

## ORIGINAL RESEARCH

# Investigating population-scale allelic differential expression in wild populations of *Oithona similis* (Cyclopoida, Claus, 1866)

Romuald Laso-Jadart<sup>1,2</sup>  | Kevin Sugier<sup>1</sup> | Emmanuelle Petit<sup>3</sup> | Karine Labadie<sup>3</sup> | Pierre Peterlongo<sup>4</sup> | Christophe Ambroise<sup>5</sup> | Patrick Wincker<sup>1,2</sup> | Jean-Louis Jamet<sup>6</sup> | Mohammed-Amin Madoui<sup>1,2</sup> 

<sup>1</sup>Génomique Métabolique, Genoscope, Institut François Jacob, CEA, CNRS, Univ Evry, Université Paris-Saclay, Evry, France

<sup>2</sup>Research Federation for the study of Global Ocean Systems Ecology and Evolution, FR2022/Tara Oceans GO-SEE, Paris, France

<sup>3</sup>CEA, Genoscope, Institut de Biologie François Jacob, Université Paris-Saclay, Evry, France

<sup>4</sup>CNRS, Inria, IRISA – UMR 6074, Univ Rennes, Rennes, France

<sup>5</sup>LaMME, CNRS, Université Paris-Saclay, Univ Evry, Evry, France

<sup>6</sup>Mediterranean Institute of Oceanology (MIO), AMU-UTLN UM110, CNRS UMR7294, IRD, UMR235, Equipe Ecologie Marine et Biodiversité (EMBIO), Université de Toulon, Toulon Cedex 9, France

## Correspondence

Romuald Laso-Jadart and Mohammed-Amin Madoui, Génomique Métabolique, Genoscope, Institut François Jacob, CEA, CNRS, Univ Evry, Université Paris-Saclay, Evry, France.  
Emails: rlasojad@genoscope.cns.fr; amadou@genoscope.cns.fr

## Funding information

Centre National de la Recherche Scientifique; European Molecular Biology Laboratory; Genoscope/Commissariat à l'Energie Atomique; French Government "Investissements d'Avenir" programmes OCEANOMICS, Grant/Award Number: ANR, -11-, BTBR and -0008; FRANCE GENOMIQUE, Grant/Award Number: ANR-10-INBS-09-08; Veolia Environment Foundation; Region Bretagne, World Courier, Illumina, Cap L'Orient, the Electricite de France (EDF) Foundation EDF Diversiterre; Fondation pour la Recherche sur la Biodiversité; Prince Albert II de Monaco Foundation; Etienne Bourgois and the Tara schooner and its captain and crew

## Abstract

Acclimation allowed by variation in gene or allele expression in natural populations is increasingly understood as a decisive mechanism, as much as adaptation, for species evolution. However, for small eukaryotic organisms, as species from zooplankton, classical methods face numerous challenges. Here, we propose the concept of allelic differential expression at the population-scale (psADE) to investigate the variation in allele expression in natural populations. We developed a novel approach to detect psADE based on metagenomic and metatranscriptomic data from environmental samples. This approach was applied on the widespread marine copepod, *Oithona similis*, by combining samples collected during the *Tara Oceans* expedition (2009–2013) and de novo transcriptome assemblies. Among a total of 25,768 single nucleotide variants (SNVs) of *O. similis*, 572 (2.2%) were affected by psADE in at least one population (FDR < 0.05). The distribution of SNVs under psADE in different populations is significantly shaped by population genomic differentiation (Pearson  $r = 0.87$ ,  $p = 5.6 \times 10^{-30}$ ), supporting a partial genetic control of psADE. Moreover, a significant amount of SNVs (0.6%) were under both selection and psADE ( $p < .05$ ), supporting the hypothesis that natural selection and psADE tends to impact common loci. Population-scale allelic differential expression offers new insights into the gene regulation control in populations and its link with natural selection.

## KEYWORDS

allelic expression, Arctic seas, copepod, metagenomics, metatranscriptomics, selection, structure, *Tara Oceans*, Zooplankton

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

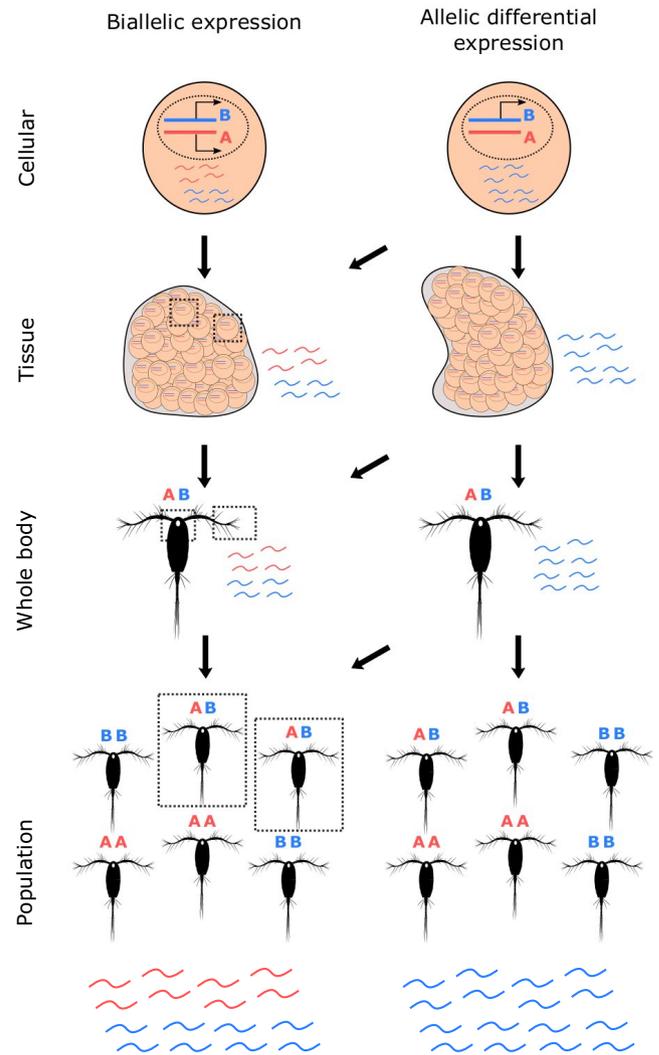
© 2020 The Authors. *Ecology and Evolution* published by John Wiley & Sons Ltd

## 1 | INTRODUCTION

Variation in gene expression within and between individuals or natural populations is an important mechanism for the acclimation of species (Fay & Wittkopp, 2008; Hutter, Saminadin-Peter, Stephan, & Parsch, 2008; Li, Liu, Kim, Min, & Zhang, 2010; Whitehead, 2012). This variation can nonexclusively be driven by selective genetic factors (Zhang et al. 2008; Fraser, 2013; Sato, Makino, & Kawata, 2016) or be induced by environmental cues and gradients, for varying periods of time (Passow et al. 2017; Brown, Arias-Rodriguez, Yee, Tobler, & Kelley, 2018). Particularly, marine species populations belonging to plankton face a large panel of physico-chemical changes in open ocean, (Guinder & Molinero, 2013; Pelejero, Calvo, & Hoegh-Guldberg, 2010) and gene expression variations have been observed in several studies (Lauritano, Procaccini, & Ianora, 2012; Salazar et al. 2019). However, taxonomic identification (Cepeda, Sabatini, Scioscia, Ramírez, & Viñas, 2016), DNA and mRNA extraction of small marine eukaryote species due to their complex genomes (Bucklin et al., 2018) still constitute an obstacle to conduct proper studies focusing on gene expression and selection in natural populations.

In the present study, we proposed to measure the population-scale allelic differential expression (psADE) of genes. psADE depends on the difference between alleles abundance at the genomic and transcriptomic level. At population scale, it aggregates differential expression between the two alleles at smaller scales (Figure 1). To measure psADE on small organisms, it would require the sequencing of several individuals at genomic and transcriptomic levels separately. An alternative approach could be to take advantage of the recent advances in metagenomic and metatranscriptomic sequencing of environmental samples, which offer a direct population-scale insight. In this context, polymorphic sites of a single species have to be extracted, allowing to evaluate whether the population-scale relative expression of an allele deviates from its genomic frequency.

Copepods, and particularly species belonging to the *Oithona* genus, are small crustaceans forming the most abundant metazoan on Earth (Gallienne, 2001; Humes, 1994; Kiørboe, 2011). This abundance, reflecting strong adaptive capacities to environmental fluctuations, together with large hypothetic effective population size (Peijnenburg & Goetze, 2013; Riginos, Crandall, Liggins, Bongaerts, & Trembl, 2016; Madoui et al. 2017; Arif et al. 2018) make this species suitable for population genomics analyses. In addition, they play an ecological key role in biogeochemical cycles and in the marine trophic food chain (Wassmann et al. 2006). In this study, we propose to detect psADE by focusing on the widespread epipelagic copepod, *Oithona similis* (Cyclopoida, Claus, 1866). We used environmental samples collected by the *Tara* Oceans expedition (Karsenti et al. 2011; Pesant et al. 2015) during its Arctic phase (2013) for which both metagenomic and metatranscriptomic data are available. Arctic Seas is an area where *O. similis* is known to be highly abundant (Blachowiak-Samolyk, Kwasniewski, Hop, & Falk-Petersen, 2008; Castellani et al., 2016; Dvoretzky, 2012; Zamora-Terol, Nielsen, & Saiz, 2013). First, variants of *O. similis* were extracted and the



**FIGURE 1** Allelic differential expression at population-scale. (a) Different scales of allele-specific expression detection for a heterozygous gene, from population to cellular levels. For a heterozygous genotype, ADE is understood as the difference in expression between two alleles of a single gene, opposed to strict biallelic expression. For clarity, the example of ADE presented here is monoallelic expression

genetic structure between the populations was studied. Then, we detected loci under psADE, under selection and under both psADE and selection. From these results, we tried to decipher the potential links between psADE, genomic differentiation and natural selection. Lastly, we investigated the molecular functions and biological processes of candidate loci under psADE and selection.

## 2 | MATERIALS AND METHODS

### 2.1 | Variant calling using *Tara* Oceans metagenomic and metatranscriptomic data

We used metagenomic and metatranscriptomic reads generated from samples of the size fraction 20–180  $\mu\text{m}$  collected in seven *Tara*

Oceans stations (TARA\_155, 158, 178, 206, 208, 209, and 210) (Figure 2a) according to protocols fully described by Alberti et al. (2017) (Table S1). Because of the lack of a reference genome, the reference-free variant caller *DiscoSNP++* (Uricaru et al. 2014; Peterlongo et al. 2017) was used to extract SNVs simultaneously from raw metagenomic and metatranscriptomic reads and was run using parameter  $-b 1$ . Only SNVs corresponding to biallelic loci with a minimum of 4x of depth of coverage in all stations were initially selected. Then, to capture loci belonging to *Oithona similis*, SNVs were clustered based on their loci co-abundance across samples using density-based clustering algorithm implemented in the R package *dbscan* (Ester, Kriegel, Sander, & Xu, 1996; Ram, Jalal, Jalal, & Kumar, 2010) and ran with parameters  $\epsilon = 10$  and  $\text{minPts} = 10$ . This generated three SNVs clusters, the largest of which contained 102,258 SNVs. To ensure only the presence of *O. similis* SNVs, we observed the fitting of the depth of coverage to the expected negative binomial distribution in each population (Figure S3). As the variant calling step is reference-free, the alternative allele (here, B-allele) is arbitrary set by *DiscoSNP++*. For each variant in each population, the B-allele frequency (BAF) and the population-level B-allele relative expression (BARE) were computed as follows,  $\text{BAF} = \frac{G_B}{G_B + G_A}$  and  $\text{BARE} = \frac{T_B}{T_B + T_A}$ , with  $G_A, G_B$  the metagenomic read counts supporting

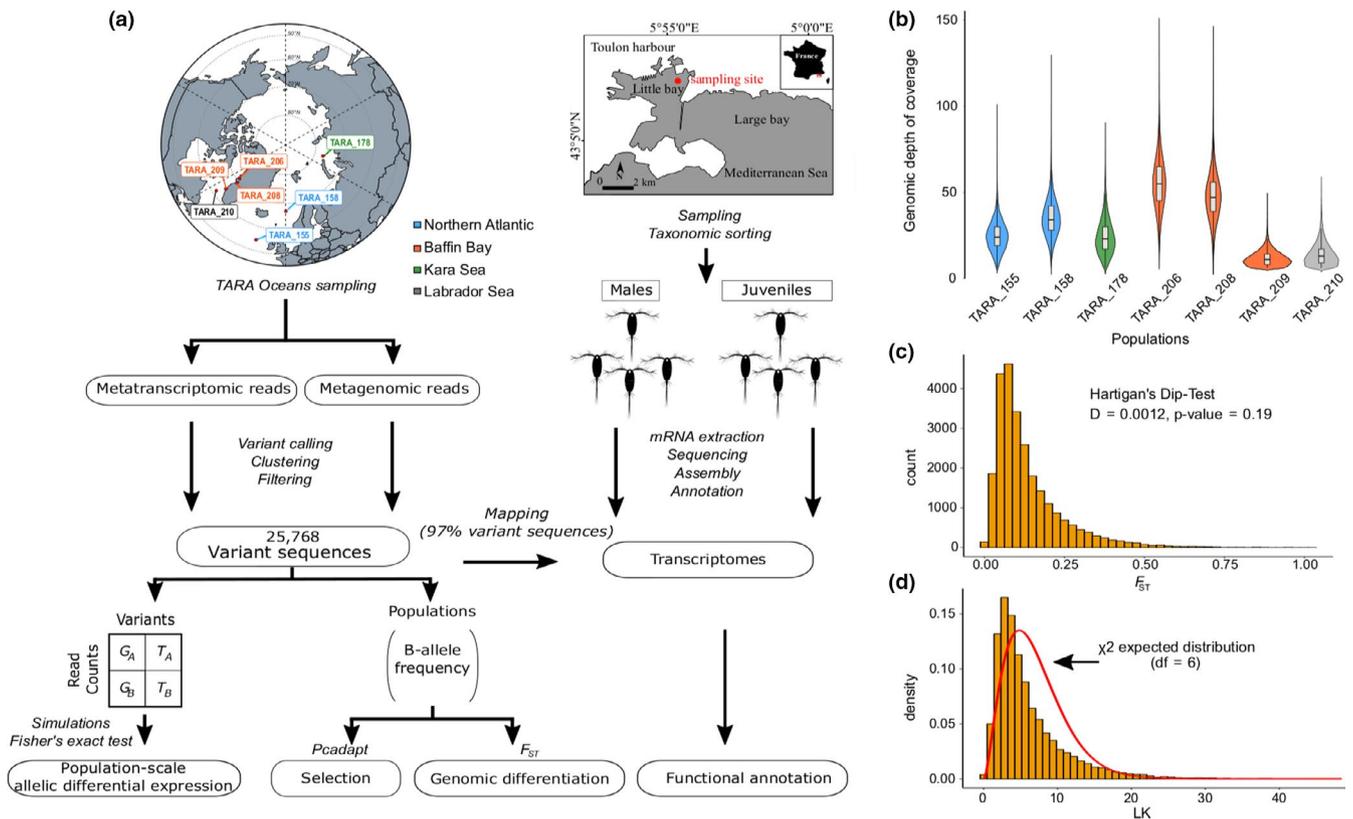
the reference and alternative alleles respectively and  $T_A, T_B$  the metatranscriptomic read counts supporting the reference and alternative alleles, respectively.

Biallelic loci were then filtered based on their metagenomic coverage. For each sample, the median and standard deviation  $\sigma$  of the distribution of metagenomic coverage of all biallelic loci were estimated. Biallelic loci must be characterized by a metagenomic coverage between a limit of  $\text{median} \pm 2\sigma$ , with a minimum and maximum of 5x and 150x coverage in each sample to avoid low covered and multicopy genomic regions. To keep out rare alleles and potential calling errors, only variants characterized by a BAF comprised between 0.9 and 0.1, and a BARE between 0.95 and 0.05 in at least one population were chosen for the final dataset resulting in 25,768 biallelic loci (Table S5).

To ensure that these loci belong to *O. similis*, the global  $F$ -statistics (or Wright's fixation index (Wright, 1951; Lewontin & Krakauer 1973) over the seven populations was computed as follows,

$$F_{ST} = \frac{\sigma^2}{\bar{p}(1-\bar{p})}$$

with  $\bar{p}$  and  $\sigma^2$  being the mean allele frequency and the related variance, and its distribution was tested for unimodality via a Hartigan's dip test (Hartigan & Hartigan, 1985). Moreover, LK statistics (Lewontin & Krakauer 1973) was computed as follows,  $LK = \frac{n-1}{F_{ST}}$  and compared with the expected chi-squared distribu-



**FIGURE 2** Genomic polymorphism data of *O. similis*. (a) Scheme representing the study, from samplings to analyses. (b) Genomic depth of coverage distributions of the set of 25,768 variants by sample. (c)  $F_{ST}$  distribution across the seven samples. (d) LK distribution. The red line represents chi-squared theoretical distribution ( $df = 6$ )

tion with  $df = n-1$ , with  $n$  being the number of populations and  $\overline{F_{ST}}$  being the mean  $F_{ST}$  across all loci.

## 2.2 | Population-scale ADE detection using metagenomic and metatranscriptomic data

In each population, we first selected heterozygous loci variants ( $BAF \neq \{0,1\}$ ). We tested the correlation between BAF and BARE and modeled their relationship by a linear regression. Then, we computed  $D = BAF - BARE$  and estimated the distribution parameters  $\mu$  and  $\sigma^2$  by fitting a normal distribution via *fitdist* function from *fitdistrplus* R package (Delignette-Muller & Dutang, 2015).

We then tested the psADE of each variant using a two-sided Fisher's exact test on a  $2 \times 2$  table containing the read counts  $G_A$ ,  $G_B$ ,  $T_A$ , and  $T_B$ . Given the large number of tests, we applied the Benjamini and Hochberg  $p$ -value correction (Benjamini & Hochberg, 1995) to control the False Discovery Rate (FDR). This generated seven sets of candidate loci under psADE, one set for each population.

## 2.3 | Noise detection in population-scale ADE using simulated data

To account for noise originating from potential sampling bias during sequencing, simulations were performed by generating sets of variants: (a) We modeled the distributions of the genomic depth of coverage of the loci (i.e., the sum of  $G_A$  and  $G_B$ ) from each of the seven samples separately (Table S3) by a negative binomial (NB) distribution (Robinson & Smyth, 2007) and estimated seven  $\mu$  and  $\theta$  (the NB mean and shape parameters); (b) The relationship among the seven samples between the observed  $\mu$  and  $\theta$  by a linear regression (Figure S4), allowing us to estimate a shape parameter  $\theta$  for any given mean  $\mu$ ; (c) A-allele frequencies followed a U-shaped distribution, approximated by a beta distribution of shape parameters  $\alpha$  and  $\beta$ ; (d) The expression level (i.e., the sum of  $T_A$  and  $T_B$ ) was modeled by a gamma distribution of shape and rate parameters  $a$  and  $b$ . These estimations were performed with *fitdist* function of *fitdistrplus* R package (Delignette-Muller & Dutang, 2015) and are presented in Figure S3.

To simulate the genomic A-allele frequency for one biallelic loci in a given population, we extracted one random deviates from its beta distribution of shape parameters  $\alpha$ ,  $\beta$  estimated previously for each population in (c). The A-allele frequency was multiplied by the estimated  $\mu$  to obtain an expected genomic read count of allele A,  $\overline{G}_A$ . In the same way,  $\overline{G}_B$  was obtained by the difference between  $\mu_n$  and  $\overline{G}_A$ . Simulated genomic read counts for allele A and B were then obtained by generating a random value from negative binomial distributions of parameters  $\mu_A = \overline{G}_A$  and  $\mu_B = \overline{G}_B$  respectively, and the corresponding size parameters  $\theta_A$  and  $\theta_B$  estimated using the linear regression between  $\theta$  and  $\mu$  as described above in (b).

A locus expression level was then computed by generating random deviates following a gamma distribution of shape  $a$  and rate  $b$

estimated previously for each population in (d). Under the null hypothesis, the allele abundance at genomic and transcriptomic level is the same (see Table S3), the expected transcriptomic read count of allele A,  $\overline{T}_A$ , was generated by multiplying the genomic A-allele frequency previously computed and the locus expression level. In the same way,  $\overline{T}_B$  was obtained by the difference between the locus level expression and  $\overline{T}_A$ . Finally, simulated transcriptomic read count  $T_A$  and  $T_B$  were obtained by generated random values from a Poisson distribution of parameter  $\lambda_A = \overline{T}_A$  and  $\lambda_B = \overline{T}_B$ , respectively. All simulations were performed using *lm*, *rbeta*, *rbinom*, *rgamma*, and *rpois* R functions.

To formally test for psADE due to noise (i.e., under the null hypothesis), seven sets of 50,000 loci were simulated using the seven sets of parameters learnt for each sample. Fisher's exact test was performed only on heterozygous loci with a non-null expression level, and simulated variants with a  $q$ -value  $< 0.05$  (Benjamini-Hochberg correction) were considered as noisy psADE, as described above.

## 2.4 | Estimation of genomic differentiation and detection of variants under selection

Pairwise- $F_{ST}$  was computed as follows,  $F_{ST} = \frac{\sigma_i^2}{p_i(1-p_i)}$ , for each locus

between each pair of populations  $i$  and the median pairwise- $F_{ST}$  was retained to measure the genomic differentiation between each population. A Mantel test was performed to test for isolation-by-distance between median pairwise- $F_{ST}$  and geographic Euclidean distances using *vegan* v2.5-2 (Oksanen et al. 2018) and *geosphere* v1.5-7 (Hijmans, 2017) R packages. The *pcadapt* R package v4.0.2 (Luu, Bazin, & Blum, 2017) was used to detect selection among populations from the B-allele frequency matrix. The computation was run on "Pool-seq" mode, with a minimum allele frequency of 0.05 across the populations, and variants with a corrected Benjamini and Hochberg  $p$ -value  $< .05$  were considered under selection.

## 2.5 | Modeling psADE with population differentiation

The seven sets of candidate loci under psADE were crossed to identify variants under psADE in several populations (named "shared psADEs") and all nonempty, nonoverlapping crossings between populations were represented by an upset plot.

To test whether populations characterized by weak genetic differentiation tend to share more loci under psADE than genetically distant populations, we modeled the number of shared psADEs between populations by genomic differentiation using a nonlinear model:  $y = a e^{-bx} + c$ , with  $y$  being the number of shared ADE and  $x$  the genomic differentiation. The latter was estimated by computing the median  $F_{ST}$  of each nonempty crossing set of populations as described above, with  $i$  being here the considered set of populations. To find the starting values, the model was linearized as follows,

$\log(y-c_0) \approx \log(a) + bx$ , with  $c_0 = \min(y)^{0.5}$  and  $a$  and  $b$  parameters were estimated via the *lm* R function. The nonlinear model was then applied, and least squares estimates were used via the *nls* R function. Pearson's correlation between the fitted and empirical values was then computed via the *cor.test* R function.

## 2.6 | psADE and link with natural selection

To identify alleles under both psADE and natural selection, the set of variants under psADE in each population was crossed with the set of loci detected under selection. The size of the intersection was tested by a hypergeometric test,  $H(q, m, n, k)$ , with  $q$  being number of alleles under psADE in the population and under selection (size of intersection),  $m$  being the total number of alleles under selection,  $n$  being the total number of variants under neutral evolution, and  $k$  being the total number of alleles under psADE in the tested population. We considered that, in a given population, the number of alleles under both psADE and selection was significantly higher than expected by chance for  $p$ -value  $< .05$ .

## 2.7 | Material sampling, mRNA extraction, and Mediterranean *O. similis* transcriptomes sequencing

To conduct a functional analysis, Mediterranean *O. similis* transcriptomes were produced. *Oithona similis* specimens were sampled at the North of the Large Bay of Toulon, France (Lat 43°06' 02.3" N and Long 05°56' 53.4" E). Sampling took place in November 2016. The samples were collected from the upper water layers (0–10 m) using zooplankton nets with a mesh of 90  $\mu$ m and 200  $\mu$ m (0.5 m diameter and 2.5 m length). Samples were preserved in 70% ethanol and stored at  $-4^{\circ}\text{C}$ . From the Large Bay of Toulon samples, *O. similis* individuals were isolated under the stereomicroscope (Nishida, 1985; Rose, 1933). We selected two different development stages: four copepodites (juveniles) and four adult males. Each individual was transferred separately and crushed, with a tissue grinder (Axygen) into a 1.5 ml tube (Eppendorf). Total mRNAs were extracted using the 'RNA isolation' protocol from NucleoSpin RNA XS kit (Macherey-Nagel) and quantified on a Qubit 2.0 with a RNA HS Assay kit (Invitrogen) and on a Bioanalyzer 2100 with a RNA 6000 Pico Assay kit (Agilent). cDNA was constructed using the SMARTer-Seq v4 Ultra low Input RNA kit (ClonTech). The libraries were built using the NEBNext Ultra II kit for paired-end sequencing with an Illumina HiSeq2500. After adaptors trimming, only reads with a mean Phred score  $> 20$  were kept.

## 2.8 | Transcriptomes assembly and annotation

Each read set was assembled with Trinity v2.5.1 (Haas et al. 2013) using default parameters and transcripts were clustered using cd-hit v4.6.6 (Fu, Niu, Zhu, Wu, & Li, 2012) using  $-c 0.9 -a 0.8 -aL 0.8$  parameters (Table S2). To ensure the classification of the sampled individuals, each ribosomal read

set were detected with SortMeRNA (Kopylova, Noé, & Touzet, 2012) and mapped with bwa v0.7.15 using default parameters (Li & Durbin, 2009) to 82 ribosomal 28S sequences of *Oithona* species used in Cornils, Wend-Heckmann, & Held, 2017 (Figure S2). The transcriptome assemblies were annotated with Transdecoder v5.1.0 (Haas et al. 2013) using default settings to predict the open reading frames (ORFs) and protein sequences (Table S2). In parallel, homology searches were also included as ORF retention criteria for Transdecoder; the peptide sequences of the longest ORFs were aligned on *Oithona nana* proteome (Madoui et al. 2017) using DIAMOND v0.9.22 (Buchfink, Xie, & Huson, 2014). Protein domain annotation was performed on the final ORF predictions with Interproscan v5.17.56.0 (Jones et al. 2014) and a threshold of  $E$ -value  $< 10^{-5}$  was applied for Pfam annotations. Finally, homology searches of the predicted proteins were done against the nonredundant NCBI protein database, restricted to Arthropoda (taxid: 6656), with DIAMOND v0.9.22.

## 2.9 | Variant functional annotation

The variant functional annotation was conducted in two steps. First, the variant sequences were mapped on the previously annotated *O. similis* transcripts using the "VCF\_creator.sh" program of DiscoSNP++. Secondly, a variant annotation was carried out with SNPEff (Cingolani et al. 2012) to identify the location of variants within transcripts (i.e., exon or UTR) and to estimate their effect on the proteins (missense, synonymous or nonsense). The excess of candidate variant annotations was tested for the different classes of SNPEff, for the three sets of variants under selection, psADE, and both, by comparing to the total sets of annotated variants. A significant excess was considered for a hypergeometric test  $q$ -value  $< 0.05$ , after Benjamini-Hochberg correction.

## 2.10 | Gene enrichment analysis

To identify putative biological function or processes associated to the variants, a domain-based analysis was conducted. The Pfam annotation of the transcripts carrying variants categorized as psADE and selection was used and compared to Pfam annotation of the total sets of variants with a hypergeometric test for enrichment. A significant excess was declared for a  $q$ -value  $< 0.05$  (Benjamini-Hochberg correction). To complete the domain-based analysis, the functional annotations obtained from the homology searches against the nr were manually curated.

## 3 | RESULTS

### 3.1 | Extracting polar *Oithona similis* variants from environmental samples

From metagenomic and metatranscriptomic raw data of seven sampling stations (Figure 2a), we identified 102,258 variants using a

reference-free approach. Among them, 25,768 expressed *O. similis* variants were retrieved after filtering. To ensure that the variants belonged to *O. similis*, we performed three different analyses. First, in each sample, the distributions of variable loci depth of coverage were unimodal (Figure 2b) and fitted the expected negative binomial distributions (Figure S3). Secondly, 97% of 25,768 variants were mapped on Mediterranean *O. similis* transcriptomes (Figure 2a). Third, the global distribution of  $F_{ST}$  of the seven populations was unimodal (Hartigans' dip test,  $D = 0.0012$ ,  $p$ -value = .19) with a low median  $F_{ST}$  at 0.1 (Figure 2c), confirmed by the pairwise- $F_{ST}$  distributions (Figure S6d). Finally, the LK distribution over all the loci followed the expected chi-squared distribution (Figure 2d), showing that most of the loci follow the neutral evolution model, as expected in a single species.

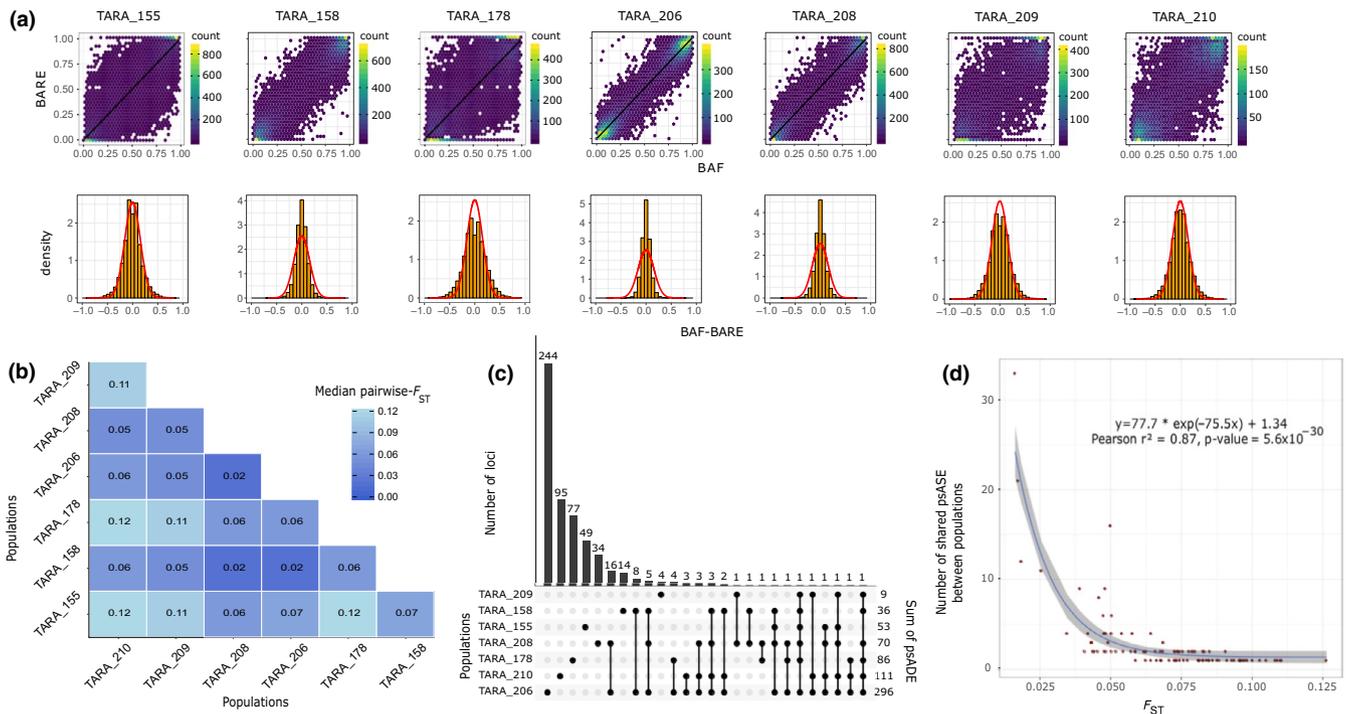
### 3.2 | *Oithona similis* genomic differentiation in Arctic Seas

The seven populations were globally characterized by a weak to moderate differentiation, with a maximum median pairwise- $F_{ST}$  of 0.12 between populations from TARA\_210 and 155/178 (Figure 3b, Figure S6d). Populations from stations TARA\_158 (Norway Current), 206 and 208 (Baffin Bay) were genetically closely related, with the lowest median pairwise- $F_{ST}$  (0.02), despite TARA\_158 did not

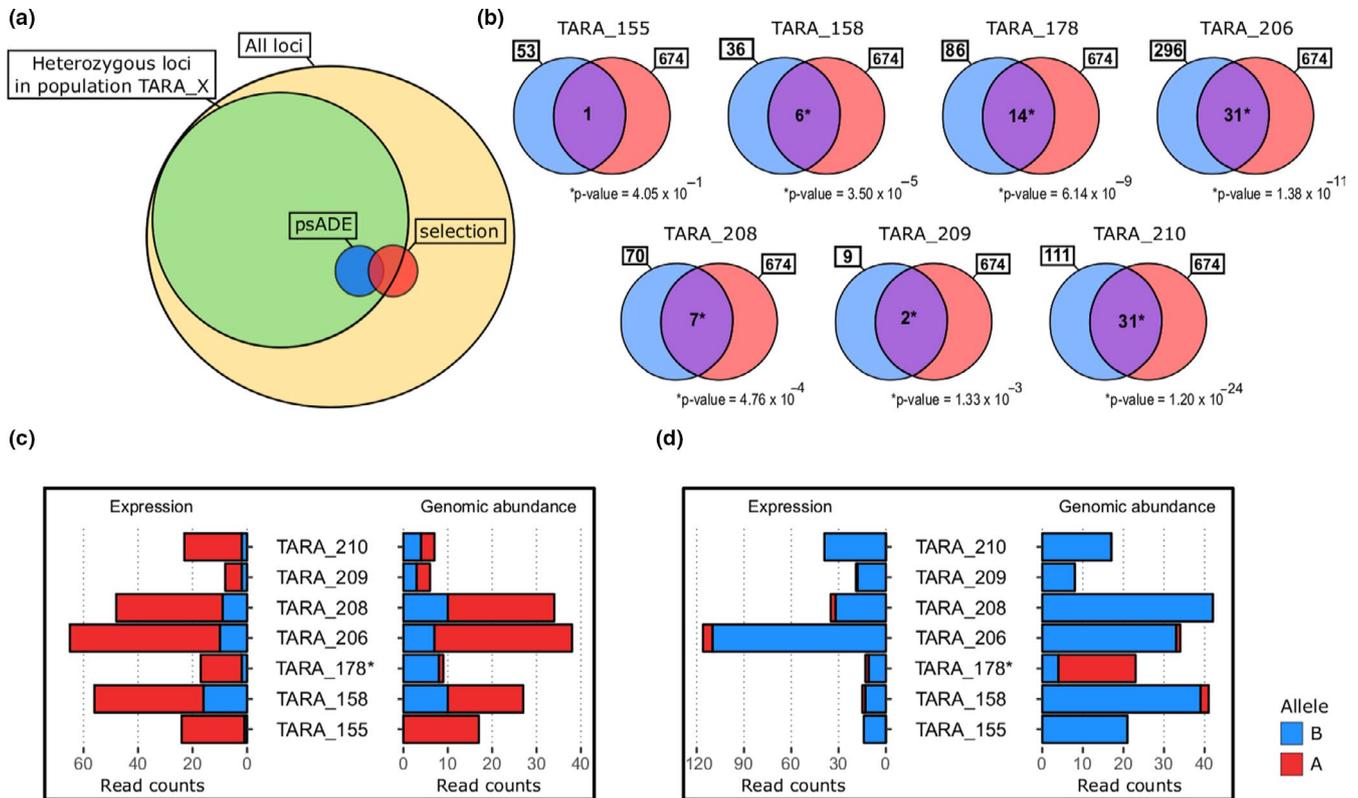
co-geolocalize with the two other stations. The four other populations (TARA\_155, 178, 209, and 210) were equally distant from each other (0.1–0.12). Finally, TARA\_158, 206, and 208 on one side and TARA\_155, 178, 210, and 209 on the other side showed the same pattern of differentiation (0.05–0.07). A Mantel test was performed and revealed no correlation between  $F_{ST}$  and geographic distances ( $r = 0.34$ ,  $p$ -value = .13; Figure S7).

### 3.3 | Detection of population-scale ADE

As expected, most of the loci presented a strong correlation between B-allele frequency and B-allele relative expression (Figure 3a, Table S3). Thus, we observed the difference between BAF and BARE, which followed a Gaussian distribution centered on 0 in each population (Figure 3a, Table S3). The number of SNVs tested for psADE varied between 13,454 and 22,578 for TARA\_210 and 206 respectively. We found a significant amount of variants under psADE in each population under a Fisher's exact test (Figure 3c, Table S4). Potential noise due to sample bias during sequencing was estimated by simulations for each of the seven population and was relatively negligible compared with real data. Distributions of simulated  $p$ -values under null hypothesis (i.e., no psADE) and *O. similis* empirical  $p$ -values (Figure S5) show higher amount of significant *O. similis*



**FIGURE 3** Population-scale allelic differential expression and its link with genomic differentiation. a, Each column corresponds to a population. Upper panel represents the relation between BAF and BARE, each hexagone corresponds to an area containing the number of variants indicated by the color scale. Black lines are the linear regression curves. Lower panel represents the distribution of BAF-BARE. The red lines correspond to the Gaussian distribution estimated from the data. b, Pairwise  $F_{ST}$  matrix. The median (mean) of each pairwise- $F_{ST}$  distribution computed on allele frequencies is indicated c, Upset plot of psADE detection in the seven populations. Each bar of the upper plot corresponds to the number of variants under psADE in the combination of population(s) indicated by black dots in the lower plot. d, Genomic differentiation and shared psADE. Each dot is a combination of population as presented in the lower panel of the upset plot. The blue line represents the nonlinear regression curve estimated from the data and 95% confidence interval in gray



**FIGURE 4** Crossing psADE and selection. (a) Scheme representing the method; in yellow the total dataset of SNVs, in green the tested SNVs in one population, in blue SNVs under psADE in this population, in red the SNVs targeted by selection. (b) Crossing the candidate variants under psADE (blue) and those under selection (red) for each population. The hypergeometric *p*-value corresponds to the significance of the amount of variants both under psADE in the considered population and under selection (purple). (c, d) Genomic abundance and expression profiles of variants 20,286,969 and 1,522,691. The first one is under psADE in TARA\_178 for allele A (\*), which is fixed in TARA\_155. The second one is under psADE in favor of allele B in TARA\_178 (\*) and reaching near fixation for the same allele in all the other populations

*p*-values compared with the simulated ones, resulting in a proportion of true-positives psADEs varying from 70% to 100% (Table S4).

Overall, we found 572 variants under psADE, including 513 population-specific psADEs, and 59 psADEs shared by several populations (Figure 3c). Remarkably, 29 psADEs out of the 59 were present only in the populations from TARA\_158, 206, and 208 that correspond to the genetically closest populations, leading us to compare the relationship between sharing psADEs and population differentiation. By comparing the number of shared psADEs in the different sets of populations to their genomic differentiation, we found a negative trend between the two (with a strong negative exponential slope estimate), illustrated by a significant correlation between nonlinear fitted and empirical values ( $0.87$ , *p*-value  $5.6 \times 10^{-30}$ , Figure 3d). This modeling shows that genetically close populations tend to share more variants under psADE.

### 3.4 | Loci under population-level ADE and selection in Arctic populations

The set of variants was tested for selection using *pcadapt*. The PCA decomposed the genomic variability in six components

(Figure S6a–c); the first two components discriminated TARA\_155 and 178 from the others (32% and 28.1% variance explained, respectively), and the third component differentiated TARA\_210 and 209 (19.5%). The fourth principal component separated TARA\_209 and 210 from 158/206/208 (11.3%), with the last two concerning TARA\_158/206/208. Globally, these results dovetailed with the  $F_{ST}$  analysis, with details discussed later. Finally, we detected 674 variants under selection, representing 2.6% of the dataset (corrected *p*-value  $< .05$ ).

The seven sets of variants under psADE were crossed with the set of variants under selection (Figure 4a). The size of the intersections ranged from 1 to 31 variants (TARA\_155 and 206/210) and was significantly higher than expected by chance for all the populations but TARA\_155 (Figure 4b, hypergeometric test *p*-value  $< .05$ ). It represented a total of 84 unique variants under selection and psADE in at least one population, corresponding to 15% and 12% of variants under psADE and under selection, respectively. Two main different genomic and expression profiles can be observed (Figure S8). First, as illustrated in Figure 4c, loci can show an allele under psADE but not fixed in one population (here allele A in TARA\_178) and fixed in another (TARA\_155). A second observed pattern, more extreme, concerns variants as exemplified in Figure 4d, which presents an allele

avored by psADE (allele B in TARA\_178), with a low genomic frequency and fixed in nearly all the other populations.

### 3.5 | Functional analysis of transcripts under population-scale ADE and selection

The full dataset of variants was positioned on the eight transcriptomes to extract putative functional information, with a total of 25,048 variants (97% of total) successfully mapped on 16,272 transcripts. First, SNPEff was used to estimate the localization of variants inside the transcripts, and an enrichment was estimated for all categories, and for three sets of variants categorized respectively as under selection, psADE and both (Figure S9). Overall, the two first sets showed a significant excess of variants in 3' UTR. Plus, among the variants under psADE, one was categorized as a "stop gained" and one as a "stop retained". However, variants under psADE and selection did not show any excess of specific effect.

Among the 84 loci identified under psADE and selection, 80 were located on *O. similis* transcripts (Table S6). Amid these transcripts, 64 (76%) were linked to at least one Pfam domain (61 different domains) and 59 (70%) to a functional annotation from the nr database. From this total of 61 Pfam domains, 23 presented a significant excess compared with domains present in the global set of transcripts, corresponding to 21 transcripts (Figure S10). On the latter, three transcripts were involved in nervous system features: omega-amidase NIT2, vang-like 2B protein, and 5-oxoprolinase.

## 4 | DISCUSSION

### 4.1 | Genomic and transcriptomic variation data belong to a single *Oithona similis* lineage

Because genomes of small animals like copepods are difficult to reconstruct, we used *DiscoSNP++*, a reference-free variant caller to extract variants from metagenomic data, that already showed its accuracy on *Tara* Oceans metagenomic data (Arif et al. 2018).

Global populations of *O. similis* are known to be composed of cryptic lineages across oceanic basins (Cornils et al., 2017). It is also known that this species is highly abundant among other copepods in Arctic Seas (Blachowiak-Samolyk et al., 2008; Castellani et al., 2016; Dvoretzky, 2012; Zamora-Terol et al., 2013). Thus, the assessment that the extracted variants from the seven samples used in our study belongs to the same *O. similis* cryptic lineage was a prerequisite for further analyses. Three different analyses support this assumption. First, the distribution of depth of coverage in each of the seven samples followed the expected negative binomial distribution (Supplementary Figure S3). Indeed, the possibility to observe these patterns in the presence of different species would require them to be equally co-abundant, which is unlikely. Thus, this covariation of the depth of coverage of these variants supports the single species genome origin. Secondly, the high

proportion of variants (97%) mapped on the Mediterranean *O. similis* transcriptomes, another cryptic lineage (Cornils et al., 2017), showed that the variant clustering method was efficient to re-group loci of *O. similis*. Finally, the unimodal distribution of  $F_{ST}$  showed that these populations of *O. similis* belong to the same polar cryptic species, and that most of the loci are under neutral evolution. Altogether, these results show that we were able to retrieve polymorphic data of a single species, *O. similis*, on which population differentiation analyses and psADE detection can be undertaken.

### 4.2 | *Oithona similis* populations are weakly differentiated within the Arctic Seas

We observed that the seven populations examined showed low genomic differentiation, despite the large distances separating them, which was illustrated by a nonsignificant Mantel test for isolation-by-distance (Figure S7).  $F_{ST}$  and *pcadapt* analyses both showed the same patterns of genomic differentiation. First, the differentiation of populations from TARA\_155 and 178 is relatively high compared with the others. Secondly, the geographically close populations from TARA\_210 and 209 present a relatively high differentiation (median pairwise- $F_{ST}$  of 0.11, PC3). This could be explained by the West Greenland current acting as a physical barrier between the populations, which could lead to reduced gene flow (Myers, Donnelly, & Ribergaard, 2008). At last, the strong link between TARA\_158 from Northern Atlantic current and TARA\_206/208 from the Baffin Bay is the most intriguing. Despite the large distances that separate the first one from the others, these three populations are well connected. Based on this weak structure and that most of loci follows a neutral evolution (Figure 2d), outliers detected by *pcadapt* probably are truly under selection and not due to specific population differentiation.

Metagenomic data enable to draw the silhouette of the gene flow between populations but lacks resolution when dealing with intrapopulation structure. However, our findings are concordant with previous studies underpinning the large-scale dispersal, interconnectivity of marine zooplankton populations in other oceans, at diverse degrees (Blanco-Bercial & Bucklin, 2016; Goetze, 2005; Höring, Cornils, Auel, Bode, & Held, 2017; Peijnenburg & Goetze, 2013). Weak genetic structure in the polar region was highlighted for other major Arctic copepods like *Calanus glacialis* (Weydmann, Coelho, Serrão, Burzyński, & Pearson, 2016) and *Pseudocalanus* species (Aarbakke, Bucklin, Halsband, & Norrbin, 2014). The absence of structure was explained by ancient diminutions of effective population size due to past glaciations (Aarbakke et al., 2014; Bucklin & Wiebe, 1998; Edmands, 2001), or high dispersal and connectivity between the present-day populations due to marine currents (Weydmann et al., 2016). Using Lagrangian travel time or dispersal probabilities could help to estimate how much marine currents explain this observed genomic differentiation.

### 4.3 | Population-scale ADE in *O. similis* populations and its link with differentiation and selection

We were able to detect variants under psADE in the seven populations. First, allele frequency and relative expression are strongly correlated in the data, showing as expected that the more an allele is observed at the genomic level, the more this allele is expressed. Simulations performed showed that although this sequencing bias noise is present in our data, it does not significantly affect psADE detection. Among the variants under psADE, a large part was population-specific and a minority was under psADE in several populations. From the latter, we showed that closely related populations tended to share more variants under psADE than other more differentiated populations, meaning that a genetic basis partially shapes these psADEs. This result shows that psADE allows the observation of similar allelic expression variation in individuals sampled from geographically very distant populations (>4,000 km for TARA\_206 and 158) but having very strong genetic relatedness. This finding clearly suggests that these variations of expression are probably controlled by the same regulatory genetic drivers.

We estimated the number of loci under psADE that were also targeted by natural selection. A significant amount of SNVs (84) was subject to selection among the seven populations and to psADE in at least one population, meaning that in Arctic populations of *O. similis*, psADE and natural selection affect the same genomic regions.

Two main patterns can be observed in the candidate loci. The first one (as in Figure 4c) shows one allele under psADE and selection in different populations (here allele A in TARA\_178 and TARA\_155 respectively). A second observed pattern concerns loci as exemplified in Figure 4d (see also Figure S8), where an allele (here B) is nearly fixed in six populations and under psADE in one population (here TARA\_178), with this psADE due to a low abundance and high expression. In this population, the allele B sees its frequency decreasing because another allele appears in this population. However, since the latter is under-expressed, it could mean that it is a deleterious mutation, and strong regulatory elements or molecular mechanisms repress its expression, or that even a small expression enables a higher fitness for individuals carrying it, or that the allele favored by psADE is the one enabling higher fitness, leading to fixation in other populations. Ultimately, although determining how psADE and selection interact remains beyond the scope of this study, we can hypothesize from these observations that the action of the two mechanisms on a locus can be (a) independent, psADE and selection acting separately, (b) sequential, with psADE acting before, while or after selection occurs.

The process of acclimation through gene expression and the link with genetic variation and adaptation have been studied widely in several organisms (Fay & Wittkopp, 2008; Signor & Nuzhdin, 2018; Williams, Chan, Cowley, & Little, 2007). In a first study in human, a link has been established between gene expression and selection, affecting particular genes and phenotypes, looking at *cis*-acting SNPs (Fraser, 2013). In a second study, the team was able to detect loci under ADE and selection at the same time in different human populations (Tian et al. 2018). Also, approaches in a plant model, *Capsella*

*grandiflora*, a species characterized by weak population structure and large effective population size, emphasized the relative impact of purifying selection and positive selection on *cis*-regulatory variation in populations (Josephs, Lee, Stinchcombe, & Wright, 2015; Steige, Laenen, Reimegård, Scofield, & Slotte, 2017). Our study, by focusing on whole populations of numerous individuals, offers further clues to understand the link between gene expression variation and selection.

Further investigations including replicates, more populations, and the production of a genome and genotypes would help to confirm our results, disentangle the different causes of psADE, and question the link between psADE and selection.

### 4.4 | Functional insights of natural selection and population-scale ADE in *Oithona similis*

By analyzing the functional annotation of the variants with SNPEff, we found a significant excess of variants under selection and variants under psADE located in 3'UTR, but no excess in variants under selection and psADE. Variations in these regions are known to both affect transcription efficiency through mRNA secondary structures, stability, and location (Matoulkova, Michalova, Vojtesek, & Hrstka, 2012; Mignone, Gissi, Liuni, & Pesole, 2002), leading to affect the function of the gene. However, no clear pattern was observed among the candidate variants.

On the 674 loci under selection, some variants were located in transcripts annotated by homology search (Table S6) as pantothenate kinase, glycine receptors/GABA receptors, and FMRFamide receptor. The same genes are also under selection in *Oithona nana* populations of Mediterranean (Madoui et al. 2017), stressing their evolutive importance. To date, variants located in transcripts linked to FMRFamide, glycine, and GABA receptors are also found among the 572 variants under psADE, but not in the 84 candidate variants under psADE and selection.

Glutamate, GABA and glycine are known neurotransmitters in arthropods motor neurons (Smarandache-Wellmann, 2016). Pantothenate kinase, an enzyme catalyzing the phosphorylation of vitamin B5, constituting the first step in Coenzyme-A biosynthesis pathway, is linked to neurodegenerative diseases in human and *Drosophila* (Pandey, Turm, Bekenstein, Shifman, & Kadener, 2013). Among the transcripts under psADE and selection, three are of specific interests, as they are also involved in nervous system. Omegaamidase NIT2 is an enzyme that produces  $\alpha$ -ketoglutarate, a precursor of glutamate and GABA. The 5-oxoprolinase produces glutamate from 5-oxoproline. Finally, vang-like protein 2B is involved in the formation of ommatidies in *Drosophila* (Leung et al. 2016). Other studies focusing on axon myelination in calanoid species illustrate how nervous system can play an important role in copepod evolution (Lenz, 2012; Weatherby, Davis, Hartline, & Lenz, 2000). From our results, more functional analyses would allow a better characterization of these genes, but it reveals the potential evolutionary importance of nervous system in copepods.

## 5 | CONCLUSION

Gene expression variation is thought to play a crucial role in acclimation and adaptive history of natural populations. Herein, we integrated metagenomic and metatranscriptomic data to detect ADE at the population level in populations of copepods. Then, we demonstrated the link between psADE and population differentiation on one hand and with natural selection on the other hand, by providing a quantitative observation of this phenomenon and its impact on specific biological features of copepods. In the future, we will try to expand these observations to other organisms and question the nature of the link between psADE and natural selection.

## ACKNOWLEDGMENTS

This work was supported by sponsors who participated in the Tara Oceans Expedition 2009–2013: Centre National de la Recherche Scientifique, European Molecular Biology Laboratory, Genoscope/Commissariat à l'Energie Atomique, the French Government "Investissements d'Avenir" programmes OCEANOMICS (ANR-11-BTBR-0008), FRANCE GENOMIQUE (ANR-10-INBS-09-08), Agnes b., the Veolia Environment Foundation, Region Bretagne, World Courier, Illumina, Cap L'Orient, the Electricite de France (EDF) Foundation EDF Diversiterre, Fondation pour la Recherche sur la Biodiversite, 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.tara-expeditions.org). We acknowledge the support of Vincent Segura and Leila Tirichine for fruitful scientific discussions and support on the analyses and manuscript. This is contribution number 106 from Tara Oceans.

## CONFLICT OF INTEREST

The authors declare no competing interests.

## AUTHOR CONTRIBUTION

**Romuald Laso-Jadart:** Conceptualization (equal); Formal analysis (lead); Investigation (lead); Methodology (equal); Writing-original draft (lead). **Kevin Sugier:** Resources (lead). **Emmanuelle Petit:** Resources (lead). **Karine Labadie:** Resources (lead). **Pierre Peterlongo:** Software (supporting); Writing-review & editing (supporting). **Christophe Ambroise:** Methodology (supporting); Validation (supporting); Writing-review & editing (equal). **Patrick Wincker:** Funding acquisition (lead); Supervision (supporting); Writing-review & editing (supporting). **Jean-Louis Jamet:** Resources (lead); Writing-review & editing (supporting). **Mohammed-Amin Madoui:** Conceptualization (lead); Formal analysis (supporting); Investigation (equal); Methodology (equal); Supervision (lead); Validation (equal); Writing-original draft (supporting); Writing-review & editing (lead).

## DATA AVAILABILITY STATEMENT

All data are available at ENA (European Nucleotide Archive); see Table S1.

## ORCID

Romuald Laso-Jadart  <https://orcid.org/0000-0001-8410-1121>

Mohammed-Amin Madoui  <https://orcid.org/0000-0003-4809-2971>

## REFERENCES

- Aarbakke, O. N. S., Bucklin, A., Halsband, C., & Norrbin, F. (2014). Comparative phylogeography and demographic history of five sibling species of *Pseudocalanus* (Copepoda: Calanoida) in the North Atlantic Ocean. *Journal of Experimental Marine Biology and Ecology*, 461, 479–488. <https://doi.org/10.1016/j.jembe.2014.10.006>
- Alberti, A., Poulain, J., Engelen, S., Labadie, K., Romac, S., Ferrera, I., ... Wincker, P. (2017). Viral to metazoan marine plankton nucleotide sequences from the Tara Oceans expedition. *Scientific Data*, 4, 170093. <https://doi.org/10.1038/sdata.2017.93>
- Arif, M., Gauthier, J., Sugier, K., Ludicone, D., Jaillon, O., Wincker, P., ... Madoui, M.-A. (2018). Discovering Millions of Plankton Genomic Markers from the Atlantic Ocean and the Mediterranean Sea. *Molecular Ecology Resources*, 0–3. <https://doi.org/10.1111/1755-0998.12985>
- Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B*, 57, 289–300. <https://doi.org/10.2307/2346101>
- Blachowiak-Samolyk, K., Kwasniewski, S., Hop, H., & Falk-Petersen, S. (2008). Magnitude of mesozooplankton variability: A case study from the Marginal Ice Zone of the Barents Sea in spring. *Journal of Plankton Research*, 30, 311–323. <https://doi.org/10.1093/plankt/fbn002>
- 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. <https://doi.org/10.1111/mec.13581>
- Brown, A. P., Arias-Rodriguez, L., Yee, M. C., Tobler, M., & Kelley, J. L. (2018). Concordant changes in gene expression and nucleotides underlie independent adaptation to hydrogen-sulfide-rich environments. *Genome Biology and Evolution*, 10, 2867–2881. <https://doi.org/10.1093/gbe/evy198>
- Buchfink, B., Xie, C., & Huson, D. H. (2014). Fast and sensitive protein alignment using DIAMOND. *Nature Methods*, 12, 59–60. <https://doi.org/10.1038/nmeth.3176>
- Bucklin, A., Divito, K. R., Smolina, I., Choquet, M., Questel, J. M., Hoarau, G., & O'Neill, R. (2018). Population Genomics of Marine Zooplankton. 0–66. [https://doi.org/10.1007/13836\\_2017\\_9](https://doi.org/10.1007/13836_2017_9)
- Bucklin, A., & Wiebe, P. H. (1998). Low mitochondrial diversity and small effective population sizes of the copepods *Calanus finmarchicus* and *Nannocalanus minor*: Possible impact of climatic variation during recent glaciation. *Journal of Heredity*, 89, 383–392. <https://doi.org/10.1093/jhered/89.5.383>
- Castellani, C., Licandro, P., Fileman, E., Di Capua, I., & Mazzocchi, M. G. (2016). Contribution to the Themed Section: 'The role of zooplankton in marine biogeochemical cycles: From fine scale to global marine zooplankton and the metabolic theory of ecology: Is it a predictive tool?' *Journal of Plankton Research*, 38, 762–770.
- Cepeda, G. D., Sabatini, M. E., Scioscia, C. L., Ramirez, F. C., & Viñas, M. D. (2016). On the uncertainty beneath the name *Oithona similis* Claus, 1866 (Copepoda, Cyclopoida). *ZooKeys*, 2016, 1–15. <https://doi.org/10.3897/zookeys.552.6083>
- Cingolani, P., Platts, A., Wang, L. 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*, 6(2), 80–92. <https://doi.org/10.4161/fly.19695>
- Claus, C. (1866). Die Copepoden-Fauna von Nizza. Ein Beitrag zur Charakteristik der Formen und deren Abänderungen 'im Sinna Darwin's'. *Elwert'sche Univ. Marbg. Leipzig*, 1(1), 34.

- 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. <https://doi.org/10.1016/j.ympev.2016.12.019>
- Delignette-Muller, M. L., & Dutang, C. (2015). fitdistrplus: An R Package for Fitting Distributions. *Journal of Statistical Software*, 64, 1–34. <https://doi.org/10.18637/jss.v064.i04>
- Dvoretzky, V. G. (2012). Seasonal mortality rates of *Oithona similis* (Cyclopoida) in a large Arctic fjord. *Polar Science*, 6, 263–269. <https://doi.org/10.1016/j.polar.2012.09.001>
- Edmands, S. (2001). Phylogeography of the intertidal copepod *Tigriopus californicus* reveals substantially reduced population differentiation at northern latitudes. *Molecular Ecology*, 10, 1743–1750. <https://doi.org/10.1046/j.0962-1083.2001.01306.x>
- Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). *A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise*. [www.aaai.org](http://www.aaai.org). Accessed January 8, 2019.
- Fay, J. C., & Wittkopp, P. J. (2008). Evaluating the role of natural selection in the evolution of gene regulation. *Heredity (Edinb.)*, 100, 191–199. <https://doi.org/10.1038/sj.hdy.6801000>
- Fraser, H. B. (2013). Gene expression drives local adaptation in humans. *Genome Research*, 23(7), 1089–1096. <https://doi.org/10.1101/gr.152710.112>
- Fu, L., Niu, B., Zhu, Z., Wu, S., & Li, W. (2012). CD-HIT: Accelerated for clustering the next-generation sequencing data. *Bioinformatics*, 28, 3150–3152. <https://doi.org/10.1093/bioinformatics/bts565>
- Gallienne, C. P. (2001). Is *Oithona* the most important copepod in the world's oceans? *Journal of Plankton Research*, 23, 1421–1432. <https://doi.org/10.1093/plankt/23.12.1421>
- Goetze, E. (2005). Global Population Genetic Structure and Biogeography of the Oceanic Copepods *Eucalanus Hyalinus* and *E. Spinifer*. *Evolution*, 59, 2378. <https://doi.org/10.1554/05-077.1>
- Guinder, V. A., & Molinero, J. C. (2013). Climate change effects on marine phytoplankton. *Marine Ecology in a Changing World*, 68–90. <https://doi.org/10.1201/b16334-4>
- Haas, B. J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P. D., Bowden, J., ... Regev, A. (2013). De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nature Protocols*, 8, 1494–1512. <https://doi.org/10.1038/nprot.2013.084>
- Hartigan, J. A., & Hartigan, P. M. (1985). The dip test of unimodality. *Annals of Statistics*, 13, 70–84. <https://doi.org/10.1214/aos/1176346577>
- Hijmans, R. J. (2017). geosphere: Spherical Trigonometry. <https://cran.r-project.org/package=geosphere>
- Höring, F., Cornils, A., Auel, H., Bode, M., & Held, C. (2017). Population genetic structure of *Calanoides natalis* (Copepoda, Calanoida) in the eastern Atlantic Ocean and Benguela upwelling system. *Journal of Plankton Research*, 39, 618–630. <https://doi.org/10.1093/plankt/fbx035>
- Humes, A. G. (1994). How many copepods? *Hydrobiologia*, 293, 1–7. <https://doi.org/10.1007/BF00229916>
- Hutter, S., Saminadin-Peter, S. S., Stephan, W., & Parsch, J. (2008). Gene expression variation in African and European populations of *Drosophila melanogaster*. *Genome Biology*, 9, 1–15. <https://doi.org/10.1186/gb-2008-9-1-r12>
- 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. <https://doi.org/10.1093/bioinformatics/btu031>
- Josephs, E. B., Lee, Y. W., Stinchcombe, J. R., & Wright, S. I. (2015). Association mapping reveals the role of purifying selection in the maintenance of genomic variation in gene expression. *Proceedings of the National Academy of Sciences*, 112, 15390–15395. <https://doi.org/10.1073/pnas.1503027112>
- Karsenti, E., Acinas, S. G., Bork, P., Bowler, C., De Vargas, C., Raes, J., ... Wincker, P. (2011). A holistic approach to marine eco-systems biology. *PLoS Biology*, 9, e1001177. <https://doi.org/10.1371/journal.pbio.1001177>
- Kjørboe, T. (2011). What makes pelagic copepods so successful? *Journal of Plankton Research*, 33, 677–685. <https://doi.org/10.1093/plankt/fbq159>
- Kopylova, E., Noé, L., & Touzet, H. (2012). SortMeRNA: Fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. *Bioinformatics*, 28, 3211–3217. <https://doi.org/10.1093/bioinformatics/bts611>
- Lauritano, C., Procaccini, G., & Ianora, A. (2012). Gene expression patterns and stress response in marine copepods. *Marine Environment Research*, 76, 22–31. <https://doi.org/10.1016/j.marenvres.2011.09.015>
- Lenz, P. H. (2012). The biogeography and ecology of myelin in marine copepods. *Journal of Plankton Research*, 34, 575–589. <https://doi.org/10.1093/plankt/fbs037>
- Leung, V., Iliescu, A., Jolicoeur, C., Gravel, M., Apuzzo, S., Torban, E., ... Gros, P. (2016). The planar cell polarity protein Vangl2 is required for retinal axon guidance. *Developmental Neurobiology*, 76, 150–165. <https://doi.org/10.1002/dneu.22305>
- Li, H., & Durbin, R. (2009). Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics*, 26, 589–595. <https://doi.org/10.1093/bioinformatics/btp698>
- Li, J., Liu, Y., Kim, T. H., Min, R., & Zhang, Z. (2010). Gene expression variability within and between human populations and implications toward disease susceptibility. *PLoS Computational Biology*, 6, <https://doi.org/10.1371/journal.pcbi.1000910>
- Luu, K., Bazin, E., & Blum, M. G. B. (2017). pcadapt: An R package to perform genome scans for selection based on principal component analysis. *Molecular Ecology Resources*, 17, 67–77. <https://doi.org/10.1111/1755-0998.12592>
- Madoui, M.-A., Poulain, J., Sugier, K., Wessner, M., Noel, B., Berline, L., ... Wincker, P. (2017). New insights into global biogeography, population structure and natural selection from the genome of the epipelagic copepod *Oithona*. *Molecular Ecology*, 26, 4467–4482. <https://doi.org/10.1111/mec.14214>
- Matoulkova, E., Michalova, E., Vojtesek, B., & Hrstka, R. (2012). The role of the 3' untranslated region in post-transcriptional regulation of protein expression in mammalian cells. *RNA Biology*, 9, 563–576. <https://doi.org/10.4161/rna.20231>
- Mignone, F., Gissi, C., Liuni, S., & Pesole, G. (2002). Untranslated regions of mRNAs. *Genome Biology*, 3, 0004–1. <https://doi.org/10.1186/gb-2002-3-3-reviews0004>
- Myers, P. G., Donnelly, C., & Ribergaard, M. H. (2008). Structure and variability of the West Greenland Current in Summer derived from 6 repeat standard sections. *Progress in Oceanography*, 80(1-2), 93–112. <https://doi.org/10.1016/j.pocean.2008.12.003>
- Nishida, S. (1985). Taxonomy and distribution of the family Oithonidae (Copepoda, Cyclopoida) in the Pacific and Indian Oceans. *Bull. Ocean Res. Institute. Bulletin of the Ocean Research Institute, University of Tokyo*, 20, 1–167.
- Oksanen, J., Guillaume Blanchet, F., Kindt, R., & Legendre, P. (2018). *vegan: Community Ecology Package*. <https://cran.r-project.org/package=vegan>
- Pandey, V., Turm, H., Bekenstein, U., Shifman, S., & Kadener, S. (2013). A new in vivo model of pantothenate kinase-associated neurodegeneration reveals a surprising role for transcriptional regulation in pathogenesis. *Frontiers in Cellular Neuroscience*, 7, 1–14. <https://doi.org/10.3389/fncel.2013.00146>
- Passow, C. N., Henpita, C., Shaw, J. H., Quackenbush, C. R., Warren, W. C., Schartl, M., ... Tobler, M. (2017). The roles of plasticity and evolutionary change in shaping gene expression variation in natural populations of extremophile fish. *Molecular Ecology*, 26, 6384–6399. <https://doi.org/10.1111/mec.14360>

- Peijnenburg, K. T. C. A., & Goetze, E. (2013). High evolutionary potential of marine zooplankton. *Ecology and Evolution*, 3, 2765–2781. <https://doi.org/10.1002/ece3.644>
- Pelejero, C., Calvo, E., & Hoegh-Guldberg, O. (2010). Paleo-perspectives on ocean acidification. *Trends in Ecology & Evolution*, 25, 332–344. <https://doi.org/10.1016/j.tree.2010.02.002>
- Pesant, S., Not, F., Picheral, M., Kandels-Lewis, S., Le Bescot, N., Gorsky, G., ... Searson, S. (2015). Open science resources for the discovery and analysis of Tara Oceans data. *Scientific Data*, 2, 1–16. <https://doi.org/10.1038/sdata.2015.23>
- Peterlongo, P., Riou, C., Drezen, E., & Lemaitre, C. (2017). DiscoSnp++: de novo detection of small variants from raw unassembled read set(s). *bioRxiv*. 209965. <https://doi.org/10.1101/209965>
- 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. <https://doi.org/10.1186/1475-925X-13-94>
- Ram, A., Jalal, S., Jalal, A. S., & Kumar, M. (2010). A density based algorithm for discovering density varied clusters in large spatial databases. *International Journal of Computers and Applications*, 3, 1–4. <https://doi.org/10.5120/739-1038>
- Riginos, C., Crandall, E. D., Liggins, L., Bongaerts, P., & Tremblay, E. A. (2016). Navigating the currents of seascape genomics: How spatial analyses can augment population genomic studies. *Current Zoology*, 62, 581–601. <https://doi.org/10.1093/cz/zow067>
- Robinson, M. D., & Smyth, G. K. (2007). Moderated statistical tests for assessing differences in tag abundance. *Bioinformatics*, 23, 2881–2887. <https://doi.org/10.1093/bioinformatics/btm453>
- Rose (1933). *Faune de France - Copépodes pélagiques*. Lechevalier, Paris. [www.faunedefrance.org](http://www.faunedefrance.org), 132, 767. <https://doi.org/10.1038/132767a0>
- Salazar, G., Paoli, L., Alberti, A., Huerta-Cepas, J., Ruscheweyh, H.-J., Cuenca, M., ... Wincker, P. (2019). Gene expression changes and community turnover differentially shape the global ocean meta-transcriptome. *Cell*, 179(5), 1068–1083. <https://doi.org/10.1016/j.cell.2019.10.014>
- Sato, M. P., Makino, T., & Kawata, M. (2016). Natural selection in a population of *Drosophila melanogaster* explained by changes in gene expression caused by sequence variation in core promoter regions. *BMC Evolutionary Biology*, 16, 1–12. <https://doi.org/10.1186/s12862-016-0606-3>
- Signor, S. A., & Nuzhdin, S. V. (2018). The evolution of gene expression in cis and trans. *Trends in Genetics*, 34(7), 532–544. <https://doi.org/10.1016/j.tig.2018.03.007>
- Smarandache-Wellmann, C. R. (2016). Arthropod neurons and nervous system. *Current Biology*, 26, R960–R965. <https://doi.org/10.1016/j.cub.2016.07.063>
- Steige, K. A., Laenen, B., Reimegård, J., Scofield, D. G., & Slotte, T. (2017). Genomic analysis reveals major determinants of cis-regulatory variation in *Capsella grandiflora*. *Proceedings of the National Academy of Sciences of the USA*, 114, 1087–1092. <https://doi.org/10.1073/pnas.1612561114>
- Tian, L., Khan, A., Ning, Z., Yuan, K., Zhang, C., Lou, H., ... Xu, S. (2018). Genome-wide comparison of allele-specific gene expression between African and European populations. *Human Molecular Genetics*, 27, 1067–1077. <https://doi.org/10.1093/hmg/ddy027>
- Uricaru, R., Rizk, G., Lacroix, V., Quillery, E., Plantard, O., Chikhi, R., ... Peterlongo, P. (2014). Reference-free detection of isolated SNPs. *Nucleic Acids Research*, 43(2), e11. <https://doi.org/10.1093/nar/gku1187>
- Wassmann, P., Reigstad, M., Haug, T., Rudels, B., Carroll, M. L., Hop, H., ... Pavlova, O. (2006). Food webs and carbon flux in the Barents Sea. *Progress in Oceanography*, 71, 232–287. <https://doi.org/10.1016/j.pocean.2006.10.003>
- Weatherby, T. M., Davis, A. D., Hartline, D. K., & Lenz, P. H. (2000). The need for speed. II. Myelin in calanoid copepods. *Journal of Comparative Physiology A: Sensory, Neural, and Behavioral Physiology*, 186(4), 347–357. <https://doi.org/10.1007/s003590050435>
- Weydmann, A., Coelho, N. C., Serrão, E. A., Burzyński, A., & Pearson, G. A. (2016). Pan-Arctic population of the keystone copepod *Calanus glacialis*. *Polar Biology*, 39, 2311–2318. <https://doi.org/10.1007/s00300-016-1898-x>
- Whitehead, A. (2012). Comparative genomics in ecological physiology: Toward a more nuanced understanding of acclimation and adaptation. *Journal of Experimental Biology*, 215, 884–891. <https://doi.org/10.1242/jeb.058735>
- Williams, R. B. H., Chan, E. K. F., Cowley, M. J., & Little, P. F. R. (2007). The influence of genetic variation on gene expression. *Genome Research*, 17, 1707–1716. <https://doi.org/10.1101/gr.6981507>
- Wright, S. (1951). The genetical structure of populations. *Annals of Eugenics*, 15, 323–354. <https://doi.org/10.1111/j.1469-1809.1949.tb02451.x>
- Zamora-Terol, S., Nielsen, T. G., & Saiz, E. (2013). Plankton community structure and role of *Oithona similis* on the western coast of Greenland during the winter-spring transition. *Marine Ecology Progress Series*, 483, 85–102. <https://doi.org/10.3354/meps10288>
- Zhang, W., Duan, S., Kistner, E. O., Bleibel, W. K., Huang, R. S., Clark, T. A., ... Dolan, M. E. (2008). Evaluation of genetic variation contributing to differences in gene expression between populations. *American Journal of Human Genetics*, 82, 631–640. <https://doi.org/10.1016/j.ajhg.2007.12.015>

## SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

**How to cite this article:** Laso-Jadart R, Sugier K, Petit E, et al. Investigating population-scale allelic differential expression in wild populations of *Oithona similis* (Cyclopoida, Claus, 1866). *Ecol Evol*. 2020;10:8894–8905. <https://doi.org/10.1002/ece3.6588>