BMC Genomics

RESEARCH ARTICLE

Open Access

Evidence of multiple genome duplication events in *Mytilus* evolution

Ana Corrochano-Fraile[1], Andrew Davie[1], Stefano Carboni[1,2*] and Michaël Bekaert[1]

Abstract

Background: Molluscs remain one significantly under-represented taxa amongst available genomic resources, despite being the second-largest animal phylum and the recent advances in genomes sequencing technologies and genome assembly techniques. With the present work, we want to contribute to the growing efforts by filling this gap, presenting a new high-quality reference genome for *Mytilus edulis* and investigating the evolutionary history within the Mytilidae family, in relation to other species in the class Bivalvia.

Results: Here we present, for the first time, the discovery of multiple whole genome duplication events in the Mytilidae family and, more generally, in the class Bivalvia. In addition, the calculation of evolution rates for three species of the Mytilinae subfamily sheds new light onto the taxa evolution and highlights key orthologs of interest for the study of *Mytilus* species divergences.

Conclusions: The reference genome presented here will enable the correct identification of molecular markers for evolutionary, population genetics, and conservation studies. Mytilidae have the capability to become a model shellfish for climate change adaptation using genome-enabled systems biology and multi-disciplinary studies of interactions between abiotic stressors, pathogen attacks, and aquaculture practises.

Keywords: *Mytilus edulis*, Whole-genome duplication, Evolution, Positive selection, Paleogenomics

Introduction

The family Mytilidae constitutes a diverse group of bivalves, broadly distributed in marine environments. *Mytilus edulis* and *Mytilus galloprovincialis* are the common species cultivated in Europe and both hybridise with *Mytilus trossulus* where their geographical distribution overlaps [1, 2] forming the European *Mytilus* Species Complex [3]. Nonetheless, a number of environmental and genetic barriers work together to maintain genetic discontinuities between the different species of the complex [4]. *M. edulis* and *M. galloprovincialis* can be considered cosmopolitan species while *M. trossulus* remains more geographically confined to the northernmost regions of the Pacific and Atlantic oceans and to the Baltic Sea [5]. At a finer geographical scale, mussel species display an extraordinary capability of environmental adaptation, extending from high inter-tidal to sub-tidal regions, from estuary to fully marine conditions, and from sheltered to extremely wave-exposed shores. Mussels are furthermore exposed to a wide range of potentially pathogenic microorganisms and pollutants, and yet they display a remarkable resilience to stress and infections [6]. Of particular interest are observations of a relatively high heterozygosity, rapid evolutionary responses to environmental threats, including predation [7], and recent suggestions that widespread relaxed selection in "low locomotion" molluscs, such as bivalves, and high copy number variants [8] could underpin observed high resilience and rapid adaptation to new environments [9].

The Phylum Mollusca remains significantly under-represented amongst those with available genomic

*Correspondence: s.carboni@fondazioneimc.it
[2] International Marine Centre, Loc. Sa Mardini snc, 09170 Torre Grande, OR, Italy
Full list of author information is available at the end of the article

© The Author(s) 2022. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

resources [10]. To date, only two high-quality genomes and associated gene models are available within the Mytilidae family: *M. galloprovincialis*, first sequenced by Murgarella et al. [11] and then improved by Gerdol et al. [6], and *M. coruscus* recently sequenced by Li et al. [12] and Yang et al. [13]. Comparative genomics provides an opportunity to investigate the "signatures" that natural selection has left on the genomes of related species. By analysing the frequency distribution of synonymous substitution per synonymous sites, it is possible to identify major evolutionary events, including Whole-Genome Duplications (WGDs). While WGDs are rare within animal lineages, they deeply shaped vertebrate evolution and represent important evolutionary landmarks from which some major lineages have diversified [14]. Furthermore, comparisons among related species adapted to contrasting niches, can provide an opportunity to investigate how their genomes diverge in response to different habitat conditions [15]. In cases where specific amino acids are known to affect protein function, analyses of intra-specific polymorphism and divergence can be used to directly study function variation in natural populations [16, 17]. Both whole genome duplication analysis and positive selection genome-wide analysis can therefore expose the strength and direction of selection on genes' functional variation and corroborate the adaptive significance of the loci under study [18, 19].

With this work, we want to contribute to the growing efforts in filling the gap in Molluscs genomic resources [20] by presenting a new high-quality reference genome for *M. edulis* and investigating the evolutionary history within the Mytilidae family and in relation to other species in the class Bivalvia. Here we present a new reference genome for *M. edulis*; we introduce the first evidence of WGD events in the Mytilidae family and in Bivalvia more generally; finally, we identify gene clusters under significant positive selection within each species of the Mytilidae family for which suitable reference genomes are available (*M. edulis*, *M. galloprovincialis* and *M. coruscus*).

The availability of this reference genome will not only increase interest in Mytilidae as a model for ecological and evolutionary research, but will be also a valuable tool for breeding programmes [21]. The discovery of multiple duplication events will enable the correct identification of molecular markers for evolutionary, population genetics, and conservation studies. The Mytilidae have the capability to become a model shellfish for climate change adaptation using genome-enabled systems biology and multi-disciplinary studies of interactions between abiotic stressors, pathogen attacks, and aquaculture practises.

Results

Sequencing results

After sequencing with the PromethION platform, a total of 15.95 million (111.65 Gb) long-reads were generated and used for the genome assembly. The mean length of the sequences was 7002 nt. The Illumina HiSeq X Ten platform produced 652.47 million (195.74 Gb) paired-ended short reads (150 nt). Based on the presumption that the genome size will be similar to that of closely related taxa: *M. galloprovincialis* [11] and *M. coruscus* [12] with an estimated genome size of 1.60 Gb and 1.90 Gb respectively; therefore, the estimated sequencing coverage was 64x and 113x, for long and short reads respectively (Table S1).

De novo assembly of the *M. edulis* genome

Using Jellyfish, the frequency of 17-mers and 23-mers in the Illumina filtered data were determined (Fig. S1) and followed the theoretical Poisson distribution typical of a diploid species [22]. The proportion of heterozygosity in the *M. edulis* genome was evaluated as being 3.69 and 4.84% respectively, and the genome size was estimated as 1.01 Gb and 1.10 Gb, with a repeat content of 68.13 and 39.91% respectively (Table 1).

Long-reads were assembled, polished with Racon and short-read sequence were corrected with Pilon, creating

Table 1 Statistics of the genome assembly of *M. edulis*

Category	Number/length
K-mer = 17	
Estimated genome size	1,010,184,781 nt
Estimated repeats	688,190,885 nt
Estimated hetorizygosity	3.69%
K-mer = 23	
Estimated genome size	1,096,306,163 nt
Estimated repeats	437,569,400 nt
Estimated hetorizygosity	4.84%
Number of contigs	3339
Total length	1,827,085,763 nt
Total repeats	1,029,206,554 nt
Observed hetorizygosity	0.48%
Largest contig	10,529,124 nt
N_{50}	1,097,279 nt
GC	32.17%
Read Mapped	91.35%
Avg. coverage depth	152x
Coverage over 10x	99.99%
N's per 100 kbp	13.73
BUSCO recovered	98.9%
Predicted rRNA genes	132
Predicted protein coding genes	69,246

an assembled genome of 3339 contigs with a total length and contig N_{50} of 1.83 Gb and 1.10 Mb, respectively Table 1). The realignment of the short-reads also provided descriptions of the mean observed heterozygosity of 0.48%, which is consistent with the most recent evidence collected from de novo Restriction site associated DNA (RAD) analysis [23] and microsatellite loci study [24]. A second *M. edulis* genome was recently released, NCBI Accession GCA_019925275.1. This chromosome level de novo assembly was only based on long-reads (PacBio Sequel platform), where fewer error corrections are possible; but produced a comparable genome size of 1.65 Gb and contig N_{50} of 0.49 Mb.

Repeat sequences and gene models

The transposable elements and repetitive sequences have been annotated using RepeatMasker and LTR-Finder. In total, we have found 1.03 Gb (56.33%) of repetitive sequences (Table S2). We used a combined method that integrates ab initio gene prediction and RNA-seq-based prediction to annotate the protein-coding genes in *M. edulis* genome. A total of 69,246 distinct gene models and 73,842 transcripts were annotated. The completeness of gene regions was further assessed using BUSCO using a Metazoa (release 10) benchmark of 954 conserved Metazoa genes, of which 93.8% had complete gene coverage (including 29.4% duplicated ones), 5.1% were fragmented and only 1.1% were classified as missing (Fig. 1A). These data largely support a high-quality *M. edulis* genome assembly, which can be used for further investigation. The predicted proteins from the reconstructed genes were subjected to BlastP similarity searches against SwissProt, Pfam, InterPro, KEGG and GO databases. Of the total 69,246 gene models annotated by at least one database, 9005 (13.0%) were annotated in all five databases used (Fig. 1B and Table S3). A total of 31,620 predicted genes were annotated to three major GO classes: "biological processes", "cellular components" and "molecular functions" (Fig. 1C).

Mitochondrial genome

The mitochondrial genome was retrieved manually from the genome assembly. The sequence of 16.74 kb was validated for continuity and circularity, and fully annotated. The full mitochondrial genome (Fig. 2A) was compared to the reference *M. edulis* genome [25]. Only one haplotype was recovered which is identical at 99% (85 SNPs) with the reference genomes (EBI Accession NC_006161.1) and is consistent with a female mitotype [26]. Complete annotated mitochondrial genomes for all Mytilinae (subfamily) to date (11 species; Table S4) were collected. Concatenated alignments constructed from all mitochondrial shared CDS sequences were used to construct a phylogenetic tree (Fig. 2B). This phylogenetic tree is consistent with the species relationships observed in previous studies [27].

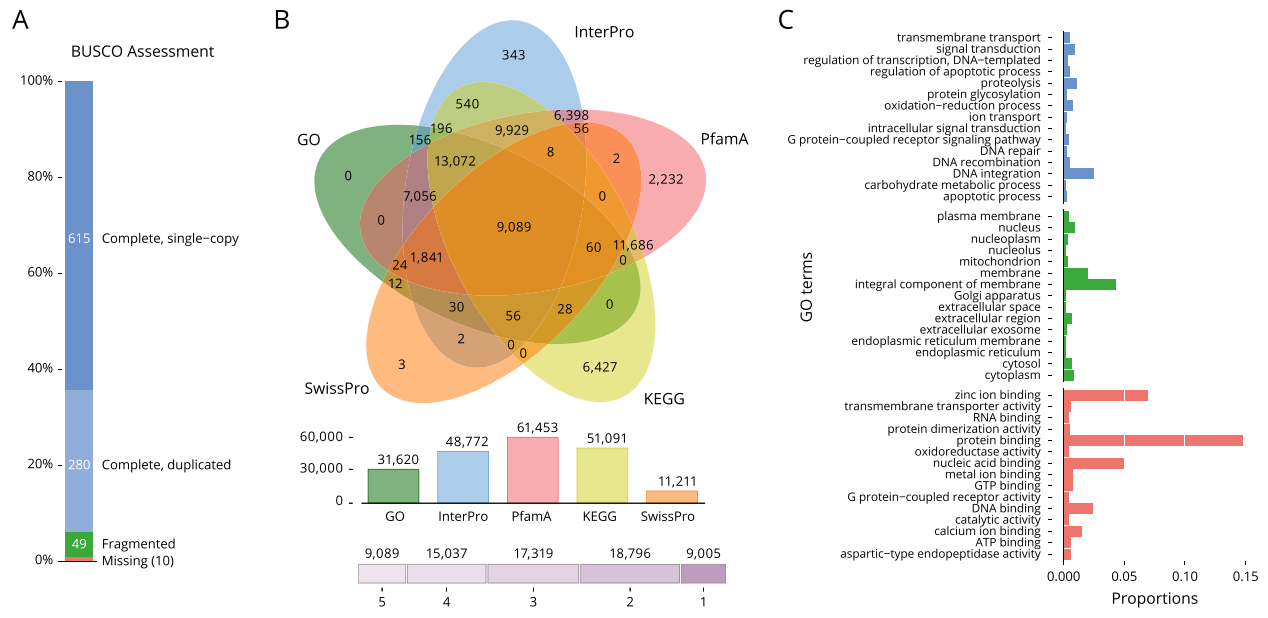

Fig. 1 Gene composition and annotation estimations. **A** BUSCO evaluation (Metazoa database; number of framework genes 954), 98.9% of the gene were recovered; **B** A five-way Venn diagram. The figure shows the unique and common genes displaying predicted protein sequence similarity with one or more databases (details in Supplementary Table S3); **C** Level 2 GO annotations using the gene ontology of assembled transcripts

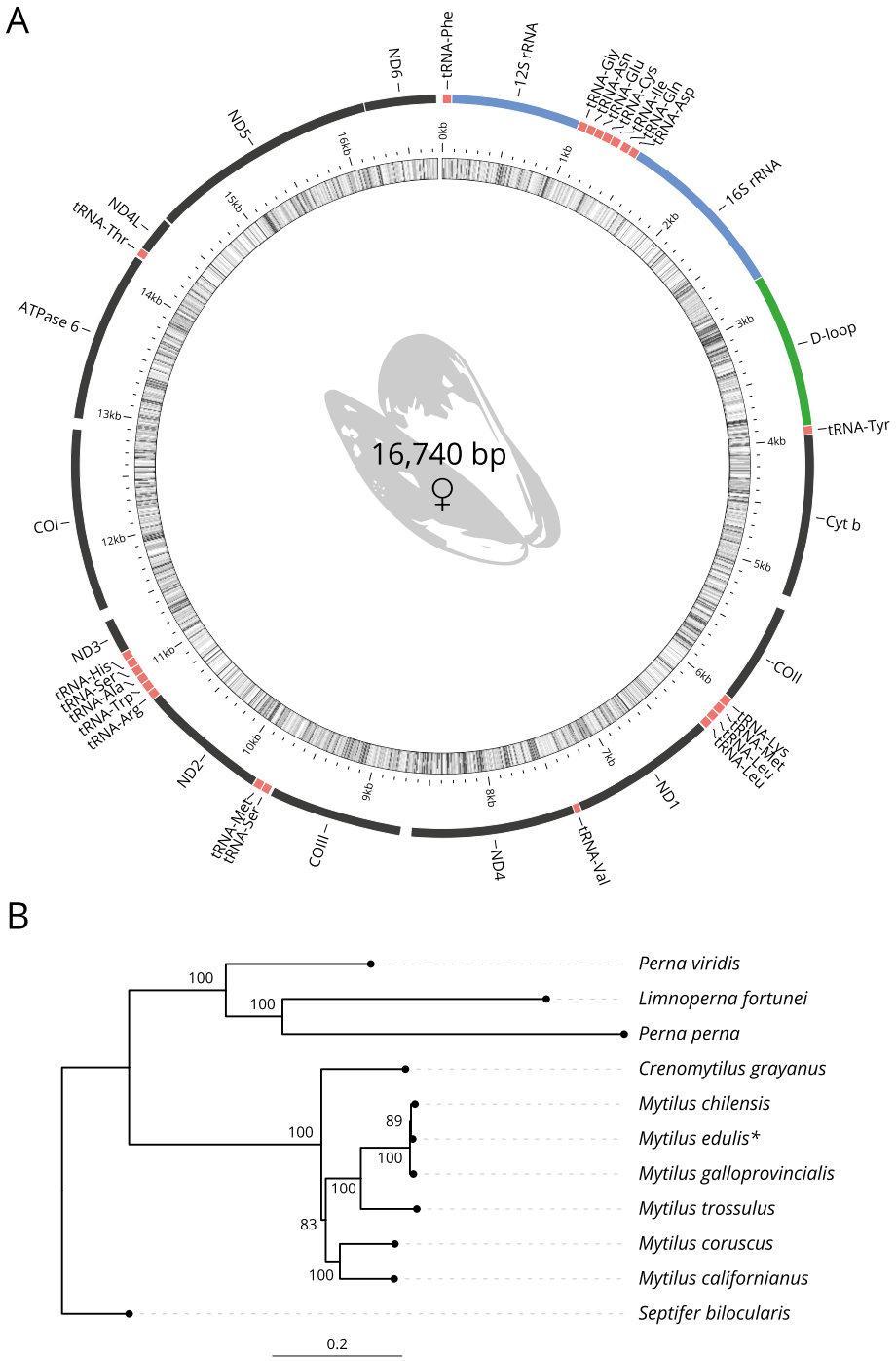

Fig. 2 Genome assembly. **A** *M. edulis* annotated mitochondrial genome; **B** Phylogenetic tree inferred from the mitochondrial gene

Detecting whole genome duplications

To assess the paleo-history of the Mytilinae (subfamily), we performed a comparative genomic investigation. A total of 2293 gene duplications younger than Ks = 5 were inferred across the total data set of 16,291 assembled unigene clusters in Mytilinae (*M. coruscus*, *M. edulis* and *M. galloprovincialis*). The histograms of duplication ages for each species analysed demonstrated evidence of two large-scale duplications (Fig. 3A). Mixture model analyse of Ks distributions (Fig. 3B and Table S5) to identify ancient whole genome duplications [28, 29] were consistent with the two consecutive whole genome

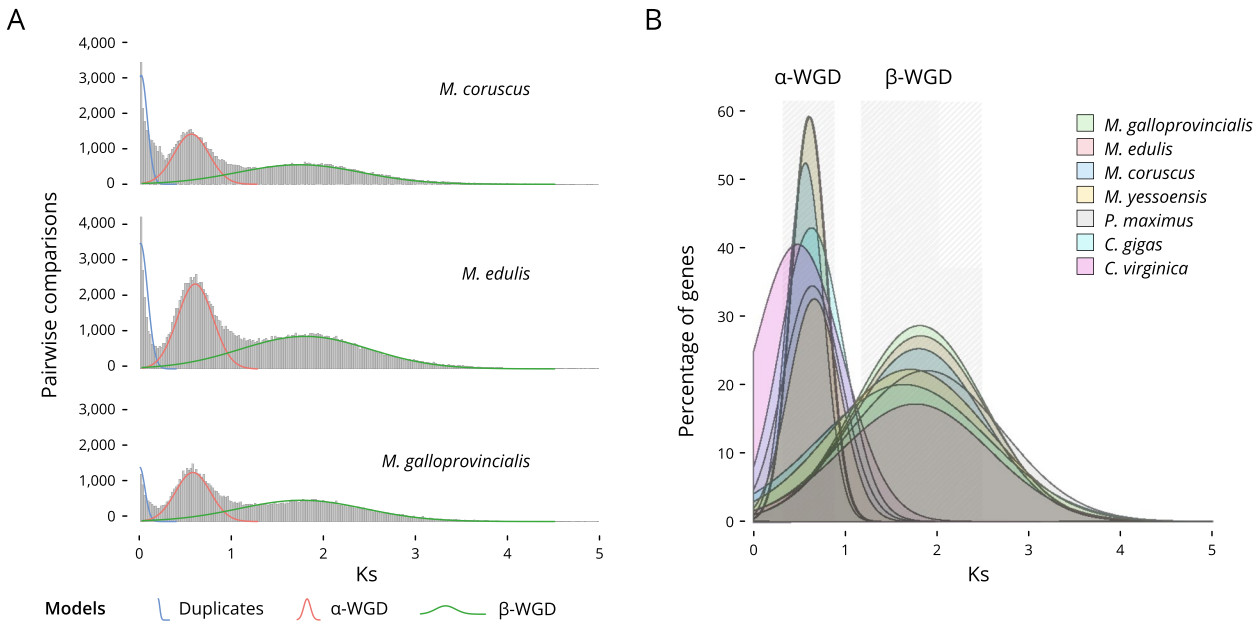

Fig. 3 Ka and Ks analysis. **A** Distribution of the Ks values of the duplicate pairs in *M. coruscus*, *M. edulis* and *M. galloprovincialis*; **B** Bivalvia (class) ortholog Ks distribution and multiple WGDs. Combined Ks plot of the gene age distributions of seven species (see Tables S5 and S6). The median peaks for these plots are highlighted. Analyses of ortholog divergence indicated that these taxa diverged after their most recent WGDs

duplication events (α-WGD and β-WGD). The duplication distributions each contained evidence of two peaks of similar synonymous divergences. For example, in *M. edulis* these peaks are located at median Ks of 0.6132 and 1.8196 (Table S5). We extended the analysis to all Bivalvia (class). Out of the 46 whole genomes available, only 7 (including the three Mytilinae) have gene models allowing further analyses (Table S6). All exhibit evidence of α-WGD and β-WGD (Table S5). The median value, for Ks peaks, is compatible with a shared WGD event (compatible age) indicating that these taxa diverged after their most recent WGDs.

Identification and functional analysis of positively selected genes

Figure 4 shows the mean Ka/Ks ratios for each orthologs, unigene cluster, in *M. edulis* and *M. galloprovincialis* (Fig. 4A); *M. edulis* and *M. coruscus* (Fig. 4B) and *M.*

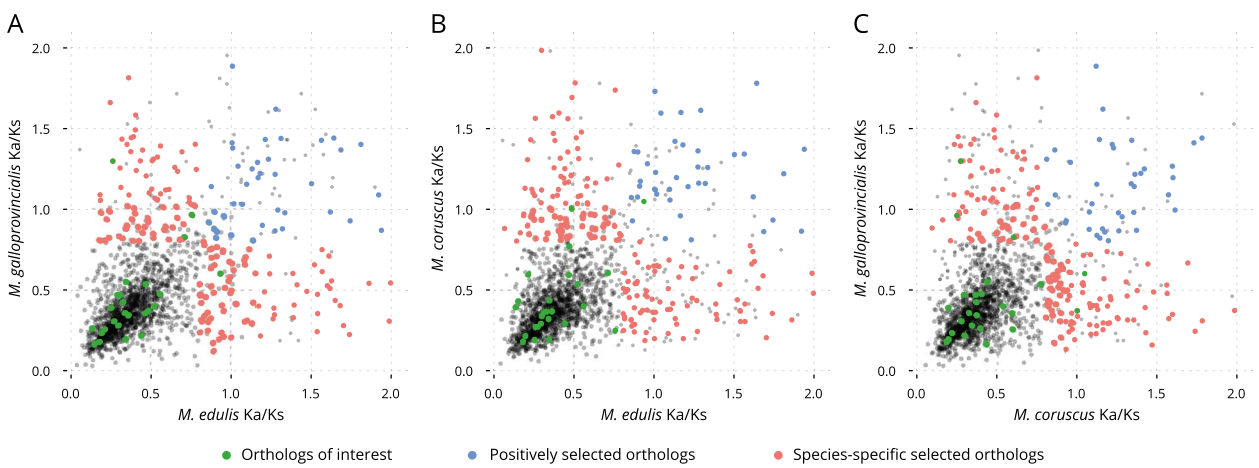

Fig. 4 Mean Ka/Ks ratios for each orthologs cluster in the **A** *M. edulis* and *M. galloprovincialis*; **B** *M. edulis* and *M. coruscus*; **C** *M. coruscus* and *M. galloprovincialis*

coruscus and *M. galloprovincialis* (Fig. 4C). In each figure, genes clusters are colour coded to indicate groups of orthologs positively selected in both species (blue), species-specific positively selected orthologs (red), and groups of orthologs of interest (green) involved in immunity, stress response and shell formation. Collectively, the data in Fig. 4 provide a new insight into positive selection occurring in the three *Mytilus* species object of this study.

The functional analysis of positively selected orthologs has also allowed for the identification of gene duplications within assembled unigene clusters involved in key physiological processes. Here, we provide an overview of the identified functions of genes under positive selection. When *M. edulis* and *M. galloprovincialis* Ka/Ks ratios are plotted (Fig. 4A), several gene clusters can be observed to be under positive selection in both species (in blue). Of these, four were identified as contributing to immunity, stress response or shell formation: the WNT Inhibitory Factor 1 (WIF1), the nucleotide exchange factor SIL1, the kelch-like protein and the midline (MID1) protein. WIF1 contributes to several immune response functions [30], and presented a Ka/Ks value of 1.4 and 1.8 for *M. galloprovincialis* and *M. edulis* respectively. SIL1 is a protein that interacts with Heat shock protein 70 (Hsp70) during stress responses [31], and showed a Ka/Ks value of 1.4 and 1.6 for *M. galloprovincialis* and *M. edulis*, respectively. The kelch-like protein facilitates protein binding and dimerisation [32], and presented a Ka/Ks value of 1.4 and 1.7 for *M. galloprovincialis* and *M. edulis*, respectively. Finally, the MID1 protein, presenting E3 ubiquitin ligase activity [33], showed a Ka/Ks value of 1.4 and 1.6 for *M. galloprovincialis* and *M. edulis*, respectively.

Many of the gene clusters were also found to be under intense positive selection in *M. galloprovincialis*, but under intense purifying selection in *M. edulis* or vice versa (in red). These genes are of particular interest as they indicate relatively rapid divergence between the two species. In Fig. 4, the genes with highly divergent selection are: the Glycolipid transfer protein, a ubiquitous protein characterised by their ability to accelerate the intermembrane transfer of glycolipids [34]; The SGNH Hydrolase-Like Protein for which no function has been identified [35]; The vitellogenic carboxypeptidase-like protein, involved in key developmental processes [36]. Furthermore, two unknown proteins with Ka/Ks values of 1.7 and 1.5 for *M. galloprovincialis* and Ka/Ks value of 0.2 and 0.4 for *M. edulis* have also been identified. On the other hand, gene clusters with high Ka/Ks values for *M. edulis* but low for *M. galloprovincialis* included: Mucolipin, which promotes calcium homeostasis and is involved in stress response functions [37]. The KAT8 regulatory NSL complex with developmental and cellular homeostasis function [38]. The TolB-like protein, involved in a tol-dependent translocation system [39]. The RING finger protein 170, which mediates the ubiquitination and degradation of inositol 1,4,5-trisphosphate receptors, and it is involved in immune response functions [33] and finally, Fibropellin, a cell adhesion protein [40]. Only a limited number of gene clusters appear to be under positive selection amongst those commonly used in immunity, stress response and shell formation comparative studies (in green), and the majority of the genes clusters showed negative selection with the exception of four, which resulted to be all involved in immune response pathways [41]. Of these, three were positively selected in *M. galloprovincialis* and conserved in *M. edulis* (membrane-bound C-type lectin, Galectin 3, MAP kinase 4-like) and one was positively selected in *M. edulis* and conserved in *M. galloprovincialis* (TNF ligand-like 2).

The comparison of Ka/Ks values between *M. coruscus* and *M. edulis* (Fig. 4B) shows a similar picture to that of *M. edulis* and *M. galloprovincialis* (Fig. 4A). Positively selected orthologs in both species (with Ka/Ks values ranging from 1.3 to 1.9) include: the nucleotide exchange factor SIL1; the archease-like protein, related to stress response functions [42]; the MID1 protein and the thiosulfate/3-mercaptopyruvate sulfotransferase protein involved in developmental and stress response functions [43]. The ortholog clusters positively selected in *M. coruscus* but conserved in *M. edulis* (Ka/Ks values from 1.6 to 2.0 and from 0.3 to 0.7 respectively) include: the purine-nucleoside phosphorylase, which encodes an enzyme which reversibly catalyses the phosphorolysis of purine nucleosides [44]; the palmitoyl-protein thioesterase, which facilitates the morphological development of neurons and synaptic function in mature cells [45], and the ADAR protein which is an RNA-binding protein and has antiviral immunity in marine molluscs [46]. Two further proteins with unknown associated functions and with Ka/Ks values of 1.8 and 1.7 for *M. coruscus* and of 0.3 and 0.5 for *M. edulis* were also identified. Similarly, orthologs positively selected in *M. edulis* but conserved in *M. coruscus* were functionally characterised and resolved to be the same as those described for *M. edulis* and *M. galloprovincialis* (Fig. 4A). Only two gene clusters showed positive selection amongst those of physiological interest (Fig. 4B, in green). C-type lectin 7 (immunity) was positively selected in *M. coruscus* with a Ka/Ks value of 1.0 and 0.4 for *M. coruscus* and *M. edulis*, respectively; while TNF ligand-like 2 (immunity) is presenting positive selection in both species, with Ka/Ks values of 1.0 and 0.9 for *M. coruscus* and *M. edulis*, respectively.

In the *M. galloprovincialis* and *M. coruscus* comparison (Fig. 4C), positively selected proteins in both

species (in blue) include: the nucleotide exchange factor SIL1, the importin-7 protein, involved in nuclear import of histones and the homeobox protein cut-like (CUTL), involved in cell-cell adhesion interactions that are required for normal development [47]. The ortholog clusters positively selected in *M. galloprovincialis* (in red) but conserved in *M. coruscus* (Ka/Ks values from 1.4 to 1.8 and from 0.2 to 0.7 respectively) are the same as described in *M. edulis* and *M. galloprovincialis* (Fig. 4A). The ortholog clusters positively selected in *M. coruscus* (in red) but conserved in *M. galloprovincialis* (Ka/Ks values from 1.6 to 1.9 and 0.3 to 0.4 respectively) are again the same as described in *M. coruscus* and *M. edulis* (Fig. 4B). The proteins positively selected for *M. galloprovincialis* and conserved for *M. coruscus* (Fig. 4C, in green) are: the membrane-bound C-type lectin (immunity) with a Ka/Ks value of 1.3 and 0.3 for *M. galloprovincialis* and *M. coruscus*, respectively. The galectin 3 protein (immunity) with a Ka/Ks value of 1.0 and 0.3 for *M. galloprovincialis* and *M. coruscus*, respectively. And the MAP kinase 4-like protein (immunity) with a Ka/Ks value of 0.8 and 0.6 for *M. galloprovincialis* and *M. coruscus*, respectively. Finally, two genes showed positive selection among those of physiological interest (in green) in favour of *M. coruscus*, and conserved for *M. galloprovincialis*. The TNF ligand-like 2 (immunity) with a Ka/Ks value of 1.0 and 0.6 for *M. coruscus* and *M. galloprovincialis*, respectively. And C-type lectin 7 (immunity) with a Ka/Ks value of 1.0 and 0.4 for *M. coruscus* and *M. galloprovincialis*, respectively.

The vast majority of our orthologs of interest (in green) selected from the literature have not shown a substantial number of proteins under positive selection for genes related to immunity, stress response, and shell formation.

For completeness, all genes involved in immunity, stress response and shell formation under positive selection in any of the three species examined here, were identified and grouped by species (Table S7). For *M. galloprovincialis*, 6, 10 and 3 genes related to immunity, stress response and shell formation, respectively were detected. For *M. edulis*, 13, 6 and 4 genes related to immunity, stress response and shell formation, respectively were detected, and for *M. coruscus*, 10, 5 and 2 genes related to immunity, stress response and shell formation, respectively were detected.

Discussion

Reference genome and whole genome duplication

Mussels are also known as poor man's shellfish as they are inexpensive and abundant. These features have perhaps contributed to a relative neglect in the investigation of this species' genomic structural variation, and whether such structural changes can play a significant role in their evolution and ecological adaptations [20]. In the wild, mussels thrive on rocks and stones along the coast, but the majority of mussels consumed are farmed in coastal waters providing food security and employment opportunities to a multitude of fragile coastal communities worldwide [48]. Similarly, to several other molluscs classes, genomic research into *M. edulis* has been hampered by the lack of a reference genome. This bottleneck is historically linked with the technical difficulties in extracting high molecular weight genomic DNA from Molluscan tissues and thus, allow for long reads sequencing techniques to be successfully applied [20]. In addition, the relatively large genome size and a high level of heterozygosity further complicates the assembly of high-quality reference genomes in the phylum.

With the aim of shedding new light onto the genomic structure and evolution of the class Bivalvia, we sequenced the blue mussel genome, and we introduced the first evidence of WGD events in the Mytilidae family and in Bivalvia more generally. Finally, we identify genes within key physiological pathways under significant positive selection. Taken together, our results provide new insights into the Mytilidae family genome structure and introduce new genomic resources for the investigation of Bivalves evolution, population genetics and for future selective breeding activities. The genome was assembled into 3339 scaffolds with a total length of 1.83 Gb, a GC content of 32.17% and a scaffold N_{50} of 1.10 Mb. In addition, we found 1.03 Gb (56.33% of the assembly) of repeat content, 69,246 protein-coding genes, 132 rRNAs and a heterozygosity of 0.48% (Table 1). The results are equivalent with the other *Mytilus* genomes: Genome size between 1.90 Gb and 1.28 Gb, and repetitive sequences between 52.83 and 58.56% [6, 11–13]. In addition, transcriptomic data and the derived gene models are comparable with the other available Mytilidae transcriptomes. Phylogeny of the *Mytilus* (based on the mitochondrial genomes) confirms the position of *M. edulis* in the genus, with the *M. edulis* and *M. galloprovincialis* (sympatric species) separate from *M. trossulus* and *M. coruscus* (which group with *M. californianus*; Fig. 2B).

Our analysis provides, for the first time, genomic evidence for paleopolyploidy in the class Bivalvia. Combining our gene age distribution and phylogenomic analyses, we found evidence for two significant, episodic bursts of gene duplication. While some of these duplication events may be caused by other processes of gene duplication, they are compatible with WGDs observed using similar methods in plants [49] and animals [14, 50]. Ks analysis showed that an ancient WGD event and a more modern WGD event occurred before the divergence of the Bivalvia. This explains why bivalves, and molluscs more generally, present large genomes [20]. The genomic

information for *M. edulis* presented here, will help clarify the evolutionary processes in Bivalvia species and contribute to improving the understanding of the physiological and morphological diversity of Bivalvia species. Our discovery of WGDs in the ancestry of Bivalvia raises questions about the role of gene and genome duplication in their evolution. After duplication, the most likely fate of duplicated genes is the loss of one of the duplicates through non-functionalisation that occurs by accumulation of deleterious mutations [51, 52]. While common after WGD, gene loss could however play a key role in speciation [53], through a process known as divergent resolution [54]. In addition, duplicated genes may also be retained in two copies [55] and either specialise by the partitioning of ancestral gene functions (i.e. sub-functionalisation) or by the acquisition of a novel function (i.e. neo-functionalisation).

Incomplete genetic data (draft genomes and transcriptomes), as well as reduced datasets (Enzymes, RAD, or EST), made it impossible to correctly detect WGDs and duplicated genes in Bivalvia, before now. In the absence of complete genomes and the full picture of WGD events, duplicated sequences are often overlooked or wrongly interpreted. This can lead to artefacts such as high heterozygosity [23], pseudogenes, and a rapid rate of gene acquisition and loss [6]. The discovery of several events of WGD in the Bivalvia phylogeny suggests the prospect that large-scale duplications are consistent with the evolution of novelty and diversity in the physiology of mass spawners like Bivalvia. However, dating of such events remains difficult due to the lack of annotated genomes deeper in the phylogeny, which still is a priority to fully elucidate molluscan evolution.

Identification and functional analysis of positively selected genes

The functional analysis of positively selected orthologs has allowed us to compare our results with studies related to the identification of gene involved in key physiological processes. When identifying gene clusters under positive selection in both species (blue dots in Fig. 4; *M. galloprovincialis-M. edulis*, *M. coruscus-M. edulis*, and *M. galloprovincialis-M. coruscus*), we find the predominant functions for those genes are mainly related to immunity, stress responses and developmental processes. Our results agree with past studies confirming that genes related to immunity are under selection in multiple lineages, likely via adaptive evolution mechanisms linked to host-pathogens co-evolution [15]. The stress response genes presenting positive selection are related or are interacting with Hsp proteins (SIL1 and the archease-like protein). Since the marine environment has considerable concentration of bacteria and viruses, molluscs depend on cellular and molecular mediated immune responses that help them to survive under challenging conditions [56]. That is why filter-feeding animals such as bivalves rely on the intervention of shock proteins which synthesis depends on environmental stressful conditions such as temperature, salinity, hypoxia, heavy metal, and infectious pathogens [57].

Genes presenting intense positive selection in one species but intense purifying selection in the others are of interest because they indicate rapid divergence between species. Once again, the three species have their maximum Ka/Ks values in genes related to developmental processes, immunity, and stress response. Overall, genes identified as being under positive selection in this study, are consistent with the defence system of bivalves depending on the innate immune response against stressful conditions such as environmental stressors, pollution and disease outbreaks.

The identification of all the genes involved in immunity, stress response and shell formation under positive selection in any of the three species (Table S7) has provided us with relevant information that could be used in future studies to identify markers for future comparative physiology and evolution studies. Our results for *M. galloprovincialis* have shown a considerable amount of stress response proteins (10 proteins) under positive selection compared to the other two species. A significant amount of those stress response genes has documented roles in heat tolerance or direct associations to heat-stress responses, e.g., zinc finger MYM-type protein 2-like (ZMYM2), mitogen-activated protein kinase 6 (MAPK6), heat shock protein 22 (HSPB8). This is also supported by past studies [58, 59] were genomic functions previously linked to divergent temperature adaptation reflected accelerated molecular divergence between warm-adapted *M. galloprovincialis* and cold-adapted congeners, such as *M. edulis*. Molecular divergence of *M. galloprovincialis* is consistent with warm-temperature adaptation demonstrated by physiological studies. *M. galloprovincialis* also has more positive selection in stress response proteins related to heavy metal detection, transport and metal binding (e.g., arylesterase / paraoxonase, pyruvate dehydrogenase E1 component alpha subunit, inositol polyphosphate 1-phosphatase, Solute carrier family 12), than *M. edulis* and *M. coruscus*. *M. edulis* and *M. galloprovincialis* presented positively selected shell formation proteins, in the EF-hand domains, which appears to be evolving faster in the two species, albeit in different gene clusters: EF-hand domain protein, EF-Hand, calcium-binding site for *M. edulis* and EF-hand calcium-binding domain-containing protein for *M. galloprovincialis*. In bivalves, the Ca^{2+} binding EF hand domains include a Calmodulin-like protein (CaLP),

a multifunctional calcium sensor that belongs to a new member of the CaM (cell adhesion molecules) superfamily, localised in the organic layer sandwiched between nacre and prismatic (aragonite) layer (calcite) in *Pinctada fucata* [60]. Studies have shown that CalP might be involved in the growth of nacre layer and prismatic layer [61]. Our results suggest and support past studies indicating that closely related bivalves use different secretory repertoires to construct their shell [62] which might lead to positive selection at a gene level as reflected in our results. Also, shell dissolution and decreased shell growth caused by ocean acidification have been described in marine bivalves [63] forcing the need for a fast environmental adaptation. Taking in account current alterations in precipitation patterns as well as stronger and more frequent heat waves and fluctuating sea surface salinities [64], our results suggest that *M. galloprovincialis* appears to be better equipped than *M. edulis* and *M. coruscus* to adapt to higher temperatures, aquatic toxicity, and contamination.

Conclusions

The recruitment, settlement, and grow-out phase of bivalve aquaculture and more precisely, in *Mytilus* spp. is strongly dependant with the environmental conditions. Therefore, the implications of climate change are not restricted to wild populations. Strong changes in local environmental conditions may limit production and force the relocation of grow-out sites to suitable areas. Thinner and weaker shells will facilitate their rupture during transportation and increase losses due to predation. Genomic selection studies and identification of molecular markers can favour the development of genetically improved lines for multiple traits and facilitate the management of genetic variability. The development of high-quality assembled genomes, as provided by the current research, will favour the identification of genomic regions linked to traits responsible for environmental resilience, which will support the long-term sustainable management and exploitation of the species.

Methods

Material collection

The *M. edulis* used in this work was obtained from a female adult blue mussel gill tissue from St Andrews Bay (Scotland, UK), a location that previously reported only the presence of a pure *M. edulis* population [3]. Gill tissue was dissected, stored in 95% ethanol and shipped to Novogene Ltd. (Cambridge, UK) for DNA extraction and sequencing. A sub-sample was tested to confirm the species identification using Wilson et al. [3] test panel and protocol.

Library construction and sequencing

High-quality DNA was used for subsequent library preparation and sequencing using both the PromethION and Illumina platforms at Novogen UK (Novogene UK Company Ltd., UK). To obtain long non-fragmented sequence reads, 15 µg of genomic DNA was sheared and size-selected (30–80 kb) with a BluePippin and a 0.50% agarose Gel cassette (Sage Science, USA). The selected fragments were processed using the Ligation Sequencing 1D Kit (Oxford Nanopore, UK) as directed by the manufacturer's instructions and sequenced using the PromethION DNA sequencer (Oxford Nanopore, UK) for 48 h. For the estimation and correction of genome assembly, an Illumina DNA paired-end library with an insert size of 350 bp was built in compliance with the manufacturer's protocol and sequenced on an Illumina HiSeq X Ten platform (Illumina Inc., USA) with paired-end 150 nt read layout.

RNA isolation, cDNA library construction and sequencing

The total RNA was extracted using the TRIzol reagent (Invitrogen, USA) according to the manufacturer's instructions. The preparation and sequencing reactions of cDNA library were done by Novogene Ltd. Briefly, the poly (A) messenger RNA was isolated from the total RNA with oligo (dT) attached magnetic beads (Illumina Inc., USA). Fragmentation was carried out using divalent cations under elevated temperature in Illumina proprietary fragmentation buffer. Double-stranded complimentary DNAs (cDNAs) were synthesised, and sequencing adaptors were ligated according to the Illumina manufacturer's protocol (Illumina Inc., USA). After purification with AMPureXP beads, the ligated products were amplified to generate high quality cDNA libraries. The cDNA libraries were sequenced on an Illumina Hiseq 4000 platform (Illumina Inc., USA) with paired-end reads of 150 nucleotides.

Genome assembly

Reads from the two types of sequencing libraries were used independently during assembly stages. Long-reads were filtered for length (> 5000 nt) and complexity (entropy over 15), while all short reads were filtered for quality (QC > 25), length (150 nt), absence of primers / adaptors and complexity (entropy over 15) using fastp v0.20.1 [65]. Using Jellyfish v2.3.0 [66], the frequency of 17-mers and 23-mers in the Illumina filtered data was calculated with a 1 bp sliding window [67] to evaluate genome size. Long-reads were then assembled using wtdbg2 v2.5 [68] which uses fuzzy Bruijn graph as well as raven v1.5.0 [69]. As it assembles raw reads without error correction and then creates a consensus from the

intermediate assembly outputs, several error corrections, gap closing, and polishing steps have been implemented. The initial outputs have been combined using Quickmerge v0.3 [70]. The combined output was re-aligned to the long-read and polished using Minimap2 v2.17 [71] and Racon v1.4.3 [72], first with filtered reads, to bridge potential gaps, then with the filtered reads to correct for error. Finally, Pilon v1.23 [73] was used to polish and correct for sequencing error using the short-reads. The redundant contigs due to diploidy were reduced by aligning the long reads back to the assembly with Minimap2 v2.17 [71] and by passing the alignment through the Purge Haplotigs pipeline v1.1.1 [74]. This reduced the artefact scaffolds and created the final haploid representation of the genome. Scaffolds were ordered with Medusa v1.6 [75] using *M. galloprovincialis* [6, 11] and *M. coruscus* [12] genome scaffolds. Mitochondrial genome was annotated using MITOS r999 [76] and manually curated.

Repeat sequences

The transposable elements were annotated using a de novo prediction using RepeatModeler v2.0.1 [77] and LTR-Finder v1.07 [78]. The repetitive sequences yielded from these two programs were combined into a non-redundant repeat sequence library. With this library, the *M. edulis* genome was scanned using RepeatMasker v4.10 [79] for the representative sequences.

Gene models

RNA-sequencing (RNA-seq) reads of poor quality i.e. with an average quality score less than 20) or displaying ambiguous bases or too short and PCR duplicates were discarded using fastp v0.20.1 [65]. Ribosomal RNA was further removed using SortMeRNA v3.0.2 [80] against the Silva version 119 rRNA databases [81]. The cleaned RNA-seq reads were pooled and mapped to the genome using the using HiSat2 v2.2.0 [82]. A combined approach that integrates ab initio gene prediction and RNA-seq-based prediction was used to annotate the protein-coding genes in *M. edulis* genome. We used Braker v2.1.5 [83] to make de novo gene predictions. The accuracy and sensitivity of the predicted model was improved by applying iterative self-training with transcripts. The predicted coding sequences were been annotated using the InterProscan v5.46–81.0 [84, 85], Swiss-Prot release 2020_02 [86] and Pfam release 33.1 [87] databases. For classification, the transcripts were handled as queries using Blast+/BlastP v2.10.0 [88], E-value threshold of 10^{-5}, against Kyoto Encyclopedia of Genes and Genomes (KEGG) r94.1 [89]. Gene Ontology [90] was recovered from the annotations of InterPro, KEGG and SwissProt. Subsequently, the classification was performed using R v4.0.0 [91] and the Venn diagram was produced by jvenn [92]. The completeness of gene regions was further tested using BUSCO v4.0.2 [93] with a Metazoa (release 10) benchmark of 954 conserved Metazoa genes.

Phylogenetic tree

Concatenated alignments constructed from all mitochondrial shared CDS sequences were used to construct a phylogenetic tree. Sequences were aligned using MACSE v10.02 [94]. A Maximum Likelihood (ML) tree was inferred under the GTR model with gamma-distributed rate variation (Γ) and a proportion of invariable sites (I) using a relaxed (uncorrelated log-normal) molecular clock in RAxML v8.2.12 [95].

Calculating Ka, Ks, and Ka/Ks values

Complete annotated nuclear genomes of Bivalvia (class) were collected. Blast+/BlastP v2.10.0 [88] was used to search for duplicated sequences in protein-coding genes between each genome. Duplicate pairs were identified as sequences that demonstrated over 70% sequence similarity, mutual protein coverage > 80%, protein length > 30 amino-acid from an all-against-all search. Duplicated sequences were aligned accounting for codon and coding frame, using MACSE v10.02 [94]. Finally, the Ka (number of non-synonymous substitutions per non-synonymous site) and Ks (number of synonymous substitutions per synonymous site) for each pair was calculated using an MPI version of KaKs_Calculator [96] under the MLPB model [97]. The Ks values > 5.0 were excluded from further analysis due to the saturated substitutions at synonymous sites. Univariate mixture models were fitted to the distributions of Ks by expectation maximisation that uses the finite mixture expectation maximisation algorithm [29, 98].

Estimation of evolution rates

A set of core-orthologs was constructed from the three complete annotated nuclear genomes of Mytilinae (subfamily), *M. edulis*, *M. galloprovincialis* [6] and *M. coruscus* [12] and were used to identity cluster of orthologous genes with a 1:1:1 ratio. Ka and Ks estimation were reported pairwise between species after MACSE v10.02 [94] and KaKs_Calculator [96] under the MLPB model [97] as above.

From a literature search of comparative mussel biology studies, we identified genes relevant to core physiological functions, specifically immunity, stress response and shell formation. Subsequently, a local BLAST search was conducted on NCBI to identify the genes of interest in the available genomes of the three species with a cut-off point of 80% sequences similarity.

Abbreviations
cDNA: Complimentary DNA; *CDS*: Coding DNA Sequence; *EBI*: European Bioinformatics Institute; *GO*: Gene Ontology; *GTR*: General Time-Reversible model; *Ka*: Non-synonymous mutation rate; *KEGG*: Kyoto Encyclopedia of Genes and Genomes; *Ks*: Synonymous mutation rate; *ML*: Maximum Likelihood; *NCBI*: National Center for Biotechnology Information; *RAD*: Restriction site associated DNA; *RNA-seq*: RNA-sequencing; *SNP*: Single Nucleotide Polymorphism; *WGD*: Whole Genome Duplication.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12864-022-08575-9.

Additional file 1: Figure S1. The k-mer distribution used for the estimation of genome size. The heterozygous and homozygous peaks of k-mer depth are clearly markers, suggesting a high-complexity genome. A The 17-mer distribution. Predicted genome size, 1,010,184,781 nt; B The 23-mer distribution. Predicted genome size, 1,096,306,163 nt. **Table S1.** Sequencing data, summary statistics. _ Estimation based on *M. galloprovincialis* and *M. coruscus* genomes size. **Table S2.** RepeatMasker statistics. _ repeats fragmented by insertions or deletions have been counted as one element. y LTR Finder results: 255,413 LTR pairs over 4932 regions and 151,703,353 bp. **Table S3.** Summary of annotation results for *M. edulis* gene models using a range of databases. _InterPro covers 12 databases (CDD-3.17,Coils-2.2.1, Gene3D-4.2.0, Hamap-2020 01, MobiDBLite-2.0, PANTHER-14.1, PRINTS-42.0, ProSitePatterns-2019 11, ProSitePro_les-2019 11, SFLD-4, SMART-7.1, SUPERFAMILY-1.75, TIGRFAM-15.0). **Table S4.** Mytilinae (subfamily) mitochondrial genomes. **Table S5.** Bivalvia (class) genome where Ka & Ks estimations were possible: All exhibit evidences of _-WGD and _-WGD. _ this study. **Table S6.** Bivalvia (class) genome and availability of gene models and annotations. _ this study. **Table S7.** Genes involved in immunity, stress response and shell formation under positive selection in *M. galloprovincialis*, *M. edulis* and *M. coruscus*.

Acknowledgements
A special thank you goes to Dr. Nick Lake and Dr. Eleanor Adamson and to the team at Novogene Ltd. For the purpose of open access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

Authors' contributions
S.C. sourced the funding, co-developed the conceptual idea, supervised the findings of this work and edited the manuscript. M.B. co-developed the concept with S.C., curated all the computational and bioinformatics elements of the study and edited the manuscript. A.C.F. sourced and prepared the biological material, wrote the first draft of the manuscript and conducted the functional analysis of all orthologs, A.D. extracted the RNA and edited the final manuscript. The author(s) read and approved the final manuscript.

Funding
The NERC SUPER Doctoral Training Program, Fishmongers' Company, Association of Scottish Shellfish Growers, the Sustainable Aquaculture Innovation Centre, and the University of Stirling funded this work. The funding bodies played no role in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

Availability of data and materials
The raw sequencing reads of all libraries are available from EBI/ENA via the accession numbers ERR4296957-ERR4296958 (long reads), ERR4296954-ERR4296955 (short reads) and ERR4172341 (RNA-seq). The assembled genome (MEDL1; GCA_905397895.1) is available in EBI with the accession numbers ERS4576331, project PRJEB38403.

Declarations

Ethics approval and consent to participate
This work was approved by the University of Stirling Ethics Committee (Animal Welfare and Ethics Review Board). Animal handling and collection in this study was carried out following its approved guidelines and regulations.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Author details
[1]Faculty of Natural Sciences, Institute of Aquaculture, University of Stirling, Stirling FK9 4LA, UK. [2]International Marine Centre, Loc. Sa Mardini snc, 09170 Torre Grande, OR, Italy.

Received: 25 August 2021 Accepted: 20 April 2022
Published online: 02 May 2022

References

1. Riginos C, Cunningham CW. Local adaptation and species segregation in two mussel (*Mytilus edulis* × *Mytilus trossulus*) hybrid zones. Mol Ecol. 2004;14:381–400.
2. Gosling EM. Systematics and geographic distribution of *Mytilus*. Dev Aquac Fish Sci. 1992;25:1–20.
3. Wilson J, Matejusova I, McIntosh RE, Carboni S, Bekaert M. New diagnostic SNP molecular markers for the *Mytilus species* complex. PLoS One. 2018;13:e0200654.
4. El Ayari T, Trigui El Menif N, Hamer B, Cahill AE, Bierne N. The hidden side of a major marine biogeographic boundary: a wide mosaic hybrid zone at the Atlantic-Mediterranean divide reveals the complex interaction between natural and genetic barriers in mussels. Heredity. 2019;122:770–84.
5. Gosling EM. Speciation and species concepts in the marine environment. In: Beaumont AR, editor. Genetics and evolution of aquatic organisms. London: Chapman and Hall; 1994. p. 1–14.
6. Gerdol M, Moreira R, Cruz F, Gómez-Garrido J, Vlasova A, Rosani U, et al. Massive gene presence-absence variation shapes an open pan-genome in the Mediterranean mussel. Genome Biol. 2020;21:275.
7. Freeman AS, Byers JE. Divergent induced responses to an invasive predator in marine mussel populations. Science. 2006;313:831–3.
8. Modak TH, Literman R, Puritz JB, Johnson KM, Roberts EM, Proestou D, et al. Extensive genome-wide duplications in the eastern oyster (*Crassostrea virginica*). Philos Trans R Soc B Biol Sci. 2021;376. https://doi.org/10.1098/rstb.2020.0164.
9. Sun S, Li Q, Kong L, Yu H. Limited locomotive ability relaxed selective constraints on molluscs mitochondrial genomes. Sci Rep. 2017;7:10628.
10. Sigwart JDD, Lindberg DRR, Chen C, Sun J. Molluscan phylogenomics requires strategically selected genomes. Philos Trans R Soc B Biol Sci. 2021;376:20200161.
11. Murgarella M, Puiu D, Novoa B, Figueras A, Posada D, Canchaya C. A first insight into the genome of the filter-feeder mussel *Mytilus galloprovincialis*. PLoS One. 2016;11:e0151561.
12. Li R, Zhang W, Lu J, Zhang Z, Mu C, Song W, et al. The whole-genome sequencing and hybrid assembly of *Mytilus coruscus*. Front Genet. 2020;11:440.
13. Yang J-L, Feng D-D, Liu J, Xu J-K, Chen K, Li Y-F, et al. Chromosome-level genome assembly of the hard-shelled mussel *Mytilus coruscus* , a widely distributed species from the temperate areas of East Asia. Gigascience. 2021;10:giab024.
14. Berthelot C, Brunet F, Chalopin D, Juanchich A, Bernard M, Noël B, et al. The rainbow trout genome provides novel insights into evolution after whole-genome duplication in vertebrates. Nat Commun. 2014;5:3657.
15. Oliver TA, Garfield DA, Manier MK, Haygood R, Wray GA, Palumbi SR. Whole-genome positive selection and habitat-driven evolution in a shallow and a deep-sea urchin. Genome Biol Evol. 2010;2:800–14.
16. Dean AM, Thornton JW. Mechanistic approaches to the study of evolution: the functional synthesis. Nat Rev Genet. 2007;8:675–88.
17. Storz JF, Bridgham JT, Kelly SA, Garland T. Genetic approaches in comparative and evolutionary physiology. Am J Phys Regul Integr Comp Phys. 2015;309:R197–214.

18. Linnen CR, Kingsley EP, Jensen JD, Hoekstra HE. On the origin and spread of an adaptive allele in deer mice. Science. 2009;325:1095–8.
19. Natarajan C, Hoffmann FG, Lanier HC, Wolf CJ, Cheviron ZA, Spangler ML, et al. Intraspecific polymorphism, interspecific divergence, and the origins of function-altering mutations in deer mouse hemoglobin. Mol Biol Evol. 2015;32:978–97.
20. Davison A, Neiman M. Pearls of wisdom-a Theo murphy issue on molluscan genomics. Philos Trans R Soc B Biol Sci. 2021;376. https://doi.org/10.1098/rstb.2020.0151.
21. Regan T, Bean TP, Ellis T, Davie A, Carboni S, Migaud H, et al. Genetic improvement technologies to support the sustainable growth of UK aquaculture. Rev Aquac. 2021;13:1958–85.
22. Benadelmouna A, Saunier A, Ledu C, Travers M-A, Morga B. Genomic abnormalities affecting mussels (*Mytilus edulis-galloprovincialis*) in France are related to ongoing neoplastic processes, evidenced by dual flow cytometry and cell monolayer analyses. J Invertebr Pathol. 2018;157:45–52.
23. Vendrami DLJ, De Noia M, Telesca L, Brodte E-M, Hoffman JI. Genome-wide insights into introgression and its consequences for genome-wide heterozygosity in the *Mytilus* species complex across Europe. Evol Appl. 2020;13:2130–42.
24. Coolen JWP, Boon AR, Crooijmans R, Pelt H, Kleissen F, Gerla D, et al. Marine stepping-stones: connectivity of *Mytilus edulis* populations between offshore energy installations. Mol Ecol. 2020;29:686–703.
25. Boore JL, Medina M, Rosenberg LA. Complete sequences of the highly rearranged molluscan mitochondrial genomes of the Scaphopod *Graptacme eborea* and the bivalve *Mytilus edulis*. Mol Biol Evol. 2004;21:1492–503.
26. Breton S, Burger G, Stewart DT, Blier PU. Comparative analysis of gender-associated complete mitochondrial genomes in marine mussels (*Mytilus* spp.). Genetics. 2006;172:1107–19.
27. Lee Y, Kwak H, Shin J, Kim S-C, Kim T, Park J-K. A mitochondrial genome phylogeny of Mytilidae (Bivalvia: Mytilida). Mol Phylogenet Evol. 2019;139:106533.
28. Shi T, Huang H, Barker MS. Ancient genome duplications during the evolution of kiwifruit (Actinidia) and related Ericales. Ann Bot. 2010;106:497–504.
29. Tiley GP, Barker MS, Burleigh JG. Assessing the performance of Ks plots for detecting ancient whole genome duplications. Genome Biol Evol. 2018;10:2882–98.
30. Capt C, Bouvet K, Guerra D, Robicheau BM, Stewart DT, Pante E, et al. Unorthodox features in two venerid bivalves with doubly uniparental inheritance of mitochondria. Sci Rep. 2020;10:1087.
31. Fu X, Sun Y, Wang J, Xing Q, Zou J, Li R, et al. Sequencing-based gene network analysis provides a core set of gene resource for understanding thermal adaptation in Zhikong scallop *Chlamys farreri*. Mol Ecol Resour. 2014;14:184–98.
32. Shi X, Xiang S, Cao J, Zhu H, Yang B, He Q, et al. Kelch-like proteins: physiological functions and relationships with diseases. Pharmacol Res. 2019;148:104404.
33. Zanchetta ME, Meroni G. Emerging roles of the TRIM E3 ubiquitin ligases MID1 and MID2 in cytokinesis. Front Physiol. 2019;10:274.
34. Brown RE, Mattjus P. Glycolipid transfer proteins. Biochim Biophys Acta Mol Cell Biol Lipids. 2007;1771:746–60.
35. Le LTHL, Yoo W, Jeon S, Kim KK, Kim TD. Characterization and immobilization of a novel SGNH family esterase (LaSGNH1) from *lactobacillus acidophilus* NCFM. Int J Mol Sci. 2019;21:91.
36. Sui Y-P, Liu X-B, Chai L-Q, Wang J-X, Zhao X-F. Characterization and influences of classical insect hormones on the expression profiles of a molting carboxypeptidase a from the cotton bollworm (*Helicoverpa armigera*). Insect Mol Biol. 2009;18:353–63.
37. Zhang X, Chen W, Gao Q, Yang J, Yan X, Zhao H, et al. Rapamycin directly activates lysosomal mucolipin TRP channels independent of mTOR. PLoS Biol. 2019;17:e3000252.
38. Radzisheuskaya A, Shliaha PV, Grinev VV, Shlyueva D, Damhofer H, Koche R, et al. Complex-dependent histone acetyltransferase activity of KAT8 determines its role in transcription and cellular homeostasis. Mol Cell. 2021;81:1749–65 e8.
39. Carr S, Penfold CN, Bamford V, James R, Hemmings AM. The structure of TolB, an essential component of the Tol-dependent translocation system, and its protein–protein interaction with the translocation domain of colicin E9. Structure. 2000;8:57–66.
40. Nie H, Wang H, Jiang K, Yan X. Transcriptome analysis reveals differential immune related genes expression in *Ruditapes philippinarum* under hypoxia stress: potential HIF and NF-κB crosstalk in immune responses in clam. BMC Genomics. 2020;21:318.
41. Gerdol M, Venier P. An updated molecular basis for mussel immunity. Fish Shellfish Immunol. 2015;46:17–38.
42. Auxilien S, El Khadali F, Rasmussen A, Douthwaite S, Grosjean H. Archease from *Pyrococcus abyssi* improves substrate specificity and solubility of a tRNA m5C methyltransferase. J Biol Chem. 2007;282:18711–21.
43. Mao G, Wang R, Guan Y, Liu Y, Zhang S. Sulfurtransferases 1 and 2 play essential roles in embryo and seed development in *Arabidopsis thaliana*. J Biol Chem. 2011;286:7548–57.
44. Stoeckler JD, Poirot AF, Smith RM, Parks RE, Ealick SE, Takabayashi K, et al. Purine nucleoside phosphorylase. 3. Reversal of purine base specificity by site-directed mutagenesis. Biochemistry. 1997;36:11749–56.
45. Koster KP. AMPAR palmitoylation tunes synaptic strength: implications for synaptic plasticity and disease. J Neurosci. 2019;39:5040–3.
46. Green TJ, Rolland J-L, Vergnes A, Raftos D, Montagnani C. OsHV-1 countermeasures to the Pacific oyster's anti-viral response. Fish Shellfish Immunol. 2015;47:435–43.
47. Pérez-Parallé ML, Carpintero P, Pazos AJ, Abad M, Sánchez JL. The HOX gene cluster in the bivalve mollusc *Mytilus galloprovincialis*. Biochem Genet. 2005;43:417–24.
48. Smaal AC, Ferreira JG, Grant J, Petersen JK, Strand Ø. Goods and Services of Marine Bivalves. Cham: Springer International Publishing; 2019.
49. Badouin H, Gouzy J, Grassa CJ, Murat F, Staton SE, Cottret L, et al. The sunflower genome provides insights into oil metabolism, flowering and Asterid evolution. Nature. 2017;546:148–52. https://doi.org/10.1038/nature22380.
50. Li Z, Tiley GP, Galuska SR, Reardon CR, Kidder TI, Rundell RJ, et al. Multiple large-scale gene and genome duplications during the evolution of hexapods. Proc Natl Acad Sci. 2018;115:4713–8.
51. Nei M, Roychoudhury AK. Probability of fixation and mean fixation time of an overdominant mutation. Genetics. 1973;74:371–80.
52. Takahata N, Maruyama T. Polymorphism and loss of duplicate gene expression: a theoretical study with application of tetraploid fish. Proc Natl Acad Sci. 1979;76:4521–5.
53. Lynch M. The evolutionary fate and consequences of duplicate genes. Science. 2000;290:1151–5.
54. Taylor JS, Van de Peer Y, Meyer A. Genome duplication, divergent resolution and speciation. Trends Genet. 2001;17:299–301.
55. Ohno S. Evolution by gene duplication. Berlin, Heidelberg: Springer-Verlag; 1970.
56. Pourmozaffar S, Tamadoni Jahromi S, Rameshi H, Sadeghi A, Bagheri T, Behzadi S, et al. The role of salinity in physiological responses of bivalves. Rev Aquac. 2020;12:1548–66.
57. Wan Q, Whang I, Lee J. Molecular and functional characterization of HdHSP20: a biomarker of environmental stresses in disk abalone Haliotis discus discus. Fish Shellfish Immunol. 2012;33:48–59.
58. Saarman NP, Kober KM, Simison WB, Pogson GH. Sequence-based analysis of thermal adaptation and protein energy landscapes in an invasive bluemussel (*Mytilus galloprovincialis*). Genome Biol Evol. 2017;9:2739–51.
59. Popovic I, Riginos C. Comparative genomics reveals divergent thermal selection in warm- and cold-tolerant marine mussels. Mol Ecol. 2020;29:519–35.
60. Yan Z, Fang Z, Ma Z, Deng J, Li S, Xie L, et al. Biomineralization: functions of calmodulin-like protein in the shell formation of pearl oyster. Biochim Biophys Acta Gen Subj. 2007;1770:1338–44.
61. Feng D, Li Q, Yu H, Kong L, Du S. Identification of conserved proteins from diverse shell matrix proteome in *Crassostrea gigas*: Characterization of genetic bases regulating shell formation. Sci Rep. 2017;7:1–12.
62. Peterson KJ, Cotton JA, Gehling JG, Pisani D. The Ediacaran emergence of bilaterians: congruence between the genetic and the geological fossil records. Philos Trans R Soc Lond Ser B Biol Sci. 2008;363:1435–43.
63. Melzner F, Stange P, Trübenbach K, Thomsen J, Casties I, Panknin U, et al. Food supply and seawater pCO2 impact calcification and internal shell dissolution in the blue mussel *Mytilus edulis*. PLoS One. 2011;6:e24223.

64. Steeves LE, Filgueira R, Guyondet T, Chassé J, Comeau L. Past, present, and future: performance of two bivalve species under changing environmental conditions. Front Mar Sci. 2018;5:1–14.
65. Chen S, Zhou Y, Chen Y, Gu J. Fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics. 2018;34:i884–90.
66. Marçais G, Kingsford C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics. 2011;27:764–70.
67. Vurture GW, Sedlazeck FJ, Nattestad M, Underwood CJ, Fang H, Gurtowski J, et al. GenomeScope: fast reference-free genome profiling from short reads. Bioinformatics. 2017;33:2202–4.
68. Ruan J, Li H. Fast and accurate long-read assembly with wtdbg2. Nat Methods. 2020;17:155–8.
69. Vaser R, Šikić M. Raven: a *de novo* genome assembler for long reads. bioRxiv. 2021. https://doi.org/10.1101/2020.08.07.242461.
70. Solares EA, Chakraborty M, Miller DE, Kalsow S, Hall K, Perera AG, et al. Rapid low-cost assembly of the drosophila melanogaster reference genome using low-coverage, long-read sequencing. G3 (Bethesda). 2018;8:3143–54.
71. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 2018;34:3094–100.
72. Vaser R, Sović I, Nagarajan N, Šikić M. Fast and accurate *de novo* genome assembly from long uncorrected reads. Genome Res. 2017;27:737–46.
73. Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, et al. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS One. 2014;9:e112963.
74. Roach MJ, Schmidt SA, Borneman AR. Purge Haplotigs: allelic contig reassignment for third-gen diploid genome assemblies. BMC Bioinformatics. 2018;19:460.
75. Bosi E, Donati B, Galardini M, Brunetti S, Sagot MF, Lió P, et al. MeDuSa: a multi-draft based scaffolder. Bioinformatics. 2015;31:2443–51.
76. Bernt M, Donath A, Jühling F, Externbrink F, Florentz C, Fritzsch G, et al. MITOS: Improved de novo metazoan mitochondrial genome annotation. Mol Phylogenet Evol. 2013;69:313–9.
77. Flynn JM, Hubley R, Goubert C, Rosen J, Clark AG, Feschotte C, et al. RepeatModeler2 for automated genomic discovery of transposable element families. Proc Natl Acad Sci. 2020;117:9451–7.
78. Stanke M, Diekhans M, Baertsch R, Haussler D. Using native and syntenically mapped cDNA alignments to improve *de novo* gene finding. Bioinformatics. 2008;24:637–44.
79. Smit AFA, Hubley R, Green P. RepeatMasker; 2019.
80. Kopylova E, Noé L, Touzet H. SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. Bioinformatics. 2012;28:3211–7.
81. Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, et al. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Res. 2012;41:D590–6.
82. Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nat Biotechnol. 2019;37:907–15.
83. Hoff KJ, Lomsadze A, Borodovsky M, Stanke M. Whole-genome annotation with BRAKER. In: Kollmar M, editor. Gene prediction: methods and protocols. New York: Springer New York; 2019. p. 65–95.
84. Jones P, Binns D, Chang H-Y, Fraser M, Li W, McAnulla C, et al. InterProScan 5: genome-scale protein function classification. Bioinformatics. 2014;30:1236–40.
85. Mitchell AL, Attwood TK, Babbitt PC, Blum M, Bork P, Bridge A, et al. InterPro in 2019: improving coverage, classification and access to protein sequence annotations. Nucleic Acids Res. 2019;47:D351–60.
86. Bateman A, Martin MJ, O'Donovan C, Magrane M, Alpi E, Antunes R, et al. UniProt: the universal protein knowledgebase. Nucleic Acids Res. 2017;45:D158–69.
87. El-Gebali S, Mistry J, Bateman A, Eddy SR, Luciani A, Potter SC, et al. The Pfam protein families database in 2019. Nucleic Acids Res. 2019;47:D427–32.
88. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990;215:403–10. https://doi.org/10.1016/S0022-2836(05)80360-2.
89. Kanehisa M, Sato Y, Furumichi M, Morishima K, Tanabe M. New approach for understanding genome variations in KEGG. Nucleic Acids Res. 2019;47:D590–5.
90. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet. 2000;25:25–9. https://doi.org/10.1038/75556.
91. R Core Team. R: a language and environment for statistical computing. Vienna; 2021. https://www.r-project.org/.
92. Bardou P, Mariette J, Escudié F, Djemiel C, Klopp C. jvenn: an interactive Venn diagram viewer. BMC Bioinformatics. 2014;15:293. https://doi.org/10.1186/1471-2105-15-293.
93. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. 2015;31:3210–2.
94. Ranwez V, Harispe S, Delsuc F, Douzery EJP. MACSE: multiple alignment of coding SEquences accounting for frameshifts and stop codons. PLoS One. 2011;6:e22594.
95. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 2014;30:1312–3.
96. Zhang Z, Li J, Zhao X-Q, Wang J, Wong GK-S, Yu J. KaKs_Calculator: Calculating Ka and Ks through model selection and model averaging. Genomics Proteomics Bioinformatics. 2006;4:259–63.
97. Tzeng Y-H. Comparison of three methods for estimating rates of synonymous and nonsynonymous nucleotide substitutions. Mol Biol Evol. 2004;21:2290–8.
98. Benaglia T, Chauveau D, Hunter DR, Young D. mixtools: an R package for analyzing finite mixture models. J Stat Softw. 2009;32:1–29.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions