ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/317819489

New insights into global biogeography, population structure and natural selection from the genome of the epipelagic copepod Oithona

Article *in* Molecular Ecology · June 2017

DOI: 10.1111/mec.14214

CITATIONS
5

READS
296

12 authors, including:

Mohammed-Amin Madoui
Genoscope - Centre National de Séquençage
30 PUBLICATIONS **543** CITATIONS

SEE PROFILE

Julie Poulain
Atomic Energy and Alternative Energies Commission
352 PUBLICATIONS **17,316** CITATIONS

SEE PROFILE

Kevin Sugier
Genoscope - Centre National de Séquençage
2 PUBLICATIONS **5** CITATIONS

SEE PROFILE

Benjamin Noel
CEA-Institut de Génomique
67 PUBLICATIONS **5,688** CITATIONS

SEE PROFILE

Some of the authors of this publication are also working on these related projects:

Project Caracterization of new viruses from Arthropods View project

Project TARA Expeditions : http://oceans.taraexpeditions.org/en/m/science/goals/ View project

All content following this page was uploaded by Mohammed-Amin Madoui on 26 July 2017.

The user has requested enhancement of the downloaded file.

Received: 22 January 2017 | Revised: 22 March 2017 | Accepted: 24 May 2017

DOI: 10.1111/mec.14214

ORIGINAL ARTICLE

WILEY MOLECULAR ECOLOGY

New insights into global biogeography, population structure and natural selection from the genome of the epipelagic copepod *Oithona*

Mohammed-Amin Madoui[1,2,3] | Julie Poulain[1] | Kevin Sugier[1,2,3] | Marc Wessner[1] | Benjamin Noel[1] | Leo Berline[4] | Karine Labadie[1] | Astrid Cornils[5] | Leocadio Blanco-Bercial[6] | Lars Stemmann[7] | Jean-Louis Jamet[8] | Patrick Wincker[1,2,3]

[1]Commissariat à l'Energie Atomique (CEA), Institut de Biologie François Jacob, Genoscope, Evry, France

[2]Centre National de la Recherche Scientifique, UMR 8030 Université d'Evry val d'Essonne, Evry, France

[3]Université d'Evry Val D'Essonne, Evry, France

[4]CNRS/INSU/IRD, Mediterranean Institute of Oceanography (MIO), Aix-Marseille Université, Marseille, France

[5]Alfred-Wegener-Institut Helmholtz-Zentrum für Polar- und Meeresforschung, Polar Biological Oceanography, Bremerhaven, Germany

[6]Bermuda Institute of Ocean Sciences, St. George's, Bermuda

[7]INSU-CNRS, Laboratoire D'Océanographie de Villefranche, UPMC Univ Paris 06, Sorbonne Universités, Villefranche-Sur-Mer, France

[8]Laboratoire PROTEE-EBMA E.A. 3819, Université de Toulon, La Garde Cedex, France

Correspondence
Mohammed-Amin Madoui, Commissariat à l'Energie Atomique (CEA), Institut de Biologie François Jacob, Genoscope, Evry, France.
Email: amadoui@genoscope.cns.fr

Funding information
Centre National de la Recherche Scientifique; European Molecular Biology Laboratory; Génoscope/Commissariat à l'Energie Atomique; the French Government "Investissements d'Avenir", Grant/Award Number: ANR-11-BTBR-0008; FRANCE GENOMIQUE, Grant/Award Number: ANR-10-INBS-09-08; Agnès b.; Veolia Environment Foundation; Region Bretagne; World Courier; Illumina; Cap L'Orient; Eléctricité de France (EDF) Foundation EDF Diversiterre; Fondation pour la Recherche sur la Biodiversité; Prince Albert II de Monaco Foundation; Etienne Bourgois

Abstract

In the epipelagic ocean, the genus *Oithona* is considered as one of the most abundant and widespread copepods and plays an important role in the trophic food web. Despite its ecological importance, little is known about *Oithona* and cyclopoid copepods genomics. Therefore, we sequenced, assembled and annotated the genome of *Oithona nana*. The comparative genomic analysis integrating available copepod genomes highlighted the expansions of genes related to stress response, cell differentiation and development, including genes coding Lin12-Notch-repeat (LNR) domain proteins. The *Oithona* biogeography based on 28S sequences and metagenomic reads from the *Tara* Oceans expedition showed the presence of *O. nana* mostly in the Mediterranean Sea (MS) and confirmed the amphitropical distribution of *Oithona similis*. The population genomics analyses of *O. nana* in the Northern MS, integrating the *Tara* Oceans metagenomic data and the *O. nana* genome, led to the identification of genetic structure between populations from the MS basins. Furthermore, 20 loci were found to be under positive selection including four missense and eight synonymous variants, harbouring soft or hard selective sweep patterns. One of the missense variants was localized in the LNR domain of the coding region of a male-specific gene. The variation in the B-allele frequency with respect to the MS circulation pattern showed the presence of genomic clines between *O. nana* and another undefined *Oithona* species possibly imported through Atlantic waters. This study provides new approaches and results in zooplankton population genomics through the integration of metagenomic and oceanographic data.

© 2017 John Wiley & Sons Ltd

KEYWORDS
genome, Mediterranean Sea, phylogeography, selection

1 | INTRODUCTION

Oceanic global changes are thought to have a great impact on zooplankton communities, notably through long timescale observations that have shown significant changes in copepod populations (Beaugrand, Reid, Ibanez, Lindley, & Edwards, 2002). The study of pelagic copepod populations at the molecular level helps to identify environmental factors that drive the appearance and fixation of adaptive traits. Current approaches applied to pelagic copepods typically use ribosomal genes, mitochondrial cytochrome oxidase subunit I and II genes and microsatellites markers to identify species, genotypes and haplotypes (e.g., Blanco-Bercial, Álvarez-Marqués, & Bucklin, 2011; Blanco-Bercial, Cornils, Copley, & Bucklin, 2014; Cornils, Wend-Heckmann, & Held, 2017; Goetze, Andrews, Peijnenburg, Portner, & Norton, 2015; Hirai, Kuriyama, Ichikawa, Hidaka, & Tsuda, 2015). With appropriate sampling, the calculation of within- and between-population genetic distances can then be used to infer copepod population structure and connectivity (Kozol, Blanco-Bercial, & Bucklin, 2012). These approaches applied on the mesopelagic copepod *Haloptilus longicornis* at a large spatial scale demonstrated structure stability among the North and South Atlantic gyres and demonstrated the structure stability across 2 years (Goetze et al., 2015). Advanced high-throughput sequencing technologies like RAD-seq now allow the identification of hundreds to thousands of polymorphic loci without a reference genome (Blanco-Bercial & Bucklin, 2016). Recently applied to the calanoid copepod *Centropages typicus* in North Atlantic Ocean (NAO), this strategy permitted the identification of loci under selection and significant structure in the populations across the NAO. These results support the idea of a high evolutionary and adaptive potential of copepods in the open ocean (Peijnenburg & Goetze, 2013) and slightly modify the previous idea of a weak genetic structure in populations with a high migration rate (Helaouët & Beaugrand, 2009). Although this recent approach provides a new view in copepod population genetics, detected loci under selection could not be directly linked to biological functions due to the lack of a reference genome.

Genome-wide approaches have been applied mostly on humans, plants, animals or microorganisms of agronomic or public health interest for which reference genomes were available. These approaches provide a comprehensive view of genomic regions targeted by selection and are more informative than RAD-seq or other capture-based technologies to accurately identify selective sweeps and distinguish causal mutations from genetic draft (Andrews, Good, Miller, Luikart, & Hohenlohe, 2016). Although next-generation sequencing has reached its golden era, only a few copepod genomes (*Eurytemora affinis*, *Tigriopus californicus*, *Caligus rogercresseyi* and *Lepeophtherius salmonis*) have been sequenced and are available, but no genome is available for cyclopoid copepods. This approach remains costly and the investment depends greatly on the genome size of the organisms. Among the pelagic copepods, calanoids possess among the largest genomes, often exceeding several billions of bases (McLaren, Sevigny, & Corkett, 1988; McLaren, Sévigny, & Frost, 1989; Rasch & Wyngaard, 2006; Wyngaard & Rasch, 2000), which does not make them a practical model for a genome-wide approach, despite their ecological importance. Cyclopoids, in contrast, are known to have much smaller genomes (Wyngaard, McLaren, White, & Sévigny, 1995; Wyngaard & Rasch, 2000), although some lineages show complex patterns of genomic variation linked to life stages, likely due to chromatin diminution (Wyngaard, Rasch, & Connelly, 2011).

The cyclopoid copepod genus *Oithona* is considered very abundant and widespread in the world ocean's surface (Gallienne & Robins, 2001). In the past, its abundance was underestimated due to their small size (Clark, Frid, & Batten, 2000; Williams & Muxagata, 2006). This genus plays, however, an important ecological role as grazers and secondary producers in the marine trophic food chain (Turner, 2004), sustaining the growth of commercially important larval fishes such as anchovy (Viñas & Ramirez, 1996) and Argentine hake (Viñas & Santos, 2000).

Within this genus, three species are described as particularly widespread based on morphological identifications (Nishida, 1985): *Oithona similis* Claus, 1866, *Oithona atlantica* Farran, 1908 and *Oithona nana* Giesbrecht, 1892. The present knowledge on the biogeography of *Oithona* species has been mainly conducted through morphological identification of specimens collected by independent and geographically restricted studies. *Oithona similis* has been identified from all oceans and climate zones (Razouls, de Bovée, Kouwenberg, & Desreumaux, 2016) and prefers temperatures below 20°C (Castellani, Licandro, Fileman, di Capua, & Grazia Mazzocchi, 2015), which strengthens the speculation that its occurrence in tropical regions may be based on misidentifications (Nishida, 1985). A recent finding has also shown that *O. similis* is a complex of independent lineages with distinct biogeographies linked to climate zones, and not a single cosmopolitan species (Cornils et al., 2017). *Oithona atlantica* is also widely distributed and occurs in the temperate and polar oceanic regions of the Atlantic and Pacific Ocean (Cepeda, Blanco-Bercial, Bucklin, Beron, & Vinas, 2012; Nishida, 1985). While the former two species are abundant in temperate to polar waters, *O. nana* has also been found extensively in tropical and subtropical zones, mostly in coastal regions (Temperoni, Viñas, Diovisalvi, & Negri, 2011; Williams & Muxagata, 2006). The small size of *Oithona* (<1 mm) and its subtle morphological species-specific traits that can only be observed by microscopy (Nishida, Omori, & Tanaka, 1977) represent a serious difficulty for studies involving a large number of samples. Ribosomal 28S and mitochondrial genes have been successfully used to characterize *Oithona* species (Cepeda et al., 2012; Cornils et al., 2017; Ueda, Yamaguchi, Saitoh, Orui Sakaguchi, &

Tachihara, 2011) but also to identify invasive species (Cornils & Wend-Heckmann, 2015). This combination of morphological identification and molecular analyses provides robust molecular resources that can be used as a reference for species identification using molecular analysis-only approaches including genome-wide approaches (i.e., metagenomics).

In this study, a global phylogeography for *Oithona* species is proposed based on published records and new metagenomic data produced under the *Tara* Oceans consortium (de Vargas et al., 2015; Vannier et al., 2016). Focusing on *O. nana*, its ~85 Mb genome is presented and compared to other available copepod genomes. Based on its genomic variation landscape throughout the spatial range, the population structure within the Mediterranean Sea is determined, and multiple loci under selection are identified by developing a metagenomic-adapted framework. Finally, the information from the *O. nana* genome and the metagenomic and oceanographic data from the Mediterranean Sea are combined to study the relationships between variations in the allele frequency and the circulation patterns.

2 | MATERIALS AND METHODS

2.1 | Genome sequencing

Oithona nana individuals were sampled in the small harbour of Toulon, France, in June 2014 with a 90-μm-mesh net and stored in 100% ethanol. For genome sequencing, 2,000 adult individuals were isolated under the stereomicroscope and washed individually in a physiological saline solution. The individuals were transferred by pools of 40 individuals into 1.5-ml tubes (50 tubes in total) containing the alkaline lysis buffer adjusted to pH = 8 (for 10 ml, 9.5 ml H_2O, 25 μl NaOH 10 M, 4 μl EDTA 0.5 M, 226.5 μl HCl 100% at 1/10, 244.5 H_2O) and ground with a tissue grinder and cooled with liquid nitrogen. The tubes were then incubated for 35 min at 95°C and cooled on ice for 5 min. A total of 20 μl of neutralizing solution (Tris-HCl 40 mM, pH 5.0) was added to the tubes. The tubes were then vortexed, centrifuged on a mini centrifuge at 6,400 rpm and kept for 10 min on ice. DNA was purified using Agencourt® AMPure beads by adding 1.5 volumes of Ampure and vortexing. The tubes were left for 5 min at room temperature, then transferred to a magnetic holder; the supernatant was removed and two washes with EtOH 70% were carried out. The ethanol wash was left 30 s before removal. After the ethanol steps, tubes were left open to dry for 10 min before elution in 25 μl of DNA-free ultrapure water. Water and beads were mixed and left for 3 min at room temperature and for 2 min on the magnetic holder. After a total of 5 min, the genomic DNA was retrieved with the supernatant and quantified on a Qubit 2.0 (Invitrogen).

To prepare the overlapping paired-end library, 30 ng of genomic DNA from a single vial (40 individuals) was sonicated to a 100–800 base pairs (bp) size range using an E210 Covaris instrument (Covaris, Inc., USA). Fragments were end-repaired and then 3′-adenylated. Illumina adapters were added by NEBNext Sample Reagent Set (New England Biolabs). Ligation products were purified with 1:1 Ampure XP beads (Beckmann Coulter), and DNA fragments (>200 bp) were PCR-amplified using Illumina adapter-specific primers and Platinum Pfx DNA polymerase (Invitrogen). Amplified library fragments were size-selected to around 300 bp on a 3% agarose gel. After library profile analysis using an Agilent 2100 Bioanalyzer (Agilent Technologies, USA) and qPCR quantification (MxPro, Agilent Technologies, USA), the library was sequenced using 101-bp paired-end read chemistry in a single flow cell on the Illumina MiSeq (Illumina, USA). Raw reads were trimmed for adapters and low-quality bases (Phred value under 20); only trimmed reads longer than 30 bp were kept, producing a final read set of 24.7 × 10^6 paired-end reads corresponding to 2.4 × 10^9 base pairs.

The DNA from the remaining 49 tubes were pooled and used to build three long insert libraries. The three mate pair libraries were prepared following the Nextera protocol (Nextera Mate Pair sample preparation kit, Illumina). Briefly, genomic DNA was simultaneously enzymatically fragmented and tagged with a biotinylated adaptor. Fragments were size-selected (3–5, 5–8 and 8–11 kb) through regular gel electrophoresis and circularized overnight with a ligase. Linear, noncircularized DNA fragments were digested, and circularized DNA was fragmented to 300–1,000 bp size range using the Covaris E210. Biotinylated DNA was immobilized on streptavidin beads, end-repaired, 3′-adenylated, and Illumina adapters were added. DNA fragments were PCR-amplified using Illumina adapter-specific primers and then purified with Ampure XP. Libraries were quantified by qPCR, and library profiles were evaluated using an Agilent 2100 Bioanalyzer (Agilent Technologies, USA). Each library was sequenced using 150-bp paired-end read chemistry on a single flow cell on the Illumina MiSeq. Read sets were trimmed for cleaning as previously described and the sequencing produced finally 4.8 × 10^6, 3.5 × 10^6 and 3.3 × 10^6 mate pair reads for 3- to 5-kb, 5- to 8-kb and 8- to 11-kb libraries, respectively.

2.2 | Genome assembly

To estimate the genome size of *O. nana* based on the sequencing data, a k-mer spectrum of the genome was built with Kmergenie 1.5692 (Chikhi & Medvedev, 2014) on the paired-end reads. This estimated the *O. nana* genome size around 85 Mb (Appendix S1). The k-mer profile of the genome did not have the two distinct peaks corresponding to the homozygous (right peak) and heterozygous (left peak) k-mer and thus did not correspond to the canonical profile of a heterozygous genome. This is explained by the pooling of individuals that tends to dilute less frequent heterozygous alleles and thus lowers the first peak. The paired-end reads were assembled with DIPSPADES 3.5 (Safonova, Bankevich, & Pevzner, 2015) using the –diploid option of the program that merges homologous contigs into one single contig. Contigs over 500 bp were selected for scaffolding by integration of the three Nextera libraries using BESST 1.3.7 (Sahlin, Vezzi, Nystedt, Lundeberg, & Arvestad, 2014) with default parameters. Gaps in scaffolds were closed using GAPCLOSER 1.12-6 with paired-end and mate pair reads. The scaffolds longer than 2 kb were kept in the final assembly. The gene completeness of the assembly was estimated with CEGMA 2.4.010312 (Parra, Bradnam, & Korf, 2007).

2.3 | mRNA sequencing and assembly

Additionally, 150 males and 50 females were isolated from the sample of the harbour of Toulon, France (see above), on the stereomicroscope and pooled by sex for total mRNA extraction. mRNA was extracted using NucleoSpin RNAXS from Machery-Nagel. cDNA were then constructed using the Illumina SMRTer Ultra low RNA kit. One paired-end library was built for each sex using the NEB-Next DNA sample PrepReagent Set1 kit from Ozyme. cDNA libraries were sequenced in an Illumina MiSeq. Reads were trimmed as previously described and assembled with VELVET 1.2.07 (Zerbino & Birney, 2008) followed by OASES 0.2.08 (Schulz, Zerbino, Vingron, & Birney, 2012). Redundant contigs were clustered with CD-HIT 4.6.1 using a 95% identity cut-off (Li & Godzik, 2006) producing 40,011 and 39,237 contigs for the female and male transcriptomes, respectively.

2.4 | Genome annotation

Assembled transcripts were aligned against the *O. nana* genome using BLAT 36 (Kent, 2002), and refined alignments were produced locally with EST2GENOME 5.2 (Mott, 1997) to produce a first biological evidence for gene prediction. Proteomes from other sequenced and annotated copepod and crustacean genomes were downloaded from public resources. This included the genomic data of *Tigriopus californicus*, *Eurytemora affinis* and *Daphnia pulex* (Colbourne et al., 2011). The proteomes were aligned against the *O. nana* genome with BLAT and realigned locally with GENEWISE 2.2 (Birney & Durbin, 2000) to produce a second gene prediction. In addition, crustacean proteins from Uniprot (Apweiler et al., 2004) were downloaded and aligned similarly. Gene predictions from mRNA and proteomes were used by GMOVE (Dubarry et al., 2016) to build the gene models. The gene set was represented by 15,359 genes with 23.35% of monoexonic genes. Predicted proteins were translated from the gene models, and a domain search was performed with INTERPROSCAN 5.8–49.0 (Jones et al., 2014).

2.5 | Comparative genomic analysis of copepods

Genomes, gene annotations and proteomes of two copepods (*T. californicus* and *E. affinis*) and one branchiopod (*D. pulex*) were downloaded from public databases and used for comparative genomic analysis of the *O. nana* genome. Genome and gene structure metrics like number of exon/intron per gene, exon/intron size and monoexonic gene rate were calculated from the ggf files (Appendix S2). Eleven nonredundant metrics were analysed by principal component analysis (PCA) to identify the specific genomic structure of the four crustaceans. Orthologous gene pairs were formed by best reciprocal blast hit (BRH) of each protein against the proteomes using two filters: an alignment length/protein length ratio over 50% and an identity over 30%. The matrix containing the number of BRH between each organism was used to cluster the organisms and to represent their relative distance based on their gene homologies on a dendrogram using hierarchical clustering. Functional protein domains were detected with InterProScan on each proteome, which provided the domain annotations and their possible Gene Ontology (GO). For each proteome, the number of genes classified in GO terms was used to calculate a distance matrix of the crustaceans that was taken as input for hierarchical clustering. The protein domains having a relatively high difference in occurrence among the three copepod proteomes (i.e., with standard deviation of the occurrence among the three copepods over 10) were represented on a heatmap to identify clusters of domains having the same occurrence pattern. An equivalent analysis was performed using the GO terms.

2.6 | Ribosomal sequences analysis in the *Tara* Oceans samples

Ribosomal 28S corresponding to Oithonidae were downloaded from the NCBI (Appendix S3). To avoid the use of 28S sequences that corresponded to taxonomic assignation mistakes, only sequences associated with published research articles were selected. Nonredundant and sequences without undetermined nucleotides were selected. Ten additional sequences of other Cyclopoida were also added to the data set as an outgroup. Sequences were aligned with MAFFT 7 (Katoh & Standley, 2013) using the progressive method, and most left and most right alignment regions were trimmed to produce a 582-bp alignment. Based on this alignment, a phylogenetic tree was built on the MEGA 5.2 platform (Kumar, Stecher, & Tamura, 2016) with PHYML 3 (Guindon et al., 2010) using a generalized time-reversible model and 100 bootstraps to obtain branch support (Appendix S4). The per-base Shannon entropy was calculated along the alignment to define conserved and variable regions that corresponded to *Oithona*-specific and species-specific regions, respectively (Appendix S5). A matrix identity was built to define the minimum identity threshold to use for species specificity for further metagenomic reads alignment (Appendix S6). A reasonable specificity was reached for a 98% identity threshold for *Oithona similis* and *O. nana*, but the discrimination of *Oithona atlantica* from *O. plumifera* was not possible. To minimize the sensitivity loss in species identification, we did not use a higher threshold. Each sequence was then considered as a reference and indexed independently with BWA 0.7.12 (Li & Durbin, 2010). Metagenomic reads from *Tara* Ocean samples corresponding to the fraction size of 20–180 μm were aligned with "bwa mem" against each 28S reference and only reads with identity ≥98% were selected. The maximum depth of coverage in the conserved regions was counted as the total *Oithona* abundance. The proportion of each species was then estimated and the difference between the total *Oithona* abundance and the sum of all species abundance was counted as undefined species (Appendix S7). The copy number per genome of the 28S gene was assumed to be fairly uniform for all *Oithona* species.

2.7 | Genomic variant analysis in the *Oithona nana* populations

Metagenomic reads from *Tara* Oceans samples corresponding to the 20–180 μm fraction size collected in the surface and deep chlorophyll maximum (DCM; Pesant et al., 2015) were aligned against the *O. nana* genome using "bwa mem" with a 17-bp seed and stored in one single binary alignment multiple file per station. To avoid spurious mapping, reads with low complexity were discarded with Dust (Morgulis, Gertz, Schaffer, & Agarwala, 2006). Metagenomic reads with an identity cut-off over 80% were selected. Genome coverage at each position was calculated with BEDTOOLS 2.24.0 (Quinlan, 2014). Samples having a mean identity below 95%, or a bimodal distribution of the identity, were discarded (Appendix S8). A second identity filter of 97% was applied, and samples having a mean genomic coverage above 4× were kept. The five selected stations that passed these filters (TARA_10, 11, 12, 24, 26) corresponded to stations where *O. nana* were previously identified using the 28S sequences (see Section 3). Based on the metagenomic reads alignments, the correlation between the 28S coverage and the genomic coverage was calculated to validate the metagenomic reads mapping and filtering. Variable genomic sites were detected using the samtools/bcftools pipeline (Li et al., 2009), and loci with a maximum of two alleles were kept. For each sample, the variants were annotated with SNPEFF (Cingolani et al., 2012) in order to assess their genomic location (i.e., exon, intron, UTR, intergenic) and their impact on the predicted proteins (i.e., missense, synonymous variant and nonsense). The distribution of the proportion of the average synonymous and missense variants by gene was modelled by a gamma and exponential distribution, respectively, and outliers (i.e., gene that presented more variant than expected) were tested and p-values were corrected using a strict Bonferroni procedure.

Biallelic sites with a total coverage between 4× and 80× in the five populations were merged in a single vcf file containing 221,018 genomic positions. The B-allele frequency (BAF) was calculated and used for the pairwise F_{ST} calculation at each locus; the mean and median pairwise F_{ST} were used to estimate the genetic distance between the five populations. As the individual genotypes were not accessible, the intrapopulation variance could not be estimated and thus the significance of the F_{ST} could not be determined. A PCA on the BAF of the five populations was performed to cluster the populations.

The BAF calculated previously was used as input to calculate the F_{LK} statistic (Bonhomme et al., 2010), which is an improvement of the LK statistic (Lewontin & Krakauer, 1973). Compared to F_{ST} and LK, F_{LK} uses as prior a kinship matrix of the populations based on the allele frequencies. As no population could be considered as outgroup, the "midpoint" option was used to build the kinship matrix. The F_{LK} use assumes that the majority of the variants are under the neutral model and thus the F_{LK} distribution has to follow a chi-square distribution with a degree of freedom of $n-1$ (n = number of populations). The neutrality of the variants was tested by comparing the F_{LK} distribution to the chi-square distribution with $df = 4$. To control the *p*-value inflation due to multiple testing, we applied a Benjamini–Hochberg correction to obtain a *q*-value for each variable locus. Loci with a *q*-value under .2 were considered to be under selection. The genomic scaffolds that contained loci under selection were represented on a Manhattan plot. These loci were then compared to the genome variant annotation to provide a list of genes and biological functions under selection.

2.8 | *Oithona* spp. genomic variation and the Mediterranean circulation patterns

Previously, we identified four samples originating from stations located in the southern part of the MS (TARA_7, 8, 17 and 18) having an identity around 95% and a bimodal distribution of identity. These samples were thought to contain an *Oithona* species closely related to *O. nana* but genetically distant enough to be not considered as *O. nana* (mean identity around 95%). We used the metagenomic reads alignment from the ten stations (TARA_7, 8, 10, 11, 12, 14, 17, 18, 24 and 26) with a cut-off identity of 90% to detect the genomic variants and selected 754,669 biallelic loci with a valid call in the 10 samples. We calculated the BAF and observed the variation in the BAF with respect to the hydrodynamical connectivity starting from stations located on the Algerian Current (Stations 7 and 8). The stations were sorted from upstream to downstream along the main circulation patterns (Algerian, Lybio-Egyptian and Northern currents) in two possible ways. Way1 follows the order TARA_7, 8, 14, 12 and 11 and is located only in the western basin. Way2 follows the order TARA_7, 8, 17, 18, 26 and 24, and starts from the western basin to the Levantine basin and the Adriatic Sea. The variation in the BAF was used to identify genomic clines in the *Oithona* populations along the two proposed ways. We analysed the variation in the BAF as a function of the Lagrangian distance from TARA_7. Briefly, the Lagrangian distance was computed as the mean travel time between each area over a large ensemble of Lagrangian particles simulation (Berline, Rammou, Doglioli, Molcard, & Petrenko, 2014).

3 | RESULTS

3.1 | Genome assembly and annotation of *Oithona nana*

The *O. nana* somatic genome was sequenced and assembled in 7,375 contigs with a N50 of 39.6 kb that were linked in 4,626 scaffolds with a N50 of ~400 kb (Appendix S9). After gap closing, only 3.7% of undetermined bases remained. The final genome sequence contains 89% of the 458 conserved core eukaryotic proteins used by CEGMA with an average of 1.2 copies per gene. Thus, the sequencing strategy (i.e., pooled individuals) and the assembly based on dipSpades successfully merged the different haplotypes present in the initial genomic DNA into a single representative haplotype suitable for gene annotation. Low-complexity and repetitive elements represented 3.6% of the genome sequence and were masked prior to the genome annotation. The genome annotation procedure applied with GMOVE predicted 15,359 genes with 23% monoexonic genes. The resulting proteome was scanned to identify conserved protein domains and 72.3% of the proteins harboured at least one conserved domain present in the INTERPRO database.

3.2 | Comparative analysis of the *Oithona nana* genome

To identify *O. nana* specific genomic features, the structures of the *O. nana* genome and genes were compared to other available annotated copepod genomes, including the genome of the harpacticoid *Tigriopus californicus* (which also belongs to Podoplea), the calanoid *Eurytemora affinis*, and to another crustacean genome, the branchiopod *Daphnia pulex*. Some copepod somatic genomes are known to be subject to chromatin reduction (Clower, Holub, Smith, & Wyngaard, 2016; Drouin, 2006; Rasch & Wyngaard, 2006; Sun, Wyngaard, Walton, Wichman, & Mueller, 2014; Zagoskin, Marshak, Mukha, & Grishanin, 2010), so for the two copepod genomes used for the comparative analysis, we refer only to their somatic genomes. The PCA of the four crustacean genomes and gene metrics (Figure 1a) showed that *O. nana* and *T. californicus* share similar gene and genome structures compared to *E. affinis* and *D. pulex*, which form two separated groups (Figure 1b). The podoplean group is characterized by larger exon sizes, higher number of monoexonic genes, lower number of exon by genes, smaller intron sizes, less genes and smaller genes and genome. The construction of orthologous gene pairs (Figure 1c) showed that the podoplean shared more orthologous genes between them (2,911 genes) than with the two other crustaceans. *Eurytemora affinis* has a large proportion of genes (20,796) that do not have any orthologues with the two podopleans. The clustering of the four crustaceans (Figure 1d) based on the number of orthologous gene pairs produced a dendrogram indicating that calanoids indeed diverged earlier than *T. californicus* and *O. nana*. The PCA based on gene and genome metrics and the orthologous gene pair analysis showed that within copepods, the *E. affinis* genome evolved differently than the podoplean genomes through two possible main events: the appearance of both new introns and genes. These results also showed that podopleans have more compact somatic genomes.

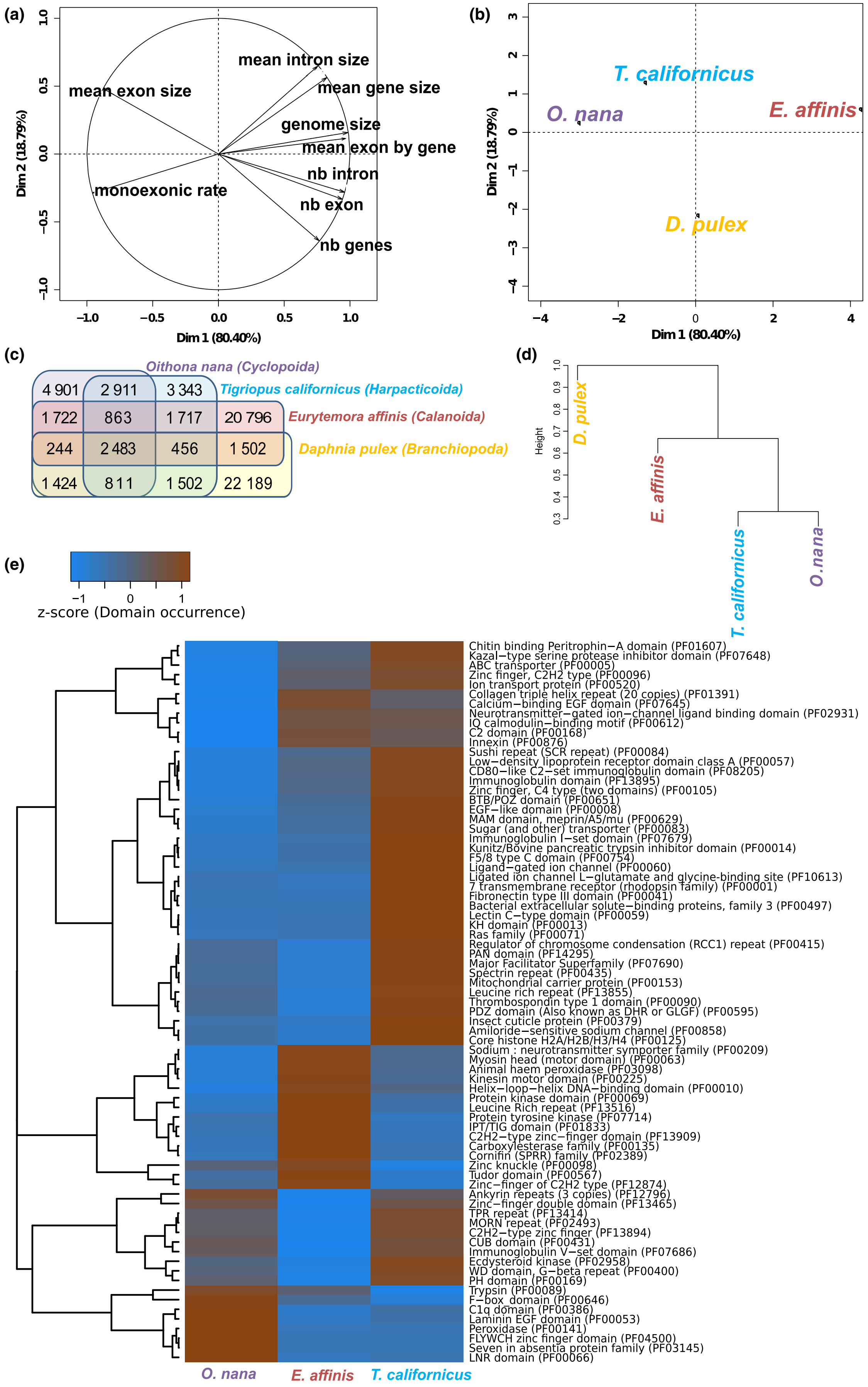
3.3 | Functional protein domains

The Pfam protein domains found in the three copepods proteomes were linked to the GO database and the distributions of high-level GO terms did not show any specific differences. However, the analysis based directly on the Pfam domain occurrences (Figure 1e) and their related GO biological process (Appendix S10) represented on a heatmap provided a list of domains and biological processes overrepresented in the proteomes. These domains were clustered in five groups: the *O. nana* group, the *E. affinis* group, the *T. californicus* group, the *T. californicus* and *E. affinis* group and the *O. nana* and *T. californicus* group. The latter two groups contained domains overrepresented in the two species. The *O. nana* group (Table 1) was represented especially by domains involved in cell differentiation (LNR and seven in absentia domains), multicellular organism development (laminin EGF domain), proteolysis (trypsin and F-box domains) and response to oxidative stress (peroxidase). The proteins harbouring LNR domains (32 proteins) presented domain associations with metalloproteinase (six proteins), or with one or two trypsin domains (six proteins). Furthermore, 18 LNR domain-containing proteins were found to have only LNR domains. Three of the LNR domain-containing proteins found in *O. nana* were coded by genes localized in a cluster of six genes and all coded trypsin domain-containing proteins, which may result from multiple gene duplications. Other LNR domain-containing proteins are located in different scaffolds. As certain proteins containing only LNR domains reached more than 1,000 amino acids, they may contain unknown functional protein domains.

3.4 | Global biogeography of *Oithona* species using metagenomic data

Metagenomic reads sequenced from *Tara* Oceans samples (Karsenti et al., 2011; Pesant et al., 2015) using the 20–180 μm fraction size collected from the surface and DCM waters were aligned on a manually curated 28S sequence database. The database contained ribosomal sequences from seven different *Oithona* species (*O. similis*, *O. atlantica*, *O. plumifera*, *O. nana*, *O. simplex*, *O. davisae* and *O. brevicornis*). Aligned reads with an identity over 98% were selected. This allowed the identification of *O. similis* and *O. nana* without ambiguity. Among the Oithonidae, a conserved region of the 28S was identified (Appendix S4). The total coverage obtained in this region was then considered as the total Oithonidae abundance. This allowed the quantification of undefined *Oithona* species (see Section 2). The *Oithona* biogeography in the surface waters (Figure 2a) showed a amphitropical distribution for *O. similis*, which was present in the temperate waters of Northern and Southern hemispheres and Antarctic polar waters. The OTU1 was mostly identified in the tropical and subtropical waters with an exception in one sub-Antarctic sample from station 85. *Oithona nana* was mostly found in the Mediterranean Sea but was also present sporadically in the Indian (TARA_39) and Pacific Oceans (TARA_140). Seven stations located in the Gambier Islands (SPO) showed the presence of *Oithonidae* but the species could not be determined. Several stations presented an assemblage of *O. similis* and OTU1 especially in the NAO stations localized along the Gulf Stream, the Brazil Current and also in the

FIGURE 1 Comparative genomic analysis of the *Oithona nana* genome. (a) principal component analysis (PCA) of the variables based on 11 genomic and gene metrics. (b) PCA of the individuals based on 11 genomic and gene metrics. (c) Venn diagram of the orthologous gene pairs. Numbers indicate the number of best reciprocal blast hit. (d) Hierarchical clustering based on orthologous gene pairs. (e) Heatmap of Pfam domains overrepresented in copepod species. The Pfam domains are clustered by their occurrence *z*-score


(a)
mean intron size
mean gene size
mean exon size
genome size
mean exon by gene
nb intron
monoexonic rate
nb exon
nb genes
Dim 2 (18.79%)
Dim 1 (80.40%)
1.0
0.5
0.0
−0.5
−1.0
(b)
T. californicus
O. nana
E. affinis
D. pulex
Dim 2 (18.79%)
Dim 1 (80.40%)
3
2
1
0
−1
−2
−3
−4
4
(c)
Oithona nana (Cyclopoida)
Tigriopus californicus (Harpacticoida)
Eurytemora affinis (Calanoida)
Daphnia pulex (Branchiopoda)
4 901
2 911
3 343
1 722
863
1 717
20 796
244
2 483
456
1 502
1 424
811
1 502
22 189
(d)
Height
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
D. pulex
E. affinis
T. californicus
O.nana
(e)
−1
0
1
z-score (Domain occurrence)
Chitin binding Peritrophin−A domain (PF01607)
Kazal−type serine protease inhibitor domain (PF07648)
ABC transporter (PF00005)
Zinc finger, C2H2 type (PF00096)
Ion transport protein (PF00520)
Collagen triple helix repeat (20 copies) (PF01391)
Calcium−binding EGF domain (PF07645)
Neurotransmitter−gated ion−channel ligand binding domain (PF02931)
IQ calmodulin−binding motif (PF00612)
C2 domain (PF00168)
Innexin (PF00876)
Sushi repeat (SCR repeat) (PF00084)
Low−density lipoprotein receptor domain class A (PF00057)
CD80−like C2−set immunoglobulin domain (PF08205)
Immunoglobulin domain (PF13895)
Zinc finger, C4 type (two domains) (PF00105)
BTB/POZ domain (PF00651)
EGF−like domain (PF00008)
MAM domain, meprin/A5/mu (PF00629)
Sugar (and other) transporter (PF00083)
Immunoglobulin I−set domain (PF07679)
Kunitz/Bovine pancreatic trypsin inhibitor domain (PF00014)
F5/8 type C domain (PF00754)
Ligand−gated ion channel (PF00060)
Ligated ion channel L−glutamate and glycine-binding site (PF10613)
7 transmembrane receptor (rhodopsin family) (PF00001)
Fibronectin type III domain (PF00041)
Bacterial extracellular solute−binding proteins, family 3 (PF00497)
Lectin C−type domain (PF00059)
KH domain (PF00013)
Ras family (PF00071)
Regulator of chromosome condensation (RCC1) repeat (PF00415)
PAN domain (PF14295)
Major Facilitator Superfamily (PF07690)
Spectrin repeat (PF00435)
Mitochondrial carrier protein (PF00153)
Leucine rich repeat (PF13855)
Thrombospondin type 1 domain (PF00090)
PDZ domain (Also known as DHR or GLGF) (PF00595)
Insect cuticle protein (PF00379)
Amiloride−sensitive sodium channel (PF00858)
Core histone H2A/H2B/H3/H4 (PF00125)
Sodium : neurotransmitter symporter family (PF00209)
Myosin head (motor domain) (PF00063)
Animal haem peroxidase (PF03098)
Kinesin motor domain (PF00225)
Helix−loop−helix DNA−binding domain (PF00010)
Protein kinase domain (PF00069)
Leucine Rich repeat (PF13516)
Protein tyrosine kinase (PF07714)
IPT/TIG domain (PF01833)
C2H2−type zinc−finger domain (PF13909)
Carboxylesterase family (PF00135)
Cornifin (SPRR) family (PF02389)
Zinc knuckle (PF00098)
Tudor domain (PF00567)
Zinc−finger of C2H2 type (PF12874)
Ankyrin repeats (3 copies) (PF12796)
Zinc−finger double domain (PF13465)
TPR repeat (PF13414)
MORN repeat (PF02493)
C2H2−type zinc finger (PF13894)
CUB domain (PF00431)
Immunoglobulin V−set domain (PF07686)
Ecdysteroid kinase (PF02958)
WD domain, G−beta repeat (PF00400)
PH domain (PF00169)
Trypsin (PF00089)
F−box domain (PF00646)
C1q domain (PF00386)
Laminin EGF domain (PF00053)
Peroxidase (PF00141)
FLYWCH zinc finger domain (PF04500)
Seven in absentia protein family (PF03145)
LNR domain (PF00066)
O. nana
E. affinis
T. californicus

eastern basin of the MS. Besides *O. similis*, *O. nana* and OTU1, other *Oithona* species present in our reference database have not been identified in any *Tara* Oceans samples.

TABLE 1 Pfam domains overabundance in the *Oithona nana* genome compared to other copepods

Pfam	*Oithona nana*	*Tigriopus californicus*	*Eurytemora affinis*	*Daphnia pulex*
Peroxidase (PF00141)	32	3	2	0
LNR domain (PF00066)	47	4	3	6
Seven in absentia protein family (PF03145)	33	9	9	2
Trypsin (PF00089)	170	146	161	268
Laminin EGF domain (PF00053)	161	75	59	85
C1q domain (PF00386)	38	14	8	153
F-box domain (PF00646)	31	5	14	9
FLYWCH zinc finger domain (PF04500)	20	2	2	9

Fewer samples were sequenced from the DCM than from the surface waters (Appendix S11). Comparing the relative abundance of *Oithona* species for the eight stations (TARA_7, 8, 10, 12, 17, 18, 23 and 25) for which sequencing data from DCM and surface waters were available (Figure 2b), *Oithona similis* was more abundant in the DCM except for station TARA_23. No specific trend was observed for *O. nana* and OTU1.

3.5 | *Oithona nana* genomic variations in the Mediterranean Sea

The genomic variation landscape of *O. nana* was investigated in five *Tara* Oceans stations (TARA_10, 11, 12, 24 and 26) located in the northern part of the MS. The genomic coverage obtained from the *O. nana* genome was strongly correlated with the 28S coverage (Pearson's $R^2 = .99$; Figure 3a), which validates the choice of the parameters and the filters applied to select the metagenomic reads corresponding to *O. nana* and also to select the stations that did not provide population admixtures with other *Oithona* species.

Oithona nana populations had variants ranging from 608,559 variants in the station TARA_26, which corresponds to the station

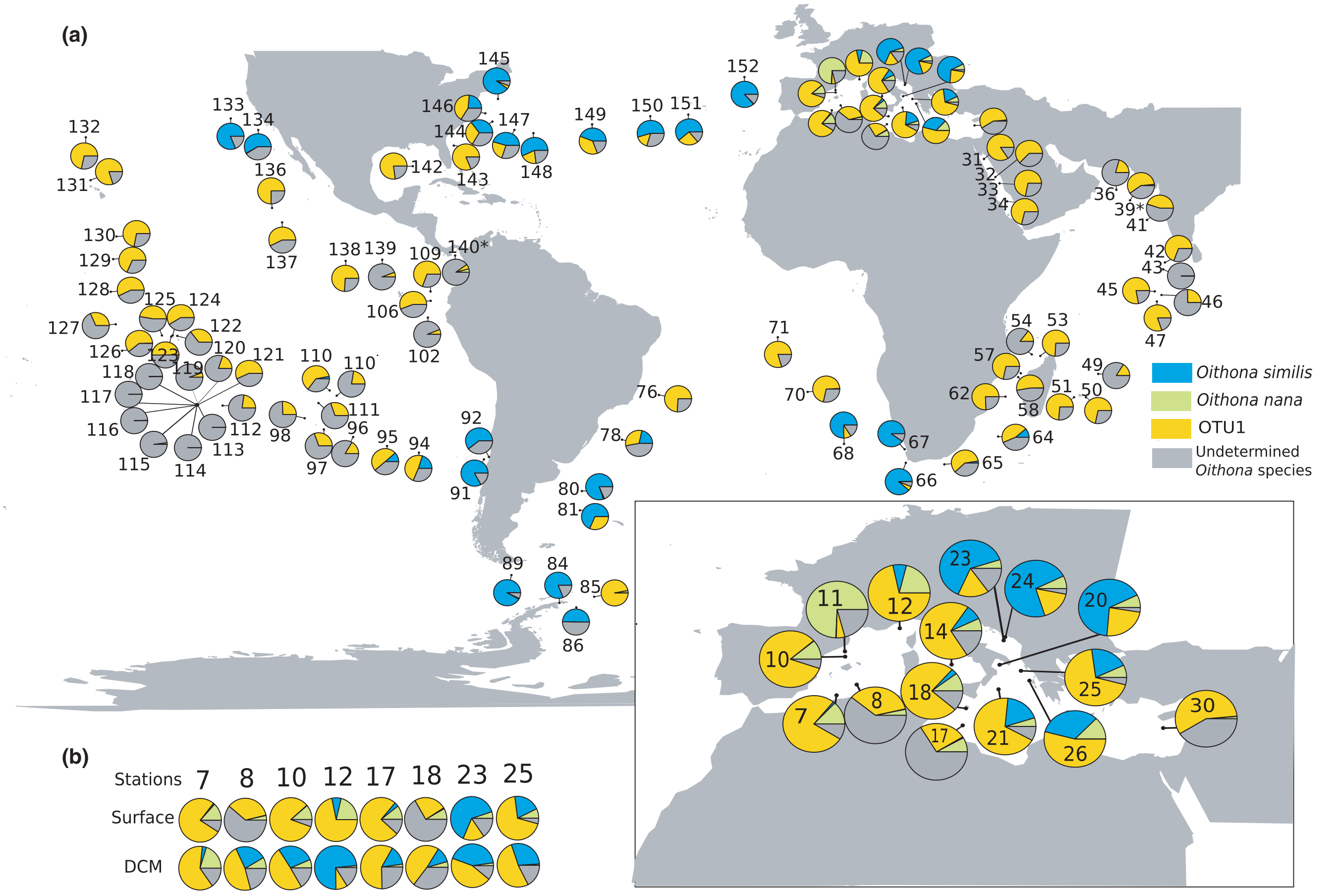


FIGURE 2 Biogeography of *Oithona* species. (a) Global biogeography of the *Oithona* species using the *Tara* Oceans metagenomic data and 28S sequences. The proportion of each species is represented by a pie chart; asterisks correspond to *Oithona nana* presence with a very low abundance. The MS region is enlarged for clarity. (b) Variation in *Oithona* species abundance depending on the sampling depth in the Mediterranean Sea. The proportion of each species is represented by a pie chart. DCM corresponds to deep chlorophyll maximum

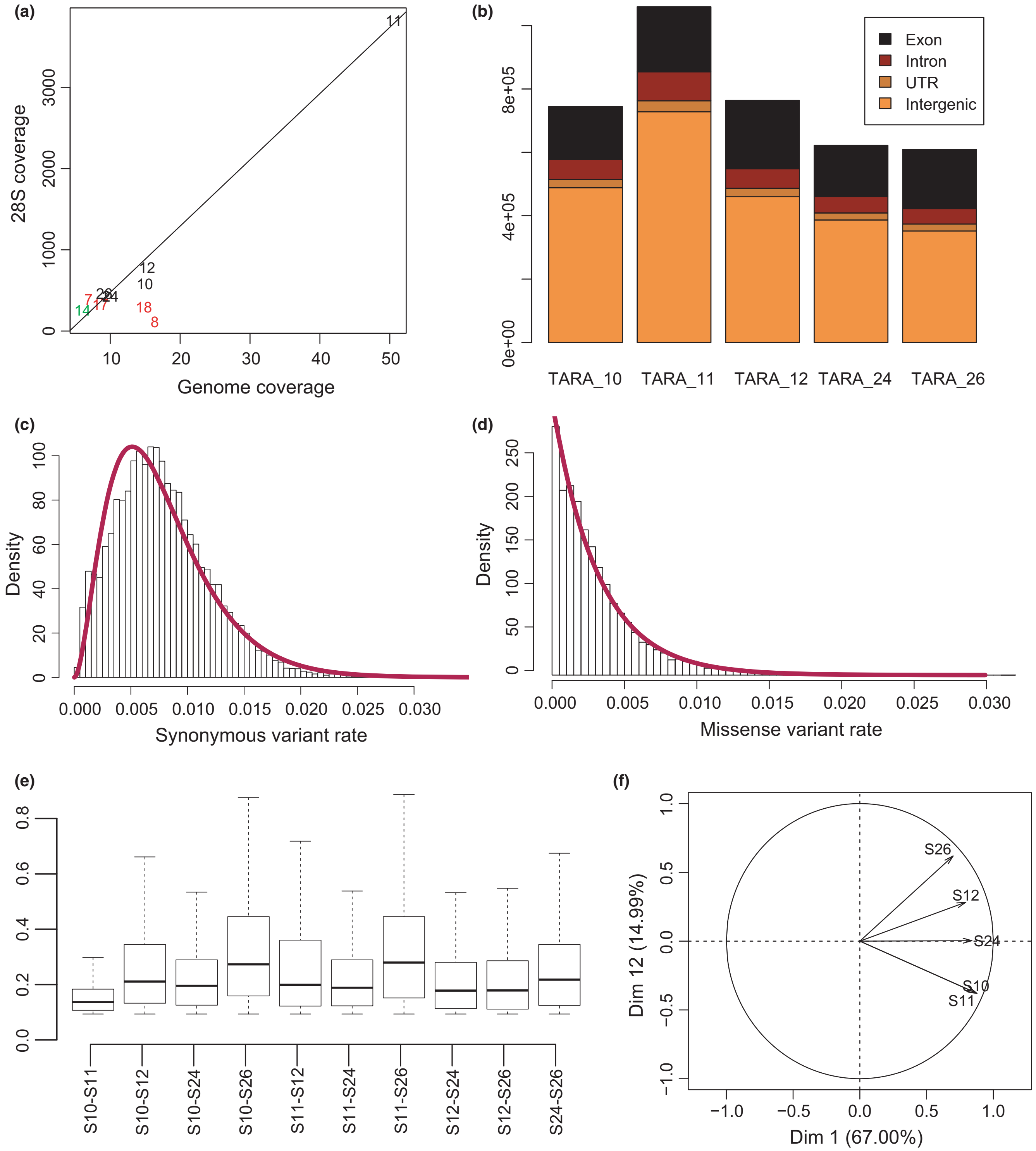


FIGURE 3 Genomic variants in *Oithona nana* populations of the Mediterranean Sea. (a) Correlation between the 28S and the genomic coverage. Station numbers are plotted; numbers in red correspond to stations discarded due to a bimodal distribution of the identity percentage. Stations in green correspond to stations discarded due to a low genomic coverage. (b) Distribution of the variants in the genome. (c) Distribution of the synonymous variant rate. The red curve corresponds to the gamma distribution estimated from the data. (d) Distribution of the missense variant rate. The red curve corresponds to the exponential distribution estimated from the data. (e) Distribution of the pairwise F_{ST}. (f) principal component analysis results of the populations based on the B-allele frequency

where *O. nana* was the less abundant over the five selected stations to 1,059,550 variants (Appendix S12) for the station TARA_11, which corresponds to the sample where *O. nana* was the more abundant over the five selected stations. This indicates that the amount of called variants depends on the genomic coverage. Introns represent 20.8% of the genome, and the annotation of the variants

TABLE 2 Median pairwise F_{ST} distances between *Oithona nana* populations sampled in five stations of the Mediterranean Sea

	Western MS basin			Eastern MS basin	
	TARA_10	TARA_11	TARA_12	TARA_24	TARA_26
TARA_10	–	0.05	0.13	0.1	0.22
TARA_11		–	0.11	0.1	0.22
TARA_12			–	0.1	0.09
TARA_24				–	0.14
TARA_26					–

showed less intronic variants than expected comparing to a random distribution of the variants, suggesting that the 97% identity filter used to select the metagenomic reads was too stringent for the intronic regions (Figure 3b). The synonymous variant rate by gene followed a gamma distribution (shape = 2.94, rate = 379.6) and no outliers (genes with more synonymous variants than expected) were detected (Figure 3c). The missense variants rate by gene followed an exponential distribution (rate = 314) and no outliers were detected (Figure 3d).

Among all variable loci detected previously, we selected 221,018 biallelic loci and calculated the BAF and pairwise F_{ST} to estimate the genomic distance between the five populations. The pairwise F_{ST} between stations (Figure 3e, Table 2) showed that the populations from the Adriatic Sea (TARA_24 and 26) were structured and presented moderate differentiation comparing to two populations from the western basin (TARA_10, 11). We observed a moderate genetic structure within the western basin population group. The lower value was obtained for TARA_10 and 11 (F_{ST} = 0.048), which was expected considering the relatively short distance between the two stations (<100 km). The PCA based on the BAF clustered the Adriatic populations with the one from TARA_12 (Figure 3f) and separated the population from TARA_10 and TARA_11. This suggested that the *O. nana* populations are structured between in the MS basins but also within the two basins.

3.6 | *Oithona nana* genomic loci under positive selection in the Mediterranean populations

To identify *O. nana* genomic loci under selection, the F_{LK} statistics was calculated based on the BAFs from the five populations selected previously. The LK and the F_{LK} globally fitted a χ^2 (*df* = 4) distribution (Appendix S13) and thus supported the neutral model. Among the 221,018 biallelic loci tested to be under a χ^2 (*df* = 4) distribution, 20 loci had a *q*-value <.2 and were considered under positive selection (Table 3). The BAFs of loci under positive selection showed that most of the loci were under positive selection in populations from TARA_24 and TARA_26 stations (Appendix S14). Different patterns of selective sweeps were observed (Appendix S15). Two loci presented a clear soft selective sweep (scaffold_75 and scaffold 2085); these loci were shown as under selection in the TARA_24 and 26 populations. Five loci had a hard selective sweep signature (two on scaffold_4, one on scaffold_7 and two on scaffold_17); for these five loci, the selection occurred only in the TARA_26 population.

The alleles under selection corresponded to eight synonymous, four missense, five intergenic, two intron and one 3′ variant. One of the missense variants (Figure 4a) is located in the GSONAT00014698001 gene (scaffold_2085). The Manhattan plot around this variant presented two drafted variants, one located in the first exon and one in the first intron, suggesting a soft selective sweep signature (Figure 4b,c). The gene product was a 686-amino acid protein presenting one signal peptide and five LNR domains (Figure 4d). The variant was localized on the most *N*-terminal LNR domain and changes a threonine to a proline (Figure 4e). The RNA-seq reads from the male and female transcriptomes were mapped on the genes and, based on the RPKM values, the expression of GSONAT00014698001 was likely to be male specific. Two other missense variants were located in the first exon of GSONAT00014305001 (scaffold_1229) that codes for a FMRFamide receptor, which is a G protein-coupled receptor (IPR017452) of FMRFamide neuropeptides. The BAF of these two variants showed that selection was occurring in the populations from TARA_24 and TARA_26. The scaffold_1229 contained only three variable loci; thus, the selective sweep pattern around this locus could not be determined. The last missense variant is located in the fourth exon of the GSONAT00006046001 gene (scaffold_31), which codes a hypothetical protein without known domain. The BAFs of this variant showed that the selection occurs only in the population from TARA_26 and the Manhattan plot suggests a hard selective sweep around this locus (Appendix S15).

3.7 | Genomic clines between *Oithona* species in the Mediterranean Sea

We extended the previous analysis to integrate five stations (TARA_7, 8, 14, 17 and 18) located in the southern part of the MS that suggested the presence of an *Oithona* species closely related to *O. nana* but having lower identity (~95%). From the metagenomic reads alignments we selected 754,669 biallelic loci with coverage between 4× and 80× in all 10 stations to calculate the BAFs, having an identity percentage >90%. Based on the topology of the 28S phylogeny built with consensus sequences from the alignment of metagenomic reads from TARA_8 (with branching outside the *O. nana* clade; Appendix S16), and on the bimodal F_{LK} distribution calculated from the BAFs (Appendix S17), we confirmed the presence of another *Oithona* species, closely related, but distinct to *O. nana*.

The populations from stations located in the Southern part of the MS showed BAF medians ranging from 0.54 to 0.82 (Table 4). Populations from stations located in the northern part of the MS presented lower BAF medians ranging from 0 to 0.43. The population from TARA_11 had a unimodal BAF distribution with a peak at BAF = 0; meanwhile, all other stations had a bi- or trimodal distribution with two peaks at BAF = 0 and BAF = 1, and sometimes a third wide peak between BAF = 0 and 1. There was a decrease in the

TABLE 3 Genomic location and functional annotation of loci under positive selection in the *Oithona nana* populations

Scaffold	Position	Ref	Alt	Gene model	Variant	AA modification	Annotation
4	471967	G	A	GSONAT00000985001	Synonymous	p.Ser743Ser	Zinc finger, C2H2 protein
4	723596	A	C	GSONAT00001049001	Synonymous	p.Gly198Gly	Pantothenate kinase
7	175799	T	A	GSONAT00001735001	Synonymous	p.Arg312Arg	PH domain-like protein
10	237870	T	A	GSONAT00002382001	Synonymous	p.Ile2197Ile	Dynein beta chain
10	247465	G	A	GSONAT00002383001	Synonymous	p.Phe19Phe	Hypothetical protein
17	96647	G	A	GSONAT00003499001	Intron		LIM domain protein
17	350150	T	C	GSONAT00003546001	Synonymous	p.Tyr195Tyr	LIMS1
24	706636	G	A	GSONAT00005133001	3 prime UTR		Hypothetical protein
31	326964	G	T	GSONAT00006046001	Missense	p.Cys153Phe	Hypothetical protein
56	319206	T	A	GSONAT00008543001	Upstream gene		Fork head domain protein
61	332115	T	C	GSONAT00008966001	Synonymous	p.Val255Val	GABA/glycine receptor
62	372141	C	T	GSONAT00009071001	Synonymous	p.Ser51Ser	Hypothetical protein
75	293884	A	G	GSONAT00009905001	Upstream gene		Hypothetical protein
207	24666	G	A	GSONAT00013216001	Upstream gene		Hypothetical protein
1229	680	T	A	GSONAT00014305001	Missense	p.Leu64His	FMRFamide receptor
1229	689	C	T	GSONAT00014305001	Missense	p.Ala67Val	FMRFamide receptor
2085	2429	T	G	GSONAT00014698001	Missense	p.Thr104Pro	LNR domains protein
4053	516	T	A		Intergenic		
4053	597	T	C		Intergenic		
4053	1311	A	G		Intergenic		

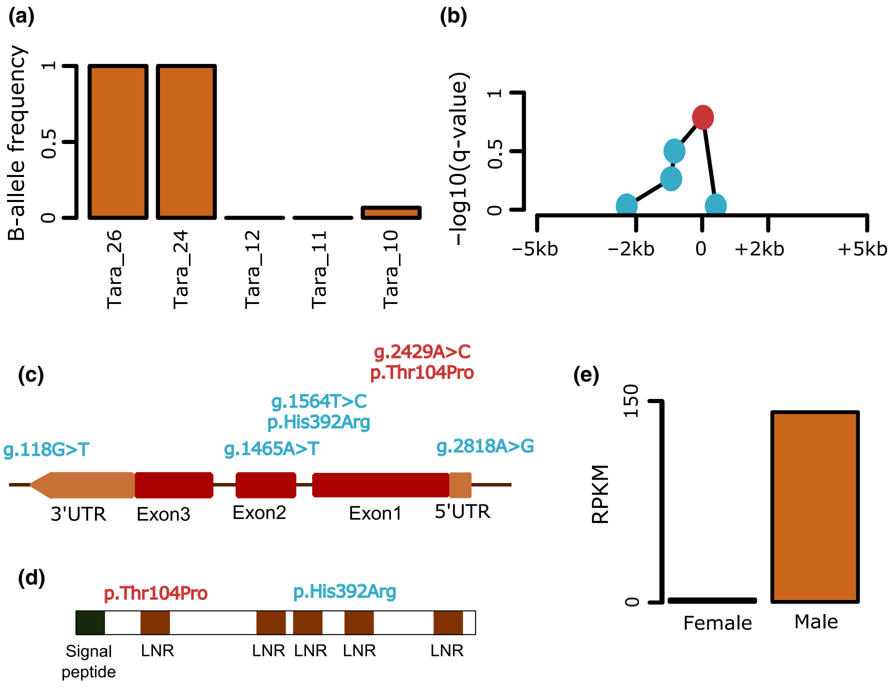


FIGURE 4 Positive selection of variant in the GSONAT00014698001 gene. (a) B-allele frequency value of the loci under selection in the northern stations. (b) Manhattan plot in a 2-kb window around the loci under selection. The dot in red is considered under selection (q-value < 0.2). (c) GSONAT00014698001 gene structure and variable sites. (d) GSONAT00014698001 protein structure. (e) RPKM values of the GSONAT00014698001 gene in males and females

median BAF and of the peak height at BAF = 1 along the Algerian Current (from TARA_8 to TARA_17), along the Northern Current (from TARA_14 to TARA_11) and also from the Southern Adriatic Sea to its centre (from TARA_26 to TARA_24). This peak decrease was associated with an increase in the peak height at BAF = 0 and a shift of the third peak to lower BAF values. The highest BAF values were observed in the Algerian Current, which transports the surface waters coming from the NAO into the MS through the Strait of Gibraltar. From these results, we hypothesized that the *Oithona* populations from these ten stations contained the other *Oithona* species that was imported from the NAO into the MS through the Strait of Gibraltar. This putative species was more abundant than *O. nana*

TABLE 4 Lagrangian distances between stations of the Mediterranean Sea. The distance is calculated from the TARA_7 according to Berline et al. (2014)

Stations	Latitude	Longitude	Way	Median BAF	Mean BAF	Lagrangian distance (days)
TARA_7	37.048	1.9402	1 and 2	0.54	0.54	38.3
TARA_8	38.004	3.9777	1 and 2	0.75	0.84	98.8
TARA_10	40.641	2.8772	1	0	0.09	592.3
TARA_11	41.168	2.7996	1	0	0.05	538.3
TARA_12	43.351	7.8994	1	0.33	0.34	449.5
TARA_14	39.903	12.8686	1	0.3	0.32	269.8
TARA_17	36.271	14.3061	2	0.75	0.65	344.5
TARA_18	35.749	14.2947	2	0.82	0.71	298.7
TARA_24	42.457	17.9428	2	0.2	0.22	579.9
TARA_26	38.449	20.1813	2	0.43	0.41	671.5

BAF, B-allele frequency.

from the Algerian Current to the beginning of the Lybio-Egyptian Current. The *O. nana* proportion in the populations increased along the Northern Current and after entering in the Adriatic Sea, suggesting a cline between these two species between the northern and the southern MS.

The variation in the BAF distribution was analysed along the Mediterranean currents as a function of the Lagrangian distance (Berline et al., 2014) from the TARA_7 (Table 4). This variation was analysed along two possible trajectories following the main circulation patterns (Figure 5d). The first trajectory (way1) is located in the western basin; it starts at the beginning of the Algerian Current, going through the Tyrrhenian Sea, merges with the Northern Current and ends in the Balearic Sea; it corresponds to stations TARA_7, 8, 14, 12, 10 and 11. Along this trajectory, we observed a genomic cline between *O. nana* and the other species (Figure 5e) with an admixture in the populations from stations TARA_12 and 14. The second trajectory (way2) goes through the Strait of Sicily and ends in the Adriatic Sea; it corresponds to stations TARA_7, 8, 18, 17, 26 and 24. Along this trajectory, we also observed a genomic cline located in the southern part of the Adriatic Sea with an admixture in the populations from the stations TARA_26 and 24 (Figure 5f).

4 | DISCUSSION

4.1 | Towards the building of new copepod reference genomes

The availability of the *Oithona nana* genome opens the door to study the genetics of this highly abundant and widespread genus, and its interactions across the trophic web at the molecular level. The ability to build efficient data sets for genome assembly depends partly on the abundance and quality of the DNA that can be obtained from a single genotype. In the case of small-sized copepods, the DNA obtained from one individual is not sufficient to build small or large insert Illumina libraries (~1 ng/individual). Thus, the assembly of sequences from a pool of individuals of different genotypes and the merging of the different haplotypes in a single "chimeric" one remains the best strategy. In the case of *O. nana*, it led to a final genome assembly with a good completeness and provided a first high-quality genomic reference for cyclopoids. In future, this approach can be used to create new reference genomes for small-sized zooplankton species.

The comparative genomic analysis provided a first insight into copepod genomic evolution. Compared to other available somatic genomes of copepods, the *O. nana* somatic genome is compact and contains less, but larger, introns. However, in our study, no information concerning differences between the somatic and the germline genome is provided. Only a small fraction of the *O. nana* genome is represented by repetitive elements. This could be explained by chromatin reduction during the early differentiation of the *O. nana* embryos, which is known from some freshwater cyclopoid species, and that might be an important driver in genome rearrangement and evolution (Grishanin, 2014). To better know which genomic elements (if any) are excised during the *O. nana* chromatin reduction, further genomic studies should target germline genome analysis.

4.2 | Protein domains overabundance and new structures in the *O. nana* proteome

Thirty-two genes coding LNR-containing proteins have been identified in the *O. nana* proteome. The association of LNR domains with metalloproteinase domains is well documented especially in the human pappalysins 1 and 2, which both contain three LNR domains. In humans, pappalysin is known to form a homodimeric complex that lyses insulin growth factor binding protein (IGFPB) and regulates insulin-like growth factor (IGF) availability and its downstream effects, which includes male sexual differentiation (Ventura & Sagi, 2012). The association of LNR with trypsin domains and proteins containing only LNR domains are new characteristics that differentiate *O. nana* from other copepods used in the study. The role of these proteins is unknown, but one possibility would be that LNR-

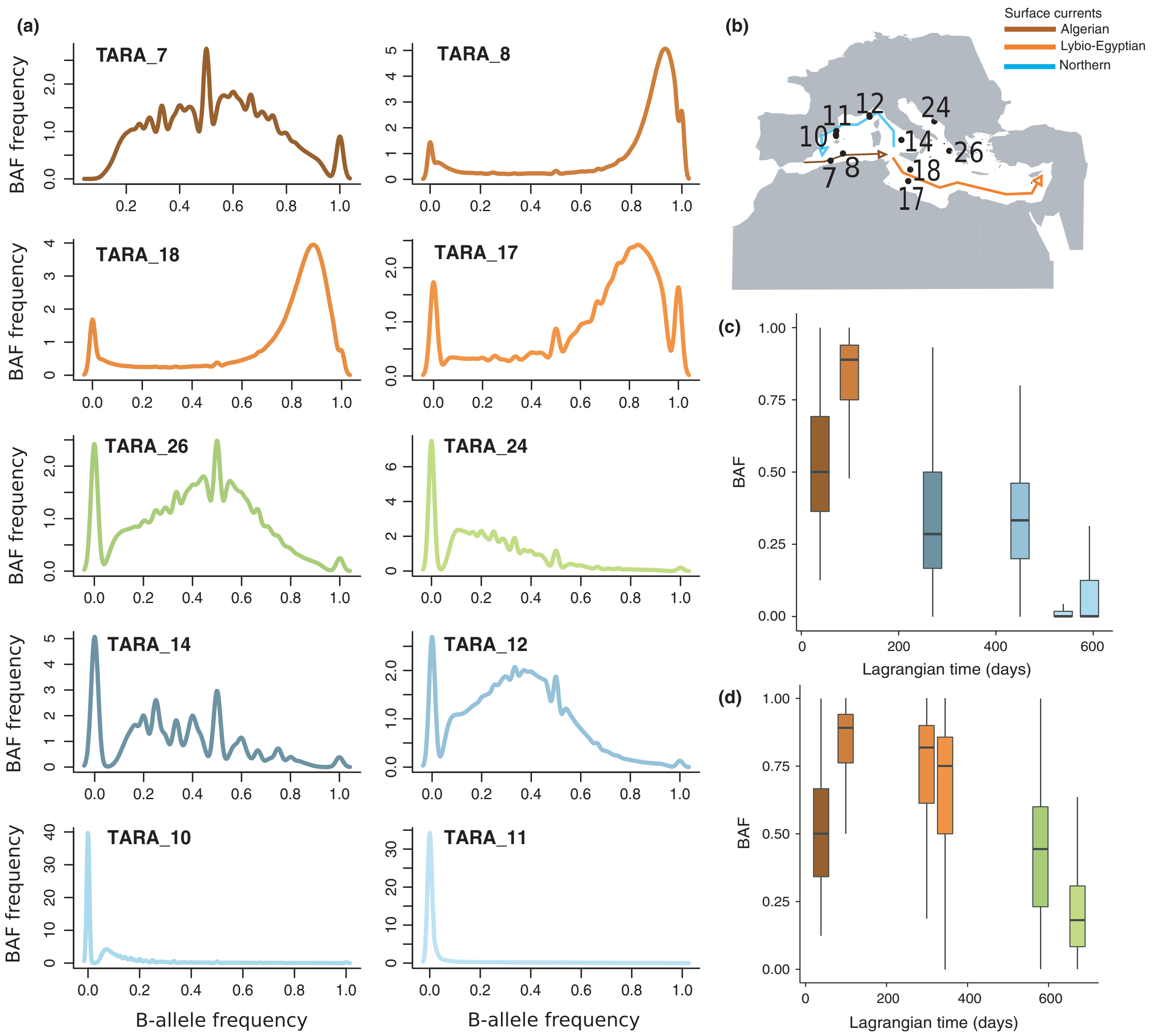


FIGURE 5 *Oithona* population genomics in the Mediterranean Sea. (a) B-allele frequency (BAF) distribution of *Oithona* variants in the southern stations (TARA_7, 8, 17 and 18), northwestern stations (TARA_10, 11, 12 and 14) and northeastern stations (TARA_24 and 26). (b) Dominant Mediterranean surface currents. (c) BAF variation along the surface currents of the way 1. *X*-axis is the Lagrangian distance in days. *Y*-axis is the BAF. (d) BAF variation along the surface current of the way 2

containing proteins might be involved in the development and sex differentiation of *O. nana* through the proteolysis of IGFPBs. To validate this hypothesis, RNA-seq data analysis including different developmental stages (larvae, copepodite, male and female stages) should be produced. The high number of peroxidases in the *O. nana* proteomes suggests that *O. nana* could be efficient in detoxifying a wide range of products through peroxidation. This could partly explain its capacity to survive in highly polluted areas such as the harbour of Toulon, where the specimens were sampled from. RNA-seq experiments analysing the response of *O. nana* to the presence of different pollutants could be performed to identify toxin-specific expression of peroxidase genes.

4.3 | *Oithona* biogeography using the *Tara* Oceans metagenomic data

Until recently, we lacked a global genetic data set to study the distribution of *Oithona* species across the world. Taking advantage of the *Tara* Oceans metagenomic effort, we extracted reads from the 28S DNA marker from the fraction size compatible with the *Oithona* presence. The analysis of Oithonidae 28S sequences showed that 28S resolution varied depending on the species and that OTU1 contained at least two species. The mitochondrial lineages found recently for *O. similis* (Cornils et al., 2017) were also not visible in 28S. Unfortunately, this does not allow a comprehensive biogeography of *Oithona* species by this method but provides still some new aspects of the *Oithona* biogeography. The predominance of *O. similis* in temperate and polar waters illustrates its ecological relevance as a small-sized zooplankton in these oceanic areas. Its absence in tropical regions also confirms previous distribution assumptions (Nishida, 1985). Thus, *O. similis* is a strong candidate for future genomic efforts. *Oithona nana* was identified only in a few *Tara* Oceans samples and mostly in the MS. Despite its worldwide distribution (Razouls et al., 2016), *O. nana* is limited to coastal waters and, unfortunately, the *Tara* Oceans project did not focus on such environments. Several samples were identified as containing undefined *Oithona* spp.; the lack of resolution was partially due to the general lack of references for most of the 44 *Oithona* species. More efforts have to be undertaken for generating new references for under-represented species in the public databases.

4.4 | Population genomics using metagenomic data and the *O. nana* genome

The use of metagenomic data to identify zooplankton genomic variants at the whole genome level has been performed for the first time, to our knowledge, in this study. Considering the limits of the metagenomic data, we have developed a framework and proposed good practices that will ensure robust and reliable results. The populations of *O. nana* in the MS seem to be structured between and within basins; however, the intrapopulation genetic differentiation could not be measured.

The loci under selection corresponded to alleles mostly present in the TARA_26 population, which also supports the genomic differentiation occurring in the population at this station. The relatively low amount of loci under selection detected can be partly due to our conservative approach. The presence of loci under selection in the gene coding a LNR domain protein with a sex-biased expression indicated that this gene might play an important role in the *O. nana* fitness in the populations of TARA_24 and 26 and thus would be a good candidate for functional genomic analyses.

The identification of genomic clines in the ocean has already been proposed along linear trajectories, and tools have been developed to work at single loci level (Derryberry, Derryberry, Maley, & Brumfield, 2014). Here, however, we proposed a new way to visualize population admixtures by modelling the variation in the BAF by the Lagrangian distance, which enables the modelling of BAF variation along nonlinear trajectories. This approach is more likely to follow genomic variation along gyres. It allowed the identification of two genomic clines located outside the Algerian Current. A question out of reach for our data set is estimating the temporal stability of these clines. Indeed, although the metagenomic data provided by the *Tara* Oceans expedition brought a snapshot of the population structure in the MS, a time series would greatly help to assess the temporal stability of the observations made, and the consequences for the species connectivity.

5 | CONCLUSIONS

The approaches proposed in this study are likely to be efficient on any zooplankton taxon having a tractable genome size and being abundant in samples used for metagenomic sequencing. The availability of the global *Tara* Oceans data set will allow such analyses to be performed on many important components of the zooplankton community, helping to reveal their distribution and functional properties. The availability of full genomes to study the population structure of zooplankton is a great advantage. In combination with metagenomic data and with the appropriate approach, it will allow for the identification of population structure and loci under selection on different organisms. This would accelerate our knowledge of zooplankton population genomics and provide a better understanding of the molecular mechanisms involved in the adaptation of species to environmental conditions and changes. Another advantage of this approach is the lack of a need to individually check each sample for the species of interest. This characteristic is even more critical when considering small, difficult to identify taxa, allowing the study of any species, and not only those for which identification is possible or easy.

ACKNOWLEDGEMENTS

We thank the people and sponsors who participated in the *Tara* Oceans Expedition 2009–2013: Centre National de la Recherche Scientifique, European Molecular Biology Laboratory, Génoscope/Commissariat à l'Energie Atomique, the French Government "Investissements d'Avenir" programmes OCEANOMICS (ANR-11-BTBR-0008), FRANCE GENOMIQUE (ANR-10-INBS-09-08), Agnès b., the Veolia Environment Foundation, Region Bretagne, World Courier, Illumina, Cap L'Orient, the Eléctricité de France (EDF) Foundation EDF Diversiterre, Fondation pour la Recherche sur la Biodiversité, the Prince Albert II de Monaco Foundation, Etienne Bourgois and the *Tara* schooner and its captain and crew. *Tara* Oceans would not exist without continuous support from 23 institutes (oceans.taraexpeditions.org). We thank Amy Maas for the manuscript improvement. We also thank Christian Sardet and Maria Grazia Mazzocchi, for help at the start of this project. This is *Tara* Oceans Contribution Number 56.

DATA ACCESSIBILITY

The metagenomic data from *Tara* Oceans are available at ENA (Appendix S18). The *Oithona nana* genome sequence and annotation are available at ENA with the study Accession no. PRJEB18938.

AUTHOR CONTRIBUTIONS

J.-L.J. and J.P. collected the samples. J.P. performed the molecular analyses. L.B.-B. provided the oceanographic data. M.-A.M., K.S. and M.W. performed the bioinformatics analyses. M.-A.M., L.B.B., A.C., L.S., S.G. and P.W. participated in the data interpretation. M.-A.M., L.B.B., A.C. and P.W. contributed to the manuscript. M.-A.M. and P.W. supervised the study.

REFERENCES

Andrews, K. R., Good, J. M., Miller, M. R., Luikart, G., & Hohenlohe, P. A. (2016). Harnessing the power of RADseq for ecological and evolutionary genomics. *Nature Reviews Genetics*, *17*, 81–92.

Apweiler, R., Bairoch, A., Wu, C. H., Barker, W. C., Boeckmann, B., Ferro, S., … Yeh, L. S. (2004). UniProt: The universal protein knowledgebase. *Nucleic Acids Research*, *32*, D115–D119.

Beaugrand, G., Reid, P. C., Ibanez, F., Lindley, J. A., & Edwards, M. (2002). Reorganization of North Atlantic marine copepod biodiversity and climate. *Science*, *296*, 1692–1694.

Berline, L., Rammou, A. M., Doglioli, A., Molcard, A., & Petrenko, A. (2014). A connectivity-based eco-regionalization method of the Mediterranean Sea. *PLoS One*, *9*, e111978.

Birney, E., & Durbin, R. (2000). Using Gene Wise in the *Drosophila* annotation experiment. *Genome Research*, *10*, 547–548.

Blanco-Bercial, L., Álvarez-Marqués, F., & Bucklin, A. (2011). Comparative phylogeography and connectivity of sibling species of the marine copepod *Clausocalanus* (Calanoida). *Journal of Experimental Marine Biology and Ecology*, *404*, 108–115.

Blanco-Bercial, L., & Bucklin, A. (2016). New view of population genetics of zooplankton: RAD-seq analysis reveals population structure of the North Atlantic planktonic copepod *Centropages typicus*. *Molecular Ecology*, *25*, 1566–1580.

Blanco-Bercial, L., Cornils, A., Copley, N., & Bucklin, A. (2014). DNA barcoding of marine copepods: Assessment of analytical approaches to species identification. *PLoS Currents*, *6*.

Bonhomme, M., Chevalet, C., Servin, B., Boitard, S., Abdallah, J., Blott, S., & Sancristobal, M. (2010). Detecting selection in population trees: The Lewontin and Krakauer test extended. *Genetics*, *186*(1), 241–262.

Castellani, C., Licandro, P., Fileman, E., di Capua, I., & Grazia Mazzocchi, M. (2015). Oithona similis likes it cool: Evidence from two long-term time series. *Journal of Plankton Research*, *38*, 703–717.

Cepeda, G. D., Blanco-Bercial, L., Bucklin, A., Beron, C. M., & Vinas, M. D. (2012). Molecular systematic of three species of *Oithona* (Copepoda, Cyclopoida) from the Atlantic Ocean: Comparative analysis using 28S rDNA. *PLoS One*, *7*, e35861.

Chikhi, R., & Medvedev, P. (2014). Informed and automated k-mer size selection for genome assembly. *Bioinformatics*, *30*, 31–37.

Cingolani, P., Platts, A., Wang le, L., Coon, M., Nguyen, T., Wang, L., … Ruden, D. M. (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 (Austin)*, *6*, 80–92.

Clark, R., Frid, C., & Batten, S. D. (2000). A critical comparison of two long-term zooplankton time series from the Central-west North Sea. *Journal of Plankton Research*, *3*, 27–39.

Clower, M. K., Holub, A. S., Smith, R. T., & Wyngaard, G. A. (2016). Embryonic development and a quantitative model of programmed DNA elimination in *Mesocyclops Edax* (S. A. Forbes, 1891) (Copepoda: Cyclopoida). *Journal of Crustacean Biology*, *36*, 661–674.

Colbourne, J. K., Pfrender, M. E., Gilbert, D., Thomas, W. K., Tucker, A., Oakley, T. H., … Boore, J. L. (2011). The ecoresponsive genome of *Daphnia pulex*. *Science*, *331*, 555–561.

Cornils, A., & Wend-Heckmann, B. (2015). First report of the planktonic copepod *Oithona davisae* in the northern Wadden Sea (North Sea): Evidence for recent invasion. *Helgoland Marine Research*, *69*, 243–248.

Cornils, A., Wend-Heckmann, B., & Held, C. (2017). Global phylogeography of *Oithona similis* s.l. (Crustacea, Copepoda, Oithonidae) – A cosmopolitan plankton species or a complex of cryptic lineages? *Molecular Phylogenetics and Evolution*, *107*, 473–485.

de Vargas, C., Audic, S., Henry, N., Decelle, J., Mahé, F., Logares, R., … Karsenti, E. (2015). Eukaryotic plankton diversity in the sunlit ocean. *Science*, *348*, 1–12.

Derryberry, E. P., Derryberry, G. E., Maley, J. M., & Brumfield, R. T. (2014). HZAR: Hybrid zone analysis using an R software package. *Molecular Ecology Resources*, *14*, 652–663.

Drouin, G. (2006). Chromatin diminution in the copepod *Mesocyclops edax*: Diminution of tandemly repeated DNA families from somatic cells. *Genome*, *49*, 657–665.

Dubarry, M., Noel, B., & Rukwavu, T., & Aury, J. M. (2016) *Gmove a tool for eukaryotic gene predictions using various evidences*. Retrieved from https://github.com/institut-de-genomique/gmove

Gallienne, C. P., & Robins, D. B. (2001). Is *Oithona* the most important copepod in the world's oceans? *Journal of Plankton Research*, *23*, 1421–1432.

Goetze, E., Andrews, K. R., Peijnenburg, K. T., Portner, E., & Norton, E. L. (2015). Temporal stability of genetic structure in a mesopelagic copepod. *PLoS One*, *10*, e0136087.

Grishanin, A. (2014). Chromatin diminution in Copepoda (Crustacea): Pattern, biological role and evolutionary aspects. *Comparative Cytogenetics*, *8*, 1–10.

Guindon, S., Dufayard, J. F., Lefort, V., Anisimova, M., Hordijk, W., & Gascuel, O. (2010). New algorithms and methods to estimate maximum-likelihood phylogenies: Assessing the performance of PhyML 3.0. *Systematic Biology*, *59*, 307–321.

Helaouët, P., & Beaugrand, G. (2009). Physiology, ecological niches and species distributions. *Ecosystems*, *12*, 1235–1245.

Hirai, J., Kuriyama, M., Ichikawa, T., Hidaka, K., & Tsuda, A. (2015). A metagenetic approach for revealing community structure of marine planktonic copepods. *Molecular Ecology Resources*, *15*, 68–80.

Jones, P., Binns, D., Chang, H. Y., Fraser, M., Li, W., McAnulla, C., … Hunter, S. (2014). InterProScan 5: Genome-scale protein function classification. *Bioinformatics*, *30*, 1236–1240.

Karsenti, E., Acinas, S. G., Bork, P., Bowler, C., De Vargas, C., Raes, J., & Tara Oceans Consortium Coordinators. (2011). A holistic approach to marine eco-systems biology. *PLoS Biology*, *9*, e1001177.

Katoh, K., & Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Molecular Biology and Evolution*, *30*, 772–780.

Kent, W. J. (2002). BLAT–the BLAST-like alignment tool. *Genome Research*, *12*, 656–664.

Kozol, R., Blanco-Bercial, L., & Bucklin, A. (2012). Multi-gene analysis reveals a lack of genetic divergence between *Calanus agulhensis* and *C. sinicus* (Copepoda; Calanoida). *PLoS One*, *7*, e45710.

Kumar, S., Stecher, G., & Tamura, K. (2016). MEGA7: Molecular evolutionary genetics analysis version 7.0 for bigger datasets. *Molecular Biology and Evolution*, *33*, 1870–1874.

Lewontin, R. C., & Krakauer, J. (1973). Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms. *Genetics*, *74*, 175–195.

Li, H., & Durbin, R. (2010). Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics*, *26*, 589–595.

Li, W., & Godzik, A. (2006). Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, *22*, 1658–1659.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., … 1000 Genome Project Data Processing Subgroup. (2009). The sequence alignment/map format and SAM tools. *Bioinformatics*, *25*, 2078–2079.

McLaren, I. A., Sevigny, J. M., & Corkett, C. J. (1988). Body sizes, development rates, and genome sizes among *Calanus* species. *Hydrobiologia*, *167–168*, 275–284.

McLaren, I. A., Sévigny, J. M., & Frost, B. W. (1989). Evolutionary and ecological significance of genome sizes in the copepod genus *Pseudocalanus*. *Canadian Journal of Zoology*, *67*, 565–569.

Morgulis, A., Gertz, E. M., Schaffer, A. A., & Agarwala, R. (2006). A fast and symmetric DUST implementation to mask low-complexity DNA sequences. *Journal of Computational Biology*, *13*, 1028–1040.

Mott, R. (1997). EST_GENOME: A program to align spliced DNA sequences to unspliced genomic DNA. *Computer Applications in the Biosciences*, *13*, 477–478.

Nishida, S. (1985). Taxonomy and distribution of the family Oithonidae (Copepoda, Cyclopoida) in the Pacific and Indian oceans. *Bulletin of the Ocean Research Institute-University of Tokyo*, *20*, 1–167.

Nishida, S., Omori, M., & Tanaka, O. (1977). Cyclopoid copepods of the family Oithonidae in Suruga Bay [Japan] and adjacent waters. *Bulletin of Plankton Society of Japan*, *24*, 119–158.

Parra, G., Bradnam, K., & Korf, I. (2007). CEGMA: A pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics*, *23*, 1061–1067.

Peijnenburg, K. T., & Goetze, E. (2013). High evolutionary potential of marine zooplankton. *Ecology and Evolution*, *3*, 2765–2781.

Pesant, S., Not, F., Picheral, M., Kandels-Lewis, S., Le Bescot, N., Gorsky, G., . . . The Tara Oceans Consortium Coordinators. (2015). Open science resources for the discovery and analysis of *Tara* Oceans data. *Scientific Data*, *2*, 150023.

Quinlan, A. R. (2014). BEDTools: The swiss-army tool for genome feature analysis. *Current Protocols in Bioinformatics*, *47*, 11–12.

Rasch, E. M., & Wyngaard, G. A. (2006). Genome sizes of cyclopoid copepods (Crustacea): Evidence of evolutionary constraint. *Biological Journal of the Linnean Society*, *87*, 625–635.

Razouls, C., de Bovée, F., Kouwenberg, J., & Desreumaux, N. (2016) *Diversity and Geographic Distribution of Marine Planktonic Copepods*. Retrieved from http://copepodes.obs-banyuls.fr/en

Safonova, Y., Bankevich, A., & Pevzner, P. A. (2015). dipSPAdes: Assembler for highly polymorphic diploid genomes. *Journal of Computational Biology*, *22*, 528–545.

Sahlin, K., Vezzi, F., Nystedt, B., Lundeberg, J., & Arvestad, L. (2014). BESST–efficient scaffolding of large fragmented assemblies. *BMC Bioinformatics*, *15*, 281.

Schulz, M. H., Zerbino, D. R., Vingron, M., & Birney, E. (2012). Oases: Robust de novo RNA-seq assembly across the dynamic range of expression levels. *Bioinformatics*, *28*, 1086–1092.

Sun, C., Wyngaard, G., Walton, D. B., Wichman, H. A., & Mueller, R. L. (2014). Billions of basepairs of recently expanded, repetitive sequences are eliminated from the somatic genome during copepod development. *BMC Genomics*, *15*, 186.

Temperoni, B., Viñas, M. D., Diovisalvi, N., & Negri, R. (2011). Seasonal production of *Oithona nana* Giesbrecht, 1893 (Copepoda: Cyclopoida) in temperate coastal waters off Argentina. *Journal of Plankton Research*, *33*, 729–740.

Turner, J. T. (2004). The importance of small *Planktonic* copepods and their roles in pelagic marine food webs. *Zoological studies*, *43*, 255–266.

Ueda, H., Yamaguchi, A., Saitoh, S., Orui Sakaguchi, S., & Tachihara, K. (2011). Speciation of two salinity-associated size forms of *Oithona dissimilis* (Copepoda: Cyclopoida) in estuaries. *Journal of Natural History*, *45*, 2029–2079.

Vannier, T., Leconte, J., Seeleuthner, Y., Mondy, S., Pelletier, E., Aury, J. M., . . . Jaillon, O. (2016). Survey of the green picoalga *Bathycoccus* genomes in the global ocean. *Scientific Reports*, *6*, 37900.

Ventura, T., & Sagi, A. (2012). The insulin-like androgenic gland hormone in crustaceans: From a single gene silencing to a wide array of sexual manipulation-based biotechnologies. *Biotechnology Advances*, *30*, 1543–1550.

Viñas, M. D., & Ramirez, F. C. (1996). Gut analysis of first-feeding anchovy larvae from the Patagonian spawning areas in relation to food availability. *Archive of Fishery and Marine Research*, *43*, 231–256.

Viñas, M. D., & Santos, B. A. (2000). First-feeding of hake (*Merluccius hubbsi*) larvae and prey availability in the North Patagonian spawning area – Comparison with anchovy. *Archive of Fishery and Marine Research*, *48*, 242–254.

Williams, J. A., & Muxagata, E. (2006). The seasonal abundance and production of *Oithona nana* (Copepoda:Cyclopoida) in Southampton Water. *Journal of Plankton Research*, *28*, 1055–1065.

Wyngaard, G. A., McLaren, I. A., White, M. M., & Sévigny, J.-M. (1995). Unusually high numbers of ribosomal RNA genes in copepods (Arthropoda: Crustacea) and their relationship to genome size. *Genome*, *38*, 97–104.

Wyngaard, G. A., & Rasch, E. M. (2000). Patterns of genome size in the Copepoda. *Hydrobiologia*, *417*, 43–56.

Wyngaard, G. A., Rasch, E. M., & Connelly, B. A. (2011). Unusual augmentation of germline genome size in *Cyclops kolensis* (Crustacea, Copepoda): Further evidence in support of a revised model of chromatin diminution. *Chromosome Research*, *19*, 911–923.

Zagoskin, M. V., Marshak, T. L., Mukha, D. V., & Grishanin, A. K. (2010). Chromatin diminution process regulates rRNA gene copy number in freshwater copepods. *Acta Naturae*, *2*, 52–57.

Zerbino, D. R., & Birney, E. (2008). Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. *Genome Research*, *18*, 821–829.

SUPPORTING INFORMATION

Additional Supporting Information may be found online in the supporting information tab for this article.

How to cite this article: Madoui M-A, Poulain J, Sugier K, et al. New insights into global biogeography, population structure and natural selection from the genome of the epipelagic copepod *Oithona*. *Mol Ecol*. 2017;00:1–16. https://doi.org/10.1111/mec.14214

View publication stats