Linking Allele-Specific Expression And Natural Selection In Wild Populations

Romuald Laso-Jadart, Kevin Sugier, Emmanuelle Petit, Karine Labadie, Pierre Peterlongo, Christophe Ambroise, Patrick Wincker, Jean-Louis Jamet, Mohammed-Amin Madoui

To cite this version:
Romuald Laso-Jadart, Kevin Sugier, Emmanuelle Petit, Karine Labadie, Pierre Peterlongo, et al.. Linking Allele-Specific Expression And Natural Selection In Wild Populations. 2019. hal-02275928

HAL Id: hal-02275928
https://inria.hal.science/hal-02275928
Preprint submitted on 2 Sep 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
Linking Allele-Specific Expression And Natural Selection In Wild Populations

Romuald Laso-Jadart¹,⁶*, Kevin Sugier¹, Emmanuelle Petit², Karine Labadie², Pierre Peterlongo⁵, Christophe Ambroise⁴, Patrick Wincker¹,⁶, Jean-Louis Jamet⁵, Mohammed-Amin Madoui¹,⁶*

¹Génomique Métabolique, Genoscope, Institut François Jacob, CEA, CNRS, Univ Evry, Université Paris-Saclay, Evry, France.
²CEA, Genoscope, Institut de Biologie François Jacob, Université Paris-Saclay, Evry, 91057, France.
³Univ Rennes, CNRS, Inria, IRISA - UMR 6074, F-35000 Rennes.
⁴LaMME, CNRS, Univ Evry, Université Paris-Saclay, Evry, France
⁵Université de Toulon, Aix-Marseille Université, CNRS/INSU/IRD, Mediterranean Institute of Oceanology MIO UMR 110, CS 60584, 83041 Toulon cedex 9, France.

*Corresponding authors. Emails: rlasojad@genoscope.cns.fr & amadoui@genoscope.cns.fr
Abstract

Allele-specific expression (ASE) is now a widely studied mechanism at cell, tissue and organism levels. However, population-level ASE and its evolutive impacts have still never been investigated. Here, we hypothesized a potential link between ASE and natural selection on the cosmopolitan copepod *Oithona similis*. We combined metagenomic and metatranscriptomic data from seven wild populations of the marine copepod *O. similis* sampled during the *Tara Oceans* expedition. We detected 587 single nucleotide variants (SNVs) under ASE and found a significant amount of 152 SNVs under ASE in at least one population and under selection across all the populations. This constitutes a first evidence that selection and ASE target more common loci than expected by chance, raising new questions about the nature of the evolutive links between the two mechanisms.
Introduction

Allele-specific expression (ASE), or allelic imbalance, refers to the difference of expression between two alleles of a locus in a heterozygous genotype due to genetic or epigenetic polymorphism. Through DNA methylation or histone modifications, epigenetics could repress a disadvantageous or a specific parental allele, leading in some cases to monoallelic expression, as demonstrated in a variety of organisms including mouse, maize or bumblebee\(^1\)\(^-\)\(^4\). On the other hand, ASE may have a genetic origin through, for example, mutations in transcription factor binding sites\(^5\)\(^,\)\(^6\), or post-transcriptional mechanisms like non-sense mediated decay\(^7\)\(^-\)\(^9\). Recently, several studies led to a better understanding of ASE thanks to the development of advanced tools allowing their detection at the individual, tissue and cell levels\(^7\)\(^,\)\(^9\)\(^-\)\(^15\). ASE has been investigated in the context of cis- and trans-regulation of gene expression\(^16\), expression evolution\(^17\) and association between gene expression and human diseases\(^18\). First approaches in natural populations of primates and flycatchers have been undertaken with individual-level data\(^19\)\(^-\)\(^21\). Moreover, studies began to question the relative contribution of genetics and environment on gene expression using ASE in human\(^22\)\(^-\)\(^25\) and fruit flies\(^26\).

However, population-level ASE in several wild populations of one species and its potential evolutive origins and consequences remain largely uninvestigated. The need for numerous individual RNA-seq and whole-genome genotyping data constitutes the main obstacle for population-scale analyses. Today, the advances of next-generation sequencing technologies allow integrating large metagenomic and metatranscriptomic data from environmental samples, and new approaches can now be considered using whole population information.

In natural populations, we expect most loci to be under neutral evolution and balanced expression (Fig. 1a)\(^27\)\(^,\)\(^28\). When selection occurs on a specific locus, the selected allele tends to homozygosity.
creating a specific population-level expression pattern of the selected allele (Fig. 1b). In the absence of selection, if the same allele is favored by ASE in most of the individuals, the observed population-level expression pattern (Fig. 1c) will be similar to the one observed in the case of selection (Fig. 1b). Considering that both mechanisms impact fitness, we hypothesized that ASE and natural selection could preferentially target the same loci, in different populations, showing a possible link between the two mechanisms.

In this study, we focus on the widespread epipelagic, temperate and cold water small-sized copepod, *Oithona similis* (Cyclopoida, Claus 1866), notably known to be highly abundant in Arctic30–33. Copepods, and particularly *Oithona*, are small crustaceans forming the most abundant metazoan on Earth, reflecting strong adaptive capacities to environmental fluctuations34–36. They play a key ecological role in biogeochemical cycles and in the marine trophic food chain37; therefore copepods constitute an ideal model to study wild population evolution38–41.

The first goal of our study was to identify loci under selection before demonstrating that population-level ASE can be detected with metagenomic and metatranscriptomic data collected by the *Tara* Oceans expedition42 during its Arctic phase. Then we provided evidence of a quantitative link between ASE and natural selection.

**Material and Methods**

**Material sampling, mRNA extraction and transcriptome sequencing**

*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 (0-10m) using zooplankton nets with a mesh of 90µm.
and 200µm (0.5 m diameter and 2.5 m length). Samples were preserved in 70% ethanol and stored at -4°C. From the Large Bay of Toulon samples, *O. similis* individuals were isolated under the stereomicroscope. 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 were constructed using the SMARTer-Seq v4 Ultra low Input RNA kit (ClonTech). The libraries were constructed using the NEBNext Ultra II kit, and were sequenced with an Illumina HiSeq2500 (Supplementary Fig. 1).

Transcriptomes assembly and annotation

Each read set was assembled with Trinity v2.5.1 using default parameters and transcripts were clustered using cd-hit v4.6.1 (Supplementary Table 1). To ensure the classification of the sampled individuals, each ribosomal read set were detected with SortMeRNA and mapped with bwa v0.7.15 using default parameters to 82 ribosomal 28S sequences of *Oithona* species used in Cornils et al., 2017 (Supplementary Fig. 2). The transcriptome assemblies were annotated with Transdecoder v5.1.0 to predict the open reading frames (ORFs) and protein sequences (Supplementary Table 1). In parallel, homology searches were also included as ORF retention criteria; the peptide sequences of the longest ORFs were aligned on *Oithona nana* proteome using DIAMOND v0.9.22. Protein domain annotation was performed on the final ORF predictions with Interproscan v5.17.56.0 and a threshold of e-value <10\(^{-5}\) was applied for Pfam annotations. Finally, homology searches of the predicted proteins were done against the nr NCBI
Variant calling using *Tara Oceans* metagenomic and metatranscriptomic data

We used metagenomic and metatranscriptomic reads generated from samples of the size fraction 20–180 µm collected in seven *Tara Oceans* stations TARA_155, 158, 178, 206, 208, 209 and 210 (Supplementary Table 2), according to protocols described in Alberti et al. 2017. The reference-free variant caller *DiscoSNP++* was used to extract SNVs simultaneously from raw metagenomic and metatranscriptomic reads, and ran using parameters –b 1. Only SNVs corresponding to biallelic loci with a minimum of 4x of depth of coverage in all stations were initially selected. Then, SNVs were clustered based on their loci co-abundance across samples using density-based clustering algorithm implemented in the R package dbscan and ran with parameters epsilon = 10 and minPts = 10. This generated three SNVs clusters, the largest of which contained 102,258 SNVs. To ensure the presence of *O. similis* SNVs without other species, we observed the fitting of the depth of coverage to the expected negative binomial distribution in each population (Supplementary Fig. 3). For each variant in each population, the B-allele frequency (BAF) and the population-level B-allele relative expression (BARE) were computed; \( BAF = \frac{G_B}{G_B + G_A} \) and \( BARE = \frac{T_B}{T_B + T_A} \), with \( G_A \) and \( G_B \) the metagenomic read counts of the reference and alternative alleles respectively, \( T_A \) and \( T_B \) the metatranscriptomic read counts of the reference and alternative alleles respectively.

Variant filtering and annotation

SNVs were filtered based on their metagenomic coverage. Those with a metagenomic coverage lying outside a threshold of median ± 2 \( \sigma \) in at least one population, with a minimum and
maximum of 5x and 150x coverage were discarded. 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 SNVs (Supplementary Fig. 4).

The variant annotation was conducted in two steps. First, the variant sequences were relocated on the previously annotated *O.similis* transcripts using the “VCF_creator.sh” program of *DiscoSNP++*. Secondly, a variant annotation was carried with SNPeff 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 in the following classes: missenses, synonymous, 5’ and 3’ UTR. A significant excess was considered for a hypergeometric test p-value < 0.05.

**Genomic differentiation and detection of selection**

The differentiation among the seven populations was investigated through the computation of the $F_{ST}$ metric or Wright’s fixation index. For each locus, global $F_{ST}$ including the seven populations and pairwise-$F_{ST}$ between each pair of population was computed, using the corresponding BAF matrix. For the global $F_{ST}$ computation, a Hartigans’ dip test for unimodality was performed. We retained the median pairwise-$F_{ST}$ as a measure of the genomic differentiation between each population. The *pcadapt* R package v4.0.2 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 < 0.05 were considered under selection.
Population-level ASE detection using metagenomic and metatranscriptomic data

In each population, we first selected variants for $\text{BAF} \neq \{0,1\}$. Then, we computed $D = \text{BAF} - \text{BARE}$, as the deviation between the BAF and the BARE. In the absence of ASE, $D$ is close to 0, as most of the biallelic loci are expected to have a balanced expression\textsuperscript{7,28,61} we expect the $D$ distribution to follow a Gaussian distribution centered on 0. We estimated the Gaussian distribution parameters and tested the probability of a variant to belong to this distribution (“$D$-test” or “deviation test”). Given the large number of tests, we applied the Benjamini and Hochberg approach to control the False Discovery Rate (FDR).

We also computed a “low expression bias” test by comparing the read counts $T_A$ and $T_B$ to the observed metagenomic proportion $1 - \text{BAF}$ and $\text{BAF}$ respectively with a chi-square test and applied the Benjamini and Hochberg correction for multiple testing. These two tests were applied to BAFs, BAREs and read counts of the seven populations separately and the seven sets of candidate loci targeted by ASE (deviation test q-value < 0.1 and low expression bias test q-value < 0.1) were crossed to identify loci under ASE in different populations, or shared ASEs.

To identify alleles targeted by both ASE and selection, the set of variants under ASE in each population was crossed with the set of loci detected under selection. The size of the intersection was tested by a hypergeometric test, $\text{Hypergeometric}(q,m,n,k)$, with $q$ being number of alleles under ASE 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 ASE in the tested population. We considered that, in a given population, the number of alleles under both ASE and selection was significantly higher than expected by chance for p-value < 0.05.
Gene enrichment analysis

To identify specific biological function or processes associated to the variants, a domain-based analysis was conducted. The Pfam annotation of the transcripts carrying variants targeted by ASE and selection was used as entry for dcGO Enrichment. A maximum of the best 300 GO-terms were chosen based on their z-score and FDR p-value (<10^-3) in each ontology category. To reduce redundancy, these selected GO-terms were processed using REVIGO, with a similarity parameter of 0.5 against the whole Uniprot catalogue under the SimRel algorithm. To complete the domain-based analysis, the functional annotations obtained from the homology searches against the nr were manually curated.

Results

Oithona similis genomic differentiation and selection in Arctic Seas

From metagenomic and metatranscriptomic raw data of seven sampling stations (Fig. 2a), we identified 25,768 expressed variants. Among them, 97% were relocated on O. similis transcriptomes.

The global distribution of $F_{ST}$ of the seven populations was unimodal (Hartigans’ dip test, D=0.0012, p-value=0.99) with a median-$F_{ST}$ at 0.1, confirmed by the pairwise-$F_{ST}$ distributions (Supplementary Fig. 5). 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 (Fig. 2c,d). 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).

The PCA decomposed the genomic variability in six components; the first two components discriminated TARA_155 and 178 from the others (32% and 28.1% variance explained respectively, Fig. 2b), 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 (Supplementary Fig. 5). 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 < 0.05).

**Loci targeted by population-level ASE and selection in Arctic populations**

The number of SNVs tested for ASE varied between 13,454 and 22,578 for TARA_210 and 206 respectively. As expected, the $D$ deviation, representing the deviation between B-allele frequency and B-allele relative expression, followed a Gaussian distribution in each population (Fig. 3a and Supplementary Fig. 6). Variants under ASE (i.e. having a $D$ significantly higher or lower than expected) were found in every population, ranging from 26 to 162 variants for TARA_178 and 206 respectively (Table 1). Overall, we found 587 variants under ASE, including 535 population-specific ASEs, and 52 ASEs shared by several populations (Fig.3b). Remarkably, 30 ASEs out of the 52 were present in the populations from TARA_158, 206 and 208 that correspond to the genetically closest populations. The seven sets of variants under ASE were crossed with the set of variants under selection, as illustrated for TARA_209 (Fig. 3c). The size of the intersection ranged from 5 to 42 variants (TARA_155 and 210/206) and was significantly higher than expected by chance for all the populations (hypergeometric test p-value < 0.05). It represented a total of 152 unique variants under selection and ASE in at least one population (Table 1, 21, 22).
Supplementary Table 3), corresponding to 23% and 26% of variants under ASE and under selection respectively.

**Functional analysis of genes targeted by ASE and selection**

Among the 152 loci targeted by ASE and selection, 145 were relocated on *O. similis* transcripts (Supplementary Table 4). Amid these transcripts, 137 (90%) had a predicted ORF, 97 (64%) were linked to at least one Pfam domain and 90 (59%) to a functional annotation. Fifteen SNVs were missense variations, 59 synonymous, 31 and 29 were located in 5' and 3' UTR, without any significant excess (Supplementary Table 4 and 5). Based on homology searches (Supplementary Table 4), eight genes were linked to nervous system (Table 2). Among them, two genes were involved in glutamate metabolism (omega-amidase NIT2 and 5-oxoprolinase), three were predicted to be glycine, γ-amino-butyric acid (GABA) and histamine neuroreceptors. Finally, four were also implicated in arthropods photoreceptors. The domain-based analysis confirmed these results, with an enrichment in GO-terms biological process also linked to nervous system (Supplementary Fig. 7).

**Discussion**

*O. similis* populations are weakly structured within the Arctic Seas

Global populations of *O. similis* are known to be composed of cryptic lineages. Thus the assessment that the seven populations used in our study belong to the same *O. similis* cryptic lineage was a prerequisite for further analyses. The high proportion of variants mapped on the Mediterranean transcriptomes (97%) showed that the variant clustering method was efficient to regroup loci of an *O. similis* cryptic species. Plus, the unimodal distribution of *F*<sub>ST</sub> showed that these populations of *O. similis* belong to the same polar cryptic species.
Secondly, we see that the seven populations examined are well connected with low median pairwise-$F_{ST}$, despite the large distances separating them. Weak genetic structure in the polar region was already highlighted for other major Arctic copepods like *Calanus glacialis* and *Pseudocalanus* species. The absence of structure was explained by ancient diminutions of effective population size due to past glaciations, or high dispersal and connectivity between the present-day populations due to marine currents.

Going into details, three different cases can be described. First, the differentiation of populations from TARA_155 and 178 compared to the others could be explained by isolation-by-distance. Secondly, the geographically close populations from TARA_210 and 209 present higher differentiation (median pairwise-$F_{ST}$ of 0.11). This could be explained by the West Greenland current acting as a physical barrier between the populations, which could lead to reduced gene flow. 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.

Metagenomic data enable to draw the silhouette of the population genetics but lacks resolution when dealing with intra-population 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.

**Toward the link between ASE and natural selection**

Usually, at the individual level, the ASE analyses are achieved by measuring the difference in RNA-seq read counts of a heterozygous site. But at the population level, obtaining a large number of individuals remains a technical barrier especially for uncultured animals, or when the
amount of DNA retrieved from a single individual is not sufficient for high-throughput sequencing. Here, the detection of ASE at population level was possible by comparing the observed frequencies of the alleles based on metagenomic and metatranscriptomic data, which by passes the obstacles previously described.

In our study, the amount of detected ASE in each population was always lower than 1% of tested heterozygous variants, which altogether correspond to 2% of the total set of variants. In humans, baboons and flycatchers, 17%, 23% of genes and 7.5% of transcripts were affected by ASE respectively. The difference with our results can be explained by one main reason. The detection of population-level ASE identifies only the ASE present in a large majority of individuals, which can be considered as “core ASEs”.

These core ASEs constitute the majority of detected ASEs and are population-specific, meaning the main drivers of this expression pattern are local conditions like different environmental pressures or population dynamics including, for example, the proportion of each developmental stage and sex, known to vary between populations and across seasons. Another result is the presence of a small amount of variants affected by ASE in different populations. Most of these variants are under ASE in at least two of the three closest populations from TARA_158, 206 and 208. First, the genetic closeness and large geographic distances between these three populations suggest that their shared ASEs are under an environmental independent genetic control. Secondly, the number of variants tested for ASE is higher in these three populations than the others, leading to a greater proportion of ASEs detected which also elevates the chances for a variant to be declared under ASE in several populations.
A significant amount of SNVs (152) were subject to selection among the seven populations and to ASE in at least one population. We confirmed our first hypothesis (Fig. 1), as exampled with the variant 841109 (Fig. 3d), characterized by an ASE in favor of the B-allele in TARA_209 and fixation of this allele in TARA_210. Three main features of ASE can be under selection. First, the observed variant can be in linkage with another variation in upstream cis-regulatory elements like transcription factors fixation sites, or epialleles. Secondly, the annotation of candidate variants with SNPeff revealed a majority of variants located in 5' and 3'UTRs, which are variations known to both affect transcription efficiency through mRNA secondary structures, stability and location. For variants located in exons, a majority were identified as synonymous mutations, growingly described as potential target of selection by codon usage bias, codon context, mRNA secondary structure or transcription and translation dynamics. Finally, fifteen missense mutations were spotted, but with moderate predicted impact on protein amino acid composition. However, we did not find premature nonsense mutation, even if variants under ASE has been described to trigger or escape potential nonsense-mediated decay, but the possibility that the causal variation is located in introns cannot be ruled out.

The process of adaptation through gene expression was suspected in human populations and investigated thanks to the large and accessible amount of data. In a first study, a link has been established between gene expression and selection, affecting particular genes and phenotypes, looking at cis-acting SNPs. In a second study, the team was able to detect ASE in different populations and to quantify genetic differentiation and selection. They found particularly one gene that shows strong differentiation between European and African populations and under ASE in Europeans and not in Africans. However, they did not quantify this phenomenon. Both emphasized the impact of selection on gene expression. In the same way, another approach
showed that ASE or expression variations with high effect size were rare in the populations, based on intra-population analyses in *Capsella grandiflora* and human. This situation is presumably encountered in our analysis, as exampled with the B-allele of variant 20760212, under ASE and with a low genomic frequency in TARA_210, but fixed in the others (Fig. 3d). Our results complete previous analyses, as they quantify the link between ASE and selection in populations and reveal the evolutive potency of ASE, for the first time at the population-level. It remains to understand the nature of the association between ASE and selection. To address this question, we formulate the hypothesis that they impact chronologically the same loci, following constant or increasing selective pressure as well as environmental changes (Fig. 4).

**Nervous system and visual perception are important targets of the natural selection and ASE in *O. similis***

This evolutive link between ASE and selection is supported by the biological functions associated to the targeted genes, which are involved notably in the copepods nervous system in two ways. The first result is the presence of genes implicated in glutamate metabolism and glycine and/or GABA receptors. Glutamate and GABA are respectively excitatory and inhibitory neurotransmitters in arthropods motor neurons. Plus, glycine and GABA receptors have already been described as a target of selection in *O. nana* in Mediterranean Sea. Secondly, the functional analysis revealed also the importance of the eye and visual perception in the *O. similis* evolution.

Copepod nervous system constitutes a key trait for its reproduction and survival, and based on our data, a prime target for evolution, allowing higher capacity of perceiving and fast reacting leading to more efficient predator escape, prey catching and mating. This can explain the great evolutive success of these animals.
Conclusion

Gene expression variation is thought to play a crucial role in evolutive and adaptive history of natural populations. Herein, we developed proper methods integrating metagenomic and metatranscriptomic data to detect ASE at the population-level for the first time. Then, we demonstrated the link between ASE and natural selection by providing a quantitative observation of this phenomenon and its impact on specific biological features of copepods. In the future, we will try to generalize these observations to other organisms. Then, we will understand the nature of the link between ASE and natural selection by questioning the chronology between the two mechanisms.

Acknowledgments

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, 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). This is contribution number XX from Tara Oceans.
Author’s contributions

Individuals for transcriptome production were sampled by J-LJ and KS. KS extracted RNA, EP and KL prepared the libraries and sequencing, MAM assembled the reads and RLJ annotated transcriptomes. PP and CA gave expertise support on DiscoSNP++ and statistical framework respectively. RLJ and MAM performed the analyses and wrote the manuscript. MAM designed and supervised the study. J-LJ and PW offered scientific support.

Competing interests

The authors declare no competing interests.

References


70. Höring, F., Cornils, A., Auel, H., Bode, M. & Held, C. Population genetic structure of


### Tables

**Table 1**: Allele-specific expression detection and link with selection by population

**Table 2**: Functional annotations of variants targeted by ASE and selection implicated in nervous system

**Supplementary Table 1**: *Oithona similis* Mediterranean transcriptomes summary

**Supplementary Table 2**: *Tara* Oceans and *Oithona similis* Mediterranean transcriptomes samples accession numbers

**Supplementary Table 3**: Variants targeted by ASE and selection statistics

**Supplementary Table 4**: Variants targeted by ASE and selection functional annotations
Supplementary Table 5: Variant annotation by SNPeff

Figures

Figure 1: Population genomic and transcriptomic profiles of a biallelic locus in a case of a, Neutral evolution and balanced expression; b, Selection in favor of the B-allele; c, ASE in favor of the B-allele.

Figure 2: Genomic differentiation of *O. similis* populations from Arctic Seas. a, Geographic locations of the seven Tara Oceans sampling sites: Northern Atlantic (blue), Kara Sea (green), Baffin Bay (orange) and Labrador Sea (grey). b, Principal Component Analysis (PCA) computed by pcadapt based on allele frequencies. c, Pairwise-$F_{ST}$ matrix. The median (mean) of each pairwise-$F_{ST}$ distribution computed on allele frequencies is indicated. d, Graph representing the genomic differentiation of the seven populations of *O. similis*. The nodes represent the populations; their width reflects their centrality in the graph. The edges correspond to the genetic relatedness based on the median pairwise-$F_{ST}$ between each pair of population; 0.02 (large solid line), 0.05 to 0.07 (thin solid line) and 0.11 to 0.12 (dashed line).

Figure 3: Population Allele-specific expression detection and link with natural selection. a, The deviation $D$ distribution in TARA_209. The red line corresponds to the Gaussian distribution estimated from the data. b, Upset plot of the ASE detection in the seven populations. Each bar of the upper plot corresponds to the number of variants under ASE in the population(s) indicated by black dots in the lower plot. c, Crossing ASE and selection. The yellow circle represents the total set of variants. In green, the number of heterozygous variants tested for ASE in TARA_209. In blue and red, the amount of detected variants under ASE in TARA_209 and under selection among the populations respectively. In purple, the intersection comprising variants under ASE in TARA_209 and under selection, with its hypergeometric test p-value. d, Metagenomic and metatranscriptomic profiles of variants 841109 and 20760212. Each population is indicated on the x-axis, with the associated B-allele frequency (red) and B-allele relative expression (blue). The frequency is shown on the y-axis. The asterisks mean ASE was detected in the corresponding population.

Figure 4: From Allele-specific expression to natural selection. a, Evolution of allele frequency and allele relative expression over time. b, Evolution of selective pressure over time

Supplementary Fig 1: Method pipeline overview

Supplementary Fig 2: Validation of taxonomic assignation

Supplementary Fig 3: *Oithona similis* depth of coverage of biallelic loci in seven *Tara* Oceans samples
Supplementary Fig 4: Metagenomic coverage distribution of the seven Tara Oceans samples

Supplementary Fig 5: Genomic differentiation of Arctic Seas Oithona similis populations

Supplementary Fig 6: Allele-specific expression detection

Supplementary Fig 7: Functional analysis of O. similis transcripts targeted by ASE and selection
Table 1: Allele-specific expression detection and link with selection by population

<table>
<thead>
<tr>
<th>Population</th>
<th>Genomic median depth of coverage</th>
<th>Number of tested variants</th>
<th>Number of variants under ASE</th>
<th>Number of variants under ASE and selection</th>
<th>Hypergeometric test p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>TARA_155</td>
<td>25</td>
<td>18,812</td>
<td>91</td>
<td>6</td>
<td>9.89E-3*</td>
</tr>
<tr>
<td>TARA_158</td>
<td>35</td>
<td>21,476</td>
<td>131</td>
<td>29</td>
<td>5.06E-20*</td>
</tr>
<tr>
<td>TARA_178</td>
<td>24</td>
<td>18,145</td>
<td>26</td>
<td>9</td>
<td>5E-11*</td>
</tr>
<tr>
<td>TARA_206</td>
<td>55</td>
<td>22,578</td>
<td>162</td>
<td>42</td>
<td>4.82E-31*</td>
</tr>
<tr>
<td>TARA_208</td>
<td>48</td>
<td>21,469</td>
<td>133</td>
<td>14</td>
<td>2.2E-6*</td>
</tr>
<tr>
<td>TARA_209</td>
<td>12</td>
<td>13,956</td>
<td>62</td>
<td>24</td>
<td>1.05E-23*</td>
</tr>
<tr>
<td>TARA_210</td>
<td>14</td>
<td>13,454</td>
<td>69</td>
<td>42</td>
<td>8.89E-51*</td>
</tr>
<tr>
<td>Overall</td>
<td>-</td>
<td>25,768</td>
<td>587 (2.3%)</td>
<td>152 (0.59%)</td>
<td>-</td>
</tr>
</tbody>
</table>
Table 2: Functional annotations of variants targeted by ASE and selection implicated in nervous system

<table>
<thead>
<tr>
<th>VarID</th>
<th>Ref</th>
<th>Alt</th>
<th>Homology search</th>
<th>Pfam</th>
<th>SnpEff Localization</th>
<th>SnpEff Impact</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>722267</td>
<td>A</td>
<td>G</td>
<td>histamine H1 receptor</td>
<td>PF00001</td>
<td>3' UTR</td>
<td>MODIFIER</td>
<td>85–87</td>
</tr>
<tr>
<td>9665345</td>
<td>T</td>
<td>G</td>
<td>chaoptin</td>
<td>PF13306</td>
<td>PF13855</td>
<td>synonymous variant</td>
<td>LOW</td>
</tr>
<tr>
<td>15623788</td>
<td>G</td>
<td>A</td>
<td>eye-specific diacylglycerol kinase</td>
<td>PF13637</td>
<td>synonymous variant</td>
<td>LOW</td>
<td>90,91</td>
</tr>
<tr>
<td>23795359</td>
<td>A</td>
<td>T</td>
<td>vang-like protein 2-B</td>
<td>PF06638</td>
<td>synonymous variant</td>
<td>LOW</td>
<td>92,93</td>
</tr>
<tr>
<td>1276227</td>
<td>C</td>
<td>T</td>
<td>glycine receptor subunit alpha-2 / gamma-aminobutyric acid receptor subunit alpha-6</td>
<td>PF02932</td>
<td>PF2931</td>
<td>3' UTR</td>
<td>MODIFIER</td>
</tr>
<tr>
<td>1404415</td>
<td>G</td>
<td>C</td>
<td>omega-amidase NIT2</td>
<td>PF00795</td>
<td>3' UTR</td>
<td>MODIFIER</td>
<td>-</td>
</tr>
<tr>
<td>11174785</td>
<td>A</td>
<td>G</td>
<td>5-oxoprolinase</td>
<td>PF02538</td>
<td>PF05378</td>
<td>PF01968</td>
<td>5' UTR</td>
</tr>
<tr>
<td>11690229</td>
<td>A</td>
<td>T</td>
<td>glycine receptor subunit alpha-2</td>
<td>PF02931</td>
<td>synonymous variant</td>
<td>LOW</td>
<td>-</td>
</tr>
</tbody>
</table>
Figure 1: Population genomic and transcriptomic profiles of a biallelic locus in a case of a, Neutral evolution and balanced expression; b, Selection in favor of the B-allele; c, ASE in favor of the B-allele.
Figure 2: Genomic differentiation of *O. similis* populations from Arctic Seas. a, Geographic locations of the seven Tara Oceans sampling sites: Northern Atlantic (blue), Kara Sea (green), Baffin Bay (orange) and Labrador Sea (grey). b, Principal Component Analysis (PCA) computed by *pcaAdapt* based on allele frequencies. c, Pairwise-$F_{ST}$ matrix. The median (mean) of each pairwise-$F_{ST}$ distribution computed on allele frequencies is indicated. d, Graph representing the genomic differentiation of the seven populations of *O. similis*. The nodes represent the populations; their width reflects their centrality in the graph. The edges correspond to the genetic relatedness based on the median pairwise-$F_{ST}$ between each pair of population; 0.02 (large solid line), 0.05 to 0.07 (thin solid line) and 0.11 to 0.12 (dashed line).
Figure 3: Population allele-specific expression detection and link with natural selection. a, The deviation $D$ distribution in TARA_209. The red line corresponds to the Gaussian distribution estimated from the data. b, Upset plot of the ASE detection in the seven populations. Each bar of the upper plot corresponds to the number of variants under ASE in the population(s) indicated by black dots in the lower plot. c, Crossing ASE and Selection. The yellow circle represents the total set of variants. In green, the number of heterozygous variants tested for ASE in TARA_209. In blue and red, the amount of detected variants under ASE in TARA_209 and under selection among the populations respectively. In purple, the intersection comprising variants under ASE in TARA_209 and under selection, with its hypergeometric test p-value. d, Metagenomic and metatranscriptomic profiles of variants 841109 and 20760212. Each population is indicated on the x-axis, with the associated B-allele frequency (red) and B-allele relative expression (blue). The frequency is shown on the y-axis. The asterisks mean ASE was detected in the corresponding population.
Figure 4: From allele-specific expression to natural selection. **a**, Evolution of allele frequency and allele relative expression over time. **b**, Evolution of selective pressure over time.