Annelid functional genomics reveal the origins of bilaterian life cycles

Francisco M. Martín-Zamora[1,§], Yan Liang[1,§], Kero Guynes[1], Allan M. Carrillo-Baltodano[1], Billie E. Davies[1], Rory D. Donnellan[1], Yongkai Tan[2], Giacomo Moggioli[1], Océane Seudre[1], Martin Tran[1,3], Kate Mortimer[4], Nicholas M. Luscombe[2], Andreas Hejnol[5,6], Ferdinand Marlétaz[7,*], José M. Martín-Durán[1,*]

§ These authors contributed equally.

[1]School of Biological and Behavioural Sciences, Queen Mary University of London, London, United Kingdom

[2]Genomics and Regulatory Systems Unit, Okinawa Institute of Science and Technology Graduate University, Okinawa, Japan

[3]Department of Infectious Disease, Imperial College London, London, United Kingdom

[4]Department of Natural Sciences, Amgueddfa Cymru – Museum Wales, Cardiff, United Kingdom

[5]Department of Biological Sciences, University of Bergen, Bergen, Norway

[6]Institute of Zoology and Evolutionary Research, Faculty of Biological Sciences, Friedrich Schiller University Jena, Jena, Germany

[7]Department of Genetics, Evolution and Environment, University College London, London, United Kingdom

* Correspondence: José M. Martín-Durán (chema.martin@qmul.ac.uk), Ferdinand Marlétaz (f.marletaz@ucl.ac.uk)

Indirect development with an intermediate larva exists in all major animal lineages[1], making larvae central to most scenarios of animal evolution[2-11]. Yet how larvae evolved remains disputed. Here we show that temporal shifts, i.e., heterochronies, in trunk formation underpin the diversification of larvae and bilaterian life cycles. Combining chromosome-scale genome sequencing in the annelid *Owenia fusiformis* with transcriptomic and epigenomic profiling during the life cycles of this and two other annelids, we found that trunk development is deferred to pre-metamorphic stages in the feeding larva of *O. fusiformis* but starts after gastrulation in the non-feeding larva with gradual metamorphosis of *Capitella teleta* and the direct developing embryo of *Dimorphilus gyrociliatus*. Accordingly, the embryos of *O. fusiformis* develop first into an enlarged anterior domain that forms larval tissues and the adult head[12]. Notably, this also occurs in the so-called "head larvae" of other bilaterians[13-17], with whom the *O. fusiformis* larva shows extensive transcriptomic similarities. Together, our findings suggest that the temporal decoupling of head and trunk formation, as maximally observed in "head larvae", allowed larval evolution in Bilateria, which diverges from prevailing scenarios that propose either co-option[9,10] or innovation[11] of gene regulatory programmes to explain larva and adult origins.

Many animal embryos develop into a larva that metamorphoses into a sexually competent adult[1]. Larvae are morphologically and ecologically diverse and given their broad phylogenetic distribution, they are central to major scenarios of animal evolution[2-11]. However, these scenarios dissent on whether larvae are ancestral[2-6] or secondarily evolved[9,10] and on the mechanisms that facilitated the evolution of larvae[2,9-11]. Therefore, larval origins—and their importance to explain animal evolution—are still contentious.

The trochophore is a widespread larval type characterised by an apical sensory organ and a pre-oral locomotive ciliary band[18] that is typically assigned to Annelida and Mollusca. Annelids, however, show diverse life cycles and larval morphologies, including species with direct and indirect development and either planktotrophic or lecithotrophic larvae[19]. Notably, the groups Oweniidae and Magelonidae—which form Oweniida, the sister taxon to all other annelids[20]—have distinctive planktotrophic larvae (Fig. 1a; Extended Data Fig. 1a). In particular, the oweniidae larva, referred to as "mitraria"[12], has an enlarged pre-oral region and a bundle of posterior chaetae, as well as a pair of nephridia and a long monociliated ciliary band alike those of phylogenetically distant larvae of echinoderms and hemichordates[21,22]. Yet oweniids show many developmental characters considered ancestral to Annelida, and even Spiralia as a whole[23,24], including similarities in larval molecular patterns with other trochophore and bilaterian larvae[22,23,25,26]. Therefore, the diversity of life cycles and larval forms but generally conserved early embryogenesis and adult body plans of Annelida is an excellent model to investigate how larval traits evolve and to formulate and assess hypotheses on the origin of larvae and animal life cycles.

O. fusiformis *has a conserved genome*

To investigate how larvae evolved in Annelida, we first generated a chromosome-scale reference assembly for the oweniid *Owenia fusiformis* (Fig. 1b, inset). The haploid assembly spans 505.8 Mb and has 12 chromosome-scale scaffolds (Supplementary Fig. 1). Almost half of the assembly (43.02%) consists of repeats (Extended Data Fig. 1b, c) and we annotated 26,966 protein-coding genes and 31,903 transcripts, representing a nearly complete (97.5%) set of metazoan BUSCO genes (Supplementary Fig. 1). Gene family reconstruction and gene content analysis nested *O. fusiformis* with other non-annelid spiralians and taxa with slow-evolving genomes (Fig. 1b; Extended Data 1d, e), supporting that *O. fusiformis* has fewer

gene family gains and losses and retains more ancestral metazoan orthogroups than other annelid taxa (Fig. 1c; Extended Data Fig. 1f, g). Indeed, *O. fusiformis* has a *chordin* ortholog, a bone morphogenetic protein (BMP) inhibitor involved in dorsoventral patterning thought to be lost in annelids[27] and that is asymmetrically expressed around the blastopore of the gastrula and larval mouth in *O. fusiformis* (Extended Data Fig. 2). Moreover, *O. fusiformis* has globally retained the ancestral bilaterian linkage, exhibiting chromosomal fusions that are present in molluscs and even nemerteans, and less lineage-specific chromosomal rearrangements than other annelids (Fig. 1d; Extended Data Fig. 1h, i). Therefore, *O. fusiformis* shows a more complete gene repertoire and ancestral syntenic chromosomal organisation than other annelids, which together with its phylogenetic position and conserved early embryogenesis[23,24] makes it a key lineage to reconstruct the evolution of Annelida, and Spiralia generally.

Heterochronies in gene expression

To identify transcriptomic changes underpinning distinct life cycles in Annelida, we compared temporal series of embryonic, larval and competent/juvenile transcriptomes of *O. fusiformis* and *C. teleta*, two indirect developers with planktotrophic and lecithotrophic[28] larvae, respectively, and *D. gyrociliatus*, a direct developer[29,30] (Fig. 2a). Transcriptional dynamics during early embryogenesis are overall similar among these species (Supplementary Fig. 3). While *C. teleta* and *D. gyrociliatus* show increasing transcriptomic divergence with each other as they develop into adult stages, the maximal transcriptomic divergence between these annelids and *O. fusiformis* occurs at the mitraria stage (Extended Data Fig. 3a, b). Soft clustering of all expressed transcripts produced 12 distinct groups of temporally co-regulated genes in *O. fusiformis* and *C. teleta*, and 9 clusters in *D. gyrociliatus* (Extended Data Fig. 3c–e), expressed gradually along the life cycle of all three species. Only

one cluster in each species shows a bimodal activation at early embryogenesis and in the competent larva, juvenile or adult forms, consistently involving genes enriched for core cellular processes (Extended Data Fig. 3f). Indeed, translation and metabolism predominate in clusters of early development in the three annelids, while cell communication and signalling, morpho- and organogenesis are enriched in later stages of development (Extended Data Fig. 3f). Therefore, regardless of the life cycle, transcriptional dynamics are generally conserved during annelid development, yet adults and the planktotrophic larva are the most transcriptionally distinct stages.

To identify the genes underlying the transcriptional differences at larval and adult stages, we performed pairwise inter-species comparisons of gene and transcription factor composition among clusters of temporally co-regulated genes (Fig. 2b, c; Extended Data Fig. 4a, b). Early followed by late clusters are the most conserved in the three comparisons when all genes are considered (Extended Data Fig. 4c, d). However, transcription factors used in post-larval stages in indirect development are consistently shifted to early embryogenesis in direct development (Fig. 2c; Extended Data Fig. 4c, e). In both *O. fusiformis* and *C. teleta*, this shift involves 28 transcription factors that act in a variety of developmental processes, from nervous system (e.g., *pax6*[31]) and mesoderm (e.g., *foxF*[26]) formation to axial patterning (e.g., *Hox1* and *Hox4*[32]) (Supplementary Fig. 12). Notably, the overall expression of these 28 genes is also temporally shifted between indirect developing annelids, with the maximum level of expression occurring earlier in *C. teleta* than in *O. fusiformis* (Fig. 2d). Additionally, 2,583 genes also exhibit temporal shifts between the larvae of *O. fusiformis* and *C. teleta* (Fig. 2e), including 105 transcription factors, but mostly enzymes and structural genes that likely reflect the different biology of these two larvae (Extended Data Fig. 4f, g; Supplementary Fig. 13–16). Therefore, temporal shifts (i.e., heterochronies) in the use of shared genetic

programmes and regulatory genes correlate with and might account for life cycle and larval differences in Annelida.

Different timings of trunk development

Homeodomain transcription factors are the largest class among the 28 transcription factors with temporal expression shifts between direct and indirect developing annelids (Supplementary Fig. 12). Indeed, homeodomain genes are enriched in the competent larva in *O. fusiformis* but are prevalent from stage 5 larva onwards in *C. teleta* (Extended Data Fig. 4h). Accordingly, *Hox* genes, which regionalise the bilaterian trunk along the anteroposterior axis[33], are strongly upregulated in the competent mitraria larva (Extended Data Fig. 5a, b). *Owenia fusiformis* has a conserved complement of 11 *Hox* genes—like *C. teleta*[32]—arranged as a compact, ordered cluster in chromosome 1, except for *Post1*, which is located downstream on that same chromosome (Extended Data Fig. 5c, d). *Capitella teleta* and *D. gyrociliatus* start expressing *Hox* genes along their trunks[30,32] during or soon after gastrulation (Extended Data Fig. 5e). *Owenia fusiformis*, however, does not express *Hox* genes during embryogenesis but in the trunk rudiment during larval growth, already in an anteroposterior staggered pattern, as later observed in the juvenile (Fig. 3a; Extended Data Fig. 5e–h). This late activation of *Hox* genes is not unique to *O. fusiformis*, but also occurs for most *Hox* genes in the planktotrophic trochophore of the echiuran annelid *Urechis unicinctus*[34] (Extended Data Fig. 5e). Therefore, the spatially collinear *Hox* code along the trunk is established at distinct developmental stages depending on the life cycle mode in Annelida.

To determine whether the difference in timings of trunk patterning is limited to the expression of *Hox* genes, we used tissue-specific adult transcriptomes to define a set of 1,655

anterior and 407 posterior and trunk genes in *O. fusiformis* (Extended Data Fig. 6a–d). While anterior genes are significantly more expressed during embryogenesis, posterior and trunk genes are upregulated at the mitraria stage and significantly outweigh the expression dynamics of anterior genes from that stage onwards (Fig. 3b; Extended Data Fig. 6e–f). Moreover, anterior, trunk, and posterior genes with spatially resolved expression follow different temporal dynamics in *O. fusiformis*, *C. teleta*, and *D. gyrociliatus*. In *O. fusiformis*, trunk[25] and posterior[24,26] genes concentrate in a small ventral area and around the anal opening of the larva and increase in spatial range and expression levels as the trunk forms (Extended Data Fig. 6g, h). Anterior genes[26,35], however, pattern most of the mitraria and their expression remains stable during development (Extended Data Fig. 6g, h). In contrast, posterior and anterior genes follow similar dynamics in *C. teleta,* and trunk genes upregulate already post-gastrula in both *C. teleta* and *D. gyrociliatus* (Extended Data Fig. 6i–l). Therefore, trunk development, which initially occurs from lateral growth of the trunk rudiment[12,28], is deferred to pre-metamorphic stages in planktotrophic annelid trochophores compared to annelids with lecithotrophic larvae and direct developers.

Heterochronies in Hox *regulation*

To investigate the genomic regulatory basis for the heterochronies in trunk development between annelid larvae, we profiled open chromatin regions at five equivalent developmental stages in *O. fusiformis* and *C. teleta* (Fig. 2a) and identified 63,726 and 44,368 consensus regulatory regions, respectively. In both species, open chromatin is more abundant within gene bodies (Extended Data Fig. 7a). There is, however, a general increase in promoter peaks (in *O. fusiformis*) and distant intergenic regulatory elements (in both species) during development (Extended Data Fig. 7b), and the largest changes in peak accessibility occur in the mitraria in *O. fusiformis* and stage 5 larva in *C. teleta* (Supplementary Fig. 18). In

O. fusiformis, most regulatory regions act before the start of trunk formation, while the numbers of accessible regions with a maximum of accessibility before and after the onset of trunk development are comparable in *C. teleta* (Extended Data Fig. 7c). Accordingly, regulation of genes involved in morpho- and organogenesis, as well as neurogenesis, concentrates in late clusters in *O. fusiformis*, but unfolds more continuously in *C. teleta* (Supplementary Fig. 23). Therefore, different dynamics of chromatin accessibility occur during development and larva formation in these two annelids.

To investigate the regulatory programmes controlling larva development in *O. fusiformis* and *C. teleta*, we predicted transcription factor-binding motifs on ATAC-seq peaks and identified 33 motifs common to both species that are robustly assigned to a known transcription factor class (Supplementary Fig. 29). Notably, the binding dynamics of these 33 motifs revealed a temporal shift of regulatory motifs acting between the mitraria and competent larva in *O. fusiformis* to the stage 4tt larva of *C. teleta* (Fig. 3c; Extended Data Fig. 7d–f). Seven motifs follow this pattern (Extended Data Fig. 7g, Supplementary Fig. 29), including one with high similarity to the human HOX/CDX/EVX motif archetype (Fig. 3d, e) that is overrepresented and upregulated based on its binding score at the competent stage in *O. fusiformis* (Extended Data Fig. 7h; Supplementary Fig. 30). Indeed, motif binding dynamics in regulatory elements assigned to *Hox* genes support a change of global regulation of the *Hox* cluster at the competent and stage 4tt larva stages in *O. fusiformis* and *C. teleta*, respectively (Fig. 3f; Supplementary Fig. 31), mirroring the transcriptional onset of these genes and the start of trunk development in the two species[32]. Motifs assigned to NKX and GATA factors, which are expressed in the developing trunk in both species[25,36], are amongst the most abundant bound motifs in the *Hox* cluster in both species (Extended Data Fig. 7i). However, only 39 one-to-one orthologs with bound HOX/CDX/EVX motifs at the maximum

of motif binding are common to *O. fusiformis* and *C. teleta* (Extended Data Fig. 7j). Therefore, different regulatory dynamics of the *Hox* cluster—possibly triggered by a reduced common set of upstream regulators—underpin temporal variability in *Hox* activity and downstream targets. These shifts likely promoted the developmental and morphological differences in trunk formation between planktotrophic and lecithotrophic annelid larvae.

Different dynamics of novel genes

Novel genes, which account for a significant proportion of some larval transcriptomes[6,37], could also contribute to and explain transcriptomic differences between annelid larvae. In *O. fusiformis*, *C. teleta* and *D. gyrociliatus*, genes of metazoan and pre-metazoan origin tend to peak, dominate, and be enriched at early development, whereas younger genes are more highly expressed in competent and juvenile stages (Extended Data Fig. 8a–e). Species-specific genes follow, however, lineage-specific dynamics (Supplementary Fig. 32), being, for instance, more expressed in the juveniles of *O. fusiformis* and *D. gyrociliatus*, but in the blastula and gastrula of *C. teleta* (and to some extent also at the blastula stage in *O. fusiformis*; Extended Data Fig. 8a, c, d). Species-specific genes are only enriched and overrepresented at larval stages in the case of *C. teleta* (Extended Data Fig. 8f–h). Therefore, genes of different evolutionary origins contribute to the development of annelid larvae, suggesting that the increased use of novel genes in some lophotrochozoan larvae[6,37] might be due to the evolution of lineage-specific larval traits.

Similarities between bilaterian larvae

To assess whether the transcriptional dynamics found in annelids are also observed in other metazoans, we extended our comparative transcriptomic approach to nine other animal lineages. In relative terms, global transcriptional dynamics between *O. fusiformis* and other

225 animals tend to be more dissimilar at early development than at juvenile and adult stages
226 (Fig. 4a; Extended Data Fig. 9a, b; Extended Data Fig. 10a). The exception is the direct
227 developer *Danio rerio*, for which the mitraria larva is the most dissimilar stage (Fig. 4a), as is
228 also the case when comparing *O. fusiformis* with the direct-developing annelid
229 *D. gyrociliatus* (Extended Data Fig. 3b). Notably, *O. fusiformis* shares maximal
230 transcriptomic similarities during larval phases with bilaterian species with planktotrophic
231 ciliated larvae and even cnidarian planulae (Fig. 4a; Extended Data Fig. 9a–e). Genes
232 involved in core cellular processes directly contribute to these similarities, likely reflecting
233 common structural and ecological needs of metazoan larvae (Extended Data Fig. 9f, g) but
234 transcription factor expression levels are also maximally similar between those species at
235 larval phases (Extended Data Fig. 9a, b, e). Therefore, adult development is generally more
236 similar[9] than early embryogenesis across major animal lineages, but phylogenetically distant
237 animal larvae also exhibit unexpected genome-wide transcriptional—and potentially
238 regulative—similarities.
239
240 **Discussion**
241 Our study provides an unprecedented perspective on life cycle evolution in Bilateria. The
242 planktotrophic larva of *O. fusiformis* defers trunk differentiation to late pre-metamorphic
243 stages and largely develops from anterior ectodermal domains. This occurs in other feeding
244 annelid larvae[38] (Extended Data Fig. 5f), and likely in Chaetopteriformia[39,40] too, and thus the
245 late differentiation of the adult trunk might be an ancestral trait to Annelida (Extended Data
246 Fig. 10b). Delaying trunk development to post-larval stages also occurs in phylogenetically
247 distant clades within Spiralia[16,17], Ecdysozoa[14,41], and Deuterostomia[15,42,43], whose larvae are
248 generally referred to as "head larvae"[13,14]. By contrast, non-feeding larvae[32,44] and direct
249 developers[30] in both Annelida and other bilaterian taxa[45,46] start to pattern their trunks with or

straight after the onset of anterior/head patterning, which always takes place before gastrulation in bilaterians[47,48]. Therefore, heterochronies in trunk development correlate with, and possibly account for, the evolution of different life cycles in animals (Fig. 4b). This differs from previously proposed mechanisms to explain the origins of animal life cycles, namely co-option of adult genes into larval-specific regulatory programmes[9,10] and independent evolution of adult gene regulatory modules[2,49].

Bilaterian "head larvae" could be lineage-specific innovations associated with the evolution of maximal indirect development[13,14,16] that evolved convergently by delaying trunk differentiation and *Hox* patterning (Fig. 4c). The similarities in larval molecular patterns[5,15,16] would then reflect ancient gene regulatory modules that were independently co-opted to develop analogous cell types and larval organs. Alternatively, the post-embryonic onset of trunk differentiation and *Hox* expression might be the most parsimonious ancestral state for Bilateria (Extended Data Fig. 10c, d). This could have facilitated the evolution of larvae, which would then originally share anterior genetic modules for their development (Fig. 4c). Regardless of the scenario and despite their limitations, our datasets highlight the importance of heterochronic changes for the diversification of bilaterian life cycles, uncovering a reduced set of candidate genes and regulatory motifs that might influence life cycle differences in Annelida, and perhaps even Bilateria. In the future, comparative functional studies of these and other genes will reveal how temporal changes in gene expression and regulation have shaped the evolution of animal larvae and adults.

Main References

1 Hall, B. K. & Wake, M. H. in The Origin and Evolution of Larval Forms (eds Brian K. Hall & Marvalee H. Wake) Ch. 1, 1-19 (Academic Press, 1999).

2 Nielsen, C. Animal phylogeny in the light of the trochaea theory. Biological Journal of the Linnean Society 25, 243-299 (2008).

3 Garstang, W. The origin and evolution of larval forms. 77-98 (British Association for the Advancement of Science, London, 1928).

4 Jägersten, G. Evolution of the Metazoan Life Cycle. (Academic Press, 1972).

5 Marlow, H. et al. Larval body patterning and apical organs are conserved in animal evolution. BMC Biol 12, 7 (2014).

6 Wang, J. et al. Evolutionary transcriptomics of metazoan biphasic life cycle supports a single intercalation origin of metazoan larvae. Nat Ecol Evol 4, 725-736 (2020).

7 Salvini-Plawen, L. v. On the origin and evolution of the lower Metazoa. Z Zool Syst Evol Forsch 16, 40-87 (1978).

8 Hazsprunar, G., Salvini-Plawen, L. v. & Rieger, R. M. Larval Planktotrophy — A Primitive Trait in the Bilateria? Acta Zoologica 76, 141-154 (1995).

9 Raff, R. A. Origins of the other metazoan body plans: the evolution of larval forms. Philos Trans R Soc Lond B Biol Sci 363, 1473-1479 (2008).

10 Sly, B. J., Snoke, M. S. & Raff, R. A. Who came first - larvae or adults? Origins of bilaterian metazoan larvae. Int J Dev Biol 47, 623-632 (2003).

11 Davidson, E. H., Peterson, K. J. & Cameron, R. A. Origin of bilaterian body plans: evolution of developmental regulatory mechanisms. Science 270, 1319-1325 (1995).

12 Wilson, D. P. IV. On the Mitraria Larva of *Owenia fusiformis* Delle Chiaje. Philosophical Transactions of the Royal Society of London. Series B, Containing Papers of a Biological Character 221, 231-334 (1932).

13 Lacalli, T. C. Protochordate body plan and the evolutionary role of larvae: old controversies resolved? Can. J. Zool. 83, 216-224 (2005).

14 Strathman, R. Multiple origins of feeding head larvae by the Early Cambrian. Can. J. Zool. 98, 761-776 (2020).

15 Gonzalez, P., Uhlinger, K. R. & Lowe, C. J. The Adult Body Plan of Indirect Developing Hemichordates Develops by Adding a Hox-Patterned Trunk to an Anterior Larval Territory. Curr Biol 27, 87-95 (2017).

16 Gasiorowski, L. & Hejnol, A. Hox gene expression during development of the phoronid *Phoronopsis harmeri*. EvoDevo 11, 2, (2020).

17 Hiebert, L. S. & Maslakova, S. A. Hox genes pattern the anterior-posterior axis of the juvenile but not the larva in a maximally indirect developing invertebrate, *Micrura alaskensis* (Nemertea). BMC Biol 13, 23 (2015).

18 Rouse, G. W. Trochophore concepts: ciliary bands and the evolution of larvae in spiralian Metazoa. Biological Journal of the Linnean Society 66, 411-464 (2008).

19 Rouse, G. W. Polychaetes have evolved feeding larvae numerous times. Bulletin of Marine Science 67, 391-409 (2000).

20 Rouse, G. W., Pleijel, F. & Tilic, E. Annelida. (Oxford University Press, 2022).

21 Smart, T. I. & Von Dassow, G. Unusual development of the mitraria larva in the polychaete *Owenia collaris*. Biol Bull 217, 253-268 (2009).

22 Gasiorowski, L. et al. Molecular evidence for a single origin of ultrafiltration-based excretory organs. Curr Biol 31, 3629-3638 e3622 (2021).

23 Carrillo-Baltodano, A. M., Seudre, O., Guynes, K. & Martin-Duran, J. M. Early embryogenesis and organogenesis in the annelid *Owenia fusiformis*. EvoDevo 12, 5 (2021).

24 Seudre, O., Carrillo-Baltodano, A. M., Liang, Y. & Martín-Durán, J. M. ERK1/2 is an ancestral organising signal in spiral cleavage. Nat Commun 13, 2286 (2022).

25 Martin-Duran, J. M. et al. Convergent evolution of bilaterian nerve cords. Nature 553, 45-50 (2018).

26 Martin-Duran, J. M., Passamaneck, Y. J., Martindale, M. Q. & Hejnol, A. The developmental basis for the recurrent evolution of deuterostomy and protostomy. Nat Ecol Evol 1, 5 (2016).

27 Tan, S., Huan, P. & Liu, B. Molluscan dorsal-ventral patterning relying on BMP2/4 and Chordin provides insights into spiralian development and evolution. Mol Biol Evol (2021).

28 Seaver, E. C., Thamm, K. & Hill, S. D. Growth patterns during segmentation in the two polychaete annelids, *Capitella* sp. I and *Hydroides elegans*: comparisons at distinct life history stages. Evol Dev 7, 312-326 (2005).

29 Kerbl, A., Fofanova, E. G., Mayorova, T. D., Voronezhskaya, E. E. & Worsaae, K. Comparison of neuromuscular development in two dinophilid species (Annelida) suggests progenetic origin of *Dinophilus gyrociliatus*. Front Zool 13, 49 (2016).

30 Martin-Duran, J. M. et al. Conservative route to genome compaction in a miniature annelid. Nat Ecol Evol 5, 231-242 (2021).

31 Klann, M. & Seaver, E. C. Functional role of *pax6* during eye and nervous system development in the annelid *Capitella teleta*. Dev Biol 456, 86-103 (2019).

32 Fröbius, A. C., Matus, D. Q. & Seaver, E. C. Genomic organization and expression demonstrate spatial and temporal Hox gene colinearity in the lophotrochozoan *Capitella* sp. I. PLoS One 3, e4004 (2008).

33 Pearson, J. C., Lemons, D. & McGinnis, W. Modulating Hox gene functions during animal body patterning. Nat Rev Genet 6, 893-904, doi:10.1038/nrg1726 (2005).

34 Wei, M. et al. Echiuran Hox genes provide new insights into the correspondence between Hox subcluster organization and collinearity pattern. Proc Biol Sci 289, 20220705 (2022).

35 Seudre, O. et al. The Fox gene repertoire in the annelid *Owenia fusiformis* reveals multiple expansions of the foxQ2 class in Spiralia. bioRxiv, 2022.2003.2002.482670 (2022).

36 Boyle, M. J. & Seaver, E. C. Developmental expression of *foxA* and *gata* genes during gut formation in the polychaete annelid, *Capitella* sp. I. Evol Dev 10, 89-105 (2008).

37 Wu, L., Ferger, K. E. & Lambert, J. D. Gene Expression Does Not Support the Developmental Hourglass Model in Three Animals with Spiralian Development. Mol Biol Evol 36, 1373-1383 (2019).

38 Woltereck, R. Trochophora-Studien I. Histogie der larve und die Entstehung des Annelids bei den Polygordius-Arten der Nordsee. Zoologica 13 (1902).

39 Irvine, S. Q. & Martindale, M. Q. Expression patterns of anterior Hox genes in the polychaete *Chaetopterus*: correlation with morphological boundaries. Dev Biol 217, 333-351 (2000).

40 Peterson, K. J., Irvine, S. Q., Cameron, R. A. & Davidson, E. H. Quantitative assessment of Hox complex expression in the indirect development of the polychaete annelid Chaetopterus sp. Proc Natl Acad Sci U S A 97, 4487-4492 (2000).

41 Brenneis, G., Bogomolova, E. V., Arango, C. P. & Krapp, F. From egg to "no-body": an overview and revision of developmental pathways in the ancient arthropod lineage Pycnogonida. Front Zool 14, 6 (2017).

42 Arenas-Mena, C., Cameron, A. R. & Davidson, E. H. Spatial expression of Hox cluster genes in the ontogeny of a sea urchin. Development 127, 4631-4643 (2000).

43 Arenas-Mena, C., Martinez, P., Cameron, R. A. & Davidson, E. H. Expression of the Hox gene complex in the indirect development of a sea urchin. Proc Natl Acad Sci U S A 95, 13062-13067 (1998).

44 Kulakova, M. et al. Hox gene expression in larval development of the polychaetes *Nereis virens* and *Platynereis dumerilii* (Annelida, Lophotrochozoa). Dev Genes Evol 217, 39-54 (2007).

45 Aronowicz, J. & Lowe, C. J. Hox gene expression in the hemichordate *Saccoglossus kowalevskii* and the evolution of deuterostome nervous systems. Integr Comp Biol 46, 890-901 (2006).

46 Pascual-Anaya, J. et al. Broken colinearity of the amphioxus Hox cluster. Evodevo 3, 28 (2012).

47 Carrillo-Baltodano, A. M. & Meyer, N. P. Decoupling brain from nerve cord development in the annelid *Capitella teleta*: Insights into the evolution of nervous systems. Dev Biol 431, 134-144 (2017).

48 Sur, A., Magie, C. R., Seaver, E. C. & Meyer, N. P. Spatiotemporal regulation of nervous system development in the annelid *Capitella teleta*. Evodevo 8, 13 (2017).

49 Nielsen, C. Animal Evolution. Interrelationships of the Living Phyla. 3rd edn, (Oxford University Press, 2012).

Figure Legends

Figure 1 | *Owenia fusiformis* has a unique larva and a conservatively evolving genome. **a**, The larvae of oweniids and magelonids are unlike other annelid larvae. Differential interface contrast (DIC) images and z-stack confocal laser scanning views of a *O. fusiformis* mitraria and a *Magelona* spp. larva stained for DAPI and acetylated α-tubulin. **b,** Principal component analysis of metazoan gene complements demonstrates that *O. fusiformis* (image of an *O. fusiformis* adult in the inset) clusters with other lineages with conservatively evolving gene complements. See Extended Data Fig. 1e for a fully labelled graph. **c,** Percentage of retained pre-metazoan and metazoan orthogroups per species. Dotted vertical line represents the value for *O. fusiformis*. **d,** Karyotypic correspondence between *O. fusiformis* and *Pecten maximus*, which exemplifies the ancestral spiralian chromosome complement. Each colour represents an ancestral bilaterian linkage group. Schematic drawings are not to scale. at: apical tuft; an: anus; ch: chaetae; he: head; mo: mouth; pt: prototroch; tt: telotroch. Scale bars, 50 μm in **a**, 2.5 mm in **b**.

Figure 2 | Heterochronies in gene regulatory programmes underpin annelid life cycle diversification. a, Experimental design of the comparative developmental RNA-seq and ATAC-seq time courses. Orange circles highlight stages of *O. fusiformis*, *C. teleta*, and *D. gyrociliatus* development sampled for bulk RNA-seq. Orange circles with a blue inner dot highlight developmental stages sampled for ATAC-seq. **b**, **c**, Similarity heatmaps showcasing the orthogroup overlap between the transcription factors contained in clusters of co-regulated genes obtained by soft *k*-means clustering, between all three studied annelid taxa. Time points associated to key clusters are shown for all three species. Dotted black lines in **c** highlight the sharp timing expression differences of a significant number of transcription factors shifted from post-larval expression in indirect developers to early embryogenesis in

411 *D. gyrociliatus*. *P*-values were derived from upper-tail hypergeometric tests and Benjamini-
412 Hochberg-adjusted. **d,** Average expression dynamics of the 28 single copy ortholog
413 transcription factors shifted from late expression in both *O. fusiformis* and *C. teleta* to early
414 expression in *D. gyrociliatus*. Curves are locally estimated scatterplot smoothings, coloured
415 shaded areas represent standard error of the mean. **e**, Heatmap of relative similarity based on
416 whole genome orthogroup overlap analysis by quadrants between pre-larval (early) and post-
417 larval (late) clusters in *O. fusiformis* and *C. teleta*. Dotted black lines denote the groups of
418 genes and transcription factors under heterochronies between both species.
419
420 **Figure 3 | Trunk development is delayed to pre-metamorphosis in *O. fusiformis*.**
421 **a,** Representative images from three independent analyses of *Hox* gene expression by whole
422 mount *in situ* hybridisation at the mitraria larva, pre-competent larva, and juvenile stages of
423 *O. fusiformis*. Only *Hox3* is expressed at the mitraria stage (white arrow). *Hox* genes show
424 spatial collinearity along the anteroposterior axis at the developing trunk of the pre-competent
425 larva (white arrows), and in the juvenile. Dotted lines in the competent larva panels indicate
426 background from the midgut. Black arrowheads in the juvenile panels indicate head to trunk
427 boundary. cs: chaetal sack; mg: mid gut; mo: mouth. **b,** Average expression dynamics of
428 anterior (n = 1,655), and posterior and trunk genes (n = 407) expressed in corresponding adult
429 tissues during *O. fusiformis* development. *P*-values were derived from two-tailed Student's *t*-
430 tests and adjusted with the Bonferroni method for multiple testing correction. ***: *P*-
431 value < 0.001; n.s.: not significant. Centre lines in boxplots are the median, box is the
432 interquartile range (IQR), and whiskers are the first or third quartile ± 1.5 × IQR.
433 **c,** Correlation matrices of transcription factor binding score (TFBS). Dotted black line
434 highlights the high TFBS correlation and heterochrony between the mitraria and competent
435 larvae of *O. fusiformis* and the stage 4tt larva of *C. teleta*. **d**, Sequence logo of the annelid

436 archetype (top) shows substantial similarity to the human homolog (bottom). **e,** TFBS
437 dynamics for the annelid HOX/CDX/EVX motif during *O. fusiformis* (purple) and *C. teleta*
438 (blue) development. **f,** Average TFBS dynamics of all motifs in the peaks of the *Hox* cluster.
439 Curves are locally estimated scatterplot smoothings, coloured shaded areas represent standard
440 error of the mean. Scale bars in **a**, 50 µm in larval stages and 100 µm in the juvenile.
441
442 **Figure 4 | The evolution of life cycles in Annelida and Bilateria. a,** Heatmaps of pairwise
443 normalised Jensen-Shannon divergence (JSD) between *O. fusiformis* and *C. gigas*,
444 *S. purpuratus*, *D. rerio* and *N. vectensis.* Asterisks indicate the stages of minimal JSD of each
445 species to the larval phase of *O. fusiformis*. Larval phases are highlighted in green. Average
446 relative JSD of the stages of minimal divergence to each *O. fusiformis* stage is shown on top
447 of each heatmap. Confidence intervals represent standard deviation from 250 bootstrap
448 resamplings of the ortholog sets. **b**, Schematic drawings of the three main types of life cycles
449 and the timing of *Hox* gene expression in bilaterians. Compared to indirect development with
450 feeding larvae, lineages with non-feeding larvae and direct development pre-displace (i.e.,
451 initiate earlier) trunk differentiation and *Hox* gene expression. Larval organs are reduced in
452 non-feeding larvae and absent in direct development. **c**, Proposed alternative scenarios for the
453 evolution of maximal indirect development with “head larvae” in Bilateria. Top, head larvae
454 evolved convergently by repeatedly shifting trunk development (as seen by *Hox* gene
455 expression) to pre-metamorphic stages. Bottom, head and trunk development were ancestrally
456 temporally decoupled, which could have facilitated the evolution of head larvae in different
457 bilaterian lineages.

Methods

Adult culture, spawning and in vitro fertilisation

Sexually mature *O. fusiformis* adults were collected from subtidal waters near the Station Biologique de Roscoff and cultured in the lab as described before[23]. *In vitro* fertilisations and collections of embryonic and larval stages were performed as previously described[23]. *Capitella teleta* Blake, Grassle & Eckelbarger, 2009 was cultured, grown, and sifted, and its embryos and larvae were collected following established protocols[28]. *Magelona* spp. were collected in muddy sand from the intertidal of Berwick-upon-Tweed, Northumberland, NE England (~55.766781, -1.984587) and kept initially in aquaria at the National Museum Cardiff before their transfer to Queen Mary University of London, where they were kept in aquaria with artificial sea water.

Genome size measurements

To estimate the haploid DNA nuclear content of *O. fusiformis*, we used a flow cytometer Partex CyFlow Space fitted with a Cobalt Samba green laser (532 nm, 100 mW) and the built-in software FloMax v.2.82, as described for the annelid *Dimorphilus gyrociliatus*[23], with adult individuals of *Drosophila melanogaster* as reference. Additionally, we used Jellyfish v.2.3[50] to count and generate a 31-mer histogram from adaptor-cleaned, short-read Illumina reads (see section below), and GenomeScope 2.0[51] to obtain an in-silico estimation of the genome size and heterozygosity of *O. fusiformis*.

Genome sequencing, assembly, and quality check

Ultra-high molecular weight (UHMW) genomic DNA (gDNA) was extracted following the Bionano genomics IrysPrep agar-based, animal tissue protocol using sperm from a single *O. fusiformis* male. UHMW gDNA was cleaned up using a salt:chloroform wash following

PacBio's recommendations before long-read sequencing using PacBio v3.0 chemistry at the University of California Berkeley. A total of 16 SMRT cells of PacBio Sequel were used for sequencing with 600 min movie time, producing a total of 170.07 Gb of data (10.72 million reads, N50 read length between 25.75 kb and 30.75 kb). In addition, we used UHMW gDNA of that same individual to generate a 10x Genomics linked reads library, which we sequenced in an Illumina HiSeq4000 at Okinawa Institute of Science and Technology (OIST) to produce 28.62 Gb of data (141.66 million read pairs). PacBio reads were assembled with CANU v.8.3rc2[52] assuming 'batOptions="-dg 3 -db 3 -dr 1 -ca 500 -cp 50' and 'correctedErrorRate=0.065'. Pacbio reads were remapped using pbalign v.0.3.2 and the assembly polished once using Arrow (genomicconsensus, v2.3.2). Then Illumina paired end reads generated with the 10x Genomics linked reads were extracted, remapped using bwa mem v.0.7.17[53] and used for polishing with Racon v.1.16[54]. Bionano Genomics optical mapping data was used to scaffold the PacBio-based assembly, which was de-haploidised with purge_haplotigs v.1.0.4[55] setting cut-offs at 35, 85 and 70x coverages to reconstruct a high-quality haploid reference assembly. HiC-based chromosome scaffolding was performed as described below. Merqury v.1.1[56] and BUSCO v.5[57] were used to assess genome completeness and evaluate the quality of the assembly (Supplementary Fig. 1).

Transcriptome sequencing

Fourteen samples spanning key developmental time points of *O. fusiformis* life cycle, including active oocyte, zygote, 2-cell, 4-cell, and 8-cell stages, 3 hours post-fertilisation (hpf), 4 hpf, coeloblastula (5 hpf), gastrula (9 hpf), axial elongation (13 hpf), early larva (18 hpf), mitraria larva (27 hpf), pre-metamorphic competent larva (3 weeks post-fertilisation, wpf) and post-metamorphic juvenile were collected in duplicates (except for the latter), flash frozen in liquid nitrogen and stored at –80 °C for total RNA extraction. Samples

within replicates were paired, with each one containing ~300 embryos or ~150 larvae coming from the same *in vitro* fertilisation. Nine further samples from adult tissues and body regions (blood vessel, body wall, midgut, prostomium, head, ovary, retractor muscle, tail, and testes) were also collected as described above. Likewise, further five samples spanning post-cleavage time points of *C. teleta,* including 64 cells and gastrula stages, and stage 4tt, stage 5, and stage 7 larval stages, were also collected in duplicates. Total RNA was isolated with the Monarch Total RNA Miniprep Kit (New England Biolabs, NEB) following supplier's recommendations. Total RNA samples from developmental stages from both *O. fusiformis* and *C. teleta* were used to prep strand-specific mRNA Illumina libraries that were sequenced at the Oxford Genomics Centre (University of Oxford, UK) over three lanes of an Illumina NovaSeq6000 system in 2 × 150 bases mode to a depth of ~50 M reads (Supplementary Tables 13, 16). Adult tissue samples were sequenced at BGI on a BGISeq-500 platform in 2 × 100 bases mode to a depth of ~25 M reads (Supplementary Table 49).

Annotation of repeats and transposable elements (TEs)

RepeatModeler v.2.0.1[58] and RepBase were used to construct a *de novo* repeat library for *O. fusiformis*, which was then filtered for *bona fide* genes using the predicted proteome of *C. teleta*. Briefly, we used DIAMOND v.0.9.22[59] with an *e*-value cut-off of 1e-10 to identify sequences in the *de novo* repeat library with significant similarity to protein coding genes in *C. teleta* that are not transposable elements. Sequences with a significant hit were manually inspected to verify they were not transposable elements and if so, they were manually removed from the *de novo* repeat library. The filtered consensus repeat predictions were then used to annotate the genome assembly of *O. fusiformis* with RepeatMasker "open-4.0". We next used LTR_finder v.1.07[60], a structural search algorithm, to identify and annotate Long Tandem Repeats (LTR). Finally, we generated a consensus set of repeats by merging

RepeatMasker and LTR_finder predictions with RepeatCraft[61], using default parameters but a maximum LTR size of 25 kb (as derived from the LTR_finder annotation) (Supplementary Table 1). The general feature format (gff) and fasta files with the annotation of TEs and repeats are available in the GitHub repository (see Data Availability section).

Gene prediction and functional annotation

We used SAMtools v.1.9[62] and the annotation of repeats to soft-mask *O. fusiformis* genome assembly before gene prediction. We then mapped all embryonic and adult transcriptomes and a publicly available dataset[63] (SRR1222288) with STAR v. 2.5.3a[64] after removing low-quality read pairs and read pairs containing Illumina sequencing adapters with trimmomatic v.0.39[65]. StringTie v.1.3.6[66] was used to convert STAR alignments into gene transfer format (GTF) files and Portcullis v.1.1.2[67] to generate a curated set of splice junctions. Additionally, we generated *de novo* transcriptome assemblies for all samples with Trinity v.2.5.1[68] with default parameters, which were thereafter mapped to the soft-masked assembly with GMAP v.2020-04-08[69]. We then ran the default Mikado v.2.1 pipeline[70] to merge all transcriptomic evidence and reliable splice junctions into a single set of best-supported transcripts and gene models. From this merged dataset, we filtered full-length, non-redundant transcripts with a BLAST hit on at least 50 % of their length and at least two exons to obtain a gene set that we used to train Augustus v.3.2.3[71]. Simultaneously, we used the Mikado gene annotation and Portcullis splice junctions to generate confident sets of exon and intron hints, respectively. We also ran Exonerate v.2.4.0[72] to generate spliced alignments of the proteome of *C. teleta* proteome on *O. fusiformis* soft-masked genome assembly to obtain further gene hints. We then merged all exon and intron hints into a single dataset which we passed to Augustus v.3.2.3[71] for *ab initio* gene prediction. Finally, PASA v.2.3.3[73] was used to combine RNA-seq and *ab initio* gene models into a final gene set, from which spurious predictions with in-frame

558 STOP codons (228 gene models), predictions that overlapped with repeats (5,779 gene
559 models) and that had high similarity to transposable elements in the RepeatPeps.lib database
560 (2,450 models) were removed. This filtered gene set includes 26,966 genes, encompassing
561 31,903 different transcripts (Supplementary Fig. 1). To assess the completeness of this
562 annotation, we ran BUSCO v.5[57] in proteome mode, resulting in 97.7 % of the core genes
563 present. Moreover, 31,678 out of the 31,903 (99.29%) of the filtered transcripts are supported
564 by RNA-seq data and 80.69% of the transcripts have a significant BLAST match (*e*-value
565 cut-off < 0.001) to a previously annotated annelid gene (database containing non-redundant
566 proteomes of the high-quality annelid genomes of *C. teleta*, *D. gyrociliatus*, *E. andrei*,
567 *L. luymesi*, *P. echinospica*, *R. pachyptila* and *S. benedicti*). A similar functional annotation
568 approach was followed to re-annotate the genome of *C. teleta* with the new RNA-seq data,
569 using as starting assembly the soft masked version available at Ensembl Metazoa. This
570 resulted in 41,221 transcripts, 39,814 of which have RNA-seq support (96.59%).
571 Additionally, 80.47% of the transcripts have a significant BLAST match (*e*-value cut-off <
572 0.001) to other well-annotated annelid genomes (see above).
573
574 Protein homologies for the filtered transcripts of *O. fusiformis* and *C. teleta* were annotated
575 with BLAST v.2.2.31+[74] on the UniProt/SwissProt database provided with Trinotate v.3.0[75].
576 We used HMMER v.2.3.2[76] to identify protein domains using Trinotate's PFAM-A database
577 and signalP v.4.1[77] to predict signal peptides. These functional annotations were integrated
578 into a Trinotate database, which retrieved Gene Onthology (GO), eggNOG and KEGG terms
579 for each transcript. In addition, we ran PANTHER HMM scoring tool to assign a
580 PantherDB[78] orthology ID to each transcript. In total, we retrieved a functional annotation for
581 22,516 transcripts (63.86 %). Functional annotation reports are provided in the GitHub
582 repository (see Data Availability section).

583

584 *Chromosome-scale scaffolding*

585 Sperm from a single *O. fusiformis* worm and an entire sexually mature male were used as
586 input material to construct two Omni-C Dovetail libraries following manufacturer's
587 recommendations for marine invertebrates. These libraries were sequenced in an Illumina
588 NovaSeq6000 at the Okinawa Institute of Science and Technology (Okinawa, Japan) to a
589 depth of 229 and 247 million reads. HiC reads were processed using the Juicer pipeline
590 r.e0d1bb7[79] to generate a list of curated contracts ('merged no dups') that was subsequently
591 employed to scaffold the assembly using 3d-dna v.180419[80]. The resulting assembly and
592 contact map were visually inspected and curated using Juicebox v.1.11.08[79] and adjustments
593 submitted for a subsequent run of optimisation using 3d-dna. Finally, repeats and TEs were
594 re-annotated in this chromosome scale assembly as described above, and the annotation
595 obtained for the PacBio-based assembly was lifted over with Liftoff v.1.6.1[81] (Supplementary
596 Fig. 1). All gene models but two were successfully re-annotated in the chromosome-scale
597 assembly.

598

599 *Gene family evolution analyses*

600 We used the AGAT suite of scripts to generate non-redundant proteomes with only the
601 longest isoform for a set of 21 metazoan proteomes (Supplementary Table 2). To reconstruct
602 gene families, we used OrthoFinder v.2.2.7[82] using MMSeqs2[83] to calculate sequence
603 similarity scores and an inflation value of 2. OrthoFinder gene families were parsed and
604 mapped onto a reference species phylogeny to infer gene family gains and losses at different
605 nodes and tips using the ETE 3 library[84], as well as to estimate the node of origin for each
606 gene family. Gene expansions were computed for each species using a hypergeometric test
607 against the median gene number per species for a given family employing previously

608 published code[30] (Supplementary Tables 3–7). Principal component analysis was performed
609 on the orthogroups matrix by metazoan lineage, given that orthogroups were present in at
610 least three of the 22 analysed species, to eliminate taxonomically restricted genes. All single
611 copy ortholog files derived from this analysis employed throughout the study are available in
612 the GitHub repository (see Data Availability section).
613
614 *Macrosynteny analyses*
615 Single copy orthologues obtained using the mutual best hit (MBH) approach generated using
616 MMseqs2[83] using the annotations of *Branchiostoma floridae*[85], *Pecten maximus*[86],
617 *Streblospio benedictii*[87], and *Lineus longissimus*[88,89] were used to generate Oxford synteny
618 plots comparing sequentially indexed orthologue positions. Plotting order was determined by
619 hierarchical clustering of the shared orthologue content using the complete linkage method as
620 originally proposed. Comparison of the karyotype of all four species was performed using the
621 Rideogram package by colouring pairwise orthologues according to the ALG assignment in
622 comparisons with *P. maximus* and *B. floridae*.
623
624 *Evolutionary analysis of* chordin *in annelids*
625 The identification of *chordin* (*chrd*) and *chordin-like* (*chrdl*) genes in *O. fusiformis* was based
626 on the genome functional annotation (see above). To mine *chrd* orthologues, 81 annelid
627 transcriptomic datasets were downloaded from SRA (Supplementary Table 8) and assembled
628 with Trinity v.2.5.1[68] to create BLAST local nucleotide databases. We also created a
629 nucleotide database for *C. teleta* using its annotated genome[90] (ENA accession number
630 GCA_000328365.1). Human and *O. fusiformis* CHRD proteins were used as queries to find
631 *chrd* orthologues following the MBH approach (*e*-value $\leq 10^{-3}$), obtaining 103 unique
632 candidate *chrd* transcripts that were then translated (Supplementary Table 9). A single

633 candidate CHRD protein for *Themiste lageniformis* (unpublished data, provided by Michael J
634 Boyle) was included *ad hoc* at this step. In addition, 15 curated CHRD and CHRDL protein
635 sequences (and an outgroup) were fetched from various sources (Supplementary Table 10)
636 and aligned together with *O. fusiformis* CHRD and CHRDL sequences in MAFFT v.7[91] with
637 the G-INS-I iterative refinement method and default scoring parameters. From this mother
638 alignment further daughter alignments were obtained using "mafft --addfragments"[92], the
639 accurate "--multipair" method, and default scoring parameters. For orthology assignment,
640 two phylogenetic analyses were performed on selected candidate sequences, which included
641 the longest isoform for each species-gene combination, given that it included a 10-residue or
642 longer properly aligned fragment in either the CHRD domains or the von Willebrand factor
643 type C (VWFC) domains. vWFC and CHRD domains were trimmed and concatenated using
644 domain boundaries defined by ProSITE domain annotation for the human chordin precursor
645 protein (UniProt: Q9H2X0). Either all domains or the VWFC domains only were used for
646 phylogenetic inference (Extended Data Figure 2c, d, Supplementary Tables 11, 12) with a
647 WAG amino acid replacement matrix[93] to account for transition rates, the FreeRate
648 heterogeneity model (R4)[94] to describe sites evolution rates, and an optimization of amino
649 acid frequencies using maximum likelihood (ML) using IQ-TREE v.2.0.3[95]. 1,000 ultrafast
650 bootstraps (BS)[96] were used to extract branch support values. Bayesian reconstruction in
651 MrBayes v.3.2.7a[97] were also performed using the same WAG matrix but substituting the R4
652 model for the discrete gamma model[98], with 4 rate categories (G4). All trees were composed
653 in FigTree v.1.4.4. Alignment files are available in the GitHub repository (see Data
654 Availability section).
655
656 *Gene expression profiling*

We profiled gene expression dynamics from blastula to juvenile stages for *O. fusiformis*, from 64-cell to competent larva stages for *C. teleta* (Supplementary Fig. 2), from early development to female adult stages for *D. gyrociliatus*, and across the 9 adult tissues samples of *O. fusiformis*. Sequencing adaptors were removed from raw reads using trimmomatic v.0.39[65]. Cleaned reads were pseudo-aligned to the filtered gene models using kallisto v.0.46.2[99] and genes with an expression level above an empirically defined threshold of 2 transcripts per million (TPM) were deemed expressed. For each species, the DESeq2 v.1.30.1 package[100] was used to normalise read counts across developmental stages (Supplementary Tables 13–21) and adult tissues (Supplementary Tables 49–51) and to perform pairwise differential gene expression analyses between consecutive developmental stages. *P*-values were adjusted using the Benjamini-Hochberg method for multiple testing correction. We defined a gene as significantly upregulated for a $\log_2$(fold-change) (LFC) > 1 or downregulated for a LFC < 1, given that adjusted *p*-value < 0.05. Principal component analyses were performed on the variance stabilising-transformed matrices of the normalised DESeq2 matrices. For the *O. fusiformis* adult tissues samples, genes specifically expressed (TPM > 2) in both the head and head plus two anteriormost segments samples only were classified as adult anterior genes, and those expressed in both the tail and the body wall only were classified as adult trunk and posterior genes (Supplementary Tables 52, 53). For all 3 annelid taxa, anterior, trunk, and posterior markers were defined as genes whose spatial expression pattern has been validated through *in situ* hybridisation in the literature (Supplementary Tables 54–56). TPM and DESeq2 gene expression matrices of developmental and adult tissue samples are also available in the GitHub repository (see Data Availability section).

Gene clustering and co-expression network analyses

682 Transcripts were clustered according to their normalised DESeq2 expression dynamics
683 through soft *k*-means clustering (or soft clustering) using the mfuzz v.2.52 package[101]
684 (Supplementary Tables 23–26). Out of the total number of transcripts, we discarded those
685 which were not expressed at any developmental stage (225 out of 31,903 for *O. fusiformis*,
686 1,407 out of 41,221 for *C. teleta*, and 200 out of 17,388 for *D. gyrociliatus*). We then
687 determined an optimal number of 12 clusters (*O. fusiformis* and *C. teleta*) and 9 clusters
688 (*D. gyrociliatus*) for our datasets by applying the elbow method to the minimum centroid
689 distance as a function of the number of clusters. For the construction of the gene co-
690 expression networks for *O. fusiformis* and *C. teleta*, we used the WGCNA package v.1.70–
691 3[102]. All transcripts expressed at any developmental stage were used to build a signed
692 network with a minimum module size of 300 genes and an optimised soft-thresholding power
693 of 16 and 8, for *O. fusiformis* and *C. teleta*, respectively. Block-wise network construction
694 returned 15 gene modules for *O. fusiformis*, from which one module was dropped due to poor
695 intramodular connectivity, and 19 gene modules for *C. teleta* (Supplementary Tables 23, 24).
696 The remaining 14 gene modules of *O. fusiformis* (A–N) and 19 gene modules of *C. teleta* (A–
697 O, W–Z) were labelled with distinct colours with unassigned genes labelled in grey. Random
698 subsets consisting of the nodes and edges of 30 % of the transcripts were fed to Cytoscape
699 v.3.8.2[103] for network visualisation (Supplementary Fig. 9). Module eigengenes were chosen
700 to summarise the gene expression profiles of gene modules. Gene ontology (GO) enrichment
701 analysis of each gene cluster and gene module was performed using the topGO v.2.44
702 package. We performed a Fisher's exact test and listed the top 30 (soft *k*-means clusters) or
703 top 15 (WGCNA modules) significantly enriched GO terms of the class biological process
704 (Supplementary Tables 27–31, Supplementary Fig. 4–6, 10, 11). To ease visualisation, all
705 486 non-redundant enriched GO terms from the 33 soft *k*-means clusters from all 3 species
706 were clustered through *k*-means clustering by semantic similarity using the

simplifyEnrichment v.1.2.0 package[104] (Supplementary Fig. 7, 8). Full network nodes and edges files and the random 30 % network subset files are available in the GitHub repository (see Data availability section).

Transcription factor repertoire analysis

We selected a custom set of 36 transcription factor classes from all 9 transcription factor superclasses from the TFClass database[105]. Transcripts in *O. fusiformis*, *C. teleta*, and *D. gyrociliatus* were deemed transcription factors and classified into one or more of the 36 classes if they were a match for any of the corresponding PANTHER identifiers (Supplementary Tables 32–33, Supplementary Fig. 3). Over- and underrepresentation of the different transcription factor classes in the gene expression clusters was tested through pairwise two-tailed Fisher's exact tests, for which we then adjusted the *p*-values using the Benjamini-Hochberg correction for multiple testing.

Orthogroup overlap analysis

We performed pairwise comparisons between each possible combination of soft *k*-means clusters of all 3 annelid taxa. The numbers of overlapped orthogroups between either the full clusters or the transcription factors belonging to each cluster only were subjected to upper-tail hypergeometric tests. *P*-values were then adjusted using the Benjamini-Hochberg method for multiple testing correction. For the simplified analyses by quadrants, clusters were classed as early/pre-larval (*O. fusiformis*: 1–6; *C. teleta*: 1–5; *D. gyrociliatus*: 1–3) or late/pre-larval (*O. fusiformis*: 8–12; *C. teleta*: 7–12; *D. gyrociliatus*: 5–7), thus rendering 4 different quadrants for each species pairwise comparison: $early_{species\ A}$ – $early_{species\ B}$, $early_{species\ A}$ – $late_{species\ B}$, $late_{species\ A}$ – $early_{species\ B}$, and $late_{species\ A}$ – $late_{species\ B}$. Clusters corresponding to

731 female adult expression in *D. gyrociliatus* (8 and 9) were discarded for comparison purposes.
732 Relative similarity (RS) for each of the four quadrants was computed as the following ratio:

733 $$RS = \frac{\text{mean}(-\log_{10}(adj\ p\text{-}value)_{\text{quadrant}})}{\text{mean}(-\log_{10}(adj\ p\text{-}value)_{\text{total}})}$$

734 Values above 1 indicate a higher orthogroup overlap than average, whereas values below 1
735 represent a lower overlap than average. For genes under heterochronic shifts – i.e., with
736 distinct temporal expression dynamics – between indirect and direct development, a gene set
737 was constructed with the genes with a single copy ortholog in both *O. fusiformis* and *C. teleta*
738 whose expression was shifted from post-larval clusters (*O. fusiformis*: 7–12; *C. teleta*: 8–12)
739 to early clusters 2 and 3 in *D. gyrociliatus* (see Fig. 2b) (Supplementary Tables 34, 35;
740 Supplementary Fig. 12). For the characterisation of genes under heterochronic shifts between
741 planktotrophic and lecithotrophic larvae, two gene sets were generated with the genes with
742 early$_{\textit{O. fusiformis}}$–late$_{\textit{C. teleta}}$ and late$_{\textit{O. fusiformis}}$–early$_{\textit{C. teleta}}$ dynamics, as described above
743 (Supplementary Tables 36–39; Supplementary Fig. 13, 14). Gene ontology (GO) enrichment
744 analysis of both gene sets was performed using the topGO v.2.44 package. We performed a
745 Fisher's exact test and listed the top 15 significantly enriched GO terms of the class
746 biological process (Supplementary Table 40). BlastKOALA[106] server was used to assign a
747 KEGG orthology number to one-to-one orthologs showing heterochronic sifts and KEGG
748 mapper[107] to analyse the annotations (Supplementary Tables 41, 42).
749
750 *Pathway analyses*
751 Human genes involved in the animal autophagy pathway (map04140) were obtained from the
752 KEGG pathway database[108]. *D. melanogaster* and *Saccharomyces cerevisiae* genes involved
753 in the chitin synthesis pathway were fetched from FlyBase[109] and SGD[110] based on the
754 enzyme nomenclature (EC) numbers of the pathway enzymatic activities[111]. Orthology in
755 *O. fusiformis* and *C. teleta* for the autophagy pathway genes was determined from the single

copy ortholog sets to the human genes, where one for both species existed (Supplementary Tables 43, 44). For the chitin synthesis pathway, and due to the high number of paralogs and expansions/losses of enzymatic activities of the chitin synthesis pathway, orthology was inferred from PANTHER family/subfamily identifiers to the corresponding enzymatic activities (Supplementary Tables 45, 46). We then used this orthology to reconstruct the chitin synthesis pathway in annelids. Timing across both species and the presence or lack thereof of heterochronic shifts between *O. fusiformis* and *C. teleta* was determined as described above (Supplementary Fig. 15, 16).

Hox *genes orthology assignment*

129 curated Hox sequences were retrieved from various databases (Supplementary Table 47) and aligned with *O. fusiformis* Hox proteins with MAFFT v.7 in automatic mode. Poorly aligned regions were removed with gBlocks v.0.91b[112] yielding the final alignments. Maximum likelihood trees were constructed using RAxML v.8.2.11.9[113] with an LG substitution matrix[114] and 1,000 ultrafast BS. All trees were composed in FigTree v.1.4.4. Alignment files are available in the GitHub repository (see Data Availability section).

Whole mount in situ *hybridisation and immunohistochemistry*

Fragments of *chordin* and *Hox* genes were isolated as previously described[24] using gene-specific oligonucleotides and a T7 adaptor. Riboprobes were synthesise with the T7 MEGAscript kit (ThermoFisher, AM1334) and stored at a concentration of 50 ng/µl in hybridisation buffer at –20 ºC. Whole mount *in situ* hybridisation in embryonic, larval, and juvenile stages were conducted as described elsewhere[24,26]. Antibody staining in larval stages of *O. fusiformis*, *Magelona* spp. and *C. teleta* was carried out as previously described[23,115] using the following antibodies: Mouse anti-acetyl-alpha tubulin Antibody, clone 6-11B-1,

1:800 dilution (Sigma-Aldrich Cat# MABT868, RRID:AB_2819178), and Goat anti-Mouse IgG (H+L) Cross-Adsorbed Secondary Antibody, Alexa Fluor 647, 1:800 dilution (Thermo Fisher Scientific Cat# A-21235, RRID:AB_2535804). DIC images of the colorimetric *in situs* were obtained with a Leica 560 DMRA2 upright microscope equipped with an Infinity5 camera (Lumenera). Fluorescently stained samples were scanned with a Nikon CSU-W1 Spinning Disk Confocal.

Assay for Transposase-Accessible Chromatin using sequencing (ATAC-seq)

We performed two replicates of ATAC-seq from samples containing ~50,000 cells at the blastula (~900 embryos), gastrula (~500), elongation (~300), mitraria larva (~150 larvae) and competent larva (~40) stages for *O. fusiformis*, and the 64-cells stage (~500 embryos), gastrula (~200), stage 4tt larva (~120 larvae), stage 5 larva (~90) and stage 8 larva (~50) for *C. teleta* following the omniATAC protocol[116], but gently homogenising the samples with a pestle in lysis buffer and incubating them on ice for 3 min. Tagmentation was performed for 30 min at 37°C with an in-house purified Tn5 enzyme[117]. After DNA clean-up, ATAC-seq libraries were amplified as previously described. Primers used for both PCR and qPCR are listed in Supplementary Tables 57 and 59. Amplified libraries were purified using ClentMag PCR Clean Up Beads as indicated by the supplier and quantified and quality checked on a Qubit 4 Fluorometer (Thermo-Fisher) and an Agilent 2200 TapeStation system before pooling at equal molecular weight. Sequencing was performed on an Illumina HiSeq4000 platform in 2 × 75 bases mode at the Oxford Genomics Centre (University of Oxford, United Kingdom) (blastula, elongation and mitraria larva stages, and one replicate of the gastrula sample of *O. fusiformis*, as well as the 64 cells, gastrula, and stage 4tt larva stages of *C. teleta*) and on an Illumina NovoSeq6000 in 2 × 150 bases mode at Novogene (Cambridge,

United Kingdom) (one replicate of gastrula and the two replicates of competent larva stages of *O. fusiformis* and the two replicates of stage 5 and stage 8 larva of *C. teleta*).

Chromatin accessibility profiling

We used cutadapt v.2.5[118] to remove sequencing adaptors and trim reads from libraries sequenced in 2 × 150 bases mode to 75 bases reads. Quality filtered reads were mapped using NextGenMap v.0.5.5[119] in paired-end mode, duplicates were removed using samtools v.1.9[120] and mapped reads were shifted using deepTools v.3.4.3[121] (Supplementary Tables 58, 60). Fragment size distribution was estimated from resulting BAM files and transcription start site (TSS) enrichment analysis was computed using computeMatrix and plotHeatmap commands in deepTools v.3.4.3. Peak calling was done with MACS2 v.2.2.7.1[122,123] (-f BAMPE --min-length 100 --max-gap 75 and -q 0.01). Reproducible peaks were identified by irreproducible discovery rates (IDR) (IDR < 0.05) v.2.0.4. at each developmental stage. Peaks from repetitive regions were filtered with BEDtools v.2.28.0[124] at each developmental stage. Next, we used DiffBind v.3.0.14[125] to generate a final consensus peak set of 63,732 peaks in *O. fusiformis* and 46,409 peaks in *C. teleta*, which were normalised using DESeq2 method (Supplementary Fig. 17). Peak clustering according to accessibility dynamics was performed as described above for RNA-seq, using the same number of 12 clusters to make both profiling techniques comparable. Principal component analysis and differential accessibility analyses between consecutive developmental stages were also performed as described above. An LFC > 0 and a LFC < 0 indicates whether a peak opens or closes, respectively, given that the adjusted *p*-value < 0.05. Stage-specific and constitutive peaks were determined using UpSetR v.1.4.0[126] and both the consensus peak set and the stage-specific peak sets were classified by genomic region using HOMER v.4.11[127] and further curated. Visualisation of peak tracks and gene structures was conducted with pyGenomeTracks v.2.1[128] and deepTools v.3.4.3[121]. To

correlate chromatin accessibility and gene expression, this genomic region annotation was used to assign peaks to their closest gene (63,726 peaks were assigned to 23,025 genes in *O. fusiformis* and 44,368 peaks were assigned to 23,382 genes in *C. teleta*). Pearson correlation coefficient between chromatin accessibility and gene expression was computed individually by peak with two-sided tests (Supplementary Fig. 18). Gene ontology (GO) enrichment analyses of the gene sets regulated by peak clusters was performed using the topGO v.2.44 package. We performed a Fisher's exact test and listed the top 30 significantly enriched GO terms of the class biological process (Supplementary Fig. 19, 20). To ease visualisation, all 242 non-redundant enriched GO terms were clustered through *k*-means clustering by semantic similarity using the simplifyEnrichment v.1.2.0 package[104] (Supplementary Tables 61–71; Supplementary Fig. 21–23). Coverage files and peak set files are available in the GitHub repository (see Data Availability section).

Motif identification, clustering, matching and curation

To identify transcription factor-binding motifs in chromatin accessible regions in the two species, we first used HOMER[127] (v.4.1) to identify known and *de novo* motifs in the consensus peak sets, which yielded 456 motifs for *O. fusiformis* and 364 motifs for *C. teleta* (Supplementary Tables 72, 73). Significance of motifs was derived from binomial tests from cumulative binomial distributions. We then used GimmeMotifs v.0.16.1[129] , with a 90% similarity cut-off to cluster the motifs predicted in *O. fusiformis* and *C. teleta* into 141 consensus motifs, which we matched against four motif databases to assign their putative identity (Gimme vertebrate 5.0[129], HOMER[127], CIS-BP[130] and a custom JASPAR2022[131] core motifs without plant and fungi motifs, Supplementary Fig. 24). We then used the human non-redundant TF motif database (https://resources.altius.org/~jvierstra/projects/motif-clustering-v2.0beta/) to manually curate the annotation. After removing motifs that likely represented

sequence biases, we finally obtained 95 motif archetypes for *O. fusiformis* and 91 for *C. teleta* (Supplementary Table 74), which we then used to perform motif counts in peaks (Supplementary Tables 75, 76) and motif accessibility estimation (Supplementary Tables 77, 78) with GimmeMotifs v.0.16.1[129]. Data clustering was performed with mfuzz v.2.52[101] (Supplementary Fig. 25, 27). Over- and underrepresentation of counts of the common curated motif archetypes in the peak accessibility soft clusters (see above) was tested through pairwise two-tailed Fisher's exact tests, for which we then adjusted the *p*-values using the Bonferroni correction for multiple testing.

Transcription factor footprinting and Hox *gene regulatory network exploration*

To predict transcription factor binding, as a proxy of activity, we conducted footprinting analysis with TOBIAS[132] v.0.12.0 during development in the 95 and 91 motif archetypes for *O. fusiformis* and *C. teleta*, respectively (Supplementary Tables 79, 80). Bound/unbound sites were first estimated by fitting a two-component gaussian-mixture model, and significance was then tested by a one-tail test from the right-most normal distribution. Transcription factor binding scores (TFBS) were clustered with mfuzz v.2.52[101]. Pearson correlation coefficients of motif accessibility and TFBS were calculated by stage and by motif separately based on the 33 common, curated motif archetypes (Supplementary Fig. 26, 28–30). To reconstruct potential upstream regulators and downstream effectors of the *Hox* genes, we first subset ATAC-seq peaks annotated to the *Hox* genes in the *Hox* cluster (i.e., all but *Post1*) in *O. fusiformis* and *C. teleta* and extracted the bound motifs on those peaks (Supplementary Tables 81, 82). TFBS were sum up for each motifs to obtain global dynamics, and their temporal dynamics were then clustered with mfuzz v.2.52[101] (Supplementary Fig. 31). For the downstream genes regulated by *Hox*, we obtained genes annotated to ATAC-seq peaks with a bound HOX/EVX/CDX motif at the competent stage in *O. fusiformis* and stage 4tt larva in *C.*

teleta (Supplementary Tables 83, 84). One-to-one orthologs were used to identified shared targets and PANTHER IDs to obtain their functional annotation.

Phylostratigraphy

To evaluate gene expression dynamics by phylostratum and developmental stage in all 3 annelid lineages, we used the OrthoFinder gene families and their inferred origins. We deemed all genes originating before and with the Cnidarian-Bilaterian ancestor of pre-metazoan and metazoan origin (Supplementary Tables 85–87). We then applied a quantile normalisation onto the DESeq2 normalised matrices of gene expression. The 75 % percentile of the quantile-normalised gene expression levels was used as the summarising measure of the gene expression distribution by developmental stage. Over- and underrepresentation of the different phylostrata in the gene expression clusters was tested through pairwise two-tailed Fisher's exact tests, for which we then adjusted the *p*-values using the Bonferroni correction for multiple testing. Gene expression dynamics of novel genes and genes of pre-metazoan and metazoan origin across selected metazoan lineages (see Comparative transcriptomics section below) were also evaluated as described above (Supplementary Fig. 32)..

Comparative transcriptomics

Publicly available RNA-seq developmental time courses for the development of *Amphimedon queenslandica*, *Clytia hemisphaerica*, *Nematostella vectensis*, *Strongylocentrotus purpuratus*, *Branchiostoma lanceolatum*, *Danio rerio*, *Drosophila melanogaster*, *Caenorhabditis elegans*, *Crassostrea gigas*, *Dimorphilus gyrociliatus*, and two stages of *Capitella teleta* were downloaded from the SRA using SRA-Toolkit v.2.11.3 (Supplementary Table 88), cleaned for adaptors and low-quality reads with trimmomatic v.0.39[65] and pseudo-aligned to their

respective non-redundant genome-based gene repertoires – i.e., with a single transcript isoform, the longest, per gene model – using kallisto v.0.46.2[99]. We then performed a quantile transformation of TPM values using scikit-learn v.1.0.2[133] and calculated the Jensen-Shannon divergence (JSD) from (i) all single copy orthologs, (ii) the set single copy transcription factor orthologs, and (iii) the set of common single copy orthologs across all lineages, either between all possible one-to-one species comparisons (i) or between all species and *O. fusiformis* (ii, iii), using the philentropy v.0.5.0 package[134]:

$$JSD_{\mathrm{raw}}(P \parallel Q) = \frac{1}{2}\sum_{i=0}^{n} p_i \times \log_2\left(\frac{p_i}{\frac{1}{2}(p_i + q_i)}\right) + \frac{1}{2}\sum_{i=0}^{n} q_i \times \log_2\left(\frac{q_i}{\frac{1}{2}(p_i + q_i)}\right)$$

Transcriptomic divergences were calculated based on 250 bootstrap replicates, from which statistically robust mean values and standard deviations were obtained. Raw mean JSD values (JSD_{raw}) were adjusted (JSD_{adj}) by dividing by the number of single copy orthologs (i), single copy transcription factor orthologs (ii), or common single copy orthologs (iii) of each comparison (Supplementary Tables 22, 89, 90), and normalised using the minimum and maximum adjusted JSD values from all one-to-one species comparisons as follows:

$$JSD_{\mathrm{norm}}(P \parallel Q) = \frac{JSD_{\mathrm{adj}}(P \parallel Q) - \min JSD_{\mathrm{adj}}}{\max JSD_{\mathrm{adj}} - \min JSD_{\mathrm{adj}}};\ JSD_{\mathrm{norm}} \in [0, 1]$$

Relative JSD values were obtained equally, using minimum and maximum adjusted JSD values from each one-to-one species comparison instead. Gene-wise JSD ($gwJSD$) between five key one-to-one larval stages comparisons was computed as follows:

$$gwJSD(P \parallel Q) = \frac{1}{2} \times p_i \times \log_2\left(\frac{p_i}{\frac{1}{2}(p_i + q_i)}\right) + \frac{1}{2} \times q_i \times \log_2\left(\frac{q_i}{\frac{1}{2}(p_i + q_i)}\right)$$

Similarity-driving genes – i.e., those with very low gwJSD – were subset as those below the threshold defined as 25 % of the point of highest probability density of the gwJSD distributions. Gene ontology (GO) enrichment analysis of the similarity-driving gene sets was

927 performed using the topGO v.2.44 package. We performed a Fisher's exact test and listed the
928 top 30 significantly enriched GO terms of the class biological process (Supplementary
929 Table 91). To ease visualisation, all 51 non-redundant enriched GO terms from the 5 gene
930 sets were clustered through *k*-means clustering by semantic similarity using the
931 simplifyEnrichment v.1.2.0 package[104]. The subsets of similarity-driven transcription factors
932 of each pairwise comparison are listed in Supplementary Table 92. For comparative *Hox* gene
933 expression dynamics profiling in metazoan lineages, the same non-redundant gene expression
934 matrices were normalised using the DESeq2 v.1.30.1 package[100] (Supplementary Fig. 33),
935 unless *Hox* gene models where missing, in which case they were manually added *ad hoc* to
936 the non-redundant genome-based gene repertoires (Supplementary Table 94). *Hox* gene
937 expression profiling in *Urechis unicinctus* was performed as described for the rest of taxa but
938 using the available reference transcriptome[135] instead (Supplementary Table 48). All gene
939 expression matrices are available in the GitHub repository (see Data Availability section).
940
941 **Data availability**
942 Accession codes and unique identifiers to previously publicly available datasets we used for
943 this study are listed in Supplementary Table 2 (genome files used in gene family evolution
944 analyses), Supplementary Table 8 (transcriptomes used in the evolutionary analysis of
945 chordin in annelids), Supplementary Tables 41 and 43 (gene identifiers used in pathway
946 analyses), Supplementary Table 47 (sequence identifiers used in *Hox* genes orthology
947 assignment), Supplementary Table 48 (RNA-seq datasets used for *Hox* gene expression
948 profiling in *U. unicinctus*) and Supplementary Table 88 (RNA-seq datasets used for
949 comparative annelid and metazoan transcriptomics and *Hox* gene expression profiling).
950 Repetitive elements database RepBase can be accessed at https://www.girinst.org/repbase/.
951 Transcription factor public database TFClass can be found at http://tfclass.bioinf.med.uni-

goettingen.de/. All sequence data associated with this project are available at the European Nucleotide Archive (project PRJEB38497) and Gene Expression Omnibus (accession numbers GSE184126, GSE202283, GSE192478, GSE210813 and GSE210814). Genome assemblies, transposable element annotations, genome annotation files used for RNA-seq and ATAC-seq analyses, WGCNA nodes and edges files, alignment files used in orthology assignment, and other additional files are publicly available in https://github.com/ChemaMD/OweniaGenome.

Code availability

All code used in this study is available in https://github.com/ChemaMD/OweniaGenome.

Method References

50 Marçais, G. & Kingsford, C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics 27, 764-770 (2011).

51 Ranallo-Benavidez, T. R., Jaron, K. S. & Schatz, M. C. GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. Nat Commun 11, 1432 (2020).

52 Koren, S. et al. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome Res 27, 722-736, doi:10.1101/gr.215087.116 (2017).

53 Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv 1303.3997 (2013).

54 Vaser, R., Sovic, I., Nagarajan, N. & Sikic, M. Fast and accurate de novo genome assembly from long uncorrected reads. Genome Res 27, 737-746 (2017).

55 Roach, M. J., Schmidt, S. A. & Borneman, A. R. Purge Haplotigs: allelic contig reassignment for third-gen diploid genome assemblies. BMC Bioinformatics 19, 460 (2018).

56 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).

57 Manni, M., Berkeley, M. R., Seppey, M., Simao, F. A. & Zdobnov, E. M. BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes. Mol Biol Evol 38, 4647-4654 (2021).

58 Flynn, J. M. et al. RepeatModeler2 for automated genomic discovery of transposable element families. Proc Natl Acad Sci U S A 117, 9451-9457 (2020).

59 Buchfink, B., Reuter, K. & Drost, H. G. Sensitive protein alignments at tree-of-life scale using DIAMOND. Nat Methods 18, 366-368 (2021).

60 Xu, Z. & Wang, H. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. Nucleic Acids Res 35, W265-268 (2007).

61 Wong, W. Y. & Simakov, O. RepeatCraft: a meta-pipeline for repetitive element de-fragmentation and annotation. Bioinformatics 35, 1051-1052 (2019).

62 Danecek, P. et al. Twelve years of SAMtools and BCFtools. Gigascience 10 (2021).

63 Weigert, A. et al. Illuminating the base of the annelid tree using transcriptomics. Mol Biol Evol 31, 1391-1401 (2014).

64 Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29, 15-21 (2013).

65 Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30, 2114-2120 (2014).

66 Pertea, M. et al. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nat Biotechnol 33, 290-295 (2015).

67 Mapleson, D., Venturini, L., Kaithakottil, G. & Swarbreck, D. Efficient and accurate detection of splice junctions from RNA-seq with Portcullis. Gigascience 7 (2018).

68 Grabherr, M. G. et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol 29, 644-652 (2011).

69 Wu, T. D. & Watanabe, C. K. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. Bioinformatics 21, 1859-1875 (2005).

70 Venturini, L., Caim, S., Kaithakottil, G. G., Mapleson, D. L. & Swarbreck, D. Leveraging multiple transcriptome assembly methods for improved gene structure annotation. Gigascience 7 (2018).

71 Stanke, M., Schoffmann, O., Morgenstern, B. & Waack, S. Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. BMC Bioinformatics 7, 62 (2006).

72 Slater, G. S. & Birney, E. Automated generation of heuristics for biological sequence comparison. BMC Bioinformatics 6, 31 (2005).

73 Haas, B. J. et al. Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. Nucleic Acids Res 31, 5654-5666 (2003).

74 Camacho, C. et al. BLAST+: architecture and applications. BMC Bioinformatics 10, 421 (2009).

75 Bryant, D. M. et al. A Tissue-Mapped Axolotl De Novo Transcriptome Enables Identification of Limb Regeneration Factors. Cell Rep 18, 762-776 (2017).

76 Eddy, S. R. Accelerated Profile HMM Searches. PLoS Comput Biol 7, e1002195 (2011).

77 Petersen, T. N., Brunak, S., von Heijne, G. & Nielsen, H. SignalP 4.0: discriminating signal peptides from transmembrane regions. Nat Methods 8, 785-786 (2011).

78 Mi, H. et al. PANTHER version 16: a revised family classification, tree-based classification tool, enhancer regions and extensive API. Nucleic Acids Res 49, D394-D403 (2021).

79 Durand, N. C. et al. Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. Cell Syst 3, 95-98 (2016).

80 Dudchenko, O. et al. De novo assembly of the Aedes aegypti genome using Hi-C yields chromosome-length scaffolds. Science 356, 92-95 (2017).

81 Shumate, A. & Salzberg, S. L. Liftoff: accurate mapping of gene annotations. Bioinformatics 37, 1639-1643 (2021).

82 Emms, D. M. & Kelly, S. OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biol 20, 238 (2019).

83 Steinegger, M. & Soding, J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. Nat Biotechnol 35, 1026-1028 (2017).

84 Huerta-Cepas, J., Serra, F. & Bork, P. ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data. Mol Biol Evol 33, 1635-1638 (2016).

85 Marletaz, F. et al. Amphioxus functional genomics and the origins of vertebrate gene regulation. Nature 564, 64-70 (2018).

86 Zeng, Q. et al. High-quality reannotation of the king scallop genome reveals no 'gene-rich' feature and evolution of toxin resistance. Comput Struct Biotechnol J 19, 4954-4960 (2021).

87 Zakas, C., Harry, N. D., Scholl, E. H. & Rockman, M. V. The genome of the poecilogonous annelid Streblospio benedicti. Genome Biology and Evolution (2022).

88 Kwiatkowski, D. et al. The genome sequence of the bootlace worm, Lineus longissimus (Gunnerus, 1770). Wellcome Open Res 6, 272 (2021).

89 Cannon, J. T. et al. Xenacoelomorpha is the sister group to Nephrozoa. Nature 530, 89-93 (2016).

90 Simakov, O. et al. Insights into bilaterian evolution from three spiralian genomes. Nature 493, 526-531 (2013).

91 Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol 30, 772-780, doi:10.1093/molbev/mst010 (2013).

92 Katoh, K. & Frith, M. C. Adding unaligned sequences into an existing alignment using MAFFT and LAST. Bioinformatics 28, 3144-3146 (2012).

93 Whelan, S. & Goldman, N. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. Mol Biol Evol 18, 691-699 (2001).

94 Soubrier, J. et al. The influence of rate heterogeneity among sites on the time dependence of molecular rates. Mol Biol Evol 29, 3345-3358 (2012).

95 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).

96 Hoang, D. T., Chernomor, O., von Haeseler, A., Minh, B. Q. & Vinh, L. S. UFBoot2: Improving the Ultrafast Bootstrap Approximation. Mol Biol Evol 35, 518-522 (2018).

97 Ronquist, F. & Huelsenbeck, J. P. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19, 1572-1574 (2003).

98 Yang, Z. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. J Mol Evol 39, 306-314 (1994).

99 Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. Nat Biotechnol 34, 525-527 (2016).

100 Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol 15, 550 (2014).

101 Futschik, M. E. & Carlisle, B. Noise-robust soft clustering of gene expression time-course data. J Bioinform Comput Biol 3, 965-988 (2005).

102 Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics 9, 559 (2008).

103 Shannon, P. et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res 13, 2498-2504 (2003).

104 Gu, Z. & Hübschmann, D. simplifyEnrichment: an R/Bioconductor package for Clustering and Visualizing Functional Enrichment Results. bioRxiv, 2020.2010.2027.312116 (2021).

105 Wingender, E., Schoeps, T., Haubrock, M., Krull, M. & Donitz, J. TFClass: expanding the classification of human transcription factors to their mammalian orthologs. Nucleic Acids Res 46, D343-D347 (2018).

106 Kanehisa, M., Sato, Y. & Morishima, K. BlastKOALA and GhostKOALA: KEGG Tools for Functional Characterization of Genome and Metagenome Sequences. J Mol Biol 428, 726-731 (2016).

107 Kanehisa, M., Sato, Y. & Kawashima, M. KEGG mapping tools for uncovering hidden features in biological data. Protein Sci 31, 47-53 (2022).

108 Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res 28, 27-30 (2000).

109 Gramates, L. S. et al. FlyBase: a guided tour of highlighted features. Genetics 220 (2022).

110 Cherry, J. M. et al. Saccharomyces Genome Database: the genomics resource of budding yeast. Nucleic Acids Res 40, D700-705 (2012).

111 Merzendorfer, H. The cellular basis of chitin synthesis in fungi and insects: common principles and differences. Eur J Cell Biol 90, 759-769 (2011).

112 Talavera, G. & Castresana, J. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. Syst Biol 56, 564-577 (2007).

113 Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30, 1312-1313 (2014).

114 Le, S. Q. & Gascuel, O. An improved general amino acid replacement matrix. Mol Biol Evol 25, 1307-1320 (2008).

115 Meyer, N. P., Carrillo-Baltodano, A., Moore, R. E. & Seaver, E. C. Nervous system development in lecithotrophic larval and juvenile stages of the annelid Capitella teleta. Front Zool 12, 15 (2015).

116 Corces, M. R. et al. An improved ATAC-seq protocol reduces background and enables interrogation of frozen tissues. Nat Methods 14, 959-962 (2017).

117 Picelli, S. et al. Tn5 transposase and tagmentation procedures for massively scaled sequencing projects. Genome Res 24, 2033-2040 (2014).

118 Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. 2011 17, 3 (2011).

119 Sedlazeck, F. J., Rescheneder, P. & von Haeseler, A. NextGenMap: fast and accurate read mapping in highly polymorphic genomes. Bioinformatics 29, 2790-2791 (2013).

120 Li, H. et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25, 2078-2079 (2009).

121 Ramirez, F. et al. deepTools2: a next generation web server for deep-sequencing data analysis. Nucleic Acids Res 44, W160-165 (2016).

122 Gaspar, J. M. Improved peak-calling with MACS2. bioRxiv, 496521 (2018).

123 Zhang, Y. et al. Model-based analysis of ChIP-Seq (MACS). Genome Biol 9, R137 (2008).

124 Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26, 841-842 (2010).

125 Ross-Innes, C. S. et al. Differential oestrogen receptor binding is associated with clinical outcome in breast cancer. Nature 481, 389-393 (2012).

126 Conway, J. R., Lex, A. & Gehlenborg, N. UpSetR: an R package for the visualization of intersecting sets and their properties. Bioinformatics 33, 2938-2940 (2017).

127 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).

128 Lopez-Delisle, L. et al. pyGenomeTracks: reproducible plots for multivariate genomic datasets Bioinformatics 37, 422-423 (2020).

129 van Heeringen, S. J. & Veenstra, G. J. GimmeMotifs: a de novo motif prediction pipeline for ChIP-sequencing experiments. Bioinformatics 27, 270-271 (2011).

130 Weirauch, M. T. et al. Determination and inference of eukaryotic transcription factor sequence specificity. Cell 158, 1431-1443 (2014).

131 Castro-Mondragon, J. A. et al. JASPAR 2022: the 9th release of the open-access database of transcription factor binding profiles. Nucleic Acids Res 50, D165-D173 (2022).

132 Bentsen, M. et al. ATAC-seq footprinting unravels kinetics of transcription factor binding during zygotic genome activation. Nat Commun 11, 4267 (2020).

133 Pedregosa, F. et al. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research 12, 2825−2830 (2011).

134 Drost, H. G. Philentropy: Information Theory and Distance Quantification with R. Journal of Open Source Software 3, 765 (2018).

135 Park, C. et al. The developmental transcriptome atlas of the spoon worm Urechis unicinctus (Echiurida: Annelida). Gigascience 7, 1-7 (2018).

Acknowledgements

We thank Alex de Mendoza, Daria Gavriouchkina, and Steve Rossiter for their support and valuable comments on the manuscript, as well as the staff at Station Biologique de Roscoff for their help with collections and animal supplies, and the Oxford Genomics Centre at the Wellcome Centre for Human Genetics (funded by Wellcome Trust grant reference 203141/Z/16/Z) for the generation and initial processing of RNA-seq and ATAC-seq sequencing data. We also thank Michael J Boyle for providing the *chordin* sequence for *Themiste lageniformis*, Joseph Deane for initial help with *Hox* gene characterisation in *O. fusiformis*, and the core technical staff at the Department of Biology at Queen Mary University of London for their constant support. This work used computing resources from Queen Mary University of London's Apocrita HPC facilities. This work was funded by the Horizon 2020 Framework Programme to JMM-D (European Research Council Starting Grant action number 801669) and AH (European Research Council Consolidator Grant action number 648861), a Royal Society University Research Fellowship (URF\R1\191161) and a Japan Society for the Promotion of Science Kakenhi grant (JP 19K06620) to FM, and a Biotechnology and Biological Sciences Research Council LIDo iCASE PhD studentship (BB/T008709/1) to JMM-D and BED.

Author Contributions

JMM-D, FM, YL and FMM-Z conceived and designed the study; YL collected RNA-seq samples for *O. fusiformis* and *C. teleta*, performed ATAC-seq experiments and contributed to all data analyses; FMM-Z performed *chordin* orthology studies and contributed to all data analyses; KG conducted *in situ* hybridisation analyses of *Hox* genes; AC-B collected RNA-seq samples for *C. teleta*, performed immunostainings on larvae and gene expression analyses of *chordin*; BED and RDD contributed to computational analyses; YT performed OMNI-C libraries; GM performed repeat annotations and analyses; OS identified and performed *in silico* analyses of *Hox* genes; MT performed genomic extractions and optical mapping; KM collected *Magelona* spp.; AH and NML contributed to sequencing efforts; FM and JMM-D assembled and annotated the genome and contributed to data analyses; YL, FMM-Z and JMM-D drafted the manuscript and all authors critically read and commented on the manuscript.

Competing Interests

The authors declare no competing interests.

Additional Information

Supplementary Information The online version contains supplementary information

Correspondence and request for materials should be addressed to José M. Martín-Durán or Ferdinand Marlétaz.

Extended Data Figure Legends

Extended Data Figure 1 | The genome of *Owenia fusiformis* is conservatively evolving. **a,** Differential interface contrast (DIC) images and z-stack confocal laser scanning views of a *C. teleta* trochophore larva stained for DAPI and acetylated α-tubulin. **b, c,** Pie charts of the transposable element content and Kimura substitution plots of transposable element divergence for *O. fusiformis* and other selected annelid species belonging to different annelid clades as depicted in **c**. Unlike *H. robusta* and *L. luymesi*, which show bursts of transposable elements, *O. fusiformis* shows more steady rates of expansion. **d**, Gene family evolution analysis across 22 metazoan lineages under a consensus tree topology. Gains are shown in green, losses in violet. Gene family losses in *O. fusiformis* are like those of slow-evolving lineages. **e**, Principal component analysis from Fig. 1b, showing the full set of species. **f, g,** *O. fusiformis* has the lowest number of gene losses of all sampled annelids (**e**), and the least gene expansions (**f**) after the extremely compact genome of *D. gyrociliatus*. **h**, Macrosynteny analysis between *O. fusiformis*, and from top to bottom, the cephalochordate *Branchiostoma floridae*, the bivalve *Pecten maximus*, and the annelid *Streblospio benedicti*. *Owenia fusiformis* retains ancestral linkage groups but also exhibits annelid- and species-specific chromosomal arrangements. However, the karyotype of *O. fusiformis* is more conserved than that of the annelid *S. benedicti*. **i**, Macrosynteny analysis between the bivalve *P. maximus* and the nemertean worm *L. longissimus*. *Lineus longissimus* exhibits conserved ancestral bilaterian linkage groups, including three potential lophotrochozoan-specific chromosomal rearrangements (H+Q, J2+L and K+O2), plus a nemertean-specific fusion (G+C1). Scale bar in **a**, 50 μm.

Extended Data Figure 2 | *chordin* was lost multiple times in annelids. a, Domain organisation of Chordin (CHRD) and Chordin-like (CHRDL1/2) proteins, as inferred from

human orthologs. **b,** Public AlphaFold protein structure prediction for human Chordin (UniProt: Q9H2X0) and Chordin-like 1 (UniProt: Q9BU40) revealed a previously unknown and uncharacterised domain in CHRDL1 and CHRDL2 (also depicted in **a**). **c**, **d**, Orthology assignment of *chordin* annelid candidates. From the multiple sequence alignment, candidate annelid sequences with a 10-residue or longer fragment aligned against either the CHRD (**c**; i.e., bona fide *chordin* genes) or the vWFC domains (**d**; i.e., putative *chordin* genes) were kept for further analysis. CHRDL cluster is shaded in blue; CHRD cluster, in red. Bootstrap support values (top) and posterior probabilities (bottom) are shown at both key nodes. Sequences in red and blue are curated CHRD and CHRDL sequences, respectively. **e, f**, Summary phylogenetic trees of presence or absence of *chordin* (red) or putative *chordin* (light brown) across Annelida. **g,** RNA-seq expression levels of *chordin* in *O. fusiformis*, which peaks at the blastula and gastrula stages, after the specification and inductive activity of the embryonic organiser. Curve is a locally estimated scatterplot smoothing, coloured shaded area represents standard error of the mean. **h,** Whole mount *in situ* hybridisation of *chordin* at the blastula (5 hours post fertilisation, hpf), gastrula (9 hpf), and mitraria larva (27 hpf) stages of *O. fusiformis*. Asterisks mark the animal/anterior pole. gp: gastral plate; bp: blastopore, mo: mouth. Representative results of three independent analyses. Scale bar in **h**, 50 μm.

Extended Data Figure 3 | Gene expression dynamics during annelid life cycles.

a, b, Heatmaps of average pairwise transcriptomic Jensen–Shannon Divergence (JSD) between *O. fusiformis* and *C. teleta* (**a**), and between *D. gyrociliatus* and either *O. fusiformis* (**b**, left) or *C. teleta* (**b**, right). Average relative JSD of the *C. teleta* or *O. fusiformis* stages of minimal divergence to each corresponding stage is shown on top. Confidence intervals represent standard deviation from 250 bootstrap resamplings of the ortholog sets. **c–e,** Soft *k*-

means clustered heatmap of all transcripts whose expression was not null in at least one developmental stage into an optimal number of 12 clusters (*O. fusiformis*, **c**; and *C. teleta*, **d**) and 9 clusters (*D. gyrociliatus*, **e**). Soft clustering considerably increased temporal resolution for the RNA-seq time course of *D. gyrociliatus*. On the right of each heatmap, gene-wise expression dynamics (grey lines) and locally estimated scatterplot smoothing (coloured lines) for each cluster. Coloured shaded areas represent standard error of the mean. **f,** Enrichment analysis of biological process gene ontology (GO) terms for RNA-seq clusters. Each line represents a single GO term, for which the $-\log_{10}(p\text{–value})$ for each RNA-seq cluster is shown in a colour-coded scale. GO terms were clustered into 15 distinct clusters based on semantic similarity (see Supplementary Fig. 7, 8). Clusters are shown on the bottom of the heatmaps. For the full list of GO terms and clusters, see Supplementary Fig. 4–6. *P*-values were derived from upper-tail Fisher's exact tests.

Extended Data Figure 4 | Heterochronic shifts in gene regulatory programmes between annelid life cycles. a, b, Similarity heatmaps showcasing the orthogroup overlap between the clusters of co-regulated genes (see Extended Data Fig. 3c–e), between the three annelids. *P*-values were derived from upper-tail hypergeometric tests and Benjamini-Hochberg-adjusted. **c,** Explanation of the orthogroup overlap analysis by quadrants. Clusters were classed as "early" (before dotted lines) or "late" (after dotted lines). Clusters of the female adult of *D. gyrociliatus* were disregarded. **d, e,** Heatmaps of relative similarity by quadrants of the orthogroup overlap analyses of the whole genomes (**d**) and transcription factors only (**e**). Colour scale in **d** and **e** is the same as in **c**. **f,** KEGGbrite characterisation of the gene sets under heterochronic shifts (surrounded by dotted black lines in Fig. 2e) between *O. fusiformis* and *C. teleta*. **g,** Bar plots depicting *p*-values of top biological process GO terms of genes shifted from late expression in *O. fusiformis* to early expression in *C. teleta*. *P*-values were

derived from upper-tail Fisher's exact tests. Full list is available in Supplementary Fig. 13. **h,** Enrichment analysis of the number of transcription factors per class in clusters of co-transcribed genes of *O. fusiformis* (left), *C. teleta* (centre) and *D. gyrociliatus* (right). For each cluster and class combination, the Bonferroni-adjusted *p*–value from the two-sided Fisher's exact test is shown. Cells in red represent overrepresented classes (odds ratio, OR > 1; adjusted *p*–value < 0.05); cells in blue, underrepresented classes (OR < 1, adjusted *p*–value < 0.05). Dotted lines highlight clusters of maximal enrichment of the homeodomain class. n.s.: not significant.

Extended Data Figure 5 | The *Hox* gene complement and expression in *O. fusiformis*. a, Orthology assignment of *O. fusiformis Hox* genes through maximum likelihood phylogenetic inference. Bootstrap support values are shown for major gene groups. Of: *O. fusiformis*. **b,** Volcano plot of the mitraria to competent larva transition, highlighting the marked upregulation of *Hox* genes. LFC: $\log_2$(fold-change). *P*-values were derived from the described DESeq2 pipeline and Benjamini-Hochberg-adjusted. **c**, Chromosomal location of the *Hox* cluster and *Post1* gene in *O. fusiformis* (top) and schematic comparison of *Hox* cluster organisation in annelids and a mollusc (bottom). Arrows denote direction of transcription. **d,** Schematic representation to scale of the genomic loci and intron–exon composition of *Hox* genes in *O. fusiformis*. **e,** Heatmaps of *Hox* gene expression during the development of *C. teleta*, *O. fusiformis* and the echiuran annelid *Urechis unicinctus*. In the two annelid species with planktotrophic larvae, *Hox* genes only become expressed at the larval stage (dotted vertical line), and not during embryogenesis, as observed in *C. teleta*. **f,** Whole mount *in situ* hybridisation of *Hox* genes in the gastrula (lateral views) and in the mitraria larva, pre-competent larva, and juvenile stages of *O. fusiformis* (ventral views). The area encircled by a dotted white line at the pre-competent stage highlights a region of probe

trapping from ingested food content. bp: blastopore; mo: mouth. Representative results of three independent analyses. **g, h** Schematic representations of the expression of *Hox* genes in the trunk rudiment of the competent larva (**g**) and juvenile trunk (**h**). A: anterior; P: posterior. Drawings are not to scale, and schematic expression domains are approximate. Scale bars in **f**, 50 µm in gastrulae and larvae, and 100 µm in juvenile.

Extended Data Figure 6 | Transcriptomic dynamics of anteroposterior genes.

a, Schematic drawing of the adult body regions used to define anterior and posterior and trunk genes. **b,** Correlation matrix of RNA-seq experiments from all nine adult tissues, calculated from a variance stabilising-transformed matrix of the normalised DESeq2 matrix. **c,** Venn diagram showing the number of tissue-specific and shared expressed genes (TPM > 2). Gene sets highlighted with red text were defined as adult anterior, and adult posterior and trunk genes. **d,** Phylostratigraphic classification of adult anterior, and adult posterior and trunk genes, compared to the whole genome and a random subset of 1,000 genes. **e, f,** Expression dynamics of each phylostratum by developmental stage in the adult anterior (**e**), and adult posterior and trunk gene sets (**f**), calculated from the 75 % percentile of a quantile-normalised matrix of gene expression levels. Adult anterior genes of most phylostrata peak at the blastula, while the maximum expression of adult trunk/posterior genes of most phylostrata peak at post-larval stages. **g–l,** Average expression dynamics of *in situ* hybridisation-validated anterior, trunk, and posterior markers throughout *O. fusiformis* (**g, h**), *C. teleta* (**i, j**), and *D. gyrociliatus* (**k, l**) development. For boxplots in **g**, **i**, and **k**, centre lines, median; box, interquartile range (IQR); whiskers, first or third quartile ± 1.5 × IQR. Lower whiskers are sometimes not apparent due to the distribution skewness towards zero. Curves in **h**, **j,** and **l** are locally estimated scatterplot smoothings. Coloured shaded areas represent standard error of the mean. n = 23, 8, and 17 anterior markers, 10 and 3 posterior markers,

and 15, 10, and 8 trunk markers, for *O. fusiformis*, *C. teleta*, and *D. gyrociliatus*, respectively. Key stages where expression of trunk markers is incipient are shown for both *O. fusiformis* and *C. teleta*.

Extended Data Figure 7 | Chromatin dynamics during annelid development. a, Genomic feature annotation of the consensus ATAC-seq peaks. **b,** Stacked bar plots showing the proportion of called peaks per developmental stage classified by genomic feature. **c,** Heatmap of normalised peak accessibility of the soft clustered consensus ATAC-seq peak sets. **d,** Self-correlation matrices of normalised motif accessibility and transcription factor binding score, revealing distinct chromatin regulatory dynamics throughout development. **e,** Correlation matrices of normalised motif accessibility to transcription factor binding score during annelid development. **f,** Correlation matrix of normalised motif accessibility between both species. **d–f** further validate the non-triviality of the results obtained in Fig. 3c. Pearson correlation coefficients in **d–f** were derived from two-tailed tests. **g,** Heatmap of normalised motif accessibility and transcription factor binding dynamics for each of the common annotated annelid motif archetypes during *O. fusiformis* and *C. teleta* development. Colour scale denotes transcription factor binding score dynamics, bubble size represents motif accessibility dynamics, both in a z-score scale. Motif archetypes highlighted in red are representative examples of the heterochronic shifts shown in bulk in Fig. 3c. **h,** Enrichment analysis of the number of occurrences of the common annotated annelid motif archetypes in the peak clusters inferred through soft *k*-means clustering and shown in **c**, for *O. fusiformis* (top) and *C. teleta* (bottom). For each cluster and motif combination, the Bonferroni-adjusted *p*–value of the two-tailed Fisher's exact test is shown. Red cells represent significantly overrepresented lineages (odds ratio, OR > 1, adjusted p–value < 0.05). Blue cells denote significantly underrepresented lineages (OR < 1, adjusted p–value < 0.05). **i,** Most abundant

bound motifs in peaks of the *Hox* clusters. **j,** Downstream regulated genes by transcription factors bound to the HOX/CDX/EVX motif archetype.

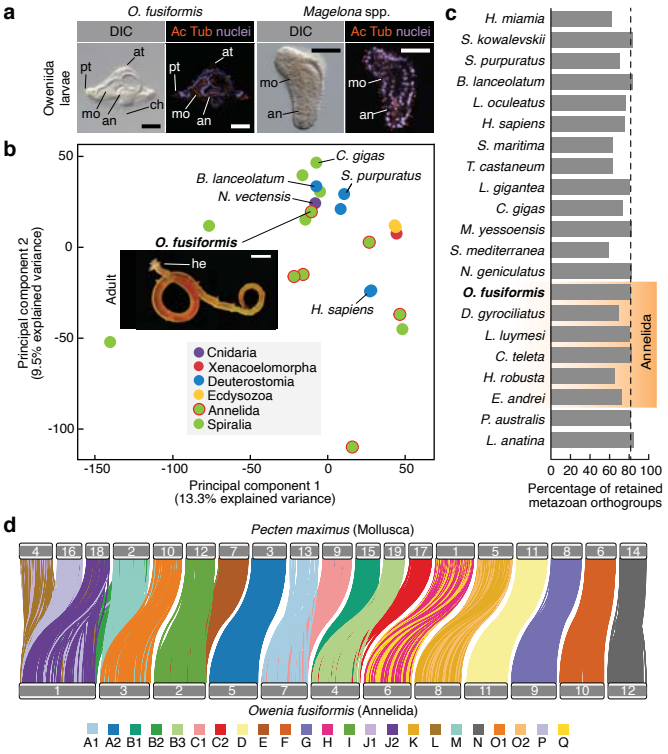
Extended Data Figure 8 | Phylostratigraphy analyses in annelid life cycles. **a**, **b** Expression dynamics (**a**) and expression contribution (**b**) of each phylostratum by developmental stage in all three annelids, calculated from the 75% percentile of a quantile-normalised matrix of gene expression levels. Older genes are expressed at the highest levels across annelid development. **c–e,** Boxplots of quantile-normalised expression levels of genes classified by phylostratum across *O. fusiformis* (**c**), *C. teleta* (**d**), and *D. gyrociliatus* (**e**) development. A random subset of 2,000 genes is shown as a negative control. *n* denotes number of genes per phylostratum. **h–j,** Enrichment analysis of the number of genes per phylostratum in clusters of co-transcribed genes as inferred through soft *k*-means clustering and shown in Extended Data Fig. 3c–e, for *O. fusiformis* (**f**), *C. teleta* (**g**), and *D. gyrociliatus* (**h**). For each cluster and phylostratum combination, the Bonferroni-adjusted *p*–value of the two-tailed Fisher's exact test is shown. Upper tables include significantly overrepresented lineages (odds ratio, OR > 1, adjusted p–value < 0.05). Lower tables include significantly underrepresented lineages (OR < 1, adjusted p–value < 0.05). Shaded grey areas indicate clusters of genes with peak expression at the mitraria larva, for *O. fusiformis*; and stage 4tt through stage 7 larval stages, for *C. teleta*.

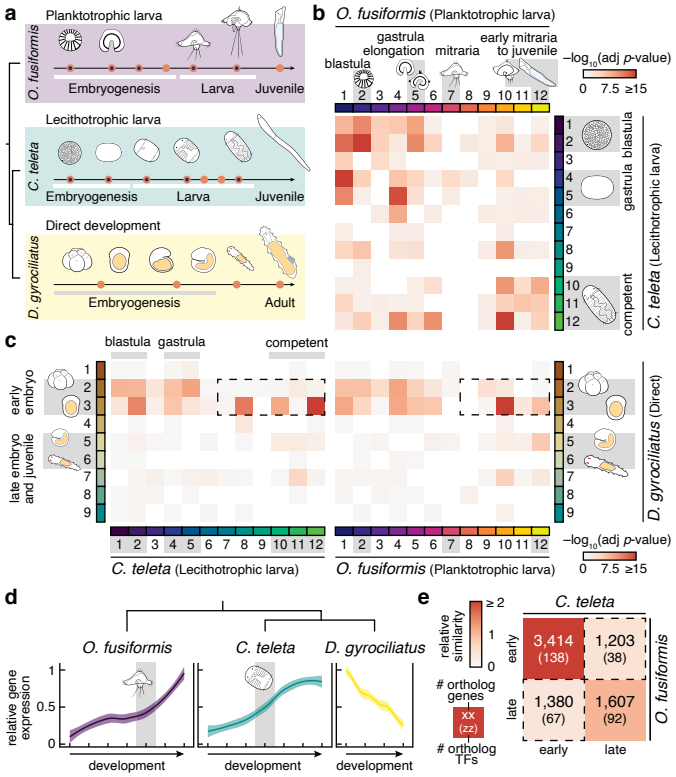
Extended Data Figure 9 | Bilaterian planktotrophic larvae and cnidarian larvae share maximal transcriptional similarity. a, Heatmaps of normalised transcriptomic Jensen–Shannon divergence (JSD) from pairwise comparisons of all single copy one-to-one orthologs (left), the set of common orthologs to all species (centre), and all single copy one-to-one transcription factor orthologs (right), between *O. fusiformis* and ten other metazoan

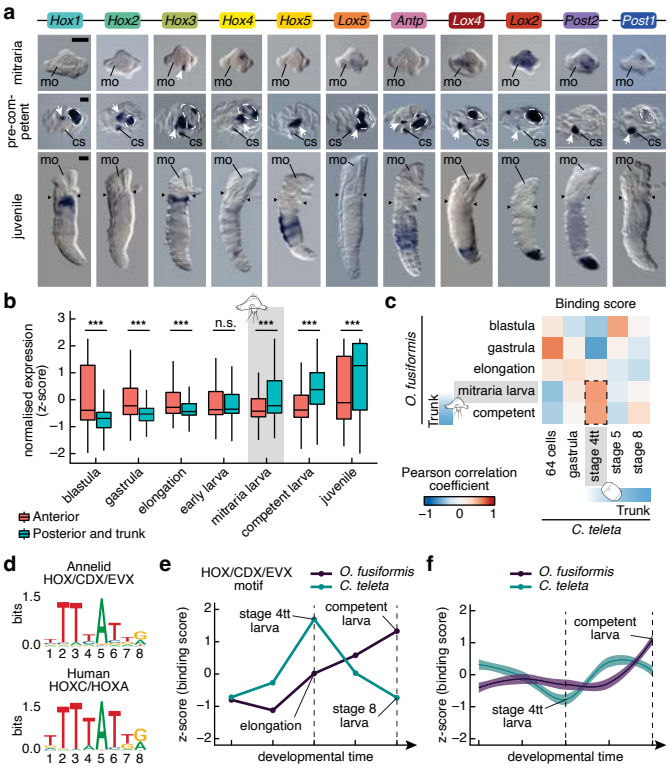
lineages with different life cycles. Larval stages are highlighted in green. **b,** Average relative JSD for the datasets shown in **a**, from stages of minimal JSD to each *O. fusiformis* stage. Confidence intervals represent the standard deviation from 250 bootstrap resamplings of the ortholog sets. **c–e,** Stages of minimal JSD to each *O. fusiformis* stage, calculated from the one-to-one ortholog set (**c**), the common ortholog set (**d**), and the one-to-one transcription factor ortholog set (**e**). Larval stages are highlighted in green. **f**, Violin plots of the gene-wise Jensen Shannon divergence (gwJSD) distributions for the pairwise comparisons of the one-to-one ortholog sets between the mitraria larva of *O. fusiformis* and the stages of minimal transcriptomic divergence as in **c**. for *C. gigas* (n = 6,737 single copy orthologs), *C. hemisphaerica* (n = 4,691), *C. teleta* (n = 7,651), *N. vectensis* (n = 5,254), and *S. purpuratus* (n = 5,015). Boxes represent mean estimate ± standard deviation. Dotted lines mark the point of highest probability density. Genes below ¼ of this point were subset as similarity-driving genes. **g,** Biological process GO terms enrichment of the five similarity-driving gene sets. GO terms were clustered by semantic similarity into 4 clusters. Each row represents a single GO term, for which the $-\log_{10}(p\text{–value})$ for each gene set is shown in a colour-coded scale.

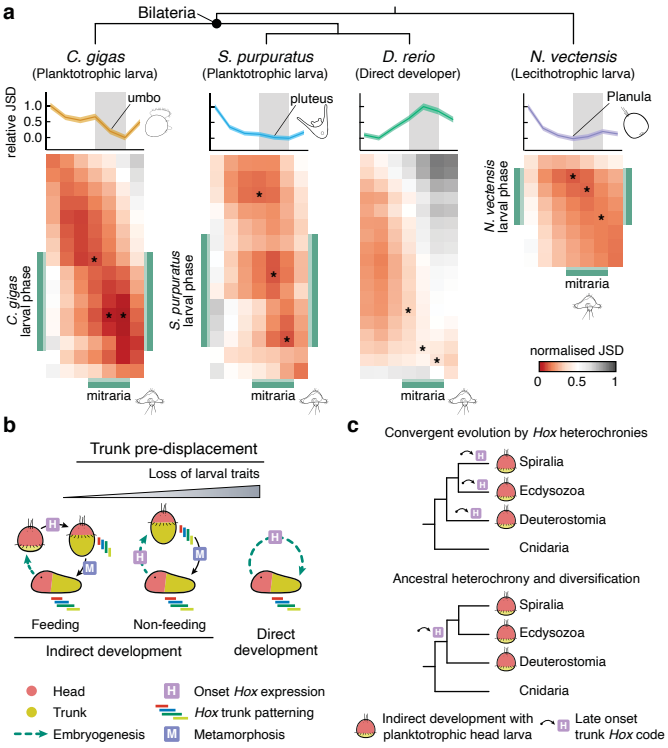
Extended Data Figure 10 | Comparative transcriptomic analysis of metazoan life cycles. **a,** Matrix of heatmaps of normalised transcriptomic Jensen–Shannon divergence (JSD) from pairwise comparisons of all single copy one-to-one orthologs between all eleven metazoan lineages. From top to bottom and left to right: the annelids *O. fusiformis* and *C. teleta*, the bivalve *C. gigas*, the nematode *C. elegans*, the insect *D. melanogaster*, the vertebrate *D. rerio*, the cephalochordate *B. lanceolatum*, the sea urchin *S. purpuratus*, the cnidarians *N. vectensis* and *C. hemisphaerica*, and the poriferan *A. queenslandica*. **b,** Proposed evolutionary scenario for larval and life cycle evolution in Annelida. Post-embryonic trunk

patterning is likely an ancestral condition with the convergent pre-displacement of trunk differentiation to embryogenesis concurring with the evolution of indirect development with feeding larva and direct development. Drawings are not to scale. **c,** Expression dynamics of *Hox* genes across the developmental RNA-seq time courses of all eleven species from **a** and the echiuran annelid *U. unicinctus*. Heatmaps were vertically aligned at the blastula, gastrula, and juvenile stages for all species. Lophotrochozoan lineages with trochophore larvae were also vertically aligned at the trochophore stage. Dotted lines encompass the larval stages of species with ciliated larvae. See Extended Data Figure 5e and Supplementary Figure 33 for the fully labelled and non-deformed heatmaps. **d**, Alternative evolutionary scenarios for the deployment of *Hox* genes (as proxy for trunk patterning and assuming the staggered expression along the directive axis of cnidarians and anteroposterior axis of bilaterians is homologous, which does not necessarily imply homology of the two axes). Given our current understanding of *Hox* gene deployment in cnidarian and bilaterian taxa, a late post-embryonic *Hox* patterning ancestral to Bilateria and Cnidaria, as seen in extant lineages with maximal indirect development, is a more parsimonious scenario (on the right).


a
O. fusiformis
Magelona spp.
DIC
Ac Tub nuclei
Owenida larvae
at
pt
mo
an
ch
b
C. gigas
B. lanceolatum
S. purpuratus
N. vectensis
O. fusiformis
he
Adult
H. sapiens
Cnidaria
Xenacoelomorpha
Deuterostomia
Ecdysozoa
Annelida
Spiralia
Principal component 2 (9.5% explained variance)
Principal component 1 (13.3% explained variance)
c
H. miamia
S. kowalevskii
S. purpuratus
B. lanceolatum
L. oculeatus
H. sapiens
S. maritima
T. castaneum
L. gigantea
C. gigas
M. yessoensis
S. mediterranea
N. geniculatus
O. fusiformis
D. gyrociliatus
L. luymesi
C. teleta
H. robusta
E. andrei
P. australis
L. anatina
Annelida
Percentage of retained metazoan orthogroups
d
Pecten maximus (Mollusca)
Owenia fusiformis (Annelida)
A1 A2 B1 B2 B3 C1 C2 D E F G H I J1 J2 K L M N O1 O2 P Q


a
Planktotrophic larva
O. fusiformis
Embryogenesis
Larva
Juvenile
Lecithotrophic larva
C. teleta
Embryogenesis
Larva
Juvenile
Direct development
D. gyrociliatus
Embryogenesis
Adult
b
O. fusiformis (Planktotrophic larva)
blastula
gastrula elongation
mitraria
early mitraria to juvenile
−log10(adj p-value)
0 7.5 ≥15
1 2 3 4 5 6 7 8 9 10 11 12
blastula
gastrula
competent
C. teleta (Lecithotrophic larva)
c
blastula
gastrula
competent
early embryo
late embryo and juvenile
D. gyrociliatus (Direct)
C. teleta (Lecithotrophic larva)
O. fusiformis (Planktotrophic larva)
−log10(adj p-value)
0 7.5 ≥15
d
O. fusiformis
C. teleta
D. gyrociliatus
relative gene expression
1
0.5
0
development
development
development
e
relative similarity
≥2
1
0
ortholog genes
xx
(zz)
ortholog TFs
C. teleta
early
late
3,414
(138)
1,203
(38)
1,380
(67)
1,607
(92)
early
late
O. fusiformis


a
Hox1
Hox2
Hox3
Hox4
Hox5
Lox5
Antp
Lox4
Lox2
Post2
Post1
mitraria
pre-competent
juvenile
mo
cs
b
normalised expression (z-score)
***
n.s.
blastula
gastrula
elongation
early larva
mitraria larva
competent larva
juvenile
Anterior
Posterior and trunk
c
Binding score
O. fusiformis
blastula
gastrula
elongation
mitraria larva
competent
Trunk
64 cells
gastrula
stage 4tt
stage 5
stage 8
Trunk
C. teleta
Pearson correlation coefficient
−1
0
1
d
Annelid HOX/CDX/EVX
Human HOXC/HOXA
bits
e
HOX/CDX/EVX motif
O. fusiformis
C. teleta
stage 4tt larva
competent larva
elongation
stage 8 larva
z-score (binding score)
developmental time
f
O. fusiformis
C. teleta
competent larva
stage 4tt larva
z-score (binding score)
developmental time


a
Bilateria
C. gigas
(Planktotrophic larva)
S. purpuratus
(Planktotrophic larva)
D. rerio
(Direct developer)
N. vectensis
(Lecithotrophic larva)
relative JSD
1.0
0.5
0.0
umbo
pluteus
Planula
C. gigas larval phase
S. purpuratus larval phase
N. vectensis larval phase
mitraria
normalised JSD
0
0.5
1
b
Trunk pre-displacement
Loss of larval traits
Feeding
Non-feeding
Indirect development
Direct development
Head
Trunk
Embryogenesis
Onset Hox expression
Hox trunk patterning
Metamorphosis
c
Convergent evolution by Hox heterochronies
Spiralia
Ecdysozoa
Deuterostomia
Cnidaria
Ancestral heterochrony and diversification
Indirect development with planktotrophic head larva
Late onset trunk Hox code