

# The genomic origin of the unique chaetognath body plan

<https://doi.org/10.1038/s41586-025-09403-2>

Received: 22 October 2024

Accepted: 14 July 2025

Published online: 13 August 2025

Open access

 Check for updates

Laura Piovani<sup>1</sup>, Daria Gavriouchkina<sup>2</sup>, Elise Parey<sup>1</sup>, Luke A. Sarre<sup>3,4</sup>, Katja T. C. A. Peijnenburg<sup>5,6</sup>, José María Martín-Durán<sup>3,4</sup>, Daniel S. Rokhsar<sup>7,8,9</sup>, Noriyuki Satoh<sup>10</sup>, Alex de Mendoza<sup>3,4</sup>, Taichiro Goto<sup>11</sup> & Ferdinand Marlétaz<sup>1✉</sup>

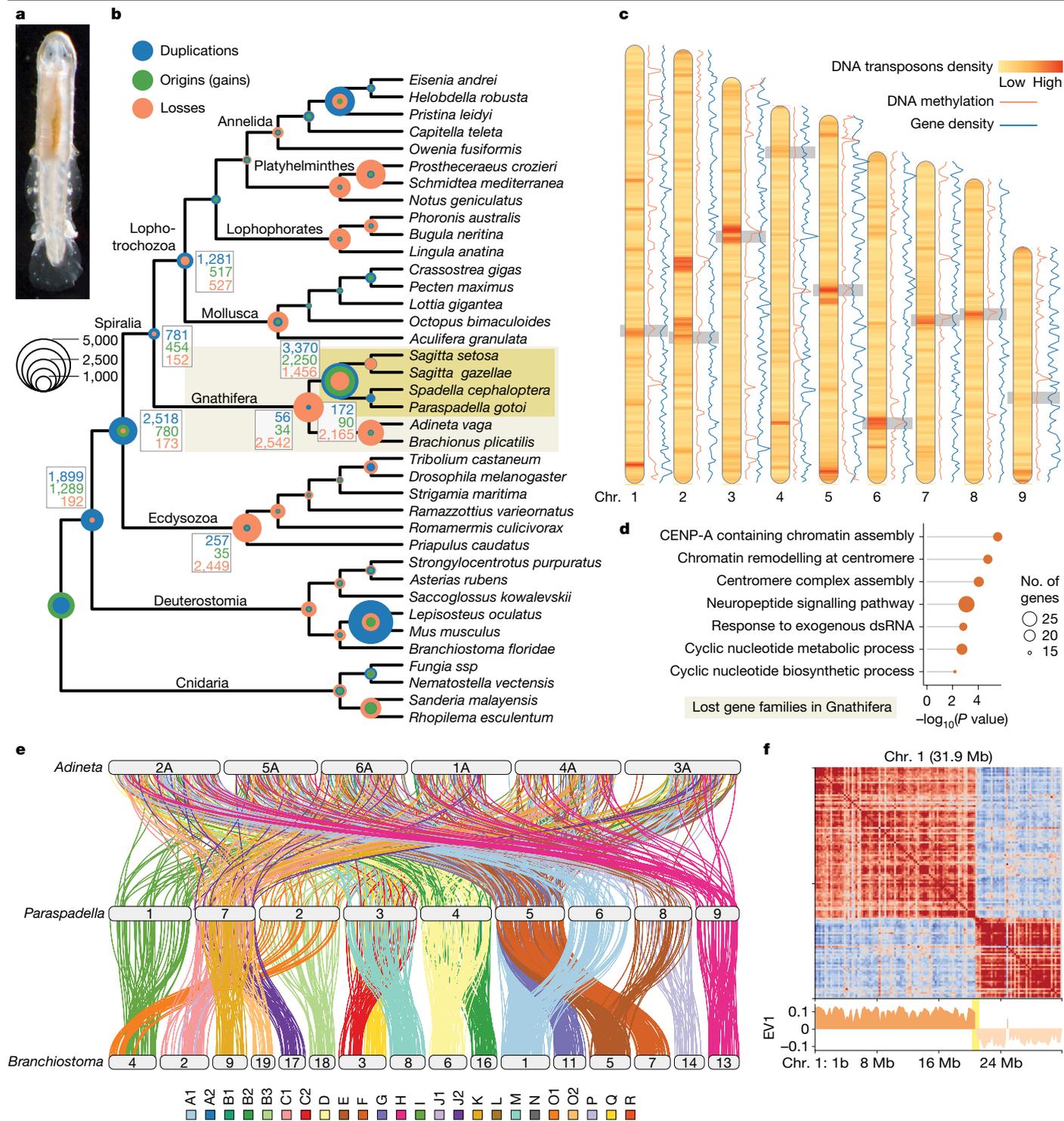
The emergence of animal phyla, each with their unique body plan, was a rapid event in the history of animal life, yet its genomic underpinnings are still poorly understood<sup>1</sup>. Here we investigate at the genomic, regulatory and cellular levels, the origin of one of the most distinctive animal phyla, the chaetognaths, whose organismal characteristics have historically complicated their phylogenetic placement<sup>2,3</sup>. We show that these characteristics are reflected at the cell-type level by the expression of genes that originated in the chaetognath lineage, contributing to adaptation to planktonic life at the sensory and structural levels<sup>4</sup>. Similarly to other members of gnathiferans (which also include rotifers and several other microscopic phyla)<sup>5,6</sup>, chaetognaths have undergone accelerated genomic evolution with gene loss and chromosomal fusions<sup>7,8</sup>. Furthermore, they secondarily duplicated thousands of genes<sup>9,10</sup>, without evidence for a whole-genome duplication, yielding, for instance, tandemly expanded Hox genes, as well as many phylum-specific genes. We also detected repeat-rich highly methylated neocentromeres and a simplified DNA methylation toolkit that is involved in mobile element repression rather than transcriptional control. Consistent with fossil evidence<sup>11,12</sup>, our observations suggest that chaetognaths emerged after a phase of morphological simplification through a reinvention of organ systems paralleled by massive genomic reorganization, explaining the uniqueness of their body plan.

Whereas different scenarios have been proposed to account for the emergence of animal phyla such as chordates<sup>13</sup>, arthropods<sup>14</sup> or annelids<sup>15</sup>, other lineages resist attempts to do so. A prominent example is the chaetognaths, whose phylogenetic position in the tree of bilaterians has long been disputed<sup>2</sup>. Although chaetognaths do not share obvious similarities with any other extant animal group and their ancient fossil record does not provide obvious clues into their origins, their body plan has remained remarkably stable since the Cambrian period<sup>11,12</sup>. Patterns of embryonic development (enterocoely and secondary mouth opening)<sup>16</sup> originally suggested an association with deuterostomes, whereas other traits were reminiscent of protostomes (chitinous grasping spines and nervous system)<sup>3,4</sup>. After years of sometimes contradictory molecular analyses<sup>10,17</sup>, chaetognaths have been assigned to the clade Gnathifera, which represents the sister group to the rest of lophotrochozoan animals within spiralian<sup>5,6</sup> (Fig. 1a,b). This placement was then corroborated by the shared presence of the *MedPost* Hox gene—a unique Hox gene bearing median and posterior molecular signatures—with rotifers, another prominent gnathiferan group<sup>18,19</sup>, and was further supported by fossil interpretation<sup>11,12,20</sup>.

Understanding the genomic events that led to the emergence of such an unusual body plan thus represents a key challenge to understanding

the origin of animals itself. Recent studies have shown how genomic and regulatory changes have sometimes (for example, vertebrates<sup>21</sup> and annelids<sup>7</sup>) but not always (for example, echinoderms<sup>22</sup>) accompanied the emergence of novel bodyplans. Many lophotrochozoans, such as annelids, molluscs or lophophorates, retained ancestral bilaterian genomic characters, such as gene content, introns<sup>23</sup>, or conserved chromosomal ancestral linkage<sup>7,24</sup>. However, within their sister group (the Gnathifera) the only lineage characterized from a genomic standpoint are the rotifers, which underwent fast genomic evolution associated with the acquisition of asexuality in bdelloids<sup>25,26</sup>. Determining how the evolution of genes and genomes shaped the cell-type complement of gnathiferan lineages is thus crucial to understand the evolution of organismal complexity<sup>27</sup>. To try to determine whether gnathiferans preserved some of the ancestral traits uncovered in other spiralian and to grasp the origin of the enigmatic chaetognath body plan, we focused on the free-living predatory chaetognath *Paraspadella gotoi*, and generated a chromosome-scale genome, regulatory profiling resources (Hi-C-seq, assay for transposase-accessible chromatin with high-throughput sequencing (ATAC-seq) and methylome data) as well as a single-cell sequencing atlas, which, together, reveal how changes in the gene repertoire and gene regulation underlined the cell-type bases

<sup>1</sup>Centre for Life's Origin and Evolution, Department of Genetics, Evolution and Environment, University College London, London, UK. <sup>2</sup>UK Dementia Research Institute, University College London, London, UK. <sup>3</sup>School of Biological and Behavioural Sciences, Queen Mary University of London, London, UK. <sup>4</sup>Centre for Epigenetics, Queen Mary University of London, London, UK. <sup>5</sup>Marine Evolution and Ecology, Naturalis Biodiversity Center, Leiden, The Netherlands. <sup>6</sup>Department of Freshwater and Marine Ecology, Institute for Biodiversity and Ecosystem Dynamics, University of Amsterdam, Amsterdam, The Netherlands. <sup>7</sup>Molecular Genetics Unit, Okinawa Institute of Science and Technology Graduate University, Okinawa, Japan. <sup>8</sup>Department of Molecular and Cell Biology, University of California, Berkeley, CA, USA. <sup>9</sup>Chan-Zuckerberg BioHub, San Francisco, CA, USA. <sup>10</sup>Marine Genomics Unit, Okinawa Institute of Science and Technology Graduate University, Okinawa, Japan. <sup>11</sup>Department of Biology, Faculty of Education, Mie University, Tsu, Japan. ✉e-mail: f.marletaz@ucl.ac.uk



**Fig. 1 | Evolution of gene complement and genome architecture in gnathiferans and chaetognaths. a**, Adult *P. gotoi* (photograph credit: Taichiro Goto). **b**, Summary of gene family reconstruction in selected metazoan genomes (Supplementary Table 2) including inference of duplication events using gene tree–species tree reconciliation<sup>63</sup>. **c**, Chromosomes of *P. gotoi*, highlighting the binned density of DNA transposons elements, the level of DNA methylation and the gene density, and the position of centromeres inferred from Hi-C data (highlighted with a grey box). Chr., chromosome. **d**, Gene ontology terms (biological functions) enriched in gene families lost in the gnathiferan lineage (Fisher’s exact test, two-sided). **e**, Evolution of gene linkage

in gnathiferans, highlighting the fusions that took place in chaetognath and the extensive chromosomal scrambling in the *Adineta* rotifer. Links between chromosomes showing significant reciprocal enrichment are displayed and coloured according to their bilateral linkage group as defined by a comparison with amphioxus. Amphioxus chromosomes not showing an enrichment with *P. gotoi* are omitted. **f**, Pearson correlation matrix and corresponding eigenvector (EV1) derived from Hi-C data for chromosome 1. The shift in contacts associated with position of centromere is highlighted (for other chromosomes, see Extended Data Fig. 4).

of body plan evolution in this active predator with elaborate sensory organs and complex behaviours<sup>28</sup>.

## Fast genome evolution in Gnathifera

We sequenced the genome of the chaetognath *P. gotoi* by using long and short reads from a five-generation inbred line and scaffolded the assembly to chromosome-scale level using proximity ligation data (Fig. 1c and Extended Data Fig. 1a). We obtained a relatively small assembly size of 257 Mb and found 9 major chromosome-size scaffolds, consistent with cytogenetic data in other species<sup>2,29</sup>. Repeats make up to 20% of the genome, with the major class being long terminal repeat (LTR) retrotransposons (8.6% of repeats), followed by long interspersed nuclear elements (LINEs) (4.6%) and DNA transposons (3.4%) and a large fraction of unclassified repeats (around 80%) (Extended Data Fig. 1b). Of note, several DNA transposons are enriched in localized putative centromeric regions (Fig. 1c). We annotated 22,072 protein-coding genes using long and short-read transcriptomic data (Supplementary Table 1).

To characterize the genomic evolution of gnathiferans, we compared our chaetognath genome and transcriptomes with those of other spiralian<sup>7,25,30</sup>. Gene family analysis uncovered extensive gene loss (2,542 ancestral gene families) in the branch leading to Gnathifera (Fig. 1b and Supplementary Tables 2 and 3). This observation is confirmed by the distribution of PANTHER families (Extended Data Fig. 2a) and maximum-likelihood estimation of gene loss (Extended Data Fig. 2b), and is accompanied with contraction of neural- and sensory-related gene families (Extended Data Fig. 2c,d). Among gnathiferans, both rotifers and to a lesser extent chaetognaths further experienced lineage-specific gene losses (2,165 and 1,456 families lost, respectively) and contractions of multiple gene families<sup>31</sup> (172 and 116 families, respectively; Extended Data Fig. 2c). Estimation of these gene family events is robust to alternative species tree<sup>5,6</sup> (Supplementary Fig. 1). Functional enrichment analyses surprisingly indicate that nearly all genes involved in CenH3 (also known as CENP-A) centromeric chromatin assembly have been lost in Gnathifera (12 out of 20, including the CenH3 and centromere protein T (CENP-T) genes; Fig. 1d and Supplementary Table 4), a condition that has been observed in insects exhibiting non-localized holocentromeres<sup>32</sup>.

Absence of CenH3 centromeres has previously been associated with accelerated genomic rearrangements<sup>33</sup>. We thus investigated the retention in chaetognath and rotifers of bilaterian ancestral linkage groups (BLGs), which are conserved throughout bilaterian animal evolution, from chordates to annelids<sup>7,8</sup>. We found that most BLGs remain statistically detectable when comparing chaetognaths with amphioxus, sea star and sea scallop despite multiple fusions, whereas they appear completely scrambled in rotifers (Fig. 1e and Extended Data Fig. 3a–d). Most chaetognath chromosomes seem to derive from 2 to 4 fused BLGs with detected BLGs only represented once, and A2, B2, L, N or Q not detected. Among the four fusions-with-mixing (⊗) previously reported in spiralian, namely H⊗Q, J2⊗L, K⊗O2, and O1⊗R<sup>7,8,34</sup>, we identified the presence of K⊗O2 and O1⊗R in chaetognaths, indicating that they are likely to be shared by all spiralian (Fig. 1e and Extended Data Fig. 3). We could not detect Q and L, which could mean either that J2⊗L and H⊗Q fusions took place after the divergence of gnathiferans (particularly as we detected H and J2 alone) or that the signal is too limited. No chromosomal correspondence is detectable between chaetognaths and rotifers (Fig. 1e and Extended Data Fig. 3d), indicating a similar fast pace of chromosomal evolution in bdelloid and monogonont rotifers (Extended Data Fig. 3e,f). Together, these results suggest an accelerated rate of chromosomal rearrangement in the gnathiferan lineages.

The loss of CenH3 has been associated in other lineages with the presence of holocentromeres and correlated with an absence of chromatin compartmentalization at the level of chromosome arms in plants<sup>33</sup>. Here we conversely found a bimodal compartmentalization of the three-dimensional chromatin architecture for all *P. gotoi* chromosomes

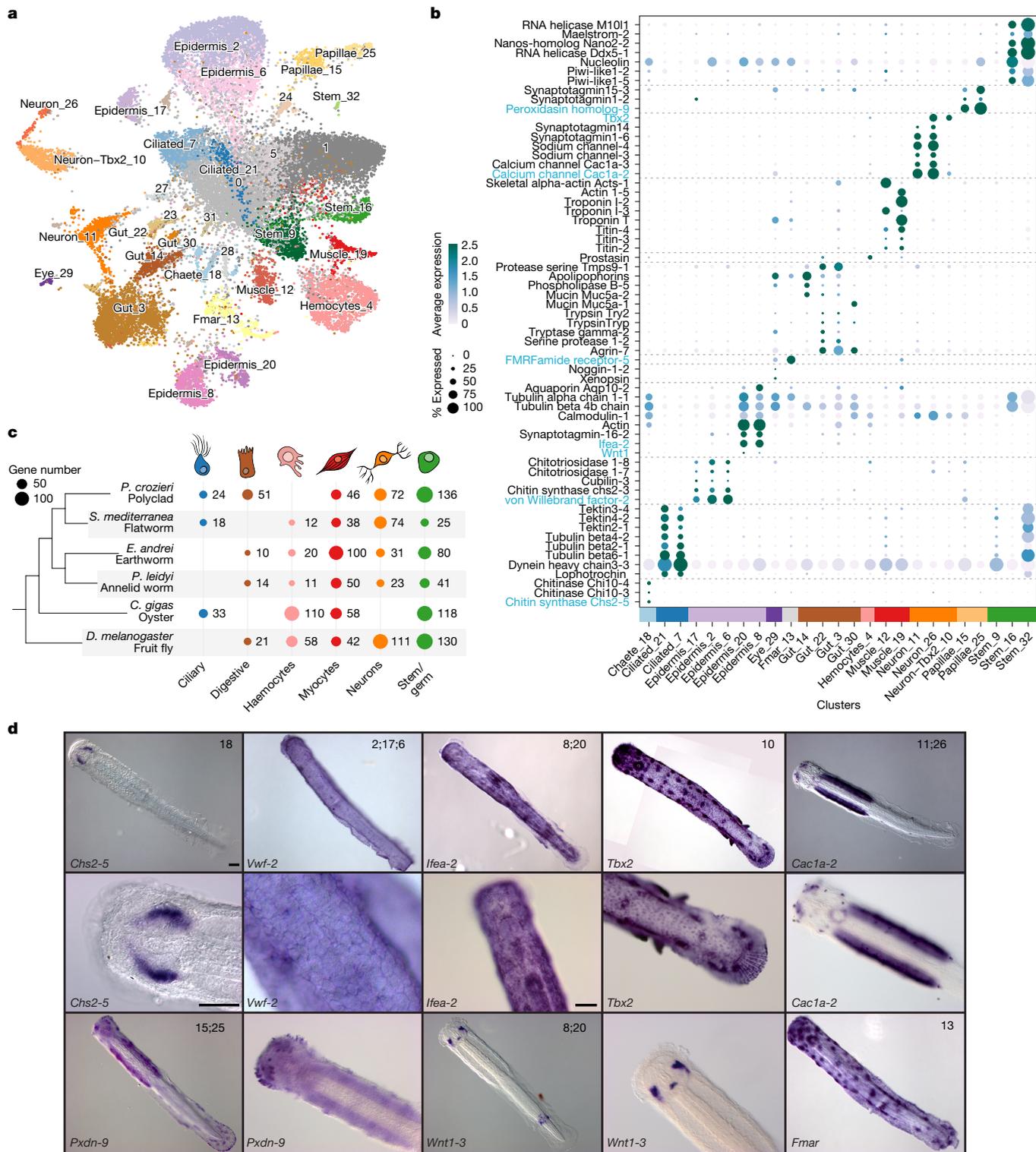
(Fig. 1f and Extended Data Fig. 4). The borders between these two interacting compartments are enriched in several mobile element categories observed in centromeric regions of other lineages, such as DNA transposons (for example, PIF/Harbinger and ISL2EU transposons; Fig. 1c) as well as LINE/SINE retrotransposons (Supplementary Table 5) and they show increased CpG methylation (Fig. 1c). These regions therefore exhibit classical features of centromeres indicating that chaetognaths possess localized centromeres despite lacking CenH3 and other molecular actors of centromere assembly<sup>35</sup> and corroborating prior microscopic observations<sup>2,29</sup>. Unlike chaetognaths, the rotifer *Adineta vaga* are likely to possess holocentromeres<sup>25,35,36</sup>, suggesting that centromeres have broadly diverged in the gnathiferan lineage and have been replaced in chaetognaths by neocentromeres making use of alternative molecular components.

At a smaller scale, chaetognath chromosomes appear to display very limited large-scale A/B compartmentalization (Extended Data Fig. 4a,e) and we did not observe an aggregated signal consistent with the presence of local topologically associating domains<sup>37</sup> (TADs) (Extended Data Fig. 4c,d).

## Ancestral and novel cell types

To better understand how the observed changes in genome architecture and gene complement have shaped the unusual body of chaetognaths, we generated a single-cell atlas of *P. gotoi* juveniles and adults comprising almost 30,000 cells classified into roughly 30 differentiated cell types (Fig. 2a,b and Extended Data Fig. 5a–d). We used a combination of in situ hybridization, literature surveys and cross-species comparisons to characterize these cell types and found neuronal, epidermal and muscle cell types, ciliary cells, stem cells, chaetal cells, gut cells and eye cells. To determine their evolutionary origin, we compared chaetognath cell types to those of selected representatives of main animal lineages<sup>38–40</sup> (Extended Data Fig. 6). We found multiple cell types that are likely to be derived from the bilaterian ancestor (namely, neurons, muscles, gut, ciliary or germ cells), as supported by the shared expression of multiple orthologous genes (Fig. 2c). However, several cell types appear to be chaetognath-specific, such as their characteristic grasping spines, sensory papillae or ciliary sense organs.

Among ancestral cell types, we identified two neuronal clusters (clusters 11 and 26) that exhibit gene expression of the neuronal marker *Elav*, several synaptotagmins, calcium (for example *Cac1a-2*) and sodium (for example, *Scna-3*) channels and a dopamine-synthesizing enzyme (*Ddc*), and can be distinguished by distinct gene expression of glutamate receptors (*Grik2* and *Grm*) and degrading enzymes (*Glna*) (Fig. 2b). *Cac1a-2* is expressed in both the ventral and cephalic ganglia as well as in some peripheral neurons, an abundant cell type in the chaetognath nervous system<sup>41</sup> (Fig. 2d). These neuronal clusters map well to those of other protostomes (shared expression of the transcription factor genes *Bc11a*, *Cot1*, *Csde1* and *Meis* and the splicing factor gene *Rfox3*; Fig. 2c and Supplementary Table 9). Identified sensory cells comprise a photoreceptor cell-type in the eye (cluster 29) that expresses xenopsin 1 (*Opsd-1*) as described<sup>42</sup>. Notably, *P. gotoi* ciliary cell types express Lophotrochin, a gene that was previously shown to be expressed in ciliary bands of lophotrochozoans larvae and adults, including the corona of cilia of rotifers<sup>39,43</sup>. Two muscle cell types express titin and troponin as well as distinct actin paralogues, corroborating prior expression surveys of alpha actin (*Acts-1*), localized in longitudinal muscles (cluster 12), and Actin 1 (*Act1-5*), in the muscles of the gnatha (cluster 19)<sup>44</sup>. Finally, two putative germline clusters (clusters 16 and 32; Fig. 2b) express stem cell markers, multiple copies of *Vasa* (*Ddx*) previously showed to be present in chaetognath germ cells<sup>45</sup>, other germline markers such as *piwi*, *Maelstrom* and *Elav*, as well as several enzymes controlling DNA methylation (DNMT3.1, DNMT3.4 and TETb). More than a third of the total of cell-type marker genes are specific to these germline clusters (Extended Data Fig. 6), indicating a massive



**Fig. 2 | Single-cell repertoire in the chaetognath *P. gotoi*.** **a**, Uniform manifold approximation and projection (UMAP) dimensionality reduction of cell expression profiles from pooled juvenile and adult libraries (Extended Data Fig. 5) with inferred clusters coloured and labelled. **b**, Dot plot of selected cell-type marker genes for each cluster, with genes selected as markers for in situ hybridization highlighted in blue, and dotted lines separating markers characteristics of broad cell-type categories. **c**, Summary of SAMap comparisons between chaetognath and selected bilaterians. Dot plot and text labels indicate the number of conserved cell-type genes between *P. gotoi* and selected species (full species name in Fig. 1b). **d**, Microphotographs of whole-mount in situ hybridization staining for selected markers in juvenile stages (3 days post-

fertilization (dpf)). All in situ hybridization experiments were carried out on at least three distinct individuals. The second row shows higher magnification of images in the first row. Numbers indicate clusters in **a** that are represented by the labelled cells. *Chs2-5* is indicative of the chaete cells located near the grasping spines. *Vwf-2* is expressed in the epidermal cells, in a subset of which *Wnt1-3* seems to be localized. *Tbx2* and *Fmar-5* are expressed in ciliary organs. *Cac1a-2* is expressed in the main ventral neural ganglion as well as other cephalic cells. *Pxdn-9* marks the papillae sensory cells in the head. A full list of cell-type markers is provided in Supplementary Table 6 and molecular probes are listed in Supplementary Table 7. Scale bars, 50  $\mu$ m.

expression of maternal transcripts during oogenesis, as previously observed in other species<sup>46</sup>. Chaetognath stem cells co-express multiple chromatin modifiers (for example, *Atrx*, *Chd5* and *Dek*) or splicing factors (*Tra2b* and *Srrm1*) as well as the transcription factors *Yboxh-2* with other protostomes (Supplementary Table 9). Overall, we observed a small overlap in transcription factor co-expression among protostome cell types, which agrees with similar comparisons<sup>39,40,47</sup> and was further confirmed by a 1:1 comparison between the earthworm and a flatworm dataset (*Schmidtea mediterranea*; Supplementary Fig. 2).

Chaetognath-specific cell types—that is, those that appear unrelated to any other bilaterian cell types in our cross-species comparison—are associated with phylum-specific sensory and structural roles. Among these, we find two peripheral sensory papillae cell types (expressing neuronal genes such as synaptotagmins and acetylcholine receptors *Acha7-1*) (clusters 15 and 25) positioned around the mouth and interspersed in the epidermis throughout the body (see *Pxdn-9* (ref. 48) in Fig. 2d). We also recover two cell types (clusters 10 and 13) that are associated with ciliary sense organs, mechanoreceptors distributed on the body of the animal that constitute their primary sensory system for swimming and predation, as indicated by expression of *Tbx2* and *Fmar-5* in these structures<sup>48</sup> (Fig. 2b,d). The expression of chitin synthase (*Chs2-5*) near chaetognath grasping spines indicates that these are probably ensuring the growth and synthesis of the grasping spines (cluster 18), whereas the gene *Vwf-2* (clusters 2, 6 and 17) and *Ifea-2* (clusters 8 and 20) show expression in the epidermal layer (Fig. 2d). The presence of multiple epidermal cell types is consistent with the presence of a complex pluristratified epidermis in chaetognaths, a condition rarely found outside of vertebrates<sup>49</sup>. All these structural cell types express various chitin synthase and chitinases, with alternative paralogues present in the ‘chaete’ cells (clusters 18: *Chs2-5*, *Chi10-1*, *Chi10-3* and *Chi10-4*) and epidermal cells (clusters 2, 6 and 17: *Chs2-3*, *Chit1-7* and *Chit1-8*).

To further understand the origin and diversification of chaetognath cell types, we then investigated the evolution of the genes involved in their making.

## Gene turnover and cell-type innovation

Phylogenetic reconciliation indicated that chaetognaths share a large burst of gene duplication, but whether this event results from a whole-genome duplication or not remains unclear<sup>9,10</sup> (Fig. 1b). The magnitude of this event (3,379 families) is comparable in extent to other gene duplication events such as the vertebrate 2R (4,203 families) or the clitellate-specific gene expansion event<sup>34</sup> (Fig. 1b). Functional enrichment in *P. gotoi* duplicated genes highlighted terms related to ion transmembrane transport, as well as developmental genes (Fig. 3a), which is also corroborated with phylogenetic estimation of gene family expansion in chaetognaths (Extended Data Fig. 2c,d). This gene expansion is linked to cell-type novelty, as for instance, *P. gotoi* has eight copies of the piezo gene, a mechanosensory ion channel (with one paralogue specifically expressed in the Papillae cluster 25), which is present as a single copy in limpet or sea urchin<sup>50</sup>, or up to eight copies of *Elav* (expressed in neurons and germline cells), which is present as a single copy in most non-vertebrates (Supplementary Table 3).

To determine whether gene duplications occurred at the same evolutionary time, we examined the distribution of transversions at four-fold degenerate sites (FDTV) in pairwise gene comparison. We observe a single FDTV peak for genes that belong to duplicated subfamilies showing distinct numbers of paralogues, indicating that they probably originated in a single event (Fig. 3c). Duplicated genes mostly occur in pairs, but higher copy numbers make up more than 50% of these duplicates (3 or more copies; Extended Data Fig. 7a). Next, we examined the location of duplicates across the genome. The chromosomal distribution of duplicates does not match the distribution of ancestral linkages across chromosomes as only one copy of bilaterian linkage

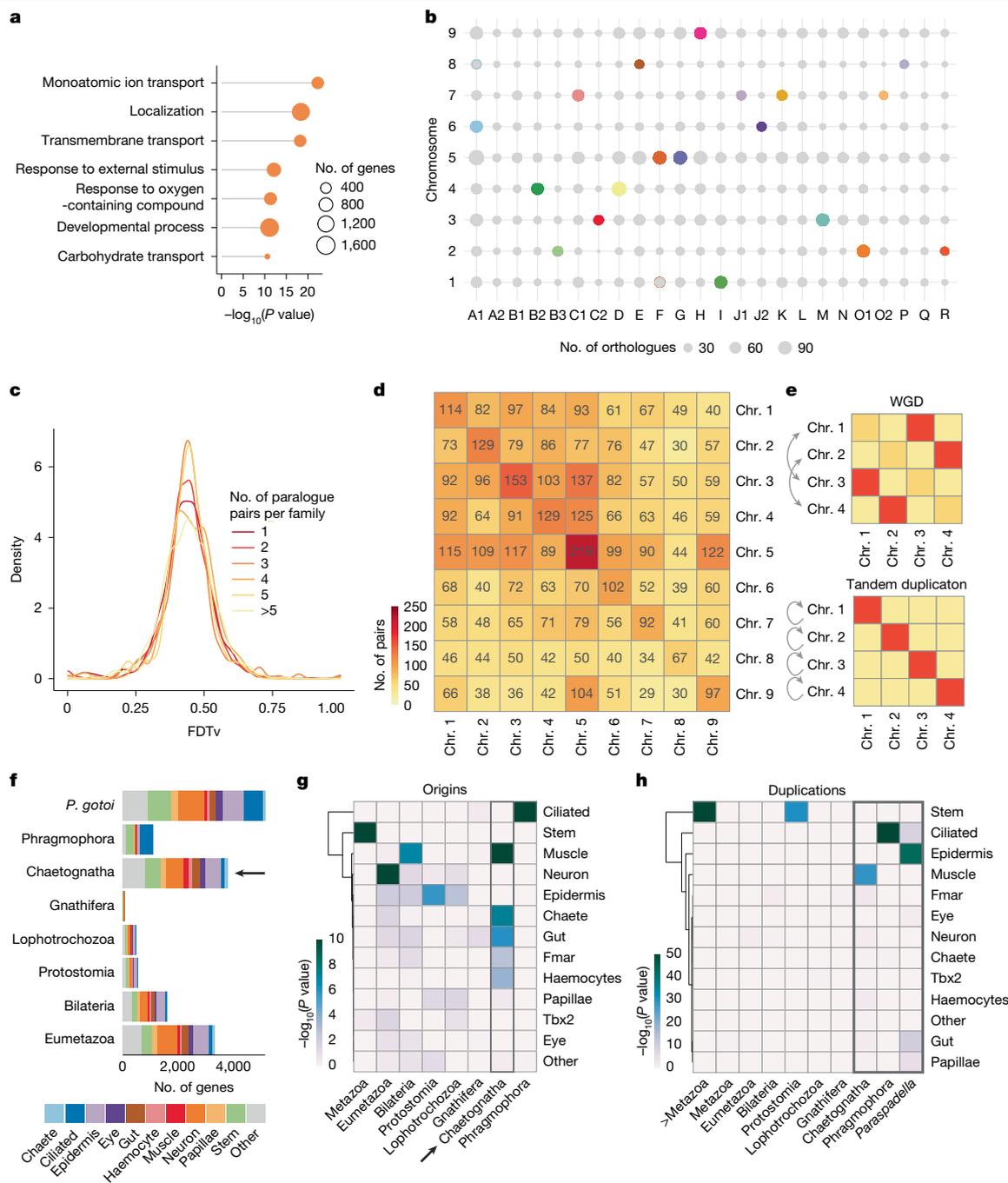
groups is detected in *P. gotoi* chromosomes (Fig. 3b). We also found that about 25% of duplicate pairs are located on the same chromosomes, and that no chromosome pair displays a mutual enrichment in duplicates, as would be expected in a whole-genome duplication event (Fig. 3d,e). The separation between duplicate pairs shows a bimodal distribution, with some pairs approximately 10 kb apart and some around 10 Mb apart (Extended Data Fig. 7b). Similarly, we did not detect fewer introns in members of duplicated gene families, ruling out a burst of retrotransposition (Extended Data Fig. 7c). Together, these observations are more consistent with a burst of tandem duplication than with a single whole-genome duplication event, which is a rather unique condition among animals. The unusual expansion of median and posterior Hox genes of chaetognaths is an example of such tandem expansion (Fig. 4a).

This reshuffling of the chaetognath gene repertoire is not limited to gene duplications. In total, up to 2,250 gene families appear to be specific to chaetognaths (that is, found in at least two of the four chaetognath species considered, but not outside chaetognaths) (Fig. 1a). This is quite a remarkable condition for a bilaterian phylum; by contrast, we detected only 157 and 124 lineage-specific gene families in molluscs or echinoderms, respectively<sup>51</sup> (Fig. 1a). In *P. gotoi*, these families include 1,832 genes, 8.3% of all predicted genes; to ascertain these numbers, we further confirmed that 1,307 of these genes do not have detectable homology outside chaetognaths in a direct phylostratigraphy assignment, and we observed Pfam domains in only 397 of them (Supplementary Table 1). These distinct lines of evidence support that this observation is not a reconstruction artefact such as those pointed out by recent studies<sup>31</sup>. These novel genes appear to have a regulated expression and putative functional role, as 1,471 of them are among cell-type markers in at least a single-cell cluster (Fig. 3f,g and Supplementary Table 1). We did not observe similar bursts of gene novelties in similarly fast-evolving lineages such as platyhelminths or ecdysozoans, which further rules out the possibility of homology detection failure (Fig. 1a).

We found that both novel and duplicated gene populations have a key role in chaetognath cell types, particularly novel ones, as many of them are cell-type-specific (for example, ‘chaete’, *Fmar*, ciliary and epidermis; Fig. 3g,h). We estimated the phylostratigraphic age of gene sets specific to each cell type by leveraging gene phylogenetic age and noticed that ciliary cells appear among the ones with the most recent profile (Extended Data Fig. 8a,b). Notably, such an incorporation of phylum-specific genes among cell-type markers is observed only in chaetognaths when compared to five other investigated species in which most marker genes are either ancient or species-specific (Fig. 3f and Extended Data Fig. 8c). We therefore hypothesize that after extensive gene loss in the gnathiferan lineage, particularly that affecting gene families involved in sensory and nervous functions, the evolution of cell types in chaetognaths relied on newly evolved genes as well as lineage-specific tandemly duplicated genes. This reshuffling of the gene repertoire therefore probably had an instrumental role in the establishment of their distinctive body plan of chaetognaths.

## Expansion of the chaetognath Hox cluster

As a primary example of tandem duplication, we found that *P. gotoi* possesses an increased complement of 14 Hox genes<sup>19</sup>, related to an expansion of both median (7 genes, 3 copies of the PG6/8 paralogue group) and posterior Hox genes (5 genes), an exceptional situation outside vertebrates (Fig. 4a and Supplementary Fig. 3). The Hox cluster of *P. gotoi* is unusually large, spanning 2.4 Mb (Extended Data Fig. 9a), with genes arranged in the expected ‘colinear’ order, and also includes 15 interspersed non-Hox genes (Extended Data Fig. 9a). Remarkably, the *MedPost* gene, originally described as an oddity<sup>19</sup> and later as a synapomorphy of Gnathifera<sup>18</sup>, is consistently located in the cluster between median and bona fide posterior Hox genes, as originally predicted; its



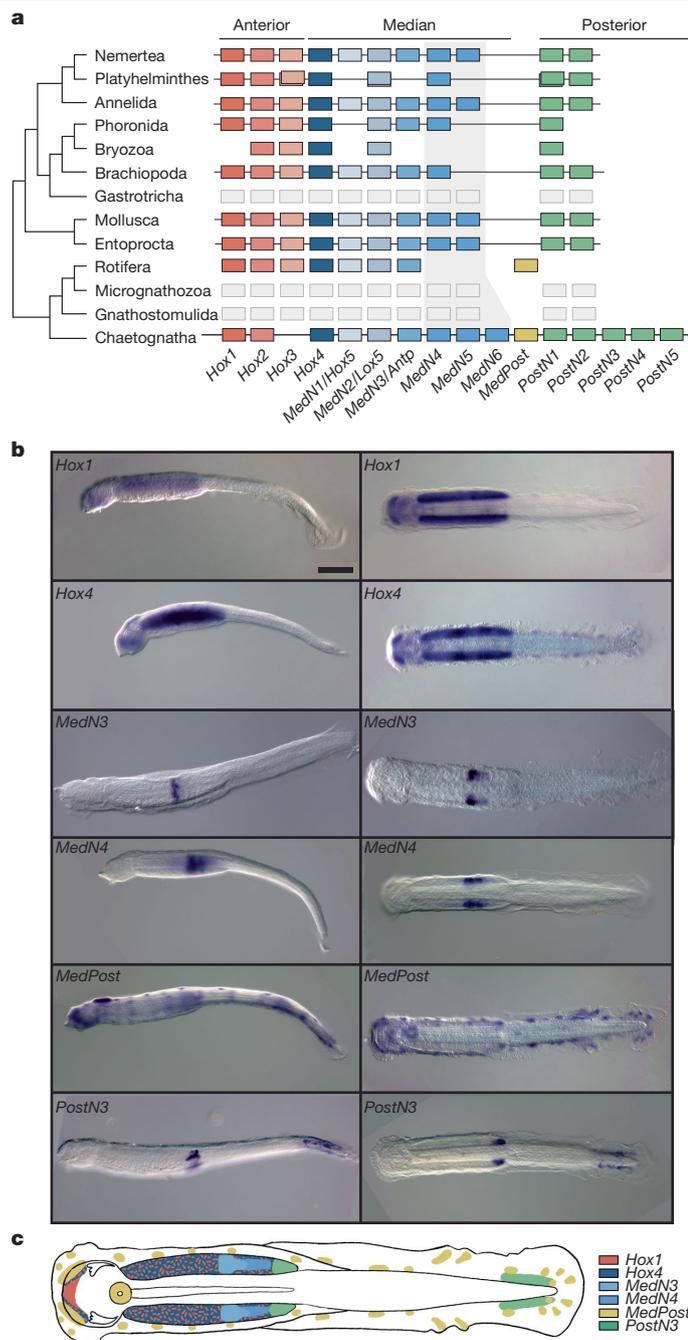
**Fig. 3 | Gene duplication and gene novelty in chaetognaths. a**, Overrepresented gene ontology terms in chaetognath duplicated genes (Fisher’s exact test, two-sided). **b**, Distribution of ancestral BLGs in chaetognath chromosomes. Each bubble represents the number of shared orthologues between a pair of chromosomes, which is shown in colour (Fig. 1e) if the enrichment is significant (Fisher test) and is otherwise shown in grey. **c**, Distribution of synonymous transversions at fourfold degenerate sites (FDTVs) between pairs of paralogues in sets of duplicates showing distinct numbers of chaetognath-specific paralogues. **d**, Number of paralogue pairs shared across chaetognath

chromosomes indicate the absence of pairwise chromosome clustering. Each duplicated gene was only counted in paralogue pair to avoid giving excess weight to large paralogous families, thus the matrix is asymmetrical. **e**, Expected patterns of paralogue pairing in chromosomes in case of whole-genome duplication (WGD; top) or tandem duplication (bottom). **f**, Phylostratigraphic distribution of cell-type marker genes for *P. gotoi* (arrow). **g, h**, Hypergeometric enrichment of genes that originated (**g**) or duplicated (**h**) at successive phylostrata in broad cell types of chaetognaths. Box and arrow highlight recently evolved and duplicated genes, respectively.

location could not be corroborated in rotifers in which the Hox cluster is dispersed<sup>25</sup>.

Hox genes are expressed in several cell types, mainly epidermal, neuronal and papillae cell types (Extended Data Fig. 9b). In situ hybridization on six Hox genes in *P. gotoi* juveniles (Fig. 4b, c and Supplementary Table 7) indicates that whereas anterior Hox genes show a broad

expression in the chaetognath nervous system (paired ventral and cephalic ganglions for *Hox1* and *Hox4*), median and posterior Hox genes show a more classical staggered expression in the ventral ganglion that appears coherent with the paralogy group (*MedN3* is more anterior than *MedN4*, and *PostN4* in the posterior-most part of the ganglion)<sup>19</sup> (Extended Data Fig. 9c). *PostN4* presents a dual domain, both in the



**Fig. 4 | Hox genes and their expression in *P. gotoi*.** **a**, Distribution of Hox genes in lophotrochozoans assuming the phylogeny of ref. 5, highlighting the expansion of different paralogy groups in distinct spiralian lineages. Anterior genes are coloured in shades of red and pink, median genes are in blue, medpost genes are in yellow and posterior genes are in green; each shade represent a paralogy group. Grey boxes indicate an unknown gene complement. A horizontal line crossing Hox genes indicates evidence for physical linkage. **b**, Microphotographs of whole-mount in situ hybridization for 6 Hox genes in lateral (left) and apical (right) view of *P. gotoi* 3 dpf juveniles. All in situ hybridization experiments were carried out on at least three distinct individuals. *Hox1* and *Hox4* are expressed in paired ventral and cephalic ganglions, whereas *MedN3*, *MedN4* and *PostN3* show a staggered expression in the nervous system. *MedPost* appears to be expressed in an array of sensory cells that are interspersed in the epidermis. **c**, Schematic of a juvenile *P. gotoi* with overlaid Hox expression domains. Scale bar, 100  $\mu$ m.

posterior-most part of the ventral ganglion and in the postanal region of the juvenile (Fig. 4b). Surprisingly, *MedPost* is localized in papillae peripheral sensory neurons interspersed in the epidermis (Fig. 4b and Extended Data Fig. 9b). The expression of *MedPost* in sensory neurons throughout the body is different from that observed in rotifers, where it is expressed specifically in the postanal components of the nervous system (Extended Data Fig. 9c). In chaetognaths, posterior Hox genes are expressed in this postanal region; these genes have been lost in rotifers<sup>18</sup>. The cell-type-specific expression of *MedPost* suggests that acquisition of new expression domains could have participated in the emergence of novelties at the body plan or cell-type level.

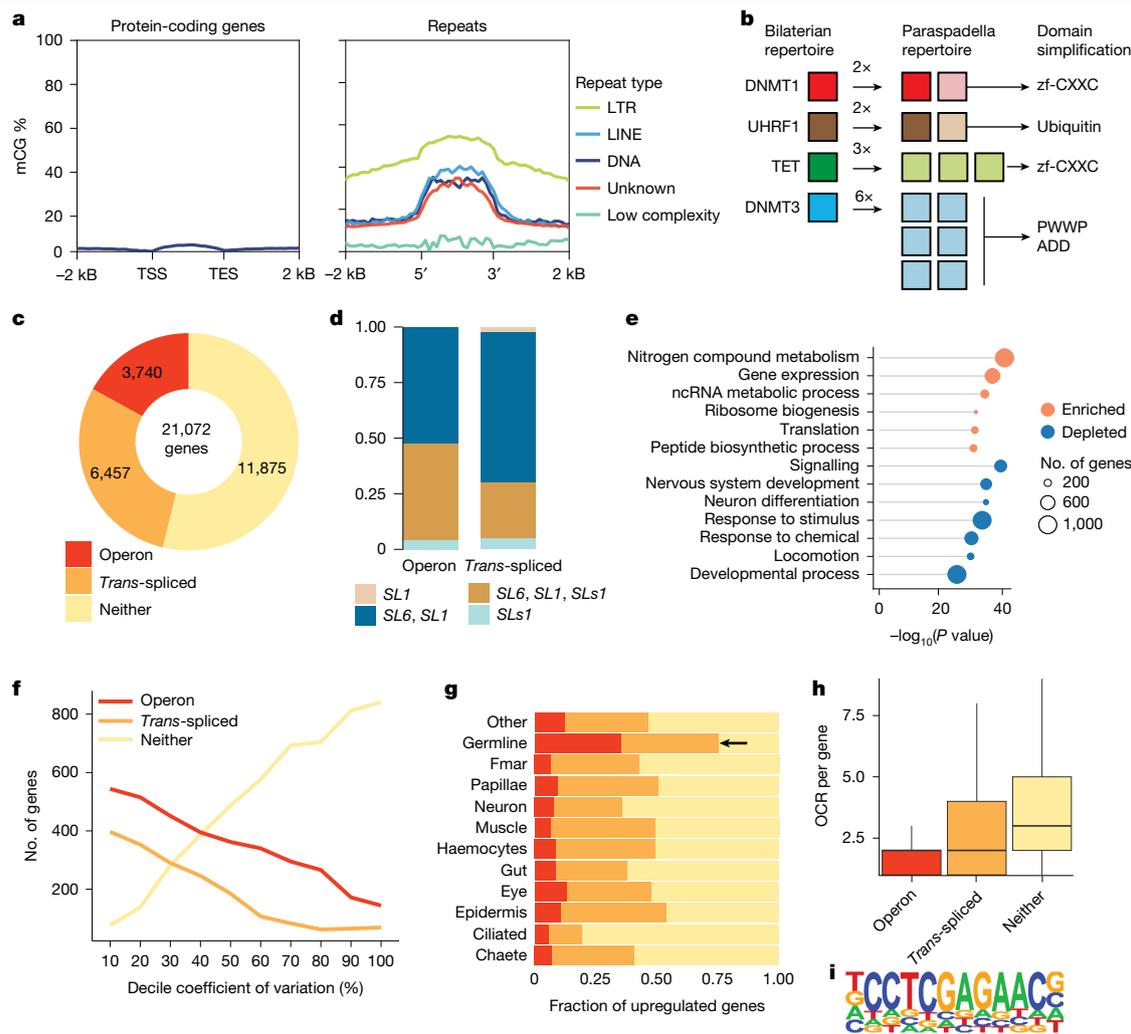
## Retargeting of DNA methylation

We set out to understand whether the reshuffling of chaetognath gene complement had an effect on their gene regulation<sup>52</sup>. We found that *P. gotoi* represents an outlier in terms of DNA methylation among animals. Whereas most non-vertebrate lineages have highest levels of DNA methylation on gene bodies, usually associated with robustly expressed genes<sup>53</sup>, the vast majority of *P. gotoi* genes are not methylated (Fig. 5a and Supplementary Table 10), and the few that are contain protein domains that are typical of retrotransposons and have depleted chromatin accessibility, suggesting that they are silent genes with transposon origins (Extended Data Fig. 10a,b). Moreover, whereas transposable elements in most invertebrates are rarely directly methylated, *P. gotoi* shows a clear targeting of DNA methylation on transposons—most acutely on LTR retrotransposons, but also on other classes—but none on low-complexity repeats that are less likely to be parasitic (Fig. 5a and Extended Data Fig. 10c). This pattern of exclusive transposable element methylation and absence of gene body methylation represents a pattern reminiscent of fungi and other unicellular eukaryotes, and has been observed in only few animal lineages, including early diverging nematodes and ctenophores<sup>54,55</sup>.

To explore how *P. gotoi* has modified its methylome compared to other animals, we examined the genetic toolkit responsible for DNA methylation in this species. We found that *DNMT1* and its partner *UHRF1* underwent gene duplications, with their paralogues undergoing probable sub-functionalization through the loss of ancestral protein domains such as zinc-finger CXXC in *DNMT1* or the ubiquitin domain in *UHRF1* (Fig. 5b and Extended Data Fig. 10d). Similarly, none of the three copies of *TET* (including a retrocopy) encode the ancestral zinc-finger CXXC domain combination that is probably shared by most spirilians<sup>56</sup>. Finally, six copies of *DNMT3* are present in chaetognaths, which all lost the ADD and PWWP domains; the latter drives DNA methylation to gene bodies by binding to H3K36me3 in other animals, a histone post translational modification that is typical of gene bodies<sup>57</sup>. Most DNA methylation genes are expressed in germ line cell types similarly to what has been observed in mice<sup>58</sup> (Extended Data Fig. 10e). In sum, *P. gotoi* represents an independent example of DNA methylation retargeting from gene bodies to transposable elements, associated with protein domain simplification of the DNA methylation toolkit.

## Trans-splicing and operons

Many genes in chaetognaths have been reported to undergo *trans*-splicing, a post-transcriptional spliceosomal modification that is associated with operonic transcription and resolves polycistronic transcripts<sup>10,59</sup>. At the genome scale, we found that *trans*-splicing is prevalent, occurring in nearly half the genes (10,197 out of 21,072 predicted genes), and 18% of all genes are organized in operons resolved by *trans*-splicing (Fig. 5c). *Trans*-splicing involves spliceosomal addition of splice leaders to messenger RNAs, with the splice leader sequence derived from genes often located in the ribosomal DNA clusters far from their target genes<sup>60</sup>. We identified three distinct splice leaders in *P. gotoi*, of which one (SLs1) is shared with *Spadella cephaloptera*<sup>10</sup>.



**Fig. 5 | Gene regulation, DNA methylation and *trans*-splicing.** **a**, DNA methylation in *P. gotoi* has reverted to exclusive transposon targeting. The metaplots show average methylation levels at CpG dinucleotides (mCG%) at protein-coding genes and repeats, classified for major transposable element type. Only repeats larger than 500 bp were used. TES, transcriptional end site. **b**, Diagram showing the expansion and simplification of the ancestral bilaterian DNA methylation toolkit in *P. gotoi*. Paralogues with simplified domain architectures are coloured in a lighter shade than those that retained the ancestral state. Full domain architectures of each gene are shown in Extended Data Fig. 10d. *zf*, zinc finger. **c**, Number of genes receiving *trans*-splicing and classified as belonging to an operon in *P. gotoi*. **d**, Occurrence of different annotated splice leaders and their combination in *trans*-splicing and operonic

genes. **e**, Gene ontology enrichment (orange) and depletion (blue) for genes belonging to operons (Fisher's exact test, two-sided). ncRNA, noncoding RNA. **f**, *Trans*-spliced and operonic genes exhibit a lower expression specificity as measured by coefficient of variation calculated using cell-type averaged expression. **g**, Fraction of operonic, *trans*-spliced and other genes among marker genes in single-cell clusters with enrichment in germline marked by an arrow. **h**, Number of open chromatin regions associated with the different gene categories (box plot shows median, quartiles and extrema,  $n = 2,643, 5,586$  and  $8,851$  genes for operons, *trans*-spliced genes and other genes, respectively), indicating the higher number of putative regulatory elements in genes that are not *trans*-spliced. **i**, Example of an enriched motif in open chromatin regions related to operons.

*Trans*-splicing and operonic transcription have a patchy distributions across animals, which suggests independent evolution in distinct lineages (nematodes, tunicates and rotifers)<sup>10,60</sup>. However, despite its proposed role in recovery from growth arrest<sup>59</sup> or germline maintenance<sup>61</sup>, the evolution role of *trans*-splicing remains debated. Here, multiple splice leaders are found among the various transcripts of a given gene, and contrary to nematodes<sup>60</sup>, we did not identify operon-specific splice leaders (Fig. 5d).

Functionally, operonic genes are preferentially related to nitrogen and protein metabolism and they are negatively associated with development, nervous system and signalling, suggesting that finely regulated genes are excluded from this process (Fig. 5e). For instance, the longest operon gathers nine genes, including genes encoding ribosomal proteins, ubiquitin and proteasome subunits (Supplementary Table 1). *Trans*-spliced genes have a higher expression level (Extended

Data Fig. 7f) and have, in general, a lower expression specificity at the cell-type level than those not subjected to splice leader addition (Fig. 5f); however, operonic and *trans*-spliced genes make up most of cell-type marker genes in the germline clusters (Fig. 5g). *Trans*-spliced genes are also not restricted to a phylogenetic origin and belong to all phylostrata (Extended Data Fig. 7g). Consistently, *trans*-spliced genes have a smaller number of putative regulatory elements (for example, as indicated by ATAC-seq peaks), in agreement with their lower regulatory and expression complexity (Fig. 5h). The presence of a large fraction of *trans*-spliced genes could have contributed to the overall compact regulatory landscape of *P. gotoi* compared to other lineages, with open chromatin regions located closer to the transcription start site (TSS)<sup>721</sup> (Extended Data Fig. 7h and Supplementary Table 11). However, we identified conserved distal regulatory regions and transcription factor binding properties with that of other animals (Extended

Data Fig. 7h–k). Interestingly, this also suggests that non operonic *trans*-spliced genes share these functional properties with operon genes rather than non-*trans*-spliced ones. We also detected enriched motifs associated with *trans*-spliced genes (for example, Fig. 5i), which do not correspond to transcription factors characterized in other lineages, corroborating the idea of independent origin of *trans*-splicing and operonic transcription in metazoans<sup>10,60</sup>.

## Conclusion

Our results show that chaetognaths stand out from other animal lineages at the molecular and genomic levels, and provide clues to explain why their unique body plan appears so distinctive from other animal phyla<sup>2</sup>. Chaetognaths have undergone an extensive reshuffling of their gene complement as well as major fusions of ancestral chromosomal linkages (Fig. 1). This extensive duplication of genes does not appear to be related to a single whole-genome duplication event, but rather a burst of tandem duplications (Fig. 3). In contrast to the microscopic rotifers and other gnathiferans, large-bodied chaetognaths have an active predatory lifestyle with an array of sensory receptors, such as ciliary organs reminiscent of the lateral line<sup>4</sup>, that are associated with innovation and diversification at the cell-type level (papillar, Fmar, ciliated and epidermal).

The gnathiferan lineage underwent accelerated genomic evolution, either as a single event or as multiple lineage-specific events that could plausibly have been coupled with morphological and body size reduction from their ancestors. In this view, chaetognaths would have secondarily evolved complex organ systems by mobilizing lineage-specific duplicated and novel genes instead of mainly relying on genes from the ancestral bilaterian toolkit, which had been decimated in this lineage, unlike what is observed in other animals (Fig. 3g,h and Extended Data Fig. 7). Of note, rotifers also acquired new genes, albeit through the distinct process of horizontal gene transfer, that were involved in the evolution of novel traits such as resistance to desiccation or ionizing radiation<sup>62</sup>. The paucity of Cambrian fossils for other gnathiferans such as rotifers or gnathostomulids, and their relatively large body size (for example, *Amiskwia*<sup>12</sup>) does not provide palaeontological support for a small-bodied gnathiferan ancestor. Further characterization of genomes and gene expression in other gnathiferan lineages (such as micrognathozoans or gnathostomulids) would help to further resolve organismal and genomic evolution in the gnathiferan lineage.

Chaetognaths also shed some light on the relationship between gene regulation and gene and genome evolution. The reshuffling of the gene repertoire in chaetognaths coincides with the evolution of novel strategies of gene regulation, such as the acquisition of *trans*-splicing<sup>10</sup>. Consistent with a previously proposed model<sup>61</sup>, *trans*-spliced genes are particularly enriched in the germinal cells and show limited regulation. *Trans*-splicing may have a role in stabilizing housekeeping gene expression when gene body methylation is lost, suggesting a recurrent evolutionary process. Of note, *trans*-splicing is present in ctenophores, chaetognaths and nematodes<sup>53,54</sup>, the three animal phyla in which gene body methylation has been repurposed for transposable element targeting. However, other evolutionary forces may have a role in the loss of gene body methylation, such as the DNA repair burden imposed by DNA methyltransferases<sup>54</sup>. Notably, lineages with *trans*-splicing, such as nematodes, tunicates, rotifers, platyhelminths and potentially ctenophores, are also prone to loss of ancestral linkage<sup>8,60</sup>. The absence of detectable sequence-insulated chromatin domains in chaetognaths, which have recently been shown to be a bilaterian innovation<sup>55</sup>, would require further confirmation (for example, using micro-C<sup>55</sup>), but could suggest that such structures are less widespread than anticipated.

Our analysis of the genome, cell types and gene regulation provides clues to the origin of one of the most enigmatic bilaterian lineages, and a molecular explanation for its derived state and distinctive characteristics.

## 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/s41586-025-09403-2>.

1. Erwin, D. H. The origin of animal body plans: a view from fossil evidence and the regulatory genome. *Development* **147**, dev182899 (2020).
2. Bone, Q., Kapp, H. & Pierrot-Bults, A. C. *The Biology of Chaetognaths* (Oxford Univ. Press, 1991).
3. Rieger, V. et al. Immunohistochemical analysis and 3D reconstruction of the cephalic nervous system in Chaetognatha: insights into the evolution of an early bilaterian brain? *Invertebr. Biol.* **129**, 77–104 (2010).
4. Müller, C. H. G., Rieger, V., Perez, Y. & Harzsch, S. Immunohistochemical and ultrastructural studies on ciliary sense organs of arrow worms (Chaetognatha). *Zoomorphology* **133**, 167–189 (2014).
5. 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.e3 (2019).
6. Laumer, C. E. et al. Revisiting metazoan phylogeny with genomic sampling of all phyla. *Proc. Biol. Sci.* **286**, 20190831 (2019).
7. Martin-Zamora, F. M. et al. Annelid functional genomics reveal the origins of bilaterian life cycles. *Nature* **615**, 105–110 (2023).
8. Simakov, O. et al. Deeply conserved synteny and the evolution of metazoan chromosomes. *Sci. Adv.* **8**, eabi5884 (2022).
9. Telford, M. J. & Holland, P. W. Evolution of 28S ribosomal DNA in chaetognaths: duplicate genes and molecular phylogeny. *J. Mol. Evol.* **44**, 135–144 (1997).
10. Marlétaz, F. et al. Chaetognath transcriptome reveals ancestral and unique features among bilaterians. *Genome Biol.* **9**, R94 (2008).
11. Park, T.-Y. S. et al. A giant stem-group chaetognath. *Sci. Adv.* **10**, eadi6678 (2024).
12. Vinther, J. & Parry, L. A. Bilateral jaw elements in *Amiskwia sagittiformis* bridge the morphological gap between Gnathiferans and Chaetognaths. *Curr. Biol.* **29**, 881–888.e1 (2019).
13. Satoh, N. *Chordate Origins and Evolution* (Elsevier, 2016).
14. Budd, G. E. & Telford, M. J. The origin and evolution of arthropods. *Nature* **457**, 812–817 (2009).
15. Chen, H. et al. A Cambrian crown annelid reconciles phylogenomics and the fossil record. *Nature* **583**, 249–252 (2020).
16. John, C. C. Memoirs: habits, structure, and development of *Spadella cephaloptera*. *Q. J. Microsc. Sci.* **75**, 625–696 (1933).
17. Telford, M. J. & Holland, P. W. The phylogenetic affinities of the chaetognaths: a molecular analysis. *Mol. Biol. Evol.* **10**, 660–676 (1993).
18. Fröblius, A. C. & Funch, P. Rotiferan Hox genes give new insights into the evolution of metazoan bodyplans. *Nat. Commun.* **8**, 9 (2017).
19. Papillon, D., Perez, Y., Fasano, L., Le Parco, Y. & Caubit, X. Hox gene survey in the chaetognath *Spadella cephaloptera*: evolutionary implications. *Dev. Genes Evol.* **213**, 142–148 (2003).
20. Bekkouche, N. & Gąsiorowski, L. Careful amendment of morphological data sets improves phylogenetic frameworks: re-evaluating placement of the fossil *Amiskwia sagittiformis*. *J. Syst. Palaeontol.* **20**, 1–14 (2022).
21. Marlétaz, F. et al. Amphioxus functional genomics and the origins of vertebrate gene regulation. *Nature* **564**, 64–70 (2018).
22. Parey, E. et al. The brittle star genome illuminates the genetic basis of animal appendage regeneration. *Nat. Ecol. Evol.* **8**, 1505–1521 (2024).
23. Simakov, O. et al. Insights into bilaterian evolution from three spiralian genomes. *Nature* **493**, 526–531 (2013).
24. Luo, Y.-J. et al. The Lingula genome provides insights into brachiopod evolution and the origin of phosphate biomineralization. *Nat. Commun.* **6**, 8301 (2015).
25. Simion, P. et al. Chromosome-level genome assembly reveals homologous chromosomes and recombination in asexual rotifer *Adineta vaga*. *Sci. Adv.* **7**, eabg4216 (2021).
26. Flot, J.-F. et al. Genomic evidence for asexual evolution in the bdelloid rotifer *Adineta vaga*. *Nature* **500**, 453–457 (2013).
27. Arendt, D. et al. The origin and evolution of cell types. *Nat. Rev. Genet.* **17**, 744–757 (2016).
28. Goto, T. & Yoshida, M. The mating sequence of the benthic arrowworm *Spadella schizoptera*. *Biol. Bull.* **169**, 328–333 (1985).
29. Ren-feng, W. Analysis of chromosome karyotypes in Chaetognath *Sagitta crassa*. *J. Dalian Fish. Univ.* **26**, 260–263 (2011).
30. Lewin, T. D. et al. Fusion, fission, and scrambling of the bilaterian genome in Bryozoa. *Genome Res.* **35**, 78–92 (2025).
31. Guijarro-Clarke, C., Holland, P. W. H. & Paps, J. Widespread patterns of gene loss in the evolution of the animal kingdom. *Nat. Ecol. Evol.* **4**, 519–523 (2020).
32. Senaratne, A. P. et al. Formation of the CenH3-deficient holocentromere in Lepidoptera avoids active chromatin. *Curr. Biol.* **31**, 173–181.e7 (2021).
33. Hofstätter, P. G. et al. Repeat-based holocentromeres influence genome architecture and karyotype evolution. *Cell* **185**, 3153–3168.e18 (2022).
34. Lewin, T. D., Liao, I. J.-Y. & Luo, Y.-J. Annelid comparative genomics and the evolution of massive lineage-specific genome rearrangement in bilaterians. *Mol. Biol. Evol.* **41**, msae172 (2024).
35. Muller, H., Gil, J. Jr & Drinnenberg, I. A. The impact of centromeres on spatial genome architecture. *Trends Genet.* **35**, 565–578 (2019).
36. Houtain, A. et al. Transgenerational chromosome repair in the asexual bdelloid rotifer *Adineta vaga*. Preprint at *bioRxiv* <https://doi.org/10.1101/2024.01.25.577190> (2024).

37. Lieberman-Aiden, E. et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289–293 (2009).
38. Álvarez-Campos, P. et al. Annelid adult cell type diversity and their pluripotent cellular origins. *Nat. Commun.* **15**, 3194 (2024).
39. Piovani, L. et al. Single-cell atlases of two lophotrochozoan larvae highlight their complex evolutionary histories. *Sci. Adv.* **9**, eadg6034 (2023).
40. Li, J. et al. Deep learning of cross-species single-cell landscapes identifies conserved regulatory programs underlying cell types. *Nat. Genet.* **54**, 1711–1720 (2022).
41. Rieger, V. et al. Development of the nervous system in hatchlings of *Spadella cephaloptera* (Chaetognatha), and implications for nervous system evolution in Bilateria. *Dev. Growth Differ.* **53**, 740–759 (2011).
42. Wollesen, T., Rodríguez Monje, S. V., Oel, A. P. & Arendt, D. Characterization of eyes, photoreceptors, and opsins in developmental stages of the arrow worm *Spadella cephaloptera* (Chaetognatha). *J. Exp. Zool. B* **340**, 342–353 (2023).
43. Wu, L. et al. Genes with spiralian-specific protein motifs are expressed in spiralian ciliary bands. *Nat. Commun.* **11**, 4171 (2020).
44. Yasuda, E., Goto, T., Makabe, K. W. & Satoh, N. Expression of actin genes in the arrow worm *Paraspadella gotoi* (Chaetognatha). *Zoolog. Sci.* **14**, 953–960 (1997).
45. Carré, D., Djediat, C. & Sardet, C. Formation of a large Vasa-positive germ granule and its inheritance by germ cells in the enigmatic Chaetognaths. *Development* **129**, 661–670 (2002).
46. Piovani, L. & Marlétaz, F. Single-cell transcriptomics refuels the exploration of spiralian biology. *Brief. Funct. Genomics* **22**, 517–524 (2023).
47. Sebé-Pedrós, A. et al. Early metazoan cell type diversity and the evolution of multicellular gene regulation. *Nat. Ecol. Evol.* **2**, 1176–1188 (2018).
48. Goto, T. & Yoshida, M. in *Nervous Systems in Invertebrates* (ed. Ali, M. A.) 461–481 (Springer, 1987).
49. Ahnelt, P. Chaetognatha. in *Biology of the Integument: Invertebrates* (eds. Bereiter-Hahn, J., Matoltsy, A. G. & Richards, K. S.) 746–755 (Springer, 1984).
50. Valencia-Montoya, W. A., Pierce, N. E. & Bellono, N. W. Evolution of sensory receptors. *Annu. Rev. Cell Dev. Biol.* **40**, 353–379 (2024).
51. Vakirlis, N., Carvunis, A.-R. & McLysaght, A. Synteny-based analyses indicate that sequence divergence is not the main source of orphan genes. *eLife* **9**, e53500 (2020).
52. Maeso, I., Acemel, R. D. & Gómez-Skarmeta, J. L. Cis-regulatory landscapes in development and evolution. *Curr. Opin. Genet. Dev.* **43**, 17–22 (2017).
53. de Mendoza, A. et al. Convergent evolution of a vertebrate-like methylome in a marine sponge. *Nat. Ecol. Evol.* **3**, 1464–1473 (2019).
54. Rošić, S. et al. Evolutionary analysis indicates that DNA alkylation damage is a byproduct of cytosine DNA methyltransferase activity. *Nat. Genet.* **50**, 452–459 (2018).
55. Kim, I. V. et al. Chromatin loops are an ancestral hallmark of the animal regulatory genome. *Nature* **642**, 1097–1105 (2025).
56. Guynes, K. et al. Annelid methylomes reveal ancestral developmental and aging-associated epigenetic erosion across Bilateria. *Genome Biol.* **25**, 204 (2024).
57. Schwaiger, M. et al. Evolutionary conservation of the eumetazoan gene regulatory landscape. *Genome Res.* **24**, 639–650 (2014).
58. Barau, J. et al. The DNA methyltransferase DNMT3C protects male germ cells from transposon activity. *Science* **354**, 909–912 (2016).
59. Zaslaver, A., Baugh, L. R. & Sternberg, P. W. Metazoan operons accelerate recovery from growth-arrested states. *Cell* **145**, 981–992 (2011).
60. Douris, V., Telford, M. J. & Averof, M. Evidence for multiple independent origins of trans-splicing in Metazoa. *Mol. Biol. Evol.* **27**, 684–693 (2010).
61. Danks, G. B. et al. Trans-splicing and operons in metazoans: translational control in maternally regulated development and recovery from growth arrest. *Mol. Biol. Evol.* **32**, 585–599 (2015).
62. Wilson, C. G., Pieszko, T., Nowell, R. W. & Barraclough, T. G. Recombination in bdelloid rotifer genomes: asexuality, transfer and stress. *Trends Genet.* **40**, 422–436 (2024).
63. 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).

**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) 2025

## Methods

### DNA isolation and sequencing

An inbred line was generated for five generations by crossing the progeny of an adult *P. gotoi* collected in the tidal zone near Amakusa (Kyushu prefecture, Japan). To this aim, juvenile chaetognaths were fed with increasingly older nauplii of the copepod *Tigriopus japonicus* to match their size until capable of feeding on adults according to previously established culture protocol<sup>64</sup>. DNA was extracted from about 20 adults using phenol-chloroform<sup>65</sup>. Illumina paired-end (25.9 Gb, about 100×), mate-pair (4 libraries: 3, 5, 7 and 10 kb insert) and Pacbio long-reads (5 Gb, about 20×) were sequenced from libraries prepared with this material. A proximity ligation library was constructed from about 20 adults collected at the same locality near Amakusa. In brief, animals were crosslinked in 1% paraformaldehyde, and chromatin was subsequently extracted, immobilized on solid-phase reversible immobilization beads, washed and digested with DpnII as described<sup>66</sup>. After end-labelling, proximity ligation was carried out using T4 DNA ligase and cross-linking reversed using Proteinase K, removed from the beads and the DNA fragments were purified again on SPRI beads. Sequencing library was constructed using the NEB Ultra library preparation kit (New England Biolabs) and sequenced on a NovaSeq6000 instrument.

### Genome assembly

The genome was assembled using meraculous (v.2.2.2.5) with a *k*-mer size of 41 and diploid mode set to '1' to enable haplotypes merging<sup>67</sup>. Primary contigs were screened for contamination using Blobtools (v.1.0.1) and three contaminant contigs were removed at this stage. Pacbio long-reads were then incorporated in the assembly using PBjelly with blasr alignment performed with the parameters '-minMatch 8 -sdpTupleSize 8 -minPctIdentity 75 -bestn 1 -nCandidates 10 -maxScore -500'. The resulting assembly was scaffolded using HiRise with the Hi-C data<sup>68</sup> as previously described<sup>69</sup>. The assembly was evaluated with BUSCO (v.5.4.1) with a final score of C:94.3% [S:91.3%, D:3.0%], F:1.3%, M:4.4% (n:954, metazoa\_odb10). Examination of coverage and *k*-mer distribution using Merqury (v.1.3)<sup>70</sup> and BUSCO statistics<sup>71</sup> indicate that a minimal fraction of residual haplotypes has been retained (Extended Data Fig. 1). This is possibly due to the loss of polymorphisms during the inbreeding procedure.

### Annotation

Illumina RNA-seq and Pacbio Isoseq reads were generated using RNA extracted from pooled adult individuals. RNA-seq was aligned to the genome using STAR (v.2.5.2b), assembled using stringtie (v.1.3.3b) and also assembled as de novo transcripts using Trinity (v.2.5.1)<sup>72</sup>. For Isoseq reads, full length transcripts were obtained after circular consensus, detection of SMART adaptors, and polishing, as recommended by the manufacturer (<https://github.com/PacificBiosciences/IsoSeq>). These transcripts, as well as the trinity contigs, were aligned to the genome using GMAP (v.2017-09-05) and processed according to the Mikado pipeline (v.1.2.1) that leverages distinct transcriptomes, detects putative gene fusions by similarity search against Swissprot, enforces adequate splice-junctions from a curated set inferred using Portcullis (v.1.0.2) and inferred coding sequence using Trans-decoder<sup>73</sup>. High-quality transcripts derived from Mikado were used as a training set to generate a species-specific profile in the Augustus de novo prediction tool<sup>74</sup>. Intron and exon positions were subsequently used as hints to perform a gene prediction with Augustus, together with proteins of *Lottia gigantea* aligned using Exonerate (v.2.2.0). Gene models were further refined using PASA adding untranslated region (UTR) regions and isoforms by relying on the Mikado transcriptomes. We obtained a total of 22,082 genes (Supplementary Table 1). Repeat families were reconstructed using RepeatModeller (2.0.1) and subsequently used to annotate repetitive regions and compute repeat divergence in the genome with RepeatMasker (4.1.0) (Extended Data Fig. 1b). Rideogram was used for karyotype representation of gene and transposable elements contents

(Fig. 1c). Functional annotation was performed using Egnog-mapper (v.2.1.5)<sup>75</sup> and Gene Ontology (GO) enrichments were performed using the goatools Python package<sup>76</sup>. For GO enrichment associated with family losses, we assigned non-redundant GO terms for each family from the emapper annotation obtained for selected species specified in Supplementary Table 2 and computed enrichment at the family level.

### Gene family and synteny analyses

Gene families were first reconstructed using Broccoli (v.1.1)<sup>77</sup> using the proteomes from the species listed in Supplementary Table 2. We constructed sequence alignments for all families including more than 6 genes, more than 3 species and fewer than 400 sequences in total using MAFFT (v.7.471)<sup>78</sup> filtered with CLIPKIT (v.1.1.6, -m gappy)<sup>79</sup> and an initial tree reconstructed with IQ-TREE (v.2.1.1) assuming the LG + R model<sup>80</sup>. These alignments and trees were then used as input for GeneRax (v.2.1.3)<sup>63</sup> with duplication and loss model (UndatedDL) and a per family rate. Reconciled trees in XML format were subsequently parsed to infer gene loss and gene duplication events for each node of the species tree (Fig. 1a and Extended Data Fig. 2b). Gene gains and losses were inferred following parsimony principles and using the ete3 Python library to process phylogenetic trees. For relevant lost families, we assessed whether the best blastp match in relevant species belonged to another gene family to disclose potential gene fragmentation related to evolutionary acceleration. Gene absence was also corroborated by tblastn against the genome of *P. gotoi*. For complementary phylostratigraphic analyses, we used GenEra (v.1.0.2)<sup>81</sup> with default options using the NCBI NR database. To compute FDTVs, we performed a self-alignment of the chaetognath proteome using a MMSeqs2 (r12-113e3)<sup>82</sup>, and then reverse translated the alignments and computed the corresponding substitutions using custom code. For the CAFé analysis, a dated tree was obtained by extracted phylogenetic markers described in ref. 5 for the selected proteomes included in the analyses. An ultrametric dated tree was reconstructed using Phylobayes (v.4.1e) assuming a CAT + GTR model, a CIR relaxed clock model, birth-death priors on divergence times (0.5) and the following calibrations<sup>83–85</sup> (lower–upper bounds in million years): euarthropods (531–514), ecdysozoans (636–519), Gnathifera (636–518.5), bilaterians (636–521) and deuterostomes (635–516). CAFÉ (v.5.1) was then applied using gene families previously defined from Broccoli assuming a gamma model with one rate category (*k* = 1) and no error model<sup>86</sup>.

For synteny comparisons, orthologous pairs were inferred by mutual best hit using MMSeqs2 (r12-113e3)<sup>82</sup>. Joint orthologue genome coordinates in chromosomal location were used for 'Oxford grid' plots (Fig. 1f and Extended Data Fig. 3) and Fisher's exact test was used to determine mutual enrichment of orthologues between chromosomes. Ribbon plots connecting orthologues in multiple species (Fig. 1d) were generated using Rideogram (v.0.2.2).

Hox genes were annotated based on similarity search toward previously annotated chaetognath Hox genes<sup>19,87</sup>, inspection of diagnostic residues and phylogenetic analysis of selected bilaterian homeodomain genes (Supplementary Fig. 3) using IQ-TREE with a LG4X model and 1000 ultra-fast bootstraps.

### Hi-C processing

Hi-C reads were mapped using bwa mem (v.0.7.17) with parameters -5SP -TO. Ligation events in the aligned sequences were then identified with pairtools (v.1.0.2)<sup>88</sup>. In brief, we first used pairtools parse with parameters --min-mapq 30 --max-inter-align-gap 30 --no-flip --add-columns mapq --walks-policy 5unique, sorted resulting contact pairs with pairtools sort and removed duplicates with pairtools dedup to obtain the final filtered contact pairs file. Using juicer\_tools (v.1.22.01)<sup>89</sup>, filtered contact pairs were exported to a multi-resolution Hi-C matrix (2.5 Mb, 1 Mb, 500 kb, 250 kb, 100 kb, 50 kb, 25 kb, 15 kb, 10 kb, 5 kb) and normalized with the Knight-Ruiz (KR) procedure<sup>90</sup>. Contact map resolution was estimated using the approach proposed

in ref. 66. In brief, for each bin size stored in the multi-resolution Hi-C matrix, we count the number of bins with >1,000 contacts. The map resolution is given as the smallest bin size where >80% of bins have >1,000 contacts (5 kb here). All downstream analyses were performed using the fanc Python package (v.0.9.23)<sup>91</sup>.

Putative centromere positions were inferred from Hi-C (Extended Data Fig. 4f): for each chromosome, we: (1) computed Pearson contact correlation matrices; (2) visually selected for each chromosome the eigenvector that best captured centromere information; and (3) identified centromere as the most insulating region (highest eigenvector value differential). We tested for the enrichment of specific repeats at inferred centromere location using a permutation procedure. Specifically, for each repeat family or subfamily 'r' overlapping at least 1 predicted centromere, and for each chromosome 'c', we: (1) generated  $n = 1,000$  randomizations of 'r' over 'c' and (2) computed empirical permutation-based *P* values based on the random background overlap of repeats at centromere.

To assess the organization of the *P. gotoi* genome into TADs, we computed insulation scores on the Hi-C matrix at 25 kb resolution using a window size of 250 kb. Putative TAD boundaries were called from insulation scores and TADs with boundary strength >0.5 (that is, discarding the first peak of weak boundary scores) were retained for the aggregated TAD and Hox cluster plots. Finally, organization into A/B compartments was assessed by: (1) selecting, for each chromosome, eigenvectors that best separate active and inactive regions on the basis of ATAC peak number; (2) orienting eigenvectors with ATAC peaks so that entries >0 correspond to active regions and <0 inactive; (3) representing interactions strength across active and inactive regions as a saddle plot (Extended Data Fig. 4e) using 50 eigenvector value categories.

### ATAC-seq

ATAC-seq was performed as previously described<sup>92</sup> following the Omni-ATAC protocol including digitonin (Promega)<sup>93</sup>. Tn5 tagmentase (Illumina) was applied to 50k nucleus and library subsequently constructed by PCR. ATAC-seq reads were aligned to the genome using bowtie2 (v.2.4.2)<sup>94</sup> with the parameters `-very-sensitive` and `-k 10`. Open chromatin regions (OCR or peaks) were called using Genrich (v.0.6) (<https://github.com/jsh58/Genrich>) using ATAC model (-j), keeping unpaired alignments (-y), removing PCR duplicates (-r) and excluding reads mapped to mitochondria (-e MT). Unified OCRs were computed using pyBEDtools (github) as well as their relative location toward genes and 42,810 peaks in total were identified (Supplementary Table 11) and compared with their distribution in other species<sup>721,92</sup>. Mapping statistics and fraction of reads in peaks, calculated using bedtools (v.2.30.0), are provided in Supplementary Table 12. Motif enrichment was performed using HOMER (v.5) for both known and de novo motifs and plant derived motifs were excluded from results relying on known motifs<sup>95</sup>.

### Trans-splicing and operons

Low-input RNA-seq libraries were generated from gastrula, hatchling and 3 days post-hatchling juveniles using low-input protocol (NEBNext Low Input RNA). We used SLIDR to perform a de novo annotation of splice leaders from RNA-seq data using default parameters<sup>96</sup> and compared the identified list with previously identified splice leaders in other chaetognath species<sup>10</sup>. We found that *P. gotoi* uses three major splice leaders of which two are shared with other chaetognath species. We performed a BLAST similarity search on Isoseq full length transcripts to annotate *trans*-spliced genes. To quantify *trans*-splicing in RNA-seq data, we aligned each set of RNA-seq reads to the genome using STAR (v.2.7.8a), detected splice leaders in reads using BBDuk from BBTtools to with `k = 25` and `hdist = 3` and set them aside as separate BAM (<https://jgi.doe.gov/data-and-tools/software-tools/bbttools/bb-tools-user-guide/>). We then quantified reads for our gene models using featureCount<sup>97</sup>. We considered a gene *trans*-spliced if at least five reads including a splice leader aligned to it. We annotated operons in the genome as sets of two

or more genes separated by less than 250 bp, which yielded 2,151 candidate operons containing 5,061 genes. Of those, 1,470 validated operons include at least 1 gene with annotated splice leader involving 3,565 genes.

### Single-cell transcriptomics

Adults and 3 dpf juveniles of *P. gotoi* were dissociated in filtered natural sea water in presence of 1% Pronase with gentle trituration, cells were washed, debris removed using 40- $\mu$ m and 70- $\mu$ m cell strainers (Flowmi) and resuspended in 50% Leibowitz L15 medium. Single-cell suspension quality and cell concentration was evaluated using C-Chip Disposable haemocytometer (NanoEntek) and viability was assessed using Trypan blue (Thermo Fisher). Around 15,000 cells were loaded into a 10x chip (10x Genomics, v.2) and 3 libraries (2 for adults and 1 for 3 dpf juveniles) were prepared and sequenced on a NovaSeq6000 and NovaSeq X instrument (Illumina). Reads were mapped onto the *P. gotoi* genome using Cell Ranger (v.7.0.1) after extending the 3' UTR of the genome annotation by 900 bp with the script `utr_extension.py`. Library saturation is 59.8% for adult 1, 74.2% for adult 2, and 89.5% for 3 dpf juveniles, and read depth is 24,583 mean reads per cell for adult 1, 37,173 for adult 2 and 58,191 for juveniles. Counts were analysed using Seurat (v.5)<sup>98</sup> and all cells with fewer than 500 or more than 20,000 unique molecular identifiers, fewer than 200 genes and mitochondrial content higher than 20% were filtered out. Our final dataset contains 29,513 cells (21,777 adult cells and 7,736 juvenile cells) expressing 19,367 genes. These datasets were merged, normalized, scaled and subjected to principal components analysis (using the top 50 principal components) followed by canonical correlation analysis integration were performed on the object layers. Our final resolution for clustering was set to 1, resulting in a total of 33 cell clusters (Rscript in Supplementary Information). Gene expression specificity such as the tau metrics or the coefficient of variation were computed as described<sup>99</sup>.

SAMap (v.1.0.2) was applied between the *P. gotoi* datasets and the following species: the polyclad flatworm *Prostheceraeus crozieri*<sup>39</sup>, the flatworm *S. mediterranea*<sup>100</sup>, the oyster *Crassostrea gigas* (larval stages<sup>39</sup>), the earthworm *Eisenia andrei*<sup>40</sup>, the annelid worm *Pristina leidyi*<sup>38</sup> and the fruit fly *Drosophila melanogaster*<sup>40</sup>. Cell-type annotation for all species can be found in Supplementary Table 8; for a list of genes shared between dataset per cell type, see Supplementary Table 9.

### In situ hybridization

Synthetic probes were designed for selected genes (Supplementary Table 7) flanked by T7 or SP6 adaptor (IDT). Riboprobes were then synthesized using the DIG labelling kit (Roche). In situ hybridization was performed as described<sup>7,101,102</sup> with a shortened Proteinase K digestion (2–3 min). Stained embryos and juveniles were imaged with a Zeiss AxioImager M1 using bright field Nomarski optics. Unfortunately, surveying the expression of other Hox genes than the one we investigate here in *P. gotoi* proved challenging as the high AT percentage and low-complexity stretches in their transcripts made it impossible to synthesize or clone a probe.

### Enzymatic Methyl-seq analysis

Using the same genomic DNA obtained for genome sequencing, 100 ng of *P. gotoi* gDNA were spiked in with unmethylated lambda phage and CpG methylated pUC19 control before fragmenting it to about 300 bp with a BioRuptor (Covaris). The resulting sheared DNA was then used to prepare an Enzymatic Methyl-seq library according to the manufacturer's instructions (NEBNext Enzymatic Methyl-seq Kit). The library was sequenced with a NovaSeq instrument using 150 paired-end reads to an average coverage of  $\sim 14\times$  per CpG site. Reads were mapped to the reference genome, including also a phage lambda and pUC19 sequences, using BS-Seeker2 (v.2.1.8) with bowtie2 as back-end. Duplicate reads were removed with Sambamba (v.1.0.1), and methylation calls were obtained using CGmapTools (v.0.1.3). The lambda genome showed a non-conversion rate of 0.11%, and the pUC19 methylated control showed

97.88% global methylation at CpG methylation sites, as expected for this technology and validating the quality of the library. The rest of the data were read into R using the *bsseq* Bioconductor package, which was used to obtain methylation average per gene and genomic windows used for chromosome-level methylation. Bigwig files were constructed from the CGmapTools output, and were used as input for DeepTools2 to detect methylation levels on top of protein-coding genes and the repeat annotation (Extended Data Fig. 10). Methylation genes were obtained using human DNMT1, DNMT3A, TET1 and UHRF1 sequences as input for a BLASTP search against the predicted proteome of *P. gotoi*, and were subsequently analysed in the InterProScan web server (<https://www.ebi.ac.uk/interpro>). The top hits were manually inspected and contrasted with the phylogenetic trees obtained from the GeneRax search.

## Reporting summary

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

## Data availability

The genome assembly and associated data have been deposited to the NCBI database under BioProject PRJNA1169515. Functional genomics data have been deposited in the Gene Expression Omnibus (GEO) GSE279488. Processed data files including genome annotation have also been deposited at Zenodo (<https://doi.org/10.5281/zenodo.13936459> (ref. 103)).

## Code availability

Code used is available at Zenodo (<https://doi.org/10.5281/zenodo.13936459> (ref. 103)).

- Goto, T. & Yoshida, M. Growth and reproduction of the benthic arrowworm *Paraspadella gotoi* (Chaetognatha) in laboratory culture. *Invertebr. Reprod. Dev.* **32**, 201–207 (1997).
- Green, M. R. & Sambrook, J. *Molecular Cloning. A Laboratory Manual* 4th edn (2012).
- Rao, S. S. P. et al. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**, 1665–1680 (2014).
- Chapman, J. A. et al. Meraculous: de novo genome assembly with short paired-end reads. *PLoS ONE* **6**, e23501 (2011).
- Putnam, N. H. et al. Chromosome-scale shotgun assembly using an in vitro method for long-range linkage. *Genome Res.* **26**, 342–350 (2016).
- Simakov, O. et al. Deeply conserved synteny resolves early events in vertebrate evolution. *Nat. Ecol. Evol.* **4**, 820–830 (2020).
- Rhie, A., Walenz, B. P., Koren, S. & Phillippy, A. M. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biol.* **21**, 245 (2020).
- Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–3212 (2015).
- 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).
- 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).
- Klopfenstein, D. V. et al. GOATOOLS: a Python library for Gene Ontology analyses. *Sci Rep.* **8**, 10872 (2018).
- Derelle, R., Philippe, H. & Colbourne, J. K. Broccoli: combining phylogenetic and network analyses for orthology assignment. *Mol. Biol. Evol.* **37**, 3389–3396 (2020).
- Katoh, K., Misawa, K., Kuma, K.-I. & Miyata, T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* **30**, 3059–3066 (2002).
- Steenwyk, J. L., Buida, T. J. 3rd, Li, Y., Shen, X.-X. & Rokas, A. ClipKIT: a multiple sequence alignment trimming software for accurate phylogenomic inference. *PLoS Biol.* **18**, e3001007 (2020).
- Nguyen, L.-T., Schmidt, H. A., von Haeseler, A. & Minh, B. Q. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* **32**, 268–274 (2015).
- Barrera-Redondo, J., Lotharukpong, J. S., Drost, H.-G. & Coelho, S. M. Uncovering gene-family founder events during major evolutionary transitions in animals, plants and fungi using GenEra. *Genome Biol.* **24**, 54 (2023).
- Steinegger, M. & Söding, J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat. Biotechnol.* **35**, 1026–1028 (2017).
- Benton, M. J., Donoghue, P. C. J. & Asher, R. J. in *The Timetree Of Life* (ed. Kumar, S. B. H.) 35–86 (Oxford Univ. Press, 2009).

- Rota-Stabelli, O., Daley, A. C. & Pisani, D. Molecular timetrees reveal a Cambrian colonization of land and a new scenario for ecdysozoan evolution. *Curr. Biol.* **23**, 392–398 (2013).
- Vannier, J., Steiner, M., Renvoise, E., Hu, S.-X. & Casanova, J.-P. Early Cambrian origin of modern food webs: evidence from predator arrow worms. *Proc. Biol. Sci.* **274**, 627–633 (2007).
- Mendes, F. K., Vanderpool, D., Fulton, B. & Hahn, M. W. CAFE 5 models variation in evolutionary rates among gene families. *Bioinformatics* **36**, 5516–5518 (2021).
- Matus, D. Q., Halanych, K. M. & Martindale, M. Q. The Hox gene complement of a pelagic chaetognath, *Flaccisagitta enflata*. *Integr. Comp. Biol.* **47**, 854 (2007).
- Open2C, et al. Pairtools: from sequencing data to chromosome contacts. *PLoS Comput. Biol.* **20**, e1012164 (2024).
- Durand, N. C. et al. Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. *Cell Syst.* **3**, 95–98 (2016).
- Knight, P. A. & Ruiz, D. A fast algorithm for matrix balancing. *IMA J. Numer. Anal.* **33**, 1029–1047 (2013).
- Kruse, K., Hug, C. B. & Vaquerizas, J. M. FAN-C: a feature-rich framework for the analysis and visualisation of chromosome conformation capture data. *Genome Biol.* **21**, 303 (2020).
- Marlétaz, F. et al. Analysis of the *P. lividus* sea urchin genome highlights contrasting trends of genomic and regulatory evolution in deuterostomes. *Cell Genomics* **3**, 100295 (2023).
- Corces, M. R. et al. An improved ATAC-seq protocol reduces background and enables interrogation of frozen tissues. *Nat. Methods* **14**, 959–962 (2017).
- Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
- Heinz, S. et al. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol. Cell* **38**, 576–589 (2010).
- Wenzel, M. A., Müller, B. & Pettitt, J. SLIDR and SLOPPR: flexible identification of spliced leader *trans*-splicing and prediction of eukaryotic operons from RNA-seq data. *BMC Bioinformatics* **22**, 140 (2021).
- 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).
- Hao, Y. et al. Dictionary learning for integrative, multimodal and scalable single-cell analysis. *Nat. Biotechnol.* **42**, 293–304 (2024).
- Kryuchkova-Mostacci, N. & Robinson-Rechavi, M. A benchmark of gene expression tissue-specificity metrics. *Brief. Bioinformatics* **18**, 205–214 (2016).
- García-Castro, H. et al. ACME dissociation: a versatile cell fixation-dissociation method for single-cell transcriptomics. *Genome Biol.* **22**, 89 (2021).
- Hejnal, A. & Martindale, M. Q. Acoel development indicates the independent evolution of the bilaterian mouth and anus. *Nature* **456**, 382–386 (2008).
- Hejnal, A. In situ protocol for embryos and juveniles of *Convolutriloba longifissura*. *Protoc. Exch.* <https://doi.org/10.1038/nprot.2008.201> (2008).
- Marlétaz, F. et al. The genomic origin of the unique chaetognath body plan [Data set]. *Zenodo* <https://doi.org/10.5281/zenodo.13936459> (2024).
- Gąsiorowski, L., Martín-Durán, J. M. & Hejnal, A. in *Hox Modules in Evolution and Development* (ed. Ferrier, D. E. K.) 177–194 (CRC, 2023).

**Acknowledgements** F.M. and L.P. are supported by a Royal Society University Research Fellowship (URF/R1/191161), a Biotechnology and Biological Sciences Research Council research grant (BB/V01109X/1), a Leverhulme Trust research grant (RPG-2021-436) and a Japan Society for the Promotion of Science Kakenhi grant (JP 19K06620). E.P. is supported by a Newton International Fellowship from the Royal Society (NIF/R1/222125). Work at the Okinawa Institute of Science and Technology (OIST) Molecular Genetics Unit (D.S.R., D.G. and F.M.) and Marine Genomics Unit (N.S.) was supported by OIST internal funds. We acknowledge the OIST sequencing facility and its members for support, especially N. Arakaki and H. Goto. 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. We thank the OIST Sequencing Section and acknowledge the support of OIST supercomputing. A.d.M. and J.M.M.-D. are funded by the Horizon 2020 Framework Programme (European Research Council Starting Grants action number 950230 to A.d.M. and 801669 to J.M.M.-D.). L.A.S. was funded by a QMUL PhD studentship and the ERC Starting Grant 950230. The authors thank H. Miyamoto for his help; M. Telford for insightful comments on the manuscript; K. Valentinova for help dealing with overdetermination; and Y. Le Parco for prior inspiration and discussions. The authors acknowledge the pioneering work of the late Q. Bone on chaetognaths and zooplankton.

**Author contributions** F.M., T.G. and N.S. conceived the study. D.G. and L.P. performed the single-cell RNA-sequencing experiments and analyses. L.P. carried out in situ hybridization, phylostratigraphic analyses and SAMap comparisons. J.M.M.-D. helped to set up and interpret in situ hybridization in chaetognaths. F.M., N.S. and D.S.R. sequenced, assembled and annotated the genome. T.G. generated the inbred line of *P. gotoi*. D.G. carried out ATAC-seq experiments and analyses. T.G., K.T.C.A.P. and F.M. collected and spawned adult chaetognaths to obtain embryos. E.P. analysed proximity ligation data and performed the gene family expansion analyses. L.A.S. constructed the Enzymatic Methyl-seq library. A.d.M. analysed the methylation gene toolkit and methylation signal. F.M. conducted comparative genomics and gene family analyses. F.M. and L.P. wrote the paper with input from all authors.

**Competing interests** D.S.R. is a shareholder and advisor to Cantata Bio. The other authors declare no competing interests.

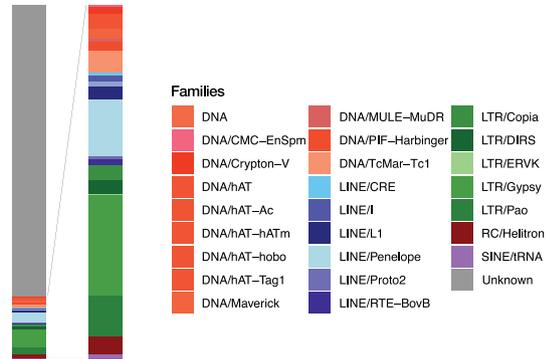
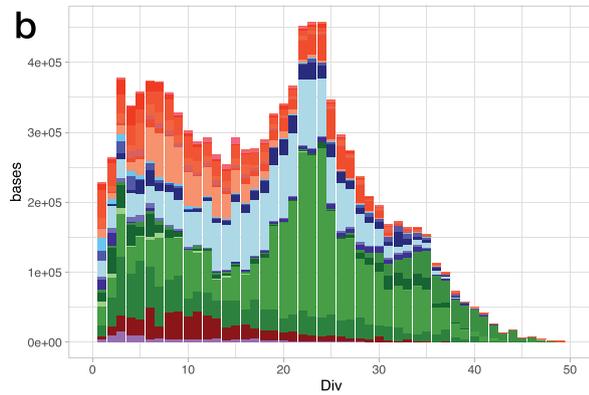
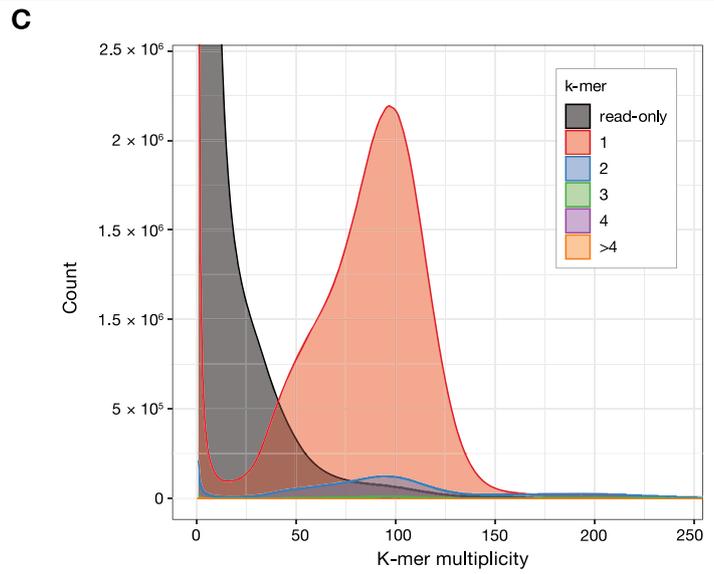
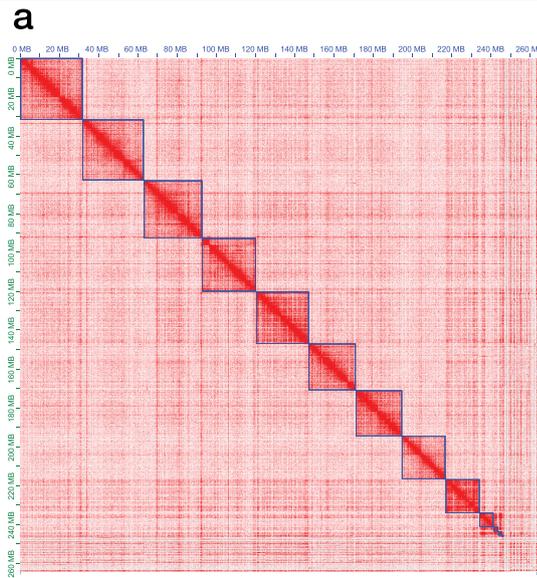
## Additional information

**Supplementary information** The online version contains supplementary material available at <https://doi.org/10.1038/s41586-025-09403-2>.

**Correspondence and requests for materials** should be addressed to Ferdinand Marlétaz.

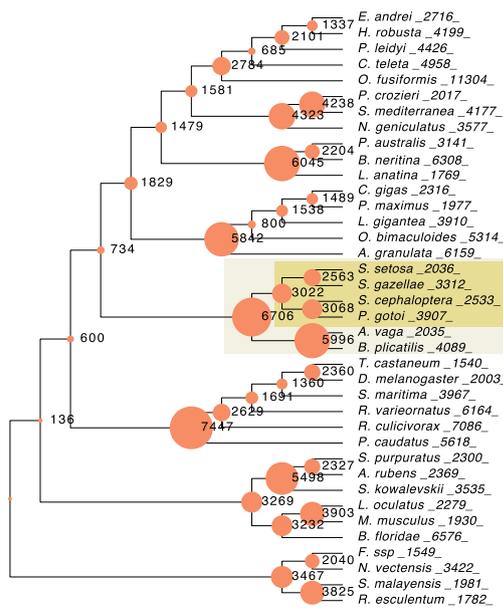
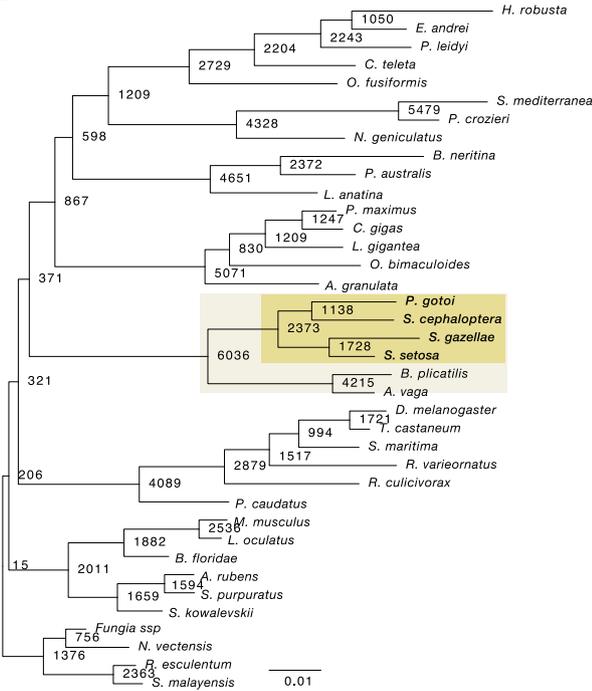
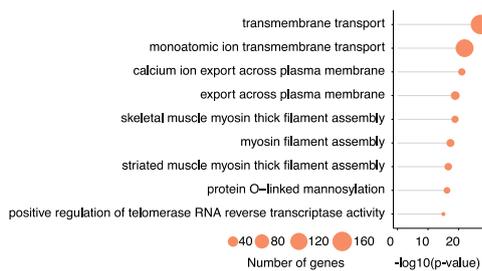
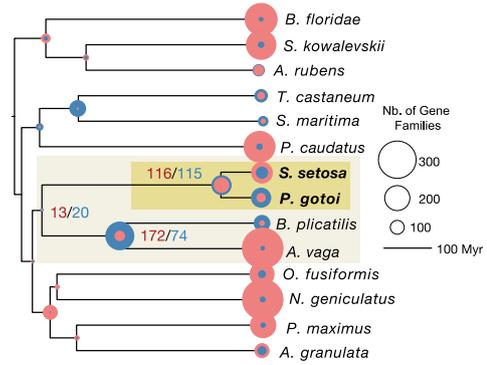
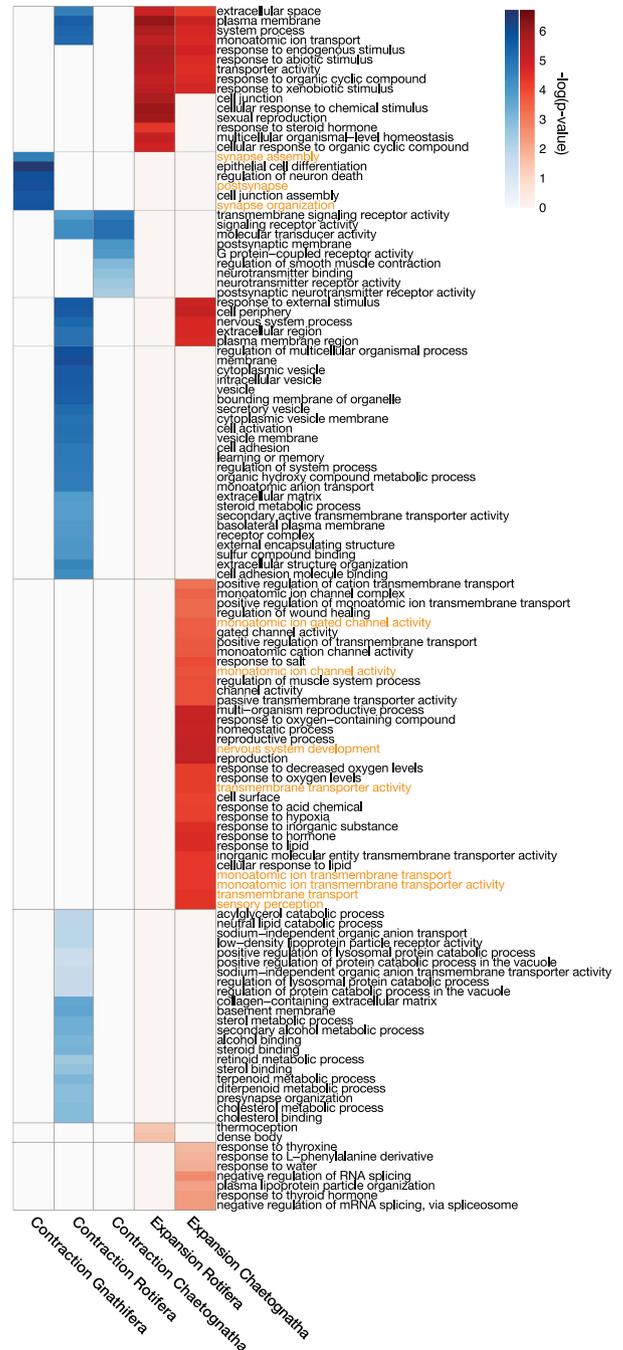
**Peer review information** Nature thanks Yi-Jyun Luo, Katrine Worsaae and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

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



**Extended Data Fig. 1 | Genome assembly and repetitive landscape for *P. gotoi*.** **a**, Hi-C contact map representing binned contact density across the genome and highlighting the 9 chromosome-size scaffolds of *P. gotoi* consistent with reported karyotype in other chaetognath species. **b**, Repeat landscape summarising the fraction of regions diverging from consensus

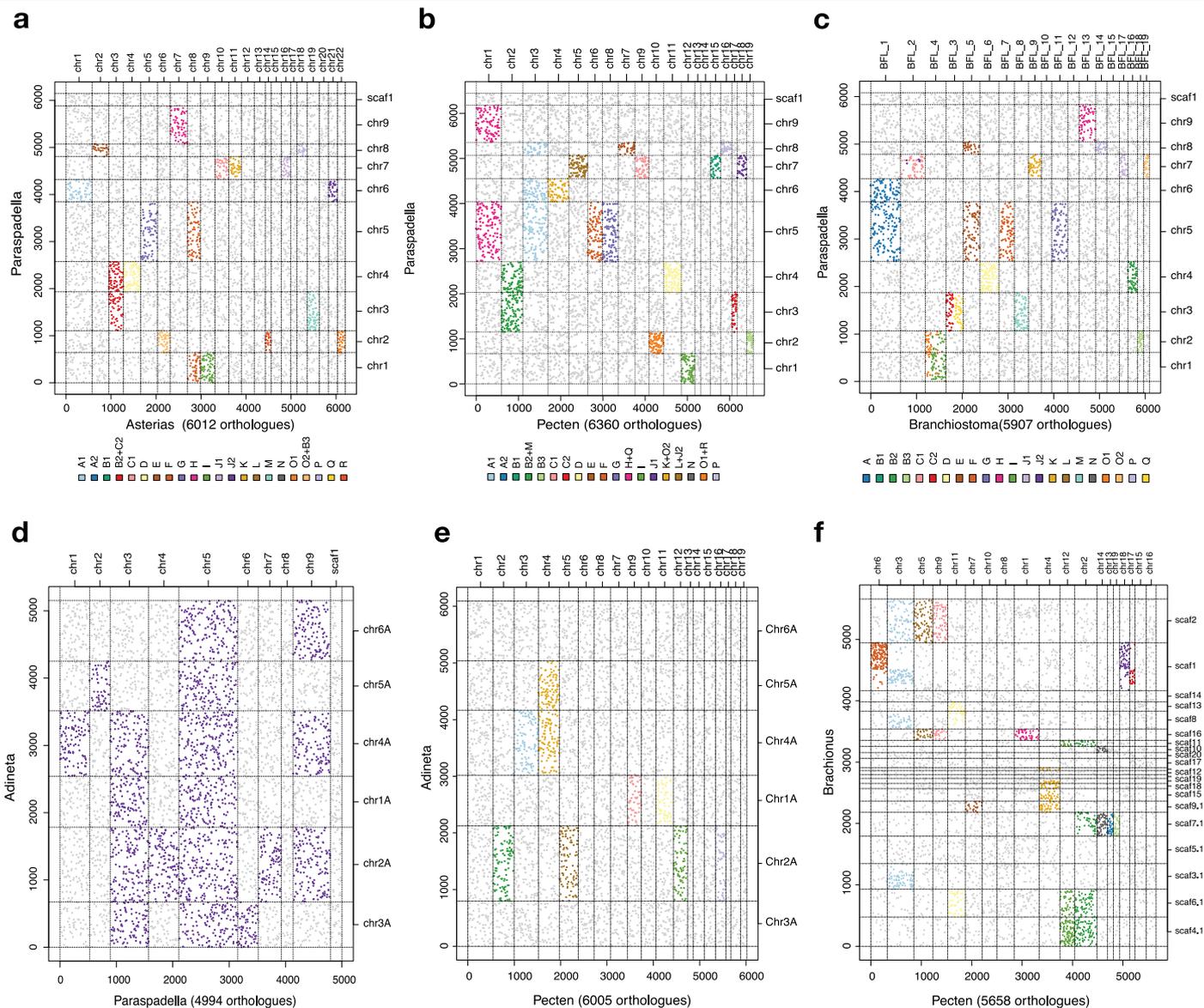
repeats at varying levels of divergence (Kimura 2-parameters) and the proportion of different repeats annotated in *P. gotoi*. **c**, k-mer spectrum assuming a 21-mer featuring the coverage at various k-mer multiplicity in the assembly and supporting polymorphism homogenisation.

**a****b****e****c****d****Extended Data Fig. 2** | See next page for caption.

# Article

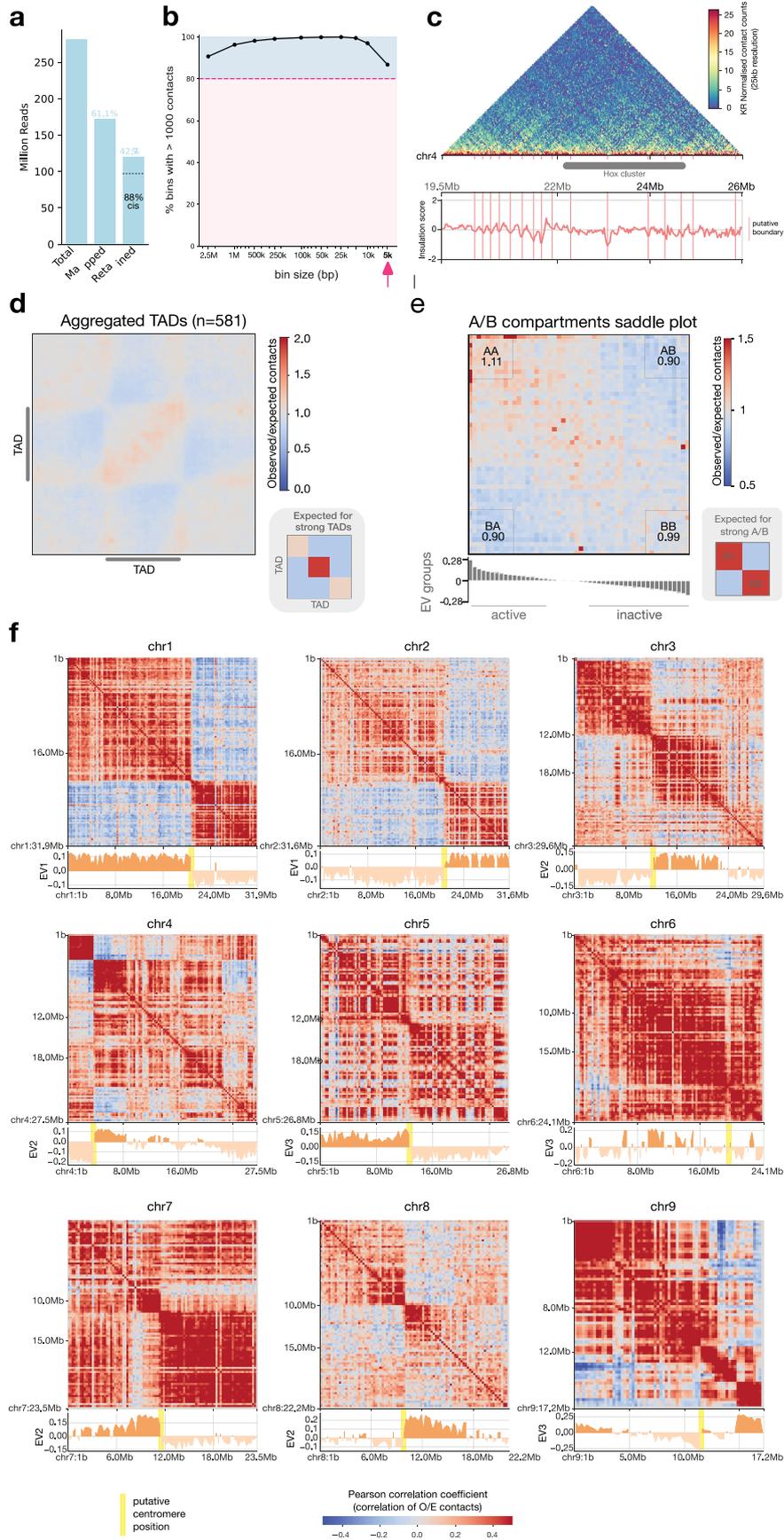
**Extended Data Fig. 2 | Evolution of gene complement.** **a.** Loss of PANTHER families inferred on the metazoan tree according to parsimony. **b.** Maximum-likelihood rates of gene loss inferred from GeneRax gene tree/species tree reconciliation. **c.** Number of expanded (red) and contracted (blue) gene families inferred in the CAFE analysis mapped on a time tree. **d.** Gene ontology enrichment for gene families expanded (red) or contracted (blue) at different

gnathiferan nodes highlighting terms involved in neuronal function in yellow that appears reduced in the gnathiferan node and expanded secondarily in chaetognaths (Fisher's exact test, two-sided). **e.** Gene ontology enrichment calculated for genes belonging to families expanded in chaetognaths (Fisher's exact test, two-sided).



**Extended Data Fig. 3 | Synteny evolution in Gnathifera.** ‘Oxford’ synteny plots showing the pairwise synteny relationship between the chaetognath *P. gotoi*, the sea star *Asterias rubens* (a), the sea scallop *Pecten maximus* (b), the amphioxus *Branchiostoma Floridae* (c), the rotifer *Adineta vaga* (d). A direct comparison between rotifers (e, *Adineta saga*; f, *Brachionus manjavacas*) and

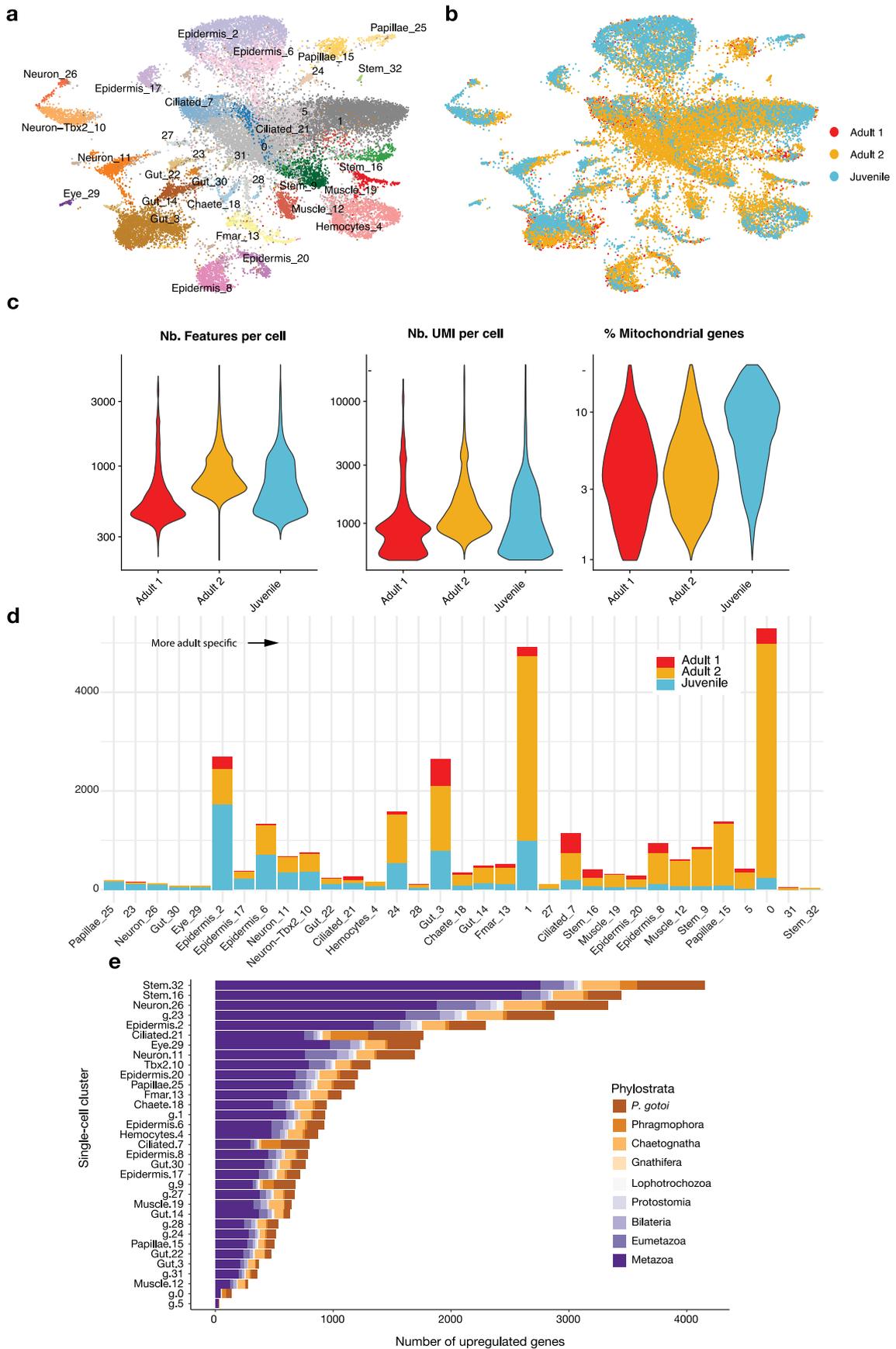
sea scallop highlights the lack of direct correspondence and the loss of ALGs in rotifers. Orthologous genes are computed as mutual-best-hits (MBH) and coloured according to the inferred ALGs. Dots in pairs of chromosomes that show a reciprocal enrichment using Fisher’s exact test ( $p < 0.05$ , one-sided) are coloured according to their ALG assignments (a, b, d) or arbitrarily (c).



Extended Data Fig. 4 | See next page for caption.

**Extended Data Fig. 4 | Limited compartmentalisation of *P. gatoi* chromatin.** **a**, Mapping statistics of the Hi-C library highlighting the number of mapped and inferred contacts used to generate contacts maps including those in *cis*- **b**, Resolution of Hi-C library estimated by computing the number of bins containing >1000 contacts at increasingly higher resolutions. **c**, Hi-C contacts map in the Hox cluster region showing insulation scores and putative TAD boundaries revealing the limited structuration of chromatin in the Hox region. **d**, TAD metaplot showing the limited domain structuration of *P. gatoi* chromatin (average within TADs observed/expected contact ratio = 1.05) as

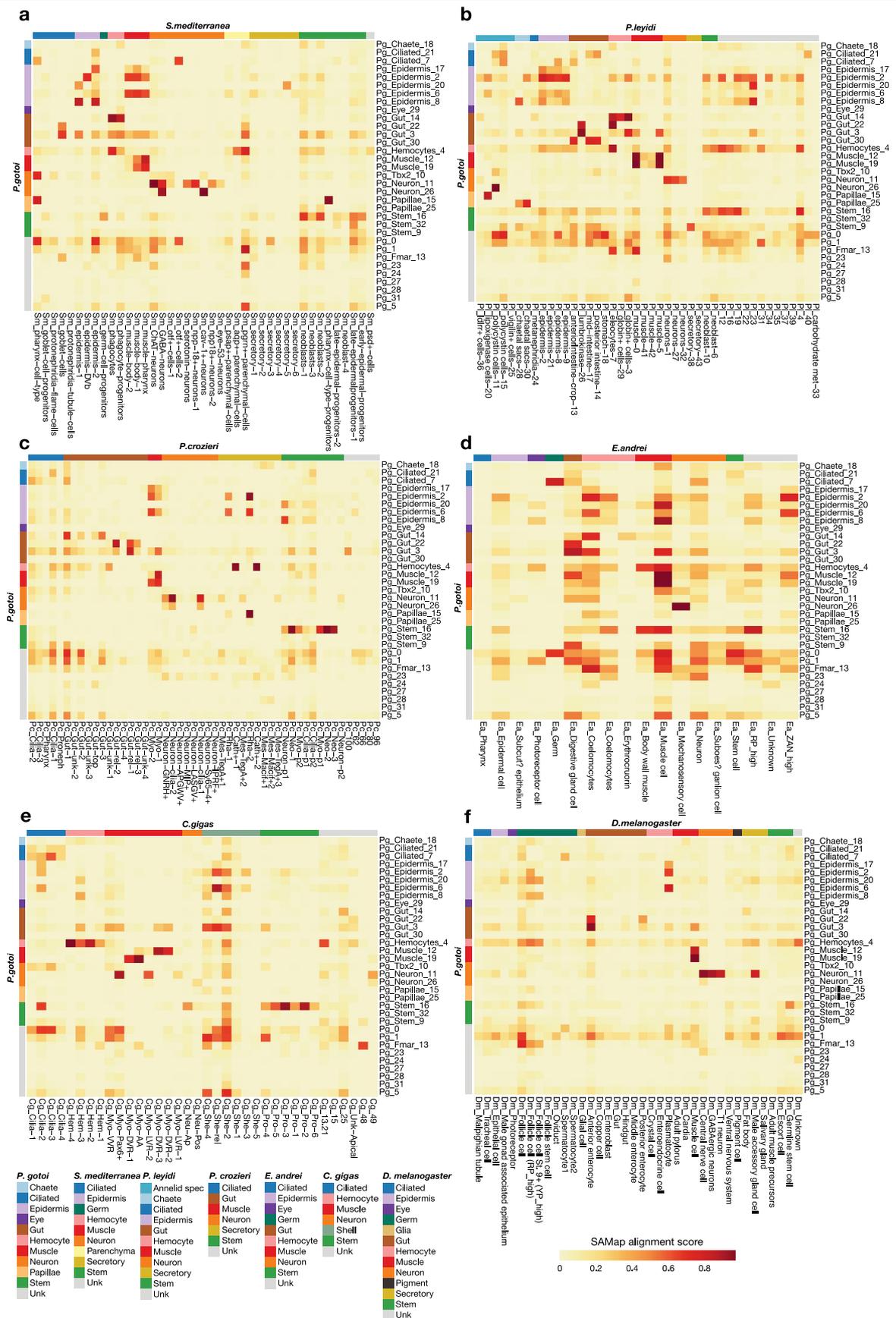
well as the expected pattern in case of strong TAD structuration shown as the bottom right inset. **e**, Saddle plot showing interactions strength across active and inactive regions with inset showing expected pattern for strong A/B compartmentalisation. **f**, Pearson correlation matrices and associated eigenvectors that best capture centromere position, calculated at 250 kb resolution for the 9 chromosomes of *P. gatoi*. Putative centromere positions are indicated (highest eigenvector value differential). Hi-C maps suggest a lack of A/B structures (except possibly for chr5).



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

**Extended Data Fig. 5 | Single cell of juvenile and adult *P. gatoi*.** **a**, UMAP of all cells with annotated cluster, **b**, same UMAP with individual cells annotated from the sample origin. **c**, Quality control statistics across single-cell samples, including number of features (genes) per cell, number of reads per cell, and percentage of mitochondrial reads per cell. **d**, Number of cells derived from the

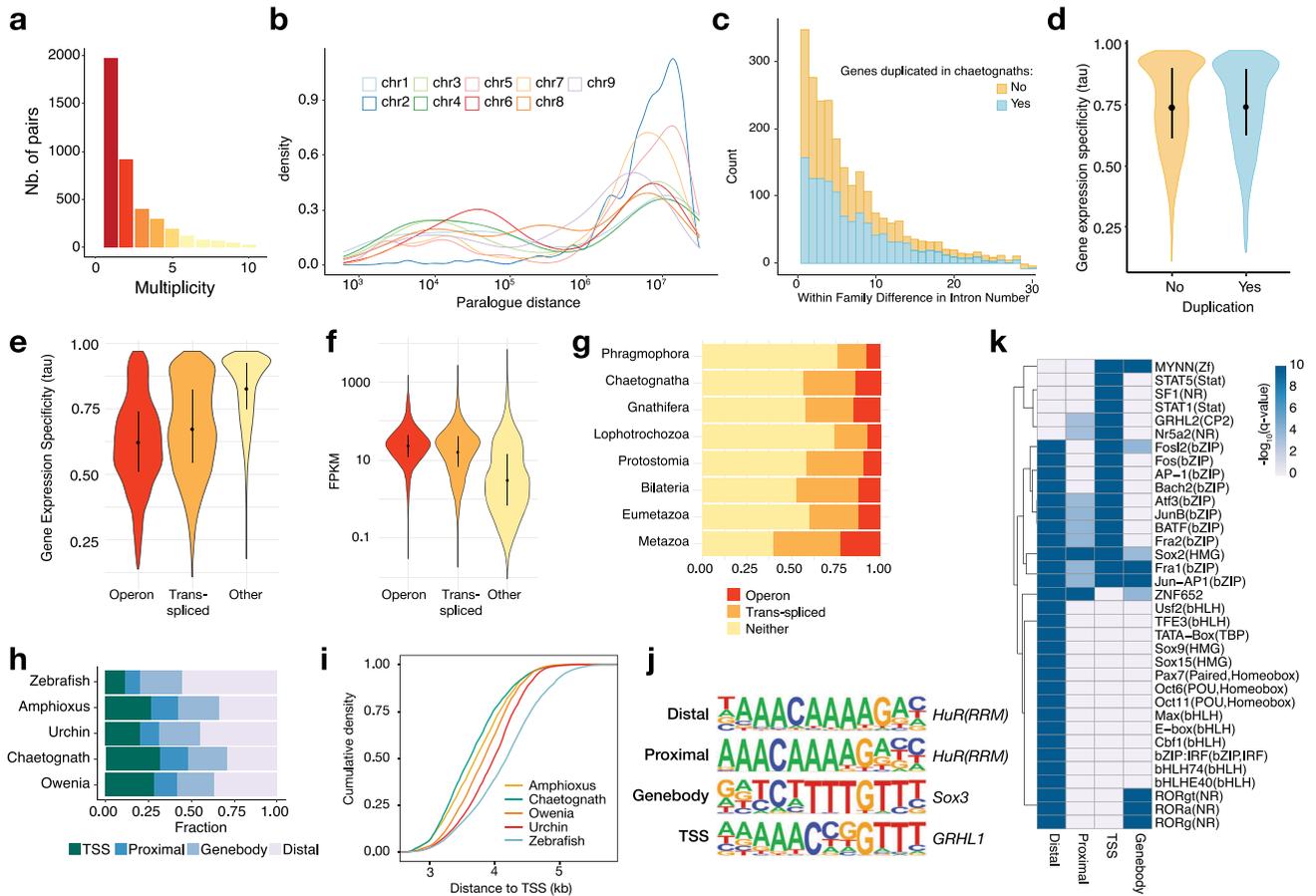
3 different samples in each single-cell cluster. Most cell types are shared across juveniles and adults except for one germline cell cluster (cluster 32) and cluster 31 which are specific to adults, while papillae cluster 25 and cluster 23 are mostly composed of juveniles cells. **e**, Number of genes from distinct phylostratigraphic origin in each single-cell cluster.



Extended Data Fig. 6 | See next page for caption.

**Extended Data Fig. 6 | Pairwise SAMap comparison between selected scRNA-seq datasets and *P. gotoi*.** **a**, platyhelminth *Schmidtea mediterranea* from (García-Castro et al.<sup>100</sup>). **b**, annelid *Pristina leyidi* from (Álvarez-Campos et al.<sup>38</sup>). **c**, the larva of the flatworm *Prostheceraeus crozieri* from (Piovani et al., 2023)<sup>39</sup> **d**, earthworm *Eisenia andrei* (annelid) from (Li et al.<sup>40</sup>). **e**, the larva of the oyster *Crassostrea gigas* from (Piovani et al., 2023)<sup>39</sup> and **f**, the fruit fly

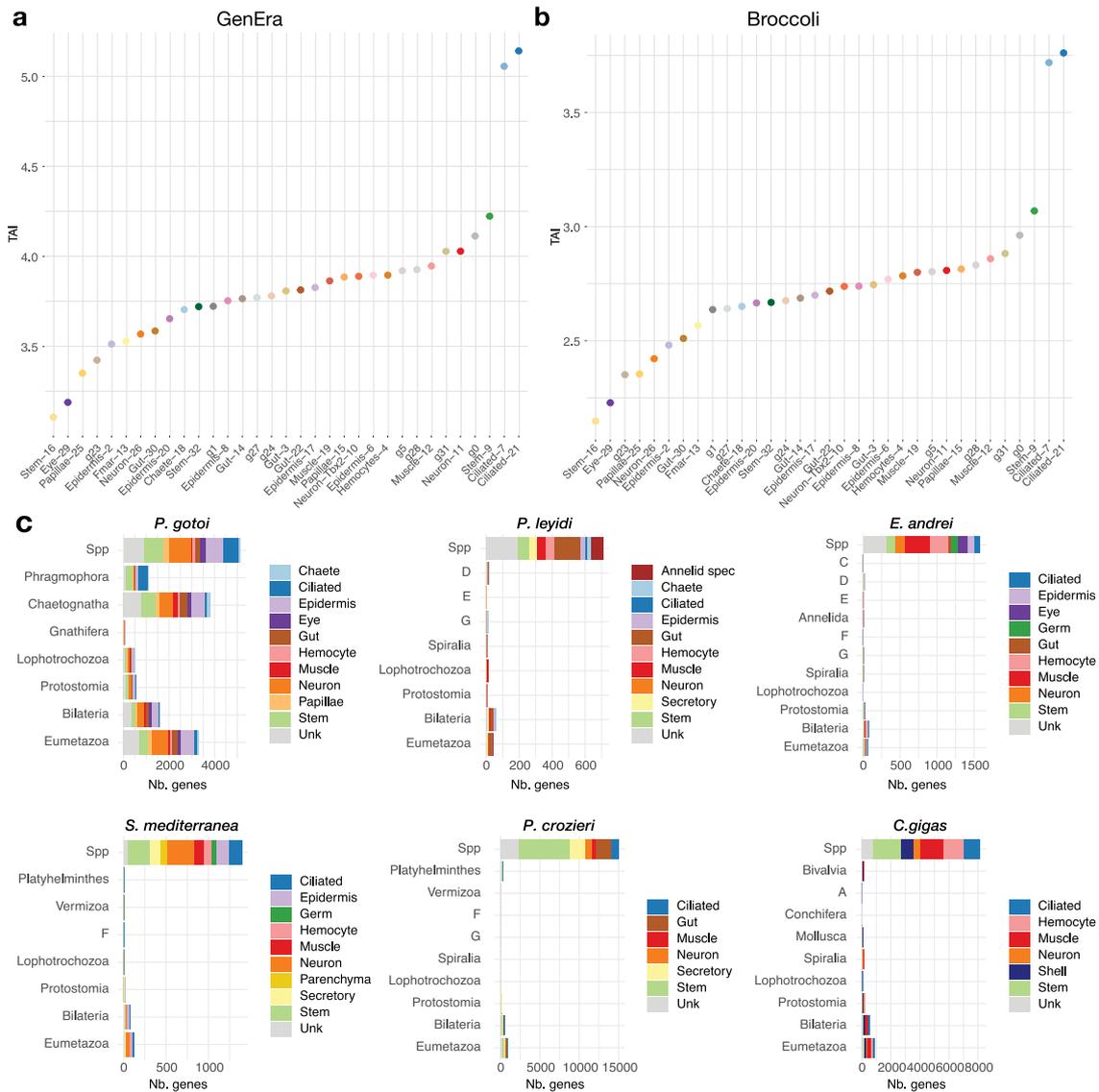
*Drosophila melanogaster* (Li et al.<sup>40</sup>). The cell-type colour code corresponds to broad putatively homologous cell type categories. The scale bar indicates the alignment score which is defined as the average number of mutual nearest cross-species neighbours of each cell relative to the maximum possible number of neighbours.



**Extended Data Fig. 7 | Genome duplication and regulation in *P. gotoi*.**

**a**, Number of paralogous genes (multiplicity) derived from each chaetognath-specific duplication. **b**, Distance between paralogous pairs belonging to a single chromosome. **c**, Difference between exon number of paralogues belonging to a given gene family for gene family containing chaetognath-specific duplicates or not. **d-e**, Gene expression specificity ( $\tau$ ) computed using averaged expression of main chaetognath cell types for chaetognath duplicates (**d**,  $n = 15060$  and  $4250$  genes for duplicated and non-duplicated, respectively), as well as depending on their trans-splicing status (**e**,  $n = 3612$ ,  $6342$  and  $9356$  genes for operon, trans-spliced and other genes, respectively). In **d-f**, violin plots with mean and quartiles as error bars.

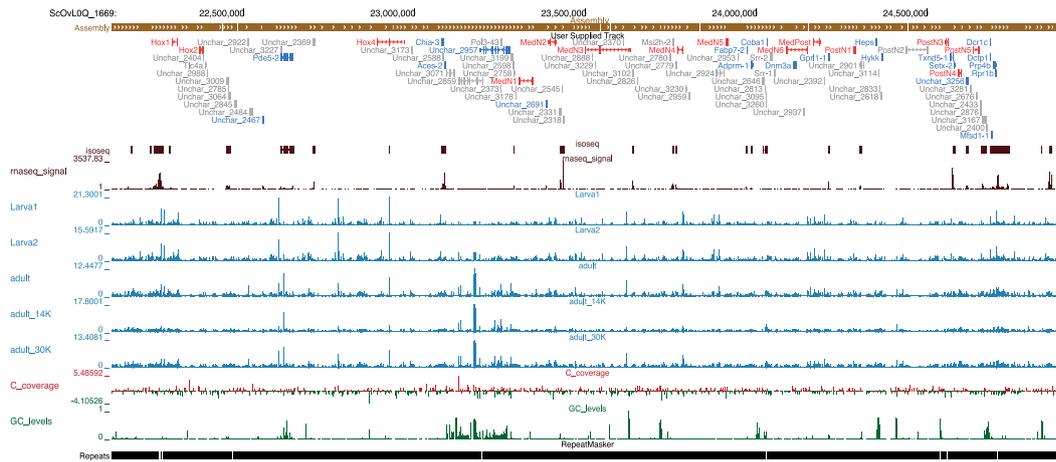
**f**, Expression level (FPKM) of genes depending on their trans-splicing status ( $n = 3740$ ,  $6457$  and  $11885$  genes for operon, trans-spliced and other genes, respectively); **g**, Distribution of genes of distinct trans-splicing status across phylostratigraphic levels. **h**, Number of Open Chromatin Regions (OCR) assigned to various genomic locations (proximal is  $<5$  kb) toward the genome. **i**, Distance to TSS of ATAC-based OCRs in a select species. **j**, De novo motifs detected by HOMER for OCR localised at different genomic positions. **k**, Enriched known motifs detected by HOMER in the different peak categories highlighting the conservation of some transcription factors properties, with bHLH and homeodomains preferentially targeting distal OCRs (one-sided hypergeometric test).



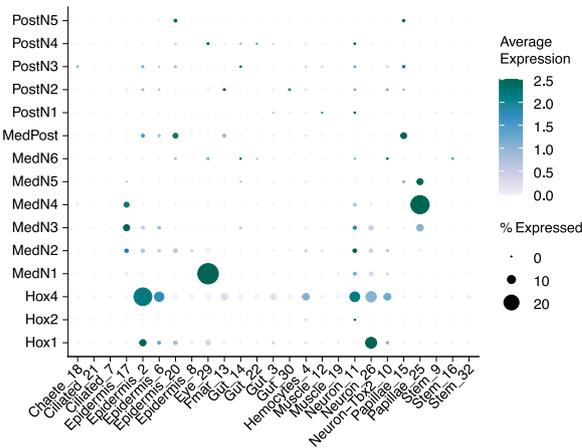
**Extended Data Fig. 8 | Phylostratigraphic analysis of single-cell gene expression clusters. a-b**, Transcriptome Age Index (AI) calculated using gene age derived from GenEra that relies on blast similarity search (a) and from the age of inferred Broccoli gene families (b). **c**, Phylostratigraphic distribution of

cell-type marker genes in *P. gotoi* and other species considered according to the age of Broccoli gene families. **d**, Number of cell-types belonging to broad functional categories in the different species for which we compared single-cell datasets.

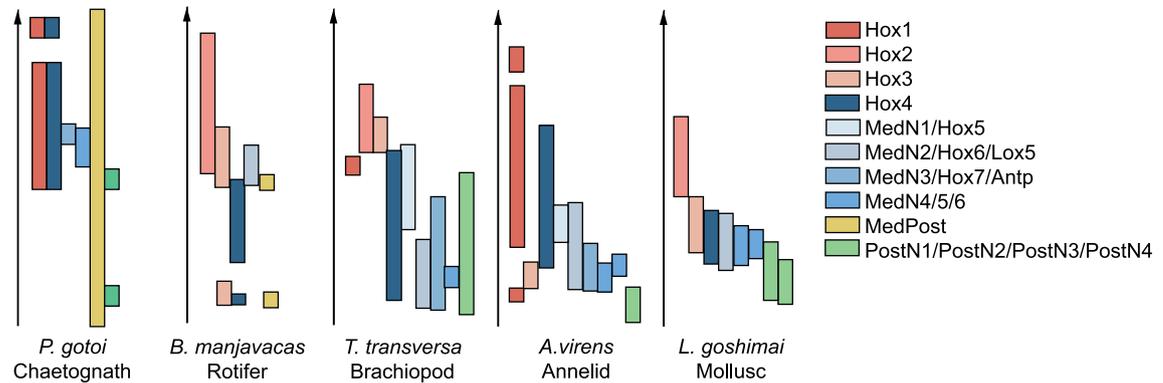
**a**



**b**

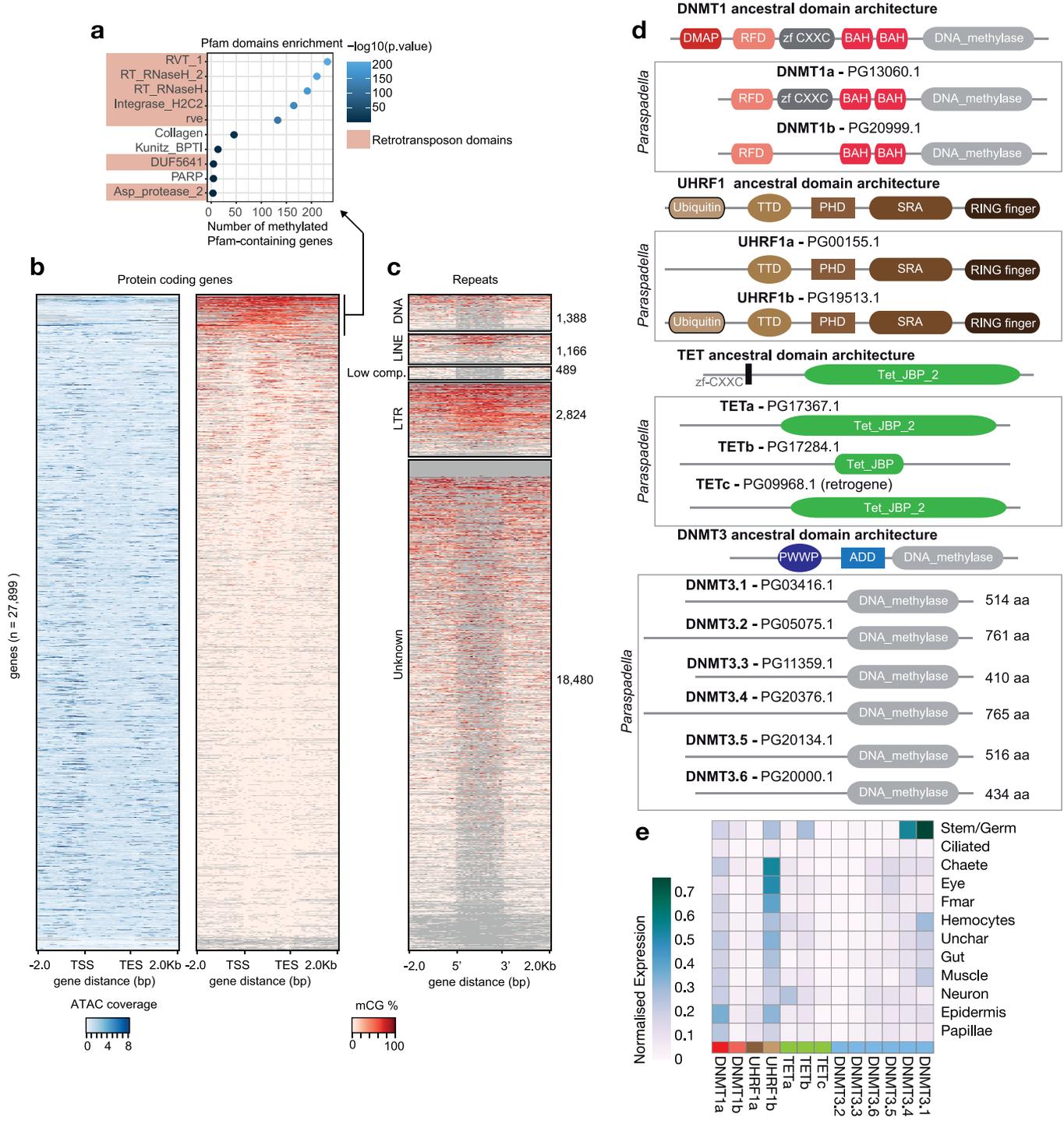


**c**



**Extended Data Fig. 9 | Hox cluster and Hox expression in *P. gotoi*.** **a**, Structure of *P. gotoi* Hox cluster with Hox genes labeled in red, high-confidence genes with high expression (FPKM > 5) and homology outside chaetognath in blue and other putative genes in grey. Such high-confidence genes include multi-exonic genes encoding for collagen-related or ion-channel peptides.

The RNA-seq (dark red), the ATAC-seq signal (blue) and the methylated GC (GC levels) are also displayed. **b**, Dotplot of expression of Hox genes in single-cell clusters showing cluster-specific expression of some Hox genes. **c**, Collinearity schematics for selected spiralian from this work and from Fröbius et al.<sup>18</sup> and Gąsiorowski et al.<sup>104</sup>.



**Extended Data Fig. 10 | A simplified methylation toolkit controls transposable elements in *P. gotoi*.** **a**, Methylation is only found on transposon-derived sequences. **a**) Number of methylated genes in *Paraspadella* (mCG > 20%) that contains enriched Pfam domains (Two sided Fisher exact test p-value < 0.01). Only the top most abundant domains are shown, with all those known to be associated with retrotransposons shaded in red. **b**, Heatmap showing the ATAC-seq and Enzymatic Methyl-seq signal across all gene models, sorted by methylation intensity. **c**, Heatmap showing the methylation levels on

the major types of repetitive elements in the *Paraspadella* genome. Only >500 bp repeats/inserts are displayed. Domain architecture of DNA methylation toolkit in *Paraspadella*. **d**, The domain architecture of DNMT1, UHRF1, TET and DNMT3 genes in an ancestral configuration and in *P. gotoi*, defined by PfamA domain presence/absence. RFD (PF12047), zf-CXXC (PF02008), BAH (PF01426), DNA\_methylase (PF00145), TTD (PF12148), PHD (PF00628), SRA (PF02182), Ring Finger (PF00097), TET\_JBP (PF12851). **e**, Expression of *P. gotoi* methylation toolkit in broad cell types.