Title: **A punctuated burst of massive genomic rearrangements and the origin of non-marine annelids**

Carlos Vargas-Chávez[1], Lisandra Benítez-Álvarez[1], Gemma I. Martínez-Redondo[1], Lucía Álvarez-González[2,3], Judit Salces-Ortiz[1], Klara Eleftheriadi[1], Nuria Escudero[1], Nadège Guiglielmoni[4], Jean-François Flot[5,6], Marta Novo[7], Aurora Ruiz-Herrera[2,3], Aoife McLysaght[8] & Rosa Fernández[1*]

[1] Metazoa Phylogenomics Lab, Institute of Evolutionary Biology (CSIC-UPF), Barcelona, Spain

[2] Departament de Biologia Cel·lular, Fisiologia i Immunologia, Universitat Autònoma de Barcelona, Cerdanyola del Vallès, Spain

[3] Genome Integrity and Instability Group, Institut de Biotecnologia i Biomedicina, Universitat Autònoma de Barcelona, Spain

[4] Universität zu Köln, Institut für Zoologie, Cologne, Germany

[5] Service Evolution Biologique et Ecologie, Université libre de Bruxelles (ULB), Brussels, Belgium

[6] (IB)² – Interuniversity Institute of Bioinformatics in Brussels, Brussels, Belgium

[7] Departamento de Biodiversidad, Ecología y Evolución, Universidad Complutense de Madrid, Spain

[8] Department of Genetics, Trinity College Dublin, Dublin 2, Ireland

* Corresponding author: rosa.fernandez@ibe.upf-csic.es

The genomic basis of cladogenesis and adaptive evolutionary change has intrigued biologists for decades. The unique insights from a genome-level perspective have revealed a striking pattern of conserved macrosynteny within chromosomes across huge evolutionary distances in animals, yet progress in many lineages has been hampered by the absence of genome-level data. Here, we show that the tectonics of genome evolution in clitellates, a clade composed of most freshwater and all terrestrial species of the phylum Annelida, is characterised by extensive genome-wide scrambling that resulted in a massive loss of macrosynteny between marine annelids and clitellates, to the point that ancient bilaterian linkage groups (ie, groups of genes inherited as a block in most bilaterian phyla) are fully disrupted. These massive rearrangements included the formation of putative neocentromeres with newly-acquired transposable elements, and preceded a further period of genome-wide reshaping events including whole-genome duplications and massive macrosyntenic reshuffling between clitellate lineages, potentially triggered by the loss of genes involved in genome stability or homeostasis of cell division. Notably, while these rearrangements broke short-range interactions observed between genes in marine annelids (including *Hox*), they were reformed as long-range interactions in clitellates. These genomic rearrangements gave rise to new genetic elements that may have facilitated the adaptation to freshwater and terrestrial environments in clitellates. Our findings provide evidence of a massive genomic reshaping through genome-wide chromosome scrambling in clitellates consistent with the punctuated equilibrium model, which sharply contrasts with the highly conserved chromosomal architecture observed in most animal taxa. Our study thus suggests that the genomic landscape of this animal lineage resulted from a rare burst of genomic changes that ended a long period of stability that persists across large evolutionary distances.

Main

Understanding the genomic basis of lineage origination and adaptation is key to uncovering how life diversifies and thrives in ever-changing ecosystems, providing invaluable insights into the mechanisms driving evolutionary success and the emergence of biodiversity. The understanding of how these processes unfold in most animal lineages has been curtailed by the lack of high-quality genomic resources, restricting comparative genomic studies to the investigation of

smaller-scale events. Genome-wide features and patterns, such as macrosynteny or genome architecture, have not been open to exploration.

The cornucopia of high-quality genomic resources generated by current initiatives aimed at bridging this gap is now revealing surprising information about animal evolution that can help us understand the origin of new lineages and how they adapt to changing environments, moving beyond previously known mechanisms such as point mutations, gene repertoire evolution (eg, gene gain, duplication and loss[1]) or whole-genome duplication. For instance, Nakatani and McLysaght[2], Nakatani et al.[3] and Simakov et al.[4] reported chromosome-scale conservation of gene linkages (ie, macrosynteny patterns, referred to as ancestral linkage groups, ALGs[4]) across distantly related animal phyla, revealing that many genes have remained on the same chromosome for hundreds of millions of years, and remain together in animal lineages as divergent as sponges, cnidarians, mollusks and cephalochordates. This phenomenon suggests a fundamental importance of genome organisation, perhaps for gene expression regulation or other regulatory functions.

In this context, the observation of substantial genome rearrangements at the macrosyntenic level elicits great interest. Recent studies in some animal phyla, including for instance vertebrates (eg[3,5,6]), lepidopterans[7], bryozoans[8], cephalopods[9] or tunicates[10], demonstrated the presence of extended rearrangements. Regarding invertebrates, in lepidopterans, Wright et al.[7] showed that macrosynteny (inferred from lepidopteran-specific linkage groups) has remained intact across lepidopteran lineages spanning 250 My of evolution, with the exception of some lineages in which extensive reorganisation through fusion and fission was observed. Similarly, in bryozoans Lewin et al.[8] recently found extensive macrosyntenic rearrangements in ALGs that are otherwise highly conserved across most animal phyla, revealing that in bryozoans they often underwent fusion or fission. Cephalopods[9] and tunicates[10] also show genomic rearrangements at the level of ALGs, but to a lesser extent compared to bryozoans for instance. Notably, even in such cases of extensive chromosomal rearrangement ALGs remain recognisable, suggesting that macrosynteny is an enduring and important feature of animal genomes.

While previous comparative studies in vertebrates have suggested some mechanisms of genome architecture remodelling[11–15], both the patterns and mechanisms of this phenomenon are still poorly understood in most invertebrate phyla. While most studies reporting genomic

rearrangements focus mostly on the pattern, little is yet known about the mechanisms driving these rearrangements. For instance, lepidopterans have holocentric chromosomes and thus rearrangements could be in principle more frequent and straightforward[16]. By contrast, karyotype studies of bryozoans have reported monocentric chromosomes[17], suggesting different mechanisms may be at play in different lineages. Likewise, the architectural and functional consequences of such massive genomic changes are poorly explored in most animal phyla. Understanding these mechanisms across invertebrate phyla and their consequences for genome architecture, gene expression and function, is therefore critical for understanding the evolutionary forces that shape animal genomes across the Metazoa Tree of Life.

Here, we report the complete loss of macrosynteny in lineages from the same phylum (Annelida), coinciding with the transition from marine to non-marine annelids (those included in the class Clitellata, comprising earthworms, leeches, potworms and their kin), to the point that ALGs are no longer recognizable. This massive genome rearrangement resulted in a complete restructuring of the genome that cannot be simply explained by typical rates of fusion and fission and rather results from genome-wide scrambling. These genomic tectonics affected intrachromosomal interactions that were rewired as interchromosomal ones after chromosome mixing. We found that neocentromeres in earthworms may have evolved through the co-option of newly-acquired transposable elements after scrambling, pointing to a fast centromere evolution in these clitellates. Furthermore, we identified some genetic elements that arose through this genome scrambling process that are putatively involved in adaptation to freshwater and terrestrial environments in leeches and earthworms, respectively, indicating that the massive scrambling may have been a catalyst or facilitator to the major habitat transitions that occurred around the same time. Our findings not only shed light on the remarkable genomic transformations at the within-phylum level in annelids but also suggest that clitellate evolution aligns with the punctuated equilibrium model[18–20], rather than resulting from stepwise gradual genomic changes.

A complete loss of macrosynteny between marine and non-marine annelids resulting in genome-wide chromosome scrambling

We examined the macrosynteny relationships of orthologous genes across several chromosome-level annelid genomes (two of them newly generated for this study). While the marine annelids displayed almost complete macrosynteny conservation with respect to the

ALGs (visible as ribbons of predominantly one colour per chromosome in Fig 1a), this relationship was shattered between marine and nonmarine annelids (the clitellates)(Fig. 1a,b). Although chromosome-level genomes are only available for earthworms and leeches within clitellates, we explored macrosynteny conservation with the draft genome of another clitellate lineage (a potworm or enchytraeid), confirming the loss of macrosynteny compared to marine annelids, and therefore indicating that the loss of macrosynteny occurred earlier than the divergence of earthworms and leeches (Fig. 1b). After this rampant genomic reorganisation event, macrosynteny in leeches and earthworms was characterised by additional massive rearrangements occurring between both groups, which are due to chromosome fusion and fission rather than genome-wide chromosome scrambling (Fig. 1c,d, Fig. 2b,c).

To test the extent of chromosome mixing resulting from these extremely unusual genomic tectonics, we investigated whether the scrambling of ALGs followed a random distribution. We defined clitellate-specific linkage groups (ClitLGs hereafter) as groups of genes whose presence on the same chromosome was conserved across leeches and earthworms. We inspected these to check if they were enriched in any ALGs, and found a complete lack of statistical enrichment, thus supporting a tectonic process characterised by random genome-wide fusion-with-mixing and indicating that ALGs and ClitLGs do not show any algebraic relationship (Fig. 1c,d; Supplementary Table 1). These observations support a scenario of chromosome scrambling randomly distributed across the genome, and disfavour typical models of chromosomal fusion and/or fission where parts of the constituent chromosomes are still recognizable, as occurs in most animal phyla where extended rearrangements have been reported[7,8,10]. In addition, the genomes were unalignable, ie, alignment-based ancestral genome reconstruction recovered a highly fragmented sequence for the most recent common ancestor of clitellates and marine annelids, reconstructing only 9.38 Mb scattered across 1139 contigs (Supplementary Table 2), which also supports the random genome-wide chromosome mixing.

To test whether the observed rearrangements could be the result of differential gene loss after whole genome duplication, we inferred the presence of such events through the analysis of synonymous substitution (Ks). Several rounds of whole genome duplication (WGD) were detected, with the most likely scenario pointing to two rounds on the branch leading to earthworms and one on the branch leading to Clitellata, with the latter only recovered by some analyses (Fig. 2a, Supplementary Data 1 and 2). We considered whether the observed genome structure could be explained by a process of differential gene loss following WGD, but the

genomes do not contain any of the expected relationships as described in Kellis et al.[21], briefly that (i) almost every region in the outgroups of clitellates would correspond to two sister regions in clitellates, (ii) each of the two regions in clitellates would have an ordered subset of the genes from the corresponding region in the outgroups, with each region having ca. half of the genes and together covering nearly all of them, and (iii) almost every region in the clitellates would correspond to a region in the outgroups, and thus could not be simply explained under this scenario.

Loss of genes involved in genome stability during chromosome scrambling and higher frequency of rare genome-reshaping events in leeches and earthworms

Clitellates encompass other lineages beyond earthworms, leeches and potworms, such as the families Naididae, Lumbriculidae and other early-splitting interstitial species[22], none of them is represented by reference or draft genomes so far, which hampers our understanding on the precise branch or branches where these genomic changes may have occurred. In the absence of genomic resources for these lineages to pinpoint with more precision in which branch of the Annelida Tree of Life these rearrangements may have occurred, our findings point to catastrophic genomic restructuring either on the branch leading to potworms, leeches and earthworms, or somewhere between the origin of clitellates and their split with marine annelids (ie, either at the origin of clitellates or somewhere in its surroundings in the Annelida Tree of Life, encompassing species that remain unsampled for genome sequencing). In any case, both scenarios are consistent with a model of punctuated equilibrium[18–20], in which a burst of genomic changes is observed in a short period after a long period of stasis (measured not in time units but as relative evolutionary distance between lineages).

Despite the lack of availability of genomes for several clitellate lineages, high-quality transcriptomes of all main lineages are available[22–24], which enables the exploration of gene repertoire evolution. Given the extensive genome-wide chromosome scrambling that we have observed, we expect the generation of gene fragments that are no longer functional[25] during the genome fragmentation process. Thus, we would expect an increase in gene loss in the branches where the rearrangements occurred. To test whether they originated at the dawn of all clitellates or are specific to the clade comprising enchytraeids, leeches and earthworms, we explored gene repertoire evolution across an extended clitellate dataset including all main lineages. Gene loss in the branch leading to Clitellata was ca. 25% higher than in the

surrounding branches (Fig. 2d), which is consistent with the scenario of a catastrophic event in the branch leading to clitellates as the origin of such massive rearrangements.

To test the potential functional consequences of the increased gene loss at the origin of clitellates, we investigated the putative functions of the lost genes in that node. Lost genes were largely enriched in functions related to cell division, DNA replication and DNA repair. Examples include the Slx1-Slx4 complex (an endonuclease complex involved in processing diverse DNA damage intermediates, including resolution of Holliday junctions (provoked by double strand breaks), collapse of stalled replication forks and removal of DNA flaps[26]), the GINS complex (a component of the eukaryotic DNA replication machinery required both for the initiation of chromosome replication and for the normal progression of DNA replication forks[27]), the MutSalpha complex (DNA mismatch repair relies on this complex for mismatch recognition and strand-specific nuclease recruitment to remove mispaired bases in daughter strands[28]), and the SHREC complex (that regulates nucleosome positioning to assemble higher-order chromatin structures critical for heterochromatin functions[29]) (Fig. 2e,f; Supplementary Data 3-6). We hypothesise that the loss of these genes could be the underlying cause of a high number of rare genome-reshaping events observed in leeches and earthworms. For instance, at least three more rounds of whole genome duplication were observed in earthworms (two more 'ancient' on the branch leading to earthworms and one more 'recent' in the family Megascolecidae, Fig. 2a,c; Supplementary Data 1 and 2), massive genomic rearrangements were observed between leeches and earthworms despite being relatively closely related (Fig. 2b), common fission and fusion of clitellate-specific linkage groups was detected in leeches (Fig. 2c) and high levels of gene loss, duplications and rearrangements were observed in *Hox* genes in both lineages, as discussed below (Fig. 2d,e,f). Our results therefore may suggest that the presence of relaxed selective constraints (in this case, less efficient DNA repair mechanisms) can facilitate the occurrence and fixation of genome reshuffling, as recently proposed in rodents[30]. Furthermore, these results are congruent with the observed high concentration of aneuploidy found in oligochaetes: Pavlíček et al[31] reported common aneuploidy and Robertsonian translocations in multiple species of clitellates (including earthworms, leeches and species from the families Naididade, Lumbriculidae and Branchiobdellidae). Since aneuploidy is strongly associated with chromosomal and genome rearrangements and is often recognised as a direct outcome of genome instability[32–34], our results may support the hypothesis of a single catastrophic event in the branch leading to clitellates resulting in the loss of genes associated to genome stability and chromatin reshaping, which may have resulted in common aneuploidy in clitellate.

Domestication of transposable elements absent in marine annelids into putative neocentromeres in earthworms

Given the extent and the apparent speed of the observed genome rearrangements, we considered the effect this might have had on centromeres, and more specifically whether these rearrangements might have necessitated the acquisition of new centromeres to ensure faithful chromosomal segregation during cell division. To investigate whether the composition of centromeres (which are known to be unique in all annelids including clitellates, ie, they are monocentric) changed after this period of massive rearrangement, we inferred the transposable element (TE) landscape across all annelids in our datasets, as well as that of satellite DNA, since both elements are the ones that constitute centromeres[35,36]. Among the more than a hundred TE superfamilies explored, only one was exclusively present in clitellates and absent in marine annelids and outgroups: the CMC-Chapaev-3 superfamily, belonging to DNA TEs (Supplementary Table 3). Notably these TEs also happened to be significantly enriched in the centromeric areas in earthworms (ie, with a significantly higher density (Fisher's exact test, p-value < 0.05) in 85% of the centromeric regions in all chromosomes, Fig. 3a; Supplementary Data 7). Regarding the inference of satellite DNA, we failed to identify a predominant motif coinciding with the putative centromeres, and in addition the identified ones did not coincide with the location of the centromere in clitellates while it did in marine annelids, which may be indicative as well of neocentromere formation[37] (Fig. 3b). Due to the monocentric nature of earthworm chromosomes[24], this could be considered as evidence of neocentromere formation, but further functional analyses would be needed to actually test this hypothesis. To understand where these TEs could have come from, we inferred a maximum likelihood phylogenetic tree with the sequence of the CMC-Chapaev-3 transposable element transposases, including as many species as possible as retrieved from the public databases (see Methods). In the case of earthworms, their transposases were frequently closely related to distant animal phyla such as arthropods, cnidarians or vertebrates but also viruses, indicating that they may have been acquired through horizontal gene transfer or via viral infection (Fig. 3c; Supplementary Data 8).

Chromosome scrambling resulted in a reshaping of the genome architecture and favoured intrachromosomal interactions

A detailed comparative analysis of the genome architecture in representative annelids species (a marine annelid, an earthworm and a leech) (Fig. 4a) revealed distinct patterns of genome-wide chromosomal interactions, both at the chromosomal level and at the sub-Megabase pair (sub-Mbp) scale. At the chromosomal level, annelids showed clustering of centromeres (Fig. 4b), mirroring previous observations in invertebrates from other phyla (i.e., mosquitos[38]) and marsupials[11]. At the sub-Mbp scale, annelids lacked clear A/B compartments and exhibited compartmentalised TADs defined by low insulator scores (Fig. 4c). Remarkably, annelids showed higher log2 interaction ratios per chromosome (Fig. 4d) when compared to model species (i.e, the fruit fly and chicken). These differences were also detected when analysing distance-dependent interaction frequencies represented as curves of contact probability as a function of genomic distance [P(s)] (Fig. 4e), suggesting that annelids are characterised by 'floppy' and relaxed chromosomes (i.e., lower intra-chromosomal interactions than heterologous interactions). Additionally, annelids were heavily enriched in long-range interactions compared to model organisms (chicken and fruit fly), indicating a significant level of genomic interactions between different chromosomes (Fig. 4c-d).

When comparing marine annelids and clitellates, chromosome scrambling resulted in a shift in the inter-intrachromosomal interactions in annelids, regardless of genome size, with a lower value in clitellates indicating a reduction of the inter-chromosomal interaction frequency and an increase in the intra-chromosomal ones, a reflection of more compacted chromosomes. This pattern was maintained when comparing ancestral linkage groups (ALGs) (Fig. 4f).

An exception to this pattern was the case of *Hox* genes: they showed a canonical clustered organisation in marine annelids, but in clitellates (at least in earthworms) they were both duplicated and highly rearranged in different chromosomes (Fig. 4h, i; Supplementary Data 9). However, *Hox* genes from the same ancestral cluster maintained clear long-range interactions between chromosomes, indicating that they interact at the 3D level despite not being physically placed in a gene cluster (Fig. 4g). In leeches, however, many *Hox* genes were lost and interchromosomal interactions between *Hox* genes could not be detected, indicating a deeper reshaping of the *Hox* gene repertoire.

Genes potentially involved in adaptation to life on land and freshwater environments arose via chromosome scrambling

Differential gene expression analysis revealed that several genes involved in response to abiotic stress in clitellates (ie, involved in oxygen homeostasis, response to visible and UV-B light, osmoregulation, etc.; see Methods) either were relocated in the genome or arose as a result of the massive chromosome mixing during that event (Fig. 5a, b). In the first case, new gene interactions were formed as a result of this change of position, possibly because of the operation of different regulatory mechanisms that were newly in proximity with the gene (Fig. 5c). For the case of the newly emerged genes, they mostly belonged to hierarchical orthologous groups (HOGs) that resulted from fission of 'ancient' HOGs (ie, that preexisted the origin of clitellates), as expected under an scenario of chromosome shattering. They had homology with small heat shock proteins, transcription factors, proteins involved in arousal from lethargus in *C. elegans*, photoreception and antistatins (anticoagulants firstly discovered in leeches), among others (Fig. 5d; Supplementary Table 4). Some others had homology with genes involved in cytokinesis and DNA damage control, such as a homolog to RhoB (that mediates apoptosis in neoplastically transformed cells after DNA damage) or protein regulator of cytokinesis 1 (an important regulator during mitotic spindle assembly and cytoplasmic division); these could represent pieces of the genes in HOGs lost during chromosome scrambling involved in cell division that were co-opted to newer functions related to the adaptation to new environments after mixing and changing position in the genomes. Only one HOG could be identified as originating de novo bona fide (ie, not the result of fusion and fission from any existing HOGs), and was involved in sodium-potassium cell trafficking (Fig. 5d; Supplementary Table 4).

Discussion

Our study revealed massive genomic rearrangements at the origin of the clitellates, a clade of non-marine annelids. These chromosomal tectonic shifts completely eliminated the conservation of ancestral linkage groups that is otherwise present throughout metazoans, and did so in a comparatively short evolutionary window (measured not as absolute time but rather as evolutionary distance between the lineages explored). This extent and speed of genome restructuring is incompatible with regular models of chromosome fusion and fission, instead suggesting a process of genome-wide chromosome scrambling. While Simakov et al. [39] and Moggioli et al.[40] reported extensive reorganisation in the genome of a freshwater leech relative to the last common spiralian ancestor based on the analysis of draft genomes, the generation and availability of chromosome-level genomes of several clitellate lineages has revealed the

timing and extent of these massive changes at the within-phylum level, potentially coinciding with the split between marine annelids and clitellates and their habitat transition towards freshwater and land.

Our results provide an example of a complete loss of macrosynteny genome-wide at the within-phylum level of a higher magnitude that previously seen in other animal phyla such as bryozoans[8], cephalopodes[9] or tunicates[10]; from a macrosynteny point of view, genome structure is much more divergent between a marine annelid and a clitellate annelid than between a marine annelid and animals as distantly related as a sponge or a mollusk. Even though no chromosome-level genomes of early-splitting clitellates are available so far, the gene repertoire evolution analyses presented here (including both an extended clitellate dataset representing all main lineages and closely related marine annelids) suggest that this punctuated burst of genomic structural change may coincide with the origin of a singular clade within Annelida - the clitellates. This lineage is characterised by a series of evolutionary novelties including changes in their reproduction mode (marine annelids are mostly dioic while clitellates are hermaphrodite and some parthenogenetic), frequent polyploidy (eg, some earthworms can range between x2 to x12 even at the within-species level[41,42]), a new organ (the clitellum, a 'collar' that forms a reproductive cocoon during part of their life cycles), the loss of parapodia (lateral fleshy protrusions in marine annelids) and the exit from the sea towards freshwater and terrestrial environments (with the exception of some marine leeches from the family Piscicolidae)[43–45]. Notably, clitellates are also characterised by common aneuploidy and an apparent lack of canonical cell division[31], which is congruent with our findings of a peak of gene loss in clitellated enriched in functions putatively related to regulation of the cell cycle, genome stability and chromatin reshaping.

Regarding the functional consequences of these massive rearrangements, here we show that the punctuated burst of complex genome-wide chromosomal rearrangements occurring either in the phylogenetic proximity of clitellates or at the origin of the clade resulted in the generation of some novel genomic elements with a putatively adaptive role in the colonisation of freshwater and terrestrial habitats from marine ancestors. One outstanding question is how this profound genome reshaping event did not result in extinction. The answer may be in the particular genome architecture of marine annelids, which seem not to be organised in compartments, and to be much more floppy than other animal systems explored so far, which may have resulted in a high resilience to the deep genome reshaping occurring after chromosome scrambling.

While the timing of these genomic rearrangement remains unclear, we argue that the genomic hallmarks observed in clitellates are highly unlikely to have arisen via a mechanism of accumulation of pairwise interchromosomal rearrangements over time (particularly given the fact that the branch leading to clitellates is relatively short[22]) and we hypothesise that instead probably occurred during a single cellular catastrophe (for instance, due to a clastogenic event related to a change of environmental factors such as oxygen or radiation levels, resulting in chromosome shattering), or either a series of interconnected translocations mediated by DNA double-strand breaks and repair processes accumulated in relatively short evolutionary distances. The pattern of profound genomic change observed in clitellates may be seen as analogous in a macroevolutionary context to chromoanagenesis, defined as massive and complex chromosomal rearrangements based on the chaotic shattering and restructuring of chromosomes identified in cancer cells[46–48]. Its discovery over the last decade has changed our perception of the formation of chromosomal abnormalities and their aetiology, as it demonstrates that profound genome reshaping via genome-wide massive rearrangements in a single generation is biologically possible[47–49]. Understanding if such a mechanism may be in place in animals beyond vertebrates is particularly revealing in the field of biodiversity genomics, where the role of chromosomal rearrangements in the generation of new forms of life and the mechanisms underlying genomic change at a macroevolutionary scale are still largely unexplored[50].

All in all, our study illustrates how saltational rather than gradual changes played an important role during the evolution of an animal lineage characterised by a series of morphological and ecological innovations, providing new insights into the mode and tempo of macroevolution in animals. These results thus position clitellates as excellent model systems for further investigation into the mechanisms leading to massive genomic rearrangements and their consequences at the architectural and functional level, as well as their potential role as catalyzers of cladogenesis at a macroevolutionary scale in short evolutionary windows.

FIGURES

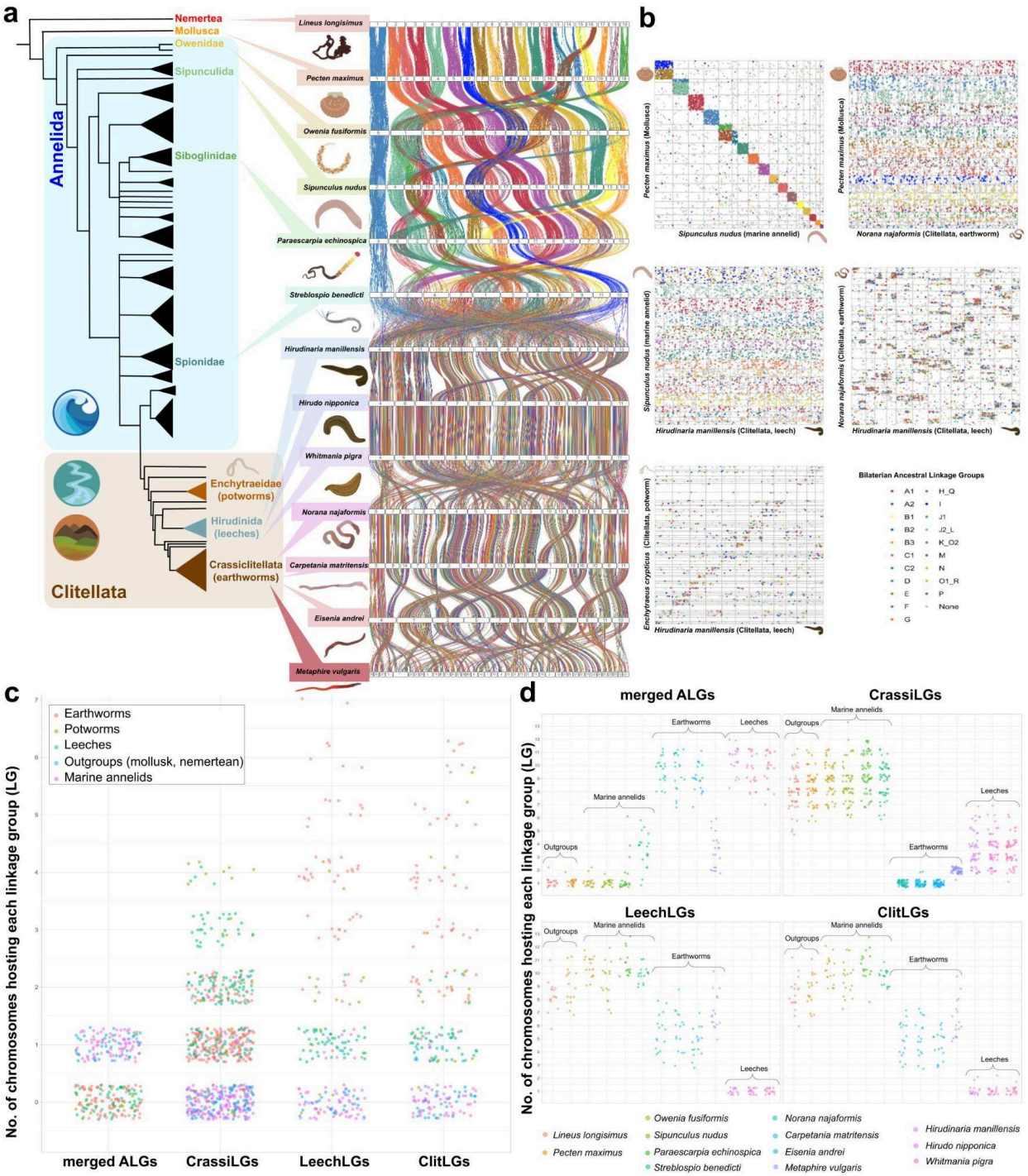
a
Nemertea
Mollusca
Owenidae
Sipunculida
Annelida
Siboglinidae
Spionidae
Enchytraeidae (potworms)
Hirudinida (leeches)
Crassiclitellata (earthworms)
Clitellata
Lineus longisimus
Pecten maximus
Owenia fusiformis
Sipunculus nudus
Paraescarpia echinospica
Streblospio benedicti
Hirudinaria manillensis
Hirudo nipponica
Whitmania pigra
Norana najaformis
Carpetania matritensis
Eisenia andrei
Metaphire vulgaris
b
Pecten maximus (Mollusca)
Sipunculus nudus (marine annelid)
Norana najaformis (Clitellata, earthworm)
Hirudinaria manillensis (Clitellata, leech)
Enchytraeus cryptlicus (Clitellata, potworm)
Bilaterian Ancestral Linkage Groups
A1
A2
B1
B2
B3
C1
C2
D
E
F
G
H_Q
I
J1
J2_L
K_O2
M
N
O1_R
P
None
c
Earthworms
Potworms
Leeches
Outgroups (mollusk, nemertean)
Marine annelids
No. of chromosomes hosting each linkage group (LG)
merged ALGs
CrassiLGs
LeechLGs
ClitLGs
d
merged ALGs
CrassiLGs
LeechLGs
ClitLGs
Outgroups
Marine annelids
Earthworms
Leeches
No. of chromosomes hosting each linkage group (LG)
Lineus longisimus
Pecten maximus
Owenia fusiformis
Sipunculus nudus
Paraescarpia echinospica
Streblospio benedicti
Norana najaformis
Carpetania matritensis
Eisenia andrei
Metaphire vulgaris
Hirudinaria manillensis
Hirudo nipponica
Whitmania pigra

Figure 1. Macrosyntenic evolution of clitellates. 1a. Left, Annelida Tree of Life showing all main lineages and the position of the species included in the present study, based on the topology reported in Capa et al. (Annelida)[51] and Erseus et al. (Clitellata)[22]. Centre, ribbon plot of the chromosome-scale ancestral gene linkage across annelids, with a mollusk and a nemertean as outgroups. For each species, each white rectangle represents a chromosome and the number inside represents the rank order by size of that chromosome in that species. The vertical lines (ribbons) connect the orthologous genes from each genome (see Methods). Colour code follows the Bilaterian-Cnidarian-Sponge Linkage Groups (BCnS LGs) as in[4]. **1b.** Oxford dotplots of the chromosome-scale ancient gene linkage in an early-splitting marine annelid (*Sipunculus nudus*), two clitellates (the earthworm *Norana najaformis* and the enchytraeid *Enchytraeus crypticus,* draft genome) compared to a mollusk (*Pecten maximus*) and a leech (*Hirudinaria manillensis*), showing high macrosyntenic genome conservation between outgroups and marine annelids and the complete rupture of macrosynteny between marine annelids and clitellates. Vertical and horizontal lines define the border of the chromosomes of each pair of species being compared. Each dot represents an orthology relationship between genes at the x and y coordinates. Colour code is as per Fig. 1a. **1c.** Scatterplot showing the number of chromosomes from each main lineage hosting each linkage group (LG) from four datasets: merged ancestral linkage groups (mergedALGs), Crassiclitellata (CrassiLGs), Hirudinida (LeechLG) and Clitellata (ClitLGs). **1d.** Scatterplot showing the number of chromosomes from each species hosting each Linkage Group from the four datasets described in Fig. 1c. All artwork was designed for this study by Gemma I. Martínez-Redondo. [*High-quality figure available here: 🗎 Panel_Fig_1 .pdf]*

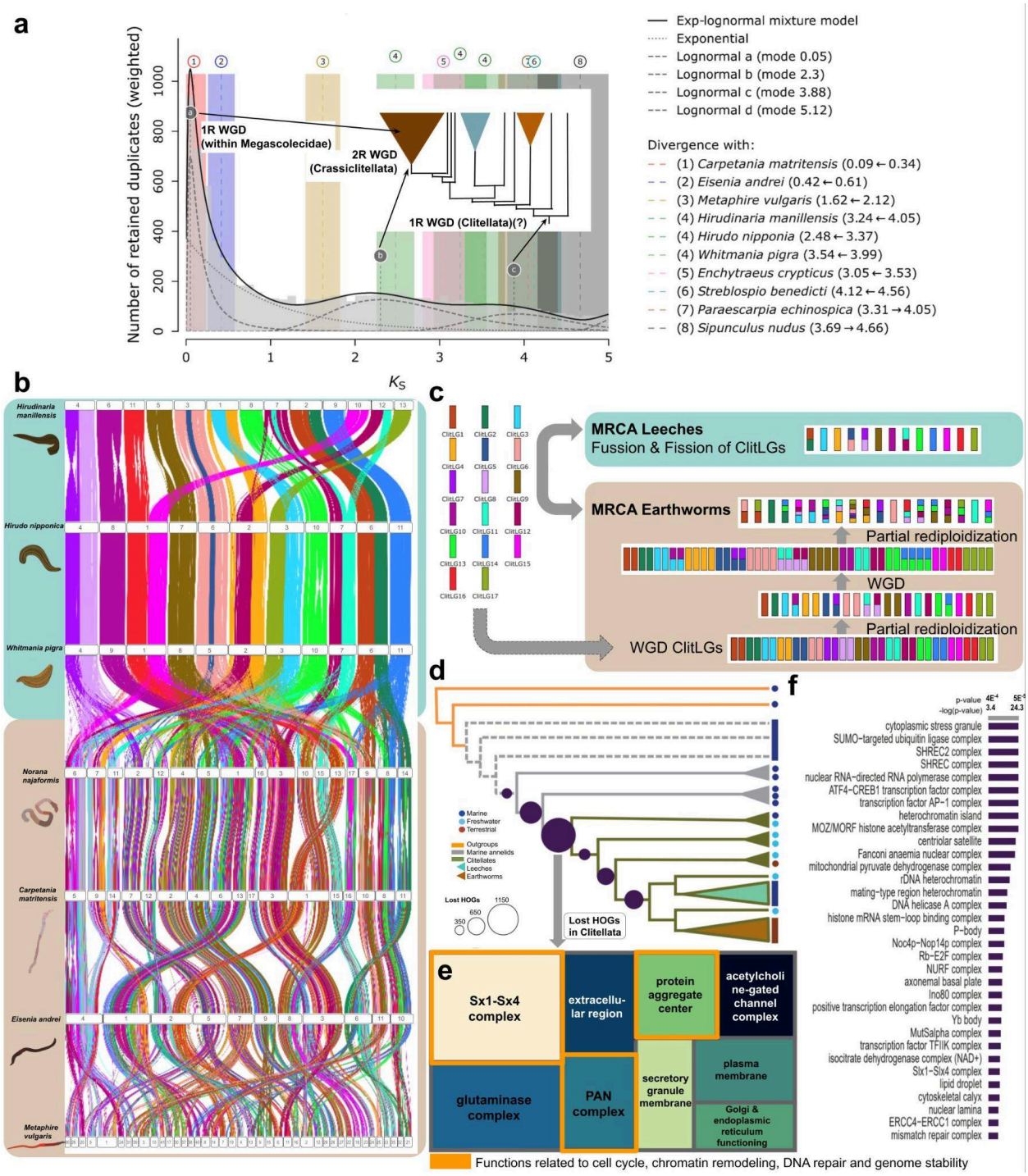

Figure 2. Frequent rare large-scale genomic changes among clitellates after genome-wide chromosome scrambling. a. Substitution-rate-adjusted mixed paralog–ortholog Ks plot for the earthworm *N. najaformis*. The inferred three putative WGD events are indicated with dashed lines labelled a, b and c corresponding to the Ks-based WGD age estimates. The speciation events calculated using the rate-adjusted mode estimates of the ortholog

Ks-distributions between *N. najaformis* and the other species are drawn as long-dashed lines labelled with numbers between 1 and 8 surrounded by coloured boxes ranging from one standard deviation below and above the mean mode estimate. Lines representing the same speciation event share number and colour. Question mark at the WGD indicated in Clitellata reflects the lack of consensus obtained by different analyses (see Supplementary Data 1 and 2). **2b.** Chromosome-scale ribbon plot showing macroevolutionary patterns of Clitellate Linkage Groups (ClitLGs) between leeches and earthworms. The vertical lines (ribbons) connect the orthologous genes from each genome (see Methods). Colour code follows the legend in Fig. 2c. **2c.** Ancestral genome reconstruction of the macroevolutionary events leading to the ancestral genome of the Most Recent Common Ancestor (MRCA) of leeches and earthworms, respectively. WGD, whole genome duplication. **2d.** Phylostratigraphic pattern of gene loss (measured as loss of Hierarchical Orthologous Groups, HOGs) in clitellates and the surrounding nodes. **2e.** Simplified treemap representation of putative functions enriched in lost HOGs in Clitellata, cellular component (p-value<0.05). The size of the square is proportional to the p-value of the enrichment. The most general term per square is shown for simplicity (see Suppl. Mat. for further information). Orange squares comprise functions related to cell cycle, chromatin remodelling, DNA repair and genome stability. **2f.** Extended list of functions lost in Clitellata related to cell cycle, chromatin remodelling, DNA repair and genome stability. Terms are shown when the p-value of the enrichment is lower than $4e^{-4}$ (see also Supplementary Data 10 and the manuscript's GitHub repository). *[High-quality figure available here: 🗎 Panel_Fig_2.pdf]*

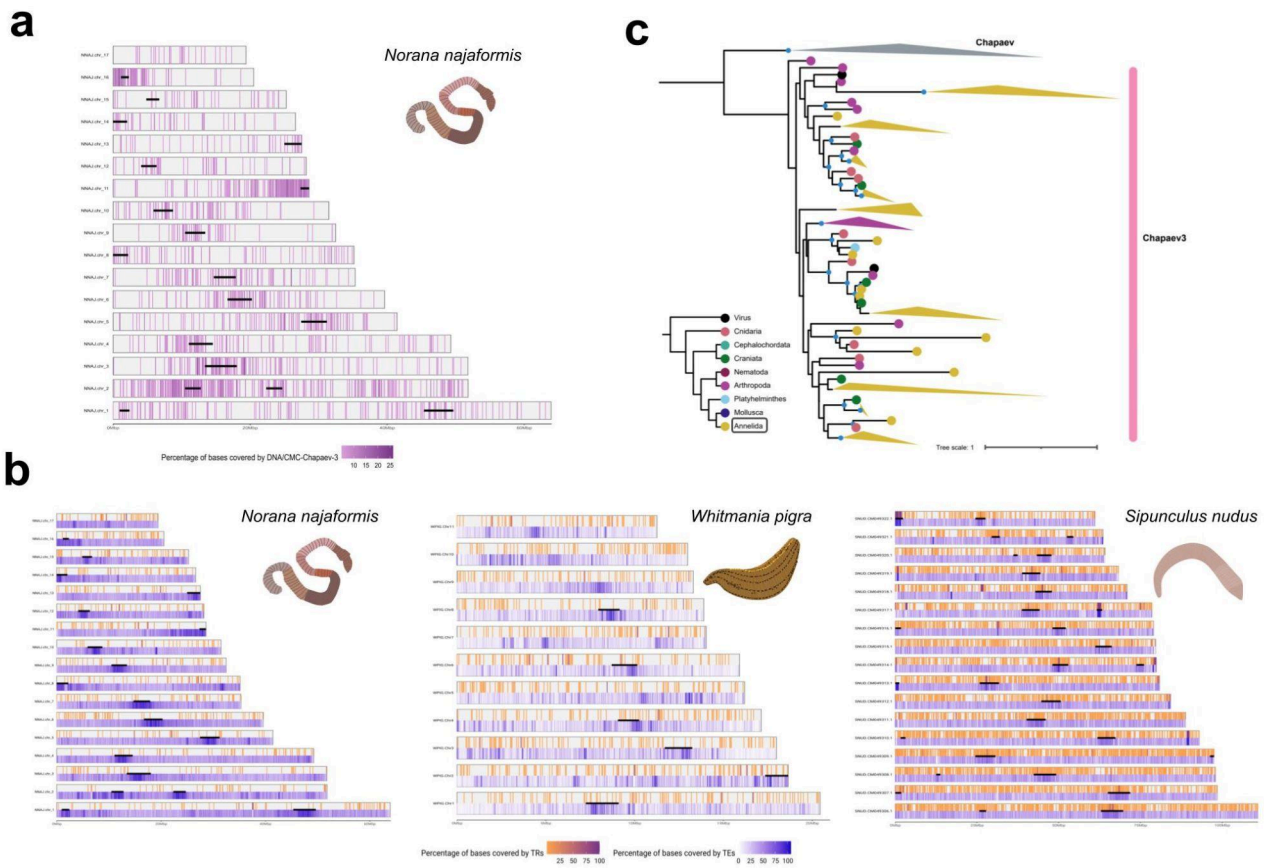

Figure 3. Centromere composition in clitellates. 3a. Distribution of Chapaev-3 transposable element insertions in the genome of *N. najaformis*. The genome was divided in bins of 50kb and the percentage of bases covered by members of the Chapaev-3 superfamily in each bin is shown. Each horizontal bar represents a chromosome. The putative centromeres are shown with a bold black line when they could be inferred with confidence (see Methods). **3b.** Distribution of tandem repeats including satellite DNA (orange) and transposable elements of all superfamilies (purple) in the genomes of *N. najaformis* (earthworm), *W. pigra* (leech) and *S. nudus* (marine annelid), providing evidence or their monocentricity. The putative centromeres are shown with a bold black line. **3c.** Maximum likelihood phylogenetic tree of Chapaev-3 transposases. Small blue dots represent clades with SH-aLRT support >=80% and ultrafast bootstrap support >=95%. Clitellate sequences are indicated in yellow. *[High-quality figure available here: Panel_Fig_3.pdf]*

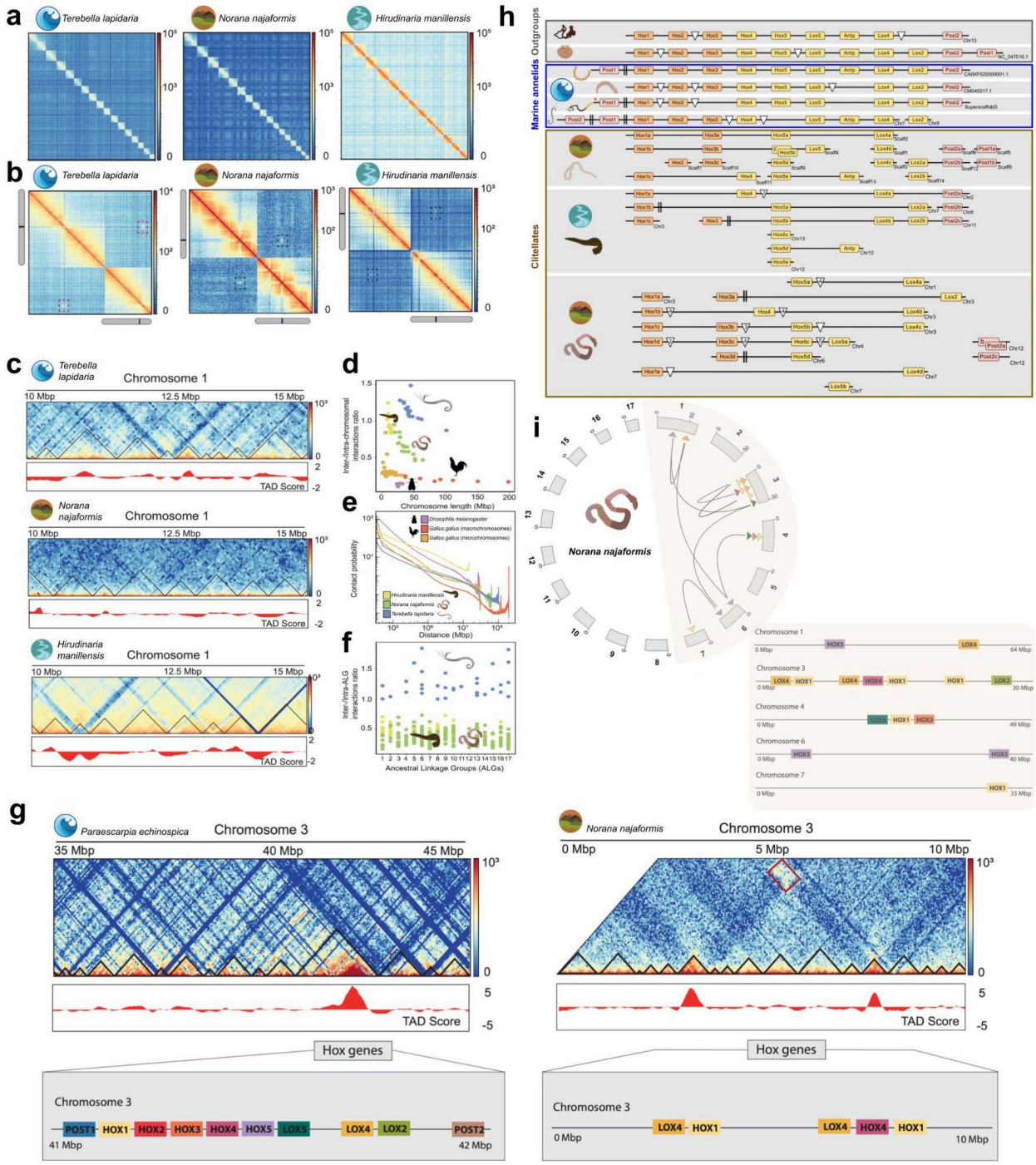

Figure 4. Genome architecture changes in marine annelids and clitellates. 4a. Whole-genome Hi-C contact maps for *Terebella lapidaria*, *Norana najaformis* and *Hirudinaria manillensis*. **4b.** Hi-C contact maps representing a pair of chromosomes/scaffolds depicting centromeric heterologous interactions in all three species. **4c.** Chromosome 1 region-specific 500 Kbp heatmaps and insulator (TAD) score for the same three species. **4d.** Inter-/intra-chromosomal interactions according to chromosome length (in Mbp) in the same

three annelid species, together with chicken (*Gallus gallus*[52]) and the fruit fly (*Drosophila melanogaster*[53]). In the case of the chicken, colour intensity differentiates between macro- (red color) and micro- (orange color) chromosomes. **4e.** Chromosome-specific contact probability P(s) as a function of genomic distance in *T. lapidaria*, *N. najaformis*, *H. manillensis*, chicken and the fruit fly. **4f** Inter-/intra-chromosomal interactions according to ancestral linkage groups (ALG) for the annelids *Paraescarpia echinospica*, *N. najaformis* and *H. manillensis*. **4g.** Examples of Hi-C contact maps depicting an schematic representation of *Hox* genes in *P. echinospica* (left panel) and *N. najaformis* (right panel). *[Artwork either designed explicitly for this study by Gemma I. Martínez-Redondo or retrieved from PhyloPic with a Universal Public Domain Dedication licence in the case of the fly and the chicken]* **4h.** Representation of the main *Hox* cluster across annelids and outgroups. Rearrangements are primarily observed in clitellates. Rectangular boxes represent *Hox* genes, while chromosomes or scaffolds (in the case of the draft genome of the enchytraeid) are symbolised as distinct horizontal lines. Tandem duplications are depicted by duplicated rectangles and are named with consecutive letters. Genes not classified as a *Hox* gene are represented by triangles, with the number of such genes depicted inside the triangle. Separations greater than 300kb between two non-consecutive *Hox* genes are represented by two vertical lines (||). Chromosomes are split in different lines when there are multiple copies of several genes of the *Hox* cluster in the same chromosome. **4i.** Circos plot depicting the long-range interactions involving the *Hox* genes in *N. najaformis*. *[High-quality figure available here: Panel_Fig_4.pdf]*

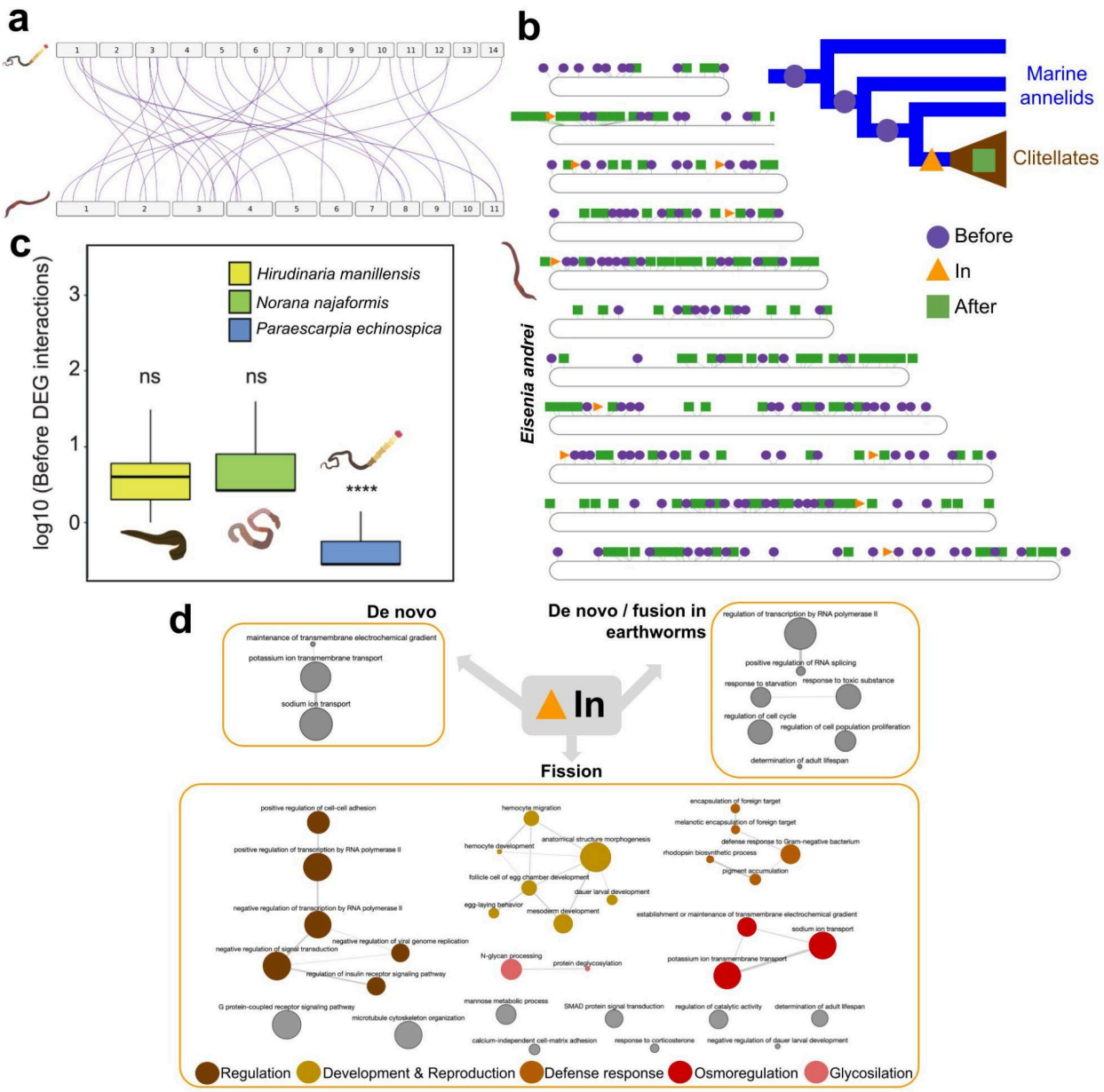

Figure 5. New genes and gene interactions involved in adaptation to freshwater and terrestrial environments arose as a consequence of genome-wide chromosome scrambling. 5a. Genomic position of the differentially expressed genes (DEGs) in *E. andre*i in response to different types of abiotic stress related to habitat change that arose before the origin of clitellates. The position of orthologous genes in a marine annelid (*P. echinospica*) are indicated. Only those with orthologous genes in *P. echinospica* are shown. **5b**. Evolutionary origin of these DEGs in *E. andrei*. Genes are separated into three groups, those which arose

before Clitellata (purple circle), those which arose in the branch leading to Clitellata (ie, during the event of chromoanagenesis; yellow triangle)), and those which appeared after (green square). **5c.** Boxplot displaying log10 interactions between differentially expressed genes under abiotic stress in a leech and an earthworm whose HOGs arose before the diversification of clitellates and that were put in close proximity due to genome reshuffling. The boxplot shows significantly more interactions between them than the corresponding ortholog genes in a marine annelid (*P. echinospica*; two-sided t test, $p < 0.001$). **5d.** Putative enriched function of genes arising during chromoanagenesis (biological process). Coloured networks represent clusters of functions related ontologically; a general biological process is provided for these. *[High-quality figure available here: 🗎 Panel_Fig_5.pdf).*

METHODS

Specimen collection

Norana najaformis specimens (for chromosome-level genome sequencing) were collected in The Sierra de Ordal (Barcelona) under the collection permit number SF0156/22 given by Generalitat de Catalunya. Samples from *Carpetania matritensis* (for chromosome-level genome sequencing) were collected in El Molar (Madrid, Spain) under Nagoya permit ESNC59. Specimens from *Eisenia andrei* and *Hirudo medicinalis* (for stress experiments and differential gene expression, see below) were purchased from local vendors. Datasets for the additional clitellate species and outgroups used in this study were downloaded from available public databases (Supplementary Table 5).

Dissection and stress experiments

All specimens of the species used for wet lab experiments were in starvation for at least 24 hours to minimise any food residue in the gut. For genome sequencing of *N. najaformis* and *C. matritensis* (including Hi-C and RNAseq), specimens were dissected on ice and either flash frozen until tissue was extracted or fresh tissue was used directly for genomic DNA extraction. Several tissues were used for cDNA library construction for genome annotation (regenerated tail, pharynx, seminal vesicle, circulatory system, ganglia, typhlosol, body wall). Samples from *E. andrei* (Clitellata, earthworm) and *H. medicinalis* (Clitellata, leech) were subjected to several

abiotic stress experiments: visible and UV-B light, hyperoxia, hypoxia, and osmoregulation, plus control specimens (see Supplementary Methods). All samples were dissected into head, mid body and tail, flash frozen in liquid nitrogen and kept <-70ºC until further processing.

Long-read whole-genome sample preparation and sequencing

For *N. najaformis*, high molecular weight (HMW) gDNA extraction was performed from a freshly dissected piece from the middle body using the Genomic-tip 20/G (Qiagen) following manufacturer's instructions. Concentration and purity was measured by Qubit DNA BR Assay kit (Thermo Fisher Scientific) and Nanodrop 2000 (Thermo Fisher Scientific) respectively. HMW integrity was assessed by using a Femto Pulse instrument (Agilent technologies). Libraries were prepared using the SMRT-bell library kit (PacBio) and sequenced in a 8M SMRT cell in a PacBio Sequel II under the CCS mode. For *C. matritensis,* DNA was isolated using the MagAttract HMW DNA Kit (Qiagen) with some modifications. DNA was quantified using the Qubit High Sensitivity dsDNA Assay (Thermo Fisher Scientific). After that, the recommended SMRTbell Express Template Prep Kit 2.0 (PacBio) was used to prepare the libraries, following the manufacturer's instructions. Libraries were sequenced in a Sequel II (PacBio), using a 8M SMRT cell, under the CLR mode. For Illumina library preparation, the Illumina DNA Prep kit was used strictly following the manufacturer's instructions. The library was sequenced in a fraction of a NovaSeq PE150 flow cell, aiming for a total output of 90 GB.

Chromatin conformation capture sample preparation and sequencing

A flash-frozen piece from each species was used to perform chromatin conformation capture sequencing (Hi-C). Hi-C libraries were prepared using the Hi-C High-coverage kit (Arima Genomics) in the Metazoa Phylogenomics Lab (Institute of Evolutionary Biology (CSIC-UPF), following manufacturer's instructions including custom modifications when necessary. Sample concentration was assessed by Qubit DNA HS Assay kit (Thermo Fisher Scientific) and library preparation was carried out using the KAPA Hyper Prep kit (Roche) and using the NEBNext® Multiplex Oligos for Illumina (New England Biolabs). Library amplification was carried out with the KAPA HiFi DNA polymerase (Roche). The amplified libraries were sequenced on a

HiSeqXten (Illumina) with a read length of 2x150bp and a 30Gb coverage, resulting in ca. 200M reads per library.

RNA extraction, library preparation and sequencing

RNA extraction from different tissues of *N. najaformis* was obtained from flash-frozen dissected tissues using HigherPurity™ Total RNA Extraction Kit (Canvax Biotech) and following manufacturer's instructions. DNAse treatment was performed with the On-membrane DNase I Set (RNase-Free) (Canvax Biotech). RNA extractions from *E. andrei* and *H. medicinalis* after stress experiments were performed using the TRIzol® reagent (Invitrogen, USA) method following the manufacturer's instructions. Concentration of all samples was assessed by Qubit RNA BR Assay kit (Thermo Fisher Scientific). Samples were subjected to either Illumina's TruSeq Stranded mRNA library preparation kit or Truseq stranded Total RNA Library with Ribo-Zero depending on the tissue and species, and sequenced on a NovaSeq 6000 (Illumina, 2 × 150 bp) for a 6Gb coverage.

Genome assembly

PacBio HiFi reads of *N. najaformis* were assembled using NextDenovo v2.4[54] with parameters read_type=hifi genome_size=600m read_cutoff = 1k. Uncollapsed haplotypes were removed by mapping PacBio HiFi reads to the draft assembly using minimap2 v2.24[55] with parameter -x map-hifi followed by purge_dups v1.0.1. PacBio CLR reads of *C. matritensis* were assembled using NextDenovo v2.4[54] with parameters read_type=clr genome_size=600m read_cutoff = 1k. Uncollapsed haplotypes were removed by mapping PacBio CLR reads to the draft assembly using minimap2 v2.24[55] with parameter -x map-pb followed by purge_dups v1.0.1[56]. Hi-C reads were trimmed using Cutadapt v2.9[57] and pre-processed using bowtie2 2.3.5.1[58] and hicstuff v2.3.0[59] with parameters -e DpnII,HinfI -m iterative. Assemblies were subsequently scaffolded based on Hi-C contacts using instaGRAAL v0.1.6[59] no-opengl branch with parameters --level 5 --cycles 50 and curated with the module instagraal-polish with parameters -m polishing -j NNNNNNNNNN. Gaps in the *N. najaformis* scaffolds were filled using TGS-GapCloser v1.0.1[60] with parameters --ne --tgstype pb and a modification of mapping parameters to -x asm20.

PacBio HiFi reads were then mapped with minimap2 v2.24[55] with parameters -ax map-hifi --MD --secondary=no and provided as input for polishing using HyPo v1.0.3[61] with parameters -c 28 -s 600m -k ccs. Gaps in the *C. matritensis* scaffolds were filled using TGS-GapCloser v1.0.1[60] with parameters --ne --tgstype pb. PacBio CLR reads were mapped with minimap2 v2.24[55] with parameters -ax map-pb --MD --secondary=no. Illumina reads were trimmed using Cutadapt v2.9[57] and mapped using bowtie2 v2.3.5.1[58]. The mapped reads were provided as input to HyPo v1.0.3[61] with parameters -c 140 -s 600m. *k*-mer content was assessed in both final assemblies using KAT comp v2.4.2[62] with the PacBio HiFi dataset for *N. najaformis* and the Illumina dataset for *C. matritensis*. BUSCO v5.0.0[63] was run against the Metazoa odb10 lineage in genome mode. Contact maps were built using hicstuff as previously described and visualised using the module hicstuff view with parameter -b 1000. A complete genome report for both species is available in Supplementary Data 11.

Annotation of transposable elements

Transposable elements (TEs) were *de novo* identified within each genome assembly using RepeatModeler v2.0.3[64] with the LTRStruct flag to identify long terminal repeat (LTR) retrotransposons. Subsequently, the generated library of TE sequences was employed to mask the corresponding genome using RepeatMasker v4.1.2[65]. RepeatMasker was run with the following flags: "-s -a -nolow -no_is -xsmall".

Gene annotation

RNAseq reads were trimmed using Cutadapt v2.9[57] and assembled using Trinity v2.11.0[66]. The transcriptome assembly and the RNA-seq reads were provided as input to Funannotate v1.8.13[67] for training using HISAT2 v2.2.1[68] to map the reads, PASA v2.5.2[69] and TransDecoder v5.5.0[70]. Genes were then predicted using a combination of the training outputs, Augustus v3.3.3[71] and GeneMark-ES v4.68[72], and processed using EVidenceModeler v1.1.1[73].

Differential Gene Expression

Quality assessment of the raw Illumina RNA-seq reads of *H. medicinalis* and *E. andrei* for all experiments mentioned above was performed using FastQC v0.11.9

(https://www.bioinformatics.babraham.ac.uk/projects/fastqc) and adapters and ambiguous bases were trimmed using Trimmomatic v0.39[74] (MINLEN: 75, SLIDINGWINDOW: 4:15, LEADING: 10, TRAILING: 10, AVGQUAL: 30). The trimmed RNA-seq reads were also assessed with FastQC before further analysis. For *H. medicinalis*, a de novo reference transcriptome assembly generated for each one as described above and was used as the reference in differential gene expression analysis, as no genome is available for this species. The indexing and quantification of the transcripts was performed using Salmon v1.5.2[75] for each sample. For each gene, the transcripts per million (TPM) value was calculated and a perl script (align_and_estimate_abundance.pl) included in the Trinity package was used to generate the counts and trimmed mean of M-values (TMM) scaling normalized expression matrices. For *E. andrei*, since there is a high-quality chromosome-level genome available[76], the trimmed RNA-seq reads were aligned to the genome using the splice-aware aligner STAR v2.7.10[77] and featureCounts v2.0.5[78] was used for read summarization. The generated counts matrix was used as input in the perl script run_TMM_normalization_write_FPKM_matrix.pl of the Trinity package to generate the TMM normalised matrix. Differential gene expression analysis was conducted using the perl script run_DE_analysis.pl also included in Trinity toolkit and the edgeR package[79]. Finally, the Benjamini-Hochberg correction method[80] was applied with the script analyze_diff_expr.pl with cut-offs of adjusted p-value at 0.001 (FDR) and a four-fold change.

Inference of ancestral linkage groups

Macrosynteny between chromosome-level genomes of 11 annelid species fully assembled and annotated (including the two genomes of Hormogastridae earthworms generated for this study plus two outgroups (a nemertean - *Lineus longissimus* - and a mollusk - *Pecten maximus*) was explored with odp v0.3.0[81] (Supplementary Data 5). The Bilaterian-Cnidarian-Sponge Linkage Groups (BCnS LGs[4]) were inferred to describe the relation between these linkage groups and the chromosomes of the species in our dataset. For clarity, we merged the linkage groups that were always together in the outgroups and the marine annelids as follows: A1 comprises A1a and A1b, E comprises Ea and Eb, H_Q comprises H and Q, J2_L comprises J2 and L, K_O2 comprises K and O2, O1_R comprises O1 and R and Q comprises Qa, Qb, Qc and Qd. We referred to these linkage groups as merged linkage groups (mergedLGs). Additionally, we used odp v0.3.0[81] to infer linkage groups specific for the leeches (named LeechLGs) using the genomes of the three leeches, and for Crassiclitellata (ie, earthworms, named CrassiLGs) using the genomes of one earthworm per family represented (*N. najaformis* representing

Hormogastridae, *E. andrei* representing Lumbricidae and *Metaphire vulgaris* representing Megascolecidae). Given that *M. vulgaris* experienced a recent genome duplication, pairs of linkage groups corresponding to homologous chromosomes and the same chromosomes in the other two earthworms, were merged (see Fig. 1a). To infer linkage groups for Clitellata (ClitLGs), the LeechLGs and the CrassiLGs were manually merged. To test for the enrichment of a specific linkage group in the set of chromosomes of a species, a Fisher test was used. The list of p-values was corrected using the Benjamini-Hochberg correction method. The ribbon plots and the oxford dotplots were generated using custom R scripts using the ggplot2 and RIdeogram packages and the output of odp. To identify the genomic regions that corresponded to each linkage group in each species, a custom script was used. Briefly, the region between a pair of genes corresponding to the same linkage group was considered to belong to that linkage group as long as there was not more than one gene belonging to a different linkage group. These scripts are available at the GitHub repository https://github.com/MetazoaPhylogenomicsLab/Vargas-Chavez_et_al_2024_Chromosome_shattering_clitellate_origins.

Whole genome alignment and whole genome duplication analysis

An extended dataset of chromosome-level assemblies (including some non-annotated ones of other marine annelid lineages that therefore could not be included for macrosynteny analysis) were aligned using progressiveCactus v2.6.0[82] (Supplementary Table 2). Statistics for the alignment and the ancestral genomes for each node were extracted using halStats from HAL tools[83]. To identify whole genome duplications, ksrates v1.1.3[84] was used as a manual pipeline with default parameters without the collinearity analysis, with all the Clitellates used as focal species plus the draft genome of *Enchytraeus crypticus*. Additionally, Tree2GD v1.0.43[85] was run with default parameters using all annotated chromosome-level annelids plus the draft genome of *E. crypticus* as well.

Transposable element, satellite DNA and centromere identification and analysis

TEs were *de novo* identified within each of the available genome assemblies using RepeatModeler[64] and the obtained library was used to mask the corresponding genome as previously described (see Annotation of transposable elements). Putative centromeres were

identified with custom scripts using their TE density. Briefly, each genome was divided into bins of 50 kb. The fraction of bases covered by TEs was calculated for each bin. These values were smoothed using the mean value of the 50 bins surrounding each bin. Next, the 10% of the bins with the highest coverage were identified and joined when the adjacent ones. Finally, merged bins longer than 1 Mb were selected as the putative centromeres. Putative centromere positions were checked in the contact maps of the chromosome-level assemblies of some species to confirm our inference (a marine annelid - *Terebella lapidaria,* an earthworm - *N. najaformis*, and a leech, *Hirudinaria manillensis*), since our results support the existence of centromere interactions in annelids (see Results). Custom scripts and a more detailed description is available at the GitHub repository (https://github.com/MetazoaPhylogenomicsLab/Vargas-Chavez_et_al_2024_Chromosome_shattering_clitellate_origins). To identify tandem repeats and satellite DNA, Spectral Repeat Finder (SRF[86]) was used. The density throughout the genome was plotted using custom R scripts available at the GitHub repository associated with this manuscript.

Gene repertoire evolutionary dynamics

High-quality genomic data from thirty-five annelid species and two outgroups (Nemertea and Mollusca) were used to infer gene repertoire evolutionary dynamics across the Annelida phylum (Supplementary Table 5), including the newly generated data described above and following the current systematic classification of the phylum[22,23,87–89]. The pipeline described in the MATEdb2 database[90] was used to retrieve the longest isoform for each species. Hierarchical orthologous groups (HOGs) were inferred with OMA v2.6[91]. Gene repertoire evolutionary dynamics across nodes were estimated with pyHam [92] and curated with custom scripts following the workflow available at the GitHub repository (https://github.com/MetazoaPhylogenomicsLab/Vargas-Chavez_et_al_2024_Chromosome_shattering_clitellate_origins). Longest isoforms in protein mode were functionally annotated with both homology-based methods (eggNOG-mapper v2[93]) and natural language processing ones (FANTASIA[94]). Using the Gene Ontology (GO) annotations inferred with FANTASIA, GO enrichment analysis of genes loss in each internode were calculated using the Fisher test and the elim algorithm as implemented in the topGO package[95]. Additionally, to confirm that genes lost in Clitellata were enriched in pathways involved in genome stability, DNA repair, cell cycle

and chromatin remodelling, the genes from those HOGs present in non-clitellate species (*Sipunculus nudus, Streblospio benedicti, Paraescarpia echinospica* and *Pecten maximus* as outgroup) were retrieved for each species, gene names were extracted from the eggNOG-mapper annotations and analysed in the software Reactome (https://reactome.org/, Supplementary Data 3-6).

Comparative genomics analyses of genes involved in response to abiotic stress

To put the genes involved in response to abiotic stress in an evolutionary context, the longest isoforms of the references used in differential gene expression analyses (DEGs) of *H. medicinalis* and *E. andrei* were included in the analysis of gene repertoire evolution. DEGs for each species and experiment were assigned to HOGs and mapped into the phylogeny to have a phylostratigraphic profile of their evolutionary origin. DEGs of *E. andrei* and *H. medicinalis* were classified in three groups based on their origin with respect to Clitellata: 1) 'before', 2) 'in', and 3) 'after' (ie, they arose before the origin of Clitellata, in the branch leading to this clade - influenced by chromoanagenesis - or after). In the case of *H. medicinalis,* due to the lack of a genome we identified the coordinates of orthologous sequences in HOGs inferred from the chromosome-level genome of *Hirudinaria manillensis*[96] corroborated high homology (reciprocal best hits in BLAST+[97]).

Composite gene analyses

An all vs all BLASTp of all our proteomes was performed to create a sequence similarity network with the cleanBlastp command in CompositeSearch[98]. The correspondence between the filtered HOGs and the genes was used, together with the output of cleanBlastp, as input for CompositeSearch[98] to identify composite and component genes. Parameter values suggested in the tutorial were used (E-value of 1e-5, 30% identity, 80% coverage, maximum overlap of 20). The 12 DEGs that originated in Clitellata were then explored to characterise their origin based on the criteria used by Mulhair et al.[99]. A HOG was considered to be a composite HOG when more than half of the genes belonging to that HOG were composites. A composite HOG was considered to have originated from a fusion when most of its component HOGs were inferred to

have originated before the origin of the composite HOG. Similarly, a fission origin is inferred when the age of origin of the component HOGs is younger than the age of origin of the component HOG. In the case of component HOGs, it was considered part of a fusion when all the genes in the HOG were components, and most of the composite genes these genes were components of had an origin younger than the origin of the component HOG. Contrarily, when the origin of the composite HOG predated the origin of the component HOG, it was considered to have originated from fission.

Hox genes inference

Transposase phylogenetic trees

For the *Chapaev-3* family, the consensus sequences for all members of this superfamily from each species were extracted from the species specific TE libraries generated by RepeatModeler[64]. Next, previously identified transposases (together with sequences belonging to *Chapaev* that would be used as outgroups for rooting the trees[100]) were obtained from the peptide RepeatPeps database from RepeatMasker[65]. Clitellate *Chapaev-3* transposase DNA consensus sequences were used as queries in BLASTx homology searches against the *Chapaev-3* sequences from RepeatPeps. Only the best hits for each TE consensus sequence with an aligned region larger than 50 bp were kept and the full sequence was translated using the reading frame from the best hit with custom scripts.

As Zhang et al.[101] suggested that members of the *Chapaev* superfamily could be horizontally transferred between distantly related species using bracoviruses as vectors, we searched RepeatPeps *Chapaev-3* sequences against the NCBI Virus protein database (downloaded on the 5th October 2023) and ViruSite (2023_2 database version[102]) using BITACORA[103]. Only one hit was obtained (BDT63348.1, family *Nimaviridae*). Bracovirus (family Nudiviridae) *Chapaev-3* from Zhang et al.[101] was also included in the analysis. All annelid transposase, RepeatPeps, and identified viral sequences were aligned using MAFFT v7[104]. Resulting alignment was trimmed using the automated option from trimAl v1.4.1[105] and sequences that after trimming had less than 100 amino acids were removed. Phylogenetic tree was then inferred using IQ-TREE 2.2.2.2[106]. ModelFinder[107] established WAG+F+R5 as the best model.

Genomic architecture analysis

For *N. najaformis* we used the Hi-C data newly generated in this project (available from the European Nucleotide Archive, ENA, under accession number ERR11011328) from a section of the midbody. The following Hi-C sequencing datasets processing following the same protocol (and thus providing comparable results, which largely vary depending on the enzymes included in the Hi-C protocol) were downloaded from ENA: *Paraescarpia echinospica* (accession number SRR15733960), *T. lapidaria* (accession number ERR10851521)(both marine annelids) and *H. manillensis* (accession number SRR15881149, a leech). Three datasets were selected to explore inter- and intra- chromosomal interactions, ensuring that they were of similar quality and therefore their comparative analysis did not result in any biases since low quality Hi-C data may result in an increase in the inter/intrachromosomal interactions ratio[108] (see Suppl. Table S6). The first was using the coordinates for the identified *Hox* genes in *H. manillensis*, *P. echinospica,* and *N. najaformis*. The second contained the genomic regions of the mergedALGs coordinates in the aforementioned genomes (see above). The third contained the positions of genes involved in response to abiotic stress previously identified in the three genomes. This was done by identifying the reciprocal best hits to the DEGs of *E. andrei* from the 'before' and 'in' groups (above) using the easy-rbh module from MMSeqs2[109].

Hi-C matrices were built at 500 Kbp resolution by remapping Hi-C reads against the final assembly using Juicer[110] with default parameters. Small scaffolds (< 30 Mbp) were discarded and only chromosome-level superscaffolds, organised by size, were included. Lastly, matrices were normalized, corrected, and plotted using hicNomalize, hicCorrect and hicPlotMatrix from HiCExplorer v.3.7b[111,112]. Using as an input the corrected 500 Kbp matrices normalised to 50M counts, contact probability vs distance curves (P(s)) were calculated using 'hicPlotDistvsCounts' from HiCExplorer v3.7 [111,112]. The curves were plotted using Rstudio and maximum distance of 1 Gbp. Using the module "hicConvertMatrix" from HiCExplorer v3.7b[111,112], intra- and inter-chromosomal interactions were obtained by converting the corrected and normalised 500 Kbp matrices from h5 to interactions. The inter-chromosome/intra-chromosome interaction ratio was then determined using Rstudio to calculate the mean number of inter- and intra-chromosomal interactions per chromosome. The correspondence between ALGs and chromosomes was used to analyse in the same way ALGs interactions distribution. TAD

insulator score was obtained using Hi-C analysis tools package FAN-C v0.9.1[105]. The module 'fanc insulation' was used to calculate the insulator score on normalised 50 Kbp matrices. Finally, inter- and intra-chromosomal long-range contact were identified by extracting interactions 2-fold higher than the inter- or intra- chromosomal interactions mean, respectively. For intra-chromosomal interactions only contacts between loci separated from at least 2Mb were considered long-range interactions. The long-range interactions connecting Hox genes were plotted on the circos plot.

Data availability

The sequencing reads used to assemble the genome of *C. matritensis* are available in the ENA database under the accession number PRJEB74758. Those used to assemble and annotate the genome of *N. najaformis* are available under the accession number PRJEB60177. The annotated genome of *C. matritensis* is available under project PRJEB74757 and that of *N. najaformis* under PRJEB74664. The sequencing reads for the stress experiments in *H. medicinalis* and *E. andrei* are under the accession numbers PRJEB74906 and PRJEB74907 respectively. Data retrieved from public repositories is available under accession numbers reported in Supplementary Data 5.

Code availability

Custom scripts are available in our GitHub repository (https://github.com/MetazoaPhylogenomicsLab/Vargas-Chavez_et_al_2024_Chromosome_shattering_clitellate_origins).

Acknowledgements

We dedicate this manuscript to Darío J. Díaz Cosín (PhD advisor of MN and RF), who devoted his career to the study of earthworm biology and was always fascinated by them; little did we know then how truly special (from an evolutionary perspective) these creatures are. We thank Aureliano Bombarely, Rita Rebollo, Clément Goubert and Francesco Cicconardi for insightful

discussions on genome rearrangements, transposable elements and tips on whole genome alignment, respectively; Pau Balart and Leandro Aristide for helping on the capture of the *Norana najaformis* specimens; Koryu Kin and Gonzalo Bercedo for kindly allowing us to use the Hypoxylab; and Natasha Tilikj and Luis Cunha for aiding in data generation for the *Carpetania matritensis* genome. GIM-R acknowledges the support of Secretaria d'Universitats i Recerca del Departament d'Empresa i Coneixement de la Generalitat de Catalunya and ESF Investing in your future (grant 2021 FI_B 00476). LA-G was supported by an FPI predoctoral fellowship from the Ministry of Economy and Competitiveness (PRE-2018-083257). MN acknowledges support from Ramón y Cajal fellowship (RYC2018-024654-I) and by Grant PGC2018-094112-A-I00 (which provided funding for the genome of *C. matritensis*) both from MCIN/AEI/10.13039/501100011033 and by "ESF: Investing in your future" and "ERDF: A way of making Europe" respectively. ARH acknowledges support from the Spanish Ministry of Science and Innovation (PID2020-112557GB-I00) funded by AEI/10.13039/501100011033, the Secretaria d'Universitats i Recerca del Departament d'Economia i Coneixement de la Generalitat de Catalunya (AGAUR 2021-SGR00122) and the Catalan Institution for Research and Advanced Studies (ICREA). AMcL was supported by funding from the European Research Council grant agreement 771419. RF acknowledges support from the following sources of funding: Ramón y Cajal fellowship (grant agreement no. RYC2017-22492 funded by MCIN/AEI /10.13039/501100011033 and ESF 'Investing in your future'), the European Research Council (this project has received funding from the European Research Council (ERC) under the European's Union's Horizon 2020 research and innovation programme (grant agreement no. 948281)), the Catalan Biogenome Project (which provided funding for sequencing the genome of *N. najaformis*) and the Secretaria d'Universitats i Recerca del Departament d'Economia i Coneixement de la Generalitat de Catalunya (AGAUR 2021-SGR00420). We also thank Centro de Supercomputación de Galicia and CSIC for access to computer resources (CESGA and DRAGO respectively).

Author contributions

CV-C and RF designed the study and coordinated data analysis. CV-C conducted analysis on macrosynteny, transposable element evolution and satellite repeat identification, whole genome duplication and ancestral genome reconstructions. LB-A carried out analyses on gene repertoire evolution. GIM-R performed composite gene analysis and transposase and *Hox* gene phylogenies. KE conducted stress experiments and differential gene expression analysis. LA-G

and AR-H led analysis of genome architecture. JS-O and NE conducted wet lab experiments and stress experiments. NG and CV-C led genome assembly and annotation. JFF assisted with genome assembly. MN generated new data for genome assembly. CV-C, LB-A, GIM-R, KE, JS-O, LA-G, AR-H, AMcL and RF interpreted the data and discussed results. RF wrote the initial version of the manuscript, with input from all authors. RF provided resources and supervised the study. All authors revised and approved the final version of the manuscript.

Competing interests

The authors declare no competing interests.

References

1. Fernández, R. & Gabaldón, T. Gene gain and loss across the metazoan tree of life. *Nat Ecol Evol* **4**, 524–533 (2020).
2. Nakatani, Y. & McLysaght, A. Genomes as documents of evolutionary history: a probabilistic macrosynteny model for the reconstruction of ancestral genomes. *Bioinformatics* **33**, i369–i378 (2017).
3. Nakatani, Y. *et al.* Reconstruction of proto-vertebrate, proto-cyclostome and proto-gnathostome genomes provides new insights into early vertebrate evolution. *Nat. Commun.* **12**, 4489 (2021).
4. Simakov, O. *et al.* Deeply conserved synteny and the evolution of metazoan chromosomes. *Sci Adv* **8**, eabi5884 (2022).
5. Marlétaz, F. *et al.* The hagfish genome and the evolution of vertebrates. *Nature* **627**, 811–820 (2024).
6. Yu, D. *et al.* Hagfish genome elucidates vertebrate whole-genome duplication events and their evolutionary consequences. *Nat Ecol Evol* **8**, 519–535 (2024).
7. Wright, C. J., Stevens, L., Mackintosh, A., Lawniczak, M. & Blaxter, M. Comparative

genomics reveals the dynamics of chromosome evolution in Lepidoptera. *Nat Ecol Evol* (2024) doi:10.1038/s41559-024-02329-4.

8. Lewin, T. D. *et al.* Fusion, fission, and scrambling of the bilaterian genome in Bryozoa. *bioRxiv* 2024.02.15.580425 (2024) doi:10.1101/2024.02.15.580425.
9. Albertin, C. B. *et al.* Genome and transcriptome mechanisms driving cephalopod evolution. *Nat. Commun.* **13**, 1–14 (2022).
10. Plessy, C. *et al.* Extreme genome scrambling in marine planktonic Oikopleura dioica cryptic species. *Genome Res.* **34**, 426–440 (2024).
11. Álvarez-González, L. *et al.* Principles of 3D chromosome folding and evolutionary genome reshuffling in mammals. *Cell Rep.* **41**, 111839 (2022).
12. Waters, P. D. *et al.* Microchromosomes are building blocks of bird, reptile, and mammal chromosomes. *Proc. Natl. Acad. Sci. U. S. A.* **118**, (2021).
13. Damas, J. *et al.* Evolution of the ancestral mammalian karyotype and syntenic regions. *Proc. Natl. Acad. Sci. U. S. A.* **119**, e2209139119 (2022).
14. Damas, J., Kim, J., Farré, M., Griffin, D. K. & Larkin, D. M. Reconstruction of avian ancestral karyotypes reveals differences in the evolutionary history of macro- and microchromosomes. *Genome Biol.* **19**, 155 (2018).
15. Farré, M., Robinson, T. J. & Ruiz-Herrera, A. An Integrative Breakage Model of genome architecture, reshuffling and evolution: The *Integrative Breakage Model* of genome evolution, a novel multidisciplinary hypothesis for the study of genome plasticity. *Bioessays* **37**, 479–488 (2015).
16. Escudero, M., Marques, A., Lucek, K. & Hipp, A. L. Genomic hotspots of chromosome rearrangements explain conserved synteny despite high rates of chromosome evolution in a holocentric lineage. *Mol. Ecol.* (2023) doi:10.1111/mec.17086.
17. Backus, B. T. Some bryozoan karyotypes and chromosome numbers. *Genetical Res.* **29** (3): 187-191 (1977).

18. Eldredge, N. *Time Frames: The Rethinking of Darwinian Evolution and the Theory of Punctuated Equilibria*. (Simon & Schuster, 1986).
19. Eldredge, N. *Time Frames: The Evolution of Punctuated Equilibria*. (Princeton University Press, 2014).
20. Eldredge, N., & Gould, S.J. Punctuated equilibria: An alternative to phyletic gradualism. In Schopf, T.J.M. (ed.). Models in Paleobiology. San Francisco, CA: Freeman Cooper. pp. 82–115 (1972).
21. Kellis, M., Birren, B. W. & Lander, E. S. Proof and evolutionary analysis of ancient genome duplication in the yeast Saccharomyces cerevisiae. *Nature* **428**, 617–624 (2004).
22. Erséus, C. *et al.* Phylogenomic analyses reveal a Palaeozoic radiation and support a freshwater origin for clitellate annelids. *Zool. Scr.* **49**, 614–640 (2020).
23. Andrade, S. C. S. *et al.* Articulating 'Archiannelids': Phylogenomics and Annelid Relationships, with Emphasis on Meiofaunal Taxa. *Mol. Biol. Evol.* **32**, 2860–2875 (2015).
24. Muldal, S. The chromosomes of the earthworms: I. The evolution of polyploidy. *Heredity* **6**, 56–76 (1952).
25. Ly, P. & Cleveland, D. W. Rebuilding chromosomes after catastrophe: emerging mechanisms of chromothripsis. *Trends Cell Biol.* **27**, 917–930 (2017).
26. Xu, X. *et al.* Structure specific DNA recognition by the SLX1-SLX4 endonuclease complex. *Nucleic Acids Res.* **49**, 7740–7752 (2021).
27. Labib, K. & Gambus, A. A key role for the GINS complex at DNA replication forks. *Trends Cell Biol.* **17**, 271–278 (2007).
28. Ortega, J., Lee, G. S., Gu, L., Yang, W. & Li, G.-M. Mispair-bound human MutS-MutL complex triggers DNA incisions and activates mismatch repair. *Cell Res.* **31**, 542–553 (2021).
29. Sugiyama, T. *et al.* SHREC, an effector complex for heterochromatic transcriptional silencing. *Cell* **128**, 491–504 (2007).

30. Álvarez-González, L. *et al.* 3D chromatin remodelling in the germ line modulates genome evolutionary plasticity. *Nat. Commun.* **13**, 2608 (2022).
31. Pavlíček, T., Cohen, T., Yadav, S., Glasstetter, M., Král, P. & Pearlson, O. Aneuploidy occurrence in Oligochaeta. *Ecology and Evolutionary Biology*, ***1*(3)**, 57-63 (2016).
32. Baker, T. M., Waise, S., Tarabichi, M. & Van Loo, P. Aneuploidy and complex genomic rearrangements in cancer evolution. *Nat Cancer* **5**, 228–239 (2024).
33. Mazzagatti, A., Engel, J. L. & Ly, P. Boveri and beyond: Chromothripsis and genomic instability from mitotic errors. *Mol. Cell* **84**, 55–69 (2024).
34. Michael Ghadimi, B. & Ried, T. *Chromosomal Instability in Cancer Cells*. (Springer, 2015).
35. Malik, H. S. & Henikoff, S. Major evolutionary transitions in centromere complexity. *Cell* **138**, 1067–1082 (2009).
36. Malik, H. S. & Henikoff, S. Conflict begets complexity: the evolution of centromeres. *Curr. Opin. Genet. Dev.* **12**, 711–718 (2002).
37. Henikoff, S., Ahmad, K. & Malik, H. S. The centromere paradox: stable inheritance with rapidly evolving DNA. *Science* **293**, 1098–1102 (2001).
38. Hoencamp, C. *et al.* 3D genomics across the tree of life reveals condensin II as a determinant of architecture type. *Science* **372**, 984–989 (2021).
39. Simakov, O. *et al.* Insights into bilaterian evolution from three spiralian genomes. *Nature* **493**, 526–531 (2012).
40. Moggioli, G. *et al.* Distinct genomic routes underlie transitions to specialised symbiotic lifestyles in deep-sea annelid worms. *Nat. Commun.* **14**, 2814 (2023).
41. [No title]. https://www.nature.com/articles/hdy19524.pdf.
42. Casellato, S. & Others. On polyploidy in Oligochaetes, with particular reference to lumbricids. in *On earthworms* 75–87 (Mucchi Editore., 1987).
43. Brusca, R. C. & Brusca, G. J. XXVI, Invertebrados, 2\textordfeminine edición, Madrid y otros. Preprint at (2005).

44. Encyclopaedia britannica. *Am. J. Surg.* **10**, 164–166 (1930).

45. Bleidorn, C., Helm, C., Weigert, A., & Aguado, M. Annelida. In A. Wanninger (Ed.), Evolutionary developmental biology of invertebrates 2 (pp. 193–230). Vienna: Springer (2015).

46. Stephens, P. J. *et al.* Massive genomic rearrangement acquired in a single catastrophic event during cancer development. *Cell* **144**, 27–40 (2011).

47. Pellestor, F., Gaillard, J. B., Schneider, A., Puechberty, J. & Gatinois, V. Chromoanagenesis, the mechanisms of a genomic chaos. *Semin. Cell Dev. Biol.* **123**, 90–99 (2022).

48. Holland, A. J. & Cleveland, D. W. Chromoanagenesis and cancer: mechanisms and consequences of localized, complex chromosomal rearrangements. *Nat. Med.* **18**, 1630–1638 (2012).

49. Heng, H. H. *Genome Chaos: Rethinking Genetics, Evolution, and Molecular Medicine*. (Academic Press, 2019).

50. Lucek, K. *et al.* The impact of chromosomal rearrangements in speciation: From micro- to macroevolution. *Cold Spring Harb. Perspect. Biol.* **15**, (2023).

51. Capa, M., Aguado, M. T., & Bleidorn, C. Annelida, Polychaeta: Revisión sistemática actualizada: cambios acontecidos entre 2001 y 2007. In Fauna Ibérica. Madrid, Spain: CSIC-Museo Nacional de Ciencias Naturales (2018).

52. Fishman, V. *et al.* 3D organization of chicken genome demonstrates evolutionary conservation of topologically associated domains and highlights unique architecture of erythrocytes' chromatin. *Nucleic Acids Res.* **47**, 648–665 (2019).

53. Arzate-Mejía, R. G., Josué Cerecedo-Castillo, A., Guerrero, G., Furlan-Magaril, M. & Recillas-Targa, F. In situ dissection of domain boundaries affect genome topology and gene transcription in Drosophila. *Nat. Commun.* **11**, 894 (2020).

54. Hu, J. *et al.* An efficient error correction and accurate assembly tool for noisy long reads.

bioRxiv 2023.03.09.531669 (2023) doi:10.1101/2023.03.09.531669.

55. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
56. Guan, D. *et al.* Identifying and removing haplotypic duplication in primary genome assemblies. *Bioinformatics* **36**, 2896–2898 (2020).
57. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17**, 10–12 (2011).
58. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
59. Matthey-Doret, C., Baudry, L., Bignaud, A., Montagne R, Cournac, A., Guiglielmoni, N., Foutel-Rodier, T. and Scolari V.F. hicstuff: Simple library/pipeline to generate and handle Hi-C data . Zenodo. http://doi.org/10.5281/zenodo.4066363 (2020).
60. Xu, M. *et al.* TGS-GapCloser: A fast and accurate gap closer for large genomes with low coverage of error-prone long reads. *Gigascience* **9**, (2020).
61. Kundu, R., Casey, J. & Sung, W.-K. HyPo: Super Fast & Accurate Polisher for Long Read Genome Assemblies. *bioRxiv* 2019.12.19.882506 (2019) doi:10.1101/2019.12.19.882506.
62. Mapleson, D., Garcia Accinelli, G., Kettleborough, G., Wright, J. & Clavijo, B. J. KAT: a K-mer analysis toolkit to quality control NGS datasets and genome assemblies. *Bioinformatics* **33**, 574–576 (2017).
63. Manni, M., Berkeley, M. R., Seppey, M., Simão, F. A. & Zdobnov, E. M. BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes. *Mol. Biol. Evol.* **38**, 4647–4654 (2021).
64. Flynn, J. M. *et al.* RepeatModeler2 for automated genomic discovery of transposable element families. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 9451–9457 (2020).
65. Tarailo-Graovac, M. & Chen, N. Using RepeatMasker to identify repetitive elements in

genomic sequences. *Curr. Protoc. Bioinformatics* **Chapter 4**, 4.10.1–4.10.14 (2009).

66. Haas, B. J. *et al.* De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat. Protoc.* **8**, 1494–1512 (2013).
67. Palmer, J. & Stajich, J. nextgenusfs/funannotate: funannotate v1.5.3 (1.5.3). Zenodo. https://doi.org/10.5281/zenodo.2604804 (2019).
68. Kim, D., Paggi, J. M., Park, C., Bennett, C. & Salzberg, S. L. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol.* **37**, 907–915 (2019).
69. Haas, B. J. *et al.* Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res.* **31**, 5654–5666 (2003).
70. Haas, B., Papanicolaou, A., Yassour, M. & Others. TransDecoder. *TransDecoder* (2017).
71. Keller, O., Kollmar, M., Stanke, M. & Waack, S. A novel hybrid gene prediction method employing protein multiple sequence alignments. *Bioinformatics* **27**, 757–763 (2011).
72. Lomsadze, A., Ter-Hovhannisyan, V., Chernoff, Y. O. & Borodovsky, M. Gene identification in novel eukaryotic genomes by self-training algorithm. *Nucleic Acids Res.* **33**, 6494–6506 (2005).
73. Haas, B. J. *et al.* Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biol.* **9**, R7 (2008).
74. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
75. Patro, R., Duggal, G., Love, M. I., Irizarry, R. A. & Kingsford, C. Salmon provides fast and bias-aware quantification of transcript expression. *Nat. Methods* **14**, 417–419 (2017).
76. Shao, Y. *et al.* Genome and single-cell RNA-sequencing of the earthworm *Eisenia andrei* identifies cellular mechanisms underlying regeneration. *Nat. Commun.* **11**, 2656 (2020).
77. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
78. Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for

assigning sequence reads to genomic features. *Bioinformatics* **30**, 923–930 (2014).

79. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).
80. Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. R. Stat. Soc. Series B Stat. Methodol.* **57**, 289–300 (1995).
81. Schultz, D. T. *et al.* Ancient gene linkages support ctenophores as sister to other animals. *Nature* **618**, 110–117 (2023).
82. Armstrong, J. *et al.* Progressive Cactus is a multiple-genome aligner for the thousand-genome era. *Nature* **587**, 246–251 (2020).
83. Hickey, G., Paten, B., Earl, D., Zerbino, D. & Haussler, D. HAL: a hierarchical format for storing and analyzing multiple genome alignments. *Bioinformatics* **29**, 1341–1342 (2013).
84. Sensalari, C., Maere, S. & Lohaus, R. ksrates: positioning whole-genome duplications relative to speciation events in KS distributions. *Bioinformatics* **38**, 530–532 (2022).
85. Chen, D., Zhang, T., Chen, Y., Ma, H. & Qi, J. Tree2GD: a phylogenomic method to detect large-scale gene duplication events. *Bioinformatics* **38**, 5317–5321 (2022).
86. Zhang, Y., Chu, J., Cheng, H. & Li, H. De novo reconstruction of satellite repeat units from sequence data. *Genome Res.* **33**, 1994–2001 (2023).
87. Struck, T. H. Direction of evolution within Annelida and the definition of Pleistoannelida. *J. Zoolog. Syst. Evol. Res.* **49**, 340–345 (2011).
88. Struck, T. H. *et al.* The evolution of annelids reveals two adaptive routes to the interstitial realm. *Curr. Biol.* **25**, 1993–1999 (2015).
89. Weigert, A. *et al.* Illuminating the base of the annelid tree using transcriptomics. *Mol. Biol. Evol.* **31**, 1391–1401 (2014).
90. Martínez-Redondo, G. I. *et al.* MATEdb2, a collection of high-quality metazoan proteomes

across the Animal Tree of Life to speed up phylogenomic studies. *bioRxiv* 2024.02.21.581367 (2024) doi:10.1101/2024.02.21.581367.

91. Altenhoff, A. M. *et al.* OMA standalone: orthology inference among public and custom genomes and transcriptomes. *Genome Res.* **29**, 1152–1163 (2019).
92. Train, C.-M., Pignatelli, M., Altenhoff, A. & Dessimoz, C. iHam and pyHam: visualizing and processing hierarchical orthologous groups. *Bioinformatics* **35**, 2504–2506 (2019).
93. Cantalapiedra, C. P., Hernández-Plaza, A., Letunic, I., Bork, P. & Huerta-Cepas, J. eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale. *Mol. Biol. Evol.* **38**, 5825–5829 (2021).
94. Martínez-Redondo, G. I., Barrios-Núñez, I., Vázquez-Valls, M., Rojas, A. M. & Fernández, R. Illuminating the functional landscape of the dark proteome across the Animal Tree of Life through natural language processing models. *bioRxiv* 2024.02.28.582465 (2024) doi:10.1101/2024.02.28.582465.
95. TopGO. *Bioconductor* https://bioconductor.org/packages/release/bioc/html/topGO.html.
96. Zheng, J. *et al.* Molecular mechanisms underlying hematophagia revealed by comparative analyses of leech genomes. *Gigascience* **12**, (2022).
97. Camacho, C. *et al.* BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421 (2009).
98. Pathmanathan, J. S., Lopez, P., Lapointe, F.-J. & Bapteste, E. CompositeSearch: A Generalized Network Approach for Composite Gene Families Detection. *Mol. Biol. Evol.* **35**, 252–255 (2018).
99. Mulhair, P. O. *et al.* Bursts of novel composite gene families at major nodes in animal evolution. *bioRxiv* 2023.07.10.548381 (2023) doi:10.1101/2023.07.10.548381.
100. Novikova, O. S. & Blinov, A. G. Origin, evolution, and distribution of different groups of non-LTR retrotransposons among eukaryotes. *Russ. J. Genet.* **45**, 129–138 (2009).
101. Zhang, H.-H., Feschotte, C., Han, M.-J. & Zhang, Z. Recurrent horizontal transfers of

Chapaev transposons in diverse invertebrate and vertebrate animals. *Genome Biol. Evol.* **6**, 1375–1386 (2014).

102. Stano, M., Beke, G. & Klucar, L. viruSITE-integrated database for viral genomics. *Database* **2016**, (2016).

103. Vizueta, J., Sánchez-Gracia, A. & Rozas, J. bitacora: A comprehensive tool for the identification and annotation of gene families in genome assemblies. *Mol. Ecol. Resour.* **20**, 1445–1452 (2020).

104. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).

105. Capella-Gutiérrez, S., Silla-Martínez, J. M. & Gabaldón, T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972–1973 (2009).

106. Minh, B. Q. *et al.* IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Mol. Biol. Evol.* **37**, 1530–1534 (2020).

107. Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., von Haeseler, A. & Jermiin, L. S. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat. Methods* **14**, 587–589 (2017).

108. Lajoie, B. R., Dekker, J. & Kaplan, N. The Hitchhiker's guide to Hi-C analysis: practical guidelines. *Methods* **72**, 65–75 (2015).

109. Steinegger, M. & Söding, J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat. Biotechnol.* **35**, 1026–1028 (2017).

110. Durand, N. C. *et al.* Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. *Cell Syst* **3**, 95–98 (2016).

111. Ramírez, F. *et al.* High-resolution TADs reveal DNA sequences underlying genome organization in flies. *Nat. Commun.* **9**, 189 (2018).

112. Wolff, J. *et al.* Galaxy HiCExplorer 3: a web server for reproducible Hi-C, capture Hi-C and

single-cell Hi-C data analysis, quality control and visualization. *Nucleic Acids Res.* **48**, W177–W184 (2020).

113. Kruse, K., Hug, C. B. & Vaquerizas, J. M. FAN-C: a feature-rich framework for the analysis and visualisation of chromosome conformation capture data. *Genome Biol.* **21**, 303 (2020).