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 & Genome Evolution 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 gained from a genome-level perspective have revealed a striking pattern of conserved macrosynteny within chromosomes across huge phylogenetic 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 and homeostasis of cell division. Notably, while these rearrangements broke short-range interactions observed between *Hox* genes in marine annelids, they were reformed as long-range interactions in clitellates. These genomic rearrangements led to the relocation of genes and the formation of new chimeric genetic elements, both of which may have contributed to the adaptation of clitellates to freshwater and terrestrial environments. Our findings provide evidence of a massive genomic reshaping within clitellates at 2D and 3D levels, and suggest that synteny may not limit the structural evolution of this animal lineage. Our study thus suggests that the genomic landscape of Clitellata resulted from a rare burst of genomic changes that ended a long period of stability that persists across large phylogenetic distances.

Introduction

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. However, 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. In particular, 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 (e.g., gene gain, duplication and loss[1]) or whole-genome duplication. For instance, Nakatani et al.[2] and Simakov et al.[3] reported chromosome-scale conservation of gene linkages (ie, macrosynteny patterns, referred to as ancestral linkage groups, ALGs[3]) 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, molluscs 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 (e.g.[2,4,5]), lepidopterans[6], bryozoans[7], cephalopods[8] or tunicates[9], demonstrated the presence of extensive rearrangements. Regarding invertebrates, in lepidopterans, Wright et al.[6] 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.[7] 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[8] and tunicates[9] also show genomic rearrangements at the level of ALGs, but to a lesser extent compared for instance to bryozoans. 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[10–14], both the patterns and mechanisms of this phenomenon are still poorly understood in most invertebrate phyla. Most studies reporting genomic

rearrangements focus mostly on the pattern; however, 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[15]. By contrast, karyotype studies of bryozoans have reported monocentric chromosomes[16], suggesting that 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.

In this study, we generated chromosome-level genome assemblies of two earthworms from the family Hormogastridae (*Norana najaformis* and *Carpetania matritensis*), and compared them with nine annelid genomes to better understand the genomic changes leading to the colonisation of freshwater and terrestrial environments in this animal phylum. 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 chromosomal interactions, resulting in a shift towards an increase in intrachromosomal 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 either arose or changed their position through this genome scrambling process with a putative adaptive role in the colonisation of 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, resulting from a punctual burst of genomic reshaping rather than stepwise gradual genomic changes, but also suggest that clitellate genome evolution is not limited by syntenic constraints.

Results

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

To identify genomic signatures that could inform the genetic and physiological basis of the adaptations to non-marine ecological niches in clitellates, we used long PacBio HiFi and Hi-C reads to assemble chromosome-level genomes of two earthworms of the family Hormogastridae (*N. najaformis* and *C. matritensis*) and compared them with other clitellate and marine annelid genomes available in public databases. The assembled genomes span 608 Mb in *N. najaformis* (1,842 scaffolds with 626 gaps and an N50 of 36.9 Mb) and 588 Mb in *C. matritensis* (59 scaffolds with 519 gaps and an N50 of 36.1 Mb), and include 17 pseudo-chromosomes for both species (Supplementary Data 1). The BUSCO completeness is 93.6% of single-copy and 2.6% of duplicated genes in *N. najaformis*, and 92.9% of single-copy and 3.2% of duplicated genes in *C. matritensis*.

We examined the macrosynteny relationships of orthologous genes across eleven chromosome-level annelid genomes, including the newly generated ones for the hormogastrid earthworms (Table 1). 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, *Enchytraeus crypticus*), 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 indicating that ALGs and ClitLGs do not show any of the algebraic relationships described in[3] (i.e., ClitLGs (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[6,7,9]. 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 1,139 contigs (Supplementary Table 2), which also suggests genome-wide chromosome mixing.

To test whether the observed rearrangements could be the result of differential gene loss after whole-genome duplication (WGD), we inferred the presence of such events through the analysis of synonymous substitutions (Ks) taking into account synteny and gene tree - species tree reconciliation methods. None of the methods supported a WGD event at Clitellata (Fig. 2a; Supplementary Data 2), therefore discarding that differential gene loss accounts for the observed genome structure. In earthworms, both methods recovered evidence of WGD. Within clitellates, gene tree-species tree reconciliation methods clearly supported two rounds of WGD at the node of Crassiclitellata (earthworms) for all earthworm species (Fig. 2a). On the contrary, the methods based on Ks using synteny blocks as anchors provided conflicting results, pointing to different numbers of potential WGD events depending on the species analysed (between one and two events in earthworms) and, more importantly, the node in which they happened varied largely, ranging from species-specific WGD to WGD events in nodes where different earthworm families diverged (Supplementary Data 2). We further explored the stoichiometry between ClitLGs and CrassiLGs, as under a scenario of two rounds of WGD we would expect to observe a 1:4 relationship between them. While observing a higher density of 1:2 and 1:4 relationships, other relationships such as 1:3, 1:5 and 1:6 could also be observed (Fig. 1c). We argue that while there is evidence pointing to one or several rounds of WGD, the fast dynamics of macrosyntenic changes within clitellates (potentially linked to the loss of genes in genome stability, as discussed below) make their robust inference challenging.

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

Despite the lack of availability of genome assemblies for several clitellate lineages, high-quality transcriptomes of all main lineages are available[17–19], enabling 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[20] 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 genes lost at that branch. Lost genes were largely enriched in functions related to cell division, DNA replication and DNA repair. Examples include several complexes such as Slx1-Slx4, GINS, MutSalpha or SHREC. The Slx1-Slx4 is 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[21] (Fig. 2d). The GINS complex is 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[22]. The MutSalpha complex plays a crucial role in DNA mismatch repair by recognizing mismatches and recruiting strand-specific nucleases to remove mispaired bases from daughter strands[23]. Finally, the SHREC complex regulates nucleosome positioning to assemble higher-order chromatin structures critical for heterochromatin functions[24] (Fig. 2d; Supplementary Data 3, 4; Supplementary Table 3). 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, several putative WGD were observed in earthworms as described above (in addition to the recent WGD event in the family Megascolecidae, Fig. 2b; Supplementary Data 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. Our results therefore may suggest that the presence of relaxed selective constraints (in this case, resulting in a less efficient DNA repair mechanisms) can facilitate the occurrence and fixation of genome reshuffling, as recently proposed in rodents[25].

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

In order to explore if clitellate genomes show a different transposable element (TE) blueprint compared to marine annelids and its potential relationship with the observed rearrangements, we investigated TE organisation and evolutionary dynamics across the 11 annelid species. Leech genomes contained a smaller percentage of TEs compared to other clitellates and marine annelids, whereas enchytraeids and earthworms had percentages similar to those of marine annelids (Fig. 3a). TE landscapes differed considerably in clitellates and marine annelids (Fig. 3a,b; Supplementary Table 4). Enchytraeids and earthworms exhibited similar profiles, characterised by the expansion of several TE superfamilies including DNA/hAT-Charlie and a large-scale expansion of LINE/L2, the latter previously described for earthworms[26] (Fig. 3b). TE superfamily composition differed considerably in leeches compared to marine annelids and the other clitellates, with the most prominent expansions being DNA/hAT-Charlie, DNA/hAT-AC, DNA/hAT-Tip100 and LINE/CRE. We found one leech-exclusive TE superfamily (LINE/Dong-R4) that accounted for ca. 5% of genome coverage (Fig, 3b). Remarkably, one leech species (*Whitmania pigra*) showed a rather dissimilar pattern compared to the two other species included in this study (*Hirudo nipponia* and *Hirudinaria manillensis)*(Fig. 3b). We investigated whether synteny breakpoints were disproportionately associated with specific repetitive sequences, but due to the massive genome scrambling we failed to detect any significant signal indicating a higher or lower presence of specific TE superfamilies than expected from random iterations. Therefore, the role of TEs in the extreme genome scrambling observed in clitellates remains unclear.

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 ensuring faithful chromosomal segregation during cell division. To investigate whether the composition of

centromeres, which are known to be unique in all annelids including clitellates (i.e., they are monocentric), changed after this period of massive rearrangement, we analysed the content of repetitive elements across all annelids in our datasets. In particular, we inferred the transposable element (TE) landscape as well as that of satellite DNA, since they both constitute centromeres[27,28]. Among the more than 100 TE superfamilies explored, only two were exclusively present in clitellates and absent in marine annelids and outgroups: the CMC-Chapaev-3 superfamily, belonging to DNA TEs, and the CRE superfamily, belonging to LINE TEs (Fig. 3b; Supplementary Table 4). Notably, these TEs also happened to be significantly enriched in the centromeric areas in earthworms, i.e., with a significantly higher density in 85% of the centromeric regions in all chromosomes (Fisher's exact test, p-value < 0.05; Fig. 3c). TEs from the CMC-Chapaev-3 superfamily were inferred as only present in clitellates and absent in marine annelids and leeches, and TEs from the CRE superfamily were present in enchytraeids and earthworms and absent in leeches, marine annelids and outgroups (Fig. 3b, Supplementary Table 4). 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 the way they did in marine annelids, which may be indicative as well of neocentromere formation in clitellates[29]. 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 clitellate-exclusive 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 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. 3d; Supplementary Data 5). Unlike earthworms, leech centromeres did not show enrichment of any specific TE superfamily, suggesting that centromere evolution in leeches may follow a different mechanism, potentially relying on other genomic elements or structural factors for centromere function and stability.

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

A detailed comparative analysis of the genome architecture in representative annelid species (two marine annelids, *Terebella lapidaria* and *Paraescarpia echinospica*; an earthworm, *Norana najaformis* and a leech, *Hirudinaria manillensis*) revealed distinct patterns of genome-wide chromosomal interactions, both at the chromosomal level and at the sub-Megabase pair (sub-Mbp) scale (Fig. 4a-c). At the chromosomal level, annelids showed clustering of centromeres (Fig. 4a-c, center), mirroring previous observations in invertebrates from other phyla (e.g., mosquitos[30]) and marsupials[10]. At the sub-Mbp scale, annelids lacked clear A/B compartments and exhibited compartmentalised topologically associated domains (TADs) defined by low insulator scores (Fig. 4a-c, right). Notably, annelids exhibit lower first eigenvector values than both chickens and flies (Fig. 5c), indicating that annelids may have their chromosomal regions interacting differently, potentially reflecting unique genome architecture or regulatory mechanisms in this animal phyla. While we cannot be entirely certain of their absence, we suggest that both A/B compartments and TADs are attenuated in annelids.

To assess the robustness of the detected TADs, we recalculated them using TADbit[31], which provides a score from 0 to 10, with 10 indicating very robust TADs (see Methods). Our results show that although at least half of the genome in annelids were compacted with strong boundary scores (>7), a significant proportion of TADs had weaker scores (<7)(Fig. 5a; Supplementary Data 6). Additionally, when creating an aggregated TAD plot, we observed a clear decrease in interactions at TAD boundaries (Fig. 5b). Interestingly, this depletion of interactions is only enriched at the boundaries, suggesting a less organised intra-chromosomal structure that favours interactions between domains. Altogether, these results suggest that while annelids do exhibit TADs, their boundaries may be less well-defined, leading to a more dynamic or permissive interaction landscape within the genome. This could reflect a difference in chromosomal organisation compared to vertebrate genomes, potentially linked to unique aspects of annelid biology.

Remarkably, annelids showed higher inter/intra-interaction ratios per chromosome (Fig. 5d) 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. 5e), suggesting that annelids are characterised by ‘floppy’ and relaxed chromosomes (i.e., lower intra-chromosomal interactions than inter-chromosomal ones). Additionally, annelids were heavily enriched in inter-chromosomal interactions compared to model organisms (chicken and fruit fly), indicating a significant level of genomic interactions between different chromosomes (Fig. 5d).

When comparing marine annelids and clitellates, chromosome scrambling resulted in a shift in 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 of intra-chromosomal ones, a reflection of more compacted chromosomes. This pattern was maintained when comparing ancestral linkage groups (ALGs) (Fig. 5f; Supplementary Data 7).

An exception to this pattern was the case of *Hox* genes: they showed a canonical clustered organisation in marine annelids (Fig. 6a), but in clitellates (at least in earthworms) they were both duplicated and highly rearranged in different chromosomes (Fig. 6a,b; Supplementary Data 8). However, *Hox* genes from the same ancestral cluster maintained clear long-range interactions both between two separated regions of the same chromosome and between chromosomes, indicating that they interact at the 3D level despite not being physically placed in a gene cluster (Fig. 6b,d,e). In leeches, however, many *Hox* genes were lost and inter-chromosomal interactions between *Hox* genes could not be detected, indicating a deeper reshaping of the *Hox* gene repertoire (Fig. 6a).

Genes potentially involved in adaptation to life on land and freshwater environments emerged through chromosome scrambling

Even though such a genome-wide pattern of scrambling will be expected to largely impact non-genic regions, it can give rise to new genetic elements with a potential adaptive role in two ways: by relocation of genes to a different genomic region, or by de novo appearance by splitting and merging gene fragments in a new order. A gene's position change in the genome can impact its regulation, expression, and interactions, potentially boosting adaptability to new environments (e.g.[32]). This flexibility may drive evolutionary adaptation by altering gene expression, creating new functions, and avoiding harmful effects. On the other hand, chromosome scrambling can give rise to chimeric genes resulting from the fusion of older fragments that may end up acquiring coding potential (e.g.[33,34]), allowing them to create novel proteins or regulatory functions that can help organisms adapt to new environmental challenges.

To test whether chromosomal rearrangements played a role in the adaptation of clitellates to freshwater and terrestrial environments, we first explored the evolutionary origin of all hierarchical orthologous groups (HOGs) through phylostratigraphy and categorised them in

three groups: (i) arising in the branches before the genome mixing (i.e., those comprised between the branch leading to clitellates and the root of the tree, and comprising genes from marine annelids and clitellates), which represent genes that were physically relocated during the chromosome scrambling and were therefore potentially subject to shifts in gene expression, selection, regulation or interaction with other genetic elements (hereafter 'before' HOGs); (ii) arising in the branch leading to Clitellata, and therefore coinciding with the genome-wide chromosome scrambling observed at that branch and comprising only clitellate sequences (hereafter referred to as 'in' HOGs); and (iii) arising after the genome scrambling, composed as well only of clitellate sequences (hereafter 'after' HOGs) (Table 2). To test whether chromosomal rearrangements resulted in non-neutral evolution in the genes that were relocated during these events (i.e., those found in 'before' HOGs), we examined the selection pressures acting on these genes. For that, we used Pelican[36], which detects shifts in the direction of selection at specific amino acid positions by analysing amino acid profiles, offering an enhanced approach to understanding adaptive changes at the molecular level. Our analysis revealed that 79.6% (8,108 of 10,188; Supplementary Table 5) of the orthogroups containing genes that underwent chromosomal rearrangements in clitellates exhibited significant evidence of shifts in directional selection. This unexpectedly high proportion sharply contrasts with what is predicted under neutral evolutionary models, where chromosomal rearrangements are generally expected to be selectively neutral. Instead, this finding suggests that the chromosomal relocations were adaptive and played a central role in the clitellates' transition to freshwater and terrestrial environments.

In order to further test whether genome-wide chromosomal scrambling facilitated adaptation to freshwater and terrestrial environments, we first conducted a series of experiments aimed at identifying putative adaptive genes by analysing differential gene expression in response to abiotic conditions characteristic of freshwater and terrestrial environments in leeches (*Hirudo medicinalis*) and earthworms (*Eisenia andrei*), respectively. (Fig. 7a; see Methods; summary statistics, density plots, volcano plots, expression heatmaps and matrices of expression values for all differentially expressed genes are available as Supplementary Data 9, 10 and Supplementary Tables Supplementary Tables 6, 7, 8). The number of differentially expressed coding genes (DEGs hereafter) was very similar in both species (Table 2). We categorised DEG-harbouring HOGs containing DEGs in the same three groups as explained before, based on their phylostratigraphic origin (i.e., 'before' HOGs, 'in' HOGs and 'after' HOGs; Fig. 7a). For both species, approximately half of HOGs containing DEGs arose before the origin of Clitellata,

while the other half arose after, being therefore lineage-specific (*E. andrei*: 'before' DEGs, 43.25%; 'after' DEGs, 52.59%; *H. medicinalis*: 'before' DEGs, 42.86%, 'after' DEGs, 52.04%)(Table 2, Supplementary Table 9). This shows that the gene repertoire involved in response to abiotic stress is composed of both 'ancient' genes (arising early in the evolutionary history of these organisms) and more recently evolved, lineage-specific genes. Only 5 and 12 HOGs in *H. medicinalis* and *E. andrei* respectively, arose in the branch leading to clitellates (Table 2). Remarkably, the number of DEG-harbouring HOGs in both species showed virtually no overlap (only 3 HOGs were shared; Supplementary Table 10), meaning that both species leverage a completely different gene repertoire to face environmental stress.

We next explored whether these newly emerged genes HOGs in Clitellata containing DEGs (i.e., 'in' DEGs) were chimeric, as expected under an scenario of chromosome shattering. Our analyses suggested that these HOGs resulted from fission of 'ancient' HOGs (ie, that preexisted the origin of clitellates). 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. 7b; Supplementary Table 11). 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. Only one HOG could be identified as originating de novo bona fide (i.e., not the result of fusion and fission from any existing HOGs), and was involved in sodium-potassium cell trafficking (Supplementary Table 11).

To test whether relocated DEGs resulted in new gene interactions, we examined the Hi-C interactions between DEGs in leeches and earthworms, comparing these with those of the orthologous genes in marine annelids. At the genome architecture level, DEGs in leeches and earthworms showed significantly more interactions between them than their orthologs in marine annelids, suggesting that their relocation after genome scrambling favoured new genetic interactions among them (Fig. 7c). In order to test a potential adaptive role in relocated DEGs, we next investigated whether DEGs that were relocated after chromosome scrambling (i.e, 'before' DEGs) were subjected to shifts in the regime of directional selection, since such shifts could indicate that these genes have undergone evolutionary changes to enhance fitness in response to new environmental pressures. In particular, we compared shifts in selection

regimes between clitellates and non-clitellates to assess whether relocated DEGs in clitellates experienced adaptive changes distinct from their non-clitellate counterparts. Regardless of the method used to detect directional selection (see Methods), a high percentage of DEG-containing HOGs showed significant shifts in directional selection regimes both in earthworm and leeches (69.6% and 53.6% in *E. andrei* and 54.8% and 45.3% in *H. medicinalis* with Pelican[36] and HyPhy[37], respectively)(Fig. 7d). Genes with significant shifts in directional selection in the leech were involved in response to all abiotic stresses tested (visible light, UV-B light, osmotic stress, hypoxia and hyperoxia). On the contrary, more than half of the genes with significant changes in directional selection regimes in the earthworm were differentially expressed in specimens recovering after exposure to UV-B light, and therefore putatively involved in UV-induced DNA damage repair (Supplementary Table 10). The main categories based on the Clusters of Orthologous Genes (COG) database[38] look different in both species, with the most represented categories in *E. andrei* being inorganic ion transport metabolism; posttranslational modification/protein turnover/chaperones; replication, recombination and repair; and lipid transport and metabolism, and the most represented ones in *H. medicinalis* being translation, ribosomal structure and biogenesis; coenzyme transport and metabolism; and transcription (Fig. 7d).

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 are otherwise present throughout metazoans, and did so in a comparatively short evolutionary window (measured not as absolute time but rather as phylogenetic 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. Similar findings have been reported in two pieces of work contemporary to ours[41,42], which strengthens the robustness of the results presented here.

Our results provide an example of a complete loss of macrosynteny genome-wide at the within-phylum level of a higher magnitude than previously seen in other animal phyla such as bryozoans[7], cephalopodes[8] or tunicates[9]. 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 mollusc, suggesting that clitellate genome evolution is not constraint by synteny. Clitellates encompass other lineages beyond earthworms, leeches and potworms, such as the families Naididae, Lumbriculidae and other early-splitting interstitial species[17], none of which 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[43–45], 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 phylogenetic distance between lineages). Even though no chromosome-level genome sequences 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. These include changes in their reproduction mode (marine annelids are mostly dioic while clitellates are hermaphrodite and some parthenogenetic) and frequent polyploidy (e.g., some earthworms can range between diploid (×2) to dodecaploid (×12) even at the within-species level[46,47]). Another key feature is the development of a new organ called clitellum, a “collar” that forms a reproductive cocoon during part of their life cycles. Additionally, clitellates have lost parapodia (lateral fleshy protrusions in marine annelids) and have adapted to freshwater and terrestrial environments, with the exception of some marine leeches from the family Piscicolidae[48–50]. Notably, clitellates are also characterised by common

aneuploidy and an apparent lack of canonical cell division. Pavlíček et al[51] reported common aneuploidy and Robertsonian translocations in multiple species of clitellates, including earthworms, leeches and species from the families Naididade, Lumbriculidae and Branchiobdellidae, 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. Since aneuploidy is strongly associated with chromosomal and genome rearrangements and is often recognised as a direct outcome of genome instability[52–54], 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 clitellates.

The emergence of genes potentially involved in adaptation to freshwater and terrestrial environments through chromosome scrambling highlights the dynamic nature of clitellate genome architecture in response to environmental pressures. Our findings that ca. 80% of the relocated genes ('before' HOGs and DEGs) were subject to significant shifts in directional selection provide compelling evidence that these chromosomal rearrangements were not neutral but adaptive. This suggests that the physical relocation of genes during chromosomal scrambling exposed them to new selective pressures, potentially reshaping their regulatory landscapes and optimising their functionality for the new ecological niches that clitellates encountered. Furthermore, due to their extensive nature and the high degree of selective pressure in genes relocated due to the massive chromosome scrambling, these genome-wide rearrangements may have also driven the cladogenesis of clitellates, contributing to their divergence from other annelid lineages.

Chromosome rearrangements not only facilitate the relocation of genes, which can potentially impact their regulatory landscapes and expression, but also contribute to the generation of chimeric genes, creating novel functional elements that can drive evolutionary innovation. The identification of both ‘ancient’ and lineage-specific DEGs suggests that adaptation to abiotic stress in Clitellata is supported by a mix of conserved and newly evolved mechanisms. The significant shifts in directional selection observed in a substantial proportion of relocated DEGs (i.e., ‘before’ DEGs) underscore the role of chromosomal scrambling in exposing these genes to new selective pressures, likely enhancing their adaptive potential in response to terrestrial and freshwater environments. These findings align with the hypothesis that chromosomal reshuffling

may act as a key evolutionary mechanism, potentially enabling organisms to refine their gene regulatory networks and functional repertoires in the face of changing environmental challenges.

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. Our results also suggest that genome organisation in three-dimensional space in annelids may be different to vertebrates and model organisms, since they seem to lack key structural units such as clear A/B compartments and clearly-defined TADs. This, together with the fact that genome evolution in clitellates may not be constrained by synteny, positions them as excellent models to further our understanding on genome evolution across the animal kingdom. A recent study found as well the lack of TADs and of syntenic constraints in freshwater and parasitic platyhelminthes[55], suggesting that our current knowledge on animal genome evolution is currently incomplete and that a further investigation of the genome architecture of lesser-studied invertebrates is likely to result in unexpected insights into the diversity and plasticity of genomic organisation.

While the timing of these genomic rearrangement remains unclear, we argue that the genomic hallmarks observed in clitellates (that exhibit the highest rearrangement indices across bilaterians[41]) 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[17]. Instead, we hypothesise that they 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 phylogenetic 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[56–58]. 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[57–59].

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[60].

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 catalysers of cladogenesis at a macroevolutionary scale in short evolutionary windows.

TABLES

Table 1. Species included in the macrosynteny analysis.

Species name	Phylum	Class	Order	Family	Habitat	Data type	Genome size (Mbp)	Haploid karyotype (n)	Reference
Owenia fusiformis Delle Chiaje, 1844	Annelida	Polychaeta	-	Oweniidae	Marine	Genome	500	12	Martín-Zamora et al., 2023
Sipunculus nudus Linnaeus, 1766	Annelida	Polychaeta	Sipuncula	Sipunculidae	Marine	Genome	1427	17	Zheng et al., 2023
Paraescarpia echinospica Southward, Schulze & Tunnicliffe, 2002	Annelida	Polychaeta	Sabellida	Siboglinidae	Marine	Genome	1102	14	Sun et al., 2021
Streblospio benedicti Webster, 1879	Annelida	Polychaeta	Spionida	Spionidae	Marine	Genome	701	11	Zakas et al., 2022
Enchytraeus crypticus Westheide & Graefe, 1992	Annelida	Clitellata	Enchytraeida	Enchytraeidae	Terrestrial	Genome	525	-	Amorim et al., 2021
Hirudo nipponia Whitman, 1886	Annelida	Clitellata	Arhynchobdellida	Hirudinidae	Freshwater	Genome	204	11	Zheng et al., 2023
Hirudinaria manillensis (Lesson, 1842)	Annelida	Clitellata	Arhynchobdellida	Hirudinidae	Freshwater	Genome	157	13	Zheng et al., 2023
Whitmania pigra (Whitman, 1844)	Annelida	Clitellata	Arhynchobdellida	Haemopidae	Freshwater	Genome	181	11	Zheng et al., 2023
Norana najaformis Marchán, Fernández, Díaz Cosín & Novo, 2018	Annelida	Clitellata	Crassiclitellata	Hormogastridae	Terrestrial	Genome	608	17	This project
Carpetania matritensis Marchán, Fernández, Díaz Cosín & Novo, 2018	Annelida	Clitellata	Crassiclitellata	Hormogastridae	Terrestrial	Genome	588	17	This project
Eisenia andrei Bouché, 1972	Annelida	Clitellata	Crassiclitellata	Lumbricidae	Terrestrial	Genome	1315	11	Shao et al., 2020
Metaphire vulgaris (Chen, 1930)	Annelida	Clitellata	Crassiclitellata	Megascolecidae	Terrestrial	Genome	729	41	Jin et al., 2020
Pecten maximus (Linnaeus, 1758)	Mollusca	Bivalvia	Pectinida	Pectinidae	Marine	Genome	918	19	Kenny et al., 2020
Lineus longissimus (Gunnerus, 1770)	Nemertea	Pilidiophora	Heteronemertea	Lineidae	Marine	Genome	391	19	Kwiatkowski et al., 2021

Table 2. Metrics of Hierarchical OrthoGroups (HOGs) and differentially expressed genes (DEGs). Top, total number of HOGs inferred in the gene repertoire evolutionary dynamics analysis. Number of HOGs that arose before, in and after Clitellata are indicated. The number of HOGs with a balanced taxonomic representation (i.e., HOGs containing genes from >20% of marine annelid and 20% of clitellate species) is shown for the 'before' phylostratigraphic category only, as they were subjected to further analysis to test directional selection. Bottom, total number of DEGs inferred for *E. andrei* and *H. manillensis* under each of the stress experiments performed. Number of coding DEGs, of DEGs assigned to HOGs and the phylostratigraphic category of these HOGs ('before', 'in, 'after') is also indicated.

	Before	In	After
Total no. HOGs	18454	1645	23785
No. HOGs balanced taxonomic representation	10188		

Experiment-Regulation	DGEs	Coding DGEs	Species-specific	Assigned to HOGs	TOTAL HOGs	Assigned to HOGs 'in'	Assigned to HOGs 'before'	Assigned to HOGs 'after'
Eisenia andrei								
Hyperoxia-UP	32	32	11	21	21	3	9	9
Hypoxia-UP	6	6	2	4	4	0	0	4
Osmo-UP	45	45	8	37	35	1	22	14
UV24D-UP	146	146	26	120	113	5	50	65
VL-UP	10	10	4	6	5	0	1	5
Hyperoxia-Down	78	78	22	56	53	1	21	34
Hypoxia-Down	22	22	5	17	15	1	6	10
Osmo-Down	34	34	12	22	20	2	13	7
UV24D-Down	93	93	33	60	58	0	33	27
VL-Down	15	15	7	8	8	0	2	6
TOTAL HOGs					332	13	157	181
Unique HOGs					289	12	125	152
% HOGs						4.15	43.25	52.60
Hirudinaria manillensis								
Hyperoxia-UP	141	27	20	7	7	0	4	3
Hypoxia-UP	188	23	17	6	6	1	2	3
Osmo-UP	245	59	47	12	12	1	5	6
UV24D-UP	189	43	27	16	16	0	9	7
VL-UP	85	12	6	6	6	0	3	3
Hyperoxia-Down	175	58	25	33	30	2	13	18
Hypoxia-Down	247	65	39	26	23	0	10	16
Osmo-Down	174	42	26	16	16	0	6	10
UV24D-Down	187	48	28	20	18	0	9	11
VL-Down	124	28	21	7	7	1	3	3
TOTAL HOGs					141	5	64	80
Unique HOGs					98	5	42	51
% HOGs						5.10	42.86	52.04

FIGURES

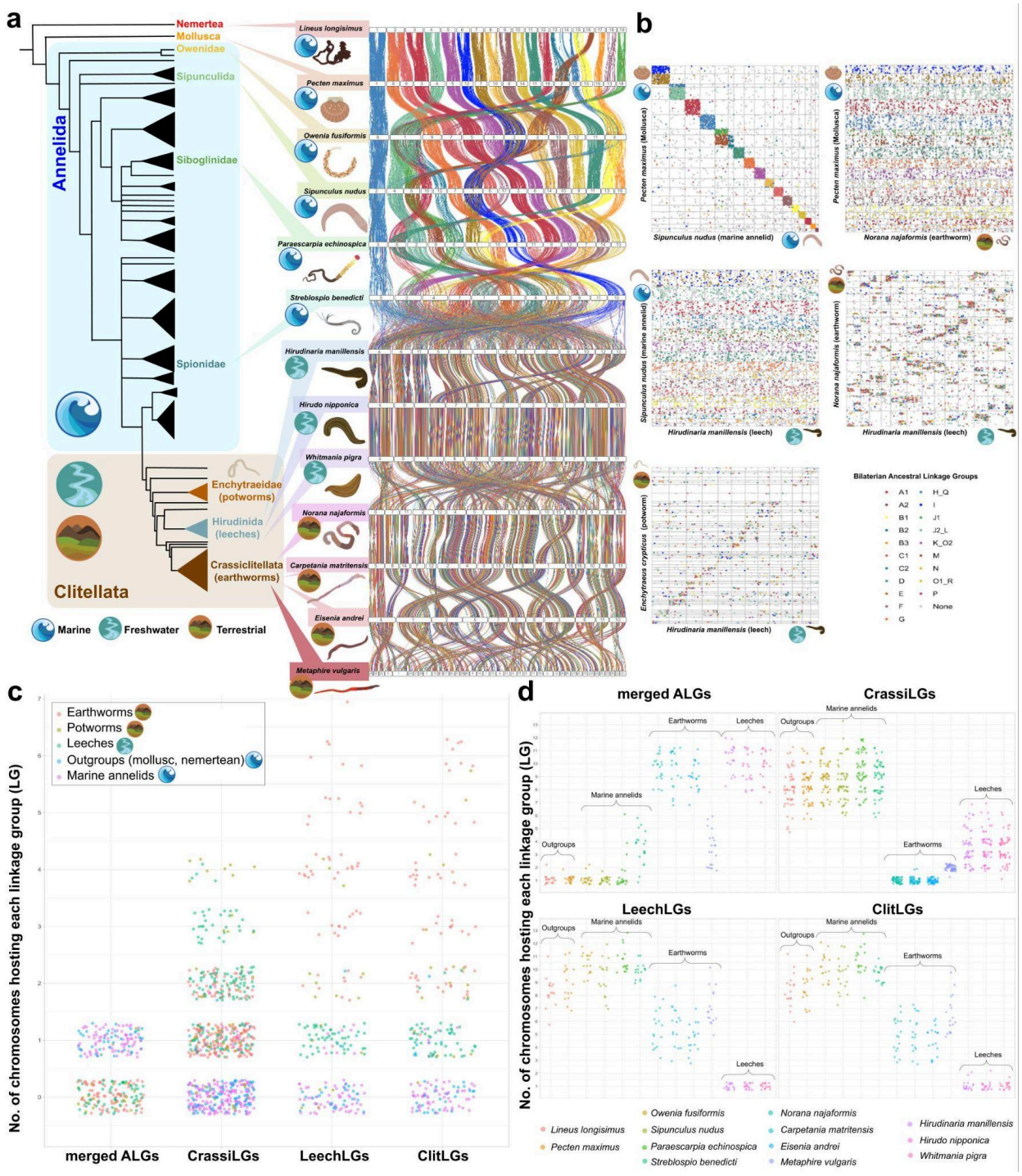

Figure 1. Macrosyntenic evolution of clitellates. 1a. Left, Annelida Tree of Life showing all main lineages and the position of the species included in the present study, based on the topology reported in Capa et al. (Annelida)[61] and Erseus et al. (Clitellata)[17]. Centre, ribbon plot of the chromosome-scale ancestral gene linkage across annelids, with a mollusc 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[3]. **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 mollusc (*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_v2.pdf]*

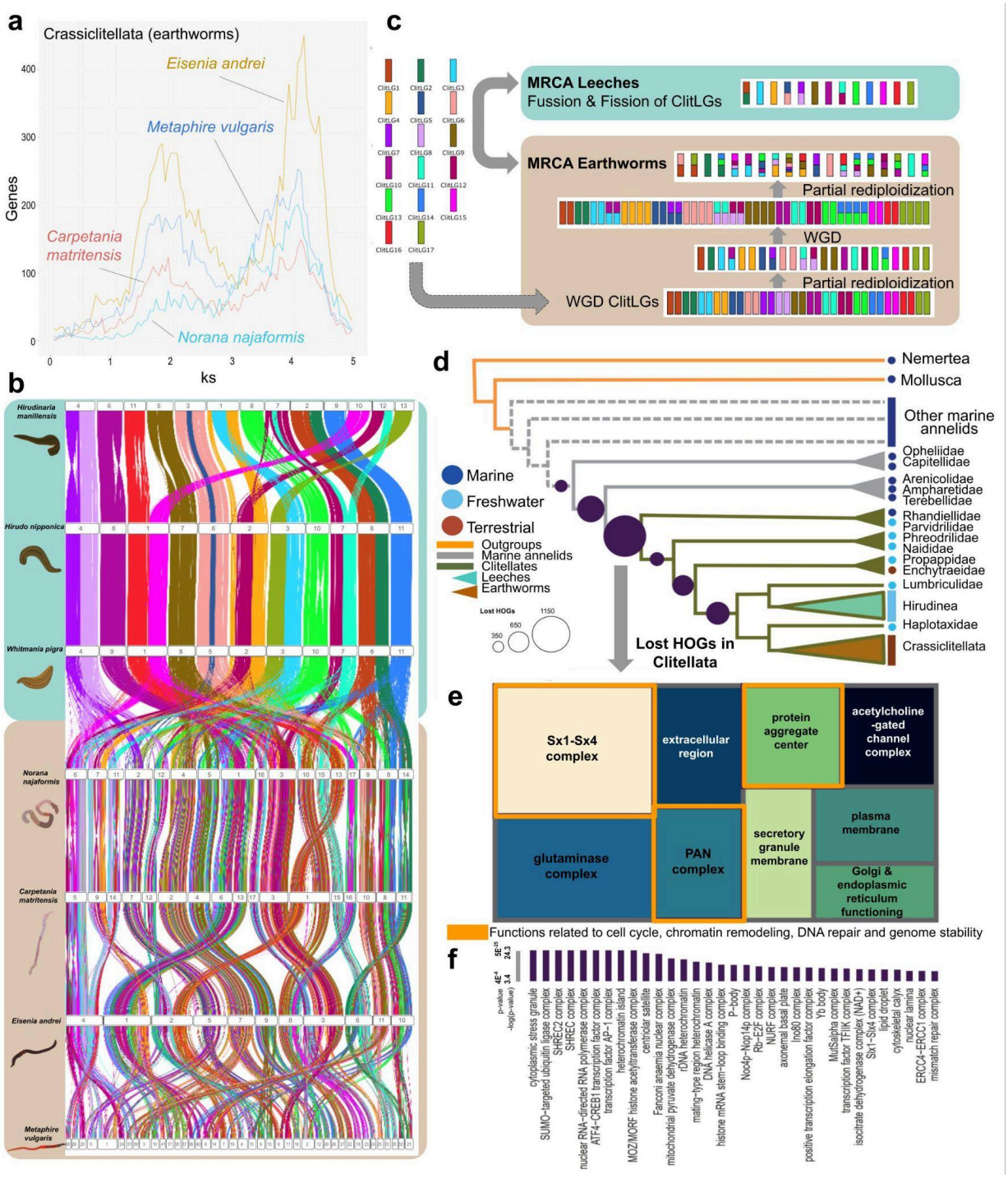

Figure 2. Frequent rare large-scale genomic changes among clitellates after genome-wide chromosome scrambling. a. Substitution-rate-adjusted mixed paralog–ortholog Ks plot for the node of Crassiclitelata (earthworms). The inferred two putative WGD events are indicated by the two prominent peaks corresponding to the Ks-based WGD age estimates. **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. Taxon families are shown at the tips (see also Supplementary Table 12). **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 4 and the manuscript's GitHub repository). *[High-quality figure available here: Panel_Fig_2_v2.pdf]*

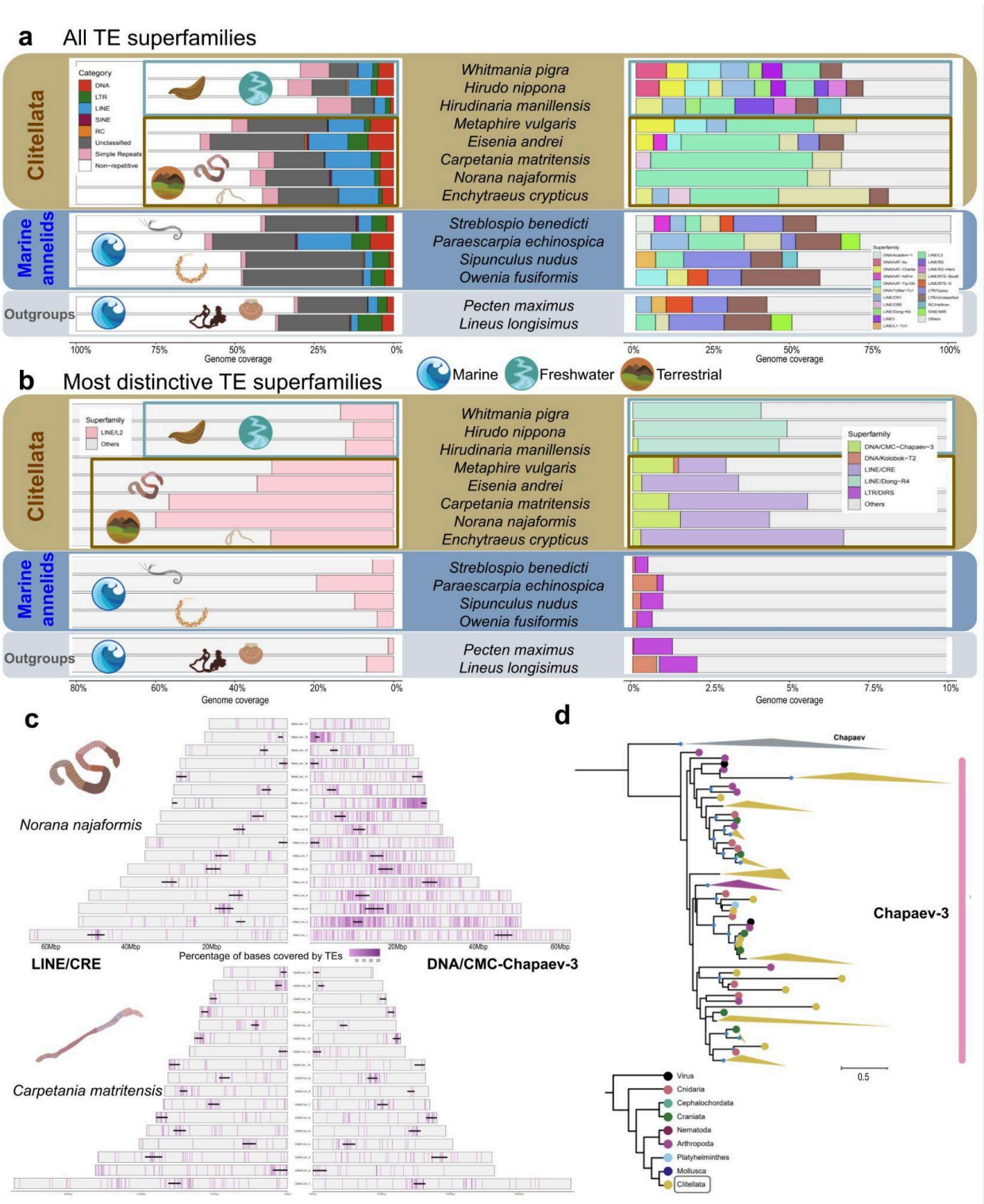

Figure 3. Transposable element landscape and centromere composition in clitellates. 3a. Left, percentage of the genome of each species covered by transposable elements, divided in main categories. Right, transposable element superfamily distribution per species. The most

abundant superfamilies are shown. **3b.** Left, genome coverage by the L2 superfamily. Right, genome coverage of the most qualitatively distinctive transposable element superfamilies (i.e., pattern of presence/absence in the different lineages). **3c.** Distribution of CMC-Chapaev-3 and CRE transposable element insertions in the genomes of *N. najaformis* (top) and *C. matritensis* (bottom). The genome was divided in bins of 50kb and the percentage of bases covered by members of the CMC-Chapaev-3 (right) and CRE (left) superfamilies 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). **4d.** Maximum likelihood phylogenetic tree of CMC-Chapaev-3 transposases. Small blue dots at nodes 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_v2.pdf].*

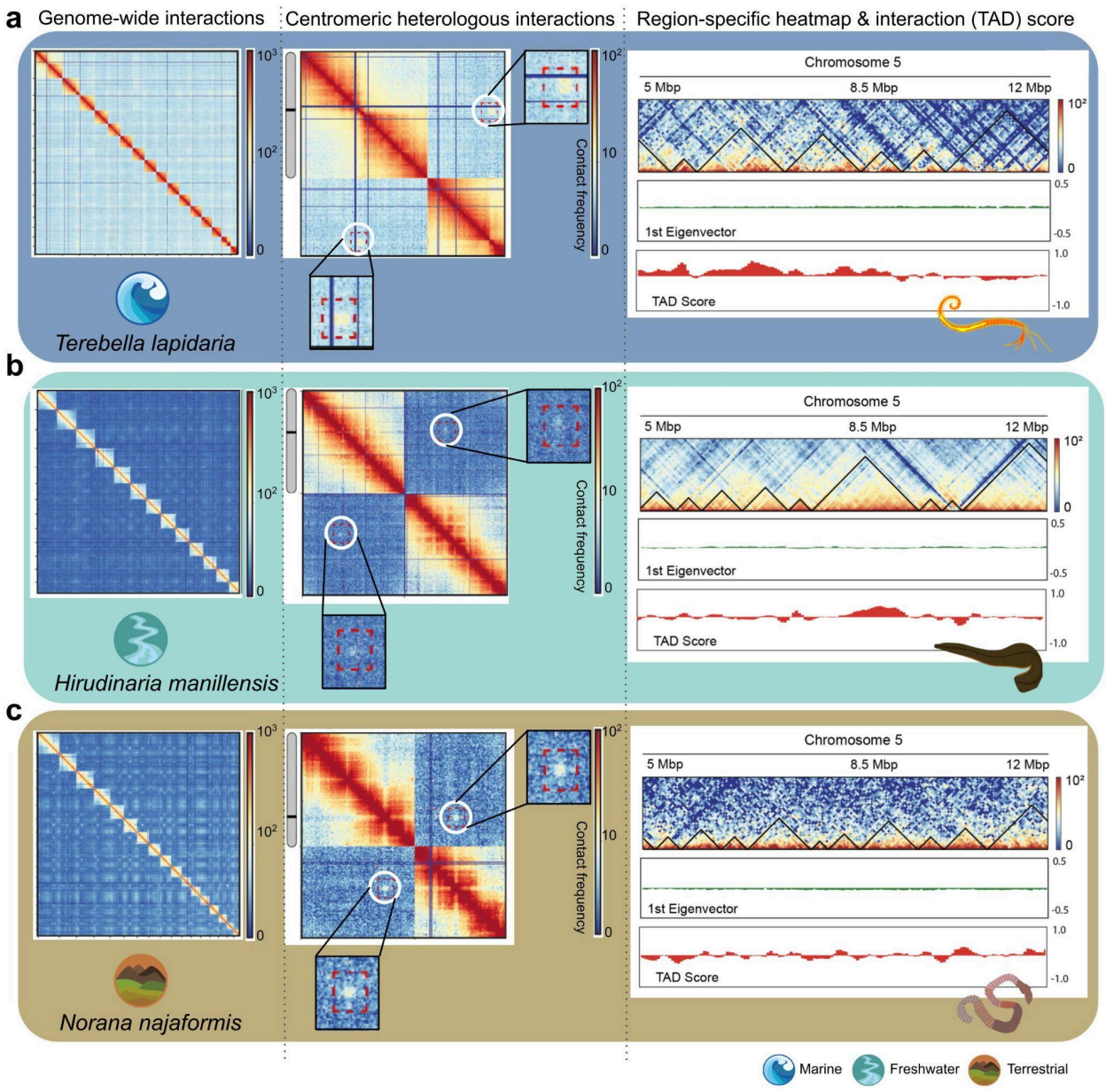

Figure 4. Genome architecture organisation in marine annelids and clitellates. 4a-c. Left, whole-genome Hi-C contact maps for *Terebella lapidaria* (**4a**), *Hirudinaria manillensis* (**4b**) and *Norana najaformis.* (**4c**). Center, Hi-C contact maps representing a pair of chromosomes/scaffolds depicting centromeric heterologous interactions in all three species. Right, Chromosome 1 region-specific 500 Kbp heatmaps and insulator (TAD) score for the same three species. *[High-quality figure available here:* Panel_Fig_4_v2.pdf *]*

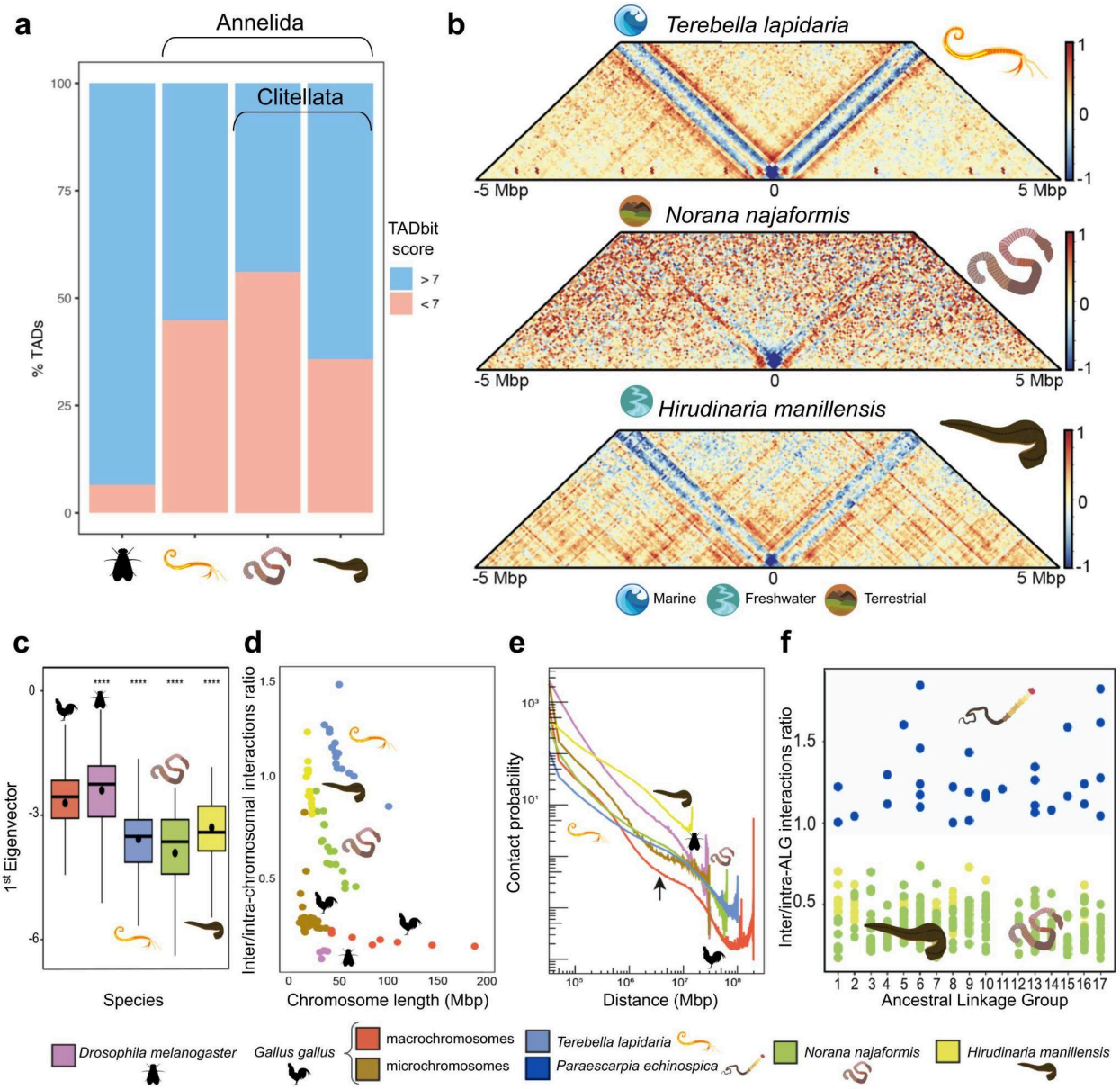

Figure 5. Annelid TAD organisation and inter-intra-chromosomal interactions. 5a. Genome-wide TAD scores per species as inferred by TADbit[31]. **5b.** Aggregated TAD plot of annelid species, showing a decrease in interactions at TAD boundaries. **5c.** Statistical analysis of the first eigenvector in the species analysed, showing that they are significantly lower in annelids compared to model organisms (two-sided t-test, ****p<0.001). **5d.** Inter-/intra-chromosomal interactions according to chromosome length (in Mbp) in the same three annelid species, together with chicken (*Gallus gallus*[62]) and the fruit fly (*Drosophila melanogaster*[63]). **5e.** Chromosome-specific contact probability P(s) as a function of genomic distance in *T. lapidaria*, *N. najaformis*, *H. manillensis*, chicken and the fruit fly. **5f.** Inter-/intra-chromosomal interactions according to ancestral linkage groups (ALG) for the annelids *Paraescarpia echinospica*, *N. najaformis* and *H. manillensis*. *[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][High-quality figure available here: Panel_Fig_5_v2.pdf].

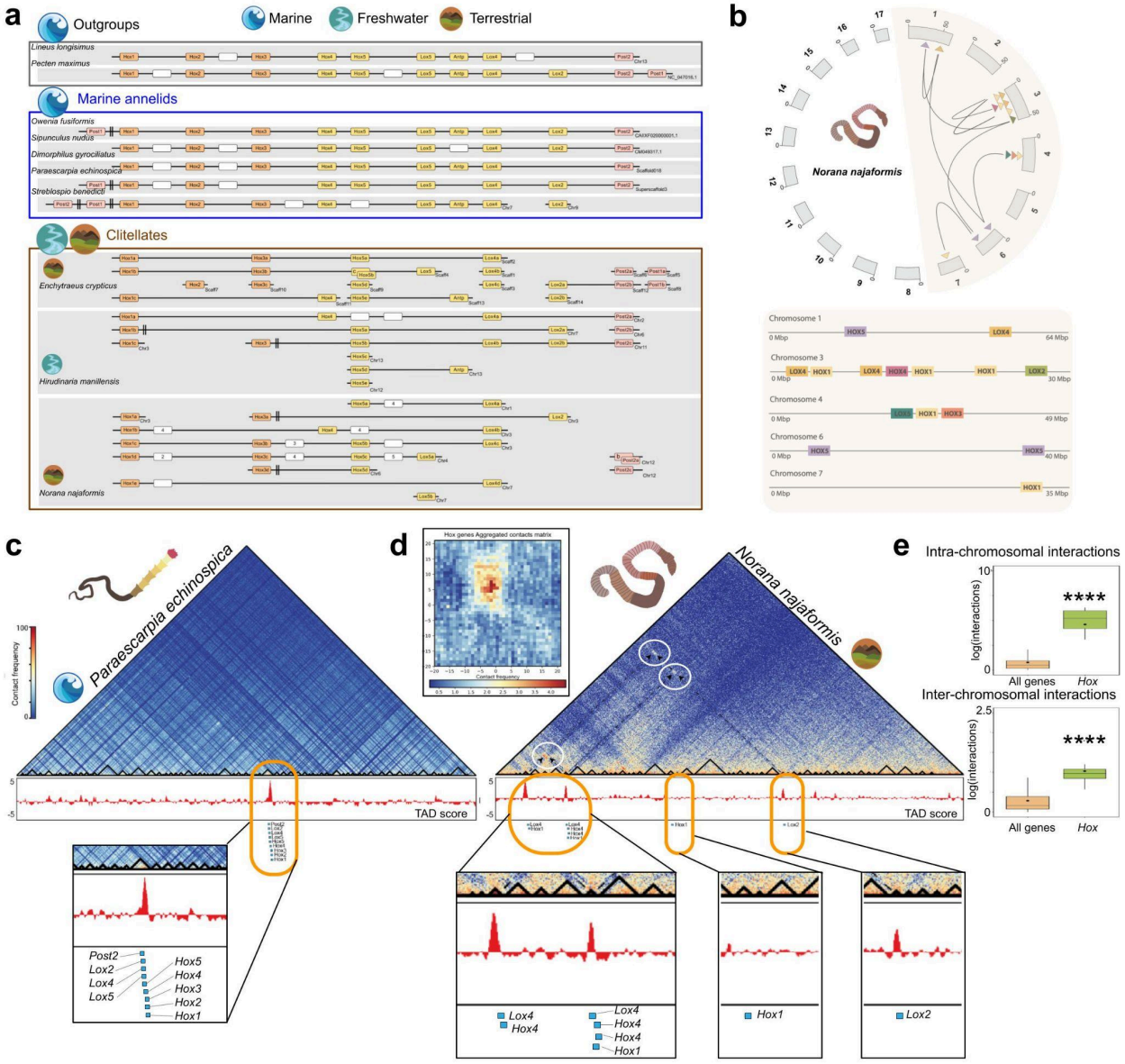

Figure 6. *Hox* gene repertoire and 3D interactions in annelids. 5a. Representation of the main *Hox* cluster across annelids and outgroups. Rearrangements are primarily observed in clitellates. Rectangular boxes represent *Hox* genes, while chromosomes or scaffolds (in the case of the draft genome of 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 empty rectangles, 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. **5b.** Circos plot depicting the long-range interactions involving the *Hox* genes in *N.*

najaformis. **5c,d.** Chromosome-wide Hi-C map of *P. echinospica* (**5c**) and *N. najaformis* (**5d**), where clear long-range intra-chromosomal interactions between *Hox* genes can be seen in *N. najaformis* (highlighted with white circles). *Hox* genes aggregated contact matrix is shown for *N. najaformis*. **5e.** Box-plot reflecting *Hox* genes inter-chromosomal and intra-chromosomal interactions using genome-wide interactions as reference (two-sided t-test, ****p<0.001). *[High-quality figure available here: Panel_Fig_6_v2.pdf].*

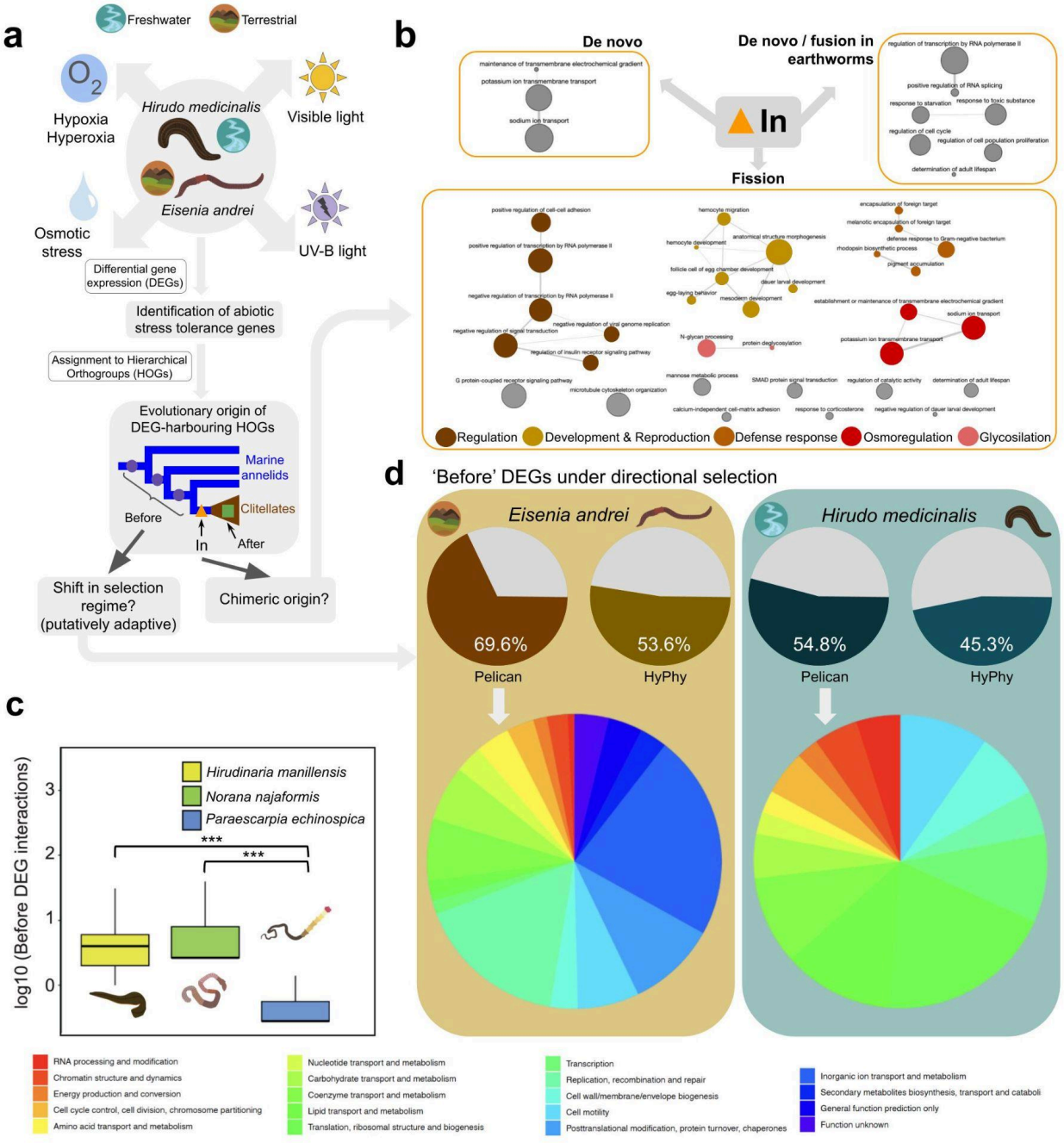

Figure 7. New genes and gene interactions involved in adaptation to freshwater and terrestrial environments arose as a consequence of genome-wide chromosome scrambling. 7a. Schematic representation of the experimental design for the investigation of differential gene expression on *H. medicinalis* (leech) and *E. andrei* (earthworm) under abiotic stress conditions. **7b.** Putative enriched function of genes arising during chromosome scrambling (biological process). Coloured networks represent clusters of functions related ontologically; a general biological process is provided for these. **7c.** 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, $p < 0.001$). **7d.** Percentage of genes under significant shifts in directional selection for both species. Results for both methods tested are shown (Pelican and HyPhy, see Methods). Main categories of DEGs significantly under directional selection in *E. andrei* and *H. medicinalis*. *[High-quality figure available here: Panel_Fig_7_v2.pdf).*

References (Main text and figures)

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. *et al.* Reconstruction of proto-vertebrate, proto-cyclostome and proto-gnathostome genomes provides new insights into early vertebrate evolution. *Nat. Commun.* **12**, 4489 (2021).
3. Simakov, O. *et al.* Deeply conserved synteny and the evolution of metazoan chromosomes. *Sci Adv* **8**, eabi5884 (2022).
4. Marlétaz, F. *et al.* The hagfish genome and the evolution of vertebrates. *Nature* **627**, 811–820 (2024).
5. Yu, D. *et al.* Hagfish genome elucidates vertebrate whole-genome duplication events and their evolutionary consequences. *Nat Ecol Evol* **8**, 519–535 (2024).
6. 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.
7. 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.
8. Albertin, C. B. *et al.* Genome and transcriptome mechanisms driving cephalopod evolution. *Nat. Commun.* **13**, 1–14 (2022).
9. Plessy, C. *et al.* Extreme genome scrambling in marine planktonic *Oikopleura dioica* cryptic

species. *Genome Res.* **34**, 426–440 (2024).

10. Álvarez-González, L. *et al.* Principles of 3D chromosome folding and evolutionary genome reshuffling in mammals. *Cell Rep.* **41**, 111839 (2022).
11. Waters, P. D. *et al.* Microchromosomes are building blocks of bird, reptile, and mammal chromosomes. *Proc. Natl. Acad. Sci. U. S. A.* **118**, (2021).
12. Damas, J. *et al.* Evolution of the ancestral mammalian karyotype and syntenic regions. *Proc. Natl. Acad. Sci. U. S. A.* **119**, e2209139119 (2022).
13. 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).
14. 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).
15. 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.
16. Backus, B. T. Some bryozoan karyotypes and chromosome numbers. *Genetical Res.* **29** (3): 187-191 (1977).
17. Erséus, C. *et al.* Phylogenomic analyses reveal a Palaeozoic radiation and support a freshwater origin for clitellate annelids. *Zool. Scr.* **49**, 614–640 (2020).
18. Andrade, S. C. S. *et al.* Articulating “archiannelids”: Phylogenomics and annelid relationships, with emphasis on meiofaunal taxa. *Mol. Biol. Evol.* **32**, 2860–2875 (2015).
19. Struck, T. H. *et al.* Phylogenomic analyses unravel annelid evolution. *Nature* **471**, 95–98 (2011).
20. Ly, P. & Cleveland, D. W. Rebuilding chromosomes after catastrophe: Emerging

mechanisms of chromothripsis. *Trends Cell Biol.* **27**, 917–930 (2017).

21. Xu, X. *et al.* Structure specific DNA recognition by the SLX1-SLX4 endonuclease complex. *Nucleic Acids Res.* **49**, 7740–7752 (2021).
22. Labib, K. & Gambus, A. A key role for the GINS complex at DNA replication forks. *Trends Cell Biol.* **17**, 271–278 (2007).
23. 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).
24. Sugiyama, T. *et al.* SHREC, an effector complex for heterochromatic transcriptional silencing. *Cell* **128**, 491–504 (2007).
25. Álvarez-González, L. *et al.* 3D chromatin remodelling in the germ line modulates genome evolutionary plasticity. *Nat. Commun.* **13**, 2608 (2022).
26. Shao, Y. *et al.* Genome and single-cell RNA-sequencing of the earthworm *Eisenia andrei* identifies cellular mechanisms underlying regeneration. *Nat. Commun.* **11**, 1–15 (2020).
27. Malik, H. S. & Henikoff, S. Major evolutionary transitions in centromere complexity. *Cell* **138**, 1067–1082 (2009).
28. Malik, H. S. & Henikoff, S. Conflict begets complexity: the evolution of centromeres. *Curr. Opin. Genet. Dev.* **12**, 711–718 (2002).
29. Henikoff, S., Ahmad, K. & Malik, H. S. The centromere paradox: stable inheritance with rapidly evolving DNA. *Science* **293**, 1098–1102 (2001).
30. 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).
31. Serra, F. *et al.* Automatic analysis and 3D-modelling of Hi-C data using TADbit reveals structural features of the fly chromatin colors. *PLoS Comput. Biol.* **13**, e1005665 (2017).
32. Cheng, L. *et al.* Large-scale genomic rearrangements boost SCRaMbLE in Saccharomyces cerevisiae. *Nat. Commun.* **15**, 770 (2024).

33. Landweber, L. F., Kuo, T.-C. & Curtis, E. A. Evolution and assembly of an extremely scrambled gene. *Proceedings of the National Academy of Sciences* **97**, 3298–3303 (2000).
34. Kawashima, T. *et al.* Domain shuffling and the evolution of vertebrates. *Genome Res.* **19**, 1393–1403 (2009).
35. Bhutkar, A., Russo, S. M., Smith, T. F. & Gelbart, W. M. Genome-scale analysis of positionally relocated genes. *Genome Res.* **17**, 1880–1887 (2007).
36. Duchemin, L., Lanore, V., Veber, P. & Boussau, B. Evaluation of methods to detect shifts in directional selection at the genome scale. *Mol. Biol. Evol.* **40**, (2023).
37. Kosakovsky Pond, S. L. *et al.* HyPhy 2.5-A customizable platform for evolutionary hypothesis testing using phylogenies. *Mol. Biol. Evol.* **37**, 295–299 (2020).
38. Tatusov, R. L., Koonin, E. V. & Lipman, D. J. A genomic perspective on protein families. *Science* **278**, 631–637 (1997).
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. Lewin, T. D., Liao, I. J.-Y. & Luo, Y.-J. Annelid comparative genomics and the evolution of massive lineage-specific genome rearrangement in bilaterians. *Mol. Biol. Evol.* **41**, (2024).
42. Schultz, D. T. *et al.* Acceleration of genome rearrangement in clitellate annelids. *bioRxiv* (2024) doi:10.1101/2024.05.12.593736.
43. Eldredge, N. *Time Frames: The Rethinking of Darwinian Evolution and the Theory of Punctuated Equilibria*. (Simon & Schuster, 1986).
44. 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).
45. Brosius, J. & Tiedge, H. Reverse transcriptase: mediator of genomic plasticity. *Virus Genes*

11, 163–179 (1995).

46. Muldal, S. The chromosomes of the earthworms: I. The evolution of polyploidy. *Heredity* **6**, 56–76 (1952).
47. Casellato, S. & Others. On polyploidy in Oligochaetes, with particular reference to lumbricids. in *On earthworms* 75–87 (Mucchi Editore., 1987).
48. Brusca, R. C. & Brusca, G. J. XXVI, Invertebrados, 2a edición, Madrid y otros. Preprint at (2005).
49. Kaygorodova, I. & Matveenko, E. Diversity of the *Piscicola* Species (Hirudinea, Piscicolidae) in the Eastern Palaearctic with a Description of Three New Species and Notes on Their Biogeography. *Diversity*, *15*(1), 98 (2023).
50. 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).
51. 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).
52. Baker, T. M., Waise, S., Tarabichi, M. & Van Loo, P. Aneuploidy and complex genomic rearrangements in cancer evolution. *Nat Cancer* **5**, 228–239 (2024).
53. Mazzagatti, A., Engel, J. L. & Ly, P. Boveri and beyond: Chromothripsis and genomic instability from mitotic errors. *Mol. Cell* **84**, 55–69 (2024).
54. Michael Ghadimi, B. & Ried, T. *Chromosomal Instability in Cancer Cells*. (Springer, 2015).
55. Ivanković, M. *et al.* A comparative analysis of planarian genomes reveals regulatory conservation in the face of rapid structural divergence. *Nat. Commun.* **15**, 8215 (2024).
56. Stephens, P. J. *et al.* Massive genomic rearrangement acquired in a single catastrophic event during cancer development. *Cell* **144**, 27–40 (2011).
57. 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).

58. Holland, A. J. & Cleveland, D. W. Chromoanagenesis and cancer: mechanisms and consequences of localized, complex chromosomal rearrangements. *Nat. Med.* **18**, 1630–1638 (2012).
59. Heng, H. H. *Genome Chaos: Rethinking Genetics, Evolution, and Molecular Medicine*. (Academic Press, 2019).
60. Lucek, K. *et al.* The impact of chromosomal rearrangements in speciation: From micro- to macroevolution. *Cold Spring Harb. Perspect. Biol.* **15**, (2023).
61. 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).
62. 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).
63. 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).

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 12). 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 RNA-seq), specimens were dissected on ice and either flash frozen until tissue was extracted or fresh tissue was used directly for genomic DNA extraction. A detailed list of species used for each analysis is provided in Supplementary Table 12.

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* Several (regenerated tail, pharynx, seminal vesicle, circulatory system, ganglia, typhlosol, body wall) 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). 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[64] 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[65] with parameter -x map-hifi followed by purge_dups v1.0.1. PacBio CLR reads of *C. matritensis* were assembled using NextDenovo v2.4[64] 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[65] with parameter -x map-pb followed by purge_dups v1.0.1[66]. Hi-C reads were trimmed using Cutadapt v2.9[67] and pre-processed using bowtie2 2.3.5.1[68] and hicstuff

v2.3.0[69] with parameters -e DpnII,HinfI -m iterative. Assemblies were subsequently scaffolded based on Hi-C contacts using instaGRAAL v0.1.6[69] 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[70] with parameters --ne --tgstype pb and a modification of mapping parameters to -x asm20. PacBio HiFi reads were then mapped with minimap2 v2.24[65] with parameters -ax map-hifi --MD --secondary=no and provided as input for polishing using HyPo v1.0.3[71] with parameters -c 28 -s 600m -k ccs. Gaps in the *C. matritensis* scaffolds were filled using TGS-GapCloser v1.0.1[70] with parameters --ne --tgstype pb. PacBio CLR reads were mapped with minimap2 v2.24[65] with parameters -ax map-pb --MD --secondary=no. Illumina reads were trimmed using Cutadapt v2.9[67] and mapped using bowtie2 v2.3.5.1[68]. The mapped reads were provided as input to HyPo v1.0.3[71] with parameters -c 140 -s 600m. *k*-mer content was assessed in both final assemblies using KAT comp v2.4.2[72] with the PacBio HiFi dataset for *N. najaformis* and the Illumina dataset for *C. matritensis*. BUSCO v5.0.0[73] 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 1.

Annotation of transposable elements

Transposable elements (TEs) were de novo identified within each genome assembly using RepeatModeler v2.0.3[74] 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[75]. RepeatMasker was run with the following flags: "-s -a -nolow -no_is -xsmall".

Gene annotation

RNA-seq reads were trimmed using Cutadapt v2.9[67] and assembled using Trinity v2.11.0[76]. The transcriptome assembly and the RNA-seq reads were provided as input to Funannotate v1.8.13[77] for training using HISAT2 v2.2.1[78] to map the reads, PASA v2.5.2[79] and TransDecoder v5.5.0[80]. Genes were then predicted using a combination of the training outputs, Augustus v3.3.3[81] and GeneMark-ES v4.68[82], and processed using EVidenceModeler v1.1.1[83].

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 mollusc - *Pecten maximus*) was explored with odp v0.3.0[84] (Table 1). The Bilaterian-Cnidarian-Sponge Linkage Groups (BCnS LGs[3]) 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[84] to infer linkage groups specific for the leeches (named LeechLGs) using the genomes of the three leeches, and for Crassiclitellata (i.e., 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), we intersected the LeechLGs and the CrassiLGs. For every gene, we determined its corresponding LeechLG and CrassiLG and then evaluated if the overlap between each LeechLG and each CrassiLG was significant using a Fisher test. We corrected for multiple comparisons using the Benjamini-Hochberg method and only overlaps with corrected p-values under 1e-10 were considered as candidate ClitLGs. This process was repeated for the aforementioned six species. We then compared these candidates across species to identify species-specific ClitLGs and to merge or split the candidate ClitLGs as needed. 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 (Supplementary Table 1). 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 between them. 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[85] (Supplementary Table 2). Statistics for the alignment and the ancestral genomes for each node were extracted using halStats from HAL tools[86]. To identify whole-genome duplications, ksrates v1.1.3[87] was run as a manual pipeline with default parameters and activating the collinearity analysis. *N. najaformis*. *E. andrei*, *M. vulgaris* and *H. nipponia* were used as focal species. Additionally, Tree2GD v1.0.43[88] was run with default parameters using all annotated chromosome-level annelids plus the draft genome of *Enchytraeus 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[74] 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 adjacent. 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_shat

tering_clitellate_origins). To identify tandem repeats and satellite DNA, Spectral Repeat Finder (SRF[89]) was used. The density throughout the genome was plotted using custom R scripts available at the GitHub repository associated with this manuscript.

We assessed whether synteny breakpoints were disproportionately associated with specific repetitive sequences using the R package GenomicRanges (v1.46.1)[90]. We defined 10 kb flanking windows for each ALG fragment in the genome of *Norana najaformis*. Due to the high level of ALG fragmentation, we almost did not retrieve any windows with clear borders (i.e., without mixing), and hence considered that clear breakpoints cannot be inferred with confidence in clitellates.

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[74]. Next, previously identified transposases (together with sequences belonging to *Chapaev* that would be used as outgroups for rooting the trees[91]) were obtained from the peptide RepeatPeps database from RepeatMasker[75]. 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.[92] 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[93]) using BITACORA[94]. Only one hit was obtained (BDT63348.1, family *Nimaviridae*). Bracovirus (family Nudiviridae) *Chapaev-3* from Zhang et al.[92] was also included in the analysis. All annelid transposase, RepeatPeps, and identified viral sequences were aligned using MAFFT v7[95]. Resulting alignment was trimmed using the automated option from trimAl v1.4.1[96] 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[97]. ModelFinder[98] established WAG+F+R5 as the best model.

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 12), including the newly generated data described above and following the current systematic classification of the phylum[17,18,99–101]. The pipeline described in the MATEdb2 database[102] was used to retrieve the longest isoform for each species. Hierarchical orthologous groups (HOGs) were inferred with OMA v2.6[103]. Gene repertoire evolutionary dynamics across nodes were estimated with pyHam [104] 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[105]) and natural language processing ones (FANTASIA[106]). 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[107]. 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).

Hox genes inference

Homeobox genes were inferred as previously described in Zwarycz et al.[108] on the protein sequences of all species with chromosome-level genomes, plus the draft genome of *Enchytraeus crypticus*, since it was the only representative of Enchytraeidae, a clitellate lineage different to earthworms and leeches. To further identify *Hox* genes, phylogenetic trees were inferred with IQ-TREE v2.2.2.2[97] under the LG+F+R10 model as selected by ModelFinder[98]. A second tree (model LG+F+R9) was inferred with the subset of putative *Hox* genes to classify

each sequence into its *Hox* subfamily (Supplementary Data 11). *Hox* clusters were defined using the genomic coordinates of the putative *Hox* genes. If a pair of *Hox* genes was less than 300kb apart and had less than 10 genes in between, they were considered as part of the same cluster. We consider this a conservative approach reflecting the overall *Hox* gene distance observed in our dataset on one side (e.g., in *S. nudus*), and the observations reporting for other chromosome-level genomes on the other (e.g., distance between the *lab* gene at the posterior end of lepidopteran *Hox* clusters is separated by a distance between 1.4Mb and 24Mb, containing numerous non-*Hox* genes[109].

Hi-C data processing

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).

Hi-C data processing was performed using the TADbit pipeline v1[31] (version 1.0). Briefly, reads were mapped in windows from 15bp to 75bp in 5bp steps using GEM3-Mapper v3[110]. Only valid pairs of mapped reads were considered to avoid noisy contacts. To achieve this, we used the following filters provided by TADbit v1[31] to remove artefacts such as “self-circle,” “dangling-end,” “error,” “extra dangling-end,” “too short,” “too large,” “duplicated,” and “random breaks.” (see Supplementary Table 13). Subsequently, contact matrices were created at 50Kbp resolution and normalised to 50M contacts, following the ICE method using hicNormalize from HiCExplorer v3.7[111]. Finally, matrices were corrected and plotted using hicCorrect and hicPlotMatrix from HiCExplorer v3.7[111].

Inter-chromosome/intra-chromosome interaction ratio and averaged contact probability curves

Intra- and inter-chromosomal interactions were analysed by converting the corrected and normalised 50 Kbp matrices from h5 to interactions, using the tool 'hicConvertMatrix' from HiCExplorer v.3.7[111]. This format is a dataframe that contains interaction frequency data in bin pairs (bin1, bin2 and contact frequency). Using this data, the sum of inter- chromosomal interactions (taking contacts of bins in different chromosomes) and intra- chromosomal interactions (taking contacts of bins in the same chromosome) per chromosome was calculated. To obtain the inter-chromosome/intra-chromosome interaction ratio, the sum of inter-chromosomal interactions was divided by the sum of intra-chromosomal interactions per chromosome. This calculation was performed using Rstudio. The correspondence between ALGs and chromosomes was used to analyse ALGs interaction distribution in the same manner.

Using the corrected and normalised 50 Kbp matrices as an input, contact probability vs distance curves (P(s)) were calculated using 'hicPlotDistvsCounts' from HiCExplorer v3.7[111]. The curves were plotted using Rstudio and maximum distance of 1 Gbp.

A/B compartments and TAD inference

The A/B compartments were estimated by an eigenvector analysis of the genome contact matrix after normalisation by the observed–expected method[112] Specifically, boundary changes between the two compartments occur where the entries of the first eigenvector change sign[113]. The 1st eigenvector and insulator score values were obtained using the Hi-C analysis tools package, FAN-C v0.91[114]. The 1st eigenvector was calculated with the 'fanc compartments' tool on normalised 50Kbp matrices using default settings. The 'fanc insulation' tool was used to calculate the insulator score on normalised 50Kbp matrices. TADbit v1[31] was employed to calculate TAD boundary strength, assigning values from 1 to 10.

Aggregated TAD plots were created using the "hicAverageRegions" tool from HiCExplorer v.3.7[111], with TADbit TAD boundary coordinates and normalised 50 Kbp contact matrices as inputs. Plots displaying contact heatmaps, along with 1st eigenvector and insulator score tracks, were generated using the pyGenomeTracks tool (https://github.com/deeptools/pyGenomeTracks).

Long-range interactions

Inter- and intra-chromosomal long-range contacts were identified using ginteraction dataframes by extracting interactions that were at least twice the mean of inter- or intra-chromosomal interactions, respectively. For intra-chromosomal interactions, only contacts between loci separated by at least 2Mb were considered long-range. The long-range interactions connecting Hox genes were visualised using a circos plot.

Genome-wide selection analyses

We explored signatures of selection in the HOGs that arose before Clitellata and that changed their position due to the chromosome scrambling to test if the change of genomic location resulted in shifts in selection regimes. For that, we selected HOGs that arose before Clitellata and that contained a minimum of 20% of the species of marine annelids and a minimum of 20% of clitellate species, in order to ensure an adequate taxon representation (n=10,188; see also Supplementary Table 5).

To detect differential selection in protein sequence alignments at the genome-wide level in clitellates potentially facilitating the transition from marine to freshwater or terrestrial environments, we used the software Pelican[36]. Based on a maximum likelihood approach, Pelican aims at detecting significant shifts in amino acid preferences between lineages. The software scans alignments for sites differing in amino acid preferences depending on a trait or condition that is specified on a phylogenetic tree (in our case, marine vs non marine). We used the option pelican scan discrete and –alphabet=AA. Pelican produces a file containing one p-value per site across all genes and assigns a p-value of 1 to amino acids that are shared across all species. In order to get gene-level predictions, we used the Gene-wise Truncated Fisher's method (GTF) to obtain a score for all HOGs as implemented in the R package GTFisher (https://gitlab.in2p3.fr/phoogle/pelican/-/wikis/Gene-level-predictions). To calculate p-values, we considered the best 10 site-specific p-values in each alignment (k=10) and corrected for a false discovery rate (FDR) of 0.05 with the Benjamini & Hochberg (1995) (BH) method. We considered as significantly under directional selection those HOGs with FDR-adjusted p-values < 0.05 (Supplementary Table 5).

Stress experiments

Specimens from *E. andrei* and *H. medicinalis* were subjected to several types of abiotic stress related to their terrestrial/freshwater ecological niches, including exposure to visible and UV-B light, hyperoxia, hypoxia, and osmotic stress. All samples were left in quarantine in an empty plastic container with wet Whatman paper for 24 hours before any experiment to empty gut content. Five biological replicates per species and experiment were included, as well as five control specimens for each species (n=30 per species). Samples under visible light were kept under natural light (i.e., close to a window in the laboratory) for 15 minutes in an empty petri dish with no space to hide. For exposure to UV-B light, specimens were exposed under a UV-B lamp (302 nm) for 2 minutes. Animals were allowed to recover in darkness 24 hours in a dark chamber at a constant temperature of 16ºC in order to activate the UV-induced DNA damage repair genetic machinery. Hyperoxia experiments were carried out by adding pure oxygen into a controlled sealed chamber until reaching 38-44% oxygen for 20 minutes. Oxygen concentration was continuously monitored by an oxygen sensor located at the interior of the hyperoxia chambers (Presens OXY-1 ST Fiber). Hypoxia experiments were done in a HypoxyLab™ (Oxford Optronix) at 8% oxygen concentration for *H. medicinalis* and 15% oxygen concentration for *E. andrei* for 20 minutes, based on previous information on what constitutes hypoxic conditions in freshwater or terrestrial environments without causing mortality[115,116]. For the osmoregulation experiment, *H. medicinalis* was immersed in sea water for 2 minutes to trigger osmotic stress, after making sure that the exposure time was causing stress but not mortality. Specimens of *E. andrei* were dried superficially with Whatman paper and left in a 9 cm petri dish for 15 minutes exposed to air. Control samples were directly processed without any additional treatment after quarantine. After the experiments, all samples were dissected and tails (body part after segment 35 in *H. medicinalis* and after the clitellum in *E. andrei*) were flash frozen in liquid nitrogen and kept at -70ºC until further processing. RNA extractions after stress experiments were performed using the TRIzol® reagent (Invitrogen, USA) method following the manufacturer's instructions. Library preparation and sequencing was done as described above.

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[117] (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 was generated 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[118] 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 normalised expression matrices. For *E. andrei*, since there is a high-quality chromosome-level genome available[26], the trimmed RNA-seq reads were aligned to the genome using the splice-aware aligner STAR v2.7.10[119] and featureCounts v2.0.5[120] 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[121]. Finally, the Benjamini-Hochberg correction method[122] 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.

Comparative genomics and selection analyses of genes involved in response to abiotic stress

To investigate the evolutionary origin of the genes involved in response to abiotic stress, the longest protein isoforms of the references used in differential gene expression analyses of *H. medicinalis* and *E. andrei* were included in the analysis of gene repertoire evolution as described above to infer which HOGs contained these genes. 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 - coinciding with the chromosome scrambling - or after). In the case of *H. medicinalis,* due to the lack of an available genome sequence, we identified the coordinates of orthologous sequences in HOGs inferred from the chromosome-level genome of *Hirudinaria manillensis*[123] with corroborated high homology (reciprocal best hits in BLAST+[124]).

In order to test the potential adaptive role of genes that changed position during the chromosome scrambling in the colonisation of new environments, we tested if DEGs within HOGs that changed position during the genome-wide fusion-with-mixing (i.e., those within the 'before' group as described above) were subjected to directional selection. We used the software Pelican[36] as described above (see 'Genome-wide selection analyses'). As a second selection analysis, we tested whether DEG-harbouring HOGs from the 'before' category were subject to positive selection. A codon based alignment for each HOG was obtained using HyPhy version 2.5.42[37]. This alignment was used to obtain a phylogenetic tree using IQTREE version 1.6.12[97] selecting the MFP substitution model. Clitellate branches were tested for positive selection using aBSREL[125]. Putative function of DEGs was explored with the Cluster of Orthologous Genes (COG) database in NCBI[38] to assign COG annotations.

Composite gene analyses

To understand the chimeric nature of genes that arose during the genome scrambling and that are differentially expressed in response to abiotic stress in *E. andrei* and *H. medicinalis*, the HOGs containing the differentially expressed genes that originated in Clitellata as defined in the previous section (i.e., 'in' group) were explored to characterise their chimeric origin based on the criteria used by Mulhair et al.[126]. An all vs all BLASTp of all our proteomes was performed to create a sequence similarity network with the cleanBlastp command in CompositeSearch[127]. The correspondence between the filtered HOGs and the genes was used, together with the output of cleanBlastp, as input for CompositeSearch[127] 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). 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 composite 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.

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 Table 12.

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). NG was supported by the European Union's Horizon 2020 research and innovation programme under Marie Skłodowska-Curie grant agreement 764840 to JFF (ITN IGNITE; www.itn-ignite.eu) as well as by a Deutsche Forschungsgemeinschaft (DFG) grant (458953049). 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 (Methods)

64. 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.

65. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).

66. Guan, D. *et al.* Identifying and removing haplotypic duplication in primary genome assemblies. *Bioinformatics* **36**, 2896–2898 (2020).

67. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17**, 10–12 (2011).

68. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).

69. 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).

70. 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).

71. 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.

72. 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).
73. Manni, M., Berkeley, M. R., Seppey, M., Simão, F. A. & Zdobnov, E. M. BUSCO opdate: 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).
74. 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).
75. 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).
76. 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).
77. Palmer, J. & Stajich, J. nextgenusfs/funannotate: funannotate v1.5.3 (1.5.3). Zenodo. https://doi.org/10.5281/zenodo.2604804 (2019).
78. 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).
79. Haas, B. J. *et al.* Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res.* **31**, 5654–5666 (2003).
80. Haas, B., Papanicolaou, A., Yassour, M. & Others. TransDecoder. *TransDecoder* (2017).
81. Keller, O., Kollmar, M., Stanke, M. & Waack, S. A novel hybrid gene prediction method employing protein multiple sequence alignments. *Bioinformatics* **27**, 757–763 (2011).
82. 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).

83. Haas, B. J. *et al.* Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biol.* **9**, R7 (2008).
84. Schultz, D. T. *et al.* Ancient gene linkages support ctenophores as sister to other animals. *Nature* **618**, 110–117 (2023).
85. Armstrong, J. *et al.* Progressive Cactus is a multiple-genome aligner for the thousand-genome era. *Nature* **587**, 246–251 (2020).
86. 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).
87. Sensalari, C., Maere, S. & Lohaus, R. ksrates: positioning whole-genome duplications relative to speciation events in KS distributions. *Bioinformatics* **38**, 530–532 (2022).
88. 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).
89. Zhang, Y., Chu, J., Cheng, H. & Li, H. De novo reconstruction of satellite repeat units from sequence data. *Genome Res.* **33**, 1994–2001 (2023).
90. Lawrence, M., Huber, W., Pagès, H., Aboyoun, P., Carlson, M., Gentleman, R., Morgan, M., Carey, V. . Software for computing and annotating genomic ranges. *PLoS Comp. Biol.* **9**, e1003118 (2013) http://www.ploscompbiol.org/article/info%3Adoi%2F10.1371%2Fjournal.pcbi.1003118.
91. 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).
92. 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).
93. Stano, M., Beke, G. & Klucar, L. viruSITE-integrated database for viral genomics. *Database* **2016**, (2016).

94. 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).

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

96. 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).

97. 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).

98. 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).

99. Struck, T. H. Direction of evolution within Annelida and the definition of Pleistoannelida. *J. Zoolog. Syst. Evol. Res.* **49**, 340–345 (2011).

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

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

102. 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.

103. Altenhoff, A. M. *et al.* OMA standalone: orthology inference among public and custom genomes and transcriptomes. *Genome Res.* **29**, 1152–1163 (2019).

104. Train, C.-M., Pignatelli, M., Altenhoff, A. & Dessimoz, C. iHam and pyHam: visualizing and processing hierarchical orthologous groups. *Bioinformatics* **35**, 2504–2506 (2019).

105. 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).

106. 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.

107. TopGO. *Bioconductor* https://bioconductor.org/packages/release/bioc/html/topGO.html.

108. Zwarycz, A. S., Nossa, C. W., Putnam, N. H. & Ryan, J. F. Timing and scope of genomic expansion within Annelida: Evidence from homeoboxes in the Ggenome of the earthworm *Eisenia fetida*. *Genome Biol. Evol.* **8**, 271–281 (2015).

109. Mulhair, P. O. & Holland, P. W. H. Evolution of the insect Hox gene cluster: Comparative analysis across 243 species. *Semin. Cell Dev. Biol.* **152-153**, 4–15 (2024).

110. Marco-Sola, S., Sammeth, M., Guigó, R. & Ribeca, P. The GEM mapper: fast, accurate and versatile alignment by filtration. *Nat. Methods* **9**, 1185–1188 (2012).

111. 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).

112. Lieberman-Aiden, E. *et al.* Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289–293 (2009).

113. Fortin, J.-P. & Hansen, K. D. Reconstructing A/B compartments as revealed by Hi-C using long-range correlations in epigenetic data. *Genome Biol.* **16**, 180 (2015).

114. 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).

115. Hildebrandt, J. P. & Zerbst-Boroffka, I. Osmotic and ionic regulation during hypoxia in the medicinal leech, *Hirudo medicinalis* L. *J. Exp. Zool.* **263**, 374–381 (1992).

116. Šustr, V. & Václav Pižl, V. Oxygen consumption of the earthworm species *Dendrobaena mrazeki*. *Eur. J. Soil Biol.* **45**, 478–482 (2009).

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

118. 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).

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

120. 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).

121. 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).

122. 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).

123. Zheng, J. *et al.* Molecular mechanisms underlying hematophagia revealed by comparative analyses of leech genomes. *Gigascience* **12**, (2022).

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

125. Smith, M. D. *et al.* Less is more: an adaptive branch-site random effects model for efficient detection of episodic diversifying selection. *Mol. Biol. Evol.* **32**, 1342–1353 (2015).

126. 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.

127. 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).