The Crown Pearl: a draft genome assembly of the European freshwater pearl mussel *Margaritifera margaritifera* (Linnaeus, 1758)

André Gomes-dos-Santos[1,2*], Manuel Lopes-Lima[1,3,4*], André M. Machado[1], António Marcos Ramos[5,15], Ana Usié[5,15], Ivan N. Bolotov[6], Ilya V. Vikhrev[6], Sophie Breton[7], L. Filipe C. Castro[1,2], Rute R. da Fonseca[8], Juergen Geist[9], Martin E. Österling[10], Vincent Prié[11], Amílcar Teixeira[12], Han Ming Gan[13], Oleg Simakov[14], Elsa Froufe[1*]

[1] CIIMAR/CIMAR — Interdisciplinary Centre of Marine and Environmental Research, University of Porto, Terminal de Cruzeiros do Porto de Leixões, Avenida General Norton de Matos, S/N, P 4450-208 Matosinhos, Portugal; [2] Department of Biology, Faculty of Sciences, University of Porto, Rua do Campo Alegre 1021/1055, 4169-007 Porto, Portugal; [3] CIBIO/InBIO - Research Center in Biodiversity and Genetic Resources, Universidade do Porto, Campus Agrário de Vairão, Rua Padre Armando Quintas, 4485-661 Vairão, Portugal; [4] IUCN SSC Mollusc Specialist Group, c/o IUCN, David Attenborough Building, Pembroke St., Cambridge, England; [5] Centro de Biotecnologia Agrícola e Agro-alimentar do Alentejo (CEBAL) / Instituto Politécnico de Beja (IPBeja), Beja 7801-908 Beja, Portugal; [6] Federal Center for Integrated Arctic Research, Russian Academy of Sciences, Arkhangelsk, Severnoy Dviny emb. 23 163000 Russia; [7] Department of Biological Sciences, University of Montreal, Campus MIL, Montreal, Canada; [8] Center for Macroecology, Evolution and Climate, GLOBE Institute, University of Copenhagen, 2100 Copenhagen, Denmark; [9] Aquatic Systems Biology Unit, Technical University of Munich, TUM School of Life Sciences, Mühlenweg 22, D-85354 Freising, Germany; [10] Department of Environmental and Life Sciences – Biology, Karlstad University, Universitetsgatan 2, 651 88 Karlstad, Sweden; [11] Research Associate, Institute of Systematics, Evolution, Biodiversity (ISYEB), National Museum of Natural History (MNHN), CNRS, SU, EPHE, UA CP 51, 57 rue Cuvier, 75005 Paris, France; [12] Centro de Investigação de Montanha (CIMO), Instituto Politécnico de Bragança, Bragança, Portugal; [13] GeneSEQ Sdn Bhd, Bandar Bukit Beruntung, Rawang 48300 Selangor, Malaysia; [14] Department of Neurosciences and Developmental Biology, University of Vienna, Universitätsring 1, 1010 Wien, Vienna, Austria; [15] MED — Mediterranean Institute for Agriculture, Environment and Development, CEBAL — Centro de Biotecnologia Agrícola e Agro-Alimentar do Alentejo, 7801-908 Beja, Portugal

* To whom correspondence should be addressed. Tel. +351223401889.

andrepousa64@gmail.com; elsafroufe@gmail.com; lopeslima.ciimar@gmail.com

The Crown Pearl: a draft genome assembly of the European freshwater pearl mussel *Margaritifera margaritifera* (Linnaeus, 1758)

Abstract

Since historical times, the inherent human fascination with pearls turned the freshwater pearl mussel *Margaritifera margaritifera* (Linnaeus, 1758) into a highly valuable cultural and economic resource. Although pearl harvesting in *M. margaritifera* is nowadays residual, other human threats have aggravated the species conservation status, especially in Europe. This mussel presents a myriad of rare biological features, e.g. high longevity coupled with low senescence and Doubly Uniparental Inheritance of mitochondrial DNA, for which the underlying molecular mechanisms are poorly known. Here, the first draft genome assembly of *M. margaritifera* was produced using a combination of Illumina Paired-end and Mate-pair approaches. The genome assembly was 2,4 Gb long, possessing 105,185 scaffolds and a scaffold N50 length of 288,726 bp. The *ab initio* gene prediction allowed the identification of 35,119 protein-coding genes. This genome represents an essential resource for studying this species' unique biological and evolutionary features and ultimately will help to develop new tools to promote its conservation.

1. Introduction

Pearls are fascinating organic gemstones that have populated the human beauty imaginary for millennia. Legend says that the famous Egyptian ruler Cleopatra, to display her wealth to her lover Marc Antony, dissolved a pearl in a glass of vinegar and drank it. The human use of pearls or their shell precursor material, the mother-of-pearl (nacre), is ancient. The earliest known use of decorative mother-of-pearl dates to 4200 BC in Egypt, with pearls themselves only becoming popular around 600 BC. Before the arrival of marine pearls to Europe, most were harvested from a common and widespread freshwater bivalve, the freshwater pearl mussel *Margaritifera margaritifera* L. 1758, where generally one pearl is found per 3,000 mussels leading to massive mortality [1]. In Europe, during the Roman Empire period, pearls were a desirable luxury, so that it is believed that one of the reasons that persuaded Julius Caesar to invade Britain was to access its vast freshwater pearl resources [2]. *Margaritifera margaritifera* freshwater pearls were extremely valuable being included in many royal family jewels such as the British, Scottish, Swedish, Austrian and German crown jewels and even in the Russian city's coat of arms [2–5]. Although over-harvesting represented a serious threat to the species for centuries (mostly in Europe and Russia), there has been a decrease in interest and demand for freshwater pearls in the 20th century [4]. However, the global industrialization process introduced stronger threats to the survival of the species [6–8]. In fact, *M. margaritifera* belongs to one of the most threatened taxonomic groups on earth, the Margaritiferidae [6]. The species was once abundant in cool oligotrophic waters throughout most of northwest Europe and northeast North America [6–8]. However, habitat degradation, fragmentation and pollution have resulted in massive population declines [8]. Consequently, the Red List of Threatened Species from the International Union for Conservation of Nature (IUCN) has classified *M. margaritifera* as Endangered globally and Critically Endangered in Europe [7,9]. Population declines are particularly concerning in Europe, where a lot of investment has been done in rehabilitation and propagation projects aimed at improving the species conservation status [9,10]. North America and Russia seem to be able to control populations sizes by maintaining more isolated and less threat exposed populations [7].

Besides being able to produce pearls, *M. margaritifera* presents many other remarkable biological characteristics, e.g. is among the most longest-living invertebrates, reaching up to 280 years [6,11]; displays very weak signs of senescence, referred as the concept of "*negligible senescence*" [12]; has an obligatory parasitic larval stage on salmonid fishes used for nurturing and dispersion [8,10]; and, like many other bivalves (see Gusman et al. [13] for a recent enumeration), shows an unusual mitochondrial DNA inheritance system, called Doubly Uniparental Inheritance or DUI [14,15]. Although these biological features are well described, the molecular mechanisms underlying their regulation and functioning are poorly studied and practically unknown.

Thus, a complete genome assembly for *M. margaritifera* is critical for developing the molecular resources required to improve our knowledge of such mechanisms. These resources can then be used in multiple fields, such as in conservation biology (e.g. to overcome the main bottlenecks of propagation programs); in freshwater pearl production industry (e.g. to better understand biomineralization mechanisms); in biomedical applications (e.g. to study bone regeneration); and in ageing and senescence studies among others.

In the last decade, the decreasing cost of next generation sequencing, coupled with improved bioinformatic tools, has facilitated the generation of genomic resources for non-model organisms. Several Mollusca genomes are currently available and new assemblies are released every year at an increasing trend (reviewed in [16–18]). Despite this, to date, only two Unionida mussel genomes have been published, those of *Venustaconcha ellipsiformis* (Conrad, 1836) [19] and *Megalonaias nervosa* (Rafinesque, 1820) [20]. These represent valuable comparative resources and are among the largest bivalve genomes sequenced to date (1.80 Gb and 2.36 Gb respectively) [16,18]; known Bivalvia genome sizes range from 0.559 Gb (*Crassostrea gigas*, Ostreida [21]) to 2.38 Gb (*Modiolus philippinarum*, Mytiloida [22]). Unlike many other bivalves (e.g. [23,24]), both *V. ellipsiformis* and *M. nervosa* genomes revealed relatively low levels of heterozygosity, with estimated values per site of 0.0060 and 0.0077, respectively [19,20].

The current study presents the first draft genome assembly of the freshwater pearl mussel *M. margaritifera*. The genome assembly was performed combining Illumina Paired-End short reads with Illumina Mate-Paired reads. The assembled genome has a total length of 2,4 Gb distributed throughout 105,185 scaffolds, with a GC content of 35.42% and scaffold N50 length of 288,726 bp. More than half of the genome was found to be composed of repetitive elements (i.e. 59.07%) and 35,119 protein coding genes were predicted in this initial annotation.

2. Material and Methods

2.1. Sample collection, DNA extraction and sequencing

One *M. margaritifera* (Linnaeus, 1758) specimen was collected from the River Tua, Douro basin in the North of Portugal (permit 284/2020/CAPT and fishing permit 26/20 issued by ICNF - Instituto de Conservação da Natureza e das Florestas). The whole individual is stored in 96% ethanol at the Unionoid DNA and Tissue Databank, CIIMAR, University of Porto. Genomic DNA (gDNA) was extracted from the foot tissue using DNeasy Blood and Tissue Kit (Qiagen, Hilden, Germany) according to the manufacturer's instructions.

Two distinct NGS libraries and sequencing approaches were implemented i.e. Illumina Paired-end reads (PE) and Illumina long insert size Mate-pair reads (MP). For Illumina PE library preparation, approximately 100ng of gDNA as measured using Qubit Broad-Range Kit (Invitrogen, Santa Clara, CA, USA) was sheared to approximately 300-400 bp using the Qsonica Q800R system (Qsonica, Newton, CT, USA). The sheared DNA was then prepared for sequencing using NEB Ultra DNA kit with standard Illumina adapter and sequenced in an Illumina machine NovaSEQ6000 system located at Deakin Genomics Centre using a run configuration of 2×150 bp. Illumina MP library preparation and sequencing were performed by Macrogen Inc., Korea, where a 10kb insert size Nextera Mate Pair Library was constructed and subsequently sequenced in a NovaSeq6000 S4 using a run configuration of 2x150 bp.

2.2. Genome size and heterozygosity estimation

The overall characteristics of the genome were accessed using PE reads. Firstly, the general quality of the reads was evaluated with FastQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/). The raw reads were then quality trimmed with Trim Galore v.0.4.0 (https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/), allowing the trimming of adapter sequences and removal of low-quality reads using Cutadapt [25] followed by another quality check with FastQC. Clean reads were used for genome size estimation in a two-step approach: (i) using Jellyfish v.2.2.10 for counting and histogram construction of k-mer frequency distributions, with k-mers length of 25 and 31 and (ii) using histograms of frequency distribution to estimate the genome size, heterozygosity rate and repeat content of the genome, with GenomeScope2 [26,27].

2.3. Genome assembly and quality assessment

Long range Illumina mate paired reads quality processing was as described above and both PE and MP cleaned reads were used for whole genome assembly. The assembly was produced by running Meraculous v.2.2.6 with several distinct k-mer sizes (meraculoususing) [28]. This allowed determining the optimal kmer size of 101. Genome assembly metrics were estimated using QUAST v5.0.2 [29]. Assembly completeness, heterozygosity and collapsing of repetitive regions were evaluated through analysis of k-mer distribution using PE reads, with the tool "KAT comp" from the K-mer analysis toolkit [30]. Furthermore, PE reads were aligned to the genome assembly using BBMap [31]. BUSCO v. 3.0.2 [32] was used to provide a quantitative measure of the assembly completeness, with a curated list (i.e. OrthoDB) of near-universal single-copy orthologs. Here, both eukaryotic (303 single-copy orthologs) and metazoan (978 single-copy orthologs) libraries profiles were used to test the genome assembly completeness.

2.4. Repeat Sequences and Gene Models predictions

Given the generally high composition of repetitive elements in Mollusca genomes (e.g. [16]) they should be identified and masked before proceeding to genome annotation. An annotated library

of repetitive elements was created for *M. margaritifera* genome assembly, using RepeatModeler v.2.0.1. [33] (excluding sequences <2.5 kb). Afterwards, repetitive elements were soft masked using RepeatMasker v.4.0.7. [34] combining the repetitive elements for the taxa "Bivalvia", from the RepeatMasker database (comprising the databases Dfam_consensus-20170127 and RepBase-20181026), with the newly constructed library for *M. margaritifera* genome.

BRAKER2 pipeline v2.1.5 [35,36] was used for gene prediction in the genome. First, all RNA-seq data of *M. margaritifera* [37,38] available on GenBank were downloaded, assessed with FastQC v.0.11.8 (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/), quality-controlled and uniformized with Trimmomatic v.0.38 [39] (Parameters, LEADING:5 TRAILING:5 SLIDINGWINDOW:4:20 MINLEN:36), the sequencing errors corrected with Rcorrector v.1.0.3. Afterwards, the RNA-seq data was aligned to the masked genome assembly, using Hisat2 v.2.2.0 with the default settings [40]. Secondly, the complete proteomes of 13 mollusc species, one Chordata (*Ciona intestinalis*) and one Echinodermata (*Strongylocentrotus purpuratus*) were downloaded from distinct public databases (Supplementary Table S1) and used as additional evidence for gene prediction. The BRAKER2 pipeline was applied with the parameters (--etpmode; --softmasking; --UTR=off; --crf; --cores=30) and following the authors' instructions [35,36]. The resulting gene predictions (i.e. gff3 file) were renamed, cleaned and filtered using AGAT v.0.4.0 [41], correcting coordinates of overlapping gene prediction, removing predicted coding sequence regions (CDS) with less than 100 amino acid (in order to avoid a high rate of false positive predictions) and removing incomplete gene predictions (i.e. without start and/or stop codons). Functional annotation was first conducted by searching for protein domain information using InterProScan v.5.44.80 [42], and afterwards, a protein blast search was conducted using DIAMOND v. 0.9.32 [43] against SwissProt (Download at 2/07/2020), TREMBL (Download at 2/07/2020) and RefSeq-NCBI (Download at 3/07/2020) [44,45].

2.5. Phylogenetic analyses

For the phylogenetic assessment, the proteomes of 12 molluscan species were downloaded from distinct public databases (Supplementary Table S2) and included 11 Autobranchia bivalves and

two outgroup species, i.e. the Cephalopoda *Octopus bimaculoides* and Gastropoda *Biomphalaria glabrata* (Figure 3). To retrieve single-copy orthologs between these 12 species and *M. margaritifera*, the protein sets were first clustered into families, using OrthoFinder v2.4.0 [46] specifying msa as the method of gene tree inference (-M). The resulting 118 single copy orthologous sequences were individually aligned using MUSCLE v3.8.31 [47], with default parameters and subsequently trimmed with TrimAl v.1.2 [48] specifying a gap threshold of 0.5 (-gt). Trimmed sequences were then concatenated using FASconCAT-G (https://github.com/PatrickKueck/FASconCAT-G). The best molecular evolutionary model was estimated using ProTest v.3.4.1 [49]. Phylogenetic inferences were conducted in IQ-Tree v.1.6.12 [50] for Maximum Likelihood analyses (with initial tree searches followed by ten independent runs and 10000 ultra-bootstrap replicates) and MrBayes v.3.2.6 [51] for Bayesian Inference (two independent runs, 1,000,000 generations, sampling frequency of one tree per 1000 generations). All phylogenetic analyses were applied using the substitution model LG+I+G.

2.6. Hox and ParaHox gene identification and phylogeny

To identify the repertoire Hox and ParaHox genes *in M. margaritifera*, a similarity search by BLASTn [52] of the CDS of *M. margaritifera* genome, was conducted using the annotated homeobox gene set of *Crassostrea gigas* [53,54]. Candidate CDSs were further validated for the presence of the homeodomain by CD-Search [55]. Finally, each putative CDS identity was verified by BLASTx and BLASTp [52] searches in Nr-NCBI nr database and phylogenetic analyses. Since the search was conducted in the annotated genome (i.e. scaffolds over 2.5kb), when genes were not found, a new search was conducted in the remaining scaffolds. At the end, any genes still undetected were search in the Transcriptome assembly of the species (Bioproject: PRJNA369722) [37]. Due to the phylogenetic proximity and for comparative purposes, Hox and ParaHox genes were also searched in the genome assembly of *Megalonaias nervosa* [20].

For phylogenetic assessment of Hox and Parahox genes, amino acid sequences of homeodomain of the genes from *M. margaritifera* and *M. nervosa*, were aligned with other Mollusca orthologs

([56,57] and references within; Supplementary File1). Molecular evolutionary models and Maximum Likelihood phylogenetic analyses were obtained using IQ-TREE v.1.6.12 [50,58].

3. Results and Discussion

3.1. Sequencing Results

A total of 494 Gb (~209x) of raw PE and 76 Gb (~32x) of raw MP data were generated, which after trimming and quality filtering were reduced by 0.3% and 10% respectively (Table 1). GenomeScope2 model fitting of the k-mer distribution analysis estimated a genome size between 2.31-2.36 Gb and very low heterozygosity between 0.127-0.105% (Figure 2). Although larger than the genome of *V. ellipsiformis* (i.e. 1.80 Gb), the size estimation of the *M. margaritifera* genome is in line with the recently assembled Unionida mussel *M. nervosa* [20] (i.e. 2.38 Gb). The estimated heterozygosity is the lowest observed within Unionida genomes [19,20] and one of the lowest in Mollusca [16], which is remarkable considering it refers to a wild individual. This low value is likely a consequence of population bottlenecks during glaciations events, which have been shown to shape the evolutionary history of many freshwater mussels (e.g. [19,59,60]) and may also be enhanced by recent human-mediated threats.

3.2. *Margaritifera margaritifera* de novo genome assembly

The Meraculous assembly and scaffolding yield a final genome size of 2.47 Gb with a contig N50 of 16,899 bp and a scaffold N50 of 288,726 bp (Table 1). Both N50 values are significantly higher than *V. ellipsiformis* genome assembly, i.e. 3,117 bp and 6,523 bp, respectively [19]. Presently, this *M. margaritifera* genome assembly reveals the highest scaffold N50 of the three Unionida genomes currently available [19,20]. On the other hand, *M. nervosa* genome assembly contig N50, i.e. 51,552 bp, is higher than *M. margaritifera*, which is expected given the use of Oxford Nanopore ultra-long reads libraries in the assembly produced by Rogers et al [20]. BUSCOs scores of the final assembly indicate a fairly complete genome assembly (Table 1) and although the contiguity is lower when compared with other recent Bivalve genome assemblies,

the low percentage of fragmented genes (i.e. 5.9% for Eukaryota and 4.9% for Metazoa) gives further support to the quality of the genome assembly. Similarly, the slight difference observed between the genome size and the initial size estimation is unlikely to be a consequence of erroneous assembly duplication, as duplicated BUSCOs scores are also low (i.e. 1% for Eukaryota and 1.1% for Metazoa). The quality of the genome assembly is further supported by the high percentages of PE reads mapping back to the genome (i.e. 97.75%, Table 1), as well as the KAT k-mer distribution spectrum (Figure 3), which demonstrates that almost no read information was excluded from the final assembly. Overall, these statistics indicate that the *M. margaritifera* draft genome assembly here presented is fairly complete, nonredundant, and useful resource for various applications.

3.3. Repeat Identification and Masking and Gene Models Prediction

The use of the custom repetitive library combined with the RepBase [61] "Bivalvia" library, resulted in masking repetitive elements in more than half of the genome assembly, i.e. 59.07% (Table 2). Most of the annotated repetitive elements were unclassified (31.86%), followed by DNA elements (16.00%), long interspersed nuclear elements (LINEs) (6.13%), long terminal repeats (LTRs) (3.72%) and short interspersed nuclear elements (SINEs) (0.79%). After masking, gene prediction resulted in the identification of 35,119 protein-coding genes, with an average gene length of 25,712 bp and average CDS length of 1,287 bp (Supplementary Table S3). Furthermore, 26,836 genes were functionally annotated by similarity to at least one of the three databases used in the annotation (Table 1). The number of predicted genes is in accordance to those observed in other bivalves (and Mollusca) genome assemblies, which although highly variable, in average have around 34,949 predicted genes (calculated from Table 2 of Gomes-dos-Santos et al. [16]). Although the number of genes predicted within the three Unionida genomes is highly variable, i.e. 123,457 in *V. ellipsiformis*, 49,149 in *M. nervosa* and 35,119 in *M. margaritifera*, a direct comparison should be taken with caution, given the considerable differences in genome qualities and the different gene predictions strategies applied in the three assemblies.

3.4. Single Copy Orthologous Phylogeny

Both Maximum Likelihood and Bayesian Inference phylogenetic trees revealed the same topology with high support for all nodes (Figure 3). The phylogeny recovered the reciprocal monophyletic groups Pteriomorphia (represented by Orders Ostreida, Mytilida, Pectinida and Arcida) and Heteroconchia (represented by Orders Unionida and Venerida). These results are in accordance with recent comprehensive bivalve phylogenetic studies [38,62–64]. The only difference is observed within Pteriomorphia, where two sister clades are present, one composed by Arcida and Pectinida and the other by Mytilida and Osteida (Figure 3), while accordingly to the most recent phylogenomic studies, Arcida appears basal to all other Pteriomorphia [38,63,64]. It is noteworthy that Arcida and Pectinida clade is the less supported in the phylogeny, which together with the fact that many Pteriomorphia clades are missing in the present study, should explain these discrepant results. Heteroconchia is divided into monophyletic Palaeoheterodonta and Heterodonta (here only represented by two Euheterodonta bivalves). As expected, the two Unionida species, i.e. *M. nervosa* and the newly obtained *M. margaritifera*, are placed within Palaeoheterodonta.

3.5. Hox and ParaHox gene repertoire and phylogeny

Homeobox genes refer to a family of homeodomain-containing transcription factors with important roles in Metazoan development by specifying anterior-posterior axis and segment identity (e.g. [65,66]). Many of these genes are generally found in tight evolutionary conserved physical clusters (e.g. [67,68]). Hox genes are typically arranged into tight physical clusters, showing temporal and spatial collinearity [69]. Consequently, Hox genes provide useful information for understanding the emergence of morphological novelties, understanding the historical evolution of the species, infer ancestral genomic states of genes/clusters and even study genome rearrangements, such as whole-genome duplications (e.g. [65,66,70]). Given the disparate body plans in molluscan classes, the study of Hox cluster composition, organization and gene expression has practically become a standard in Mollusca genome assembly studies [21,22,78–82,57,71–77]. Homeobox genes are divided into four classes, of which the Antennapedia

(ANTP)-class (Hox, ParaHox, NK, Mega-homeobox, SuperHox) is the best studied, particularly the Hox and ParaHox clusters [57,70,75]. The number of genes from these two clusters is relatively well conserved across Lophotrochozoa, with Hox cluster being composed of 11 genes (3 anterior, 6 central and 2 posterior) and ParaHox cluster composed of 3 genes. Although several structural and compositional differences have been observed within Mollusca ANTP-class (e.g. Bivalvia [21], Cephalopoda [72], Gastropoda [74] and Polyplacophora [80], most Bivalvia seem to retain the gene composition expected for lophotrochozoans: Hox1, Hox2, Hox3, Hox5, Lox, Antp, Lox4, Lox2, Post2, Post1 for the Hox cluster and Gsx, Xlox, Cdx for the ParaHox cluster [81]. Consequently, the identification of these genes on a bivalve genome assembly represent further validation of the genome completeness and overall correctness. Furthermore, to the best of our knowledge, this study reports for the first time the Hox and ParaHox genes were identified Unionida. A single copy of the 3 ParaHox and 10 Hox genes were found in the *M. margaritifera* genome assembly (Supplementary Table S4). Despite an intensive search, no evidence of the presence of Hox4 was detected. However, the gene was identified in the *M. margaritifera* transcriptome, thus confirming its presence in the species. All genes, apart from *Antp* and Lox5, were scattered in different scaffolds, with Hox5, Post1 and Gsx being present in scaffolds smaller than 2.5kb (Supplementary Table S4). Both the small proximity between Antp and Lox5 and the fact that both genes are expressed in the same direction are in accordance with the results observed in other bivalves, including in the phylogenetically closest species (from which Hox cluster has been characterized), i.e. the Venerida clam *Cyclina sinensis* (Gmelin, 1791) [57]. The fact that the remaining genes were scattered in the different scaffolds is likely a consequence of the low contiguity of the genome assembly since the distances between Bivalvia Hox genes within a cluster can be as high as 9.9 Mb [57]. Conversely, 3 Hox and 1 ParaHox genes were found in the *M. margaritifera* transcriptome assembly and 9 Hox and 1 ParaHox gene were found in *M. nervosa* genome assembly (Supplementary Table S4). Finally, to further validate the identity of the identified Hox and ParaHox genes, a phylogenetic analysis using the homeodomains (encoded 60–63 amino acid domain) of several Mollusca species was conducted (Figure 5). All Hox and ParaHox genes of *M. margaritifera* (as well as *M. nervosa*) were well

positioned within their respective orthologous genes from other Mollusca species (Figure 4), thus confirming their identity.

3.6. Conclusion and future perspectives

Unionida freshwater mussels are a worldwide distributed and diverse group of organisms with 6 recognized families and around 800 described species [83,84]. These organisms play fundamental roles in ecosystems, such as water filtration, nutrient cycling and sediment bioturbation and oxygenation [85,86], allowing to maintain and support freshwater communities [10]. However, as a consequence of several anthropogenic threats, freshwater mussels are experiencing a global-scale decline [10,87]. *Margaritifera margaritifera* belongs to the most threatened of the 6 Unionida families, i.e. Margaritiferidae. Despite all this, our understanding of the genetics of this species is still to date restricted to a few mtDNA markers phylogenetic and restricted phylogeographical studies [6,88–90] as well as neutral genetic markers (SSR) [89,91,92], making the availability of the present genome a timely resource. Being the first representative genome of the family Margaritiferidae, it will help launch both basic and applied genomic-level research on the unique biological and evolutionary features characteristic of this emblematic group.

Funding

AGS was funded by the Portuguese Foundation for Science and Technology (FCT) under the grant SFRH/BD/137935/2018. This research was developed under ConBiomics: the missing approach for the Conservation of freshwater Bivalves Project No NORTE-01-0145-FEDER-030286, co-financed by COMPETE 2020, Portugal 2020 and the European Union through the ERDF, and by FCT through national funds. Additional strategic funding was provided by FCT UIDB/04423/2020 and UIDP/04423/2020. Authors interaction and writing of the paper was promoted and facilitated by the COST Action CA18239: CONFREMU - Conservation of freshwater mussels: a pan-European approach.

Data Availability

All the raw sequencing data are available from GenBank via the accession numbers SRR13091478, SRR13091479 and SRR13091477. The assembled genomes are available in the assession number JADWMO000000000, under the BioProject PRJNA678877 and BioSample SAMN16815977 (Supplementary Table S5). Fasta alignment of homeodomain amino acid sequences from Hox and ParaHox genes used in gene tree construction is available in Additional File 1. The scaffolds in which homeodomains were detected (as described in Supplementary Table S4) are available as Supplementary File 2. The repeat masked genome assembly, BRAKER2 prediction statistic and prediction gff files, as well as all predicted genes, transcripts and amino acid sequence files are available at Figshare: https://doi.org/10.6084/m9.figshare.13333841

Conflict of interest

None declared.

References

1. Hessling, T. von. 1859, *Die Perlnmuscheln und thre Perlen (Naturwissen-schaftlich und geschichtlich mit, Berueecksichtigung der Perlgewaesser Baerns)*. Forgotten Books, Leipzig.

2. Strack, E. 2015, European Freshwater Pearls: Part 1-Russia. *J. Gemmol.*, **34**, 580–92.

3. Bespalaya, Y. V., Bolotov, I. N., Makhrov, A. A., and Vikhrev, I. V. 2012, Historical geography of pearl fishing in rivers of the Southern White Sea Region (Arkhangelsk Oblast). *Reg. Res. Russ.*, **2**, 172–81.

4. Makhrov, A., Bespalaya, J., Bolotov, I., et al. 2014, September 17, Historical geography of pearl harvesting and current status of populations of freshwater pearl mussel *Margaritifera margaritifera* (L.) in the western part of Northern European Russia. *Hydrobiologia*. Springer International Publishing, pp. 149–59.

5. Schlüter, J., and Rätsch, C. 1999, *Perlen und Perlmutt*. Ellert und Richter, Hamburg.

6. Lopes-Lima, M., Bolotov, I. N., Do, V. T., et al. 2018, Expansion and systematics redefinition of the most threatened freshwater mussel family, the Margaritiferidae. *Mol. Phylogenet. Evol.*, **127**, 98–118.

7. Moorkens, E., Cordeiro, J., Seddon, M., von Proschwitz, T., and Woolnough, D. 2018, *Margaritifera margaritifera* (errata version published in 2018). *IUCN Red List Threat. Species 2018*, e.T12799A128686456.

8. Geist, J. 2010, May 7, Strategies for the conservation of endangered freshwater pearl mussels (*Margaritifera margaritifera* L.): A synthesis of conservation genetics and ecology. *Hydrobiologia*. Springer Netherlands, pp. 69–88.

9. Moorkens, E. A. 2018, Short-term breeding: releasing post-parasitic juvenile *Margaritifera* into ideal small-scale receptor sites: a new technique for the augmentation of declining populations. *Hydrobiologia*, **810**, 145–55.

10. Lopes-Lima, M., Sousa, R., Geist, J., et al. 2017, Conservation status of freshwater mussels in Europe: state of the art and future challenges. *Biol. Rev.*, **92**, 572–607.

11. Bauer, G. 1992, Variation in the Life Span and Size of the Freshwater Pearl Mussel. *J. Anim. Ecol.*, **61**, 425.

12. Hassall, C., Amaro, R., Ondina, P., Outeiro, A., Cordero-Rivera, A., and San Miguel, E. 2017, Population-level variation in senescence suggests an important role for temperature in an endangered mollusc. *J. Zool.*, **301**, 32–40.

13. Gusman, A., Lecomte, S., Stewart, D. T., Passamonti, M., and Breton, S. 2016, Pursuing the quest for better understanding the taxonomic distribution of the system of doubly uniparental inheritance of mtdna. *PeerJ*, **2016**, e2760.

14. Breton, S., Stewart, D. T., Shepardson, S., et al. 2011, Novel Protein Genes in Animal mtDNA: A New Sex Determination System in Freshwater Mussels (Bivalvia: Unionoida)? *Mol. Biol. Evol.*, **28**, 1645–59.

15. Breton, S., Beaupré, H. D., Stewart, D. T., Hoeh, W. R., and Blier, P. U. 2007, The unusual system of doubly uniparental inheritance of mtDNA: isn't one enough? *Trends Genet.*, **23**, 465–74.

16. Gomes-dos-Santos, A., Lopes-Lima, M., Castro, L. F. C., and Froufe, E. 2020, November 11, Molluscan genomics: the road so far and the way forward. *Hydrobiologia*. Springer International Publishing, pp. 1705–26.

17. Hollenbeck, C. M., and Johnston, I. A. 2018, Genomic Tools and Selective Breeding in Molluscs. *Front. Genet.*, **9**, 253.

18. Takeuchi, T. 2017, Molluscan Genomics: Implications for Biology and Aquaculture. *Curr. Mol. Biol. Reports*, **3**, 297–305.

19. Renaut, S., Guerra, D., Hoeh, W. R., et al. 2018, Genome Survey of the Freshwater Mussel *Venustaconcha ellipsiformis* (Bivalvia: Unionida) Using a Hybrid De Novo Assembly Approach Martin, B., (ed.), . *Genome Biol. Evol.*, **10**, 1637–46.

20. Rogers, R. L., Grizzard, S. L., Bockrath, K., et al. 2020, Gene family amplification facilitates adaptation in freshwater Unionid bivalve *Megalonaias nervosa*. *arXiv*. 2008.00131v2

21. Zhang, G., Fang, X., Guo, X., et al. 2012, The oyster genome reveals stress adaptation and complexity of shell formation. *Nature*, **490**, 49–54.

22. Sun, J., Zhang, Y., Xu, T., et al. 2017, Adaptation to deep-sea chemosynthetic environments as revealed by mussel genomes. *Nat. Ecol. Evol.*, **1**, 0121.

23. Calcino, A. D., De Oliveira, A. L., Simakov, O., et al. 2019, The quagga mussel genome and the evolution of freshwater tolerance. *DNA Res.*, **26**, 411–22.

24. Uliano-Silva, M., Dondero, F., Dan Otto, T., et al. 2018, A hybrid-hierarchical genome assembly strategy to sequence the invasive golden mussel, *Limnoperna fortunei*. *Gigascience*, **7**, 1–10.

25. Martin, M. 2011, Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*, **17**, 10.

26. Ranallo-Benavidez, T. R., Jaron, K. S., and Schatz, M. C. 2020, GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. *Nat. Commun.*, **11**, 1–10.

27. Vurture, G. W., Sedlazeck, F. J., Nattestad, M., et al. 2017, GenomeScope: fast reference-free genome profiling from short reads Berger, B., (ed.), . *Bioinformatics*, **33**, 2202–4.

28. Chapman, J. A., Ho, I., Sunkara, S., Luo, S., Schroth, G. P., and Rokhsar, D. S. 2011, Meraculous: De novo genome assembly with short paired-end reads Salzberg, S. L., (ed.), . *PLoS One*, **6**, e23501.

29. Gurevich, A., Saveliev, V., Vyahhi, N., and Tesler, G. 2013, QUAST: quality assessment tool for genome assemblies. *Bioinformatics*, **29**, 1072–5.

30. Mapleson, D., Accinelli, G. G., Kettleborough, G., Wright, J., and Clavijo, B. J. 2017, KAT: A K-mer analysis toolkit to quality control NGS datasets and genome assemblies. *Bioinformatics*, **33**, 574–6.

31. Bushnell, B., and Rood, J. 2018, BBTools. BBMap. Joint Genome Institute.

32. Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., and Zdobnov, E. M. 2015, BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, **31**, 3210–2.

33. Smit, A., and Hubley, R. 2015, RepeatModeler. USA:Institute for Systems Biolog, Seattle.

34. Smit, A., and Hubley, R. 2015, RepeatMasker. USA:Institute for Systems Biolog, Seattle.

35. Hoff, K. J., Lange, S., Lomsadze, A., Borodovsky, M., and Stanke, M. 2016,

BRAKER1: Unsupervised RNA-Seq-Based Genome Annotation with GeneMark-ET and AUGUSTUS: Table 1. *Bioinformatics*, **32**, 767–9.

36. Hoff, K. J., Lomsadze, A., Borodovsky, M., and Stanke, M. 2019, Whole-genome annotation with BRAKER. *Methods in Molecular Biology*. Humana Press Inc., pp. 65–95.

37. Bertucci, A., Pierron, F., Thébault, J., et al. 2017, Transcriptomic responses of the endangered freshwater mussel *Margaritifera margaritifera* to trace metal contamination in the Dronne River, France. *Environ. Sci. Pollut. Res.*, **24**, 27145–59.

38. Gonzalez, V. L., Andrade, S. C. S., Bieler, R., et al. 2015, A phylogenetic backbone for Bivalvia: an RNA-seq approach. *Proc. R. Soc. B Biol. Sci.*, **282**, 20142332–20142332.

39. Bolger, A. M., Lohse, M., and Usadel, B. 2014, Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, **30**, 2114–20.

40. Kim, D., Langmead, B., and Salzberg, S. L. 2015, HISAT: A fast spliced aligner with low memory requirements. *Nat. Methods*, **12**, 357–60.

41. Dainat, J., Hereñú, D., and Pucholt, P. AGAT: Another Gff Analysis Toolkit to handle annotations in any GTF/GFF format. Zenodo.

42. Quevillon, E., Silventoinen, V., Pillai, S., et al. 2005, InterProScan: Protein domains identifier. *Nucleic Acids Res.*, **33**, W116–20.

43. Buchfink, B., Xie, C., and Huson, D. H. 2015, Fast and sensitive protein alignment using DIAMOND. *Nat. Methods*, **12**, 59–60.

44. Boeckmann, B. 2003, The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.*, **31**, 365–70.

45. Pruitt, K. D., Tatusova, T., and Maglott, D. R. 2007, NCBI reference sequences (RefSeq): A curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.*, **35**, D61–5.

46. Emms, D. M., and Kelly, S. 2019, OrthoFinder: Phylogenetic orthology inference for comparative genomics. *Genome Biol.*, **20**, 238.

47. Edgar, R. C. 2004, MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, **32**, 1792–7.

48. Capella-Gutiérrez, S., Silla-Martínez, J. M., and Gabaldón, T. 2009, trimAl: A tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics*, **25**, 1972–3.

49. Abascal, F., Zardoya, R., and Posada, D. 2005, ProtTest: Selection of best-fit models of protein evolution. *Bioinformatics*, **21**, 2104–5.

50. Nguyen, L. T., Schmidt, H. A., Von Haeseler, A., and Minh, B. Q. 2015, IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.*, **32**, 268–74.

51. Ronquist, F., Teslenko, M., van der Mark, P., et al. 2012, MrBayes 3.2: Efficient Bayesian Phylogenetic Inference and Model Choice Across a Large Model Space. *Syst. Biol.*, **61**, 539–42.

52. Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. 1990, Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–10.

53. Barton-Owen, T. B., Szabó, R., Somorjai, I. M. L., and Ferrier, D. E. K. 2018, A revised

spiralian homeobox gene classification incorporating new polychaete transcriptomes reveals a diverse TALE class and a divergent hox gene. *Genome Biol. Evol.*, **10**, 2151–67.

54. Paps, J., Xu, F., Zhang, G., and Holland, P. W. H. 2015, Reinforcing the egg-timer: Recruitment of novel Lophotrochozoa homeobox genes to early and late development in the Pacific oyster. *Genome Biol. Evol.*, **7**, 677–88.

55. Lu, S., Wang, J., Chitsaz, F., et al. 2020, CDD/SPARCLE: The conserved domain database in 2020. *Nucleic Acids Res.*, **48**, D265–8.

56. Huan, P., Wang, Q., Tan, S., and Liu, B. 2020, Dorsoventral decoupling of Hox gene expression underpins the diversification of molluscs. *Proc. Natl. Acad. Sci. U. S. A.*, **117**, 503–12.

57. Li, Y., Nong, W., Baril, T., et al. 2020, Reconstruction of ancient homeobox gene linkages inferred from a new high-quality assembly of the Hong Kong oyster (*Magallana hongkongensis*) genome. *BMC Genomics*, **21**, 713.

58. Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., Von Haeseler, A., and Jermiin, L. S. 2017, ModelFinder: Fast model selection for accurate phylogenetic estimates. *Nat. Methods*, **14**, 587–9.

59. Froufe, E., Gonçalves, D. V, Teixeira, A., et al. 2016, Who lives where? Molecular and morphometric analyses clarify which *Unio* species (Unionida, Mollusca) inhabit the southwestern Palearctic. *Org. Divers. Evol.*, **16**, 597–611.

60. Froufe, E., Prié, V., Faria, J., et al. 2016, Phylogeny, phylogeography, and evolution in the Mediterranean region: News from a freshwater mussel (*Potomida*, Unionida). *Mol. Phylogenet. Evol.*, **100**, 322–32.

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

62. Bieler, R., Mikkelsen, P. M., Collins, T. M., et al. 2014, Investigating the Bivalve Tree of Life – an exemplar-based approach combining molecular and novel morphological characters. *Invertebr. Syst.*, **28**, 32.

63. Lemer, S., Bieler, R., and Giribet, G. 2019, Resolving the relationships of clams and cockles: dense transcriptome sampling drastically improves the bivalve tree of life. *Proc. R. Soc. B Biol. Sci.*, **286**, 20182684.

64. Lemer, S., González, V. L., Bieler, R., and Giribet, G. 2016, Cementing mussels to oysters in the pteriomorphian tree: a phylogenomic approach. *Proc. R. Soc. B Biol. Sci.*, **283**, 20160857.

65. Ferrier, D. E. K., and Holland, P. W. H. 2001, January, Ancient origin of the Hox gene cluster. *Nat. Rev. Genet.* Nature Publishing Group, pp. 33–8.

66. Holland, P. W. H. 2013, January 1, Evolution of homeobox genes. *Wiley Interdiscip. Rev. Dev. Biol.* John Wiley & Sons, Ltd, pp. 31–45.

67. Pollard, S. L., and Holland, P. W. H. 2000, Evidence for 14 homeobox gene clusters in human genome ancestry. *Curr. Biol.*, **10**, 1059–62.

68. Castro, L. F. C., and Holland, P. W. H. 2003, Chromosomal mapping of ANTP class homeobox genes in amphioxus: Piecing together ancestral genomes. *Evol. Dev.*, **5**, 459–65.

69. Ferrier, D. E. K., and Holland, P. W. H. 2002, Ciona intestinalis ParaHox genes: Evolution of Hox/ParaHox cluster integrity, developmental mode, and temporal

colinearity. *Mol. Phylogenet. Evol.*, **24**, 412–7.

70. Brooke, N. M., Garcia-Fernàndez, J., and Holland, P. W. H. 1998, The ParaHox gene cluster is an evolutionary sister of the Hox gene cluster. *Nature*, **392**, 920–2.

71. Albertin, C. B., Simakov, O., Mitros, T., et al. 2015, The octopus genome and the evolution of cephalopod neural and morphological novelties. *Nature*, **524**, 220–4.

72. Da Fonseca, R. R., Couto, A., Machado, A. M., et al. 2020, A draft genome sequence of the elusive giant squid, *Architeuthis dux*. *Gigascience*, **9**.

73. Li, Y., Sun, X., Hu, X., et al. 2017, Scallop genome reveals molecular adaptations to semi-sessile life and neurotoxins. *Nat. Commun.*, **8**, 1721.

74. Liu, C., Ren, Y., Li, Z., et al. 2020, Giant African snail genomes provide insights into molluscan whole-genome duplication and aquatic-terrestrial transition. *bioRxiv*, 2020.02.02.930693.

75. Pérez-Parallé, M. L., Pazos, A. J., Mesías-Gansbiller, C., and Sánchez, J. L. 2016, Hox, Parahox, Ehgbox, and NK Genes in Bivalve Molluscs: Evolutionary Implications. *J. Shellfish Res.*, **35**, 179–90.

76. Simakov, O., Marletaz, F., Cho, S.-J., et al. 2012, Insights into bilaterian evolution from three spiralian genomes. *Nature*, **493**, 526–31.

77. Sun, J., Chen, C., Miyamoto, N., et al. 2020, The Scaly-foot Snail genome and implications for the origins of biomineralised armour. *Nat. Commun. 2020 111*, **11**, 1–12.

78. Sun, J., Mu, H., Ip, J. C. H., et al. 2019, Signatures of divergence, invasiveness, and terrestrialization revealed by four apple snail genomes Russo, C., (ed.), . *Mol. Biol. Evol.*, **36**, 1507–20.

79. Takeuchi, T., Koyanagi, R., Gyoja, F., et al. 2016, Bivalve-specific gene expansion in the pearl oyster genome: implications of adaptation to a sessile lifestyle. *Zool. Lett.*, **2**, 3.

80. Varney, R., Speiser, D., McDougall, C., Degnan, B., and Kocot, K. 2020, The iron-responsive genome of the chiton *Acanthopleura granulata*. *bioRxiv*, 2020.05.19.102897.

81. Wang, S., Zhang, J., Jiao, W., et al. 2017, Scallop genome provides insights into evolution of bilaterian karyotype and development. *Nat. Ecol. Evol.*, **1**, 0120.

82. Yan, X., Nie, H., Huo, Z., et al. 2019, Clam Genome Sequence Clarifies the Molecular Basis of Its Benthic Adaptation and Extraordinary Shell Color Diversity. *iScience*, **19**, 1225–37.

83. Bogan, A. E. 2008, Global diversity of freshwater mussels (Mollusca, Bivalvia) in freshwater. *Hydrobiologia*. Springer Netherlands, Dordrecht, pp. 139–47.

84. Graf, D. L., and Cummings, K. S. 2007, Review of the systematics and global diversity of freshwater mussel species (Bivalvia: Unionoida). *J. Molluscan Stud.*, pp. 291–314.

85. Howard, J. K., and Cuffey, K. M. 2006, The functional role of native freshwater mussels in the fluvial benthic environment. *Freshw. Biol.*, **51**, 460–74.

86. Vaughn, C. C. 2018, March 28, Ecosystem services provided by freshwater mussels. *Hydrobiologia*. Springer International Publishing, pp. 15–27.

87. Böhm, M., Dewhurst-Richman, N. I., Seddon, M., et al. 2020, The conservation status of the world's freshwater molluscs. *Hydrobiologia*, 1–24.

88. Bolotov, I. N., Vikhrev, I. V., Bespalaya, Y. V., et al. 2016, Multi-locus fossil-calibrated

phylogeny, biogeography and a subgeneric revision of the Margaritiferidae (Mollusca: Bivalvia: Unionoida). *Mol. Phylogenet. Evol.*, **103**, 104–21.

89. Zanatta, D. T., Stoeckle, B. C., Inoue, K., et al. 2018, High genetic diversity and low differentiation in north american *Margaritifera margaritifera* (bivalvia: Unionida: Margaritiferidae). *Biol. J. Linn. Soc.*, **123**, 850–63.

90. Araujo, R., Schneider, S., Roe, K. J., Erpenbeck, D., and Machordom, A. 2017, The origin and phylogeny of Margaritiferidae (Bivalvia, Unionoida): a synthesis of molecular and fossil data. *Zool. Scr.*, **46**, 289–307.

91. Bouza, C., Castro, J., Martínez, P., et al. 2007, Threatened freshwater pearl mussel *Margaritifera margaritifera* L. in NW Spain: low and very structured genetic variation in southern peripheral populations assessed using microsatellite markers. *Conserv. Genet.*, **8**, 937–48.

92. Geist, J., and Kuehn, R. 2005, Genetic diversity and differentiation of central European freshwater pearl mussel (*Margaritifera margaritifera* L.) populations: Implications for conservation and management. *Mol. Ecol.*, **14**, 425–39.

Tables legends.

Table 1 – *Margaritifera margaritifera* sequencing, genome assembly, read alignment, gene prediction and annotation general statistics.

Table 2 – Statistics of the content of repetitive elements in the *M. margaritifera* genome assembly. Values were produced by RepeatMasker using a RepeatModeler's custom build *M. margaritifera* repeat library (abbreviated with "Marmar") combined with the RepBase Biavalve repeat library (RepeatMasker option -lib).

Figure legends

Figure 1 – *Margaritifera margaritifera* specimen in its natural habitat.

Figure 2 - GenomeScope2 k-mer (25 and 31) distribution displaying estimation of genome size (len), homozygosity (aa), heterozygosity (ab), mean kmer coverage for heterozygous bases (kcov), read error rate (err), the average rate of read duplications (dup), k-mer size used on the run (k:) and ploidy (p:).

Figure 3 – *Margaritifera margaritifera* genome assembly assessment using KAT comp tool to compare the Illumina Paired-end reads k-mer content within the genome assembly. Different colours represent the read k-mer frequency in the assembly.

Figure 4 – Maximum Likelihood phylogenetic tree based on concatenated alignments of 118 single copy orthologous amino acid sequences retrieved by OrthoFinder. * above the nodes refer to bootstrap and posterior probabilities support values above 99%.

Figure 5 - Hox and ParaHox Maximum Likelihood gene tree constructed using Mollusca homeodomain amino acid sequences. Bootstrap values are presented above the nodes.

Supplementary data

Supplementary Table S1 – List of proteomes used for BRAKER2 gene prediction pipeline.

Supplementary Table S2 – List of proteomes used to retrieve single-copy orthologs in OrthoFinder v2.4.0.

Supplementary Table S3 – BRAKER2 gene prediction complete report.

Supplementary Table S4 – Genomic locations of Hox and ParaHox genes in the genome assemblies of *M. margaritifera* and *M. nervosa* and trancriptome assembly of *M. margaritifera*.

Supplementary Table S5 – Descriptors and acession numbers of tissue samples, raw data and assemblies of *Margaritifera margaritifera*

Supplementary File 1 – Fasta alignment of homeodomain amino acid sequences from Hox and ParaHox genes used in gene tree construction. Sequences used include the Hox and ParaHox homeodomains obtained in the current study as well as other Mollusca homeodomain sequences retrieved from (Huan et al., 2020; Li et al., 2020) and references within.

Supplementary File 2 – Scaffolds fasta sequences in which homeodomains were detected (as described in Table S4).

Table 1 – *Margaritifera margaritifera* sequencing, genome assembly, read alignment, gene prediction and annotation general statistics.

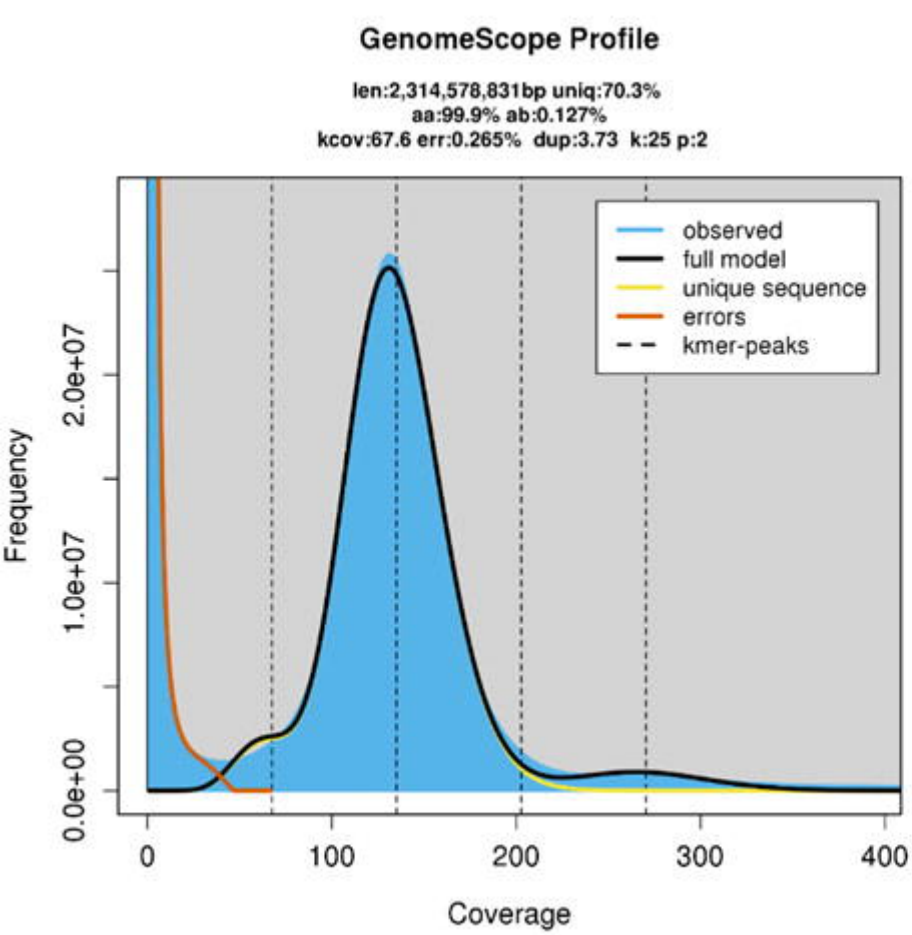
	Contig *	Scaffold *
Raw Data Stats		
Raw sequencing reads (PE 150bp)		3,298,603,550
Raw sequencing reads (MP-10kb 150bp)		511,524,202
Clean sequencing reads (PE 150bp)		3,286,495,504
Clean sequencing reads (MP-10kb 150bp)		459,166,278
Assembly Stats		
Total number of Sequences (>= 1,000 bp)	265,718	105,185
Total number of Sequences (>= 10,000 bp)	66,019	15,384
Total number of Sequences (>= 25,000 bp)	18,725	11,583
Total number of Sequences (>= 50,000 bp)	4,284	9,265
Total length (>= 1,000 bp)	2,230,001,992	2,472,078,101
Total length (>= 10,000 bp)	1,523,143,239	2,293,496,118
Total length (>= 25,000 bp)	789,559,702	2,236,013,546
Total length (>= 50,000 bp)	299,796,296	2,152,307,394
N50 length (bp)	16,899	288,726
L50 length	34,910	2,393
Maximum length	209,744	2,510,869
GC content, %	35.42	35.42
Clean Paired-end (PE) Reads Alignment Stats		
Mapped PE %	-	97.754
Proper pairs PE %	-	90.653
Average PE coverage %	-	181.968
Scaffolds with any coverage %	-	100.00
Total BUSCOS found (% bp)		
# Euk database	-	C:86.8% [S:85.8%, D:1.0%], F:5.9%
# Met database	-	C:84.9% [S:83.8%, D:1.1%], F:4.9%
Gene Prediction and Annotation Stats ±		
Protein coding genes (CDS)	-	35,119
Transcripts (mRNA)	-	40,544
Protein Coding genes Functional Annotated	-	26,836
Transcripts Functional Annotated	-	31,584
Total gene length (bp)	-	902,994,752
Total mRNA length (bp)	-	1,101,526,909
Total CDS length (bp)	-	52,211,391
Total exon length (bp)	-	52,211,391
Total intron length (bp)	-	1,024,450,311

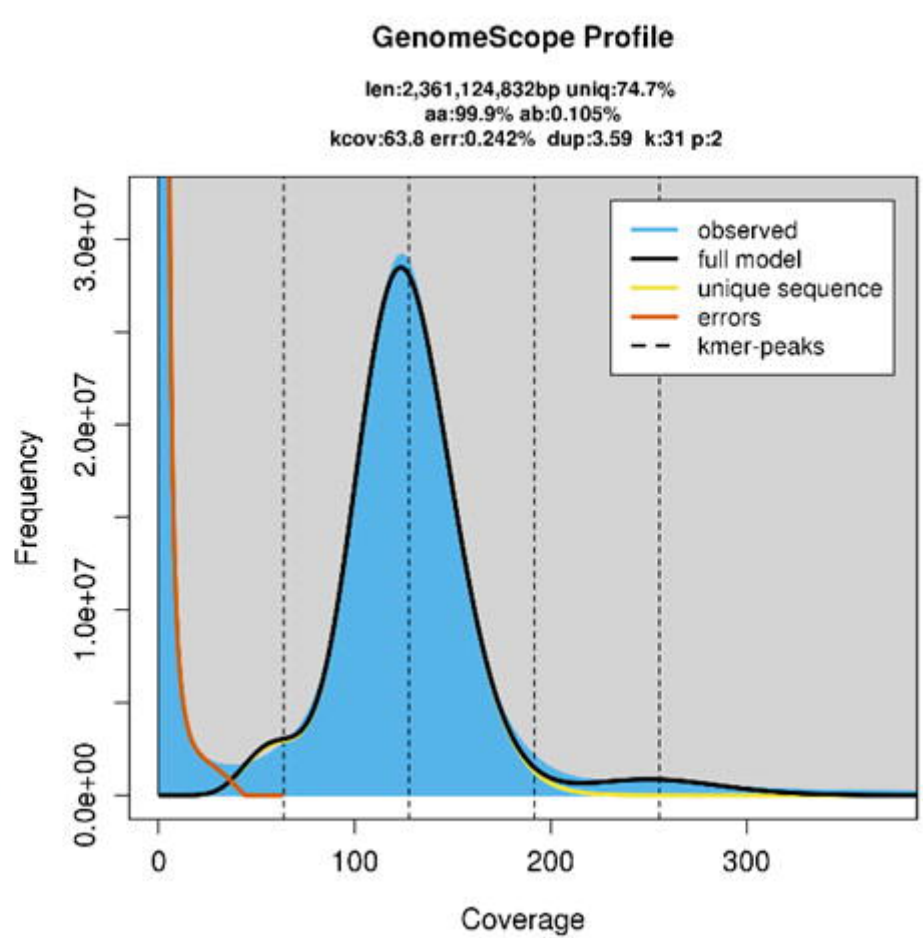
*All statistics are based on contigs/scaffolds of size ≥1,000bp.
Euk: From a total of 303 genes of Eukaryota library profile.
Met: From a total of 978 genes of Metazoa library profile.
C: Complete; S: Single; D: Duplicated; F: Fragmented
± All statistics are based on contigs/scaffolds of size ≥2,500bp

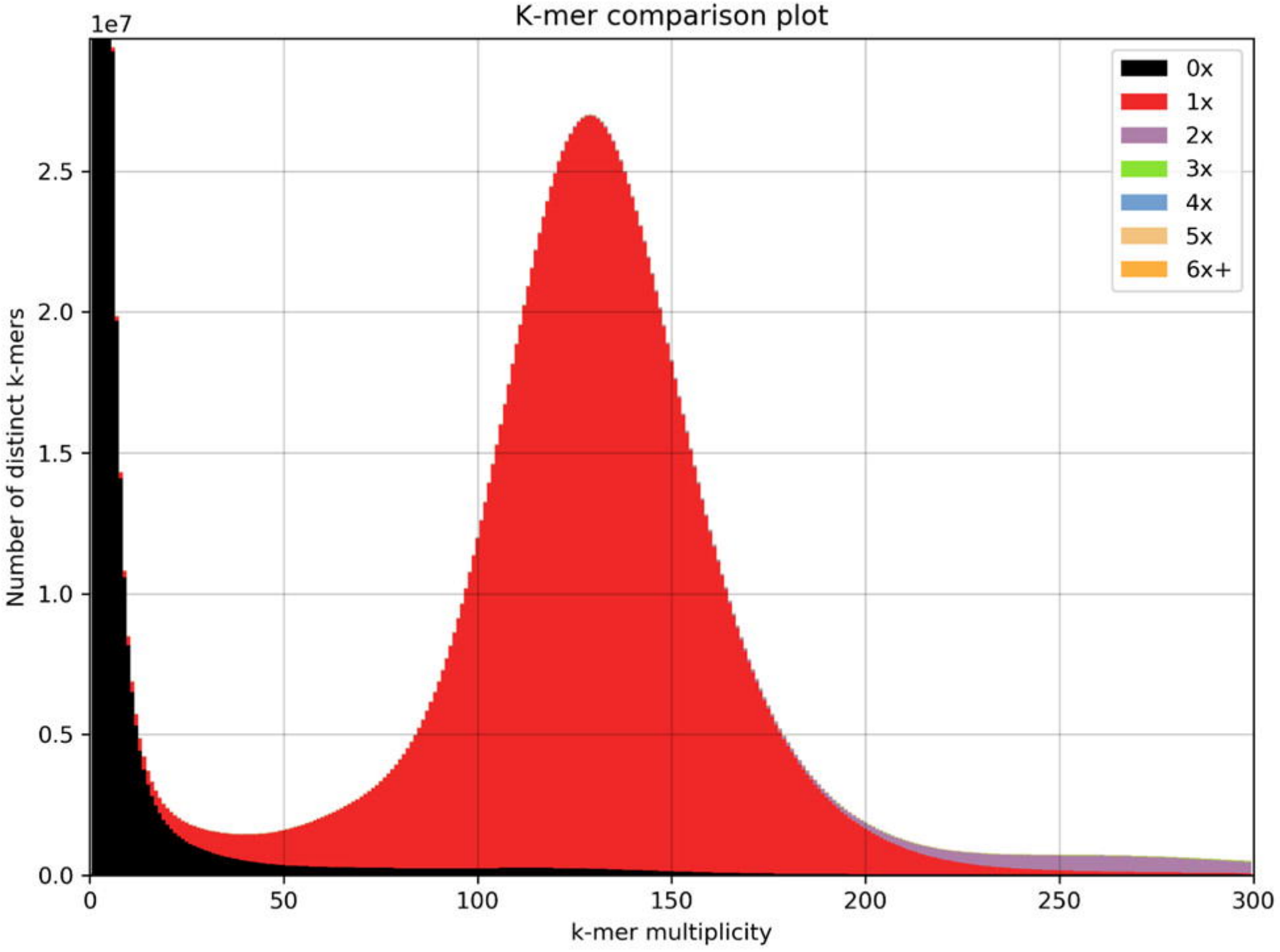
Table 2 – Statistics of the content of repetitive elements in the *M. margaritifera* genome assembly. Values were produced by RepeatMasker using a RepeatModeler's custom build *M. margaritifera* repeat library (abbreviated with "Marmar") combined with the RepBase Biavalve repeat library (RepeatMasker option -lib).

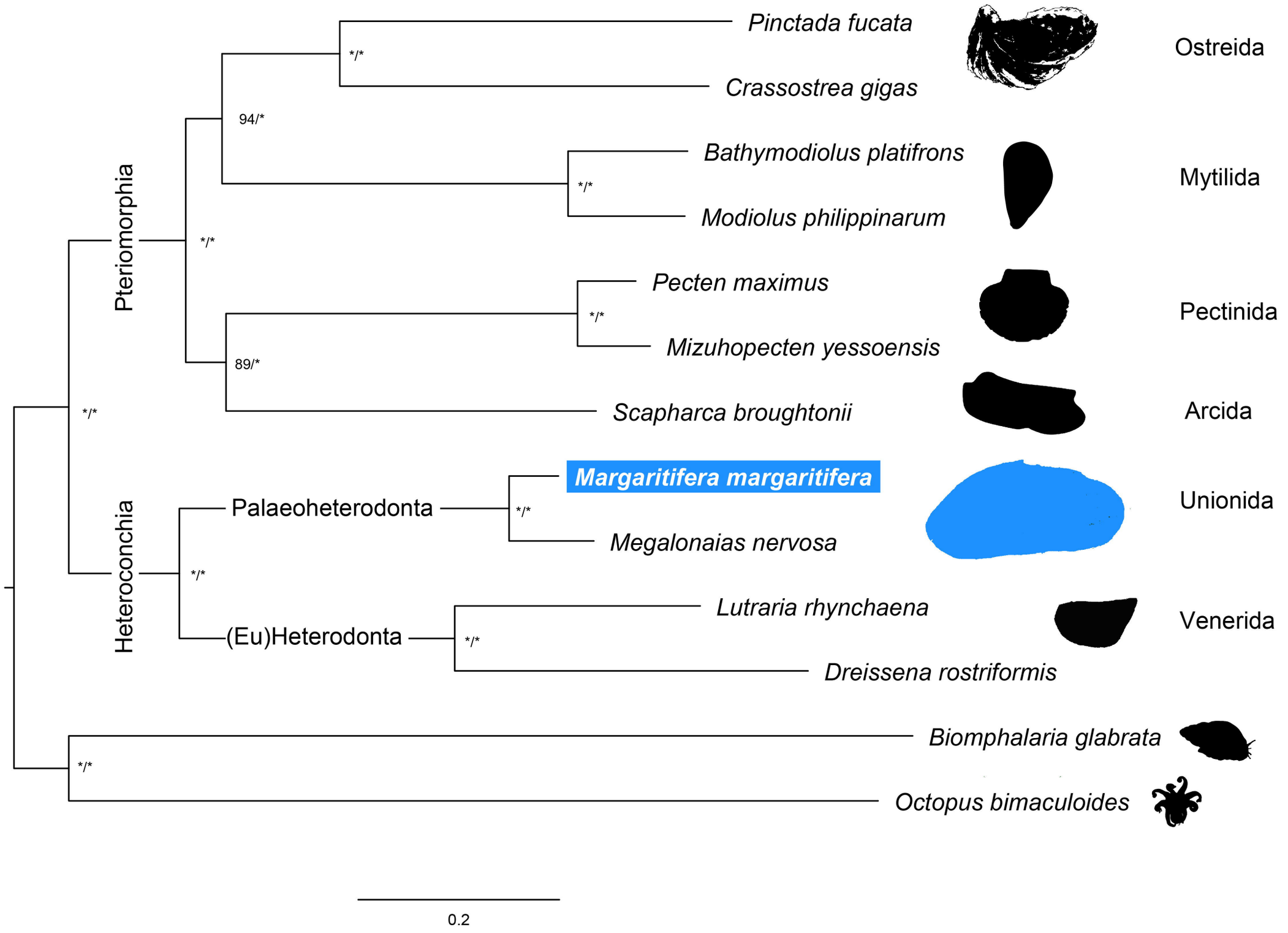
		Number of elements*	Length occupied (bp)	Percentage of sequence (%)
		Marmar + Bivalvia	**Marmar + Bivalvia**	**Marmar + Bivalvia**
SINEs:		108986	17810092	0.79%
	ALUs	0	0	0%
	MIRs	51807	7321859	0.33%
LINEs:		395376	137422770	6.13%
	LINE1	7854	2661360	0.12%
	LINE2	108179	29801298	1.33%
	L3/CR1	13806	3697570	0.17%
LTR elements:		174445	83417191	3.72%
	ERVL	0	0	0%
	ERVL-MaLRs	0	0	0%
	ERV_classI	2849	481472	0.02%
	ERV_classII	1072	286047	0.01%
DNA elements:		1208077	358545022	16.00%
	hAT-Charlie	22178	3778430	0.17%
	TcMar-	54446	15068283	0.67%

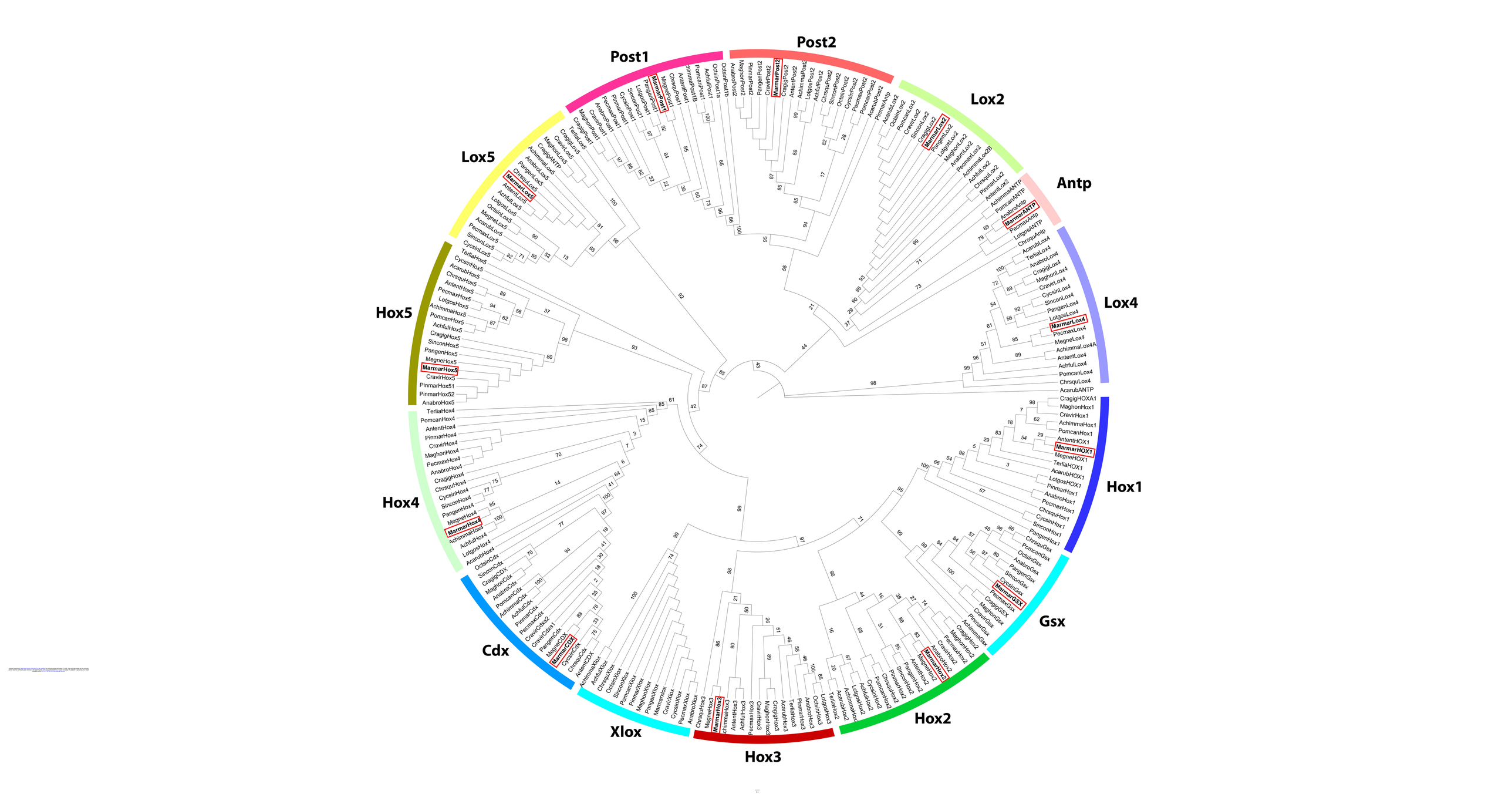
	Tigger		
Unclassified:	3057728	713890849	31.86%
Total interspersed repeats:		1311085924	58.51%
Small RNA:	51767	7672478	0.34%
Satellites:	24005	4250110	0.19%
Simple repeats:	64021	8534185	0.38%
Low complexity:	970	115583	0.01%
Total masked		**1323560844**	**59.07%**


GenomeScope Profile
len:2,314,578,831bp uniq:70.3%
aa:99.9% ab:0.127%
kcov:67.6 err:0.265% dup:3.73 k:25 p:2
observed
full model
unique sequence
errors
kmer-peaks
Frequency
0.0e+00
1.0e+07
2.0e+07
0
100
200
300
400
Coverage


GenomeScope Profile
len:2,361,124,832bp uniq:74.7%
aa:99.9% ab:0.105%
kcov:63.8 err:0.242% dup:3.59 k:31 p:2
observed
full model
unique sequence
errors
kmer-peaks
Frequency
0.0e+00
1.0e+07
2.0e+07
3.0e+07
0
100
200
300
Coverage


K-mer comparison plot
1e7
2.5
2.0
1.5
1.0
0.5
0.0
Number of distinct k-mers
0
50
100
150
200
250
300
k-mer multiplicity
0x
1x
2x
3x
4x
5x
6x+


Pinctada fucata
Crassostrea gigas
Bathymodiolus platifrons
Modiolus philippinarum
Pecten maximus
Mizuhopecten yessoensis
Scapharca broughtonii
Margaritifera margaritifera
Megalonaias nervosa
Lutraria rhynchaena
Dreissena rostriformis
Biomphalaria glabrata
Octopus bimaculoides
Ostreida
Mytilida
Pectinida
Arcida
Unionida
Venerida
Pteriomorphia
Heteroconchia
Palaeoheterodonta
(Eu)Heterodonta
94/*
89/*
/
0.2


Post1
Post2
Lox2
Antp
Lox4
Hox1
Gsx
Hox2
Hox3
Xlox
Cdx
Hox4
Hox5
Lox5
MarmarPost1
MarmarPost2
MarmarLox2
MarmarANTP
MarmarLox4
MarmarHOX1
MarmarGSX
MarmarHox2
MarmarHox3
MarmarCDX
MarmarHox4
MarmarHox5
MarmarLox5