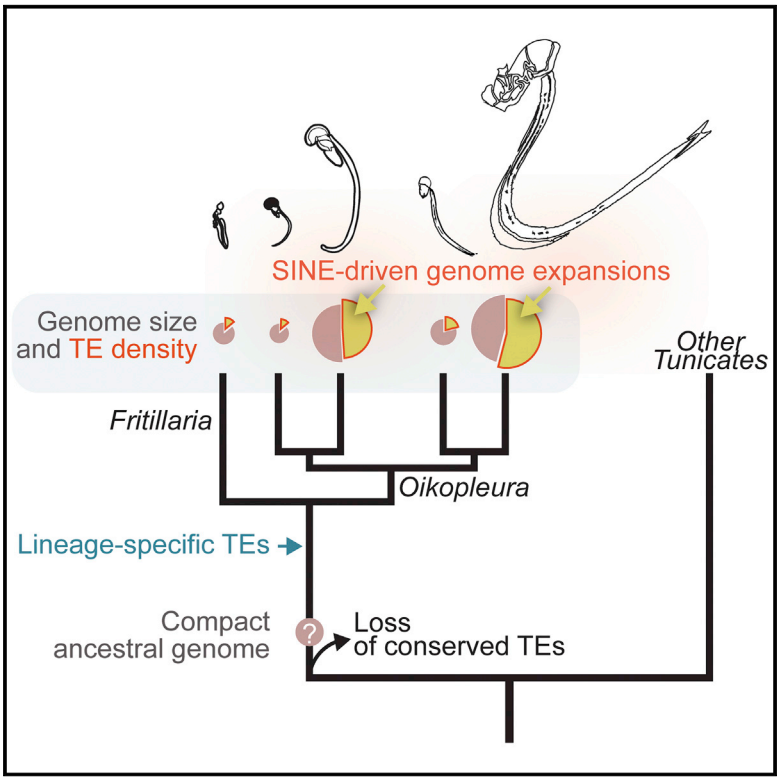
Report

Current Biology

Massive Changes of Genome Size Driven by Expansions of Non-autonomous Transposable Elements

Graphical Abstract



Highlights

- Genome size varies up to 12× in larvaceans, chordates with a distinctive anatomy
- Small and large species have the smallest and largest genomes, respectively
- Transposable elements have driven multiple independent genome expansions
- Genomes mainly increased through accumulations of non-autonomous elements (SINEs)

Authors

Magali Naville, Simon Henriet, Ian Warren, Sara Sumic, Magnus Reeve, Jean-Nicolas Volff, Daniel Chourrout

Correspondence

jean-nicolas.volff@ens-lyon.fr (J.-N.V.), daniel.chourrout@uib.no (D.C.)

In Brief

The causes of genome size variation and their relation with life traits are poorly understood. Naville, Henriet, et al. show that, in tunicate larvaceans, close relatives of vertebrates, multiplications of non-autonomous transposable elements were the main drive of remarkable genome expansions.

Naville et al., 2019, Current Biology *29*, 1161–1168
April 1, 2019
https://doi.org/10.1016/j.cub.2019.01.080

CellPress

Current Biology
Report

Massive Changes of Genome Size Driven by Expansions of Non-autonomous Transposable Elements

Magali Naville,[1,5] Simon Henriet,[2,5] Ian Warren,[1] Sara Sumic,[2] Magnus Reeve,[2,4] Jean-Nicolas Volff,[1,*] and Daniel Chourrout[2,3,6,*]
[1]Institut de Génomique Fonctionnelle de Lyon, Univ Lyon, CNRS UMR 5242, Ecole Normale Supérieure de Lyon, Universite Claude Bernard Lyon 1, allée d'Italie, F-69364 Lyon, France
[2]Sars International Centre for Marine Molecular Biology, Thormøhlensgt. 55, 5006 Bergen, Norway
[3]Key Laboratory of Marine Genetics and Breeding, Ocean University of China, Ministry of Education, Qingdao 266003, China
[4]Present address: Institute of Marine Research, Postbox 1870 Nordnes, 5817 Bergen, Norway
[5]These authors contributed equally
[6]Lead Contact
*Correspondence: jean-nicolas.volff@ens-lyon.fr (J.-N.V.), daniel.chourrout@uib.no (D.C.)
https://doi.org/10.1016/j.cub.2019.01.080

SUMMARY

In eukaryotes, genome size correlates little with the number of coding genes or the level of organismal complexity (C-value paradox). The underlying causes of variations in genome size, whether adaptive or neutral, remain unclear, although several biological traits often covary with it [1–5]. Rapid increases in genome size occur mainly through whole-genome duplications or bursts in the activity of transposable elements (TEs) [6]. The very small and compact genome of *Oikopleura dioica*, a tunicate of the larvacean class, lacks elements of most ancient families of animal retrotransposons [7, 8]. Here, we sequenced the genomes of six other larvaceans, all of which are larger than that of *Oikopleura* (up to 12 times) and which increase in size with greater body length. Although no evidence was found for whole-genome duplications within the group of species, the global amount of TEs strongly correlated with genome size. Compared to other metazoans, however, the TE diversity was reduced in all species, as observed previously in *O. dioica*, suggesting a common ancestor with a compacted genome. Strikingly, non-autonomous elements, particularly short interspersed nuclear elements (SINEs), massively contributed to genome size variation through species-specific independent amplifications, ranging from 3% in the smallest genome up to 49% in the largest. Variations in SINE abundance explain as much as 83% of interspecific genome size variation. These data support an indirect influence of autonomous TEs on genome size via their ability to mobilize non-autonomous elements.

RESULTS AND DISCUSSION

Larvaceans are planktonic tunicates with a very distinctive body plan. They include three families: the Oikopleuridae; the Fritillaridae; and the Kowalevskiidae. The diversification of larvacean species (and possibly other tunicates) involved a great deal of relatively rapid body size variation, which is associated with unusual size changes for cells, nuclei, and, in the end, genomes. Seven species were sampled here, including six oikopleurids and one fritillarid (*Fritillaria borealis*). Adult length varies at least ten-fold between the seven species studied [9] (F. Lombard, personal communication; Figure 1A).

Phylogenetic analysis based on ribosomal protein coding genes (Figure 1A) confirmed that the two giant species *Bathochordaeus sp.* and *Mesochordaeus erythrocephalus* are oikopleurids and might be renamed accordingly. According to the phylogeny, oikopleurids consists of two subgroups: first, an "*O. dioica* group" containing *O. dioica*, *O. albicans*, and *O. vanhoeffeni*, the two latter diverging last, and second, an "*O. longicauda* group," consists of *O. longicauda* and both giant species, with the giant species diverging last. As expected, the fritillarid *F. borealis* is found distant from all oikopleurids. The very high substitution rate earlier noted for *O. dioica* [8] seems to be a common trait among larvaceans, based on the systematically long branches in the tree.

Sequencing and assembly of the genomes of the seven larvacean species revealed strong differences of genome size, from 72 Mb for *O. dioica* up to 874 Mb for *M. erythrocephalus*. Interestingly, a relationship was found between genome and body size. *F. borealis*, *O. dioica*, and *O. longicauda*, all small species, have relatively small genomes, although much larger genomes are found in the larger species *O. albicans* and *O. vanhoeffeni* and in the so-called giant species *Bathochordaeus sp.* and *M. erythrocephalus* [10] (Figure 1A).

In order to reveal the origin of such an exceptional genome size variation, we first looked for evidence of genome duplication events. Two hundred highly diverse protein-coding genes, which

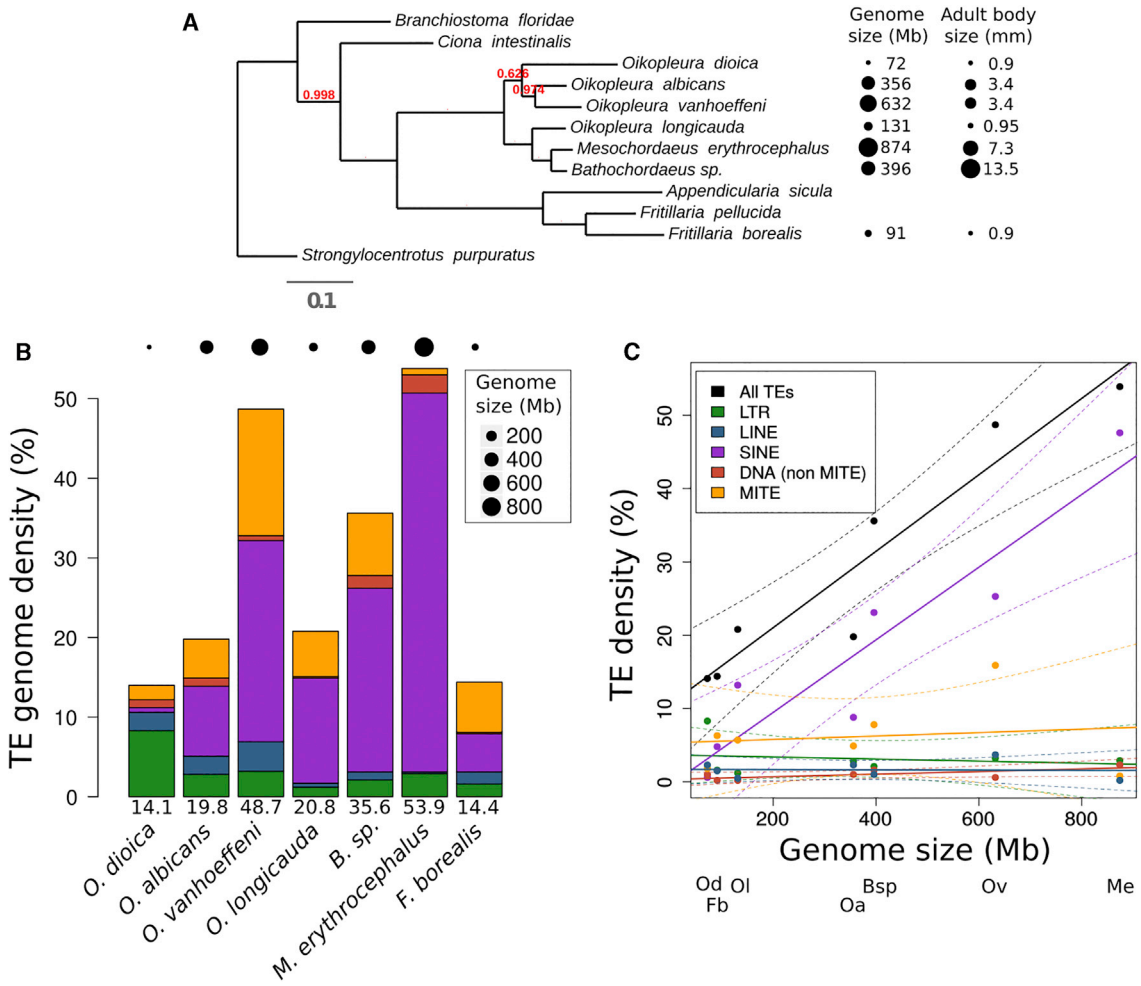


Figure 1. Larvacean Phylogeny, Genome Sizes, and Genomic Density of the Main Categories of Transposable Elements

(A) Phylogenetic analysis of ribosomal protein genes from chordates, including nine larvaceans. Estimations of genome size (in Gb) are indicated for each species used in the current study (see STAR Methods for more details). Two species whose genomes were more recently sequenced (*Appendicularia sicula* and *Fritillaria pellucida*) were added in order to increase the taxonomic density of fritillarids, though their genome sequences were not annotated for TEs. Two species were used as outgroups: the sea urchin *Strongylocentrotus purpuratus*, as a non-chordate deuterostome, and the lancelet *Branchiostoma floridae*, belonging to cephalochordates. Body sizes are classically estimated by the adult trunk length, measured for most species in wild specimens and for *O. dioica* on laboratory bred animals.

(B) Proportion of each TE class in the genomes surveyed. The sizes of black dots above the bar plot are proportional to genome sizes. Numbers under the bar plot correspond to total TE density percentage in each species. TE density ranges from ca. 14% in the compact genomes of *O. dioica* and *F. borealis* to 54% in the expanded genome of *M. erythrocephalus* are shown.

(C) Percentage of total TEs (in black) or of different TE classes (in colors) as a function of genome size. Lines correspond to linear regressions; dotted lines correspond to the 95% confidence intervals around the mean predictions. TE density, and more precisely SINE density, correlates with genome size.

See also Figure S1 and Table S1.

were putatively unique in the bilaterian ancestor [11, 12], were selected based on their high level of conservation in larvaceans. A minority of these was found to be duplicated in one or the other larvacean lineage, but the proportions of duplicated genes were similar among different species. This rules out the existence of differential whole-genome, or other large-scale, duplications that could have led to an expansion of some of the genomes (Figure S1).

In contrast, transposable elements (TEs) had an obvious and very significant impact on genome size. This is first evident when comparing their global abundance. TEs cover highly variable proportions of the genomes, from 14% in *O. dioica* and *F. borealis* up to 54% in *M. erythrocephalus* (Figure 1B; Table S1), and TE density is strongly correlated with genome size (Pearson's test corrected by the phylogeny; adj. $r^2 = 0.88$; p value = 1.1e−03; Figure 1C, black line). Among TE density estimations in other chordates, these proportions are in the low and high range, respectively [13, 14]. However, for a given genome size, TE coverage is much higher in larvaceans than in vertebrates (Figure 2).

To gain insight into the past and present activity of TEs, we calculated the degree of similarity between different copies in each family (Figure 3A). The resulting "TE landscapes" show the largest proportions of TE copies with high similarity levels

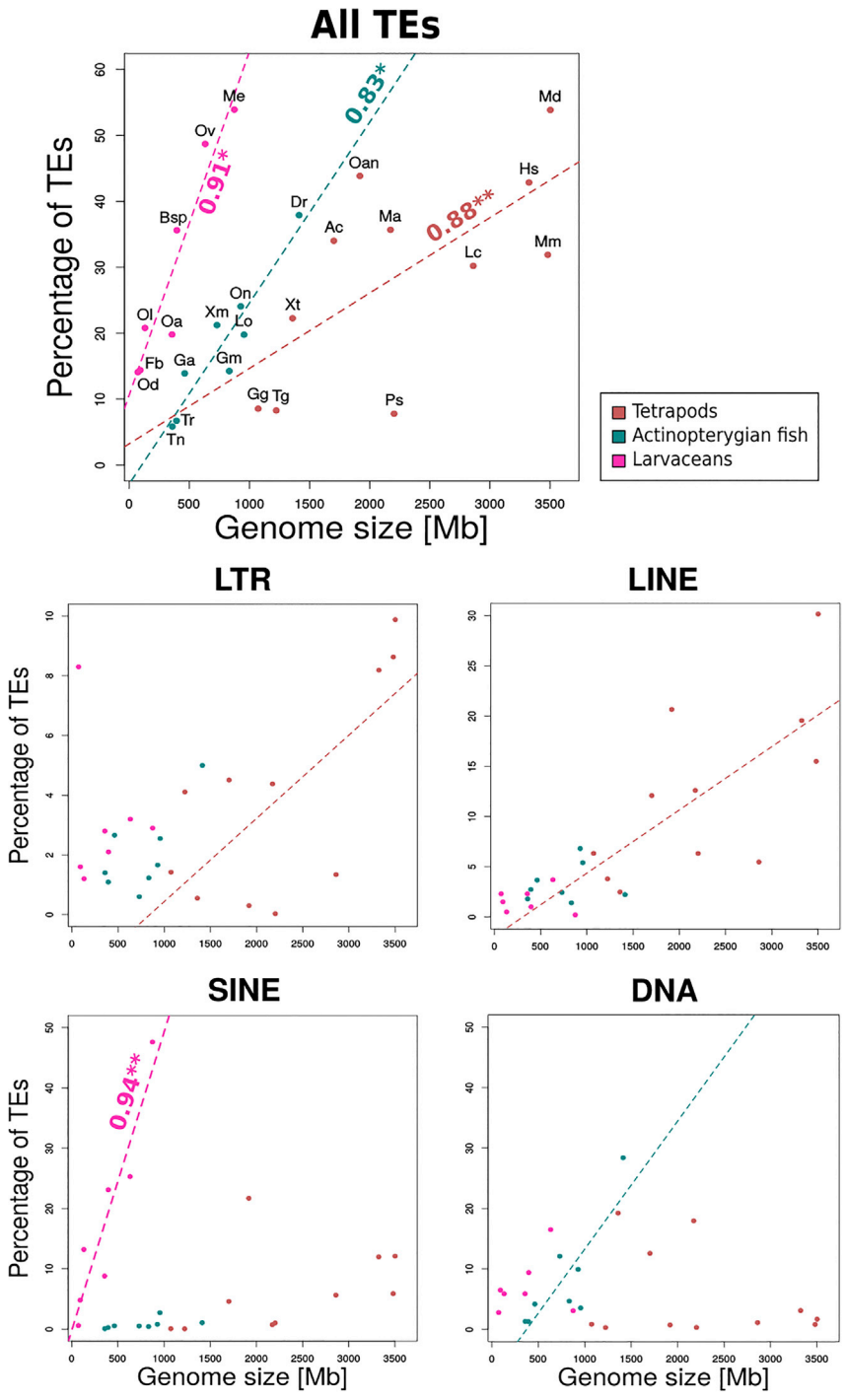


Figure 2. Relationship between TE Coverage and Genome Size in Different Phylogenetic Groups

Numbers on linear regressions correspond to adjusted r^2 coefficients (Pearson's test corrected by the phylogeny). $^{**}10^{-3} \leq$ p value < 0.01; $^{*}0.01 \leq$ p value < 0.05. Aa, *Anguilla anguilla* (European eel); Ac, *Anolis carolinensis* (green anole); Am; *Alligator mississippiensis* (American alligator); Bsp, *Bathochordaeus sp.*; Dr, *Danio rerio* (zebrafish); Fb, *Fritillaria borealis*; Ga, *Gasterosteus aculeatus* (stickleback); Gg, *Gallus gallus* (chicken); Gm, *Gadus morhua* (Atlantic cod); Hs, *Homo sapiens* (human); Lo, *Lepisosteus oculatus* (spotted gar); Md, *Monodelphis domestica* (opossum); Me, *Mesochordaeus erythrocephalus*; Mm, *Mus musculus* (mouse); Oa, *Oikopleura albicans*; Oan, *Ornithorhynchus anatinus* (platypus); Od, *Oikopleura dioica*; Ol, *Oikopleura longicauda*; On, *Oreochromis niloticus* (tilapia); Ov, *Oikopleura vanhoeffeni*; Ps, *Pelodiscus sinensis* (Chinese soft-shell turtle); Tg, *Taeniopygia guttata* (zebra finch); Tn, *Tetratodon nigroviridis* (Tetraodon); Tr, *Takifugu rubripes* (fugu); Xm, *Xiphophorus maculatus* (platyfish); Xt, *Xenopus tropicalis* (western clawed frog). Pearson's correlations were estimated, taking into account only tetrapods (red dashed line), or only larvaceans (fuchsia dashed line), and after correction to take into account phylogenetic relationships (see STAR Methods). Non-larvacean data are adapted from Chalopin et al. [14]. TE coverage significantly correlates with genome size in larvaceans as well as in tetrapods and fish. The distinction of the different TE classes impact shows that, in larvaceans, this correlation is mainly due to SINEs, although it appears more influenced by LTRs and LINEs in tetrapods and by DNA elements in fish, although these last correlations are not sustained when taking into account the phylogeny.
See also Figure S2.

(80%–100%) for *Bathochordaeus sp.* and *O. albicans* and the smallest proportion for *O. dioica*. This suggests that the most recent bursts of transposition occurred during the evolution of the species with larger genomes. An exception in the *O. dioica* TE landscape is for a group of almost identical copies of the well-characterized Ty3/gypsy *Oikopleura* retrotransposon (TOR) long terminal repeat (LTR) retrotransposons [7, 17]. Although the other species are hermaphrodites, *O. dioica* is the only dioecious larvacean and has a large Y chromosome [8]. The Y chromosome contains 43% of the TOR insertions (99/222), representing a 9× enrichment compared to the rest of the genome (p value $< 2.2e{-}16$; proportion test).

We then surveyed the phylogenetic diversity of all TEs. Members of all retrotransposon clades commonly found in other animals, but absent in *O. dioica* [14], were also found to be absent in the other oikopleurids (Table S1). A couple of those clades are, however, found in *F. borealis*, indicating more systematic losses in the oikopleurid lineage. The relatively poor representation of ancient retrotransposon clades in *Ciona intestinalis* suggests that some losses were either frequent or very old during the history of tunicates and possibly related to genome compaction events [14]. However, as observed for *O. dioica* [7], lineage-specific TE clades were found in the other larvacean species. All oikopleurids possess TOR LTR retrotransposons, as well as some *Dictyostelium* intermediate repeat sequence 1 (DIRS) retrotransposons, a group of

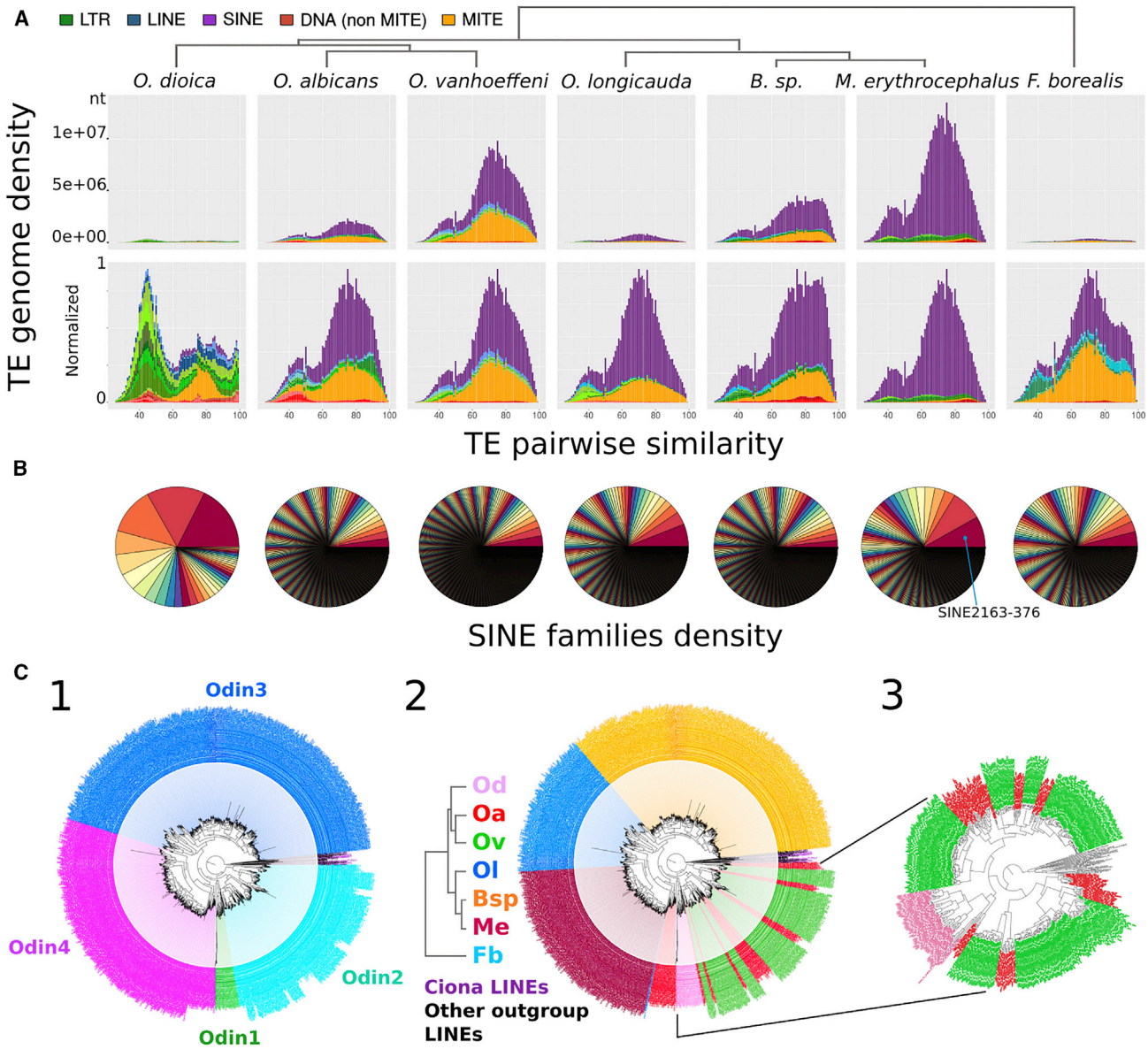


Figure 3. Importance of the SINEs among TEs in the Larvacean Genomes

(A) Landscapes of TEs in the different species. Histograms represent the proportion of pairwise sequence similarities obtained when comparing all sequences with each other in each TE family. An ancient burst of LTR elements occurred in *O. dioica*, and the other species landscapes are mainly dominated by bursts of non-autonomous elements.

(B) Genome density distribution of SINE elements in the different larvacean species. Except in *O. dioica*, where 50% of the region covered by SINEs is made of four different SINE elements, SINEs are highly diversified in the different species, with no particular SINE being highly spread compared to the others. Different colors show different families of SINEs. SINE2163-376, the most prominent family in the genome of *M. erythrocephalus*, is shown by an arrow (see main text).

(C) Phylogenies of LINE/Odin elements. (C1 and C2) Phylogenetic reconstruction of 1,950 Odin insertions based on 463 sites of the reverse transcriptase. (C1) Sequences are colored according to their subfamily. (C2) Sequences are colored according to the species they belong to. (C3) Subtree obtained with the subset of Odin1 and Odin2 sequences only found in *O. dioica*, *O. albicans*, and *O. vanhoeffeni* is shown. Proteins were aligned using Mafft [15] and phylogenies reconstructed using FastTree [16].

See also Table S2.

elements with atypical long repeats. As in *O. dioica*, TOR elements in the five newly sequenced oikopleurid genomes frequently have a third open reading frame (ORF) that encodes an envelope-like gene [17]. *F. borealis* possesses a new superfamily of Gypsy-like LTR retrotransposons that we named FbGypsy. Odin, R2-like LINEs (long interspersed nuclear elements) as well as Penelope-like elements are the only families of retrotransposons without LTRs identified in oikopleurids (Figure 3C). *F. borealis* has no Odin elements but has members of other LINE families. These consist of already known families, such as RTE or CR1 [18, 19], as well as new families that we named ULF1 and ULF2 (for "unknown LINEs of *Fritillaria*";

Table S1). SINEs (short interspersed nuclear elements), non-autonomous retroelements that do not encode any protein, were predicted in all species, and all SINE families were species-specific (see a more detailed description below). Predicted SINEs possess the classical box A and box B sequences of polIII promoters separated by a 25- to 50-nt sequence and a 3′ poly(A) stretch [20]. Finally, a more heterogeneous picture resulted from surveying DNA transposons. Several ancient families of autonomous elements [14, 21] were found in either all or only some larvacean species. As for SINEs, MITEs (miniature inverted transposable elements), which are non-autonomous DNA transposons, were detected in variable proportions in all species, and none of them was shared between distinct species. Overall, the qualitative survey of TE diversity showed a drastic impoverishment of retrotransposon diversity particularly in oikopleurids, only partly ''compensated'' by the acquisition of new lineage-specific families. The fact that the four largest larvacean genomes do not possess ancient retrotransposon families suggests that the genome of their common ancestor was small, possibly connected to the inactivation of these TE families, and then ''re-expanded'' in some lineages.

Although intact or nearly intact LTR elements prevail in the smallest genome (56% of the TE part of *O. dioica* genome), non-autonomous SINEs and MITEs dominate in all the other species (Figure 1B). In the three largest ones (*O. vanhoeffeni*, *Bathochordaeus sp.*, and *M. erythrocephalus*), MITEs and SINEs occupy 41%, 30%, and 48% of the total and 84%, 84%, and 89% of the TE-containing part of the genomes, respectively. The level of abundance for SINEs correlates more strongly with the genome size (Pearson's test corrected by the phylogeny; cor = 0.94; p value = 4.5e−03) than does that of MITEs (cor = 0.18; p value = 0.73) and of autonomous elements (cor = 0.72; p value = 0.11; Figures 1B, 1C, purple line, 2, and S2). When considering the phylogeny of larvaceans, the differences of SINE content account for 83% of the genome size variation (adjusted r^2; linear regression).

Strikingly, larvacean genomes contain a high diversity of SINE elements with low genomic density, with no clearly dominant families (Figure 3B; Table S2). For instance, the SINE-rich genome of *M. erythrocephalus* contains as many as 868 SINE families, with SINE2163-376, the most prominent family (ca. 190,000 copies), covering only 4% of the genome (Figure 3B). One exception is the SINE-poor genome of *O. dioica*, where four major families make up as much as 50% of the SINE content but with a total genome density of less than 1% (Table S1). Comparisons with SINEbase [22] suggest that these elements are larvacean specific. SINEs from one larvacean species show no significant homology with SINEs from other larvacean species, suggesting independent origins and amplifications, although we cannot exclude very high sequence divergence after common ancestry.

Only a small proportion of SINE families were found to be related to tRNA genes, with the highest proportion in the SINE-poor genome of *O. dioica* (23%) compared to between 1% and 7% in the other species. No similarity with tRNA sequences were found in the most prominent families (Table S2). No relationship was found with rRNA, small nuclear RNA (snRNA), or other non-coding RNAs, and SINEs identified as tRNA derived covered less than 3% of all genomes considered (Table S1). This indicated either that the tRNA motif was too degenerated to be recognized in most SINE elements by the methods used or that they are derived from other RNA genes not identified here.

We could not detect any obvious similarity between larvacean SINE elements and LINE-Penelope elements present in the same genome. This might suggest that SINE mobilization mostly occurred through interactions of the non-LTR retrotransposon enzymatic machinery with the poly(A) tail of the SINEs, as observed for the L1 retrotransposons in mammals [23], rather than with a retrotransposon-related 3′ sequence [24]. In such a model, one family of non-LTR retrotransposons might transpose several SINE families, as observed for L1 in mouse, which mobilizes both B1 (7SL RNA-derived) and B2 (tRNA-derived) SINE families [25].

In order to test whether non-LTR retrotransposons might mobilize larvacean SINE elements, we examined, in species from the *O. dioica* group, whether SINE abundance positively correlates with the similarity of copies of LINEs, which might serve as an indicator of their activity. A molecular phylogeny of Odin reverse transcriptase proteins revealed shorter branches in *O. vanhoeffeni* compared to *O. dioica* and *O. albicans* (average branch lengths of 2.57 and 3.24 for *O. dioica* and *O. albicans*, respectively, compared to 1.03 in *O. vanhoeffeni*; p value [OA+OD versus OV] = 0.03; t test; Figures 3A and 3C). Accordingly, the average substitution rate in the RT region between Odin elements was lower in *O. vanhoeffeni* (1.43) than in *O. dioica* (3.18) and *O. albicans* (4.72; p value = 0.04; t test). These results indicate that Odin elements were more recently active in *O. vanhoeffeni* than in *O. dioica* and *O. albicans*. This might explain the high SINE content in *O. vanhoeffeni* compared to *O. albicans* (25.3% versus 8.8%; p value = 3.6e−03; χ^2) and to *O. dioica* (SINEs are almost absent in this genome). Although Penelope-like elements are much more spread in *O. vanhoeffeni*, analysis of the average similarity between copies did not reveal any obvious difference among the three species (Table S1; data not shown).

That non-autonomous TEs could prevail to such an extent over autonomous elements to increase genome size is an intriguing phenomenon, which to our knowledge has not been reported for other taxonomic groups. Before proposing mechanistic interpretations, possible trivial technical biases, such as differences in the methods used to annotate TEs in a variety of genomes (especially non-autonomous elements), should be ruled out. We did not reannotate the SINEs in other genomes but revisited in the literature the global SINE density estimations for two vertebrate groups, ray-finned fish and tetrapods. No correlation between SINE abundance and genome size could be observed in these two groups (Figures 2 and S2). We did find correlations in tetrapods between the genome size and LTR or LINE densities and in fish with the DNA elements, but none of them subsisted when taking into account the phylogeny of the sampled species (Figure 2).

In conclusion, the originality of the larvacean context must be emphasized. There may be no better example of direct relationship between genome size and a measurable phenotypic trait, the body length, itself probably correlated to longevity and other life history parameters [26]. Interestingly, it was shown that, despite major changes of body length, the number of cells hardly differs among larvacean species, at least in one tissue [27].

Consequently, the body length strongly correlates with cell and nucleus size, whose relationship with genome size appears more direct.

For such changes of genome size, although active autonomous retrotransposons are maintained at a low copy number, non-autonomous elements may become essential dynamic contributors.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- CONTACT FOR REAGENT AND RESOURCE SHARING
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
- METHOD DETAILS
  - Genome sequencing and assembly
  - Phylogeny of larvacean species
  - Proportions of duplicated genes
  - Identification of TE-containing loci
  - Merging and confirmation of putative transposable elements
  - Phylogenetic reconstructions of TE families
  - Localization of TEs in the genomes and quantification of their genomic density
  - Correlation with genome size and phylogenetic contrasts
  - TE landscapes
- DATA AND SOFTWARE AVAILABILITY
  - Oikopleura longicauda
  - Bathochorddaeus sp
  - Mesochordaeus erythrocephalus
  - Oikopleura albicans
  - Oikopleura vanhoeffeni
  - Fritillaria borealis

SUPPLEMENTAL INFORMATION

Supplemental Information includes two figures, two tables, and two data files and can be found with this article online at https://doi.org/10.1016/j.cub.2019.01.080.

ACKNOWLEDGMENTS

We thank Marie Sémon for her expertise and help on phylogenetic contrast analysis and Emmanuel Quemener for helping us with the computational resources of the Centre Blaise Pascal at ENS de Lyon (http://www.cbp.ens-lyon.fr/doku.php). We thank Anne Aasjord for excellent technical assistance in the Sars Centre *Oikopleura* facility and the following scientists for coordinating or helping in the collection of other larvacean species: Steven Haddock (Monterey Bay Aquarium Research Institute); Don Deibel (Memorial University of Newfoundland); Fabien Lombard and Gaby Gorsky (Observatoire Oceanologique de Villefranche sur Mer); and Linda Holland (Scripps Institution of Oceanography). We thank the Genecore facility of EMBL (Heidelberg) for most Illumina sequencing and the Norwegian Sequencing Centre (University of Oslo) for additional PacBio sequencing. This project has been funded by two major grants of the Research Council of Norway, of which D.C. is the PI: 250005 accelerated evolution in chordates and the origin of larvaceans and 234817 Sars International Centre for Marine Molecular Biology Research, 2013–2022. This work was further supported by a grant overseen by the French National Research Agency (ANR-16-CE92-0019—EVOBOOSTER).

AUTHOR CONTRIBUTIONS

M.R. collected most larvacean species in Europe and North America, S.H. organized the genome sequencing, and S.S. carried out all steps of genome assembly. D.C. carried out analyses of the larvacean gene complement and species phylogeny. I.W. annotated most of the transposable elements in the seven genomes. M.N. synthesized the data and performed final analyses and comparisons of the TE content. S.H. performed part of the analysis of LTR elements. J.-N.V. and D.C. coordinated the work. The manuscript was co-written by M.N., J.-N.V., and D.C.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: August 12, 2018
Revised: November 26, 2018
Accepted: January 31, 2019
Published: March 14, 2019

REFERENCES

1. Petrov, D.A. (2001). Evolution of genome size: new approaches to an old problem. Trends Genet. *17*, 23–28.
2. Cavalier-Smith, T. (2005). Economy, speed and size matter: evolutionary forces driving nuclear genome miniaturization and expansion. Ann. Bot. *95*, 147–175.
3. Lynch, M. (2007). The frailty of adaptive hypotheses for the origins of organismal complexity. Proc. Natl. Acad. Sci. USA *104* (*Suppl 1*), 8597–8604.
4. Lynch, M., and Conery, J.S. (2003). The origins of genome complexity. Science *302*, 1401–1404.
5. Dufresne, F., and Jeffery, N. (2011). A guided tour of large genome size in animals: what we know and where we are heading. Chromosome Res. *19*, 925–938.
6. Piegu, B., Guyot, R., Picault, N., Roulin, A., Sanyal, A., Kim, H., Collura, K., Brar, D.S., Jackson, S., Wing, R.A., and Panaud, O. (2006). Doubling genome size without polyploidization: dynamics of retrotransposition-driven genomic expansions in *Oryza australiensis*, a wild relative of rice. Genome Res. *16*, 1262–1269.
7. Volff, J.-N., Lehrach, H., Reinhardt, R., and Chourrout, D. (2004). Retroelement dynamics and a novel type of chordate retrovirus-like element in the miniature genome of the tunicate *Oikopleura dioica*. Mol. Biol. Evol. *21*, 2022–2033.
8. Denoeud, F., Henriet, S., Mungpakdee, S., Aury, J.-M., Da Silva, C., Brinkmann, H., Mikhaleva, J., Olsen, L.C., Jubin, C., Cañestro, C., et al. (2010). Plasticity of animal genome architecture unmasked by rapid evolution of a pelagic tunicate. Science *330*, 1381–1385.
9. Bone, Q. (1998). The Biology of Pelagic Tunicates (Oxford University Press).
10. Sherlock, R.E., Waltz, K.R., and Robison, B.H. (2016). The first definitive record of the giant larvacean, *Bathochordaeus charon*, since its original description in 1900 and a range extension to the northeast Pacific Ocean. Mar. Biodivers. Rec. *9*, 79.
11. Putnam, N.H., Butts, T., Ferrier, D.E.K., Furlong, R.F., Hellsten, U., Kawashima, T., Robinson-Rechavi, M., Shoguchi, E., Terry, A., Yu, J.-K., et al. (2008). The amphioxus genome and the evolution of the chordate karyotype. Nature *453*, 1064–1071.
12. Simakov, O., Marletaz, F., Cho, S.-J., Edsinger-Gonzales, E., Havlak, P., Hellsten, U., Kuo, D.-H., Larsson, T., Lv, J., Arendt, D., et al. (2013). Insights into bilaterian evolution from three spiralian genomes. Nature *493*, 526–531.
13. Canapa, A., Barucca, M., Biscotti, M.A., Forconi, M., and Olmo, E. (2015). Transposons, genome size, and evolutionary insights in animals. Cytogenet. Genome Res. *147*, 217–239.

14. Chalopin, D., Naville, M., Plard, F., Galiana, D., and Volff, J.-N. (2015). Comparative analysis of transposable elements highlights mobilome diversity and evolution in vertebrates. Genome Biol. Evol. *7*, 567–580.
15. Katoh, K., and Standley, D.M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol. Biol. Evol. *30*, 772–780.
16. Price, M.N., Dehal, P.S., and Arkin, A.P. (2009). FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. Mol. Biol. Evol. *26*, 1641–1650.
17. Henriet, S., Sumic, S., Doufoundou-Guilengui, C., Jensen, M.F., Grandmougin, C., Fal, K., Thompson, E., Volff, J.-N., and Chourrout, D. (2015). Embryonic expression of endogenous retroviral RNAs in somatic tissues adjacent to the *Oikopleura* germline. Nucleic Acids Res. *43*, 3701–3711.
18. Malik, H.S., and Eickbush, T.H. (1998). The RTE class of non-LTR retrotransposons is widely distributed in animals and is the origin of many SINEs. Mol. Biol. Evol. *15*, 1123–1134.
19. Silva, R., and Burch, J.B. (1989). Evidence that chicken CR1 elements represent a novel family of retroposons. Mol. Cell. Biol. *9*, 3563–3566.
20. Wenke, T., Döbel, T., Sörensen, T.R., Junghans, H., Weisshaar, B., and Schmidt, T. (2011). Targeted identification of short interspersed nuclear element families shows their widespread existence and extreme heterogeneity in plant genomes. Plant Cell *23*, 3117–3128.
21. Bouallègue, M., Rouault, J.D., Hua-Van, A., Makni, M., and Capy, P. (2017). Molecular evolution of piggyBac superfamily: from selfishness to domestication. Genome Biol. Evol. *9*, 323–339.
22. Vassetzky, N.S., and Kramerov, D.A. (2013). SINEBase: a database and tool for SINE analysis. Nucleic Acids Res. *41*, D83–D89.
23. Doucet, A.J., Wilusz, J.E., Miyoshi, T., Liu, Y., and Moran, J.V. (2015). A 3′ poly(A) tract is required for LINE-1 retrotransposition. Mol. Cell *60*, 728–741.
24. Okada, N., Hamada, M., Ogiwara, I., and Ohshima, K. (1997). SINEs and LINEs share common 3′ sequences: a review. Gene *205*, 229–243.
25. Dewannieux, M., and Heidmann, T. (2005). L1-mediated retrotransposition of murine B1 and B2 SINEs recapitulated in cultured cells. J. Mol. Biol. *349*, 241–247.
26. Gregory, T.R., Hebert, P.D., and Kolasa, J. (2000). Evolutionary implications of the relationship between genome size and body size in flatworms and copepods. Heredity (Edinb) *84*, 201–208.
27. Spriet, E. (1997). Studies on the house building epithelium of *Oikopleurid appendicularia* (Tunicata): early differentiation and description of the adult pattern of oikoplast cells. Master's thesis (University of Bergen).
28. Sherlock, R.E., Walz, K.R., Schlining, K.L., and Robison, B.H. (2017). Morphology, ecology, and molecular biology of a new species of giant larvacean in the eastern North Pacific: *Bathochordaeus mcnutti* sp. nov. Mar. Biol. *164*, 20.
29. Llorens, C., Futami, R., Covelli, L., Domínguez-Escribá, L., Viu, J.M., Tamarit, D., Aguilar-Rodríguez, J., Vicente-Ripolles, M., Fuster, G., Bernet, G.P., et al. (2011). The Gypsy Database (GyDB) of mobile genetic elements: release 2.0. Nucleic Acids Res. *39*, D70–D74.
30. Gnerre, S., Maccallum, I., Przybylski, D., Ribeiro, F.J., Burton, J.N., Walker, B.J., Sharpe, T., Hall, G., Shea, T.P., Sykes, S., et al. (2011). High-quality draft assemblies of mammalian genomes from massively parallel sequence data. Proc. Natl. Acad. Sci. USA *108*, 1513–1518.
31. Andrews, S. (2010). FASTQC. A quality control tool for high throughput sequence data. http://www.bioinformatics.babraham.ac.uk/projects/fastqc.
32. Kajitani, R., Toshimoto, K., Noguchi, H., Toyoda, A., Ogura, Y., Okuno, M., Yabana, M., Harada, M., Nagayasu, E., Maruyama, H., et al. (2014). Efficient de novo assembly of highly heterozygous genomes from whole-genome shotgun short reads. Genome Res. *24*, 1384–1395.
33. Huang, X., and Madan, A. (1999). CAP3: a DNA sequence assembly program. Genome Res. *9*, 868–877.
34. Boetzer, M., Henkel, C.V., Jansen, H.J., Butler, D., and Pirovano, W. (2011). Scaffolding pre-assembled contigs using SSPACE. Bioinformatics *27*, 578–579.
35. Boetzer, M., and Pirovano, W. (2012). Toward almost closed genomes with GapFiller. Genome Biol. *13*, R56.
36. Nadalin, F., Vezzi, F., and Policriti, A. (2012). GapFiller: a de novo assembly approach to fill the gap within paired reads. BMC Bioinformatics *13* (*Suppl 14*), S8.
37. Kumar, S., Jones, M., Koutsovoulos, G., Clarke, M., and Blaxter, M. (2013). Blobology: exploring raw genome data for contaminants, symbionts and parasites using taxon-annotated GC-coverage plots. Front. Genet. *4*, 237.
38. Bolger, A.M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics *30*, 2114–2120.
39. Bankevich, A., Nurk, S., Antipov, D., Gurevich, A.A., Dvorkin, M., Kulikov, A.S., Lesin, V.M., Nikolenko, S.I., Pham, S., Prjibelski, A.D., et al. (2012). SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J. Comput. Biol. *19*, 455–477.
40. Chakraborty, M., Baldwin-Brown, J.G., Long, A.D., and Emerson, J.J. (2016). Contiguous and accurate de novo assembly of metazoan genomes with modest long read coverage. Nucleic Acids Res. *44*, e147.
41. Koren, S., Walenz, B.P., Berlin, K., Miller, J.R., Bergman, N.H., and Phillippy, A.M. (2017). Canu: scalable and accurate long-read assembly via adaptive *k*-mer weighting and repeat separation. Genome Res. *27*, 722–736.
42. Boetzer, M., and Pirovano, W. (2014). SSPACE-LongRead: scaffolding bacterial draft genomes using long read sequence information. BMC Bioinformatics *15*, 211.
43. English, A.C., Richards, S., Han, Y., Wang, M., Vee, V., Qu, J., Qin, X., Muzny, D.M., Reid, J.G., Worley, K.C., and Gibbs, R.A. (2012). Mind the gap: upgrading genomes with Pacific Biosciences RS long-read sequencing technology. PLoS ONE *7*, e47768.
44. Dereeper, A., Guignon, V., Blanc, G., Audic, S., Buffet, S., Chevenet, F., Dufayard, J.-F., Guindon, S., Lefort, V., Lescot, M., et al. (2008). Phylogeny.fr: robust phylogenetic analysis for the non-specialist. Nucleic Acids Res. *36*, W465–W469.
45. Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. J. Mol. Biol. *215*, 403–410.
46. Eddy, S.R. (2011). Accelerated profile HMM searches. PLoS Comput. Biol. *7*, e1002195.
47. Ellinghaus, D., Kurtz, S., and Willhoeft, U. (2008). LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. BMC Bioinformatics *9*, 18.
48. Xu, Z., and Wang, H. (2007). LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. Nucleic Acids Res. *35*, W265–W268.
49. Han, Y., and Wessler, S.R. (2010). MITE-Hunter: a program for discovering miniature inverted-repeat transposable elements from genomic sequences. Nucleic Acids Res. *38*, e199.
50. Lowe, T.M., and Chan, P.P. (2016). tRNAscan-SE On-line: integrating search and context for analysis of transfer RNA genes. Nucleic Acids Res. *44*, W54–W57.
51. Smit, A.F.A., Hubley, R., and Green, P. (2013). RepeatMasker Open-4.0. http://www.repeatmasker.org/.
52. Hopcroft, R.R., and Robison, B.H. (1999). A new mesopelagic larvacean, *Mesochordaeus erythrocephalus*, sp. nov., from Monterey Bay, with a description of its filtering house. J. Plankton Res. *21*, 1923–1937.
53. Troedsson, C., Ganot, P., Bouquet, J.M., Aksnes, D.L., and Thompson, E.M. (2007). Endostyle cell recruitment as a frame of reference for development and growth in the urochordate *Oikopleura dioica*. Biol. Bull. *213*, 325–334.
54. Fenaux, R. (1966). Les Appendiculaires des mers d'Europe et du Bassin Méditerranéen (Masson).

55. Choe, N., and Deibel, D. (2011). Life history characters and population dynamics of the boreal larvacean *Oikopleura vanhoeffeni* (Tunicata) in Conception Bay, Newfoundland. J. Mar. Biol. Assoc. U.K. *91*, 1587–1598.

56. Ganot, P., Bouquet, J.M., and Thompson, E.M. (2006). Comparative organization of follicle, accessory cells and spawning anlagen in dynamic semelparous clutch manipulators, the urochordate Oikopleuridae. Biol. Cell *98*, 389–401.

57. Flood, P.R. (2003). House formation and feeding behaviour of *Fritillaria borealis* (Appendicularia: Tunicata). Mar. Biol. *143*, 467–475.

58. Lamble, S., Batty, E., Attar, M., Buck, D., Bowden, R., Lunter, G., Crook, D., El-Fahmawi, B., and Piazza, P. (2013). Improved workflows for high throughput library preparation using the transposome-based Nextera system. BMC Biotechnol. *13*, 104.

59. Bao, W., Kojima, K.K., and Kohany, O. (2015). Repbase update, a database of repetitive elements in eukaryotic genomes. Mob. DNA *6*, 11.

60. Wheeler, T.J., Clements, J., Eddy, S.R., Hubley, R., Jones, T.A., Jurka, J., Smit, A.F.A., and Finn, R.D. (2013). Dfam: a database of repetitive DNA based on profile hidden Markov models. Nucleic Acids Res. *41*, D70–D82.

61. Feschotte, C., and Pritham, E.J. (2007). DNA transposons and the evolution of eukaryotic genomes. Annu. Rev. Genet. *41*, 331–368.

62. Xiong, W., He, L., Lai, J., Dooner, H.K., and Du, C. (2014). HelitronScanner uncovers a large overlooked cache of Helitron transposons in many plant genomes. Proc. Natl. Acad. Sci. USA *111*, 10263–10268.

63. Wicker, T., Sabot, F., Hua-Van, A., Bennetzen, J.L., Capy, P., Chalhoub, B., Flavell, A., Leroy, P., Morgante, M., Panaud, O., et al. (2007). A unified classification system for eukaryotic transposable elements. Nat. Rev. Genet. *8*, 973–982.

64. Di Tommaso, P., Moretti, S., Xenarios, I., Orobitg, M., Montanyola, A., Chang, J.M., Taly, J.F., and Notredame, C. (2011). T-Coffee: a web server for the multiple sequence alignment of protein and RNA sequences using structural information and homology extension. Nucleic Acids Res. *39*, W13–W17.

65. Sievers, F., and Higgins, D.G. (2014). Clustal omega. Curr. Protoc. Bioinformatics *48*, 3.13.1–3.13.16.

66. Yang, Z. (2007). PAML 4: phylogenetic analysis by maximum likelihood. Mol. Biol. Evol. *24*, 1586–1591.

67. Amemiya, C.T., Alföldi, J., Lee, A.P., Fan, S., Philippe, H., Maccallum, I., Braasch, I., Manousaki, T., Schneider, I., Rohner, N., et al. (2013). The African coelacanth genome provides insights into tetrapod evolution. Nature *496*, 311–316.

68. Herrero, J., Muffato, M., Beal, K., Fitzgerald, S., Gordon, L., Pignatelli, M., Vilella, A.J., Searle, S.M.J., Amode, R., Brent, S., et al. (2016). Ensembl comparative genomics resources. Database (Oxford) *2016*, bav096.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Critical Commercial Assays		
Nextera DNA Library Prep Kit	Illumina	FC-121-1030
MiSeq Reagent Kit v3 (600-cycle)	Illumina	MS-102-3003
Deposited Data		
Raw and analyzed data	This paper	GenBank: SCLD00000000-SCLH01000000; SDII01000000
		Data S1
		Data S2
coxI DNA sequences for *Bathochordaeus*	[10]	GenBank: KX599256-KX599281; KT881543–KT881545
	[28]	
Oikopleura dioica genome	[8]	http://www.genoscope.cns.fr/externe/Download/Projets/Projet_HG/data/assembly/
Gypsy Database	[29]	http://www.gydb.org
REPBASE	Genetic Information Research Institute	https://www.girinst.org/repbase/
GtRNAdb	The Lowe lab	http://gtrnadb.ucsc.edu/
RFAM	EMBL-EBI	http://rfam.xfam.org/
SINE Base	Engelhardt Institute of Molecular Biology	http://sines.eimb.ru/
Biological Samples		
Bathochordaeus sp.	Steven Haddock, Monterey Bay	N/A
Mesochordaeus erythrocephalus	Steven Haddock, Monterey Bay	N/A
Oikopleura vanhoeffeni	Don Deibel, Witless Bay	N/A
Fritillaria borealis	Marine Biological Station, Espegrend and Rosslandspollen, Rossland	N/A
Oikopleura albicans	Fabien Lombard and Gaby Gorsky, Villefranche-sur-mer	N/A
Oikopleura longicauda	Linda Holland, La Jolla	N/A
Software and Algorithms		
ALLPATHS-LG	[30]	ftp://ftp.broadinstitute.org/pub/crd/ALLPATHS/Release-LG/
FASTQC	[31]	https://www.bioinformatics.babraham.ac.uk/projects/download.html#fastqc
PLATANUS	[32]	http://platanus.bio.titech.ac.jp/platanus-assembler/platanus-1-2-4
CAP3	[33]	http://seq.cs.iastate.edu/cap3.html
SSPACE	[34]	https://www.baseclear.com/services/bioinformatics/basetools/sspace-standard/
GapFiller	[35, 36]	https://www.baseclear.com/services/bioinformatics/basetools/gapfiller/
Blobology	[37]	https://github.com/blaxterlab/blobology
Trimmomatic	[38]	http://www.usadellab.org/cms/index.php?page=trimmomatic
Spades	[39]	https://github.com/ablab/spades
Quickmerge	[40]	https://github.com/mahulchak/quickmerge
CANU	[41]	https://github.com/marbl/canu
SSPACE-LongRead	[42]	https://www.baseclear.com/services/bioinformatics/basetools/sspace-longread/
PBSuite	[43]	https://sourceforge.net/p/pb-jelly/wiki/Home/

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
MUSCLE	[44]	http://www.phylogeny.fr/simple_phylogeny.cgi
Gblocks	[44]	http://www.phylogeny.fr/simple_phylogeny.cgi
PhyML	[44]	http://www.phylogeny.fr/simple_phylogeny.cgi
BLAST	[45]	https://blast.ncbi.nlm.nih.gov/Blast.cgi?CMD=Web&PAGE_TYPE=BlastDocs&DOC_TYPE=Download
HMMER	[46]	
LTRHarvest	[47]	http://genometools.org/tools/gt_ltrharvest.html
LTRFinder	[48]	
MITE-Hunter	[49]	http://target.iplantcollaborative.org/mite_hunter.html
Sine-Finder	[20]	
tRNAscan-SE 2.0	[50]	http://lowelab.ucsc.edu/tRNAscan-SE/
RepeatMasker	[51]	http://www.repeatmasker.org
Mafft	[15]	https://mafft.cbrc.jp/alignment/software/

CONTACT FOR REAGENT AND RESOURCE SHARING

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Daniel Chourrout (daniel.chourrout@uib.no).

EXPERIMENTAL MODEL AND SUBJECT DETAILS

The *Oikopleura dioica* genome dataset was produced from an inbred line resulting from successive sib-matings [8]. Six new larvacean species were collected for this study: *Oikopleura albicans* was collected near Villefranche-sur-Mer (France) and sequencing was performed on DNA from a single adult male. Similarly, DNA from single animals were used for *Oikopleura vanhoeffeni* (collected in Newfoundland, Canada), *Bathochordaeus sp.* and *Mesochordaeus erythrocephalus* (collected in Monterey, California). Due to the small size of individuals, pools of animals had to be used for DNA extraction in *Fritillaria borealis* (approx. 110 specimens collected in Bergen, Norway) and *Oikopleura longicauda* (47 specimens collected at La Jolla, California) to obtain a sufficient amount of DNA for constructing genomic libraries.

The "giant" appendicularians from genera *Mesochordaeus* and *Bathochordaeus* were collected using a remotely operated underwater vehicle. *Mesochordaeus erythrocephalus* was identified based on its distinctive red pigmentation [52]. At the time of collection, the classification of the *Bathochordaeus* genus was still under debate and the captured animal could not be clearly identified. A recent work allowed us to distinguish the species *B. charon*, *B. mcnutti* and *B. stygius* based on the comparison of several specimens [10, 28]. The homology to molecular barcodes (mitochondrial cytochrome *c* oxidase subunit I gene) suggests that our *Bathochordaeus* specimen might correspond to *B. stygius*. For *Oikopleura longicauda*, the genome assembly revealed that two very closely related species or subspecies, which were not recognized morphologically, were mixed before DNA extraction and sequencing. As the sequencing coverage for each of them seemed similar, the genome size was estimated by half of the total genome assembly size.

Animal body sizes refer to adult trunk length, measured in various studies: 0.9 mm for *O. dioica* [53], 3.4 mm for *O. albicans* [54], 3.4 mm for *O. vanhoeffeni* [55], 0.95 mm for *O. longicauda* [56], 7.3 mm for *M. erythrocephalus* [52] and 0.9 mm for *F. borealis* [57]. For *Bathochordaeus sp.*, we used the size reported for *B. stygius* [28].

METHOD DETAILS

Genome sequencing and assembly

Genomic DNA was purified with the the Chomczynski procedure. Sequencing libraries were prepared at Genecore (EMBL, Heidelberg) and sequenced on HiSeq 2000, producing paired-end Illumina reads with average length of 100bp.

Error Correction script, which is a part of the ALLPATHS-LG software [30], was used to trim the paired Illumina reads from adaptor sequences and perform reads correction. The quality of the resulting reads was confirmed with FASTQC [31]. PLATANUS assembler was chosen to assemble the Illumina reads, given its ability to perform well on highly polymorphic datasets [32]. Upon manual inspection, certain artifacts were detected in the genome assemblies, likely due to the high rate of polymorphism and pooled samples. Artifacts consisted of concatenated haplotypes connected with a single nucleotide gap (N).

In-house scripts were used to break those scaffolds at single nucleotide gaps, merging the adjacent contigs when their ends overlapped with CAP3 software [33]. The resulting contigs and scaffolds were further scaffolded with the trimmed and corrected Illumina

reads through several rounds of scaffolding with SSPACE software [34]. The remaining gaps were filled using GapFiller [35, 36]. The assemblies were subsequently checked for contamination using the Blobology tool [37], removing scaffolds that were suspected to be a product of contamination. The resulting assembly sizes for *F. borealis, O. longicauda, Bathochordaeus sp. and M. erythrocephalus* were 143.1Mb, 308.8Mb, 396.5Mb and 874.0Mb, respectively. Scaffold N50 values were 2.8kb, 2.4kb, 1.3kb and 1.9kb, respectively.

That a certain level of redundancy can remain after assembling genomic reads with usual tools is expected due to the allelic polymorphism, which is increased if a pool of individuals and not a single animal has been used to provide the genomic DNA. We measured in each genome the number of occurrences of individual homeobox genes and indeed found a significant redundancy for *F. borealis*. To precisely determine the genome size and improve the assembly, we later on sequenced the genome of a single animal. In this purpose, we prepared Nextera libraries using a low-input protocol [58] and we sequenced the DNA with the MiSeq System. Obtained paired-end reads were trimmed from adaptor sequences using the Trimmomatic tool [38]. The quality of the resulting reads was assessed with FASTQC [31]. All reads whose length was at least 36bp were subsequently assembled with the Spades genome assembler [39]. The assembly was then checked for contamination using Blobology [37], removing 8.64Mb due to the suspected contamination. After several rounds of scaffolding with SSPACE [34] and gap-filling with GapFiller [35, 36], the assembly size was 84Mb and the N50 4532bp. To consolidate the two *F. borealis* assemblies, the Quickmerge program was used [40] with the single animal assembly as a query. The final assembly was obtained by scaffolding with SSPACE [34] and gap-filling with GapFiller [35, 36], producing a 92.4Mb assembly with a N50 of 10.6kb.

To further improve the assembly contingency of *O. albicans* and *O. vanhoeffeni*, long jumping distance (LJD) libraries were produced by sequencing DNA samples of pooled adult animals (20 and 7 specimens, respectively). The sequencing was done by Eurofins Genomics (Ebersberg, Germany). For *O. albicans*, 3kb and 8kb Illumina mate pair reads were obtained by HiSeq 2000 sequencing, while 3kb, 8kb and 20kb were obtained for *O. vanhoeffeni*. The libraries were trimmed from adapters using Trimmomatic [38] and subsequently used to scaffold the assemblies with the SSPACE software [34]. The gaps were filled with all the available Illumina reads for each animal using 10 rounds of GapFiller [35, 36].

PacBio libraries were also obtained for *O. albicans* and *O. vanhoeffeni* by sequencing single adult animals. Libraries were prepared using Pacific Biosciences 10-20 kb library preparation protocol for low input DNA. The library was sequenced on Pacific Biosciences RS II instrument using P6-C4 chemistry. The reads were corrected and trimmed using the CANU assembler [41]. The resulting coverage was not sufficient for PacBio-only or hybrid PacBio-Illumina assembly, so the reads were used for scaffolding and gap-filling only. Several rounds of scaffolding were done using the SSPACE-LongRead software [42]. The remaining gaps were filled with Illumina reads using GapFiller [35, 36] and PacBio reads using the PBJelly tool from the PBSuite software [43]. As for other species, most gaps were filled using Illumina reads. Consequently, using PacBio for two of the genome assemblies cannot have biased the comparison of TE coverage among species. The size of the resulting *O. albicans* assembly was 356.9Mb and *O. vanhoeffeni* assembly 632.2Mb. Scaffold N50 values were 209.7Kb and 255.4Kb, respectively.

Phylogeny of larvacean species

The phylogeny of larvaceans was obtained using a set of 13 concatenated ribosomal proteins annotated in the different genomes (RPS6, RPS7, RPS8, RPS19, RPL5, RPL8, RPL10A, RPL11, RPL15, RPL18, RPL23A, RPL26, RPL27A; supp. materials), submitted to the Phylogeny.fr website (‘‘one click’’ analysis with default parameters: multiple alignment using Muscle, site selection using Gblocks, and phylogenetic reconstruction using PhyML) [44]. After site selection, a set of 1982 amino acids was finally used for the phylogeny (Figure 1A). Data from two additional species recently sequenced but not addressed in this study were added to increase the robustness of the tree (*Fritillaria pellucida* and *Appendicularia sicula*). They as expected grouped with *F. borealis*. Sea urchin (as a non-chordate deuterostome) and lancelet (as a cephalochordate) were used as outgroups in this phylogenetic analysis.

Proportions of duplicated genes

To check whether or not large scale or whole genome duplications may have caused the expansion of some of the larvacean genomes, the duplication status of 200 conserved genes was examined for each species except *O. longicauda* (due to the mixture of two (sub-) species prior to sequencing) using reciprocal Blast search. These genes were searched using query sequences that are part of a dataset of 7368 putative proteins from *Nematostella vectensis* assumed to be conserved in ancestral bilaterians and identified through the hierarchical ‘‘metazome’’ clustering approach [11, 12]. To avoid multi-gene families, the dataset was first strongly reduced to protein sequences showing no significant homology with others of the dataset, using BlastP. It was further restricted to segments of sequences highly conserved in *O. dioica* (TBlastN). These two reductions of the dataset led to a collection of 533 query sequences that were aligned with each of six larvacean genome assemblies *(O. dioica, O. albicans, O. vanhoeffeni, Bathochordaeus sp., M. erythrocephalus, F. borealis)* using TBlastN. Segments of scaffolds from the genome assemblies that matched this collection (TBlastN hit longer than 30 residues with at least 50% identity) were retrieved for reciprocal BlastX. An important proportion of the hits were multiple due to their fragmentation by introns. In almost all cases, reciprocal BlastX on the entire bilaterian dataset showed that the selected scaffold segment contained orthologs of the initial query sequence from the reduced collection. That query sequence often matched more than one scaffold in a given genome. Multiple matching scaffolds could represent either true gene duplicates or alleles retained during the genome assembly process, imposing a careful examination of each hit. Decision for gene duplication was based on the detection of at least two amino-acid substitutions in one exon (alignment ends, near the intron-exon border are not considered). In most cases, gene duplication was ruled out because the exon sequence was invariable. Genes

were removed from the survey in ambiguous cases. A parallel BlastX on NCBI nr-aa database led to exclude a few genome scaffolds that gave top hits with non-animal sequences and thus might result from cross-genome contaminations. We ended up with a sample of 200 sequences for which orthologs were found in the six species (Figure S1).

Identification of TE-containing loci

Two general strategies were adopted to identify TEs within genomes: i1) comparative approaches using protein sequences from known transposable elements, and i2) *de novo* approaches detecting structural features of TEs. The second approach is particularly important for elements such as SINEs and MITEs that do not encode proteins.

Comparative approaches using protein sequences from known TEs

An exhaustive collection of protein sequences representing superfamilies listed in Chalopin et al. [14] was compiled, composed of: an in house list of sequences, consensus protein sequences for domains found in LTR retroelements from the Gypsy Database (http://gydb.org [29];), and sequences from REPBASE from closely related species [59]. These were searched against the genomes from the seven species using TBlastN (E-value < 10^{-5}). DNA sequences available from previous studies on *O. dioica* [7] were also used as queries for a BlastN search (E-value < 10^{-5}). In addition, Hidden Markov Models (HMMs) for conserved amino-acid motifs found in TEs were collected from the Gypsy Database and Dfam [60]. These were searched against the genomes using HMMER against with a threshold of e^{-5} [46]. Hits obtained with TBlastN, BlastN and HMMs were finally merged.

De novo approaches using structural features

Potential TEs can be detected using structural features such as the LTRs found in LTR retroelements or inverted repeats in DNA elements. LTRHarvest [47] and LTRFinder [48] were used to identify LTR elements. Potential MITEs were located with MITE-Hunter [49]. These predicted MITEs present structural homogeneity, terminal inverted repeats flanked by target site duplications, and they show multiple interspersed copies. Comparing these MITE sequences with autonomous DNA elements, we could not connect any of them. This is not completely surprising as such functional link can be restricted to few nucleotides of the TIRs only [61]. Sine-Finder [20] allowed detecting potential SINE elements using predicted RNA structures. HelitronScanner [62] was used to identify Helitrons using 5′ and 3′ terminal motifs. Different filters were applied to the outputs. For MITE-Hunter, singlet and compound elements were excluded, as recommended by the authors. For Sine-Finder, potential elements were compared against the tRNA database (http://gtrnadb.ucsc.edu/) and also passed through the software tRNAscan-SE 2.0 to identify any genuine tRNAs [50]. We also analyzed each genome with tRNAscan-SE 2.0 using Eukaryotic search mode with Infernal First Pass. The output was used for searching tRNA-like motifs in our SINE database with BlastN. Hits were considered positive only if the similarity to a tRNA was conserved among several copies of the same SINE family. rRNA (including 5S) and snRNA genes were identified by BlastN analysis in genomes and compared with SINE sequences. Similarity with other non-coding RNA molecules were also searched using the RFAM database (http://rfam.xfam.org/). The putative SINE elements were searched against genomes using BlastN; single/very low copy (< 10) putative SINEs were excluded. SINEs were further analyzed by Blast comparison to SINEBase [22] and to larvacean non-LTR retrotransposon sequences identified in this work.

Merging and confirmation of putative transposable elements

All the putative TE loci identified in step one were collected together using custom python scripts. Overlapping genomic loci with evidence of TEs were merged. The DNA sequences were then extracted for all loci and compared to each other using BlastN. Loci that shared 80% (or more) sequence identity in at least 80% of their sequences were grouped together in a same TE family [63]. Loci that did not group with any other sequence and were under 500bp long were excluded from further analysis.

In the first step of confirming and identifying the putative TEs as real TEs, predicted coding sequences were extracted for each loci using different protein-coding sequences as templates to predict the location and open reading frames. These included: RTs for TOR, Ty3/Gypsy and DIRS1 LTR retro-elements; RTs from LINE2, Nimb, I, Penelope, CR1, and R2 non-LTR retroelements; transposases from piggybac, mariner, and hAT DNA transposons; integrase genes from Polintons/Mavericks; and various conserved protein sequences from Helitrons. Predicted protein sequences were also searched using BlastP against the NCBI non-redundant protein database and protein sequences from REPBASE to provide further evidence that they were coding for transposable element sequences. In order to classify them, proteins were aligned with known TE proteins using T-coffee [64] and phylogenetic trees were drawn using FastTree [16]. Representative sequences from each TE family (individual sequences that are the most complete according to the structural features of the family) are provided as supplementary materials.

Once the superfamily for each element had been identified, the previously identified groups of TEs were used to determine the full length of the elements, to identify structural features, and to pull out further open reading frames. Loci within a group of TEs along with up to 10kb of their flanking sequences were searched against each other to identify the limits of the elements. These were verified by manual inspection of alignments with T-Coffee [64] (or Clustal-Omega for larger alignments [65]). Target site duplications (TSDs) were identified where possible. The loci were also searched against themselves using BlastN in order to identify inverted repeats and LTRs, and these data were combined with the outputs from LTRFinder and LTRHarvest to locate such features.

Phylogenetic reconstructions of TE families

Protein sequences of TEs were obtained by using previously predicted proteins as queries for a TBlastN search against the genomic insertions; targeted protein sequences were then extracted from the Blast result file. Multiple sequence alignments were computed

with Mafft [15]. Phylogenetic reconstructions were obtained using FastTree [16] with default parameters. Branch lengths and evolutionary rates in the different clusters of Odin1/Odin2 were computed using the PAML codeml program [66].

Codeml.ctl parameters file

```
seqfile = sites.phy * sequence data filename
treefile = tree * tree structure file name
outfile = outfile * main result file name
noisy = 9 * 0,1,2,3,9: how much rubbish on the screen
verbose = 1 * 0: concise; 1: detailed, 2: too much
runmode = 0 * 0: user tree; 1: semi-automatic; 2: automatic
* 3: StepwiseAddition; (4,5):PerturbationNNI; −2: pairwise
seqtype = 1 * 1:codons; 2:AAs; 3:codons→AAs
CodonFreq = 2 * 0:1/61 each, 1:F1X4, 2:F3X4, 3:codon table
* ndata = 10
clock = 0 * 0:no clock, 1:clock; 2:local clock; 3:CombinedAnalysis
aaDist = 0 * 0:equal, +:geometric; -:linear, 1-6:G1974,Miyata,c,p,v,a
aaRatefile = dat/jones.dat * only used for aa seqs with model = empirical(_F)
* dayhoff.dat, jones.dat, wag.dat, mtmam.dat, or your own
model = 0
* models for codons:
* 0:one, 1:b, 2:2 or more dN/dS ratios for branches
* models for AAs or codon-translated AAs:
* 0:poisson, 1:proportional, 2:Empirical, 3:Empirical+F
* 6:FromCodon, 7:AAClasses, 8:REVaa_0, 9:REVaa(nr = 189)
NSsites = 0 * 0:one w;1:neutral;2:selection; 3:discrete;4:freqs;
* 5:gamma;6:2gamma;7:beta;8:beta&w;9:betaγ
* 10:beta&gamma+1; 11:beta&normal > 1; 12:0&2normal > 1;
* 13:3normal > 0
icode = 0 * 0:universal code; 1:mammalian mt; 2-10:see below
Mgene = 0
* codon: 0:rates, 1:separate; 2:diff pi, 3:diff kapa, 4:all diff
* AA: 0:rates, 1:separate
fix_kappa = 0 * 1: kappa fixed, 0: kappa to be estimated
kappa = 2 * initial or fixed kappa
fix_omega = 0 * 1: omega or omega_1 fixed, 0: estimate
omega = 0.4 * initial or fixed omega, for codons or codon-based AAs
fix_alpha = 1 * 0: estimate gamma shape parameter; 1: fix it at alpha
alpha = 0. * initial or fixed alpha, 0:infinity (constant rate)
Malpha = 0 * different alphas for genes
ncatG = 8 * # of categories in dG of NSsites models
getSE = 0 * 0: don't want them, 1: want S.E.s of estimates
RateAncestor = 1 * (0,1,2): rates (alpha > 0) or ancestral states (1 or 2)
Small_Diff = 0.5e-6
cleandata = 1 * remove sites with ambiguity data (1:yes, 0:no)?
* fix_blength = −1 * 0: ignore, −1: random, 1: initial, 2: fixed
method = 0 * Optimization method 0: simultaneous; 1: one branch a time
* Genetic codes: 0:universal, 1:mammalian mt., 2:yeast mt., 3:mold mt.,
* 4: invertebrate mt., 5: ciliate nuclear, 6: echinoderm mt.,
* 7: euplotid mt., 8: alternative yeast nu. 9: ascidian mt.,
* 10: blepharisma nu.
* These codes correspond to transl_table 1 to 11 of GENEBANK.
```

Localization of TEs in the genomes and quantification of their genomic density

In order to quantify the total amount of TEs in each species and to get TE positions, representative TEs identified in the previous part were masked for low complexity regions and subsequently used as repeat database for a RepeatMasker search in the genomes (http://www.repeatmasker.org [51],). The masking of low complexity regions in the database sequences avoids non-specific hits when masking the genome. Hits shorter than 100 bp (70 bp for SINEs and MITEs) were also filtered out before calculating TE densities.

Correlation with genome size and phylogenetic contrasts

Correlation between genome size and TE content was tested by taking into account the phylogenetic relationships of the species studied. The phylogeny of tetrapods was retrieved from Amemiya et al. [67], the phylogeny of fish from Ensembl Compara data [68]; the phylogeny of larvaceans was reconstructed as explained previously. These phylogenies are provided below:

Larvacean tree

(SP:0.08993338355,(((((FB:0.1174961415,FP:0.0750021485)1.0000000000:0.0644759946,AS:0.1792786955)1.0000000000:0.2189967213,(((BC:0.0237717394,ME:0.0323069774)1.0000000000:0.0285556477,OL:0.0504900656)1.0000000000:0.0423280470,((OV:0.0697663542,OA:0.0472063752)0.9740000000:0.0197939737,OD:0.1437188750)0.6260000000:0.0269701350)1.0000000000:0.1606273136)1.0000000000:0.0850913913,CI:0.1395921377)0.9980000000:0.0644955878,BF:0.1400621921):0.08993338355);

Tetrapods tree

(((((((H.sapiens:0.042,M.musculus:0.115):0.142,M.domestica:0.148):0.039,O.anatinus:0.158):0.124,((T.guttata:0.091,G.gallus:0.067):0.112,A.carolinensis:0.255):0.045):0.097,X.tropicalis:0.5):0.155,L.chalumnae:0.279):0.03,((tilapia:0.115,puffer:0.2):0.23,zebrafish:0.267):0.533):0.015,shark:0.524);

Fish tree

(((((((Xiphophorus_maculatus:0.10185,Poecilia_formosa:0.10185):0.10185,Oryzias_latipes:0.14):0.0483,Gasterosteus_aculeatus:0.1836):0.0010,Oreochromis_niloticus:0.1728):0.0319,(Takifugu_rubripes:0.1096,Tetraodon_nigroviridis:0.1227):0.136):0.0411,Gadus_morhua:0.1965):0.0574,(Astyanax_mexicanus:0.1487,Danio_rerio:0.2053):0.0653):0.107,Lepisosteus_oculatus:0.107);

The effect of the phylogeny on the genome size versus TE content relation was tested using the R *caper* package (https://CRAN.R-project.org/package=caper). Fitting a linear model on the data, we obtained a *lambda* parameter of 0 for tetrapods and larvaceans, meaning that covariances between taxa are negligible in these clades and thus that correlations are not biased by the phylogeny. In contrast, the *lambda* parameter obtained in fish was 1, indicating that evolutionary relationships between some of the fish species surveyed induced a bias in the correlation calculation. We then computed corrected Pearson's correlation using the R *ape* package (https://CRAN.R-project.org/package=ape). The R *phytools* package was used to plot the modeled evolution of variables along trees (https://CRAN.R-project.org/package=phytools).

TE landscapes

For each TE family identified previously, all genomic insertions were retrieved and aligned together using Mafft [15]. Global DNA sequence identity was then computed for each possible pair of sequences, excluding gaps. Landscape graphs were drawn by reporting the total number of pairwise comparisons for a given family to the total genomic density of this family.

DATA AND SOFTWARE AVAILABILITY

Oikopleura longicauda

This Whole Genome Shotgun project has been deposited at DDBJ/ENA/GenBank under the accession SCLD00000000. The version described in this paper is version SCLD01000000.

Bathochorddaeus sp

This Whole Genome Shotgun project has been deposited at DDBJ/ENA/GenBank under the accession SCLE00000000. The version described in this paper is version SCLE01000000.

Mesochordaeus erythrocephalus

This Whole Genome Shotgun project has been deposited at DDBJ/ENA/GenBank under the accession SCLF00000000. The version described in this paper is version SCLF01000000.

Oikopleura albicans

This Whole Genome Shotgun project has been deposited at DDBJ/ENA/GenBank under the accession SCLG00000000. The version described in this paper is version SCLG01000000.

Oikopleura vanhoeffeni

This Whole Genome Shotgun project has been deposited at DDBJ/ENA/GenBank under the accession SCLH00000000. The version described in this paper is version SCLH01000000.

Fritillaria borealis

This Whole Genome Shotgun project has been deposited at DDBJ/ENA/GenBank under the accession SDII00000000. The version described in this paper is version SDII01000000.

Representative TE sequences have been deposited to Repbase.