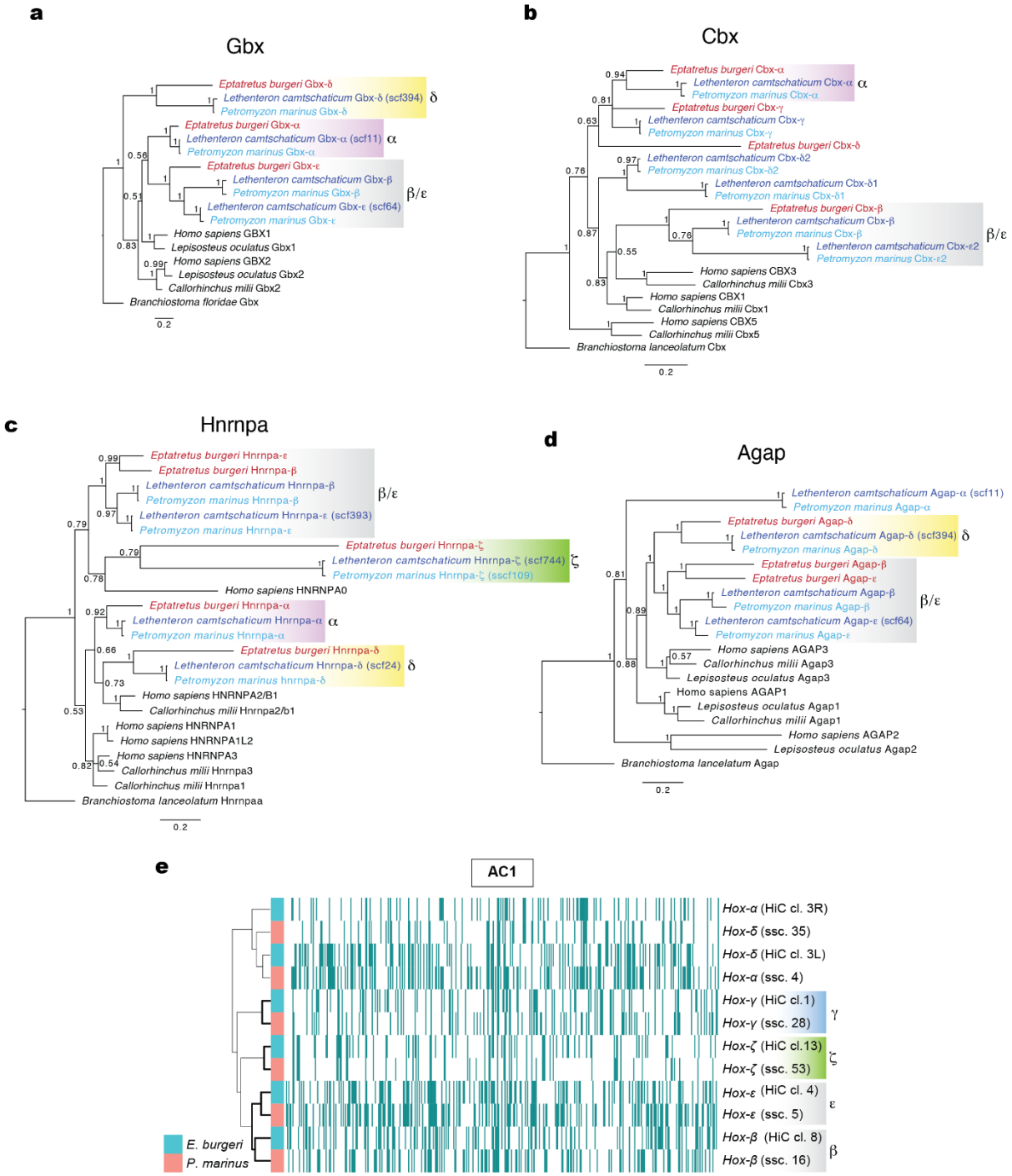
separate codon positions. We also calculated the GC content for each coding sequence, which is equal to the GC content of all three codon positions. **g**, Violin plots of size distribution of intron (left), intergenic (middle) and gene body (right) lengths over a logarithmic scale of the hagfish (*E. burgeri*), two lamprey species (sea lamprey, *P. marinus*; Arctic lamprey, *L. camtschaticum*), six gnathostome vertebrates (human, *H. sapiens*; frog, *X. tropicalis*; coelacanth, *L. chalumnae*; chicken, *G. gallus*; spotted gar, *L. oculatus*; and the elephant shark, *C. milii*) and two invertebrate deuterostomes (sea cucumber, *A. japonicus*; amphioxus, *B. belcheri*). For each genomic feature and each species, the median and IQR (interquartile range) length statistics are indicated with a white rectangle. Dashed vertical line indicates median size of *E. burgeri* features; *** $P < 2.2 \times 10^{-16}$, two-sided Wilcoxon sum-rank test.



Extended Data Figure 2. Phylogenomic analysis of chordates and gene duplication rates across metazoan evolution. **a,** Bayesian Inference tree of 31 chordate species was built using a protein alignment (see Methods). All nodes were recovered with a posterior probability of 1. Cyclostome monophyly was unequivocally supported. Scale bar indicates 0.1 substitutions per site. **b, c,** The exact number of gene duplication events inferred using OrthoFinder2 with greater than 50% support (**b**) and the number of events per million years per branch (**c**) are shown. In each case, the colour of the branch represents the value according to the key on the upper-left of each pane. Hagfish is highlighted in red font in all panes.

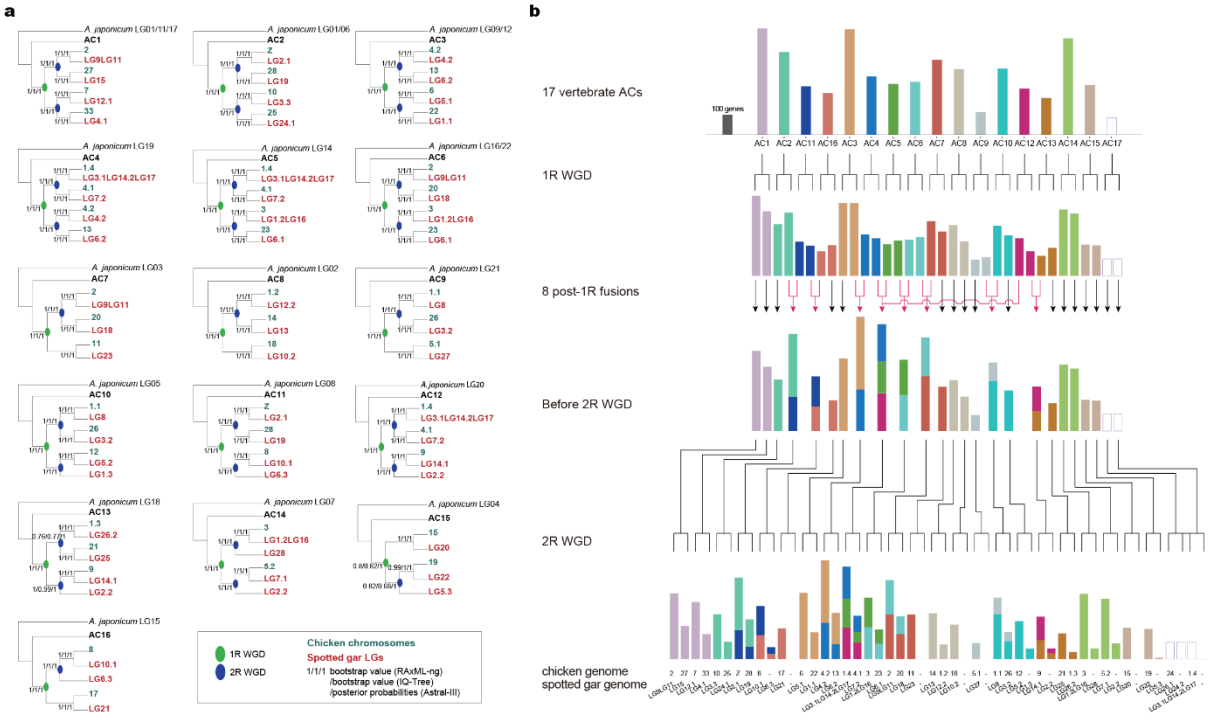


Extended Data Figure 3. **Reconstruction of ancestral gene content**. **a**, Cladogram showing the phylogenetic relationships of 45 species representatives of all major eumetazoan taxa with species of gnathostomes and cyclostomes highlighted in blue and orange, respectively. Gene family gains and losses are indicated in selected nodes: green, novel homology groups (HG); blue, novel core HGs; red, lost HGs. **b**, Top 14 Protein Class GO hits for novel homology groups (HG) gained across different nodes of chordates, color coded by taxa (legend at the bottom right) and sorted by the Vertebrata node. The largest GO enriched terms are 'transmembrante signal receptor' and 'intercellular signal molecule' in vertebrates, and 'defense/immunity protein' in gnathostomes.

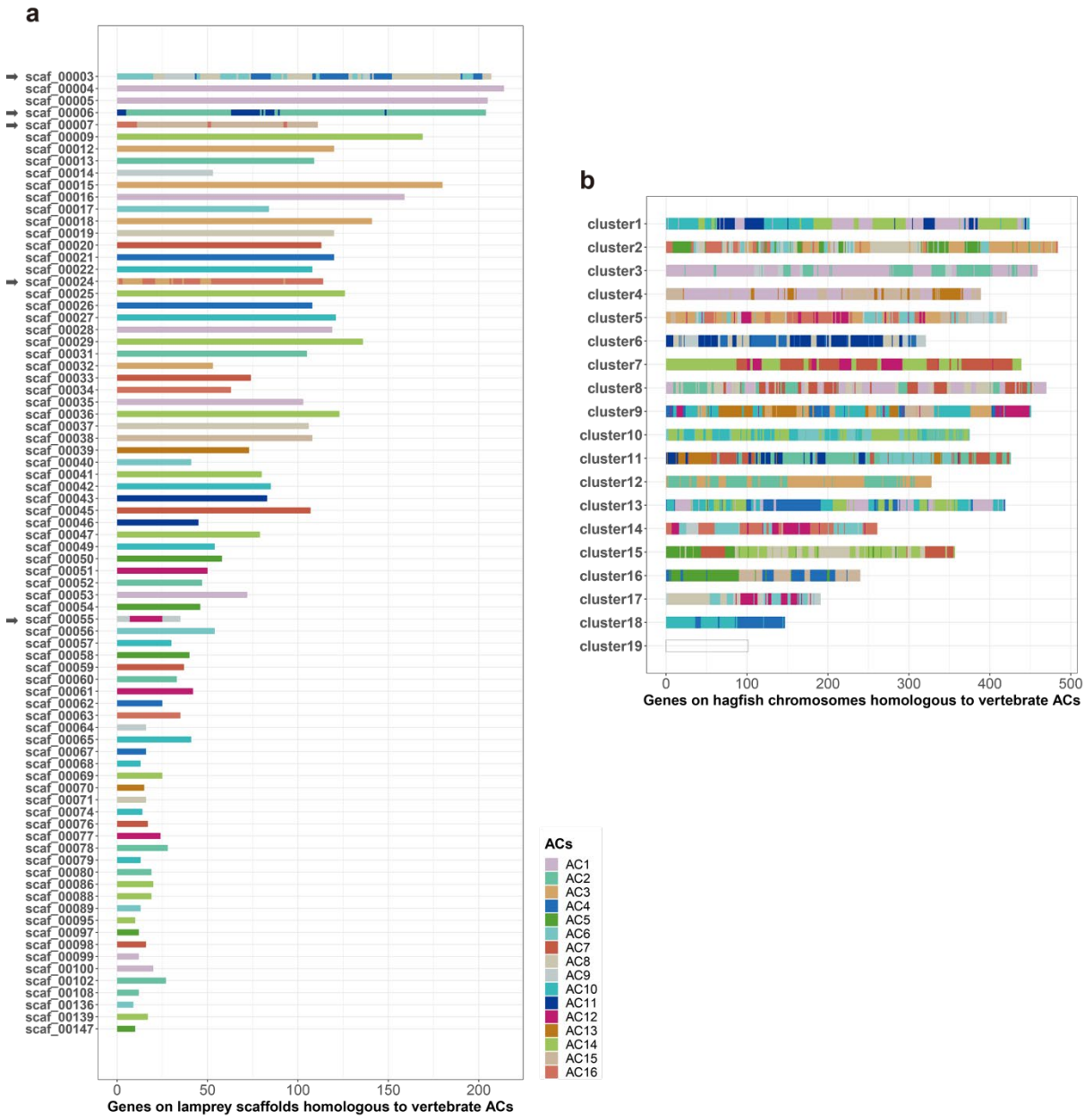


Extended Data Figure 4. Phylogenetic and retention profile clustering analyses of Hox syntenic regions. a-d, Bayesian inference phylogenetic trees of amino acid sequences of 4 non-*Hox* syntenic genes to *Hox* clusters, Gbx (**a**), Cbx (**b**), Hnrnpa (**c**) and Agap (**d**), of the inshore hagfish (in red), the sea lamprey (in light blue), the Arctic lamprey (in dark blue) and selected gnathostomes. Orthologs from the European amphioxus *Branchiostoma lanceolatum* were used as outgroup to root the trees. Posterior probability is indicated in each node. Scales indicate number of substitutions per site. Phylogenetic analyses of *Hox* genes generally fail to determine orthology due their high conservation and short alignments. The phylogenetic trees

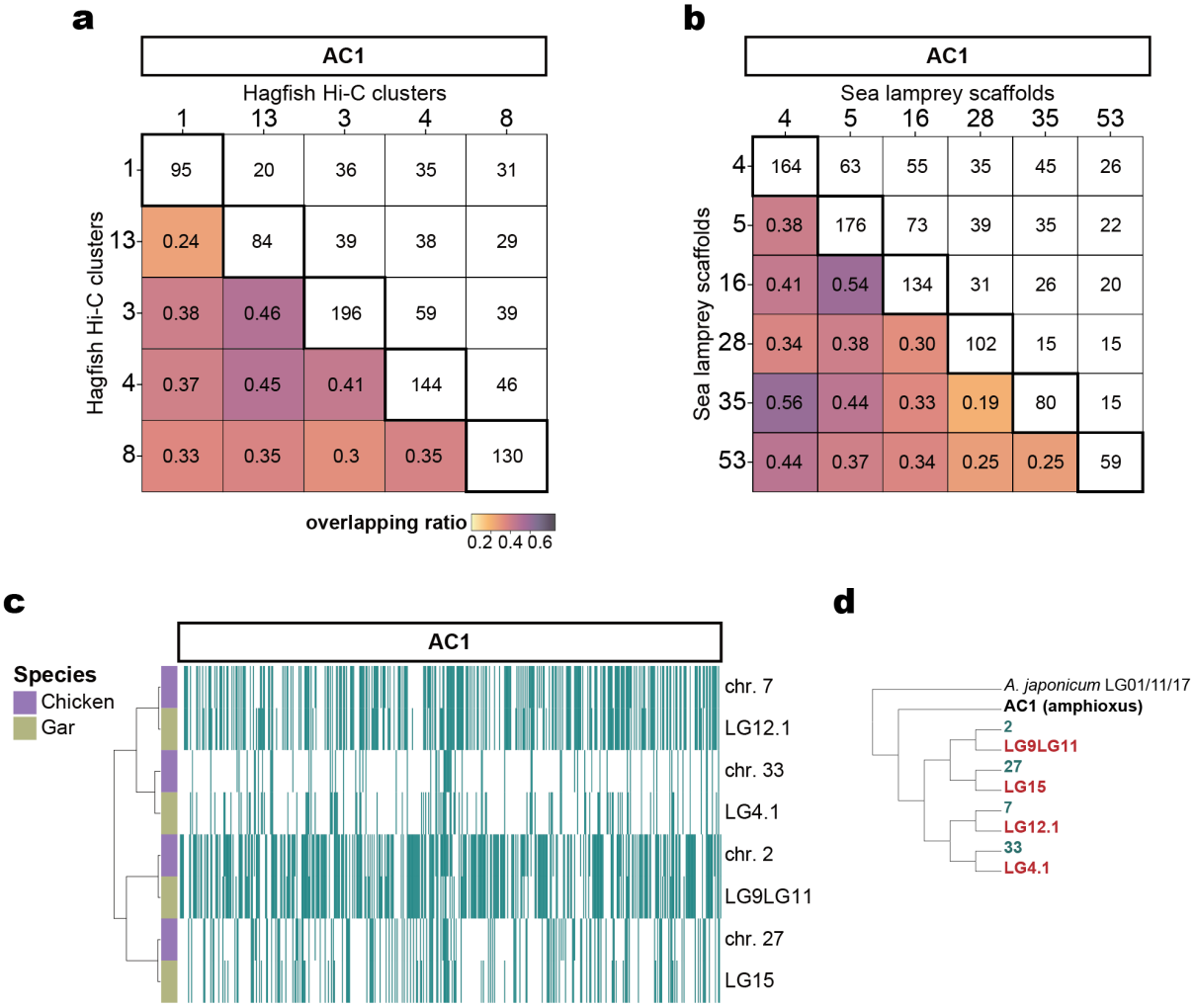
of these non-*Hox* linked genes clearly support the orthology of *Hox-α* (Gbx, Cbx and Hnrnpa), *Hox-δ* (Gbx, Hnrnpa and Agap), and *Hox-ζ* (Hnrnpa) clusters, while *β* and *ε* genes always group together, as previously observed for the lamprey[17]. The alignments used to build the trees, together with the MrBayes parameters and number of generations used to build each tree are provided as Supplementary Files 16-19. **e**, clustering analysis of retention profiles (see main text) resolved the orthology relationships of *Hox-β*, *Hox-ε*, *Hox-γ*, as well as *Hox-ζ* clusters. Supported orthologies in each analysis are marked with color-coded rectangles. The location of each cluster is indicated in parenthesis in e (ssc, super scaffold; HiC cl, Hi-C contact cluster, or chromosome). For the clustering analysis of AC1-derived chromosomes in the lamprey and hagfish, we split Hi-C cluster 3 into two halves, each containing one Hox cluster: 3L (coordinates 0-107.78 Mb), 3R (107.78-194 Mb).



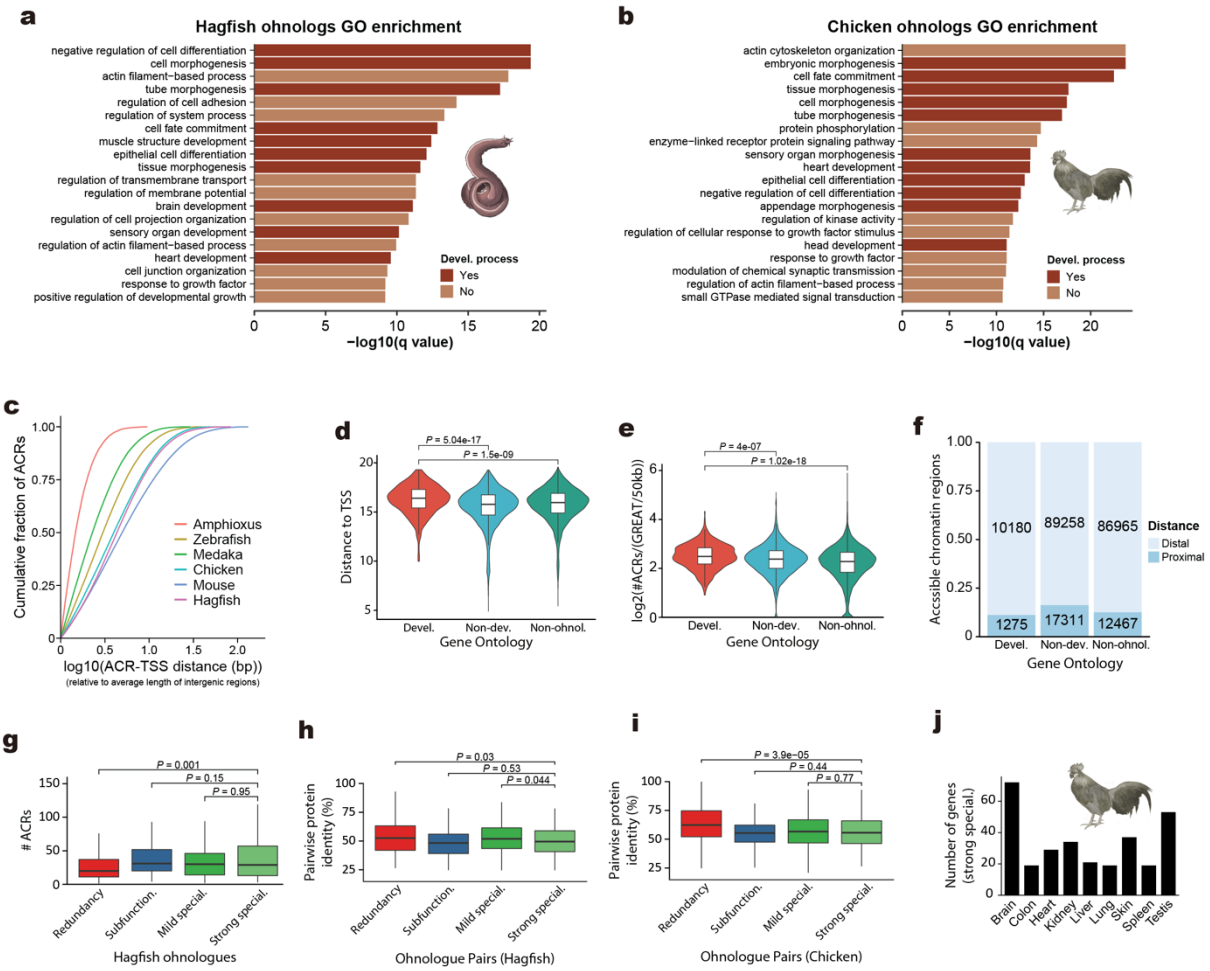
Extended Data Figure 5. **Emergence of gnathostome karyotypes via 2R**. **a**, Chromosomal level phylogenetic trees demonstrate the occurrence of 2R WGD. The two rounds of WGD are color-coded. Cyan and red denote genes encoded by corresponding chicken chromosomes and gar LGs, respectively. Bootstrap support and posterior probability values of three methods are marked on branches. **b**, Pre- and post-1R chromosomal fusions contributed to the evolution of ancestral gnathostome karyotypes. The size of each ancestral chromosome, except for AC17 contributed chromosomes (see Supplementry Information), is proportional to the number of retained genes from ACs. The pre-1R fusion leading to AC3 is drawn as a dotted line due to its uncertainty.



Extended Data Fig. 6. Contributions of vertebrate ACs to the genomes of hagfish and lamprey. a, Sea Lamprey scaffolds are generally homologous to one single AC except five scaffolds labeled with an arrow. Sea lamprey scaffolds including scaf_00001, scaf_00002, scaf_00008, scaf_00010, scaf_00011 and scaf_00023, confounded by missassembly, are not presented here. **b**, The distribution of homologous genes to ACs on 19 hagfish cluster. Only significant homologous relationships between hagfish clusters and ACs are shown. Because hagfish cluster 19 is not homologous to any AC, it is presented as a blank block. Genes are color-coded according to its homologous AC.



Extended Data Figure 7. Overlapping ratio and clustering analysis of ohnologous and orthologous chromosomes. a,**b**, AC1 corresponds to five and six mutually paralogous chromosomes in hagfish (**a**) and sea lamprey (**b**) genomes, respectively. Numbers in colour-coded cells (bottom left triangle) indicate the *OR* between two chromosomes. Numbers in white cells (top right triangle) indicate the number of shared retained genes between two chromosomes. Numbers on the diagonal line from top left to bottom right (thick-lined cells) indicate the total number of retained genes of a chromosome. Overlapping rations corresponding to all hagfish chromosomes and lamprey scaffolds are provided in Supplementary File 9. **c**, Retention profile clustering analysis of gnathostome orthologous chromosomes deriving from AC1. Retained genes are denoted by dark cyan lines. Four orthologous chromosome pairs are defined. **d**, The clustering found in (**c**) is the same as that found in the phylogenetic analysis (Extended Data Fig. 5a), demonstrating the reliability of the *OR* approach.



Extended Data Figure 8. Fate and *cis*-regulatory evolution of ohnologs after WGD. a, **b**, Gene Ontology enrichment analysis of ohnologs in the hagfish (**a**) and the chicken (**b**). Top 20 terms are shown, majority of which are related with development. **c**, Cumulative distribution of distance of ACRs from the closest TSSs normalized by the average length of intergenic regions in each genome. **d**, Distribution of the distance from ACRs to closest TSS of developmental ohnologs (Devel.), non-developmental ohnologs (Non-dev.) and non-ohnologous (Non-ohnol.) genes. **e**, Distribution of the number of ACRs per gene, normalized by the GREAT region length, for developmental ohnologs (Devel.), non-developmental ohnologs (Non-dev.) and non-ohnologous (Non-ohnol.) genes. **f**, Proportion of distal ACRs across different gene functional categories. Within a GREAT-defined region, proximal regulatory sequences were defined as those from 5 kb upstream to 1 kb downstream of a TSS, and the rest of the region was treated as distal. **g**, Distribution of the number of ACRs of hagfish ohnologs for each category (special., specialization). **h**, Distribution of pairwise protein identity for hagfish ohnologous pairs for each category. **i**, Distribution of pairwise protein identity for chicken ohnologous pairs for each category. **j**, Number of ohnologues with strong

specialization in chicken expressed in each tissue. Only the gene in a pair with narrower expression breadth is analyzed. *P* values in **d**, **e**, **g**, **h**, **i** correspond to two-sided Wilcoxon sum-rank test between the indicated groups.