Conservative route to extreme genome compaction in a miniature annelid

José M. Martín-Durán[1,3,§], Bruno C. Vellutini[1,4,*], Ferdinand Marlétaz[5,6*], Viviana Cetrangolo[1], Nevena Cvetesic[7], Daniel Thiel[1,8], Simon Henriet[1], Xavier Grau-Bové[9], Allan M. Carrillo-Baltodano[3], Wenjia Gu[3], Alexandra Kerbl[10,11], Yamile Marquez[12], Nicolas Bekkouche[10], Daniel Chourrout[1], Jose Luis Gómez-Skarmeta[13], Manuel Irimia[12,14,15], Boris Lenhard[7], Katrine Worsaae[10,†], Andreas Hejnol[1,2,§,†]

[1]Sars International Centre for Marine Molecular Biology, University of Bergen. Thormøhlensgate 55, N-5006, Bergen, Norway.

[2]Department of Biological Sciences, University of Bergen. Thormøhlensgate 55, N-5006, Bergen, Norway.

[3]School of Biological and Chemical Sciences, Queen Mary University of London. London E1 4NS, United Kingdom.

[4]Max Planck Institute of Molecular Cell Biology and Genetics, Dresden, Germany.

[5]Molecular Genetics Unit, Okinawa Institute of Science and Technology, Graduate University. Onna, Okinawa 904-0495, Japan.

[6]Present address: Centre for Life's Origins and Evolution, Department of Genetics, Evolution and Environment, University College London, London, United Kingdom.

[7]Institute for Clinical Sciences, Faculty of Medicine, Imperial College London, and MRC London Institute of Medical Sciences. London W12 0NN, United Kingdom.

[8]Present address: Living Systems Institute, University of Exeter, Stocker Road, Exeter, EX4 4QD, United Kingdom.

[9]Department of Vector Biology, Liverpool School of Tropical Medicine. Pembroke Place, L3 5QA, Liverpool, United Kingdom.

[10]Department of Biology, University of Copenhagen. 2100 Copenhagen, Denmark.

[11]Present address: Centrum für Naturkunde, Universität Hamburg, Martin-Luther-King-Platz 3, 20146 Hamburg, Germany.

[12]Centre for Genomic Regulation, Barcelona Institute of Science and Technology (BIST). 88 Dr Aiguader, Barcelona 08003, Spain.

[13]Centro Andaluz de Biología del Desarrollo (CABD), Consejo Superior de Investigaciones Cientificas-Universidad Pablo de Olavide-Junta de Andalucía.

[14]Universitat Pompeu Fabra, Barcelona, Spain.

[15]ICREA, Barcelona, Spain.

*These authors contributed equally to this work

[§]Corresponding authors

[†]Senior authors

Summary

Animal genomes vary in size by orders of magnitude[1]. While genome size expansion relates to transposable element mobilisation[2-5] and polyploidisation[6-9], the causes and consequences of genome reduction are unclear[1]. This is because our understanding of genome compaction relies on animals with extreme lifestyles, such as parasites[10,11], and free-living animals with exceptionally high rates of evolution[12-15]. Here, we decode the extremely compact genome of the annelid *Dimorphilus gyrociliatus*, a morphologically miniature meiobenthic segmented worm[16]. With a ~68 Mb size, *Dimorphilus* genome is the second smallest ever decoded for a free-living animal. Yet, it retains many traits classically associated with larger and slower-evolving genomes, such as an ordered, intact Hox cluster, a generally conserved developmental toolkit, and traces of ancestral

bilaterian linkage. Unlike animals with small genomes, the analysis of *Dimorphilus* epigenome revealed canonical features of genome regulation, excluding the presence of operons and *trans*-splicing. Instead, the gene dense *Dimorphilus* genome presents divergent kynurenine and Myc pathways, key physiological regulators of growth, proliferation and genome stability in animal cells that can cause small body size when impaired[17-21]. Altogether, our results uncover a novel, conservative route to extreme genome compaction, suggesting a mechanistic relationship between genome size reduction and morphological miniaturisation in animals.

Animals, and eukaryotes generally, exhibit a striking range of genome sizes across species[1], seemingly uncorrelated with morphological complexity and gene content, which has been deemed the "C-value enigma"[22]. Animal genomes often increase in size mobilising their transposable element (TE) repertoire (e.g. in rotifers[2], chordates[3,4] and insects[5]) and through chromosome rearrangements and polyploidisation (e.g. in vertebrates and teleosts[6-8], and insects[9]), which is usually counterbalanced through TE removal[23], DNA deletions[24,25] and rediploidisation[26]. Although the adaptive impact of these changes is complex and probably often influenced by neutral nonadaptive population dynamics[27], genome expansions might also increase the evolvability of a lineage by providing new genetic material that can stimulate species radiation[6] and the evolution of new genome regulatory contexts[28] and gene architectures[29]. By contrast, the adaptive value of genome compaction is more debated and hypotheses are often based on correlative associations[1], e.g. with changes in metabolic[30] and developmental rates[31], cell sizes[1,32], and the evolution of radically new lifestyles (e.g. powered flight in birds and bats[25,33], and parasitism in nematodes[11] and orthonectids[10]). Besides, extreme genomic compaction leading to minimal genome sizes, as in some free-living species of nematodes[34], tardigrades[35] and appendicularians[4,36], co-occurs with

prominent changes in gene repertoire[37,38], genome architecture (e.g. loss of macrosynteny[12]) and genome regulation (e.g. *trans*-splicing and operons[13-15]). However, these divergent genomic features are also present in closely related species with larger genomes[4,39,40], and thus it is still unclear whether these are fundamental genomic mechanisms essential for drastic genomic streamlining or genomic specialisations unrelated to genome compaction.

The marine annelid *Dimorphilus gyrociliatus* (O. Schmidt, 1857) (formerly *Dinophilus gyrociliatus*) has a C-value (i.e. haploid genome size) of only 0.06–0.07 pg (~59–68 Mb)[41], the smallest ever reported for an annelid[42]. *Dimorphilus* is a miniature free-living meiobenthic species[16] whose adults exhibit a strong sexual dimorphism, evident already during oogenesis (Fig. 1a). The adult females are about 1mm long and display a typical albeit simplified annelid segmental body plan[43], with only six segments, reduced coelom, and no appendages, parapodia and chaetae (Supp. Note 1). *Dimorphilus* males are however only 50 µm long, comprise just a few hundreds of cells, lack a digestive system, but still possess highly specialised sensing and copulatory organs[44]. Despite their miniature size, *Dimorphilus* retains ancestral annelid traits, such as a molecularly regionalised nervous system in the female[45,46] and the typical quartet spiral cleavage[47]. Although there are only a handful of genomes sequenced (Supp. Table 1), annelids have retained ancestral spiralian and bilaterian genomic features[48]. Therefore, *Dimorphilus*, with its reduced genome size and small body, emerges as a unique system to investigate the genome architectural and regulatory changes associated with genome compaction and assess the interplay between genomic and morphological miniaturisation.

Genome sequencing, TE simplification and intron loss

We performed long-read PacBio sequencing (Extended Data Fig. 1a) to generate a highly contiguous (N50, 2.24Mb) and complete (95.8% BUSCO genes) ~78Mb-long haploid assembly, comparable in quality to other published annelid genomes (Extended Data Fig. 1b, c; Supp. Table 1). Flow cytometry measurements and k-mer based analyses estimated *Dimorphilus* genome size in ~72Mb and ~65Mb genome, respectively (Fig. 1c, d), concurring with previous estimations[41]. While their simple morphology originally prompted to consider them early-branching annelids[49] ("Archiannelida"), molecular phylogenies later placed *Dimorphilus* either within Sedentaria[50] or as sister to Errantia+Sedentaria[51], the two major annelid clades (Supp. Note 2). Gathering an extensive dataset of annelid sequences[52], we robustly placed *Dimorphilu*s together with *Trilobodrilus axi*, *Dinophilus vorticoides* and *Lobatocerebrum* sp. – all miniature annelids – in a clade we name Dinophiliformia that is sister to Errantia+Sedentaria, thus confirming previous hypotheses[51] (Fig. 1e; Extended Data Fig. 2). Given the larger bodies and genome sizes found in all other annelid lineages outside Dinophiliformia (Fig. 1e), and that *T. axi* also exhibits a ~87Mb compact genome (Fig. 1d), our data suggest a common event of genome compaction and morphological miniaturisation in the lineage leading to *Dimorphilus* and its relatives.

To assess how changes in repetitive content contributed to genome reduction in *Dimorphilus*, we annotated the landscape of transposable elements (TEs), uncovering a much lower percentage (4.87%) than in other annelid genomes (Fig. 2a; Extended Data Fig. 3a, b). Most TEs (91.5%) group in four classes and as in the annelid *Helobdella*[48], TEs are either old copies or very recent expansions (Fig. 2b). The most abundant TE class is a Ty3-*gypsy*-like LTR retrotransposon, whose *envelope* (*env*) protein shows similarities with that of animal Herpesviridae and appears to be an annelid- or *Dimorphilus*-specific subfamily (Extended Data Fig. 3c–e). Compared to species with minimal genome sizes, *Dimorphilus* TE load is 3–

4x lower than in the appendicularian *Oikopleura dioica* and the tardigrade *Ramazzottius varieornatus*, but around 4x larger than in insects with larger, but still compact genomes (~100 Mb) (Supp. Table 5). Therefore, TE depletion contributed to genome compaction in *Dimorphilus*, 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 employed transcriptomic data and *ab initio* predictions to annotate 14,203 protein coding genes in *Dimorphilus* genome, a smaller gene repertoire than that of other annelids (Fig. 2c; Extended Data Fig. 1b, c; Supp. Table 1). However, the gene number is comparable to free-living species with similar genome sizes, such as *O. dioica*[36] (~15,000 genes) and *R. varieornatus*[39] (~14,000 genes). With a gene density (208.86 genes/Mb) double than in the annelids *Capitella teleta* (99.96 genes/Mb) and *Helobdella robusta* (97.5 genes/Mb), *Dimorphilus* has shorter intergenic regions and transcripts, but comparable exon lengths and even larger UTRs (Extended Data Fig. 4a, b, d–f), suggesting that intron shortening might have contributed to genome compaction. However, although *Dimorphilus* shows overall very short introns and has adapted its splicing for short intron sizes (Extended Data Fig. 4i), introns are not shorter on average than in *Capitella* and even the centipede *Strigamia maritima* (Fig. 2d; Extended Data Fig. 4h), both with larger genomes than *Dimorphilus*. Instead, *Dimorphilus* has fewer introns than other annelids (Fig. 2e) and exhibits an intron density and ancestral intron retention comparable to other animals with small genome sizes, such as *O. dioica*, *C. elegans*, and *R. varieornatus* (Extended Data Fig. 4j, k). Therefore, gene and intron loss, rather than short intron size – which was likely a pre-existing condition – correlates with genome compaction in *Dimorphilus*.

Gene losses, Hox cluster and macrosynteny

To investigate how gene loss shaped *Dimorphilus* genome and morphology, we first reconstructed gene families and estimated gene loss and gain rates. Over 80% of *Dimorphilus* genes are assigned to shared gene families, the highest in any annelid sequenced so far (Extended Data Fig. 5a). In 38.9% of the retained gene families, there is only one *Dimorphilus* gene, and *Dimorphilus* has the smallest average gene family size among annelids (1.63 genes/orthogroup; Supp. Table 7). Although the rate of gene family loss is greater than in *Capitella*, which is the most conservative sequenced annelid genome[48], gene loss in *Dimorphilus* is similar to *Helobdella* and *Hydroides elegans*, annelids with larger genomes (Fig. 2f; Extended Data Fig. 5b). Therefore, our data suggests that gene repertoire streamlining outweighs complete gene family loss, and thus likely underpins the reduced total gene number of *Dimorphilus*.

Consistent with the streamlining of its gene repertoire, we detected only nine expanded gene families in *Dimorphilus* (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). Besides, *Dimorphilus* shows canonical repertoires of gene families expanded in other annelids, such as G protein-coupled receptors (GPCRs) and epithelial sodium channels (ENaCs)[48] (Extended Data Fig. 6a, b; Supp. Table 8). The GPCR complement is indeed dynamic and often linked to specific (neuro)physiological adaptations, as seen in lineages with miniature genomes that have experienced either losses (e.g. *O. dioica* lacks Class C, glutamate receptors) or expansions (e.g. *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

Dimorphilus nervous system is functionally equivalent to, though morphologically smaller than, those of larger annelids[45,46].

Despite its miniature body plan, *Dimorphilus* has an overall conserved developmental toolkit, both at the level of transcription factors and signalling pathways (Extended Data Fig. 5f, g). *Dimorphilus*, and Dinophiliformia generally, exhibit a limited repertoire of certain extracellular signalling molecules (e.g. Wnt and TGF-b ligands) and lacks *bona fide* FGF and VEGF ligands (Extended Data Fig. 5g, h). However, these simplifications do not affect the receptor repertoire (Extended Data Fig. 5j). Unlike appendicularians[53], tardigrades[39] and nematodes[39] with compact genomes, *Dimorphilus* exhibits a compact, ordered Hox cluster, only lacking *post1* (Fig. 3a; Extended Data Fig. 7a, b). In other annelids[54,55], *post1* is separate from the main Hox cluster, and as in brachiopods[56], it is expressed in chaetoblasts[55], supporting the homology of this novel cell-type[56]. Remarkably, the distantly related *Helobdella* and *Dimorphilus* 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. Stage-specific transcriptomics (Extended Data Fig. 8) and gene expression analyses during *Dimorphilus* life cycle indicate that all Hox genes but *Hox5* and *Antp* become simultaneously transcribed in early development, after gastrulation (Fig. 3a; Extended Data Fig. 7c). During body elongation and segment formation, Hox genes are expressed in staggered anteroposterior domains along the developing trunk, in patterns resembling those of *Capitella*[55], further supporting that *Dimorphilus* retains the ancestral annelid molecular body patterning (Fig. 3a; Extended Data Fig. 7d). Therefore, *Dimorphilus* Hox genes show only spatial collinearity (Extended Data Fig. 7e), providing a compelling case where temporal collinearity is not driving Hox cluster compaction and maintenance[57].

Animal groups with reduced genome sizes show altered gene orders, as exemplified by their disorganised Hox clusters[53,58] and the loss of conserved gene linkage blocks that represent the ancestral chromosomal organisation[12,48]. In *O. dioica*, this loss has been related to the loss of the classical Non-Homologous End Joining (c-NHEJ) double strand DNA break (DSB) repair pathway[59]. In addition to an ordered Hox cluster, *Dimorphilus* shows residual conservation of ancestral linkage blocks, which appear eroded but still visible (Fig 3b). These blocks are almost intact in *Capitella* but completely lost in *Helobdella* (Fig. 3b; Extended Data Fig. 7f). Moreover, *Dimorphilus* has a conserved DSB repertoire (Supp. Table 9), with the exception of BRCA1, which is however also absent in other invertebrates capable of homologous recombination, such as *D. melanogaster*[60]. Therefore, mutation prone DSB repair mechanisms that can increase DNA loss do not underpin extreme genomic compaction in *Dimorphilus*, which occurred without drastic genome architecture rearrangements.

We identified however gene losses that could impair cell growth and genomic stability, thus accounting for body miniaturisation and genome compaction. *Dimorphilus*, but not other members of Dinophiliformia, lacks the kynurenine pathway (Fig. 3c), which converts endogenous tryptophan into nicotinamide adenine dinucleotide (NAD^+), an essential cofactor in redox metabolism[17] involved together with sirtuins in DNA damage and cellular senescence[18]. Besides, the Myc transduction pathway, which regulates growth and proliferation[19], lacks the regulators *mad* (in *Dimorphilus*) and *mnt* (in all Dinophiliformia) (Fig. 3d). In Dinophiliformia, MYC additionally has a W135 point mutation in the broadly conserved MYC box II (MBII) transactivation domain that has been shown to impair MYC function in human cells, in particular its ability to repress growth arrest genes[61] (Fig. 3e). Myc downregulation in vertebrates and flies causes hypoplasia[20], which could explain the miniature size of dinophiliformids, and slows down DNA replication[21], which could act as a

selective pressure favouring smaller genomes. Additionally, low levels of NAD^+ in *Dimorphilus* might induce further genomic instability[18] (hence its smaller genome size), premature senescence[17,18] (*Dimorphilus* has the fastest life cycle in Dinophiliformia), and further impaired growth[17] (*Dimorphilus* is the smallest and only dinophiliformid with dwarf non-feeding males). Although the full extent of these genomic changes is hard to evaluate given the poor understanding of cell and organ growth in annelids, our data suggest a link between gene loss, genome compaction and morphological miniaturisation in *Dimorphilus*, previously unrecognised in any other free-living animal with a miniature genome.

The regulatory genome

To investigate how compaction affected genome regulation, we first used ATAC-seq to identify ~10,000 reproducible open chromatin regions in adult *Dimorphilus* females (Extended Data Fig. 9a–d). Open chromatin regions are short in *Dimorphilus* and mostly found in promoters (Fig. 4a, b), consistent with its small genome size and small intergenic regions. Despite the generally short intron size in *Dimorphilus*, 944 ATAC-seq peaks were in intronic regions significantly larger than non-regulatory introns (Fig. 4c). We recovered a canonical metagene regulatory profile (Fig. 4d), which together with the lack of putative splice leaders in 5' UTRs (Extended Data Fig. 4g), suggest that *trans*-splicing and operons do not occur in *Dimorphilus*, similar to other annelids[62]. The CTCF DNA-binding motif was the most abundant in active regulatory regions, located mostly in promoters and as single motifs (Fig. 4e; Extended Data Fig. 9e–h), supporting that, unlike nematodes with compact genomes[63], CTCF plays a major regulatory role in *Dimorphilus* genome, as observed in vertebrates[64]. Thus, our data indicates that *Dimorphilus* has retained a conserved genomic regulatory profile, streamlining regulatory regions, and potentially losing distal intergenic *cis*-regulatory elements with genome compaction.

Since most regulatory information is restricted to promoter regions (<1kb upstream the transcription start site, TSS), we applied CAGE-seq to characterise promoter architecture (Extended Data Fig. 10a). Promoters are narrow (<150bp) in *Dimorphilus* and use pyrimidine-purine dinucleotides as preferred initiators (Fig. 4 f, g; Extended Data Fig. 10e). Upstream TA and downstream GC enrichment, respectively, revealed the presence of TATA-box and downstream promoter elements (DPE) in *Dimorphilus*, with TATA-box generally associated with narrow promoters (Fig. 4h; Extended Data Fig. 10f). Similar to vertebrates[65], strength of nucleosome positioning correlates with promoter broadness in *Dimorphilus* (Fig. 4i) and thus narrow TATA-box dependent promoters have lower +1 nucleosome occupancy than wide non-TATA-box promoters (Fig. 4j). As in other eukaryotes, TATA-box containing *Dimorphilus* promoters have somewhat higher expression levels, while promoters with DPE motif have no particular features, indicating this element might be non-functional (Fig. 4k). Therefore, the general *Dimorphilus* promoter architecture resembles that of other bilaterians (Extended Data Fig. 10f), further supporting that extreme genomic compaction did not alter genome regulation.

Discussion

Our study demonstrates that extreme genome compaction and morphological miniaturisation are specificities of *Dimorphilus gyrociliatus* (Fig. 1e), grounded in a derived phylogenomic position within Annelida, TE depletion, intergenic region shortening, intron loss and streamlining of the gene complement and genome regulatory landscape (Fig. 2a, e, f; Fig. 4a, f). Traditionally, morphological miniaturisation in *Dimorphilus* and Dinophiliformia has been considered a case of progenesis (underdevelopment)[43,50], yet the exact underlying mechanisms were unknown. As in other animal lineages[37,38,66], our data supports that

morphological change might be partially explained by gene loss in *Dimorphilus* (Fig. 4l), as we identified a reduced repertoire of extracellular signalling ligands, which could explain the limited elaboration of its segmental body, and the loss of developmental genes related to missing organs, such as chaetae (*post1* and FGF ligand) and mesodermal derivatives like coeloms (VEGF ligand). However, *cis*-regulation of gene expression is mostly restricted to the proximal regions in *Dimorphilus* (Fig. 4b), suggesting that coordinated long-distance gene regulation is limited. Therefore, simplification or dismantlement of the gene regulatory networks required to develop adult characters might also underpin morphological miniaturisation in *Dimorphilus*, and cause progenesis generally.

Unlike in other cases of extreme genomic compaction[4,12-15,34-36], our work provides compelling evidence that genome miniaturisation did not trigger drastic changes in genome architecture and regulation in *Dimorphilus* (Fig. 3b; Fig. 4c, e, h, m). Therefore, the genomic features observed in appendicularians, tardigrades and nematodes are lineage specificities that might have eventually facilitated genome compaction, but that are not always associated with genome size reduction, thus questioning the assumed causal link between fast evolving genomic traits and genome compaction. Instead, *Dimorphilus* exhibits losses and genetic modifications of metabolic (kynurenine pathway) and cellular (Myc) pathways involved in growth, proliferation, and senescence, and that when deregulated cause genomic instability, but also hypoplasia, dwarfism, and premature aging. As such, these genomic changes indicate the first known mechanistic link between genome size reduction and organism miniaturisation, suggesting Myc and sirtuins as potential candidates connecting genome size with metabolism[30], developmental rate[31], and body/cell size[1,32]. Altogether, our study characterises an alternative, more conservative route to extreme genome compaction, and

furthermore provides an exciting new system and genomic resources to investigate the evolutionary plasticity and function of core cellular mechanisms in animals.

Methods

Detailed methods are available in the Supplementary Information File.

Data availability

All new raw sequence data associated to this project is available under project with primary accession PRJEB37657 in the European Nucleotide Archive (ENA).

Code availability

All custom code used in this study is freely available upon request.

References

1 Gregory, T. R. in *The Evolution of the Genome* (ed T. Ryan Gregory) Ch. Chapter 1, 3-87 (Academic Press, 2005).

2 Blommaert, J., Riss, S., Hecox-Lea, B., Mark Welch, D. B. & Stelzer, C. P. Small, but surprisingly repetitive genomes: transposon expansion and not polyploidy has driven a doubling in genome size in a metazoan species complex. *BMC Genomics* **20**, 466 (2019).

3 Sun, C. *et al.* LTR retrotransposons contribute to genomic gigantism in plethodontid salamanders. *Genome Biol Evol* **4**, 168-183 (2012).

4 Naville, M. *et al.* Massive Changes of Genome Size Driven by Expansions of Non-autonomous Transposable Elements. *Curr Biol* **29**, 1161-1168 e1166 (2019).

5 Talla, 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).

6 Ohno, S. *Evolution by gene duplication*. (Springer-Verlag, 1970).

7 Sacerdot, C., Louis, A., Bon, C., Berthelot, C. & Roest Crollius, H. Chromosome evolution at the origin of the ancestral vertebrate genome. *Genome Biol* **19**, 166 (2018).

8 Braasch, I. & Postlethwait, J. H. in *Polyploidy and Genome Evolution* (eds Pamela S. Soltis & Douglas E. Soltis) 341-383 (Springer Berlin Heidelberg, 2012).

9 Li, Z. *et al*. Multiple large-scale gene and genome duplications during the evolution of hexapods. *Proc Natl Acad Sci U S A* **115**, 4713-4718 (2018).

10 Slyusarev, G. S., Starunov, V. V., Bondarenko, A. S., Zorina, N. A. & Bondarenko, N. I. Extreme Genome and Nervous System Streamlining in the Invertebrate Parasite Intoshia variabili. *Current Biology* (2020).

11 Abad, P. *et al*. Genome sequence of the metazoan plant-parasitic nematode *Meloidogyne incognita*. *Nat Biotechnol* **26**, 909-915 (2008).

12 Denoeud, F. *et al*. Plasticity of animal genome architecture unmasked by rapid evolution of a pelagic tunicate. *Science* **330**, 1381-1385 (2010).

13 Ganot, P., Kallesoe, T., Reinhardt, R., Chourrout, D. & Thompson, E. M. Spliced-leader RNA trans splicing in a chordate, *Oikopleura dioica*, with a compact genome. *Mol Cell Biol* **24**, 7795-7805 (2004).

14 Guiliano, D. B. & Blaxter, M. L. Operon conservation and the evolution of *trans*-splicing in the phylum Nematoda. *PLoS Genet* **2**, e198 (2006).

15 Danks, G. B. *et al. Trans*-splicing and operons in metazoans: translational control in maternally regulated development and recovery from growth arrest. *Mol Biol Evol* **32**, 585-599 (2015).

16 Worsaae, K., Kerbl, A., Vang, A. & Gonzalez, B. C. Broad North Atlantic distribution of a meiobenthic annelid - against all odds. *Sci Rep* **9**, 15497 (2019).

17 Ying, W. NAD+/NADH and NADP+/NADPH in cellular functions and cell death: regulation and biological consequences. *Antioxid Redox Signal* **10**, 179-206 (2008).

18 Fang, E. F. *et al.* Nuclear DNA damage signalling to mitochondria in ageing. *Nat Rev Mol Cell Biol* **17**, 308-321 (2016).

19 Grandori, C., Cowley, S. M., James, L. P. & Eisenman, R. N. The Myc/Max/Mad network and the transcriptional control of cell behavior. *Annu Rev Cell Dev Biol* **16**, 653-699 (2000).

20 Trumpp, A. *et al.* c-Myc regulates mammalian body size by controlling cell number but not cell size. *Nature* **414**, 768-773 (2001).

21 Dominguez-Sola, D. *et al.* Non-transcriptional control of DNA replication by c-Myc. *Nature* **448**, 445-451 (2007).

22 Gregory, T. R. Coincidence, coevolution, or causation? DNA content, cellsize, and the C-value enigma. *Biological Reviews* **76**, 65-101 (2001).

23 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).

24 Kapusta, A. & Suh, A. Evolution of bird genomes-a transposon's-eye view. *Ann N Y Acad Sci* **1389**, 164-185, doi:10.1111/nyas.13295 (2017).

25 Kapusta, A., Suh, A. & Feschotte, C. Dynamics of genome size evolution in birds and mammals. *Proc Natl Acad Sci U S A* **114**, E1460-E1469 (2017).

26 Robertson, F. M. *et al*. Lineage-specific rediploidization is a mechanism to explain time-lags between genome duplication and evolutionary diversification. *Genome Biol* **18**, 111 (2017).

27 Lynch, M. & Conery, J. S. The origins of genome complexity. *Science* **302**, 1401-1404 (2003).

28 Sundaram, V. & Wysocka, J. Transposable elements as a potent source of diverse *cis*-regulatory sequences in mammalian genomes. *Philos Trans R Soc Lond B Biol Sci* **375**, 20190347 (2020).

29 Feschotte, C. Transposable elements and the evolution of regulatory networks. *Nature Reviews Genetics* **9**, 397-405 (2008).

30 Kozlowski, J., Konarzewski, M. & Gawelczyk, A. T. Cell size as a link between noncoding DNA and metabolic rate scaling. *Proc Natl Acad Sci U S A* **100**, 14080-14085 (2003).

31 Pagel, M. & Johnstone, R. A. Variation across species in the size of the nuclear genome supports the junk-DNA explanation for the C-value paradox. *Proc Biol Sci* **249**, 119-124 (1992).

32 Cavalier-Smith, T. Economy, speed and size matter: evolutionary forces driving nuclear genome miniaturization and expansion. *Ann Bot* **95**, 147-175 (2005).

33 Wright, N. A., Gregory, T. R. & Witt, C. C. Metabolic 'engines' of flight drive genome size reduction in birds. *Proc Biol Sci* **281**, 20132780 (2014).

34 Consortium, C. e. S. Genome sequence of the nematode *C. elegans*: a platform for investigating biology. *Science* **282**, 2012-2018 (1998).

35 Hashimoto, T. *et al.* Extremotolerant tardigrade genome and improved radiotolerance of human cultured cells by tardigrade-unique protein. *Nat Commun* **7**, 12808 (2016).

36 Seo, H. C. *et al.* Miniature genome in the marine chordate *Oikopleura dioica*. *Science* **294**, 2506 (2001).

37 Fernández, R. & Gabaldón, T. Gene gain and loss across the metazoan tree of life. *Nature Ecology & Evolution* (2020).

38 Guijarro-Clarke, C., Holland, P. W. H. & Paps, J. Widespread patterns of gene loss in the evolution of the animal kingdom. *Nature Ecology & Evolution* (2020).

39 Yoshida, Y. *et al.* Comparative genomics of the tardigrades *Hypsibius dujardini* and *Ramazzottius varieornatus*. *PLoS Biol* **15**, e2002266 (2017).

40 Dieterich, C. *et al.* The *Pristionchus pacificus* genome provides a unique perspective on nematode lifestyle and parasitism. *Nat Genet* **40**, 1193-1198 (2008).

41 Gambi, M. C., Ramella, L., Sella, G., Protto, P. & Aldieri, E. Variation in Genome Size in Benthic Polychaetes: Systematic and Ecological Relationships. *J. mar. biol. Ass. U.K.* **77**, 1045-1057 (1997).

42 Gregory, T. R. *et al.* Eukaryotic genome size databases. *Nucleic Acids Res* **35**, D332-338 (2007).

43 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).

44 Windoffer, R. & Westheide, W. The Nervous System of the Male *Dinophilus gyrociliatus* (Annelida: Polychaeta). I. Number, Types and Distribution Pattern of Sensory Cells. *Acta Zoologica* **69**, 55-64 (1988).

45 Kerbl, A., Martin-Duran, J. M., Worsaae, K. & Hejnol, A. Molecular regionalization in the compact brain of the meiofaunal annelid *Dinophilus gyrociliatus* (Dinophilidae). *Evodevo* **7**, 20 (2016).

46 Kerbl, A., Conzelmann, M., Jekely, G. & Worsaae, K. High diversity in neuropeptide immunoreactivity patterns among three closely related species of Dinophilidae (Annelida). *J Comp Neurol* **525**, 3596-3635 (2017).

47 Nelson, J. A. The Early Development of *Dinophilus*: A Study in Cell-Lineage. *Proceedings of the Academy of Natural Sciences of Philadelphia* **56**, 687-737 (1904).

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

49 Hermans, C. O. The Systematic Position of the Archiannelida. *Systematic Zoology* **18**, 85-102 (1969).

50 Struck, T. H. *et al.* The evolution of annelids reveals two adaptive routes to the interstitial realm. *Curr Biol* **25**, 1993-1999 (2015).

51 Andrade, S. C. *et al.* Articulating "Archiannelids": Phylogenomics and Annelid Relationships, with Emphasis on Meiofaunal Taxa. *Mol Biol Evol* **32**, 2860-2875 (2015).

52 Helm, C. *et al.* Convergent evolution of the ladder-like ventral nerve cord in Annelida. *Front Zool* **15**, 36 (2018).

53 Seo, H. C. *et al. Hox* cluster disintegration with persistent anteroposterior order of expression in *Oikopleura dioica*. *Nature* **431**, 67-71 (2004).

54 Hui, J. H. *et al.* Extensive chordate and annelid macrosynteny reveals ancestral homeobox gene organization. *Mol Biol Evol* **29**, 157-165 (2012).

55 Frobius, 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).

56 Schiemann, S. M. *et al.* Clustered brachiopod Hox genes are not expressed collinearly and are associated with lophotrochozoan novelties. *Proc Natl Acad Sci U S A* **114**, E1913-E1922 (2017).

57 Duboule, D. The rise and fall of Hox gene clusters. *Development* **134**, 2549-2560 (2007).

58 Smith, F. W. *et al.* The Compact Body Plan of Tardigrades Evolved by the Loss of a Large Body Region. *Curr Biol* **26**, 224-229 (2016).

59 Deng, W., Henriet, S. & Chourrout, D. Prevalence of Mutation-Prone Microhomology-Mediated End Joining in a Chordate Lacking the c-NHEJ DNA Repair Pathway. *Curr Biol* **28**, 3337-3341 e3334 (2018).

60 Sekelsky, J. DNA Repair in *Drosophila*: Mutagens, Models, and Missing Genes. *Genetics* **205**, 471-490 (2017).

61 Oster, S. K., Mao, D. Y., Kennedy, J. & Penn, L. Z. Functional analysis of the N-terminal domain of the Myc oncoprotein. *Oncogene* **22**, 1998-2010 (2003).

62 Hastings, K. E. SL trans-splicing: easy come or easy go? *Trends Genet* **21**, 240-247 (2005).

63 Heger, P., Marin, B. & Schierenberg, E. Loss of the insulator protein CTCF during nematode evolution. *BMC Mol Biol* **10**, 84 (2009).

64 Vietri Rudan, M. *et al.* Comparative Hi-C reveals that CTCF underlies evolution of chromosomal domain architecture. *Cell Rep* **10**, 1297-1309 (2015).

65 Haberle, V. *et al.* Two independent transcription initiation codes overlap on vertebrate core promoters. *Nature* **507**, 381-385 (2014).

66 Albalat, R. & Cañestro, C. Evolution by gene loss. *Nat Rev Genet* **17**, 379-391 (2016).

Acknowledgements

We thank all members of the Hejnol lab for their continuous support, as well as the staff at the Norwegian Sequencing Centre (NSC) and GeneCore (EMBL) for their technical support. We also thank Rafa Domínguez Acemel, Marta Silva Magri and Sandra Jimenez Gancedo for their help with establishing ATAC-seq in *D. gyrociliatus*, Juan Tena-Aguilar for his help with ATAC-seq bioinformatic analyses, and Marina Ezcurra for providing live *C. elegans* worms for flow cytometry analyses. Finally, we thank Alexandre de Mendoza for a critical read to this manuscript. 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).

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. and N.B. performed gene expression 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. A.M.C.-B. performed gene expression analyses on *C. teleta*. J.M.M.-D., S.H., D.T., D.C., M.I. and X.G.-B. performed computational analyses. All authors contributed to interpretation of the results, and J.M.M.-D., K.W. and A.H. wrote the manuscript.

Additional information

Supplementary information is available for this paper. Correspondence and request for materials should be addressed to chema-martin@qmul.ac.uk and andreas.hejnol@uib.no. Reprints and permissions information is available at www.nature.com/reprints.

Figures and Figure Legends

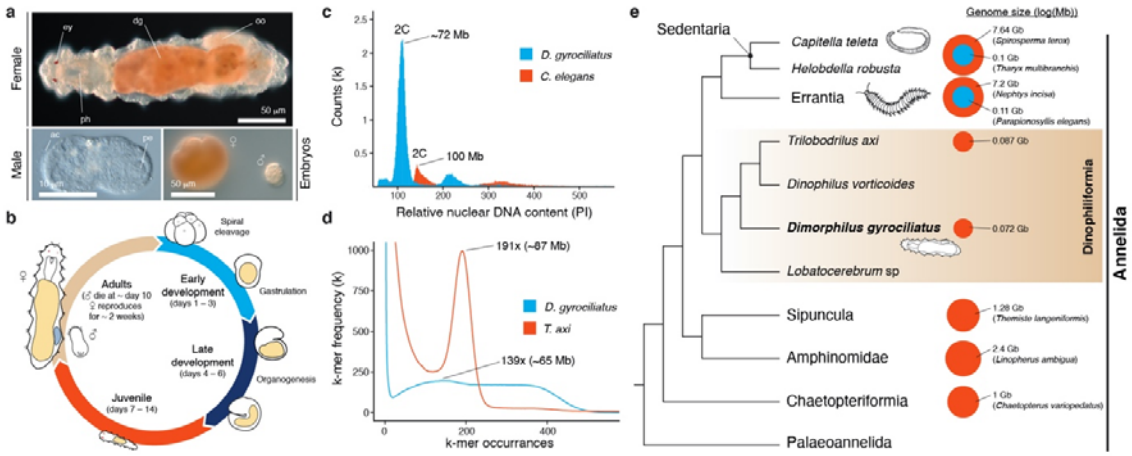


Figure 1 | *D. gyrociliatus* has the smallest annelid genome. (**a**) *D. gyrociliatus* adults are miniature annelid worms with an extreme sexual dimorphism, already apparent during early embryogenesis. (**b**) The life cycle of *D. gyrociliatus* comprises a 6-days long embryogenesis with a canonical early spiral cleavage programme, followed by a juvenile and an adult, reproductively active stage. (**c**) Flow cytometry analysis using the nematode *C. elegans* as reference and propidium iodide (PI) nuclear intensity estimates the genome size of *D. gyrociliatus* in ~72 Mb, while k-mer counts (**d**) estimates the genome size of *D. gyrociliatus* and *T. axi* to be 64.8 Mb and 87 Mb, respectively. (**e**) *D. gyrociliatus* and *T. axi* belong to Dinophiliformia, the sister group to Sedentaria and Errantia, and their genome sizes are the smallest known among annelids. ac, apical cells; dg, digestive system; ey, eye; oo, oocyte; pe, penis; ph, pharynx. Drawings are not to scale.

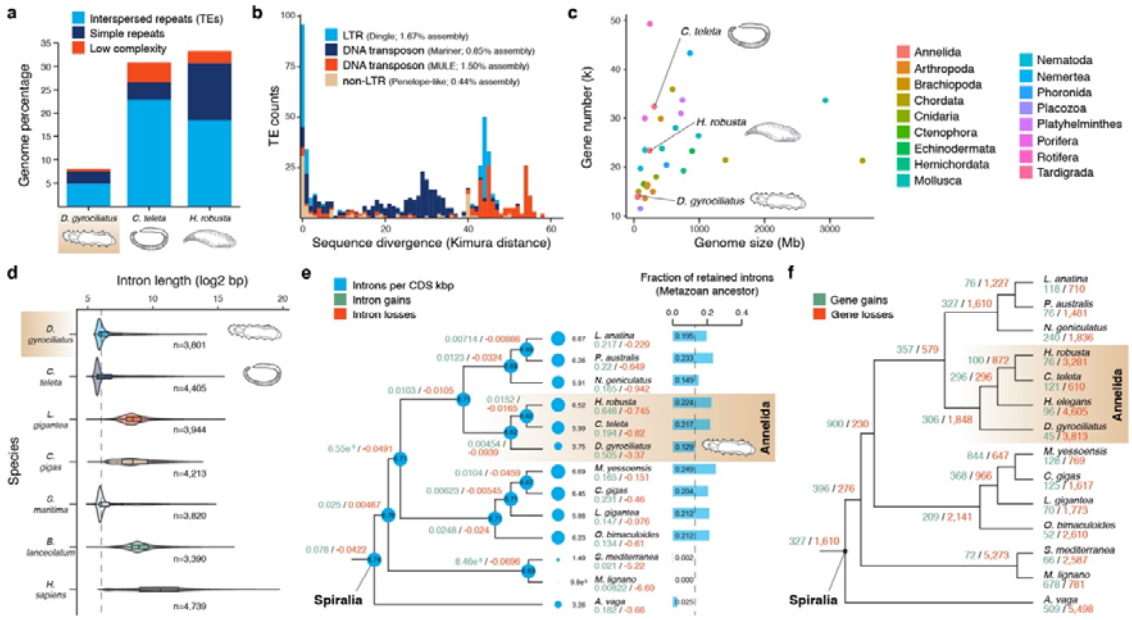


Figure 2 | *D. gyrociliatus* has a reduced TE landscape and gene repertoire. (**a**) *D. gyrociliatus* has considerably less TEs and simple repeats than other annelid genomes. (**b**) TE expansions are limited in size and correspond to either very recent bursts or old elements. (**c**) *D. gyrociliatus* has a reduced gene repertoire compared to other annelids, but comparable to other animals of similar genome size. (**d**) The size distribution of orthologous introns is comparable between *D. gyrociliatus* and the annelid *C. teleta* and the centipede *S. maritima*, which are both slow evolving lineages with larger genomes. However, *D. gyrociliatus* has lost introns (**e**), yet at a much lower rate and preserving many more ancestral animal introns than other fast evolving spiralian lineages, such as flatworms and rotifers. (**f**) Indeed, gene family loss in *D. gyrociliatus* is similar or lower than that observed in other fast evolving spiralian lineages.

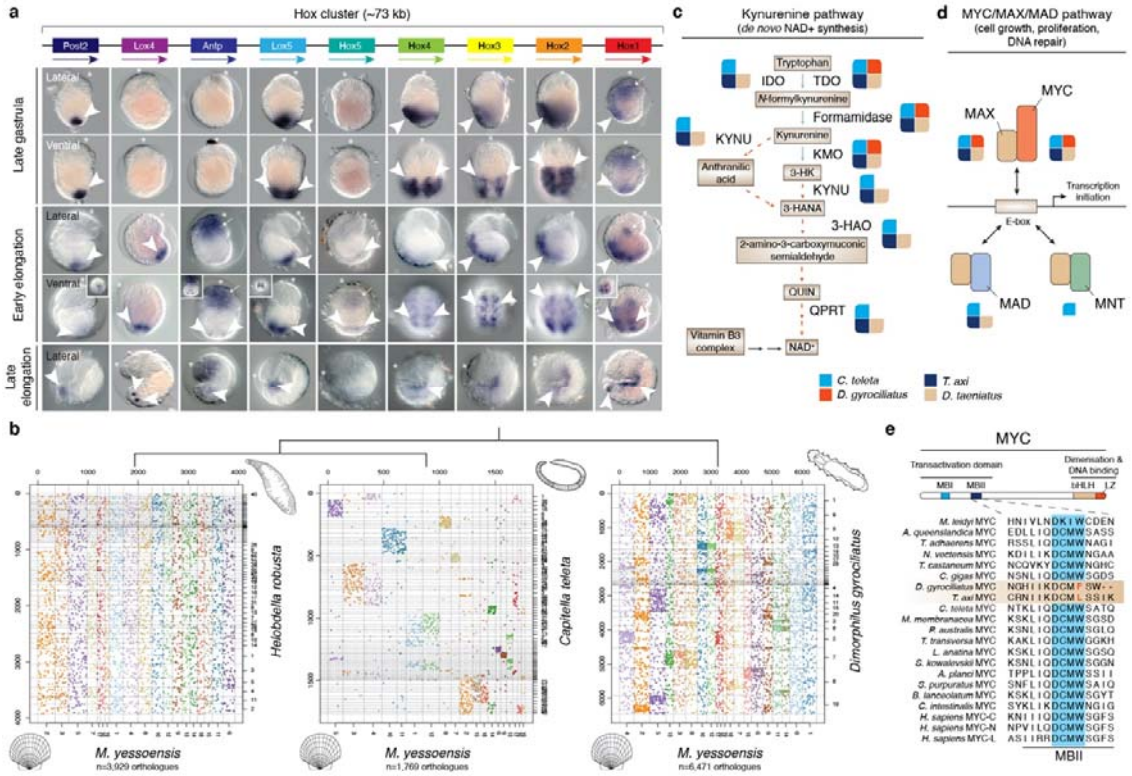


Figure 3 | *D. gyrociliatus* has conserved TF repertoires, but divergent or simplified signalling pathways. (**a**) *D. gyrociliatus* has a conserved Hox complement, organised in a compact cluster. Whole mount *in situ* hybridisation during embryogenesis reveals that Hox genes exhibit spatial, but not temporal collinear expression domains (arrowheads) along the trunk region, with *Hox1*, *Hox5* and *Antp* further exhibiting anterior head expression domains (arrows). (**b**) The presence of an organised Hox cluster correlates with the preservation of some macrosyntenic blocks in *D. gyrociliatus*, which are lost in the fast-evolving *H. robusta*. (**c**) *D. gyrociliatus* lacks the key enzymes of the kynurenine pathway to produce *de novo* NAD+ and has a divergent MYC/MAX/MAD pathway (**d**), with the absence of MAD and MNT (the latter also absent in other dinophiliformids), and a divergent MBII repressor domain in MYC (**e**).

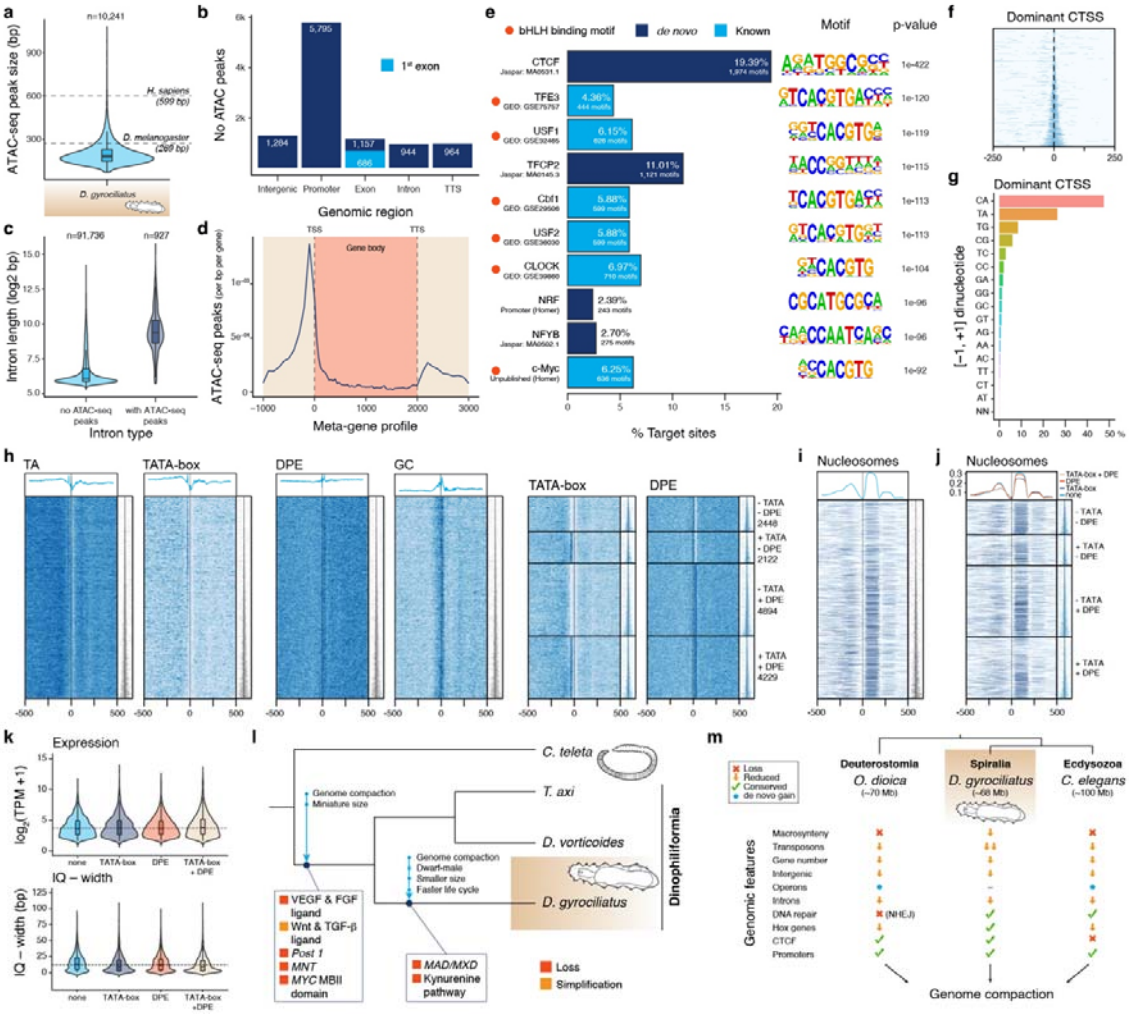
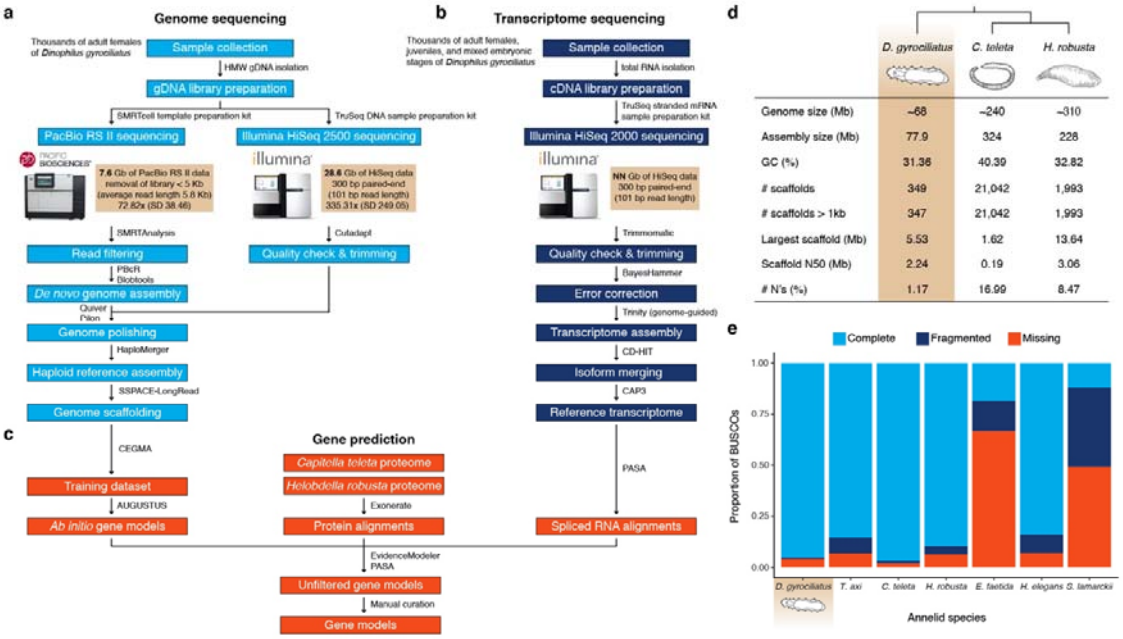
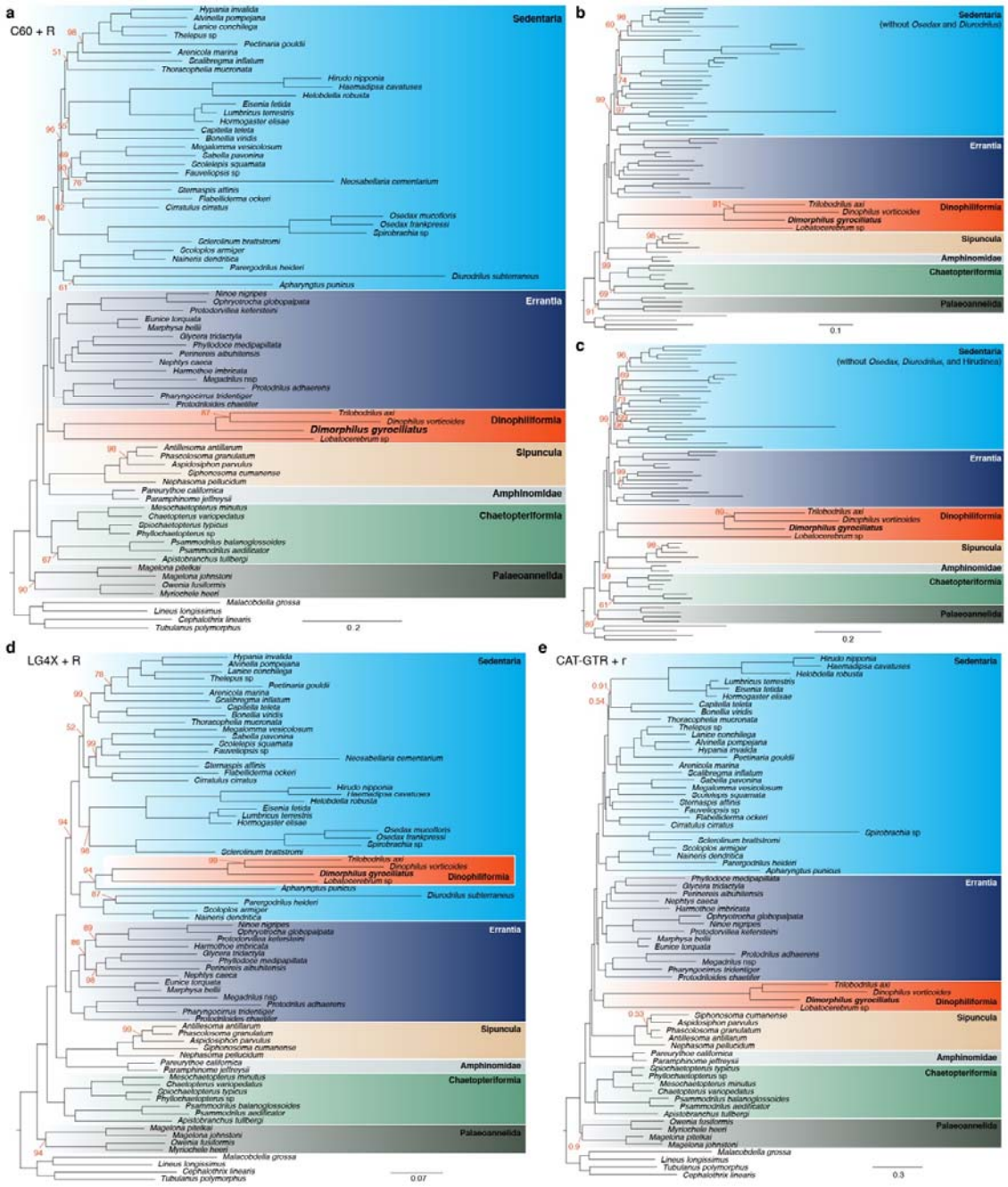


Figure 4 | The regulatory genomic landscape of *D. gyrociliatus*. (**a**) The open chromatin regions are shorter in *D. gyrociliatus* than in other animal genomes, but they exhibit a canonical distribution around gene bodies (**b**), with most of the open chromatin regions in promoters, intergenic regions and (first) introns (**c**). Importantly, the presence/absence of open chromatin regions in introns (**d**) correlates with the size of the intron. (**e**) The most abundant motif in open chromatin regions corresponds to CTCF. (**f**) Tag-clusters centred on the dominant CAGE-supported TSS (CTSS) are usually narrow (based on interquantile range q0.1-q0.9) and (**g**) retain the canonical metazoan polymerase II initiation pyrimidine (C, T)/purine (A, G) dinucleotides. (**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, regardless of the promoter type (**j**). (**k**) While genes with a TATA-box tend to be slightly narrower on average, there are not major differences in expression levels between genes with different promoter elements. (**l**) Schematic diagram of the genomic changes which occurred during genome compaction and morphological miniaturisation in *D. gyrociliatus* and Dinophiliformia. (**m**) *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 *O. dioica* and *C. elegans*.

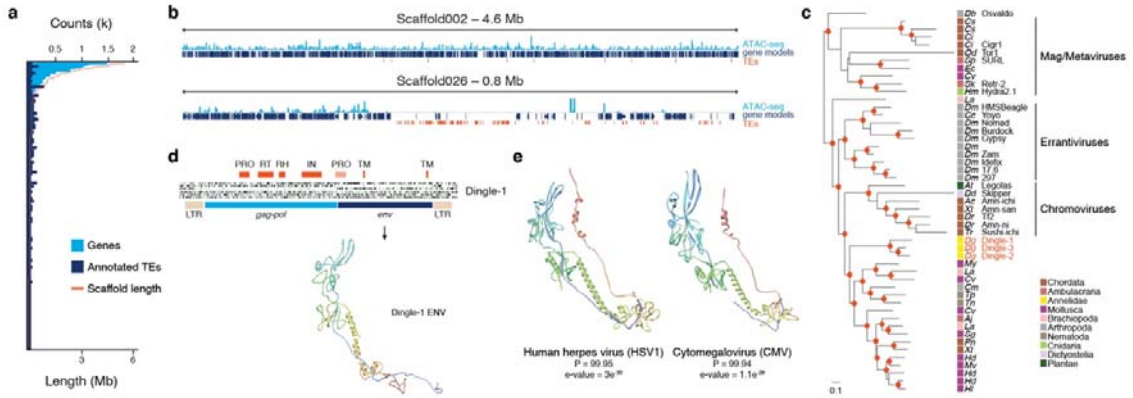


Extended Data Figure 1 | Sequencing approach and assembly statistics. (**a**–**c**) Diagram of the approach taken to sequence and assemble *D. gyrociliatus* genome and transcriptome, and to annotate coding genes in the genome. (**d**) Comparison of genome assembly statistics between *D. gyrociliatus* 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.

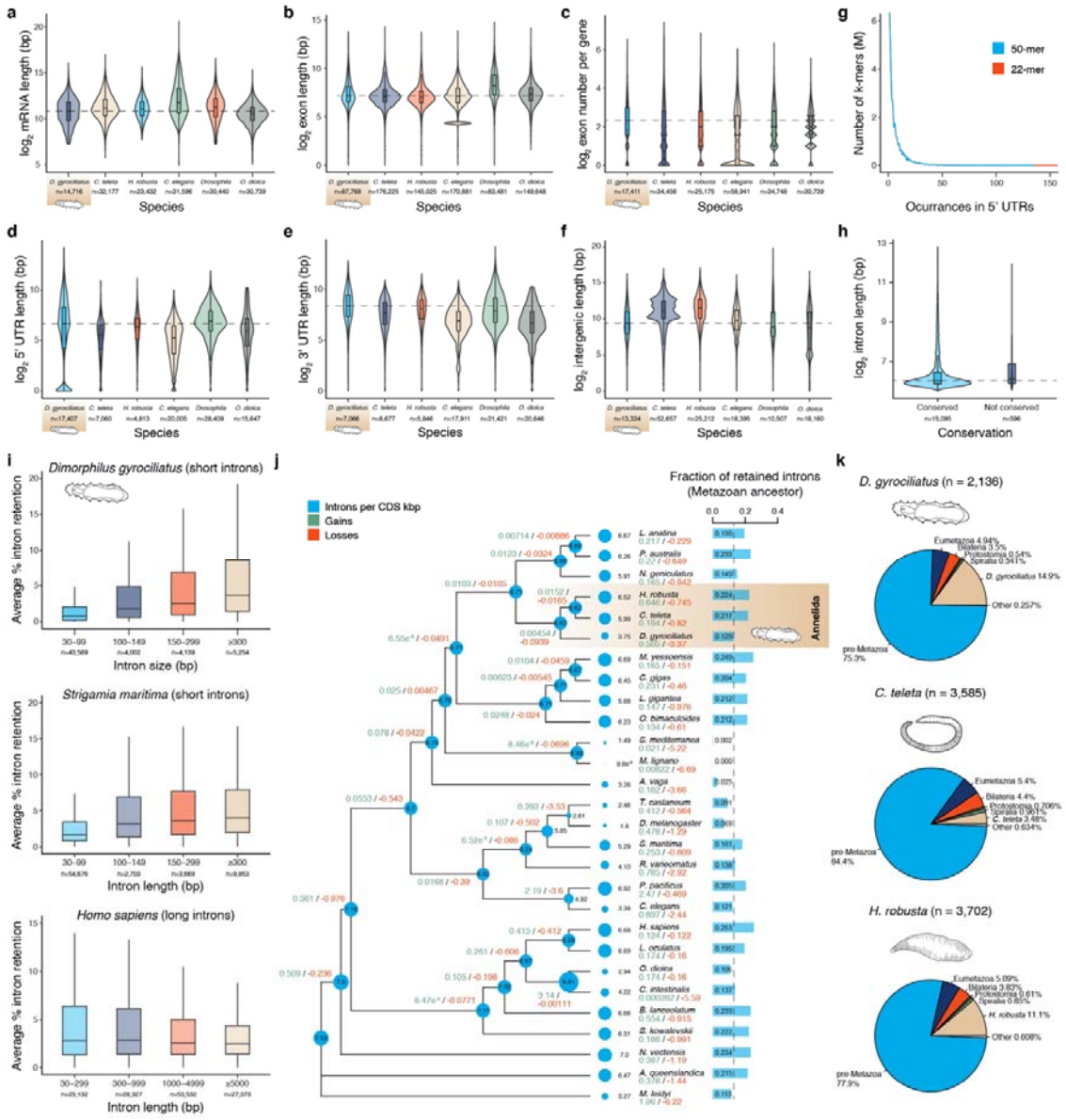


Extended Data Figure 2 | *D. gyrociliatus* phylogenetic position. (**a**–**c**) Maximum likelihood phylogenetic tree using the site heterogeneous model of protein evolution C60 + R and the entire annelid dataset (**a**), excluding the fast evolving *Osedax* and *Diurodrilus* lineages (**b**), and additionally excluding Hirudinea (**c**). In all cases, *D. gyrociliatus* 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 heterogenous model and the same dataset. In all trees, only values other than 100 bootstrap or 1 posterior probability are shown.

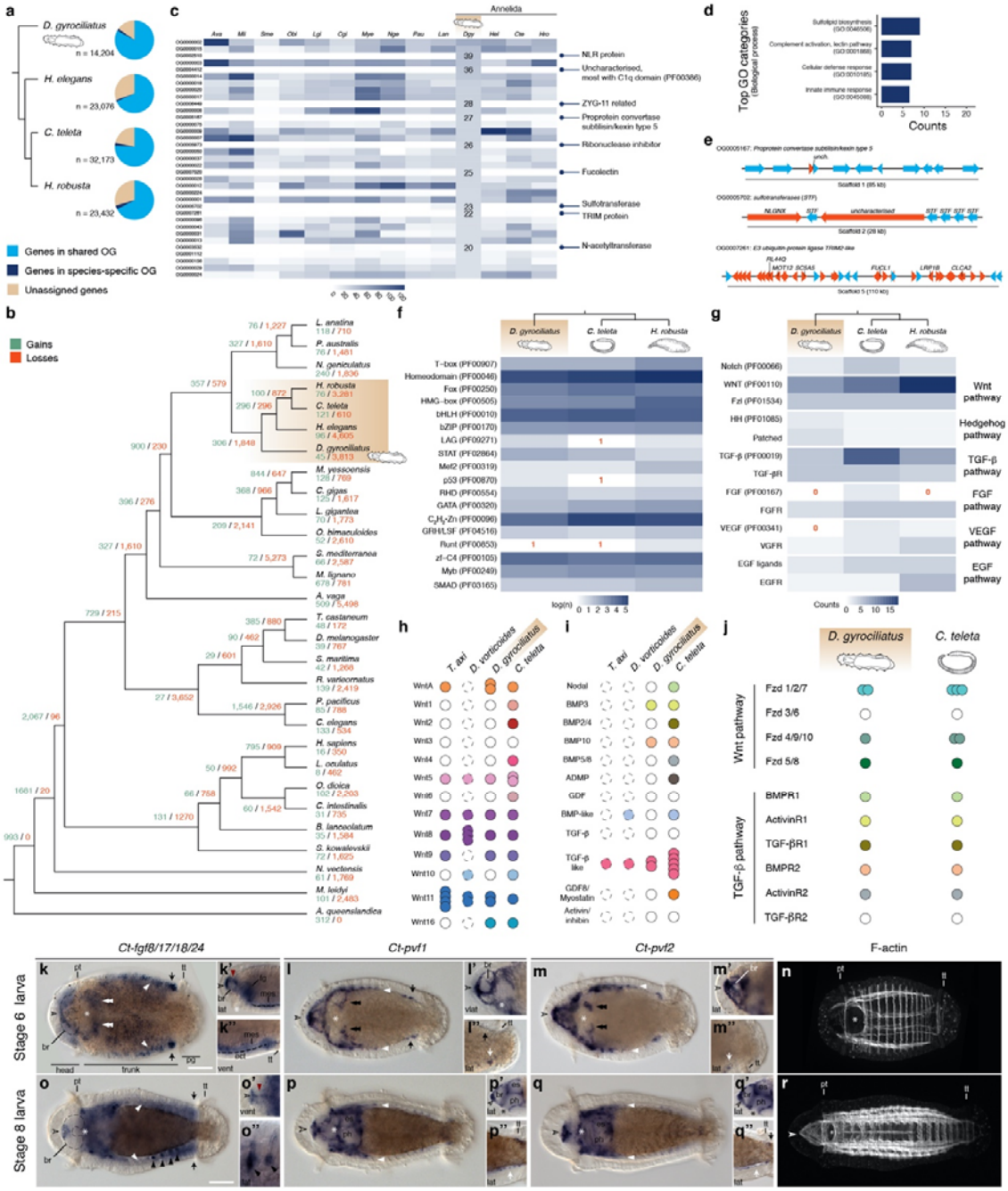


Extended Data Figure 3 | The transposable element repertoire of *D. gyrociliatus*. (**a**) Graph showing the number of genes and TEs per scaffold. (**b**) Diagram of Scaffold002 and Scaffold026 illustrating how transposable elements (TEs, in red) often concentrate in gene-free (dark blue boxes) and closed 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**). 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*.



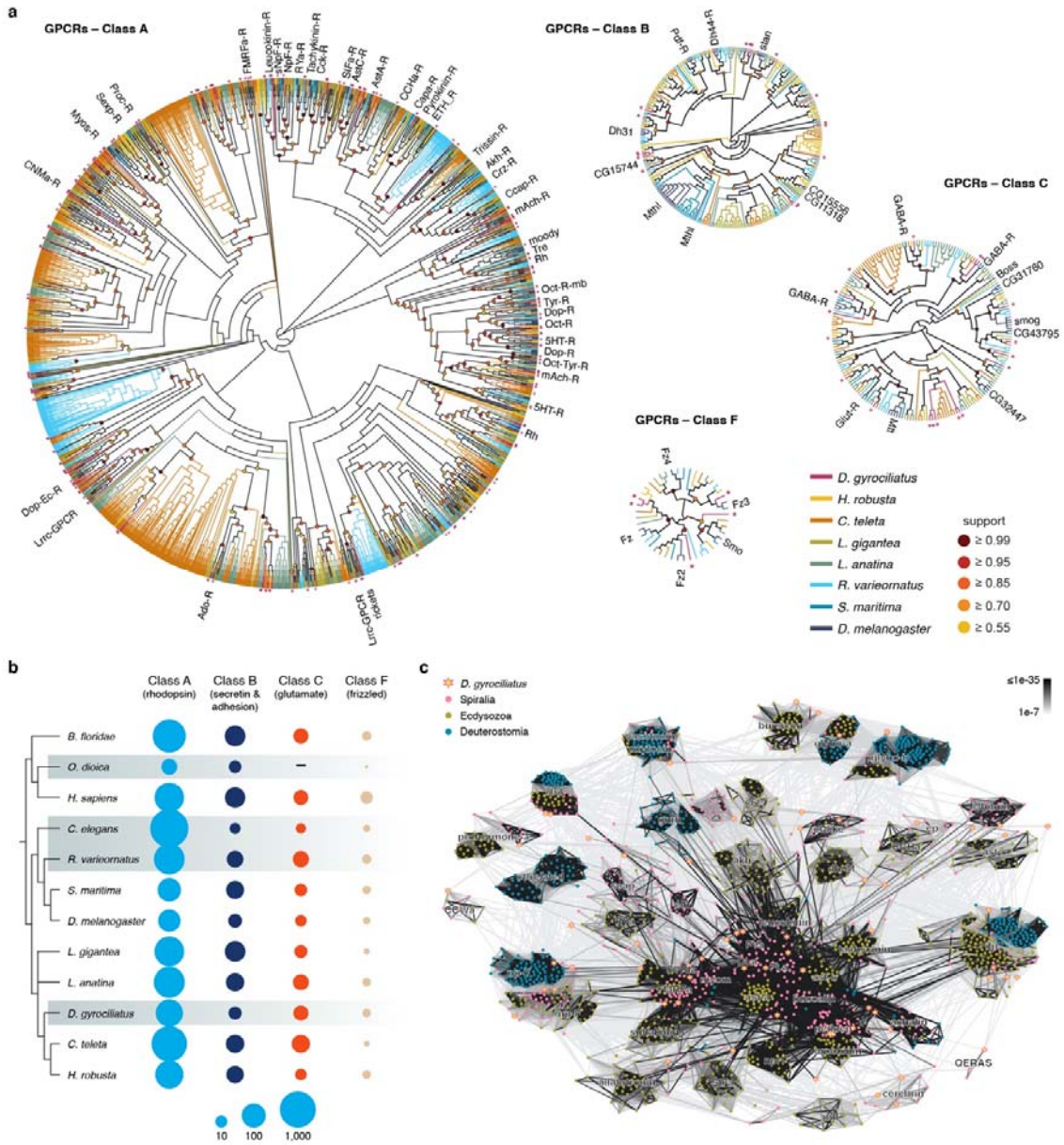
Extended Data Figure 4 | Comparative analyses of gene structure features in the *D. gyrociliatus* genome. (**a**–**f**) Violin plots showing the genome-wide distribution of mRNA and exon lengths, exon numbers per gene, and the lengths of 5' UTR, 3'UTR and intergenic regions in *D. gyrociliatus*, the 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 the 5' UTR regions of *D. gyrociliatus* does not indicate the presence of overrepresented sequences that could act as splice leaders. (**h**) The

distribution of intron sizes is consistent 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 comparable to other animal lineages with much larger genomes. (**k**) Inferred origin of the intron sites in *D. gyrociliatus* and the annelid *C. teleta* and *H. robusta*, expressed as the sum of gain probabilities at their respective ancestral nodes.



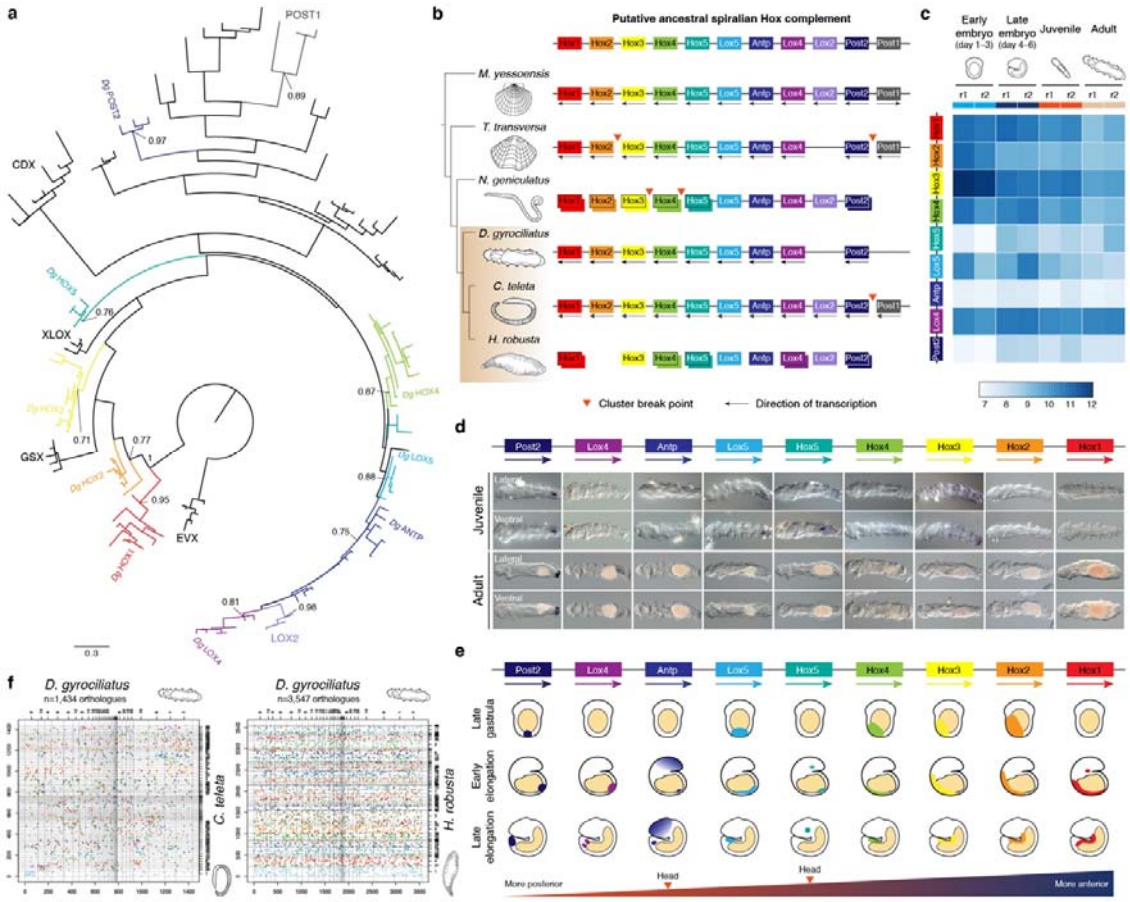
Extended Data Figure 5 | Expansions and gene losses in the genome of *D. gyrociliatus*. (**a**) Percentage of genes in shared orthogroups (OG; light blue) and species-specific (dark blue and light brown) in the four studied annelid species. (**b**) Metazoan-wide analysis of gene gain and loss, indicating 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 at the tips. (**c**) Heatmap of

the 35 largest OG in *D. gyrociliatus*, indicating those that correspond to lineage-specific expansions (OG size indicated in the cell, and its putative orthology). (**d**) The most abundant gene ontology (GO) terms of expanded families indicate that these are mostly involved in immunity. (**e**) Most of the expanded OGs are local copies (light blue). (**f**, **g**) Heatmaps depicting the repertoire of transcription factors (TFs) and ligands and receptors in *D. gyrociliatus* and the annelids *C. teleta* and *H. robusta*. *D. gyrociliatus* has representatives of all major families of TFs and in numbers comparable to other annelids. We could not find a clear ortholog of the FGF and VEGF ligand in *D. gyrociliatus*. Although *D. gyrociliatus* 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 dinophiliformids. (**j**) However, *D. gyrociliatus* has a conserved repertoire of frizzled and TGF-β receptors. (**k**–**m, o**–**q**) Differential interference contrast (DIC) micrographs of whole-mount *in situ* hybridisation of *Capitella teleta* larvae of the FGF (*Ct-fgf8/17/18/24*) and VEGF (*Ct-pvf1* and *Ct-pvf2*) ligands and phalloidin staining at these points (**n**, **r**). In early and late larval stages FGF and VEGF ligands are expressed mostly 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). Interestingly, we detected expression of FGF in well-developed and nascent chaetoblasts (black closed arrowheads). Br, brain; es, esophagus; fg, foregut; lat, lateral; pg, pygidium; ph, pharynx; pt, prototroch; tt, telotroch; vent, ventral; vlat, ventrolateral. Scale bars, 50 μm. Asterisks mark the stomodeum.



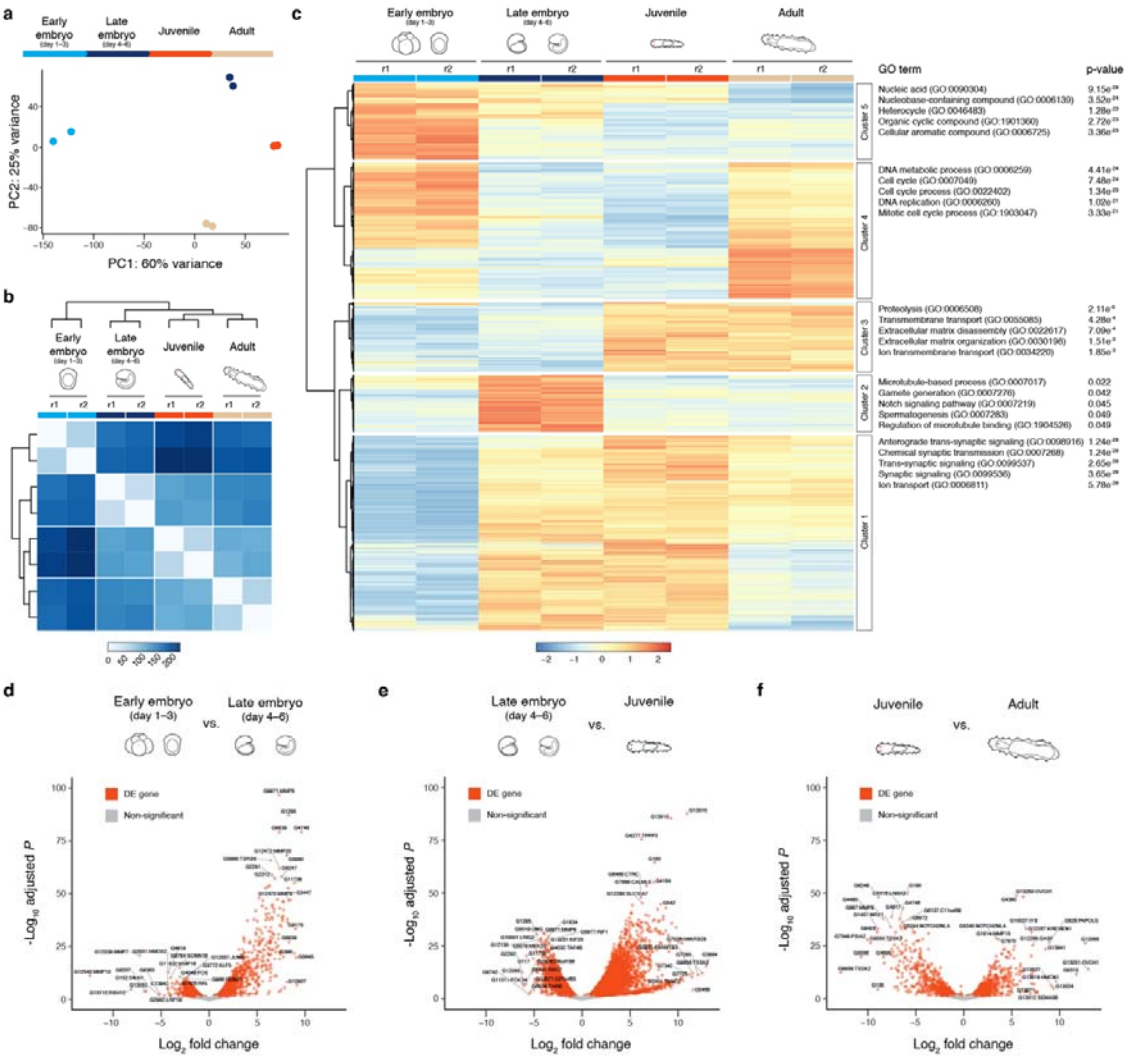
Extended Data Figure 6 | The GPCR and neuropeptide repertoire of *D. gyrociliatus*. (**a**) Orthology analyses of G-protein coupled receptors (GPCRs) for each class. The magenta asterisks highlight *D. gyrociliatus* receptors, and the annotations are given based on the *D. melanogaster* orthology. In (**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 miniaturised 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.



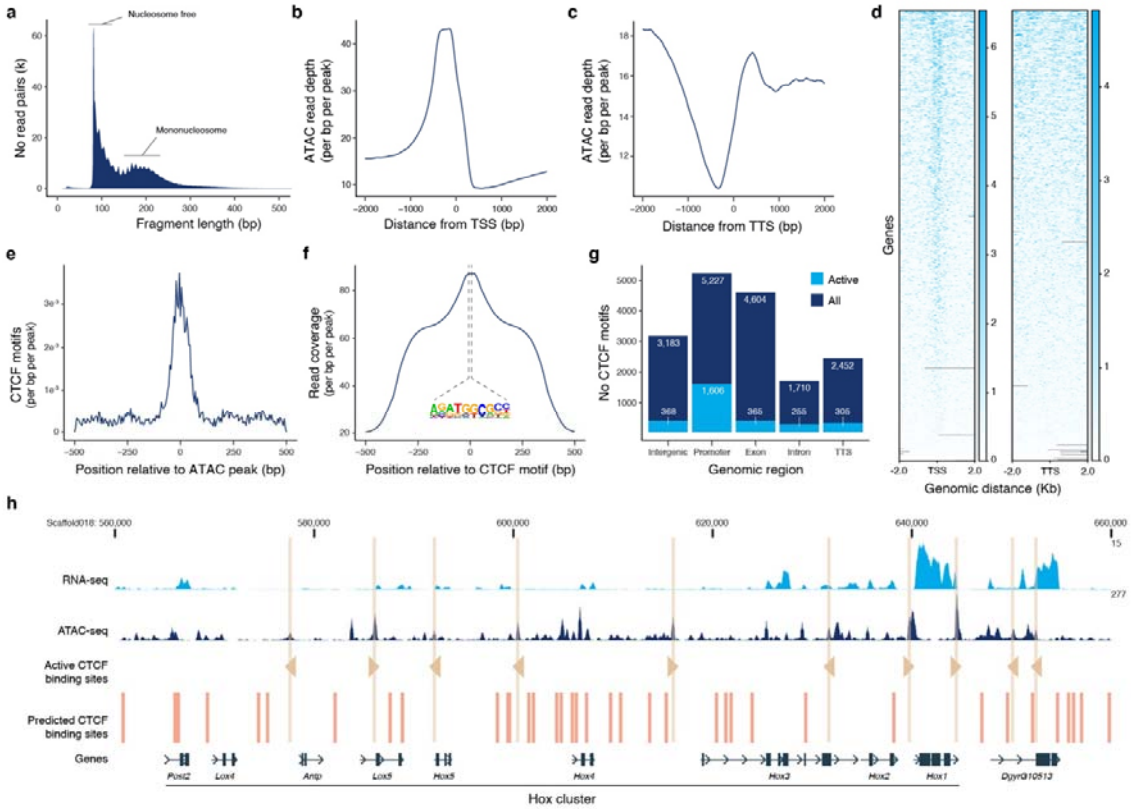
Extended Data Figure 7 | *Dimorphilus* has a conserved Hox cluster with spatial collinearity. (**a**) Maximum likelihood tree of Hox and ParaHox genes, with Evx proteins as outgroup, to assign orthology relationships of *Dimorphilus* Hox genes (indicated in the tree). Only bootstrap values for main orthogroups are shown. (**b**) Schematic representation of Hox complements and Hox genomic organisations 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 hybridisation 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 organisation 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**) Macrosyntenic relationships are little conserved between annelid worms, indicating lineage-independent large-scale genomic reorganisations.

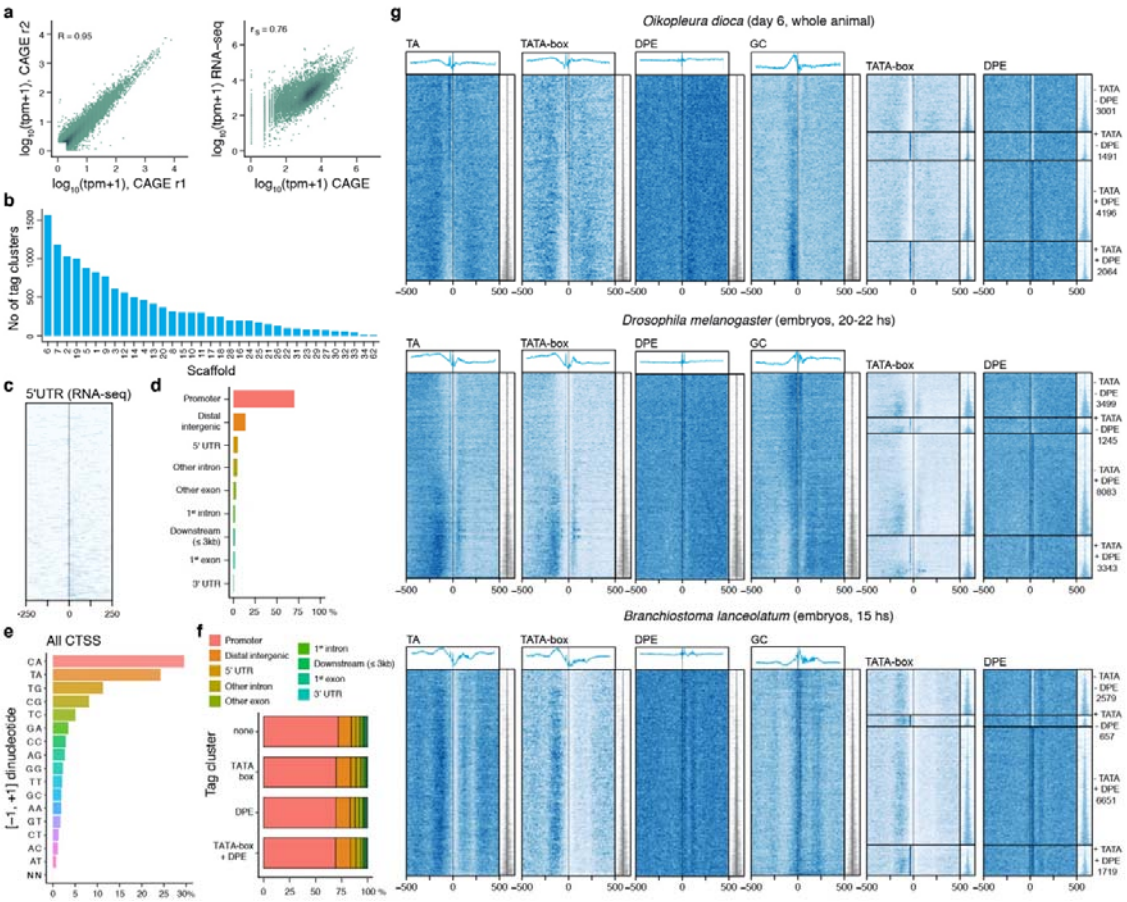


Extended Data Figure 8 | Differential expression analyses during the life cycle of *D. gyrociliatus*. (**a**) Principal component analysis of the stage-specific RNA-seq samples using the top eight thousand most-variable genes. The raw count data was transformed to homogenize the variance and normalized 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 genes with lowest p-adjusted values and highest log fold change are labelled.



Extended Data Figure 9 | CTCF-binding motifs are the most abundant in open chromatin regions. (**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**) The majority of CTCF binding motifs in open chromatin regions (i.e. "active") are in promoters, with more or less similar numbers of in intergenic, exonic, intronic and TTS regions. (**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 Figure 10 | General features and comparative aspects of CAGE-seq derived promoters in *D. 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 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. (**c**) Genomic locations of dominant CAGE-supported TSS. (**e**) Dinucleotide composition of all CTSSs identified in *D. gyrociliatus* CAGE-seq libraries. (**f**) Genomic locations of tag clusters identified to contain a TATA-box or 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.