

# The hagfish genome and the evolution of vertebrates

<https://doi.org/10.1038/s41586-024-07070-3>

Received: 17 April 2023

Accepted: 15 January 2024

Published online: 23 January 2024

Open access

 Check for updates

Ferdinand Marlétaz<sup>1,2</sup>✉, Nataliya Timoshevskaya<sup>3</sup>, Vladimir A. Timoshevskiy<sup>3</sup>, Elise Parey<sup>1</sup>, Oleg Simakov<sup>2,4</sup>, Daria Gavriouchkina<sup>2,10</sup>, Masakazu Suzuki<sup>5</sup>, Kaoru Kubokawa<sup>6</sup>, Sydney Brenner<sup>7,11</sup>, Jeramiah J. Smith<sup>3</sup>✉ & Daniel S. Rokhsar<sup>2,8,9</sup>✉

As the only surviving lineages of jawless fishes, hagfishes and lampreys provide a crucial window into early vertebrate evolution<sup>1–3</sup>. Here we investigate the complex history, timing and functional role of genome-wide duplications<sup>4–7</sup> and programmed DNA elimination<sup>8,9</sup> in vertebrates in the light of a chromosome-scale genome sequence for the brown hagfish *Eptatretus atami*. Combining evidence from syntenic and phylogenetic analyses, we establish a comprehensive picture of vertebrate genome evolution, including an auto-tetraploidization (1R<sub>v</sub>) that predates the early Cambrian cyclostome–gnathostome split, followed by a mid–late Cambrian allo-tetraploidization (2R<sub>v</sub>) in gnathostomes and a prolonged Cambrian–Ordovician hexaploidization (2R<sub>CY</sub>) in cyclostomes. Subsequently, hagfishes underwent extensive genomic changes, with chromosomal fusions accompanied by the loss of genes that are essential for organ systems (for example, genes involved in the development of eyes and in the proliferation of osteoclasts); these changes account, in part, for the simplification of the hagfish body plan<sup>1,2</sup>. Finally, we characterize programmed DNA elimination in hagfish, identifying protein-coding genes and repetitive elements that are deleted from somatic cell lineages during early development. The elimination of these germline-specific genes provides a mechanism for resolving genetic conflict between soma and germline by repressing germline and pluripotency functions, paralleling findings in lampreys<sup>10,11</sup>. Reconstruction of the early genomic history of vertebrates provides a framework for further investigations of the evolution of cyclostomes and jawed vertebrates.

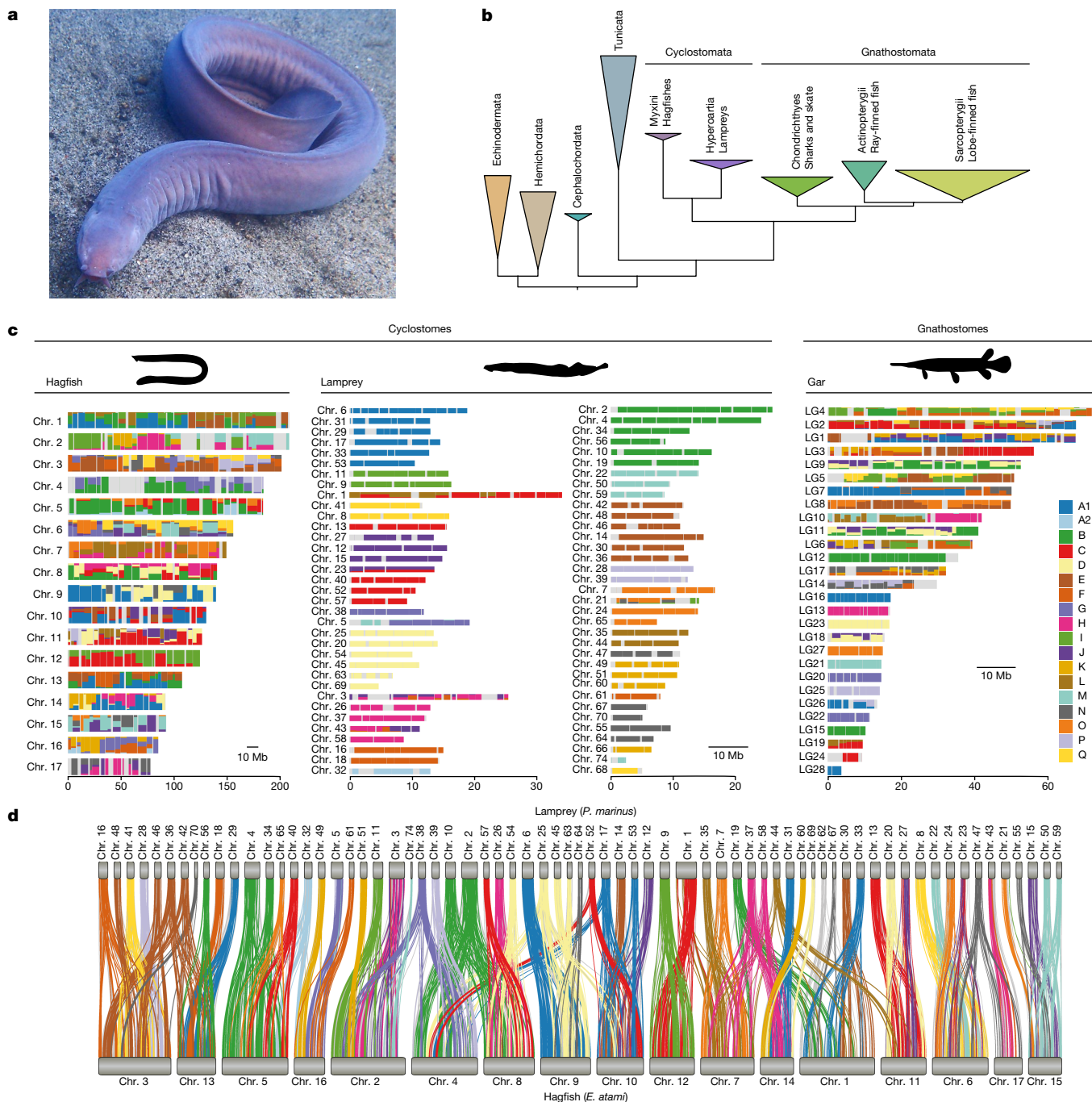
Hagfishes are deep-sea scavengers with a prodigious capacity for producing slime<sup>12</sup> (Fig. 1a). As one of only two surviving lineages of jawless fishes, hagfishes provide a unique comparative perspective on early vertebrate evolution. Both hagfishes and lampreys stand apart from jawed vertebrates (gnathostomes) in the absence of jaws, bone and dentine<sup>3</sup>, and they have been grouped together as cyclostomes<sup>13</sup>, the sister group to gnathostomes. However, hagfishes lack several key characteristics that are shared by lampreys and gnathostomes, including definitive vertebrae<sup>14</sup>, lensed eyes with oculomotor control and electroreceptive sensory organs<sup>1,2</sup>. The relative simplicity of the hagfish body plan suggests an alternative hypothesis whereby hagfishes diverged before a craniate clade that groups lampreys with jawed vertebrates<sup>3</sup>. Early molecular phylogenies (albeit with limited sequence datasets and taxonomic sampling) have consistently supported cyclostome monophyly<sup>15–17</sup>, which implies that hagfishes are secondarily simplified. But the molecular bases of this derived body plan in the light of the tumultuous genomic history of vertebrates are poorly understood.

Early vertebrate evolution was punctuated by multiple polyploidizations, although the nature and timing of these ancient events,

and their effects on vertebrate biology, remain elusive<sup>4–6</sup>. An early duplication preceding the gnathostome–cyclostome split (1R<sub>v</sub>) is generally accepted but has not been clearly resolved by molecular phylogenetics<sup>18</sup>. A gnathostome-specific allo-tetraploidization (2R<sub>v</sub>) was definitively established on the basis of chromosomal rearrangements observed in gnathostomes but not lampreys<sup>11,19,20</sup>, leading to the rejection of the hypothesis that two rounds of genome duplication (1R and 2R) occurred before the cyclostome–gnathostome split<sup>21–23</sup>. Conversely, lampreys experienced additional independent duplication(s) not found in gnathostomes<sup>11,19,20</sup>. A hexaploidy inferred in lampreys was further hypothesized to be ancestral to cyclostomes<sup>20</sup>, on the basis of the observation that lampreys and hagfishes both possess six Hox clusters, although their respective orthology remains unclear<sup>24</sup>. Definitive resolution of the duplication history of cyclostome genomes must also account for the disparate karyotypes of hagfishes ( $2n \approx 34$  somatic chromosomes) and lampreys ( $2n \approx 168$ ) (refs. 7,11,25,26).

Notably, hagfishes<sup>25,27,28</sup> and lampreys<sup>10,11,29</sup> perform programmed elimination of germline-specific chromosomes from the genomes of

<sup>1</sup>Centre for Life's Origins and Evolution, Department of Genetics, Evolution and Environment, University College London, London, UK. <sup>2</sup>Molecular Genetics Unit, Okinawa Institute of Science and Technology Graduate University, Okinawa, Japan. <sup>3</sup>Department of Biology, University of Kentucky, Lexington, KY, USA. <sup>4</sup>Department for Neurosciences and Developmental Biology, University of Vienna, Vienna, Austria. <sup>5</sup>Department of Science, Graduate School of Integrated Science and Technology, Shizuoka University, Shizuoka, Japan. <sup>6</sup>Ocean Research Institute, The University of Tokyo, Tokyo, Japan. <sup>7</sup>Comparative and Medical Genomics Laboratory, Institute of Molecular and Cell Biology, A\*STAR, Biopolis, Singapore, Singapore. <sup>8</sup>Department of Molecular and Cell Biology, University of California, Berkeley, Berkeley, CA, USA. <sup>9</sup>Chan Zuckerberg Biohub, San Francisco, CA, USA. <sup>10</sup>Present address: UK Dementia Research Institute, University College London, London, UK. <sup>11</sup>Deceased: Sydney Brenner. ✉e-mail: f.marletaz@ucl.ac.uk; jjsmit3@uky.edu; dsrokhsar@gmail.com



**Fig. 1 | Phylogenetic relationships and syntenic architecture of cyclostomes and gnathostomes.** **a**, The brown hagfish, *Eptatretus atami* (photo credit, M. Suzuki). **b**, Summary of deuterostome phylogeny based on 176 selected genes (61,939 positions) using a site-heterogeneous model (CAT+GTR). This topology is robust to compositional heterogeneity and similar to what was obtained with 1,467 genes using a site-homogeneous partitioned model (see Methods, Supplementary Note 1 and Extended Data Fig. 2). **c**, Karyograms showing the ancestry of hagfish, lamprey and gar chromosomes in terms

of chordate linkage groups (CLGs A1, A2 and B–Q) described previously<sup>19,35</sup> (see also ref. 20 and Supplementary Note 2). Coloured bins contain 20 genes and only genes from CLGs with significant enrichment (Fisher’s exact test) are counted (Methods). Hagfish, lamprey and gar silhouettes downloaded from PhyloPic (credit to Gareth Monger for lamprey). **d**, Conserved syntenies show that hagfish chromosomes are typically fusions of multiple lamprey chromosomes. Lines connect orthologous genes and are coloured according to the ancestral chordate linkage groups (colour legend in c).

somatic cells<sup>8,9</sup>. In lampreys, germline-specific chromosomes encode numerous genes with putative functions in the maintenance and development of the germline<sup>10,11</sup>. Although hagfishes were the first vertebrate species shown to experience developmentally programmed DNA elimination<sup>25</sup>, only germline-enriched satellite repeats have been characterized<sup>30–33</sup>. Because no germline-specific protein-coding genes have been reported in hagfish so far, their possible germline functions,

evolutionary origin and relationship to germline-specific genes in lampreys have not been addressed.

Here we report a chromosome-scale genome assembly for the brown hagfish *E. atami*. Using a synteny-based phylogenetic approach, we definitively resolve and date the timing of duplication and divergence events that shaped the genomes of extant vertebrate lineages, and assess rediploidization after duplication. We further dissect the effect

of these events on the emergence of genes that are involved in vertebrate and hagfish characteristics, and find that hagfish genomes are derived by extensive gene loss, consistent with their morphological simplification. We also find that hagfish genes that are programmatically eliminated during early embryonic development contribute to several aspects of germ cell biology, and reveal the evolution of vertebrate germline-specific chromosomes.

## Evolution of cyclostome genomes

We sequenced the germline genome of the brown hagfish *E. atami* (formerly *Paramyxine atami*) using a combination of long and short reads from testes, and organized the assembly into chromosomes using proximity ligation data from somatic tissue (Supplementary Table 1). Our *E. atami* assembly spans 2.52 Gb and includes 17 large chromosomal scaffolds, consistent with the expected somatic karyotype ( $2n = 34$ ) (Extended Data Fig. 1a and Supplementary Table 2). The length of the assembly is intermediate between the fluorescence-based estimates of genome size for somatic (2.01 Gb) and germline (3.37 Gb) cells<sup>25,28</sup>, consistent with *k*-mer estimates (2.02 Gb and 3.28 Gb, respectively, Extended Data Fig. 1b and Methods). The *E. atami* germline genome also includes seven highly repetitive chromosomes that are completely eliminated during development and whose sequences are present in our assembly as sub-chromosomal fragments, similar to what is seen in the highly repetitive germline-specific chromosomes of lampreys<sup>7,11</sup> and songbirds<sup>34</sup>. We annotated 28,469 genes, of which 22,663 show similarity with the protein-coding complement of other species.

We first used our hagfish gene set to test the monophyly of cyclostomes, extending pioneering early studies<sup>16,17</sup> by introducing (i) broader taxonomic sampling of cyclostomes (including new data for the Atlantic hagfish *Myxine glutinosa*) and (ii) improved modelling of site heterogeneity and compositional bias (Methods, Supplementary Note 1 and Supplementary Table 3). A new set of 1,467 orthologues informed by complete hagfish and lamprey genomes alleviates possible paralogy issues, and includes eightfold more markers than earlier studies did (Methods). These analyses confirm the monophyly of cyclostomes with both partitioned analysis (Extended Data Fig. 2a) and site-heterogeneous model analysis (Fig. 1b and Extended Data Fig. 2b). Robustness to compositional heterogeneity is further supported by six-category amino acid recoding validated by posterior predictive tests (Extended Data Fig. 2c,d and Supplementary Table 4).

Despite their disparate karyotypes, the chromosomes of hagfish and lamprey are simply related (Fig. 1c,d) although after around 457 million years of independent evolution, gene order is highly scrambled (Extended Data Fig. 3a,b) and repetitive landscapes are distinct (Extended Data Fig. 1c,d). In general, each hagfish chromosome is typically orthologous to a fusion of between two and six lamprey chromosomes and, conversely, each lamprey chromosome is typically associated with a single hagfish chromosome, with few exceptions (Extended Data Fig. 3b). To differentiate between possible chromosomal fusions in the hagfish lineage and/or fissions (or duplication) in the lamprey lineage, we used the previously reconstructed ancestral chordate linkage groups (CLGs A1, A2 and B-Q)<sup>20,35</sup> (Supplementary Note 2). Although lamprey chromosomes are typically derived from single CLGs (consistent with previous analyses of lamprey genomes<sup>19,20</sup>), hagfish chromosomes are evidently derived from these ancestral elements by irreversible fusions<sup>35</sup>, analogous to but distinct from the fusions observed on the gnathostome stem lineage<sup>19</sup> (Fig. 1c). The direct and largely one-to-one segmental correspondence between hagfish and lamprey is consistent with the previous assumption<sup>20,24</sup> that the cyclostomes share the same duplication history, although more detailed phylogenetic analysis is required to rule out scenarios of convergent duplications.

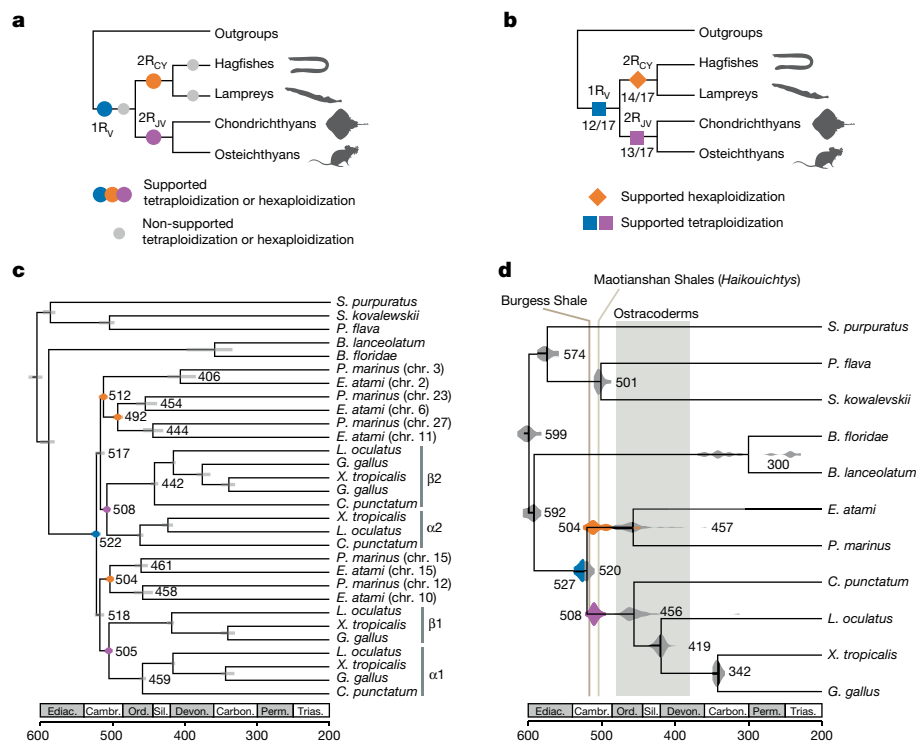
## Genome duplications in early vertebrates

We used two complementary phylogenetic approaches to fully resolve the sequence of early vertebrate polyploidization events: (i) model-based polyploidization inference from a large number of individual gene trees and (ii) concatenation of genes with similar evolutionary histories on the basis of chromosome-scale synteny. We first tested alternative scenarios for the sequence of whole-genome duplications (WGDs) based on 8,931 individual gene trees by probabilistic reconciliation of gene and species trees using WHALE<sup>36</sup> (Methods). This analysis provided significant support for the occurrence of a single genome duplication in the vertebrate stem lineage ( $1R_v$ ), followed by independent polyploidizations on the gnathostome ( $2R_{gv}$ ) and cyclostome ( $2R_{cv}$ ) stem lineages (all Bayes factors  $BF_{Null,vs.WGD} < 10^{-3}$ ) (Fig. 2a and Extended Data Fig. 4). By contrast, we found no support for a second round of polyploidization on the vertebrate stem lineage ( $2R_v$ ) or for polyploidization events specific to the lamprey or hagfish lineages (Fig. 2a and Extended Data Fig. 4), consistent with synteny-based analysis<sup>19,20</sup> and Fig. 1d.

We also developed a synteny-based approach that takes advantage of the shared evolutionary history of persistently linked genes to enhance the limited phylogenetic signal of individual gene trees and avoid the confounding effects of differential gene loss<sup>21,37,38</sup>. In this approach, we determined the duplication history of each CLG by concatenating genes from its derivatives in hagfish, lamprey and several jawed vertebrates (for example, chromosomes or chromosomal segments of the same colour in Fig. 1d; Methods). Within each species, the paralogous chromosome segments are called ‘paralogons’<sup>39</sup>. Because the CLGs are preserved in diverse invertebrates<sup>35</sup> corresponding sets of chromosomally linked genes can be concatenated to provide outgroups for phylogenetic and molecular dating. We reconstructed paralogon-based molecular phylogenies for 17 of the 18 CLGs or proto-vertebrate chromosomes (PVCs) (Supplementary Table 5); the 18th group (CLG A2 in the notation of ref. 35, and PVC18 in ref. 20) contains a relatively small number of genes consistently linked across vertebrate taxa and has anomalous properties in both lampreys and gnathostomes<sup>20</sup>.

Paralogon-based molecular phylogenies support a single early vertebrate auto-tetraploidization ( $1R_v$ ) before the cyclostome–gnathostome split, followed by a later gnathostome-specific allo-tetraploidization ( $2R_{gv}$ ) and a cyclostome-specific polyploidy ( $2R_{cv}$ ) (for example, CLGJ in Fig. 2b; Extended Data Fig. 5 and Supplementary Fig. 1). The  $1R_v$  duplication node precedes cyclostome–gnathostome speciation in 12 of the 14 CLG paralogon phylogenies with bootstrap support  $BP > 60$  (Supplementary Table 5 and Supplementary Fig. 1). A single shared tetraploidization on the vertebrate stem ( $1R_v$ ) is therefore consistent with both probabilistic inference of genome duplications from single gene trees and paralogon-based phylogenies. Molecular dating indicates that these duplication and speciation events occurred in close succession, and we estimate that the divergence of  $1R_v$  paralogons was completed by around 527 million years ago (Ma) and that the cyclostome–gnathostome split occurred around 520 Ma.

Our estimated date for the divergence of  $1R_v$  paralogons corresponds to the cessation of homeologous recombination (rediploidization) rather than the WGD itself, as noted in a previous study<sup>40</sup>. We tested for lineage-specific rediploidization across CLGs (relative to the gnathostome–cyclostome divergence) by comparing the likelihoods of gene trees under the ancestral and lineage-specific rediploidization models, as previously proposed<sup>41–45</sup> (Fig. 3a and Extended Data Fig. 6). We found that ancestral rediploidization after  $1R_v$  was supported by a larger number of significant gene trees for all CLGs (Fig. 3b), indicating that meiotic rediploidization was essentially complete by the time of the cyclostome–gnathostome split. This contrasts with other more recent vertebrate auto-polyploidizations, in which a number of homeologous chromosomes have maintained tetrasomic inheritance through subsequent speciation events<sup>41–44</sup>.



**Fig. 2 | History of genome duplications in vertebrates.** **a**, Probabilistic inference of polyploidization events in early vertebrate evolution on the basis of gene tree–species tree reconciliation (WHALE<sup>36</sup>; Extended Data Fig. 4, Supplementary Table 8 and Methods) supports an initial tetraploidization shared by all vertebrates (1R<sub>v</sub>), a jawed-vertebrate-specific tetraploidization (2R<sub>jv</sub>) and a cyclostome-specific polyploidization (2R<sub>cy</sub>). Supported polyploidization events (Bayes factors BF<sub>Null vs WGD</sub> < 10<sup>-3</sup>) are shown in colour (1R<sub>v</sub>, 2R<sub>jv</sub> and 2R<sub>cy</sub>) and non-supported ones in grey (2R<sub>v</sub>, hagfish-specific, lamprey-specific). The WHALE method cannot distinguish between tetraploidization and hexaploidization events. **b**, Paralogon-based polyploidization inference using molecular phylogenies reconstructed for each of the 17 informative CLGs (Supplementary Fig. 1). Successive polyploidization events during vertebrate evolution are shown as coloured polygons and the proportion of CLG trees

displaying these duplication nodes is indicated below. **c**, Sample paralogon tree for CLGJ. As for gene trees, in paralogon trees some nodes correspond to speciation events (grey) and others to duplication events (coloured); both types of events can be dated using a molecular clock. Species and datasets used are listed in Supplementary Table 8, and dating was performed with PhyloBayes (Methods) using fossil calibrations reported in Supplementary Table 7. **d**, Molecular dating of the polyploidizations and speciation events in early vertebrate evolution. Divergence times are indicated for speciation (grey) and duplication nodes (coloured as in **a**) are indicated. In **c, d**, each node is labelled with the mean divergence time across CLGs. Ediac., Ediacaran; Cambr., Cambrian; Ord., Ordovician; Sil., Silurian; Devon., Devonian; Carbon., Carboniferous; Perm., Permian; Trias., Triassic.

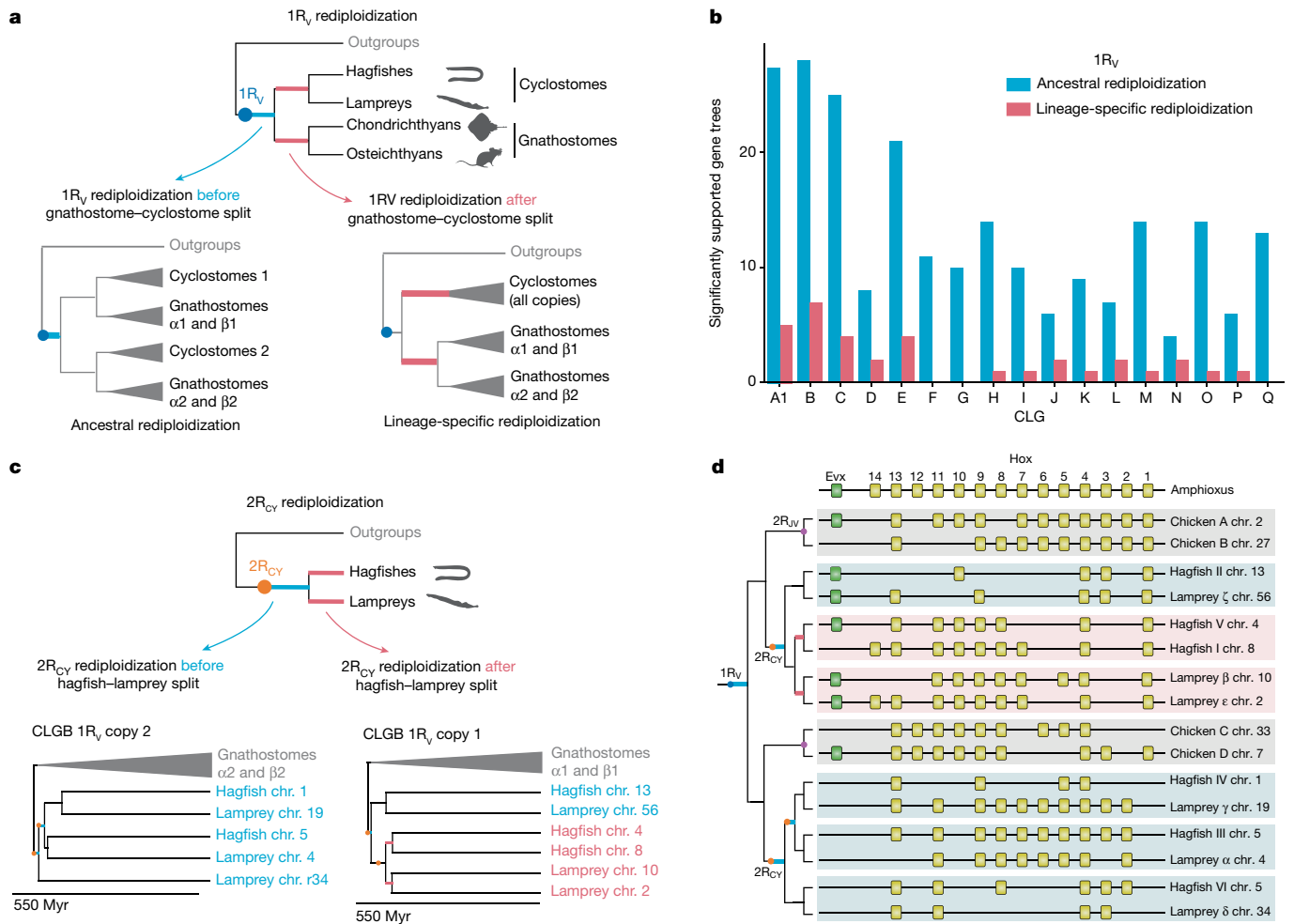
Unfortunately, molecular phylogenetics can only estimate a later bound on the timing of the 1R<sub>v</sub> auto-tetraploidization event itself, because it is likely to be obscured by a period of homoeologous recombination of unknown duration: the 1R<sub>v</sub> duplication event could have predated the divergence time of 1R<sub>v</sub> ohnologues (around 527 Ma) by millions of years<sup>40</sup>.

### Distinct duplications in cyclostomes

Paralogon-based molecular phylogenies also strongly support and refine the 2R<sub>jv</sub> allo-tetraploidization scenario<sup>19,20</sup> (Supplementary Table 6). Molecular dating of paralogon trees places the split of the pre-2R<sub>jv</sub> alpha and beta progenitors in the middle Cambrian, around 508 Ma (Fig. 2d, Extended Data Fig. 5, Supplementary Table 7 and Methods). The allo-tetraploidization event itself (the hybridization of alpha and beta progenitors and subsequent associated genome doubling), however, occurred some time after the alpha–beta divergence and cannot itself be precisely placed using molecular phylogenies. In recent vertebrate allo-tetraploids such as *Xenopus*<sup>46</sup> and goldfish<sup>47</sup>, hybridization occurred within 10–15 million years of the divergence of progenitors. If we take these as analogies for gnathostome allo-tetraploidy, then 2R<sub>jv</sub> probably occurred in the late Cambrian, long before the origin of crown-group gnathostomes around 456 Ma near the middle–late Ordovician boundary.

Among cyclostomes, paralogon trees confirm the general orthology of hagfish and lamprey chromosome segments (Fig. 2c), as suggested above (Supplementary Table 5 and Supplementary Fig. 1). We typically observed one or two duplication nodes for each CLG, indicating shared cyclostome genome-wide duplications that took place before the hagfish–lamprey split around 457 Ma. Although the nature of the cyclostome-specific duplications is difficult to decipher owing to extensive losses, the net effect appears to be hexaploidization<sup>20</sup> without the obvious patterns of differential gene retention that are typical of allo-polyploidy and are observed in the gnathostome lineage<sup>19,20</sup> (Extended Data Fig. 7).

The bimodal distribution of divergence times observed between homoeologous (that is, paralogous) cyclostome chromosomes (peaks at around 511 and around 493) Ma; see Extended Data Fig. 5a) is consistent with two-step hexaploidy through the hybridization of diploids and related tetraploids (Extended Data Fig. 5b), as seen in sturgeons<sup>48</sup>. Although the near one-to-one relationship between orthologous hagfish and lamprey chromosome segments (Fig. 1d) suggests that rediploidization after 2R<sub>cy</sub> was largely completed by the origination of crown-group cyclostomes, we formally tested for lineage-specific rediploidization (Fig. 3c). From the estimated paralogon divergence times and concatenated paralogon tree topologies, we identified a single case of lineage-specific rediploidization that affected CLGB paralogons in hagfish and lampreys after 2R<sub>cy</sub>. Specifically, the paralogon



**Fig. 3 | Limited lineage-specific rediploidization after vertebrate genome duplications.** **a**, Gene-tree topologies expected under the ancestral rediploidization (left) and lineage-specific rediploidization (right) models after the  $1R_v$  (Methods and Extended Data Fig. 6). In the ancestral rediploidization scenario, paralogous gene sequences diverge before the cyclostome–gnathostome split and thus group by duplicated gene copy. In the lineage-specific rediploidization scenario, paralogue sequences diverge independently in the stem gnathostome and cyclostome lineages, and thus genes are grouped by lineage. **b**, Number of significantly supported gene trees in favour of ancestral

and lineage-specific rediploidization scenarios after  $1R_v$ , for each of 17 informative ancestral linkage groups (CLGs). **c**, Tree topologies expected under the ancestral and lineage-specific rediploidization models after  $2R_{CY}$ . The CLGB paralogon tree shows an ancestral rediploidization topology for  $1R_v$  copy 2, but lineage-specific rediploidization for  $1R_v$  copy 1, where two hagfish (chr. 4 and chr. 5) and two lamprey (chr. 10 and chr. 2) paralogons independently rediploidized. Myr, million years. **d**, Evolutionary history of vertebrate Hox gene clusters resolved by the CLGB paralogon phylogeny (see bottom of **c**).

### Evolution of vertebrate Hox clusters

Notably, CLGB contains the Hox cluster, a key locus in early analyses of vertebrate WGD (ref. 40). A more targeted phylogenetic analysis of concatenated Hox and bystander genes also recovered the same lineage-specific rediploidization tree topology as the full CLG paralogon analysis (Extended Data Fig. 8). Through our CLGB paralogon and Hox-plus-bystander trees, we fully resolve the evolutionary history of the four gnathostome and six cyclostome Hox clusters after the  $1R_v$ ,  $2R_{CY}$  and  $2R_{JV}$  events (Fig. 3d). Although previous studies did not

identify one-to-one relationships between Hox clusters in lampreys and hagfishes<sup>24</sup>, we report unambiguous orthologies between lamprey Hox  $\zeta$ ,  $\alpha$ ,  $\gamma$  and  $\delta$  and hagfish Hox II, III, IV and VI, with groupings as follows:  $\zeta$ –II,  $\alpha$ –III,  $\gamma$ –IV and  $\delta$ –VI. By contrast, after lineage-specific rediploidization of CLGB, no true (that is, one-to-one) orthology relationships exist between lamprey Hox  $\beta$  and  $\epsilon$  and hagfish Hox V and I, which should be considered as ‘tetralogues’<sup>49</sup>.

### Origin of neural crest

Our paralogon-based classification makes it possible to robustly assign paralogues to specific duplications, and this sheds some light on the relative origin, with respect to WGDs, of vertebrate characteristics such as neural crest, placodes and hormone systems<sup>50</sup>. As highlighted previously<sup>51</sup>, establishing whether both paralogous branches retain (or partition) the ancestral role can help to pinpoint whether a character is likely to have emerged before or after  $1R_v$ .

To assess the origin of neural crest, we considered a set of 22 gene families involved in the specification and migration of neural crest<sup>52,53</sup> (Fig. 4a and Supplementary Table 8) and we asked whether

corresponding  $1R_V$  paralogues perform neural crest cell (NCC)-related functions in gnathostomes and in lampreys on the basis of the literature<sup>54–56</sup> and available RNA sequencing (RNA-seq) data<sup>56</sup>. We find that for many of these gene families, including *Tfap2*, *SoxE*, *EdnR*, *Twist1* and *Gata3*, paralogues on both  $1R_V$  branches are involved in neural-crest-related functions (Supplementary Table 9). This pattern indicates that NCC-related functions were inherited from pre- $1R_V$  genes in the vertebrate ancestor, and thus suggests that the neural crest originated before  $1R_V$ . Post-WGD subfunctionalization had a limited role in its emergence, contrary to other gnathostome novelties such as limbs<sup>57</sup>. Consistently, in lampreys, an alternative  $1R_V$  paralogue also seems to be involved in NCCs, involving, for example, *Gata3*, *Six1*, *Msx1* and possibly *FoxD* (Fig. 4a).

By contrast, the establishment of the trunk and cranial NCCs seems to differ among cyclostomes, osteichthyans and even amniotes, with distinct genes being involved<sup>56</sup>. Some but not all of the genes involved in this process seem to show a more recent occurrence of subfunctionalization. For instance, *Lhx5*, *Id3*, *Gid2* and *Dmbx*, which have a role in gnathostome cranial NCCs, do not have  $1R_V$  or  $2R_{JV}$  paralogues with a similar function, whereas *Tfap2* and *SoxE*, which are involved in the ancestral specification of both cranial and trunk NCCs, have paralogues on both of the  $1R_V$  branches that are involved in this function. Lampreys show marked differences: neither *RhoB* nor *Ets* are involved or expressed in lamprey migratory NCCs, and *Lhx5*, *Dmbx* and *Ets1* are expressed in later NCC derivatives (Extended Data Fig. 9c and Supplementary Table 10). Despite the extensive gene loss experienced in the hagfish lineage (see below), we recovered homologues for most of the NCC-related genes that we investigated. Further functional studies will be necessary to determine whether subsequent  $2R_{CV}$  paralogues in hagfish were incorporated in NCC-related functions specific to this lineage<sup>52</sup>.

### A distinct fate for paralogues

Paralogues retained in gnathostomes after two rounds of genome duplication were previously shown to be functionally associated with the regulation of development and nervous system activity<sup>18</sup>. To determine whether similar genes were retained preferentially in multiple copies in the cyclostome lineage after  $1R_V$  and/or after cyclostome-specific  $2R_{CV}$ , we tested paralogue sets that show distinct retention patterns for functional enrichment (Fig. 4b). We recovered gene ontology terms that were previously found to be enriched in gnathostome paralogues (for example, axon guidance and embryonic organ development), but, notably, we found that they were preferentially associated with paralogues retained after the pan-vertebrate  $1R_V$  regardless of their post-gnathostome duplication  $2R_{JV}$  fate (Fig. 4b), suggesting that  $1R_V$  had a key role in the early elaboration of the vertebrate nervous system. In cyclostomes, however, these terms are preferentially associated with paralogues that were systematically retained after all polyploidizations ( $1R_V$  and  $2R_{CV}$ ); this suggests a distinct path of paralogue evolution at the functional level, possibly coupled with an increased retention after  $2R_{CV}$  compared with  $2R_{JV}$ .

The fate of paralogues after WGD is often related to their acquisition of more specific expression domains that can explain subfunctionalization and functional innovation<sup>58–60</sup>. To examine patterns of divergence in gene expression in gnathostomes and cyclostomes, we compared paralogues across a consistent set of six organs in amphioxus, gar, lampreys and hagfish. Considering 3,009 gene families, we found a higher level of gene-expression specificity in gar than in lampreys and hagfish, with the hagfish showing the least specificity (Extended Data Fig. 10b). We then counted the number of expression patterns that were gained or lost in the same gene family between amphioxus and the different vertebrate species, which also indicated a lower level of subfunctionalization in cyclostomes than in gnathostomes (Fig. 4c). Finally, we asked whether particular organs show a significant enrichment of

paralogous genes using gene-expression clustering (weighted gene co-expression network analysis (WGCNA) (Methods and Extended Data Fig. 9a,b). Of note, we found that only neural tissue exhibits enrichment in both gnathostomes (for example, gar) and hagfish, whereas many recently duplicated genes are expressed in an organ-specific manner (Extended Data Fig. 9b). Together, these results imply that cyclostomes—and, to a greater extent, hagfish—show more limited subfunctionalization or specialization of expression patterns than do gnathostomes.

### Gene loss and hagfish novelties

Hagfish underwent the most extensive gene loss among vertebrates, with 1,386 missing gene families, of which 892 were present in the deuterostome ancestor (Fig. 4d and Extended Data Fig. 9d). Hagfishes stand out as having lost all members of several entire gene families, rather than exhibiting just an increased loss of paralogues (Extended Data Fig. 10c).

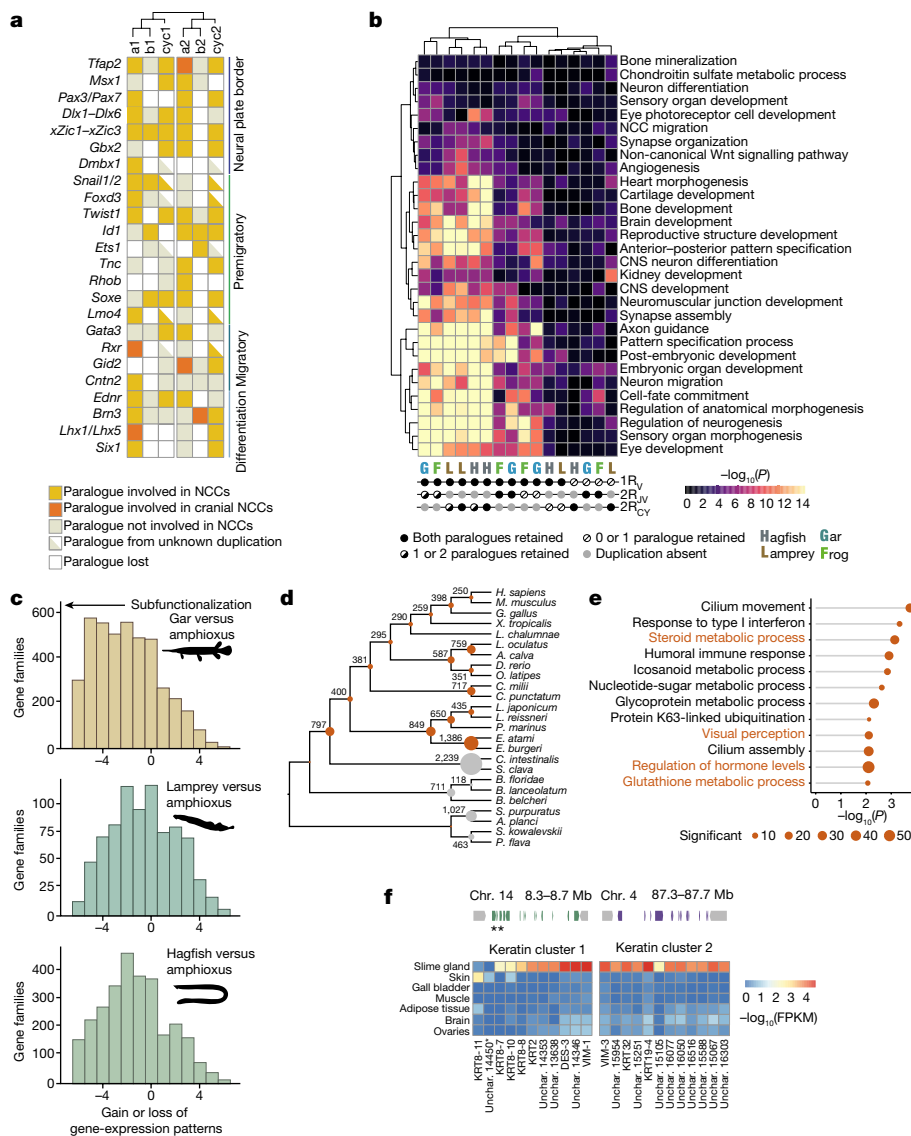
Several gene families lost in hagfish are functionally enriched for roles associated with missing characters in hagfish (Fig. 4e). For instance,  $\gamma$ -crystallins, which make up the lenses of vertebrate eyes, are absent in hagfish (but independently expanded in lampreys and gnathostomes), as are the *EYS* (eyes shut homologue) and *RBP3* (retinol-binding protein 3) genes that are involved in photoreceptor maintenance and development<sup>61</sup> (Supplementary Table 11). Several genes that are involved in bone development and its hormonal control in other vertebrates<sup>62</sup> are absent in hagfish: two members of the RANK–osteoprotegerin pathway that control osteoclast proliferation in gnathostomes<sup>63</sup>, as well as the genes encoding the parathyroid hormones (*PTH* and *PTLH*), which have a role in the regulation of calcium metabolism (their receptor is still present)<sup>64</sup>. These genes are present in lampreys and their loss in the hagfish lineage could be associated with the limited condensation of the hagfish vertebral cartilage.

Hagfish have also gained new traits, most notably their prodigious ability to secrete a highly viscous slime that acts as a defence against predators. We found two clusters of genes that are specifically and highly expressed in the slime gland (Extended Data Fig. 9e) and are related to intermediate filaments ( $\alpha$ -keratin)<sup>65</sup>. One of these clusters contains a gene that is expressed mainly in the skin but not in the slime gland, consistent with the recent suggestion that the keratin threads of hagfish slime could have originated as elements of the skin<sup>66</sup> (Fig. 4f). We found that the most highly expressed glycoproteins in the slime gland included von Willebrand A and D domains, rather than mucin-type domains as previously hypothesized<sup>67</sup>.

### Programmed DNA and gene elimination

The somatic and germline cells of hagfish exhibit distinct karyotypes, owing apparently to the loss of germline-specific chromosomes through embryonically programmed DNA elimination. On the basis of *k*-mer counts (Extended Data Fig. 1b), we estimate that around 1.3 Gb is lost from the approximately 3.3-Gb germline genome of *E. atami*, consistent with cytofluorometry<sup>25,28</sup>. Analysis of the genome assembly identified a large number of germline-specific genes and confirmed that germline-specific regions contain large numbers of complex repetitive elements<sup>30–33</sup>, including one newly identified repeat that accounts for 4% of the genome (Fig. 5, Extended Data Figs. 11 and 12 and Supplementary Note 3).

So far, no germline-specific genes have been identified in any hagfish species. We identified germline-specific genes in *E. atami* by comparing the read depth of germline and somatic reads across low- to medium-copy regions of the assembled genome (Methods and Fig. 5d). We discovered 81 Mb of germline-specific sequence that encode 1,654 genes, 226 of which have identifiable human homologues (to 121 non-redundant human genes) (Supplementary Table 12). We confirmed



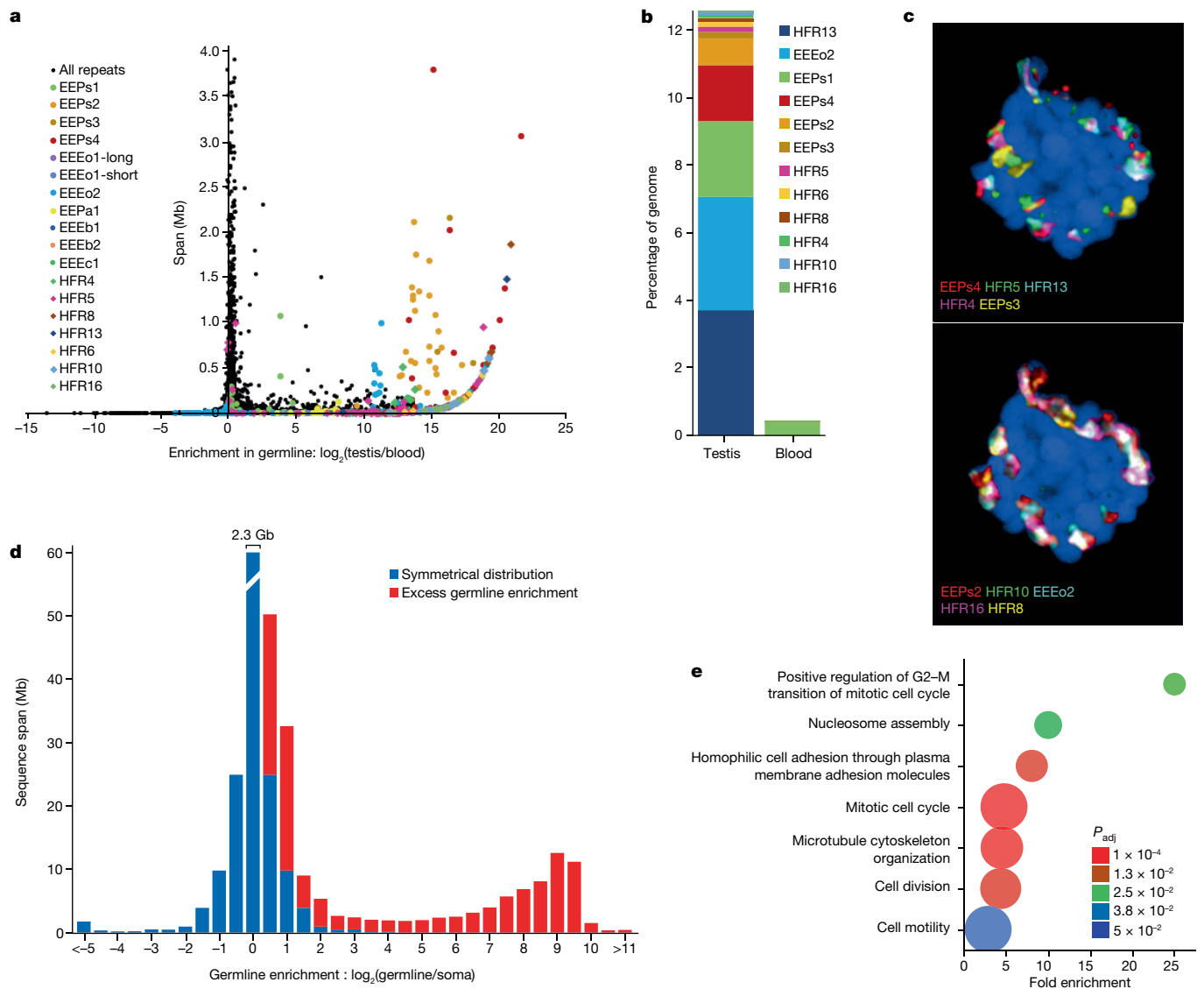
**Fig. 4 | Functional effects of vertebrate WGD and gene loss in vertebrates.**

**a**, Key neural-crest-related gene families with members classified according to their functional role (colour) and paralogy status relative to  $1R_v$  and  $2R_{v/y}$ . The involvement of paralogues derived from both copies of the  $1R_v$  in NCC-related function, in both gnathostomes and lampreys, supports the hypothesis that NCCs predate  $1R_v$ . **b**, Enrichment of functional annotation terms (gene ontology) in sets of genes showing a specific pattern of retention after vertebrate WGDs. Each column corresponds to a set of paralogous genes with a specific pattern of post-duplication retention in a given species. We distinguished cases in which both paralogues can be assigned to a specific duplication and are retained, cases in which at least one of the paralogues is retained and cases in which at least one of the two copies is lost. CNS, central nervous system. **c**, Distribution of the difference of positive organ-specific expression domains between selected vertebrate species and the amphioxus outgroup for ohnologous gene

that 44 of 46 tested germline-specific intervals can be PCR-amplified from testes but not blood DNA (95.7% validation rate) (Supplementary Table 13). Germline-specific genes in hagfish are enriched in several biological functions on the basis of gene ontology analyses, including functions related to cell cycle, cell motility and chromatin or DNA repair (Fig. 5e and Supplementary Tables 14 and 15). Similar functions were also enriched among germline-specific genes in sea lamprey, and support the hypothesis that somatically eliminated genes generally perform functions that benefit the development and maintenance of the germline<sup>11</sup>.

families<sup>59</sup>. A shift to the left in the distribution (as seen for the gar) indicates an extensive subfunctionalization through the restriction of gene-expression domains in vertebrates. **d**, Gene-family loss in deuterostomes, highlighting the severe loss in the hagfish lineage relative to that seen in other vertebrates and deuterostomes (grey). Species abbreviations are provided in Supplementary Table 8. **e**, Functional enrichment (gene ontology) for gene families lost in the hagfish lineages, highlighting a simplification of visual and hormonal systems (labels in orange). **f**, Structure of the two clusters of  $\alpha$ -keratin genes on chromosomes 14 and 4, and their expression in the slime gland and the skin shown as a heat map (gene expression expressed as fragments per kilobase per million reads (FPKM)). Unchar is the prefix used for naming genes that did not receive a gene name by homology search. Genes are shown in the same order in the heat map as they are located in the two clusters. Stars indicate the two genes that are expressed preferentially in the skin (Extended Data Fig. 10).

The broad functional similarity between germline-specific genes of hagfish and lamprey suggests that DNA elimination could be a shared ancestral feature of the cyclostome lineage<sup>8</sup>. To attempt to identify the vestiges of theoretical ancestral germline-specific chromosomes in the cyclostome lineage, we examined patterns of orthology and paralogy for eliminated genes. Despite the general functional similarity of eliminated genes in hagfish and sea lampreys, few orthologous genes were found to be eliminated in both genomes. In total, 7 of 121 non-redundant hagfish gene families were also eliminated in sea lampreys (*CDH1*, *CDH2* and *CDH4*; *GJCI*; *MSH4*; *NCAM1*; *SEMA4B* and *SEMA4C*; *WNT5A*, *WNT5B*,



**Fig. 5 | Germline-specific and enriched sequences and genes in hagfish.** **a**, Plot showing the degree of germline enrichment and estimated span of all predicted repetitive elements in the *E. atami* genome, focusing on elements with a cumulative span of less than 4 Mb (per family member). Previously identified elements<sup>30,33</sup> are highlighted by coloured circles and newly identified high-copy elements are highlighted by coloured diamonds. Additional higher copy repeats are visible in Extended Data Fig. 12m,n. The colouring scheme is the same in **b** and in Extended Data Fig. 12m,n. **b**, Estimated cumulative span of the eight most highly abundant repeats shown as the percentage of the genome covered. **c**, Fluorescence in situ hybridization (FISH) of high-copy

germline-specific repeats to a testes metaphase plate showing their distinct spatial clustering within chromosomes (blue counterstaining is NucBlue: Hoechst 33342; individual pairs of probes are shown in Extended Data Fig. 12m,n). **d**, Comparison of the sequence depth of DNA extracted from germline (testes) versus somatic (blood) tissues identifies a large number of genomic intervals with evidence for strong enrichment in the germline. The bin representing no enrichment contains a total of 2.3 Gb of the assembly. **e**, Genes encoded within germline-specific regions are enriched for several ontology terms related to regulation of cell cycle and cell motility (Panther Biological Processes: most specific subclass shown; Supplementary Table 14).

*WNT7A* and *WNT7B*; and *YTHDC2*; Extended Data Fig. 11). An analysis of gene trees indicates that three of these (orthologues of *MSH4*, *WNT7A* and *YTHDC2*; Extended Data Fig. 11) share a last common ancestor that can be traced to a single lineage after the basal vertebrate divergence and duplication events. This small set of genes might reflect the vestiges of shared germline-specific sequences that were eliminated early in the cyclostome lineage, or, alternatively, these genes might have been independently recruited to the germline-specific fraction during the early evolution of both lineages.

Germline-specific chromosomes in songbirds and lampreys are continuously capturing duplicates of somatic genes, establishing new germline-specific genes that often evolve rapidly owing to their unique selective and mutational genomic environment<sup>7,34</sup>. In *E. atami*,

we observe several germline-specific genes that have undergone extra rounds of duplication after duplicating or translocating to germline chromosomes. The genes with the highest germline-specific copy numbers are homologues of *FBXL4*, a modulator of E3 ubiquitin ligase that regulates the proteasomal turnover of the histone demethylase KDM4A (ref. 68) (25 copies); and *TRRAP*, a component of several histone acetyltransferase complexes (18 copies) (Extended Data Fig. 11 and Supplementary Table 12). The *FBXL4* orthogroup also contains 45 paralogues in the draft genome of the closely related hagfish *Eptatretus burgeri* (ref. 69), indicating that the origin of germline-specific *FBXL4* and the expansion of this gene family predates the split between the two hagfish species, with additional lineage-specific expansions and losses underlying differences in paralogue numbers over the past few



million years. These gene families seem to have undergone substantial expansion even in the recent past, emphasizing their high rate of turnover.

The accumulation of epigenetic silencing marks and regulated degradation has been implicated in the cellular mechanisms that underlie the elimination of lamprey germline-specific chromosomes<sup>8,70</sup>. This suggests that some components of hagfish DNA elimination mechanisms might be encoded by the germline-specific chromosomes themselves, or contribute to other aspects of hagfish germ cell development. Other genes involved in the same pathways as *FBXL4* and *TRRAP* have also duplicated in the context of the *E. atami* germline-specific chromosomes, albeit to a lesser extent. These include *KLHL10*, a component of the E3 ubiquitin ligase complex involved in spermatogenesis (4 copies); *SIN3*, a transcriptional repressor whose human homologue is highly expressed in the testis (4 copies); and *DNMT1*, the primary enzyme responsible for maintaining silencing DNA methylation marks after DNA replication (2 copies). Notably, each of these five families of germline-specific genes also possesses at least one somatically retained paralogue, indicating that germline-specific expansion of gene families related to ubiquitination and regulation of chromatin state has evolved in the context of largely intact ancestral somatic pathways.

## Conclusion

Early vertebrate evolution was accompanied by a series of ancient polyploidization events that have been difficult to unambiguously resolve using conventional sequence-based molecular phylogenetics. Challenges include the antiquity of these events and the relatively short intervals between them<sup>18,21</sup>, as well as lineage-specific evolution and gene loss after duplication<sup>21,37,38</sup>. We used the hagfish genome and an approach focused on chromosome-scale phylogenetics to fully resolve this history of ancient vertebrate polyploidies (Fig. 2 and Extended Data Fig. 5). The earliest duplication,  $1R_v$ , occurred on the vertebrate stem lineage in the early Cambrian (around 527 Ma), around 10 million years before the appearance of *Haikouichthys* and *Myllokunmingia* (ref. 3) the earliest vertebrate fossils. Whether the similarity in timing is coincidental or causal remains to be seen.

After the shared duplication, cyclostomes and gnathostomes experienced independent polyploidizations during the late Cambrian–early Ordovician, coinciding with a gap in the vertebrate fossil record. We can, however, begin to relate early genomic events to the emergence and elaboration of vertebrate innovations by correlating the contemporary functions of gene duplicates with their appearance at specific duplication events. For example, we find that when one gene functions in neural crest, its  $1R_v$  paralogues also do, as expected if the neural crest regulatory circuits already existed before  $1R_v$ . More speculatively, we note that *Evx* homeobox genes, which have a role in the development and patterning of paired fins and limbs, duplicated at  $1R_v$  with both lineages being retained in gnathostomes (*Evx1*–*HoxA* and *Evx2*–*HoxD*), but that cyclostomes are missing *HoxC* and *HoxD*-associated *Evx* paralogues owing to lineage-specific loss. This observation suggests that  $1R_v$  duplicates might have acquired roles in fin bud development and patterning very early in the evolution of the gnathostome lineage, consistent with the observation of paired fin fold morphologies in early diverging galeaspid<sup>71</sup>.

Finally, analysis of *E. atami* germline-specific chromosomes in comparison with other vertebrates supports the hypothesis that these chromosomes encode functions that are advantageous for the development of germ cells and the production of gametes, and indicates that rapid turnover of germline-specific gene content might be a common feature across highly divergent lineages. As with other features of their biology, differences in the gene content of lamprey and hagfish germline-specific chromosomes might reflect their long history of independent evolution and the marked differences in their reproductive, ecological and developmental biology that have accumulated over

the approximately 460 million years since the last common cyclostome ancestor.

## Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41586-024-07070-3>.

1. Shimeld, S. M. & Donoghue, P. C. J. Evolutionary crossroads in developmental biology: cyclostomes (lamprey and hagfish). *Development* **139**, 2091–2099 (2012).
2. Miyashita, T. et al. Hagfish from the Cretaceous Tethys Sea and a reconciliation of the morphological–molecular conflict in early vertebrate phylogeny. *Proc. Natl Acad. Sci. USA* **116**, 2146–2151 (2019).
3. Janvier, P. Facts and fancies about early fossil chordates and vertebrates. *Nature* **520**, 483–489 (2015).
4. Ohno, S. *Evolution by Gene Duplication* (Springer, 1970).
5. Holland, P. W., Garcia-Fernández, J., Williams, N. A. & Sidow, A. Gene duplications and the origins of vertebrate development. *Dev. Suppl.* **1994**, 125–133 (1994).
6. Donoghue, P. C. J. & Purnell, M. A. Genome duplication, extinction and vertebrate evolution. *Trends Ecol. Evol.* **20**, 312–319 (2005).
7. Timoshevskaya, N. et al. An improved germline genome assembly for the sea lamprey *Petromyzon marinus* illuminates the evolution of germline-specific chromosomes. *Cell Rep.* **42**, 112263 (2023).
8. Smith, J. J., Timoshevskiy, V. A. & Saraceno, C. Programmed DNA elimination in vertebrates. *Annu. Rev. Anim. Biosci.* **9**, 173–201 (2021).
9. Drotos, K. H. I., Zagoskin, M. V., Kess, T., Gregory, T. R. & Wyngaard, G. A. Throwing away DNA: programmed downsizing in somatic nuclei. *Trends Genet.* **38**, 483–500 (2022).
10. Smith, J. J., Antonacci, F., Eichler, E. E. & Amemiya, C. T. Programmed loss of millions of base pairs from a vertebrate genome. *Proc. Natl Acad. Sci. USA* **106**, 11212–11217 (2009).
11. Smith, J. J. et al. The sea lamprey germline genome provides insights into programmed genome rearrangement and vertebrate evolution. *Nat. Genet.* **50**, 270–277 (2018).
12. Spitzer, R. H. & Koch, E. A. in *The Biology of Hagfishes* (eds Jørgensen, J. M. et al.) 109–132 (Springer, 1998).
13. Duméril, A. M. C. *Dissertation sur la famille des poissons cyclostomes, pour démontrer leurs rapports avec les animaux sans vertèbres* (Didot, 1812).
14. Ota, K. G., Fujimoto, S., Oisi, Y. & Kuratani, S. Identification of vertebra-like elements and their possible differentiation from sclerotomes in the hagfish. *Nat. Commun.* **2**, 373 (2011).
15. Mallatt, J. & Sullivan, J. 28S and 18S rDNA sequences support the monophyly of lampreys and hagfishes. *Mol. Biol. Evol.* **15**, 1706–1718 (1998).
16. Kuraku, S. & Kuratani, S. Time scale for cyclostome evolution inferred with a phylogenetic diagnosis of hagfish and lamprey cDNA sequences. *Zool. Sci.* **23**, 1053–1064 (2006).
17. Delsuc, F., Tsagkogeorga, G., Lartillot, N. & Philippe, H. Additional molecular support for the new chordate phylogeny. *Genesis* **46**, 592–604 (2008).
18. Putnam, N. H. et al. The amphioxus genome and the evolution of the chordate karyotype. *Nature* **453**, 1064–1071 (2008).
19. Simakov, O. et al. Deeply conserved synteny resolves early events in vertebrate evolution. *Nat. Ecol. Evol.* **4**, 820–830 (2020).
20. Nakatani, Y. et al. Reconstruction of proto-vertebrate, proto-cyclostome and proto-gnathostome genomes provides new insights into early vertebrate evolution. *Nat. Commun.* **12**, 4489 (2021).
21. Kuraku, S., Meyer, A. & Kuratani, S. Timing of genome duplications relative to the origin of the vertebrates: did cyclostomes diverge before or after? *Mol. Biol. Evol.* **26**, 47–59 (2009).
22. Smith, J. J. et al. Sequencing of the sea lamprey (*Petromyzon marinus*) genome provides insights into vertebrate evolution. *Nat. Genet.* **45**, 415–421 (2013).
23. Sacerdot, C., Louis, A., Bon, C., Berthelot, C. & Roest Crollius, H. Chromosome evolution at the origin of the ancestral vertebrate genome. *Genome Biol.* **19**, 166 (2018).
24. Pascual-Anaya, J. et al. Hagfish and lamprey Hox genes reveal conservation of temporal colinearity in vertebrates. *Nat. Ecol. Evol.* **2**, 859–866 (2018).
25. Nakai, Y., Kubota, S. & Kohno, S. Chromatin diminution and chromosome elimination in four Japanese hagfish species. *Cytogenet. Cell Genet.* **56**, 196–198 (1991).
26. Caputo Barucchi, V., Giovannotti, M., Nisi Cerioni, P. & Splendiani, A. Genome duplication in early vertebrates: insights from agnathan cytogenetics. *Cytogenet. Genome Res.* **141**, 80–89 (2013).
27. Nakai, Y. & Kohno, S. Elimination of the largest chromosome pair during differentiation into somatic cells in the Japanese hagfish, *Myxine garmani* (Cyclostomata, Agnatha). *Cytogenet. Genome Res.* **45**, 80–83 (1987).
28. Nakai, Y. et al. Chromosome elimination in three Baltic, south Pacific and north-east Pacific hagfish species. *Chromosome Res.* **3**, 321–330 (1995).
29. Smith, J. J., Baker, C., Eichler, E. E. & Amemiya, C. T. Genetic consequences of programmed genome rearrangement. *Curr. Biol.* **22**, 1524–1529 (2012).
30. Kubota, S., Kuro-o, M., Mizuno, S. & Kohno, S. Germ line-restricted, highly repeated DNA sequences and their chromosomal localization in a Japanese hagfish (*Eptatretus okinoseanus*). *Chromosoma* **102**, 163–173 (1993).
31. Goto, Y., Kubota, S. & Kohno, S. Highly repetitive DNA sequences that are restricted to the germ line in the hagfish *Eptatretus cirrhatus*: a mosaic of eliminated elements. *Chromosoma* **107**, 17–32 (1998).
32. Nabeyama, M., Kubota, S. & Kohno, S. Concerted evolution of a highly repetitive DNA family in eptatretidae (Cyclostomata, agnatha) implies specifically differential homogenization and amplification events in their germ cells. *J. Mol. Evol.* **50**, 154–169 (2000).

33. Kojima, N. F. et al. Whole chromosome elimination and chromosome terminus elimination both contribute to somatic differentiation in Taiwanese hagfish *Paramyxine sheni*. *Chromosome Res.* **18**, 383–400 (2010).
34. Kinsella, C. M. et al. Programmed DNA elimination of germline development genes in songbirds. *Nat. Commun.* **10**, 5468 (2019).
35. Simakov, O. et al. Deeply conserved synteny and the evolution of metazoan chromosomes. *Sci. Adv.* **8**, eabi5884 (2022).
36. Zwaenepoel, A. & de Van de Peer, Y. Inference of ancient whole-genome duplications and the evolution of gene duplication and loss rates. *Mol. Biol. Evol.* **36**, 1384–1404 (2019).
37. Holland, P. W. H., Marlétaz, F., Maeso, I., Dunwell, T. L. & Paps, J. New genes from old: asymmetric divergence of gene duplicates and the evolution of development. *Philos. Trans. R. Soc. B* **372**, 20150480 (2017).
38. Kuraku, S. Palaeophylogenomics of the vertebrate ancestor—impact of hidden paralogy on hagfish and lamprey gene phylogeny. *Integr. Comp. Biol.* **50**, 124–129 (2010).
39. Coulier, F., Popovici, C., Villet, R. & Birnbaum, D. MetaHox gene clusters. *J. Exp. Zool.* **288**, 345–351 (2000).
40. Furlong, R. F. & Holland, P. W. H. Were vertebrates octoploid? *Philos. Trans. R. Soc. B* **357**, 531–544 (2002).
41. Robertson, F. M. et al. Lineage-specific rediploidization is a mechanism to explain time-lags between genome duplication and evolutionary diversification. *Genome Biol.* **18**, 111 (2017).
42. Gundappa, M. K. et al. Genome-wide reconstruction of rediploidization following autopolyploidization across one hundred million years of salmonid evolution. *Mol. Biol. Evol.* **39**, msab310 (2022).
43. Parey, E. et al. An atlas of fish genome evolution reveals delayed rediploidization following the teleost whole-genome duplication. *Genome Res.* **32**, 1685–1697 (2022).
44. Redmond, A. K., Casey, D., Gundappa, M. K., Macqueen, D. J. & McLysaght, A. Independent rediploidization masks shared whole genome duplication in the sturgeon-paddlefish ancestor. *Nat. Commun.* **14**, 2879 (2023).
45. Lien, S. et al. The Atlantic salmon genome provides insights into rediploidization. *Nature* **533**, 200–205 (2016).
46. Session, A. M. et al. Genome evolution in the allotetraploid frog *Xenopus laevis*. *Nature* **538**, 336–343 (2016).
47. Chen, Z. et al. De novo assembly of the goldfish (*Carassius auratus*) genome and the evolution of genes after whole-genome duplication. *Sci. Adv.* **5**, eaav0547 (2019).
48. Fontana, F. et al. Evidence of hexaploid karyotype in shortnose sturgeon. *Genome* **51**, 113–119 (2008).
49. Martin, K. J. & Holland, P. W. H. Enigmatic orthology relationships between Hox clusters of the African butterfly fish and other teleosts following ancient whole-genome duplication. *Mol. Biol. Evol.* **31**, 2592–2611 (2014).
50. Shimeld, S. M. & Holland, P. W. Vertebrate innovations. *Proc. Natl Acad. Sci. USA* **97**, 4449–4452 (2000).
51. Wada, H. & Makabe, K. Genome duplications of early vertebrates as a possible chronicle of the evolutionary history of the neural crest. *Int. J. Biol. Sci.* **1449**, 228 (2006).
52. Martik, M. L. & Bronner, M. E. Riding the crest to get a head: neural crest evolution in vertebrates. *Nat. Rev. Neurosci.* **22**, 616–626 (2021).
53. Simões-Costa, M. & Bronner, M. E. Establishing neural crest identity: a gene regulatory recipe. *Development* **142**, 242–257 (2015).
54. Sauka-Spengler, T., Meulemans, D., Jones, M. & Bronner-Fraser, M. Ancient evolutionary origin of the neural crest gene regulatory network. *Dev. Cell* **13**, 405–420 (2007).
55. Hockman, D. et al. A genome-wide assessment of the ancestral neural crest gene regulatory network. *Nat. Commun.* **10**, 4689 (2019).
56. Martik, M. L. et al. Evolution of the new head by gradual acquisition of neural crest regulatory circuits. *Nature* **574**, 675–678 (2019).
57. Minguillon, C., Gibson-Brown, J. J. & Logan, M. P. *Tbx4/5* gene duplication and the origin of vertebrate paired appendages. *Proc. Natl Acad. Sci. USA* **106**, 21726–21730 (2009).
58. Lynch, M. & Conery, J. S. The evolutionary fate and consequences of duplicate genes. *Science* **290**, 1151–1155 (2000).
59. Marlétaz, F. et al. Amphioxus functional genomics and the origins of vertebrate gene regulation. *Nature* **564**, 64–70 (2018).
60. Force, A. et al. Preservation of duplicate genes by complementary, degenerative mutations. *Genetics* **151**, 1531–1545 (1999).
61. Dong, E. M. & Allison, W. T. Vertebrate features revealed in the rudimentary eye of the Pacific hagfish (*Eptatretus stoutii*). *Proc. Biol. Sci.* **288**, 20202187 (2021).
62. Venkatesh, B. et al. Elephant shark genome provides unique insights into gnathostome evolution. *Nature* **505**, 174–179 (2014).
63. Theill, L. E., Boyle, W. J. & Penninger, J. M. RANK-L and RANK: T cells, bone loss, and mammalian evolution. *Annu. Rev. Immunol.* **20**, 795–823 (2002).
64. Poole, K. E. S. & Reeve, J. Parathyroid hormone—a bone anabolic and catabolic agent. *Curr. Opin. Pharmacol.* **5**, 612–617 (2005).
65. Fudge, D. S. et al. From ultra-soft slime to hard  $\alpha$ -keratins: the many lives of intermediate filaments. *Integr. Comp. Biol.* **49**, 32–39 (2009).
66. Zeng, Y. et al. Epidermal threads reveal the origin of hagfish slime. *eLife* **12**, e81405 (2023).
67. Fudge, D. S., Levy, N., Chiu, S. & Gosline, J. M. Composition, morphology and mechanics of hagfish slime. *J. Exp. Biol.* **208**, 4613–4625 (2005).
68. Van Rechem, C. et al. The SKP1-Cul1-F-box and leucine-rich repeat protein 4 (SCF-FbxL4) ubiquitin ligase regulates lysine demethylase 4A (KDM4A)/Jumonji domain-containing 2A (JMJD2A) protein. *J. Biol. Chem.* **286**, 30462–30470 (2011).
69. Nishimura, O. et al. Inference of a genome-wide protein-coding gene set of the inshore hagfish *Eptatretus burgeri*. *F1000Res.* **11**, 1270 (2022).
70. Timoshevskiy, V. A., Herdy, J. R., Keinath, M. C. & Smith, J. J. Cellular and molecular features of developmentally programmed genome rearrangement in a vertebrate (sea lamprey: *Petromyzon marinus*). *PLoS Genet.* **12**, e1006103 (2016).
71. Gai, Z. et al. Galeaspid anatomy and the origin of vertebrate paired appendages. *Nature* **609**, 959–963 (2022).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2024

## Methods

### Genome sequencing and assembly

DNA was extracted from a testis from a male *E. atami* individual and extracted using proteinase K digestion and phenol:chloroform extraction<sup>72</sup>. Animals were sampled in Suruga Bay, off Yaizu (300–330-m depth) and maintained in seawater aquariums at 11–13 °C. In agreement with procedures authorized by the Guidelines for Proper Conduct of Animal Experiments by the Science Council of Japan (2006), animals were anaesthetized using Tricaine (MS222, Sigma) before euthanasia and dissection. Paired-end and mate-pairs Illumina libraries were generated using Illumina Truseq and Nextera Mate-pair kits and sequenced on HiSeq2000 and HiSeq2500 instruments (Supplementary Table 1). The Illumina dataset was assembled using Meraculous (v.2.2.2.5) with a *k*-mer of 71 and ‘diploid mode’ set to ‘1’ to attempt the merging haplotypes<sup>73</sup>, and subsequently scaffolded using mate-pairs information (Supplementary Table 2). PacBio long-reads data at around 35× coverage were generated on a PacBio RSII instrument (Supplementary Table 2) and incorporated using PBJelly (v.15.8.24)<sup>74</sup>. PBJelly aligns the PacBio reads to the assembly using the Blasr aligner and collects reads surrounding and spanning gaps. Sequences assembled from these spanning reads are used to fill gaps and extend scaffolds. We used the parameters ‘-minMatch 8 -sdpTupleSize 8 -minPctIdentity 75 -bestN 1 -nCandidates 10 -maxScore -500’ for Blasr alignment.

The gap-filled assembly was further scaffolded using proximity ligation information. We used both Chicago libraries relying on syntenic reconstructed chromatin and Hi-C libraries capturing the native chromatin contacts, and scaffolding was performed using the HiRise package<sup>75</sup>. Hagfish liver was cross-linked in 1% paraformaldehyde, and chromatin was subsequently extracted, immobilized on SPRI beads, washed and digested with DpnII (ref. 76). After end-labelling, proximity ligation was performed using T4 DNA ligase and cross-linking was reversed using proteinase K. The DNA fragments were removed from the beads and then purified again on SPRI beads. The sequencing library was constructed using the NEB Ultra Library Preparation Kit (New England Biolabs).

The genome-wide heterozygosity was estimated to be 0.9%. The final BUSCO score (Metazoa) is C:90.0% (S:89.8%, D:0.2%, F:4.0%, M:6.0%, n:954). The size of the hagfish genome was estimated by counting 21-mers with Meryl (v.1.1)<sup>77</sup>. Using a fitting four-peak model as implemented in Genomescope2, the estimated size is 2.02 Gb and 3.28 Gb using sequencing data from blood and testis DNA, respectively<sup>78</sup> (Extended Data Fig. 1b).

### Transcriptome and genome annotation

We generated RNA-seq data for 13 organs with an average depth of 25 million reads. We aligned the reads to the genome using STAR (v.2.5.2b) with an average 78.7% uniquely mapping reads<sup>79</sup>. These alignments were used to assemble transcriptomes for each organ using StringTie (v.1.3.3b) and subsequently merged together using Taco<sup>80</sup>. In parallel, a de novo assembly of the bulk RNA-seq data was performed using Trinity (v.2.11.0) both in reference-free and genome-guided mode<sup>81</sup>.

We also sequenced full-length cDNA from brain RNA on eight cells of Pacbio RSII. Following the Iso-Seq protocol, circular consensus of subreads were calculated and validated as full length on the basis of the presence of SMART adaptors at both extremities. Full-length transcripts were clustered and polished using all circular consensus reads with Quiver (v.2.0.0), yielding 23,343 high-quality transcripts.

Assembled transcripts from de novo and genome-guided Trinity and high-quality Iso-Seq transcripts were aligned to the genome using GMAP (v.2018-03-25). Mikado (v.1.2.1) was used to generate a high-quality reference transcriptome leveraging (i) the aligned Trinity de novo and genome-based transcriptomes; (ii) the Iso-Seq transcripts; (iii) the StringTie transcriptomes merged with Taco and a set of curated splice junctions generated from RNA-seq alignments using Portcullis

(v.1.0.2). Putative fusion transcripts were detected by Blast comparison against Swiss-Prot and ORFs were annotated using TransDecoder<sup>82</sup>. Transcripts derived from the reference transcriptome were selected to train the Augustus de novo gene prediction tool<sup>83</sup>. Intron positions and exon positions were converted into hints for Augustus gene prediction.

Finally, we constructed a database of repetitive elements using RepeatModeler (v.1.0.11) and used it for masking repetitive sequences with RepeatMasker (v.4.0.7). Gene models with half or more of their exons showing 50% overlap with repeats were discarded, yielding 46,822 filtered gene models. Alternative transcripts and UTRs were subsequently incorporated using the PASA pipeline<sup>82</sup>. These gene models contain a total number of 4,915 distinct PFAM domains.

### Phylogenomics and molecular dating

To obtain sequences from a previously unsampled hagfish group, we extracted RNA from *M. glutinosa* liver preserved in RNAlater using the RNAeasy kit (Qiagen). The RNA-seq library was constructed using the NEBNext Ultra II Directional RNA Library Prep Kit for Illumina (NEB) and sequenced on a Novaseq6000 instrument (SRR). The transcriptome was assembled using Trinity (v.2.11.0)<sup>81</sup>, enabling read trimming, and was translated using TransDecoder (v.5.5.0)<sup>82</sup>.

We inferred a set of 1,467 single-copy orthologues suitable for phylogenetic reconstruction by applying the OMA tool (v.2.4.1)<sup>84</sup> to a subset of deuterostome proteomes including lamprey and the newly generated hagfish gene models (Supplementary Table 8). Selected transcriptomes were assembled using Trinity (v.2.11.0) and translated using TransDecoder (v.5.5.0)<sup>82</sup>. We built hidden Markov model (HMM) profiles using Hmmer (v.3.1b2) for each orthologue family and extracted orthologues for phylogenetic reconstruction using a previously described approach<sup>85</sup>. Subsequent sequences were aligned using Msaprobs<sup>86</sup>, mistranslated stretches were filtered out using HmmerCleaner<sup>87</sup> and diverging regions intractable for phylogenetic analysis were removed using BMGE (-g 0.9)<sup>88</sup>. Phylogenetic trees were reconstructed for each alignment using IQ-TREE (v.2.1.1) with a LGX+R model<sup>89</sup>. For computational intensive analyses, such as site-heterogenous reconstruction with CAT+GTR, we selected the 20% orthologues with the lowest saturation. Molecular dating analysis was conducted using PhyloBayes (v.4.1e)<sup>90</sup> using the CAT+GTR+G4 model and the CIR relaxed clock (with soft-bound) assuming fossil calibrations<sup>2,91,92</sup> (Supplementary Table 7).

### Synteny reconstruction

Pairs of orthologous genes were obtained by mutual best hit after reciprocal proteome comparison using MMSeqs2 (r12-113e3), and were used to create a system of joint coordinates to plot orthologue position in two species. Fisher's exact test was used to determine mutual enrichment of orthologues between chromosomes, and only significant enrichments were incorporated in binned orthologous content representations (Fig. 1c). Plots connecting orthologues in multiple species (Fig. 1d) were generated using Rideogram (v.0.2.2).

### Gene-family analyses and phylogenetic analysis of paralogs

We reconstructed gene families using Broccoli (ref. 93) for a set of genomes from deuterostome species (Supplementary Table 6). For gene families that included at least 6 genes and 3 species but fewer than 450 sequences in total, we applied GeneRax to infer the losses and duplications that affected a given gene family<sup>94</sup>. To do that, we generated individual alignments using MAFFT (v.7.305)<sup>95</sup>, filtered them using BMGE and reconstructed a tree using IQ-TREE and an LG+R model<sup>89</sup>. These curated alignments and trees were used as input for GeneRax (v.1.2.2) assuming a D+L (duplication plus loss model). Reconciled trees in the RecPhyloXML format were parsed to estimate the duplications and lineage-specific losses at each node of the species tree<sup>96</sup> as seen in Extended Data Fig. 10c. Reconciled trees were split if they showed a duplication at the ‘deuterostomia’ node indicative of a deep paralogy relationship.

For each gene family, we first assigned the CLG by considering the location of amphioxus and sea urchin genes and the corresponding CLG-to-chromosome assignment, and then evaluated the occurrence of the paralogues derived from the 1R and 2R<sub>v</sub> in gnathostomes on the basis of the vertebrate classification that was previously established<sup>19</sup> and has been revised in this study (Supplementary Table 6). Selected species for gene families including derivatives of the 1R paralogons and at least three out of four possible paralogons for gnathostomes ( $\alpha 1, \alpha 2, \beta 1, \beta 2$ ) were collected (Supplementary Table 5). These genes were concatenated for each CLG on the basis of their paralogon identity in gnathostomes, and the chromosomal identity of the CLG derivatives in cyclostomes. Two datasets were generated: a ‘strict’ one, in which at least three distinct gnathostome paralogons were required for each retained gene family; and a ‘relaxed’ one, in which only two or more gnathostome paralogons were required (Supplementary Table 5). A similar approach was used to classify individual genes depending on the duplication events from which they derive. We collected gene ontology terms and functional classification information by applying eggNOG (ref. 97) on the proteome of our interest species and term enrichment analysis conducted using the TopGO package (v.2.50.0).

For analyses of gain and loss, we used gene-family reconstruction that incorporated the gene models of the related hagfish *E. burgeri*<sup>69</sup> to assess recent gene-family expansions or contractions in the hagfish lineage. Gene functions were assigned using the PANTHER classification<sup>98</sup>.

### Tests of WGD hypotheses on the vertebrate phylogeny

We used WHALE (v.2.1.0)<sup>36</sup> to rigorously test WGD hypotheses on a reduced vertebrate species tree (Fig. 2a and Extended Data Fig. 4). We leveraged a total of 8,931 gene families in this analysis, selected to contain at least one gene copy in each clade from the root, in compliance with the assumption of WHALE that genes were acquired in a common ancestor of all included species. We further filtered large families to reduce the computational burden. For each of the 8,931 retained families, we built a multiple sequence alignment based on the amino acid sequences with MAFFT (v.7.508)<sup>95</sup> and reconstructed 1,000 bootstrap trees with IQ-TREE (v.2.2.0.3)<sup>89</sup> under the LG+G model. We summarized clade conditional distribution (CCD) from bootstrapped trees using the ALEobserve tool from the ALE software<sup>99</sup>. We ran WHALE on the dated species trees and CCD data to test five different WGD hypotheses on the vertebrate species tree: 1R<sub>v</sub>, 2R<sub>v</sub>, 2R<sub>cy</sub>, a hagfish-specific duplication and a lamprey-specific duplication (Extended Data Fig. 4a). We used the variable rate DLWGD WHALE model, which models independent duplication and loss rates across branches. We assumed a normal distribution  $N(\log(0.15), 2)$  on the mean log-scaled duplication and loss rate, an exponential distribution (mean = 0.1) prior on its variance, a Beta (3, 1) hyper prior on the  $\eta$  parameter (distribution of the number of genes at the root) and uniform priors on the retention parameters (q parameter) for all WGDs. We obtained significant Bayes factors ( $BF_{\text{Null vs. WGD}} < 10^{-3}$ ) in support of large-scale duplication (post-duplication retention parameter  $q \neq 0$ ) for the 1R<sub>v</sub>, 2R<sub>v</sub> and 2R<sub>cy</sub> events (Extended Data Fig. 4b). These results were reproduced using the simpler constant rate DLWGD model. We similarly tested an alternative scenario with two duplications on the vertebrate stem (1R<sub>v</sub> and 2R<sub>v</sub>; Extended Data Fig. 4c,d). In this configuration, and using uniform priors on the retention parameters, we observed that WHALE could not distinguish retention parameters for 1R<sub>v</sub> and 2R<sub>v</sub>: this is revealed by the bimodality of the estimated posterior distribution for each of these two parameters. We found that using distinct priors on retention parameters allows the estimation of distinct retention parameters for 1R<sub>v</sub> and 2R<sub>v</sub> and shows support for a single 1R<sub>v</sub> event (Extended Data Fig. 4e). This investigation of alternative priors was conducted on a pilot run of 1,000 randomly selected gene families, to alleviate computational time (1,000 families were previously suggested to be sufficient for parameter estimation<sup>36</sup>).

### Ancestral and lineage-specific meiotic rediploidization

We selected a set of 1,247 gene families, including genes of 6 vertebrate species (bamboo shark *Chiloscyllium plagiosum*, spotted gar *Lepisosteus oculatus*, chicken *Gallus gallus*, western clawed frog *Xenopus tropicalis*, brown hagfish *E. atami* and sea lamprey *Petromyzon marinus*) and the closest outgroup (depending on taxonomic availability), to test for ancestral and lineage-specific rediploidization after the 1R genome duplication. These 1,247 families were selected so as to result in distinct tree topologies under the ancestral and lineage-specific rediploidization models, on the basis of the following criteria: (i) at least one gnathostome species has retained both 1R\_1 (that is, alpha1 and/or beta1) and 1R\_2 (that is, alpha2 and/or beta2) gene copies; (ii) at least one hagfish gene and one lamprey gene; (iii) at least one non-vertebrate outgroup gene; and (iv) a non-prohibitive number of hagfish and lamprey genes so that a maximum of 10 possible ancestral rediploidization topologies can be derived for the family (Extended Data Fig. 6). For each gene family, we designed constrained tree topologies as expected under the lineage-specific and ancestral rediploidization models (Fig. 3a). More specifically, for the constrained ancestral rediploidization topologies, we built constrained topologies as follows: we first placed 1R\_1 and 1R\_2 gnathostome gene copies in two different clades following 1R and then derived possible combinations of hagfish and lamprey genes to be placed on the 1R\_1 and 1R\_2 clades, using well-supported hagfish and lamprey chromosomal orthologies to limit the number of combinations (Extended Data Fig. 6 and Supplementary Table 5). Next, for each of these 1,247 families, we built gene trees using RAxML (v.8.2.12)<sup>100</sup>, with 10 distinct starting trees and the PROTGAMMAJTT model, for: the unconstrained maximum likelihood (ML) tree; the constrained ancestral rediploidization topologies; and the constrained lineage-specific rediploidization topology. We then used the AU-test implemented in CONSEL (ref. 101) to test for significant differences in log-likelihoods reported by RAxML (ref. 100). A tree topology was rejected when significantly less likely than the ML tree at  $\alpha = 0.05$ .

We used the same approach to test for lineage-specific rediploidization in lampreys and hagfish on CLGB-1R after the 2R cyclostome hexaploidization. We ran likelihood tests on 30 informative gene families, constraining the lineage-specific rediploidization gene-tree topology as presented in Fig. 3c and constraining the ancestral rediploidization topologies according to the two possible ways of grouping hagfish and lamprey genes together (that is, either grouping genes from hagfish chr. 4 with lamprey chr. 10 and hagfish chr. 8 with lamprey chr. 2, or hagfish chr. 8 with lamprey chr. 10 and hagfish chr. 4 with lamprey chr. 8).

The code to reproduce the analysis, as well as the associated resulting gene trees, have been deposited in GitHub (<https://github.com/fmarletaz/hagfish/tree/main/rediploidization>).

### Phylogenetic tree based on concatenation of Hox clusters

We investigated the phylogenetic relationships between Hox clusters and bystander genes in seven genomes: amphioxus, sea lamprey, hagfish, human, mouse, chicken and spotted gar. We identified Hox and bystander genes in three steps: (i) starting from human gene names, we searched for orthologues in the other species using our set of reconciled gene trees (GeneRax trees); (ii) we used NCBI blastp (ref. 102) to confirm identified hox genes and further search for Hox genes missed by the gene-trees approach; and (iii) we used miniprot (v.0.5-r179)<sup>103</sup> with the sets of human and *E. burgeri* Hox proteins<sup>24</sup> to search for Hox genes missing from genome annotations of other species. We next aligned each gene family using their amino acid sequence with MAFFT (v.7.508)<sup>95</sup> and concatenated alignment from each cluster. Finally, we used the concatenation matrix to build a phylogenetic tree with RAxML-NG (v.1.1)<sup>100</sup> using the LG+G4+F model, 10 different starting parsimony trees and 100 bootstrap replicates.

## Comparative transcriptomics

RNA-seq reads for hagfish (this study), lamprey *Lampetra japonica* (PRJNA354821, PRJNA349779 and PRJNA312435), gar *Lepisosteus oculatus* (PRJNA255881) and the cephalochordate amphioxus (PRJNA416977) were aligned with STAR (v.2.5.2b)<sup>79</sup>, and counts for annotated genes were obtained using featureCount from the subreads package (v.1.6.3)<sup>104</sup>. Counts were converted to FPKM in the R package for subsequent analyses: WGCNA (v.1.7.0) was used to cluster gene expression in the full organ set and, after filtering out genes with limited variance and coverage, the 'softpower' parameter was estimated to be 20, and clustering was run with a 'signed' network type<sup>105</sup>. The gene-expression specificity index (or  $\tau$ ) was calculated as described previously<sup>106</sup> on sets of organs (brain or neural tube; gills; heart; intestine; kidney; liver or hepatic tissues; ovary or female gonad; skin or epidermis; and muscle). For comparative analyses, gene families with paralogues derived from the vertebrate WGD were selected on the basis of their duplication history, and the gene-expression specificity index was compared across species for the same gene families (Extended Data Fig. 10b). We also compared gain and losses of expression domains for a given gene family by binarizing gene expression across a reduced set of six organs (brain, gills, intestine, liver, muscle and ovary) and counting expression patterns gains or losses between genes belonging to a given gene including paralogues and outgroup. The number of gain and loss events is then plotted as a distribution centred around zero (Fig. 4c).

The expression of paralogous genes in lamprey neural crest was assessed by quantifying gene expression using Salmon (v.1.10.0)<sup>107</sup> from RNA-seq data generated in a previous study<sup>56</sup> on dissected cranial and trunk dissected tissues using the latest lamprey genome and annotation<sup>7</sup>. Paralogy status and expression is specified in Supplementary Table 10.

## Detection of germline-enriched and germline-specific regions

DNA was extracted from testes and blood by phenol-chloroform extraction<sup>72</sup>. To enrich for germ cells, testes tissue was ground gently with a plastic pestle in a 1.5-ml microfuge tube and residual connective tissues were discarded before proteinase K digestion. Outsourced library prep and Illumina sequencing (HiSeq2500 V4, 150-bp paired-end reads) were performed by Hudson Alpha Genome Services Laboratory.

Sequence data were aligned to the *E. atami* genome assembly using BWA-mem (v.0.7.5a-r416)<sup>108</sup> with option -a and filtered by samtools view<sup>108</sup> with option -F2308. Only primary alignments with mapping quality 5 and higher were retained for further analysis. The resulting files were processed using DifCover (v.3.0.1)<sup>71,34</sup> to calculate the degree of germline enrichment across all discontinuous 500-base intervals of low-copy sequence using modal coverages for testes and blood of 32 $\times$  and 54 $\times$  respectively, low-coverage masking of regions with a read depth of less than 1/3 $\times$  in both samples and high-coverage masking of sequences with a read depth greater than 3 $\times$  modal coverage in both samples. To identify germline-specific genes that are present at a higher copy number, we ran DifCover using low-coverage masking with a read depth of less than 10 $\times$  in both samples and high-coverage masking of sequences with a read depth greater than 30 $\times$  modal coverage.

## PCR validation of germline-enriched loci

Primers were designed using a coverage-masked version of the *E. atami* genome using Primer3 (version 4.1.0)<sup>109</sup>. Amplification of PCR validation reactions was performed using GoTaq DNA polymerase (Promega, 1.2 units per 50  $\mu$ l reaction), Colorless GoTaq Reaction Buffer, 1  $\mu$ g genomic DNA template and 100 ng oligonucleotide primer. PCR cycling conditions included a 3-min initial denaturation step at 95  $^{\circ}$ C, 34 cycles of a three-step thermal cycling consisting of a 30-s denaturation at 95  $^{\circ}$ C, a 30-s primer annealing step at 55–65  $^{\circ}$ C (Supplementary Table 13) and a 30-s extension step at 72  $^{\circ}$ C. A final extension at 72  $^{\circ}$ C was performed on all reactions to ensure that full-length amplicons were produced.

Amplification was assessed by agarose gel electrophoresis. Eight primer pairs with an ambiguous signal in the first round of PCR were redesigned and retested (Supplementary Table 13). We note that some PCR markers might not be fully diagnostic with respect to germline specificity, as somatic gene duplicates are continuously captured by the germline-specific chromosomes in both lamprey<sup>71,34</sup> and songbird lineages<sup>34</sup>.

## Computational prediction of germline-enriched and highly abundant somatic repeats

Abundant  $k$ -mers ( $k = 31$ ) were identified from testes and blood DNA-sequencing data using Jellyfish (v.2.2.4)<sup>110</sup>. Minimal copy-number thresholds for defining abundant  $k$ -mers were set at 3 $\times$  the modal copy number: 72 for testes and 120 for blood. Abundant  $k$ -mers were extracted and assembled into a set of de-novo-assembled repetitive sequences using Velvet (v.1.2.10)<sup>111</sup> with a hash length of 29. These sequences were aligned (blastn with -word\_size 17) to repetitive elements generated from the *E. atami* genome assembly by RepeatModeler<sup>112</sup> and sequences that aligned with less than 90% identity or less than 80% of their length were added to the set of reference-derived repeats to form a union set.

Enrichment analysis was performed by separately aligning paired-end reads from testes and blood to the union set of repeats. Primary alignments, identified by samtools view<sup>113</sup> with option -F2308, were also filtered to retain only alignments that either cover more than 80% of a repeat or have more than 80% of read bases aligned. Enrichment scores were calculated with DifCover pipeline (v.3)<sup>11</sup>. Stage 2 of the pipeline was run with parameters  $v=10000$ ,  $l=0$ ,  $a=b=10$ ,  $A=B=10^8$ . Stage 3 of the pipeline was modified by using a subroutine from DNACopy<sup>114</sup> without 'smoothing' the data before analysis. From a set of 180,032 intervals generated by DifCover, we chose 138 highly abundant and germline-specific sequences with enrichment scores of more than 10 and an estimated span size of more than 100 kb. The estimated genomic span of these repeats was computed as [(testes coverage/modal testes coverage)  $\times$  (number of bases with read depth coverage > 10)], in which modal testes coverage = 32.

Clustering of 138 highly abundant and germline-specific sequences was performed using CD-HIT-EST (v.4.6, with parameters: -c0.8, -G0, -aS 0.3, -aL 0.3, -sc1, -g1, -b4)<sup>115</sup>, resulting in the formation of 38 clusters that were further merged to 24 by manual curation and cross-alignment of sequences from the initial clusters. For characterization of repetitive structures and identification of motifs, representative sequences from each cluster were mapped to the assembly (blastn, -word\_size 15) and to a collection of published hagfish repeats. We found that 4 of 24 clusters have sequences that are homologous to the published repeats of *Paramyxine sheni* EEPs2, EEPs3 and EEPs4 and *Eptatretus okinoseanus* EEEO2 (refs. 30,33). Primers for these and representatives of 7 other clusters were designed with the Primer3 (v.0.4.0) tool (Supplementary Table 16).

To facilitate FISH visualization, we also searched for possible candidates for centromeric repeats. Such candidates are expected to be (1) highly abundant in both somatic and germline sequence and (2) enriched in a 'centromeric' region of every chromosome. From the union set, we chose repeats with blood coverage > 10<sup>5</sup> or span > 1 Mb and aligned them to the assembly (blastn -word\_size 15, p>75, coverage > 80%). Repeats with more than 200 hits in a 1-Mb window were grouped to three families labelled Soma1–Soma3. Soma2 seemed to be homologous to the *P. sheni* repeat EEPs1 (ref. 33) and Soma1 and Soma3 to the *E. burgeri* contigs LC047612.1 and LC047003.1. FISH analysis confirmed the in silico prediction that EEPs1 is highly abundant in both testes and blood DNA of *E. atami*.

To estimate more accurately the genomic span of chosen germline-enriched and somatic repeats, we realigned reads from blood and testes to the sequences of these repeats or to the sequence extended as a tandem repetition of repeat motifs spanning at least 150 bp

# Article

(Supplementary Table 12), and applied all described previously steps for filtering and coverage and span estimation.

## In situ hybridization

**Slide preparation.** Snap-frozen samples of blood and testes were used for slide preparation of somatic and germline cells for validation of the presence and specificity of repeats in different cell types. A small amount of blood (about 20 mg) was gently thawed on ice, mixed with 2 ml buffered hypotonic solution (0.4% KCl, 0.01 M HEPES, pH 6.8) and incubated for 30 min at room temperature. The cells were prefixed by gently mixing the suspension with several drops of fixative solution (methanol:acetic acid 3:1). After centrifugation (5,000g for 10 min), the supernatant was removed and cells were resuspended and fixed with methanol:acetic acid 3:1. Three further fixative solution changes were performed to ensure that cells were fully equilibrated to fixative solution. Fixed cells were stored at  $-20^{\circ}\text{C}$ . One fixative change was made before spreading the cell suspension onto slides. A drop of about 20  $\mu\text{l}$  was applied to a steamed slide, which was immediately placed on a heating block in a humidity chamber at  $60^{\circ}\text{C}$  for 1–2 min. After air drying, slides were examined with a microscope using a low condenser position to aid in viewing unstained nuclei and metaphases. Slides were aged for 1–3 days on a warming stage at  $37^{\circ}\text{C}$  before hybridization. For germline cells, a piece of testis (30–40 ng) was minced with a razor blade, placed in a homogenizer and disaggregated in hypotonic solution. Testis cell suspensions were filtered through a 40–50- $\mu\text{m}$  cell strainer to remove excess tissue. Subsequent steps of fixation and slide preparation for testis tissue were as described for blood.

**Probe labelling.** Probes for FISH were generated using a modified conventional PCR: the reaction mix with a final volume of 25  $\mu\text{l}$  contained 0.1 mM each of unlabelled dATP, dCTP and dGTP and 0.03 mM of dTTP; 0.5  $\mu\text{l}$  one fluorophore conjugated dUTP (cyanine 3-dUTP (Enzo), cyanine 5-dUTP (Enzo) or fluorescein-12-dUTP (Thermo Fisher Scientific));  $1\times$  Taq-buffer; and 0.625 U GoTaq DNA Polymerase (Promega). Each PCR amplification was performed using 0.5  $\mu\text{g}$  of genomic DNA template from testes, 34 PCR cycles and a 30-s extension step to obtain appropriately sized probes for FISH. After cycling the reaction, 25  $\mu\text{l}$  PCR mix was combined with 5  $\mu\text{l}$  sheared salmon sperm DNA ( $1\text{ mg ml}^{-1}$ ; Thermo Fisher Scientific), 3  $\mu\text{l}$  3 M sodium acetate, pH 5.2 and 80  $\mu\text{l}$  100% cold ethanol, and kept overnight at  $-20^{\circ}\text{C}$  for probe precipitation. After spinning and supernatant removal, the pellet was dissolved in 25–30  $\mu\text{l}$  of 50% formamide and stored at  $-20^{\circ}\text{C}$  before use.

**FISH.** FISH on chromosome preparations was performed according to a standard protocol for chromosome spreads<sup>116</sup> with modifications<sup>117</sup>. Before hybridization, slides were incubated in  $2\times$  SSC for 30 min at  $37^{\circ}\text{C}$ , passed through an ethanol series (70%, 80%, 100%), dried and denatured in formamide (70% in  $2\times$  SSC) for 2 min, prewarmed to  $70^{\circ}\text{C}$ . After the formamide denaturation, slides were placed immediately in cold ( $-20^{\circ}\text{C}$ ) 70% ethanol, further dehydrated in 80% and 100% ethanol, and kept on a slide warmer at  $37^{\circ}\text{C}$  until the hybridization mix with probe was applied.

Differently labelled hybridization probes were mixed (1  $\mu\text{l}$  of each per slide) with hybridization master mix (60% formamide, 10% dextran sulfate and  $1.2\times$  SSC) to a final volume of 10  $\mu\text{l}$ . The hybridization mix was denatured at  $95^{\circ}\text{C}$  for 7 min, cooled in ice, prewarmed to  $37^{\circ}\text{C}$ , applied to the slide, coverslipped and sealed with rubber cement. After overnight incubation in a humidity chamber at  $37^{\circ}\text{C}$ , slides were washed in  $0.4\times$  SSC and 0.3% NP-40 for 3 min at  $70^{\circ}\text{C}$  and in  $2\times$  SSC, 0.1% NP-40 for 5 min at room temperature. One drop of ProLong Glass Antifade Mountant with NucBlue Stain was placed in the centre of an area to be examined and covered with a coverslip.

**Microscopy and image analysis.** Slides were analysed with an Olympus-BX63 microscope using filter sets for DAPI, FITC, Cy3 and

Cy5. Images were captured using CellSens software (Olympus) and processed with Adobe Photoshop CC 2019 and ImageJ 1.53k (NIH).

## Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

## Data availability

Raw and processed sequences have been deposited in the NCBI Sequence Read Archive (SRA) (PRJNA953751) and Gene Expression Omnibus (GEO) (GSE230176). The RNA-seq data for *M. glutinosa* are available on the SRA (SRR25213276). The resequenced somatic tissues are also available on the SRA (blood, SRR24133795; testes, SRR24130678). RNA-seq datasets used for comparative analyses are publicly available for Japanese lamprey (PRJNA354821, PRJNA349779 and PRJNA312435), gar (PRJNA255881), amphioxus (PRJNA416977) and sea lamprey (PRJNA497902). The read data are available at PRJNA953751. The genome and its annotation are also deposited in zenodo: <https://zenodo.org/records/10227719>. Source data are provided with this paper.

## Code availability

The code used is available at <https://github.com/fmarletaz/hagfish>.

- Green, M. R. & Sambrook, J. *Molecular Cloning: A Laboratory Manual* 4th edn (Cold Spring Harbor Laboratory Press, 2012).
- Chapman, J. A. et al. Meraculous: de novo genome assembly with short paired-end reads. *PLoS ONE* **6**, e23501 (2011).
- English, A. C. et al. Mind the gap: upgrading genomes with Pacific Biosciences RS long-read sequencing technology. *PLoS ONE* **7**, e47768 (2012).
- Putnam, N. H. et al. Chromosome-scale shotgun assembly using an in vitro method for long-range linkage. *Genome Res.* **26**, 342–350 (2016).
- Meyer, M. & Kircher, M. Illumina sequencing library preparation for highly multiplexed target capture and sequencing. *Cold Spring Harb. Protoc.* **2010**, pdb.prot5448 (2010).
- Miller, J. R. et al. Aggressive assembly of pyrosequencing reads with mates. *Bioinformatics* **24**, 2818–2824 (2008).
- Vurture, G. W. et al. GenomeScope: fast reference-free genome profiling from short reads. *Bioinformatics* **33**, 2202–2204 (2017).
- Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
- Niknafs, Y. S., Pandian, B., Iyer, H. K., Chinnaiyan, A. M. & Iyer, M. K. TACO produces robust multisample transcriptome assemblies from RNA-seq. *Nat. Methods* **14**, 68–70 (2017).
- Grabherr, M. G. et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* **29**, 644–652 (2011).
- Haas, B. J. et al. Automated eukaryotic gene structure annotation using EvidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biol.* **9**, R7 (2008).
- Stanke, M. et al. AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Res.* **34**, W435–W439 (2006).
- Altenhoff, A. M. et al. OMA standalone: orthology inference among public and custom genomes and transcriptomes. *Genome Res.* **29**, 1152–1163 (2019).
- Marlétaz, F., Peijnenburg, K. T. C. A., Goto, T., Satoh, N. & Rokhsar, D. S. A new spiralian phylogeny places the enigmatic arrow worms among gnathiferans. *Curr. Biol.* **29**, 312–318 (2019).
- Liu, Y., Schmidt, B. & Maskell, D. L. MSAProbs: multiple sequence alignment based on pair hidden Markov models and partition function posterior probabilities. *Bioinformatics* **26**, 1958–1964 (2010).
- Di Franco, A., Poujol, R., Baurain, D. & Philippe, H. Evaluating the usefulness of alignment filtering methods to reduce the impact of errors on evolutionary inferences. *BMC Evol. Biol.* **19**, 21 (2019).
- Criscuolo, A. & Gribaldo, S. BMGE (block mapping and gathering with entropy): a new software for selection of phylogenetic informative regions from multiple sequence alignments. *BMC Evol. Biol.* **10**, 210 (2010).
- Minh, B. Q. et al. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. *Mol. Biol. Evol.* **37**, 1530–1534 (2020).
- Rodrigue, N. & Lartillot, N. Site-heterogeneous mutation-selection models within the PhyloBayes-MPI package. *Bioinformatics* **30**, 1020–1021 (2014).
- Irisarri, I. et al. Phylotranscriptomic consolidation of the jawed vertebrate timetree. *Nat. Ecol. Evol.* **1**, 1370–1378 (2017).
- Kuraku, S., Ota, K. G. & Kuratani, S. In *The Timetree of Life* (eds Blair Hedges, S. & Kumar, S.) 317–319 (Oxford Univ. Press, 2009).
- Derelle, R., Philippe, H. & Colbourne, J. K. Broccoli: combining phylogenetic and network analyses for orthology assignment. *Mol. Biol. Evol.* **37**, 3389–3396 (2020).
- Morel, B., Kozlov, A. M., Stamatakis, A. & Szöllösi, G. J. GeneRax: a tool for species-tree-aware maximum likelihood-based gene family tree inference under gene duplication, transfer, and loss. *Mol. Biol. Evol.* **37**, 2763–2774 (2020).
- Kato, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).

96. Duchemin, W. et al. RecPhyloXML: a format for reconciled gene trees. *Bioinformatics* **34**, 3646–3652 (2018).
97. Huerta-Cepas, J. et al. eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res.* **47**, D309–D314 (2019).
98. Thomas, P. D. et al. PANTHER: making genome-scale phylogenetics accessible to all. *Protein Sci.* **31**, 8–22 (2022).
99. Szöllösi, G. J., Rosikiewicz, W., Boussau, B., Tannier, E. & Daubin, V. Efficient exploration of the space of reconciled gene trees. *Syst. Biol.* **62**, 901–912 (2013).
100. Kozlov, A. M., Darriba, D., Flouri, T., Morel, B. & Stamatakis, A. RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. *Bioinformatics* **35**, 4453–4455 (2019).
101. Shimodaira, H. & Hasegawa, M. CONSEL: for assessing the confidence of phylogenetic tree selection. *Bioinformatics* **17**, 1246–1247 (2001).
102. Johnson, M. et al. NCBI BLAST: a better web interface. *Nucleic Acids Res.* **36**, W5–W9 (2008).
103. Li, H. Protein-to-genome alignment with miniprot. *Bioinformatics* **39**, btad014 (2023).
104. Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923–930 (2014).
105. Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinf.* **9**, 559 (2008).
106. Yanai, I. et al. Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. *Bioinformatics* **21**, 650–659 (2005).
107. Patro, R., Duggal, G., Love, M. I., Irizarry, R. A. & Kingsford, C. Salmon provides fast and bias-aware quantification of transcript expression. *Nat. Methods* **14**, 417–419 (2017).
108. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
109. Koressaar, T. & Remm, M. Enhancements and modifications of primer design program Primer3. *Bioinformatics* **23**, 1289–1291 (2007).
110. Marçais, G. & Kingsford, C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* **27**, 764–770 (2011).
111. Zerbino, D. R. & Birney, E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res.* **18**, 821–829 (2008).
112. Flynn, J. M. et al. RepeatModeler2 for automated genomic discovery of transposable element families. *Proc. Natl Acad. Sci. USA* **117**, 9451–9457 (2020).
113. Li, H. et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
114. Seshan, V. E. & Olshen, A. DNACopy: DNA copy number data analysis. R package version 1.76.0 <https://bioconductor.org/packages/DNACopy> (2023).
115. Li, W. & Godzik, A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**, 1658–1659 (2006).
116. Rooney, D. E. *Human Cytogenetics: Constitutional Analysis* (Oxford Univ. Press, 2001).
117. Timoshevskiy, V. A., Sharma, A., Sharakhov, I. V. & Sharakhova, M. V. Fluorescent in situ hybridization on mitotic chromosomes of mosquitoes. *J. Vis. Exp.* **67**, e4215 (2012).

**Acknowledgements** We thank B. Venkatesh and S. Kuraku for early discussions; N. Segi, T. Suzuki, B. Muramatsu and H. Dohra for technical support; H. Hasegawa and K. Hasegawa for hagfish supply; and M. Levine, M. Martik, H. van Mullem, C. Amemiya and L. Piovani for comments on the manuscript. Work at the OIST Molecular Genetics Unit (D.S.R., O.S., D.G. and F.M.) was supported by OIST internal funds. F.M. is supported by the Royal Society Fellowship URF\R1\191161 and the BBSRC grant BB/V01109X/1. D.S.R. is a Chan Zuckerberg Biohub Investigator and is supported by the Marthella Foskett Brown Family Chair of Biological Sciences at the University of California, Berkeley. J.J.S. is supported by grants from the National Institutes of Health (NIH) (R35GM130349) and the National Science Foundation (NSF) (MCB1818012). M.S. was in part supported by the Field Science Center and Research Institute of Green Science and Technology, Shizuoka University. E.P. is supported by a Newton International Fellowship from the Royal Society (NIF\R1\222125). We thank the OIST Sequencing Section for DNA and RNA sequencing and acknowledge the support of OIST supercomputing and the University of Kentucky High-Performance Computing complex.

**Author contributions** D.S.R., O.S., F.M., J.J.S. and S.B. conceived the study, which was led by F.M., J.J.S. and D.S.R. F.M. and O.S. sequenced, assembled and annotated the genome. N.T., V.A.T. and J.J.S. performed the DNA elimination analysis. F.M., O.S. and D.S.R. performed synteny analyses. F.M. completed phylogenetic and paralogon analysis. E.P. contributed duplication phylogenetic modelling, lineage-specific rediploidization and Hox cluster analysis. D.G. analysed transcriptomic data. M.S. provided biological samples and figure material. K.K. provided biological material and took part in transcriptomic analyses. D.S.R., F.M., E.P. and J.J.S. wrote the paper with input from N.T. and V.A.T. All authors read and approved the manuscript.

**Competing interests** The authors declare no competing interests.

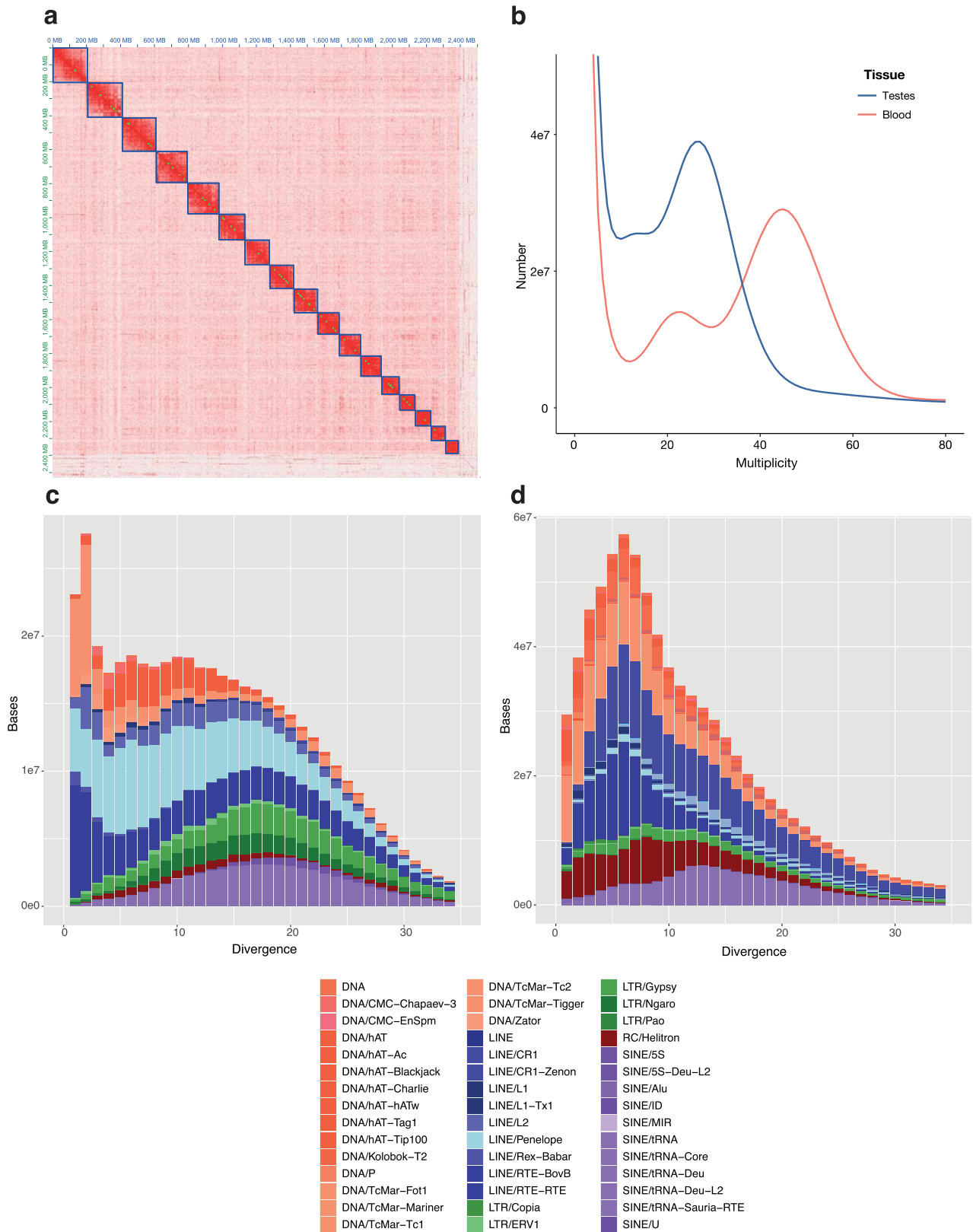
#### Additional information

**Supplementary information** The online version contains supplementary material available at <https://doi.org/10.1038/s41586-024-07070-3>.

**Correspondence and requests for materials** should be addressed to Ferdinand Marlétaz, Jeremiah J. Smith or Daniel S. Rokhsar.

**Peer review information** Nature thanks Sebastian Shimeld and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

**Reprints and permissions information** is available at <http://www.nature.com/reprints>.

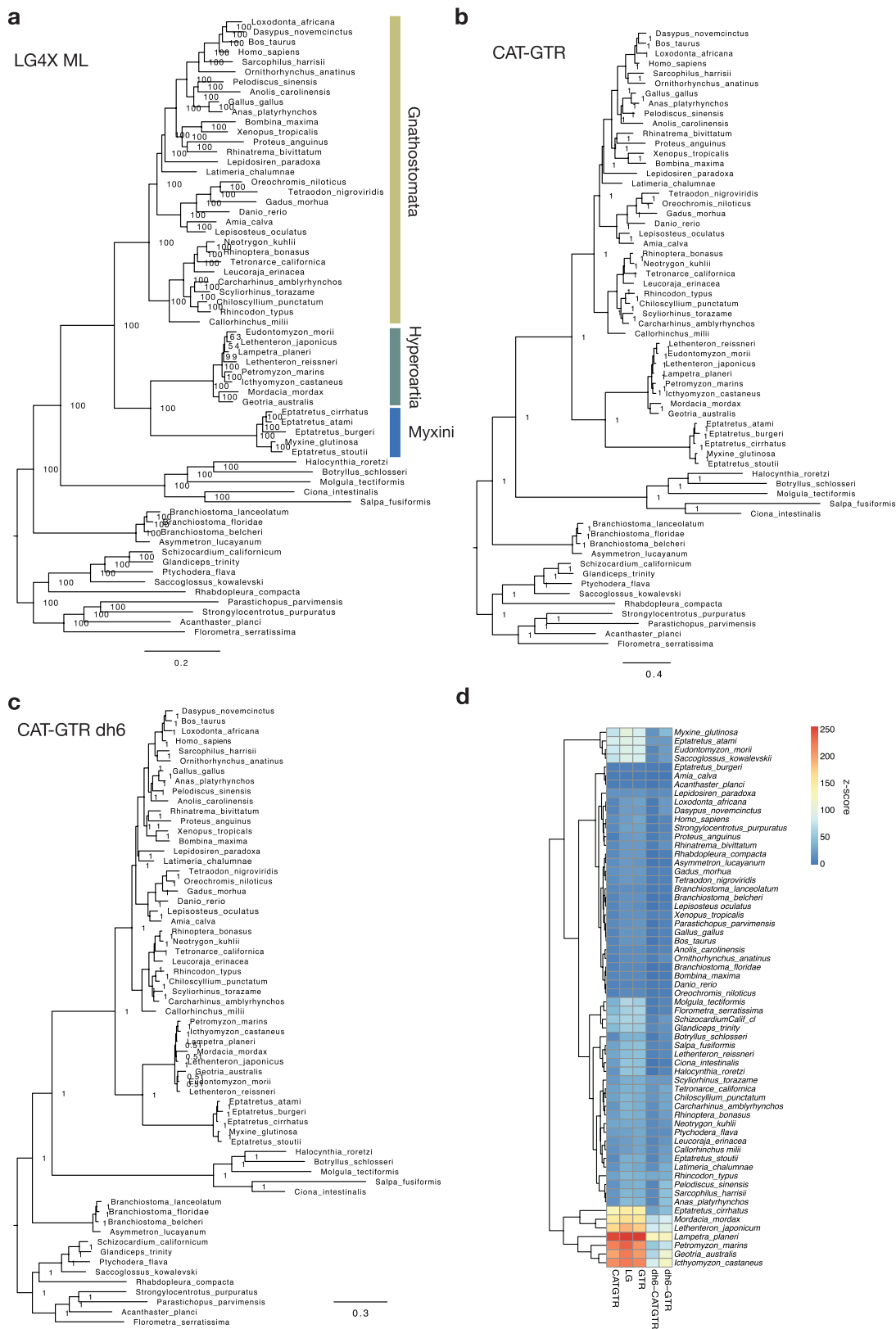


**Extended Data Fig. 1 | Genome content and architecture of *E. atami*.**

**a**, Hi-C contact map visualizing the density of interactions between binned genomic regions in the proximity ligation data. The high contact regions are consistent with the 17 somatic chromosomes. **b**, Density of 21-mer of increasing multiplicity in the somatic (blood) and germline (testes) shotgun sequence

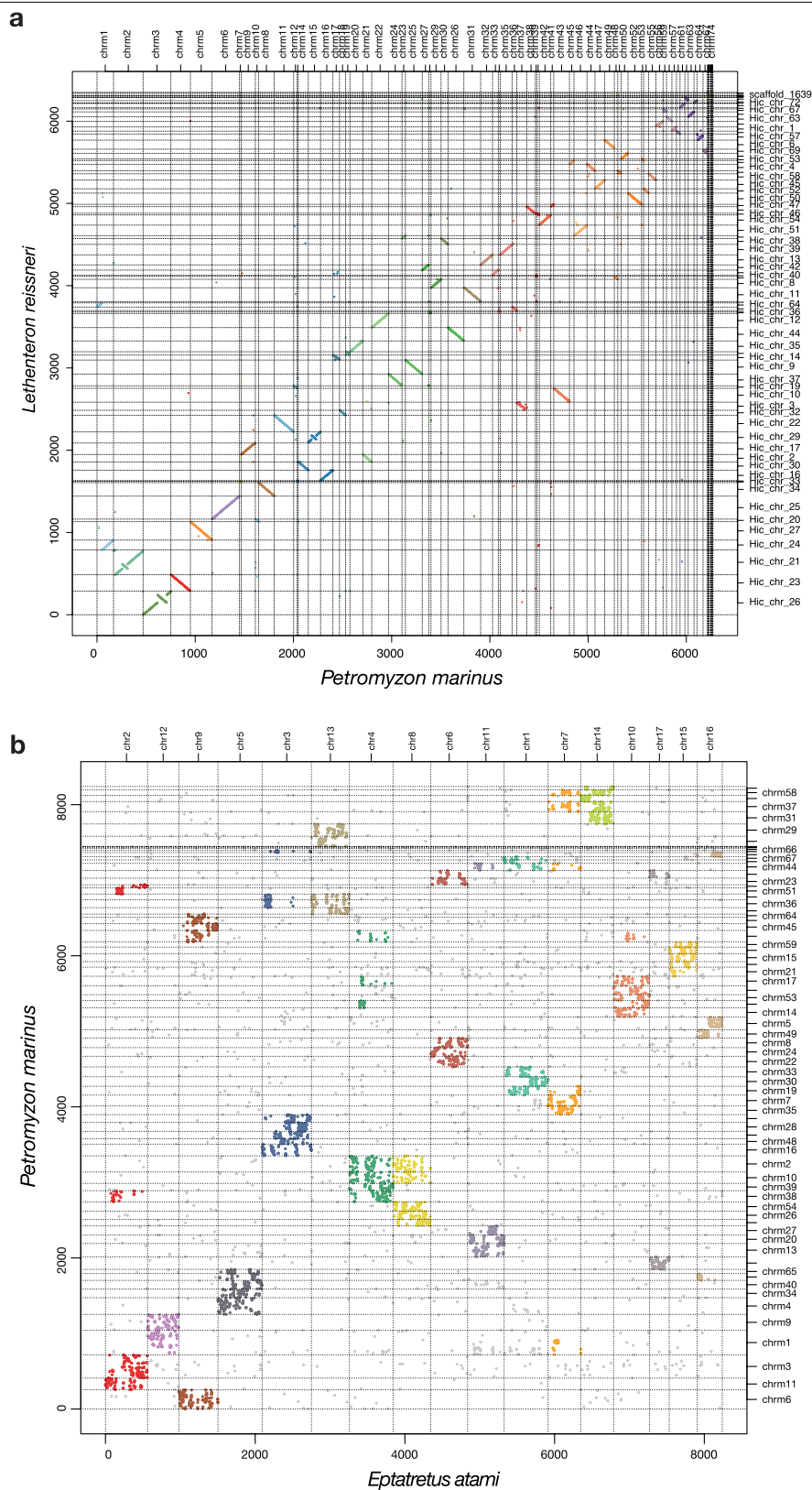
data indicated an estimated genome sizes of 2.02 and 3.28 Gb, respectively. **c,d**, Repeat landscape summarizing the fraction of regions diverging from consensus repeats at varying levels of divergence (Kimura 2-parameter distance) in lampreys (**c**) and hagfish (**d**). Lamprey and hagfish show a markedly different profile with respect to the number and diversity of repetitive element classes.





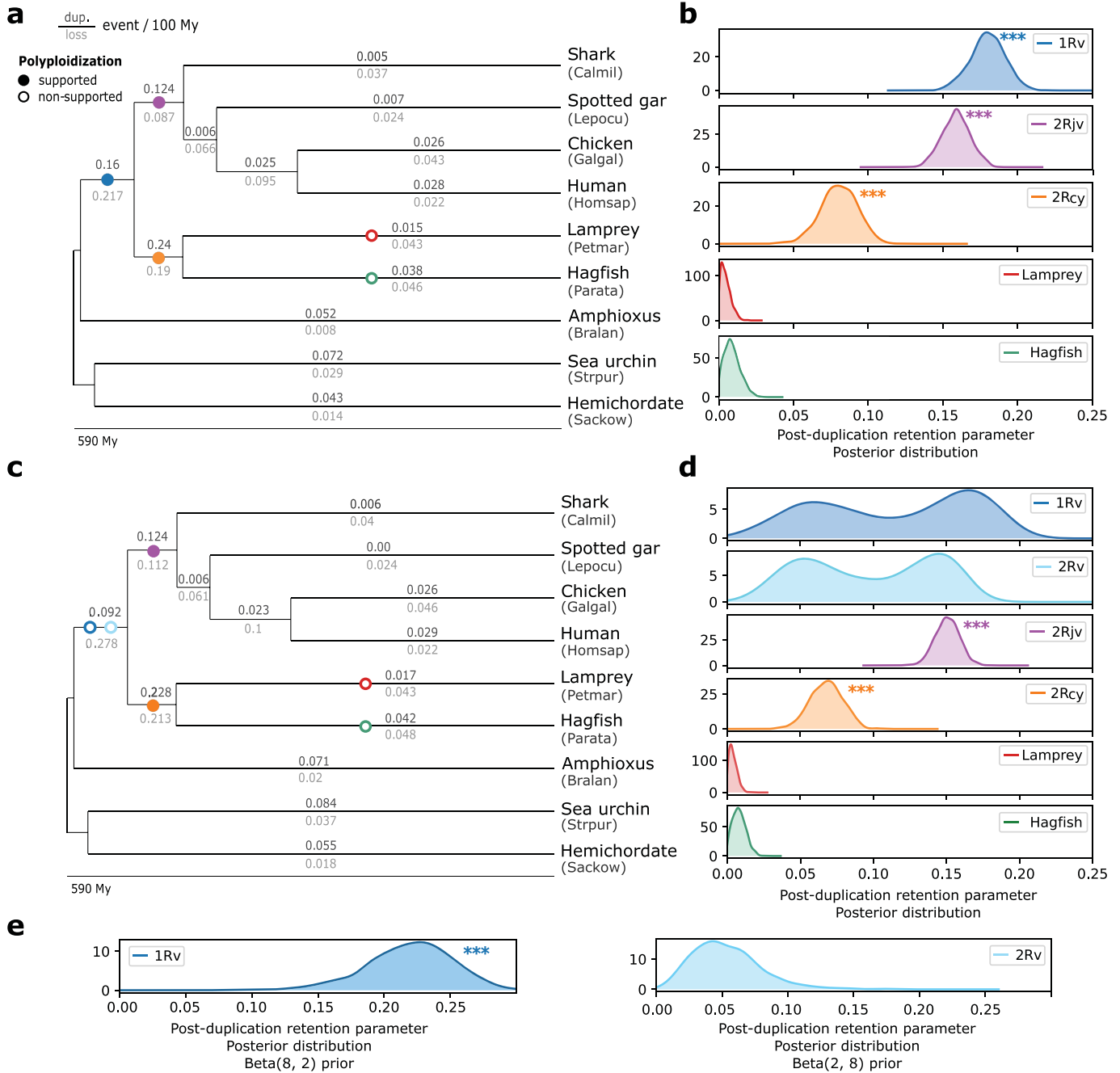
**Extended Data Fig. 2 | Phylogenetic reconstruction of deuterostome relationships with a focus on cyclostome position.** **a**, Tree reconstructed with IQ-TREE assuming LG4X model using a dataset of 1,467 single-copy orthologues and a partitioned model. **b**, Tree reconstructed using PhyloBayes and a CAT+GTR+G4 model using a subset of 176 orthologues showing the lowest saturation (see methods). **c**, Tree reconstructed using the same set of orthologues after Dayhoff 6 categories amino acid recoding to account for

possible compositional heterogeneity due to high GC% in cyclostome genomes. **d**, z-score of posterior predictive analyses to assess composition heterogeneity. Positive z-scores indicate that average amino acid diversity is underestimated (negative z-scores indicate an overestimation) which highlights the composition bias existing in some lamprey and hagfish species and shows that recoding (Dayhoff 6) alleviates these biases.



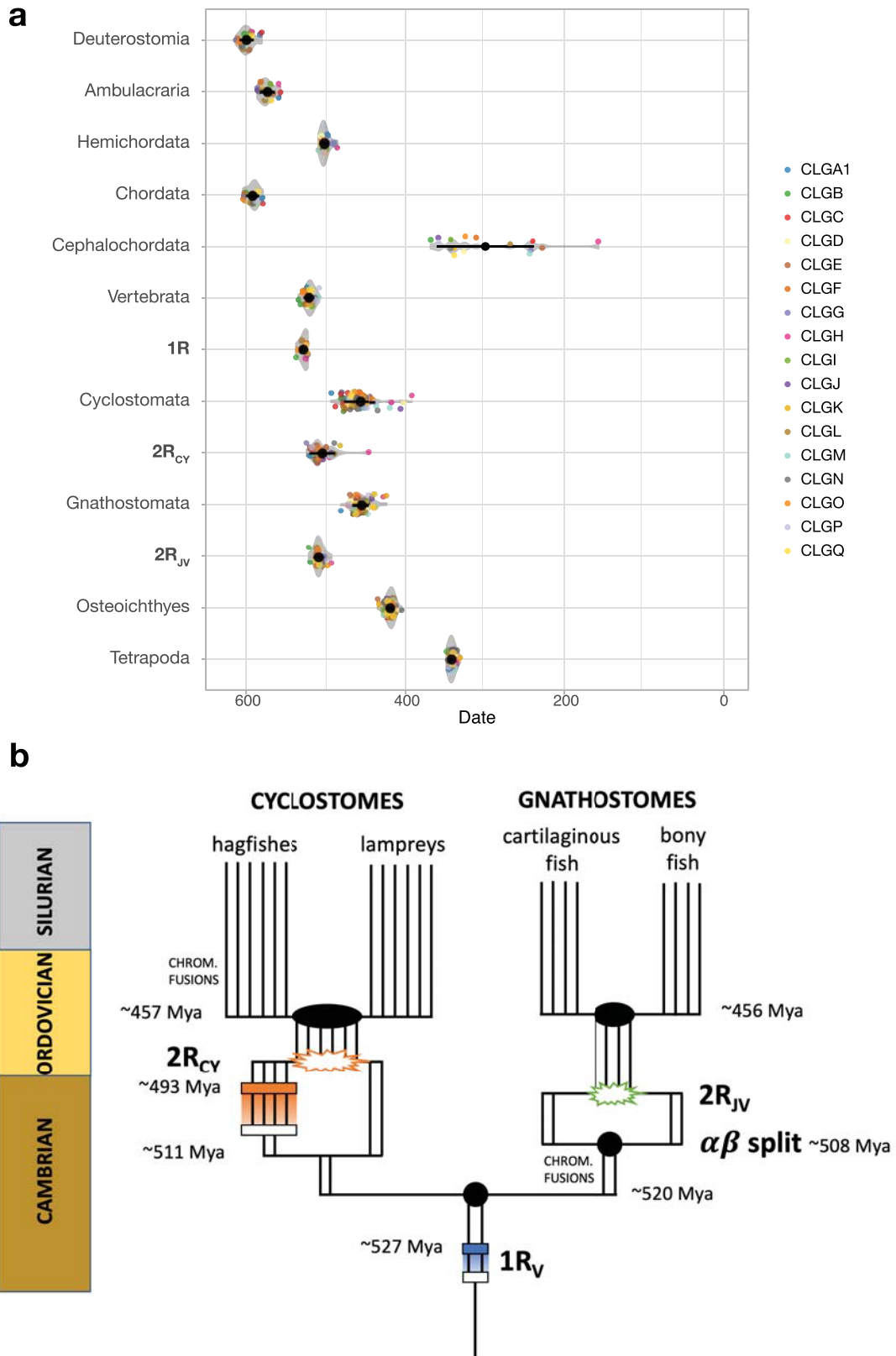
**Extended Data Fig. 3 | Comparison of the chromosomal architectures of cyclostome genomes. a.** Comparison between two lampreys (*Lethenteron reissneri* and *P. marinus*) highlighting the conservation of both chromosomal identity and extensive collinear segments. **b.** Comparison between the hagfish *E. atami* and the lamprey *P. marinus*. In both panels, dots show the relative

location of orthologous genes between two species, coloured if the chromosome: chromosome enrichment is significant by Fisher's exact test (Methods); others shown in grey. The colours in **a** and **b** are based on *P. marinus* and *E. atami* chromosomes, respectively. In **b**, *P. marinus* chromosomes are sorted to aid in visualizing many-to-one mappings shown in Fig. 1d.



**Extended Data Fig. 4 | Tests of genome duplication hypotheses on the vertebrate tree.** **a**, Species phylogeny and polyploidization hypotheses tested with WHALE<sup>36</sup> using 8,931 gene families (Methods, see Supplementary Table 8 for details of the genomes used in the analysis). Polyploidization hypotheses are indicated by circles on the corresponding branches, with supported polyploidizations indicated with solid circles. Inferred background gene duplication and loss rates are presented on the branches. **b**, Posterior distribution obtained for the WHALE post-duplication retention parameter  $q$ , for each hypothesis presented in **a**. Stars indicate distributions significantly different from 0 (Bayes factors  $BF_{\text{Null vs. WGD}} < 10^{-3}$ ), which correspond to the supported polyploidization events. **c**, Alternative set of polyploidization

hypotheses tested, as in **a**, but with two successive duplications proposed in the ancestral vertebrate lineage (1R<sub>v</sub> and 2R<sub>v</sub>). **d**, Posterior distribution obtained for the WHALE post-duplication retention parameter, for each hypothesis presented in **c**. Here, the posterior distribution for retention parameters of the 1R<sub>v</sub> and 2R<sub>v</sub> events are bimodal, suggesting that the method cannot effectively separate parameters estimated for 1R<sub>v</sub> and 2R<sub>v</sub> when starting from identical priors. **e**, Use of distinct priors on 1R<sub>v</sub> (Beta(8, 2)) and 2R<sub>v</sub> (Beta(2, 8)) separates the estimated posterior distribution into distinct unimodal posterior distributions and provides support for a single shared 1R<sub>v</sub> event in the vertebrate stem lineage. This analysis was performed on a random subset of 1,000 gene families, to reduce computational time (Methods).

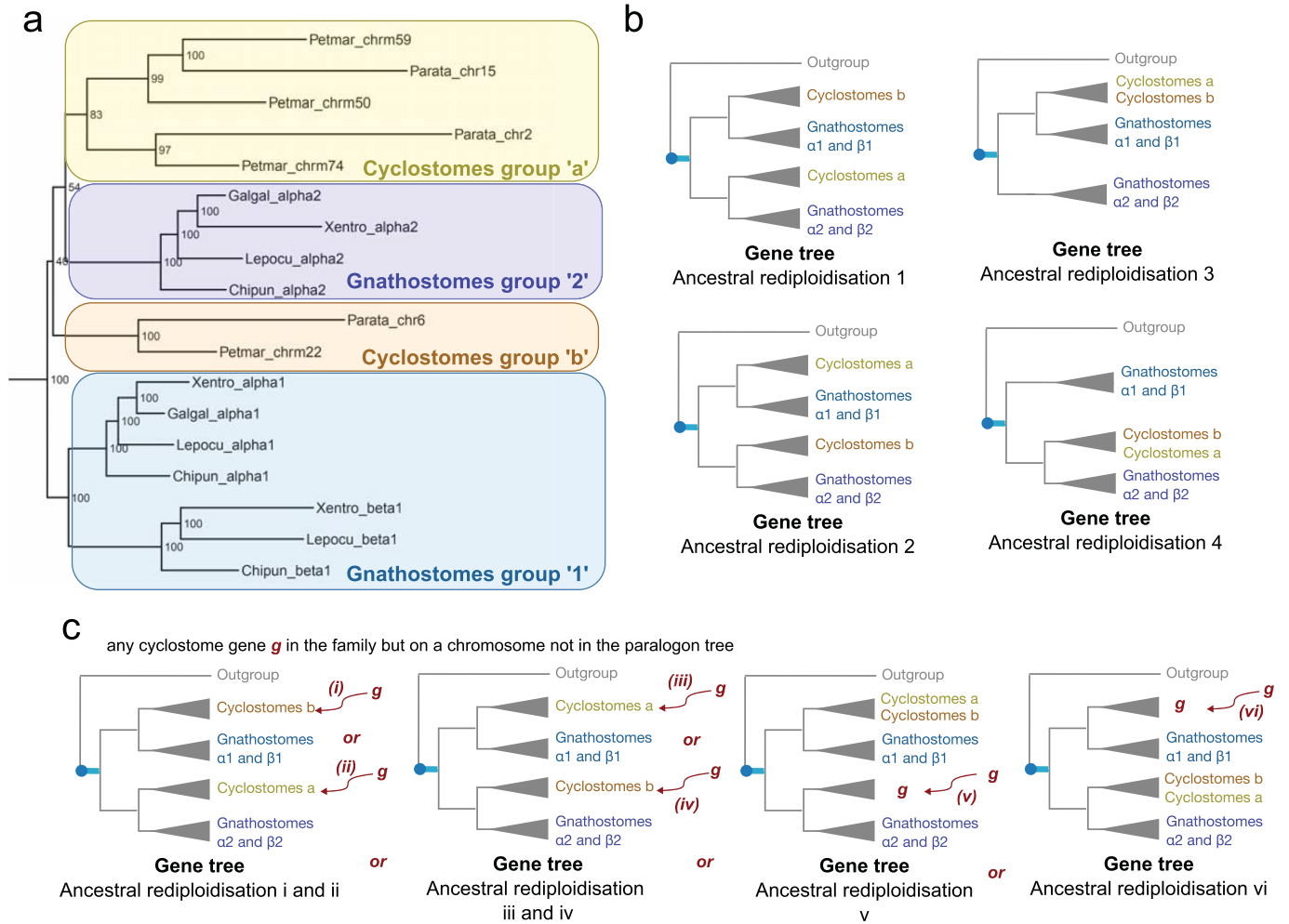


Extended Data Fig. 5 | See next page for caption.

**Extended Data Fig. 5 | Timescale of vertebrate genome evolution.**

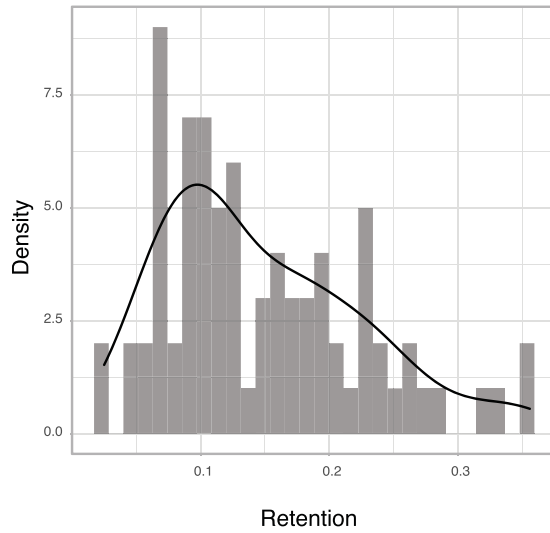
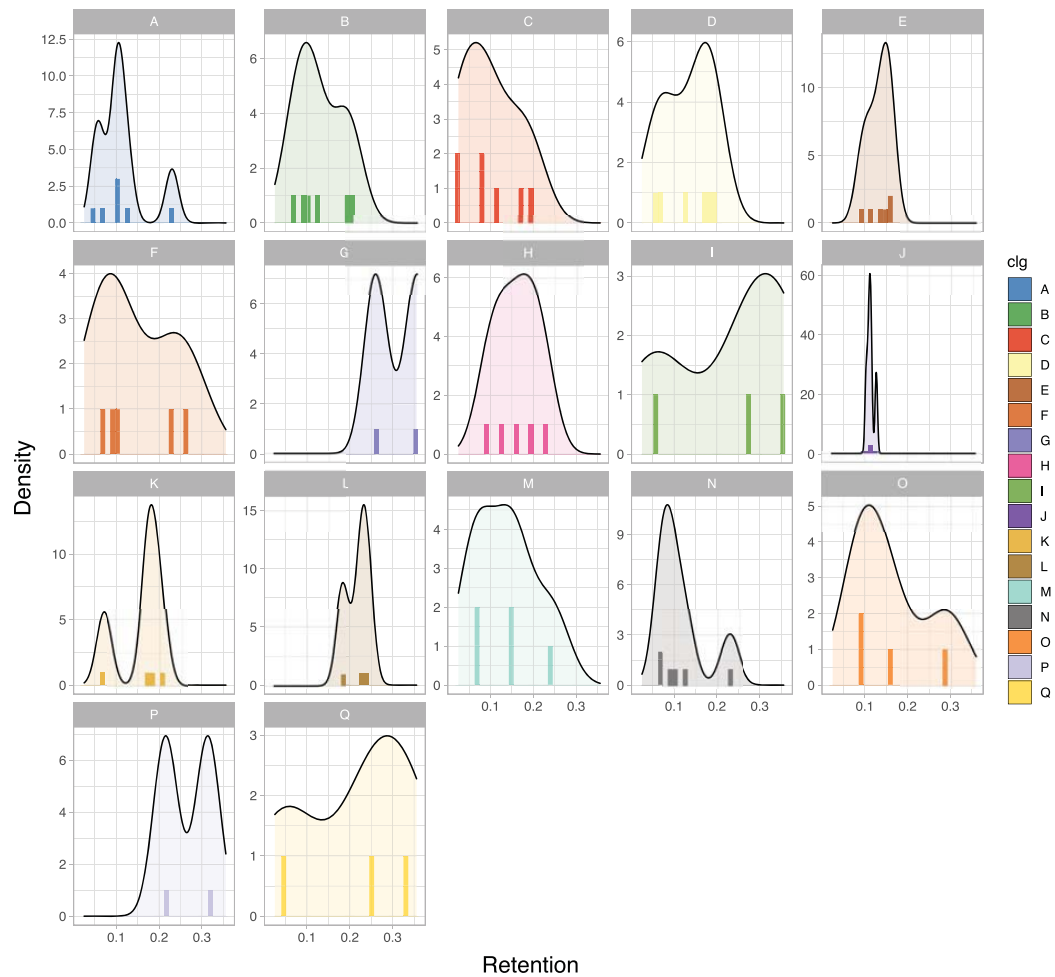
**a.** Distributions of timings for speciation and duplication events derived from paralogon phylogenies, showing details of the distributions indicated in Fig. 3c. **b.** Scenario for genome duplication and speciation events during early vertebrate evolution. Filled black circles or ovals mark speciation events; horizontal rectangles indicate presumptive auto-tetraploidies; starbursts indicate allo-polyploidies arising from hybridization of distinct progenitors (for example, alpha–beta in gnathostomes). Timings are based on **a**. Note that although speciation times (for example, the split between gnathostome progenitors alpha and beta, divergence of lamprey and hagfish lineages) can be estimated from gene or paralogon trees, hybridization times (for example,  $2R_{JV}$ , shown as green starburst) cannot be estimated from gene-tree analysis. Similarly, homoeologous recombination after auto-tetraploidization implies that the auto-tetraploidization event itself cannot be timed, but only the cessation of homoeologous recombination. Thus, the estimate of around 527 Ma for  $1R_V$  (horizontal blue rectangle) represents the cessation of recombination after this presumptive auto-tetraploidy (open rectangle on vertebrate stem)

with homologous recombination represented by blue shading. The absolute timing of  $1R_V$  itself is unknown. (Auto-tetraploidy is suggested by the lack of differential gene loss between the two paralogous branches after  $1R$ , as noted previously<sup>19</sup>.) The rough estimate of a 10-million-year interval between the alpha–beta split and  $2R_{JV}$  allo-tetraploidy is based on analogy with recent vertebrate allo-tetraploidies in frogs and goldfish. Cyclostome hexaploidization  $2R_{CV}$  is shown as a two-step process culminating in the hybridization of diploid and tetraploid stem cyclostomes (orange starburst). This scenario follows the recent model of hexaploidy in sturgeon in which auto-tetraploids and diploid species coexist and hybridize<sup>48</sup>. In this scenario, the earliest divergences among cyclostome paralogues occurs around 511 Ma when the diploid and future tetraploid lineages split, which could be coincident with the early tetraploidization itself. Homoeologous recombination (shown as orange shading) is largely complete by around 493 Ma, defining a second peak in paralogue divergence (horizontal orange rectangle). Not shown is ongoing homoeologous recombination in CLGB which continues into the stem hagfish and lamprey lineages, as discussed further in the main text.



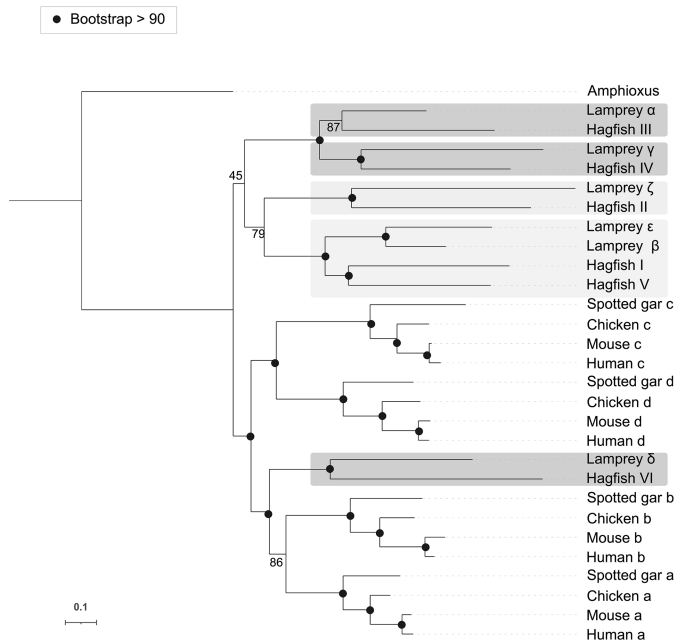
**Extended Data Fig. 6 | Method for construction of post-1R ancestral rediploidization constrained gene-tree topologies, using CLGM as an example.** **a**, Gnathostome 1R 1 and 1R2 copies can be confidently identified and serve as a skeleton to build ancestral rediploidization tree topologies (blue-purple groups). Hagfish and lamprey chromosomes confidently grouped in a clade from the CLGM paralogon tree are defined as potential 1R-derived paralogs (yellow-orange groups) and kept together in the constrained ancestral rediploidization tree topology (see **b**). All sets of cyclostome chromosomes that were kept together for other CLGs are indicated in Supplementary Table 5. **b**, Possible groupings of hagfish and lamprey genes with

gnathostome genes based on their chromosomal location, following 1R ancestral rediploidization. **c**, Genes located on hagfish and lamprey chromosomes that are not considered in the reconstructed paralogon tree (due to low representation because of small-scale rearrangements displacing them on different chromosomes) can each be placed on either side of the duplication in the absence of any prior information. In the presented scenario, this results in six different possible ancestral rediploidization (i to vi) constrained tree topologies. Only topologies with a maximum of three lamprey genes and three hagfish genes on each side of the IR are permitted, to remove possibly confounding effects of complex multicopy gene families.

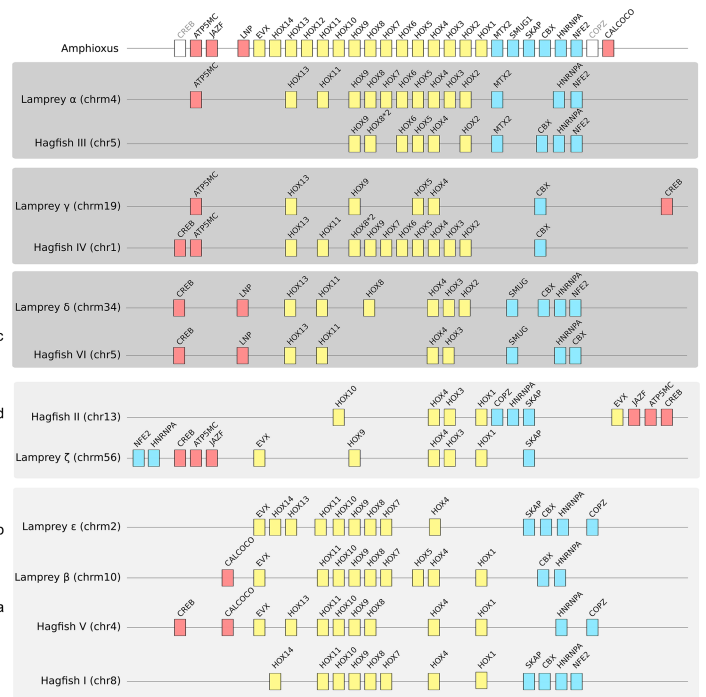
**a****b**

**Extended Data Fig. 7 | Orthologue retention rates after  $2R_{CV}$ .** Retention is computed for each *P. marinus* chromosomal segment derived from a single CLG as the fraction of orthologues maintained on the segment in a comparison with the total number of orthologues for the same CLG in *Branchiostoma floridae*.

**a**, Distribution of retention rates plotted for all CLGs. **b**, Distribution of retention rates plotted for each CLG. These distributions are not distinctly bimodal, in contrast to the finding for  $2R_{IV}$ <sup>19</sup>. Lamprey is used because it more closely preserves ancestral cyclostome state, as proposed previously<sup>20</sup>.

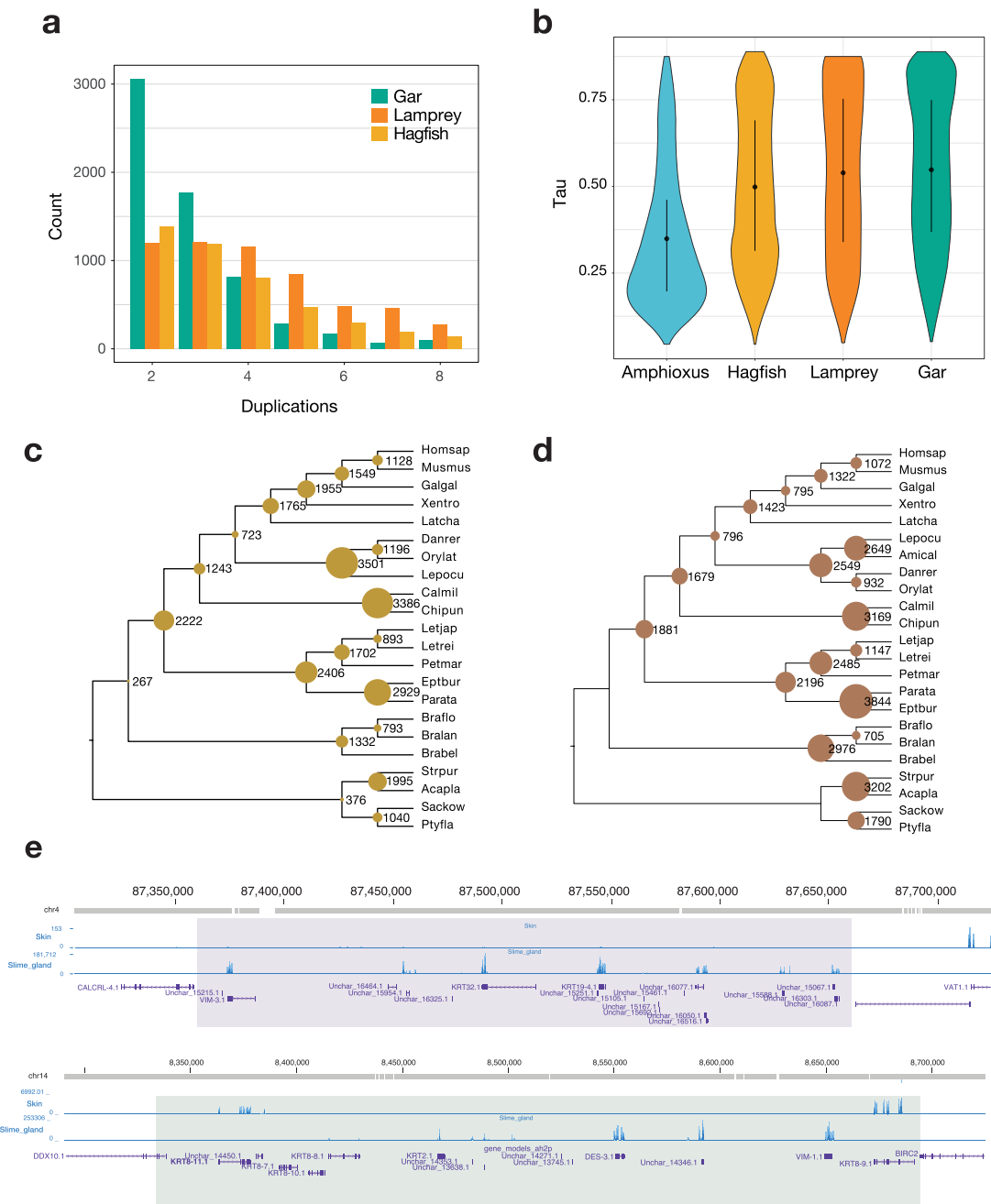


**Extended Data Fig. 8 | Phylogeny of the Hox clusters on the basis of a concatenation of Hox genes and bystanders.** Left, phylogeny of the Hox clusters, with node bootstrap support. One-to-one orthologies for gnathostome clusters are well-supported, similarly for cyclostome clusters with the exception of hagfish V-I and lamprey β-ε. Dark grey boxes highlight cyclostomes clusters that are expected to be orthologous to gnathostome



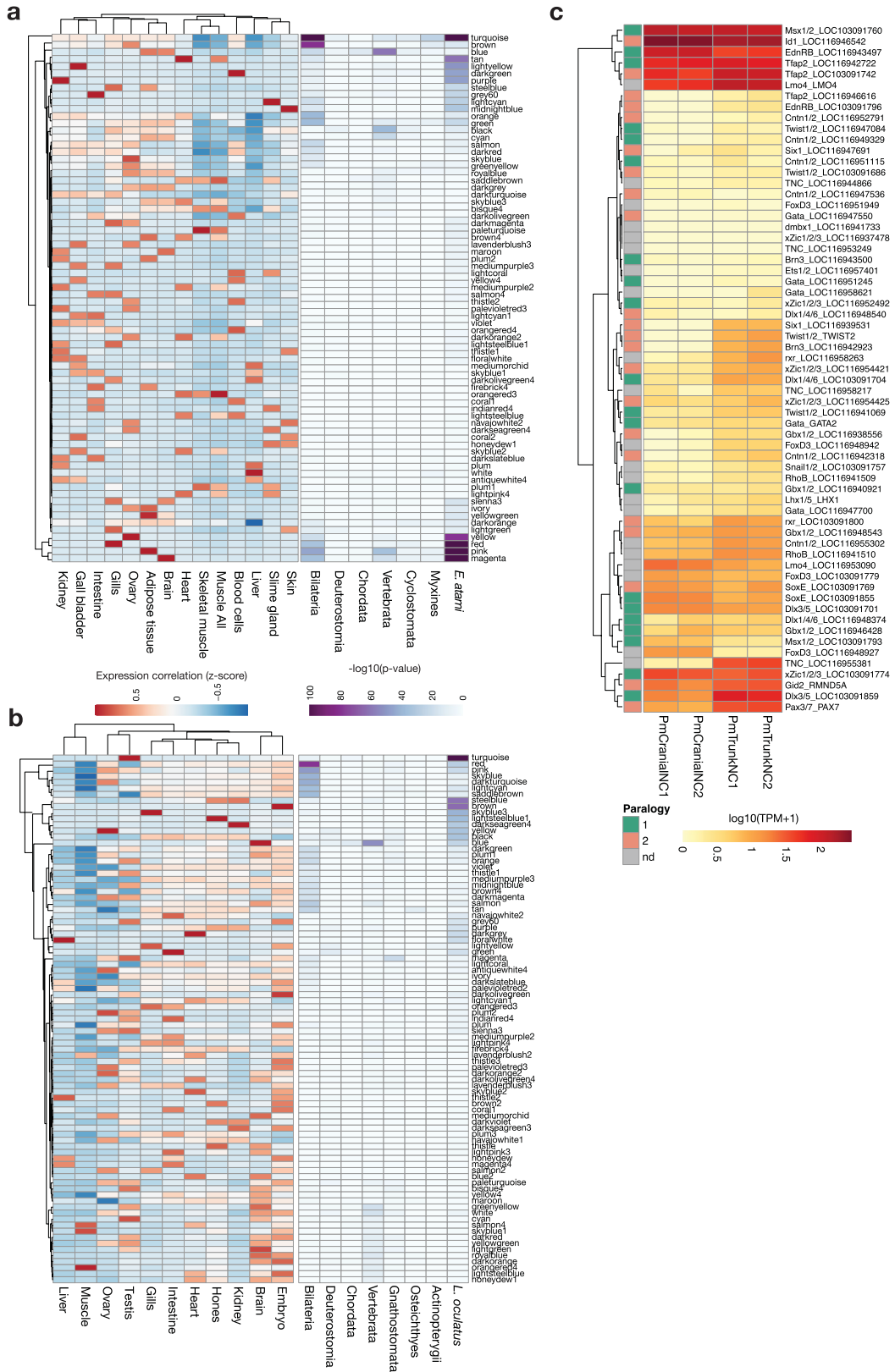
clusters A/B based on chromosomal orthology (Fig. 3d), similarly light grey boxes are for expected orthologs to gnathostome clusters C/D. Right, schematic representation of cyclostome and gnathostome Hox clusters. Hox genes are shown as yellow boxes, 5' bystanders as red boxes and 3' bystanders as blue boxes. The order of genes reflects the actual arrangement of genes in each species.





**Extended Data Fig. 9 | Evolution of duplicated genes and gene families in hagfish. a,** Counts of gene families containing the specified number of retained paralogues in gar, lamprey (*P. marinus*) and hagfish (*E. atami*). **b,** Comparison of the tissue-specificity of gene expression (tau index) for ohnologue gene families in lampreys, hagfish, gar and the (unduplicated) amphioxus outgroup (Methods). The distribution of the maximal tau value for each gene family is

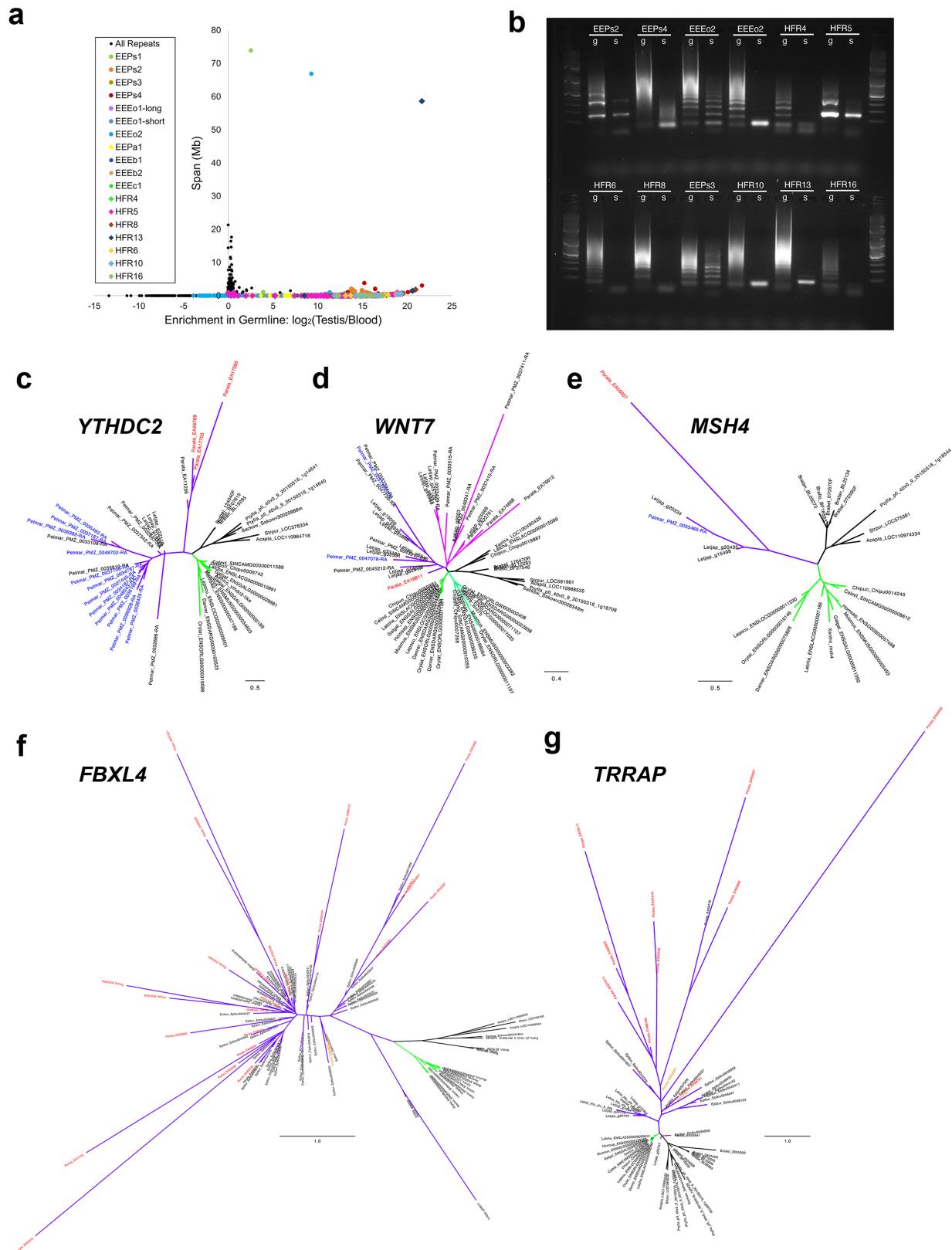
shown. **c,** Node-specific gene loss events inferred by GeneRax in a species–gene tree reconciliation framework (Methods). Species labels are specified in Supplementary Table 8. **d,** Loss of Panther families across deuterostomes species inferred as the most parsimonious events from gene-family composition. **e,** Genome structure of the two clusters of expanded keratin genes, with mRNA expression in slime gland and skin (blue track).



Extended Data Fig. 10 | See next page for caption.

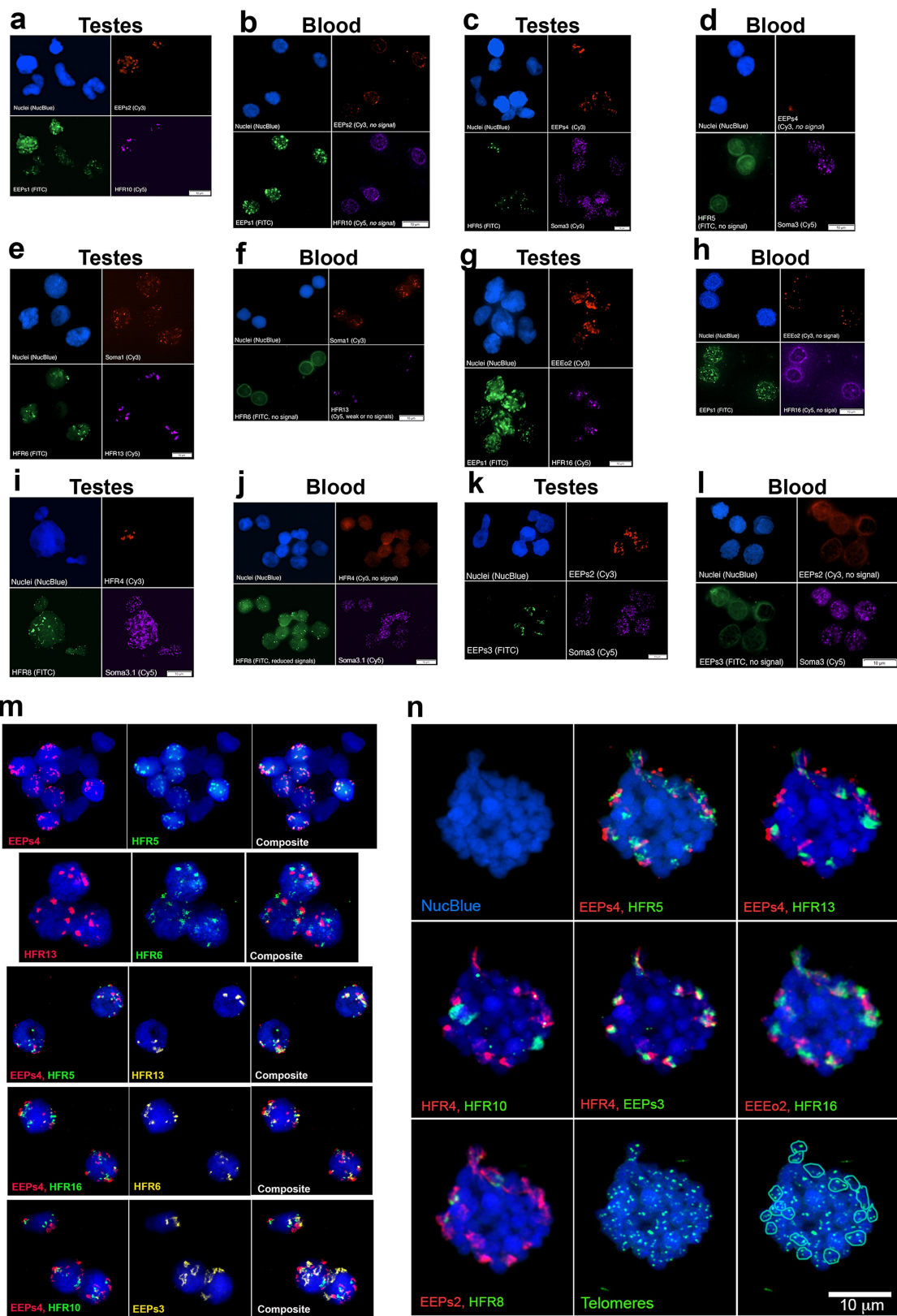
**Extended Data Fig. 10 | Gene expression and gene duplications in vertebrates. a, b.** Weighted gene co-expression network analysis (WGCNA) among organs for hagfish (**a**) and gar (**b**). Each row corresponds to a WGCNA cluster (with an arbitrary colour name) and its expression specificity is shown in selected tissues on the left (**a**, hagfish, **b**, gar). The enrichment of gene duplicated at successive phylogenetic nodes in each WGCNA cluster is indicated on the right as the p-value ( $-\log_{10}$ ) of hypergeometric tests. A significant enrichment is observed in gene with strong neural expression (brain, blue cluster). **c.** Expression of selected paralogues involved in neural crest

specification and migration in cranial and trunk neural crest tissues from lamprey *P. marinus*. RNA-seq data from a previous study<sup>56</sup> was quantified using the latest version of the lamprey genome and RefSeq annotation (kPetMar1). For each gene family, all paralogues derived from the vertebrate polyploidization event ( $1R_v$  and  $2R_{cy}$ ) are considered and classified (see Supplementary Tables 9 and 10). As denoted in inset, 1 (green cells) and 2 (pink cells) refer to the two original paralog branches derived from  $1R_v$  (see main Fig. 4a). Grey groups could not be definitively assigned.



**Extended Data Fig. 11 | Eliminated genes and repeats identified in the hagfish genome.** **a**, Plot showing the degree of germline enrichment and estimated span of all predicted repetitive elements. Previously identified elements<sup>30,33</sup> are highlighted by coloured circles and new high-copy elements are highlighted by coloured diamonds. **b**, PCR validation illustrating germline enrichment and tandem repetition of predicted satellite elements. g: germline (testes) DNA used as template, s: somatic (blood) DNA used as template. **c-e**, Gene trees for homologues that are eliminated in both lamprey and hagfish.

Gnathostome clades are highlighted in shades of green and cyclostome clades are highlighted in shades of purple. Individual germline-specific genes are highlighted in red (hagfish) or blue (lamprey). **c**, Tree for YTHDC2 homologues. **d**, Tree for WNT7 homologues. **e**, Tree for MSH4 homologues. **f, g**, Gene trees for homologues that are highly duplicated in hagfish. Gnathostome clades are highlighted in shades of green and cyclostome clades are highlighted in shades of purple. Individual germline-specific genes are highlighted in red. **f**, Tree for FBXL4 homologues. **g**, Tree for TRRAP homologues.



Extended Data Fig. 12 | See next page for caption.

# Article

**Extended Data Fig. 12 | FISH of repeats to germline and somatic interphase nuclei.** Nuclei are labelled with the DNA stain NucBlue (blue) and for all panels labelled “no signal” fluorescence images are overexposed to both show background signal and aid in confirming the location of nuclei in those images. **a,b**, Germline enriched EEPs2 (red) and HFR10 (magenta), and the somatic repeat EEPs1 (green) are hybridized to nuclei isolated from **a**, germline: testes and **b**, soma: blood. **c,d**, Germline enriched repeats EEPs4 (red) and HFR5 (green), and the somatic repeat Soma3 (magenta) are hybridized to nuclei isolated from **c**, germline: testes and **d**, soma: blood. **e,f**, Germline enriched repeats HFR13 (magenta) and HFR6 (green), and the somatic repeat Soma1 (red) are hybridized to nuclei isolated from **e**, germline: testes and **f**, soma: blood. **g,h**, Germline enriched repeats EEEo2 (red) and HFR16 (magenta), and the somatic repeat EEPs1 (green) are hybridized to nuclei isolated from **g**,

germline: testes and **h**, soma: blood. **i,j**, Germline enriched repeats HFR4 (red) and HFR8 (green), and the somatic repeat Soma3.1 (magenta) are hybridized to nuclei isolated from **i**, germline: testes and **j**, soma: blood. **k,l**, Germline enriched repeats EEPs2 (red) and EEPs3 (green), and the somatic repeat Soma3 (magenta) are hybridized to nuclei isolated from **k**, germline: testes and **l**, soma: blood. **m,n**, In situ Hybridization of probes for ten germline-enriched satellite sequences. **m**, Probes are hybridized to germline interphase nuclei. **n**, Probes are hybridized to germline interphase nuclei. The location of hybridization signals for telomere probes and approximate bounds of 18 germline-specific dyads, corresponding to nine distinct germline-specific chromosomes. For all images, pairs of repeats are shown to aid in visualizing the relative location of individual probes.

## Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our [Editorial Policies](#) and the [Editorial Policy Checklist](#).

### Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a Confirmed

- The exact sample size ( $n$ ) for each experimental group/condition, given as a discrete number and unit of measurement
- A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- The statistical test(s) used AND whether they are one- or two-sided  
*Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
- A description of all covariates tested
- A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- For null hypothesis testing, the test statistic (e.g.  $F$ ,  $t$ ,  $r$ ) with confidence intervals, effect sizes, degrees of freedom and  $P$  value noted  
*Give  $P$  values as exact values whenever suitable.*
- For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- Estimates of effect sizes (e.g. Cohen's  $d$ , Pearson's  $r$ ), indicating how they were calculated

*Our web collection on [statistics for biologists](#) contains articles on many of the points above.*

### Software and code

Policy information about [availability of computer code](#)

Data collection

Data analysis https://github.com/fmarletaz/hagfish

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio [guidelines for submitting code & software](#) for further information.

## Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our [policy](#)

Raw and processed sequences have been deposited in NCBI SRA (PRJNA953751 ) and Gene Expression Omnibus (GSE230176). The RNA-seq of *Myxine glutinosa* is available on the SRA (SRR25213276). The resequenced somatic tissues are also available on the SRA (Blood: SRR24133795 and Testes: SRR24130678). RNA-seq datasets used for comparative analyses are publicly available for Japanese lamprey (PRJNA354821, PRJNA349779, PRJNA312435), gar (PRJNA255881) and amphioxus (PRJNA416977) as well as the sea lamprey (PRJNA497902).

## Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

Life sciences       Behavioural & social sciences       Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see [nature.com/documents/nr-reporting-summary-flat.pdf](https://www.nature.com/documents/nr-reporting-summary-flat.pdf)

## Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	No statistical method was used for determining sample size.
Data exclusions	No data exclusions.
Replication	No such experiments.
Randomization	No randomization was used
Blinding	Blinding was not relevant to our study as comparison were performed by computer software not influenced by investigator.

## Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

### Materials & experimental systems

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> Antibodies
<input checked="" type="checkbox"/>	<input type="checkbox"/> Eukaryotic cell lines
<input checked="" type="checkbox"/>	<input type="checkbox"/> Palaeontology and archaeology
<input type="checkbox"/>	<input checked="" type="checkbox"/> Animals and other organisms
<input checked="" type="checkbox"/>	<input type="checkbox"/> Human research participants
<input checked="" type="checkbox"/>	<input type="checkbox"/> Clinical data
<input checked="" type="checkbox"/>	<input type="checkbox"/> Dual use research of concern

### Methods

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> ChIP-seq
<input checked="" type="checkbox"/>	<input type="checkbox"/> Flow cytometry
<input checked="" type="checkbox"/>	<input type="checkbox"/> MRI-based neuroimaging

## Animals and other organisms

Policy information about [studies involving animals](#); [ARRIVE guidelines](#) recommended for reporting animal research

Laboratory animals	<i>For laboratory animals, report species, strain, sex and age OR state that the study did not involve laboratory animals.</i>
Wild animals	Animals were samples in Suruga Bay, off Yaizu (300-330m depth) and maintained in seawater aquariums at 11-13°C.
Field-collected samples	<i>For laboratory work with field-collected samples, describe all relevant parameters such as housing, maintenance, temperature, photoperiod and end-of-experiment protocol OR state that the study did not involve samples collected from the field.</i>



## Ethics oversight

In agreement with procedures authorized by Guidelines for Proper Conduct of Animal Experiments by the Science Council of Japan (2006), animals were anaesthetized using Tricaine (MS222, Sigma) before sacrifice and dissection.

Note that full information on the approval of the study protocol must also be provided in the manuscript.