ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

DOTTORATO DI RICERCA IN

SCIENZE DELLA TERRA, DELLA VITA E DELL'AMBIENTE

Ciclo 36

Settore Concorsuale: 05/B1 - ZOOLOGIA E ANTROPOLOGIA

Settore Scientifico Disciplinare: BIO/05 - ZOOLOGIA

NEGLECTED GENOMIC ELEMENTS IN OVERLOOKED TAXONOMIC GROUPS: DIVERSITY, EVOLUTION, AND GENOMIC IMPACT OF TRANSPOSABLE ELEMENTS IN BIVALVE MOLLUSCS

Presentata da: Jacopo Martelossi

Coordinatore Dottorato

Barbara Cavalazzi

Supervisore

Fabrizio Ghiselli

Co-supervisore

Liliana Milani

Esame finale anno 2024

"In today's vastly expanded scientific enterprise, obsessed with impact factors and competition, we will need much more night science to unveil the many mysteries that remain about the workings of organisms." – François Jacob, Science, 2011

Abstract

Transposable elements (TEs) are intriguing features found in eukaryotic genomes, capable of replicating independently within the host cell and spreading throughout the genome. They exhibit high diversity across various eukaryotic clades and even among closely related species. While traditionally overlooked, advancements in long-read sequencing technologies have revitalized TE studies. TEs serve are significant sources of genetic variation in natural populations, potentially leading to species differentiation and local adaptations. However, their distribution, evolutionary trajectories, and biological consequences remain poorly understood in non-model species. Bivalves (Class: Bivalvia) represent one of such overlooked taxonomic group. These ancient and diverse filter-feeding aquatic molluscs diversified around 500 million years ago during the early Cambrian. Bivalves are considered as emerging as model systems for studying general biology and human health. Yet, our understanding of their biology will potentially remain limited until a comprehensive understanding of their mobilome is achieved. This thesis makes a first attempt to addresses this gap by exploring the distribution, evolution, and genomic impacts of TEs in bivalves. In Chapter I, I generated the first long reads-based genome assembly for the Manila clam *Ruditapes philippinarum*, focusing on removing potential false duplications due to high heterozygosity. In Chapters II and III I utilized a wide sampling of bivalve genomes to study the distribution and evolution of LINEs, SINEs, and DDE/D DNA transposons. Bivalves were found to host a high diversity of different transposons, with multiple bivalve-specific amplifications likely associated with their diversification. Additionally, I compared TE evolutionary trajectories with other eukaryotic clades, highlighting common and different evolutionary trends. In Chapter IV, I characterized the genomic impacts of TEs and related Structural Variants (SVs) among the economically important oysters. Here I found that up to 14% of the oyster genome exhibits structural differences between haplotypes in terms of insertions and deletions, with TE insertions outnumbering genomic deletions. TEs and SVs were also found to be significant contributors to population differentiation in the Estuarine oyster *C. ariakensis*, potentially providing substrates for local adaptations to varying ocean salinity and temperatures. As a secondary outcome of this thesis, I significantly increased the availability of high-quality TE resources for bivalves by depositing hundreds of novel sequences in the curated database of DFAM. I hope this thesis will help inspiring further research into characterizing transposons and their effects in non-model species. The post-genomics era presents an unparalleled opportunity for scientists to understand genome composition and evolution, and we do not have to miss it.

Content

1. Introduction

Introduction

One of the most prominent features in many eukaryotic genomes is the presence of repetitive DNA (Charlesworth et al., 1994; Wessler, 2006; Wells and Feschotte, 2020), which occupies more than 50% of the human genome (Lander et al., 2001) and reach the 92.45% in the Antarctic krill (Shao et al., 2023). These repetitive genomic regions are commonly categorized into two main types: tandem repeats and interspersed repeats, depending on their nature and organization (Charlesworth et al., 1994). Tandem repeats consist of highly homogeneous DNA stretches where two or more copies of a monomer are repeated consecutively, forming a tandem array. In contrast, interspersed repeats are dispersed throughout the genome and are nonadjacent. Most interspersed repeats correspond to transposable elements (TEs). TEs are selfish genetic elements widespread across almost all eukaryotes and able of actively moving throughout the genome replicating themself independently from the host cell (Wells and Feschotte, 2020). Barbara McClintock firstly discovered TEs in the middle of the last century while studying the variability in the colour patterns of maize kernels. She found that the phenotype is dependent on the interplay between a TE and a pigment gene in what is called *Ac/Ds* system (McClintock, 1951). This discovery suggests for the first time that organism genomes are not static entities but have instead a fluid organization subject to rearrangements both between and within individuals (*e.g* between different cell types).

With the advent of the genomic era and the rapid advances in sequencing technologies over the last two decades, this concept has gained increasing importance (Wellenreuther et al., 2019). Transposable elements, once greatly overlooked, are now becoming more and more subject of study in the context of evolutionary processes, also thanks to our increased ability to identify them thanks to long read technologies and highly contiguous genome assemblies (Shahid and Slotkin, 2020, Peona et al., 2021). Indeed, because of their repetitive nature, TEs can impose challenges during the assembly process, and high-copy number recently mobilized elements can be impossible to represent correctly using short-reads technologies alone, resulting in their underrepresentation in the genome assembly.

Due to their selfish nature and ability to move across the genome, transposons are significant mutagenic agents in natural populations (McDonald, 1993). Typically, most transposable element insertions are deleterious or neutral for the host organism (McDonald, 1993; Bennetzen and Wang, 2014), and their accumulation can decrease its fitness, as recently observed in maize (Stitzer et al., 2023). Despite this, over both short and long evolutionary timescales, transposable elements can contribute to the emergence of evolutionary innovations through processes such as exaptation, domestication, and changing in host gene regulation (Schrader and Schmitz, 2019). Some notable examples include the evolution of introns (Rogers and Bendich, 2023), the evolution of adaptive immune systems in jawed vertebrates (Kapitonov and Koonin, 2015), the contribution of TE-derived exons in the formation of alternative splicing variants promoting proteome diversity (Schmitz and Brosius, 2011), and the role of TE insertions in providing pesticide resistances in *Drosophila* (Salces-Ortiz et al., 2020) as well as in other insect species (Gilbert et al., 2021). Moreover, transposons contribute to genome evolution in multiple other ways beyond *de novo* insertions. Indeed, both recently accumulated TEs and old, immobilized copies are actively removed from genomes or act as substrates for other types of structural variations through homologous recombination events, such as non-allelic homologous recombination (NAHR) (Startek et al., 2015), and non-homologous repair mechanisms (Balachandran et al., 2022). These processes can induce changes in genome structure, affecting its 3D conformation and shaping the chromatin landscape (Lawson et al., 2023). Moreover, they are also significant contributors to genome size evolutionary dynamics across metazoans, in accordance with an “accordion model” of genome size evolution where increasing in genome size mainly caused by transposition activity is counteracted by genomic deletions (Kapusta et al., 2017).

Transposons are not only widespread across eukaryotes but also highly diverse (Wells and Feschotte, 2020). They are commonly subdivided into two main classes: Class I and Class II, depending on their replication mechanism (Finnegan, 1989). Class I elements, also known as retrotransposons, replicate via an RNA intermediate, while Class II elements use a single or double-stranded DNA intermediate. For this reason, they are also called DNA transposons. For most TEs of both classes, the transposition event is detectable in the genome by the identification of two small direct repeats called target site duplications (TSD). For some TE groups, the length of the TSDs can be used as a diagnostic feature to classify the transposon into different groups (Wicker et al., 2007; Feschotte and Pritham, 2007). Almost all type of transposons can exist as autonomous elements or as non-autonomous counterparts. Autonomous transposons are those carrying all features necessary for their own transposition whereases non-autonomous lack coding capacity. A particular case is that of the Class I short interspersed nuclear elements (SINEs) that will be introduced later.

Within Class II, there are at least three heterogeneous modes of replication (Wells and Feschotte, 2020). The most common one is the 'cut-and-paste' mechanism, involving the complete excision of the transposon from its original location and insertion into a novel genomic region. These elements can be further subdivided into DDE/D elements, whose transposition is catalyzed by a DDE/D transposase, and Tyrosine recombinase (YR)-mediated transposons. DDE/D-derived insertions are always characterized by target site duplications (TSD), whereas YR is variable in this regard. Both DDE/D and YR transposons usually, but not always, feature a single open reading frame (ORF) flanked by terminal inverted repeats (TIRs). The TIRs contain binding sites recognized by the transposase enzyme and initiate the transposition of the element. Importantly, during the transposition process, 'cut-and-paste' transposons can undergo internal deletions, giving rise to shorter, non-autonomous counterparts called miniature inverted repeats (MITEs) (Hsia and Schnable, 1996). Despite the absence of coding capacity, these elements can still preserve the transposase binding sites along the TIRs, allowing them to be mobilized *in trans* until an autonomous counterpart survives across the genome

The second group of DNA transposons are Mavericks, also called Polintons, which replicate through a "self-synthesizing" process (Kapitonov and Jurka, 2006). As the name suggests, these elements are able to directly synthesize their DNA copy thanks to the presence of a protein-primed family-B DNA polymerase (pPolB). The close relationship of this protein to those of adenovirus and the ability to encode for double and single jelly-roll capsid-like proteins suggest that these elements represent endogenous viruses or virophages. Similarly to DDE/D Class II transposons, Mavericks are also characterized by TIRs and TSDs.

Finally, the third group of DNA transposons are Rolling circle elements (RC), also called Helitrons. These enigmatic TEs use a particular replication mechanism called rolling-circle-like replication (Kapitonov and Jurka, 2001). Unlike all other class II elements, RC lacks terminal inverted repeats and encodes for a DNA helicase similar to those encoded by known rolling-circle replicons. Moreover, they never generate TSDs upon duplication. Due to the lack of any commonality in both the replication mechanism and structure, RC elements are sometimes separated from other DNA transposons when reporting statistics about genome-wide TE coverage. I also adopted the same approach in all chapters of this thesis.

The replication mechanism of Class I retrotransposons involves the retrotranscription of their RNA followed by the reintegration of the resulting DNA sequence. The main elements included in this class are LTR and non-LTR retrotransposons with both of them leaving TSDs upon transposition.

LTR retrotransposons possess long terminal repeats and are closely related to retroviruses, sharing similar structural features and replication mechanisms (Eickbush and Malik, 2007). These elements encode two or three open reading frames (ORFs). The *gag* and *pol* genes are always present and required for the replication cycle and transposition (Eickbush and Jamburuthugoda, 2008). The *pol* gene includes the reverse transcriptase (RT), integrase (IN), and ribonuclease H enzyme, which catalyze the reverse transcription and integration of the cDNA. LTRs can commonly be found in their non-autonomous version (solo LTRs), where, due to ectopic recombination between the LTR portions, the internal coding region is removed from the genome, leaving only one of the two LTRs.

Non-LTR retrotransposons include Long Interspersed Nuclear Elements (LINEs) and Short Interspersed Nuclear Elements (SINEs). The formers are characterized by one or two open reading frames (ORFs), with one always encoding reverse transcriptase (RT) and integrase (IN) enzymes that catalyze target-primed reverse transcription (TPRT) (Luan et al., 1993). The 5' terminal region of LINEs includes the Pol II promoter necessary for their expression. After translation, the LINE RNA moves back from the cytoplasm into the nucleus together with its protein products. Here, the IN enzyme catalyzes a single strand break in the host DNA, and the LINE RNA hybridizes at its 3' end, followed by retrotranscription and integration of the cDNA. TPRT is a highly imprecise process that often leads to the premature termination of reverse transcription (Kazazian and Goodier, 2002). These 5'-truncated copies usually lose the Pol III promoter, preventing them from further propagation. SINEs are non-autonomous transposons that rely on the protein machinery of their LINE counterparts. These elements constitute a heterogeneous group of short transposons organized in a modular manner, typically with a length < 700 bps (Vassetzky and Kramerov, 2013). Their structure includes a head, a body, and a tail region, with the body possibly absent in some elements. The head is invariably associated with one of the three RNA types synthesized by RNA pol III (tRNAs, 5S rRNA, or 7SL RNA), which is hijacked for their transcription. In contrast, the tail is homologous to the 3' ends of LINEs and serves as recognition for the LINE-derived reverse transcriptase (RT). Unlike MITEs, their origin is only partially dependent on their autonomous counterparts. A detailed description of the structure of SINEs is presented in the introduction of chapter three.

Beyond these primary classifications based on the replication mechanism, it is widely accepted that lower-level classifications should reflect the phylogenetic relationships of the elements (Wicker et al., 2007). In this context, phylogenies based on RT and DDE/D protein segments have been the most commonly utilized due to the relatively high conservation of the protein sequences (Arkhipova, 2017). However, the deep-divergence time of most transposons and the short length of the sequences pose challenges in establishing robust phylogenies. This is particularly true for DDE/D transposases, as there are relatively short and the divergence of most groups predates the diversification of eukaryotes (Arkhipova, 2017). Currently, the most widely used classification for both retrotransposons and Class II DDE/D elements is based on superfamilies, for which sufficient evidence of monophyly has been found (e.g., Yuan and Wessler, 2011; Kapitonov et al., 2009). Ideally, clustering patterns within superfamilies should further subdivide them into clades. However, in practice, this is not always the case. For instance, within LINEs, the two primary TE databases, RepBase (Jurka et al., 2005) and DFAM (Storer et al., 2021), employ different classification schemes. These differences may pose challenges when comparing TE annotations produced with different libraries. Finally, SINE elements can be subdivided into superfamilies based on the presence of highly conserved domains (HCD) in their body regions (Gilbert and Labuda, 1999; Ogiwara et al., 2002; Nishihara et al., 2006; Nishihara et al., 2016).

Due to their deleterious effects, eukaryotes have evolved a wide range of repression mechanisms to control transposable element (TE) activity. Host TE repression is mainly achieved through microRNAs (Rozhkov et al., 2013), DNA methylation (Molaro and Malik, 2016), Krab-Zinc finger proteins (Yang et al., 2017), and repeat-induced mutations in fungi (Gladyshev, 2017). However, TEs have also evolved mechanisms to control their own proliferation, such as the overproduction inhibition process observed in *Drosophila* Mariner DNA transposons (Lohe and Hartl, 1996). Self-regulating repression mechanisms are believed to increase with an elevated copy number of TEs, suggesting that the more active a TE is, the more it is expected to be self-repressed (Rouzic and Deceliere, 2005). Additionally, many transposons exhibit suboptimal transposition efficiency in natural settings (Lampe et al., 1999). Complex relationships involving genetic drift, TE deletion and replication rates, host suppression mechanisms, TE self-regulation, horizontal transfer of transposons (i.e., movement of transposons between individuals and even different species), and competition between different TEs have been extensively incorporated into mathematical models to study the short and long-term evolution of TEs (e.g., Rouzic and Capy, 2006; Abrusán and Krambeck, 2006; Le Rouzic et al., 2007; Szitenberg et al., 2016). The concept of transposons as individuals within a population, coupled with the observation of varying TE richness and diversity across different host species and a non-random distribution of different TEs across genomes, leads to the intriguing concept of the genome as an ecosystem (Venner et al., 2009). In this context, TE evolution is not only dependent on the properties of transposons themselves (the individuals) but also on the properties of the environment (the host), such as effective population size and recombination rate.

The distribution, evolutionary trajectories, and biological consequences of TE activity have been extensively explored across vertebrates, particularly among mammals (e.g., Brandt et al., 2005; Hellen and Brookfield, 2013; Ricci et al., 2018; Senft and Macfarlan, 2021; Osmanski et al., 2023), fishes (e.g., Gao et al., 2016; Shao et al., 2019; Symonová and Suh, 2019; Carleton et al., 2020; Chang et al., 2022; Mallik et al., 2023), and birds (e.g., Suh, 2015; Suh et al., 2016; Kapusta and Suh, 2017; Kapusta et al., 2017; Manthey et al., 2018; Galbraith et al., 2021). In invertebrates, aside from some model species like Dros*ophila melanogaster, Heliconius spp.*, and *Caenorhabditis elegans*, very little is known about their TE diversity (Sproul et al., 2023). Understanding the distribution, diversity, and richness of TEs across the tree of life is not only interesting *per se* but could also help generate evolutionary hypotheses that can be tested regarding a particular taxonomic group. Moreover, a possible indirect outcome of TE-centered studies is the potential to produce novel genomic resources freely available for the scientific community. These data can then be reused for more organismal-centered research, speeding up science (e.g., Osmanski et al., 2023).

Bivalves (Class: Bivalvia) is one of these greatly understudied group. They are an ancient and diversified clade of filter-feeding aquatic mollusks whose origin and diversification can be date back to the early Cambrian, around ~500 Mya (Kocot et al., 2020). They comprise ~20,000 recognized species around the world (Coen and Grizzle, 2016) and they can be subdivided into the five clades Protobranchia, Pterimorpha, Paleoheterodonta, Heterodonta, and Anomalosdesmata with Protobranchia being the first diverging clade (González et al., 2015). Most of the species live in oceans but multiple lineages have independently colonized freshwater environments as well as deep sea hydrothermal vents during their evolutionary history. Clams (Order Venerida), mussels (Order Mytilida), scallops (Order Pectinida), and oysters (Order Ostreida) represent crucial resources in aquaculture, with an estimated market value of $17.1 billion in 2015 (van der Schatte Olivier et al., 2020). However, the production of these species can be adversely affected by pollution and climate change, leading to increased ocean temperatures, reduced salinity and pH with negative repercussions for the economic sector (Rato et al., 2022). Some bivalve species exhibit unique genomic and biological features, including mitochondrial Doubly Uniparental Inheritance (Zouros et al., 1994), ancient homomorphic sex chromosomes (Han et al., 2022), and extremely variable lifespans, with the longest-lived noncolonial metazoan known so far, *Arctica islandica* (Iannello et al., 2023). They possess a highly diversified innate immune repertoire (Regan et al., 2021), their genomes have undergone multiple contractions during evolutionary history via chromosome loss (Adachi et al., 2021). Additionally, some species are affected by horizontally transmittable neoplasia (Dujon et al., 2021). As a result, their genomic resources, especially genome assemblies, have rapidly expanded in recent years. Genome sequencing projects have revealed bivalves as highly heterozygous species, generally characterized by large genomes and high repetitive content (Gomes-dos-Santos et al., 2020). These factors impose challenges in genome assembly projects leading to high assembly fragmentation and inclusion of haplotypic variants in non-phased genomes which results in false duplicated genomic regions.

The goals of my PhD were set in the context. In **Chapter I,** I contributed to the release of a novel genome assembly for the Manila clam *Ruditapes philippinarum*, utilizing long-read sequencing technology. In **Chapter II** and **Chapter III,** I significantly expanded the collection of freely available high-quality transposable element (TE) consensus sequences for bivalves, focusing on LINE, SINEs and Class II DDE/D related-transposons. These newly generated data were utilized to investigate TE diversity and evolution across in a wide range of bivalve genomes. In **Chapiter IV** I investigated within and between individual structural variants in oysters and their relationships to transposons.

References

- Abrusán, G., Krambeck, H.-J., 2006. Competition may determine the diversity of transposable elements. Theoretical Population Biology 70, 364–375. https://doi.org/10.1016/j.tpb.2006.05.001
- Adachi, K., Yoshizumi, A., Kuramochi, T., Kado, R., Okumura, S.-I., 2021. Novel insights into the evolution of genome size and AT content in mollusks. Mar Biol 168, 25. https://doi.org/10.1007/s00227-021-03826-x
- Arkhipova, I.R., 2017. Using bioinformatic and phylogenetic approaches to classify transposable elements and understand their complex evolutionary histories. Mobile DNA 8, 19. https://doi.org/10.1186/s13100-017-0103-2
- Balachandran, P., Walawalkar, I.A., Flores, J.I., Dayton, J.N., Audano, P.A., Beck, C.R., 2022. Transposable element-mediated rearrangements are prevalent in human genomes. Nat Commun 13, 7115. https://doi.org/10.1038/s41467-022-34810-8.
- Bennetzen, J.L., Wang, H., 2014. The Contributions of Transposable Elements to the Structure, Function, and Evolution of Plant Genomes. Annual Review of Plant Biology 65, 505–530. https://doi.org/10.1146/annurev-arplant-050213-035811.
- Böhne, A., Brunet, F., Galiana-Arnoux, D., Schultheis, C., Volff, J.-N., 2008. Transposable elements as drivers of genomic and biological diversity in vertebrates. Chromosome Res 16, 203–215. https://doi.org/10.1007/s10577-007-1202-6.
- Brandt, J., Schrauth, S., Veith, A.-M., Froschauer, A., Haneke, T., Schultheis, C., Gessler, M., Leimeister, C., Volff, J.-N., 2005. Transposable elements as a source of genetic innovation: expression and evolution of a family of retrotransposon-derived neogenes in mammals. Gene, Structural Approaches to Sequence Evolution: Molecules, Networks, Populations – Part 1 345, 101–111. https://doi.org/10.1016/j.gene.2004.11.022.
- Carleton, K.L., Conte, M.A., Malinsky, M., Nandamuri, S.P., Sandkam, B.A., Meier, J.I., Mwaiko, S., Seehausen, O., Kocher, T.D., 2020. Movement of transposable elements contributes to cichlid diversity. Molecular Ecology 29, 4956–4969. https://doi.org/10.1111/mec.15685
- Chalopin, D., Naville, M., Plard, F., Galiana, D., Volff, J.-N., 2015. Comparative Analysis of Transposable Elements Highlights Mobilome Diversity and Evolution in Vertebrates. Genome Biology and Evolution 7, 567–580. https://doi.org/10.1093/gbe/evv005.
- Charlesworth, B., Sniegowski, P., Stephan, W., 1994. The evolutionary dynamics of repetitive DNA in eukaryotes. Nature 371, 215–220. https://doi.org/10.1038/371215a0.

- Coen, L.D., Grizzle, R.E., 2016. Bivalve Molluscs, in: Kennish, M.J. (Ed.), Encyclopedia of Estuaries, Encyclopedia of Earth Sciences Series. Springer Netherlands, Dordrecht, pp. 89–109. https://doi.org/10.1007/978-94-017-8801-4_88.
- Dujon, A.M., Bramwell, G., Roche, B., Thomas, F., Ujvari, B., 2021. Transmissible cancers in mammals and bivalves: How many examples are there? BioEssays 43, 2000222. https://doi.org/10.1002/bies.202000222
- Eickbush, T.H., Jamburuthugoda, V.K., 2008. The diversity of retrotransposons and the properties of their reverse transcriptases. Virus Res 134, 221–234. https://doi.org/10.1016/j.virusres.2007.12.010
- Eickbush, T.H., Malik, H.S., 2007. Origins and Evolution of Retrotransposons, in: Mobile DNA II. John Wiley & Sons, Ltd, pp. 1111–1144. https://doi.org/10.1128/9781555817954.ch49
- Feschotte, C., Pritham, E.J., 2007. DNA Transposons and the Evolution of Eukaryotic Genomes. Annual Review of Genetics 41, 331–368. https://doi.org/10.1146/annurev.genet.40.110405.090448.
- Finnegan, D.J., 1989. Eukaryotic transposable elements and genome evolution. Trends Genet 5, 103–107. https://doi.org/10.1016/0168-9525(89)90039-5.
- Galbraith, J.D., Kortschak, R.D., Suh, A., Adelson, D.L., 2021. Genome Stability Is in the Eye of the Beholder: CR1 Retrotransposon Activity Varies Significantly across Avian Diversity. Genome Biology and Evolution 13, evab259. https://doi.org/10.1093/gbe/evab259.
- Gilbert, C., Peccoud, J., Cordaux, R., 2021. Transposable Elements and the Evolution of Insects. Annual Review of Entomology 66, 355–372. https://doi.org/10.1146/annurev-ento-070720-074650.
- Gilbert, N., Labuda, D., 1999. CORE-SINEs: Eukaryotic short interspersed retroposing elements with common sequence motifs. Proceedings of the National Academy of Sciences 96, 2869–2874. https://doi.org/10.1073/pnas.96.6.2869
- GLADYSHEV, E., 2017. Repeat-Induced Point Mutation (RIP) and Other Genome Defense Mechanisms in Fungi. Microbiol Spectr 5, 10.1128/microbiolspec.FUNK-0042–2017. https://doi.org/10.1128/microbiolspec.FUNK-0042-2017
- Gomes-dos-Santos, A., Lopes-Lima, M., Castro, L.F.C., Froufe, E., 2020. Molluscan genomics: the road so far and the way forward. Hydrobiologia 847, 1705–1726. https://doi.org/10.1007/s10750-019-04111-1

- González, V.L., Andrade, S.C.S., Bieler, R., Collins, T.M., Dunn, C.W., Mikkelsen, P.M., Taylor, J.D., Giribet, G., 2015. A phylogenetic backbone for Bivalvia: an RNA-seq approach. Proceedings of the Royal Society B: Biological Sciences 282, 20142332. https://doi.org/10.1098/rspb.2014.2332.
- Han, W., Liu, L., Wang, J., Wei, H., Li, Yuli, Zhang, Lijing, Guo, Z., Li, Yajuan, Liu, T., Zeng, Q., Xing, Q., Shu, Y., Wang, T., Yang, Y., Zhang, M., Li, R., Yu, J., Pu, Z., Lv, J., Lian, S., Hu, J., Hu, X., Bao, Z., Bao, L., Zhang, Lingling, Wang, S., 2022. Ancient homomorphy of molluscan sex chromosomes sustained by reversible sex-biased genes and sex determiner translocation. Nat Ecol Evol 6, 1891–1906. https://doi.org/10.1038/s41559-022-01898-6
- Hellen, E.H.B., Brookfield, J.F.Y., 2013. The Diversity of Class II Transposable Elements in Mammalian Genomes Has Arisen from Ancestral Phylogenetic Splits during Ancient Waves of Proliferation through the Genome. Molecular Biology and Evolution 30, 100–108. https://doi.org/10.1093/molbev/mss206
- Hsia, A.-P., Schnable, P.S., 1996. DNA Sequence Analyses Support the Role of Interrupted Gap Repair in the Origin of Internal Deletions of the Maize Transposon, MuDR. Genetics 142, 603–618. https://doi.org/10.1093/genetics/142.2.603.
- Iannello, M., Forni, G., Piccinini, G., Xu, R., Martelossi, J., Ghiselli, F., Milani, L., 2023. Signatures of Extreme Longevity: A Perspective from Bivalve Molecular Evolution. Genome Biology and Evolution 15, evad159. https://doi.org/10.1093/gbe/evad159
- Jurka, J., Kapitonov, V.V., Pavlicek, A., Klonowski, P., Kohany, O., Walichiewicz, J., 2005. Repbase Update, a database of eukaryotic repetitive elements. Cytogenetic and Genome Research 110, 462–467. https://doi.org/10.1159/000084979.
- Kapitonov, V.V., Jurka, J., 2001. Rolling-circle transposons in eukaryotes. Proceedings of the National Academy of Sciences 98, 8714–8719. https://doi.org/10.1073/pnas.151269298.
- Kapitonov, V.V., Jurka, J., 2006. Self-synthesizing DNA transposons in eukaryotes. Proceedings of the National Academy of Sciences 103, 4540–4545. https://doi.org/10.1073/pnas.0600833103.
- Kapitonov, V.V., Koonin, E.V., 2015. Evolution of the RAG1-RAG2 locus: both proteins came from the same transposon. Biol Direct 10, 20. https://doi.org/10.1186/s13062-015-0055-8.
- Kapitonov, V.V., Tempel, S., Jurka, J., 2009. Simple and fast classification of non-LTR retrotransposons based on phylogeny of their RT domain protein sequences. Gene, Genomic

Impact of Eukaryotic Transposable Elements 448, 207–213. https://doi.org/10.1016/j.gene.2009.07.019

- Kapusta, A., Suh, A., 2017. Evolution of bird genomes—a transposon's-eye view. Annals of the New York Academy of Sciences 1389, 164–185. https://doi.org/10.1111/nyas.13295.
- Kapusta, A., Suh, A., Feschotte, C., 2017. Dynamics of genome size evolution in birds and mammals. Proceedings of the National Academy of Sciences 114, E1460–E1469. https://doi.org/10.1073/pnas.1616702114.
- Kazazian, H.H., Goodier, J.L., 2002. LINE Drive: Retrotransposition and Genome Instability. Cell 110, 277–280. https://doi.org/10.1016/S0092-8674(02)00868-1
- Kocot, K.M., Poustka, A.J., Stöger, I., Halanych, K.M., Schrödl, M., 2020. New data from Monoplacophora and a carefully-curated dataset resolve molluscan relationships. Sci Rep 10, 101. https://doi.org/10.1038/s41598-019-56728-w
- Lampe, D.J., Akerley, B.J., Rubin, E.J., Mekalanos, J.J., Robertson, H.M., 1999. Hyperactive transposase mutants of the Himar1 mariner transposon. Proceedings of the National Academy of Sciences 96, 11428–11433. https://doi.org/10.1073/pnas.96.20.11428
- Lander, E.S., et al. Initial sequencing and analysis of the human genome. Nature 409, 860–921. https://doi.org/10.1038/35057062.
- Lawson, H.A., Liang, Y., Wang, T., 2023. Transposable elements in mammalian chromatin organization. Nat Rev Genet 24, 712–723. https://doi.org/10.1038/s41576-023-00609-6
- Le Rouzic, A., Boutin, T.S., Capy, P., 2007. Long-term evolution of transposable elements. Proceedings of the National Academy of Sciences 104, 19375–19380. https://doi.org/10.1073/pnas.0705238104
- Lohe, A.R., Hartl, D.L., 1996. Autoregulation of mariner transposase activity by overproduction and dominant-negative complementation. Molecular Biology and Evolution 13, 549–555. https://doi.org/10.1093/oxfordjournals.molbev.a025615
- Luan, D.D., Korman, M.H., Jakubczak, J.L., Eickbush, T.H., 1993. Reverse transcription of R2Bm RNA is primed by a nick at the chromosomal target site: a mechanism for non-LTR retrotransposition. Cell 72, 595–605. https://doi.org/10.1016/0092-8674(93)90078-5
- Mallik, R., Wcisel, D.J., Near, T.J., Yoder, J.A., Dornburg, A., 2023. Investigating the impact of whole genome duplication on transposable element evolution in ray-finned fishes. https://doi.org/10.1101/2023.12.22.572151

- Manthey, J.D., Moyle, R.G., Boissinot, S., 2018. Multiple and Independent Phases of Transposable Element Amplification in the Genomes of Piciformes (Woodpeckers and Allies). Genome Biology and Evolution 10, 1445–1456. https://doi.org/10.1093/gbe/evy105
- McClintock, B., 1951. Chromosome Organization and Genic Expression. Cold Spring Harb Symp Quant Biol 16, 13–47. https://doi.org/10.1101/SQB.1951.016.01.004.
- McDonald, J.F., 1993. Evolution and consequences of transposable elements. Current Opinion in Genetics & Development 3, 855–864. https://doi.org/10.1016/0959-437X(93)90005-A.
- Molaro, A., Malik, H.S., 2016. Hide and seek: how chromatin-based pathways silence retroelements in the mammalian germline. Current Opinion in Genetics & Development, Genome architecture and expression 37, 51–58. https://doi.org/10.1016/j.gde.2015.12.001
- Nishihara, H., Plazzi, F., Passamonti, M., Okada, N., 2016. MetaSINEs: Broad Distribution of a Novel SINE Superfamily in Animals. Genome Biology and Evolution 8, 528–539. https://doi.org/10.1093/gbe/evw029
- Nishihara, H., Smit, A.F.A., Okada, N., 2006. Functional noncoding sequences derived from SINEs in the mammalian genome. Genome Res. 16, 864–874. https://doi.org/10.1101/gr.5255506.
- Ogiwara, I., Miya, M., Ohshima, K., Okada, N., 2002. V-SINEs: A New Superfamily of Vertebrate SINEs That Are Widespread in Vertebrate Genomes and Retain a Strongly Conserved Segment within Each Repetitive Unit. Genome Res. 12, 316–324. https://doi.org/10.1101/gr.212302
- Osmanski, A.B., Paulat, N.S., Korstian, J., Grimshaw, J.R., Halsey, M., Sullivan, K.A.M., Moreno-Santillán, D.D., Crookshanks, C., Roberts, J., Garcia, C., Johnson, M.G., Densmore, L.D., Stevens, R.D., Zoonomia Consortium, Rosen, J., Storer, J.M., Hubley, R., Smit, A.F.A., Dávalos, L.M., Karlsson, E.K., Lindblad-Toh, K., Ray, D.A., 2023. Insights into mammalian TE diversity through the curation of 248 genome assemblies. Science 380, eabn1430. https://doi.org/10.1126/science.abn1430.
- Pasquesi, G.I.M., Perry, B.W., Vandewege, M.W., Ruggiero, R.P., Schield, D.R., Castoe, T.A., 2020. Vertebrate Lineages Exhibit Diverse Patterns of Transposable Element Regulation and Expression across Tissues. Genome Biology and Evolution 12, 506–521. https://doi.org/10.1093/gbe/evaa068.
- Peona, V., Blom, M.P.K., Xu, L., Burri, R., Sullivan, S., Bunikis, I., Liachko, I., Haryoko, T., Jønsson, K.A., Zhou, Q., Irestedt, M., Suh, A., 2021. Identifying the causes and

consequences of assembly gaps using a multiplatform genome assembly of a bird-of-paradise. Molecular Ecology Resources 21, 263–286. https://doi.org/10.1111/1755-0998.13252.

- Rato, A., Joaquim, S., Matias, A.M., Roque, C., Marques, A., Matias, D., 2022. The Impact of Climate Change on Bivalve Farming: Combined Effect of Temperature and Salinity on Survival and Feeding Behavior of Clams Ruditapes decussatus. Frontiers in Marine Science 9.
- Regan, T., Stevens, L., Peñaloza, C., Houston, R.D., Robledo, D., Bean, T.P., 2021. Ancestral Physical Stress and Later Immune Gene Family Expansions Shaped Bivalve Mollusc Evolution. Genome Biol Evol 13, evab177. https://doi.org/10.1093/gbe/evab177
- Ricci, M., Peona, V., Guichard, E., Taccioli, C., Boattini, A., 2018. Transposable Elements Activity is Positively Related to Rate of Speciation in Mammals. J Mol Evol 86, 303–310. https://doi.org/10.1007/s00239-018-9847-7.
- Rogers, S.O., Bendich, A.J., 2023. Introns are derived from transposons. https://doi.org/10.1101/2023.02.21.529479.
- Rouzic, A.L., Capy, P., 2006. Population Genetics Models of Competition Between Transposable Element Subfamilies. Genetics 174, 785–793. https://doi.org/10.1534/genetics.105.052241
- Rouzic, A.L., Deceliere, G., 2005. Models of the population genetics of transposable elements. Genet. Res. 85, 171–181. https://doi.org/10.1017/S0016672305007585
- Rozhkov, N.V., Hammell, M., Hannon, G.J., 2013. Multiple roles for Piwi in silencing Drosophila transposons. Genes Dev 27, 400–412. https://doi.org/10.1101/gad.209767.112
- Salces-Ortiz, J., Vargas-Chavez, C., Guio, L., Rech, G.E., González, J., 2020. Transposable elements contribute to the genomic response to insecticides in Drosophila melanogaster. Philosophical Transactions of the Royal Society B: Biological Sciences 375, 20190341. https://doi.org/10.1098/rstb.2019.0341.
- Schmitz, J., Brosius, J. ,2011. Exonization of transposed elements: A challenge and opportunity for evolution. Biochimie 93. https://doi.org/10.1016/j.biochi.2011.07.014.
- Schrader, L., Schmitz, J., 2019. The impact of transposable elements in adaptive evolution. Molecular Ecology 28, 1537–1549. https://doi.org/10.1111/mec.14794.
- Senft, A.D., Macfarlan, T.S., 2021. Transposable elements shape the evolution of mammalian development. Nat Rev Genet 22, 691–711. https://doi.org/10.1038/s41576-021-00385-1.

- Shahid, S., Slotkin, R.K., 2020. The current revolution in transposable element biology enabled by long reads. Current Opinion in Plant Biology, Genome studies and molecular genetics 54, 49–56. https://doi.org/10.1016/j.pbi.2019.12.012.
- Shao, C., Sun, S., Liu, K., Wang, Jiahao, Li, S., Liu, Q., Deagle, B.E., Seim, I., Biscontin, A., Wang, Q., Liu, X., Kawaguchi, S., Liu, Yalin, Jarman, S., Wang, Yue, Wang, H.-Y., Huang, G., Hu, J., Feng, B., De Pittà, C., Liu, Shanshan, Wang, R., Ma, K., Ying, Y., Sales, G., Sun, T., Wang, X., Zhang, Y., Zhao, Y., Pan, S., Hao, X., Wang, Yang, Xu, J., Yue, B., Sun, Y., Zhang, H., Xu, M., Liu, Yuyan, Jia, X., Zhu, J., Liu, Shufang, Ruan, J., Zhang, G., Yang, H., Xu, X., Wang, Jun, Zhao, X., Meyer, B., Fan, G., 2023. The enormous repetitive Antarctic krill genome reveals environmental adaptations and population insights. Cell 186, 1279-1294.e19. https://doi.org/10.1016/j.cell.2023.02.005
- Shao, F., Han, M., Peng, Z., 2019. Evolution and diversity of transposable elements in fish genomes. Sci Rep 9, 15399. https://doi.org/10.1038/s41598-019-51888-1.
- Sotero-Caio, C.G., Platt, R.N., II, Suh, A., Ray, D.A., 2017. Evolution and Diversity of Transposable Elements in Vertebrate Genomes. Genome Biology and Evolution 9, 161–177. https://doi.org/10.1093/gbe/evw264.
- Sproul, J.S., Hotaling, S., Heckenhauer, J., Powell, A., Marshall, D., Larracuente, A.M., Kelley, J.L., Pauls, S.U., Frandsen, P.B., 2023. Analyses of 600+ insect genomes reveal repetitive element dynamics and highlight biodiversity-scale repeat annotation challenges. Genome Res. 33, 1708–1717. https://doi.org/10.1101/gr.277387.122.
- Startek, M., Szafranski, P., Gambin, T., Campbell, I.M., Hixson, P., Shaw, C.A., Stankiewicz, P., Gambin, A., 2015. Genome-wide analyses of LINE–LINE-mediated nonallelic homologous recombination. Nucleic Acids Research 43, 2188–2198. https://doi.org/10.1093/nar/gku1394
- Stitzer, M.C., Khaipho-Burch, M.B., Hudson, A.I., Song, B., Valdez-Franco, J.A., Ramstein, G., Feschotte, C., Buckler, E.S., 2023. Transposable element abundance subtly contributes to lower fitness in maize. https://doi.org/10.1101/2023.09.18.557618.
- Storer, J., Hubley, R., Rosen, J., Wheeler, T.J., Smit, A.F., 2021. The Dfam community resource of transposable element families, sequence models, and genome annotations. Mobile DNA 12, 2. https://doi.org/10.1186/s13100-020-00230-y.
- Symonová, R., Suh, A., 2019. Nucleotide composition of transposable elements likely contributes to AT/GC compositional homogeneity of teleost fish genomes. Mobile DNA 10, 49. https://doi.org/10.1186/s13100-019-0195-y

- Szitenberg, A., Cha, S., Opperman, C.H., Bird, D.M., Blaxter, M.L., Lunt, D.H., 2016. Genetic Drift, Not Life History or RNAi, Determine Long-Term Evolution of Transposable Elements. Genome Biol Evol 8, 2964–2978. https://doi.org/10.1093/gbe/evw208
- van der Schatte Olivier, A., Jones, L., Vay, L.L., Christie, M., Wilson, J., Malham, S.K., 2020. A global review of the ecosystem services provided by bivalve aquaculture. Reviews in Aquaculture 12, 3–25. https://doi.org/10.1111/raq.12301.
- Vassetzky, N.S., Kramerov, D.A., 2013. SINEBase: a database and tool for SINE analysis. Nucleic Acids Res 41, D83–D89. https://doi.org/10.1093/nar/gks1263
- Venner, S., Feschotte, C., Biémont, C., 2009. Dynamics of transposable elements: towards a community ecology of the genome. Trends in Genetics 25, 317–323. https://doi.org/10.1016/j.tig.2009.05.003
- Wellenreuther, M., Mérot, C., Berdan, E., Bernatchez, L., 2019. Going beyond SNPs: The role of structural genomic variants in adaptive evolution and species diversification. Molecular Ecology 28, 1203–1209. https://doi.org/10.1111/mec.15066.
- Wells, J.N., Feschotte, C., 2020. A Field Guide to Eukaryotic Transposable Elements. Annu Rev Genet 54, 539–561. https://doi.org/10.1146/annurev-genet-040620-022145.
- Wessler, S.R., 2006. Transposable elements and the evolution of eukaryotic genomes. Proceedings of the National Academy of Sciences 103, 17600–17601. https://doi.org/10.1073/pnas.0607612103.
- Wicker, T., Sabot, F., Hua-Van, A., Bennetzen, J.L., Capy, P., Chalhoub, B., Flavell, A., Leroy, P., Morgante, M., Panaud, O., Paux, E., SanMiguel, P., Schulman, A.H., 2007. A unified classification system for eukaryotic transposable elements. Nat Rev Genet 8, 973–982. https://doi.org/10.1038/nrg2165.
- Yang, P., Wang, Y., Macfarlan, T.S., 2017. The Role of KRAB-ZFPs in Transposable Element Repression and Mammalian Evolution. Trends in Genetics 33, 871–881. https://doi.org/10.1016/j.tig.2017.08.006
- Yuan, Y.-W., Wessler, S.R., 2011. The catalytic domain of all eukaryotic cut-and-paste transposase superfamilies. Proceedings of the National Academy of Sciences 108, 7884–7889. https://doi.org/10.1073/pnas.1104208108.
- Zhang, H.-H., Peccoud, J., Xu, M.-R.-X., Zhang, X.-G., Gilbert, C., 2020. Horizontal transfer and evolution of transposable elements in vertebrates. Nat Commun 11, 1362. https://doi.org/10.1038/s41467-020-15149-4.

- Zouros, E., Oberhauser Ball, A., Saavedra, C., Freeman, K.R., 1994. An unusual type of mitochondrial DNA inheritance in the blue mussel Mytilus. Proceedings of the National Academy of Sciences 91, 7463–7467. https://doi.org/10.1073/pnas.91.16.7463

2. Summary of the Main Chapters

Chapter I

Multi-tissue RNA-Seq Analysis and Long-read-based Genome Assembly Reveal Complex Sex-specific Gene Regulation and Molecular Evolution in the Manila Clam.

In this chapter, I successfully assembled the first long-read genome of the Manila clam, *Ruditapes philippinarum,* using a combination of PacBio CLR reads and Illumina short reads. While the primary objectives of this project were focused on identifying sex-related differences in terms of SNPs, expression patterns, and splicing variants (analyses carried out by collaborators), I took care of all steps related to kmer-based genome survey, genome assembly, annotation of repeats, and whole-genome alignment with a publicly available short-reads-only *R. philippinarum* genome.

Due to the estimated high heterozygosity, special attention was given to the assembly process, with a meticulous removal of potential false duplications (haplotigs) from the primary assembly. The outcome yielded a more accurate representation of the *R. philippinarum* genome compared to previous versions, particularly in terms of contiguity and assembly size. Result of this chapter have been published in Genome Biology and Evolution under the DOI: https://doi.org/10.1093/gbe/evac171.

Chapter II and Chapter III

Chapter II - *Multiple and Diversified Transposon Lineages Contribute to Early and Recent Bivalve Genome Evolution.*

Chapter III - *Widespread HCD-tRNA derived SINEs in bivalves relies on multiple LINE partners and accumulate in gene-related genomic regions.*

Using the previously assembled *R. philippinarum* genome along with 26 other bivalve assemblies in Chapter II and over 40 in Chapter III, I conducted a comprehensive characterization of DDE/D Class II transposons, LINEs, and SINEs employing a combination of automatic and manual curation processes.

In Chapter II, I further investigated LINE evolutionary and expression patterns in a subset of species identifying multiple potentially active lineages. In Chapter III, my focus was on HCD SINEs, revealing novel SINE-LINE partnerships, as well as commonalities and species-specific differences in their genomic distribution across various genomic compartments.

These efforts mark the first attempts to characterize transposons and their evolutionary dynamics across a diverse range of bivalves and molluscs, resulting in the release of 810 LINE, 762 Class II DDE/D, and 37 SINE (with an additional 60 to be deposited within the next few weeks) manually curated consensus sequences on DFAM. These sequences, all representing full-length or nearly full-length in the case of LINEs, significantly expanded the repertoire compared to the previous availability of only 11 sequences.

The results of the **Chapter II** have been published in BMC Biology under the DOI: https://doi.org/10.1186/s12915-023-01632-z. Results of **Chapter III** are going to be submitted to a peer-review journal within the next few weeks.

Chapter IV

High within- and between- individual structural variability is related to transposable elements insertions and deletions in bivalves

In this final chapter, I delved into the genomic consequences of transposable element (TE) activity at both the individual and population levels in oysters. These economically important species are known for their TE-rich genomes and high heterozygosity, a common feature observed in bivalves. Initially, my focus was on unraveling the relationship between heterozygosity and transposons in four oyster genomes. To achieve this, I established a pipeline to identify within-individual structural variants (SVs) based on a haploid representation of a diploid genome. I rigorously benchmarked this pipeline using simulations with the following polarization of variants to distinguish between deletion and insertion events.

Subsequently, my attention shifted to assessing structural variability between individuals. I applied common population genomic methods to a high-quality dataset of SVs, gaining preliminary insights into their potential role in population differentiation within *Crassostrea ariakensis*. This species, widespread across East Asian estuaries, exhibits adaptability to a broad range of temperature and salinity conditions and displays a clear population structure based on previous SNP analyses. Therefore, it serves as an ideal case for studying the contribution of transposable elements and related structural variants to population differentiation.

3. Chapter I

Multi-tissue RNA-Seq Analysis and Long-read-based Genome Assembly Reveal Complex Sex-specific Gene Regulation and Molecular Evolution in the Manila Clam

Ran Xu*, Jacopo Martelossi*, Morgan Smits*, Mariangela Iannello*, Luca Peruzza, Massimiliano Babbucci, Massimo Milan, Joseph P. Dunham, Sophie Breton, Liliana Milani, Sergey V. Nuzhdin, Luca Bargelloni, Marco Passamonti, and Fabrizio Ghiselli

*Co-First authors

Note:
This chapter has been published as a research article on Genome Biology and Evolution under the DOI: 10.1093/gbe/evac171. All supplementary materials can be found on its online version.

Abstract

The molecular factors and gene regulation involved in sex determination and gonad differentiation in bivalve molluscs are unknown. It has been suggested that doubly uniparental inheritance (DUI) of mitochondria may be involved in these processes in species such as the ubiquitous and commercially relevant Manila clam, *Ruditapes philippinarum*. We present the first long-read-based de novo genome assembly of a Manila clam, and a RNA-Seq multi-tissue analysis of 15 females and 15 males. The highly contiguous genome assembly was used as reference to investigate gene expression, alternative splicing, sequence evolution, tissue-specific co-expression networks, and sexual contrasting SNPs. Differential expression (DE) and differential splicing (DS) analyses revealed sex-specific transcriptional regulation in gonads, but not in somatic tissues. Coexpression networks revealed complex gene regulation in gonads, and genes in gonad-associated modules showed high tissue specificity. However, male gonad-associated modules showed contrasting patterns of sequence evolution and tissue specificity. One gene set was related to the structural organization of male gametes and presented slow sequence evolution but high pleiotropy, whereas another gene set was enriched in reproduction-related processes and characterized by fast sequence evolution and tissue specificity. Sexual contrasting SNPs were found in genes overrepresented in mitochondrialrelated functions, providing new candidates for investigating the relationship between mitochondria and sex in DUI species. Together, these results increase our understanding of the role of DE, DS, and sequence evolution of sex-specific genes in an understudied taxon. We also provide resourceful genomic data for studies regarding sex diagnosis and breeding in bivalves.

Key words: long-read genome assembly, differential transcription, co-expression network, alternative splicing, tissue specificity, sexual contrasting genetic markers.

Introduction

Bivalves show an astonishing wealth of diverse life histories, adaptation, and phenotypic plasticity. Numerous species have become important biological models for monitoring pollution, studying adaptation to climate change, and developing biomedical tools (Krishnakumar et al. 2018; Harris et al. 2020). Moreover, many bivalves have a global economic importance, providing an essential source of protein through aquaculture and fishing (Wijsman et al. 2019).

Despite their important ecological and economic roles, and their biodiversity within Mollusca phylum, bivalves (and Molluscs in general) have been poorly investigated at the molecular level, compared to other animal groups. This is even more surprising if we consider that bivalves show peculiar biological features which make them ideal model systems in fields like evolutionary, molecular, and developmental biology (Ghiselli et al. 2021a). Bivalvia present a variety of sexual reproduction modes, ranging from strict gonochorism to sequential or simultaneous hermaphroditism (Breton et al. 2018). So far, no heteromorphic sex chromosomes have been found in bivalves and the molecular factors involved in sex determination and gonad differentiation are unknown: it has been proposed that the variety in reproduction modes is primarily due to modifications of the same genetic pathways (Breton et al. 2018). Given this context, investigating tissue-specific gene regulation may help identify gene networks involved in sex determination and gonad differentiation. In other animal species investigated so far, regulation of gene transcription, in terms of differential expression (DE) and differential splicing (DS), is known to be involved in resolving sexual conflicts (Ingleby et al. 2015; Ghiselli et al. 2018; Rogers et al. 2021). Indeed, most of the sex-specific characters are the result of genes that are differentially expressed between sexes (sex-biased genes), and rapid sequence evolution of sex-biased genes has been observed in animals (Ellegren and Parsch, 2007; Mank et al. 2007; Harrison et al. 2015; Lipinska et al. 2015; Dean and Mank, 2016; Ghiselli et al. 2018). Additionally, several studies revealed that a large proportion of genes undergo sex-specific splicing, indicating a role of DS in sex-specific development and physiology (Telonis-Scott et al. 2009; Griffin et al. 2013; Rogers et al. 2021). In species with sexual dimorphism, sexual selection was suggested as a driver of sex-biased patterns of gene expression and splicing, whereas gene expression breadth, protein–protein interaction, codon usage, and pleiotropy may also contribute to sex bias (Mank et al. 2008; Harrison et al. 2015; Grath and Parsch, 2016; Whittle and Extavour, 2019). If and how these factors shape the evolution of species lacking sexual dimorphism, as in the case of most bivalves, has yet to be explored.

Another interesting feature, found in more than 100 bivalve species, is the presence of the doubly uniparental inheritance (DUI) of mitochondria. In DUI species, two distinct lineages of mitochondrial DNA (mtDNA) are inherited by the offspring: one lineage (F-type) is transmitted through eggs and it is present in both sexes, the other (M-type) is transmitted through sperm and it is mainly found in males (less often in females and in lower abundance) where it is most abundant in gonads (Ghiselli et al. 2019; Ghiselli et al. 2021b). Numerous works have sought to elucidate the molecular mechanisms beyond DUI and speculated on the evolutionary process behind the maintenance of divergent mtDNA lineages within species, especially considering that heteroplasmy is generally considered an unfavorable condition, generally converging on the hypothesis that DUI might be linked to sexual differentiation (Breton et al. 2018; Capt et al. 2018). Having two different mitochondrial genomes with sex-specific and tissue-specific distribution opens up questions about the existence of tissue and sex-specific coordination of gene regulation, namely regarding nuclear genes involved in mitochondrial biology (Ghiselli et al. 2021b; Maeda et al. 2021; Xu et al. 2022).

In this work, we performed a de novo long-read genome assembly and a multi-tissue RNA-seq analysis of *R. philippinarum*, a gonochoric bivalve species with DUI, to investigate sex-specific and tissue-specific gene regulation and molecular evolution. More in detail, we compared differential gene transcription and DS across tissues for the first time in bivalves, focusing on differences between somatic tissues and gonads. We also investigated the relationship between tissue-specific co-expressed modules and protein sequence evolution. Our aim was to identify genes and gene networks that could have a major role in tissue differentiation, and characterize their patterns of evolution. We found that gonads, compared to somatic tissues, show a more complex gene regulation, as multiple coexpression submodules are present within the same tissue. Some of these submodules are also sex-specific, showing peculiar and divergent patterns of sequence evolution. We finally identified hub genes for each tissue-specific module, which are likely to be crucial for tissue specification, and we highlighted those that could have a central role in sex determination/differentiation and those that could have a possible role in DUI.

Results

Genome Sequencing, De Novo Assembly, and Whole-Genome Alignment

PacBio sequencing consisted of 54 SMRT cells that yielded ∼4 M reads (36.5 Gb) of raw sequences with a median length of ∼45 Kb. The Illumina sequencing resulted in ∼145 M reads (∼75 Gb) for the short insert library, and ∼48 M reads (∼25 Gb) for the long insert library. After trimming both Illumina libraries, a total of ∼180 M PE reads were kept (supplementary table S1, Supplementary Material online).

We estimated a genome size of ∼1.37 Gb (supplementary table S2, Supplementary Material online) giving an expected genome coverage of ∼25 × and ∼72 × for, respectively, the PacBio and Illumina libraries. The estimated genome size resulted concordant with previous kmer-based estimations which range from 1.32 Gb (Yan et al. 2019) to 1.37 Gb (Mun et al. 2017), but quite smaller from than the 1.97 Gb estimation obtained by the Feulgen method (González-Tizón et al. 2000). The heterozygosity and the repetitive content were estimated to range, respectively, from 4% to 3.7% and from 61.2% to 48.2%, depending on the kmer size (supplementary table S3 and supplementary fig. S1, Supplementary Material online). After three rounds of purging and polishing, the final version of the assembly consisted in 15,908 contigs with a N50 of 183 Kb, a total genome size of 1.41 Gb and a mean GC content of 0.32. We identified 884 out of 954 Metazoa BUSCO orthologs (92.7%), of which 802 were present as single copy (84.1%), and 82 as duplicates (8.6%). Missing genes represent 4.7% of the core gene set, whereas only 2.6% were identified as fragmented (table 1). KAT analyses show a kmer completeness of 52,48% (table 1; supplementary fig. S2, Supplementary Material online), and 95% and 98% of the short and long reads were successfully remapped on the assembly, respectively, with a median coverage depth of 53.42 and 22.69 (table 1). Blobtools identified 20 contigs as possible bacterial contaminations. Another six contigs were annotated as belonging to Priapulida, whereas only one to Zoopagomycota. These contigs cover 937,293 bp of the total assembly size (0.0007%) and were removed from the final version of the assembly. For a direct comparison between our newly produced assembly and the short-reads-only chromosome-level assembly from Yan et al. (2019), from now on "CRph genome" we performed a pairwise whole-genome alignment (WGA). Out of the 15,908 contigs that composed our assembly, 99.2% had at least one alignment block to the CRph genome with the majority of alignments involving an assembled chromosome. In total, all alignment blocks represented 80% of our assembly and 77.4% of the CRph genome (supplementary table S4, Supplementary Material online). Detailed results and discussion for genome assembly and comparison can be found in Supplementary Materials, Methods and Results, Supplementary Material online.

Table 1: Summary statistics of the long reads based manila clam assembly.

Assembly genome size	1,409,123,410 bp
Number of contigs	15,908
Average contig length	88,579.55 bp
Largest contig	1,574,940 bp
N50	182,737 bp
N90	37,082 bp
BUSCO	C:92.7% [S:84.1%, D:8.6%], F:2.6%, M:4.7%, n:954
Mapped short reads	343,975,629 (95%)
Mapped long reads	12,691,865 (98%)
Median short reads depth	53.42
Median long reads depth	22.69
Kmer completeness	52,48%
GC content	0.32

Genome Annotation

Using de novo approaches, we built up a starting consensus library composed of 5,600 sequences (3,197 and 2,403 by RepeatModeler and MITE_Tracker, respectively). We added another 1,031 TEs already characterized in molluscs and retrieved from RepBase. After removal of genes/gene fragments, tandem, and low copy number repeats (<5 good hits on the genome), we used a total of 2,332 nonredundant consensus sequences to annotate the *R. philippinarum* repeatome. Overall, 39.7% of the genome was masked by interspersed repeats with a prevalence of cut and paste (DNA + MITEs) and Rolling Circle TEs (14.7% Unknown elements; 9.23% MITEs; 6.1% Rolling circle; 3.5% DNA; 2.95% LINE; 1.84% LTR; 1.25% SINE) (supplementary table S5, Supplementary Material online).

The annotation pipeline generated 34,505 gene models with an average length of 8,053 bp (6.4 mean exons per gene; mean exon length: 212 bp). Of these, 22,103 (64%) had a positive match by blastx against the Swiss-prot database (supplementary table S5, Supplementary Material online). The Annotation Edit Distance (AED), a metric useful to measure the agreement between predicted gene models and external evidence, where a value of 0 indicates full agreement and 1 no external support (Holt, 2011), identified 29,322 (85%) gene models with an AED ≤0.5 and a mean equal to 0.18. BUSCO scores on the predicted proteomes using the Metazoa odb10 reference database resulted in C:83.4%[S:74.5%, D:8.9%], F:8.8%, M:7.8%. The percentage of RNA-seq reads mapped to the genome is reported in supplementary table S6, Supplementary Material online.

Differential Expression and Co-Expression Network

To investigate the global expression patterns in all tissues of both sexes, a PCA analysis was performed in DESeq2. As shown in figure 1A, different tissues presented distinct expression profiles, and while expression patterns between female and male somatic tissues were quite similar, large differences were found in gonads. Consistently, the number of differentially expressed genes (DEGs) between female and male adductor muscles and mantles were low (578 and 22, respectively), whereas the number of DEGs between gonads were 6,167, including 3,024 femalebiased DEGs and 3,143 male-biased DEGs (fig. 2A). The comparisons of DEGs between pairwise tissues were performed for males and females separately. Generally, the number of DEGs between somatic tissues (adductor muscle vs. mantle) was less than the number of DEGs between somatic tissue and gonad (e.g., gonad vs. mantle) (supplementary table S7, Supplementary Material online).

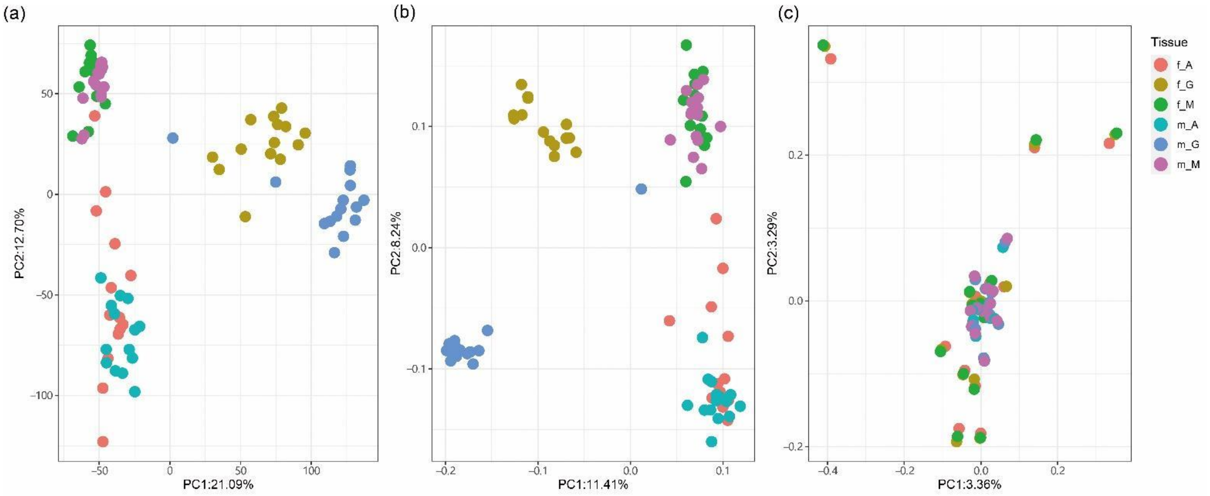


Figure 1: PCA plot for gene expression (**a**), alternative splicing (**b**) and genotype (**c**). Each dot represents a sample and each color represents a tissue type; f_A: female adductor; f_G: female gonad; f_M; female mantle; m_A: male adductor; m_G: male gonad; m_M: male mantle.

A large proportion of DEGs in females in pairwise tissue comparisons overlapped with the corresponding DEGs in males (supplementary table S8, Supplementary Material online). In all pairwise tissue comparisons, 1,787 and 2,277 genes were differentially expressed across all three tissues in females and males, respectively (supplementary fig. S3A, Supplementary Material online), and 1,009 of these DEGs were shared between females and males. Additionally, to investigate the genes showing significant sex-by-tissue interactions, we performed a DE analysis using Likelihood ratio test. The number of genes across tissues, between sexes, and between sexes across tissues was 17,802, 6,321, and 4,430, respectively.

A tissue-specific gene co-expression network was constructed to investigate gene regulatory relationships in tissue-associated modules. A total number of 8,640 genes were assigned to 10 modules (fig. 2B). The blue module (1,334 genes) and the green module (790 genes) showed high association with male gonads, whereas the pink module (417 gfenes) was associated with female gonads. Moreover, yellow (977 genes), magenta (232 genes), and purple (80 genes) modules were associated with both female and male gonads, and turquoise (2,718 genes) and brown (1,749 genes) were associated with somatic tissues. Moreover, we retrieved "hub" genes which rank in the top 5% of kWithin in each module and represent high connection with the other genes. Hub genes and functional annotations in each module are listed in supplementary table S9, Supplementary Material online. We found that the percentage of hub genes with annotation varied across modules (fig. 2C). These genes included malegonad-specific SRY-box transcription factor 30 (sox30) in the male-gonad-specific blue module, and mating-type-like protein ALPHA2 (mtlalpha2) in the female-gonad-specific pink module. For genes in the co-expression network, we measured the connectivity among genes in the same module (intramodular connectivity: kWithin), the connectivity between genes from different modules (intermodular connectivity: kOut), and the global connectivity (kTotal = kWithin + kOut). In this tissue-specific co-expression network, kWithin represents within module connectivity specific to one or multiple associated tissue types (specific connectivity), whereas kOut represents the connectivity of one gene to the genes outside the module in the other tissue types (broad connectivity). The distribution of intramodular connectivity (kWithin) and intermodular connectivity (kOut) for genes in each module is shown in figure 2D, and the statistical tests for pairwise comparisons of connectivity between modules are shown in supplementary table S10, Supplementary Material online. Generally, the gonad-associated blue module and mantle-associated turquoise module presented significantly higher kWithin compared with the overall distribution, whereas another gonad-associated green module presented significantly higher kOut (fig. 2D). A predominant number of 1,253 (93.9%), 579 (73.3%), and 397 (95.2%) genes in the blue, green, and pink modules, respectively, were also DEGs between female and male gonads. Besides, the kWithin for DEGs in gonad-associated blue, green, and pink modules were significantly higher than non-DEGs between female and male gonads (fig. 3A).

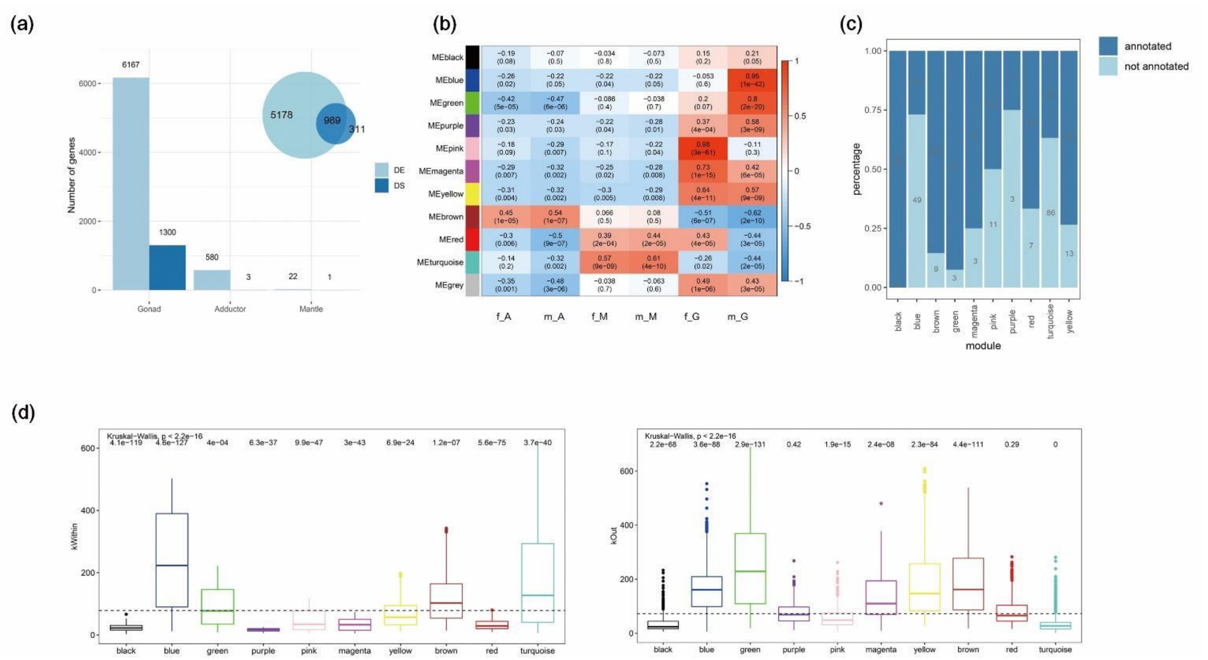


Figure 2: Differentially expressed, spliced, and co-expressed genes across tissue types. (**a**) The number of differentially expressed (DE) and spliced (DS) genes between females and males in each tissue. The Venn plot on the top-left represents the overlap between DE and DS genes in the gonad. (**b**) Module-tissue association based on the gene expression. Each row with color corresponds to a co-expression module, and each column represents a tissue-type. The correlations and p values between module and tissue are shown in each cell. (**c**) The proportion of annotated and not annotated hub genes in each module. Numbers in the bars indicate the number of hub genes for each module. (**d**) The distribution of within (kWithin) module connectivity and outside (kOut) module connectivity for genes in the co-expression modules. Wilcoxon rank-sum test with FDR corrections was used to compare the distribution of kWithin and kOut in each module to the overall distribution and the significance were shown on the top of the boxplot. ***, *P*<0.0001; **, *P*<0.001; *, *P*<0.05; ns, non-significant. The dash line indicates the median of the overall distribution.

A GO enrichment analysis was applied to explore the predicted functions of different subsets of genes, and the results are shown in supplementary table S11, Supplementary Material online. Considering the low number of DEGs between female and male mantles, we did not perform the enrichment analysis on this subset of genes. The significantly enriched GO terms in adductor muscles between males and females were related to microtubule-based process and motor activity (supplementary table S11, Supplementary Material online). Reproduction and cell cycle-related processes were significantly enriched for the DEGs between female and male gonads, and for the DEGs between sexes across tissues (supplementary table S11, Supplementary Material online). Reproduction-related processes were also enriched in the male gonad-associated blue module (supplementary table S11, Supplementary Material online). Notably, Kelch-related domains were significantly overrepresented in the blue module (supplementary table S12, Supplementary Material online).

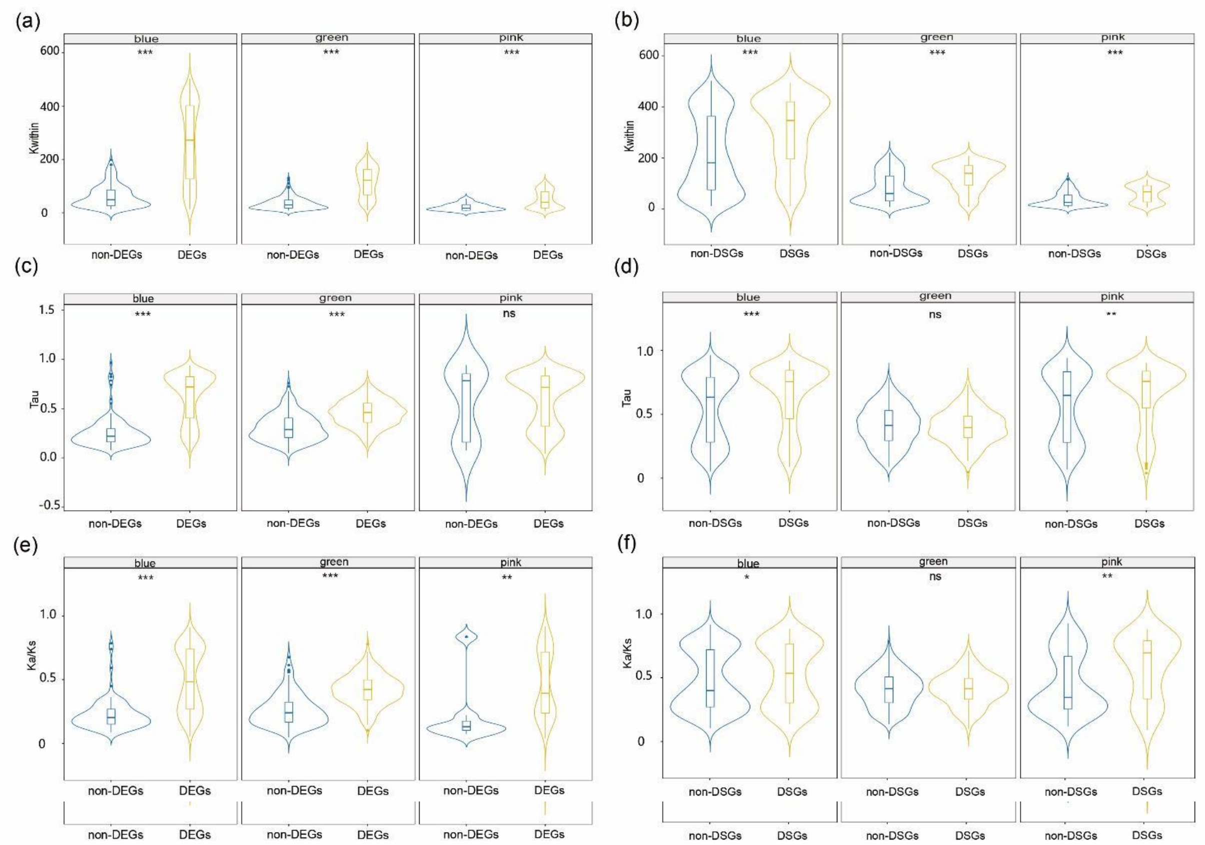


Figure 3: Comparisons between differentially expressed and non-differentially expressed genes, and between differentially spliced genes and non-differentially spliced genes in female and male gonad-associated modules. (**a**), (**c**), and (**e**) represent comparisons of the connectivity, tissue specificity, and sequence evolutionary rate between differentially expressed genes and non-differentially expressed genes. (**b**), (**d**), (**f**) represent the comparisons of the connectivity, tissue specificity and sequence evolutionary rate between differentially spliced genes and non-differentially spliced genes.

The genes co-expressed in the other male gonad-associated (green) module appeared to over-represent some general functions, with processes like “organelle assembly”, “cell project”, and “catalytic activity” being enriched. In the female gonad-associated pink module, processes related to transferase and protein metabolic activities, and homeobox-related domains were significantly enriched (supplementary table S11 and S12, Supplementary Material online). Different functional processes were enriched in the three gonad-associated modules (magenta, purple, and yellow) such as “cell adhesion” (purple and magenta modules), homeostatic related processes (purple module), and processes related to tissue development (magenta module) (supplementary table S11, Supplementary Material online). Intriguingly, for genes in the yellow module, processes related to DNA repair, DNA replication, and gene expression were significantly enriched. In mantle-associated turquoise modules, genes were overrepresented in the immune-related process and metal ion binding.

Differential Splicing Analysis

Consistent with expression profiles, splicing patterns also differed across tissues, and differences between females and males were observed in gonads but not in somatic tissues (fig. 1B). Global alternative splicing events for each tissue are shown in supplementary fig. S4, Supplementary Material online. Generally, skipping exon (SE), alternative 5′ splicing (A5), and alternative first exon (AF) accounted for a large proportion in all tissues, while retained intron (RI) and mutually exclusive exons (MXE) were the least represented events in all tissues. Moreover, alternative splicing in gonads and mantles seemed to be more frequent than in adductor muscles. Despite the pervasiveness of alternative splicing in all tissues, the number of genes showing DS between females and males in each tissue, and between pairwise tissues were far less compared with DEGs. The number of differentially spliced genes (DSGs) between female and male adductor muscles, mantles, and gonads were 3, 1, and 1,300, respectively (fig. 2A). Notably, among all the 1,300 DSGs between female and male gonads, 989 (76%) were also differentially expressed between female and male gonads. We also retrieved these DSGs in three sex-associated co-expression modules (blue, green, and pink) and we found that the DSGs in these modules showed significantly higher kWithin than non-DSGs (fig. 3B). The number of DSGs between gonads and somatic tissues was higher than that found between two somatic tissues (supplementary fig. S3 and supplementary table S6, Supplementary Material online). Moreover, in all these comparisons between different tissues, DEGs and DSGs were largely overlapping for both females and males, and around 80–90% DSGs between gonads and somatic tissues were also DEGs (supplementary table S8, Supplementary Material online). Some of these DSGs overlapped with DEGs or sex-associated modules (listed in supplementary table S9, Supplementary Material online), and the large amount of overlapping genes between DSGs and DEGs in gonads also resulted to have many processes in common, such as “microtubule-based process” and “cellular process”. Additionally, functional characterization of DSGs that did not overlap with DEGs, highlighted their involvement in chromatin remodeling and mRNA catabolic processes.

Tissue Specificity in the Co-Expression Network

The tissue specificity index Tau ranged from 0.2 to 0.8 for most genes, while only a small proportion of genes showed extremely high tissue-specific (>0.8) or broad (<0.2) expression (supplementary fig. S5, Supplementary Material online). Kruskal–Wallis test was used to assess if Tau distribution differs across modules and we found that Tau values in different co-expression modules varied markedly (Kruskal–Wallis test: $P < 0.001$). A Wilcoxon rank-sum test with FDR corrections was used to compare the distribution of Tau in each module to the overall distribution. Generally, in somatic associated red and brown modules, genes showed relatively low Tau values, indicating low tissue specificity (supplementary fig. S6, Supplementary Material online). By contrast, we found relatively high and variable Tau values in most gonad-associated modules, except for the male gonad-associated green module and gonad-associated yellow module, which had relatively low Tau values, with median values at around 0.4 and 0.3, respectively (supplementary fig. S6, Supplementary Material online). Interestingly, we found that the yellow and green modules also showed relatively high intermodular connectivities, indicating that genes in these two modules showed also high connections with other tissues (fig. 2D).

We further investigated the correlation between Tau and network connectivity using Spearman's rank sum test. We found positive correlation between whole network connectivity (kTotal) and tissue specificity (Tau) (Spearman's $R = 0.24$, $P < 2.2E\text{-}16$), and between intramodular connectivity (kWithin) and tissue specificity (Spearman's $R = 0.34$, $P < 2.2E\text{-}16$), but a weak correlation between intermodular connectivity (kOut) and tissue specificity (Spearman's $R = -0.07$, $P = 6.433E\text{-}11$). Moreover, we found significant positive correlation between tissue specificity Tau and kWithin, kTotal in most tissue-associated modules such as blue, pink, and turquoise modules, indicating that genes with high tissue specificity also presented high connection in the specific tissue type (fig. 4A and supplementary fig. S7A, Supplementary Material online). Additionally, the negative correlation between kOut and Tau was also observed in most modules except for blue, green, and yellow modules, where a positive correlation was observed (supplementary fig. S7B, Supplementary Material online). We further investigated the tissue specificity for DEGs and DSGs in the co-expression network, mainly focusing on gonad-associated modules because of the low number of DEGs and DSGs in somatic tissues. Wilcoxon rank-sum test was used to assess differences in tissue specificity and network connectivity between DEGs and non-DEGs, and between DSGs and non-DSGs. In the male gonadassociated blue module, DSGs and DEGs presented significantly higher Tau values than non-DSGs and nonDEGs (fig. 3C and D). In the green module, DEGs also presented significantly higher Tau values than non-DEGs, whereas Tau values between DSGs and non-DSGs were not significantly different from each other (3C and D). However, in the pink module, Tau showed no significant difference between DEGs and non-DEGs, but DSGs presented slightly higher Tau values than non-DSGs (fig. 3C and D).

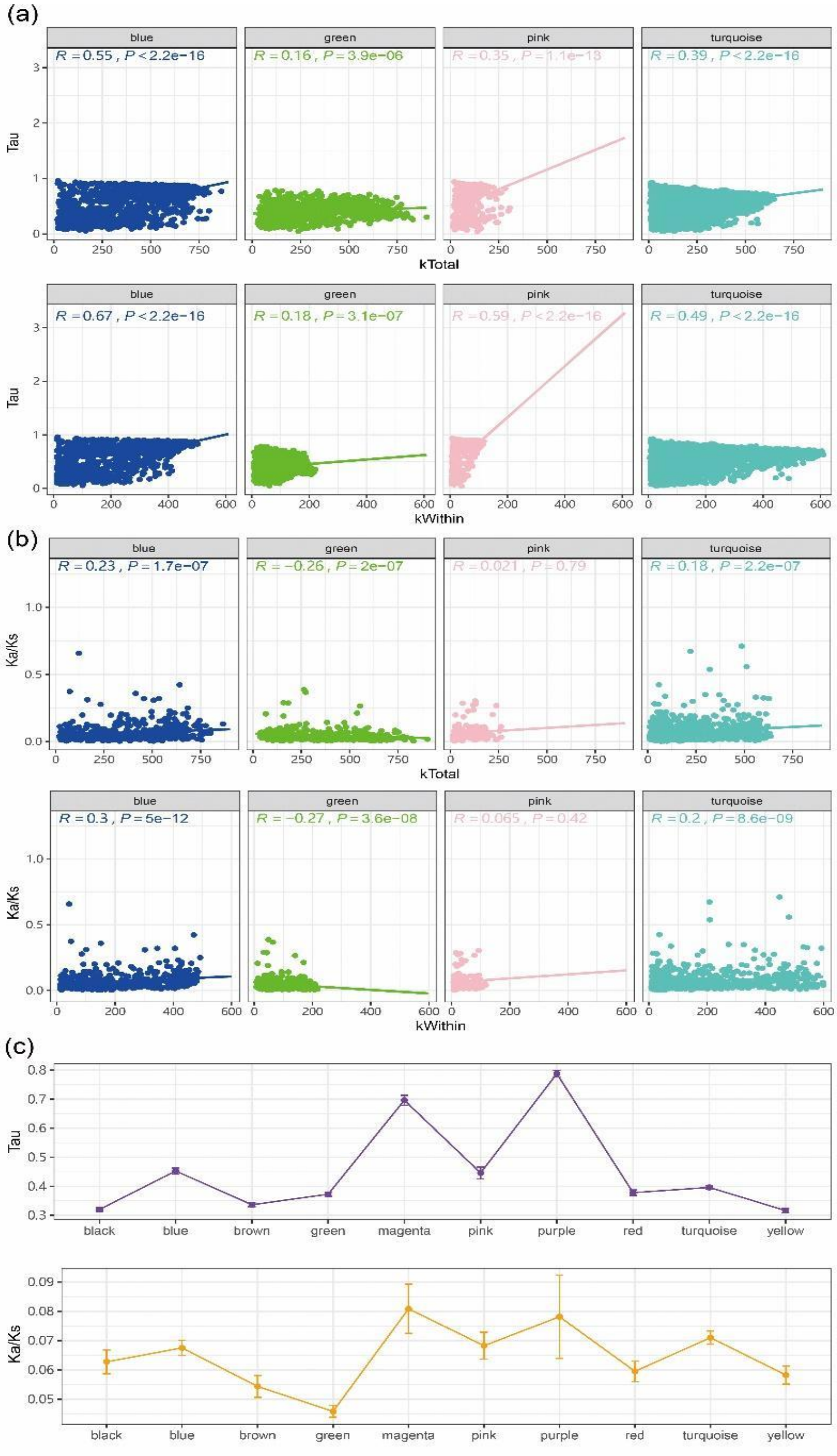


Figure 4: The relationship between network connectivity and tissue specificity, evolution rate. (**a**) The correlation between tissue specificity index (Tau) and total connectivity (kTotal), within module connectivity (kWithin) in four tissue-associated modules. (**b**) The correlation between evolutionary rate and total network connectivity (kTotal), within module connectivity (kWithin) for four tissue-associated modules. The Spearman's correlation (R) and p values were shown on the top. (**c**) The trends of tissue specificity index (Tau) and evolutionary rate (Ka/Ks) in the co-expression modules. Average value (each dot) and standard error (error bar) was used for each module.

Variation in the Rate of Sequence Evolution Across Co-Expression Modules

Kruskal–Wallis test, followed by Wilcoxon rank-sum test with FDR corrections, was used to test Ka/Ks differences across modules. Ka/Ks distribution also varied in different co-expression modules, with the male gonad-associated blue module and mantle-associated turquoise module presenting a significantly higher Ka/Ks than the overall values, and the green module presenting a significantly lower Ka/ Ks than the overall values (supplementary fig. S8, Supplementary Material online). Spearman's rank sum test was used to measure the correlation between network connectivity and Ka/Ks, and between Tau and Ka/Ks. We found no significant correlation between general network connectivity (kTotal) and evolutionary rate (kTotal: Spearman's R = −0.0044, P = 0.78). However, when we investigated this relationship in each module, we found that genes in male gonad-associated blue module and mantle-associated turquoise module showed significantly positive correlation between network connectivities (both kTotal and kWithin) and evolutionary rates, while genes in the other male gonad-associated module (green module) showed significantly negative correlation between connectivities and evolutionary rates (fig. 4B and supplementary fig. S9A, Supplementary Material online). Most modules presented no significant correlation between intermodular connectivity and evolutionary rate (supplementary fig. S9B, Supplementary Material online).

Tau was positively correlated with Ka/Ks (Spearman's R = 0.17, P < 2.2E-16) in some tissue-associated modules. Similar to the correlation between connectivity and Ka/Ks, significant positive correlation between Tau and Ka/Ks was detected in blue and turquoise modules (supplementary fig. S10, Supplementary Material online). In spite of the lack of correlation in most modules, the Tau and Ka/Ks values showed similar trends across different modules (fig. 4C). Combined with the tissue specificity analysis above, it appears that genes in the male gonad-associated blue module and mantle-associated turquoise module with high intramodular connectivity and tissue-specificities also presented high evolutionary rates, while genes in the green module with high connections to outside the modules had a lower evolutionary rate.

Wilcoxon rank-sum test was used to assess differences in Ka/Ks between DEGs and non-DEGs, and between DSGs and non-DSGs. We also observed significant differences in evolutionary rate between DEGs and non-DEGs, and between DSGs and non-DSGs in the female and male gonad-associated modules (fig. 3E and F). In all three gonad-associated modules, DEGs presented significantly higher Ka/Ks than non-DEGs. Likewise, we found that DSGs in blue and pink modules also showed significantly higher Ka/Ks than non-DSGs, but such result was not detected in the male gonad-associated green module.

Contrasting SNPs

We first retrieved SNPs for each sample separately and found that polymorphism in different tissues of the same individual was extremely low (fig. 1C). Thus, to retrieve sexspecific SNPs, we divided all the samples into female and male groups but merging the three tissues of the same individual together. We detected 750,790 total variants between male and female groups, of which 676,009 were SNPs. Of these, 252,858 SNPs were present in at least 80% of individual samples with a minimum quality score of 20. Filtered SNPs from male and female groups were analyzed using BayPass for contrast based on genotype counts, yielding 614 SNPs significantly contrasting between the two sexes ($P < 0.001$). Annovar merged the selected SNPs with the genome assembly annotation to identify the locations of each marker, specifying that of the 614 significantly contrasting SNPs, 381 were in exonic regions (supplementary table S13, Supplementary Material online). Finally, exonic SNPs from male and female groups were searched against a set of SNPs from a DNA pooled sequencing experiment of Mediterranean and Atlantic *R. philippinarum* populations (Smits et al. 2020) revealing that the two datasets contained 260 exonic SNPs in common. Genes containing contrasting SNPs are listed in supplementary table S9, Supplementary Material online, and some of them were also identified in DEGs, DSGs, or tissue-associated modules such as ankyrin repeat domain-containing protein 17 (ankrd17), double-strand-break repair protein rad21-like protein (rad21), folliculin (flcn), transcriptional regulator ATRX (atrx). Functional enrichment indicated that genes containing contrasting SNPs were also involved in processes such as “mitochondrial transmembrane transport”, “protein localization to organelle”, and “chromatin remodeling” (supplementary table S11, Supplementary Material online).

Discussion

In the present work, we sequenced and assembled a new long-read-based draft genome of the Manila clam *R. philippinarum*. Notably, this represents the first effort to sequence and assemble a wild (i.e., not inbred) specimen genome relying both on short and long-read data, and the first long-read genome assembly for this species. This genome assembly provides novel resources for Altantic populations, which has been observed to be genetically divergent to the Asian population, but very similar to the European population (Cordero et al. 2017). Additionally, the genome assembly allowed us to investigate tissue-specific gene expression and splicing patterns in *R. philippinarum*. Despite the increasing resources in terms of DNA and RNA sequences, most of the molecular pathways involved in tissue characterization are unknown in bivalves. Therefore, we also constructed a tissue-specific co-expression network whose analysis has been useful to identify candidate genes involved in the same biological processes. Genes showing the highest connection within a co-expression module are likely to have a central role in the corresponding module and are defined as “hub genes”. The analysis of hub genes has recently led to the identification of regulatory elements and biomarker targets for therapies (Grimes et al. 2019). We used high tissue specificity and high intramodular connectivity as proxies to identify networks of genes with tissue-specific functions, whereas low tissue specificity and high intermodular connectivity as a proxy of pleiotropy. We finally investigated the rate of protein evolution of genes in different modules and highlighted the complexity of gene regulation and sequence evolution in gonads.

Both Differential Expression and Differential Splicing Shape Tissue-Specific Transcriptional Profiles in Bivalves

Different expression patterns between females and males have been investigated by several studies in gonochoric and sequential hermaphroditic bivalves, but mainly focused on the reproductive tissue alone (gonads), or across developmental stages (Ghiselli et al. 2012, 2018; Capt et al. 2018, 2019; Yue et al. 2018; Broquard et al. 2021). When extending the analyses of differential expression to multiple tissues, and adding the investigation of DS, we found that in *R. philippinarum* both DE and DS separate samples according to tissues (fig. 1). This suggests that both alternative splicing and DE have a central role in shaping tissue-specific transcriptional profiles in this species, and possibly in all bivalves. Additionally, both DE and DS analyses reveal a sex-specific transcriptional regulation in gonads, which leads male and female gonads to cluster separately from each other, a pattern that was not observed in somatic tissues. In other organisms, sex-biased genes and alternative splicing are reported to be responsible for most of the phenotypic differences between sexes (Parsch and Ellegren, 2013; Harrison et al. 2015; Ingleby et al. 2015; Lipinska et al. 2015; Dean and Mank, 2016; Rogers et al. 2021). In these cases, the majority of such genes are involved in sexual dimorphism and mating behavior. Such traits are absent in most bivalves (including *R. philippinarum*), and genes with gonad-specific and sex-specific transcriptional profiles are likely to be involved in sex determination, gonad specification, and gametogenesis. A functional annotation analysis of DE and DS genes, comparing male and female gonads and comparing gonad and somatic tissues, shows an enrichment of terms involved in reproduction, cell project organization, chromatin remodeling and DNA replication. These genes can help elucidating the molecular mechanism of gonad specification and sex differentiation in bivalves (see "Contrasting SNPs and Hub Genes Potentially Involved in Sex Determination and Mitochondrial Functions").

Co-Expression Network Analysis Reveals High Complexity in Gonad Gene Regulation

Although somatic tissues are usually associated with one or two co-expression modules, gonads are characterized by multiple, sometimes sex-specific, modules with different co-expression patterns (fig. 2), revealing a more complex gene regulation. Generally, genes in gonad-associated modules are characterized by higher tissue specificity compared with somatic tissue-associated modules (with the exception of the male gonad-associated green module, see below). Additionally, sequence evolution in each module follows a similar trend to tissue specificity, and such a relationship is particularly significant in gonad-associated blue and yellow modules (fig. 3C and supplementary fig. S10, Supplementary Material online). This trend is expected, because tissue-specific genes are less constrained compared with the pleiotropic genes, and they are usually characterized by higher sequence evolution (Dean and Mank, 2016; Mank et al. 2008; Meisel 2011).

Besides the co-occurrence of multiple co-expression networks in gonads, an additional level of complexity specifically characterizes male gonads, where different networks showed opposite trends of tissue specificity and rate of protein evolution. In more detail, the blue module shows a particularly high tissue specificity and rate of protein evolution; this is a pattern in line with the higher evolutionary rates of male-biased genes observed in a wide range of animals (Grath and Parsch, 2012; Parsch and Ellegren, 2013; Harrison et al. 2015). By contrast, the green module significantly deviates from what is observed in other gonadspecific networks: genes in this green module are indeed pleiotropic and constrained by a lower rate of protein evolution.

Additionally, in contrast to other gonad-associated modules, there is no significant difference between DSGs and non-DSGs for tissue specificity and evolutionary rate in the green module, indicating that splicing may be underrepresented in highly pleiotropic genes. Such results reveal that a combination of genes with different transcription patterns, tissue specificity, and rate of protein evolution is required for male gonad differentiation.

When we looked at the functional annotation of genes belonging to the blue module, we found an enrichment of GO terms involved in reproduction. The fact that genes from this module are characterized by a faster sequence evolution is consistent with what is found in a wide range of species (Ellegren and Parsch, 2007; Grath and Parsch, 2012; Parsch and Ellegren, 2013; Harrison et al. 2015), in which male-biased genes are characterized by faster evolution. Interestingly, more than 70% of hub genes from this fast-evolving, highly tissue-specific module could not be annotated. This reveals that genes with a putative central role in male reproduction of *R. philippinarum* are mostly uncharacterized; it would be interesting to understand whether such genes show a male-biased transcription also in other bivalve or mollusc species, and investigate their evolution and role in male functions. Interestingly, most of hub genes from the other male-gonad-specific module (green module), are included in the KEGG BRITE category "cilium and associated proteins", and they include sperm flagellum proteins and motile cilium-associated proteins. This module seems therefore to be majorly involved in the "structural" component of spermatogenesis, and it is not surprising that these genes are characterized by a slower evolution, as an improper formation of spermatozoa would likely undermine reproduction. Among the hub genes in this module, it is worth mentioning the presence of three out of five tektin genes. The tektin domain is also significantly enriched in the green modules. Tektins are cytoskeletal proteins associated with microtubules, and deficiency in these proteins are known to influence sperm motility and cause male infertility (Yan, 2009).

Contrasting SNPs, Hub Genes and Domains Potentially Involved in Sex Determination and Mitochondrial Functions

Heteromorphic sex chromosomes are absent in bivalves, and sex determination is thought to be polygenic with the additional influence of environmental factors as potential triggers of sex changes (Breton et al. 2018; Dalpé et al. 2022). Identification of sex-specific SNPs is crucial for accurate sex diagnosis, breeding, and understanding of sexdetermination mechanisms. In this study, we revealed 614 high-confidence contrasting SNPs between males and females, which provide potential genetic markers for sex identification in bivalves. Interestingly, we found that genes containing contrasting SNPs were overrepresented in the processes of protein targeting and protein localization to the mitochondrion. These genes included coiled-coil-helix-coiled-coil-helix domain-containing 2 (cdchd2), mitochondrial carrier protein Rim2, mitochondrial import inner membrane translocase subunit Tim16, and mitochondrial import inner membrane translocase subunit Tom22 (supplementary table S9, Supplementary Material online), with the latter three genes being involved in translocation of nuclear-encoded proteins into mitochondria (Herrmann and Neupert, 2013). Cdchd2 was found to be involved in diverse functions in model animals, including mediating oxidative phosphorylation, responding to hypoxic stress, regulating cell migration, and mitochondrial apoptosis (Kee et al. 2021). It has been proposed that DUI bivalves might have an unconventional sex determination/differentiation system that involves mitochondrial genomes and/or their products (proteins and/or RNAs), and this system may require an appropriate recognition/discrimination process between mitochondrial and nuclear factors (Breton et al. 2011, 2018; Ghiselli et al. 2013; Milani et al. 2013; Zouros, 2020). Although finding sexspecific SNPs in genes with mitochondrial function does not serve as direct evidence of the role of mitochondria in sex determination/differentiation in bivalves, it provides interesting candidate genes for testing such hypothesis in future experiments.

We also identified candidate genes and domains potentially associated with sex determination/differentiation mechanism in bivalves that are known to have a role in such processes in model animals. Among these, SRY-box transcription factor 30 (sox30), a putative homolog to mammal sex-determining gene sry, is a hub gene of the male gonad-associated blue module (supplementary table S9, Supplementary Material online). Sox30 has been found to be differentially expressed between females and males in many bivalve species (Ghiselli et al. 2012; Zhang et al. 2014; Capt et al. 2019); our analysis confirms a possible central role in sex determination/differentiation in *R. philippinarum*. For genes in the female gonad-associated pink module, zona pellucida and homeobox domain were significantly enriched. One interesting candidate gene with the homeobox domain is PBX homeobox 4 (pbx4). In our analyses, pbx4 is a hub gene of the female gonad-associated pink module, and it is also differentially spliced between females and males (supplementary table S9, Supplementary Material online). The same gene in mammals has been found to be associated with gametogenesis (Wagner et al. 2001; Svingen and Koopman, 2007; Kawai et al. 2018). Also, pbx genes, which are characterized as hox gene co-activators, have been found to be associated with oogenesis, embryonic development, and germ cell maturation (Svingen and Koopman, 2007). Considering the female-specific transcription of pbx4 in *R. philippinarum*, further analyses will be required to understand the role of this gene in bivalves. Finally, we found that MYCBP-associated protein expressed in testis 1-like (maats1) is the hub gene in the male gonad-associated green module. This gene was previously shown to be differentially expressed during spermatogenesis (Yukitake et al. 2002) and suggested to be a candidate gene influencing the sex transformation process in the fish Monopterus albus (Chi et al. 2017). This indicates a possible role of maats1 in sex determination/differentiation in bivalves. Other hub genes such as spermatogenesis associated 17, testisspecific serine kinase 4, kelch-like family member 10 in three gonad-associated modules can be additional candidates involved in spermatogenesis, and therefore important in bivalve sex determination/differentiation system.

Conclusions

In this study, we present a long-read-based de novo genome assembly of a Manila clam from the North American Pacific Coast and an extensive RNA-Seq multi-tissue (gonad, mantle, and adductor) analysis of 15 females and 15 females, providing insights into the role of DE and splicing in bivalve tissue identity. Although DS was largely overlapping with differential gene expression, it was preferentially involved in gonad functions. Co-expression network revealed complex gene regulation in gonads. Moreover, our data showed heterogeneity in sequence evolution for male gonad-associated genes in *R. philippinarum*. Apart from a gene set that follows the common observation that male-biased genes present high sequence evolution and remain mostly uncharacterized, we detected one additional set of male gonad-associated genes showing an extremely low sequence evolution, but high pleiotropy, and with a putative central role in male reproduction in *R. philippinarum*. Together, these results increase our understanding of the role of DE, DS, and sequence evolution of sex-specific genes. We also provide resourceful genomic data for further studies regarding sex diagnosis and breeding.

Material and Methods

A detailed Materials and Methods section with all parameter sets can be found in Supplementary Materials, Methods and Results, Supplementary Material online. A brief overview is described below.

Sample Collection and Sequencing

Genomic DNA was extracted from a single male individual from the Puget Sound region (Pacific Northwest, USA) using only mantle tissue with the E.Z.N.A. Mollusc DNA Kit (Omega Bio-tek, Inc.). The PacBio library was prepared using a SMRTbell template preparation kit, and a 1050 Kb size selection was performed using a BluePippin System. Two types of Illumina libraries were prepared: a "small insert" library (insert size ∼500 bp), and a "long insert" library (insert size ∼1,500 bp). To avoid as much as possible biases in library construction, we prepared multiple replicates for each library: nine replicates for the small insert library, and ten replicates for the large insert library. Replicates were indexed and pooled, and each pool was sequenced in one separated lane of an Illumina HiSeq 2,500 with 2 × 250 bp reads at the USC Genome Core facility, University of Southern California. The long-read libraries were sequenced on a PacBio RSII using a P6-C4 chemistry at the Genomics High-Throughput Facility, University of California, Irvine. *Ruditapes philippinarum* specimens used for RNA-Seq were collected from the Northern Adriatic Sea, in the river Po delta region (Sacca di Goro, approximate GPS coordinates: 44°50′06″N, 12°17′55″E) during the spawning season (end of July). In total, 90 samples were obtained from three different tissues (adductor muscle, mantle, and gonad) of 15 males and 15 females. Total RNA was extracted with TRIzol, poly-A transcripts were isolated with magnetic beads and used as template for cDNA synthesis following the protocol as in Mortazavi et al. (2008) with modifications as in Ghiselli et al. (2012). RNAsequencing was performed on Illumina HiSeq 2,500 platform with insert size of approximately 500 bp to generate 150 bp paired-end reads.

Genome Assembly

Quality assessment and adaptor trimming of Illumina libraries were performed with Trimmomatic (Bolger et al. 2014) and FastQC. Genome size, heterozygosity, and duplication level were estimated using K-Mer Counter (Kokot et al. 2017), Genomescope 2 (Vurture et al. 2017) and kmercountexact.sh from the BBMap package (Bushnell, 2014) with different k-mer size. Contig-level genome assembly was performed using PacBio reads and wtdbg2 (Ruan and

Li, 2020). Contig correction and assembly heterozygosity reduction were performed running Hypo (Kundu et al. 2019) and purge_dups (Guan et al. 2020), respectively, for three consecutive times. Quality of the final version of the assembly was assessed with BUSCO (Seppey et al. 2019), redundans (Pryszcz and Gabaldón, 2016), and KAT (Mapleson et al. 2016). Possible contaminations in the assembly were identified and removed with Blobtools (Laetsch and Blaxter, 2017).

Manila Clam Genome Comparison

Our assembly was aligned to a previously published R. philippinarum genome assembly (short-reads only) by Yan et al. (2019; GCA_009026015.1), that we named CRph, using the mummer package (Marçais et al. 2018). The dnadiff function was used to identify and classify alignable regions between the two assemblies.

Genome Annotation

Transposable elements were annotated with RepeatModeler (Flynn et al. 2020) and MITE Tracker (Crescente et al. 2018). After removal of genes, tandem repeats and low copy number repeats, annotation of repeats was achieved running RepeatMasker (Tarailo-Graovac and Chen, 2009). Gene annotation was carried out using Maker (Cantarel et al. 2008). Three previously assembled transcriptomes of R. philippinarum, the Swiss-Prot database, and proteomes from *Crassostrea gigas* (GCF_902806645.1), *C. virginica* (GCF_002022765.2), *Lottia gigantea* (GCF_000327385.1), and *Octopus bimaculoides* (GCF_001194135.1) were used as external evidence. On these we trained SNAP (https://github.com/KorfLab/SNAP), Augustus (Stanke et al. 2008), genemark (Brů na et al. 2020), and Evidence Modeler (Haas et al. 2008). Predicted transcripts were annotated via Blastx (Altschul et al. 1990) against the full Swiss-Prot database, Pfam database and via InterProScan (Jones et al. 2014) with default options.

Gene Expression and Co-Expression Analysis

The PE reads were processed with Trimmomatic (Bolger et al. 2014) to remove adaptors and low quality reads. Then, clean reads were mapped to the genome assembly using STAR (Dobin and Gingeras, 2015) in multiple 2-pass modes. FeatureCounts (Liao et al. 2014) was used to count the number of reads in the genomic features. Samples with a low number of reads and genes with a low expression level were filtered out using NOISeq (Tarazona et al. 2015). DE analysis was performed based on the filtered data in DESeq2 using both Wald test and Likelihood ratio test (Love et al. 2014). Genes with adjusted P values <0.05 and |log2(FoldChange)| > 1 were considered as DEGs. Tissue specificity for each gene based on

Tau method was calculated using tspex (Camargo et al. 2020). Tissue specificity was estimated by Tau, an index for determining how specific or broad is gene expression. Tau ranges from 0 to 1, where 0 indicates broad expression across tissues and 1 indicates tissue-specific expression (Yanai et al. 2005). The co-expression network was constructed with Weighted Gene Co-expression Network Analysis (WGCNA) (Langfelder and Horvath, 2008). The network connectivity was retrieved from the co-expression network using the function intramodularConnectivity implemented in the WGCNA package. More in detail, for genes in the coexpression network, we measured the connectivity with genes in the same module (intramodular connectivity: kWithin), the connectivity with genes from different modules (intermodular connectivity: kOut) and its global connectivity (kTotal = kWithin + kOut). Therefore, kTotal, kWithin, and kOut in this tissue-specific co-expression network describe different properties: kTotal represents the total network connectivity and is the sum of kWithin and kOut; kWithin represents within module connectivity specific to one or multiple associated tissue types (specific connectivity); kOut represents the connection of one gene to the genes outside the module in the other tissue types (broad connectivity). Moreover, genes ranking in the top 5% of kWithin, representing high connection with the other genes in the module, were defined as the "hub" genes. The detailed parameters used in each program can be found in Supplementary Materials and Methods, Supplementary Material online.

Differential Splicing Analysis

To understand the general pattern of splicing across tissues, intron excision ratio was calculated using Leafcutter (Li et al. 2018). A PCA plot based on the intron excision ratio was produced to visualize the general splicing patterns across tissues. For the pairwise DS analysis between sexes, and between pairwise tissues, we used exon-based limma package v3.42 (Ritchie et al. 2015), which presented good performances in DS analyses with large sample sizes (Mehmood et al. 2020; Merino et al. 2019). Genes with adjusted P-value <0.05 were considered differentially spliced (DS). The bam files generated from STAR were used for genome-guided transcriptome assembly in Stringtie (Pertea et al. 2016). SUPPA (Trincado et al. 2018) was used to measure seven alternative splicing events: skipping exon (SE), alternative 5′ splicing (A5), alternative 3′ splicing (A3), retained intron (RI), alternative first exon (AF), and alternative last exon (AL).

Estimation of the Rate of Sequence Evolution

The protein coding sequences from the closely related species *Cyclina sinensis* (Family Veneridae) were retrieved from Wei et al (2020). Single-copy orthologs between C. sinensis and R. philippinarum were identified using OrthoFinder (Emms and Kelly, 2019). The orthologous protein sequences were aligned with Clustal Omega (Sievers and Higgins, 2018) and the nucleotide alignments were derived according to the protein alignments using PAL2NAL (Suyama et al. 2006). The protein evolutionary rate was estimated according to the ratio of non-synonymous to synonymous nucleotide changes (Ka/Ks), and it was calculated using KaKs_calculator2 (Wang et al. 2010).

SNP Analysis

The quality of the reads from the male/female sequencing runs was assessed using the FastQC, before being mapped to the R. philippinarum genome assembly using Rsubread (Liao et al. 2019). The resulting BAM files were used for variant calling with Freebayes (Garrison and Marth, 2012) to retain only biallelic SNPs present in at least 80% of samples using Bcftools. Next, genotypes (in 0/1 format) were extracted from the two VCF files using the Genome Analysis ToolKit (GATK) (DePristo et al. 2011) and genotype counts by population were used as input for the BayPass (Gautier 2015). SNPs that were identified by BayPass as significantly contrasted between the male and female groups were then functionally annotated using Annovar (Wang et al. 2010). The effect of SNPs was predicted with SnpEff (Cingolani et al. 2012) and the PCA plot based on the SNPs across all samples was performed with SNPRelate (Zheng et al. 2012).

Gene Set and Domain Enrichment

Gene Ontology (GO) analysis was performed for different sets of genes using topGO (Alexa 2021). The GO enrichment analysis was performed with Fisher’s exact test, and REVIGO (Supek et al. 2011) was used to reduce redundancy in the enriched GO terms. Domain enrichment analysis was performed with Fisher’s exact test in R using fisher.test function. The KEGG brite hierarchies for hub genes were performed in KAAS website (Moriya et al. 2007). Statistical Analysis Kruskal–Wallis test followed by Dunn test with FDR correction were used to assess the pairwise difference in kTOtal, kWithin, kOut, Tau, and Ka/Ks. Wilcoxon rank-sum test was used to assess if there was difference for kWithin between DEG and no-DEGs, and between DSGs and no-DSGs. Wilcoxon rank-sum test with Holm–Bonferroni correction was used to compare module-specific kTotal, kWithin, kOut, Tau, and Ka/Ks to the overall values across all the modules. The correlation between pairwise two indexes was performed with Spearman’s rank-sum test. All the tests and data visualization described above were performed in the Rstudio.

Supplementary material

Supplementary data are available at Genome Biology and Evolution online (http://www.gbe.oxfordjournals.org/).

Acknowledgements

We would like to thank the following Colleagues: Melanie Oakes and Valentina Ciobanu (Genomics High-Throughput Facility, University of California, Irvine) for PacBio library preparation and sequencing; Daniel Campo (USC Genome Core facility at the University of Southern California) for Illumina sequencing; Tevfik Hamdi Kitapci (Nuzhdin Lab, Molecular and Computational Biology, University of Southern California) for his suggestions about experimental design; Nadége Guiglielmoni for suggestions about genome assembly.

Author Contributions

Concept of the study: F.G., and M.P. Design of the study: F.G., S.V.N., and J.P.D. Acquisition of data: F.G., J.P.D., L.M., and M.I. Genome assembly: J.M. Genome annotation: L.P., and M.B. RNA-Seq: M.I., and R.X. SNP analysis: M.S. Analysis and interpretation of data: R.X., J.M., M.S., M.I., L.P., M.M., S.B., L.M., and F.G. Manuscript writing: R.X., J.M., M.S., M.I., and F.G. Manuscript critical revision: M.P., L.B., S.V.N., L.M., S.B., and M.M.

Funding

This work was supported by the Italian Ministry of Education University and Research (MIUR) FIR2013 Programme [grant number RBFR13T97A] funded to F.G., the Canziani bequest funded to F.G. and M.P., and the “Ricerca Fondamentale Orientata” (RFO) funding from the University of Bologna to F.G. and M.P. This work was made possible, in part, through access to the Genomic High-Throughput Facility Shared Resource of the Cancer Center Support Grant (CA-62203) at the University of California, Irvine and NIH shared instrumentation grants [grant numbers 1S10RR025496-01, 1S10OD010794-01].

Data Availability

PacBio and Illumina sequencing data as well as the genome assembly are deposited in the National Center for Biotechnology Information (NCBI: PRJNA807867). Genome annotation is available on https://doi.org/ 10.6084/m9.figshare.21069946.v5.

References

- Alexa A, Rahnenfuhrer J. 2021. TopGO: Enrichment Analysis for Gene Ontology. R package, version 2.41.0.
- Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. J. Mol. Biol. 215(3):403–410. doi: 10.1016/S0022-2836(05)80360-2.
- Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30(15):2114–2120. doi: 10.1093/bioinformatics/btu170.
- Breton S, et al. 2011. Novel Protein Genes in Animal mtDNA: A New Sex Determination System in Freshwater Mussels (Bivalvia: Unionoida)? Mol. Biol. Evol. 28(5):1645–1659. doi: 10.1093/molbev/msq345.
- Breton S, Capt C, Guerra D, Stewart D. 2018. Sex-Determining Mechanisms in Bivalves. In: Leonard, J. editor. Transitions Between Sexual Systems. Springer, Cham. p. 165–192. doi: 10.1007/978-3-319-94139-4_6.
- Broquard C, et al. 2021. Gonadal transcriptomes associated with sex phenotypes provide potential male and female candidate genes of sex determination or early differentiation in Crassostrea gigas, a sequential hermaphrodite mollusc. BMC Genomics. 22(1):609. doi: 10.1186/s12864-021-07838-1.
- Brůna T, Lomsadze A, Borodovsky M. 2020. GeneMark-EP+: Eukaryotic gene prediction with self-training in the space of genes and proteins. NAR Genom. Bioinform. 2(2): lqaa026. doi: 10.1093/nargab/lqaa026.
- Bushnell B. 2014. *BBMap: a fast, accurate, splice-aware aligner*. Lawrence Berkeley National Lab.(LBNL), Berkeley, CA (United States).
- Camargo AP, Vasconcelos AA, Fiamenghi MB, Pereira GAG, Carazzolle MF. 2020. tspex: a tissue-specificity calculator for gene expression data. Research Square (Preprint).
- Cantarel BL, et al. 2008. MAKER: An easy-to-use annotation pipeline designed for emerging model organism genomes. Genome Res. 18(1):188–196. doi: 10.1101/gr.6743907.
- Capt C, et al. 2018. Deciphering the Link between Doubly Uniparental Inheritance of mtDNA and Sex Determination in Bivalves: Clues from Comparative Transcriptomics. Genome Biol. Evol. 10(2):577–590. doi: 10.1093/gbe/evy019.
- Capt C, Renaut S, Stewart DT, Johnson NA, Breton S. 2019. Putative Mitochondrial Sex Determination in the Bivalvia: Insights From a Hybrid Transcriptome Assembly in Freshwater Mussels. Front. Genet. 10: 840. doi: 10.3389/fgene.2019.00840.

- Chi W, Gao Y, Hu Q, Guo W, Li D. 2017. Genome-wide analysis of brain and gonad transcripts reveals changes of key sex reversal-related genes expression and signaling pathways in three stages of *Monopterus albus.* PLoS One. 12:e0173974.
- Cingolani P, et al. 2012. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff SNPs in the genome of *Drosophila melanogaster* strain w 1118 ; iso-2; iso-3. Fly 6(2):80-92. doi: 10.4161/fly.19695.
- Crescente JM, Zavallo D, Helguera M, Vanzetti LS. 2018. MITE Tracker: an accurate approach to identify miniature inverted-repeat transposable elements in large genomes. BMC Bioinformatics. 19(1):348. doi: 10.1186/s12859-018-2376-y.
- Cordero D, Delgado M, Liu B, Ruesink J, Saavedra C. 2017. Population genetics of the Manila clam (*Ruditapes philippinarum*) introduced in North America and Europe. Sci. Rep. 7(1): 1–13.
- Dalpé A, et al. 2022. The influence of environmental conditions on sex determination in the blue mussel *Mytilus edulis*. ICES J. Mar. Sci. 79(2):394–402.
- Dean R, Mank JE. 2016. Tissue specificity and sex-specific regulatory variation permit the evolution of sex-biased gene expression. Am. Nat. 188(3):E74–E84. doi: 10.1086/687526.
- DePristo MA, et al. 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat. Genet. 43(5):491–498. doi: 10.1038/ng.806.
- Dobin A, Gingeras TR. 2015. Mapping RNA-seq Reads with STAR. Curr. Protoc. Bioinforma 51(1): 11-14. doi: 10.1002/0471250953.bi1114s51.
- Ellegren H, Parsch J. 2007. The evolution of sex-biased genes and sex-biased gene expression. Nat. Rev. Genet. 8(9):689–698. doi: 10.1038/nrg2167.
- Emms DM, Kelly S. 2019. OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biol. 20(1):238. doi: 10.1186/s13059-019-1832-y.
- Flynn JM, et al. 2020. RepeatModeler2 for automated genomic discovery of transposable element families. Proc. Natl. Acad. Sci. 117(17):9451–9457. doi: 10.1073/pnas.1921046117.
- Garrison E, Marth G. 2012. *Haplotype-based variant detection from short-read sequencing*. arXiv:1207.3907.
- Gautier M. 2015. Genome-Wide Scan for Adaptive Divergence and Association with Population-Specific Covariates. Genetics. 201(4):1555–1579. doi: 10.1534/genetics.115.181453.

- Ghiselli F, et al. 2012. De novo assembly of the Manila clam *Ruditapes philippinarum* transcriptome provides new insights into expression bias, mitochondrial doubly uniparental inheritance and sex determination. Mol. Biol. Evol. 29(2):771–786. doi: 10.1093/molbev/msr248.
- Ghiselli F, et al. 2013. Structure, transcription, and variability of metazoan mitochondrial genome: Perspectives from an unusual mitochondrial inheritance system. Genome Biol. Evol. 5(8):1535–1554. doi: 10.1093/gbe/evt112.
- Ghiselli F, et al. 2018. Comparative Transcriptomics in Two Bivalve Species Offers Different Perspectives on the Evolution of Sex-Biased Genes. Genome Biol. Evol. 10(6):1389–1402. doi: 10.6084/m9.figshare.5398618.v1.
- Ghiselli F, et al. 2019. Natural Heteroplasmy and Mitochondrial Inheritance in Bivalve Molluscs. Integr. Comp. Biol. 59(4): 1016-1032. doi: 10.1093/icb/icz061.
- Ghiselli F, et al. 2021a. Molluscan Mitochondrial Genomes Break the Rules. Philos. Trans. R. Soc. B Biol. Sci. 376(1825): 20200159.
- Ghiselli F, Iannello M, Piccinini G, Milani L. 2021b. Bivalve Molluscs as Model Systems for Studying Mitochondrial Biology. Integr. Comp. Biol. 61(5): 1699-1714. doi: 10.1093/icb/icab057.
- Grath S, Parsch J. 2012. Rate of amino acid substitution is influenced by the degree and conservation of male-biased transcription over 50 myr of *Drosophila* evolution. Genome Biol. Evol. 4(3):346–359.
- Grath S, Parsch J. 2016. Sex-Biased Gene Expression. Annu. Rev. Genet. 50:29–44. doi: 10.1146/annurev-genet-120215-035429.
- Griffin RM, Dean R, Grace JL, Rydén P, Friberg U. 2013. The Shared Genome Is a Pervasive Constraint on the Evolution of Sex-Biased Gene Expression. Mol. Biol. Evol. 30(9):2168–2176. doi: 10.1093/molbev/mst121.
- Grimes T, Potter SS, Datta S. 2019. Integrating gene regulatory pathways into differential network analysis of gene expression data. Sci. Rep. 9(1):1–12.
- Guan D, et al. 2020. Identifying and removing haplotypic duplication in primary genome assemblies. Bioinformatics. 36(9):2896–2898. doi: 10.1093/bioinformatics/btaa025.
- Haas BJ, et al. 2008. Automated eukaryotic gene structure annotation using Evidence Modeler and the Program to Assemble Spliced Alignments. Genome Biol. 9(1): 1-22. doi: 10.1186/gb-2008-9-1-r7.

- Harris RMB, et al. 2020. Biological responses to extreme weather events are detectable but difficult to formally attribute to anthropogenic climate change. Sci. Rep. 10:14067. doi: 10.1038/s41598-020-70901-6.
- Harrison PW, et al. 2015. Sexual selection drives evolution and rapid turnover of male gene expression. Proc. Natl. Acad. Sci. U. S. A. 112(14):4393–4398. doi: 10.1073/pnas.1501339112.
- Herrmann JM, Neupert W. 2013. Protein Import into Mitochondria. In: Encyclopedia of Biological Chemistry. Elsevier, Academic Press. p. 632–636. doi: 10.1016/B978-0-12-378630-2.00203-6.
- Ingleby FC, Flis I, Morrow EH. 2015. Sex-biased gene expression and sexual conflict throughout development. Cold Spring Harb. Perspect. Biol. 7(1):a017632.
- Jones P, et al. 2014. InterProScan 5: genome-scale protein function classification. Bioinformatics. 30(9):1236–1240. doi: 10.1093/bioinformatics/btu031.
- Kawai Y, Oda A, Kanai Y, Goitsuka R. 2018. Germ cell-intrinsic requirement for the homeodomain transcription factor PKnox1/Prep1 in adult spermatogenesis. PLoS One. 13:e0190702. doi: 10.1371/journal.pone.0190702.
- Kee TR, et al. 2021. Mitochondrial CHCHD2: Disease-Associated Mutations, Physiological Functions, and Current Animal Models. Front. Aging Neurosci. 13: 660843. doi: 10.3389/fnagi.2021.660843.
- Kokot M, Długosz M, Deorowicz S. 2017. KMC 3: counting and manipulating k-mer statistics. Bioinformatics. 33(17):2759–2761. doi: 10.1093/bioinformatics/btx304.
- Krishnakumar PK, Qurban MA, Sasikumar G. 2018. Biomonitoring of Trace Metals in the Coastal Waters Using Bivalve Molluscs. In Trace Elements: Human Health and Environment. InTech. doi: 10.5772/intechopen.76938.
- Kundu R, Joshua C, Sung W-K. 2019. HyPo: Super Fast & Accurate Polisher for Long Read Genome Assemblies. Biorxiv.
- Laetsch DR, Blaxter ML. 2017. BlobTools: Interrogation of genome assemblies. F1000Research. 6:1287. doi: 10.12688/f1000research.12232.1.
- Langfelder P, Horvath S. 2008. WGCNA: An R package for weighted correlation network analysis. BMC Bioinformatics. 9(1):1-13. doi: 10.1186/1471-2105-9-559.
- Li Y, et al. 2018. Annotation-free quantification of RNA splicing using LeafCutter. Nat. Genet. 50(1):151–158. doi: 10.1038/s41588-017-0004-9.

- Liao Y, Smyth GK, Shi W. 2014. FeatureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics. 30(7):923–930. doi: 10.1093/bioinformatics/btt656.
- Liao Y, Smyth GK, Shi W. 2019. The R package Rsubread is easier, faster, cheaper and better for alignment and quantification of RNA sequencing reads. Nucleic Acids Res. 47(8):e47–e47. doi: 10.1093/nar/gkz114.
- Lipinska A, et al. 2015. Sexual Dimorphism and the Evolution of Sex-Biased Gene Expression in the Brown Alga Ectocarpus. Mol. Biol. Evol. 32(6):1581–1597. doi: 10.1093/molbev/msv049.
- Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 15(12):550. doi: 10.1186/s13059-014-0550-8.
- González-Tizón, A.M., Martínez-Lage, A., Rego, I., Ausió, J., Méndez, J., 2000. DNA content, karyotypes, and chromosomal location of 18S-5.8S-28S ribosomal loci in some species of bivalve molluscs from the Pacific Canadian coast. Genome 43, 1065–1072. https://doi.org/10.1139/g00-089
- Maeda GP, Iannello M, McConie HJ, Ghiselli F, Havird JC. 2021. Relaxed selection on male mitochondrial genes in DUI bivalves eases the need for mitonuclear coevolution. J. Evol. Biol. 34(11):1722–1736. doi: 10.1111/jeb.13931.
- Mank JE, Hultin-Rosenberg L, Axelsson E, Ellegren H. 2007. Rapid Evolution of Female-Biased, but Not Male-Biased, Genes Expressed in the Avian Brain. Mol. Biol. Evol. 24(12):2698–2706. doi: 10.1093/molbev/msm208.
- Mank JE, Hultin-Rosenberg L, Zwahlen M, Ellegren H. 2008. Pleiotropic constraint hampers the resolution of sexual antagonism in vertebrate gene expression. Am. Nat. 171(1):35–43. doi: 10.1086/523954.
- Mapleson D, Garcia Accinelli G, Kettleborough G, Wright J, Clavijo BJ. 2016. KAT: a K-mer analysis toolkit to quality control NGS datasets and genome assemblies. Bioinformatics. 33(4): 574-576. doi: 10.1093/bioinformatics/btw663.
- Marçais G, et al. 2018. MUMmer4: A fast and versatile genome alignment system. PLOS Comput. Biol. 14(1):e1005944. doi: 10.1371/journal.pcbi.1005944.
- Mehmood A, et al. 2020. Systematic evaluation of differential splicing tools for RNA-seq studies. Brief. Bioinform. 21(6):2052–2065. doi: 10.1093/bib/bbz126.

- Meisel RP. 2011. Towards a More Nuanced Understanding of the Relationship between Sex-Biased Gene Expression and Rates of Protein-Coding Sequence Evolution. Mol. Biol. Evol. 28(6):1893–1900. doi: 10.1093/molbev/msr010.
- Merino GA, Conesa A, Ferná Ndez EA. 2019. A benchmarking of workflows for detecting differential splicing and differential expression at isoform level in human RNA-seq studies. Brief. Bioinform. 20(2):471–481. doi: 10.1093/bib/bbx122.
- Milani L, Ghiselli F, Nuzhdin SV, Passamonti M. 2013. Nuclear genes with sex bias in *Ruditapes philippinarum* (Bivalvia, veneridae): Mitochondrial inheritance and sex determination in DUI species. J. Exp. Zool. Part B Mol. Dev. Evol. 320(7):442–454. doi: 10.1002/jez.b.22520.
- Moriya Y, Itoh M, Okuda S, Yoshizawa AC, Kanehisa M. 2007. KAAS: an automatic genome annotation and pathway reconstruction server. Nucleic Acids Res. 35(suppl_2), W182-W185.
- Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat. Methods. 5(7):621–628. doi: 10.1038/nmeth.1226.
- Parsch J, Ellegren H. 2013. The evolutionary causes and consequences of sex-biased gene expression. Nat. Rev. Genet. 14(2):83–87. doi: 10.1038/nrg3376.
- Pertea M, Kim D, Pertea GM, Leek JT, Salzberg SL. 2016. Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. Nat. Protoc. 11(9):1650–1667. doi: 10.1038/nprot.2016.095.
- Pryszcz LP, Gabaldón T. 2016. Redundans: an assembly pipeline for highly heterozygous genomes. Nucleic Acids Res. 44(12):e113–e113. doi: 10.1093/nar/gkw294.
- Ritchie ME, et al. 2015. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 43(7):e47–e47. doi: 10.1093/nar/gkv007.
- Rogers TF, Palmer DH, Wright AE. 2021. Sex-Specific Selection Drives the Evolution of Alternative Splicing in Birds. Mol. Biol. Evol. 38(2):519–530. doi: 10.1093/molbev/msaa242.
- Ruan J, Li H. 2020. Fast and accurate long-read assembly with wtdbg2. Nat. Methods. 17(2):155–158. doi: 10.1038/s41592-019-0669-3.
- Seppey M, Manni M, Zdobnov EM. 2019. BUSCO: Assessing Genome Assembly and Annotation Completeness. Methods Mol Biol. 1962:227-245. doi: 10.1007/978-1-4939-9173-0_14.

- Sievers F, Higgins DG. 2018. Clustal Omega for making accurate alignments of many protein sequences. Protein Sci. 27:135–145. doi: 10.1002/pro.3290.
- Smits M, et al. 2020. A proteomic study of resistance to Brown Ring disease in the Manila clam, *Ruditapes philippinarum*. Fish Shellfish Immunol. 99:641–653. doi: 10.1016/j.fsi.2020.02.002.
- Stanke M, Diekhans M, Baertsch R, Haussler D. 2008. Using native and syntenically mapped cDNA alignments to improve de novo gene finding. Bioinformatics. 24(5):637–644. doi: 10.1093/bioinformatics/btn013.
- Supek F, Bošnjak M, Škunca N, Šmuc T. 2011. REVIGO Summarizes and Visualizes Long Lists of Gene Ontology Terms. PLoS One. 6(7):e21800. doi: 10.1371/journal.pone.0021800.
- Suyama M, Torrents D, Bork P. 2006. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. Nucleic Acids Res. 34:W609–W612. doi: 10.1093/nar/gkl315.
- Svingen T, Koopman P. 2007. Involvement of Homeobox Genes in Mammalian Sexual Development. Sex. Dev. 1(1):12–23. doi: 10.1159/000096235.
- Tarailo-Graovac M, Chen N. 2009. Using RepeatMasker to Identify Repetitive Elements in Genomic Sequences. Curr. Protoc. Bioinforma. 25. doi: 10.1002/0471250953.bi0410s25.
- Tarazona S, et al. 2015. Data quality aware analysis of differential expression in RNA-seq with NOISeq R/Bioc package. Nucleic Acids Res. 43(21):e140. doi: 10.1093/nar/gkv711.
- Telonis-Scott M, Kopp A, Wayne ML, Nuzhdin S V, McIntyre LM. 2009. Sex-Specific Splicing in Drosophila: Widespread Occurrence, Tissue Specificity and Evolutionary Conservation. Genetics. 181(2):421–434. doi: 10.1534/genetics.108.096743.
- Trincado JL, et al. 2018. SUPPA2: fast, accurate, and uncertainty-aware differential splicing analysis across multiple conditions. Genome Biol. 19(1): 1-11. doi: 10.1186/s13059-018-1417-1.
- Vurture GW, et al. 2017. GenomeScope: fast reference-free genome profiling from short reads. Bioinformatics. 33(14):2202–2204. doi: 10.1093/bioinformatics/btx153.
- Wagner K, Mincheva A, Korn B, Lichter P, Pöpperl H. 2001. Pbx4, a new Pbx family member on mouse chromosome 8, is expressed during spermatogenesis. Mech. Dev. 103(1-2):127–131. doi: 10.1016/S0925-4773(01)00349-5.
- Wang Dapeng, Zhang Y, Zhang Z, Zhu J, Yu J. 2010. KaKs_Calculator 2.0: A Toolkit Incorporating Gamma-Series Methods and Sliding Window Strategies. Genomics. Proteomics Bioinformatics. 8(1):77–80. doi: 10.1016/S1672-0229(10)60008-3.

- Wang K., Li M, Hakonarson H. 2010. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. 38(16):e164–e164. doi: 10.1093/nar/gkq603.
- Wei M, et al. 2020. Chromosome-Level Clam Genome Helps Elucidate the Molecular Basis of Adaptation to a Buried Lifestyle. iScience. 23(6):101148. doi: 10.1016/j.isci.2020.101148.
- Whittle CA, Extavour CG. 2019. Selection shapes turnover and magnitude of sex-biased expression in Drosophila gonads. BMC Evol. Biol. 19(1):1–20. doi: 10.1186/s12862-019-1377-4.
- Wijsman JWM, Troost K, Fang J, Roncarati A. 2019. Global production of marine bivalves. Trends and challenges. Goods Serv. Mar. bivalves. 7–26.
- Xu R, Iannello M, Havird JC, Milani L, Ghiselli F. 2022. Lack of transcriptional coordination between mitochondrial and nuclear oxidative phosphorylation genes in the presence of two divergent mitochondrial genomes. Zool. Res. 43(1):111.
- Yan W. 2009. Male infertility caused by spermiogenic defects: lessons from gene knockouts. Mol. Cell. Endocrinol. 306(1-2):24–32.
- Yan X, et al. 2019. Clam Genome Sequence Clarifies the Molecular Basis of Its Benthic Adaptation and Extraordinary Shell Color Diversity. iScience. 19:1225–1237. doi: 10.1016/j.isci.2019.08.049.
- Yanai I, et al. 2005. Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. Bioinformatics. 21(5):650–659. doi: 10.1093/bioinformatics/bti042.
- Yue C, Li Q, Yu H. 2018. Gonad Transcriptome Analysis of the Pacific Oyster Crassostrea gigas Identifies Potential Genes Regulating the Sex Determination and Differentiation Process. Mar. Biotechnol. 20(2):206–219. doi: 10.1007/s10126-018-9798-4.
- Yukitake H, Furusawa M, Taira T, Iguchi-Ariga SMM, Ariga H. 2002. AAT-1, a novel testis-specific AMY-1-binding protein, forms a quaternary complex with AMY-1, A-kinase anchor protein 84, and a regulatory subunit of cAMP-dependent protein kinase and is phosphorylated by its kinase. J. Biol. Chem. 277(47):45480–45492.
- Zhang N, Xu F, Guo X. 2014. Genomic Analysis of the Pacific Oyster (*Crassostrea gigas*) Reveals Possible Conservation of Vertebrate Sex Determination in a Mollusc. G3 Genes|Genomes|Genetics. 4(11):2207–2217. doi: 10.1534/g3.114.013904.

- Zheng X, et al. 2012. A high-performance computing toolset for relatedness and principal component analysis of SNP data. Bioinformatics. 28(24):3326–3328. doi: 10.1093/bioinformatics/bts606.
- Zouros E. 2020. Doubly uniparental inheritance of mitochondrial DNA: Might it be simpler than we thought? J. Zool. Syst. Evol. 58(2): 624-631. doi: 10.1111/jzs.12364.

4. Chapter II

Multiple and diversified transposon lineages contribute to early and recent bivalve genome evolution

Jacopo Martelossi, Filippo Nicolini, Simone Subacchi, Daniela Pasquale, Fabrizio Ghiselli, Andrea Luchetti

Note:

This chapter has been published as a research article on BMC Biology under the DOI: 10.1186/s12915-023-01632-z. All supplementary materials can be found on its online version.

Abstract

Background

Transposable elements (TEs) can represent one of the major sources of genomic variation across eukaryotes, providing novel raw material for species diversification and innovation. While considerable effort has been made to study their evolutionary dynamics across multiple animal clades, Molluscs represent a substantially understudied phylum. Here we take advantage of recent increases in their genomic resources and adopt an automated TE annotation pipeline combined with a tree based classification, as well as extensive manual curation efforts, to characterize TE repertories across 27 bivalve genomes with a particular emphasis on DDE/D Class II elements, Long Interspersed Nuclear Elements (LINEs), and their evolutionary dynamics.

Results

We found Class I elements as highly dominant in bivalve genomes with LINE elements, despite being less represented, being the most common retroposon group covering up to 10% of their genome. We mined 86,488 reverse transcriptases (RVT) containing LINE coming from 12 clades distributed across all known superfamilies and 14,275 Class II DDE/D-containing transposons coming from 16 distinct superfamilies. We uncovered a previously underestimated rich and diverse bivalve ancestral transposon complement that could be traced back to their most recent common ancestor that lived ~500 Mya. Moreover, we identified multiple instances of lineage-specific emergence and loss of different LINEs and DDE/D lineages with the interesting cases of CR1- Zenon, Proto2, RTE-X and Academ elements that underwent a bivalve-specific amplification likely associated with their diversification. Finally, we found that this LINE diversity is maintained in extant species by an equally diverse set of long-living and potentially active elements, as suggested by their evolutionary history and transcription profiles in both male and female gonads.

Conclusions

We found that bivalves host an exceptional diversity of transpons compared to other molluscs. Their LINE complement could mainly follow a 'stealth drivers' model of evolution where multiple and diversified families are able to survive and co-exist for a long period of time in the host genome, potentially shaping both recent and early phases of bivalve genome evolution and diversification. Overall, we provide not only the first comparative study of TE evolutionary dynamics in a large but largely understudied phylum such as Mollusca, but also a reference library for ORF-containing Class II DDE/D and LINE elements, which represents an important genomic resource for their identification and characterization in novel genomes.

Keyword: Mollusca, Evolutionary Genomics, TE Classification, Bivalvia, Comparative Genomics, SINE, TE evolution, TE annotation

Introduction

Transposable elements (TEs) are selfish genetic elements that replicate independently from the replication of the host genome [1, 2]. They are widespread and ubiquitous across all branches of the eukaryotic tree of life and, although showing a remarkable sequence diversity across organisms, the conservation of common catalytic domains responsible for their replication suggests that their emergence could be traced back to the eukaryotic most recent common ancestor or even predate it [3].

TE classification is not straightforward, although many efforts have been undertaken to try to reconcile their diversity in a systematic framework. Two main classes are generally recognized: class I, which includes all TEs replicating via RNA intermediates, and class II, that embodies TEs moving via DNA intermediates [4]. This latest distinction still represents the only unambiguous classification of TEs. Conversely, the within-class diversity is much more complicated to analyze, since it can be performed both with mechanistic and homology-based criteria [5]. For example, considering the way TEs replicate and reintegrate, all class I elements use a "copy-and-paste" mechanism, while class II exhibit several models: the classical "cut-and-paste", or the "peel-and-paste" (also known as rolling-circle replication) or even the "self-synthesizing" model (reviewed in [5]). The current classification scheme, which is also implemented in the main TE database, Repbase [6], is based on homology and structural similarities [7]. Class I elements mainly include Long Terminal Repeat (LTR) elements and Long Interspersed Nuclear Elements (LINEs, also indicated as non-LTR elements) which encode for a reverse transcriptase (RT), an endonuclease (EN) and other domains used to reintegrate in the host genome. Class II elements, on the other hand, include Terminal Inverted Repeat (TIR) elements, Helitrons, and Mavericks (also known as *Polintons*). In addition, both classes include non-autonomous elements (Short Interspersed Nuclear Elements, SINEs, and Miniature Inverted-repeats Transposable Elements, MITEs), TEs usually with a smaller size, which do not code for the enzymes necessary for replication/reintegration but parasitize those encoded by their autonomous counterparts [7]. Beside this commonly accepted scheme, further classification efforts are less clear. Generally speaking, when taking into consideration coding TEs, the clustering pattern after a phylogenetic analysis of their ORF(s) is taken as an indication of clades that should be considered possible families, groups of elements or clades [5].

Although as a common approach, the phylogenetic framework has limitations in this context both because of the sometimes unclear homology of TE ORFs and the genomic turnover of paralogous TE lineages blurring the phylogenetic signal [8].

The same replicative dynamics of TEs may impact on their phylogenetic clustering: in fact, based on studies on mutation distribution on non-autonomous class I Alu sequences in the human genome, two distinct models have been formulated to explain how TEs replicate [9]. The first model, named "master gene model", implies that one or few copies give origin to all other copies in the genome producing new, so-called families each time a master copy mutates. This way, new families are generated in different timeframes. On the contrary, in the other model, termed "transposons model", each new copy can produce other copies with the outcome of getting several families produced nearly at the same time.

The rate at which TEs replicate can be a function of several different factors, including the ability of the host genome to limit their uncontrolled proliferation. In particular, the successful invasion of a genome by TEs can be dependent on a complex interplay among TE features, host genome biology, repression mechanisms interfering with TE functionality, and the extent of selective pressures on the outcome of TE insertions [10]. Despite this, some TE lineages managed to reach very high copy numbers in the host genomes, apparently escaping such controlling mechanisms. A suitable model to explain these dynamics has been formulated on the well-studied human SINE family Alu and on their autonomous counterparts L1 LINEs. These elements show several subfamilies that evolved following a master gene model in different hominid lineages during the last few million years. However, their origin seems to predate their species-specific expansions by far, with little or no transposition for tens of million years. [11] hypothesized that the species-specific rise to high copy number of some subfamilies could be due to some "stealth drivers", i.e. Alu and L1 copies with a very low activity which allowed them to survive, undetected, in different host lineages and that suddenly underwent a massive replication wave in specific conditions, in given hosts.

Despite being extensively analyzed among vertebrates and arthropod genomes, TEs are surprisingly understudied in the phylum Mollusca, a large and diverse group of metazoans with many ecologically and economically important species. To date, TE studies in molluscs are limited to the characterization of one or a few elements [12-22], or to the whole mobilome, i.e. the full complement of TEs in the genome, but in a few species [23-25]. A direct consequence of the lack of genome-scale analyses of TE content in mollusc genomes is that public repositories and databases only harbor scarce information about them, making de novo assembled genome annotations less reliable [26, 27]. Therefore, beside the importance of analyzing the TE content and their relationships with host genomes in molluscs, it is also crucial for future genomic studies to get more detailed and wider TE libraries available.

In the present work, we leveraged the mollusc genome resources available currently in public databases, with a particular focus on bivalves, and carried out an extensive study of the full mobilome. In-depth analysis of Class II DDE/D-related transposons and LINEs allow us to deeply characterize an ancestral TE complement and its following expansion and contractions coupled with bivalve evolutionary history. Moreover, we manually curated a representative set of LINEs and DDE/D families that correspond to potentially recently active elements. The curated LINE library was finally used to reconstruct LINE evolutionary histories and assess their potential activity in male and female gonads of five species distributed across four different bivalve orders. The DDE/D and LINE manually curated library produced in this work could represent an important future resource for the bivalve genomic community to improve TE annotation in novel genomes.

Results

Overall TE content across molluscs using automatically generated TE sequence libraries

To analyze mollusc mobilome we compiled a dataset of 39 molluscan genomes representative of their major groups (Additional File 1: Tab. S1). Among these, 27 belong to bivalve species and represent eight different orders: Unionida, Adepedonta, Myida, Venerida, Arcida, Pectinida, Ostreida, and Mytilida. As a first step, we implemented an automatic TE annotation pipeline (See Material and Methods 5.2; Additional File 2: Fig. S1) which identified a variable number of consensus sequences, ranging from 92 elements in the annelid *Dinophilus gyrociliatus* to the 3,736 elements in the Mytilida *Modiolus philippinarum* (Additional File 3: Tab. S2). When annotating each genome with the corresponding species-specific library, as expected from an understudied phylum such as molluscs, "Unknown" elements represent a considerable proportion of the annotated repeats (mean=10.41%; Fig. 1A; Additional File 4: Tab. S3), especially in poorly studied taxa such as *Solen grandis* (16.12%) and *Mytilus coruscus* (20.07%). Segmental duplications and recently duplicated gene families could be one of the major sources of unclassified TE consensus; however, we tried to reduce their impact by removing gene and gene fragments from the repeat library, and by requiring at least 5 positive blast hits (at least with 70% of identity and query coverage) of the consensus sequence against the source genome. Unknown consensus sequences are mainly composed of short elements (median=433 bp, Additional File 5: Fig. S2A) with medium-low copy number (median=354 copies; Additional File 5: Fig. S2B). Though, it must be noted that for the well-analyzed species *Crassostrea gigas*, the percentage of unclassified elements drops down to 3.51% despite applying the same annotation pipeline (Additional File 4: Tab. S3). Overall, these results suggest that most of the unknown elements likely correspond to short, fragmented, or ancient families difficult to classify based on homology evidence alone.

The TE content also varied among and within different mollusc classes (Fig. 1A). The two TE-richest genomes were those of the pteriomorphian bivalve *M. philippinarum* (58.6%) and of the cephalopod *Octopus sinensis* (57.39%). Among bivalves, the mean TE content observed was 38.97%, with analyzed Pectinida showing a generally lower TE proportion with respect to all the other species (Fig. 1A). A significant positive correlation was observed between assembly size and TE content (Fig. 1B; Spearman's rho=0.72, p<0.01).

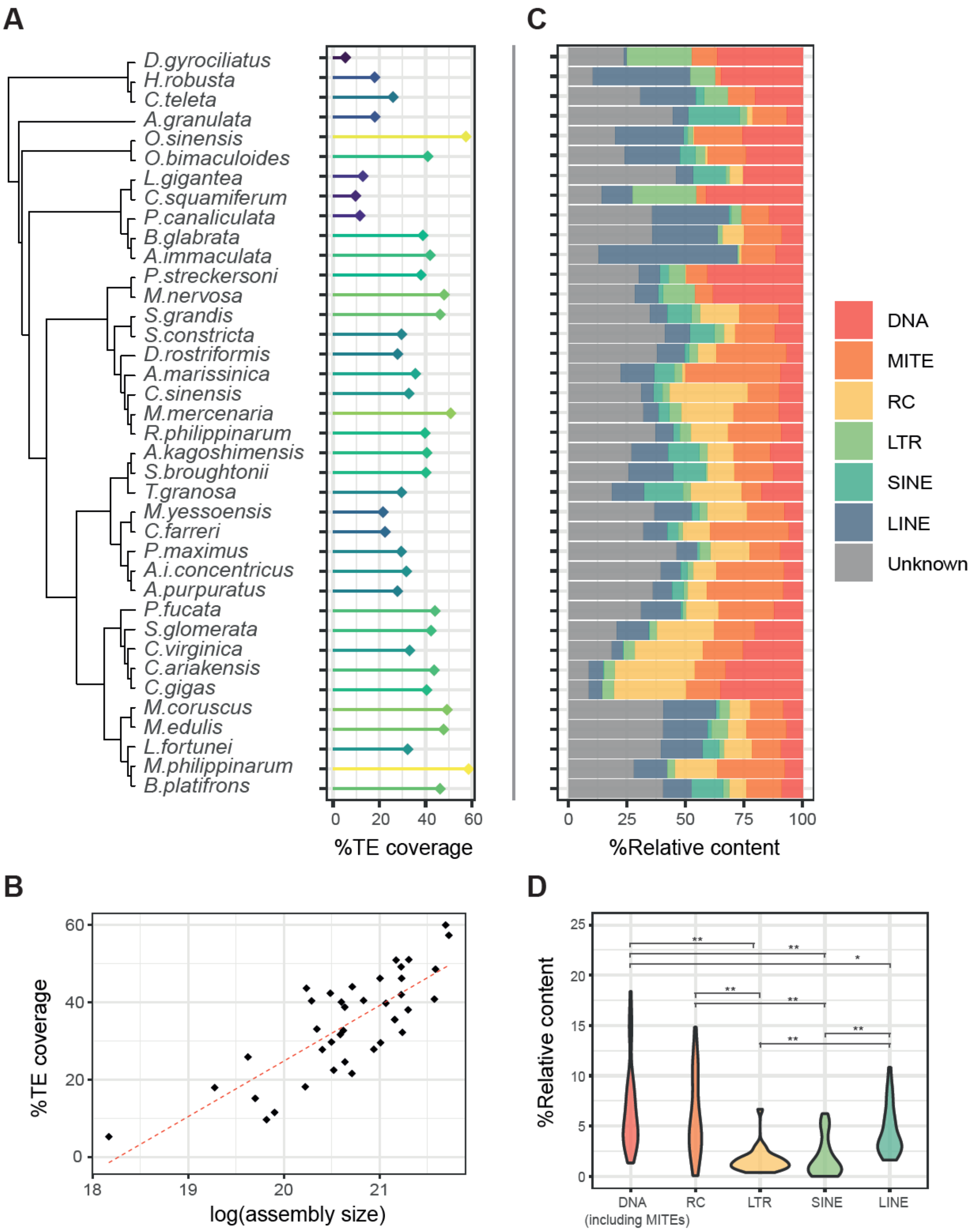


Figure 1: Transposable element content across molluscs. TE annotation results from automatically generated TE sequence libraries (see the "Genomic resources and phylogeny construction" section). (**A)** Phylogeny of the 39 analyzed genomes as retrieved from the literature and their overall transposable elements (TEs) content. (**B**) Correlation between TE coverage and assembly size as a proxy of genome size. (**C**) Relative contribution of different TE classes to the total TE content across molluscs. (**D**) Genome occupancy of each TE class in the 27 analyzed bivalves. Significant comparisons are highlighted by asterisks (pairwise Wilcoxon rank test with Bonferroni correction; *p < 0.05, **p < 0.01). A specular box plot considering all analyzed species, including other molluscan classes and annelids, is presented in Additional file 6: Fig. S3

Class-level mollusc mobilome characterization using automatically generated TE sequence libraries

When analyzing the contribution of different TE classes in the overall transposon composition across all analyzed species (Fig. 1C), after excluding unknown elements, LTR and SINE resulted significantly under-represented compared to all other groups (Kruskal-Wallis rank test; Pairwise Wilcoxon rank test with Bonferroni correction, p<0.05), but no other significant differences were identified (Additional File 6: Fig. S3). The same pattern emerged when analyzing only bivalves, but they also showed a significant overrepresentation of DNA elements, including MITEs, over LINEs (Kruskal-Wallis rank test, p< 0.05; Pairwise Wilcoxon rank test with Bonferroni correction, p<0.05; Fig. 1D).

LINEs are ubiquitous elements and constitute the most common retroposon group (mean=5.38%) but they were observed with a highly variable frequency, ranging from the 1.15% in the polyplacophora *Acanthopleura granulata* to 24.78% in the gastropod *Achatina immaculata* genome, where they dominate the TE landscape. In bivalves they represent from 1.30% of the host genome in the oyster *Crassostrea virginica* up to 10.84% in *M. coruscus*. SINEs are present across all analyzed species but always in low copy number (mean=1.69%) with a few, lineage-specific amplifications, such as in *Archivesica marissinica* (3.05%), Adepedonta order (*Sinonovacula constricta* and *S. grandis*, respectively 3.1% and 4.7%) the Arcida order (*Anadara kagoshimensis*, *Scapharca broughtonii*, and *Tegillarca granosa;* mean=5.3%), in the Mytilidae *Bathymodiolus platifrons* (6.23%) and in the Polyplacophora *A. granulata* (4%). Also LTR elements were generally found in low copy number in the analyzed species (mean=1.52%), with the exception of the Unionidae species *Megalonaias nervosa* in which LTRs account for the 6.66% of the host genome. We also observed a relatively high Rolling Circle (RC) element content (mean=5.91%) associated with bivalve diversification, reaching an average of 12% among *Crassostrea* species and 9.69% in *Cyclina sinensis*. Notable exceptions to this trend are the two Unionida *Potamilius streckersoni* and *M. nervosa* in which RC elements are greatly reduced in the former (0.05%) and absent in the latter.

General characterization of mollusc repeatome composition using automatically generated TE sequence libraries

When clustering analyzed mollusc genomes based on the number of annotated insertions for each RepeatMasker transposon type (See Material and Methods section 5.2), we found that bivalves are clearly divergent from other molluscs, both when using a hierarchical (Fig. 2) and a k-mean clustering approach with 3 centers (Additional File 7: Fig. S4). However, when looking at the relationships between and within bivalve orders, a more complex scenario emerged, with lineages belonging to different orders intermingling with each other. The only exception to this pattern were the Ostreidae, whose clustering resulted in complete agreement with their known phylogenetic relationships [28].

Concerning LINEs, the elements L2, L1-Tx1, CR1, and I are the most ubiquitous types across molluscs with representatives in respectively 36, 35, 35, and 32 species, even though their genomic occurrence can vary to a great extent. The RTE-BovB type was found greatly expanded in cephalopods and in the gastropods *A. immaculata* and *Biomphalaria glabrata* compared to other species. On the contrary, RTE-X and CR1-Zenon elements were more represented in bivalve genomes but greatly reduced or even absent in cephalopods and gastropods. Finally, R2-Hero, R4-Dong, and CRE types are identified almost exclusively in cephalopods, with only R2-Hero found in low copy number in the *A. immaculata* genome and in some bivalve species but with a patchy distribution.

Multiple SINE lineages were found, belonging to V, Meta, Core, and MIRs types with V elements that can reach up to 4.3% in the Mytilidae *B. platifrons* genome.

Regarding LTRs, Bel/Pao, DIRS, and Ngaro types were mainly found in bivalves and in the gastropods *Lottia gigantea* and *Pomacea canaliculata*, although in low copy number, and they appeared almost absent in cephalopods. On the other hand, Gypsy and Copia elements are ubiquitous across all molluscs, with the former present in higher copy numbers.

For DNA elements, different types belonging to superfamilies Mutator-like elements (MULE), Mariner, PiggyBac, CMC, Mavericks, and hAT are present across all analyzed genomes. Kolobok, Zator, and Academ superfamily types are almost exclusively found in bivalves, while Zisrupton, Novosib, and Merlin superfamilies were found almost completely restricted to the analyzed cephalopods.

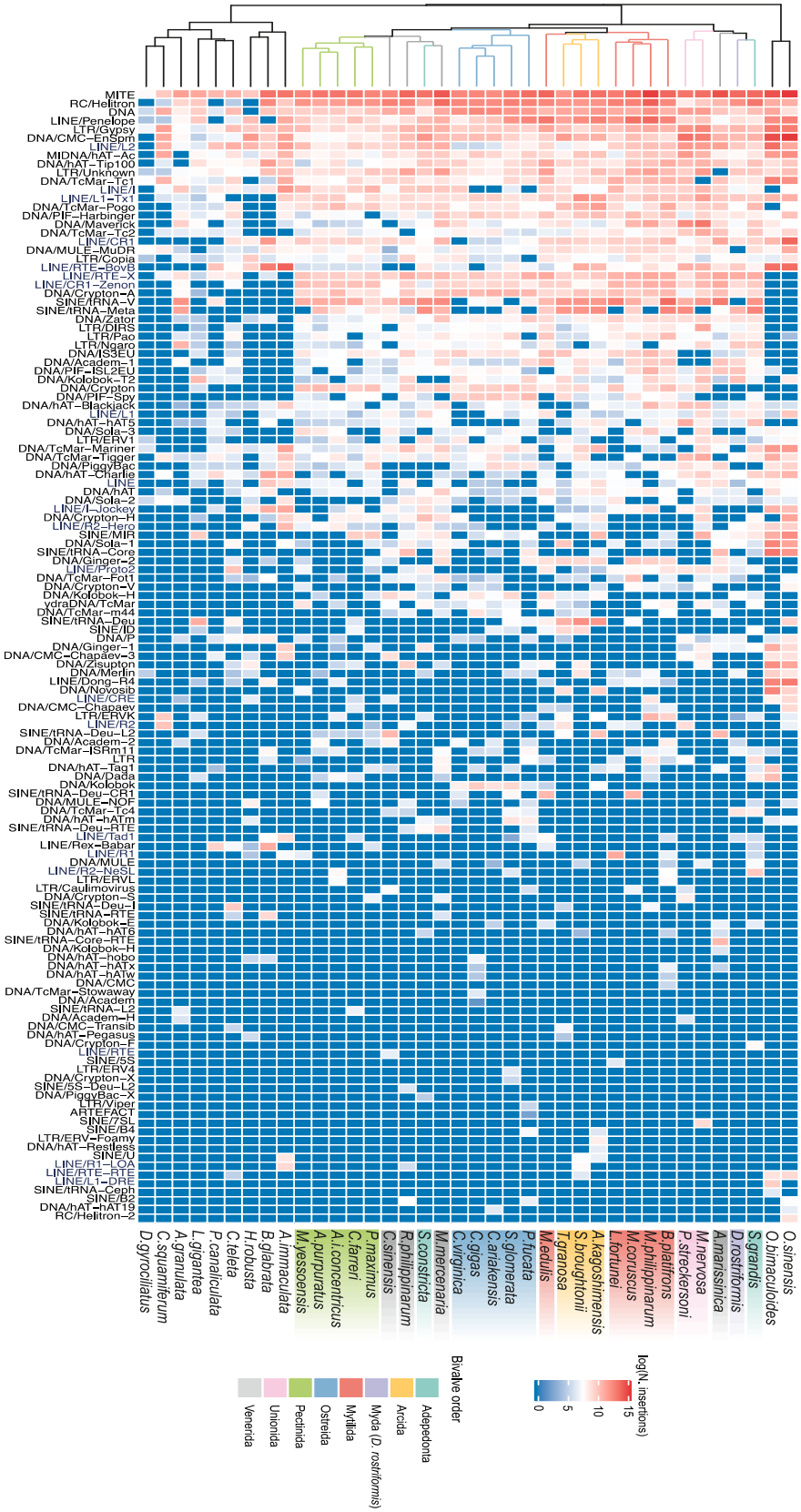


Fig. 2 Hierarchical cluster analysis on the number of insertions for each transposon type. The insertion counts were obtained after defragmentation of the TE annotation with RepeatCraft on the RepeatMasker output obtained with the automatically generated TE sequence libraries (see the "Mining and annotation of interspersed repeats" section)

Extraction and clustering of RT-containing LINEs

We decided to deeply and more confidently characterize the LINE complement implementing an ORF-based extraction and classification approach (See Material and Methods section 5.3 and 5.4). Overall, we identified a total of 86,488 LINE loci exhibiting an RT domain in an ORF longer than 300 amino acids (Additional File 8: Tab. S4). These were clustered in 13,523 groups following the 80-80 rule, and only 3,601 of them were found composed of more than 5 elements, accounting for a total of 69,763 loci (80.7%). A great variation can be observed among species in terms of both diversity and richness of clusters. Among bivalves, *A. marissinica* genome resulted as the richest one in terms of RT-containing LINEs (6,935 elements).

Overall, 8,333 LINE loci (9.6%) were annotated as putative autonomous elements, here defined as insertions showing both RT and EN domains on the same ORF, longer than 300 amino acids and without interrupting stop codons (Additional File 8: Tab. S4). As expected, we found the number of LINEs with a RT domain being positively correlated with the number of identified putative autonomous elements (Spearman's rho=0.89, $p < 0.01$; Additional File 9: Fig. S5a). The assembly contiguity, here measured as the scaffold N50 value, was also found significantly correlated to the number of identified RT-containing LINEs (Spearman's rho=0.35, $p < 0.05$, Additional File 9: Fig. S5b) as well as to the number of identified putative autonomous elements (Spearman's rho=0.34, $p < 0.05$; Additional File 9: Fig. S5c).

Phylogenetic analyses and classification of RT-containing LINEs

To classify the previously mined LINEs containing RTs in superfamilies and clades, we used a phylogenetic approach starting from amino acid consensus sequences built up from clusters with more than 4 members (See Material and Methods section 5.4). After removal of poorly aligned sequences by TrimAl, 3,252 LINE clusters were included in the phylogenetic analysis. We further added 259 reference sequences for classification purposes and annotated 111 other LINEs using RTClass1 (Additional File 10: Tab. S5). To obtain a reliable phylogeny of LINE elements useful for their annotation, we used both NJ and ML tree searches with and without topological constraints (Fig. 3A; Additional File 11: Fig. S6).

When testing all topologies in a ML framework, we obtained the highest likelihood for one of the SupFAM tree (i.e, constraining the monophyly of all superfamilies as recovered in the NJ tree; SupFAM #2; Fig. 3A; Additional File 12: Tab. S6).

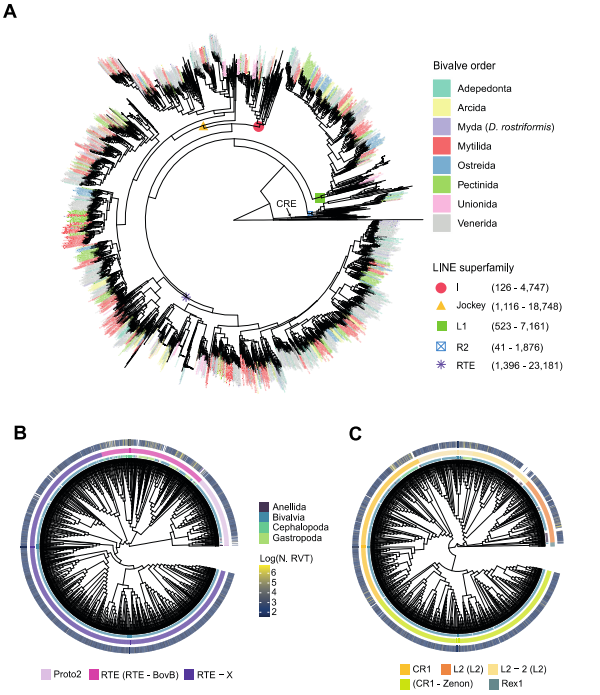


Fig. 3 Phylogeny of mollusc LINEs. Phylogenetic analyses performed on extracted RT-containing LINEs. A Maximum likelihood SupFAM tree #2 obtained by constraining the monophyly of different LINE superfamilies as recovered by the Neighbor-Joining topology. Numbers in parentheses next to the LINE superfamilies represent the number of annotated clusters and the total number of elements represented by the included clusters, respectively. All tested trees with relative bootstrap values can be found in Additional file 29: Data S1. More detailed versions of the SupFAM tree #2 subtrees can be found in Additional files 13, 14, 15, 16, and 17: Figs. S7, S8, S9, S10, and S11. B RTE and C Jockey superfamily subtrees. The inner circle represents the taxonomic annotation of mollusc classes, and the mid one is the annotation of the different clades based on reference sequences extracted from RepBase and based on [29]. Note that the L2-2 clade includes Crack, Daphne, L2A, and L2B elements. Names in parenthesis refer to the RepeatMasker type classification. The outer circle shows the log scale number of elements grouped in each cluster. Reference sequences are represented by white spaces in the inner and outer circles.

Moreover, the obtained best tree also recovered more monophyletic clades compared to all other topologies (Additional File 13-14-15-16-17: Fig S7-S8-S9-S10-S11) and resulted the most in agreement with both references [29, 30] and RepeatMasker/Dfam classification schemes. For these reasons the SupFAM #2 tree was used for LINE classification as well as for all downstream analyses.

Based on reference sequences, we managed to confidently classify all elements at the superfamily and clade level, except for 16 elements from *O. sinensis* genome, that were placed in a subclade of the I superfamily in sister relationship with I-Loa-R1 and Tad1 reference sequences (Unknown I clade; Additional File 16: Fig. S10). Moreover, as already shown (see [29]), L2A and L2B clades resulted to be paraphyletic, with polyphyletic Crack and Daphne elements clustering within them. For these reasons, henceforth we will refer to these clades as L2-2 elements, while other elements will be simply indicated as L2.

Interestingly, Proto2, RTE-X, and CR1-Zenon elements were only found in bivalves, with Proto2 also present in the annelida *Capitella teleta* (Fig. 3B-C). The complete tree-based annotation of all LINEs can be found in Additional File 27: Tab. S7. Generally speaking, the tree resulted in a complex branching pattern with multiple order-specific clades in each identified LINE clade/type, also highlighting multiple instances of expansion, contraction, and loss of LINE lineages in different bivalve orders (Fig. 3A; Additional File 13-14-15-16-17: S7-S8-S9-S10-S11). Blastp against the full RepeatPeps library and RTClass classification widely confirm our tree-based annotation with only few discordances, which mainly concerned Proto2 elements classified as RTE-X.

Richness, diversity, and distribution of RT-containing LINEs

We used Blastp against previously tree-based classified LINEs to annotate all clusters excluded from phylogenetic analyses (i.e "low-copy number", "singletons" and clusters removed by TrimAl; See Material and Methods section 5.3; Fig. 4A; Additional File 18: Fig. S12). The RTE, Jockey, and L1 superfamilies were confirmed as the richest (*i.e.*, with more elements; Fig. 4A) and most diverse (*i.e.*, with more clusters; Additional File 18: Fig. S12) across molluscs. The only R2 elements found in bivalves were classified as Hero (363 elements). Nimb and Ingi clades are the only representatives of the I superfamily across molluscs, beside the Unknown clade coming from *O. sinensis*.

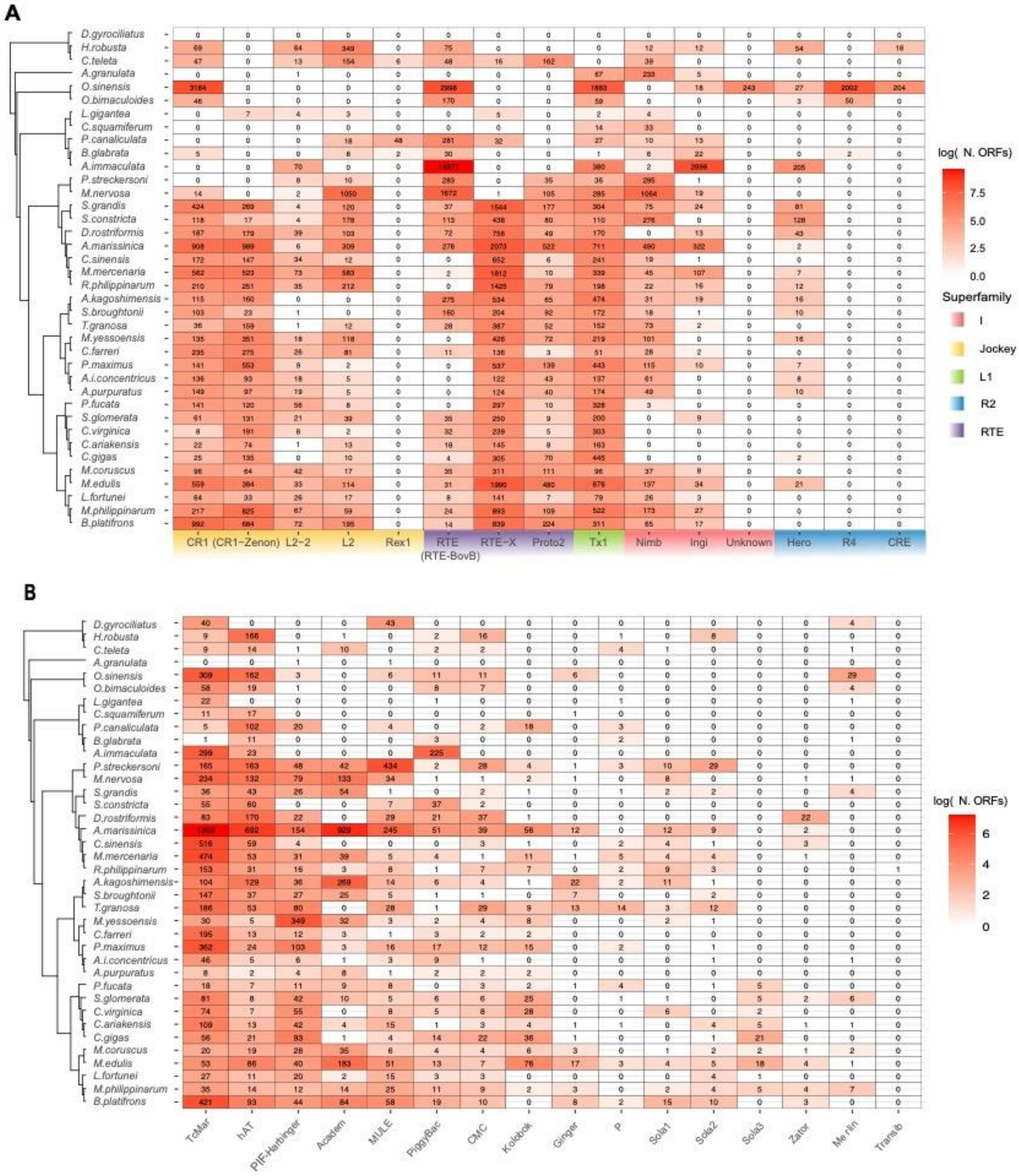


Fig. 4 Richness of mollusc class II DDE/D-related transposons and LINEs. (**A**) Number of RT-containing LINEs annotated in each analyzed genome and subdivided by clade following [29] or, when in parenthesis, by the RepeatMasker "type" classification. "Unknown" refers to elements annotated based on an O. bimaculoides clade found nested in the I superfamily but missing any reference sequence (see Additional file 16: Fig. S10). Note that the L2-2 clade includes Crack, Daphne, L2A, and L2B elements. (**B**) Number of ORF-containing DDE/D-related transposons annotated in each analyzed genome and subdivided by superfamily following [31].

The Rex1 clade was only found at a low copy number in the gastropods *P. canaliculata*, *B. glabrata* and the annelida *C. teleta*, while elements belonging to CR1-Zenon, RTE-X, and Proto2 lineages despite still being more highly represented in bivalves were also recovered at low copy numbers and/or with singleton elements in few, non-bivalve species. Bivalve genomes exhibit a high diversity of LINE lineages, hosting members from 11 out of the 14 identified clades (CR1, CR1-Zenon, L2, L2-2, RTE-X, RTE-BovB, Proto2, Tx1, Nimb, Ingi, Hero). As a comparison, the gastropod *B. glabrata*, the cephalopod *O. sinensis* and the ring worms *Helobdella robusta* and *C. teleta* showed eight different LINE clades. In *A. marissinica* all clades, with the exception of Hero and L2-2, were expanded compared to other Venerida and Imparedentia. For Arcida, Pectinida, and Ostreida we identified multiple instances of order-specific loss/contraction, such as the extreme reduction of the I superfamily in Ostreida (maximum of 9 members of the Ingi clade identified in *C. virginica*), of the RTE clade (RTE-BovB type) in Pectinida (11 members in *Chlamys farreri*) and the L2-2 clade in Arcida (only 1 element in *T. granosa*). The Unionida *M. nervosa* and *P. streckersoni* show notable differences in their LINE complement compared to all other bivalves, with a great reduction of the RTE-X and CR1/CR1-Zenon clades/type, which were found well-represented in other genomes, and an expansion of L2 and RTE-BovB elements in *M. nervosa*. The number of annotated RT-containing LINEs and the number of clusters were found significantly correlated for all superfamilies (Additional File 19: Fig. S13). Finally, the number of annotated autonomous elements is in line with previous results, but no member of the R2 superfamily was identified (Additional File 20: Fig. S14).

Distribution of Class II DDE/D-related transposons

To classify ORFs derived from DDE/D-related transposons we implemented an HMM-based approach starting from classified sequences from the 17 superfamilies described by [31] (See Materials and Methods section 5.3). Overall, we identify DDE/D class II related transposons, with an ORF longer than 300 amino acids and no interrupting stop codons, coming from 16 out of the 17 superfamilies, for a total of 14,275 elements. Their distribution approximately recapitulates what we observed with automatically generated libraries (Fig. 4A). Specifically, the TcMar resulted in the richest superfamily in 21 species, accounting for the 41% of the overall number of identified elements, followed by hAT, Academ, MULE, and PIF-Harbinger. Instead, Ginger, Sola1, Sola2, Sola3, Zator, Merlin, and Transib are less represented, with respectively 98, 95, 105, 64, 44, 63, and one element identified. Overall, Bivalves possess at least one element across all superfamilies resulting in the most diverse mollusc group here analyzed in terms of number of hosted DDE/D-related superfamilies, with Academ, Sola, and Zator elements that appear restricted to this clade. Interestingly, we found that *A. marissinca* genome hosts the highest number of DDE/D-related elements from the five superfamilies hAT, TcMar, PIF-Harbinger, Academ and CMC compared to all other bivalves, similarly to what we observed for LINE elements.

Construction of a manually curated library for LINE, SINEs, and DDE/D-related transposons

We used our annotated Class I LINEs and SINEs, and Class II DDE/D-related ORFs in a "Blast-Extend-Extract" approach to build a comprehensive and manually curated TE library of potentially or recently active elements for bivalves (See Material and Methods section 5.7). Totally, we curated 840 LINEs, 119 SINEs, and 1018 DDE/D transposons for a total of 1,917 elements. These libraries were reduced respectively to 810, 37, and 762 families after CD-HIT clustering.

For the LINE library all consensus sequences possess a RT domain, while we manage to reconstruct a RT + EN segment for 740 (91%) of them. Therefore, although we did not systematically search for full length elements due to frequent 5' truncations, most of these families may correspond to potentially active or recently active elements for which exist copies across the genome with recognizable RT and EN domains. It must also be noted that only clusters that exhibit at least one copy with an RT and EN domain on an ORF longer than 300 amino acids were selected for manual curation (See Material and Methods section 5.7). The length of the resulting consensus sequences ranges from 1,786 bp to 9,087 bp with a mean of 5,023 bp. As expected, different LINE superfamilies show different length distributions (Additional File 21: Fig. S15) with members of I and L1 superfamilies being generally longer (mean=6,122 bp and 5,851 bp, respectively), followed by Proto2 (mean=5,675 bp), CR1-Zenon (mean=5,204 bp), RTE-X (mean=4,937 bp), L2 (mean=3,991 bp), CR1 (mean=3,791 bp) and RTE-BovB (mean=3,583 bp). These values largely recapitulate the canonical length of full-length elements described in literature, as for RTE-BovB (3.2 kbp) and L1 (6-8 kbp) [32, 33], proving a successful implementation of the "Blast-Extend-Extract" approach.

The length of SINE elements varies between 174 bp and 404 bp (mean=307 bp). Nine of them were classified as V elements, eight as Meta, eight as MIR, four as Deu and on e as Core, while the other seven elements lacked a family-level classification.

For DDE/D-related elements, after checking for TIRs, flanking TSDs and the presence of an ORF longer than 300 amino acids with a significant hit against DDE/D-related HMM profiles, all curated consensus sequences correspond to autonomous full-length consensus elements. Specifically, our library includes: 332 TcMar, 133 hAT, 100 Academ, 58 PIF-Harbinger, 43 Kolobok, 39 MULE, 27 PiggyBack, 14 Sola2, eight Sola1, three Zator, three Merlin, and two CMC transposons. Also in this case, their length greatly varies between different superfamilies (Additional File 22: Fig. S16) and results concordant with known estimations [34] with Sola1 (mean=5,445 bp) and Academ (mean=5,565 bp) being generally the longest elements, and Merlin (mean=1,635 bp) and TcMar (mean=1,935 bp) being generally the shortest one.

Evolutionary and expression analyses of curated LINE and SINE families

We used the previously curated LINE library to analyze the evolutionary dynamics of potentially active or recently active LINE families (See Material and Methods section 5.9). First, the number of curated families and the number of putative autonomous elements were positively correlated to each other (Spearman's rho=0.88, p<0.01), suggesting their representativeness of the overall LINE complement. Phylogenetic analyses of curated families reflect what we observed in the full LINE tree, with elements found in the same host genome characterized by long branches and intermingling with those found in other species, even belonging to different bivalve orders (Additional File 23-24-25-26, Fig S17-S18-S19-S20). After masking the genome with RepeatMasker and both LINE and SINE curated libraries, the genomic occurrence of curated families ranges from <2% in the pectinida *C. farreri*, in the oysters *Pinctada fucata, Saccostrea glomerata,* and *C. virginica*, and in the arcid *T. granosa,* to >4% in the Unionida species *A. marissinica* and *P. streckersoni* (Fig. 5A). It must be noted that these estimations are only based on potentially active or recently active families that were selected for manual curation and therefore should not be considered as estimations of the overall LINE complement.

Repeat landscape showed similar activity profiles for CR1/Jockey, L1, and RTE superfamilies across the majority of analyzed bivalves, with one or two bursts of activity localized at low (1%-5%) but also at high (30%-50%) divergence from the consensus. However, coherently with the distribution of LINE clades, some bivalves lack the recent peak of activity (Fig. 5B). In other instances, recent lineage-specific expansion of different LINE clades/types can be observed, such as for RTE-BovB in Unionida and CR1, CR1-Zenon and RTE-X in *A. marissinica*. Moreover, using high confidence 3'-anchored insertions (*i.e.,* insertions aligning within the first 50 bp of the 3' end of the consensus and longer than 100 bp) we found a variable number of ancient LINE families that showed both recent (at least 30 copies with less than 5% divergence) and ancient (at least 5 copies with more than 30% divergence) activity (Fig. 5B).

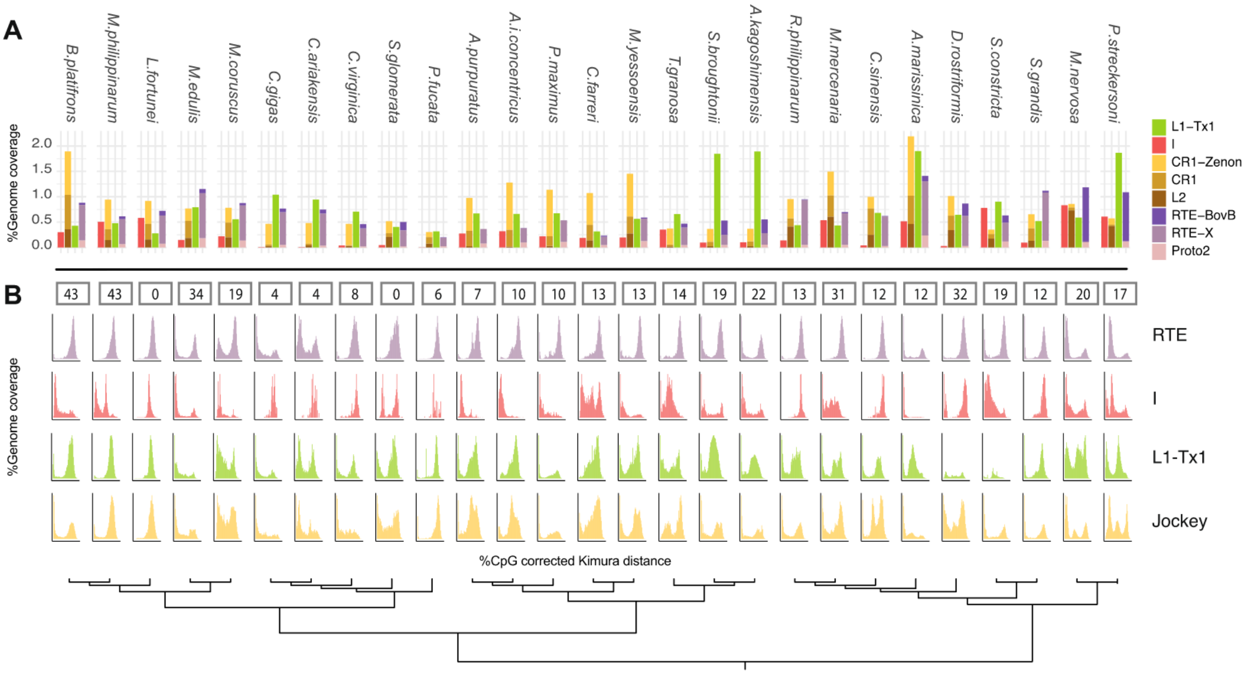


Fig. 5: Genome occurrence and evolutionary history of manually curated LINE families. RepeatMasker results obtained using our manually curated set of LINEs. (**A**) Genome coverage of curated families for each LINE clade/type. (**B**) CpG-corrected Kimura distance of each insertion from its consensus sequence as a proxy for the time of the transposition event for each LINE superfamily. The X-axes range from 0 to 50 while the Y-axes are on different scales for each specie/superfamily and represent the relative genome coverage. Numbers above the graphs represent the number of families for each species that possess insertions both in recent time (divergence < 5) and in the past (divergence > 30) requiring at least 30 annotated insertions in the recent divergence bin and 5 in the old one. Only 3′-anchored insertions longer than 100 bp were considered for this latest purpose.

All oysters as well as *L. fortunei*, *P. fucata* and *Argopecten purpuratus* possess between zero (*L. fortunei*, *S. glomerata*) and eight (*P. fucata*) ancient families while in all other bivalves their number can range between 10 in *Pecten maximus* up to 43 in the Mytilida *B. platifrons* and *M. philippinarum*. We assess the transposition potential of curated families by mapping gonads-derived RNAseq reads on 3’-anchored insertions longer than 3 kbp and extracted from five bivalve genomes. Specifically we found 96, 383, 1054, 346 and 801 insertions useful to map RNAseq reads in respectively *C. farreri*, *C. gigas*, *Mercenaria mercenaria*, *Mizuhopecten yessoensis* and *S. constricta*. The obtained transcription levels (estimated as TPM; Transcript per Million) *per* family were then tested for a correlation with the number of 3’-anchored insertions longer than 100 bp (to allow the presence of 5’ truncated copies) of the corresponding family. Across all analyzed species, tissues, and biological replicates we found a significant, positive correlation between the number of insertions and the per-family transcription level

(Spearman's rho=0.48 to 0.70, all ps<0.01; Tab. 1), a pattern consistent with an ongoing transposition of these elements [35].

Tab. 1: Spearman's rho correlation coefficients between family-based LINE transcript levels and number of insertions. Transcript levels were calculated as log2-transformed transcripts per million (TPM). Only insertions longer than 3Kb were used for mapping RNAseq reads (See Material and Methods section 5.9). Each tissue and biological replicate was separately tested for each species. MG = Male gonads; FG = Female gonads; all ps < 0.01.

Specie	**FG_1**	**FG_2**	**FG_3**	**MG_1**	**MG_2**	**MG_3**
C. farreri	0.54	0.60	0.60	0.50	0.48	0.52
C. gigas	0.46	0.50	0.50	0.40	0.48	0.51
M. mercenaria	0.64	0.61	0.60	0.57	0.60	0.60
M. yessoensis	0.70	0.68	0.67	0.63	0.60	0.62
S. constricta	0.56	0.53	0.59	0.66	0.61	0.56

We also added SINE families in the same RepeatMasker run and we obtained their reliable genome occurrence in the 13 species selected for in-depth SINE mining (See Materials and Methods section 2.6 and 2.8). SINEs genome occurrence can greatly vary between and within species belonging to different bivalve orders (Tab. 2). The genomes of *A. marissinica* (6.02%), *T. granosa* (3.69%), *S. broughtonii* (4.37%) and *B. platifrons* (4.68%) host a relatively high number of SINEs while on the contrary, we observed a great reduction in the genome of *C. gigas* (0.08%) and *S. glomerata* (0.31%). Different SINEs types successfully colonize different bivalve genomes: the Deu family was found to be dominant in *A. marissinca* (72% of the overall SINE complement), *C. sinensis* (94%) and *S. broughtonii* (55%), while the V family is dominant in the *B. platifrons* genome (67%) and the Meta in *S. constricta* (54%) and *S. grandis* (50%). Finally, in *T. granosa* both Deu and V families occupy a considerable proportion of the overall SINE complement, of respectively 30% and 46%. Finally, we did not find any evidence of significant correlation between SINEs and LINEs genomic occurrence (Spearman's rho=0.31, p=0.33).

Tab. 2: Percentage of genome occurrence of different SINE types in the 13 selected bivalves.

Specie	Meta	Core	Deu	V	Unknown	MIR	**TOT**
A. marissinica	0.22	> 0.01	4.1	1.4	0.3	> 0.01	**6.02**
C. sinensis	0.03	0.04	2.7	0.01	0.08	/	**2.86**
C. gigas	> 0.01	/	/	0.04	/	0.04	**0.08**
S. glomerata	0.3	/	> 0.01	> 0.01	/	> 0.01	**0.31**
T. granosa	0.8	> 0.01	1.13	1.7	0.05	> 0.01	**3.69**
S. broughtonii	0.1	> 0.01	2.42	1.4	0.41	0.04	**4.37**
M. coruscus	0.16	0.01	0.5	0.3	> 0.01	> 0.01	**0.97**
B. platifrons	1.48	> 0.01	> 0.01	3.19	/	/	**4.68**
S. constricta	1.34	> 0.01	> 0.01	1.08	0.01	0.04	**2.47**
S. grandis	1.34	> 0.01	> 0.01	0.94	0.35	0.04	**2.67**
M. yessoensis	0.2	/	> 0.01	0.35	> 0.01	0.26	**0.81**
P. maximus	0.23	/	> 0.01	0.50	/	0.04	**0.77**
M. nervosa	0.99	/	> 0.01	0.75	0.11	0.11	**1.96**

Discussion

A comprehensive TE annotation for bivalves

The phylum Mollusca shows a high level of organism diversity and includes species that are important for both their ecological and economic value. Although genomic studies are accumulating and comparative analyses are becoming more common for these organisms, a deep analysis of the mobilome is still limited to single genomes or to a few comparative studies with only a handful of species [24, 25]. As it could be expected, this also resulted in a scarce representativity of molluscan TEs in the public databases which makes their automated annotation less reliable. As previously shown, high-quality, manually curated repeat libraries are considered necessary for a consistent, reliable repeat annotation and characterization in novel genomes [26, 27]. In the present analysis, we decided to focus our efforts on bivalves, which represent 27 out of the 39 analyzed genomes, due to the recent, increasing genome sequencing efforts for this class. The inclusion of five gastropod genomes, representative of their major lineages, together with two cephalopods, one polyplacophoran genome, and three annelids allowed us to identify the major shifts in TEs composition that occurred during molluscan evolutionary history. To overcome limitations of automatically generated TE sequence libraries, we set up a pipeline which included both automated, ORF-based extraction and classification and manual curation approaches and that has been used consistently across the analyzed genomes. In particular, the manual curation process allowed us to provide a first freely available and manually curated repeat library for bivalves, comprising DDE/D, LINEs, and a subset of SINE elements for a total of 1,609 elements comprising all identified LINEs, with the exception of the low copy number R2 superfamily, and 12 different DDE/D-related superfamilies. These new genomic resources could help future genome annotation projects and shed novel insight on TEs evolutionary dynamics in bivalves. On the other hand, the ORF-based approach allows us to confidentiality characterize both LINEs and DDE/D-related TE complements. As a comparison, concerning LINEs in the RepBase library v. 20181026, 1,031 sequences are deposited for molluscs, with 796 of them belonging to well-characterized *C. gigas*. Fifty-nine of these are annotated as LINEs and, more specifically: one R2, two CR1, 12 CR1-Zenon, 14 L1-Tx1, 27 RTE-X. In the present analysis, we also found multiple Proto2, RTE-BovB, and L2/L2-2 elements. Regarding DDE/D transposons, out of 422 total sequences coming from RepBase for *C. gigas*, 92 possess an ORF longer than 300 amino acids and they belong to 13 different superfamilies. With our approach we manage to identify ORF-derived signatures coming from all of them, with the expectation of Zator, Merlin, and Sola1 for which only two sequences are deposited for each superfamily in RepBase. Overall, these results suggest that our ORF-based approach successfully captures in a flexible way most of the diversity of coding TEs in non-model species.

We also paid particular attention to filter out possible mis-annotations from the automatically generated TE sequence libraries, such as the inclusion of repetitive genes, tandem repeats, degenerate, and low-copy number families, which are hard to correctly annotate and classify. This approach is probably quite conservative, indeed in some instances it provided different estimates of the overall TE content compared to published genome papers. For example, in *Mytilus edulis* our study estimated 47% of TE content vs the 56% provided in [36]; the same holds for *S. glomerata* (42% in the present study vs 45% in [37]) and for *A. granulata* (18% here vs 23% in [38]). In other instances, though, our analysis provided almost the same estimates as in the previous analyses, as in *M. coruscus* (49% here vs 47% in [39]), *A. immaculata* (41% here vs 40% in [40]), and *M. mercenaria* (51% here vs 49% in [41]).

TEs have been shown to be one of the major contributors to genome size evolution in metazoan lineages, such as insects [42] and vertebrates [43], and in angiosperms as well [44]. Our analyses provided further support for this hypothesis finding a positive correlation between TE content and assembly size also in molluscs. Across bivalves, the TE content varies greatly, ranging from ~20% in the Pectinida *M. yessoensis* up to ~60% in the Mytilida *M. philippinarum.* Different sequencing technologies and sequencing depths could potentially contribute to such differences, however it must be noted that also for Illumina sequenced genome we observed a high TE content, such as for the M. *philippinarum* and *B. platifrons*. It is interesting the low TE load found across all analyzed Pectinida species. In fact, this order includes the most TE-poor bivalve species, with almost twofold less TE content compared to Mytilida and Ostreida. Similar occurrences of interspersed repeats were already observed for this lineage during whole genome sequencing projects [45-48], and transposable elements hosted by *M. yessoensis* were found to be generally less active in recent times compared to what was observed in the Pacific and pearl oysters [46]. This low TE activity was suggested to be the reason behind their conserved genome architecture that could resemble that of bilaterian ancestors [46]. However, as well-described in birds, low TE content and apparent lack of activity could also originate from nonallelic homologous recombination which could physically remove TEs and other repetitive regions from the genome without implying a general genomic stability [49].

Concerning Class I elements, LTR elements in general occupy a low proportion of host genomes as previously observed by [24], while we found LINE elements as the richest retroelements. They contribute from 1.61% to 10.84% respectively in *C. virginica* and *M. coruscus* genomes using automatically generated TE sequence libraries and between 6.18% for *A. marissinica* and 0.82% for *P. fucata* using manually curated libraries. A similar scenario occurs also for SINE elements, whose genome coverage can greatly vary between different bivalve species using both automatic and manually curated libraries. In both instances we identify the genomes of *A. marissinica, B. platifrons* and Arcida as richer in SINEs compared to other analyzed bivalves, but we did not find any evidence of a general increase of the SINE complement coupled with an increase of their autonomous counterparts LINEs.

Class II and RC elements generally outnumber other TEs, especially in bivalves where DNA elements were found significantly enriched compared to all retroposons. This is strikingly different from what observed in mammals, where retroposons constitute the most successful TE group, but similar to what is observed in actinopterygian fishes where Class II elements greatly dominate the overall TE content [43]. Moreover, we found that non-autonomous counterparts (MITEs) occupy a considerable proportion of host genomes suggesting the high proliferation of small, non-autonomous copies. Within the most rich superfamilies of DDE/D ORF-derived signatures in bivalve genomes we identified TcMariner and hAT lineages. Interestingly, the same superfamilies were also found to be the richest of ORF signatures in all other analyzed molluscs and to be ubiquitous even when using the automatically generated TE sequence libraries. Both TcMar and hAT superfamilies were found anciently expanded across cephalopods in a recent study from [25], possibly suggesting their high representativeness as a plesiomorphic state of molluscs. On the other hand we could identify notable examples of bivalve-specific expansion, such as for Academ and RC elements. The former seems to be poorly represented in non-bivalve genomes, with only few ORF identified in the ring worms *C. teleta* and *H. robusta* and few insertions annotated in non-bivalve molluscs when using automatically generated libraries. RC elements can occupy up to 12% in the analyzed *Crassostrea* species. As a comparison, RC have a more patchy distribution in arthropod genomes, generally contributing to a smaller extent of the genome size with only few, lineage-restricted expansions (*e.g. Drosophila* and *Musca domestica* [42, 50]). Also in plants, where they were firstly discovered, they are usually less represented, covering a maximum of 6% of the maize genome [51].

A highly diverse TE repertory characterizes bivalve molluscs

Both hierarchical and k-means clustering using automatically generated libraries clearly separated bivalves from other molluscs highlighting important differences in their TE complement. Although the scenario among these taxa appeared more complex, with the only case of full intra-order agreement between clustering analyses and species phylogeny in Ostreida, the analyses of both LINEs and DDE/D elements provided some notable examples of lineage-specific element differentiation.

Similarly to what has been observed in *Drosophila* [52], fishes [43, 53, 54], and other non-mammalian vertebrates [33, 55], we found that bivalves are characterized by a highly diversified DDE/D and LINEs complement. For the former, we identified ORF-related signatures coming from 17 different superfamilies while for LINEs we found 11 clades coming from all known superfamilies. Notable cases are the emergence of RTE-X, Proto2, and CR1-Zenon elements which, similarly to DDE/D Academ, appear almost limited to bivalve molluscs. Moreover, the presence of multiple, order-specific clusters across the LINE phylogeny, especially within Jockey and RTE superfamilies, suggest that these elements were already greatly diversified before the fast radiation of bivalves that occur in the early Ordovician, around 499 Mya [56, 57]. It is worth noting the underrepresentation of R2 elements across all molluscs (with exception of *O. sinensis*), a pattern strikingly different from what has been observed in other major lineages like arthropods, which are among the most successful LINEs [42, 58]. The Hero clade seems to be the only R2 element present in bivalve ancestors and the only identifiable in extant species, even though we could not identify any autonomous element.

Horizontal transposon transfer (HTT) can be a major source for the emergence of lineage-specific TE repertories, especially for aquatic species [23, 59]. In bivalves, the most studied transposon, the LTR element Steamer—a retroposon initially linked to transmissible fatal leukemia-like disease [60]—is involved in multiple HTT events [23]. The contribution of HTT in the evolution of TE repertories can be exceptionally important for DNA transposons, while LINE elements are thought to be generally transmitted through vertical inheritance [59, 61, 62]. Indeed, contrary to DNA transposons, proteins encoded by retroposons highly favor the transposition of the RNAs from which they are encoded [63]: this *cis* preference is thought to allow their long-term persistence under vertical transmission, even with the simultaneous presence of multiple, non-autonomous copies [63]. However, multiple cases of HTT involving LINE elements have been described in literature, also involving bivalves and other aquatic species [64]. Moreover, HTT events involving Harbinger elements between bivalves and sea kraits have also been recently described by [65]. Here we have not interrogated our dataset for such events and therefore their impact in the overall evolution of bivalve TE complement, and especially of DNA elements, remains poorly explored. However, our curated set of LINEs and full-length DDE/D transposons, together with the continuous rapid increase of novel genomic resources for bivalves, could represent an additional important starting point for future works.

Different bivalve orders are characterized by different LINE clades

The highly diverse ancestral bivalve LINE complement appeared to undergo multiple lineage-specific rounds of amplification and extinction/reduction events coupled with the diversification of major bivalve orders. Worthy of attention are the cases of the Unionida *M. nervosa* and *P. streckersoni* and of the chemoautotrophic symbiont-hosting *A. marissinica*. *M. nervosa* and *P. streckersoni* are characterized by an increased genome coverage of RTE-BovB elements (RTE clade) compared to other bivalves. Moreover in *M. nervosa* we identified 121 RTE-BovB autonomous elements, accounting for 44% of the total number of autonomous RTE-BovB identified across all analyzed bivalves. At the same time for both species we observed an apparent contraction of the bivalve-rich RTE-X, CR1, CR1-Zenon, and RC complements, a pattern found uniquely in this order. We can speculate that this drastic change in TE repertories could be due to their ancestral colonization of freshwater environments. Indeed, Unionida are an ancient, whole order of freshwater-only bivalves [66] and they are characterized by unique life history traits such as parental care and larval parasitism [67]. The colonization of new ecological niches and/or possible related founder effects could drive drastic change in TE content both due to alteration in the efficiency of natural selection and due to the impact of genetic drift with the stochastic loss and survival of different TE lineages [68, 69]. Similar cases of rapid LINE expansion due to genetic drift have also been observed in birds [70]. A similar scenario could also potentially occur for the deep-sea chemoautotrophic, symbiont-hosting *A. marissinica*. In this species all LINE clades appeared expanded, with a peak of activity near the present for all superfamilies and likely driven by the high amount of hosted autonomous elements (N=799) coming from 9 out of the 11 clades. Moreover, it also hosts a high number of DDE/D related transposons compared to other bivalves. Our findings are coherent with a suggested increased TE activity coupled with the diversification of pliocardiines (~70 Mya; [71]). On the other hand, we could observe multiple cases of loss/reduction of LINE representatives, for example in Ostreida (I superfamily), Pectinida (RTE-BovB clade) and Arcida (L2 and L2-2 clades).

Interestingly, oysters seem also to be generally depleted of SINEs, while Arcida, *A. marissinica*, and *B. platifrons* appeared enriched. As suggested by the absence of correlation between overall LINE and SINE genome coverage, a general estimation of the representativity of hosted LINEs is not sufficient to explain the overall variation in the genome occurrence of their non-autonomous counterparts. Until now, for only three out of eight different SINE families described in bivalves so far it has been identified their LINE donor [15] and our curated LINE library could represent an important starting point for future analyses aimed to elucidate their co-evolutionary dynamics.

Contemporary activity and long-term survival of multiple and diversified LINE lineages in bivalves

Analyzing autonomous elements, we identified multiple and diversified LINE lineages belonging to different clades that, although accounting for a relatively small proportion of the genome, co-exist within the same host. Moreover, the analysis of manually curated families showed that they may effectively be able to replicate and jump, as highlighted by the recent peak of activity identified in the repeat landscape analyses, by the presence of multiple elements showing both RT+EN domains and by the significantly positive correlation that we found between family-level transcription levels and number of insertions. Indeed, we expected to find a significantly higher amount of TE copies for highly transcribed families only when one or multiple elements are effectively able to overcome host mechanisms of post-transcriptional silencing (e.g. RNA interference). These patterns are strikingly different from what is observed in mammals where only one or few families are active at a given time and a handful of L1 lineages account for almost 20% of their genome but, again, matching what is observed in fishes and other non-mammalian vertebrates where LINEs are less dominant [55]. This mammal-specific evolutionary model is often referred to an arm race between the host and the elements and one of its landmarks is a cascade structure of the LINE phylogeny, where highly active elements are fastly replaced by new ones [33, 72, 73]. On the contrary, our results, together with the general lack of species-specific clusters with short branch lengths and high number of copies in the LINE phylogenetic trees, could highlight a reduced mobilization with multiple, less harmful 'stealth drivers' that occasionally emerge as for the previously discussed RTE-BovB elements in Unionida [33, 73-76]. At the same time this pattern could also be explained by high turnover of LINE copies due for example to ectopic recombination [33, 39] and/or lower fixation rate of recently mobilized elements [55]. Different TE evolutionary models are thought to be responsible for the different repression mechanisms adopted by the host to control transposition activity [55]. Indeed, in the arm-race scenario the host organism must quickly counteract highly active TEs through the evolution of sequence-specific repressors to limit their deleterious effect. On the contrary, in a more TE-diversified genome with multiple stealth drivers' elements could be more efficient a general process, like methylation, rather than a sequence-specific mechanism. Coherently, bivalves are characterized by a high diversity not only of LINEs but also of Class II elements and by high levels of methylation [77, 78] which could, therefore, represent the main repression mechanism.

Interestingly, across RTE, Jockey and L1 we identified an additional ancient burst of activity that seems to be shared between multiple species and multiple families were found to be active both in recent times and in the past. The ancient origin of bivalve orders makes it difficult to claim a shared activity without knowing their substitution rates and even in that case substitution saturation can obscure ancient activities. Nevertheless, the presence of both recent and ancient peaks underlies the long-term survival of these LINE lineages, as also visible in the phylogenetic trees of curated autonomous families, where their emergence tends to precede the speciation event of the host. Overall, these findings are coherent with the stealth driver model that allow TE lineages to "silently" survive over evolutionary timescales and occasionally emerge due to weakened genomic defenses, as reported in a narrower scale for the *Drosophila nasuta* species group [79], suggesting a possible important role of these elements in shaping both recent and more ancient phases of bivalve diversification.

Conclusions

In the present study we performed the first comparative analysis of transposable element evolutionary dynamics across molluscs with a particular emphasis on bivalves, an ecologically and economically important group. Despite genomic resources still being limited to few representative species compared to other clades, such as insects, the relatively low taxon sampling allowed us to deeply characterize for the first time their LINE and Class II DDE/D-related complement. Moreover, because a high-quality repeat library is essential for the analyses of new genomes, our reference set of classified LINEs and DDE/D elements, can be used to improve genome annotations and/or to easily classify novel elements across other lophotrochozoans. We also want to emphasize the necessity to extend similar analyses to other classes of transposons, empowering the scientific community with novel and high-quality genomic resources. While TEs have been hypothesized to be involved in the evolution of multiple bivalve genomic oddities, such high levels of gene presence-absence variation [80] and of hemizygosity [81], the ability to identify their possible role deeply and consistently in shaping bivalve genome evolution will be limited as long as the great majority of elements are unclassified, fragmented or not freely accessible for the scientific community.

With our approach, we discovered a diverse set of LINEs and DDE/D that were likely already greatly diversified in the most recent common ancestor of bivalves. The restricted emergence of the bivalve-rich Proto2, RTE-X, CR1-Zenon, and Academ elements could have contributed to bivalve fast radiation providing novel raw genomic material for their diversification. Moreover, we found that this LINE diversity seems to be maintained across extant species by an equally diverse set of potentially contemporary active families that could follow a stealth driver model of evolution. Indeed, multiple families seem to be able to survive and co-exist for a long period of time in the host genome without triggering the evolution of sequence-specific repression mechanisms, resembling what was previously observed in multiple non-mammalian vertebrates such as lizards and fishes. Finally, despite their relatively low genome occurrence, several LINE superfamilies/clades/type emerged and others contracted in a lineage-specific manner during the diversification of bivalves. Therefore, this highly diverse LINE complement, despite being less represented than class II elements, is a rather dynamic portion of bivalve genomes and can play important roles in local adaptations and lineage-specific evolutionary dynamics.

Material and Methods

Genomic resources and phylogeny construction

Thirty-six mollusc and three annelid genomes were downloaded from publicly available resources (NCBI, GigaDB, Dryad, MolluscDB, dbSROG and Phaidra, See Additional File 1: Tab. S1**)**, giving preference to bivalve assemblies representative of their major clades. Concerning molluscs, we selected 27 genomes belonging to bivalves, five to gastropods, two to cephalopods, and one to the polyplacophoran *A. granulata*. The species tree was manually reconstructed following the phylogenetic relationships founded in recent phylogenomic studies [82-85] as well as the reference phylogeny presented in MolluscDB [86].

Mining and annotation of interspersed repeats

For each analyzed genome we compiled species-specific repeat libraries using a combination of structural and homology-based methods. RepeatModeler v. 2.0.1 [87] with the LTR pipeline extension which include the structural-based LTRharvest [88] and LTR_retrivier packages [89], MITE Tracker [90], and HelitronScanner v. 1.1 [51] were used to build *de novo* consensus libraries. All softwares were run with default options except for HelitronScanner for which we increased the threshold of the minimum match score for both 5’ and 3’ ends from the default 5 to 10: this increases the specificity (*-ht* and *-tt*) despite decreasing the sensitivity. RepeatModeler consensus sequences were classified based on RepBase (v. 20181026) and Dfam (v. 3.1) databases, whereas MITEs were not further classified and considered only as non-autonomous DNA elements.

Bivalve genomes are characterized by high levels of duplicated genes, especially across immuno-related families (e.g. [85]) as well as by segmental duplications [91-93]. To reduce the possible inclusion of non-TE related consensus sequences, the species-specific libraries were cleaned to remove: (**a**) non-TE related genes and gene fragments, (**b**) tandem repeats, (**c**) redundancy (**d**) low copy number repeats. For the first purpose we started cleaning the reference proteomes of *H. robusta* (GCF_000326865.1), *P. canaliculata* (GCF_003073045.1), *L. gigantea* (GCF_000327385.1), *O. sinensis* (GCF_006345805.1), *M. yessoensis* (GCF_002113885.1), *C. gigas* (GCF_902806645.1), *C. virginica* (GCF_002022765.1) and *P. maximus* (GCF_902652985.1) from possible TE-related proteins. Blastp (E-value < 1E-10) was used against a reference set of transposon-related proteins covering all TE classes and obtained from the EDTA package [94] and the Repeatpeps library from the RepeatMasker package [95]. Putative TE proteins were removed, and the resulting protein set was used as a database for blastx (e-value < 1E-10) searches of our repeat libraries. Finally, ProtExcluder v. 1.1 [96] was used to remove non-TE related genes and gene fragments. For the purpose (**b**) we used the *cleanup_tandem.pl* script from the EDTA package requiring a minimum length of the consensus sequence after removing tandem repeats of 50 bp and a minimum percentage of non-ambiguous characters greater than half of the consensus length. Cleaned libraries were merged with 1,031 consensus sequences from the Mollusca RepBase library and (**c**) redundancy was reduced using CD-HIT [97] following the 80-80 rule (i.e, requiring a minimum 80% identity along the 80% of the shortest sequence; [7]) with the parameters: *-c 0.8 -n 5 -aS 0.8 -g 1 -G 0 -t 1.* As a last step (**d**), each species-specific non-redundant library was searched with blastn against the corresponding genome with a required minimum query coverage and identity of 0.7. Sequences with less than 5 hits were removed to construct our final set of consensus sequences (*i.e* 38 species-specific repeat libraries).

Annotation of repeats in each analyzed genome was achieved with running RepeatMasker v. 4.1.0 in sensitive mode (-*s*) using each of the specie-specific repeat libraries as custom database for the corresponding genome, without searching for low complexity repeats (-*nolow*) and small RNA (-*norna*). To improve the repeat annotations, the RepeatMasker output files were post-processed with RepeatCraft [98] in *loose* mode to merge closely related genomic fragments belonging to non-overlapping regions of the same consensus sequence. A hierarchical and k-means clustering of the number of TE insertions was performed respectively with the ComplexHeatmap R package v. 3.12 (Kendall's τ clustering method) and the *kmeans* function specifying 3 centers. A flowchart describing the whole workflow is presented in Additional File 2: Fig. S1**.**

ORF-based annotation of RT containing LINEs and Class II DDE/D elements

To have a more precise picture of the representation of different superfamilies and clades of both LINEs and DDE/D Class II elements we applied an ORF-based extraction and classification pipeline. Firstly, insertion sites resulting from RepeatCraft analyses were extracted with the *bedtools* suite [99] together with 1000bp at both ends to correct for possible partial/fragmented annotations due to the likely incomplete status of automated generated consensus sequences [26]. ORFinder was then used to identify and extract non-overlapping open reading frames (*-n*) with a required methionine as start codon and a minimum ORF length of at least 300 amino acids (i.e 900 nucleotides; *-ml 900*). To further characterize both Class II DDE/D related transposons and LINE elements we used an HMM-based approach. For the former, we started from the amino acid sequences corresponding to DDE/D domains found in the 17 superfamilies described in [31]. All sequences coming from each superfamily (namely hAT, Tc1/Mariner, PIF/Harbinger, CMC, Merlin, MULE, P, Kolobok, Novosib, Sola1, Sola2, Sola3, PiggyBac, Transib, Academ, Ginger, Zator) were downloaded and separately aligned with MAFFT v. 7.475 ([100]; *E-INS-i* strategy) and from each alignment we build up a superfamily-specific HMM profile using the *hmmbuild* function from the HMMER3 package [101]. The collection of all 17 profiles was then used as target database for *hmmscan* homology searches (E-value < 1E-5) against all extracted ORFs provisionally annotating each element based on the corresponding best hit. To avoid misclassification of Ginger elements due to their high homology to *Gypsy*-encoded integrases [102] and to confirm the classification of all ORFs we additionally blasted all significant hits against the full RepeatPep library (Blastp; E-value 1E-05), imitating a reciprocal best-hit approach. Sequences with a best hit against a different superfamily compared to our previous HMM-based classification were considered as miss-classified and discarded.

For LINE elements we started with an RPSblast search on the same set of extracted and translated ORFs against the complete CDD database (E-value < 1E-05). Sequences with a significant hit against RT-related profiles were considered as putative retrotransposons (see Additional File 28: Tab. S8 for a list of CDD entries). To distinguish between LTR- and LINE-derived RT-containing ORFs all LINE and LTR elements from the Repeatpeps library were extracted and separately aligned with MAFFT v. 7.475 (*l-INS-i* strategy) together with the seed sequences of the RVT_1 Pfam HMM profile (PF00078) to manually identify boundaries of the RT domain. We extracted LINE and LTR RTs from the resulting alignments and we built two class-specific HMM profiles with the *hmmbuild* function from the HMMER3 package. The two profiles were then used as target database for *hmmscan* (E-value < 1E-5) homology searches of our previously identified RT-containing ORFs. Sequences with a best hit against the LTR-specific RT profile were considered as putative LTR and therefore discarded from subsequent analyses. LINE elements were considered autonomous when both RT and EN domains (see Additional File 28: Tab. S8 for a list of CDD entries) were present on the same ORF (i.e. non intervening stop codons). Sequences missing the EN domain were classified as RT-only LINEs.

To test the interplay between assembly quality and the ability to identify RT-containing and autonomous LINEs as well as DDE/D-related transposons, we checked for correlation between number of identified elements and contig/scaffold N50 with Spearman's rank correlation tests.

All confirmed LINEs (regardless being autonomous or RT-only) and DDE/D containing transposons were clustered at the nucleotide level using CD-HIT and following the 80-80 rule (same parameter set used for repeat library construction). Therefore, hereafter we will refer to clusters as groups of TEs related by high nucleotide homology along their coding sequence to distinguish them from the canonical transposon families which ideally should take into consideration the elements along their entire length [7].

For LINE elements only we additionally called "low-copy number clusters" clusters with less than 5 members and as "singleton clusters" sequences that did not fall in any cluster. For Class II elements we avoid such classification because non-autonomous members of a family can replicate through the genome parasitizing their autonomous counterparts. Moreover, while the presence of a complete ORF can give some first insight on which superfamilies/clades could have been more active in recent/mid times, on the other hand, it must be noted that this approach is not able to identify non-autonomous elements thus greatly underestimating the number of short Class II transposons.

Tree-based classification of ORF-containing LINE elements

ORF-containing LINE elements were classified using a phylogenetic approach. We adopted the superfamily classification scheme proposed by [7] and the clade classification proposed by [29], as in [103], while we use the "type" term to refer to the RepeatMasker or Dfam classification schemes [104]. Starting from previously identified clusters (>5 members), we extracted the amino acid sequence of the RT domain based on the coordinates of the RPSblast hits. RTs segments were aligned with MAFFT v. 7.475 (*g-INS-i* strategy) and cleaned from columns with gaps in more than the 50% of the sequences using TrimAl [105]. *Cons* from the EMBOSS package [106] was then used to build up a consensus sequence from the resulting alignment setting the parameter *plurality* to 3. RT consensus sequences were then aligned together with reference LINE sequences from [29] and a subset of LTR and LINE elements from the Repeatpeps library, using MAFFT and a *g-INS-i* strategy. Poorly aligned sequences were removed from the alignment using TrimaAl (*-resoverlap* 0.75 *-seqoverlap* 80). Because of the short RT domain, the deep divergence time of LINE superfamilies and the consequently difficulties in identify stable LINE phylogenies (e.g. [29, 30, 107]) we used a combination of Neighbour-Joining, unconstrained Maximum Likelihood (ML) and constrained ML tree inferences. Each topology was then statistically tested in a ML framework to produce a confidently phylogeny useful for LINE classification. We performed (**a**) a Neighbour-Joining (NJ) clustering with Clearcut v. 1.0.9 [108], reshuffling the distance matrix and using a traditional Neighbour-Joining algorithm (*--shuffle* and *--neighbor* options, respectively); (**b**) 5 unconstrained Maximum Likelihood (ML) tree searches with IQtree v. 2.1.3 [109] and the corresponding best-fit evolutionary model identified by ModelFinder2 [110]; 6 constrained ML tree searches forcing (**c**) the full NJ topology (FullNJ constraint, one run) and (**d**) only the monophyly of LINEs superfamilies, as inferred by the NJ tree, with the exception of Jockey and I superfamilies which were constrained in a single, comprehensive monophyletic clade (SupFAM constraint, 5 runs). For the unconstrained and the SupFAM constrained ML tree inferences (analyses **b** and **d**, respectively) nodal support was estimated with 1,000 UltraFastBootstrap replicates [111]. All ML topologies were tested using Kishino-Hasegawa test [112], Shimodaira-Hasegawa test [113], expected likelihood weights [114], and approximately unbiased (AU) test [115]. As an additional confirmation of our classification and to avoid the inclusion of Penelope-like elements we (**a**) blasted each consensus RT (blastp; E-value < 1E-5) against all protein sequences from the RepeatPeps library extracting the best-hit for each query sequence and (**b**) used the online implementation of RTClass1 [29] on a random subset of 111 RT sequences covering all identified clades. Low-copy numbers, singletons, and clusters removed by TrimAl were classified based on Blastp best-hit (E-value < 0.05) against tree-based classified clusters and the whole RepeatPeps library for competing purposes. For the low-copy clusters, one representative (*i.e.* the longest) sequence was used. For bivalve species, and excluding the poorly represented R2 superfamily, the correlation between the number of RT-containing LINEs and the number of clusters in each identified LINE clade was tested for each superfamily separately with Spearman's rank correlation tests.

Additional prediction of SINEs in a subset of selected species

To have a first insight into the SINE composition of bivalves we selected 13 species (namely: *A. marissinca, C. sinensis, C. gigas, S. glomerata, T. granosa, S. broughtonii, M. coruscus, B. platifrons, S. constricta, S. grandis, P. maximus, M. yessoensis, M. nervosa*) representative of Venerida, Ostreida, Arcida, Mytilida, Adepedonta, Pectinida and Unionida, to mine additional SINE candidates using SINE_Scan v1.1.1 [116]. This software collects and validates SINE candidates based on copy number across the genome, presence of target site duplications (TSDs) and trRNA-related heads. All representative elements were merged with consensus sequences classified as SINEs by RepeatModeler in the corresponding species-specific repeat library (See Material and Methods section 2.2) and subjected to manual validation and curation as described in the following section. After this process, curated consensus sequences were annotated at the family level using the RepeatClassifier utility from the RepeatModeler package.

Manual curation of LINEs, SINEs, and DDE/D-related transposons

We selected a set of the previously found LINEs RT, SINEs and DDE/D-containing clustersfor manual refinement, following [27] guidelines. For LINEs we selected all clusters with at least one autonomous element (i.e., encoding for an ORF with both RT and EN domains without interrupting stop codons) and five other sequences (both autonomous and/or RT-only) while for DDE/D elements we required only the presence of at least five elements in the corresponding cluster. These criteria were chosen in order to prioritize the manual curation of sequences that likely possess one or more autonomous copies across the genome and thus could potentially be recently mobilized or mobilize their non-autonomous counterparts. Members of LINEs and DDE/D-related clusters were aligned at the nucleotide level using MAFFT (*--auto* strategy). CIAlign [117] was then used to remove insertions found in less than 50% of the sequences and to construct a nucleotide consensus sequence (*--remove-insertions* and *--make-consensus* option). At this set of LINEs and DDE/D preliminary consensus we also added all the aforementioned SINEs and all sequences were subjected to a “Blast-Extend-Extract” process with a minimum required query coverage and identity of 70, extending each hit by 3kb and extracting the top 25 hits for each query sequence and building up a preliminary consensus sequence using CIAlign. Resulting alignments were manually inspected to: (i) identify structural features (e.g, microsatellites for LINEs and SINEs at the 3’ end, 5’ truncations for LINEs, terminal inverted repeats and superfamily-specific motifs for DDE/D elements), (ii) identify boundaries of the elements searching for TSDs whenever possible, (iii) identify domain signatures using the CDD web server and (iv) correct and extend as long as possible the consensus sequence. Additionally, for SINE only, we also required (a) the presence of a detectable tRNA-related region at the 5’ ends and predicted with tRNAScan-SE (Sequence source: Mixed; Score cut-off 0.01; [118]) and (b) the presence of a central domain and/or a tail region after the tRNA-related head. It must be noted that the presence of TSDs to confirm the boundaries of the element was only required for SINEs and Class II superfamilies that exhibit them (thus excluding for example the SPY group from the PIF-Harbinger superfamily; see [119]), while for LINE elements their presence was checked but not required because of difficulties in finding them due to frequent 5' truncations. For LINEs we instead rely on the distinctive decay of the alignment quality towards the 5' end caused by 5' truncations [27] curating each consensus until at least 3 sequences could confidently be aligned. Relationships between the number of curated families and the number of autonomous elements identified in each species was tested using Spearman's rank correlation test.

Genome annotation of LINEs and SINEs using manually curated libraries and phylogenetic inference of curated LINE families.

After manual curation we focused our analyses to the greatly understudied LINE complement. All LINEs and SINEs libraries were merged and CD-HIT-EST was used to remove redundant copies following the 80-80 rule. The merged non-redundant library was used in an additional RepeatMasker analysis in sensitive mode and increasing the minimum score to 400 from the default value of 225 (-*cutoff* 400), to remove low scoring annotations. We tested for a correlation between genome coverage of LINEs and SINEs in the 12 selected species using Spearman's rank correlation. For LINEs only, CpG corrected Kimura distances of each copy from its consensus were calculated with the *calcDivFromAlign.pl* script from the RepeatMasker package. We define long-term survival families consensus that show both recent (<5% divergence from the consensus) and ancient (>30% divergence from the consensus) activity requiring a minimum of 30 copies in the recent and 5 in the ancient divergence bins. For this latest purpose we applied a 3' anchor-based counting method to reduce possible overestimations of the insertion number and spurious alignment between SINEs and their possible LINE counterparts. Briefly, we only count insertions that map to the first 50 nucleotides of the 3' end of each consensus sequence and with a length of at least 100 bp based on aligned query and subject coordinates reported in the RepeatMasker out file.

Finally, from each LINE consensus sequence we extracted the RT domain as previously described and, separately for each superfamily, we aligned all fragments and inferred a ML tree (MAFFT *g-INS-i* strategy; ModelFinder and IQ-TREE with 1,000 Ultrafast Bootstrap replicates).

Transcription potential of curated LINE families

To further test for activity potential of curated families in mature gonad tissues we collected from NCBI paired-ends poly (A)-enriched RNA-seq data from mature male and female samples. Three biological replicates for each tissue were selected for *C. gigas* (SRR12564937, SRR12564938, SRR12564939, SRR12564936, SRR12564935, SRR12564940), *Chlamys farreri* (SRR5130887, SRR5130883, SRR5130863, SRR5130886, SRR5130875, SRR5130872), *M. yessoensis* (SRR9157572, SRR9157579, SRR9157580, SRR9157581, SRR9157582, SRR9157588), *Mercenaria mercenaria* (SRR10951876, SRR10951875, SRR10951874, SRR10951867, SRR10951866, SRR10951865), and *Sinonovacula constricta* (SRR9937011, SRR9937009, SRR9937008, SRR9937013, SRR9937012, SRR9937010). Raw reads were trimmed and deprived of adapters using bbduck from the bbmap package [120], requiring a minimum quality of 20 (*trimq=20*) and a minimum length of the reads after trimming of 75 (*minlen=75*). We decided to map all RNAseq reads only on 3' anchored LINE insertions, as defined in the previous section, longer than 3,000 bp and extracted with bedtools. These latest filters should ensure that reads originate from families that likely possess autonomous copies across the genome. To not discard multi mapping reads, we obtained a per-family raw count for each sample using TEtools [121] and bowtie2 [122] to align reads on extracted insertions. Raw counts were then normalized by the length of the corresponding family consensus sequences and TPM values were calculated. Log2-transformed normalized counts were tested for a correlation with the number of previously identified 3' anchored insertions with a minimum length of 100 bp for the corresponding family for each species, tissue and biological replicate separately.

Availability of Data and Materials

All data generated or analysed during this study are included in this published article, its supplementary information files and publicly available repositories. Phylogenetic trees can be found in **Additional File 33: Data S1** together with the multiple sequence alignment used to generate them in **Additional File 34: Data S2**. Manually curated families can be found in **Additional File 35-36-37** with a RepeatMasker formatted style as well as in the GitHub repository https://github.com/CompBio-BO/Bivalvia_TEs and in DFAM under Creative Commons CC0 1.0 public domain license. All supplementary data have been also deposited in a figshare database under the DOI https://doi.org/10.6084/m9.figshare.22188280.v1 [123]. Scripts used to automatically generate the species-specific repeat libraries and to extract LINEs and DDE/D-related ORFs can be found in Github (https://github.com/jacopoM28/EvoTEs_BiV) and in Zenodo under the DOI 10.5281/zenodo.7944844 [124].

Authors' contributions

JM, AL and FG designed the study; JM collected the data and performed bioinformatic analyses; JM, FN, SS, and DP curated the data; JM wrote the first version of the manuscript and additional supplementary files; JM, FG, AL, and FN revised the manuscript. All authors read and approved the final version of the manuscript.

Funding

This work was supported by the Canziani bequest funded to F.G. and A.L., and the 'Ricerca Fondamentale Orientata' (RFO) funding from the University of Bologna to F.G. and A.L.

Acknowledgements

We thank the EVO·COM lab members Mariangela Iannello, Giobbe Forni and Giovanni Piccinini for useful comments and discussions about the analyses. We would also like to thank Prof. David Adelson, another anonymous reviewer, and the BMC Biology AE Dr. Ruslan Kalendar for providing insightful comments on a first version the manuscript which helped us to greatly improve the overall quality of the manuscript.

References

1. Werren JH, Nur U, Wu CI. Selfish genetic elements. Trends Ecol Evol. 1988 Nov;3(11):297-302.
2. Bourque G, Burns KH, Gehring M, Gorbunova V, Seluanov A, Hammell M, et al. Ten things you should know about transposable elements. Genome Biol. 2018 Nov 19;19(1):199.
3. Wells JN, Feschotte C. A Field guide to eukaryotic transposable elements. Annu Rev Genet. 2020 Nov 23;54:539–61.
4. Finnegan DJ. Eukaryotic transposable elements and genome evolution. Trends Genet. 1989 Jan 1;5:103–7.
5. Arkhipova IR. Using bioinformatic and phylogenetic approaches to classify transposable elements and understand their complex evolutionary histories. Mob DNA. 2017 Dec 6;8(1):19.
6. Bao W, Kojima KK, Kohany O. Repbase Update, a database of repetitive elements in eukaryotic genomes. Mob DNA. 2015 Jun 2;6(1):11.
7. Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B, et al. A unified classification system for eukaryotic transposable elements. Nat Rev Genet. 2007 Dec;8(12):973–82.
8. Luchetti A, Mantovani B. Non-LTR R2 element evolutionary patterns: phylogenetic incongruences, rapid radiation and the maintenance of multiple lineages. PLOS ONE. 2013 Feb 25;8(2):e57076.
9. Deininger PL, Batzer MA, Hutchison CA, Edgell MH. Master genes in mammalian repetitive DNA amplification. Trends Genet. 1992 Sep;8(9):307–11.
10. Kelleher ES, Barbash DA, Blumenstiel JP. Taming the turmoil within: new insights on the containment of transposable elements. Trends Genet. 2020 Jul 1;36(7):474–89.
11. Han K, Xing J, Wang H, Hedges DJ, Garber RK, Cordaux R, et al. Under the genomic radar: the stealth model of Alu amplification. Genome Res. 2005 May;15(5):655–64.
12. Šatović E, Plohl M. Tandem repeat-containing MITEs in the clam *Donax trunculus*. Genome Biol Evol. 2013;5(12):2549–59.
13. Vera M, Bello X, Álvarez-Dios JA, Pardo BG, Sánchez L, Carlsson J, et al. Screening of repetitive motifs inside the genome of the flat oyster (*Ostrea edulis*): transposable elements and short tandem repeats. Mar Genomics. 2015 Dec;24:335-341.

14. Luchetti A, Šatović E, Mantovani B, Plohl M. RUDI, a short interspersed element of the V-SINE superfamily widespread in molluscan genomes. Mol Genet Genomics. 2016 Jun 1;291(3):1419–29.

15. Nishihara H, Plazzi F, Passamonti M, Okada N. MetaSINEs: Broad distribution of a novel SINE superfamily in animals. Genome Biol Evol. 2016 Feb 12;8(3):528–39.

16. Šatović E, Plohl M. Two new miniature inverted-repeat transposable elements in the genome of the clam *Donax trunculus*. Genetica. 2017 Oct;145(4–5):379–85.

17. Biscotti MA, Barucca M, Canapa A. New insights into the genome repetitive fraction of the Antarctic bivalve *Adamussium colbecki*. PLOS ONE. 2018 Mar 28;13(3):e0194502.

18. Lee SI, Gim JA, Lim MJ, Kim HS, Nam BH, Kim NS. Ty3/Gypsy retrotransposons in the Pacific abalone *Haliotis discus hannai*: characterization and use for species identification in the genus Haliotis. Genes Genomics. 2018 Feb;40(2):177–87.

19. Puzakov MV, Puzakova LV, Cheresiz SV. An analysis of IS630/Tc1/mariner transposons in the genome of a Pacific oyster, *Crassostrea gigas*. J Mol Evol. 2018 Oct;86(8):566–80.

20. Šatović E, Luchetti A, Pasantes JJ, García-Souto D, Cedilak A, Mantovani B, et al. Terminal-Repeat Retrotransposons in Miniature (TRIMs) in bivalves. Sci Rep. 2019 Dec 27;9(1):19962.

21. Puzakov MV, Puzakova LV, Cheresiz SV. The Tc1-like elements with the spliceosomal introns in mollusk genomes. Mol Genet Genomics. 2020 May;295(3):621–33.

22. Vojvoda Zeljko T, Pavlek M, Meštrović N, Plohl M. Satellite DNA-like repeats are dispersed throughout the genome of the Pacific oyster *Crassostrea gigas* carried by Helentron non-autonomous mobile elements. Sci Rep. 2020 Sep 15;10(1):15107.

23. Metzger MJ, Paynter AN, Siddall ME, Goff SP. Horizontal transfer of retrotransposons between bivalves and other aquatic species of multiple phyla. Proc Natl Acad Sci U S A. 2018 May;115(18):E4227–35.

24. Thomas-Bulle C, Piednoël M, Donnart T, Filée J, Jollivet D, Bonnivard É. Mollusc genomes reveal variability in patterns of LTR-retrotransposons dynamics. BMC Genomics. 2018 Nov 15;19(1):821.

25. Marino A, Kizenko A, Wong WY, Ghiselli F, Simakov O. Repeat age decomposition informs an ancient set of repeats associated with coleoid cephalopod divergence. Front Genet. 2022 Mar 14;13:793734.

26. Platt RN II, Blanco-Berdugo L, Ray DA. Accurate transposable element annotation is vital when analyzing new genome assemblies. Genome Biol Evol. 2016 Feb 1;8(2):403–10.

27. Goubert C, Craig RJ, Bilat AF, Peona V, Vogan AA, Protasio AV. A beginner's guide to manual curation of transposable elements. Mob DNA. 2022 Mar 30;13(1):7.

28. Gundappa MK, Peñaloza C, Regan T, Boutet I, Tanguy A, Houston RD, et al. Chromosome-level reference genome for European flat oyster (Ostrea edulis L.). Evol Appl. 2022;15(11):1713–29.

29. Kapitonov VV, Tempel S, Jurka J. Simple and fast classification of non-LTR retrotransposons based on phylogeny of their RT domain protein sequences. Gene. 2009 Dec 15;448(2):207–13.

30. Kojima KK. Structural and sequence diversity of eukaryotic transposable elements. Genes Genet Syst. 2020 Jan 30;94(6):233-252.

31. Yuan YW, Wessler SR. The catalytic domain of all eukaryotic cut-and-paste transposase superfamilies. Proc Natl Acad Sci U S A. 2011 May 10;108(19):7884–9.

32. 1. Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J. Repbase Update, a database of eukaryotic repetitive elements. CGR. 2005;110(1–4):462–7.

33. Ivancevic AM, Kortschak RD, Bertozzi T, Adelson DL. LINEs between species: evolutionary dynamics of LINE-1 retrotransposons across the eukaryotic tree of life. Genome Biol Evol. 2016 Nov 1;8(11):3301–22.

34. Feschotte C, Pritham EJ. DNA transposons and the evolution of eukaryotic genomes. Annu Rev Genet. 2007;41:331–68.

35. Meyer A, Schloissnig S, Franchini P, Du K, Woltering JM, Irisarri I, et al. Giant lungfish genome elucidates the conquest of land by vertebrates. Nature. 2021 Feb;590(7845):284–9.

36. Corrochano-Fraile A, Davie A, Carboni S, Bekaert M. Evidence of multiple genome duplication events in *Mytilus* evolution. BMC Genomics. 2022 May 2;23(1):340.

37. Powell D, Subramanian S, Suwansa-ard S, Zhao M, O'Connor W, Raftos D, et al. The genome of the oyster Saccostrea offers insight into the environmental resilience of bivalves. DNA Res. 2018 Dec 1;25(6):655–65.

38. Varney RM, Speiser DI, McDougall C, Degnan BM, Kocot KM. The iron-responsive genome of the chiton *Acanthopleura granulata*. Genome Biol Evol. 2021 Jan 7;13(1):evaa263.

39. Yang JL, Feng DD, Liu J, Xu JK, Chen K, Li YF, et al. Chromosome-level genome assembly of the hard-shelled mussel Mytilus coruscus, a widely distributed species from the temperate areas of East Asia. GigaScience. 2021 Apr 1;10(4):giab024.

40. Liu C, Ren Y, Li Z, Hu Q, Yin L, Wang H, et al. Giant African snail genomes provide insights into molluscan whole-genome duplication and aquatic–terrestrial transition. Mol Ecol Resour. 2021;21(2):478–94.

41. Song H, Guo X, Sun L, Wang Q, Han F, Wang H, et al. The hard clam genome reveals massive expansion and diversification of inhibitors of apoptosis in Bivalvia. BMC Biology. 2021 Jan 25;19(1):15.

42. Petersen M, Armisén D, Gibbs RA, Hering L, Khila A, Mayer G, et al. Diversity and evolution of the transposable element repertoire in arthropods with particular reference to insects. BMC Ecology and Evolution. 2019 Jan 9;19(1):11.

43. Chalopin D, Naville M, Plard F, Galiana D, Volff JN. Comparative analysis of transposable elements highlights mobilome diversity and evolution in vertebrates. Genome Biol Evol. 2015 Feb 1;7(2):567–80.

44. Wang D, Zheng Z, Li Y, Hu H, Wang Z, Du X, et al. Which factors contribute most to genome size variation within angiosperms? Ecol Evol. 2021;11(6):2660–8.

45. Li Y, Sun X, Hu X, Xun X, Zhang J, Guo X, et al. Scallop genome reveals molecular adaptations to semi-sessile life and neurotoxins. Nat Commun. 2017 Nov 23;8(1):1721.

46. Wang S, Zhang J, Jiao W, Li J, Xun X, Sun Y, et al. Scallop genome provides insights into evolution of bilaterian karyotype and development. Nat Ecol Evol. 2017 Apr 3;1(5):1–12.

47. Li C, Liu X, Liu B, Ma B, Liu F, Liu G, et al. Draft genome of the Peruvian scallop Argopecten purpuratus. GigaScience. 2018 Apr 1;7(4):giy031.

48. Kenny NJ, McCarthy SA, Dudchenko O, James K, Betteridge E, Corton C, et al. The gene-rich genome of the scallop Pecten maximus. GigaScience. 2020 May 1;9(5):giaa037.

49. Kapusta A, Suh A, Feschotte C. Dynamics of genome size evolution in birds and mammals. Proc Natl Acad Sci U S A. 2017 Feb 21;114(8):E1460–9.

50. Thomas J, Vadnagara K, Pritham EJ. DINE-1, the highest copy number repeats in *Drosophila melanogaster* are non-autonomous endonuclease-encoding rolling-circle transposable elements (Helentrons). Mob DNA. 2014 Jun 4;5(1):18.

51. Xiong W, He L, Lai J, Dooner HK, Du C. HelitronScanner uncovers a large overlooked cache of Helitron transposons in many plant genomes. Proc Natl Acad Sci U S A. 2014 Jul 15;111(28):10263–8.

52. Eickbush TH, Furano AV. Fruit flies and humans respond differently to retrotransposons. Curr Opin Genet Dev. 2002 Dec 1;12(6):669–74.

53. Volff JN, Bouneau L, Ozouf-Costaz C, Fischer C. Diversity of retrotransposable elements in compact pufferfish genomes. Trends Genet. 2003 Dec;19(12):674-8.

54. Howe K, Clark MD, Torroja CF, Torrance J, Berthelot C, Muffato M, et al. The zebrafish reference genome sequence and its relationship to the human genome. Nature. 2013 Apr;496(7446):498–503.

55. Boissinot S, Sookdeo A. The evolution of LINE-1 in vertebrates. Genome Biol Evol. 2016 Dec 1;8(12):3485–507.

56. Plazzi F, Puccio G, Passamonti M. Burrowers from the past: mitochondrial signatures of Ordovician bivalve infaunalization. Genome Biol Evol. 2017 Apr 1;9(4):956–67.

57. Kocot KM, Poustka AJ, Stöger I, Halanych KM, Schrödl M. New data from Monoplacophora and a carefully-curated dataset resolve molluscan relationships. Sci Rep. 2020 Jan 9;10(1):101.

58. Luchetti A, Mantovani B. Non-LTR R2 Element Evolutionary Patterns: Phylogenetic Incongruences, Rapid Radiation and the Maintenance of Multiple Lineages. PLOS ONE. 2013 Feb 25;8(2):e57076.

59. Zhang HH, Peccoud J, Xu MRX, Zhang XG, Gilbert C. Horizontal transfer and evolution of transposable elements in vertebrates. Nat Commun. 2020 Mar 13;11(1):1362.

60. Arriagada G, Metzger MJ, Muttray AF, Sherry J, Reinisch C, Street C, et al. Activation of transcription and retrotransposition of a novel retroelement, Steamer, in neoplastic hemocytes of the mollusk *Mya arenaria*. Proc Natl Acad Sci U S A. 2014 Sep 30;111(39):14175–80.

61. Peccoud J, Loiseau V, Cordaux R, Gilbert C. Massive horizontal transfer of transposable elements in insects. Proc Natl Acad Sci U S A. 2017 May 2;114(18):4721–6.

62. Reiss D, Mialdea G, Miele V, Vienne DM de, Peccoud J, Gilbert C, et al. Global survey of mobile DNA horizontal transfer in arthropods reveals Lepidoptera as a prime hotspot. PLOS Genetics. 2019 Feb 1;15(2):e1007965.

63. Wei W, Gilbert N, Ooi SL, Lawler JF, Ostertag EM, Kazazian HH, et al. Human L1 Retrotransposition: cisPreference versus trans Complementation. Molecular and Cellular Biology. 2001 Feb 15;21(4):1429–39.

64. Ivancevic AM, Kortschak RD, Bertozzi T, Adelson DL. Horizontal transfer of BovB and L1 retrotransposons in eukaryotes. Genome Biol. 2018 Jul 9;19(1):85.

65. Galbraith JD, Ludington AJ, Sanders KL, Suh A, Adelson DL. Horizontal transfer and subsequent explosive expansion of a DNA transposon in sea kraits (*Laticauda*). Biol Lett. 2021 Sep;17(9):20210342.

66. Lydeard C, Cummings KS. Unionidae Rafinesque, 1820, and the general Unionida. In: Freshwater mollusks of the world: a distribution atlas. Baltimore, Maryland: Johns Hopkins University Press; 2019. p. 202–9.

67. Barnhart MC, Haag WR, Roston WN. Adaptations to host infection and larval parasitism in Unionoida. J North Am Benthol Soc. 2008 Jun;27(2):370–94.

68. Pritham EJ. Transposable elements and factors influencing their success in eukaryotes. J Hered. 2009 Sep-Oct;100(5):648-55.

69. Szitenberg A, Cha S, Opperman CH, Bird DM, Blaxter ML, Lunt DH. Genetic drift, not life history or RNAi, determine long-term evolution of transposable elements. Genome Biol Evol. 2016 Sep 1;8(9):2964–78.

70. Galbraith JD, Kortschak RD, Suh A, Adelson DL. Genome stability is in the eye of the beholder: CR1 retrotransposon activity varies significantly across avian diversity. Genome Biol Evol. 2021 Dec 1;13(12):evab259.

71. Ip JCH, Xu T, Sun J, Li R, Chen C, Lan Y, et al. Host–Endosymbiont Genome Integration in a Deep-Sea Chemosymbiotic Clam. Mol Biol Evol. 2021 Feb 1;38(2):502–18.

72. Furano AV. The biological properties and evolutionary dynamics of mammalian LINE-1 retrotransposons. Prog Nucleic Acid Res Mol Biol. 2000;64:255-94.

73. Cordaux R, Batzer MA. The impact of retrotransposons on human genome evolution. Nat Rev Genet. 2009 Oct;10(10):691–703.

74. Le Rouzic A, Capy P. The first steps of transposable elements invasion: parasitic strategy vs. genetic drift. Genetics. 2005 Feb;169(2):1033-43

75. Le Rouzic A, Boutin TS, Capy P. Long-term evolution of transposable elements. Proc Natl Acad Sci U S A. 2007 Dec 4;104(49):19375–80.

76. Kijima TE, Innan H. Population genetics and molecular evolution of DNA sequences in transposable elements. A simulation framework. Genetics. 2013 Nov 1;195(3):957–67.

77. Fallet M, Luquet E, David P, Cosseau C. Epigenetic inheritance and intergenerational effects in mollusks. Gene. 2020 Mar 1;729:144166.

78. Männer L, Schell T, Provataris P, Haase M, Greve C. Inference of DNA methylation patterns in molluscs. Philos Trans R Soc Lond B Biol Sci. 2021 May 24;376(1825):20200166.

79. Wei KHC, Mai D, Chatla K, Bachtrog D. Dynamics and impacts of transposable element proliferation in the *Drosophila nasuta* species group radiation. Mol Biol Evol. 2022 May 1;39(5):msac080.

80. Gerdol M, Moreira R, Cruz F, Gómez-Garrido J, Vlasova A, Rosani U, et al. Massive gene presence-absence variation shapes an open pan-genome in the Mediterranean mussel. Genome Biol. 2020 Nov 10;21(1):275.

81. Calcino AD, Kenny NJ, Gerdol M. Single individual structural variant detection uncovers widespread hemizygosity in molluscs. Philos Trans R Soc Lond B Biol Sci. 2021 May 24;376(1825):20200153.

82. González VL, Andrade SCS, Bieler R, Collins TM, Dunn CW, Mikkelsen PM, et al. A phylogenetic backbone for Bivalvia: an RNA-seq approach. Proc Biol Sci. 2015 Feb; 282(1801).

83. Combosch DJ, Collins TM, Glover EA, Graf DL, Harper EM, Healy JM, et al. A family-level Tree of Life for bivalves based on a Sanger-sequencing approach. Mol Phylogenet Evol. 2017 Feb 1;107:191–208.

84. Lemer S, Bieler R, Giribet G. Resolving the relationships of clams and cockles: dense transcriptome sampling drastically improves the bivalve tree of life. Proc Biol Sci. 2019 Feb 13;286(1896):20182684.

85. Regan T, Stevens L, Peñaloza C, Houston RD, Robledo D, Bean TP. Ancestral physical stress and later immune gene family expansions shaped bivalve mollusc evolution. Genome Biol Evol. 2021 Aug 1;13(8):evab177.

86. Liu F, Li Y, Yu H, Zhang L, Hu J, Bao Z, et al. MolluscDB: an integrated functional and evolutionary genomics database for the hyper-diverse animal phylum Mollusca. Nucleic Acids Res. 2021 Jan 8;49(D1):D1556.

87. Flynn JM, Hubley R, Goubert C, Rosen J, Clark AG, Feschotte C, et al. RepeatModeler2 for automated genomic discovery of transposable element families. Proc Natl Acad Sci U S A. 2020 Apr 28;117(17):9451–7.

88. Ellinghaus D, Kurtz S, Willhoeft U. LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. BMC Bioinformatics. 2008 Jan 14;9(1):18.

89. Ou S, Jiang N. LTR_retriever: A highly accurate and sensitive program for identification of long terminal repeat retrotransposons. Plant Physiol. 2018 Feb;176(2):1410–22.

90. Crescente JM, Zavallo D, Helguera M, Vanzetti LS. MITE Tracker: an accurate approach to identify miniature inverted-repeat transposable elements in large genomes. BMC Bioinformatics. 2018 Oct 3;19(1):348.

91. Peng J, Li Q, Xu L, Wei P, He P, Zhang X, et al. Chromosome-level analysis of the Crassostrea hongkongensis genome reveals extensive duplication of immune-related genes in bivalves. Mol Ecol Resour. 2020;20(4):980–94.

92. Qi H, Li L, Zhang G. Construction of a chromosome-level genome and variation map for the Pacific oyster *Crassostrea gigas*. Mol Ecol Resour. 2021 Jul;21(5):1670-1685.

93. Farhat S, Bonnivard E, Pales Espinosa E, Tanguy A, Boutet I, Guiglielmoni N, et al. Comparative analysis of the *Mercenaria mercenaria* genome provides insights into the diversity of transposable elements and immune molecules in bivalve mollusks. BMC Genomics. 2022 Mar 8;23(1):192.

94. Ou S, Su W, Liao Y, Chougule K, Agda JRA, Hellinga AJ, et al. Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. Genome Biol. 2019 Dec 16;20(1):275.

95. Tarailo-Graovac M, Chen N. Using RepeatMasker to identify repetitive elements in genomic sequences. Curr Protoc Bioinformatics. 2009 Mar;Chapter 4:4.10.1-4.10.14.

96. Campbell MS, Law M, Holt C, Stein JC, Moghe GD, Hufnagel DE, et al. MAKER-P: A tool kit for the rapid creation, management, and quality control of plant genome annotations. Plant Physiol. 2014 Feb 1;164(2):513–24.

97. Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics. 2012 Dec 1;28(23):3150–2.

98. Wong WY, Simakov O. RepeatCraft: a meta-pipeline for repetitive element de-fragmentation and annotation. Bioinformatics. 2019 Mar 15;35(6):1051–2.

99. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010 Mar 15;26(6):841–2.

100. Katoh K, Standley DM. MAFFT Multiple Sequence Alignment Software Version 7: improvements in performance and usability. Mol Biol Evol. 2013 Apr 1;30(4):772–80.

101. Eddy SR. Accelerated profile HMM searches. PLoS Comput Biol. 2011 Oct 20;7(10):e1002195.

102. Bao W, Kapitonov VV, Jurka J. Ginger DNA transposons in eukaryotes and their evolutionary relationships with long terminal repeat retrotransposons. Mob DNA. 2010 Jan 25;1:3.

103. Metcalfe CJ, Casane D. Modular organization and reticulate evolution of the ORF1 of Jockey superfamily transposable elements. Mob DNA. 2014 Jul 1;5(1):19.

104. Wheeler TJ, Clements J, Eddy SR, Hubley R, Jones TA, Jurka J, et al. Dfam: a database of repetitive DNA based on profile hidden Markov models. Nucleic Acids Res. 2013 Jan 1;41(D1):D70–82.

105. Capella-Gutierrez S, Silla-Martinez JM, Gabaldon T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics. 2009 Aug 1;25(15):1972–3.

106. Rice P, Longden I, Bleasby A. EMBOSS: the European Molecular Biology Open Software Suite. Trends Genet. 2000 Jun;16(6):276–7.

107. Moura Gama J, Ludwig A, Gazolla CB, Guizelini D, Recco-Pimentel SM, Bruschi DP. A genomic survey of LINE elements in Pipidae aquatic frogs shed light on Rex-elements evolution in these genomes. Mol Phylogenet Evol. 2022 Mar 1;168:107393.

108. Sheneman L, Evans J, Foster JA. Clearcut: a fast implementation of relaxed neighbor joining. Bioinformatics. 2006 Nov 15;22(22):2823–4.

109. Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A, et al. IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. Molecular Biology and Evolution. 2020 May 1;37(5):1530–4.

110. Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermiin LS. ModelFinder: fast model selection for accurate phylogenetic estimates. Nat Methods. 2017 Jun;14(6):587–9.

111. Hoang DT, Chernomor O, von Haeseler A, Minh BQ, Vinh LS. UFBoot2: improving the ultrafast bootstrap approximation. Mol Biol Evol. 2018 Feb 1;35(2):518–22.

112. Kishino H, Hasegawa M. Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in hominoidea. J Mol Evol. 1989 Aug 1;29(2):170–9.

113. Shimodaira H, Hasegawa M. Multiple comparisons of log-likelihoods with applications to phylogenetic inference. Mol Biol Evol. 1999 Aug 1;16(8):1114.

114. Strimmer K, Rambaut A. Inferring confidence sets of possibly misspecified gene trees. Proc Biol Sci. 2002 Jan 22;269(1487):137–42.

115. Shimodaira H. An approximately unbiased test of phylogenetic tree selection. Syst Biol. 2002 Jun;51(3):492-508.

116. Mao H, Wang H. SINE_scan: an efficient tool to discover short interspersed nuclear elements (SINEs) in large-scale genomic datasets. Bioinformatics. 2017 Mar 1;33(5):743–5.

117. Tumescheit C, Firth AE, Brown K. CIAlign - A highly customisable command line tool to clean, interpret and visualise multiple sequence alignments. bioRxiv; 2021. p. 2020.09.14.291484.

118. Chan PP, Lowe TM. tRNAscan-SE: searching for tRNA genes in genomic sequences. Methods Mol Biol. 2019;1962:1–14.

119. Han MJ, Xiong CL, Zhang HB, Zhang MQ, Zhang HH, Zhang Z. The diversification of PHIS transposon superfamily in eukaryotes. Mob DNA. 2015 Jun 24;6:12.

120. Bushnell B. BBMap: a fast, accurate, splice-aware aligner. Lawrence Berkeley National Lab. (LBNL), Berkeley, CA (United States); 2014 Mar. Report No.: LBNL-7065E.

121. Lerat E, Fablet M, Modolo L, Lopez-Maestre H, Vieira C. TEtools facilitates big data expression analysis of transposable elements and reveals an antagonism between their activity and that of piRNA genes. Nucleic Acids Res. 2017 Feb 28;45(4):e17.

122. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012 Mar 4;9(4):357–9.

123. Martelossi J. Transposable element characterization in molluscs. Figshare. 2023. https://doi.org/10.6084/m9.figshare.22188280.v1

124. Martelossi J. EvoTEs_BiV. Zenodo. 2023. 10.5281/zenodo.7944844

5. Chapter III

Widespread HCD-tRNA derived SINEs in bivalves relies on multiple LINE partners and accumulate in gene-related genomic regions

Jacopo Martelossi, Mariangela Iannello, Fabrizio Ghiselli, Andrea Luchetti

Note:

Results of this chapter will soon be submitted to a peer-reviewed journal. All supplementary files and their captions can be found at the end of the chapter.

Abstract

Short interspersed nuclear elements (SINEs) are non-autonomous non-LTR retrotransposons that are widespread across eukaryotes. They exist as lineage-specific, fast-evolving elements and as ubiquitous superfamilies characterized by highly conserved domains (HCD). Several of these superfamilies have been described in bivalves; however, their overall distribution and impact on host genome evolution are still unknown due to the extreme scarcity of transposon libraries for bivalves. In this study, we examined more than 40 bivalve genomes to uncover the distribution of HCD-tRNA-related SINEs, discover novel SINE-LINE partnerships, and understand their possible role in shaping bivalve genome evolution. We found that bivalve HCD SINEs have an ancient origin, and they can rely on at least four different LINE clades. Multiple species-specific SINEs were found to be highly similar between species separated by extremely long evolutionary timescales, reaching a maximum of ~400 million years in Mytilida. Studying their genomic distribution in a subset of five species, we observed different patterns of SINE enrichment in various genomic compartments as well as differences in the tendency of SINEs to form tandem-like and palindromic structures also within intronic sequences. Despite these differences, we observed that SINEs, especially older ones, tend to accumulate preferentially within or in close proximity to genes, consistent with a model of survival bias for less harmful, short non-coding transposons in euchromatic genomic regions.

Keywords: Mollusca, Transposable Element Annotation, Bivalvia, Genome evolution, Short Interspersed Nuclear Elements

Introduction

Bivalves (Class Bivalvia) are a rich and widespread clade of aquatic-only molluscs that diversified back to the early Cambrian, more than 500 Mya (Kocot et al., 2020). This class include multiple economically and ecological important species. For example, they have colonized freshwater environments (Graf, 2013) and deep-sea vents multiple times during their evolutionary history (Guo et al., 2023). They can be useful bioindicators for marine pollutants (Farrington et al., 2016) and they can represent biological models for the study of adaptations to climate change (Gazeau et al., 2013), innate immunity (Saco et al., 2023), sex determination (Nicolini et al., 2023), longevity (Blier et al., 2017, Iannello et al., 2023) and mitochondrial biology (Ghiselli et al., 2021). Moreover, they are characterized by peculiar genomic features hypothesized linked to transposable elements (TEs) activity, such as transmissible cancers (Metzger et al., 2016), high levels of hemizygosity (Calcino et al., 2021) and gene presence-absence variation (Gerdol et al., 2020).

Their important role as promising model system for addressing both general biology and human health questions, together with the increased cost-efficient accessibility of third-generation sequencing technologies, has led to a major increase in their genomic resources in recent years (Davison and Neiman, 2021). This has opened the possibility to explore, in a broader context, also usually neglected genomic components and their evolutionary dynamics such as TEs and other repetitive sequences constituting a high proportion of bivalves' genomes (**Chapter II**).

Repetitive DNA elements usually replicate in a selfish manner, independently from host's genome replication, with variety of effects on the host fitness ranging from neutral to deleterious. However, multiple cases of co-option in novel functions have been described in literature (Bourque et al., 2018). Furthermore, their evolutionary trajectory can be influenced by the dynamic of the host population, which in turn may be affected by the changes in TEs activity (Venner et al., 2009). Therefore, our understanding of TEs distribution and evolution across the tree of life represent an important step in a broader understanding of evolution of living forms.

Short Interspersed Nuclear Elements (SINEs) are a sub-class of non-autonomous, non-LTR retrotransposons that depend on the protein machinery of their autonomous counterpart LINEs (Long Interspersed Nuclear Elements) to reintegrate into the genome after transcription by RNA polymerase III (Pol III) (Wicker et al., 2007; Kramerov and Vassetzky, 2011; Vassetzky and Kramerov, 2013). Moreover, while many non-autonomous elements usually emerge from their autonomous counterparts though sequence decay or internal deletion, such as Miniature Inverted-repeat Transposable Elements (MITEs; Fattash et al., 2013) or Short Internally Deleted elements (SIDEs; Wang et al., 2019), SINEs emergence is only partially dependent from their LINE partners (Kramerov and Vassetzky, 2011). Their canonical structure comprises a head, a body and a tail region (Vassetzky and Kramerov, 2013). The head can originate from one of the three RNA type synthetized by the RNA Pol III, tRNAs, 5S rRNAs or 7SL RNAs and contain its promoter region. Even if elements originated from all three RNA types have been observed across a wide range of eukaryotes, tRNA-derived SINEs appear the most common ones (Kramerov and Vassetzky, 2011). The body, when present, contains a domain of unknown origin and function, which appear as element-specific (Vassetzky and Kramerov, 2013). However, in some instances, SINEs may carry exhibit bodies with highly conserved domains (HCD) across distinct SINE lineages and hosted by distantly related species; although the possible role of HCDs is still unclear, they have been useful for classifying SINEs at superfamily level (Luchetti and Mantovani, 2013; Luchetti and Mantovani, 2016; Nishihara et al., 2016). Finally, the 3' tail region serves as recognition for the LINE-derived reverse transcriptase (RT) and it may terminate with tandem repeats or an A-rich segment (Vassetzky and Kramerov, 2013). The SINE-LINE partnership can be specific if for RT recognition is required a LINE derived segment, usually originated from the 3‘ UTR LINE region, or aspecific when homology is not necessary (Kramerov and Vassetzky, 2011). The modular structure of SINEs suggested a characteristic evolutionary model called “mosaic evolution” under which different SINE lineages can exchange their modules through recombination (Ziętkiewicz and Labuda, 1996). This feature could allow their long-term persistence under a strict vertical inheritance evolutionary scenario in different genomic context, for example after the extinction of the original LINE partner (Luchetti and Mantovani, 2013; Luchetti and Mantovani, 2016).

A few analyses already identified some HCD SINEs in bivalves belonging to the superfamilies Core (Gilbert and Labuda, 1999; Nishihara et al., 2016), V (Ogiwara et al., 2002, Luchetti et al., 2016), Meta (Nishihara et al., 2016), Deu (Nishihara et al., 2006), and MD (Nishihara et al., 2016), where the latter are composed by a dimerization of Meta and Deu domains. Despite covering a low percentage of bivalve's genome, as observed in **Chapter II**, their emergence was traced back to their most recent common ancestor (Nishihara et al., 2016) similarly to what we hypothesized for the diversity of LINE clades in **Chapter II**. Moreover, for some of these elements Nishihara et al. (2016) and Matetovici et al. (2016) identified also their putative autonomous partners: CR1, L2 and Nimb. However, Though, these studies were limited by the limited number whole genomes available at that time and by the lack of comprehensive LINE reference library for bivalves.

Here, we leveraged the recent increase in bivalve genomic resources to comprehensively characterize HCD SINE diversity and richness across bivalve evolutionary history. The newly generated SINE library was used to screen for putative and previously unknown SINE-LINE partnerships, revealing that at least 4 different LINE lineages could act as RT donors in eight different SINE-LINE partnerships. Moreover, since SINEs can be important contributors to gene and genome evolution, we conducted a case study on a subset of 5 species to investigate the possible impact of SINE in genome evolution. Our findings showed gene-related genomic regions are enriched in SINEs, and particularly in old copies and that they can be organized in tandem-like and palindromic structures, potentially affecting gene epigenetic regulation.

Material and Methods

Genomic dataset for de-novo SINE prediction and manual curation

We selected 25 genomes from NCBI and GigaDB (Sup. Tab. 1) for the de-novo mining and manual curation of SINE elements. SINE candidates were mined from each assembly using RepeatModeler2 (Flynn et al., 2019) and SINE_Scan v1.1.1 (Mao and Wang, 2017). For each species, we merged SINE_Scan representative sequences with all TE consensus resulting from RepeatModeler2 and annotated as SINE by RepeatClassifier or by deepTE (Yan et al., 2020), which was run on "Unknown" elements to increase the chance of include as many SINEs candidate as possible. Candidates elements were then subjected to a "Blast-Extend-Extract" process (Goubert et al. 2022) blasting back each element against its source genome (Blastn v2.6.0: *qcov_hsp_perc* 70, *perc_identity* 70; Altschul et al., 1990), extracting the top 50 hits + 300bp at both ends with bedtools v2.26.0 (Quinlan and Hall, 2010) and aligned with MAFFT v7.475 (Katoh and Standley, 2013). From each alignment, we built a novel consensus sequence using the online Advance Consensus Maker tool (https://www.hiv.lanl.gov/content/sequence/CONSENSUS/consensus.html).

Boundaries of the elements were manually identified looking for the characteristic decay of the alignment towards terminal regions and the consensus sequences was curated implementing a majority rule approach, following the guidelines of Goubert et al. (2022). To confirm a candidate as a tRNA-related SINE we required: (a) the presence of a microsatellite or a poly-A region at the 3' end; (b) the presence of a tRNA- related region on the 5' end predicted by tRNAscan-SE (Lowe and Chan, 2016), through homology searches on the GtRNAdb (http://gtrnadb.ucsc.edu) or manually looking for RNA Pol III A and B boxes and (c) a length between 200 and 700 nucleotides. The presence of characteristic TSDs between 6 and 18 bps was manually checked, although it was not required to confirm a candidate as a SINE.

SINE-LINE partnerships

To identify partnerships of SINEs with their autonomous LINE counterparts, we queried all confirmed SINEs (blastn: *word size* = 7, *gap opening penality* = 2, *gap extension penality* = 2, *Match score* = 2, *Mismatch score* = -3, *evalue* = 0.01) against the bivalve specific library of LINE elements produced in **Chapter II**. All positive hits were manually checked to confirm that the homolog region fall at the 3' tail of the SINEs and within the 3' UTR of the LINE element.

Co-evolutionary dynamics of SINEs and their LINE counterparts were studied in the genomes in which we found evidence of SINE-LINE partnerships. For this purpose, we first selected all assemblies for which we identified homology between a species-specific SINE tail region and any LINE 3' UTR region. We then attempted to build a species-specific representative sequence of the LINE counterpart by blasting the original LINE element against the genome with decreasing thresholds in terms of identity and required alignment length, thereby creating a novel consensus sequence. When no homology was identified with a blastn search, we performed more sensitive tblastn searches (e-value 1e-05) of the amino acid translation of ORF2. Species-specific LINE consensus sequences were then checked for conservation of the homologous region between the species-specific SINE tail region and the LINE, as previously described. For confirmed partnerships, we used all species-specific SINE-LINE partner pairs as custom libraries for RepeatMasker in sensitive mode against the source assembly. TE landscapes, describing the divergence of each TE copy from its consensus sequence in terms of percentage of Kimura distance after CpG corrections, were calculated using the calcdivFromAlign.perl script provided with the RepeatMasker installation. Correlated activity between SINE-LINE partners was further tested for each species with Spearman's rank correlation tests between accumulation profiles (i.e number of base pairs occupied in each bin of CpG corrected Kimura divergence) of the two elements.

Superfamily and family level classification of confirmed SINEs.

For HCD SINEs classification we follow the superfamily classification scheme of Nishihara et al., (2016) based on the presence of characteristic central domains previously identified in bivalve genomes (Meta, V, Deu, Core). For the superfamily classification we started provisionally annotating each element using the RepeatClassifier utility from the RepeatModeler package. Elements that should share the same central domain were aligned using MAFFT and we then manually checked for the presence of the characteristic domain.

All elements were merged into a multi-species SINE library together with 19 elements previously described and deposited in RepBase (Sup. Tab. 2). These elements were then clustered following the 80-80 rule (Wicker et al., 2007) using cd-hit-est v4.7 (-G 0 -c 0.8 -aS 0.8 -t 1; Fu et al., 2012). Clusters were further refined into families following the definition from SINE base, where a SINE family is described as '*a set of elements sharing the same modules in the same order, excluding the tail region*' (Vassetzky and Kramerov, 2013). Note that in SINE base definition the tail region represents only the poly-A or other microsatellites at the 3' end of the SINE, whereases here we define as tail the LINE-derived region + the poly-A/microsatellite, following (CIT). Therefore, we also required the same LINE clade as putative donor of the RT. To achieve this, we ensured that each cluster contains only elements with the same modules, when this criterion was not met the original cluster was split into different families.

Copy number estimation of tRNA-related SINEs across bivalve diversity

To obtain a broader estimation of the distribution of the four SINE superfamilies across bivalve diversity, we downloaded additional N genomes from NCBI (Sup. Tab. 1) and performed homologous searches with blastn (e-value 1e-05). To avoid crossmatch with tRNA donors and LINE homologous regions, we excluded hits shorter than 150bp (i.e., approximately shorter than the 50% of the entire SINE length). After this step, we merged overlapping hits resulting from different families of the same superfamily using bedtools *merge* and counted the number of occurrences of each superfamily in each genome. A maximum of 150 random copies belonging to the superfamilies V, Meta and CORE were extracted from each genome and aligned using MAFFT in auto mode. TrimAl v1.4 (Capella-Gutierrez et al., 2009) was used to remove gap positions (--gappyout mode) and spurious sequences from the alignment (-resoverlap 0.50 -seqoverlap 55). We inferred a Maximum Likelihood tree via FastTree v2.1.10 (Price et al., 2010) using a GTR + Gamma model.

Genomic occurrence of SINEs and prediction of tandem-like SINE structures

All HCD SINEs families were used as input library for RepeatMasker v4.1.0 (Tarailo-Graovac and Chen, 2009) in sensitive mode (-s) to study their genomic occurrence in five species with available gene annotation (five-species gene set). Specifically, for *C. gigas* (Ostreida), *M. californianus* (Mytilida) and *P. maximus* (Pectinida) the RefSeq gene annotation was downloaded from NCBI repository, while for *R. philippinarum* (Venerida) and *S. broughtonii* (Arcida) they were recovered from Xu et al., (2022) and Bai et al., (2019), respectively. We considered five different features: exons, introns, annotated UTRs, 2500 bp flanking the genes and all other intergenic sequences (thus excluding 2500bp genes' flanking regions). For each feature, we counted the number of intersections with SINE insertions with Bedtools *intersect*. Over- and under-representation of SINEs in each feature was tested by constructing - with Bedtools *shuffle* - null distributions from 1000 random reshuffling iterations of all annotated SINE insertions across the genome (excluding genomic gaps). At each iteration, the number of intersections between each feature and the random intervals were counted. The observed number of intersections of SINEs in each feature was then compared to the null expectation. To directly test the hypothesis of SINEs preferential accumulation in 2500 bp gene flanking regions compared to all other intergenic regions, we split both features into intervals with a window of 500 bp with Bedtools *windows* and selected 10,000 random intervals for 100 iterations. We then counted the number of overlaps with SINE annotations at each iteration as previously described. Results for intergenic and gene-flanking genomic regions were statistically compared using t-test. Taking advantage of the high-quality repeat annotation of *C. gigas*, which repeatome is almost completely characterized (**Chapter II**), we also studied the accumulation patterns of LINEs across the same genomic intervals. Briefly, all LINEs from *C. gigas* available in RepBase, as well as those identified in **Chapter II** were combined, redundancy reduced with cd-hit-est following the 80-80 rule and used to annotate the genome with RepeatMasker. Overlaps between LINE insertions and genomic features were counted and statistically tested as previously performed for SINEs. Finally, we additionally hypothesize that gene-related regions, here defined as UTRs + exons + introns + 2500 bp gene flanking regions, are characterized by older SINEs copies compared to intergenic ones. To test this, all gene-related genomic regions were merged, and we calculated, for both gene-related and intergenic genomic regions, the percentage of Jukes-Cantor (JC) distance of each SINE copy to its consensus as a proxy for the time of insertions. Distributions were then tested with t-test. The same analyses were also performed for LINE insertions in *C. gigas*.

The same five genomes were scanned to identify presence of tandem SINE arrays. For this purpose, we only keep high-scoring SINEs, i.e RepeatMasker annotated insertions with a score higher than 400 and with a length of at least 150bp. This was necessary to remove possible miss-annotations such as host tRNA genes. We consider tandem-like SINE structures when multiple elements coming from the same family were detected one after the other.

Collection of seed alignments for DFAM submission

Novel SINE family consensus sequences were used to build up seed alignments for DFAM submission (Storer et al., 2021). For this purpose, we used the generateSeedAlignments.pl script provided with RepeatModeler installation with the flags –taxon, specifying the species name as reported in NCBI taxonomy, and –assemblyID followed by the NCBI accession number of the assembly. Resulting Stockholm files will deposited on DFAM together with the submission of this chapter to an IF journal.

Results

An improved HCD tRNA-related SINE library for Bivalves

By combining RepeatModeler, SINE_Scan, and homology searches, we identified 201 SINEs across the 25 selected bivalve genomes for the initial screening of SINE candidates (Sup. Tab. 2). All confirmed elements exhibited signatures of tRNA-related origin based on tRNA prediction analyses, homology searches against the tRNA-db, and/or manual identification of putative A and B boxes, the typical RNA polIII promoter (Sup. Tab. 2, Fig. 1). The tail region of candidate SINEs was also checked for the presence of microsatellites, and we successfully identified characteristic TSDs with sizes ranging between 6-18 bp for 181 (90%) of these elements (Sup. Tab. 2). Comparative analyses using the domains described in Nishihara et al., (2016) allowed us to subdivide these elements into the five known HCD superfamilies: Meta, V, MD, Core, and Deu. Specifically, we classified 31 elements as Core, 16 as Deu, 34 as MD, 40 as Meta, and 53 as V. Additionally, we found 27 other SINEs without clear homology to the aforementioned domains which we simply classified as tRNA-related SINEs. Within the five known HCD superfamilies, we identified 10 putative different tRNA donors, which are also shared between different superfamilies (Sup. Tab. 2).

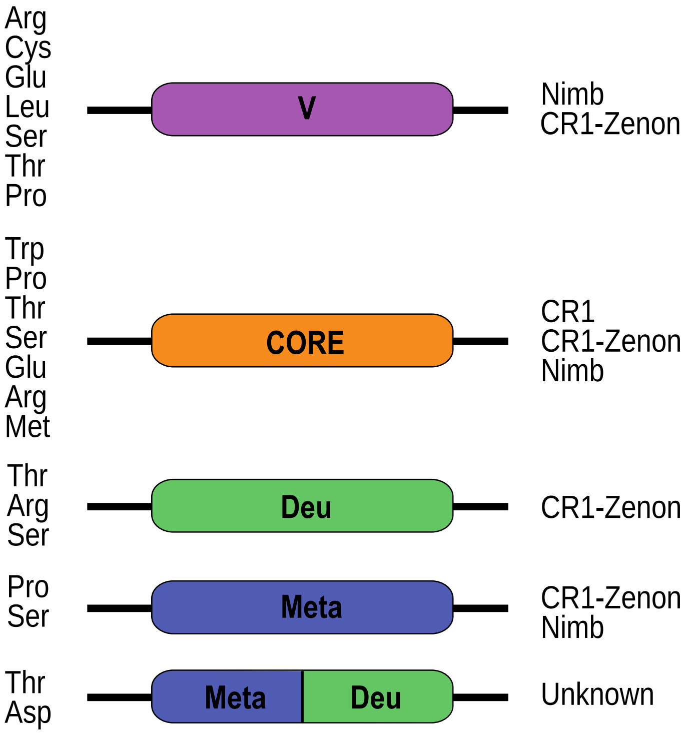


Fig. 1: Schematic representation of identified HCD tRNA-derived SINEs in bivalves. For each superfamily we reported the tRNA-related heads identified with tRNA-Scan SE and the putative LINE donor.

These 201 elements, along with 19 publicly available bivalve SINEs (Sup. Tab. 2) were clustered into 71 homology groups following the 80-80 rule. Five of these clusters were split in two different families due to different tRNA donors and an overall sequence identity just above the 80% cutoff (ranging between 80% and 84%) leading to a total of 76 distinct SINE families based on criteria of reciprocal homology and the order of SINE modules. Among these families, 17 are composed of unknown SINEs, 14 of Core SINEs, eight of Deu SINEs, six of MD SINEs, eight of Meta SINEs, and 23 of V SINEs (Sup. Tab. 2). Unknown families were discarded from following analyses. The presence of 22 families shared by multiple species (five Core, four Deu, one MD, five Meta, seven V) belonging to the same bivalve order (Sup. Tab. 2) underlies the long-term conservation of HCD SINE. Some notable examples are the families Bpla_SINE-1_Meta (tRNA head: Ser) shared between Mytilinae and Bathymodiolinae, (divergence time ~400, Lee et al., 2019), Tgra_SINE-48_V (tRNA head: Arg) shared between all Arcidae (divergence time ~177 million years; Sun and Gao, 2017) and Oden_SINE-1_CORE (tRNA head: Ser) shared between *O. denselamellosa*, *C. gigas* and *S. glomerata (*divergence time ~ 240 million years, Sun and Gao, 2017).

Bivalves HCD SINEs depend on at least 4 different LINE lineages

Using curated LINE libraries previously obtained from molluscs' genomes (**Chapter II**) together with all newly generated SINE consensus sequences, we searched for putative SINE-LINE partnerships. Our results highlights that at least four different LINE clades can match any of the SINE tails (Fig. 1; See Sup. Tab. 1 for all recognized homologies). Homologies between SINE and LINE 3' ends can be shared between different superfamilies and span between 35 bp and 61 bp with an identity ranging from 72% to 95% (Sup. Fig. 1). Nishihara et al., (2016) and Matetovici et al., (2016) found similarities between tail regions of V and CORE families with CR1 and L2 elements and between Meta SINEs and Nimb LINEs (I superfamily). Here we found that not CR1, but CR1-Zenon elements, a LINE clade closely related to CR1 and widespread in bivalves but apparently poor in other molluscs (**Chapter II**), are likely responsible for the retro-transcription/reintegration of V, CORE, Meta and Deu families, while Nimb LINEs may promote CORE, V and Meta replication. Interesting we also found one family from the Venerida *A. marissinica* genome with a tail region highly similar to 3' ends of CR1 elements (Amar_SINE-2_CORE).

To study the co-evolutionary dynamics between autonomous and non-autonomous elements, we first reconstructed, where possible, a full-length species-specific LINE counterpart. We managed to reconstruct 10 species-specific SINE-LINE partnerships across 9 different species. Repeat landscapes analyses revealed contrasting patterns between LINE and SINE activity (Sup. Fig. 2). Indeed, despite a positive and significant correlation between all accumulation profiles (all p-values < 0.05, Sup. Tab. 3) both visual inspection of repeat landscapes profiles as well as correlation analyses point to possible different evolutionary scenarios. Specifically, the accumulation patterns of BivaV-SINE2_CrGi#V, Medu_SINE-2#Meta and Sbro_SINE-2#Meta that resulted lowlier and less significantly correlated to their LINE counterparts (0.3 < Spearman's rho < 0.44; 0.002 < p-values < 0.03; Sup. Tab. 3) compared to other analysed partnerships. On the contrary the partnerships Cgig_SINE-10#CORE / Cgig-1_LINE#L2, BivaV-SINE1_MiYe#V / Myes-2_LINE#Nimb and Amar_SINE-2#CORE / Amar-1_LINE#CR1 show both strong correlations (Spearman's rho > 0.8) and overlapping activity profiles.

HCD tRNA-derived SINEs are widespread in bivalves and maintained activity after bivalve order diversification

To have a broader picture of SINEs HCD superfamilies distribution across bivalve diversity we added other 20 assemblies to our starting genomic dataset used for de-novo mining for a total of 46 analysed species representative of 11 different bivalve orders (Sup. Tab. 1): these genomes were used as database for homology searches using all previously confirmed SINEs as queries. It is to be noted that also in this analysis we found evidence of the possible long-term retention of certain SINE families despite requiring a hit length of at least 150 nucleotides. Indeed, we found blast hits across all genomes in which the putative shared families were de-novo mined (Sup. Tab. 4).

The Core superfamily was found across all members of the orders Ostreida, Cardiida - except for *Gari tellinella* – Unionida, as well as in two Adapedonta (*Solen grandis* and *Sinonovacula constricta*), two Arcida species (*Anadara kagoshimensis* and *Scapharca broughtonii*) and in three Venerida genomes (*Archivesica marissinica, Cyclina sinensis, Mactra quadrangularis*) (Fig. 2). No CORE element could be identified in Mytilida, Pectinida, Lucinida and Pteroidea representatives. However, it must be noted that while for Mytilida and Pectinida we include multiple species also in the de-novo mining, for Pteroidea and Lucinida we only analyzed one specie and only through homology searches.

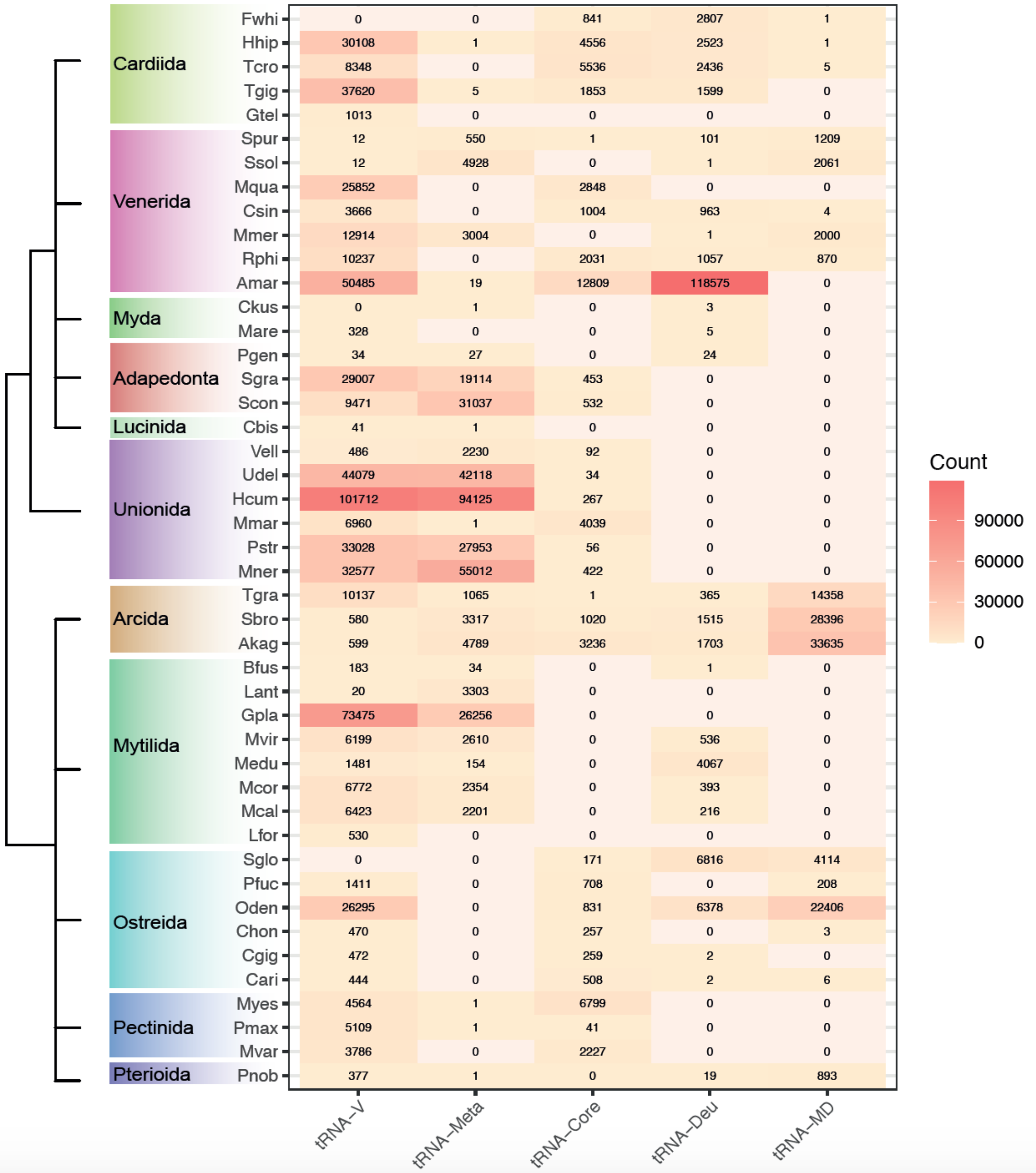


Fig. 2: Distribution of HCD superfamilies in bivalves**.** (A) Taxonomic distribution of the 5 known HCD SINE superfamilies obtained through blastn analyses of de-novo mined SINE families. Species name abbreviations refer to **Sup. Tab. 1**.

The Meta superfamily appears widespread across analyzed Acida, Mytilida, Unionida, Adapedonta and in the two Venerida *Saxidomus purpuratus* and *Spisula solida* (Fig. 2). On the other hand, the Deu superfamily has a patchier distribution compared to both Core and Meta, with elements identified in the majority of the Arcida, Cardiida and Venerida orders and in some Mytilida (*Mytilus coruscus, Mytilus edulis, Mytilus californianus and Mytilisepta virgata*) and Ostreida (*Ostrea denselamellosa*, *Saccostrea glomerata*) (Fig. 2). Interesting, the deep-sea symbiotic clam *Archivesica marissinica* hosts almost four times more Deu elements than the second richest species *A. kagoshimensis* (118,490 and 35,421 respectively) confirming an in increased of activity of specific TE groups in this lineage, possibly related to its colonization of hydrothermal vents (Ip et al., 2021; **Chapter II**).

The MD superfamily appears as the less represented one across bivalves, while the V superfamily resulted the most ubiquitous with elements identified across all species except for *S. glomerata*, *Cangeria kusceri* and *Fragum whitleyi*, confirming what was previously found by Nishihara et al., (2016) (Fig. 2). Interesting both Meta and V superfamilies are present in similar high copy number across four out of the six Unionida species here analyzed (from 27,000 to 101,712 copies) implying a possible expansion of these superfamilies in their most recent common ancestor.

Phylogenetic analyses of 150 random copies of the V and Meta superfamilies for each species (Fig. 3A-B) indicate that the great majority of elements are specific for a given bivalve order, as for Unionida, Mytilida, Arcida and Venerida, while a few other elements are shared by different bivalve orders. On the contrary, the phylogenetic pattern of the CORE superfamily is less clear as multiple groups of SINEs can be observed from the same bivalve order, and a random assemblage of other SINEs from Venerida genomes (Fig. 3C).

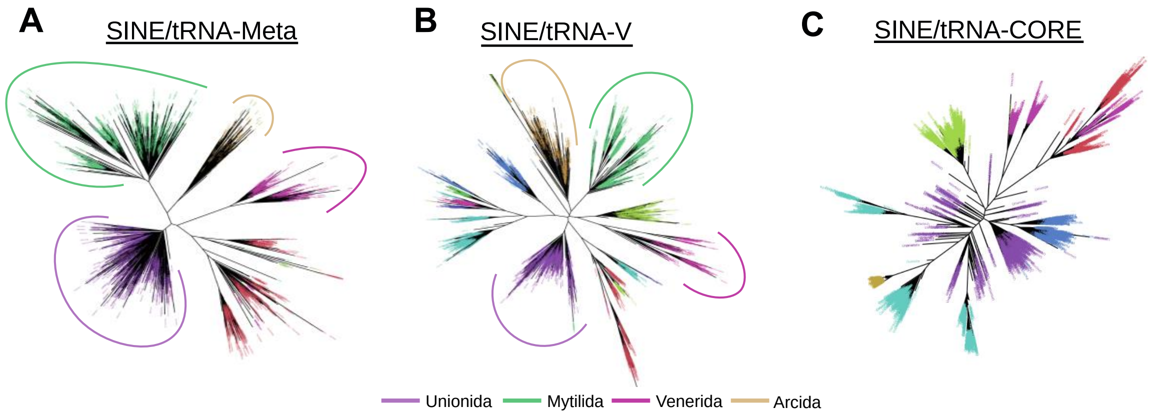


Fig. 3: Phylogenetic trees of 150 random copies extracted from each genome for the superfamilies V (**A**), Meta (**B**) and CORE (**C**). Colours of the tip labels represent bivalves order and reflect the colouring scheme of **Fig. 2.**

SINEs accumulate in gene-related genomic regions and can be organized in complex tandem-like structures

To detect potential preferences in the genomic occurrence of SINEs with respect to coding regions, we carried out a case study using five species with available gene annotation, testing the hypothesis of a higher accumulation of older SINEs in gene-related compared to intergenic genomic regions. HCD SINE insertions were found within 0.9%, 10%, 8.8%, and 4.1% of the genes in *C. gigas*, *M. californianus*, *P. maximus*, and *S. broughtonii*, respectively, but reached 41% in *R. philippinarum*. Compared to the null expectation, exons and UTRs consistently exhibited significantly fewer insertions, while gene flanking and intergenic genomic regions were generally enriched with SINEs (Table 1).

Tab. 1: Genomic distribution of observed and simulated SINEs insertions with respect to different genomic backgrounds. Gene flanking = 2500bp at both ends of genes; Intergenic = intergenic genomic regions after excluding gene flanking; SD = Standard deviation. Positive and negative Z-scores indicate more and less observed insertions compared to the null expectation, respectively.

Specie	Feature	Simulated	Observed	Z-score	P-value
C. gigas	Exons	1,260 ± 34	677	-17.1	> 0.001
	Introns	3,589 ± 49	3,425	-3.3	> 0.001
	UTRs	363 ± 19	115	-13.2	> 0.001
	Gene flaking	1,757 ± 35	2,310	15.9	> 0.001
	Intergenic	1,901 ± 36	2,001	2.8	> 0.001
M. californianus	Exons	6,959 ± 82	4,135	-34.3	> 0.001
	Introns	41,869 ± 185	36,382	-29.73	> 0.001
	UTRs	2,127 ± 45	775	-29.75	> 0.001
	Gene flaking	9,616 ± 90	12,671	33.9	> 0.001
	Intergenic	37,148 ± 149	40,557	22.9	> 0.001
P. maximus	Exons	9,289 ± 100	2,288	-70.2	> 0.001
	Introns	35,568 ± 154	29,870	-37	> 0.001
	UTRs	3,362 ± 59	1,317	-34.4	> 0.001
	Gene flaking	10,730 ± 91	11,677	10.5	> 0.001
	Intergenic	25,098 ± 129	31,465	49.4	> 0.001
R. philippinarum	Exons	7,252 ± 82	1,066	-75.6	> 0.001
	Introns	23,759 ± 147	36,981	89.9	> 0.001
	UTRs	934 ± 31	220	-23.1	> 0.001
	Gene flaking	14,992 ± 118	15,882	7.5	> 0.001
	Intergenic	91,470 ± 168	79,171	-70.2	> 0.001
S. broughtonii	Exons	1,176 ± 34	83	-32.5	> 0.001
	Introns	4,733 ± 59	5,948	20.7	> 0.001
	UTRs	99 ± 10	28	-6.98	> 0.001
	Gene flaking	1,919 ± 42	1,897	-0.53	0.7
	Intergenic	7,703 ± 59	6,745	-16.24	> 0.001

On the other hand, we observed a significant overrepresentation of insertions in the introns of *R. philippinarum* and *S. broughtonii* (Table 1), where we even found up to 61 and 145 insertions within a single gene, respectively (Table 1). These two species exhibited different accumulation patterns of SINEs within introns, with the former showing a low number of insertions in a high number of introns, while the latter showed a high number of insertions in a low number of introns (Sup. Fig. 3).

Moreover, for *C. gigas*, *M. californianus*, *R. philippinarum*, and *S. broughtonii*, gene flanking regions (2500bp flanking the gene) showed an enrichment of SINEs compared to intergenic ones (t-test; p-value < 0.01), whereas for *P. maximus*, we observed the opposite trend (Fig. 4A).

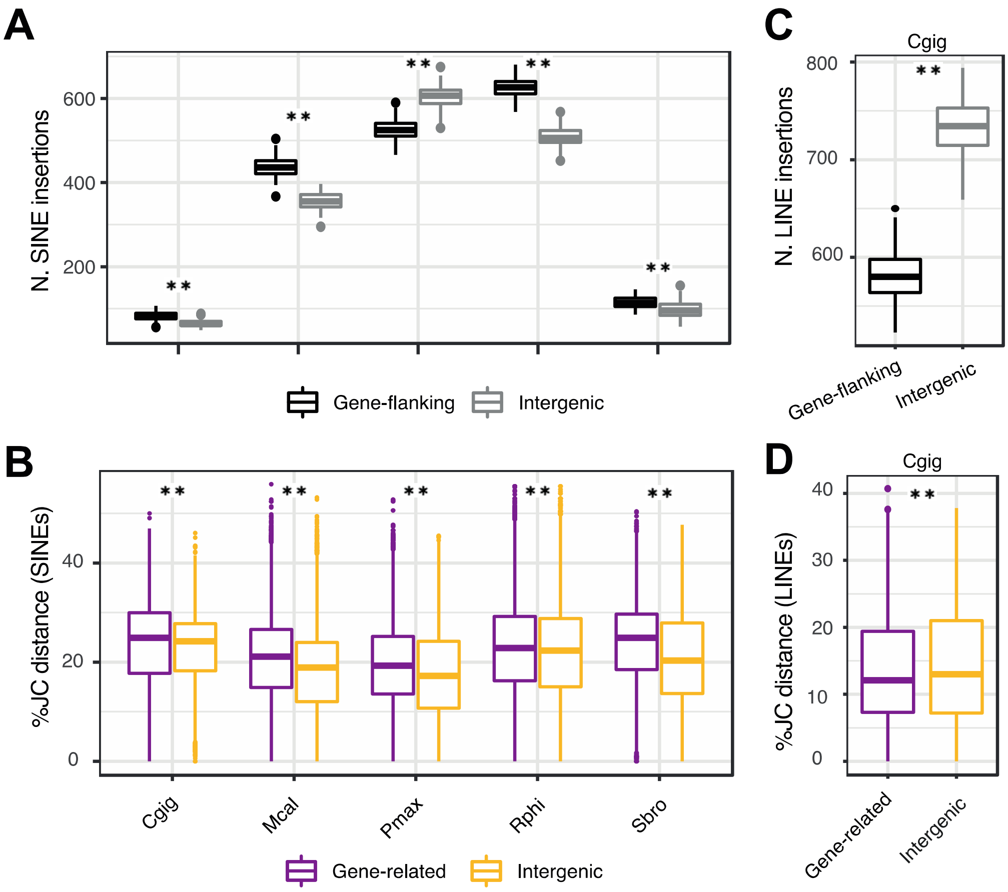


Fig. 4: Genomic occurrence of HCD SINEs. (**A**) Number of overlaps between SINE insertions with random gene flanking regions (2500 bp upstream and downstream the gene) and random intergenic genomic regions. (**B**) Jukes-Cantor (JC) distances of each SINE insertion from its consensus sequence as a proxy of the time of insertion. Gene-related = Insertions founded within genes (exons, introns, and UTRs) or in their 2500bp flanking regions. (**C**) and (**D**) are respectively specular to (**A**) and (**B**) but refer to LINE insertions in *C. gigas*. (**C**) Number of overlaps between LINE insertions with random gene flanking regions versus random intergenic genomic regions. (**D**) JC distance of LINE copies in gene-related versus intergenic genomic intervals. All comparisons are statistically significant (t-test, p-value < 0.01). Cgig = *C. gigas*, Mcal = *M. californianus*, Pmax = *P. maximus*, Rphi = *R. philippinarum*, Sbro = *S. broughtonii.*

Gene-related genomic regions (i.e., exons + introns + UTR+ gene-flanking regions) appear also characterized by older SINEs compared to intergenic ones, based on the Jukes-Cantor distance from consensus sequences across all species (Fig. 4B; t-test, all p-values < 0.001). Interestingly, we did not observe the same accumulation pattern when analysing the LINEs counterparts in *C. gigas*. Here intergenic genomic regions resulted significantly more affected by insertions compared to gene-flanking ones (Fig. 4C; t-test, p-value < 0.001; Sup. Tab. 5) and characterized by older TE insertions (Fig. 4D; t-test, p-value < 0.001).

The same five genomes were also scanned for tandem-like HCD SINEs, considering only high-scoring insertions (Fig. 5). All tandem arrays consist of two or three elements across all genomes, except for *S. broughtonii*, where we found 64 elements organized in tandem arrays of 4-15 units. Furthermore, while *C. gigas* hosts the smallest number of tandem-like SINE structures (three), in *S. broughtonii*, 3% of the high-scoring SINEs are organized in tandem arrays or palindromic structures, with 137 of them also incorporating one or multiple elements coming from a different family (Fig 5A).

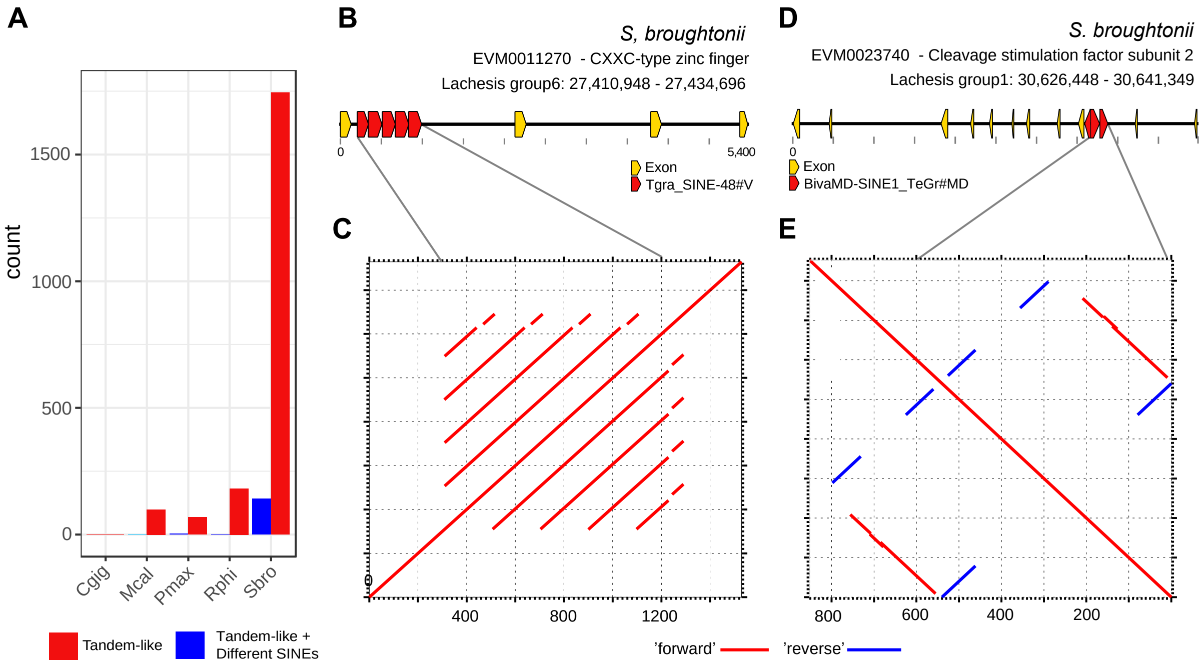


Fig. 5: Tandem-like SINE structure in bivalve genomes: **(A)** Number of tandem-like SINE structures identified in each of the five analysed bivalve genomes. Tandem-like + Different SINEs means that together with tandem SINEs structures coming from the same family, we also detected elements coming from different families. Cgig = *C. gigas*, Mcal = *M. californianus*, Pmax = *P. maximus*, Rphi = *R. philippinarum*, Sbro = *S. broughtonii.* (**B**) and (**C**) examples of respectively direct and inverted SINE repeats present in intronic sequences of the *S. broughtonii* genome.

The high number of tandem arrays in the blood clam *S. broughtonii*, even within intronic sequences (659 tandem arrays), could be an important contributor to the previously observed pattern of few introns impacted by a high number of insertions. We suggest that these SINE-rich introns could also drive the observed enrichment of SINE insertions in intronic sequences despite the low number of affected genes. Direct tandem arrays constitute most of tandemly repeated SINEs, while palindromes account for the 23% (403 structures) (Fig. 5D-E), of which 126 overlap with gene annotations and particularly within intronic sequences.

Discussion

Transposable elements (TEs) are among the most significant sources of genetic variation across the tree of life. The advancements in the sequencing field are leading to a greater appreciation of TEs in the context of understanding genome evolution, gene regulation, and species diversification. However, while genomic resources are rapidly expanding for most eukaryotic clades, accurate TE identification and annotations are still lacking in non-model species (Sproul et al., 2023), hindering our ability to comprehend their taxonomic distribution and effects on host biology. In this study, we took advantage of the increased number of bivalve genomes to comprehensively characterize HCD-containing SINEs across their diversity and to investigate their genomic distribution patterns. Our examination of 49 assemblies confirms that all known HCD-SINEs superfamilies have an ancient origin in bivalves, have been retained for a long evolutionary timescale (Nishihara et al., 2016) and can be derived from at least 10 different tRNA genes. Simultaneously, we observed important order-specific activity of the V and Meta superfamilies based on their phylogenetic clustering patterns.

Based on analyses of LINE-derived tail regions, we found that at least four different LINE lineages (CR1; CR1-Zenon; L2; Nimb) can act as RT donor to four different SINE superfamilies for a total of eight SINE-LINE relationships of which four were previously unknown (specifically the partnerships between SINE V and LINE Nimb, SINE CORE and LINE CR1, SINE CORE and Nimb and SINE Meta and CR1-Zenon). Therefore, we doubled the number of putative SINE-LINE partnerships compared to the previous studies of Nishihara et al., (2016) and Matetovici et al., (2016). Because of the strict relationship between SINEs and their LINE counterparts we might expect almost overlapping landscapes of activity in the case of partnerships between the two elements. However, in multiple instances more complex evolutionary scenarios emerge , possibly due to different competitive dynamics for the LINE-derived enzymatic machineries (Ray et al., 2019; Yang et al., 2019). Indeed, specific SINE lineages could be particularly efficient in parasitizing their LINE counterparts, preventing them from expanding. The hijacked LINE, in turn, might increase its replication rate only when the SINE partially loses its parasitizing capacity and, consequently, its replication rate. Another limitation in inferring the co-evolutionary dynamics of SINEs and LINEs using repeat landscape profiles is the inability to account for different deletion rates among various transposons. Some TEs might be more susceptible to genome elimination compared to others, resulting in their underrepresentation in older divergence bins. These phenomena could contribute to the different patterns that emerge from our analyses. Indeed, despite a consistently significant and positive correlation between accumulation profiles of SINEs and their LINE counterparts, the strength of the correlation varied significantly between species, and the repeat landscapes showed substantial overlaps only in a few instances.

Interestingly, we identified 22 HCD-SINE families with species-specific consensus sequences characterized the same highly similar modules between species separated by exceptionally long evolutionary time, up to ~400 million years in the case of the family Bpla_SINE-1_Meta, which was found in both Bathymodiolinae and Mytilinae (Lee et al., 2019). Our results represent therefore an exceptionally extreme case of what was previously observed also in grasses where SINE families were found to be far more conserved than LTR and TIR elements and retained for at least ~ 60 million years (Mao and Wang, 2017). The apparent long-term retention of TE families could also be explained by horizontal transposon transfer (HTT), which was already observed for SINEs in a few instances (Piskurek and Okada, 2007; Luchetti et al., 2016; Han et al., 2021). However, it is worth noting that we consistently identified highly similar species-specific consensus sequences among bivalves within the same order.

The ability to reconstruct consensus sequences shared between distantly related species could be favoured by the persistence of old insertions across the genome. In this context, it is interesting that, despite being underrepresented in exons and UTRs, SINE insertions tend to accumulate in gene-flanking regions (except for the Pectinida *P. maximus*) and, in the case of *R. philippinarum* and *S. broughtonii*, also within intronic sequences. The close association of HCD SINEs with gene bodies could increase the probability of their co-option as cis-regulatory elements, novel exons, or their contribution to the mRNA processing process, potentially enhancing the plasticity of tissue-specific transcripts, as recently observed in *Drosophila* (Coronado-Zamora and González, 2023). A close association between SINEs and genes was also observed in plants (Lenoir et al., 2001; Seibt et al., 2016; Mao and Wang, 2017), fishes (Luchetti et al., 2017), mammals (Buckley et al., 2017) and insect (Han et al., 2021) species. Open-chromatin genomic regions are known to be enriched in short and fragmented TEs (Ruggieri et al., 2022; Buckley et al., 2017). Furthermore, older SINE families were found to be more represented in euchromatic genomic regions compared to the younger ones both in grasses (Mao and Wang, 2017) and in the coelacanth (Luchetti et al., 2017), recapitulating our results. Indeed, we found that gene-related genomic regions (i.e., intragenic + 2,500bp gene flanking regions) are enriched in older HCD SINE insertions, in terms of distance from their consensus sequence, compared to intergenic ones. If we assume no insertion preference differences between old and young HCD SINEs, this pattern may suggest that gene-related regions could serve as safe ecological niches where short, non-coding transposons can survive. Coherently, we did not observe the same accumulation pattern when analysing LINE elements in the model bivalve *C. gigas*. One explanation is that short transposons insertions, like SINEs, could be favoured in proximity to genes by a combination of (1) reduced competition with longer, more harmful TEs and (2) lower efficiency of TE-purging processes (Mao and Wang, 2017; Devos et al., 2002). Indeed, deletions of transposable elements are mainly caused by ectopic DNA repair mechanisms, such as non-allelic homologous recombination and microhomology-mediated end joining (Hedges and Deininger, 2007; Morales et al., 2015). All these processes promote genome instability and may affect genomic flanking sequences, giving rise to complex and potentially harmful variants if a gene or a gene-interacting region is involved (Balachandran et al., 2022).

Methylation of SINE-derived direct repeats has been linked to the epigenetic regulation of downstream genes in *Arabidopsis thaliana* (Kinoshita et al., 2007), and double-stranded hairpin structures in the mRNA derived from palindromic structures, resulting from alternating orientations of SINE insertions, might serve as substrates for DICER enzymes (Seibt et al., 2016). We found that HCD SINEs in bivalves can be organized in such tandem-like and palindromic structures also within gene bodies, with an increased tendency in the Arcida *S. broughtonii*. In this species, approximately 3% of its HCD SINEs are organized in a similar manner. For comparison, in the potato genome, about 2% of the SINEs are included in tandem-like arrays (Seibt et al., 2016). The high number of SINE direct repeats identified in *S. broughtonii* raises interesting hypotheses about their potential origin and genome evolutionary dynamics of this species. Indeed, one possible outcome of unequal homologous recombination between target site duplications (TSDs) is the formation and the expansion of SINE tandem arrays (Lee et al., 2015). The high number of such structures in *S. broughtonii* could therefore implies higher recombination rates in this species compared to other analysed bivalves.

Conclusion

Here we perform for the first time a wide characterization of tRNA-related HCD SINEs in bivalves looking at their distribution, LINE partnerships and genomic occurrence. Thanks to a novel manually curated SINE library we found that bivalves HCD SINEs could derived from at least 10 different tRNAs and depend on at least four different LINE lineages. Some of these families are apparently shared between distantly related species underlying the possible long-term retention of highly similar HCD SINE linages characterized by the same tRNA-related head, central domain and LINE-derived tail. Genomic occurrence analyses across five different bivalve species highlighted their potential different effects in genome evolution. Indeed, different species show overrepresentation of SINE insertions across different genomic compartments as well as different tendencies to form tandem-like and palindromic structures which could be present in intronic sequences. Despite these differences, we found a consistent trend of accumulation of old SINEs in close proximity to genes, as previously observed in plants and other metazoan. This result suggest that evolutionary dynamics of SINEs might partially follow a common evolutionary route across eukaryotes in which euchromatic genomic regions serve as safe niches for their survival. Overall, this study represents a step forward in a broader understanding of the transposable elements' evolutionary dynamics in a highly overlooked but economically important taxonomic group like bivalves and open interesting questions about the possible role of SINEs in bivalve biology and evolution.

Availability of Supplementary Figures and Tables

All supplementary tables and figures are included as supplementary materials at the end of the chapter.

Funding

This work was supported by the Canziani bequest funded to F.G. and A.L., and the ‘Ricerca Fondamentale Orientata’ (RFO) funding from the University of Bologna to F.G. and A.L.

References

- Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J., 1990. Basic local alignment search tool. Journal of Molecular Biology 215, 403–410. https://doi.org/10.1016/S0022-2836(05)80360-2
- Bai, C.-M., Xin, L.-S., Rosani, U., Wu, B., Wang, Q.-C., Duan, X.-K., Liu, Z.-H., Wang, C.-M., 2019. Chromosomal-level assembly of the blood clam, Scapharca (Anadara) broughtonii, using long sequence reads and Hi-C. GigaScience 8, giz067. https://doi.org/10.1093/gigascience/giz067
- Balachandran, P., Walawalkar, I.A., Flores, J.I., Dayton, J.N., Audano, P.A., Beck, C.R., 2022. Transposable element-mediated rearrangements are prevalent in human genomes. Nat Commun 13, 7115. https://doi.org/10.1038/s41467-022-34810-8
- Blier, P.U., Abele, D., Munro, D., Degletagne, C., Rodriguez, E., Hagen, T., 2017. What modulates animal longevity? Fast and slow aging in bivalves as a model for the study of lifespan. Seminars in Cell & Developmental Biology, Science communication in the field of fundamental biomedical research 70, 130–140. https://doi.org/10.1016/j.semcdb.2017.07.046
- Bourque, G., Burns, K.H., Gehring, M., Gorbunova, V., Seluanov, A., Hammell, M., Imbeault, M., Izsvák, Z., Levin, H.L., Macfarlan, T.S., Mager, D.L., Feschotte, C., 2018. Ten things you should know about transposable elements. Genome Biology 19, 199. https://doi.org/10.1186/s13059-018-1577-z
- Buckley, R.M., Kortschak, R.D., Raison, J.M., Adelson, D.L., 2017. Similar Evolutionary Trajectories for Retrotransposon Accumulation in Mammals. Genome Biol Evol 9, 2336–2353. https://doi.org/10.1093/gbe/evx179
- Calcino, A.D., Kenny, N.J., Gerdol, M., 2021. Single individual structural variant detection uncovers widespread hemizygosity in molluscs. Philosophical Transactions of the Royal Society B: Biological Sciences 376, 20200153. https://doi.org/10.1098/rstb.2020.0153
- Capella-Gutierrez, S., Silla-Martinez, J.M., Gabaldon, T., 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics 25, 1972–1973. https://doi.org/10.1093/bioinformatics/btp348
- Coronado-Zamora, M., González, J., 2023. Transposons contribute to the functional diversification of the head, gut, and ovary transcriptomes across Drosophila natural strains. Genome Res. 33, 1541–1553. https://doi.org/10.1101/gr.277565.122

- Coronado-Zamora, M., González, J., 2023. Transposons contribute to the functional diversification of the head, gut, and ovary transcriptomes across Drosophila natural strains. Genome Res. 33, 1541–1553. https://doi.org/10.1101/gr.277565.122
- Davison, A., Neiman, M., 2021. Mobilizing molluscan models and genomes in biology. Philosophical Transactions of the Royal Society B: Biological Sciences 376, 20200163. https://doi.org/10.1098/rstb.2020.0163
- Emms, D.M., Kelly, S., 2019. OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biology 20, 238. https://doi.org/10.1186/s13059-019-1832-y
- Etchegaray, E., Naville, M., Volff, J.-N., Haftek-Terreau, Z., 2021. Transposable element-derived sequences in vertebrate development. Mobile DNA 12, 1. https://doi.org/10.1186/s13100-020-00229-5
- Farrington, J.W., Tripp, B.W., Tanabe, S., Subramanian, A., Sericano, J.L., Wade, T.L., Knap, A.H., 2016. Edward D. Goldberg's proposal of "the Mussel Watch": Reflections after 40years. Marine Pollution Bulletin 110, 501–510. https://doi.org/10.1016/j.marpolbul.2016.05.074
- Fattash, I., Rooke, R., Wong, A., Hui, C., Luu, T., Bhardwaj, P., Yang, G., 2013. Miniature inverted-repeat transposable elements: discovery, distribution, and activity. Genome 56, 475–486. https://doi.org/10.1139/gen-2012-0174
- Flynn, J.M., Hubley, R., Goubert, C., Rosen, J., Clark, A.G., Feschotte, C., Smit, A.F., 2019. RepeatModeler2: automated genomic discovery of transposable element families (preprint). Genomics. https://doi.org/10.1101/856591
- Fu, L., Niu, B., Zhu, Z., Wu, S., Li, W., 2012. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics 28, 3150–3152. https://doi.org/10.1093/bioinformatics/bts565
- Gerdol, M., Moreira, R., Cruz, F., Gómez-Garrido, J., Vlasova, A., Rosani, U., Venier, P., Naranjo-Ortiz, M.A., Murgarella, M., Greco, S., Balseiro, P., Corvelo, A., Frias, L., Gut, M., Gabaldón, T., Pallavicini, A., Canchaya, C., Novoa, B., Alioto, T.S., Posada, D., Figueras, A., 2020. Massive gene presence-absence variation shapes an open pan-genome in the Mediterranean mussel. Genome Biol 21, 275. https://doi.org/10.1186/s13059-020-02180-3
- Ghiselli, F., Iannello, M., Piccinini, G., Milani, L., 2021. Bivalve Molluscs as Model Systems for Studying Mitochondrial Biology. Integrative and Comparative Biology 61, 1699–1714. https://doi.org/10.1093/icb/icab057

- Gilbert, N., Labuda, D., 1999. CORE-SINEs: Eukaryotic short interspersed retroposing elements with common sequence motifs. Proceedings of the National Academy of Sciences 96, 2869–2874. https://doi.org/10.1073/pnas.96.6.2869
- Graf, D.L., 2013. Patterns of Freshwater Bivalve Global Diversity and the State of Phylogenetic Studies on the Unionoida, Sphaeriidae, and Cyrenidae *. malb 31, 135–153. https://doi.org/10.4003/006.031.0106
- Guo, Y., Meng, L., Wang, M., Zhong, Z., Li, D., Zhang, Y., Li, H., Zhang, H., Seim, I., Li, Y., Jiang, A., Ji, Q., Su, X., Chen, J., Fan, G., Li, C., Liu, S., 2023. Hologenome analysis reveals independent evolution to chemosymbiosis by deep-sea bivalves. BMC Biology 21, 51. https://doi.org/10.1186/s12915-023-01551-z
- Han, G., Zhang, N., Jiang, H., Meng, X., Qian, K., Zheng, Y., Xu, J., Wang, J., 2021. Diversity of short interspersed nuclear elements (SINEs) in lepidopteran insects and evidence of horizontal SINE transfer between baculovirus and lepidopteran hosts. BMC Genomics 22, 226. https://doi.org/10.1186/s12864-021-07543-z
- Hedges, D.J., Deininger, P.L., 2007. Inviting instability: Transposable elements, double-strand breaks, and the maintenance of genome integrity. Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis, Dedicated in memory of Dr. Tony Carrano 616, 46–59. https://doi.org/10.1016/j.mrfmmm.2006.11.021
- Ip, J.C.-H., Xu, T., Sun, J., Li, R., Chen, C., Lan, Y., Han, Z., Zhang, H., Wei, J., Wang, H., Tao, J., Cai, Z., Qian, P.-Y., Qiu, J.-W., 2021. Host–Endosymbiont Genome Integration in a Deep-Sea Chemosymbiotic Clam. Molecular Biology and Evolution 38, 502–518. https://doi.org/10.1093/molbev/msaa241
- Jin, L., Williamson, A., Banerjee, S., Philipp, I., Rape, M., 2008. Mechanism of Ubiquitin-Chain Formation by the human Anaphase-Promoting Complex. Cell 133, 653–665. https://doi.org/10.1016/j.cell.2008.04.012
- Katoh, K., Standley, D.M., 2013. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. Molecular Biology and Evolution 30, 772–780. https://doi.org/10.1093/molbev/mst010
- Kinoshita, Y., Saze, H., Kinoshita, T., Miura, A., Soppe, W.J.J., Koornneef, M., Kakutani, T., 2007. Control of FWA gene silencing in Arabidopsis thaliana by SINE-related direct repeats. Plant J 49, 38–45. https://doi.org/10.1111/j.1365-313X.2006.02936.x

- Kocot, K.M., Poustka, A.J., Stöger, I., Halanych, K.M., Schrödl, M., 2020. New data from Monoplacophora and a carefully-curated dataset resolve molluscan relationships. Sci Rep 10, 101. https://doi.org/10.1038/s41598-019-56728-w
- Kramerov, D.A., Vassetzky, N.S., 2011. Origin and evolution of SINEs in eukaryotic genomes. Heredity 107, 487–495. https://doi.org/10.1038/hdy.2011.43
- Krull, M., Brosius, J., Schmitz, J., 2005. Alu-SINE Exonization: En Route to Protein-Coding Function. Molecular Biology and Evolution 22, 1702–1711. https://doi.org/10.1093/molbev/msi164
- Lee, W., Mun, S., Kang, K., Hennighausen, L., Han, K., 2015. Genome-wide target site triplication of Alu elements in the human genome. Gene 561, 283–291. https://doi.org/10.1016/j.gene.2015.02.052.
- Lee, Y., Kwak, H., Shin, J., Kim, S.-C., Kim, T., Park, J.-K., 2019. A mitochondrial genome phylogeny of Mytilidae (Bivalvia: Mytilida). Molecular Phylogenetics and Evolution 139, 106533. https://doi.org/10.1016/j.ympev.2019.106533
- Lenoir, A., Lavie, L., Prieto, J.-L., Goubely, C., Cote, J.-C., Pélissier, T., Deragon, J.-M., 2001. The Evolutionary Origin and Genomic Organization of SINEs in Arabidopsis thaliana. Molecular Biology and Evolution 18, 2315–2322. https://doi.org/10.1093/oxfordjournals.molbev.a003778
- Li, H., 2018. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics 34, 3094–3100. https://doi.org/10.1093/bioinformatics/bty191
- Li, H., 2023. Protein-to-genome alignment with miniprot. Bioinformatics 39, btad014. https://doi.org/10.1093/bioinformatics/btad014
- Luchetti, A., Mantovani, B., 2013. Conserved domains and SINE diversity during animal evolution. Genomics 102, 296–300. https://doi.org/10.1016/j.ygeno.2013.08.005
- Luchetti, A., Mantovani, B., 2016. Rare horizontal transmission does not hide long-term inheritance of SINE highly conserved domains in the metazoan evolution. Current Zoology 62, 667–674. https://doi.org/10.1093/cz/zow095
- Luchetti, A., Plazzi, F., Mantovani, B., 2017. Evolution of Two Short Interspersed Elements in *Callorhinchus milii* (Chondrichthyes, Holocephali) and Related Elements in Sharks and the Coelacanth. *Genome Biology and Evolution*, 9(6): 1406–1417. https://doi.org/10.1093/gbe/evx094
- Luchetti, A., Šatović, E., Mantovani, B., Plohl, M., 2016. RUDI, a short interspersed element of the V-SINE superfamily widespread in molluscan genomes. Mol Genet Genomics 291, 1419–1429. https://doi.org/10.1007/s00438-016-1194-z

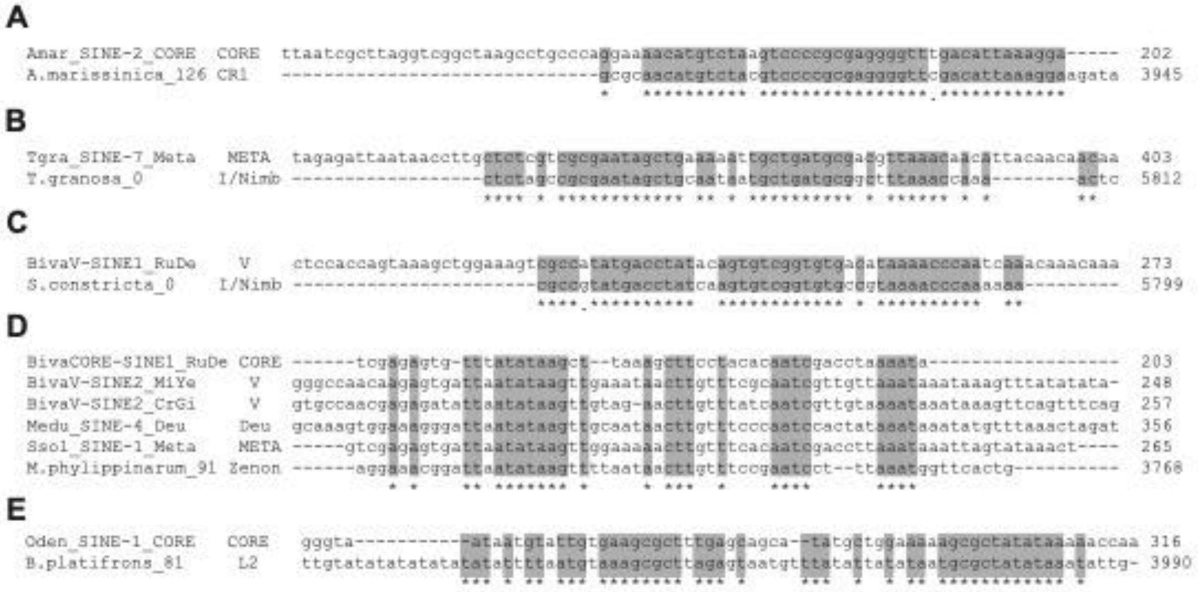
- Mao, H., Wang, H., 2017. Distribution, Diversity, and Long-Term Retention of Grass Short Interspersed Nuclear Elements (SINEs). Genome Biology and Evolution 9, 2048–2056. https://doi.org/10.1093/gbe/evx145
- Mao, H., Wang, H., 2017. SINE_scan: an efficient tool to discover short interspersed nuclear elements (SINEs) in large-scale genomic datasets. Bioinformatics 33, 743–745. https://doi.org/10.1093/bioinformatics/btw718
- Matetovici, I., Sajgo, S., Ianc, B., Ochis, C., Bulzu, P., Popescu, O., Damert, A., 2016. Mobile Element Evolution Playing Jigsaw—SINEs in Gastropod and Bivalve Mollusks. Genome Biol Evol 8, 253–270. https://doi.org/10.1093/gbe/evv257
- Metzger, M.J., Villalba, A., Carballal, M.J., Iglesias, D., Sherry, J., Reinisch, C., Muttray, A.F., Baldwin, S.A., Goff, S.P., 2016. Widespread transmission of independent cancer lineages within multiple bivalve species. Nature 534, 705–709. https://doi.org/10.1038/nature18599
- Morales, M.E., White, T.B., Streva, V.A., DeFreece, C.B., Hedges, D.J., Deininger, P.L., 2015. The Contribution of Alu Elements to Mutagenic DNA Double-Strand Break Repair. PLOS Genetics 11, e1005016. https://doi.org/10.1371/journal.pgen.1005016
- Nicolini, F., Ghiselli, F., Luchetti, A., Milani, L., 2023. Bivalves as Emerging Model Systems to Study the Mechanisms and Evolution of Sex Determination: A Genomic Point of View. Genome Biology and Evolution 15, evad181. https://doi.org/10.1093/gbe/evad181
- Nishihara, H., Plazzi, F., Passamonti, M., Okada, N., 2016. MetaSINEs: Broad Distribution of a Novel SINE Superfamily in Animals. Genome Biology and Evolution 8, 528–539. https://doi.org/10.1093/gbe/evw029
- Nishihara, H., Smit, A.F.A., Okada, N., 2006. Functional noncoding sequences derived from SINEs in the mammalian genome. Genome Res. 16, 864–874. https://doi.org/10.1101/gr.5255506
- Ogiwara, I., Miya, M., Ohshima, K., Okada, N., 2002. V-SINEs: A New Superfamily of Vertebrate SINEs That Are Widespread in Vertebrate Genomes and Retain a Strongly Conserved Segment within Each Repetitive Unit. Genome Res. 12, 316–324. https://doi.org/10.1101/gr.212302
- Price, M.N., Dehal, P.S., Arkin, A.P., 2010. FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments. PLOS ONE 5, e9490. https://doi.org/10.1371/journal.pone.0009490

- Quinlan, A.R., Hall, I.M., 2010. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26, 841–842. https://doi.org/10.1093/bioinformatics/btq033
- Ruggieri, A.A., Livraghi, L., Lewis, J.J., Evans, E., Cicconardi, F., Hebberecht, L., Ortiz-Ruiz, Y., Montgomery, S.H., Ghezzi, A., Rodriguez-Martinez, J.A., Jiggins, C.D., McMillan, W.O., Counterman, B.A., Papa, R., Belleghem, S.M.V., 2022. A butterfly pan-genome reveals that a large amount of structural variation underlies the evolution of chromatin accessibility. Genome Res. 32, 1862–1875. https://doi.org/10.1101/gr.276839.122
- Saco, A., Novoa, B., Greco, S., Gerdol, M., Figueras, A., 2023. Bivalves Present the Largest and Most Diversified Repertoire of Toll-Like Receptors in the Animal Kingdom, Suggesting Broad-Spectrum Pathogen Recognition in Marine Waters. Molecular Biology and Evolution 40, msad133. https://doi.org/10.1093/molbev/msad133
- Šatović Vukšić, E., Plohl, M., 2021. Exploring Satellite DNAs: Specificities of Bivalve Mollusks Genomes, in: Ugarković, Đ. (Ed.), Satellite DNAs in Physiology and Evolution, Progress in Molecular and Subcellular Biology. Springer International Publishing, Cham, pp. 57–83. https://doi.org/10.1007/978-3-030-74889-0_3
- Seibt, K.M., Wenke, T., Muders, K., Truberg, B., Schmidt, T., 2016. Short interspersed nuclear elements (SINEs) are abundant in Solanaceae and have a family-specific impact on gene structure and genome organization. The Plant Journal 86, 268–285. https://doi.org/10.1111/tpj.13170
- Simonti, C.N., Pavličev, M., Capra, J.A., 2017. Transposable Element Exaptation into Regulatory Regions Is Rare, Influenced by Evolutionary Age, and Subject to Pleiotropic Constraints. Molecular Biology and Evolution 34, 2856–2869. https://doi.org/10.1093/molbev/msx219
- Sorek, R., 2007. The birth of new exons: Mechanisms and evolutionary consequences. RNA 13, 1603–1608. https://doi.org/10.1261/rna.682507
- Sproul, J., Hotaling, S., Heckenhauer, J., Powell, A., Marshall, D., Larracuente, A.M., Kelley, J., Pauls, S.U., Frandsen, P.B., 2023. 600+ insect genomes reveal repetitive element dynamics and highlight biodiversity-scale repeat annotation challenges. Genome Res gr.277387.122. https://doi.org/10.1101/gr.277387.122

- Storer, J., Hubley, R., Rosen, J., Wheeler, T.J., Smit, A.F., 2021. The Dfam community resource of transposable element families, sequence models, and genome annotations. Mobile DNA 12, 2. https://doi.org/10.1186/s13100-020-00230-y
- Sun, W., Gao, L., 2017. Phylogeny and comparative genomic analysis of Pteriomorphia (Mollusca: Bivalvia) based on complete mitochondrial genomes. Marine Biology Research 13, 255–268. https://doi.org/10.1080/17451000.2016.1257810
- Tarailo-Graovac, M., Chen, N., 2009. Using RepeatMasker to Identify Repetitive Elements in Genomic Sequences. Current Protocols in Bioinformatics 25, 4.10.1-4.10.14. https://doi.org/10.1002/0471250953.bi0410s25
- Vassetzky, N.S., Kramerov, D.A., 2013. SINEBase: a database and tool for SINE analysis. Nucleic Acids Res 41, D83–D89. https://doi.org/10.1093/nar/gks1263
- Venner, S., Feschotte, C., Biémont, C., 2009. Dynamics of transposable elements: towards a community ecology of the genome. Trends in Genetics 25, 317–323. https://doi.org/10.1016/j.tig.2009.05.003
- Wang PL, Luchetti A, Alberto Ruggieri A, Xiong XM, Xu MR, Zhang XG, Zhang HH. Successful Invasions of Short Internally Deleted Elements (SIDEs) and Its Partner CR1 in Lepidoptera Insects. Genome Biol Evol. 2019 Sep 1;11(9):2505-2516. doi: 10.1093/gbe/evz174.
- Wicker, T., Sabot, F., Hua-Van, A., Bennetzen, J.L., Capy, P., Chalhoub, B., Flavell, A., Leroy, P., Morgante, M., Panaud, O., Paux, E., SanMiguel, P., Schulman, A.H., 2007. A unified classification system for eukaryotic transposable elements. Nat Rev Genet 8, 973–982. https://doi.org/10.1038/nrg2165
- Xu, R., Martelossi, J., Smits, M., Iannello, M., Peruzza, L., Babbucci, M., Milan, M., Dunham, J.P., Breton, S., Milani, L., Nuzhdin, S.V., Bargelloni, L., Passamonti, M., Ghiselli, F., 2022. Multi-tissue RNA-Seq Analysis and Long-read-based Genome Assembly Reveal Complex Sex-specific Gene Regulation and Molecular Evolution in the Manila Clam. Genome Biology and Evolution 14, evac171. https://doi.org/10.1093/gbe/evac171
- Yan, H., Bombarely, A., Li, S., 2020. DeepTE: a computational method for de novo classification of transposons with convolutional neural network. Bioinformatics 36, 4269–4275. https://doi.org/10.1093/bioinformatics/btaa519
- Yang, L., Scott, L., Wichman, H.A., 2019. Tracing the history of LINE and SINE extinction in sigmodontine rodents. Mobile DNA 10, 22. https://doi.org/10.1186/s13100-019-0164-5

- Ziętkiewicz, E., Labuda, D., 1996. Mosaic evolution of rodent B1 elements. J Mol Evol 42, 66–72. https://doi.org/10.1007/BF00163213

Supplementary Figures



A

```
Amar_SINE-2_CORE    CORE   ttaatcgcttaggtcggctaagcctgcccaggaaaacatgtctaagtccccgcgaggggtttgacattaaagga----- 202
A.marissinica_126   CR1    ------------------------------gcgcaacatgtctacgtccccgcgaggggttcgacattaaaggaagata 3945
                                                         *   ********** ****************.************
```

B

```
Tgra_SINE-7_Meta    META   tagagattaataaccttgctctcgtcgcgaatagctgaaaaattgctgatgcgacgttaaacaacattacaacaacaa 403
T.granosa_0         I/Nimb ------------------ctctagccgcgaatagctgcaataatgctgatgcggctttaaaccaaa--------actc 5812
                                             **** * ************ ** * ********** * ****** * *        **
```

C

```
BivaV-SINE1_RuDe    V      ctccaccagtaaagctggaaagtcgccatatgacctatacagtgtcggtgtgacataaaacccaatcaaacaaacaaa 273
S.constricta_0      I/Nimb -----------------------cgccgtatgacctatcaagtgtcggtgtgccgtaaaacccaaaaaa--------- 5799
                                                  ****.**********  ************ * **********  **
```

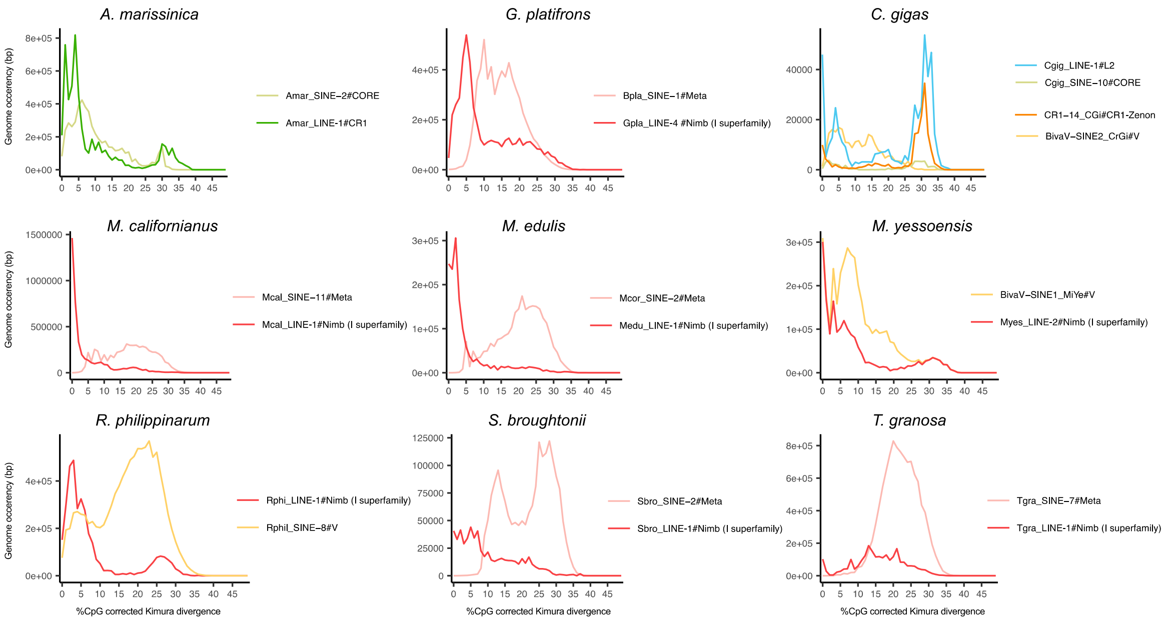
D

```
BivaCORE-SINE1_RuDe CORE   ------tcgagagtg-tttatataagct--taaagcttcctacacaatcgacctaaaata------------------ 203
BivaV-SINE2_MiYe    V      gggccaacaagagtgattaatataagttgaaataacttgtttcgcaatcgttgttaaataaataaagtttatatata- 248
BivaV-SINE2_CrGi    V      gtgccaacgagagatattaatataagttgtag-aacttgtttatcaatcgttgtaaaataaataaagttcagtttcag 257
Medu_SINE-4_Deu     Deu    gcaaagtggaaagggattaatataagttgcaataacttgtttcccaatccactataaataaatatgtttaaactagat 356
Ssol_SINE-1_Meta    META   -----gtcgagagtgattaatataagttggaaaaacttgtttcacaatcgaccttaaataaattagtataaact---- 265
M.phylippinarum_91  Zenon  ------aggaaacggattaatataagttttaataacttgtttccgaatcct--ttaaatggttcactg---------- 3768
                                    * *    ** ******* *     * ***  *    ****      ****
```

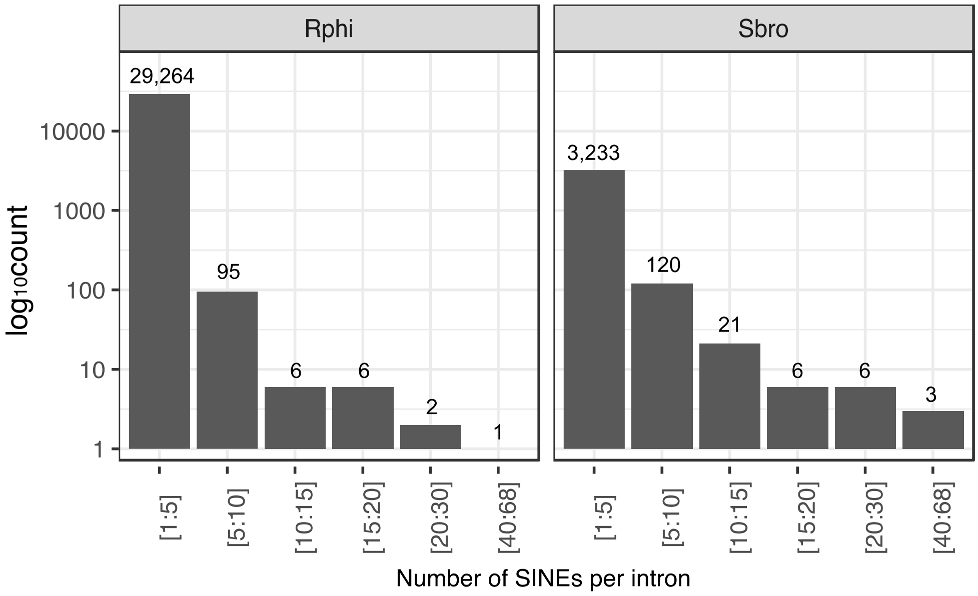
E

```
Oden_SINE-1_CORE    CORE   gggta-----------ataatgtattgtgaagcgctttgagcagca--tatgctggaaaagcgctatataaaaaccaa 316
B.platifrons_81     L2     ttgtatatatatatatatattttaatgtaaagcgcttagagtaatgttatattatataatgcgctatataaatattg- 3990
                                          *** * ** *** ******** *** *     ***  *  * ** ************ *
```

Sup. Fig. 1: SINE-LINE partnerships. Representative alignments between SINE-LINE homologues regions identified in this study. Identical nucleotides are marked in grey boxes with asterisks. All LINE families were already identified in **Chapter II**. **(A)** A.marissinica_126=LINE/CR1; **(B)** T.granosa_0=LINE/CR1-Zenon; **(C)** S.constricta_0=LINE/I; (**D**) M.phylippinarum_91=LINE/CR1-Zenon; (E) B.platrifrons_81=LINE/Nimb (I superfamily).



Sup. Fig. 2: Co-evolutionary dynamics between SINEs and their LINE counterparts. Repeat landscape profiles of species-specific SINE-LINEs partners. The plots represent the total number of base pairs (y axis) occupied in each bin of CpG corrected kimura divergence (x axis).



Sup. Fig. 3: Different SINE accumulation patterns in introns of *S. broughtonii* and *R. philippinarum*. Number of SINE insertions per intron in *R. philippinarum* (Rphi) and *S. broughtonii* (Sbro).

Supplementary Tables

Sup. Tab. 1: Species and relatively assembly accession numbers used for de-novo and homology-based mining of SINEs. Taxonomic informations were retrived from NCBI taxonomy.

Species	**Abbreviation**	**Accession number/Source**	**Taxonomy**	**SINE discovery**
Anadara kagoshimensis	Akag	GCA_021292105.1	Pteriomorpha/Arcida	De-novo
Archivesica marissinica	Amar	GCA_014843695.1	Imparidentia/Venerida	De-novo
Gigantidas platifrons	Gpla	GCA_002080005.1	Pteriomorpha/Mytilida	De-novo
Crassostrea ariakensis	Cari	GCA_020458035.1	Pteriomorpha/Ostreida	De-novo
Crassostrea gigas	Cgig	GCF_902806645.1	Pteriomorpha/Ostreida	De-novo
Crassostrea hongkongensis	Chon	GCA_015776775.1	Pteriomorpha/Ostreida	De-novo
Cyclina sinensis	Csin	GCA_012932295.1	Imparidentia/Venerida	De-novo
Limnoperna fortunei	Lfor	GCA_944474755.1	Pteriomorpha/Mytilida	Homology
Mytilus californianus	Mcal	GCF_021869535.1	Pteriomorpha/Mytilida	De-novo
Mytilus coruscus	Mcor	GCA_017311375.1	Pteriomorpha/Mytilida	De-novo
Mytilus edulis	Medu	GCA_019925275.1	Pteriomorpha/Mytilida	De-novo
Megalonaias nervosa	Mner	GCA_016617855.1	Paleoheterodonta/Unionida	De-novo
Mactra quadrangularis	Mqua	GCA_025267735.1	Imparidentia/Venerida	De-novo
Mytilisepta virgata	Mvir	GCA_028015205.1	Pteriomorpha/Mytilida	De-novo
Mizuhopecten yessoensis	Myes	GCF_002113885.1	Pteriomorpha/Pectinida	De-novo
Ostrea denselamellosa	Oden	GCA_024699665.1	Pteriomorpha/Ostreida	De-novo
Pinctada fucata	Pfuc	GCA_028142955.1	Pteriomorpha/Ostreida	De-novo
Pecten maximus	Pmax	GCF_902652985.1	Pteriomorpha/Pectinida	De-novo
Potamilus streckersoni	Pstr	GCA_016746295.1	Paleoheterodonta/Unionida	Homology
Scapharca broughtonii	Sbro	GigaDB/ http://gigadb.org/dataset/100607	Pteriomorpha/Arcida	De-novo
Sinonovacula constricta	Scon	GCA_007844125.1	Imparidentia/Adapedonta	De-novo
Saccostrea glomerata	Sglo	GCA_003671525.1	Pteriomorpha/Ostreida	De-novo
Solen grandis	Sgra	GCA_021229015.1	Imparidentia/Adapedonta	De-novo
Spisula solida	Ssol	GCA_947247005.1	Imparidentia/ Venerida	De-novo

Ruditapes philippinarum	Rphil	GCA_026571515.1	mparidentia/ Venerida	De-novo
Tridacna crocea	Tcro	GCA_943736015.1	Imparidentia/Cardiida	De-novo
Tegillarca granosa	Tgra	GCA_013375625.1	Pteriomorpha/Arcida	De-novo
Hyriopsis cumingii	Hcum	GCA_028554795.1	Palaeoheterodonta/Unionida	Homology
Mya arenaria	Mare	GCF_026914265.1	Imparidentia/Myda	Homology
Crassostrea angulata	Cang	GCF_025612915.1	Pteriomorpha/Ostreida	Homology
Panopea generosa	Pgen	GCA_029582155.1	Imparidentia/Adapedonta	Homology
Congeria kusceri	Ckus	GCA_027627225.1	Imparidentia/Myda	Homology
Mimachlamys varia	Mvar	GCA_947623455.1	Pteriomorpha/Pectinida	Homology
Gari tellinella	Gtel	GCA_922989275.2	Imparidentia/Cardiida	Homology
Tridacna gigas	Tgig	GCA_945859785.2	Imparidentia/Cardiida	Homology
Hippopus hippopus	Hhip	GCA_946811185.1	Imparidentia/Cardiida	Homology
Fragum whitleyi	Fwhi	GCA_948146395.1	Imparidentia/Cardiida	Homology
Conchocele bisecta	Cbis	GCA_029237695.1	Imparidentia/Lucinida	Homology
Saxidomus purpurata	Spur	GCA_022818135.1	Imparidentia/Venerida	Homology
Pinna nobilis	Pnob	GCA_016161895.1	Pteriomorphia/Pterioida	Homology
Unio delphinus	Udel	GCA_029339505.1	Palaeoheterodonta/Unionida	Homology
Venustaconcha ellipsiformis	Vell	GCA_003401595.1	Palaeoheterodonta/Unionida	Homology
Lithophaga antillarum	Lant	GCA_028566495.1	Pteriomorpha/Mytilida	Homology
Botula fusca	Bfus	GCA_028566455.1	Pteriomorpha/Mytilida	Homology
Margaritifera margaritifera	Mmar	GCA_015947965.1	Palaeoheterodonta/Unionida	Homology

Sup. Tab. 2: Details about confirmed SINE sequences mined with RepeatModeler2 and SINE_Scan. Asterisks in the tRNA column means that the tRNA was predicted through homology searches against GtRNAdb (http://gtrnadb.ucsc.edu), Undet means that A and B boxes were manually verified while in all other istances the tRNA donor was predicted with tRNAScan-SE. For each element we reported the species from which it was mined following the abbreviations in Sup. Tab. 1.

Name	**TSD**	**Length**	**Classification**	**Satellite**	**tRNA**	**LINE**	**Cluster**	**Family**
Akag_SINE-1	YES	354	SINE/tRNA-Core	AGATA	Trp	NA	47	Akag_SINE-1_CORE#SINE/tRNA-Core
Amar_SINE-2	NO	203	SINE/tRNA-Core	ACAT	Pro	LINE_A.marissinica_126_cons#LINE/CR1	71	Amar_SINE-2_CORE#SINE/tRNA-Core
Amar_SINE-3	YES	231	SINE/Unknown	AAACT	Leu	LINE_S.grandis_86_cons#LINE/CR1-Zenon	68	Amar_SINE-3_Unknown#SINE/Unknown
Amar_SINE-1	YES	285	SINE/tRNA-Deu	CA	Thr	NA	20	Amar_SINE-4_Deu#SINE/tRNA-Deu
Amar_SINE-4	YES	292	SINE/tRNA-Deu	CA	Thr	NA	20	
Csin_SINE-7	YES	288	SINE/tRNA-Deu	ATAG	Thr	NA	20	
Amar_SINE-5	NO	253	SINE/tRNA-V	AAACT	Undet	LINE_B.platifrons_29_cons#LINE/CR1-Zenon	64	Amar_SINE-5_V#SINE/tRNA-V
Myes_SINE-1	NO	330	SINE/tRNA-Core	ATT	Ala*	NA	17	BivaCORE-SINE2_MiYe#SINE/tRNA-Core
Pmax_SINE-3	NO	337	SINE/tRNA-Core	ATT	Ala*	NA	17	
Akag_SINE-2	YES	335	SINE/tRNA-MD	ACTC	Thr	NA	0	BivaMD-SINE1_TeGr#SINE/tRNA-Deu
Sbro_SINE-9	YES	335	SINE/tRNA-MD	ACTC	Thr	NA	0	
Sbro_SINE-10	YES	335	SINE/tRNA-MD	ACTC	Thr	NA	0	
Sbro_SINE-15	YES	336	SINE/tRNA-MD	AATC	Thr	NA	0	
Sbro_SINE-16	YES	322	SINE/tRNA-MD	AATC	Thr	NA	0	
Sbro_SINE-17	YES	336	SINE/tRNA-MD	AACTC	Thr	NA	0	
Sbro_SINE-19	YES	335	SINE/tRNA-MD	AACTC	Thr	NA	0	
Sbro_SINE-20	YES	340	SINE/tRNA-MD	AACTC	Thr	NA	0	
Sbro_SINE-22	YES	336	SINE/tRNA-MD	AACTC	Thr	NA	0	
Sbro_SINE-27	YES	334	SINE/tRNA-MD	ACTC	Thr	NA	0	

Sbro_SINE-31	YES	336	SINE/tRNA-MD	ACTC	Thr	NA	0	
Sbro_SINE-34	YES	339	SINE/tRNA-MD	AACTC	Thr	NA	0	
Sbro_SINE-35	YES	335	SINE/tRNA-MD	AATC	Thr	NA	0	
Sbro_SINE-37	YES	335	SINE/tRNA-MD	ACTC	Thr	NA	0	
Sbro_SINE-42	YES	329	SINE/tRNA-MD	AATC	Thr	NA	0	
Sbro_SINE-50	YES	336	SINE/tRNA-MD	AATC	Thr	NA	0	
Sbro_SINE-51	YES	336	SINE/tRNA-MD	AATC	Thr	NA	0	
Sbro_SINE-55	YES	340	SINE/tRNA-MD	AACTC	Thr	NA	0	
Sbro_SINE-76	YES	335	SINE/tRNA-MD	ACTC	Thr	NA	0	
Tgra_SINE-3	YES	335	SINE/tRNA-MD	AACTC	Thr	NA	0	
Tgra_SINE-9	YES	335	SINE/tRNA-MD	AATC	Thr	NA	0	
Tgra_SINE-17	YES	334	SINE/tRNA-MD	ACTC	Thr	NA	0	
Tgra_SINE-18	YES	335	SINE/tRNA-MD	CTCAA	Thr	NA	0	
Tgra_SINE-25	YES	331	SINE/tRNA-MD	TTTAA	Undet	NA	0	
Tgra_SINE-47	YES	335	SINE/tRNA-MD	AATC	Thr	NA	0	
Mner_SINE-1	YES	309	SINE/tRNA-Meta	CA	Pro	NA	19	Mner_SINE-4_Meta#SINE/tRNA-Meta
Mner_SINE-4	YES	309	SINE/tRNA-Meta	CA	Pro	NA	19	
Bpla_SINE-6	YES	259	SINE/tRNA-V	AAACT	Undet	NA	29	BivaV-SINE1_BaAz#SINE/tRNA-V
Myes_SINE-2	YES	219	SINE/tRNA-V	AAACC	Ser*	LINE_C.farreri_1_cons#LINE/I	16	BivaV-SINE1_ChFa#SINE/tRNA-V
Pmax_SINE-10	YES	218	SINE/tRNA-V	AAACC	Ser	NA	16	
Mner_SINE-5	YES	240	SINE/tRNA-V	ACA	Ser	NA	32	BivaV-SINE1_HyCu#SINE/tRNA-V
Rphil_SINE-8	YES	266	SINE/tRNA-V	AAAAC	Sup	LINE_S.constricta_0_cons#LINE/I	79	BivaV-SINE1_RuDe#SINE/tRNA-V

Cgig_SINE-11	NO	253	SINE/tRNA-V	A	Ala*	LINE_B.platifrons_35_cons#LINE/CR1-Zenon	6	BivaV-SINE2_CrGi#SINE/tRNA-V
Cari_SINE-1	YES	251	SINE/tRNA-V	AGTTC	Ala*	LINE_B.platifrons_35_cons#LINE/CR1-Zenon	6	
Cgig_SINE-1	YES	255	SINE/tRNA-V	A	Ala*	LINE_B.platifrons_35_cons#LINE/CR1-Zenon	6	
Cgig_SINE-5	YES	252	SINE/tRNA-V	A	Ala*	LINE_B.platifrons_35_cons#LINE/CR1-Zenon	6	
Chon_SINE-1	YES	250	SINE/tRNA-V	AGTTC	Ala*	LINE_B.platifrons_35_cons#LINE/CR1-Zenon	6	
Chon_SINE-5	YES	250	SINE/tRNA-V	AGTTC	Ala*	LINE_B.platifrons_35_cons#LINE/CR1-Zenon	6	
Chon_SINE-7	YES	251	SINE/tRNA-V	AGTTC	Ala*	LINE_B.platifrons_35_cons#LINE/CR1-Zenon	6	
Pfuc_SINE-4	YES	256	SINE/tRNA-V	TTTAAA	Ser	NA	31	BivaV-SINE2_PiFu#SINE/tRNA-V
Ssol_SINE-3	YES	253	SINE/tRNA-V	AAAC	Thr	NA	27	BivaV-SINE3_SpSo#SINE/tRNA-V
Bpla_SINE-1	YES	357	SINE/tRNA-Meta	CACT	Ser	LINE_B.platifrons_4_cons#LINE/I	2	Bpla_SINE-1_Meta#SINE/tRNA-Meta
Bpla_SINE-2	YES	357	SINE/tRNA-Meta	CACT	Ser	LINE_B.platifrons_4_cons#LINE/I	2	
Bpla_SINE-3	YES	357	SINE/tRNA-Meta	CACT	Ser	LINE_B.platifrons_4_cons#LINE/I	2	
Bpla_SINE-4	YES	357	SINE/tRNA-Meta	CACT	Ser	LINE_B.platifrons_4_cons#LINE/I	2	
Mvir_SINE-1	YES	350	SINE/tRNA-Meta	ATCC	Ser	NA	2	
Mcal_SINE-19	YES	352	SINE/tRNA-Meta	ATC	Ser	LINE_B.platifrons_4_cons#LINE/I	2	
Mcal_SINE-23	YES	350	SINE/tRNA-Meta	ATC	Ser	LINE_B.platifrons_4_cons#LINE/I	2	
Mcal_SINE-25	YES	351	SINE/tRNA-Meta	ATC	Ser	LINE_B.platifrons_4_cons#LINE/I	2	
Mcal_SINE-7	YES	351	SINE/tRNA-Meta	ATCA	Ser	LINE_B.platifrons_4_cons#LINE/I	2	
Mcal_SINE-14	YES	351	SINE/tRNA-Meta	ATC	Ser	LINE_B.platifrons_4_cons#LINE/I	2	
Mcal_SINE-15	YES	353	SINE/tRNA-Meta	ATC	Ser	NA	2	
Mcor_SINE-2	YES	348	SINE/tRNA-Meta	ATC	Ser	LINE_B.platifrons_4_cons#LINE/I	2	
Cari_SINE-3	NO	296	SINE/tRNA-Core	AACTT	Thr	NA	10	Cari_SINE-3_CORE#SINE/tRNA-Core

Cgig_SINE-4	YES	296	SINE/tRNA-Core	AACTT	Thr	NA	10	
Chon_SINE-2	YES	292	SINE/tRNA-Core	AACTT	Thr	NA	10	
Chon_SINE-3	YES	294	SINE/tRNA-Core	AACTT	Thr	NA	10	
Chon_SINE-6	YES	291	SINE/tRNA-Core	AACTT	Thr	NA	10	
Cgig_SINE-6	NO	341	SINE/tRNA-Core	ACCTTT	Arg*	NA	18	Cgig_SINE-6_CORE#SINE/tRNA-Core
Cari_SINE-5	YES	339	SINE/tRNA-Core	ACCTTT	Arg*	NA	18	
Chon_SINE-4	YES	339	SINE/tRNA-Core	ACCTTT	Arg*	NA	18	
Csin_SINE-1	YES	280	SINE/tRNA-V	AAAC	Thr	NA	59	Csin_SINE-1_V#SINE/tRNA-V
Csin_SINE-2	YES	272	SINE/tRNA-V	CCAAA	Asp	NA	62	Csin_SINE-2_V#SINE/tRNA-V
Csin_SINE-3	YES	216	SINE/Unknown	AAACT	Gly	LINE_P.maximus_28_cons#LINE/CR1-Zenon	35	Csin_SINE-3_Unknown#SINE/Unknown
Csin_SINE-5	YES	213	SINE/Unknown	AAACT	Gly	LINE_P.maximus_28_cons#LINE/CR1-Zenon	35	
Csin_SINE-4	YES	263	SINE/Unknown	ACTTT	Arg	NA	28	Csin_SINE-4_Unknown#SINE/Unknown
Rphil_SINE-10	YES	222	SINE/tRNA-Core	AC-rich	Thr	LINE_M.mercenaria_56_cons#LINE/CR1-Zenon	34	Csin_SINE-8_Core#SINE/tRNA-Core
Csin_SINE-8	YES	220	SINE/tRNA-Core	AATC	Thr	NA	34	
Mcal_SINE-4	YES	418	SINE/Unknown	AATC	Gly*	LINE_B.platifrons_4_cons#LINE/I	7	Mcal_SINE-4_Unknown#SINE/Unknown
Mcal_SINE-28	YES	413	SINE/Unknown	AATC	Gly*	NA	7	
Mcal_SINE-32	YES	413	SINE/Unknown	AATC	Gly*	NA	7	
Mcal_SINE-1	YES	416	SINE/Unknown	AATC	Gly*	NA	7	
Mcor_SINE-1	YES	416	SINE/Unknown	AATC	Undet	NA	7	
Mcor_SINE-16	YES	417	SINE/Unknown	AATC	Undet	NA	7	
Mcor_SINE-17	YES	416	SINE/Unknown	AATC	Undet	NA	7	
Mvir_SINE-4	NO	354	SINE/tRNA-Deu	AACT	Arg	LINE_S.glomerata_23_cons#LINE/CR1-Zenon	12	Mcor_SINE-4_Deu#SINE/tRNA-Deu

Mcal_SINE-40	YES	351	SINE/tRNA-Deu	AAATC	Undet	LINE_M.phylippinarum_91_cons#LINE/CR1-Zenon	12	
Mcor_SINE-4	YES	357	SINE/tRNA-Deu	AGAT	Undet	LINE_M.phylippinarum_91_cons#LINE/CR1-Zenon	12	
Medu_SINE-4	YES	351	SINE/tRNA-Deu	AAATC	Arg*	LINE_M.phylippinarum_91_cons#LINE/CR1-Zenon	12	
Mcal_SINE-11	YES	349	SINE/tRNA-Meta	ATC	Ser	LINE_B.platifrons_4_cons#LINE/I	13	Medu_SINE-12_Meta#SINE/tRNA-Meta
Mcal_SINE-13	YES	347	SINE/tRNA-Meta	ATC	Ser	LINE_B.platifrons_4_cons#LINE/I	13	
Mcal_SINE-3	YES	344	SINE/tRNA-Meta	ATC	Ser	LINE_B.platifrons_4_cons#LINE/I	13	
Medu_SINE-12	YES	349	SINE/tRNA-Meta	ATC	Ser*	LINE_B.platifrons_4_cons#LINE/I	13	
Mvir_SINE-2	YES	257	SINE/tRNA-V	AAAC	Undet	NA	4	Medu_SINE-8_V#SINE/tRNA-V
Mcal_SINE-9	YES	253	SINE/tRNA-V	AATC	Leu	NA	4	
Mcal_SINE-12	YES	248	SINE/tRNA-V	AATC	Leu	NA	4	
Mcal_SINE-16	YES	245	SINE/tRNA-V	AATC	Leu*	NA	4	
Mcal_SINE-5	YES	254	SINE/tRNA-V	AAATC	Leu*	NA	4	
Mcal_SINE-41	YES	253	SINE/tRNA-V	AAATC	Leu*	NA	4	
Mcor_SINE-7	YES	252	SINE/tRNA-V	AAATC	Undet	NA	4	
Medu_SINE-8	YES	265	SINE/tRNA-V	AAATC	Undet	NA	4	
Mner_SINE-2	YES	179	SINE/tRNA-Core	CCAAA	Ser	NA	38	Mner_SINE-2_CORE#SINE/tRNA-Core
Mner_SINE-6	YES	179	SINE/tRNA-Core	CCAAA	Ser	NA	38	
Mqua_SINE-1	YES	286	SINE/tRNA-Core	CTTTAA	Pro	NA	58	Mqua_SINE-1_CORE#SINE/tRNA-Core
Mqua_SINE-2	YES	281	SINE/tRNA-V	TAAA	Pro	NA	26	Mqua_SINE-2_V#SINE/tRNA-V
Mvir_SINE-3	YES	394	SINE/Unknown	AATC	Leu*	LINE_B.platifrons_4_cons#LINE/I	43	Mvir_SINE-3_Unknown#SINE/Unknown
Myes_SINE-9	NO	479	SINE/Unknown	ATT	Ala*		41	Myes_SINE-9_Unknown#SINE/Unknown

Cgig_SINE-10	NO	316	SINE/tRNA-Core	ATT	Ser*	LINE_B.platifrons_81_cons#LINE/L2	15	Oden_SINE-1_CORE#SINE/tRNA-Core
Cgig_SINE-7	NO	313	SINE/tRNA-Core	ATT	Ser*	LINE_B.platifrons_81_cons#LINE/L2	15	
Oden_SINE-1	NO	320	SINE/tRNA-Core	ATT	Ser	LINE_B.platifrons_81_cons#LINE/L2	15	
Sglo_SINE-9	NO	320	SINE/tRNA-Core	ATT	Undet	LINE_B.platifrons_81_cons#LINE/L2	15	
Oden_SINE-2	YES	334	SINE/tRNA-MD	AATC	Thr	NA	14	Oden_SINE-3_MD#SINE/tRNA-MD
Oden_SINE-3	YES	347	SINE/tRNA-MD	AATC	Thr	NA	14	
Sglo_SINE-11	YES	343	SINE/tRNA-MD	AATC	Thr	NA	14	
Sglo_SINE-12	YES	342	SINE/tRNA-MD	AATC	Thr	NA	14	
Oden_SINE-4	YES	245	SINE/tRNA-V	AATC	Ser*	NA	65	Oden_SINE-4_V#SINE/tRNA-V
Oden_SINE-5	YES	273	SINE/tRNA-Deu	AAAC	Leu	NA	25	Oden_SINE-5_Deu#SINE/tRNA-Deu
Pfuc_SINE-1	NO	329	SINE/tRNA-Core	ACCTTT	Arg	NA	52	Pfuc_SINE-1_CORE#SINE/tRNA-Core
Pfuc_SINE-2	YES	338	SINE/tRNA-MD	AAGTG	Undet	NA	50	Pfuc_SINE-2_MD#SINE/tRNA-MD
Pfuc_SINE-3	YES	168	SINE/Unknown	A	Thr	NA	72	Pfuc_SINE-3_Unknown#SINE/Unknown
Rphil_SINE-1	NO	343	SINE/tRNA-MD	AATC	Asp	NA	73	Rphil_SINE-1_MD#SINE/tRNA-MD
Rphil_SINE-3	YES	197	SINE/Unknown	AAAC	Glu	LINE_M.phylippinarum_91_cons#LINE/CR1-Zenon	81	Rphil_SINE-3_Unknown#SINE/Unknown
Rphil_SINE-2	NO	316	SINE/tRNA-V	TTAC	Pro	NA	80	Rphil_SINE-4_V#SINE/tRNA-V
Rphil_SINE-4	YES	320	SINE/tRNA-V	AAA	Ser	NA	80	
Rphil_SINE-5	YES	302	SINE/tRNA-Deu	AC	Thr	NA	78	Rphil_SINE-5_Deu#SINE/tRNA-Deu
Rphil_SINE-6	NO	302	SINE/Unknown	ACCTTT	Arg	NA	77	Rphil_SINE-6_Unknown#SINE/Unknown
Rphil_SINE-7	YES	299	SINE/Unknown	AAC	His	NA	76	Rphil_SINE-7_Unknown#SINE/Unknown
Rphil_SINE-9	YES	203	SINE/Unknown	AAAAC	Undet	NA	75	Rphil_SINE-9_V#SINE/tRNA-V
Akag_SINE-3	YES	293	SINE/tRNA-Deu	AAAGT	Arg	NA	8	Sbro_SINE-45_Deu#SINE/tRNA-Deu

Sbro_SINE-45	YES	295	SINE/tRNA-Deu	AAGAT	Arg	NA	8	
Sbro_SINE-68	YES	291	SINE/tRNA-Deu	AAGT	Arg	NA	8	
Sbro_SINE-77	YES	292	SINE/tRNA-Deu	AAAGT	Arg	NA	8	
Tgra_SINE-2	YES	297	SINE/tRNA-Deu	TAAAC	Arg	NA	8	
Akag_SINE-4	YES	290	SINE/tRNA-Core	ACTTTA	Met	NA	9	
Sbro_SINE-56	YES	296	SINE/tRNA-Core	ACTTTA	Met	NA	9	
Sbro_SINE-66	YES	293	SINE/tRNA-Core	AACTT	Met	NA	9	
Sbro_SINE-73	YES	289	SINE/tRNA-Core	ACTTTA	Met	NA	9	
Sbro_SINE-78	YES	290	SINE/tRNA-Core	ACTTTA	Met	NA	9	
Sbro_SINE-40	YES	371	SINE/Unknown	AACC	Sup	LINE_B.platifrons_4_cons#LINE/I	11	Sbro_SINE-75_Unknown#SINE/Unknown
Sbro_SINE-64	YES	370	SINE/Unknown	AACC	Sup	LINE_B.platifrons_4_cons#LINE/I	11	
Sbro_SINE-72	YES	370	SINE/Unknown	AACC	Sup	LINE_B.platifrons_4_cons#LINE/I	11	
Sbro_SINE-75	YES	390	SINE/Unknown	AACC	Sup	LINE_B.platifrons_4_cons#LINE/I	11	
Scon_SINE-6	YES	312	SINE/tRNA-Core	TTAACCTA	Arg	NA	24	Scon_SINE-1_CORE#SINE/tRNA-Core
Scon_SINE-1	YES	312	SINE/tRNA-Core	TATTAACC	Arg	NA	24	
Scon_SINE-5	YES	293	SINE/tRNA-Meta	CAAA	Undet	NA	5	Scon_SINE-2_Meta#SINE/tRNA-Meta
Scon_SINE-4	YES	311	SINE/tRNA-Meta	CAAA	Undet	NA	5	
Scon_SINE-2	YES	311	SINE/tRNA-Meta	CAAA	Undet	NA	5	
Sgra_SINE-2	YES	304	SINE/tRNA-Meta	CAA	Pro	NA	5	
Sgra_SINE-3	YES	303	SINE/tRNA-Meta	CAAA	Pro	NA	5	
Sgra_SINE-6	YES	303	SINE/tRNA-Meta	CAAA	Pro	NA	5	
Sgra_SINE-9	YES	303	SINE/tRNA-Meta	CAAA	Pro	NA	5	

Sgra_SINE-10	YES	303	SINE/tRNA-Meta	CAAA	Pro	NA	5	
Scon_SINE-3	YES	237	SINE/tRNA-V	AAAC	Undet	NA	67	Scon_SINE-3_V#SINE/tRNA-V
Sglo_SINE-1	YES	338	SINE/tRNA-MD	AATC	Thr	NA	23	Sglo_SINE-10_MD#SINE/tRNA-MD
Sglo_SINE-10	YES	340	SINE/tRNA-MD	AATC	Thr	NA	23	
Sglo_SINE-7	YES	295	SINE/tRNA-Deu	AACT	Ser	NA	25	Sglo_SINE-7_Deu#SINE/tRNA-Deu
Sgra_SINE-11	YES	419	SINE/Unknown	ACA	Ser	NA	42	Sgra_SINE-11_Unknown#SINE/Unknown
Sgra_SINE-12	YES	262	SINE/Unknown	ACTT	Pro	NA	63	Sgra_SINE-12_Unknown#SINE/Unknown
Sgra_SINE-13	NO	222	SINE/tRNA-Core	NA	NA	NA	53	Sgra_SINE-13_CORE#SINE/tRNA-Core
Sgra_SINE-5	YES	259	SINE/tRNA-Meta	CAAA	Pro	NA	30	Sgra_SINE-5_Meta#SINE/tRNA-Meta
Sgra_SINE-1	YES	245	SINE/tRNA-V	CAA	Cys	NA	33	Sgra_SINE-7_V#SINE/tRNA-V
Sgra_SINE-7	YES	245	SINE/tRNA-V	CAA	Cys	NA	33	
Ssol_SINE-2	YES	266	SINE/tRNA-Meta	AAACT	Ser	LINE_P.maximus_28_cons#LINE/CR1-Zenon	21	Ssol_SINE-2_Meta#SINE/tRNA-Meta
Ssol_SINE-1	YES	265	SINE/tRNA-Meta	AAACT	Ser	LINE_P.maximus_28_cons#LINE/CR1-Zenon	21	
Ssol_SINE-4	YES	265	SINE/tRNA-Meta	AAACT	Ser	LINE_P.maximus_28_cons#LINE/CR1-Zenon	21	
Ssol_SINE-5	YES	349	SINE/tRNA-Meta	AAAC	Gly	NA	48	Ssol_SINE-5_MD#SINE/tRNA-MD
Ssol_SINE-6	NO	239	SINE/tRNA-Meta	A	Ser	NA	30	Ssol_SINE-6_Meta#SINE/tRNA-Meta
Ssol_SINE-7	YES	357	SINE/Unknown	ACAT	Ser	NA	46	Ssol_SINE-7_Unknown#SINE/Unknown
Ssol_SINE-8	YES	270	SINE/Unknown	ACTTT	Pro	NA	28	Ssol_SINE-8_Unknown#SINE/Unknown
Tcro_SINE-1	YES	214	SINE/tRNA-Core	AAACT	Met	LINE_M.phylippinarum_91_cons#LINE/CR1-Zenon	70	Tcro_SINE-1_CORE#SINE/tRNA-Core
Tcro_SINE-2	YES	302	SINE/tRNA-V	AATC	Ser	NA	56	Tcro_SINE-2_V#SINE/tRNA-V
Tcro_SINE-3	YES	294	SINE/tRNA-Deu	AATC	Thr	NA	57	Tcro_SINE-3_Deu#SINE/tRNA-Deu
Tcro_SINE-4	YES	315	SINE/Unknown	ATC	Asp	NA	54	Tcro_SINE-4_Unknown#SINE/Unknown

Tcro_SINE-5	YES	222	SINE/tRNA-V	AAAC	Thr	NA	69	Tcro_SINE-5_V#SINE/tRNA-V
Tgra_SINE-49	YES	244	SINE/tRNA-V	TTTAAA	Ile*	LINE_S.glomerata_23_cons#LINE/CR1-Zenon	22	Tgra_SINE-33_V#SINE/tRNA-V
Tgra_SINE-33	YES	250	SINE/tRNA-V	TTTAA	Ile*	LINE_S.glomerata_23_cons#LINE/CR1-Zenon	22	
Akag_SINE-5	NO	232	SINE/tRNA-V	ACTCC	Undet	NA	1	Tgra_SINE-48_V#SINE/tRNA-V
Sbro_SINE-6	YES	231	SINE/tRNA-V	ACTCC	Arg	NA	1	
Sbro_SINE-12	YES	229	SINE/tRNA-V	ACTCC	Arg	NA	1	
Sbro_SINE-21	YES	230	SINE/tRNA-V	ACTCC	Arg	NA	1	
Sbro_SINE-23	YES	232	SINE/tRNA-V	ACTCC	Arg	NA	1	
Sbro_SINE-25	YES	239	SINE/tRNA-V	ACTCC	Arg	NA	1	
Sbro_SINE-29	YES	233	SINE/tRNA-V	ACTCC	Arg	NA	1	
Sbro_SINE-47	YES	231	SINE/tRNA-V	ACTCC	Arg	NA	1	
Sbro_SINE-58	YES	230	SINE/tRNA-V	ACTCC	Arg	NA	1	
Tgra_SINE-1	YES	240	SINE/tRNA-V	ACTCC	Arg	NA	1	
Tgra_SINE-4	YES	231	SINE/tRNA-V	ACTCC	Arg	NA	1	
Tgra_SINE-10	YES	231	SINE/tRNA-V	ACTCC	Arg	NA	1	
Tgra_SINE-13	YES	231	SINE/tRNA-V	ACTCC	Arg	NA	1	
Tgra_SINE-14	YES	241	SINE/tRNA-V	ACTCC	Arg	NA	1	
Tgra_SINE-15	YES	231	SINE/tRNA-V	ACTCC	Arg	NA	1	
Tgra_SINE-16	YES	231	SINE/tRNA-V	ACTCC	Arg	NA	1	
Tgra_SINE-48	YES	244	SINE/tRNA-V	ACTCC	Arg	NA	1	
Akag_SINE-6	YES	398	SINE/tRNA-Meta	CCAA	Ser	LINE_B.platifrons_4_cons#LINE/I	3	Tgra_SINE-7_Meta#SINE/tRNA-Meta

Sup. Tab. 3: Correlation analyses between genome accumulation profiles of SINEs and their LINE counterparts with respect to bins of size 1 of % CpG corrected Kimura divergence of each SINE copy to its consensus sequence. Rho: Spearman's rank correlation coefficient (* p-value<0.05; ** p-value<0.01; *** p-value<0.001).

Species	**SINE**	**LINE**	**rho**
A. marissinica	Amar_SINE-2#CORE	Amar-1_LINE#CR1	0.84***
G. platifrons	Gpla_SINE-1#Meta	Gpla-4_LINE#I	0.75***
C. gigas	BivaV-SINE2_CrGi#V	CR1-14_CGi#CR1-Zenon	0.32**
	Cgig_SINE-10#CORE	Cgig-1_LINE#L2	0.83***
M. californianus	Mcal_SINE-11#Meta	Mcal-1_LINE#I	0.61***
M. edulis	Medu_SINE-2#Meta	Medu-1_LINE#I	0.44**
M. yessoensis	BivaV-SINE1_MiYe#V	Myes-2_LINE#I	0.84***
R. philippinarum	Rphil_SINE-8#V	Rphi-1_LINE#I	0.6***
S. broughtonii	Sbro_SINE-2#Meta	Sbro-1_LINE#I	0.3*
T. granosa	Tgra_SINE-7#Meta	Tgra-1_LINE#I	0.79***

Sup. Tab. 4: Copy number estimation of shared HCD SINE families. Only species for which families were de-novo mined have been included. Number of copies = Copy number estimation based on blastn analyses (See: Material and Methods "Copy number estimation of tRNA-related SINEs across bivalve diversity").

Family	**Species**	**Number of copies**
Amar_SINE-4_Deu#SINE/tRNA-Deu	*Archivesica marissinica*	118937
	Cyclina sinensis	74
BivaCORE-SINE2_MiYe#SINE/tRNA-Core	*Mizuhopecten yessoensis*	6812
	Pecten maximus	41
BivaMD-SINE1_TeGr#SINE/tRNA-Deu	*Anadara kagoshimensis*	33673
	Scapharca broughtonii	28437
	Tegillarca granosa	14319
BivaV-SINE1_ChFa#SINE/tRNA-V	*Mizuhopecten yessoensis*	2348
	Pecten maximus	4441
BivaV-SINE2_CrGi#SINE/tRNA-V	*Crassostrea gigas*	471
	Crassostrea ariakensis	444
	Crassostrea hongkongensis	470
Bpla_SINE-1_Meta#SINE/tRNA-Meta	*Gigantidas platifrons*	26130
	Mytilus californianus	64
	Mytilus coruscus	60
	Mytilisepta virgata	24
Cari_SINE-3_CORE#SINE/tRNA-Core	*Crassostrea gigas*	185
	Crassostrea ariakensis	421
	Crassostrea hongkongensis	180
Cgig_SINE-6_CORE#SINE/tRNA-Core	*Crassostrea gigas*	47
	Crassostrea ariakensis	80
	Crassostrea hongkongensis	44
Csin_SINE-8_Core#SINE/tRNA-Core	*Cyclina sinensis*	1004
	Ruditapes philippinarum	1797
Mcor_SINE-4_Deu#SINE/tRNA-Deu	Mytilus californianus	216
	Mytilus coruscus	393
	Mytilisepta virgata	536
	Mytilus edulis	4069
Medu_SINE-12_Meta#SINE/tRNA-Meta	Mytilus californianus	2190
	Mytilus edulis	153
Medu_SINE-8_V#SINE/tRNA-V	Mytilus californianus	6417
	Mytilus coruscus	6752
	Mytilisepta virgata	6151
	Mytilus edulis	1481
Oden_SINE-1_CORE#SINE/tRNA-Core	*Crassostrea gigas*	27
	Ostrea denselamellosa	831
	Saccostrea glomerata	171
Oden_SINE-3_MD#SINE/tRNA-MD	*Saccostrea glomerata*	1794
	Ostrea denselamellosa	20751
Sbro_SINE-45_Deu#SINE/tRNA-Deu	*Anadara kagoshimensis*	1704

	Scapharca broughtonii	1517
	Tegillarca granosa	366
Scon_SINE-2_Meta#SINE/tRNA-Meta	*Sinonovacula constricta*	30858
	Solen grandis	11738
Tgra_SINE-48_V#SINE/tRNA-V	*Anadara kagoshimensis*	600
	Scapharca broughtonii	581
	Tegillarca granosa	9652
Tgra_SINE-7_Meta#SINE/tRNA-Meta	*Anadara kagoshimensis*	4789
	Scapharca broughtonii	3316
	Tegillarca granosa	1064

Sup. Tab. 5: Genomic distribution of observed and simulated LINE insertions in *C. gigas* with respect to different genomic backgrounds. Gene flanking = 2500bp at both ends of genes; Intergenic = intergenic genomic regions after excluding gene flanking; SD = Standard deviation. Positive and negative Z-scores indicate more and less observed insertions compared to the null expectation, respectively.

Feature	**Simulated** **mean ± SD**	**Observed**	**Z-score**	**P-value**
Exons	12552.47 ± 132.45	1640	-82.39	> 0.001
Introns	26907.344 ± 168.31	20925	-35.54	> 0.001
UTR	3427.751 ± 60.94	1073	-38.64	> 0.001
Gene flaking	12055.945 ± 95.93	13233	12.27	> 0.001
Intergenic	12464.922 ± 91.16	14979	27.58	> 0.001

6. Chapter IV

High within- and between- individual structural variability is related to transposable elements insertions and deletions in bivalves

Jacopo Martelossi, Valentina Peona, Andrea Luchetti, Alexander Suh, Fabrizio Ghiselli

Note:
The Results of this chapter are currently being integrated with additional analyses prior to journal submission. All supplementary files and their captions can be found at the end of the chapter.

Abstract

Structural variants (SVs) represent one of the most abundant sources of genetic variation across a wide diversity of eukaryotes. The increasing number of high-quality genomes has underscored their pivotal role in species diversification and evolutionary processes. Transposable elements (TEs) stand out as primary contributors to the emergence of SVs, because of their mobility and repetitive nature. Oysters, an economically important clade of bivalves, host a rich and diverse TE landscape. Despite their potential impact on genome evolution and diversification, TEs remain relatively understudied. In this study, we analyzed four high-quality oyster genomes spanning the *Ostrea* and *Crassostrea* genera together with comprehensive population dataset for the Estuarine oyster *C. ariakensis* to characterize within- and between-individual SVs and their relationship with TEs. Our findings revealed that up to 14% of oyster genomes exhibit within-individual structural variability in terms of insertions and deletions between homologous chromosomes. These variants are significantly enriched in TEs, with transposon insertions that outnumber genomic deletions. Furthermore, SVs and *de novo* TE insertions, attributed to the concurrent activity of a diverse array of transposons, segregate among distinct *C. ariakensis* populations similarly to SNPs. This suggests a possible crucial role for SVs and TEs in facilitating local adaptation phenomena. This study represents the first effort to elucidate the role of SVs in shaping bivalve genome evolution, offering empirical evidence of the successful integration of both long and short reads technologies in compiling a high-fidelity set of SVs and *de novo* TE insertions useful for population genomic analyses in non-model species.

Keywords: Genome evolution, Structural Variants, Bivalves, Transposons, Population Genomics, Ostreida

Introduction

Structural variants (SVs) encompass a diverse range of genetic variants, comprising inversions, translocations, duplications, insertions, and deletions (length ≥ 50bp), along with other complex genomic rearrangements, such as chromosomal fusion and fission (Ho et al., 2020). While many genetic and evolutionary biologists have traditionally focused on Single Nucleotide Variants (SNVs), recent years have seen a growing body of evidence linking SVs to significant evolutionary processes. Indeed, SVs affect more base pairs than SNVs in the human genome (Frazer et al., 2009) and, more generally, they may represent the most important source of genetic variation between and within species (Wellenreuther et al., 2019). While most of the SVs usually occur at low frequency across populations, coherently with their generally neutral or deleterious effect (Weissensteiner et al., 2020; Zhou et al., 2019), multiple cases reporting their involvement in population differentiation have been described. Inversions are widely studied in the context of speciation and local adaptations due to their ability to suppress recombination across large genomic regions as well as to create postzygotic barriers (Kirkpatrick and Barton, 2006). Translocations can alter the expression of involved genes due to their novel nuclear position (Harewood et al., 2010), while insertions and deletions can directly change gene structures as well as their regulatory regions (Hof et al., 2016). Transposable elements (TEs) are the richest source of SVs across many eukaryotes (Bourque et al., 2018). TE-related SVs not only include *de novo* insertions (i.e., transposable element polymorphisms), but homologous TEs can also act as substrates for ectopic DNA repair mechanisms giving rise to inversion, duplication, and deletion events (Balachandran et al., 2022). TE-derived SVs have been linked to the emergence of multiple novel phenotypes, such as the industrial melanism in the peppered moth (Hof et al., 2016), the loss of the tail in apes (Xia et al., 2021), and plumage patterns in birds (Weissensteiner et al., 2020), among the others. Moreover, TE-derived variants can be especially important for rapid adaptations, due to the big amount of genetic variation that they can introduce (Stapley et al., 2015). Therefore, integrating their analyses in population genomics studies, even in non-model species, is a fundamental step forward for a deeper understanding of their evolution.

Oysters (Order Ostreida) are a group of worldwide distributed bivalves that include numerous important species for aquaculture. Like other bivalves, oysters are mainly characterized by sessile but large populations, external fertilization, high fecundity, and high juvenile mortality rate leading to usually high genetic diversity. *Ostrea* species such as *O. edulis* and *O.*

denselamellosa exhibit a unique reproductive mode among oysters and bivalves in general. They are live-bearing species, and their larvae exhibit a usually shorter planktonic dispersal stage. Numerous oysters also show high levels of physiological plasticity (Bromley et al., 2016) and experience wide ranges of different temperature and salinity being able to adapt to highly dynamic environmental conditions (Zhou et al., 2003; Li et al., 2018; Li et al., 2021). Recent genome projects revealed that their genomes are highly heterozygous and with a high and diverse repetitive content (Peñaloza et al., 2021; Gundappa et al., 2022; **Chapter II**) and that structural variants might be important contributors to species differentiation (Qi et al., 2023). Moreover, in **Chapter II** and **Chapter III** we found numerous putatively autonomous LINEs, DDE/D Class II, and SINEs, suggesting that multiple transposons lineages could potentially be active in oysters.

Massive gene-presence absence variation has been observed in *Mytilus galloprovincialis,* with about 30% of its gene set estimated to be subject to presence-absence polymorphisms (Gerdol et al., 2020). The presence of a gene in a hemizygous state (i.e., only one of the two chromosomal pairs in a diploid species is carrying the gene) in a genome could be an indicator of the possible polymorphism of that gene in the population (Gerdol et al., 2020). High levels of hemizygosity have also been observed within other bivalve species with most of these variants related to transposable elements (Calcino et al., 2021; Takeuchi et al., 2022). However, also immune-related genes such as C1qDC, NACHT, and big defensins have been repetitively found to be enriched in such regions and subjected to presence-absence variation, also in the pacific oyster *Crassostrea gigas* (Rosa et al., 2015; Gerdol et al., 2020; Calcino et al., 2021; Takeuchi et al., 2022).

Here, we took advantage of the increasing high-quality genomic resources available for oysters to study their level of hemizygosity in terms of insertions and deletions, which we also call structural heterozygosity, quantifying the contribution of transposable elements and host genes in their emergence. For this purpose, we set up a within-individual SVs calling pipeline using multiple long-read aligners and SV callers, benchmarking our results with simulations and short reads re-genotyping. Moreover, we present some first insights into the possible important role of SVs and transposable elements in population differentiation in the estuarine oyster *Crassostrea ariakensis* using a previously published population short-read dataset comprising 107 WGS samples. This specie is present along the entire Chinese coast at different temperature and salinity, have a clear population structure and high population divergence based on previous analyses on SNPs (Li et al., 2021) and thus may represent a possible ideal case to study the evolution and impact of SVs in population differentiation.

Materials and Methods

Genomic datasets

We selected four publicly available high-quality oyster genomes to study their level of haplotypic variability: *Crassostrea* (*C. gigas*, *C. ariakensis*) and two *Ostrea* (*O. edulis* and *O. denselamellosa*) species for which a haploid representation of their diploid genome is available online (**Sup. Tab. 1**). We took care in choosing only assemblies for which a wild or farmed but not inbred sample was sequenced using both long- and short-read technologies. Two of these genomes were sequenced with PacBio Hi-Fi technology (*C. ariakensis*, *O. denselamellosa*), one with PacBio CLR (*C. gigas*) and one with Nanopore PromethION (*O. edulis*). Their contig N50s range from ~1.6 Mb for *C. gigas* to 14 Mb for *O. densemellosa*.

For population genomics and polymorphic TE insertion analyses we used a recent dataset comprising 106 *Crassostrea ariakensis* whole genomes (NCBI bioproject: PRJNA715058) sequenced with PE Illumina short reads with a read length of 150 bp at a mean coverage of ~20X (**Sup. Tab. 2**).

Transposable element annotation

For each of the four genomes, we built a *de novo* species-specific repeat library using RepeatModeler2 and the LTR pipeline extension (Flynn et al., 2020). From the resulting libraries we excluded potential host gene by firstly creating a database of oyster reference proteomes devoid of transposable element TE-derived proteins. The RefSeq annotations for *C. gigas* (GCF_902806645.1) and *O. edulis* (GCF_947568905.1) were combined and subjected to a blastp search (E-value 1E-05) against the RepeatPep database from the RepeatMasker package (Tarailo-Graovac and Chen, 2009), eliminating all queries with homology to TE-derived proteins. The resultant filtered proteome was then used as a database for blastx searches, with each species-specific repeat library serving as the query and the results were supplied to ProtExcluder. Then, the cleaned libraries were merged with the Mollusca RepBase database v. 20181026, along with a set of manually curated consensus sequences generated for bivalves in **Chapter II** and **Chapter III**. Redundancy in the merged library was eliminated following the 80-80 rule (i.e., requiring a minimum 80% identity along 80% of the shortest sequence; Wicker et al., 2007) using cd-hit-est (Fu et al., 2012). Finally, RepeatMasker was used to annotate repetitive elements in each of the five genomes with the merged non-redundant library in sensitive mode.

Heterozygosity prediction

For each genome, we calculated three distinct metrics of heterozygosity: (1) kmer-based heterozygosity, (2) SNP-based heterozygosity, and (3) structural heterozygosity, focusing on insertions and deletions between homologous chromosomes and therefore corresponding to hemizygous genomic regions.

For both kmer-based and SNP-based heterozygosity, we firstly cleaned Illumina short reads using bbduk (Bushnell, 2014), with a minimum quality threshold of 30 and a minimum length requirement of 35 bps. The filtered reads were then mapped to the genome using bwa-mem (Li, 2013), and their genomic coverage was determined using Mosdepth (Pedersen and Quinlan, 2018). Subsequently, we extracted the mapped reads and generated a kmer histogram with Jellysfish (Marçais and Kingsford, 2011), which was uploaded to Genomescope2 (Ranallo-Benavidez et al., 2020) for kmer-based heterozygosity estimation. SNPs were identified based on bwa alignments and bcftools mpileup (Li, 2011), retaining only biallelic variants with a genotype quality greater than 20, and called in genomic regions with coverage no greater than three times the median genome-wide estimation.

Structural heterozygosity, involving insertions and deletions, was calculated based on the alignment results of the long reads used for genome assembly against the assembly itself. For each genome, two alignments were generated: (1) using the minimap2 (Li, 2018) wrapper pbmm2 (https://github.com/PacificBiosciences/pbmm2) for PacBio reads or minimap directly for Nanopore reads, and (2) using LGNRM (Sedlazeck et al., 2018). In all instances the appropriate preset option was selected based on the type of reads. SVs were then called for each alignment using the PacBio variant caller pbsv (https://github.com/PacificBiosciences/pbsv) and Sniffles2 (Smolka et al., 2024), providing genomic regions corresponding to tandem repeats as predicted by RepeatMasker. This resulted in four sets of SV calls, which were then filtered and merged to obtain a filter consensus set of reliable SVs. Firstly, variants not labelled as PASS and genotyped as homozygous for the alternative allele were removed, as they likely represented assembly errors or false positives. SURVIVOR (Jeffares et al., 2017) was employed to: (1) retain variants supported by at least three SV sets (with a maximal distance of 1kb between breakpoints, considering the SV type and its strand), (2) with a length greater than 49bp, (3) supported by at least four reads, (4) at least 1kb away from assembly gaps or ends of scaffolds, and (5) corresponding to insertions or deletions only. From the resulting VCF, an additional variant set was generated, preserving the same SVs but with breakpoints forced as estimated by pbsv instead of those chosen by SURVIVOR. Based on our benchmark results for all downstream analyses we used the SV consensus set with breakpoints estimated by pbsv.

SV calling benchmarks

We ran a comprehensive benchmark of our SV calling pipeline using two distinct approaches: simulations with Sim-it (Dierckxsens et al., 2021) and short-read-based genotyping of previously identified hemizygous genomic regions with Paragraph (Chen et al., 2019).

For the simulations, we initially generated a synthetic haploid assembly of *C. ariakensis*, introducing 1,000 random deletions selected from the merged set with pbsv breakpoints inferred from the original genome. From this simulated assembly, synthetic Hi-Fi, CLR, and Nanopore reads were generated at 15X genome-wide coverage, each with the appropriate technology-specific error profile. Simultaneously, synthetic long reads at 15X coverage were generated from the original genome, and the two read sets were merged. This approach allowed us to produce a true set of hemizygous deletions relative to the original genome and the reads carrying them. The merged reads were then mapped to the original *C. ariakensis* genome, and SVs were called using the same pipeline described in Material and Methods section 2.3. Bedtools intersect (Quinlan and Hall, 2010) was used to assess the reciprocal overlap (RO) between the true set and the called set. We applied three RO thresholds to consider a variant as correctly called: 80%, 90%, and 99%. A graphical representation of the simulation process is available in Sup. Fig. 1.

Additionally, we re-genotyped all insertions and deletions of all genomes using the previously mapped Illumina short reads and Paragraph. Specifically, we re-genotyped the two consensus SV sets (i.e., with default SURVIVOR breakpoints and forcing pbsv inferred ones) and considered a SV correctly called when also Paragraph genotyped it as heterozygous.

Estimation of genomic content and origin of hemizygous genomic regions

We assessed the overlap between hemizygous deletions with transposable elements (TEs) as annotated by RepeatMasker in all four genomes. We consider a variant as TE-derived when it had an overlap with a TE annotation of at least 70% of its length. To statistically test the overrepresentation of TE-derived variants in hemizygous deletions, we compared their observed number to a null distribution generated by randomly reshuffling all hemizygous deletions 10,000 times (excluding assembly gaps) and counting at each iteration the number of TE-derived variants. Furthermore, because deletions in the reference genome—that in our case correspond to the assembled haplotype—can represent both deletions in the other haplotype or insertions in the assembled one, we polarized the variants of *C. ariakensis* using *C. gigas* as reference. Briefly PacBio reads from *C. gigas* were used to genotype hemizygous deletions of *C. ariakensis* using SVJedy-graph (Romain and Lemaitre, 2023). When a *C. ariakensis* hemizygous deletion was genotyped as homozygous for the reference allele in *C. gigas* it was consider a true deletion event and an insertion event when it was genotyped as homozygous for the alternative allele. Gene overlap with hemizygous deletions was assessed on the reference NCBI gene annotation of *C. gigas* and *O. denselamellosa* requiring an overlap of the gene with a deletion of at least 70% of its length. Homology between hemizygous genes and TE-related proteins was assessed with blastp against the RepeatPep library (E-value 1E-10).

Polymorphic transposable element insertions analyses

To identify non-reference germline TE insertions across the 106 *C. ariakensis* samples we used TEMP2 (Yu et al., 2021). Birefly, Illumina short reads were mapped to the reference genome with bwa-mem and their median genome-wide coverage calculated with mosedpeth. We ran TEMP2 with default parameters and results were filtered keeping only insertions supported by at least 10% of the median genome-wide read coverage, with read support at both ends and with a frequency in the sequenced genome ≥ 0.2. TE insertions were clustered between samples using *bedtools cluster* with a maximum allowed distance among their ends of 50bp. We considered two or more clustered insertions shared between samples when they derived from the same TE consensus sequence.

Between-individual SNP and SV calling

Single nucleotide variants (SNV), including SNPs and indels shorter than 50 bp, were jointly called across the *C. ariakensis* population dataset using Platypus (Rimmer et al., 2014), retaining only variants with a minimum mapping and base quality of 20. From the resulting VCF file we only kept biallelic SNVs called on the 10 *C. ariakensis* assembled pseudo-chromosomes and marked as PASS.

Between-individual SVs were initially called independently in each sample using Manta (Chen et al., 2016) with default parameters. Sample-specific VCF files were then filtered, retaining only variants marked as PASS, and subsequently merged into a multi-sample VCF using Jasmine (Kirsche et al., 2023), allowing a maximal Euclidean distance of 20 bp between breakpoint representations of the variants. Paragraph was then used to re-genotype the merged VCF file for each sample to obtain population-scale genotyping, skipping genomic regions with a coverage higher than 20 times the median sample coverage, as recommended by Paragraph developers. The resulting multi-sample VCF file was further filtered, retaining only insertions and deletions. Genotypes not marked as PASS by Paragraph were set as missing, and variants falling within tandem repeats were removed. Furthermore, we also removed variants genotyped as homozygous for the reference allele across all samples. Indeed, due to the strict cutoff in terms of distance between breakpoints applied during the SV merging process by Jasmine, redundancy in variants is expected in our merged VCF. Nevertheless, owing to the high sensitivity of Paragraph in breakpoint deviations (refer to Chen et al., 2019), redundantly incorrect representations are expected to be genotyped as 0/0 in all samples, thereby ensuring the accurate genotyping only of their best representation. Finally, similarly to what applied on SNPs, we only kept variants called on assembled pseudo-chromosomes.

Population genomic analyses

We ran population genomic analyses similarly for both SNVs and SVs datasets, using only variants with a minor allele frequency (MAF) ≥ 0.05 and genotyped in at least 30% of the samples. For SV-based population genomic analyses we only retained deletions since short reads could not reconstruct the inserted sequences and we could not assess the overlap with TEs. Population structure was inferred using fastStructure (Raj et al., 2014) with a parameter K (number of populations) ranging from 1 to 5. The optimal number of populations was chosen based on their marginal likelihood, as calculated by the chooseK function from the fastStructure package. PCA analyses were carried out on the same datasets using PLINK (Chang et al., 2015) with default parameters.

Results

We used two different aligners in conjunction with two distinct SV callers to obtain a confident consensus set of hemizygous genomic regions, defined as insertions and deletion between homologous chromosomes (i.e, heterozygous variants), in the four analyzed oyster assemblies. Firstly, it is important to note that each combination of SV caller and long read mapper produce quite different SV sets with between 2,799 and 24,521 variants supported by one combination only (Sup. Fig. 2). Precision (i.e., the frequency of true calls out of the total number of called variants) and recall (i.e., the frequency of correctly called variants out of the total number of simulated variants) rates, estimated through simulation-based benchmarks, highlight a strong impact of the sequencing technology and SV calling pipeline on the accurate identification of hemizygous deletions, particularly under stringent reciprocal overlap (RO) requirements (Fig. 1A-B).

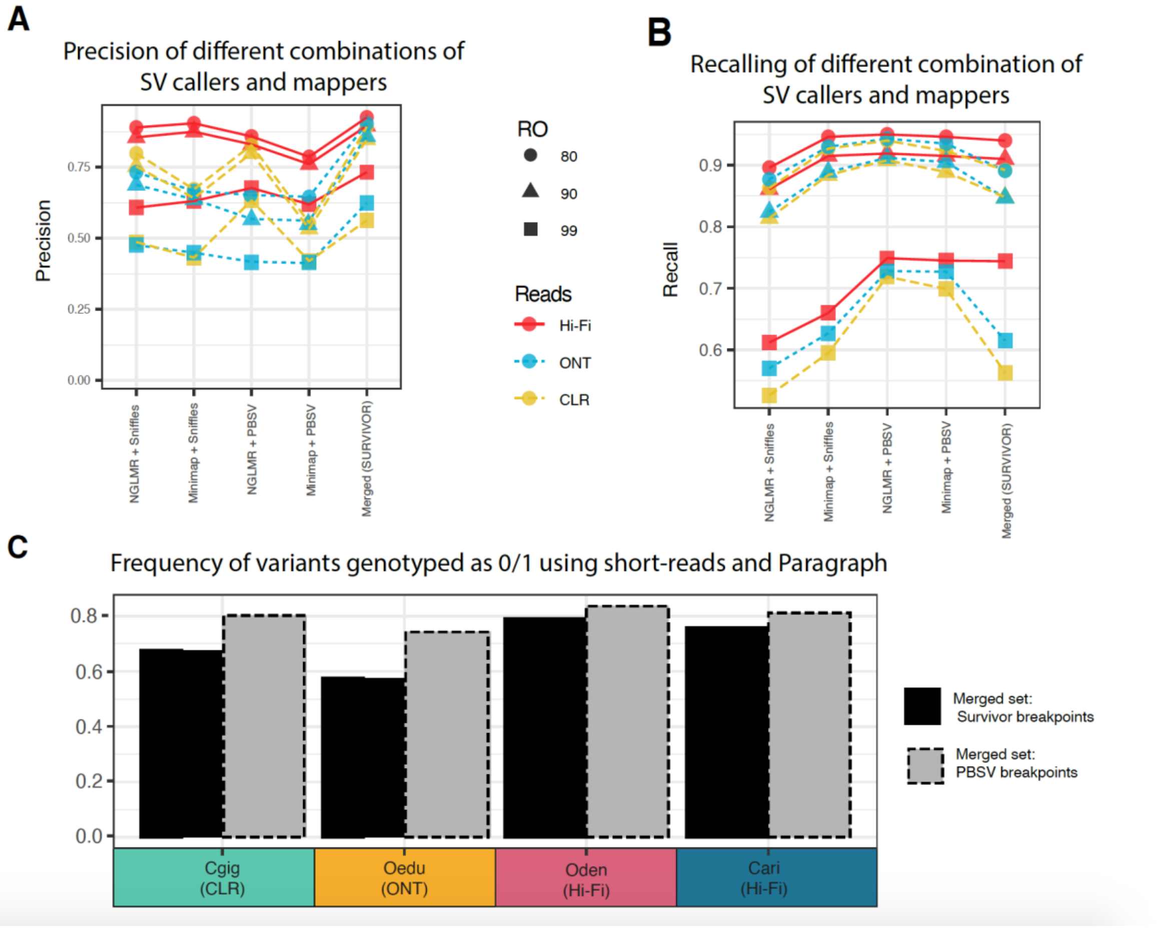


Figure 1: Benchmark results of the SV calling pipelines using simulation and short read based re-genotyping of hemizygous variants. (**A**) Precision and (**B**) recalling rates based on simulations of 1,000 hemizygous variants in the *C. ariakensis* genome. (**C**) Frequency of correctly re-genotyped variants using short reads and Paragraph. Cgig=*C. gigas*, Oedu=*O. edulis*, Oden=*O.denselamellosa*, Cari=*C. ariakensis.*

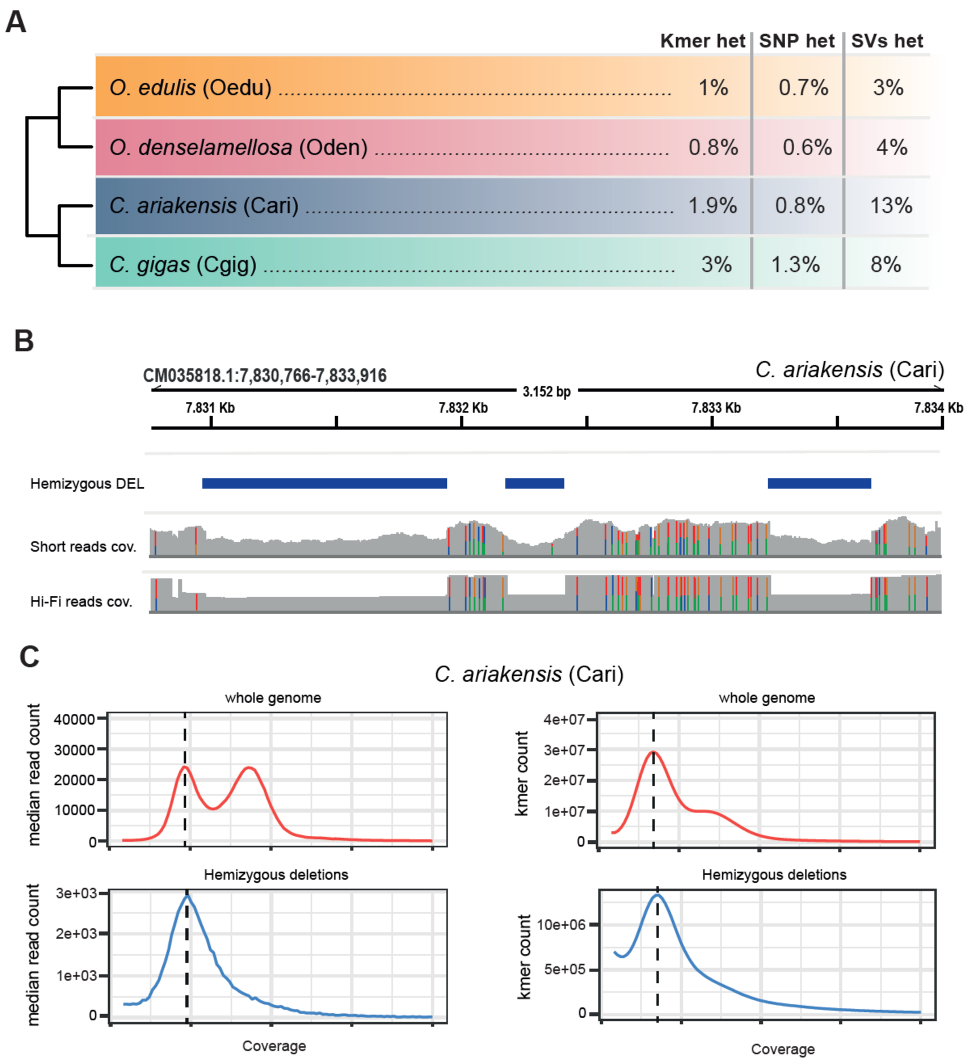
As expected, Hi-Fi reads consistently outperformed other sequencing technologies in terms of both precision and recall, regardless of the read mapper and SV caller used. Additionally, requiring support from at least three callers increases the precision rate (ranging from 0.94 to 0.89 at 80% RO, from 0.91 to 0.84 at 90% RO, and from 0.74% to 0.56 at 99% RO) but decreases the recall rate (from 0.94 to 0.89 at 80% RO, from 0.91 to 0.89 at 90% RO, and from 0.74 to 0.56 at 99% RO) compared to the single best-performing pipeline (Fig. 1A-1B). A noteworthy exception is the recall rates using CCS reads at the 99% RO requirement. In this case, we did not observe a drastic drop in the consensus set recalling rate (0.74), which is nearly identical to those obtained with pbsv (0.75 recalling rate with both NGLMR and Minimap2 aligners). Moreover, while pbsv performs less effectively than Sniffles for both CCS and ONT reads in terms of precision rates, it emerges as the best caller in terms of recall rate at the stringent 99% RO requirement, independent of the sequencing technology. These results may reflect a more accurate representation of SVs and their breakpoint using CCS reads and pbsv. We also found a striking peak of precision rate at all RO requirements when combining NGLMR and pbsv with CLR reads. This combination seems therefore to perform particularly well for long but noisy PacBio reads.

To test this hypothesis, we re-genotyped the consensus set of hemizygous variants identified in each of the four genomes using Paragraph and the short reads from the same sample used to sequence the long read dataset. Since Paragraph realigns reads to a sequence graph using stringent parameters during the genotyping process and is highly sensitive to breakpoint estimations, we expected a higher number of correctly re-genotyped variants when enforcing pbsv breakpoints. Indeed, when forcing SV breakpoints to those estimated by pbsv, we consistently observed an increase in the number of recalled variants, which always exceed the 75% (Fig. 1C). In this case as well, we observed an impact of the sequencing technology, as hemizygous variants identified in *C. ariakensis* and *O. denselamellosa*, for which Hi-Fi reads were available, were more frequently correctly genotyped using short reads, with a percentage of correctly re-genotyped variants equal to 81% and 84%, respectively. On the other hand, *C. gigas* and *O. edulis*, for which CLR and ONT reads were used, benefited more from pbsv-estimated breakpoints, with an increase in the Paragraph re-genotyping rate shifting from 68% to 80% and from 58% to 75%, respectively.

For these reasons we decided to rely on the SV consensus set with the breakpoints estimated by pbsv since a variant supported by at least three callers always include a pbsv-derived representation.

High levels of structural heterozygosity in oysters are mainly related to transposable elements

Kmer-based analyses estimated oyster heterozygosity to range from 0.8% *for O, denselamellosa* to 3% for *C. gigas* (Fig. 2A). SNP-based heterozygosity estimation consistently yielded lower values for all species, ranging from 0.6% for *O. denselamellosa* to 1.3% for *C. gigas*. Conversely, we observed high levels of structural heterozygosity, ranging from 3% for *O. edulis* to 13% for *C. ariakensis*. Given that we only considered insertions and deletions between homologous chromosomes, this indicates that a significant proportion of the oyster genome exists in a hemizygous state, where only one haplotype is present (Fig. 2B).



	Kmer het	SNP het	SVs het
O. edulis (Oedu)	1%	0.7%	3%
O. denselamellosa (Oden)	0.8%	0.6%	4%
C. ariakensis (Cari)	1.9%	0.8%	13%
C. gigas (Cgig)	3%	1.3%	8%

Figure 2: (**A**) Heterozygosity levels estimated across the four analyzed oyster genomes. Kmer het=Kmer-based heterozygosity estimation, SNP het=SNP based heterozygosity, SV het=Structural heterozygosity in terms of insertions and deletions between homologue chromosomes and thus corresponding to hemizygous genomic regions. (**B**) Example of a *C. ariakensis* genomic region characterized by three hemizygous deletions (Hemizygous DEL). These regions show about half of the coverage of flanking sequences and are depleted of any SNP. (**C**) Median read count of 1 kb sliding windows and kmer count of all mapped reads in the *C. ariakensis* genome (red) and across hemizygous deletions (blue). The dotted line represents the heterozygous peak of the whole genome. Plots for all other species are reported in **Sup. Fig. 3**.

K-mer and short read counts across hemizygous deletions reflect what we may expect from genomic regions that are present as a single copy in a diploid genome (Fig. 2C, Sup. Fig. 3). Indeed, contrary to the two-peak plots of the whole genome where both heterozygous and homozygous peaks are present, hemizygous deletions only show a single peak that overlaps with the heterozygous peak of the whole genome plot.

Short to medium-size insertions and deletions constitute most hemizygous genomic regions, with deletions ranging from 50 bp to 1 kb, comprising 79%, 75%, 83%, and 85% of the total number of variants in *C. gigas*, *C. ariakensis*, *O. edulis*, and *O. denselamellosa*, respectively (Fig. 3A).

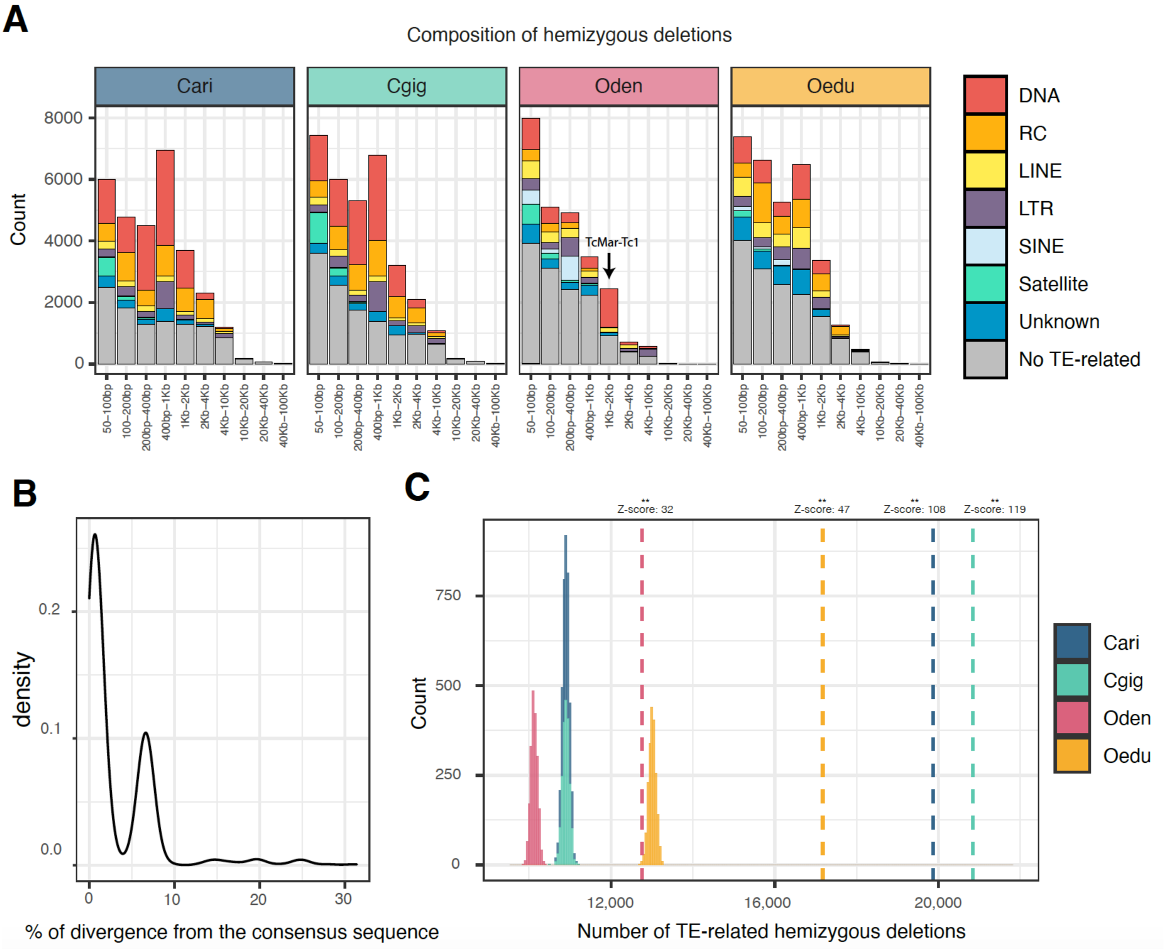


Figure 3: (**A**) Number of hemizygous deletions subdivided into different size bins and coloured by different TE types when classified as TE-related, defined as variants with an overlap with a TE annotation of at least 70%. The black arrow highlights the peak of 1Kb-2Kb hemizygous deletions overlapping with TcMar-Tc1 TEs in *O. denselamellosa*. (**B**) % of identity with the consensus sequence of TcMar-Tc1 elements found in *O. denselamellosa* hemizygous deletions (1Kb-2Kb size bin). (**C**) Null distributions of the number of TE-related hemizygous deletions generated by randomly reshuffling the variants 10,000 times across the genome and counting, at each iteration, the number of overlaps with a TE annotation, requiring at least 70% overlap. Dotted lines represent the observed number of TE-related variants. The Z-score of the observed number of TE-related variants compared to the null distribution is reported for each species. ** = p-value < 0.001. Cari = *C. ariakensis*, Cgig = *C. gigas*, Oden = *O. denselamellosa*, Oedu = *O. edulis*.

TE-related variants, defined as those having an overlap of at least 70% with a TE annotation, represent 62%, 64%, 47%, and 52% of the hemizygous deletions in the same four species (Fig. 3A). The composition of TE-related deletions generally mirrors the genome-wide distribution of different TE classes where *Crassostrea* species are dominated by DNA and RC elements (71% and 69% of the overall TE content in respectively *C. ariakensis* and *C. gigas*), whereas in the Ostrea genus also non-LTR retrotransposons can occupy a considerable proportion of the host genome (Sup. Tab. 3). Indeed, in the *Crassostrea* genus, most of the TE-related variants are DNA transposons (46% and 48% in *C. gigas* and *C. ariakensis*, respectively), followed by RC TEs (23% and 24%, respectively) (Fig. 3A). However, the latter elements outnumber DNA transposons in the size bin between 2 kb - 4 kb, where they constitute 43% and 58% of TE-related variants in *C. gigas* and *C. ariakensis*. Within the *Ostrea* genus, where fewer TE-related variants were identified, we observed a generally less pronounced predominance of specific TE classes. For example, SINEs and LTRs in *O. denselamellosa* are predominant elements in the 200 bp-400 bp size bin, constituting 32% and 24% of the variants. Despite this more homogeneous composition, the size bin between 1 kb and 2 kb in *O. denselamellosa* is almost entirely occupied by DNA transposons (83%), with Tc1-Mariner elements comprising as much as 92% of them. The distance of these elements to their consensus sequence, used as a proxy of the time of the insertion, approximate a bimodal distribution with one peak at 0% - 1% and a second one at 6% - 7% suggesting their recent accumulation (Fig. 3B). Across all four genomes we also found between 197 in *O. edulis* and 990 in *C. gigas* short hemizygous deletions with a length ranging from 50 bp to 100 bp that involved satellite DNA. Due to the high number of TE-related variants identified, we sought to determine if hemizygous genomic regions are more likely to be associated with TEs compared to the null expectation. By randomly reshuffling the original genomic intervals 10,000 times across each genome, we observed a significant overrepresentation of TE-related variants in all four species (Fig. 3B). This supports the hypothesis that transposons can serve as hotspots for the emergence of within-individual variants, leading to structural heterozygosity.

Origin of hemizygous genomic regions and relationship with TE activity

As mentioned earlier, hemizygous genomic regions can arise due to a novel insertion or the deletion of a genomic region. However, whether variants are labeled "insertions" or "deletions" by SV callers is only with respect to the reference genome. A variant labelled as a deletion could potentially represent either a genuine deletion event in the query genome or an insertion in the reference genome. To assess the impact of insertion and deletion events on the emergence of hemizygous genomic regions in *C. ariakensis*, we polarized all its deletions genotyping them with the PacBio reads from *C. gigas*. Out of the 6,435 variants (21.5% of the total) successfully genotyped, 814 (13%) were identified as homozygous for the reference allele, indicating potential deletion events in *C. ariakensis* (Fig. 4A). Another 1,443 (22%) were genotyped as heterozygous, suggesting shared hemizygous regions in both *C. ariakensis* and *C. gigas*. The remaining 4,178 (65%) were genotyped as homozygous for the alternative allele, implying they likely correspond to heterozygous insertions in the *C. ariakensis* reference genome. Coherently, most of these insertions were found to be TE-related (65%), and transposons included in such variants appear significantly younger in terms of identity to their consensus sequence compared to those included in deletion events (pairwise Wilcoxon test with Bonferroni correction: p-value = 1.4E-07; Fig. 4B).

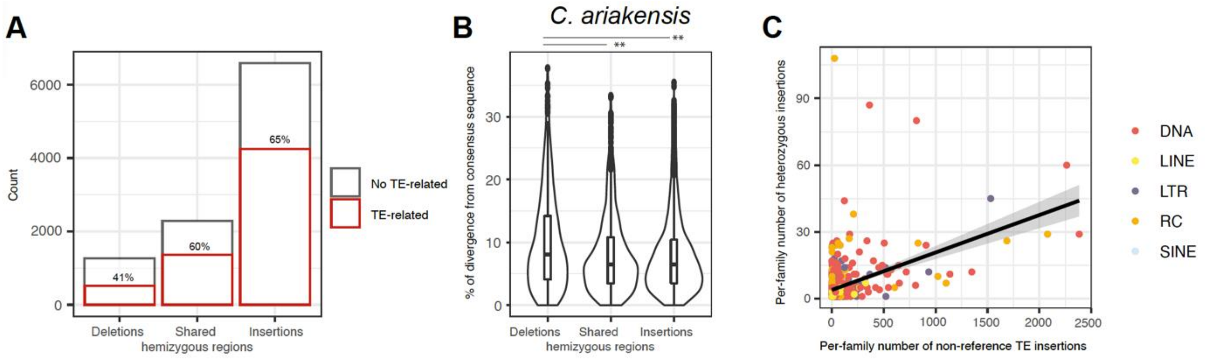


Figure 4: (**A**) Number and percentage of hemizygous deletions identified in the reference genome of *C. ariakensis*, categorized as insertions, deletions, and shared hemizygous genomic regions based on their genotype in the sister species *C. gigas*. (**B**) Percentage of identity for each TE copy compared to its consensus sequence, serving as a proxy for the time of the insertion. Straight lines indicate significant comparisons. ** = p-value < 0.01. (**C**) Scatterplot depicting the per-family number of polymorphic insertions identified in 107 *C. ariakensis* wild samples and the per-family number of heterozygous TE-related insertions identified in the *C. ariakensis* reference genome (Spearman's rank correlation rho = 0.51, p-value < 0.01).

To determine if hemizygous genomic regions can serve as indicators of TE activity across populations, we conducted *de novo* TE insertion analyses across 107 *C. ariakensis* wild samples from the whole coast of China (See Fig. 5A). We identified a total of 55,383 high-quality mobile element insertions (MEIs), ranging from a minimum of 292 to a maximum of 706 MEIs per sample (mean = 530). Most of these MEIs are related to DNA transposons (57%), and among the elements classified at the RepeatMasker type level, TcMar-Pogo, CMC-EnSpm, Kolobok, PIF-Spy, and Crypton-A contribute between 3,429 and 1,069 MEIs (Sup. Tab. 4). RC/Helitron constitute 20% of the total number of *de novo* insertions, followed by LTRs (18%)

and particularly by Gypsy transposons (7,000 insertions). This is interesting since LTRs occupy only 5.57% of the *C. ariakensis* genome, while RCs are almost three times more common (16.02% of the host genome) and may indicate recent expansion of these transposons. Regarding LINEs (3.7% of the total number of polymorphic insertions), we found 631 RTE-X, 124 L1-Tx1, and 535 CR1-Zenon insertions, coherently with the identification of autonomous copies of these superfamilies in **Chapter II**. Finally, very few SINE-related MEIs were found (0.5%), with most of them coming from the tRNA-V (129 elements) and CORE (125) superfamilies. Lastly, we observed a significant positive correlation between the number of *de novo* insertions annotated for each family (*i.e.*, consensus sequence) and the number of hemizygous genomic regions classified as insertions after polarizing the variants (Spearman's rank correlation rho = 0.51, p-value < 0.01; Fig. 4C).

TE-related structural variants contribute to population differentiation in *C. ariakensis*

The substantial within-individual variability observed in oysters may imply that TE-related structural variants could play a significant role in population differentiation within natural populations. Using the previously described population dataset of *C. ariakensis* (Fig. 5A), we identified a range of 3,092 to 15,024 insertions (mean = 7,940; sd = 3,348) and 18,151 to 32,876 deletions (mean = 25,434; sd = 3,521) per sample. After consolidating individual SV sets, performing joint genotyping with Paragraph, and filtering the resulting variants, we retained 255,140 variants, the majority of which correspond to deletions relative to the reference genome (83%). Among these deletions, 53% were associated with TEs for at least 70% of their length.

Population genomics analyses, based on 61,562 deletions (69% of which overlapping with TEs) with a MAF equal or greater than 0.05 and successfully genotyped in at least 30% of the samples, mirror the original results based on SNPs obtained by Li et al., (2021), as well as our re-analyses based on 1,285,295 SNVs (Fig. 5B), indicating the high quality of our population informative SV set. Both admixture analyses indicated that the best-fitting number of population clusters was two (Fig. 5B). PCA analyses based on SVs reflect admixture analysis results, with a clear separation of two main populations encompassing all northern (NC) and southern (SC) samples. Additionally, two clusters were identified from PCA: one containing all samples from central China and the other including individuals from the Qingdao locality (QD). Two outliers were also identified, labelled as Northern samples but clearly separated in the PCA analyses.

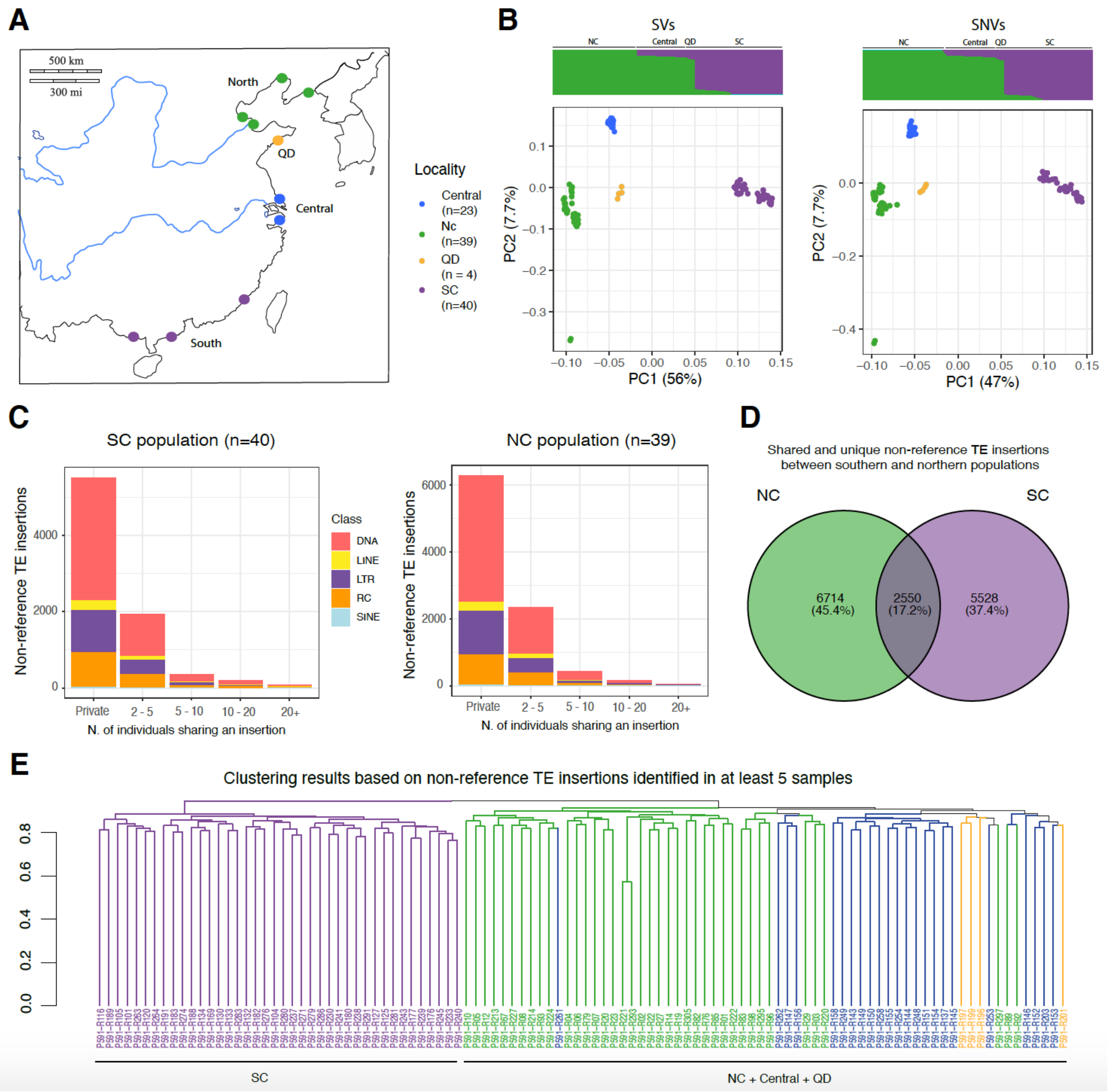


Fig. 5: (A) Geographic origin of the 107 *C. ariakensis* samples used for population genomic analyses and de-novo TE insertion discovery and sequenced by Li et al., (2021). (**B**) FastStructure and PCA analyses based on population informative SNVs and SVs. (**C**) Frequency of non-reference TE insertions across individuals of southern and northern populations. (**D**) Number of unique and shared insertions between *C. ariakensis* populations. NC = Norther China population; SC = Southern China population. (**E**) Clustering analyses based on non-reference TE insertions identified in at least five samples. Colors of tips and b ranches represent the geographic origin of the samples.Nore that the two NC samples with long branches correspond to the two outliers reported present in panel B.

When specifically examining *de novo* TE insertions, we found more MEIs segregating in the NC population compared to SC samples (Fig. 5C). Most MEIs are present as singletons or at low frequencies across SC and NC populations, chosen due to the high and similar number of samples (Fig. 5C). Indeed, among the 66 and 81 insertions identified in more than 20 individuals within their geographical area, only eight and two were found with a frequency equal to or greater than 90%, respectively. We identified an Helitron insertion present in 36 samples from the SC population (90% of the individuals) and completely absent in the NC population, whereas we did not identify any instances of the opposite scenario (*i.e*, an insertion at high frequency in NC population but absent in the SC). Additionally, we identified eight insertions at intermediate frequencies (≥ 50% of the individuals) present in SC population but absent in NC individuals (two unknown DNA transposons, two Gypsy, four Helitrons, and one PIF-Harbinger) and three presents in the NC samples but absent in the SC population (one Helitron, one Crypton-A, and one Kolobok). Finally, the 2,550 MEIs shared by the two populations in at least one individual (Fig. 5D) exhibited low frequencies (Sup. Fig. 4). However, when performing cluster analyses on MEIs shared by at least five individuals we observed again a clear separation between SC and NC samples, with Central and QD samples nested within the latter (Fig. 5E).

Discussion

In the present study, we extensively characterized within- and between-individual SVs and their relationships with transposable elements in an economically important clade of bivalves. Because the performance of SV callers might vary depending on the underlying dataset (Mahmoud et al., 2019), we firstly benchmarked the performance of two different SV callers (pbsv and Sniffles2) and two long read mapper (Minimap2 and NGLMR) based on an empirical set of SVs when analyzing within-individual structural variants. Firstly, the different results obtained with each combination underscore once again the importance of a consensus strategy between multiple calling pipelines to obtain a reliable set of SVs, even if this can increase the number of false negatives (Mahmoud et al., 2019; Dierckxsens et al., 2021; Balachandran et al., 2022). On the other hand, consistent with previous benchmarks of SV callers (Dierckxsens et al., 2021; Balachandran et al., 2022), we found pbsv to perform better in terms of breakpoint estimation with all types of reads and long-read mapper compared to Sniffles2. This highlights the importance of carefully choosing the best representation of a variant when a consensus set of SVs is produced. Indeed, Survivor chooses to report the SV representation with the start coordinate closer to the median position, while an empirically based selection could represent a better solution.

Applying our within-individual SV calling pipeline to four high-quality oyster genomes, we revealed high levels of structural heterozygosity, with between 3% to 4% of *Ostrea* spp. and 8% to 14% of *Crassostrea* spp. genomes present in a hemizygous state. Structural heterozygosity surpasses both k-mer-based and SNP-based heterozygosity estimations, indicating that long insertions and deletions contribute significantly more to haplotypic variability compared to SNVs in this clade. Similar results were obtained in previous estimations of structural heterozygosity in different bivalve species using a comparable approach. For instance, Calcino et al. (2021) estimated 10.68% of the *Pecten maximus* genome to be present in a hemizygous state, while Takeuchi et al. (2022) estimated 18.12% in the oyster *Pinctada fucata*. It must be noted that while both studies used a similar approach of re-mapping long-reads against the haploid representation of the assembly, they exclusively used pbsv without a consensus approach, which may lead to overestimations. Nonetheless, it is evident that SVs significantly contribute to haplotypic variability across a diverse range of bivalve species, and haploid representations of their genomes fail to fully capture the sequenced genome to unpredictable extents, depending on the assembler inclination to incorporate such hemizygous regions. Phased assemblies integrating Hi-C technologies, as demonstrated by Takeuchi et al. (2022), are therefore essential to obtain a comprehensive representation of bivalve genomes and particularly of their TE content. Indeed, we found a significant overrepresentation of TE-related variants and TE-derived genes among hemizygous genomic regions. Recently active transposons are more likely to be represented since a novel insertion is always present in a heterozygous state. Moreover, 'cut-and-paste' transposon insertions can potentially contribute doubly to the overall hemizygous content, as every transposition event results in the deletion of the TE from its original location. If the original TE insertion was in homozygosity, this would leave an orphan ancestral copy in heterozygosity in the sister chromatid. This may be the case for the previously described peak of TcMar-Tc1 related variants present in *O. denselamellosa*, with elements more diverging from their consensus representing the orphan degenerated element, and the more similar ones representing the novel insertions. Consistently, DNA transposons were also found to be recently active in the same species based on repeat landscape analyses (Dong et al., 2023).

Besides *de novo* TE insertion, it is well-known that genomic deletions of both old transposons and non-TE related genomic regions can frequently occur and contribute to genome size evolutionary dynamics across eukaryotes (Lagemaat et al., 2005; Sotero-Caio et al., 2017; Kapusta et al., 2017). Consistent with this 'accordion' model of genome size evolution, we found that both putative *de novo* transposon insertions and deletions of generally older TEs and other genomic regions are found in *C. ariakensis* hemizygous genomic regions, with insertion events greatly outnumbering deletion events. Interestingly, we found a relatively high number of shared hemizygous genomic regions between *C. ariakensis* and *C. gigas*, the species that we used to polarize the hemizygous variants of *C. ariakensis*. Considering that the two species separated between 30 and 40 million years ago (Li et al., 2021), this may correspond to shared ancestral polymorphisms or introgression events. However, these results may also be exaggerated by genotyping errors, which may be common even when using long read data (Duan et al., 2022). Therefore, confirming these results using, for example, a second SV genotyper like Sniffles2 and a consensus approach is required before any conclusions can be drawn. Nevertheless, we found that the number of heterozygous TE insertions in a reference genome is a good indicator of the activity of that transposon at the population level. This information can therefore be used when population level data are not available to prioritize the curation of transposons more likely to be active and for future hypothesis driven research.

Because of the high within-individual haplotypic variability found in oyster genomes, we hypothesized that transposons and structural variants in general may contribute to genetic differentiation between oyster population and particularly to the highly structured populations of *C. ariakensis* (Li et al., 2021). Thys hypothesis was confirmed based on population genomic analyses with both admixture and PCA analyses based on SVs matching almost exactly our re-analyses based on SNVs as well as the original results of Li et al., (2021). The clear separation between NC and SC population is concordant with the summer ocean currents of the Chinese coast. Generally, Northern and Southern coastal currents remain separated, and they only meet near the Yangtze River estuary, at the location of the Central population (Li et al., 2021). These results also underly that despite long reads clearly perform better and represent the gold standard for SV calling, short-reads can still be successfully used to obtain a high-quality SV set, at least for deletions with respect to the reference.

Consistent with the identification of a wide diversity of putatively active TE lineages in **Chapter II**, we found that at least CR1-Zenon, L1-Tx1, RTE-X, TcMar, Kolobok, PIF-Harbinger, and CMC superfamilies are affectively able to produce novel TE insertions in *C. ariakensis*. Also concerning SINEs, non-reference TE insertion analyses recapitulate the results of **Chapter III** where we found that *C. ariakensis* and the *Crassostrea* genus more generally harbor almost exclusively tRNA-V and tRNA-CORE SINEs. Moreover, despite the low genomic occurrence of LTR retrotransposons, Gypsy elements seem to be particularly active, and we therefore encourage further efforts also in the characterization of these elements in oysters. As we might expect, most of the *de novo* TE insertions are private to one individual reflecting their neutral or deleterious effect, as previously observed also in multiple taxa like *Drosophila* (Rech et al., 2022), and flycatchers songbirds (Suh et al., 2018). Despite this, we found a handful of insertions that have intermediate/high frequency in only the SC or NC population, representing good candidates for local adaptations. Moreover, the agreement of the clustering analyses based on MEIs with SVs and SNVs results confirms that transposons, despite occurring at low frequency, have the potential to provide the substrate for local adaptation events.

Conclusions, limitations, and future perspectives

Here we provide the first insights into the role of SVs and transposable elements in oyster genome evolution and population differentiation. Importantly, we confirm that SVs and transposons are primary contributors to the high heterozygosity observed in bivalves. The vast majority of this structural heterozygosity appears to be related to *de novo* TE insertions, while genomic deletions seem to be a less common phenomenon. Through simple population genomic analyses, we found that both SVs in general as well as MEIs segregate similarly to SNVs within different *C. ariakensis* populations, thus potentially providing the substrate for local adaptations.

Furthermore, we were able to confirm the results obtained in **Chapters II** and **Chapter III** by observing a great diversity of different TE lineages that are contemporarily active in oyster genomes. While these results are encouraging and important *per se*, we acknowledge that some parts of this work need more attention and further analyses. Firstly, in the last few months, novel oyster genomes sequenced with Hi-Fi technologies were released, including a RefSeq genome for *Saccostrea echinata* (GCF_033153115.1) and a novel reference assembly for *O. edulis* (Li et al., 2023). *S. echinata* belongs to a genus that we did not have the chance to analyze, therefore including it could be important to have a better overview across the entire oyster clade. The novel *O. edulis* assembly has replaced the previous NCBI RefSeq genome, which we used at the project outset due to the availability of the genome annotation on NCBI. Moreover, sequencing with Hi-Fi reads could increase the precision of SV calling, as demonstrated by our benchmarks. Therefore, replacing that genome could be the optimal choice.

This rapid growth in bivalve genomic resources, especially for oysters, can also enable us to polarize the hemizygous variants using a conspecific individual rather than a different species, as we chose to do here. This would increase the mappability of the reads and therefore the genotype quality. In this context, it is also important to implement a second software and a consensus approach, as we have already discussed. Finally, it could be interesting and relatively straightforward to check for other types of TE-related variants, especially for true deletions. The presence of homologous repeats across the boundaries of a deletion might imply TE-mediated rearrangements due to homologous recombination or non-homologous repair events (Balachandran et al., 2022). This, together with retro-duplication events, might help explain the presence of genes in a hemizygous state and thus giving a more mechanistic insight to our work.

Regarding population genomic analyses on *C. ariakensis*, we managed to obtain a high-quality set of SVs and MEIs; however, we have not yet had the chance to properly test putative variants under selection, aside from observing the presence of some TEs at high frequency in only one of the two main populations. In this context, it would also be important to perform a genome annotation since it is not freely available in order to identify possible host genes affected by variants under putative selection.

Availability of Supplementary Figures and Tables

All supplementary tables and figures are included as supplementary materials at the end of the chapter.

Funding

This work was supported by the Canziani bequest funded to F.G. and A.L., and the ‘Ricerca Fondamentale Orientata’ (RFO) funding from the University of Bologna to F.G. and A.L. The computations/data handling and analyses were enabled by resources in project snic2022-22-1207 provided by the National Academic Infrastructure for Supercomputing in Sweden (NAISS) at UPPMAX, funded by the Swedish Research Council through grant agreement no. 2022-06725.

References

- Balachandran, P., Walawalkar, I.A., Flores, J.I., Dayton, J.N., Audano, P.A., Beck, C.R., 2022. Transposable element-mediated rearrangements are prevalent in human genomes. Nat Commun 13, 7115. https://doi.org/10.1038/s41467-022-34810-8
- Bourque, G., Burns, K.H., Gehring, M., Gorbunova, V., Seluanov, A., Hammell, M., Imbeault, M., Izsvák, Z., Levin, H.L., Macfarlan, T.S., Mager, D.L., Feschotte, C., 2018. Ten things you should know about transposable elements. Genome Biology 19, 199. https://doi.org/10.1186/s13059-018-1577-z
- Bromley, C., McGonigle, C., Ashton, E.C., Roberts, D., 2016. Bad moves: Pros and cons of moving oysters – A case study of global translocations of Ostrea edulis Linnaeus, 1758 (Mollusca: Bivalvia). Ocean & Coastal Management 122, 103–115. https://doi.org/10.1016/j.ocecoaman.2015.12.012
- Bushnell, B., 2014. BBMap: A Fast, Accurate, Splice-Aware Aligner (No. LBNL-7065E). Lawrence Berkeley National Lab. (LBNL), Berkeley, CA (United States).
- Calcino, A.D., Kenny, N.J., Gerdol, M., 2021. Single individual structural variant detection uncovers widespread hemizygosity in molluscs. Philosophical Transactions of the Royal Society B: Biological Sciences 376, 20200153. https://doi.org/10.1098/rstb.2020.0153

- Calcino, A.D., Kenny, N.J., Gerdol, M., 2021. Single individual structural variant detection uncovers widespread hemizygosity in molluscs. Philosophical Transactions of the Royal Society B: Biological Sciences 376, 20200153. https://doi.org/10.1098/rstb.2020.0153
- Calcino, A.D., Kenny, N.J., Gerdol, M., 2021. Single individual structural variant detection uncovers widespread hemizygosity in molluscs. Philosophical Transactions of the Royal Society B: Biological Sciences 376, 20200153. https://doi.org/10.1098/rstb.2020.0153

- Chang, C.C., Chow, C.C., Tellier, L.C., Vattikuti, S., Purcell, S.M., Lee, J.J., 2015. Second-generation PLINK: rising to the challenge of larger and richer datasets. GigaScience 4, s13742-015-0047–8. https://doi.org/10.1186/s13742-015-0047-8.
- Chen, S., Krusche, P., Dolzhenko, E., Sherman, R.M., Petrovski, R., Schlesinger, F., Kirsche, M., Bentley, D.R., Schatz, M.C., Sedlazeck, F.J., Eberle, M.A., 2019. Paragraph: a graph-based structural variant genotyper for short-read sequence data. Genome Biology 20, 291. https://doi.org/10.1186/s13059-019-1909-7.
- Chen, X., Schulz-Trieglaff, O., Shaw, R., Barnes, B., Schlesinger, F., Källberg, M., Cox, A.J., Kruglyak, S., Saunders, C.T., 2016. Manta: rapid detection of structural variants and

indels for germline and cancer sequencing applications. Bioinformatics 32, 1220–1222. https://doi.org/10.1093/bioinformatics/btv710.

- Dierckxsens, N., Li, T., Vermeesch, J.R., Xie, Z., 2021. A benchmark of structural variation detection by long reads through a realistic simulated model. Genome Biology 22, 342. https://doi.org/10.1186/s13059-021-02551-4.
- Dierckxsens, N., Li, T., Vermeesch, J.R., Xie, Z., 2021. A benchmark of structural variation detection by long reads through a realistic simulated model. Genome Biol 22, 1–16. https://doi.org/10.1186/s13059-021-02551-4
- Dong, Z., Bai, Y., Liu, S., Yu, H., Kong, L., Du, S., Li, Q., 2023. A chromosome-level genome assembly of Ostrea denselamellosa provides initial insights into its evolution. Genomics 115, 110582. https://doi.org/10.1016/j.ygeno.2023.110582
- Duan, X., Pan, M., Fan, S., 2022. Comprehensive evaluation of structural variant genotyping methods based on long-read sequencing data. BMC Genomics 23, 324. https://doi.org/10.1186/s12864-022-08548-y
- Flynn, J.M., Hubley, R., Goubert, C., Rosen, J., Clark, A.G., Feschotte, C., Smit, A.F., 2020. RepeatModeler2 for automated genomic discovery of transposable element families. PNAS 117, 9451–9457.
- Frazer, K.A., Murray, S.S., Schork, N.J., Topol, E.J., 2009. Human genetic variation and its contribution to complex traits. Nat Rev Genet 10, 241–251. https://doi.org/10.1038/nrg2554
- Fu, L., Niu, B., Zhu, Z., Wu, S., Li, W., 2012. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics 28, 3150–3152. https://doi.org/10.1093/bioinformatics/bts565.
- Gerdol, M., Moreira, R., Cruz, F., Gómez-Garrido, J., Vlasova, A., Rosani, U., Venier, P., Naranjo-Ortiz, M.A., Murgarella, M., Greco, S., Balseiro, P., Corvelo, A., Frias, L., Gut, M., Gabaldón, T., Pallavicini, A., Canchaya, C., Novoa, B., Alioto, T.S., Posada, D., Figueras, A., 2020. Massive gene presence-absence variation shapes an open pan-genome in the Mediterranean mussel. Genome Biology 21, 275. https://doi.org/10.1186/s13059-020-02180-3.
- Gerdol, M., Moreira, R., Cruz, F., Gómez-Garrido, J., Vlasova, A., Rosani, U., Venier, P., Naranjo-Ortiz, M.A., Murgarella, M., Greco, S., Balseiro, P., Corvelo, A., Frias, L., Gut, M., Gabaldón, T., Pallavicini, A., Canchaya, C., Novoa, B., Alioto, T.S., Posada, D., Figueras, A., 2020. Massive gene presence-absence variation shapes an open pan-genome

in the Mediterranean mussel. Genome Biol 21, 275. https://doi.org/10.1186/s13059-020-02180-3

- Gundappa, M.K., Peñaloza, C., Regan, T., Boutet, I., Tanguy, A., Houston, R.D., Bean, T.P., Macqueen, D.J., 2022. Chromosome-level reference genome for European flat oyster (Ostrea edulis L.). Evolutionary Applications 15, 1713–1729. https://doi.org/10.1111/eva.13460

- Gundappa, M.K., Peñaloza, C., Regan, T., Boutet, I., Tanguy, A., Houston, R.D., Bean, T.P., Macqueen, D.J., 2022. Chromosome-level reference genome for European flat oyster (Ostrea edulis L.). Evolutionary Applications 15, 1713–1729. https://doi.org/10.1111/eva.13460

- Harewood, L., Schütz, F., Boyle, S., Perry, P., Delorenzi, M., Bickmore, W.A., Reymond, A., 2010. The effect of translocation-induced nuclear reorganization on gene expression. Genome Res 20, 554–564. https://doi.org/10.1101/gr.103622.109
- Ho, S.S., Urban, A.E., Mills, R.E., 2020. Structural variation in the sequencing era. Nat Rev Genet 21, 171–189. https://doi.org/10.1038/s41576-019-0180-9.
- Hof, A.E. van't, Campagne, P., Rigden, D.J., Yung, C.J., Lingley, J., Quail, M.A., Hall, N., Darby, A.C., Saccheri, I.J., 2016. The industrial melanism mutation in British peppered moths is a transposable element. Nature 534, 102–105. https://doi.org/10.1038/nature17951
- Jeffares, D.C., Jolly, C., Hoti, M., Speed, D., Shaw, L., Rallis, C., Balloux, F., Dessimoz, C., Bähler, J., Sedlazeck, F.J., 2017. Transient structural variations have strong effects on quantitative traits and reproductive isolation in fission yeast. Nat Commun 8, 14061. https://doi.org/10.1038/ncomms14061.
- Kapusta, A., Suh, A., Feschotte, C., 2017. Dynamics of genome size evolution in birds and mammals. Proceedings of the National Academy of Sciences 114, E1460–E1469. https://doi.org/10.1073/pnas.1616702114
- Kirkpatrick, M., Barton, N., 2006. Chromosome Inversions, Local Adaptation and Speciation. Genetics 173, 419–434. https://doi.org/10.1534/genetics.105.047985
- Kirsche, M., Prabhu, G., Sherman, R., Ni, B., Battle, A., Aganezov, S., Schatz, M.C., 2023. Jasmine and Iris: population-scale structural variant comparison and analysis. Nat Methods 20, 408–417. https://doi.org/10.1038/s41592-022-01753-3.
- Lagemaat, L.N. van de, Gagnier, L., Medstrand, P., Mager, D.L., 2005. Genomic deletions and precise removal of transposable elements mediated by short identical DNA segments in primates. Genome Res. 15, 1243–1249. https://doi.org/10.1101/gr.3910705

- Li, A., Dai, H., Guo, X., Zhang, Z., Zhang, K., Wang, C., Wang, X., Wang, W., Chen, H., Li, X., Zheng, H., Li, L., Zhang, G., 2021. Genome of the estuarine oyster provides insights into climate impact and adaptive plasticity. Commun Biol 4, 1–12. https://doi.org/10.1038/s42003-021-02823-6.
- Li, C., Kou, Q., Zhang, Z., Hu, L., Huang, W., Cui, Z., Liu, Y., Ma, P., Wang, H., 2021. Reconstruction of the evolutionary biogeography reveal the origins and diversification of oysters (Bivalvia: Ostreidae). Molecular Phylogenetics and Evolution 164, 107268. https://doi.org/10.1016/j.ympev.2021.107268
- Li, H., 2011. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics 27, 2987–2993. https://doi.org/10.1093/bioinformatics/btr509.
- Li, H., 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. https://doi.org/10.48550/arXiv.1303.3997.
- Li, H., 2018. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics 34, 3094–3100. https://doi.org/10.1093/bioinformatics/bty191
- Li, X., Bai, Y., Dong, Z., Xu, C., Liu, S., Yu, H., Kong, L., Li, Q., 2023. Chromosome-level genome assembly of the European flat oyster (Ostrea edulis) provides insights into its evolution and adaptation. Comparative Biochemistry and Physiology Part D: Genomics and Proteomics 45, 101045. https://doi.org/10.1016/j.cbd.2022.101045

- Li, X., Bai, Y., Dong, Z., Xu, C., Liu, S., Yu, H., Kong, L., Li, Q., 2023. Chromosome-level genome assembly of the European flat oyster (Ostrea edulis) provides insights into its evolution and adaptation. Comparative Biochemistry and Physiology Part D: Genomics and Proteomics 45, 101045. https://doi.org/10.1016/j.cbd.2022.101045

- Mahmoud, M., Gobet, N., Cruz-Dávalos, D.I., Mounier, N., Dessimoz, C., Sedlazeck, F.J., 2019. Structural variant calling: the long and the short of it. Genome Biology 20, 246. https://doi.org/10.1186/s13059-019-1828-7
- Marçais, G., Kingsford, C., 2011. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics 27, 764–770. https://doi.org/10.1093/bioinformatics/btr011.
- Pedersen, B.S., Quinlan, A.R., 2018. Mosdepth: quick coverage calculation for genomes and exomes. Bioinformatics 34, 867–868. https://doi.org/10.1093/bioinformatics/btx699.

- Peñaloza, C., Gutierrez, A.P., Eöry, L., Wang, S., Guo, X., Archibald, A.L., Bean, T.P., Houston, R.D., 2021. A chromosome-level genome assembly for the Pacific oyster Crassostrea gigas. GigaScience 10, giab020. https://doi.org/10.1093/gigascience/giab020.

- Peñaloza, C., Gutierrez, A.P., Eöry, L., Wang, S., Guo, X., Archibald, A.L., Bean, T.P., Houston, R.D., 2021. A chromosome-level genome assembly for the Pacific oyster Crassostrea gigas. GigaScience 10, giab020. https://doi.org/10.1093/gigascience/giab020

- Qi, H., Cong, R., Wang, Y., Li, L., Zhang, G., 2023. Construction and analysis of the chromosome-level haplotype-resolved genomes of two Crassostrea oyster congeners: Crassostrea angulata and Crassostrea gigas. GigaScience 12, giad077. https://doi.org/10.1093/gigascience/giad077

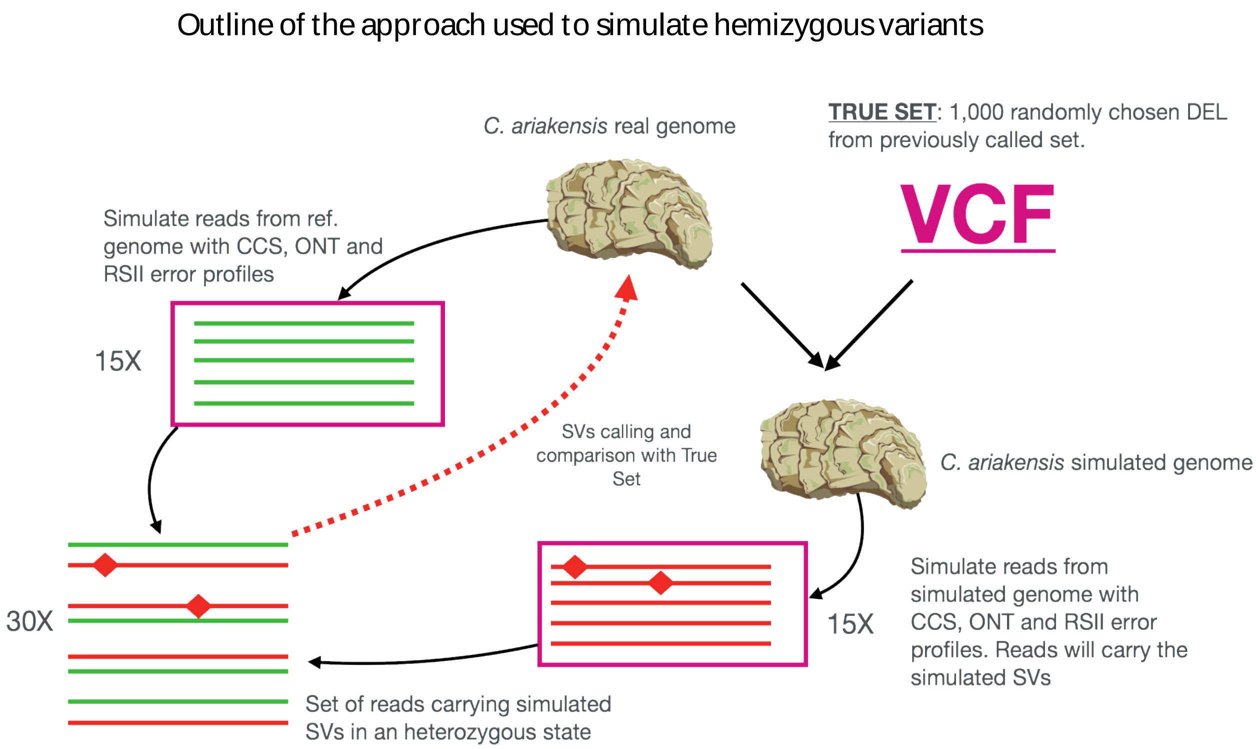
- Qi, H., Cong, R., Wang, Y., Li, L., Zhang, G., 2023. Construction and analysis of the chromosome-level haplotype-resolved genomes of two Crassostrea oyster congeners: Crassostrea angulata and Crassostrea gigas. GigaScience 12, giad077. https://doi.org/10.1093/gigascience/giad077

- Quinlan, A.R., Hall, I.M., 2010. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26, 841–842. https://doi.org/10.1093/bioinformatics/btq033.
- Raj, A., Stephens, M., Pritchard, J.K., 2014. fastSTRUCTURE: Variational Inference of Population Structure in Large SNP Data Sets. Genetics 197, 573–589. https://doi.org/10.1534/genetics.114.164350.
- Ranallo-Benavidez, T.R., Jaron, K.S., Schatz, M.C., 2020. GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. Nat Commun 11, 1432. https://doi.org/10.1038/s41467-020-14998-3.
- Rech, G.E., Radío, S., Guirao-Rico, S., Aguilera, L., Horvath, V., Green, L., Lindstadt, H., Jamilloux, V., Quesneville, H., González, J., 2022. Population-scale long-read sequencing uncovers transposable elements associated with gene expression variation and adaptive signatures in Drosophila. Nat Commun 13, 1948. https://doi.org/10.1038/s41467-022-29518-8
- Rimmer, A., Phan, H., Mathieson, I., Iqbal, Z., Twigg, S.R.F., Wilkie, A.O.M., McVean, G., Lunter, G., 2014. Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. Nat Genet 46, 912–918. https://doi.org/10.1038/ng.3036.

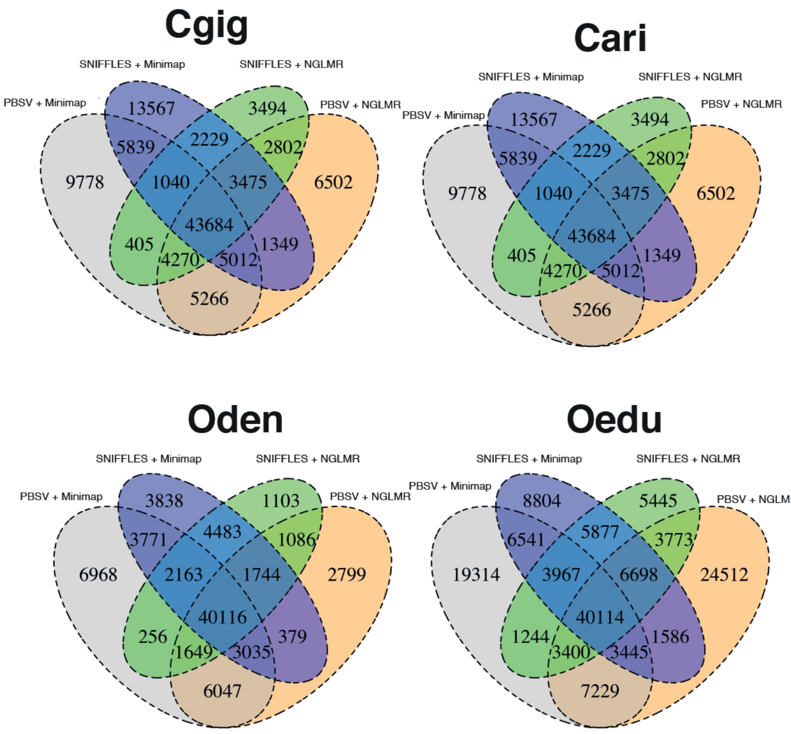
- Romain, S., Lemaitre, C., 2023. SVJedi-graph: improving the genotyping of close and overlapping structural variants with long reads using a variation graph. Bioinformatics 39, i270–i278. https://doi.org/10.1093/bioinformatics/btad237.
- Rosa, R.D., Alonso, P., Santini, A., Vergnes, A., Bachère, E., 2015. High polymorphism in big defensin gene expression reveals presence–absence gene variability (PAV) in the oyster Crassostrea gigas. Developmental & Comparative Immunology 49, 231–238. https://doi.org/10.1016/j.dci.2014.12.002
- Sedlazeck, F.J., Rescheneder, P., Smolka, M., Fang, H., Nattestad, M., von Haeseler, A., Schatz, M.C., 2018. Accurate detection of complex structural variations using single molecule sequencing. Nat Methods 15, 461–468. https://doi.org/10.1038/s41592-018-0001-7.
- Smolka, M., Paulin, L.F., Grochowski, C.M., Horner, D.W., Mahmoud, M., Behera, S., Kalef-Ezra, E., Gandhi, M., Hong, K., Pehlivan, D., Scholz, S.W., Carvalho, C.M.B., Proukakis, C., Sedlazeck, F.J., 2024. Detection of mosaic and population-level structural variants with Sniffles2. Nat Biotechnol 1–10. https://doi.org/10.1038/s41587-023-02024-y
- Sotero-Caio, C.G., Platt, R.N., II, Suh, A., Ray, D.A., 2017. Evolution and Diversity of Transposable Elements in Vertebrate Genomes. Genome Biology and Evolution 9, 161–177. https://doi.org/10.1093/gbe/evw264
- Stapley, J., Santure, A.W., Dennis, S.R., 2015. Transposable elements as agents of rapid adaptation may explain the genetic paradox of invasive species. Molecular Ecology 24, 2241–2252. https://doi.org/10.1111/mec.13089
- Suh, A., Smeds, L., Ellegren, H., 2018. Abundant recent activity of retrovirus-like retrotransposons within and among flycatcher species implies a rich source of structural variation in songbird genomes. Molecular Ecology 27, 99–111. https://doi.org/10.1111/mec.14439
- Takeuchi, T., Suzuki, Y., Watabe, S., Nagai, K., Masaoka, T., Fujie, M., Kawamitsu, M., Satoh, N., Myers, E.W., 2022. A high-quality, haplotype-phased genome reconstruction reveals unexpected haplotype diversity in a pearl oyster. DNA Res 29, dsac035. https://doi.org/10.1093/dnares/dsac035
- Tarailo-Graovac, M., Chen, N., 2009. Using RepeatMasker to Identify Repetitive Elements in Genomic Sequences. Current Protocols in Bioinformatics 25, 4.10.1-4.10.14. https://doi.org/10.1002/0471250953.bi0410s25.

- Weissensteiner, M.H., Bunikis, I., Catalán, A., Francoijs, K.-J., Knief, U., Heim, W., Peona, V., Pophaly, S.D., Sedlazeck, F.J., Suh, A., Warmuth, V.M., Wolf, J.B.W., 2020. Discovery and population genomics of structural variation in a songbird genus. Nat Commun 11, 3403. https://doi.org/10.1038/s41467-020-17195-4
- Wellenreuther, M., Mérot, C., Berdan, E., Bernatchez, L., 2019. Going beyond SNPs: The role of structural genomic variants in adaptive evolution and species diversification. Molecular Ecology 28, 1203–1209. https://doi.org/10.1111/mec.15066
- Wicker, T., Sabot, F., Hua-Van, A., Bennetzen, J.L., Capy, P., Chalhoub, B., Flavell, A., Leroy, P., Morgante, M., Panaud, O., Paux, E., SanMiguel, P., Schulman, A.H., 2007. A unified classification system for eukaryotic transposable elements. Nat Rev Genet 8, 973–982. https://doi.org/10.1038/nrg2165.
- Xia, B., Zhang, W., Wudzinska, A., Huang, E., Brosh, R., Pour, M., Miller, A., Dasen, J.S., Maurano, M.T., Kim, S.Y., Boeke, J.D., Yanai, I., 2021. The genetic basis of tail-loss evolution in humans and apes. https://doi.org/10.1101/2021.09.14.460388
- Yu, T., Huang, X., Dou, S., Tang, X., Luo, S., Theurkauf, W.E., Lu, J., Weng, Z., 2021. A benchmark and an algorithm for detecting germline transposon insertions and measuring de novo transposon insertion frequencies. Nucleic Acids Res 49, e44. https://doi.org/10.1093/nar/gkab010.
- Zhou, M., Jr, S.K.A., n.d. A Review Of Published Work On Crassostrea Ariakensis.
- Zhou, Y., Minio, A., Massonnet, M., Solares, E., Lv, Y., Beridze, T., Cantu, D., Gaut, B.S., 2019. The population genetics of structural variants in grapevine domestication. Nat. Plants 5, 965–979. https://doi.org/10.1038/s41477-019-0507-8

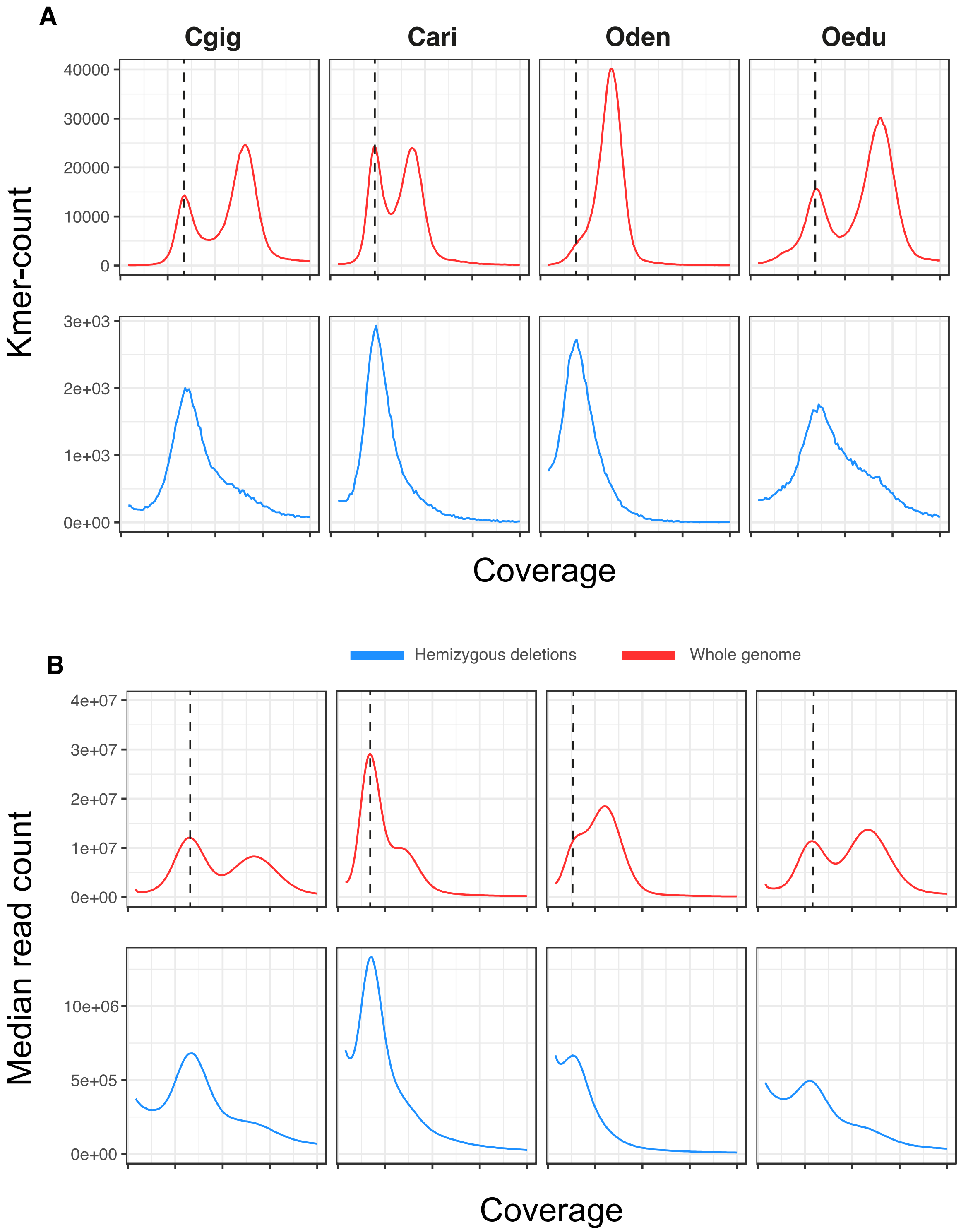
Supplementary Figures


Outline of the approach used to simulate hemizygous variants
C. ariakensis real genome
TRUE SET: 1,000 randomly chosen DEL from previously called set.
VCF
Simulate reads from ref. genome with CCS, ONT and RSII error profiles
15X
SVs calling and comparison with True Set
C. ariakensis simulated genome
30X
15X
Simulate reads from simulated genome with CCS, ONT and RSII error profiles. Reads will carry the simulated SVs
Set of reads carrying simulated SVs in an heterozygous state

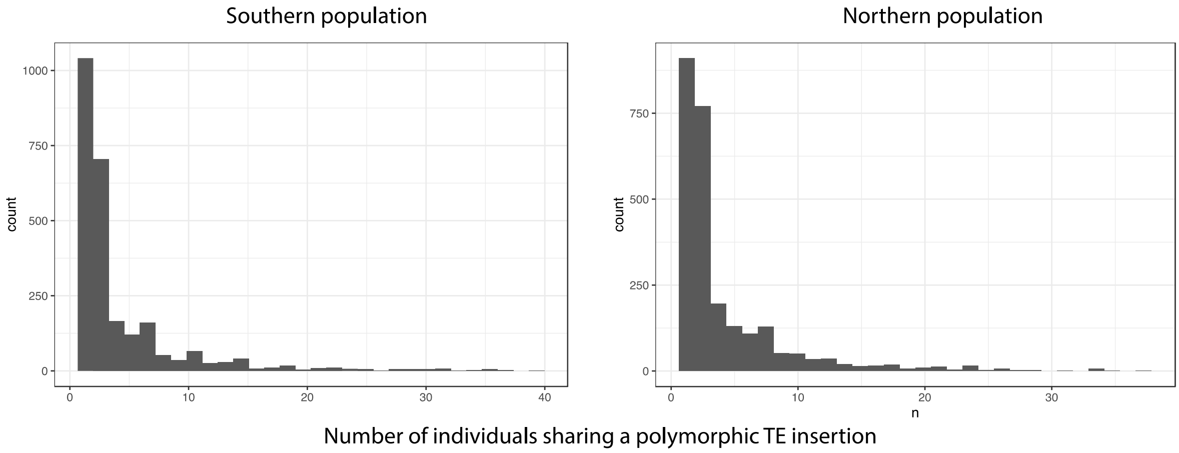
Sup. Fig. 1: Outline of the pipeline used to benchmark our within-individual SV calling pipeline.



Sup. Fig. 2: Intersections of the different SV sets obtained by each combination of SV caller and long read mapper. Cgig = *C. gigas*; Cari = *C. ariakensis*; Oden = *O. denselamellosa*; Oedu = *O. edulis*.



Sup. Fig. 3: (**A**) median genome wide short read read count over sliding windows of 1Mb and over hemizygous deletions. (**B**) Comparison between kmer count across the whole genome and considering hemizygous deletions only.Cgig = *C. gigas*; Cari = *C. ariakensis*; Oden = *O. denselamellosa*; Oedu = *O. edulis*.



Sup. Fig. 4: Frequency of number of individuals sharing a non-reference TE insertion in 107 C. ariakensis wild samples coming from the Chinese coast.

Supplementary Tables

Sup. Tab. 1: NCBI accession number and assembly statistics about the four genomes used for within-individual SV discovery and heterozygosity estimation.

Specie	Reads type	Contig N50	Scaffold N50	Assembly size	Accession
Crassostrea gigas	CLR	1.6 Mb	58.5 Mb	648 Mb	PRJNA629593 (NCBI)
Crassostrea ariakensis	CCS	5.9 Mb	66.3 Mb	663 Mb	CNP0001149 (CNGBdb)
Ostrea edulis	PromethION	1.8 Mb	95.6 Mb	935 Mb	PRJNA77211 (NCBI)
Ostrea denselamellosa	CCS	14 Mb	/	636 Mb	PRJNA838121 (NCBI)

Sup. Tab. 2: Metadata about the 107 *C. ariakensis* re-sequenced wild samples coming from the Chinese coast and downloaded from NCBI. Genome coverage was estimated after mapping the reads on the reference genome.

Biosample	SRA	Sample name	Genome coverage
SAMN19486947	SRR14865146	P591-R01	15.47
SAMN19486948	SRR14865145	P591-R02	16.98
SAMN19486949	SRR14865034	P591-R03	17.43
SAMN19486950	SRR14864959	P591-R04	17.11
SAMN19486951	SRR14864948	P591-R05	17.48
SAMN19486952	SRR14864937	P591-R06	17.32
SAMN19486953	SRR14864926	P591-R07	18.08
SAMN19486954	SRR14864915	P591-R08	17.66
SAMN19486956	SRR14864893	P591-R10	17.45
SAMN19487016	SRR14864916	P591-R101	17.45
SAMN19487019	SRR14864912	P591-R104	19.09
SAMN19487020	SRR14864911	P591-R105	17.47
SAMN19487029	SRR14864901	P591-R116	17.14
SAMN19486958	SRR14865133	P591-R12	19.26
SAMN19487033	SRR14864897	P591-R120	17.42
SAMN19487038	SRR14864891	P591-R125	17.52
SAMN19487040	SRR14864889	P591-R127	19.38
SAMN19487041	SRR14864888	P591-R128	21.64
SAMN19487043	SRR14864886	P591-R130	20.61
SAMN19487045	SRR14864884	P591-R132	26.41
SAMN19487046	SRR14864883	P591-R133	25.60

SAMN19487047	SRR14865143	P591-R134	19.91
SAMN19487050	SRR14865140	P591-R137	21.63
SAMN19486960	SRR14865111	P591-R14	18.10
SAMN19487056	SRR14865134	P591-R143	18.73
SAMN19487057	SRR14865132	P591-R144	23.04
SAMN19487058	SRR14865131	P591-R145	21.84
SAMN19487059	SRR14865130	P591-R146	28.43
SAMN19487060	SRR14865129	P591-R147	23.64
SAMN19487062	SRR14865127	P591-R149	22.68
SAMN19487063	SRR14865126	P591-R150	19.67
SAMN19487064	SRR14865125	P591-R151	21.97
SAMN19487065	SRR14865124	P591-R152	19.96
SAMN19487066	SRR14865123	P591-R153	25.60
SAMN19487067	SRR14865121	P591-R154	24.24
SAMN19487068	SRR14865120	P591-R155	21.37
SAMN19487069	SRR14865119	P591-R156	20.46
SAMN19487071	SRR14865117	P591-R158	24.21
SAMN19487082	SRR14865105	P591-R169	22.24
SAMN19487089	SRR14865097	P591-R176	18.51
SAMN19487090	SRR14865096	P591-R177	20.75
SAMN19487093	SRR14865093	P591-R180	18.71
SAMN19487095	SRR14865091	P591-R182	25.62
SAMN19487096	SRR14865090	P591-R183	21.94
SAMN19487101	SRR14865084	P591-R188	23.35
SAMN19487102	SRR14865083	P591-R189	20.89
SAMN19486965	SRR14865056	P591-R19	18.31
SAMN19487104	SRR14865081	P591-R191	20.04
SAMN19487109	SRR14865075	P591-R196	22.80
SAMN19487110	SRR14865074	P591-R197	21.47
SAMN19487112	SRR14865072	P591-R199	19.43
SAMN19486966	SRR14865045	P591-R20	17.76
SAMN19487114	SRR14865070	P591-R201	20.99
SAMN19487116	SRR14865068	P591-R203	17.91
SAMN19487119	SRR14865064	P591-R213	20.01
SAMN19487120	SRR14865063	P591-R214	20.24
SAMN19486968	SRR14865022	P591-R22	18.20
SAMN19487122	SRR14865061	P591-R220	19.60
SAMN19487123	SRR14865060	P591-R221	18.40
SAMN19487124	SRR14865059	P591-R222	21.45
SAMN19487126	SRR14865057	P591-R224	18.84
SAMN19487129	SRR14865053	P591-R227	18.40
SAMN19486969	SRR14865011	P591-R23	17.21

SAMN19487132	SRR14865050	P591-R230	19.69
SAMN19487135	SRR14865047	P591-R233	19.53
SAMN19487139	SRR14865042	P591-R237	24.52
SAMN19487140	SRR14865041	P591-R238	24.20
SAMN19487141	SRR14865040	P591-R239	18.21
SAMN19487142	SRR14865039	P591-R240	20.03
SAMN19487143	SRR14865038	P591-R241	19.87
SAMN19487145	SRR14865036	P591-R243	17.88
SAMN19487147	SRR14865032	P591-R245	20.03
SAMN19487150	SRR14865029	P591-R248	19.63
SAMN19487151	SRR14865028	P591-R249	19.71
SAMN19487155	SRR14865024	P591-R253	19.52
SAMN19487156	SRR14865023	P591-R254	22.54
SAMN19487160	SRR14865018	P591-R258	20.40
SAMN19487162	SRR14865016	P591-R261	16.94
SAMN19487163	SRR14865015	P591-R262	21.30
SAMN19487164	SRR14865014	P591-R263	19.78
SAMN19487165	SRR14865013	P591-R264	18.35
SAMN19486973	SRR14864967	P591-R27	17.79
SAMN19487171	SRR14865006	P591-R271	23.31
SAMN19487174	SRR14865003	P591-R274	19.85
SAMN19487176	SRR14865001	P591-R276	20.14
SAMN19487179	SRR14864997	P591-R279	19.11
SAMN19487180	SRR14864996	P591-R280	25.07
SAMN19487181	SRR14864995	P591-R281	17.99
SAMN19487183	SRR14864993	P591-R283	20.40
SAMN19487186	SRR14864990	P591-R286	24.70
SAMN19486975	SRR14864961	P591-R29	19.09
SAMN19487191	SRR14864984	P591-R291	18.32
SAMN19487193	SRR14864982	P591-R293	22.78
SAMN19487195	SRR14864980	P591-R295	21.13
SAMN19487197	SRR14864977	P591-R297	23.63
SAMN19487205	SRR14864969	P591-R305	23.56
SAMN19486982	SRR14864953	P591-R67	19.65
SAMN19486991	SRR14864943	P591-R76	18.18
SAMN19486994	SRR14864940	P591-R79	18.89
SAMN19486997	SRR14864936	P591-R82	17.71
SAMN19486998	SRR14864935	P591-R83	16.94
SAMN19487000	SRR14864933	P591-R85	19.23
SAMN19487004	SRR14864929	P591-R89	17.59
SAMN19487007	SRR14864925	P591-R92	19.79
SAMN19487008	SRR14864924	P591-R93	17.68

SAMN19487011	SRR14864921	P591-R96	17.30
SAMN19487013	SRR14864919	P591-R98	17.15

Sup. Tab. 3: Percentage of genome covered by each TE class as estimated by RepeatMasker for the four analysed oyster species.

	C. ariakensis	*C. gigas*	*O. edulis*	*O. denselamellosa*
DNA	22.71%	22.09%	15.31%	15.19%
RC	16.02%	14.60%	11.86%	8.02%
SINE	0.16%	0.13%	2.40%	4.15%
LINE	4.05%	3.63%	11.32%	10.24%
LTR	5.57%	6.62%	5.80%	3.67%
Unknown	6.25%	5.25%	14.43%	13.04%
Satellite	0.13%	0.13%	0.09%	0.18%
Total	54.89%	52.32%	61.12%	54.31%

Sup. Tab. 4: Number of non-reference TE insertions identified for each RepeatMasker TE type across the 107 re-sequenced *C. ariakensis* samples.

Type	Class	N. of de-novo insertions
Academ-1	DNA	4
Academ-2	DNA	55
CMC-EnSpm	DNA	1071
Copia	LTR	38
CR1	LINE	10
CR1-Zenon	LINE	535
Crypton	DNA	538
Crypton-A	DNA	3429
Crypton-V	DNA	8
DIRS	LTR	280
ERVL	LTR	13
Ginger-1	DNA	17
Gypsy	LTR	7000
hAT	DNA	31
hAT-Ac	DNA	267
hAT-Blackjack	DNA	403
hAT-Tip100	DNA	414

Helitron	RC	10961
ID	SINE	16
IS3EU	DNA	219
Kolobok	DNA	1111
Kolobok-Hydra	DNA	174
Kolobok-T2	DNA	229
L1	LINE	1
L1-Tx1	LINE	124
L2	LINE	2
Maverick	DNA	28
Merlin	DNA	3
MIR	SINE	125
MULE-MuDR	DNA	79
Ngaro	LTR	268
Pao	LTR	996
Penelope	LINE	696
PIF-Harbinger	DNA	423
PIF-ISL2EU	DNA	62
PIF-Spy	DNA	2059
PiggyBac	DNA	89
R2	LINE	1
Rex-Babar	LINE	27
RTE-BovB	LINE	13
RTE-X	LINE	631
Sola-1	DNA	29
Sola-2	DNA	48
Sola-3	DNA	33
TcMar	DNA	46
TcMar-Fot1	DNA	4
TcMar-m44	DNA	205
TcMar-Mariner	DNA	55
TcMar-Pogo	DNA	1069
TcMar-Tc1	DNA	482
TcMar-Tc2	DNA	311
TcMar-Tigger	DNA	643
tRNA	SINE	2
tRNA-V	SINE	129
Unknown	DNA	18037
Unknown	LTR	1591
Zator	DNA	249

7. Final Considerations

The diversity of living and non-living forms has always fascinated humans. Every child widens its eyes in wonder at the sight of the incredible phenotypic variability of butterflies or birds, as well as the incredible geological formations that millennia of erosion can create. I am no exception, and while as a child I collected insects in the backyard admiring their shapes and colours, during my cycles of study, I have become increasingly struck by the extreme variability of genomes, in composition and structure. This genome variability arises from random mutagenic processes and over hundreds of thousands of years, neutral and non-neutral evolutionary process may act as a molecular tinkerer (Francis Jacob, 1977), shaping the genomic landscape. Therefore, understanding the 'biodiversity' of genomes and the evolutionary forces that govern it represents a fundamental step towards a deeper understanding of organismal biodiversity.

In my thesis, I have explored the evolution, diversity, and genomic impact of one of the most important sources of mutations in eukaryotes, transposable elements (TEs), in bivalves, a frequently overlooked but rich and diverse metazoan group. Through bioinformatic work carried out during the development of the four chapters of my thesis and additional research activities over the past three years, I have gained a solid understanding of modern concepts and tools for genomics and evolutionary genomics, including genome assembly, TE annotation, RNAseq analyses, variant calling, and population genomics.

Since high-quality TE annotation requires high-quality genomes, I initially contributed to the scientific community by releasing the first high-quality long-read-based genome assembly for the highly heterozygous Manila clam *Ruditapes philippinarum*. Subsequently, I explored the diversity of transposons hosted in bivalve genomes across a wide taxonomic scale, revealing that most of the known transposon superfamilies are represented in bivalves, contrary to other analyzed molluscs. While these results are robust, it is evident that the rapid increase in mollusc genomic resources will necessitate future re-analyses to validate the observed distribution patterns. Additionally, I provided initial insights into the evolutionary trajectory of LINEs and SINEs, focusing on their commonalities and differences with other eukaryotic clades, and their potentially significant role in shaping bivalve genome evolution. The presence of multiple putatively active transposon families raises interesting questions about possible self-controlling mechanisms and relationships with host suppression machinery, enabling multiple transposons to survive for long evolutionary timescales, and why this phenomenon is not observed across all organisms. Finally, I investigated genome structural variants (SVs), their relationship with transposons, and their genomic impact at a narrower taxonomic scale, within an economically and ecologically important clade of bivalves. Here, I highlighted how TE-derived structural variants are primary contributors to the observed high heterozygosity of oysters, with *de novo* TE insertions outnumbering genomic deletions, potentially providing an important substrate for genetic variability and local adaptations. Our results highlight how also Illumina short-reads can still be used to obtain a high-quality set of SVs and polymorphic TE insertions when population-level long read data are not available, encouraging therefore their analyses in future population genomic studies.

Importantly, during these projects I have generated hundreds of novel high quality TE consensus sequences, that are now, or will be soon, deposited and freely available under the curated section of the DFAM database. I hope that these projects can be inspiring, both conceptually and methodologically, for anyone wishing to explore the 'biodiversity' of genomes and, at the same time, contribute to the growing community of scientists studying non-model organisms. Nowadays, novel genomes coming from all branches of the tree of life are sequenced at an unprecedent scale, but we know only little about their structure and composition. The post-genomics era is giving an amazing opportunity to scientist to understand genome composition and evolution, we do not have to miss it.