k-mer spectra and allelic coverage analyses reveal pervasive polymorphic duplications in *Ostrea edulis*

Lila Colston-Nepali[1,2]*, Sylvie Lapègue[1], and Nicolas Bierne[2]*

Affiliations:

[1] MARBEC, Univ Montpellier, CNRS, Ifremer, IRD, Montpellier, France

[2] ISEM, Univ Montpellier, CNRS, IRD, Montpellier, France

* corresponding authors: lila.colston-nepali@umontpellier.fr ; nicolas.bierne@umontpllier.fr

Acknowledgements

The authors are very grateful to Antonio Villalba for carrying out the sampling of the four Galician flat oyster samples used in this study. They also express their thanks to Xavier Dallaire, Pierre-Alexandre Gagnaire, Céline Reisser, and Christelle Fraïsse for valued discussion and insight.

Funding

- This work was supported by the scientific management of Ifremer and the Occitanie region as part of Lila Colston-Nepali's doctoral contract.
- Data used in this work were partly produced through the GenSeq technical facilities of the “Institut des Sciences de l’Evolution de Montpellier” with the support of LabEx CeMEB, an ANR “Investissements d’avenir” program (ANR-10-LABX-04-01).

Conflict of interest disclosure

The authors declare that they comply with the PCI rule of having no financial conflicts of interest in relation to the content of the article. The authors declare the following non-financial conflict of interest: Nicolas Bierne is a recommender for PCI.

Data, statistical scripts, command lines and simulation code

The four newly sequenced *O. edulis* whole genomes are deposited in DATAREF, the Ifremer centralized and secure storage through data archiving. They are currently in the process of obtaining a DOI and being submitted to the ENA.

ABSTRACT

Marine bivalves are known for their high genetic diversity and potentially high genetic load, as well as their structurally complex genomes. The European flat oyster, *Ostrea edulis*, is no exception. In this study we performed both *k*-mer based and reference-based analyses on short-read data with high coverage (~70-160X) from five individuals. Up to one-third of heterozygous SNPs showed allelic coverage fractions that deviated from expectation. Despite no evidence of recent whole-genome duplication, we detected significant signals for genomic regions of increased ploidy, and 15% and 25% of *k*-mer pairs distant by a single nucleotide displayed triploid-like and tetraploid-like profiles respectively. In a diploid species, this indicates the presence of heterozygote and homozygote duplications. These results were confirmed by coverage-based genotype inference, which showed that deviant SNPs are located in polymorphic duplications. While duplications affected genic and non-genic regions similarly, their genomic distribution varied across chromosomes, with a notable enrichment in non-metacentric chromosomes. The latter being prone to somatic aneuploidy, we suggest that aneuploidy could act as an indirect driver, with duplications potentially buffering the expression of recessive deleterious mutations in aneuploid cells. Finally, we consider how such widespread structural variation may complicate population genomic analyses in *O. edulis* and other marine bivalves.

INTRODUCTION

Marine bivalves often exhibit very high levels of genetic diversity, making them ideal models for population genetics since the advent of allozyme markers (Ward, Woodwark, and Skibinski 1994; Bazin, Glémin, and Galtier 2006). Early studies have revealed a number of surprising observations, including heterozygote deficiencies relative to Hardy-Weinberg equilibrium (Zouros and Foltz 1984), segregation distortions in the offspring of mating crosses (Bierne et al. 1998; Launey and Hedgecock 2001), and heterozygosity-fitness correlations (David 1998). These have led to interpretations based on profuse adaptive variation (Mitton 2000; Koehn 1983), unusual demographics such as chaotic genetic patchiness (Eldon et al. 2016; David et al. 1997) and sweepstakes reproductive variance (Hedgecock and Pudovkin 2011), and high rates of somatic aneuploidy (Thiriot-Quiévreux 1986; Leitão et al. 2001), but also technical issues such as null alleles (Foltz 1986). Although the very large size of bivalve populations could explain both the high polymorphism (Harrang et al. 2013; Takeuchi 2017) leading *de facto* to technical and cytogenetical problems, and a high genetic load (Plough 2016), not all issues have been fully resolved. The transition to DNA sequence analysis has confirmed high nucleotide diversity both at silent and amino-acid mutations in marine bivalves (Sauvage et al. 2007; Harrang et al. 2013; Romiguier et al. 2014). However, recent access to structural variation in genome sequence studies has revived the idea of potentially unique bivalve genetics underpinned by widespread presence-absence (Gerdol et al. 2020) and copy number (Modak et al. 2021; Qi, Li, and Zhang 2021) variation, which have been suggested to be adaptive (i.e. maintained by selection). As for allozyme and SNP markers, though, high population size and genetic load, possibly associated with sweepstakes reproduction, might also explain this observation similarly well.

Highly polymorphic species usually have a compact genome (Lynch and Walsh 2007), though this is not always the case, such as in marine bivalves. A high proportion of their genomes are made of repetitive sequences (Šatović and Plohl 2021). This includes pearl oysters (*Pinctada fucata*, Takeuchi et al. 2012; Takeuchi et al. 2022), Pacific oysters (*Magallana gigas*, Zhang et al. 2012), Mediterranean mussels (*Mytilus galloprovincialis*, Murgarella et al. 2016), Manila clams (*Ruditapes philippinarum*, Mun et al. 2017), deep sea and shallow-water mussels (*Bathymodiolus platifrons* and *Modiolus philippinarum*, Sun et al. 2017), scallops (*Patinopecten yessoensis*, Wang et al. 2017), and golden mussels (*Limnoperna fortunei*, Uliano-Silva et al. 2018). Although repetitive sequences include highly repeated elements such as mobile elements, satellite DNA, or tandem repeats (Šatović and Plohl 2021), segmental duplications (SDs) have been found to make up 26% of the Pacific oyster genome - the highest recorded in published genomes (Qi, Li, and Zhang 2021). Studies of resequenced genomes have also revealed presence-absence variation (PAV) throughout the Mediterranean mussel genome, with approximately 25% of genes subject to PAV (Gerdol et al. 2020), but also persistent copy number variable regions (CNVRs) consisting of over 43% of the Pacific oyster genome (Qi, Li, and

Zhang 2021), and duplication variation contributing to over 16.5% of the eastern oyster (*Crassostrea virginica*) genome (Modak et al. 2021).

However, population genomics data have been predominantly obtained with low coverage genome sequencing in order to analyze many samples within constrained project budgets. These studies focus on SNPs and ignore structural variation, even though some of these SNPs are likely to be found in these parts of the genome. Downstream analyses rely on a number of assumptions about genome assemblies, although next-generation sequencing methods have inherent technical risks of error that may violate these assumptions and lead to misinterpretation of the data. For this reason complex filtering strategies are used to improve the accuracy and reliability of genome-wide SNP datasets (O'Leary et al. 2018; Hemstrom et al. 2024). Very few studies, though, consider the effect of genome complexities on next-generation sequencing. Dallaire et al. (2023) described the effects of rediploidization and repetitive elements on low-coverage whole genome sequencing in a salmonid fish, which they hypothesize causes collapsed assembled genomic regions. As a result SNPs may be called that do not conform to expected patterns of heterozygosity and allelic coverage fractions, which they refer to as 'deviant SNPs'. For example, deviant SNPs deviated from the expected allelic coverage fraction of 1/2, with a hotspot of SNPs demonstrating an allelic coverage fraction of 1/4 (Dallaire et al. 2023). These deviant SNPs have a significant effect on downstream analyses and population genetic inferences, and most filtering approaches will fail to remove them (Dallaire et al. 2023). Consideration then, of the presence of these deviant SNPs in non-model species with complex genomes, such as bivalves, represents a potential blind spot in population genomic analyses.

The European flat oyster (*Ostrea edulis*) is a marine bivalve from the family Ostreidae, or the true oysters. *O. edulis* has highly variable reproductive success, both in the wild and in experimental conditions, and a low *Ne*/*N* ratio (Hedgecock et al. 2007; Lallias et al. 2010). *O. edulis* has been found to have high genetic load causing segregation distortions in pair crosses (Bierne et al. 1998), and up to a third of nonsynonymous mutations have been estimated to be mildly deleterious in this species (Harrang et al. 2013). The genome size of *O. edulis* is also larger than other sequenced oyster species (Boutet et al. 2022; Gundappa et al. 2022; Li et al. 2023), and contains comparable repeated elements as *M. gigas*, *C. virginica*, and *Saccostrea glomerata* (Boutet et al. 2022; Li et al. 2023). Synteny analyses indicate that *O. edulis* has not experienced a recent whole genome duplication event since divergence from Crassostreinae (Li et al. 2023), although such an event may have occurred at the origin of all the Ostreidae.

In this study we used *k*-mer and allelic coverage analyses on five very high coverage re-sequenced *O. edulis* genomes (70-160X) in order to infer the proportion of SNPs that deviate from the expectation of diploid genotypes in single copy orthologs. Our *k*-mer analyses showed that ~15% and ~25% of our SNPs are triploid-like (3X coverage) and tetraploid-like (4X coverage) respectively. Allelic coverage analyses also identified SNPs that did not adhere to

diploid expectation. Investigating how these are shared between individuals, we obtained evidence that they are located in polymorphic duplications.

METHODS

1. *Sampling and sequencing*

Five *O. edulis* samples with high coverage short-read sequencing were used: four newly sequenced samples from Galicia, Spain (sequencing depth ranging from ~70-160X coverage), and one from Brittany, France, using publicly available reads (SRA accession: SRR1723031, ~70X coverage, named ROSC in this study). Three of the Galician samples (2C4M, 6C9M, 8C17M) came from a culture raft in Cambados, Ría de Arousa, and one sample (4C6M) came from a natural bed in Ría de Ferrol. DNA was extracted from mantle tissue for all four using Machery-Nagel's NucleoSpin Tissue extraction kit. The four DNA samples were sent to the MGX sequencing platform, which produced the TruSeq DNA PCR-Free libraries, followed by sequencing on Illumina Novaseq 6000 on an S4 line.

2. *Reference-free k-mer analyses*

As a first step, reference-free *k*-mer based analyses were used. *k*-mer approaches allow users to evaluate large, and potentially complex, genomic datasets computationally efficiently and without the need for a reference genome (Jenike et al. 2025). *k*-mer frequency histograms were computed using the program FastK (https://github.com/thegenemyers/FASTK) on the raw reads, using a *k*-mer count of 21, and then two different tools were used to analyse the *k*-mer spectrum. First, GenomeScope 2.0 (Ranallo-Benavidez, Jaron and Schatz 2020) was used to analyze the coverage of the *k*-mers, and fit to a model of 2 x *p* negative binomial distributions (*p* = ploidy) representing the unique and multi-copy heterozygous and homozygous *k*-mers. A ploidy of 2, or diploid, was used. GenomeScope 2.0 was also used to estimate genome size and sequencing error in the dataset.

Next, the program Smudgeplot (v. 0.3.0) (Ranallo-Benavidez, Jaron and Schatz 2020) was used to find *k*-mer pairs that differ by exactly one nucleotide, representing either heterozygous alleles, or complex genome structures (i.e. duplications, paralogs, etc.). Total coverage of the *k*-mer pairs is plotted against the minor *k*-mer coverage to reveal hotspots that represent the haplotype structures present, and their frequency, in the genome. Input parameters for error margins and coverage ranges were set based on GenomeScope 2.0 output. Unlike GenomeScope 2.0, ploidy is not needed *a priori* to run Smudgeplot.

In order to investigate and compare genome-wide duplication events in other ostreids, GenomeScope 2.0 and Smudgeplot analyses were also performed on publicly available

short-read datasets of three oyster species: the lamellated oyster *O. denselamellosa* (SRA accession: SRR19238449), the Olympia oyster *O. lurida* (SRA accessions: SRR3306651-SRR3306656), and the Pacific oyster *M. gigas* (SRA accession: SRR7474375). To compare the duplications in the flat oyster to a species with known PAV (Gerdol et al. 2020), *k*-mer analyses were also used on the Mediterranean mussel *M. galloprovincialis* (SRA accession: SRR17154976). Further analyses were implemented on two other *Mytilus* species: the common mussel *M. edulis* (SRA accession: SRR28968509), and the thick-shelled mussel *M. coruscus* (SRA accession: SRR10502233). The SRA toolkit (https://github.com/ncbi/sra-tools) was used to download and extract FASTQ files. For *O. lurida*, six sequencing runs of the same individual were pooled.

3. *Reference-based analyses*

3.1 Alignment to reference genome and SNP calling

Sequence data was trimmed using the program fastp (Chen et al. 2018) to filter for quality and remove both polyG tails and Illumina adapters. Samples were aligned to an *O. edulis* reference genome (Li et al. 2023, GenBank assembly: GCA_032173915.1) using the program BWA-MEM (Li and Durbin 2010). Mapped reads were then filtered using the program Sambamba (Tarasov et al. 2015) to remove reads that did not pair properly, and secondary alignments. SNPs were called using bcftools mpileup and bcftools call (Danecek et al. 2021), with maximum coverage increased to 2000, resulting in 20,868,611 sites.

3.2 Dosage based inference of genotypes

Output VCF files were filtered to include only biallelic SNPs, and to remove both indels and all missing data using vcftools (Danecek et al. 2011) and bcftools (Danecek et al. 2021). SNPs mapping to unplaced scaffolds were removed in R (R Core Team 2023), with 15,106,345 remaining sites for downstream analyses. A custom python script was used to extract the reference (A) and alternative (B) allele counts, total allele count, and calculate the allele coverage fraction (site frequency) per SNP (defined as alternative reads / total reads) for each of the 5 individuals. The output CSV files were then analyzed in R (R Core Team 2023). Firstly, allelic coverage fraction distributions for heterozygous sites were assessed using histograms of alternative allele frequency for each individual. Histograms of site coverage were then used to estimate the coverage per individual, and determine the expected relative depth for 2X (diploid), 3X (triploid-like), and 4X (tetraploid-like) sites. Higher relative depth values (i.e. 5X and above) that correspond to the presence of alleles with more than two repeated copies were ignored, as they represent a small proportion of SNPs. The mode of the distribution was used as an estimate for the 2X relative depth. This estimate was then used alongside the density plots to estimate the expected 1X, 3X and 4X relative depth for each individual. Density plots were visualized with

ggplot2 (Wickham 2016) to investigate the relationship between the relative depth and the allelic coverage fraction. Gene dosage clusters were identified and circumscribed for 12 potential genotypes at biallelic sites: AA, AB, BB, AAA, AAB, BBA, BBB, AAAA, AAAB, AABB, BBBA, BBBB (see Supplementary Fig. 1). These dosage-inferred genotypes (DI-genotypes), were identified for a subset of SNPs for which we have good confidence of the inferred genotype, and were then used to further investigate variation between individuals, chromosomes, and functional categories (e.g. genic vs. non-genic).

3.2.1 Variation between individuals

To determine if the same SNPs were categorised into the same DI-genotype in different individuals, SNP lists were created for each of the 12 DI-genotypes for each individual. Comparisons were made for each individual (i.e. the percentage of AA SNPs in individual X that are present in each of the 12 categories in individual Y), leading to a total of 20 repetitions for 144 comparisons.

3.2.2 Location and density of different genomic regions

The 12 DI-genotypes were combined to represent the three types of dosage 2X, 3X and 4X (i.e. AA, AB, BB were pooled as 2X etc.). To see if the 3X and 4X SNPs are present in specific regions in the genome, the tool CMplot (Yin et al. 2021) was used to visualize SNP-density. The proportions of 2X, 3X and 4X SNPs were also calculated per chromosome for each individual.

3.2.3 Functional annotation of SNPs

The program SnpEff (Cingolani et al. 2012) was used to annotate the SNPs. A new database for the flat oyster was built using annotation, coding sequence (CDS), and protein sequence files for the same assembly that was used as a reference genome (Li et al. 2023). Annotation categories were simplified into four main groups- intergenic, intronic, synonymous and nonsynonymous variants (see Supplementary Table 1 for details). The proportion of 2X, 3X, and 4X SNPs was calculated for each of these annotation categories.

RESULTS AND DISCUSSION

***k*-mer analyses reveals a high proportion of non-diploid *k*-mers in *O. edulis* and other bivalve genomes**

The *k*-mer coverage distributions obtained with Genomescope 2.0 showed that a significant proportion of *k*-mers were classified outside of the standard diploid coverage expectation in the genomes of all five individuals (Fig. 1A and Supplementary Figures 2A-5A), with signals for

both 3X and 4X *k*-mers present. To further investigate this signal, we conducted Smudgeplot analyses that use the co-distribution of total coverage and the minor *k*-mer coverage of closely related *k*-mer pairs to determine how many copies of each allele exist across a genome. Smudgeplot allows the delineation of distinct dosage clusters that would be otherwise difficult to determine using coverage alone (even when coverage is high, as is the case here). The AAB cluster with a 3X coverage and an allelic coverage fraction of 1/3 is not expected to be found in a diploid genome, except in cases of heterozygous duplication (where one chromosome has a 1-copy allele and the other has a 2-copy allele). Smudgeplot analyses indicated strong evidence for non-diploid *k*-mer pairs, with 14-16% with a 3X (AAB) signal, and 21-26% with a 4X (AAAB or AABB) signal (Fig. 1B and Supplementary Figures 2B-5B).

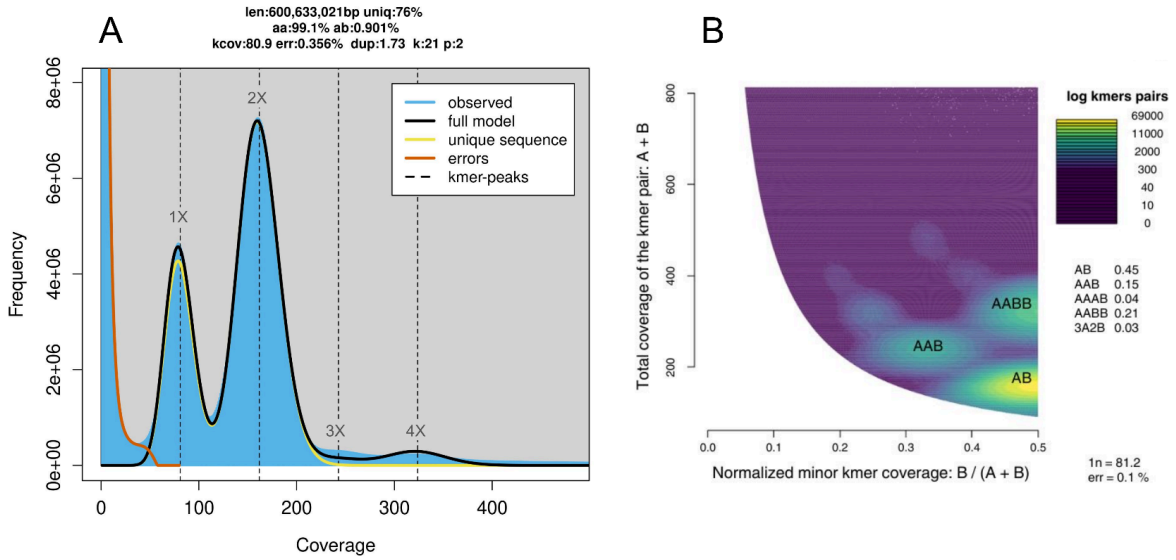

Figure 1. A) GenomeScope profile for sample 6C9M with dotted lines indicating 1X, 2X, 3X and 4X *k*-mer coverages and B) corresponding Smudgeplot analysis for the same individual, using an error threshold of 46 and a coverage range of 50-110.

Given that Smudgeplot has rarely been used in bivalves, we reanalysed several public NGS datasets in both ostreids (*Ostrea denselamellosa*, *O. lurida*, and *Magalllana gigas*), and mytilids *(Mytilus galloprovincialis*, *M. edulis*, and *M. coruscus*) in order to compare patterns of non-diploid genomic regions. We observed that 3X and 4X clusters were found in every genome, but *O. edulis* had the strongest signal for 3X and 4X regions in both Genomescope 2.0 and Smudgeplot analyses out of the species re-analysed in this study (Supplementary Figures 6-11). Further, *O. edulis* is the only species in this study that has a higher proportion of 4X than 3X *k*-mer pairs. Although *M. gigas* was described as having a very high proportion of both CNVRs and SDs (Qi, Li, and Zhang 2021), it was the species with the weakest signals amongst the ostreids analysed here. Surprisingly, a Smudgeplot analysis performed in the brooding Chilean

oyster, *O. chilensis*, indicated no *k*-mer pairs deviating from diploid expectation (see Piccoli 2024). The signal of non-diploid *k*-mer pairs in *Mytilus* mussels suggests that this may be a general feature of bivalve genomes, and that the presence or absence of one of the two copies of a duplication could be a potential driver of the presence-absence signal detected in mussels (Gerdol et al. 2020).

Reference-free *k*-mer analyses thus provide a first indication (with the high frequency of AAB, AAAB, and AABB clusters) of the presence of heterozygous duplications in *O. edulis* and other bivalve genomes. However, a reference mapping approach is needed in order to better understand the variation between individuals, how 3X and 4X SNPs cluster in the genome, how called SNPs are affected by polymorphic duplications, and to investigate the distribution of these duplications between chromosomes and between functional categories (i.e. coding vs. non-coding).

Reference mapping analysis confirms polymorphic duplications in *O. edulis*

First we investigated the fraction of allelic coverage of heterozygous SNPs and found a surprisingly large proportion (approximately 30%) of heterozygous SNPs with a coverage fraction between 0.2 and 0.4 (Fig. 2 and Supplementary Figure 12). Although routine inspection of allelic coverage should be included in best practice for NGS data analysis (Jaron et al. 2022, Bierne 2022, Ament-Velásquez 2016), it is unfortunately not reported on sufficiently in the literature. Nevertheless, it is evident that such a significant deviation from the expected distribution is highly unusual.

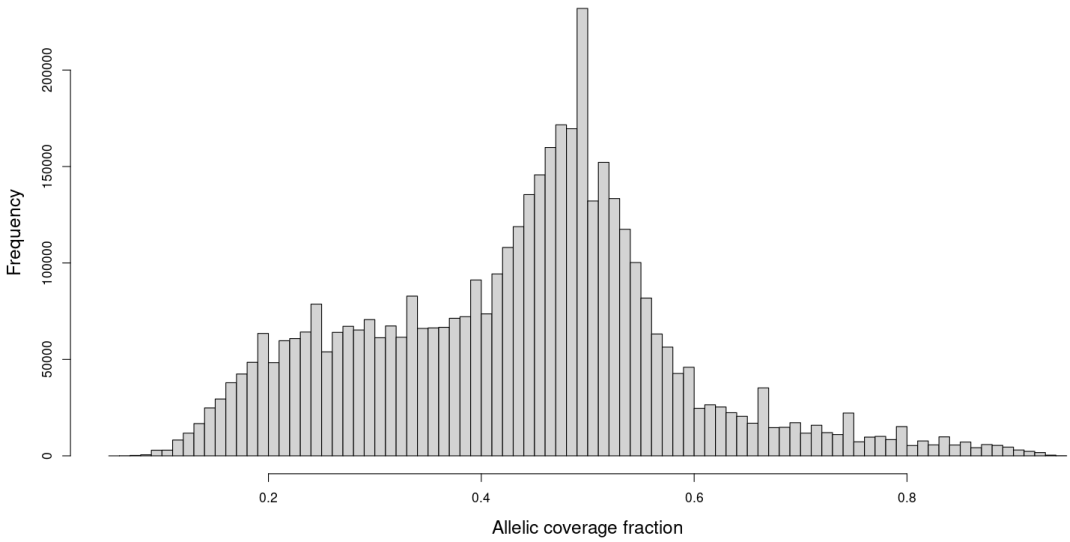

Figure 2. Histogram demonstrating the distribution of the allelic coverage fraction for called heterozygous sites in individual 6C9M.

In order to better delineate 3X and 4X SNPs, we adopted a similar approach to Smudgeplot by investigating the co-distribution of total coverage and allelic coverage fraction of heterozygote calls. SNPs adhering to the expectations for heterozygous diploid sites, AB sites (i.e. 2X relative depth and 0.5 allelic coverage fraction), had the strongest density (Fig. 3) as expected.

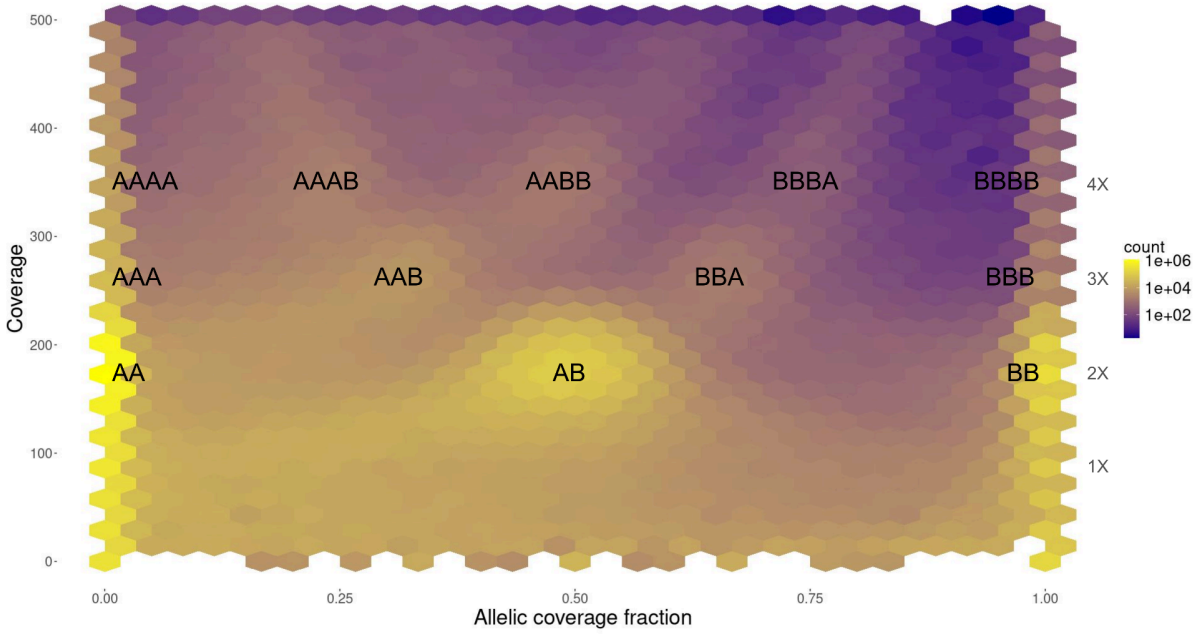

Figure 3. Density plot for sample 6C9M showing allelic coverage fraction on the x-axis and coverage on the y-axis. Marked on the right hand side of the plot is relative depth, the coverage of the SNP divided by the genome average coverage. Areas of increased density shown in brighter yellow. Locations on plot corresponding to specific dosage inferred genotypes marked.

However, corroborating the Smudgeplot results, strong density was also seen in many of the other dosage-inferred clusters corresponding to 3X and 4X genotypes as well as outside these regions (including at relative depth below 2X, as expected when one allele is less efficiently mapped with lower quality scores). This confirms that there are numerous heterozygous and homozygous duplications distributed throughout the species genome. Most SNPs that adhere to diploid expectations (2X SNPs), were also identified as regular diploid SNPs in other individuals. On average, over 95% of SNPs identified as AB in one individual were classified as AA (50.3%), AB (29.7%), or BB (15.7%) SNPs in other individuals (Fig. 4A and Supplementary Table 2).

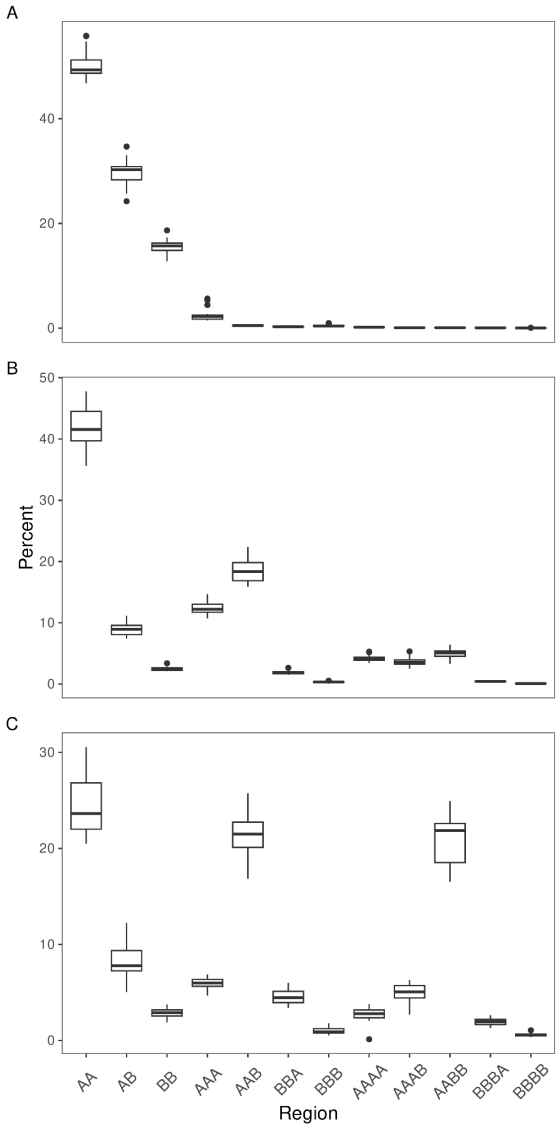

Figure 4. Boxplots showing the percentage of shared sites between SNPs categorized into **A** AB DI-genotypes, **B** AAB DI-genotypes, and **C** AABB DI-genotypes in one individual, and the 12 DI-genotypes in other individuals.

The AAB SNPs in one individual were primarily classified as AA (42%) or AAB (18.5%) in other individuals (Figure 4B). This is what we would expect when a 2-copy allele with a mutation in one copy segregates at low frequency (the duplication is mainly found at the heterozygous state). In this case individuals homozygous for the two-copy allele are rare, which could explain why only 5% of these SNPs are categorized as AABB SNPs in other individuals. The rest were classified as AAA (12.4%), AB (8.9%), AAAA (4.2%), AAAB (3.7%), and BB (2.5%) SNPs. Finally, if we look at AABB SNPs in one individual, the majority are classified as AA (24.5%), AAB (21.7%) or AABB (20.8%) SNPs in other individuals (Fig. 4C and Supplementary Table 2). SNPs switch between DI-genotypes in different individuals, suggesting

that these duplications are not fixed. This result is in better accordance with the expectation for polymorphic duplications (where the 2-copy allele has a mutation in one copy). Given that the focal individual is a homozygote for the duplication (AABB), we expect the frequency of the 2-copy allele to be higher than when the focal individual is a heterozygote for the duplication (AAB), and that we should mostly observe the three expected genotypes (AA, AAB, AABB). The rest were mainly classified as AB (8.3%), AAA (5.9%) and AAAB (5%).

The other DI-genotypes observed can be explained if we consider the possible mutational paths taken to obtain the polymorphism observed in this study (Figure 5A and 5B). First, we must have a duplication creating the 2-copy allele. The two copies are initially identical, and until a nucleotide mutation occurs, there is no SNP. A nucleotide mutation can then occur in one copy of the 2-copy allele (Figure 5A), creating a SNP at this position. This results in three alleles segregating in the population: the 1-copy allele carrying the ancestral allele of the SNP (one dose of A), the 2-copy allele carrying the ancestral allele of the SNP on both copies (two doses of A), and the 2-copy allele with the derived allele on one copy and the ancestral allele on the other (one dose of A plus one dose of B). Here allele A refers to the ancestral allele, and allele B to the derived allele. However, in the dataset A will refer to the allele carried by the reference genome, and B will refer to the alternative allele. In this scenario, we expect not only AA, AAB, and AABB DI-genotypes, but also AAA, AAAB, and AAAA DI-genotypes, which correspond to the presence of the unmutated 2-copy allele, contributing two doses of the ancestral allele in a DI-genotype. Alternatively, following duplication a mutation can occur on the 1-copy allele instead, creating an alternative mutational path (Fig. 5B). This results in three different alleles segregating in the population: the 1-copy allele carrying the ancestral allele of the SNP (one dose of A), the 2-copy allele carrying the ancestral allele of the SNP on both copies (two doses of A), and the 1-copy allele carrying the derived allele of the SNP (one dose of B). In this scenario, we expect the DI-genotypes AA, AB and BB corresponding to the nucleotide polymorphism in the 1-copy allele, but also the AAA, AAB, and AAAA genotypes corresponding to the presence of the 2-copy allele with the ancestral SNP allele on both copies, contributing two doses of A in the DI genotype.

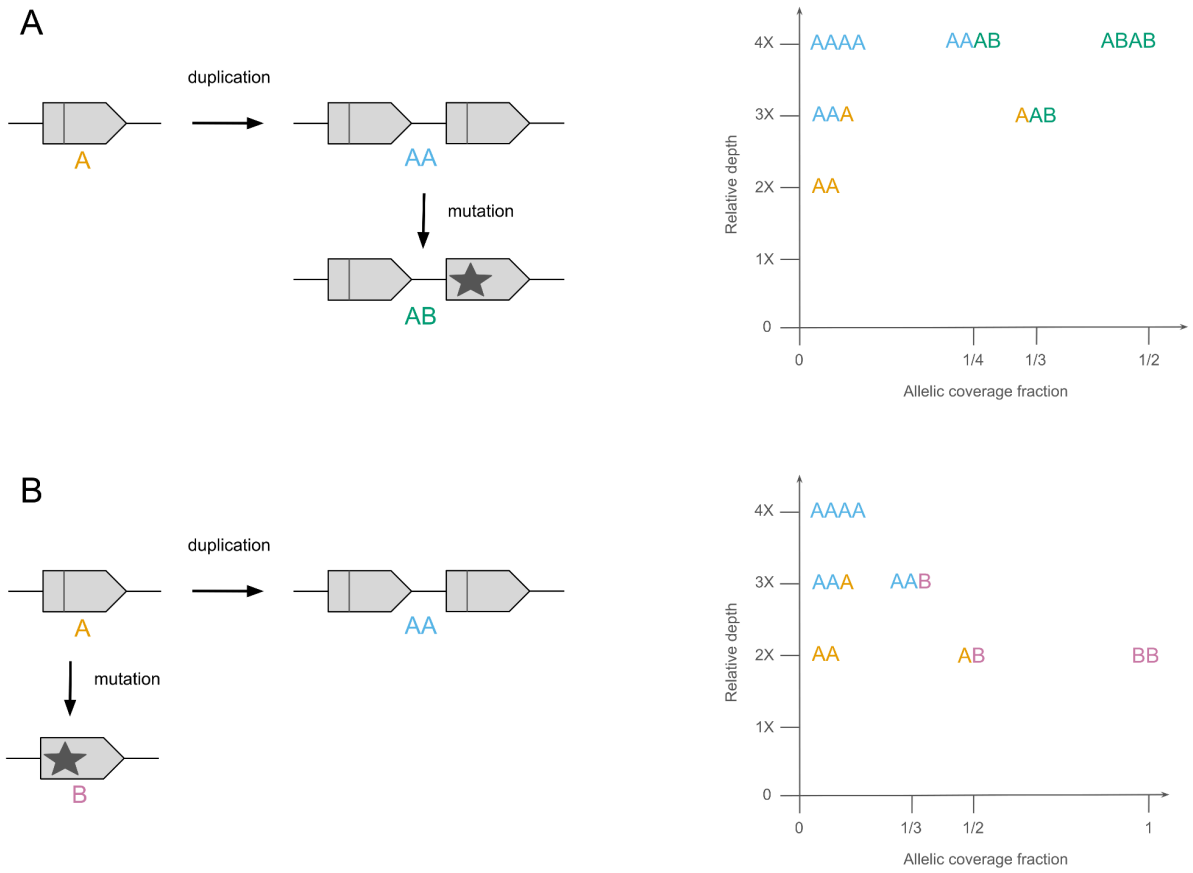

Figure 5. Two potential mutational paths leading to the polymorphic duplications seen in *O. edulis*. Each pentagon represents a portion of the genome, and the vertical lines in the pentagons represent the variant position. Stars represent mutations. Plots to the right of the mutational paths demonstrate the relative depth and allelic coverage fraction of the resulting genotypes. In mutational path **A**, a duplication event occurs, leading to a 2-copy allele. Then a mutation occurs in one of the two copies of the 2-copy allele, resulting in an ancestral variant (A) and a derived variant (B) at the focal SNP. The result is three types of allele present: a 1-copy allele with the ancestral variant (A), a 2-copy allele with the same variant at the SNP (AA), and a 2-copy allele with two different variants (AB). In mutational path **B**, a duplication event occurs, leading to a 2-copy allele. Then a mutation occurs in the 1-copy allele. The result is three types of allele present in the population: a 1-copy allele with the ancestral variant (A), a 2-copy allele with the same variant at the SNP (AA) and a 1-copy allele with a mutation (B).

These two scenarios can explain the diversity of DI-genotypes observed, and their frequencies given the genotype of the focal individual. When the alternative allele corresponds to the ancestral allele, we expect all the symmetrical DI-genotypes that we observe in the data. Of course, there are many other potential scenarios, such as when the polymorphic duplication is

due to a deletion of one copy of a fixed duplication. There may also be more complex demographic scenarios than a single population model, such as secondary contact between a lineage that has a fixed duplication, and has accumulated some divergence between the two copies (AABB), and a lineage that has kept the ancestral 1-copy allele (AA). This scenario would generate the polymorphisms we observe the most- AA, AAB, AABB. Further, we note that the duplication is not necessarily a tandem repeat, and there may be recombination between the two copies, generating recombined genotypes. However, these more complex scenarios mostly tend to remove intermediate DI-genotypes of the mutational path, and we believe that our results are well explained simply by widespread polymorphic duplication.

Duplications are not randomly distributed between chromosomes, but affect the functional categories of SNPs equally.

The density of 3X and 4X SNPs in 100kb windows varied throughout the genome (Fig. 6, Supplementary Figures 13-16), allowing us to identify windows of non-diploid SNPs corresponding to collapsed duplicated regions.

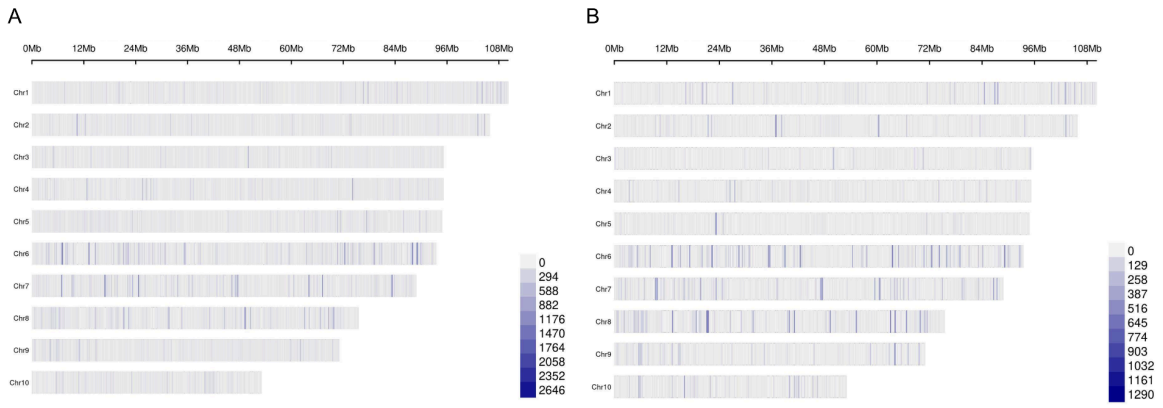

Figure 6. SNP-density plots for A) 3X SNPs and B) 4X SNPs in sample 6C9M. Density of SNPs per 100kb window shown for all 10 chromosomes. Position on the chromosome indicated on the x-axis.

Surprisingly, some chromosomes were found to have a higher density of non-diploid SNP clusters. Chromosomes 6, 7 and 8 appeared to have the highest density for all individuals, and this was verified by the average count for each chromosome (Fig. 7A). At this stage we do not have a clear explanation for this result, and further work will be needed to better understand the pattern. However, there are some different characteristics between the first five and last five chromosomes that may indicate potential drivers of this result. The first five chromosomes each have an evident putative centromere at their centre, revealed by a valley of nucleotide diversity (Supplementary Figure 17). Conversely, there are no valleys of nucleotide diversity in the central part of the last five chromosomes. Further, only chromosomes 8 and 10 have a reduction in

diversity towards their ends, perhaps because these are low-recombining regions as often found in telomeric regions (*O. edulis* does not have acrocentric chromosomes so this can not explain these regions).

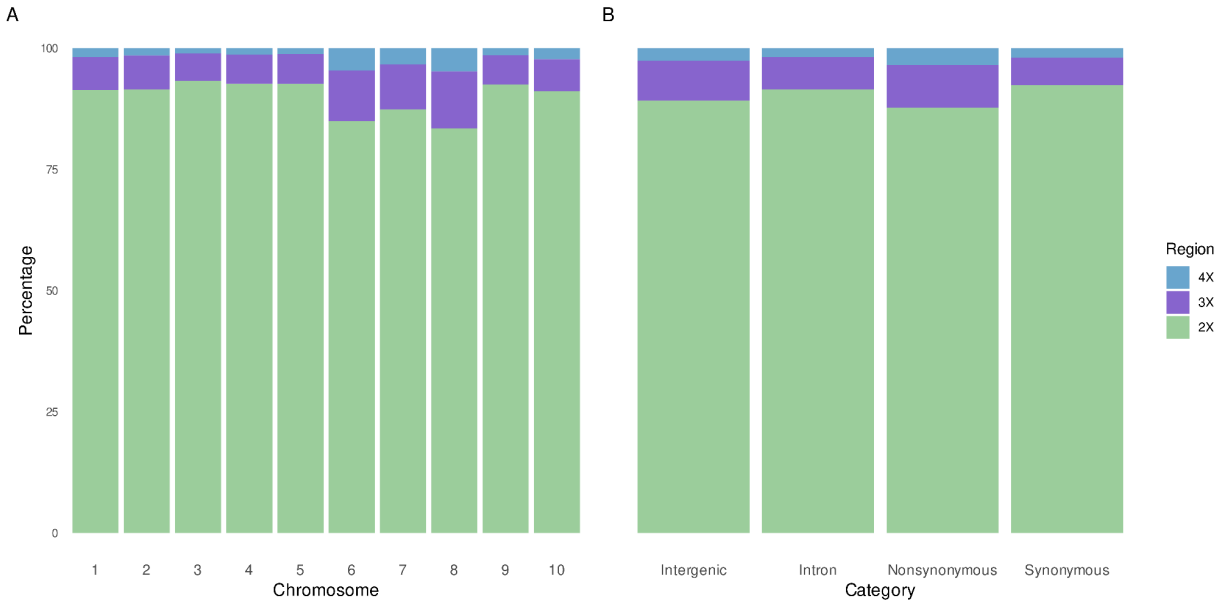

Figure 7. A Percentage of SNPs classified as 2X, 3X and 4X in all 10 chromosomes, averaged across all five individuals and **B** percentage of intergenic, intronic, nonsynonymous and synonymous variants classified as 2X, 3X and 4X SNPs, averaged across all five individuals.

Variation in chromosomal morphology has been noted in both *O. edulis* and its closely related sister species *O. angasi* (Thiriot-Quiévreux 1984 ; Li and Havenhand 1997), as well as in other ostreids (Leitão et al. 2002). In both species, the first five chromosomes are metacentric, which conforms to the position of putative centromeres that we identify using valleys of nucleotide diversity in these chromosomes. The last five shorter chromosomes consist of submetacentric and subtelocentric pairs. This conforms to the lack of valleys of nucleotide diversity in the centre of these chromosomes. Duplications are often produced by problems during recombination, or mistakes during cellular division (Ohta 2000, Zhang 2003), and these issues may be more likely to occur in chromosomes prone to cohesion defects and mispairing. Under this hypothesis, the higher density of polymorphic duplications in chromosomes 6, 7 and 8 would be caused by a higher rate of duplication mutation, owing to chromosome pairing and problems with recombination. It is not clear whether this explanation is plausible, but it would mean that duplication alleles would behave mostly neutrally and mutation rate defines their density, or that the genetic load of deleterious duplicated or deleted copy alleles would be greater on these three chromosomes. The proportion of 4X to 3X SNPs is higher in these 3 chromosomes (0.44, 0.36 and 0.42 for chromosomes 6, 7 and 8 respectively, compared with an average of 0.24 for the

other chromosomes), suggesting, as explained above, that the 2-copy alleles segregate at higher frequency in these chromosomes.

Another explanation could be linked to somatic aneuploidy, which is known to occur frequently in bivalves (Thiriot-Quievreux 2002), including *O. edulis* (Thiriot-Quiévreux 1986). Chromosome loss and gain are well known to be nonrandom and influenced by centromere size and cohesion defects (Sheltzer and Amon 2011). In marine bivalves, aneuploidy is biased towards chromosome loss, but affects only a small proportion of cells in a sample (Thiriot-Quiévreux, Pogson, and Zouros 1992), and affects specific chromosomes only (Bouilly et al. 2005; Thiriot-Quiévreux, Pogson, and Zouros 1992; Li & Havenhand 1997; Leitão et al. 2001; Khatir et al. 2022). In *O. edulis*, chromosome loss is always found at chromosomes 6 to 10 (Thiriot-Quiévreux 1986), in accordance with these chromosomes being more prone to instability during cell divisions. We do not think that somatic aneuploidy directly explains our results, as it only affects a fraction of cells in a sample, which would create dosages intermediate to our defined categories, while clusters are well-defined in our *k*-mer and reference based analyses. It may still explain the uncategorized SNPs though, which are found at significantly higher rates in these chromosomes (χ^2 test significant at $p < 2.2e-16$ for all individuals. See Supplementary Statistics). We suggest that somatic chromosome loss exposes recessive deleterious mutations to negative selection, and that extra copies producing compensation for this loss, including dosage compensation, can buffer the genetic load exposed by aneuploidy (Kondrashov 2012; Katju and Bergthorsson 2013; Basilicata and Valsecchi 2021). We investigated whether 2X, 3X, and 4X SNPs are found in similar proportions between functional categories. Intergenic and nonsynonymous sites had a slightly lower, although significant (χ^2 test significant at $p < 2.2e-16$ for all individuals. See Supplementary Statistics), proportion of 2X SNPs than intronic and synonymous sites (Fig. 7B). This result suggests that coding regions are similarly affected by duplications than non-coding regions, and that functional sites may even be slightly more affected than nonfunctional sites. This concurs with the observations made in *C. virginica* (Modak et al. 2021), and *M. galloprovincialis* (Gerdol et al. 2020).

Finally, similar patterns occurring in *M. gigas* further support our hypothesis concerning dosage compensation in chromosomes prone to aneuploidy. Qi, Li, and Zhang (2021) observe a higher rate of segmental duplications in chromosomes 1, 4, 8 and 10, and a higher rate of CNVRs in chromosomes 1, 3 and 4. Synteny analysis performed by Li et al. (2023) indicates that these chromosomes correspond to the last 5 chromosomes in *O. edulis* (see Supplementary Table 3). Bouilly et al. (2005) investigated somatic aneuploidy in *M. gigas*, and found certain chromosomes to be more prone to chromosome loss. The study predates the reference genome for *M. gigas* and it is not possible to definitively determine which chromosome numbers correspond to the chromosomes described by Qi, Li, and Zhang (2021). However, Bouilly et al. (2005) numbered the chromosomes by size, and therefore it is very likely that their chromosome

1 corresponds to chromosome 1 as numbered by Qi, Li, and Zhang (2021), as this chromosome is significantly larger than the other chromosomes. Chromosome 1, with a high rate of segmental duplications and CNVRs (Qi, Li, and Zhang 2021), was identified as experiencing the highest rates of somatic aneuploidy (Bouilly et al. 2005).

Implications for the population genetics of bivalves

Though the high proportion of non-diploid regions in *O. edulis*, demonstrated by both the *k*-mer analyses and the reference mapping approach in this study, is in-line with previous results in marine bivalves (e.g. Modak et al. 2021, Qi, Li, and Zhang 2021), the skewed distribution of individual allelic coverage fraction (~30% of SNPs between 0.2 and 0.4), is still surprising. This result raises two fundamental questions. Firstly, how can low or medium coverage data be analysed for the purposes of population structure, demographic reconstruction, and selection scan studies? Secondly, is there, as proposed regularly for decades, something potentially unique about the biology, ecology, or genetics of marine bivalves that can explain their unusual polymorphism, including rampant polymorphic duplications?

The availability of five high coverage genomes gave us the opportunity to take an unusual approach to the study of structural variation and its impact on SNP calling. Usually, only one individual used as a reference genome is deeply sequenced, while the resequenced genomes have insufficient coverage to undertake the dosage-based genotyping that we used in this study. Short indels or CNVs can be called with short-read data, but calling large duplications requires pair-mate, linked-read, or long-read sequencing, methods that are still rarely used in population genomics, especially in bivalves. Further, even with these methods, if the duplicated copy is too long for the long DNA molecules to pass through, and if the two copies are too similar to resolve duplication by assembly, a recent long duplication can easily be missed. With these issues in mind, dosage-based inference with high coverage provides valuable information. However, filtering of medium or low coverage datasets on the basis of coverage of a single individual alone can not provide the necessary information. The distributions of 2X, 3X, and 4X coverage, and 1/2, 1/3, and 1/4 allelic coverage fraction overlap too much with insufficient sequencing depth. In addition, the duplications causing the inflated coverage and skewed allelic coverage fractions occur in different regions in different individuals, making it difficult to filter specific SNPs. The analysis of the coverage of a number of individuals may help alleviate the issue. If there are enough sufficiently covered heterozygote and homozygote duplications in the sample then the excess coverage and skewed allelic coverage fraction can be detected on a sample-wide scale. This is the information that the programs ngsParalog (Linderoth 2018) or paramask (Tjeng et al. 2025) use to filter SNPs considered ‘deviant’ in multicopy regions. These methods also use heterozygote frequencies in the sample, assuming Hardy-Weinberg proportions for the former, and fitting an unknown inbreeding coefficient for the latter. These programs also rely on the expectation that deviant SNPs cluster together in collapsed duplicated regions of the genome.

The programs aim to mainly target 4X coverage and 1/4 allelic coverage fraction, with the assumption that the duplications are fixed. Therefore, these methods would need to be modified in order to better account for polymorphic duplications. These methods are likely to remove many problematic SNPs, and filtering with them is highly recommended in bivalves, and indeed any non-model species. However, they are unlikely to remove variants in rare duplicated alleles, or SNPs that do not depart much from Hardy-Weinberg proportions. In the case of *O. edulis*, however, we can advise the removal of the three most affected chromosomes (i.e. 6, 7, 8) for specific analyses such as population structure or demographic inference.

The explanation for why *O. edulis*, and perhaps bivalves in general, have such a high rate of polymorphic duplication can not be solved here. Pangenome studies will be needed to better understand the mechanistic constraints and evolutionary forces that shape these patterns. The putative whole genome duplication ancestral to Ostreidae (>50MYA), or ancestral to *Mytilus* mussels (Corrochano-Fraile et al. 2022) might have helped initiate the dynamic of gain and loss of gene copies, but cannot reasonably explain our result. Divergence between the two copies must be low, because the duplication mutation must be recent to be still polymorphic, or because the merging of duplicates in the reference genome would not occur with a strong divergence between copies. In fact, the dosage-based genotypes observed suggest that the reference genome most often has the duplication merged into a single copy, or has the 1-copy allele at the polymorphic duplication identified. This is expected given that most duplications were found to be polymorphic and should follow a classic allelic frequency spectrum (one given individual has the 1-copy allele at most of the polymorphic duplications). Follow-up studies should attempt to obtain the frequency of the 2-copy allele in different populations. This will allow testing for deviation from neutral expectations. We propose that the high burden of polymorphic duplications could be a simple consequence of the global high genetic diversity of marine bivalves with large population sizes. Indeed, a similar proportion of polymorphism in coding and non-coding regions is usually taken as evidence for neutral or nearly-neutral theory. However, adaptive or maladaptive forces may also be at play. The possible role of presence-absence and duplication variation in the evolutionary success, phenotypic variation, and local adaptation of marine bivalves has been emphasized by others (e.g. Gerdol et al. 2020; Modak et al. 2021). We suggest instead that our observations are in accordance with the high genetic load of these abundant and highly fertile organisms, which may be further enhanced by their sweepstakes reproduction (Hedgecock and Pudovkin 2011; Harrang et al. 2013; Plough 2016). Just as a high rate of segregating deleterious non-synonymous mutations have been suspected (Harrang et al. 2013), the same explanation could be proposed for mildly deleterious copy-number variants. The population genetics of accurately called duplication variants is needed to provide an answer to this question.

LITERATURE CITED

Ament-Velásquez SL, Figuet E, Ballenghien M, Zattara EE, Norenburg JL, Fernández-Álvarez FA, Bierne J, Bierne N, Galtier N. 2016. Population genomics of sexual and asexual lineages in fissiparous ribbon worms (Lineus, Nemertea): hybridization, polyploidy and the Meselson effect. *Molecular Ecology* 25:3356–3369. DOI: 10.1111/mec.13717.

Basilicata MF, Valsecchi CIK. 2021. The good, the bad, and the ugly: Evolutionary and pathological aspects of gene dosage alterations. *PLOS Genetics* 17:e1009906. DOI: 10.1371/journal.pgen.1009906.

Bazin E, Glémin S, Galtier N. 2006. Population size does not influence mitochondrial genetic diversity in animals. *Science* 312:570–572. DOI: 10.1126/science.1122033.

Bierne N. 2022. Pressing NGS data through the mill of *k*-mer spectra and allelic coverage ratios in order to scan reproductive modes in non-model species. *Peer Community in Evolutionary Biology* 1:100142. DOI: 10.24072/pci.evolbiol.100142.

Bierne N, Launey S, Naciri-Graven Y, Bonhomme F. 1998. Early effect of inbreeding as revealed by microsatellite analyses on *Ostrea edulis* larvae. *Genetics* 148:1893–1906. DOI: 10.1093/genetics/148.4.1893.

Bouilly K, Leitão A, Chaves R, Guedes-Pinto H, Boudry P, Lapègue S. 2005. Endonuclease banding reveals that atrazine-induced aneuploidy resembles spontaneous chromosome loss in *Crassostrea gigas*. *Genome* 48:177–180. DOI: 10.1139/g04-087.

Boutet I, Alves Monteiro HJ, Baudry L, Takeuchi T, Bonnivard E, Billoud B, Farhat S, Gonzales‐Araya R, Salaun B, Andersen AC, Toullec J, Lallier FH, Flot J, Guiglielmoni N, Guo X, Li C, Allam B, Pales‐Espinosa E, Hemmer‐Hansen J, Moreau P, Marbouty M, Koszul R, Tanguy A. 2022. Chromosomal assembly of the flat oyster (*Ostrea edulis* L.) genome as a new genetic resource for aquaculture. *Evolutionary Applications* 15:1730–1748. DOI: 10.1111/eva.13462.

Chen S, Zhou Y, Chen Y, Gu J. 2018. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*

(Oxford, England) 34:i884–i890. DOI: 10.1093/bioinformatics/bty560.

Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, Land SJ, Ruden DM, Lu X. 2012. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly* 6:80-92. DOI: 10.4161/fly.19695.

Corrochano-Fraile A, Davie A, Carboni S, Bekaert M. 2022. Evidence of multiple genome duplication events in Mytilus evolution. *BMC Genomics* 23:340. DOI: 10.1186/s12864-022-08575-9.

Dallaire X, Bouchard R, Hénault P, Ulmo-Diaz G, Normandeau E, Mérot C, Bernatchez L, Moore J-S. 2023. Widespread deviant patterns of heterozygosity in whole-genome sequencing due to autopolyploidy, repeated elements, and duplication. *Genome Biology and Evolution* 15:evad229 DOI: 10.1101/2023.07.27.550877.

Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, McVean G, Durbin R, 1000 Genomes Project Analysis Group. 2011. The variant call format and VCFtools. *Bioinformatics (Oxford, England)* 27:2156–2158. DOI: 10.1093/bioinformatics/btr330.

Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, Whitwham A, Keane T, McCarthy SA, Davies RM, Li H. 2021. Twelve years of SAMtools and BCFtools. *GigaScience* 10:giab008. DOI: 10.1093/gigascience/giab008.

David P. 1998. Heterozygosity–fitness correlations: new perspectives on old problems. *Heredity* 80:531–537. DOI: 10.1046/j.1365-2540.1998.00393.x.

David P, Berthou P, Noel P, Jarne P. 1997. Patchy recruitment patterns in marine invertebrates: a spatial test of the density-dependent hypothesis in the bivalve Spisula ovalis. *Oecologia* 111:331–340. DOI: 10.1007/s004420050243.

Eldon B, Riquet F, Yearsley J, Jollivet D, Broquet T. 2016. Current hypotheses to explain genetic chaos under the sea. *Current Zoology* 62:551–566. DOI: 10.1093/cz/zow094.

Foltz DW. 1986. Null alleles as a possible cause of heterozygote deficiencies in the oyster *Crassostrea*

virginica and other bivalves. *Evolution* 40:869–870. DOI: 10.1111/j.1558-5646.1986.tb00549.x.

Gerdol M, Moreira R, Cruz F, Gómez-Garrido J, Vlasova A, Rosani U, Venier P, Naranjo-Ortiz MA, Murgarella M, Greco S, Balseiro P, Corvelo A, Frias L, Gut M, Gabaldón T, Pallavicini A, Canchaya C, Novoa B, Alioto TS, Posada D, Figueras A. 2020. Massive gene presence-absence variation shapes an open pan-genome in the Mediterranean mussel. *Genome Biology* 21:275. DOI: 10.1186/s13059-020-02180-3.

Gundappa MK, Peñaloza C, Regan T, Boutet I, Tanguy A, Houston RD, Bean TP, Macqueen DJ. 2022. Chromosome‐level reference genome for European flat oyster (*Ostrea edulis* L.). *Evolutionary Applications* 15:1713–1729. DOI: 10.1111/eva.13460.

Harrang E, Lapègue S, Morga B, Bierne N. 2013. A high load of non-neutral amino-acid polymorphisms explains high protein diversity despite moderate effective population size in a marine bivalve with sweepstakes reproduction. *G3 Genes|Genomes|Genetics* 3:333–341. DOI: 10.1534/g3.112.005181.

Hedgecock D, Launey S, Pudovkin AI, Naciri Y, Lapègue S, Bonhomme F. 2007. Small effective number of parents (Nb) inferred for a naturally spawned cohort of juvenile European flat oysters *Ostrea edulis*. *Marine Biology* 150:1173–1182. DOI: 10.1007/s00227-006-0441-y.

Hedgecock D, Pudovkin AI. 2011. Sweepstakes reproductive success in highly fecund marine fish and shellfish: A Review and Commentary. *Bulletin of Marine Science* 87:971–1002. DOI: 10.5343/bms.2010.1051.

Hemstrom W, Grummer JA, Luikart G, Christie MR. 2024. Next-generation data filtering in the genomics era. *Nature Reviews Genetics*:1–18. DOI: 10.1038/s41576-024-00738-6.

Jaron KS, Hodson CN, Ellers J, Baird SJE, Ross L. 2022. Genomic evidence of paternal genome elimination in the globular springtail Allacma fusca. *Genetics* 222:iyac117. DOI: 10.1093/genetics/iyac117.

Jenike KM, Campos-Domínguez L, Boddé M, Cerca J, Hodson CN, Schatz MC, Jaron KS. 2025. k-mer approaches for biodiversity genomics. *Genome Research*. DOI: 10.1101/gr.279452.124.

Katju V, Bergthorsson U. 2013. Copy-number changes in evolution: rates, fitness effects and adaptive significance. *Frontiers in Genetics* 4. DOI: 10.3389/fgene.2013.00273.

Khatir Z, Hizam Z, Lyons B, Leitão A. 2022. Aneuploidy in the pearl oyster *Pinctada radiata* (Leach, 1814): evidence for nonrandom chromosome loss and gain in marine bivalves. *Malacologia* 65(1-2), 71-77. DOI: 10.4002/040.065.0105.

Koehn RK. 1983. 9 - Biochemical Genetics and Adaptation in Molluscs. In: Hochachka PW ed. *The Mollusca*. San Diego: Academic Press, 305–330. DOI: 10.1016/B978-0-12-751402-4.50016-9.

Kondrashov FA. 2012. Gene duplication as a mechanism of genomic adaptation to a changing environment. *Proceedings of the Royal Society B: Biological Sciences* 279:5048–5057. DOI: 10.1098/rspb.2012.1108.

Lallias D, Taris N, Boudry P, Bonhomme F, Lapègue S. 2010. Variance in the reproductive success of flat oyster *Ostrea edulis* L. assessed by parentage analyses in natural and experimental conditions. *Genetics Research* 92:175–187. DOI: 10.1017/S0016672310000248.

Launey S, Hedgecock D. 2001. High Genetic Load in the Pacific Oyster *Crassostrea gigas*. *Genetics* 159:255–265. DOI: 10.1093/genetics/159.1.255.

Leitão A, Boudry P, Thiriot-Quiévreux C. 2001. Evidence of differential chromosome loss in aneuploid karyotypes of the Pacific oyster, *Crassostrea gigas*. *Genome*, 44: 735-737.

Leitao A, Chaves R, Santos S, Boudry P, Guedes Pinto H, Thiriot Quievreux C, Santos S, Boudry P, Guedes Pinto H, Thiriot Quievreux C. 2002. Cytogenetic study of *Ostrea conchaphila* (Mollusca : Bivalvia) and comparative karyological analysis within Ostreinae. *Journal of Shellfish Research* 21:685–690.

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

Li H, Durbin R. 2010. Fast and accurate long-read alignment with Burrows-Wheeler transform.

Bioinformatics (Oxford, England) 26:589–595. DOI: 10.1093/bioinformatics/btp698.

Li XX, Havenhand JN. 1997. Karyotype, nucleolus organiser regions and constitutive heterochromatin in *Ostrea angasi* (Molluscae: Bivalvia): evidence of taxonomic relationships within the Ostreidae. *Marine Biology* 127:443–448. DOI: 10.1007/s002270050031.

Linderoth T. 2018. Identifying population histories, adaptive genes and genetic duplication from population-scale next generation sequencing. Berkeley: University of California.

Lynch M, Walsh B. 2007. The origins of genome architecture. Sinauer Associated, Sunderland, Massachusetts.

Mitton JB. 2000. Selection in natural populations. Oxford University Press.

Modak TH, Literman R, Puritz JB, Johnson KM, Roberts EM, Proestou D, Guo X, Gomez-Chiarri M, Schwartz RS. 2021. Extensive genome-wide duplications in the eastern oyster (*Crassostrea virginica*). *Philosophical Transactions of the Royal Society B: Biological Sciences* 376:20200164. DOI: 10.1098/rstb.2020.0164.

Mun S, Kim Y-J, Markkandan K, Shin W, Oh S, Woo J, Yoo J, An H, Han K. 2017. The whole-genome and transcriptome of the Manila clam (*Ruditapes philippinarum*). *Genome Biology and Evolution* 9:1487–1498. DOI: 10.1093/gbe/evx096.

Murgarella M, Puiu D, Novoa B, Figueras A, Posada D, Canchaya C. 2016. A First insight into the genome of the filter-feeder mussel *Mytilus galloprovincialis*. *PLOS ONE* 11:e0151561. DOI: 10.1371/journal.pone.0151561.

Ohta T. 2000. Evolution of gene families. *Gene* 259:45–52. DOI: 10.1016/S0378-1119(00)00428-5.

O'Leary SJ, Puritz JB, Willis SC, Hollenbeck CM, Portnoy DS. 2018. These aren't the loci you're looking for: Principles of effective SNP filtering for molecular ecologists. *Molecular Ecology* 27:3193–3206. DOI: 10.1111/mec.14792.

Piccoli GAR. 2024. *Ostrea chilensis* genomic analysis and population diversity: widespread-hemizygosity genomics. Victoria University of Wellington.

Plough LV. 2016. Genetic load in marine animals: a review. *Current Zoology* 62:567–579. DOI:

10.1093/cz/zow096.

Qi H, Li L, Zhang G. 2021. Construction of a chromosome‐level genome and variation map for the Pacific oyster *Crassostrea gigas*. *Molecular Ecology Resources* 21:1670–1685. DOI: 10.1111/1755-0998.13368.

R Core Team. 2023. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/

Ranallo-Benavidez TR, Jaron KS, Schatz MC. 2020. GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. *Nature Communications* 11:1432. DOI: 10.1038/s41467-020-14998-3.

Romiguier J, Gayral P, Ballenghien M, Bernard A, Cahais V, Chenuil A, Chiari Y, Dernat R, Duret L, Faivre N, Loire E, Lourenco JM, Nabholz B, Roux C, Tsagkogeorga G, Weber A a.-T., Weinert LA, Belkhir K, Bierne N, Glémin S, Galtier N. 2014. Comparative population genomics in animals uncovers the determinants of genetic diversity. *Nature* 515:261–263. DOI: 10.1038/nature13685.

Šatović Vukšić E, Plohl M. 2021. Exploring satellite DNAs: specificities of bivalve mollusks genomes. *Progress in molecular and subcellular biology* 60:57–83. DOI: 10.1007/978-3-030-74889-0_3.

Sauvage C, Bierne N, Lapègue S, Boudry P. 2007. Single nucleotide polymorphisms and their relationship to codon usage bias in the Pacific oyster *Crassostrea gigas*. *Gene* 406:13–22. DOI: 10.1016/j.gene.2007.05.011.

Sheltzer JM, Amon A. 2011. The aneuploidy paradox: costs and benefits of an incorrect karyotype. *Trends in Genetics* 27:446–453. DOI: 10.1016/j.tig.2011.07.003.

Sun J, Zhang Y, Xu T, Zhang Y, Mu H, Zhang Y, Lan Y, Fields CJ, Hui JHL, Zhang W, Li R, Nong W, Cheung FKM, Qiu J-W, Qian P-Y. 2017. Adaptation to deep-sea chemosynthetic environments as revealed by mussel genomes. *Nature Ecology & Evolution* 1:1–7. DOI: 10.1038/s41559-017-0121.

Takeuchi T. 2017. Molluscan genomics: implications for biology and aquaculture. *Current Molecular*

Biology Reports 3:297–305. DOI: 10.1007/s40610-017-0077-3.

Takeuchi T, Kawashima T, Koyanagi R, Gyoja F, Tanaka M, Ikuta T, Shoguchi E, Fujiwara M, Shinzato C, Hisata K, Fujie M, Usami T, Nagai K, Maeyama K, Okamoto K, Aoki H, Ishikawa T, Masaoka T, Fujiwara A, Endo K, Endo H, Nagasawa H, Kinoshita S, Asakawa S, Watabe S, Satoh N. 2012. Draft genome of the pearl oyster *Pinctada fucat*a: a platform for understanding bivalve biology. *DNA Research* 19:117–130. DOI: 10.1093/dnares/dss005.

Takeuchi T, Suzuki Y, Watabe S, Nagai K, Masaoka T, Fujie M, Kawamitsu M, Satoh N, Myers EW. 2022. A high-quality, haplotype-phased genome reconstruction reveals unexpected haplotype diversity in a pearl oyster. *DNA Research* 29:dsac035. DOI: 10.1093/dnares/dsac035.

Tarasov A, Vilella AJ, Cuppen E, Nijman IJ, Prins P. 2015. Sambamba: fast processing of NGS alignment formats. *Bioinformatics (Oxford, England)* 31:2032–2034. DOI: 10.1093/bioinformatics/btv098.

Thiriot-Quiévreux C. 1984. (Bivalvia). *Cahiers de Biologie Marine* 25:407-418.

Thiriot-Quiévreux C. 1986. Etude de l'aneuploïdie dans différents naissains d'Ostreidae (Bivalvia). *Genetica* 70:225–231. DOI: 10.1007/BF00122190.

Thiriot-Quievreux C. 2002. Review of the literature on bivalve cytogenetics in the last ten years. *Cahiers de Biologie Marine* 43:17-26.

Thiriot-Quiévreux C, Pogson GH, Zouros E. 1992. Genetics of growth rate variation in bivalves: aneuploidy and heterozygosity effects in a *Crassostrea gigas* family. *Genome* 35:39–45. DOI: 10.1139/g92-007.

Tjeng B, Arimond M, Grindeland HB, Libera AD, Fulgione A. 2024. ParaMask, a new method to identify multicopy genomic regions, corrects major biases in whole-genome sequencing data. DOI: 10.21203/rs.3.rs-4660628/v1.

Uliano-Silva M, Dondero F, Dan Otto T, Costa I, Lima NCB, Americo JA, Mazzoni CJ, Prosdocimi F, Rebelo M de F. 2018. A hybrid-hierarchical genome assembly strategy to sequence the invasive golden mussel, *Limnoperna fortunei*. *GigaScience* 7:gix128. DOI: 10.1093/gigascience/gix128.

Wang S, Zhang J, Jiao W, Li J, Xun X, Sun Y, Guo X, Huan P, Dong B, Zhang L, Hu X, Sun X, Wang J,

Zhao C, Wang Y, Wang D, Huang X, Wang R, Lv J, Li Y, Zhang Z, Liu B, Lu W, Hui Y, Liang J, Zhou Z, Hou R, Li X, Liu Y, Li H, Ning X, Lin Y, Zhao L, Xing Q, Dou J, Li Y, Mao J, Guo H, Dou H, Li T, Mu C, Jiang W, Fu Q, Fu X, Miao Y, Liu J, Yu Q, Li R, Liao H, Li X, Kong Y, Jiang Z, Chourrout D, Li R, Bao Z. 2017. Scallop genome provides insights into evolution of bilaterian karyotype and development. *Nature Ecology & Evolution* 1:1–12. DOI: 10.1038/s41559-017-0120.

Ward RD, Woodwark M; Skibinski DOF. 1994. A comparison of genetic diversity levels in marine, freshwater and anadromous fish. *Journal of Fisheries Biology* 44:213-232. DOI: 10.1111/j.1095-8649.1994.tb01200.x

Wickham H. 2016. ggplot2: elegant graphics for data analysis. Springer-Verlag New York. ISBN 978-3-319-24277-4. https://ggplot2.tidyverse.org.

Yin L, Zhang H, Tang Z, Xu J, Yin D, Zhang Z, Yuan X, Zhu M, Zhao S, Li X, Liu X. 2021. rMVP: A Memory-Efficient, Visualization-Enhanced, and Parallel-Accelerated Tool for Genome-Wide Association Study. *Genomics, Proteomics & Bioinformatics* 19:619–628. DOI: 10.1016/j.gpb.2020.10.007.

Zhang J. 2003. Evolution by gene duplication: an update. *Trends in Ecology & Evolution* 18:292–298. DOI: 10.1016/S0169-5347(03)00033-8.

Zhang G, Fang X, Guo X, Li L, Luo R, Xu F, Yang P, Zhang L, Wang X, Qi H, Xiong Z, Que H, Xie Y, Holland PWH, Paps J, Zhu Y, Wu F, Chen Y, Wang J, Peng C, Meng J, Yang L, Liu J, Wen B, Zhang N, Huang Z, Zhu Q, Feng Y, Mount A, Hedgecock D, Xu Z, Liu Y, Domazet-Lošo T, Du Y, Sun X, Zhang S, Liu B, Cheng P, Jiang X, Li J, Fan D, Wang W, Fu W, Wang T, Wang B, Zhang J, Peng Z, Li Y, Li N, Wang J, Chen M, He Y, Tan F, Song X, Zheng Q, Huang R, Yang H, Du X, Chen L, Yang M, Gaffney PM, Wang S, Luo L, She Z, Ming Y, Huang W, Zhang S, Huang B, Zhang Y, Qu T, Ni P, Miao G, Wang J, Wang Q, Steinberg CEW, Wang H, Li N, Qian L, Zhang G, Li Y, Yang H, Liu X, Wang J, Yin Y, Wang J. 2012. The oyster genome reveals stress adaptation and complexity of shell formation. *Nature* 490:49–54. DOI: 10.1038/nature11413.

Zouros E and Foltz DW. 1984. Possible explanations of heterozygote deficiency in bivalve molluscs. *Malacologia* 25:583-591.