Conservative route to genome compaction in a miniature annelid

Martín-Durán, José M.; Vellutini, Bruno C.; Marlétaz, Ferdinand; Cetrangolo, Viviana; Cvetesic, Nevena; Thiel, Daniel; Henriet, Simon; Grau-Bové, Xavier; Carrillo-Baltodano, Allan M.; Gu, Wenjia; Kerbl, Alexandra; Marquez, Yamile; Bekkouche, Nicolas; Chourrout, Daniel; Gómez-Skarmeta, Jose Luis; Irimia, Manuel; Lenhard, Boris; Worsaae, Katrine; Hejnol, Andreas

Published in:
Nature Ecology and Evolution

DOI:
10.1038/s41559-020-01327-6

Publication date:
2021

Document version
Publisher's PDF, also known as Version of record

Document license:
CC BY

Citation for published version (APA):
Martín-Durán, J. M., Vellutini, B. C., Marlétaz, F., Cetrangolo, V., Cvetesic, N., Thiel, D., Henriet, S., Grau-Bové, X., Carrillo-Baltodano, A. M., Gu, W., Kerbl, A., Marquez, Y., Bekkouche, N., Chourrout, D., Gómez-Skarmeta, J. L., Irimia, M., Lenhard, B., Worsaae, K., & Hejnol, A. (2021). Conservative route to genome compaction in a miniature annelid. Nature Ecology and Evolution, 5(2), 231-242. https://doi.org/10.1038/s41559-020-01327-6
Conservative route to genome compaction in a miniature annelid

José M. Martín-Durán, Bruno C. Vellutini, Ferdinand Marlétaz, Viviana Cetrangolo, Nevena Cvetesic, Daniel Thiel, Simon Henriet, Xavier Grau-Bové, Allan M. Carrillo-Baltodano, Wenjia Gu, Alexandre Kerbl, Yamile Marquez, Nicolas Bekkouche, Daniel Chourrout, Jose Luis Gómez-Skarmeta, Manuel Irimia, Manuel Irimia, Jose Luis Gómez-Skarmeta, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manuel Irimia, Manual
specialized sensing and copulatory organs\(^\text{a}\). Despite their miniature size, *D. gyrociliatus* retain ancestral annelid traits, such as a molecularly regionalized nervous system in the female\(^\text{b}\)\(^\text{c}\) and the typical quartet spiral cleavage (Fig. 1b). With only a few genomes sequenced (Supplementary Table 1), annelids have retained ancestral spiralian and bilaterian genomic features\(^\text{d}\). Therefore, *D. gyrociliatus*, with its reduced genome size and small body, is a unique system in which to investigate the genome architecture and regulatory changes associated with genome compaction and to assess the interplay between genomic and morphological miniaturization.

### Results

We performed long-read PacBio sequencing (Extended Data Fig. 1a) to generate a highly contiguous (N50, 2.24 Mb) and complete (95.8% BUSCO genes) ~78 Mb-long haploid assembly, comparable in quality to other published annelid genomes (Extended Data Fig. 1d,e and Supplementary Table 1). Flow cytometry measurements and K-mer based analyses estimated the size of *D. gyrociliatus* genome to be 73.82 Mb and 70.95 Mb, respectively (Fig. 1c,d), agreeing with previous estimations\(^\text{41}\). While their simple morphology originally prompted them to be considered as early-branching annelids\(^\text{51}\)\(^\text{52}\) as 73.82 Mb. *D. gyrociliatus* *T. axi*, *C. elegans* cytometry analysis using the nematode *C. elegans* as reference and propidium iodide (PI) nuclear intensity estimates the genome size of *D. gyrociliatus* as 70.95 Mb and 92.47 Mb, respectively (Fig. 1c,d), agreeing with previous estimations\(^\text{41}\). While their simple morphology originally prompted them to be considered as early-branching annelids 51,52 and even larger untranslated regions (UTRs) (Extended Data Fig. 3c). As in some insect and nematode clades\(^\text{a}\), the most abundant TE class is a Ty3-gypsy-like long terminal repeat (LTR) retrotransposon that appears to be an annelid- or *D. gyrociliatus*-specific subfamily, and thus we name it Dingle (Dinophilidae Gypsy-like elements) (Extended Data Fig. 3c). As in some insect and nematode clades\(^\text{a}\), LTR retrotransposon envelope (env) proteins are apparently related to env proteins of DNA viruses, Dingle envelope (env) protein shows similarities with envelope glycoprotein B precursors of cytomegalovirus (CMV) and herpesviridae-1 (HSV-1) (Extended Data Fig. 3d,e). Compared to species with minimal genome sizes, *D. gyrociliatus* TE load is three to four times lower than in the appendicularian *Okiepaulia dioica* and the tardigrade *Ramazzottius varieornatus* but around four times larger than in insects with larger, still compact genomes (~100 Mb) (Supplementary Table 5). Therefore, TE depletion contributed to genome compaction in *D. gyrociliatus* but this does not appear to be the main driving factor since other small animal genomes show even lower fractions of TEs.

To explore how changes in gene architecture influenced genome compaction, we used transcriptomic data and ab initio predictions to annotate 14,203 protein-coding genes in the *D. gyrociliatus* genome, a smaller gene repertoire than that of other annelids (Fig. 2c, Extended Data Fig. 3a,b). Most TEs (91.5%) group in four classes and, as in the annelid *Helobdella*, TEs are either old copies or very recent expansions (Fig. 2b). The most abundant TE class is a Ty3-gypsy-like long terminal repeat (LTR) retrotransposon that appears to be an annelid- or *D. gyrociliatus*-specific subfamily, and thus we name it Dingle (Dinophilidae Gypsy-like elements) (Extended Data Fig. 3c). As in some insect and nematode clades\(^\text{a}\), the most abundant TE class is a Ty3-gypsy-like long terminal repeat (LTR) retrotransposon that appears to be an annelid- or *D. gyrociliatus*-specific subfamily, and thus we name it Dingle (Dinophilidae Gypsy-like elements) (Extended Data Fig. 3c). As in some insect and nematode clades\(^\text{a}\), LTR retrotransposon envelope (env) proteins are apparently related to env proteins of DNA viruses, Dingle envelope (env) protein shows similarities with envelope glycoprotein B precursors of cytomegalovirus (CMV) and herpesviridae-1 (HSV-1) (Extended Data Fig. 3d,e). Compared to species with minimal genome sizes, *D. gyrociliatus* TE load is three to four times lower than in the appendicularian *Okiepaulia dioica* and the tardigrade *Ramazzottius varieornatus* but around four times larger than in insects with larger, still compact genomes (~100 Mb) (Supplementary Table 5). Therefore, TE depletion contributed to genome compaction in *D. gyrociliatus* but this does not appear to be the main driving factor since other small animal genomes show even lower fractions of TEs.

To explore how changes in gene architecture influenced genome compaction, we used transcriptomic data and ab initio predictions to annotate 14,203 protein-coding genes in the *D. gyrociliatus* genome, a smaller gene repertoire than that of other annelids (Fig. 2c, Extended Data Fig. 1b,c and Supplementary Table 1). However, the gene number is comparable to free-living species with similar genome sizes, such as *O. dioica* (~15,000 genes) and *R. varieornatus* (~14,000 genes). With a gene density (208.86 genes per Mb) double that in the annelids *Capitella teleta* (99.56 genes per Mb) and *Helobdella robusta* (97.5 genes per Mb), *D. gyrociliatus* has shorter intergenic regions and transcripts, but similar exon lengths and even larger untranslated regions (UTRs) (Extended Data Fig. 4a,b,d–f), suggesting that intron shortening might have contributed to genome compaction. However, although *D. gyrociliatus* shows overall very short introns (median 66 base pairs, bp)
and its splicing is thus more efficient at removing short intron sizes (Extended Data Fig. 4i), introns are not shorter on average than in C. teleta (median 57 bp) and even similar to the centipede Strigamia maritima (median 67 bp) (Fig. 2d and Extended Data Fig. 4h), both with larger genomes than D. gyrociliatus. Instead, D. gyrociliatus has fewer introns than other annelids (Fig. 2e) and exhibits an intron density comparable between D. gyrociliatus and the centipede S. maritima, which are both slow-evolving lineages with larger genomes. Dashed horizontal line indicates D. gyrociliatus median intron size. Rates of intron gain (green), intron loss (violet) and introns per kb of CDS (blue) in representative spiralian lineages and a consensus phylogeny. D. gyrociliatus has lost introns, yet at a much lower rate and preserving many more ancestral animal introns than other fast-evolving spiralian lineages, such as flatworms and rotifers. Note that intron densities in the platyhelminthes S. mediterranea and M. lignano are underestimated due to the low fraction of single-copy complete orthologues detected in these species for the BUSCO gene dataset.

To investigate how gene loss shaped the D. gyrociliatus genome and morphology, we first reconstructed clusters of orthologous genes using a dataset of 28 non-redundant proteomes covering major animal groups and estimated gene loss and gain rates. Over 80% of D. gyrociliatus genes are assigned to multispecies gene families; the highest percentage in any annelid sequenced so far (Extended Data Fig. 5a). However, 38.9% of the genes in D. gyrociliatus are in orthogroups where there is only one D. gyrociliatus sequence, and thus D. gyrociliatus has the smallest average gene family size among annelids (1.63 genes per orthogroup; Supplementary Table 7). Although the rate of gene family loss is greater than in C. teleta, an annelid species with a conservatively evolving genome²⁶, gene loss in D. gyrociliatus is similar to those of the annelids H. robusta and Hydrodies elegans, species with larger genomes (Fig. 3a and Extended Data Fig. 5b). Therefore, our data suggest that reduction of gene family size outweighs complete gene family loss, and thus probably underpins the reduced total gene number of D. gyrociliatus, as also observed in certain Caenorhabditis species of small genome size²⁶,³⁷.

Consistent with the streamlining of its gene repertoire, we detected only nine expanded gene families in D. gyrociliatus (but 73 and 42 in C. teleta and H. robusta, respectively), most of them corresponding to locally duplicated genes implicated in immune responses (Extended Data Fig. 5c–e). In addition, D. gyrociliatus shows canonical repertoires of gene families expanded in other annelids, such as G-protein-coupled receptors (GPCRs) and epithelial sodium channels (ENaCs)⁵⁰ (Extended Data Fig. 6a,b and Supplementary Table 8). The GPCR complement of genomes is dynamic and often linked to specific (neuro)physiological adaptations, as seen in lineages with miniature genomes that have experienced either losses (for example, O. dioica lacks Class C, glutamate...
Fig. 3 | D. gyrociatus has retained a conserved developmental toolkit and ancestral linkage blocks. a. Number of gene family gains (green) and losses (violet) in representative spiralian lineages under a consensus tree topology. Gene loss in D. gyrociatus is similar to or lower than that observed in other fast-evolving spiralian lineages. b. D. gyrociatus has a conserved Hox complement, organized in a compact cluster (top). Whole-mount in situ hybridization during embryogenesis reveals that Hox genes exhibit staggered anteroposterior domains of expression, but not temporal collinear expression domains (arrowheads) along the trunk region, with Hox1, Hox5 and Antp further exhibiting anterior head expression domains (arrows). Dashed lines in lateral views of early and late elongation timepoints demarcate the head–trunk boundary and asterisks mark the anterior end. Scale bar, 50 μm. c. Oxford dot plots of orthologous genes between the scallop M. yessoensis and three annelid genomes. Orthologous genes are coloured according to their position in M. yessoensis linkage groups. The presence of an organized Hox cluster correlates with the preservation of some macrosyntenic blocks (areas of higher density of shared orthologues) in D. gyrociatus, which are lost in the fast-evolving H. robusta.

receptors) or expansions (for example, C. elegans and R. varieornatus expanded Class A, rhodopsin receptors) (Extended Data Fig. 6b). Thus, the conserved GPCR repertoire and the canonical neuropeptide complement (Extended Data Fig. 6c) further support that D. gyrociatus nervous system is functionally equivalent to, although morphologically smaller than, that of larger annelids.

Despite its miniature body plan, D. gyrociatus has an overall conserved developmental toolkit at the level of both transcription factors and signalling pathways (Extended Data Fig. 5f,g). D. gyrociatus, and Dinophilidae generally, exhibit a limited repertoire of certain extracellular signalling molecules (for example, Wnt and TGF-β ligands) and lacks bona fide FGF and VEGF ligands (Extended Data Fig. 5g–i). However, these simplifications do not affect the receptor repertoire (Extended Data Fig. 5j). Unlike appendicularians, tardigrades and nematodes with compact genomes, D. gyrociatus exhibits a compact, ordered Hox cluster, only lacking lox2 and post1 (Fig. 3b and Extended Data Fig. 7a,b). In other annelids, post1 is separate from the main Hox cluster, and as in brachiopods, it is expressed in chaetoblasts, supporting the homology of this new cell-type. Remarkably, the distantly related H. robusta and D. gyrociatus both lack chaetae, post1 and FGF ligand (also expressed in annelid chaetoblasts, Extended Data Fig. 5k–r), suggesting that the secondary loss of chaetae followed convergent routes of gene loss in different annelid species.

To investigate whether the clustered Hox genes of D. gyrociatus exhibit temporal collinearity, we first performed comparative transcriptomics at four different stages of the D. gyrociatus female life cycle (Extended Data Fig. 8a,b). Genome-wide expression dynamics revealed five main clusters of coregulated genes (Extended Data Fig. 8c), corresponding to major developmental events, such as cell
proliferation in early development or during adult growth (clusters 5 and 4, respectively), sex differentiation (cluster 2), nervous system maturation during late embryogenesis and postembryogenesis (cluster 1) and increased metabolism after hatching (cluster 3). While there is a gradual increase in gene upregulation as embryo-progression during late embryogenesis and postembryogenesis (cluster 1) and increased metabolism after hatching (cluster 3). While there is a gradual increase in gene upregulation as embryo-progression during late embryogenesis and postembryogenesis (cluster 1) and increased metabolism after hatching (cluster 3).
sequencing (ATAC-seq) to identify ~10,000 reproducible open chromatin regions in adult D. melanogaster females (Extended Data Fig. 9a–d). Open chromatin regions are short in D. melanogaster and mostly found in promoters (Fig. 5a,b), consistent with its small genome size and small intergenic regions. Despite the generally short intron size in D. melanogaster, 944 ATAC-seq peaks were in intronic regions substantially larger than non-regulatory introns (Fig. 5c). We recovered a canonical regulatory profile (Fig. 5d), which together with the lack of putative spliced leaders in 5’ UTRs (Extended Data Fig. 4g), suggests that trans-splicing and operons

**Fig. 5 | The regulatory genomic landscape of D. gyrociliatus.** a. Violin plot depicting ATAC-seq peak size distribution in D. gyrociliatus compared to the median values in the fly D. melanogaster and humans. The open chromatin regions are shorter in D. melanogaster than in other animal genomes. b. Distribution of ATAC-seq peaks according to genomic feature. Most of the open chromatin regions are found in promoters, intergenic regions and first introns. c. Violin plots of size distributions in introns with and without ATAC-seq peaks. The presence/absence of open chromatin regions in introns correlates positively with size. d. Metagene profile of ATAC-seq signal. All gene lengths are adjusted to 2 kb. e. Top ten most-significant motifs identified in D. gyrociliatus ATAC-seq peaks. The most abundant motif in open chromatin regions corresponds to CTCF. f. Tag clusters located on the dominant Cage-supported TSS (CTSS) are usually narrow (based on interquantile range 0.1–0.9) (f) and retain the canonical metazoan polymerase II initiation pyrimidine (C, T)/purine (A, G) dinucleotides (g). h. Most (11,245 out of 13,693) of the CTSS have a TATA-box and/or a downstream promoter element (DPE). i. Nucleosomes are consistently located after the CTSS (i), regardless of the promoter type (j). k. While genes with a TATA-box tend to be slightly narrower on average (l), there are no major differences in expression levels between genes with different promoter elements (k).
Fig. 6 | A new conservative route to genome compaction in D. gyrociliatus.

a. Schematic diagram of the genomic changes which occurred during genome compaction and morphological miniaturization in D. gyrociliatus and Dinophilidae. b. D. gyrociliatus genome represents a more conservative evolutionary pathway to genome compaction compared to the more drastic genomic changes experienced by other bilaterian lineages with compact genomes, such as D. dioica and C. elegans.

Discussion

Our study demonstrates that genome compaction and morphological miniaturization are specificities of D. gyrociliatus (Fig. 1e), grounded in a nested phylogenetic position within Annelida, TE depletion, intergenic region shortening, intron loss and streamlining of the gene complement and genome regulatory landscape (Fig. 2a,e, Fig. 3a and Fig. 5a,f). Traditionally, morphological miniaturization in D. gyrociliatus and Dinophiliformia has been considered a case of progenesis (underdevelopment) and mesoderma derivatives like coeloms (VEGF ligand). However, cis-regulation of gene expression is mostly restricted to the proximal regions in Dimorphophila (Fig. 5b). Therefore, our study suggests that coordinated distal gene regulation, which is an animal innovation whose emergence has been associated with the evolution of sophisticated gene regulatory landscapes and morphological diversification, is also limited in D. gyrociliatus.

Unlike in other cases of genomic compaction, the CTCF DNA-binding motif was the most abundant in active regulatory regions, located mostly in promoters and as single motifs (Fig. 5e and Extended Data Fig. 9e–h). Unlike nematodes with compact genomes, which lack CTFC, the D. gyrociliatus genome encodes for a CTFC orthologue (Supplementary Fig. 8). However, localization of CTFC DNA-binding motifs, for the most part close to transcriptional start sites, instead of in intergenic regions, suggests that CTFC might play a role in regulating gene expression in D. gyrociliatus rather than in chromatin architecture as seen in vertebrates. Thus, our data indicate that D. gyrociliatus has retained conserved genomic regulatory features (for example, lack of operons and trans-splicing, and presence of CTFC) but streamlined regulatory regions and potentially lost distal intergenic cis-regulatory elements with genome compaction.

Since most regulatory information is restricted to promoter regions (<1 kilobase (kb) upstream of the transcription start site, TSS), we applied cap analyses gene expression (CAGE)-seq to characterize promoter architecture (Extended Data Fig. 10a). Promoters are narrow (<150 bp) in D. gyrociliatus and use pyrimidine–purine dinucleotides as preferred initiators (Fig. 5f,g and Extended Data Fig. 10e). Upstream TA and downstream GC enrichment, respectively, revealed the presence of TATA-box and downstream promoter elements (DPE) in D. gyrociliatus, with TATA-box generally associated with short promoters (Fig. 5h and Extended Data Fig. 10f). Similar to vertebrates, strength of nucleosome positioning correlates with promoter broadness in D. gyrociliatus (Fig. 5i) and thus narrow TATA-box dependent promoters have lower +1 nucleosome occupancy than wide non-TATA-box promoters (Fig. 5j). As in other eukaryotes, TATA-box containing D. gyrociliatus promoters have somewhat higher expression levels, while promoters with DPE motif have no particular features, indicating this element might be non-functional (Fig. 5k,l). Therefore, the general D. gyrociliatus promoter architecture resembles that of other bilaterians (Extended Data Fig. 10g), further supporting that genomic compaction did not alter genome regulation.

Methods

Genome sequencing and assembly. Adult females of D. gyrociliatus were used to isolate genomic DNA following standard guanidium isothiocyanate protocol and RNase A treatment. Library was prepared using Pacific Biosciences 20-kb library preparation protocol and size-selected using BluePippin with 5-kb cutoff. The library was sequenced on a Pacific Bioscience RS II instrument using P6-C4 chemistry at the Norwegian Sequencing Centre. An Illumina library of median insert size of 298 bp was sequenced in 101 bases paired end mode on an Illumina HiSeq 2500 instrument at GeneCore (EMBL). All raw sequence data associated with this project are available under primary accession PRIEB37657 in the European Nucleotide Archive.

PaChio reads were filtered with SMRTAnalysis v.2.3.0.140936 and assembled with PBCR v.8.3rc2 (refs. 90,91) using default options, except for K-mer = 14 and asmmerSize = 14. Four rounds of decontamination using Blottoools v.0.9.16 (ref. 92) were applied, removing contigs with similarity to bacteria, algae, fungi or unicellular eukaryotes. A consensus assembly was generated with Quiver and improved with Pilon v.1.16 (ref. 93) using the Illumina mate pair end reads previously filtered for adaptors with cutadapt v.1.4.2 (ref. 94). We used HaploMerger2 v.20151124 (refs. 95,96) to reconstruct a high-quality haploid reference assembly, which we further scaffolded with SPSPACE-LongRead v.1.1 (ref. 97).
We collected hundreds of adult individuals of *T. axi* Remane, 1925, at the intertidal beach of Königshafen, Sylt (Germany) and extracted genomic DNA as described above to prepare a TruSeq v3 Illumina library that was sequenced in 101 bases paired end mode on a full lane of an Illumina HiSeq 2500 instrument at GeneCore (EMBL). Before assembly, we removed adapters and low-quality regions with cutadapt v.1.4.271 (ref. 95), trimmed reads using Trimmomatic raw Illumina paired end reads of -mer-based measures, we used the K runs with at least 5,000 nuclei per run. For laser (532 nm, 100 mW) to analyse the samples, performing three independent runs with at least 5,000 nuclei per run. For K-mer-based measures, we used the raw Illumina paired end reads of *D. gyrociliatus* and *T. axi*. We removed adapters using cutadapt v.1.4.271 (ref. 95), quality trimmed the reads using Trimmmomatic and duplicate reads by using a set of protein sequences comprising the reverse transcriptase domain and the integrase core domain. The phylogeny of *T. gyrociliatus* was established with a collection of sequences from the Gypsy database, including hits obtained with TBLASTN (databases NR and TSA) using *D. gyrociliatus* sequences as queries. To localise the repeats created by transposition in *D. gyrociliatus* genome resulted in 835 complete (85.4%), 27 complete and duplicated (2.8%), 75 fragmented (7.7%) and 68 missing (6.9%) (Extended Data Fig. 1e). Finally, we used KAT v.2.4.2 (ref. 95) to estimate the completeness and contig number variation of the assemblies (Supplementary Fig. 1). Genome size measurements. For flow cytometry measurements, adult *D. gyrociulates* females and *C. elegans* worms (reference) were starved for 3–4 days before analysis. *D. gyrociulates* and *C. elegans* were chopped with a razor blade in General-Purpose Buffer272 and the resulting suspension of nuclei was filtered through a 30-µm nylon mesh and stained with propidium iodide. We used a flow cytometer Partex CyFlow Space fitted with a Cobalt Samba green laser (532 nm, 100 mW) to analyse the samples, performing three independent runs with at least 5,000 nuclei per run. For K-mer-based measures, we used the raw Illumina paired end reads of *D. gyrociulates* and *T. axi*. We removed adapters using cutadapt v.1.4.271 (ref. 95), quality trimmed the reads using Trimmmomatic and duplicate reads by using a set of protein sequences comprising the reverse transcriptase domain and the integrase core domain. The phylogeny of *T. gyrociliatus* was established with a collection of sequences from the Gypsy database, including hits obtained with TBLASTN (databases NR and TSA) using *D. gyrociliatus* sequences as queries. To localise the repeats created by transposition in *D. gyrociliatus* genome resulted in 835 complete (85.4%), 27 complete and duplicated (2.8%), 75 fragmented (7.7%) and 68 missing (6.9%) (Extended Data Fig. 1e). Finally, we used KAT v.2.4.2 (ref. 95) to estimate the completeness and contig number variation of the assemblies (Supplementary Fig. 1). Transcriptome sequencing and assembly. A publicly available dataset (Sequence Read Archive (SRA), accession number SRX2030658) was used to generate a de novo transcriptome assembly as previously described98. Redundant contigs were removed using the cd-hit-est program with default parameters of CD-HIT (ref. 99) and CAPS (ref. 100). Additionally, we used that dataset to generate a genome-guided assembly using CAP3 (ref. 110) and Trinity (ref. 111). Supplementary Table 2 shows standard statistics for the de novo and genome-guided assemblies calculated with Transrate112. Transcriptome completeness was evaluated with BUSCO v.2 (ref. 113). Stage-specific RNA-seq. Two biological replicates of four developmental stages of *D. gyrociulates* (early embryo, 1–3-days-old; late embryo, 4–6-days-old; juvenile females, 7–9-days-old; and adult females, 20–23 days-old) were used to isolate total RNA with TR1 Reagent Solution (Applied Biosystems) following manufacturer's recommendations and generate Illumina short-reads on a NextSeq 500 High platform in 75-base paired end reads mode and a -270 bp library mean insert size at GeneCore (EMBL). We pseudo-aligned reads to *D. gyrociulates* filtered gene models with Kallisto v.0.44.0 (ref. 114), and followed the standard workflow of DSEseq (ref. 115) to estimate counts, calculate size factors, estimate the data dispersion, and perform a gene-level differential expression analysis per consecutive stages (Supplementary Data 1). Datasets were first corrected for low count and high dispersion values using the apeglm log-fold change shrinkage estimator116, and then compared using Wald tests between contrasts. For clustering and visualization, we homogenized the variance across expression ranks by applying a variance-stabilizing transformation to the DESeq2 datasets. We used the pheatmap package to create heatmaps117, the package EnhancedVolcano for volcano plots118 and ggplot2 for the remaining plots119. To characterize and identify enriched gene ontology terms, we used the package clusterProfiler120. All analyses were performed in R (ref. 121) using the RStudio Desktop122. Phylogenetic analysis. Annelid transcriptomes (Supplementary Data 1) were downloaded from SRA and assembled using Trinity v.2.5.1 (ref. 123) with the Trimmeromatic read trimming option. Transcriptiones were then translated using Transdecoder v.5.0.2 (ref. 124) after searching for similarity against the metazoan Swissprot database. Predicted proteins were searched using HHMMER125 for 1,148 single-copy phylogenetic markers previously described126 using reciprocal BLAST to discard possible paralogues and character supermatrix was assembled as described before127. From this initial dataset, we selected the 264 genes with lowest saturation, yielding a concatenated alignment of 71,508 positions (as the analysis of the full dataset with site-heterogeneous models was not computationally tractable). Phylogenetic analyses were performed on the concatenated alignment using IQ-TREE22 with a GTR+G model, an LG matrix to account for transition rates within each profile, the FreeRate heterogeneity model (R4) to describe across sites evolution rates, and an optimization of amino acid frequencies using maximum likelihood. Support values were drawn from 1,000 ultrafast bootstraps with NNi optimization. We also carried out Bayesian reconstruction using the site-heterogeneous CAT + GTR + Gamma model running two chains for <1,000 iterations. We reached reasonable convergence for one of the datasets (bdppf 0.19). Annotation of repeats and transposable elements. We used RepeatModeler v.1,0,4.9 (ref. 123) and RepeatMasker open-4.0 (ref. 124) to generate an automated annotation of REs and repetitive elements on the *D. gyrociulates* genome as described before124. From this initial dataset, we selected the 264 genes with highly abundant repeats in *D. gyrociulates* and *C. elegans*. To enrich gene ontology terms, we used the package clusterProfiler120. All analyses were performed in R (ref. 121) using the RStudio Desktop122. Gene prediction and functional annotation. The predicted set of core eukaryotic genes generated by CEGMA129 was used to train and run AUGUSTUS v.3.2.1 (ref. 130). The predicted proteomes of the annelids *C. teleta* and *H. robusta* were aligned to the *D. gyrociulates* genome using EXONERATE v.2.2.0 (ref. 131) and PASA v.2.0.2 (ref. 132) was used to align the transcriptome to the genome with BLAT and GMAP aligners133,134, EvidenceModeler v.1.1.1 (ref. 135) was used to generate weighted consensus gene predictions, giving a weight of 1 to ab initio gene predictions and spotted protein alignments, and a weight of 10 to the PASA transcript assemblies. EvidenceModeler output was used to refine PASA gene models and generate alternative splice variants. Predictions with BLAST hit against transposons and/or with an overlap ≥90% on masked regions were removed. The final prediction set contains 14,203 coding-protein loci that generate 17,409 different peptides. We used ORF-Finder136 to refine TSS with CAGE-seq data. Functional annotation for the 17,409 different transcripts was performed with Trinotate v.3.0. We retrieved a functional annotation for 13,437 gene models (77.18%). Gene structure evolution. We compared genome-wide values of gene structure parameters among *D. gyrociulates*, *C. teleta*, *H. robusta*, *D. melanogaster*, *C. elegans* and *D. discoideum* (Supplementary Table 6). To identify splice leader sequences in *D. gyrociulates*, we predicted protein-coding sequences in the de novo assembled transcriptome with Transdecoder v.5.0.0 (ref. 137) and used the script nr. ORFs, gf3 file as input against transposons and/or with an overlap ≥90% on masked regions were removed. The final prediction set contains 14,203 coding-protein loci that generate 17,409 different peptides. We used ORF-Finder136 to refine TSS with CAGE-seq data. Functional annotation for the 17,409 different transcripts was performed with Transrate v.3.0. We retrieved a functional annotation for 13,437 gene models (77.18%). Intron evolution analysis. We compared distributions of intron lengths between *D. gyrociulates*, *Homo sapiens*, *C. teleta*, *Crassostrea gigas*, *Lottia gigantea*, *Strigamia maritima* and *Branchiostoma lanceolatum* (Supplementary Table 6) using only introns in genes with orthologues across the seven species (as defined by OrthoFinder; see below) and orthologous with less than four paralogues per species. To identify conserved and new *D. gyrociulates* introns, we aligned each *D. gyrociulates* protein sequence against each annotated protein isoform of each orthologous gene of the abovementioned six species and added the intron positions into the alignments138. To identify high-confidence conserved intron positions, we required that a given *D. gyrociulates* intron position was found at the exact position of the alignment and with the same phase (0 or 2 in at least four out of six other species. To define high-confidence non-conserved (probably new) introns, we required that a *D. gyrociulates* intron position did not match an intron position with the same phase within 25 alignment positions in any of the other six species. To assess the impact of intron length on splicing efficiency on *D. gyrociulates*, *S. maritima* and *H. sapiens*, we used RNA-seq based quantifications of intron
retention as previous described14 and implemented by vast-tools15. Only introns that had sufficient read coverage16,17 were used to calculate average PIR.

To quantify intron gain and loss in D. gyrociliatus we generated a database of homologous intron clusters across 28 metazoan genomes (Supplementary Table 6), obtaining one-to-one orthologous genes using BUSCO v.3 (ref. 18) (prot mode and 1×10^−4 E value) and the OrthoDB v.9 (ref. 19) dataset of 978 single-copy animal orthologues. We aligned the predicted peptides using MAFFT v.7.310 G-INS-i algorithm19 and used Malin20 to identify conserved intron sites and infer their conservation status in ancestral nodes. We extracted the loss rates of intron gain and loss in each node with Malin’s built-in model maximum-likelihood optimization procedure. We used this model to estimate the posterior probabilities of intron presence, gain and loss in extant and ancestral nodes. For each node, we calculated the intron density expressed as introns per kb of coding sequence (introns per CDS kb), as follows:

\[
\text{intron density} = \frac{\text{(num introns)/ (num genes)/ (median gene length)}}{1000}
\]

where num introns, is the number of introns present in a given node (extant or ancestral, corrected by missing sites), num genes = 978 (number of alignments of one-to-one orthologues) and median gene length = 682.5 bp (as obtained from the lengths of the seed proteins curated in the OrthoDB v.9 MetaTao database; ‘ancestral’ FASTA file). We used the same strategy to obtain the rates of intron gain and loss per node in terms of introns per CDS kb. In addition, we inferred the uncertainty of the estimated intron gains, losses and presence values with Malin and 1,000 bootstrap iterations. To visualize the evolution of intron content, we used the ape library v.5.0 (ref. 21) from the R statistical package v.3.6 (ref. 22). To calculate the percentage of ancestral metazoan introns retained in each species, we retrieved all introns present in the last common metazoan ancestor (at >99% similarity; n = 3,024) and calculated the sum of their presence probabilities in extant species.

Gene family evolution analyses. We used OrthoFinder v.2.2.7 (ref. 23) with default values to reconstruct clusters of orthologous genes between D. gyrociliatus and 27 other animal proteomes (Supplementary Table 6). OrthoFinder gene families were used to infer gene family gains and losses at different nodes using the ETE 3 library.23 Gene expansions were computed for each species using a hypergeometric test against the median gene number per species for a given family. We used the functionally annotated gene sets of D. gyrociliatus, C. teleta and H. robusta to identify their repertoire of transcription factors, ligands and receptors. If a gene was not in the annotated D. gyrociliatus genome assembly, we performed manual search via BLAST on the de novo and genome-guided transcriptome. For T. axi and D. vorticoides, gene identification was conducted on the assembled transcriptome via manual BLAST searches. To reconstruct KEGG pathways via KEGG Mapper24, we used the functional annotations obtained from Trinotate to extract KEGG IDs. GPCR sequences in D. gyrociliatus and other animals (Supplementary Table 7) were retrieved using HMMER v.3.2.1 (ref. 25) (E value cutoff <0.01) with Pfam profiles of class A (PF00001), class B (PF00002), class C (PF00003) and class F (PF00134) GPCRs (according to GRAFS classification).

Sequences from each class were tested for false positives from other classes (including cAMP-slime-mold class E GPCRs, PF05462). Phylogenetic analyses of GPCRs were performed as described elsewhere.26 Neuropeptide candidates (Supplementary Data 2) were retrieved by a combination of BLAST searches (E value cutoff <0.1) and the use of a customized script26 to detect cleavage patterns on precursors.

Orthology assignment. Multiple protein alignments were constructed with MAFFT v7.310 G-INS-i (ref. 19) poorly aligned regions were either removed by hand or with gBlocks27. Maximum likelihood trees were constructed with FastTree 2 (ref. 28) using default parameters and visualized with FigTree.

Gene expression analyses. D. gyrociliatus embryos were collected in their egg clusters and manually dissected. The embryonic eggshell was digested in a solution of 1% sodium thioglycolate (Sigma-Aldrich, T0632) and 0.05% protease (Sigma-Aldrich, P5147) in seawater, pH 8.2 for 30 min at room temperature, followed by relaxation in MgCl2 and fixation. Whole-mount in situ hybridization (WISH) was performed as described elsewhere.29 Images were taken with a Zeiss Axiocam HRc connected to a Zeiss Axioscope Ax10 using bright-field Nomarski (WMISH) was performed as described elsewhere47. Images were taken with a Zeiss microscope (MRC LMS). Demultiplexed CAGE reads (47 bp) were mapped to the D. gyrociliatus genome assembly using Bowtie2 (ref. 48) and resulting bam files were imported into R using the standard CAGEr package (v.1.20.0) and G-correction workflow.49 Normalization was performed using a referent power-law distribution50 and CAGE-derived TSSs that passed the threshold of 1 transcript per million (TPM) were clustered together using distance-based clustering (Supplementary Data 1). Genomic locations of tag clusters were determined using the ChiPseeker package and gene model annotations, where promoters were defined to include 500 bp upstream and 1000 bp downstream of the annotated transcript start site. Visualization of motifs, sequence patterns or reads coverage was performed using Heatmaps and seqPattern Bioconductor packages.

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

Data availability

All new raw sequence data associated with this project are available under primary accession PRJEB3736757 in the European Nucleotide Archive. Genome annotation files and additional datasets are available in https://github.com/ChemaMD/DimorphilusGenome.

Code availability

All custom code used in this study is freely available in https://github.com/fmarfietz/comp_genomics.

Received: 21 April 2020; Accepted: 15 September 2020; Published online: 16 November 2020

References

1. Gregory, T. R. in The Evolution of the Genome (ed. Gregory, T. R.) 3–87 (Academic Press, 2005).
2. Gregory, T. R. Coincidence, coevolution, or causation? DNA content, cell size, and the C-value enigma. Biol. Rev. 76, 65–101 (2001).
3. Blommaert, J., Riss, S., Hecox Lea, B., Mark Welch, D. B. & Stelzer, C. P. Small, but surprisingly repetitive genomes: transposon expansion and not polyplody has driven a doubling in genome size in a metazoan species complex. BMC Genomics 20, 466 (2019).
4. Sun, C. et al. LTR retrotransposons contribute to genomic gigantism in plethodontid salamanders. Genome Biol. Evol. 4, 168–183 (2012).
5. Naville, M. et al. Massive changes of genome size driven by expansions of non-autonomous transposable elements. Curr. Biol. 29, 1161–1168 (2019).
6. Tällä, V. et al. Rapid increase in genome size as a consequence of transposable element hyperactivity in wood-white (Leptidea) butterflies. Genome Biol. Evol. 9, 2491–2505 (2017).
7. Ohno, S. Evolution by Gene Duplication (Springer, 1970).
8. Sacerdott, C., Louis, A., Bon, C., Berthelot, C. & Roest Crollius, H. Chromosome evolution at the origin of the ancestral vertebrate genome. Genome Biol. 19, 166 (2018).
9. Brasch, L. & Postlethwait, J. H. in Polyplody and Genome Evolution (eds Sudits, P. S. & Sudits, D. E.) 341–383 (Springer, 2012).
10. Li, Z. et al. Massive multiple-scale gene and genome duplications during the evolution of hexapods. Proc. Natl Acad. Sci. USA 115, 4713–4718 (2018).
11. Sotero-Caio, C. G., Platt, R. N. 2nd, Suh, A. & Ray, D. A. Evolution and diversity of transposable elements in vertebrate genomes. Genome Biol. Evol. 9, 161–177 (2017).
12. Kapusta, A. & Suh, A. Evolution of bird genomes—a transposon’s-eye view. Ann. NY Acad. Sci. 1389, 164–185 (2017).
Articles

143. Paradis, E. & Schliep, K. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. Bioinformatics 35, 526–528 (2019).

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

145. Huerta-Cepas, I., Serra, F. & Bork, P. ETE 3: reconstruction, analysis, and visualization of phylogenomic data. Mol. Biol. Evol. 33, 1635–1638 (2016).

146. Kaneshia, M. & Sato, Y. KEGG Mapper for inferring cellular functions from protein sequences. Protein Sci. 29, 28–35 (2020).

147. Finn, R. D., Clements, J. & Eddy, S. R. HMMER web server: interactive sequence similarity searching. Nucleic Acids Res. 39, W29–W37 (2011).

148. Theil, D., Franz-Wachtel, M., Aguilera, F. & Hejnoń, A. Xenacoelomorph neuroepitodemes reveal a major expansion of neuroepitodemes systems during early bilaterian evolution. Mol. Biol. Evol. 35, 2528–2543 (2018).

149. Castresana, J. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol. Biol. Evol. 17, 540–552 (2000).

150. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2—a maximum-likelihood generating trees for large alignments. PLoS ONE 5, e9490 (2010).

151. 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).

152. Oulson, S., Bertrand, S. & Escriva, H. Evolution of the FGF gene family. Int. J. Evol. Biol. 2012, 298147 (2012).

153. Setiamarga, D. H. et al. An in-silico genomic survey to annotate genes coding for early development-relevant signaling molecules in the pearl oyster, Pinctada fucata. Zoo. Sci. 30, 877–888 (2013).

154. Simakov, O. et al. Deeply conserved synteny resolves early events in vertebrate evolution. Nat. Ecol. Evol. 4, 820–830 (2020).

155. Putnam, N. H. et al. Sea anemone genome reveals ancestral eumetazoan gene repertoire and genomic organization. Science 317, 86–94 (2007).

156. Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nat. Methods 10, 1213–1218 (2013).

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

158. Li, Q., Brown, J. B., Huang, H. & Bickel, P. J. Measuring reproducibility of high-throughput experiments. Ann. Appl. Stat. 5, 1752–1779 (2011).

159. 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).

160. Murata, M. et al. Detecting expressed genes using CAGE. Methods Mol. Biol. 1164, 67–85 (2014).

161. Haberle, V., Forrest, A. R., Hayashizaki, Y., Carninci, P. & Lenhard, B. CAGEr: precise TSS data retrieval and high-resolution promoterome mining for integrative analyses. Nucleic Acids Res. 43, e51 (2015).

162. Balwierz, P. J. et al. Methods for analyzing deep sequencing expression data: constructing the human and mouse promoterome with deepCAGE data. Genome Biol. 10, R79 (2009).

Acknowledgements

We thank all members of the Hejnoń laboratory for their continuous support, as well as the staff at the Norwegian Sequencing Centre and GeneCore (EMBL) for their technical support. We also thank R. Domínguez-Acemel, M. Silva Magri and S. Jimenez Gancedo for their help with establishing ATAC-seq in D. gryllicus. We thank J. Tena-Aguilar for his help with ATAC-seq bioinformatic analyses and M. Ezcurra for providing live C. elegans worms for flow cytometry analyses. Finally, we thank A. de Mendoza and C. Launer for a critical read of this manuscript, as well as M. Blaxter for comments and improvements. This study was supported by Sars Centre core budget and the European Research Council (ERC) grant agreement no. 648861 to A.H. J.M.M.-D. was additionally supported by the ERC grant agreement no. 801669, and B.C.V. by an EMBO Long-Term Fellowship (ALTF 74–2018). J.L.G.-S. received funding from the ERC (grant agreement no. 740041) and the Spanish Ministerio de Economía y Competitividad (grant no. BFU2016-74961-P) and the institutional grant Unidad de Excelencia María de Maeztu (MDM-2016-0687). We dedicate this manuscript to the loving memory of José Luis Gómez-Skarmeta.

Author contributions

J.M.M.-D., B.C.V., K.W. and A.H. conceived the study. J.M.M.-D., B.C.V. and A.H. designed experiments and analyses. J.M.M.-D. and V.C. performed collections and extractions. J.M.M.-D. and B.C.V. generated the genome and transcriptome assemblies. B.C.V. analysed the RNA-seq data. F.M. performed phylogenetic analyses and gene family evolution studies. V.C., A.K., N.B. and A.M.-C.-B. performed gene expression analyses. W.G. performed flow cytometry analyses. N.C. and B.L. performed and analysed CAGE-seq. J.M.M.-D. and J.L.G.-S. performed and analysed ATAC-seq. J.M.M.-D., S.H., D.T., D.C., M.L., X.G.-B. and Y.M. performed computational analyses. All authors contributed to interpretation of the results. J.M.M.-D., K.W. and A.H. wrote the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41559-020-01327-6.

Supplementary information is available for this paper at https://doi.org/10.1038/s41559-020-01327-6.

Correspondence and requests for materials should be addressed to J.M.M.-D. or A.H.

Peer review information Peer reviewer reports are available.

Reprints and permissions information is available at www.nature.com/reprints.

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

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

© The Author(s) 2020
Extended Data Fig. 1 | Sequencing approach and assembly statistics. **a-c.** Diagram of the approach taken to sequence and assemble *Dimorphilus gyrociliatus* genome and transcriptome, and to annotate coding genes in the genome. **d.** Comparison of genome assembly statistics between *D. gyrociliatus* and the annelids *C. teleta* and *H. robusta*. *D. gyrociliatus* genome is smaller than one third of *C. teleta*’s genome, and the assembly is contained in only ~350 scaffolds, with an N50 of 2.24 Mb, the second-best contiguity value for an annelid genome assembly to date. **e.** Genome completeness, as indicated by metazoan BUSCO repertoire, in genome assemblies of different annelid lineages. *D. gyrociliatus* completeness is comparable to *C. teleta*, the most conservative annelid genome sequence to date.
Extended Data Fig. 2 | Dimorphilus gyrociiliatus phylogenetic position. a–c, Maximum likelihood phylogenetic tree using the site-heterogeneous model of protein evolution C60 + R and the entire annelid dataset (a), excluding the fast-evolving Osedax and Diurodrilus lineages (b), and additionally excluding Hirudinea (c). In all cases, D. gyrociiliatus forms with Trilobodrilus axi, Dinophilus vorticoides and Lobatocerebrum sp. the clade Dinophiliformia, being this robustly placed as sister to Sedentaria + Errantia. d, Maximum likelihood tree using the site homogeneous model of protein evolution LG4X + R and the entire annelid dataset. This condition recapitulates Dinophiliformia, but places this group inside Sedentaria, related to other fast-evolving sedentarian lineages. e, Bayesian phylogenetic tree using the site-heterogeneous model of protein evolution CAT-GTR + Γ and excluding long branch lineages (Osedax, Diurodrilus and Hirudinea) recapitulates the maximum likelihood tree with the site heterogeneous model and the same dataset. In all trees, only values other than 100 bootstrap or 1 posterior probability are shown.
Extended Data Fig. 3 | The transposable element repertoire of *Dimorphilus gyrociliatus*. **a**, Graph showing the number of genes and transposable elements (TEs) per scaffold. **b**, Diagram of Scaffold002 and Scaffold026 illustrating how transposable elements (TEs, in red) often concentrate in gene-free (dark blue boxes) and closed chromatin (as indicated by ATAC-seq signal; light blue) islands. **c**, Maximum likelihood phylogeny of the pol gene showing that Dingle is a new family of Ty3/gypsy LTR element. The scale bar shows the number of substitutions per site and red dots are bootstrap values > 0.7. **d**, Genetic organization of Dingle, showing protein domains (top red boxes), 6-frame translations (green lines, ATG; black lines, stop codons) and the predicted protein structure of ENV, which shows resemblance to that of human herpes viruses. **e**, Genetic organization of ENV, showing protein domains (top red boxes), 6-frame translations (green lines, ATG; black lines, stop codons) and the predicted protein structure of ENV, which shows resemblance to that of human herpes viruses. In **c**, Ac, *Anolis carolinensis*; Aj, *Apostichopus japonicus*; At, *Arabidopsis thaliana*; Cc, *Ceratitis capitata*; Ci, *Ciona intestinalis*; Cm, *Callosobruchus maculatus*; Cs, *Ciona savignyi*; Cv; *Crassostrea virginica*; Db, *Drosophila buzzati*; Dd, *Dictyostelium discoideum*; Dg, *Dimorphilus gyrociliatus*; Dm, *Drosophila melanogaster*; Dr, *Danio rerio*; Ec, *Elliptio complanata*; Hd, *Haliotis discus hannai*; Hl, *Haliotis laevigata*; Hm, *Hydra magnipapillata*; La, *Lingula anatina*; Mv, *Mimachlamys varia*; My, *Mizuhopecten yessoensis*; Od, *Oikopleura dioica*; Pn, *Pundamilia nyererei*; Sg, *Saccostrea glomerata*; Sk, *Saccoglossus kowalevskii*; Sp, *Strongylocentrotus purpuratus*; Tn, *Trichinella nelson*; Tp, *Trichinella pseudospiralis*; Tr, *Takifugu rubripes*; Xt, *Xenopus tropicalis*. 
Extended Data Fig. 4 | Comparative analyses of gene structure features in the *Dimorphilus gyrociliatus* genome. a–f, Violin plots showing the genome-wide distribution of mRNA and exon lengths, exon numbers per gene, and the lengths of 5' UTR, 3' UTR and intergenic regions in *D. gyrociliatus*, the annelids *C. teleta* and *H. robusta*, and the bilaterians with compact genomes *C. elegans*, *D. melanogaster* and *O. dioica*. g, The distribution of occurrences of 22-mer and 50-mer in RNA-seq-based 5' UTR regions of *D. gyrociliatus* does not indicate the presence of over-represented sequences that could act as splice leaders. h, Violin plot showing the distribution of intron sizes between conserved and non-conserved introns in *D. gyrociliatus*. i, The percentage of intron retention according to intron size demonstrates that the splicing machinery in *D. gyrociliatus* is adapted to short introns, as it occurs in the centipede *S. maritima* (also with short introns) and inversely to what is observed in *H. sapiens*, a species with longer introns. j, Metazoan-wide analysis of intron density, intron gain and intron loss rates per lineage and their ancestors. Intron density (blue circles) are indicated at each node and terminal tip of the phylogram. Net intron gains and losses are indicated below the species name, together with the fraction of introns conserved in each extant genome, among the ones inferred to have been present at the last metazoan common ancestor. *D. gyrociliatus* shows rates of intron loss and retention of ancestral introns similar to other animal lineages with much larger genomes. k, Inferred origin of the intron sites in *D. gyrociliatus* and the annelids *C. teleta* and *H. robusta*, expressed as the sum of gain probabilities at their respective ancestral nodes.
Extended Data Fig. 5 | See next page for caption.
Extended Data Fig. 5 | Expansions and gene losses in the genome of *Dimorphilus gyrociiliatus*. a, Percentage of genes in multispecies orthogroups (OG; light blue) and species-specific orthogroups (dark blue and light brown) in the four studied annelid species. b, Metazoan-wide analysis of gene gain and loss, indicating the number of genes gained (in green) and lost (in red) at each node of the phylogram and the net value of gain/loss for each species. c, Heatmap of the 35 largest OG in *D. gyrociiliatus*, indicating those that correspond to lineage-specific expansions (OG size indicated in the cell, and its putative orthology). d, Expanded families indicate that these are mostly involved in immunity and are mostly local copies (light blue, e). f–g, Heatmaps depicting the repertoire of transcription factors (TFs) and ligands and receptors in *D. gyrociiliatus* and the annelids *C. teleta* and *H. robusta*. *D. gyrociiliatus* lacks a clear ortholog of the FGF and VEGF ligand in *D. gyrociiliatus*. Although *D. gyrociiliatus* has retained all developmental signalling pathways, it has severely simplified the ligand repertoire of the Wnt signalling pathway (h), and the TGF-β pathway (i), trend also observed in other members of Dinophilidae (dotted lines in *T. axi* and *D. vorticoides* indicate the reconstructed complements are based on transcriptomic data). j, However, *D. gyrociiliatus* has a conserved repertoire of frizzled and TGF-β receptors. k–m, o–q, Differential interference contrast (DIC) micrographs of whole-mount in situ hybridization of *Capitella teleta* larvae of the FGF (*Cap-fgf8/17/18/24*) and VEGF (*Cap-pvf1* and *Cap-pvf2*) ligands and phalloidin staining at these points n, r. FGF and VEGF ligands are expressed in mesodermal derivatives anterior (open arrowhead) and dorsal to the brain (red closed arrowhead), associated with the foregut (double arrowheads), the longitudinal bands (white closed arrowheads), and the posterior growth zone (black and white arrows). FGF is also expressed in well-developed and nascent chaetoblasts (black closed arrowheads). br, brain; es, oesophagus; fg, foregut; lat, lateral; pg, pygidium; ph, pharynx; pt, prototroch; tt, telotroch; vent, ventral; vlat, ventrolateral. Scale bars, 50 μm. Asterisks mark the stomodeum.
Extended Data Fig. 6 | See next page for caption.
Extended Data Fig. 6 | The GPCR and neuropeptide repertoire of *Dimorphilus gyrociliatus*. **a**, Orthology analyses of G-protein coupled receptors (GPCRs) for each class. The magenta asterisks highlight *D. gyrociliatus* receptors, and the annotations are given based on the *D. melanogaster* orthology. In (a) 5HT, serotonin; Ado, Adenosin; Akh, adipokinetic hormone; AstA, allatostatin A; AstC, allatostatin C; Boss, bride of sevenless; Capa, capability; Ccap, crustacean cardioactive peptide; CCHA, CCHamide; Cck, cholecystokinin; CNMa, CNMamide; Crz, corazonin; Dh, diuretic hormone; Dop, dopamine; Ec, ecdysteroid; ETH, ecdysis triggering hormone; FMRFa, FMRFamide; Fz, Frizzled; Glut, glutamate; Lrrc, leucine rich repeat containing; mACh, muscarinic acetylcholine; Mthl, methuselah; mtt, mangetout; Myos, myosuppressin; NpF, neuropeptide F; Oct, octopamine; Oct-R-mb, Octopamin receptor in mushroom bodies; Pdf, pigment dispersing factor; R, receptor; Rh, rhodopsin; RYa, RYamide; Sexp, sex peptide; SiFa, SiFamide; Smo, Smoothened; sNpF, short neuropeptide F; stan, starry night; Tre, trapped in endoderm; Tyr, tyramine. **b**, Phylogram with the number of GPCRs per class in representative animal species. Contrary to other animals with compact genomes and miniaturized morphologies, such as tardigrades, nematodes and appendicularians, *D. gyrociliatus* has a conserved GPCR repertoire. **c**, PSI-BLAST cluster map of *D. gyrociliatus* pro-neuropeptides, each dot corresponding to one sequence, their colour corresponds to the legend in upper left corner. Connections are based on E values < 1e-7 (see upper right corner). In (c), a, amide; ast, allatostatin; crz, corazonin; ct, calcitonin; dh, diuretic hormone; elh, ecdysis triggering hormone; ep, excitatory peptide; glyho-a, glycoprotein hormone alpha; glyho-b, glycoprotein hormone beta; gnrh, gonadotropin releasing hormone; ilp, insulin like peptide; myom, myomodulin; np-F, neuropeptide F; np-Y, neuropeptide Y; npl, neuropeptide-like; pdf, pigment dispersing factor; pedpep, pedal peptide; scap, short cardioactive peptide.
Extended Data Fig. 7 | The Hox cluster of *Dimorphilus gyrociliatus*. a, Maximum likelihood tree of Hox and ParaHox genes, with Evx proteins as outgroup, to assign orthology relationships of *Dimorphilus* Hox genes (indicated in the tree). Only bootstrap values for main orthogroups are shown. b, Schematic representation of Hox complements and Hox genomic organizations in representative spiralian species, with the putative ancestral spiralian Hox cluster on the top. Each Hox orthologous group is coloured differently. c, Heatmap of gene expression values of *Dimorphilus* Hox genes during the life cycle. d, Whole-mount in situ hybridization of Hox genes in *Dimorphilus* juveniles and adults. Only *Post2* and *Hox3* show conspicuous expression domains in the hindgut and posterior ectoderm of the juvenile, respectively. In adults, we only detect expression of *Post2* in the hindgut. e, Schematic summary of Hox gene expression in relation to the Hox genomic organization during *Dimorphilus* embryonic development. Hox genes exhibit an anteroposterior spatial collinearity along *Dimorphilus* trunk, with *Antp* and *Hox5* being additionally expressed in head domains. However, Hox genes do not exhibit temporal collinearity, as all but *Hox5*, *Antp*, and *Lox5* become expressed by the end of gastrulation. f, Oxford dot plots of orthologous genes between *D. gyrociliatus*, *C. teleta* and *H. robusta*. Macrosyntenic relationships are little conserved between annelid worms, indicating lineage-independent large-scale genomic reorganizations.
Extended Data Fig. 8 | Differential expression analyses during the life cycle of *Dimorphilus gyrociliatus*. a, Principal component analysis of the stage-specific RNA-seq samples using the top eight thousand most-variable genes. The raw count data was transformed to homogenize the variance and normalized using the variance-stabilizing method from DESeq2. b, Euclidean distances between the variance stabilized normalized counts of the stage-specific RNA-seq samples. c, Expression patterns for the top three thousand differentially expressed genes. Variance stabilized normalized counts were scaled around the mean value of the row to highlight changes in expression between developmental stages. Gene ontology terms associated with each cluster of expression profile are shown on the right. d-f, Differentially expressed genes from pairwise Wald tests between stage-specific RNA-seq samples. The top 18 genes with lowest p-adjusted values and highest log fold change are labelled. Considering gene expression changes significant if the adjusted p-value < 0.05, we identified 8,341 differentially expressed genes (4,543 up and 3,798 down) for ‘late embryo vs early embryo’, 1,870 genes (938 up and 932 down) for ‘juvenile vs late embryo’; and 3,746 genes (1,827 up and 1,919 down) for ‘adult vs juvenile’.
Extended Data Fig. 9 | CTCF-binding motifs are the most abundant in open chromatin regions in *Dimorphophilus gyrociliatus*. a, Insert size distribution of ATAC-seq samples in *D. gyrociliatus*. b, c, Averaged ATAC-seq read depth around transcription start sites (TSS) and transcription termination sites (TTS). d, Heatmaps of ATAC-seq read coverage around TSS (left) and TTS (right) of each annotated gene. e, Averaged location of CTCF motifs in ATAC-seq peaks. f, Aggregate ATAC-seq read coverage centred around CTCF motifs. g, Number of CTCF motifs according to genomic feature. Most CTCF-binding motifs in open chromatin regions (that is ‘active’) are in promoters. h, Genome browser snapshot showing the distribution of CTCF-binding motifs in the Hox cluster of *D. gyrociliatus* as example of the general pattern observed genome wide. Most often, there is only one CTCF motif in an open chromatin region, and there is no clear directional arrangement between consecutive or neighbouring active CTCF-binding sites.
Extended Data Fig. 10 | General features and comparative aspects of CAGE-seq derived promoters in *Dimorphilus gyrociliatus*. 

**a**, Pearson’s correlation at the CAGE-supported transcription start site (CTSS) level between CAGE-seq biological replicates (left panel) and Spearman correlation between gene-counts derived from RNA- or CAGE-seq (right panel, merged biological replicates). 

**b**, Distribution of number of tag clusters/promoters across scaffolds. 

**c**, Heatmap of tag-cluster coverage ordered by tag-cluster IQ-width from narrow (top) to broad (bottom) centred on the first nucleotide of 5’ UTRs determined by RNA-seq. 

**d**, Genomic locations of dominant CTSS. 

**e**, Dinucleotide composition of all CTSSs identified in *Dimorphilus* CAGE-seq libraries. 

**f**, Genomic locations of tag clusters identified to contain a TATA-box or downstream promoter element (DPE). 

**g**, Sequence patterns in CAGE-seq derived promoters in the appendicularian *O. dioica* (genome size ~70 Mb), the fly *D. melanogaster* (genome size ~140 Mb) and the lancelet *B. lanceolatum* (genome size ~550 Mb). All heatmaps are centred on dominant TSSs and ordered by the tag-cluster/promoter IQ-width from narrower (top) to broader (bottom). IQ-widths are shown as tag-cluster coverage in the same order as on the heatmaps (right, in grey or blue). Heatmaps (left to right) represent TA dinucleotide patterns, TATA-box or DPE density (promoter regions are scanned using a minimum of the 80th percentile match to the TATA-box or DPE position weight matrix (PWM)) or GC dinucleotide patterns. Relative signal metaplot is shown above each heatmap. Promoters are divided according to TATA-box or DPE content at –30 or +30 position relative to the dominant TSS, and a heatmap of TATA-box or DPE density across promoter categories is shown.
Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

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

- The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement
- A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- The statistical test(s) used AND whether they are one- or two-sided
- A description of all covariates tested
- A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- For null hypothesis testing, the test statistic (e.g. $F$, $t$, $r$) with confidence intervals, effect sizes, degrees of freedom and $P$ value noted

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection

No software was used to collect data for this study

Data analysis

We used the following open source software in this study: SMRTAnalysis (v.2.3.0.140936), PBcR (v.8.3rc2), Blobtools (v.0.9.16), Pilon (v.3.16), cutadapt (v.1.4.2), HaploMerger2 (v.20151124), SSPACE-LongRead (v.1.1), trimmomatic (v.0.35), SPAdes (v.3.8.2), Super_Deduper (v.2.0), Platanus (v.1.2.4), Quast (v.3.1), BWA-MEM (v.0.7.12-r1044), SAMtools (v.1.3.1), BUSCO (v2), BLAT (v.36x2), Isoblat (v.0.3), Picard tools (v.2.0.1), Jellyfish (v.2.2.3), Trinity (v.2.1.1), Bowtie2 (v.2.1.0), CD-HIT (v.4.6), CAP3 (v.02/10/15), Kallisto (v.0.44.0), DESeq2, Transdecoder (v.5.0.2), HMMER (v.2.3.2), MSAProbs (v.0.9.7), BGME, IQTREE, RepeatModeler (v.1.0.4), MITE Digger, MODELLER, CEGMA (v.2.4), AUGUSTUS (v.3.2.1), EXONERATE (v.2.2.0), GMAP (v.2015-12-31), EvidenceModeler (v.1.1.1), ORFik, BLAST (v.2.2.31+), Trinotate (v.3.0), signalP (v.4.1), tmHMM (v.2.0c), MAFFT (v.7.310), Malin, Orthofinder (v.2.2.7), FastTree (v.2), Clustal X, trimAl, MacVector (v.11.0.4), Adobe Photoshop CS6, Adobe Photoshop CC (v.14.0), Adobe Illustrator CC (v.17.0.0), MACS2 (v.2.1.1.201610309), IDRCode, HOMER (v.2), CAGER (v.1.20.0), KAT, GenomeScope2.0, BBTools

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

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:
- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

All new raw sequence data associated to this project are available under project with primary accession PRJEB37657 in the European Nucleotide Archive (ENA). Genome annotation files and additional datasets are available in https://github.com/ChemaMD/DimorphilusGenome.
Field-specific reporting

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

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

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

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

**Sample size**
Sample sizes for genomic, transcriptomic and CAGE-seq analyses were estimated based on the amount of genomic DNA and total RNA obtained per individual. For ATAC-seq analyses, sample size was estimated in order to obtain a final number of 50,000 nuclei for subsequent tagmentation.

**Data exclusions**
No data was excluded from the analyses.

**Replication**
All RNA-seq, ATAC-seq and CAGE-seq analyses were conducted in replicates.

**Randomization**
All animal collections were performed randomly.

**Blinding**
All animal collections were allocated blindly to any of the replicates of study.

Reporting for specific materials, systems and methods

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

| Materials & experimental systems | Methods |
|---------------------------------|---------|
| n/a                             | n/a     |
| ☒ Antibodies                    | ☒ ChiP-seq |
| ☒ Eukaryotic cell lines         | ☒ Flow cytometry |
| ☒ Palaeontology                 | ☒ MRI-based neuroimaging |
| ☒ Animals and other organisms   |         |
| ☒ Human research participants   |         |
| ☒ Clinical data                 |         |

Animals and other organisms

Policy information about studies involving animals: ARRIVE guidelines recommended for reporting animal research

**Laboratory animals**
This manuscript uses the laboratory strains of the annelid species Dimorphilus gyrociuliatus and Capitella teleta. For D. gyrociuliatus, we used adult females and stage-specific embryonic samples. For C. teleta, we studied larval stages.

**Wild animals**
This study does not involve wild animals.

**Field-collected samples**
This manuscript studies the annelid species Trilobodrilus axi, which was collected from the wild by the lab of Katrine Worsaae. Adult specimens were kept in filtered seawater (31 ppm) at 15 °C in the dark prior collection for genomic DNA extraction.

**Ethics oversight**
Work on annelid embryos are not subject of ethical approvals or restrictions.

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

Plots

Confirm that:
- The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).
- The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).
- All plots are contour plots with outliers or pseudocolor plots.
- A numerical value for number of cells or percentage (with statistics) is provided.

Methodology

Sample preparation
Hundreds of adult females of D. gyrociatus were collected and starved for 3–4 days before flow cytometry analysis. Worms were transferred into a petri dish, washed well in seawater to remove any contaminant, and finely chopped with a razor blade in 2 ml of General-Purpose Buffer to generate a suspension of nuclei. This suspension was filtered through a 30 μm nylon mesh and stained with propidium iodide (Sigma; 1 mg/mL) on ice.

Instrument
We used a flow cytometer Partec CyFlow Space fitted with a Cobalt Samba green laser (532nm, 100mW)

Software
We used the built-in instrument software FloMax

Cell population abundance
We used flow cytometry to estimate genome size using C. elegans as reference and thus we did not sort any cell populations. To estimate genome size from propidium iodide staining, we did three independent runs for each species analysing at least 1,000 nuclei per run.

Gating strategy
We considered all cell populations for genome size estimation, and thus no gating strategy was implemented.

☐ Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.