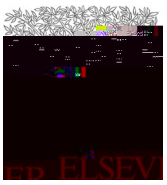
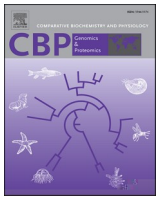
Comparative Biochemistry and Physiology - Part D: Genomics and Proteomics 45 (2023) 101045



Contents lists available at ScienceDirect

Comparative Biochemistry and Physiology - Part D: Genomics and Proteomics

journal homepage: www.elsevier.com/locate/cbpd



# Chromosome-level genome assembly of the European flat oyster (*Ostrea edulis*) provides insights into its evolution and adaptation

Xinchun Li [a,1], Yitian Bai [a,1], Zhen Dong [a], Chengxun Xu [a], Shikai Liu [a], Hong Yu [a], Lingfeng Kong [a], Qi Li [a,b,*]

[a] Key Laboratory of Mariculture, Ministry of Education, Ocean University of China, Qingdao 266003, China
[b] Laboratory for Marine Fisheries Science and Food Production Processes, Qingdao National Laboratory for Marine Science and Technology, Qingdao 266237, China

ARTICLE INFO

Edited by Chris Martyniuk

*Keywords:*
*Ostrea edulis*
Genome
Evolution
Transposable elements
*Hox* genes

ABSTRACT

The European flat oyster (*Ostrea edulis*) is an endangered and economically important marine bivalve species that plays a critical role in the coastal ecosystem. Here, we report a high-quality chromosome-level genome assembly of *O. edulis*, generated using PacBio HiFi-CCS long reads and annotated with Nanopore full-length transcriptome. The *O. edulis* genome covers 946.06 Mb (scaffold N50 94.82 Mb) containing 34,495 protein-coding genes and a high proportion of repeat sequences (58.49 %). The reconstructed demographic histories show that *O. edulis* population might be shaped by breeding habit (embryo brooding) and historical climatic change. Comparative genomic analysis indicates that transposable elements may drive lineage-specific evolution in oysters. Notably, the *O. edulis* genome has a *Hox* gene cluster rearrangement that has never been reported in bivalves, making this species valuable for evolutionary studies of molluscan diversification. Moreover, genome expansion of *O. edulis* is probably central to its adaptation to filter-feeding and sessile lifestyles, as well as embryo brooding and pathogen resistance, in coastal ecosystems. This chromosome-level genome assembly provides new insights into the genome feature of oysters, and presents an important resource for genetic research, evolutionary studies, and biological conservation of *O. edulis*.

## 1. Introduction

Oysters are filter-feeding and reef-engineer bivalve molluscs that are important to ecological system, fisheries and aquaculture (Bayne, 2017). With the rapid development of DNA sequencing technology in recent years, the genome sequences have been determined in several oyster species including four *Crassostrea* species (*Crassostrea gigas*, *C. virginica*, *C. hongkongensis* and *C. ariakensis*) and *Saccostrea glomerata* (Powell et al., 2018; Qi et al., 2021; Li et al., 2021a; Zhang et al., 2022). These genome resources are valuable for applied and fundamental basic research in oysters (Houston et al., 2020; Yang et al., 2020). However, very little is known about the genetic basis and evolutionary history of oysters in the genus *Ostrea*.

The European flat oyster *Ostrea edulis* represents an important aquaculture species native to the Atlantic and Mediterranean coasts of Europe (Fariñas-Franco et al., 2018; Colsoul et al., 2021). Although oyster aquaculture showed a steady increase in recent decades, *O. edulis* has suffered a dramatic decrease in wild populations and aquacultural production due to adverse effects of climate change, overfishing, pollutants and diseases (Colsoul et al., 2021; Gilson et al., 2021). Moreover, this decline also had a negative impact on oyster reef habitats and posed a great challenge to the marine ecosystems (Beck et al., 2011). New genomic tools for breeding programs and management provide new opportunities to improve natural resources of *O. edulis* (Vera et al., 2019; Colsoul et al., 2021). Despite the construction of EST and mapping of SNPs in *O. edulis* over the past years (Pardo et al., 2016; Gutierrez et al., 2017), the detailed genome assembly and annotation have yest to be completed that are critical for ecological and evolutionary research as well as genetic exploitation of this species.

In the present study, we report a high-quality chromosome-level genome assembly for an *O. edulis* individual from the Netherlands, utilizing the Pacific Bioscience (PacBio) single-molecule sequencing and high-throughput chromosome conformation capture (Hi-C) technologies. Furthermore, Nanopore transcriptome sequencing was used to

* Corresponding author at: Key Laboratory of Mariculture, Ministry of Education, Ocean University of China, Qingdao 266003, China.
*E-mail address:* qili66@ouc.edu.cn (Q. Li).
[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.cbd.2022.101045
Received 31 August 2022; Received in revised form 17 November 2022; Accepted 25 November 2022
Available online 29 November 2022

improve the completeness of the genome annotation. Overall, the genome information reported here, as well as two recently co-published *O. edulis* genomes from the UK and French populations (the OER-OSLIN_1.1 assembly and the Oe-Roscoff_1 assembly) (Boutet et al., 2022; Gundappa et al., 2022), will not only provide a better understanding of the genomic feature and evolutionary history of *O. edulis*, but also present new opportunities for research of conservation biology and genetic exploitation of this species.

## 2. Materials and methods

### 2.1. Sampling and sequencing

Farmed adult *O. edulis* were collected from Grevelingen and Oosterschelde in Netherlands (Fig. S1). The samples were dissected, immediately frozen in liquid nitrogen, and stored at −80 °C for library preparation and further analysis. All these individuals were verified as members of *O. edulis* using DNA fragments of cytochrome oxidase I (COI) (Danic-Tchaleu et al., 2011).

Genomic DNA (gDNA) of a female *O. edulis* individual was extracted from adductor muscle and mantle for whole genome sequencing, using the standard phenol-chloroform method. For genome survey, short paired-end DNA reads from a whole genome sequence (WGS) library with insert size of 350 bp were produced using the Illumina NovaSeq 6000 system. For long-read sequencing, a PacBio library (15–20 kb) was prepared using the SMRTbell Express Template Prep Kit 2.0 following the CCS HiFi library protocol (Pacific Biosciences, CA). Single Molecule Real Time (SMRT) sequencing was conducted on a PacBio Sequel II sequencing platform using the SMRT Cell (8 M) and Sequel II Sequencing Kit 2.0. The subreads were filtered by minimum length of 50 kb, and the HiFi reads were generated using ccs software (version 4.2.0) with the parameters of "min-passes = 3, min-rq = 0.99" (https://github.com/PacificBiosciences/ccs). The adductor muscle of another *O. edulis* individual was fixed with 1 % formaldehyde and used for Hi-C library construction by following a procedure described previously (Rao et al., 2014). *Mbo*I was used as the restriction enzyme. The library was also sequenced on an Illumina NovaSeq 6000 platform.

Total RNA was isolated from different tissues including adductor

annotated protein sequences (>50aa) of Bivalvia were downloaded from the Swiss-Prot database (Release 2022_1). A high confidence gene set was generated using Maker (version 3.01.03) (Cantarel et al., 2008) with the trained Augustus predictor, transcriptome assembly and protein sequences downloaded from NCBI and Swiss-Prot databases. To improve the gene structure annotation, we merged the gene set from Maker software (version 3.01.03) and gene models predicted with ONT FL transcriptome. Briefly, ONT cDNA reads were first filtered to remove the adapter sequences and low-quality raw reads (quality score less than 7), and then pre-processed by pychopper (version 2.5.0) (https://github.com/nanoporetech/pychopper) for the identification of FL reads, as well as trimming and orientation correction. Then, FL reads were mapped to genome using minimap2 (version 2.24-r1122) (Li, 2021). FL transcripts were assembled using StringTie2 (version 2.1.1) (Kovaka et al., 2019) in long read mode. Next, Transdecoder (version 5.5.0) (https://github.com/TransDecoder/TransDecoder) was used to identify candidate coding regions and predict gene models. Finally, gene models from Maker and ONT transcriptome-based prediction were combined using agat_sp_merge_annotations.pl script (https://github.com/NBISweden/AGAT).

Functional annotation of protein-coding genes was performed by homologous search against public databases. These include NCBI non-redundant protein (NR) (Release 2021_9_29) (Benson et al., 2000), Swiss-Prot (Bairoch and Apweiler, 2000), EggNOG (version 5.0) (Huerta-Cepas et al., 2019), Gene Ontology (GO) categories and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways (Kanehisa et al., 2017). In addition, gene motifs and domains were annotated using Interproscan (version 5.52–86.0) (Jones et al., 2014) against InterPro database (https://www.ebi.ac.uk/interpro/). The pfam_scan.pl script (ftp://ftp.ebi.ac.uk/pub/databases/Pfam/Tools/PfamScan.tar.gz) was used to align protein sequences against Pfam database (Pfam-A version 35.0) (Mistry et al., 2021).

Non-coding RNA (ncRNA) genes including transfer RNAs (tRNAs), microRNAs (miRNAs), ribosomal RNAs (rRNAs), and small nuclear RNAs (snRNAs) were annotated in the *O. edulis* genome. Briefly, tRNAs were predicted by tRNAscan-SE (version 2.0.7) (Chan et al., 2021) with parameters for eukaryotes. Screens for miRNAs, rRNAs and snRNAs were done using the Infernal (version 1.1.4) software (Nawrocki and Eddy, 2013) against the Rfam database (version 14.5) (Kalvari et al., 2018).

### 2.5. Effective population size estimation

The dynamics of effective population size (*Ne*) were estimated using the Pairwise Sequentially Markovian Coalescent (PSMC) (version 0.6.5-r67) software (Li and Durbin, 2011). The whole-genome sequencing data sets of a wild *O. edulis* from Bay of Morlaix in France and a wild Pacific oyster *C. gigas* from Dandong in China (Li et al., 2018) were downloaded from NCBI SRA database (SRA accession no. SRR17230313, SRR17226057, SRR6063159) (Table S1). Short DNA reads of *C. gigas* were mapped to the reference genome of Pacific oyster (Qi et al., 2021). The generation time (g) was assumed to be 1 year in both species, while mutation rates per nucleotide (μ) were set as 0.2e-8 for both species (Li et al., 2021a).

### 2.6. Phylogenetic analysis, gene family expansion and contraction

The orthogroups (OGs) of 14 molluscan proteomes (Table S2) were identified using Orthofinder (version 2.5.2) (Emms and Kelly, 2019). OGs from selected molluscan taxa were used for subsequent phylogenetic analysis. The phylogenetic relationships between *O. edulis* and other species were inferred based on the 1043 shared single-copy orthologous genes. Sequence alignments were performed with MAFFT (version 7.475) (Katoh and Standley, 2013) under default parameters. Species tree was constructed using FastTree 2 (Price et al., 2010). Divergence time between species was estimated using MCMCTREE from the PAML package (version 4.9j) (Yang, 2007). Eight reference-calibrated time points (Table S3) obtained from TimeTree database (http://timetree.org/) (Kumar et al., 2017) were used to constrain the nodes in the MCMC tree.

Gene family expansion and contraction were evaluated using CAFE (version 5) (Mendes et al., 2020) on the basis of the results from Orthofinder software (version 2.5.2) and species divergence time. A conditional *P* value was calculated for each gene family, and families with P value less than 0.05 were considered as having undergone significant expansion or contraction. GO enrichment and KEGG enrichment were performed for further functional analysis of expanded gene families, using the clusterprofiler R package (Yu et al., 2012).

### 2.7. Synteny analysis

The longest coding DNA sequences (CDS) for each gene, along with their coordinates, were prepared for five oyster species with chromosome-level assemblies (*O. edulis*, *O. denselamellosa*, *C. gigas, C. ariakensis*, and *C. virginica*). Next, pairwise comparisons were performed using MCscanX in the JCVI toolkit (version 1.1.12) (https://github.com/tanghaibao/jcvi) (Wang et al., 2012) to identify and visualize macrosynteny.

### 2.8. Homeobox gene analysis

The homeobox (*Hox*) genes were identified in both our *O. edulis* genome assembly and the OEROSLIN_1.1 assembly (Gundappa et al., 2022) by BLAST with an *E*-value threshold of 1e-10 against *Hox* genes in *M. yessoensis* (Wang et al., 2017) and *Lottia gigantea* (Simakov et al., 2013) genomes, respectively. The data were further confirmed by comparing to the Conserved Domains Database (http://www.ncbi.nlm.nih.gov/cdd). Genes were classified based on BLAST results and molecular phylogeny. The same approach was also used to identify homeobox genes in other molluscan genomes (*Pomacea canaliculata*, *S. glomerata*, *O. denselamellosa*, *C. gigas, C. ariakensis* and *C. virginica*). Phylogenetic trees were constructed by IQ-TREE (version 2.1.4-beta) (Minh et al., 2020), based on sequence alignments by MAFFT (version 7.475) (Katoh and Standley, 2013).

## 3. Results and discussion

### 3.1. Genome assembly and annotation

Based on Illumina sequencing data (Table S4) and 19-mer analysis, the estimated genome size of *O. edulis* is around 913 Mb with high heterozygosity of 0.78 % and repeat content of 55.85 % (Table S5). Due to remarkable genetic heterozygosity or polymorphisms, genome sequencing and assembly were inherently challenge for molluscs (Sigwart et al., 2021). To obtain a high-quality genome, we sequenced 30.44 Gb (b 33×) of PacBio HiFi long reads with an average length of 17.52 Kb (Table S4). These data were de novo assembled into 666 contigs with a N50 length of 3.23 Mb. With the aid of 132.06 Gb Hi-C data (b 145×) (Table S4), the initial assembled contigs were anchored onto 10 chromosomes (Fig. S2), consistent with the haploid karyotype of *O. edulis* (Thiriot-Quiévreux and Insua, 1992). Finally, a chromosome-level *O. edulis* genome assembly consisting of 618 scaffolds spanning 946.06 Mb is generated, with a scaffold N50 length of 94.82 Mb (Fig. 1A, Table 1, Table S5). The genome size of final assembly matched closely with our genome-size estimation by k-mer analysis and the OEROSLIN_1.1 assembly from the UK population (935.1 Mb) (Gundappa et al., 2022). The genome size of 946 Mb was slightly smaller than the 1.14 Gb genome size based on the flow cytometry results from the Spanish population (Rodr guez-Ju z et al., 1996) and the Oe-Roscoff_1 assembly (1.036 Gb) from the French population (Boutet et al., 2022). This discrepancy may be explained by population differences in genome size and high heterozygosity of *O. edulis*, which has been observed in

other bivalve genomes, including *C. gigas* (Peñaloza et al., 2021; Qi et al., 2021) and *Mytilus coruscus* (Yang et al., 2021; Sun et al., 2021).

Evaluation of the genome sequence completeness was performed based on BUSCO analysis (database: metazoa_odb10 and mollusca_odb10), and resulted in values of 95.1 % and 96.5 %, respectively (Table 1, Fig. S3). Moreover, the high mapping rate of short DNA reads of *O. edulis* individuals from different sampling sites demonstrates the high representative and accuracy of the genome assembly (Table S1). Furthermore, much of the genome comprises a high proportion of syntenic sequences compared with the OEROSLIN_1.1 assembly and the Oe-Roscoff_1 assembly (Boutet et al., 2022; Gundappa et al., 2022) (Fig. S4). Taken together, these data indicate a high-quality chromosome-scale genome assembly of *O. edulis* from this study, providing a robust framework for further exploration of oyster biology.

A total of 58.49 % bases were identified as repetitive sequences in the *O. edulis* genome (Table S6, Fig. S5), which is similar to the OEROSLIN_1.1 assembly (57.3 %) and the Oe-Roscoff_1 assembly (55.1 %) (Boutet et al., 2022; Gundappa et al., 2022). For gene annotation, we predicted 34,495 protein-coding genes with 131,967 isoforms in *O. edulis* (Table 1, Table S7). Approximately, 32,238 genes (93.46 %) were functionally annotated based on known proteins in public databases (Table S8). The gene set of *O. edulis* is larger than most oyster species sequenced to date (Li et al., 2021a; Peñaloza et al., 2021; Powell et al., 2018; Qi et al., 2021; Zhang et al., 2022), but similar to that of Eastern oyster *C. virginica*. Among three *O. edulis* genome annotations, the number of predicted gene models in our in-house annotation (Table S7) is slightly less than that of the other two in-house genome annotations of *O. edulis* (Boutet et al., 2022; Gundappa et al., 2022).

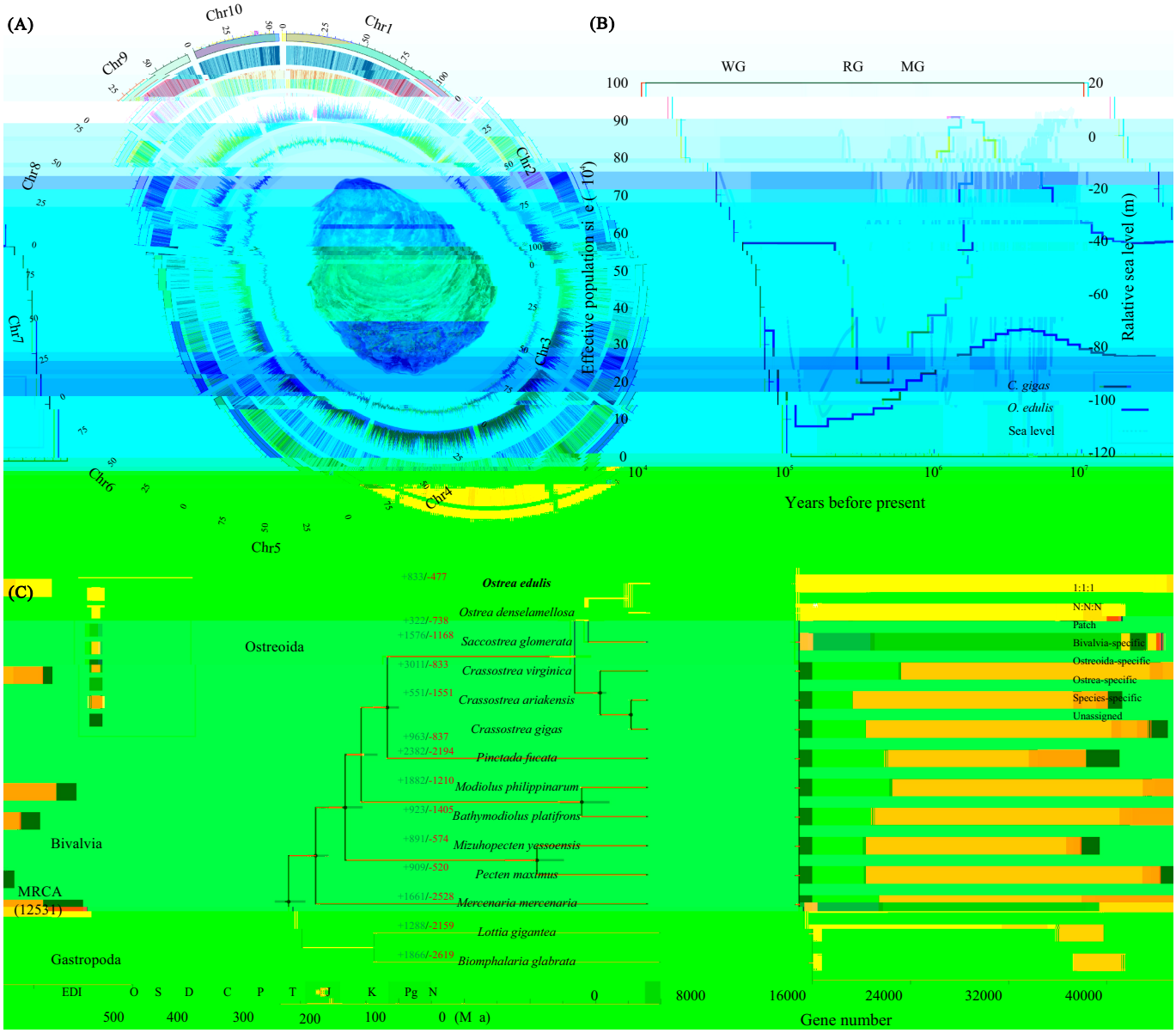



**Fig. 1.** The genome landscape and evolutionary history of *O. edulis*. (A) Circus plot of 10 chromosomes at 100-kb scale; with the European flat oyster at centre. From outer to inner ring are repeat coverage, the coverage of the longest CDS in each gene, the densities of isoforms (green) and genes (orange), and GC content. (B) Demographic histories of *O. edulis* and *C. gigas* inferred with the PSMC model. Estimates of the fluctuation of the global sea level relative to the present day were adopted from the literature (de Boer et al., 2014). (B) Phylogenetic tree and number of shared orthogroups among *O. edulis* and other molluscs. Reference-calibrated time points (Table S3) were indicated by red dots. The purple lines on the nodes indicate divergence time with a 95 % confidence interval. The numbers on each branch indicate gene family expansion (green) and contraction (red). Є, Cambrian; C, Carboniferous; D, Devonian; EDI, Ediacaran; J, Jurassic; K, Cretaceous; Ma, million years ago; N, Neogene; O, Ordovician; P, Permian; Pg, Paleogene; S, Silurian; T, Triassic.

However, the average gene length of 16,913 bp and CDS length of 1406 bp in our annotation (Table S7) are longer than the average gene length (7411 bp) and CDS length (1224 bp) of the in-house annotation of the OEROSLIN_1.1 assembly) (Gundappa et al., 2022). The average transcript length of our annotation (14,917 bp) (Table S7) is longer that of the in-house annotation of the Oe-Roscoff_1 assembly (13,916 bp) (Boutet et al., 2022). Also, 95.1 % and 98.5 % of the complete BUSCOs (database: metazoa_odb10) could be covered by the gene set and isoform set, respectively (Table 1, Fig. S3). These results indicate a high completeness of the *O. edulis* genome annotation from this study. In addition, various ncRNA sequences were also identified and annotated in the genome, including 13,963 tRNAs, 86 miRNAs, 419 rRNAs and 335 snRNAs (Table S9). Recently, a NCBI RefSeq annotation of the OEROSLIN_1.1 assembly was completed and released (https://www.ncbi.nlm.nih.gov/genome/annotation_euk/Ostrea_edulis/100//). This new assembly predicted 38,526 gene models with the average gene length of 16,710 bp and average CDS length of 1866 bp. The NCBI RefSeq annotation of *O. edulis* is comprehensive including detailed annotation of pseudogenes and non-coding transcripts. Given that our transcriptome data were generated using the Nanopore Technology that give long sequence reads, our annotation could provide another valuable dataset to enhance the power and accuracy of genome based genetic and evolutionary studies in *O. edulis*.

### 3.2. Population history and phylogenetic analyses

Through reconstruction of the demographic history, we found that the *Ne* of *O. edulis* was estimated to be smaller than that of *C. gigas* throughout the Quaternary glaciations (2.58 Mya to present) (Fig. 1B). *O. edulis* exhibits brooding by fertilizing and incubating in the mantle cavity of females (Colsoul et al., 2021). The larvae of *O. edulis* have a relatively shorter planktonic dispersal stage compared with those of *C. gigas*, and tend to be more aggregated around the parent population (Guy et al., 2019). Therefore, it is possible that the reproductive strategy plays a potential role in limiting the rapid expansion of wild *O. edulis* population. In addition, both *O. edulis* and *C. gigas* were severely affected by glaciation events as their populations expanded in a stepwise manner before the Mindel glaciation (MG), and then experienced a rapid decline over the last two glacial periods, namely the Riss (RG, 0.24–0.37 Mya) and W rm (WG, 10–120 Kya) glaciation (Fig. 1B). Notably, the population histories of the two oyster species show a remarkable correlation with the global sea-level fluctuation. For coastal marine species, sea-level drop caused by rapid cooling and glacier expansion have been viewed as a major cause of drastic decrease of available coastal habitat, resulting in genetic bottleneck, fragmented populations and population restructuring (Tsang et al., 2012; Ludt and Rocha, 2015; Li et al., 2021a). Hence, historical climatic change was potentially a main driver of demographic changes in oysters.

To infer the evolutionary history of *O. edulis*, a genome-wide phylogenetic tree was constructed based on 1043 single-copy genes from 14 molluscan genomes (Fig. 1C). The phylogenetic tree shows that *Ostrea* species and *S. glomerata* were inferred to diverge from the recent common ancestor around 91 Mya (60–130 Mya) (Fig. 1C), based on the secondary calibrations. During the Late Cretaceous period (66–100 Mya), oysters in both sides of the northern Atlantic Ocean became highly diversified, and a major burst of new genera occurred within a short period of time (Li et al., 2021b). Moreover, the divergence time of *O. edulis* from *O. denselamellosa* was dated to approximately 30 Mya (Fig. 1C), consistent with the diversity hotspot of oysters shifted from the Tethys to the Indian Ocean and East Asia from 66 to 23 Ma (Li et al., 2021b).

**Table 1**
Summary statistics of assembly and annotation of *Ostrea edulis* genome.

| Feature | Statistics |
|---|---|
| Assembled genome size (Mb) | 946.06 |
| GC content (%) | 35.50 |
| Number of chromosomes | 10 |
| Number of scaffolds | 618 |
| Longest scaffold length (Mb) | 110.13 |
| N50 scaffold length (Mb); L50 scaffold count | 94.82; 5 |
| N50 contig length (Mb); L50 contig count | 3.23; 86 |
| Repeat content (%) | 58.49 |
| Gene number | 34,495 |
| Mean gene length (bp) | 16,913 |
| Number of single-exon genes | 3869 |
| Isoform number | 131,967 |
| Number of single-exon isoforms | 8138 |
| Genome complete BUSCOs/total BUSCOs[a] | 908/954 (95.1 %) |
| Genome other BUSCOs: S; D; F; M[a] | 898; 10; 6; 40 |
| Gene set complete BUSCOs/total BUSCOs[a,b] | 907/954 (95.1 %) |
| Gene set other BUSCOs: S; D; F; M[a,b] | 884; 23; 12; 35 |
| Isoform set complete BUSCOs/total BUSCOs[a] | 939/954 (98.5 %) |
| Isoform set other BUSCOs: S; D; F; M[a] | 119; 820; 10; 5 |

[a] The lineage data set used for BUSCO assessment is metazoa_odb10 with 954 single-copy orthologues. S, Single-copy BUSCOs; D, Duplicated BUSCOs; F, Fragmented BUSCOs; M, Missing BUSCOs.

[b] Gene set includes the annotated isoforms with the longest coding sequences and protein sequences.

### 3.3. Genome structure characteristics

As observed in our genome assembly and the other two assemblies (Boutet et al., 2022; Gundappa et al., 2022), *O. edulis* has the larger haploid genome size than other previously sequenced oyster species (Table S10). High collinearity revealed by comparative genomic synteny among five oyster genomes (Fig. S6) suggests no whole genome duplication event in *O. edulis*. Comparing genome size with respect to the contents of transposons, CDS and Introns across six oyster species shows that TE content contributed to the majority of the differences across oysters (Pearson correlation $R^2 = 0.9071$, *P*-value = 3.34E-03) (Figs. S7 and S8). Thus, the expansion of the genome size of *O. edulis* is mainly caused by proliferation of TEs.

TEs are considered as drivers of genetic innovation that have a considerable impact on the genome size and genome architecture (Feschotte and Pritham, 2007; Slotkin and Martienssen, 2007; Wang et al., 2021). The oyster genomes contain 41.06–57.27 % TEs (Table S11). Notably, a genomic feature of all oyster genomes is that TE landscape is dominated by DNA transposons (11.13–21.21 %) (Fig. 2A, Table S11). This is in agreement with previous reports of other molluscan genomes (Wang et al., 2017; Yan et al., 2019; Huang et al., 2022). In addition, *Helitrons* accounts for a substantial proportion of TEs in oyster genomes (Fig. 2A, Table S11). *Helitron* rolling-circle TEs are regarded as the remnants of past activity in the evolutionary history, and might shape the evolution of oyster genomes (Peñaloza et al., 2021). Kimura distance-based divergence analysis indicates that recent TE burst events generally contributed to highly dynamic repeat content of oyster genomes, and almost no recent *Helitron* expanded in the *O. denselamellosa* genome (Fig. S9). Interestingly, long interspersed nuclear element (LINE) and short interspersed nuclear element (SINE) are abundant in the genomes of *O. eduis* and *O. denselamellosa* (Fig. 2A, Table S11), and their bursts are found in both *Ostrea* species (Fig. S9). These results suggest the potential role of retrotransposons in driving *Ostrea* genome evolution. Moreover, a large proportion of the TEs in oyster genomes (5.69–18.10 %) are lineage-specific but unclassified (Fig. 2A, Table S11), which was also observed in the OEROSLIN_1.1 assembly (37.65 %) and the Oe-Roscoff_1 assembly (10.20 %) (Boutet et al., 2022; Gundappa et al., 2022). The unclassified lineage-specific TEs may be contributors towards the genome evolution of oysters. Collectively, these data indicate that oyster genomes have been strongly influenced by the activity of TEs, and consequently provide promising models for studying host-transposon interactions.

### 3.4. Hox cluster rearrangement

*Hox* genes are known for their critical roles in the early development of metazoans as they are involved in the patterning of the anterior-posterior body axis and segment identity (Ferrier and Holland, 2001; Holland, 2013; Gaunt, 2018). The genomic organization of *Hox* genes varies considerably across molluscs (Wanninger and Wollesen, 2019; Varney et al., 2021; Huang et al., 2022). Nevertheless, an intact cluster of 11 *Hox* genes was identified in *L. gigantea* (Simakov et al., 2013) and scallops (Li et al., 2017; Wang et al., 2017; Kenny et al., 2020), supporting an 11-gene *Hox* cluster in the molluscan ancestor. The *ParaHox* cluster, on the other hand, was considered to include 3 genes in another scaffold (Wang et al., 2017). In this study, we found that the *O. edulis* genome contains 10 *Hox* genes clustered in the same chromosome (Chr 4) and 3 *ParaHox* genes in another chromosome (Chr 2) (Fig. 2B, Fig. S10). *Antp* is missing from all six oyster genomes (Fig. 2B), which was considered as a potential driver of byssus formation (Zhang et al., 2022).

Strikingly, the anterior subcluster of

expanded gene families, 78 contracted gene families and 430 species-specific

intraspecific genomic diversity of some pattern recognition receptors (C1qDCs, FREDs, and Ig domain-containing proteins) were among different populations revealed by the pan-genomic analysis of *M. galloprovincialis* (Gerdol et al., 2020). It is possible massive genetic variation associated to immune recognition may be linked to the evolutionary success in the environmental adaptation of marine invertebrates with high heterozygosity. Similar to our results (Fig. 3B), the expansions of gene families related to pathogen recognition (C-type lectins and C1qDCs) were also observed in another *O. edulis* genome assembly (Gundappa et al., 2022). *O. edulis* employs pattern recognition receptors, including lectins, TLRs and FREPs, as well as C1qDCs to detect pathogens and regulate innate immune responses during infection by *B. ostreae* (Pardo et al., 2016; Ronza et al., 2018; de la Ballina et al., 2021). Similarly, MEGF domain proteins are considered as pathogen recognition receptors and participate in immunological processes in *C. gigas* and *C. virginica* (Renault et al., 2011; Mcdowell et al., 2014; Chen et al., 2015). These expanded receptors in the canonical immune response pathways (Guo et al., 2015) and may enable *O. edulis* to develop a sophisticated innate immune system.

A previous report revealed a complete neuroendocrine-immune regulatory network in oysters, which could modulate immune response (Wang et al., 2018b). Recently, melanocyte-stimulating hormone (MSH) receptor was identified in *O. edulis* hyalinocytes, and might be able to mediate in bilateral information exchanges between the immune and neuroendocrine systems (de la Ballina et al., 2021). Therefore, expanded sets of neuronal acetylcholine receptor and MSH receptor (Fig. 3B, Table S17) could contribute to NEI regulatory network and play important roles in immune regulation in *O. edulis.* Taken together, these expanded and unique gene families related to innate immunity can provide important candidates for future investigations of disease resistance in *O. edulis*. Also, because of high heterozygosity rates in *O. edulis* genomes (Boutet et al., 2022; Gundappa et al., 2022), comparison of three genome assemblies from different populations will enrich our knowledge on the relationship of genetic va lecti 9

Colsoul, B., Boudry, P., Pérez Parallé, M.L., Bratoš Cetinić, A., Hugh Jones, T., Arzul, I., Mérou, N., Wegner, K.M., Peter, C., Merk, V., Pogoda, B., 2021. Sustainable large-scale production of European flat oyster (Ostrea edulis) seed for ecological restoration and aquaculture: a review. Rev. Aquacult. 13 (3), 1423–1468. https://doi.org/10.1111/raq.12529.

Danic-Tchaleu, G., Heurtebise, S., Morga, B., Lapegue, S., 2011. Complete mitochondrial DNA sequence of the European flat oyster Ostrea edulis confirms Ostreidae classification. BMC Res Notes 4, 400. https://doi.org/10.1186/1756-0500-4-400.

de Boer, B., Lourens, L.J., van de Wal, R.S.W., 2014. Persistent 400,000-year variability of Antarctic ice volume and the carbon cycle is revealed throughout the Plio-Pleistocene. Nat. Commun. 5 (1) https://doi.org/10.1038/ncomms3999.

de la Ballina, N.R., Villalba, A., Cao, A., 2021. Shotgun analysis to identify differences in protein expression between granulocytes and hyalinocytes of the European flat oyster Ostrea edulis. Fish Shellfish Immunol. 119, 678–691. https://doi.org/10.1016/j.fsi.2021.10.045.

Dudchenko, O., Batra, S.S., Omer, A.D., Nyquist, S.K., Hoeger, M., Durand, N.C., Shamim, M.S., Machol, I., Lander, E.S., Aiden, A.P., Aiden, E.L., 2017. De novo assembly of the Aedes aegypti genome using Hi-C yields chromosome-length scaffolds. Science 356, 92–95. https://doi.org/10.1126/science.aal3327.

Dudchenko, O., Shamim, M., Batra, S., Durand, N., Musial, N., Mostofa, R., Pham, M., Glenn St Hilaire, B., Yao, W., Stamenova, E., Hoeger, M., Nyquist, S., Korchina, V., Pletch, K., Flanagan, J., Tomaszewicz, A., McAloose, D., Estrada, C., Novak, B.J., Omer, A.D., Aiden, E., 2018. The Juicebox Assembly Tools module facilitates de novo assembly of mammalian genomes with chromosome-length scaffolds for under $1000. bioRxiv. https://doi.org/10.1101/254797.

Durand, N.C., Shamim, M.S., Machol, I., Rao, S.S.P., Huntley, M.H., Lander, E.S., Aiden, E.L., 2016. Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. Cell Syst. 3 (1), 95–98. https://doi.org/10.1016/j.cels.2016.07.002.

Emms, D.M., Kelly, S., 2019. OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biol. 20 (1) https://doi.org/10.1186/s13059-019-1832-y.

Engelsma, M.Y., Culloty, S.C., Lynch, S.A., Arzul, I., Carnegie, R.B., 2014. Bonamia parasites: a rapidly changing perspective on a genus of important mollusc pathogens. Dis. Aquat. Org. 110 (1–2), 5–23. https://doi.org/10.3354/dao02741.

Ertl, N.G., O'Connor, W.A., Brooks, P., Keats, M., Elizur, A., 2016. Combined exposure to pyrene and fluoranthene and their molecular effects on the Sydney rock oyster, Saccostrea glomerata. Aquat. Toxicol. 177, 136–145. https://doi.org/10.1016/j.aquatox.2016.05.012.

Fariñas-Franco, J.M., Pearce, B., Mair, J.M., Harries, D.B., Macpherson, R.C., Porter, J.S., Reimer, P.J., Sanderson, W.G., 2018. Missing native oyster (Ostrea edulis L.) beds in a European Marine Protected Area: Should there be widespread restorative management? Biol. Conserv. 221, 293–311. https://doi.org/10.1016/j.biocon.2018.03.010.

Ferrier, D.E.K., Holland, P.W.H., 2001. Ancient origin of the Hox gene cluster. Nat. Rev. Genet. 2 (1), 33–38. https://doi.org/10.1038/35047605.

Feschotte, C., Pritham, E.J., 2007. DNA transposons and the evolution of eukaryotic genomes. Annu. Rev. Genet. 41, 331–368. https://doi.org/10.1146/annurev.genet.40.110405.090448.

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 (17), 9451–9457. https://doi.org/10.1073/pnas.1921046117.

Gaunt, S.J., 2018. Hox cluster genes and collinearities throughout the tree of animal life. Int. J. Dev. Biol. 62 (11–12), 673–683. https://doi.org/10.1387/ijdb.180162sg.

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 (1) https://doi.org/10.1186/s13059-020-02180-3.

Gerdol, M., Manfrin, C., De Moro, G., Figueras, A., Novoa, B., Venier, P., Pallavicini, A., 2011. The C1q domain containing proteins of the Mediterranean mussel Mytilus galloprovincialis: A widespread and diverse family of immune-related molecules. Dev. Comp. Immunol. 35 (6), 635–643. https://doi.org/10.1016/j.dci.2011.01.018.

Gilson, A.R., Coughlan, N.E., Dick, J.T.A., Kregting, L., 2021. Marine heat waves differentially affect functioning of native (Ostrea edulis) and invasive (Crassostrea Magallana gigas) oysters in tidal pools. Mar. Environ. Res. 172, 105497 https://doi.org/10.1016/j.marenvres.2021.105497.

Goldstone, J.V., Hamdoun, A., Cole, B.J., Howard-Ashby, How by,

Li, Y., Sun, X., Hu, X., Xun, X., Zhang, J., Guo, X., Jiao, W., Zhang, L., Liu, W., Wang, J., Li, J., Sun, Y., Miao, Y., Zhang, X., Cheng, T., Xu, G., Fu, X., Wang, Y., Yu, X., Huang, X., Lu, W., Lv, J., Mu, C., Wang, D., Li, X., Xia, Y., Li, Y., Yang, Z., Wang, F., Zhang, L., Xing, Q., Dou, H., Ning, X., Dou, J., Li, Y., Kong, D., Liu, Y., Jiang, Z., Li, R., Wang, S., Bao, Z., 2017. Scallop genome reveals molecular adaptations to semi-sessile life and neurotoxins. Nat. Commun. 8 (1) https://doi.org/10.1038/s41467-017-01927-0.

Liu, B., Shi, Y., Yuan, J., Hu, X., Zhang, H., Li, N., Fan, W., 2013. Estimationof Genomic Characteristics by Analyzing k-mer Frequency in De Novo GenomeProjects. arXiv. org. https://doi.org/10.48550/arXiv.1308.2012 arXiv: 1308.2012.

Ludt, W.B., Rocha, L.A., 2015. Shifting seas: the impacts of Pleistocene sea-level fluctuations on the evolution of tropical marine taxa. J. Biogeogr. 42 (1), 25–38. https://doi.org/10.1111/jbi.12416.

Manni, M., Berkeley, M.R., Seppey, M., Zdobnov, E.M., 2021. BUSCO: assessing genomic data quality and beyond. Curr. Protoc. 1 (12) https://doi.org/10.1002/cpz1.323.

Marçais, G., Kingsford, C., 2011. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics 27 (6), 764–770. https://doi.org/10.1093/bioinformatics/btr011.

Mardones-Toledo, D.A., Montory, J.A., Joyce, A., Thompson, R.J., Diederich, C.M., Pechenik, J.A., Mardones, M.L., Chaparro, O.R., 2015. Brooding in the chilean oyster Ostrea chilensis: Unexpected complexity in the movements of brooded offspring within the mantle cavity. PLos One 10 (4), e122859. https://doi.org/10.1371/journal.pone.0122859.

McCartney, M.A., Auch, B., Kono, T., Mallez, S., Zhang, Y., Obille, A., Becker, A., Abrahante, J.E., Garbe, J., Badalamenti, J.P., Herman, A., Mangelson, H., Liachko, I., Sullivan, S., Sone, E.D., Koren, S., Silverstein, K.A.T., Beckman, K.B., Gohl, D.M., 2022. The genome of the zebra mussel, Dreissena polymorpha: a resource for comparative genomics, invasion genetics, and biocontrol. GD