Linking Allele-Specific Expression And Natural Selection In Wild Populations

Romuald Laso-Jadart[1,6*], Kevin Sugier[1], Emmanuelle Petit[2], Karine Labadie[2], Pierre Peterlongo[3], Christophe Ambroise[4], Patrick Wincker[1,6], Jean-Louis Jamet[5], Mohammed-Amin Madoui[1,6*]

[1]Génomique Métabolique, Genoscope, Institut François Jacob, CEA, CNRS, Univ Evry, Université Paris-Saclay, Evry, France.

[2]CEA, Genoscope, Institut de Biologie François Jacob, Université Paris-Saclay, Evry, 91057, France.

[3]Univ Rennes, CNRS, Inria, IRISA - UMR 6074, F-35000 Rennes.

[4]LaMME, CNRS, Univ Evry, Université Paris-Saclay, Evry, France

[5]Université de Toulon, Aix-Marseille Université, CNRS/INSU/IRD, Mediterranean Institute of Oceanology MIO UMR 110, CS 60584, 83041 Toulon cedex 9, France.

[6]Research Federation for the study of Global Ocean Systems Ecology and Evolution, FR2022/Tara Oceans GO-SEE, 3 rue Michel-Ange, 75016 Paris, 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 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

3

59   creating a specific population-level expression pattern of the selected allele (Fig.1b).In the

60   absence of selection, if the same allele is favored by ASE in most of the individuals, the observed

61   population-level expression pattern (Fig. 1c) will be similar to the one observed in the case of

62   selection (Fig.1b).Considering that both mechanisms impact fitness, we hypothesized that ASE

63   and natural selection could preferentially target the same loci, in different populations, showing a

64   possible link between the two mechanisms.

65   In this study, we focus on the widespread epipelagic, temperate and cold water small-sized

66   copepod, *Oithona similis* (Cyclopoida, Claus 1866), notably known to be highly abundant in

67   Arctic[30–33]. Copepods, and particularly *Oithona*, are small crustaceans forming the most abundant

68   metazoan on Earth, reflecting strong adaptive capacities to environmental fluctuations[34–36].They

69   play a key ecological role in biogeochemical cycles and in the marine trophic food chain[37];

70   therefore copepods constitute an ideal model to study wild population evolution[38–41]

71   The first goal of our study was to identify loci under selection before demonstrating that

72   population-level ASE can be detected with metagenomic and metatranscriptomic data collected

73   by the *Tara* Oceans expedition[42] during its Arctic phase. Then we provided evidence of a

74   quantitative link between ASE and natural selection.

Material and Methods

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

77   *Oithona similis* specimens were sampled at the North of the Large Bay of Toulon, France (Lat

78   43°06' 02.3" N and Long 05°56' 53.4"E). Sampling took place in November 2016. The samples

79   were collected from the upper water layers (0-10m) using zooplankton nets with a mesh of 90µm

80   and 200µm (0.5 m diameter and 2.5 m length). Samples were preserved in 70% ethanol and

81   stored at -4°C. From the Large Bay of Toulon samples, *O. similis* individuals were isolated under

82   the stereomicroscope. We selected two different development stages: four copepodites (juveniles)

83   and four adult males. Each individual was transferred separately and crushed, with a tissue

84   grinder (Axygen) into a 1.5 mL tube (Eppendorf). Total mRNAs were extracted using the 'RNA

85   isolation' protocol from NucleoSpin RNA XS kit (Macherey-Nagel) and quantified on a Qubit

86   2.0 with a RNA HS Assay kit (Invitrogen) and on a Bioanalyzer 2100 with a RNA 6000 Pico

87   Assay kit (Agilent). cDNA were constructed using the SMARTer-Seq v4 Ultra low Input RNA

88   kit (ClonTech). The libraries were constructed using the NEBNext Ultra II kit, and were

89   sequenced with an Illumina HiSeq2500 (Supplementary Fig. 1).

Transcriptomes assembly and annotation

91   Each read set was assembled with Trinity v2.5.1[43] using default parameters and transcripts were

92   clustered using cd-hit v4.6.1[44] (Supplementary Table 1).To ensure the classification of the

93   sampled individuals, each ribosomal read set were detected with SortMeRNA [45] and mapped with

94   bwa v0.7.15using default parameters [46] to 82 ribosomal 28S sequences of *Oithona* species used in

95   Cornils et al., 2017 (Supplementary Fig. 2). The transcriptome assemblies were annotated with

96   Transdecoder v5.1.0[43] to predict the open reading frames (ORFs) and protein sequences

97   (Supplementary Table 1). In parallel, homology searches were also included as ORF retention

98   criteria; the peptide sequences of the longest ORFs were aligned on *Oithona nana* proteome [40]

99   using DIAMOND v0.9.22[48]. Protein domain annotation was performed on the final ORF

100   predictions with Interproscan v5.17.56.0[49] and a threshold of e-value $<10^{-5}$ was applied for Pfam

101   annotations. Finally, homology searches of the predicted proteins were done against the nr NCBI

102   database, restricted to Arthropoda (taxid: 6656), with DIAMOND v0.9.22 (Supplementary Fig.

103   1).

104   **Variant calling using *Tara* Oceans metagenomic and metatranscriptomic data**

105   We used metagenomic and metatranscriptomic reads generated from samples of the size fraction

106   20–180 µm collected in seven *Tara* Oceans stations TARA_155, 158, 178, 206, 208, 209 and 210

107   (Supplementary Table 2), according to protocols described in Alberti et al. 2017 [50].

108   The reference-free variant caller *DiscoSNP++* [51,52] was used to extract SNVs simultaneously

109   from raw metagenomic and metatranscriptomic reads, and ran using parameters –b 1. Only SNVs

110   corresponding to biallelic loci with a minimum of 4x of depth of coverage in all stations were

111   initially selected. Then, SNVs were clustered based on their loci co-abundance across samples

112   using density-based clustering algorithm implemented in the R package dbscan [53,54] and ran with

113   parameters epsilon = 10 and minPts = 10. This generated three SNVs clusters, the largest of

114   which contained 102,258 SNVs. To ensure the presence of *O. similis* SNVs without other

115   species, we observed the fitting of the depth of coverage to the expected negative binomial

116   distribution in each population (Supplementary Fig. 3). For each variant in each population, the

117   B-allele frequency (BAF) and the population-level B-allele relative expression (BARE) were

118   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

119   the reference and alternative alleles respectively, T_A and T_B the metatranscriptomic read counts of

120   the reference and alternative alleles respectively.

121   **Variant filtering and annotation**

122   SNVs were filtered based on their metagenomic coverage. Those with a metagenomic coverage

123   lying outside a threshold of median $\pm$ 2 σ in at least one population, with a minimum and

124   maximum of 5x and 150x coverage were discarded. To keep out rare alleles and potential calling

125   errors, only variants characterized by a BAF comprised between 0.9 and 0.1, and a BARE

126   between 0.95 and 0.05 in at least one population were chosen for the final dataset resulting in

127   25,768 SNVs (Supplementary Fig. 4).

128   The variant annotation was conducted in two steps. First, the variant sequences were relocated on

129   the previously annotated *O.similis* transcripts using the "VCF_creator.sh" program of

130   *DiscoSNP++*. Secondly, a variant annotation was carried with SNPeff [55] to identify the location

131   of variants within transcripts (i.e., exon or UTR) and to estimate their effect on the proteins

132   (missense, synonymous or nonsense).The excess of candidate variant annotations was tested in

133   the following classes: missenses, synonymous, 5' and 3'UTR. A significant excess was

134   considered for a hypergeometric test p-value < 0.05.

135   **Genomic differentiation and detection of selection**

136   The differentiation among the seven populations was investigated through the computation of the

137   F_{ST} metric or Wright's fixation index [56,57].For each locus, global F_{ST} including the seven

138   populations and pairwise-F_{ST} between each pair of population was computed, using the

139   corresponding BAF matrix. For the global F_{ST} computation, a Hartigans' dip test for unimodality

140   was performed [58]. We retained the median pairwise-F_{ST} as a measure of the genomic

141   differentiation between each population. The *pcadapt* R package v4.0.2 [59] was used to detect

142   selection among populations from the B-allele frequency matrix. The computation was run on

143   "*Pool-seq*" mode, with a minimum allele frequency of 0.05 across the populations, and variants

144   with a corrected Benjamini and Hochberg [60] 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 BAF $\neq \{0,1\}$. Then, we computed $D = BAF\text{-}BARE$, 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[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-BAF and 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, 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.

167 **Gene enrichment analysis**

168 To identify specific biological function or processes associated to the variants, a domain-based

169 analysis was conducted. The Pfam annotation of the transcripts carrying variants targeted by ASE

170 and selection was used as entry for dcGO Enrichment [62]. A maximum of the best 300 GO-terms

171 were chosen based on their z-score and FDR p-value ($<10^{-3}$) in each ontology category. To

172 reduce redundancy, these selected GO-terms were processed using REVIGO[63], with a similarity

173 parameter of 0.5 against the whole Uniprot catalogue under the SimRel algorithm. To complete

174 the domain-based analysis, the functional annotations obtained from the homology searches

175 against the nr were manually curated.

Results

177 *Oithona similis* **genomic differentiation and selection in Arctic Seas**

178 From metagenomic and metatranscriptomic raw data of seven sampling stations (Fig. 2a), we

179 identified 25,768 expressed variants. Among them, 97% were relocated on *O. similis*

180 transcriptomes.

181 The global distribution of F_{ST} of the seven populations was unimodal (Hartigans' dip test,

182 D=0.0012, p-value=0.99) with a median-F_{ST} at 0.1, confirmed by the pairwise-F_{ST} distributions

183 (Supplementary Fig. 5).The seven populations were globally characterized by a weak to moderate

184 differentiation, with a maximum median pairwise-F_{ST} of 0.12 between populations from

185 TARA_210 and 155/178 (Fig. 2c,d). Populations from stations TARA_158 (Norway Current),

186 206 and 208 (Baffin Bay) were genetically closely-related, with the lowest median pairwise F_{ST}

187 (0.02), despite TARA_158 did not co-geolocalize with the two other stations. The four other

188 populations (TARA_155, 178, 209, and 210) were equally distant from each other (0.1-0.12).

189   Finally, TARA_158, 206 and 208 on one side, and TARA_155, 178, 210 and 209 on the other

190   side showed the same pattern of differentiation (0.05-0.07).

191   The PCA decomposed the genomic variability in six components; the first two components

192   discriminated TARA_155 and 178 from the others (32% and 28.1% variance explained

193   respectively, Fig. 2b), and the third component differentiated TARA_210 and 209 (19.5%). The

194   fourth principal component separated TARA_209 and 210 from 158/206/208 (11.3 %), with the

195   last two concerning TARA_158/206/208 (Supplementary Fig. 5).Globally, these results

196   dovetailed with the F_{ST} analysis, with details discussed later. Finally, we detected 674 variants

197   under selection, representing 2.6% of the dataset (corrected p-value < 0.05).

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

199   The number of SNVs tested for ASE varied between 13,454 and 22,578 for TARA_210 and 206

200   respectively. As expected, the *D* deviation, representing the deviation between B-allele frequency

201   and B-allele relative expression, followed a Gaussian distribution in each population (Fig. 3a and

202   Supplementary Fig. 6). Variants under ASE (i.e. having a *D* significantly higher or lower than

203   expected) were found in every population, ranging from 26 to 162 variants for TARA_178 and

204   206 respectively (Table 1). Overall, we found 587 variants under ASE, including 535 population-

205   specific ASEs, and 52 ASEs shared by several populations (Fig.3b). Remarkably, 30 ASEs out of

206   the 52 were present in the populations from TARA_158, 206 and 208 that correspond to the

207   genetically closest populations. The seven sets of variants under ASE were crossed with the set of

208   variants under selection, as illustrated for TARA_209 (Fig. 3c). The size of the intersection

209   ranged from 5 to 42 variants (TARA_155 and 210/206) and was significantly higher than

210   expected by chance for all the populations (hypergeometric test p-value < 0.05). It represented a

211   total of 152 unique variants under selection and ASE in at least one population (Table 1,

10

212   Supplementary Table 3), corresponding to 23% and 26% of variants under ASE and under

213   selection respectively.

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

215   Among the 152 loci targeted by ASE and selection, 145 were relocated on *O. similis* transcripts

216   (Supplementary Table 4). Amid these transcripts, 137 (90%) had a predicted ORF, 97 (64%)

217   were linked to at least one Pfam domain and 90 (59%) to a functional annotation. Fifteen SNVs

218   were missense variations, 59 synonymous, 31 and 29 were located in 5' and 3' UTR, without any

219   significant excess (Supplementary Table 4 and 5). Based on homology searches (Supplementary

220   Table 4), eight genes were linked to nervous system (Table2). Among them, two genes were

221   involved in glutamate metabolism (omega-amidase NIT2 and5-oxoprolinase), three were

222   predicted to be glycine, γ-amino-butyric acid (GABA) and histamine neuroreceptors. Finally,

223   four were also implicated in arthropods photoreceptors. The domain-based analysis confirmed

224   these results, with an enrichment in GO-terms biological process also linked to nervous system

225   (Supplementary Fig. 7).

Discussion

226

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

228   Global populations of *O. similis* are known to be composed of cryptic lineages[47]. Thus the

229   assessment that the seven populations used in our study belong to the same *O. similis* cryptic

230   lineage was a prerequisite for further analyses. The high proportion of variants mapped on the

231   Mediterranean transcriptomes (97%) showed that the variant clustering method was efficient to

232   regroup loci of an *O.similis* cryptic species. Plus, the unimodal distribution of F_{ST} showed that

233   these populations of *O. similis* belong to the same polar cryptic species.

234   Secondly, we see that the seven populations examined are well connected with low median

235   pairwise-F_{ST}, despite the large distances separating them. Weak genetic structure in the polar

236   region was already highlighted for other major Arctic copepods like *Calanus glacialis*[64], and

237   *Pseudocalanus* species[65]. The absence of structure was explained by ancient diminutions of

238   effective population size due to past glaciations[65–67], or high dispersal and connectivity between

239   the present-day populations due to marine currents[64].

240   Going into details, three different cases can be described. First, the differentiation of populations

241   from TARA_155 and 178 compared to the others could be explained by isolation-by-distance.

242   Secondly, the geographically close populations from TARA_210 and 209 present higher

243   differentiation (median pairwise-F_{ST} of 0.11). This could be explained by the West Greenland

244   current acting as a physical barrier between the populations, which could lead to reduced gene

245   flow[68]. At last, the strong link between TARA_158 from Northern Atlantic current and

246   TARA_206/208 from the Baffin Bay is the most intriguing. Despite the large distances that

247   separate the first one from the others, these three populations are well connected.

248   Metagenomic data enable to draw the silhouette of the population genetics but lacks resolution

249   when dealing with intra-population structure. However, our findings are concordant with

250   previous studies underpinning the large-scale dispersal, interconnectivity of marine zooplankton

251   populations in other oceans, at diverse degrees[38,69–71].

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

253   Usually, at the individual level, the ASE analyses are achieved by measuring the difference in

254   RNA-seq read counts of a heterozygous site. But at the population level, obtaining a large

255   number of individuals remains a technical barrier especially for uncultured animals, or when the

12

256   amount of DNA retrieved from a single individual is not sufficient for high-throughput

257   sequencing. Here, the detection of ASE at population level was possible by comparing the

258   observed frequencies of the alleles based on metagenomic and metatranscriptomic data, which by

259   passes the obstacles previously described.

260   In our study, the amount of detected ASE in each population was always lower than 1% of tested

261   heterozygous variants, which altogether correspond to 2% of the total set of variants. In humans

262   [72],baboons [21] and flycatchers [19], 17%, 23% of genes and 7.5% of transcripts were affected by

263   ASE respectively. The difference with our results can be explained by one main reason. The

264   detection of population-level ASE identifies only the ASE present in a large majority of

265   individuals, which can be considered as "core ASEs".

266   These core ASEs constitute the majority of detected ASEs and are population-specific, meaning

267   the main drivers of this expression pattern are local conditions like different environmental

268   pressures or population dynamics including, for example, the proportion of each developmental

269   stage and sex, known to vary between populations and across seasons [30,73]. Another result is the

270   presence of a small amount of variants affected by ASE in different populations. Most of these

271   variants are under ASE in at least two of the three closest populations from TARA_158, 206 and

272   208. First, the genetic closeness and large geographic distances between these three populations

273   suggest that their shared ASEs are under an environmental independent genetic control.

274   Secondly, the number of variants tested for ASE is higher in these three populations than the

275   others, leading to a greater proportion of ASEs detected which also elevates the chances for a

276   variant to be declared under ASE in several populations.

13

277 A significant amount of SNVs (152) were subject to selection among the seven populations and

278 to ASE in at least one population. We confirmed our first hypothesis (Fig. 1), as exampled with

279 the variant 841109 (Fig. 3d), characterized by an ASE in favor of the B-allele in TARA_209 and

280 fixation of this allele in TARA_210. Three main features of ASE can be under selection. First,

281 the observed variant can be in linkage with another variation in upstream *cis*-regulatory elements

282 like transcription factors fixation sites, or epialleles[7]. Secondly, the annotation of candidate

283 variants with SNPeff revealed a majority of variants located in 5' and 3'UTRs, which are

284 variations known to both affect transcription efficiency through mRNA secondary structures,

285 stability and location[74–76]. For variants located in exons, a majority were identified as

286 synonymous mutations, growingly described as potential target of selection by codon usage bias,

287 codon context, mRNA secondary structure or transcription and translation dynamics [77,78]. Finally,

288 fifteen missense mutations were spotted, but with moderate predicted impact on protein amino

289 acid composition. However, we did not find premature nonsense mutation, even if variants under

290 ASE has been described to trigger or escape potential nonsense-mediated decay [28,61,79], but the

291 possibility that the causal variation is located in introns cannot be ruled out.

292 The process of adaptation through gene expression was suspected in human populations and

293 investigated thanks to the large and accessible amount of data. In a first study, a link has been

294 established between gene expression and selection, affecting particular genes and phenotypes,

295 looking at *cis*-acting SNPs[80]. In a second study, the team was able to detect ASE in different

296 populations and to quantify genetic differentiation and selection[61]. They found particularly one

297 gene that shows strong differentiation between European and African populations and under ASE

298 in Europeans and not in Africans. However, they did not quantify this phenomenon. Both

299 emphasized the impact of selection on gene expression. In the same way, another approach

14

300   showed that ASE or expression variations with high effect size were rare in the populations,

301   based on intra-population analyses in *Capsella grandiflora* and human [28,81]. This situation is

302   presumably encountered in our analysis, as exampled with the B-allele of variant 20760212,

303   under ASE and with a low genomic frequency in TARA_210, but fixed in the others (Fig. 3d).

304   Our results complete previous analyses, as they quantify the link between ASE and selection in

305   populations and reveal the evolutive potency of ASE, for the first time at the population-level. It

306   remains to understand the nature of the association between ASE and selection. To address this

307   question, we formulate the hypothesis that they impact chronologically the same loci, following

308   constant or increasing selective pressure as well as environmental changes (Fig. 4).

309   **Nervous system and visual perception are important targets of the natural selection and**

310   **ASE in *O. similis***

311   This evolutive link between ASE and selection is supported by the biological functions associated

312   to the targeted genes, which are involved notably in the copepods nervous system in two ways.

313   The first result is the presence of genes implicated in glutamate metabolism and glycine and/or

314   GABA receptors. Glutamate and GABA are respectively excitatory and inhibitory

315   neurotransmitters in arthropods motor neurons [82]. Plus, glycine and GABA receptors have

316   already been described as a target of selection in *O. nana* in Mediterranean Sea[40,41]. Secondly, the

317   functional analysis revealed also the importance of the eye and visual perception in the *O. similis*

318   evolution.

319   Copepod nervous system constitutes a key trait for its reproduction and survival, and based on

320   our data, a prime target for evolution, allowing higher capacity of perceiving and fast reacting

321   leading to more efficient predator escape, prey catching and mating. This can explain the great

322   evolutive success of these animals [35,83,84].

15

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

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

1. Szabo, P. E. & Mann, J. R. Allele-specific expression and total expression levels of imprinted genes during early mouse development: implications for imprinting mechanism. *Genes Dev.* **9,** 3097–3108 (1995).

2. Wei, X. & Wang, X. A computational workflow to identify allele-specific expression and epigenetic modification in maize. *Genomics, Proteomics Bioinforma.* **11,** 247–252 (2013).

3. Ginart, P. *et al.* Visualizing allele-specific expression in single cells reveals epigenetic mosaicism in an H19loss-of-imprinting mutant. *Genes Dev.* **30,** 567–578 (2016).

4. Lonsdale, Z. *et al.* Allele specific expression and methylation in the bumblebee, *Bombus terrestris*. *PeerJ* **5,** e3798 (2017).

5. Bailey, S. D., Virtanen, C., Haibe-Kains, B. & Lupien, M. ABC: A tool to identify SNVs causing allele-specific transcription factor binding from ChIP-Seq experiments. *Bioinformatics* **31,** 3057–3059 (2015).

6. Cavalli, M. *et al.* Allele-specific transcription factor binding to common and rare variants associated with disease and gene expression. *Hum. Genet.* **135,** 485–497 (2016).

7. Castel, S. E., Levy-Moonshine, A., Mohammadi, P., Banks, E. & Lappalainen, T. Tools and best practices for data processing in allelic expression analysis. *Genome Biol.* **16,** 195 (2015).

8. Rivas, M. A. *et al.* Impact of predicted protein-truncating genetic variants on the human

17

370   transcriptome. *Science (80-.).* **348,** 666–669 (2015).

371   9.   Pirinen, M. *et al.* Assessing allele-specific expression across multiple tissues from RNA-
372   seq read data. *Bioinformatics* **31,** 2497–2504 (2015).

373   10.   Lu, R. *et al.* Analyzing allele specific RNA expression using mixture models. *BMC*
374   *Genomics* **16,** 566 (2015).

375   11.   Harvey, C. T. *et al.* QuASAR: Quantitative allele-specific analysis of reads.
376   *Bioinformatics* **31,** 1235–1242 (2015).

377   12.   Miao, Z., Alvarez, M., Pajukanta, P. & Ko, A. ASElux: An ultra-fast and accurate allelic
378   reads counter. *Bioinformatics* **34,** 1313–1320 (2018).

379   13.   Mayba, O. *et al.* MBASED: allele-specific expression detection in cancer tissues and cell
380   lines. *Genome Biol.* **15,** 405 (2014).

381   14.   Skelly, D. A., Johansson, M., Madeoy, J., Wakefield, J. & Akey, J. M. A powerful and
382   flexible statistical framework for testing hypotheses of allele-specific gene expression
383   from RNA-seq data. *Genome Res.* **21,** 1728–1737 (2011).

384   15.   M. Dong, Y. J. Single-Cell Allele-Specific Gene Expression Analysis. *Comput. Methods*
385   *Single-Cell Data Anal.* **1935,** (2019).

386   16.   Ge, B. *et al.* Global patterns of cis variation in human cells revealed by high-density allelic
387   expression analysis. *Nat. Genet.* **41,** 1216–1222 (2009).

388   17.   Signor, S. A. & Nuzhdin, S. V. The Evolution of Gene Expression in cis and trans. *Trends*
389   *Genet.* 1–13 (2018). doi:10.1016/j.tig.2018.03.007

390   18.   McKean, D. M. *et al.* Loss of RNA expression and allele-specific expression associated
391   with congenital heart disease. *Nat. Commun.* **7,** 1–9 (2016).

392   19.   Wang, M., Uebbing, S. & Ellegren, H. Bayesian inference of allele-specific gene
393   expression indicates abundant Cis-regulatory variation in natural flycatcher populations.
394   *Genome Biol. Evol.* **9,** 1266–1279 (2017).

395   20.   Howe, B., Umrigar, A. & Tsien, F. Chromosome Preparation From Cultured Cells. *J. Vis.*
396   *Exp.* 3–7 (2014). doi:10.3791/50203

397   21.   J. Tung, M. Y. Akinyi, S. Mutura, J. Altmann, G. A. W. and S. C. & Alberts. Allele-
398   specific gene expression in a wild nonhuman primate population. *Mol. Ecol.* **2,** 147–185
399   (2015).

400   22.   Buil, A. *et al.* Gene-gene and gene-environment interactions detected by transcriptome
401   sequence analysis in twins. *Nat. Genet.* **47,** 88–91 (2014).

402   23.   Cheung, V. G. *et al.* Monozygotic Twins Reveal Germline Contribution to Allelic
403   Expression Differences. *Am. J. Hum. Genet.* **82,** 1357–1360 (2008).

404   24.   Moyerbrailean, G. A. *et al.* High-throughput allele-specific expression across 250

405   environmental conditions. *Genome Res.* **26,** 1627–1638 (2016).

406   25.   Knowles, D. A. *et al.* Allele-specific expression reveals interactions between genetic
407   variation and environment. *Nat. Methods* **14,** 699–702 (2017).

408   26.   Leon-Novelo, L., Gerken, A. R., Graze, R. M., McIntyre, L. M. & Marroni, F. Direct
409   Testing for Allele-Specific Expression Differences Between Conditions. *G3 GENES,*
410   *GENOMES, Genet.* **8,** g3.300139.2017 (2017).

411   27.   Jensen, J. D. *et al.* The importance of the neutral theory in 1968 and 50 years on: a
412   response to Kern & Hahn 2018. *Evolution (N. Y).* 1968–1971 (2018).
413   doi:10.1111/evo.13650

414   28.   Lappalainen, T. *et al.* Transcriptome and genome sequencing uncovers functional variation
415   in humans. *Nature* **501,** 506–511 (2013).

416   29.   Claus, C. Die Copepoden-Fauna von Nizza. Ein Beitrag zur Charakteristik der Formen und
417   deren Abänderungen 'im Sinna Darwin's'. *Elwebt'sche Univ. Marbg. Leipzig* **1,** 1:34
418   (1866).

419   30.   Dvoretsky, V. G. Seasonal mortality rates of Oithona similis (Cyclopoida) in a large Arctic
420   fjord. *Polar Sci.* **6,** 263–269 (2012).

421   31.   Castellani. Contribution to the Themed Section□: ' The Role of Zooplankton in Marine
422   Biogeochemical Cycles□: From Fine Scale to Global Marine zooplankton and the
423   Metabolic Theory of Ecology□: is it a predictive tool□? *J. Plankton Res.* **38,** 762–770
424   (2016).

425   32.   Blachowiak-Samolyk, K., Kwasniewski, S., Hop, H. & Falk-Petersen, S. Magnitude of
426   mesozooplankton variability: A case study from the Marginal Ice Zone of the Barents Sea
427   in spring. *J. Plankton Res.* **30,** 311–323 (2008).

428   33.   Zamora-Terol, S., Nielsen, T. G. & Saiz, E. Plankton community structure and role of
429   Oithona similis on the western coast of Greenland during the winter-spring transition. *Mar.*
430   *Ecol. Prog. Ser.* **483,** 85–102 (2013).

431   34.   Humes, A. G. How Many Copepods? *Hydrobiologia* **293,** 1–7 (1994).

432   35.   Kiørboe, T. What makes pelagic copepods so successful? *J. Plankton Res.* **33,** 677–685
433   (2011).

434   36.   Gallienne, C. P. Is Oithona the most important copepod in the world's oceans? *J. Plankton*
435   *Res.* **23,** 1421–1432 (2001).

436   37.   Wassmann, P. *et al.* Food webs and carbon flux in the Barents Sea. *Prog. Oceanogr.* **71,**
437   232–287 (2006).

438   38.   Peijnenburg, K. T. C. A. & Goetze, E. High evolutionary potential of marine zooplankton.
439   *Ecol. Evol.* **3,** 2765–2781 (2013).

440   39.   Riginos, C., Crandall, E. D., Liggins, L., Bongaerts, P. & Treml, E. A. Navigating the

19

441 currents of seascape genomics: How spatial analyses can augment population genomic
442 studies. *Curr. Zool.* **62,** 581–601 (2016).

443 40. Madoui, M. A. *et al.* New insights into global biogeography, population structure and
444   natural selection from the genome of the epipelagic copepod Oithona. *Mol. Ecol.* **26,**
445   4467–4482 (2017).

446 41. Arif, M. *et al.* Discovering Millions of Plankton Genomic Markers from the Atlantic
447   Ocean and the Mediterranean Sea. *Mol. Ecol. Resour.* 0–3 (2018). doi:10.1111/1755-
448   0998.12985

449 42. Karsenti, E. *et al.* A Holistic Approach to Marine Eco-Systems Biology. *PLoS Biol.* **9,**
450   e1001177 (2011).

451 43. Haas, B. J. *et al.* De novo transcript sequence reconstruction from RNA-seq using the
452   Trinity platform for reference generation and analysis. *Nat. Protoc.* **8,** 1494–1512 (2013).

453 44. Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. CD-HIT: Accelerated for clustering the next-
454   generation sequencing data. *Bioinformatics* **28,** 3150–3152 (2012).

455 45. Kopylova, E., Noé, L. & Touzet, H. SortMeRNA: Fast and accurate filtering of ribosomal
456   RNAs in metatranscriptomic data. *Bioinformatics* **28,** 3211–3217 (2012).

457 46. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler
458   transform. *Bioinformatics* **26,** 589–595 (2009).

459 47. Cornils, A., Wend-Heckmann, B. & Held, C. Global phylogeography of Oithona similis
460   s.l. (Crustacea, Copepoda, Oithonidae) – A cosmopolitan plankton species or a complex of
461   cryptic lineages? *Mol. Phylogenet. Evol.* **107,** 473–485 (2017).

462 48. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using
463   DIAMOND. *Nat. Methods* **12,** 59–60 (2014).

464 49. Jones, P. *et al.* InterProScan 5: genome-scale protein function classification.
465   *Bioinformatics* **30,** 1236–1240 (2014).

466 50. Alberti, A. *et al.* Viral to metazoan marine plankton nucleotide sequences from the Tara
467   Oceans expedition. *Sci. Data* **4,** 170093 (2017).

468 51. Uricaru, R. *et al.* Reference-free detection of isolated SNPs. *Nucleic Acids Res.* (2014).
469   doi:10.1093/nar/gku1187

470 52. Peterlongo, P., Riou, C., Drezen, E. & Lemaitre, C. DiscoSnp++: de novo detection of
471   small variants from raw unassembled read set(s). *bioRxiv* 209965 (2017).
472   doi:10.1101/209965

473 53. Ester, M., Kriegel, H.-P., Sander, J. & Xu, X. *A Density-Based Algorithm for Discovering
474   Clusters in Large Spatial Databases with Noise*. (1996).

475 54. Ram, A., Jalal, S., Jalal, A. S. & Kumar, M. A Density Based Algorithm for Discovering
476   Density Varied Clusters in Large Spatial Databases. *Int. J. Comput. Appl.* **3,** 1–4 (2010).

477 55. Cingolani, P. and Platts, A. and Coon, M. and Nguyen, T. and Wang, L. and Land, S.J. and
478   Lu, X. and Ruden, D. M. A program for annotating and predicting the effects of single
479   nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain
480   w1118□; iso-2; iso-3. *Fly* **6,** 80–92 (2012).

481 56. B. S. Weir and C. Clark Cockerham. Estimating F-Statistics for the Analysis of Population
482   Structure. *Evolution (N. Y).* **38,** 1358–1370 (1984).

483 57. Wright, S. the Genetical Structure of Populations. *Ann. Eugen.* **15,** 323–354 (1951).

484 58. Hartigan, J. A. & Hartigan, P. M. The dip test of unimodality. *Ann. Stat.* **13,** 70–84 (1985).

485 59. Luu, K., Bazin, E. & Blum, M. G. B. pcadapt: an R package to perform genome scans for
486   selection based on principal component analysis. *Mol. Ecol. Resour.* **17,** 67–77 (2017).

487 60. Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate□: A Practical and
488   Powerful Approach to Multiple Testing. *J. R. Stat. Soc. Ser. B* **57,** 289–300 (1995).

489 61. Tian, L. *et al.* Genome-wide comparison of allele-specific gene expression between
490   African and European populations. *Hum. Mol. Genet.* **27,** 1067–1077 (2018).

491 62. Fang, H. & Gough, J. DcGO: Database of domain-centric ontologies on functions,
492   phenotypes, diseases and more. *Nucleic Acids Res.* **41,** (2013).

493 63. Supek, F., Bošnjak, M., Škunca, N. & Šmuc, T. Revigo summarizes and visualizes long
494   lists of gene ontology terms. *PLoS One* **6,** (2011).

495 64. Weydmann, A., Coelho, N. C., Serrão, E. A., Burzyński, A. & Pearson, G. A. Pan-Arctic
496   population of the keystone copepod Calanus glacialis. *Polar Biol.* **39,** 2311–2318 (2016).

497 65. Aarbakke, O. N. S., Bucklin, A., Halsband, C. & Norrbin, F. Comparative phylogeography
498   and demographic history of five sibling species of Pseudocalanus (Copepoda: Calanoida)
499   in the North Atlantic Ocean. *J. Exp. Mar. Bio. Ecol.* **461,** 479–488 (2014).

500 66. Edmands, S. Phylogeography of the intertidal copepod Tigriopus californicus reveals
501   substantially reduced population differentiation at northern latitudes. *Mol. Ecol.* **10,** 1743–
502   1750 (2001).

503 67. Bucklin, A. & Wiebe, P. H. Low mitochondrial diversity and small effective population
504   sizes of the copepods Calanus finmarchicus and Nannocalanus minor: Possible impact of
505   climatic variation during recent glaciation. *J. Hered.* **89,** 383–392 (1998).

506 68. Myers, P. G., Donnelly, C. & Ribergaard, M. H. Structure and variability of the West
507   Greenland Current in Summer derived from 6 repeat standard sections. *Prog. Oceanogr.*
508   (2008). doi:10.1016/j.pocean.2008.12.003

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

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

513   Calanoides natalis (Copepoda, Calanoida) in the eastern Atlantic Ocean and Benguela
514   upwelling system. *J. Plankton Res.* **39,** 618–630 (2017).

515   71.   Goetze, E. Global Population Genetic Structure and Biogeography of the Oceanic
516   Copepods Eucalanus Hyalinus and E. Spinifer. *Evolution (N. Y).* **59,** 2378 (2005).

517   72.   Zhang, S. *et al.* Genome-wide identification of allele-specific effects on gene expression
518   for single and multiple individuals. *Gene* **533,** 366–373 (2014).

519   73.   Lischka, S. & Hagen, W. Life histories of the copepods Pseudocalanus minutus, P. acuspes
520   (Calanoida) and Oithona similis (Cyclopoida) in the Arctic Kongsfjorden (Svalbard).
521   *Polar Biol.* **28,** 910–921 (2005).

522   74.   Mignone, F., Gissi, C., Liuni, S., Pesole, G. & others. Untranslated regions of mRNAs.
523   *Genome Biol* **3,** 4–1 (2002).

524   75.   Dvir, S. *et al.* Deciphering the rules by which 5'-UTR sequences affect protein expression
525   in yeast. *Proc. Natl. Acad. Sci.* **110,** E2792–E2801 (2013).

526   76.   Matoulkova, E., Michalova, E., Vojtesek, B. & Hrstka, R. The role of the 3' untranslated
527   region in post-transcriptional regulation of protein expression in mammalian cells. *RNA*
528   *Biol.* **9,** 563–576 (2012).

529   77.   Shabalina, S. A., Spiridonov, N. A. & Kashina, A. Sounds of silence: Synonymous
530   nucleotides as a key to biological regulation and complexity. *Nucleic Acids Res.* **41,** 2073–
531   2094 (2013).

532   78.   Ingvarsson, P. K. Natural Selection on Synonymous and Nonsynonymous Mutations
533   Shapes Patterns of Polymorphism in Populus tremula. *Mol. Biol. Evol.* **27,** 650–660
534   (2010).

535   79.   Rivas, M. A. *et al.* Effect of predicted protein-truncating genetic variants on the human
536   transcriptome. *Science (80-.).* **348,** 666–669 (2015).

537   80.   Fraser, H. B. Gene expression drives local adaptation in humans Gene expression drives
538   local adaptation in humans. *Genome Res.* 1089–1096 (2013). doi:10.1101/gr.152710.112

539   81.   Josephs, E. B., Lee, Y. W., Stinchcombe, J. R. & Wright, S. I. Association mapping
540   reveals the role of purifying selection in the maintenance of genomic variation in gene
541   expression. *Proc. Natl. Acad. Sci.* **112,** 15390–15395 (2015).

542   82.   Smarandache-Wellmann, C. R. Arthropod neurons and nervous system. *Curr. Biol.* **26,**
543   R960–R965 (2016).

544   83.   Svensen, C. Remote prey detection in Oithona similis: hydromechanical versus chemical
545   cues. *J. Plankton Res.* **22,** 1155–1166 (2000).

546   84.   Kiørboe, T., Andersen, A., Langlois, V. J. & Jakobsen, H. H. Unsteady motion: Escape
547   jumps in planktonic copepods, their kinematics and energetics. *J. R. Soc. Interface* **7,**
548   1591–1602 (2010).

549 85. Denno, M. E., Privman, E., Borman, R., Wolin, D. & Venton, B. J. Quantification of
550   histamine and carcinine in Drosophila melanogaster tissues. *ACS Chem Neurosci* **7,** 407–
551   414 (2016).

552 86. Monastirioti, M. Biogenic amine systems in the fruit fly Drosophila melanogaster.
553   *Microsc. Res. Tech.* **45,** 106–121 (1999).

554 87. Stuart, A. E. From fruit flies to barnacles, histamine is the neurotransmitter of arthropod
555   photoreceptors. *Neuron* **22,** 431–433 (1999).

556 88. Gurudev, N., Yuan, M. & Knust, E. chaoptin, prominin, eyes shut and crumbs form a
557   genetic network controlling the apical compartment of Drosophila photoreceptor cells.
558   *Biol. Open* **3,** 332–341 (2014).

559 89. Krantz, D. E. & Zipursky, S. L. Drosophila chaoptin, a member of the leucine-rich repeat
560   family, is a photoreceptor cell-specific adhesion molecule. *EMBO J.* **9,** 1969–77 (1990).

561 90. Masai, I., Okazakit, A., Hosoyat, T. & Hottatt, Y. Drosophila retinal degeneration A gene
562   encodes an eye-specific diacylglycerol kinase with cysteine-rich zinc-finger motifs and
563   ankyrin repeats (signal transduction/phosphatidylinositol metabolism). *Neurobiology* **90,**
564   11157–11161 (1993).

565 91. Wang, T. & Montell, C. Phototransduction and retinal degeneration in Drosophila.
566   *Pflugers Arch. Eur. J. Physiol.* **454,** 821–847 (2007).

567 92. Rawls, A. S. Strabismus requires Flamingo and Prickle function to regulate tissue polarity
568   in the Drosophila eye. *Development* **130,** 1877–1887 (2003).

569 93. Leung, V. *et al.* The planar cell polarity protein Vangl2 is required for retinal axon
570   guidance. *Dev. Neurobiol.* **76,** 150–165 (2016).

571

Tables

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

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

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

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

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

580 **Supplementary Table 4:** Variants targeted by ASE and selection functional annotations

581   **Supplementary Table 5**: Variant annotation by SNPeff

582   # Figures

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

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

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

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

610   **Supplementary Fig 1:** Method pipeline overview

611   **Supplementary Fig 2:** Validation of taxonomic assignation

612   **Supplementary Fig 3:** *Oithona similis* depth of coverage of biallelic loci in seven *Tara* Oceans
613   samples

614   **Supplementary Fig 4:** Metagenomic coverage distribution of the seven *Tara* Oceans samples

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

616   **Supplementary Fig 6:** Allele-specific expression detection

617   **Supplementary Fig 7:** Functional analysis of *O. similis* transcripts targeted by ASE and
618   selection

619

620

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

Population	Genomic median depth of coverage	Number of tested variants	Number of variants under ASE	Number of variants under ASE and selection	Hypergeometric test p-value
TARA_155	25	18,812	91	6	9.89E-3*
TARA_158	35	21,476	131	29	5.06E-20*
TARA_178	24	18,145	26	9	5E-11*
TARA_206	55	22,578	162	42	4.82E-31*
TARA_208	48	21,469	133	14	2.2E-6*
TARA_209	12	13,956	62	24	1.05E-23*
TARA_210	14	13,454	69	42	8.89E-51*
Overall	-	25,768	587 (2.3%)	152 (0.59%)	-

622

623

624

26

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

| VarID	Ref	Alt	Homology search	Pfam	SnpEff Localization	SnpEff Impact	References
722267	A	G	histamine H1 receptor	PF00001	3 ' UTR	MODIFIER	85–87
9665345	T	G	chaoptin	PF13306 | PF13855	synonymous variant	LOW	88,89
15623788	G	A	eye-specific diacylglycerol kinase	PF13637	synonymous variant	LOW	90,91
23795359	A	T	vang-like protein 2-B	PF06638	synonymous variant	LOW	92,93
1276227	C	T	glycine receptor subunit alpha-2 / gamma-aminobutyric acid receptor subunit alpha-6	PF02932 | PF2931	3 ' UTR	MODIFIER	-
1404415	G	C	omega-amidase NIT2	PF00795	3 ' UTR	MODIFIER	-
11174785	A	G	5-oxoprolinase	PF02538 | PF05378 | PF01968	5 ' UTR	MODIFIER	-
11690229	A	T	glycine receptor subunit alpha-2	PF02931	synonymous variant	LOW	-

625
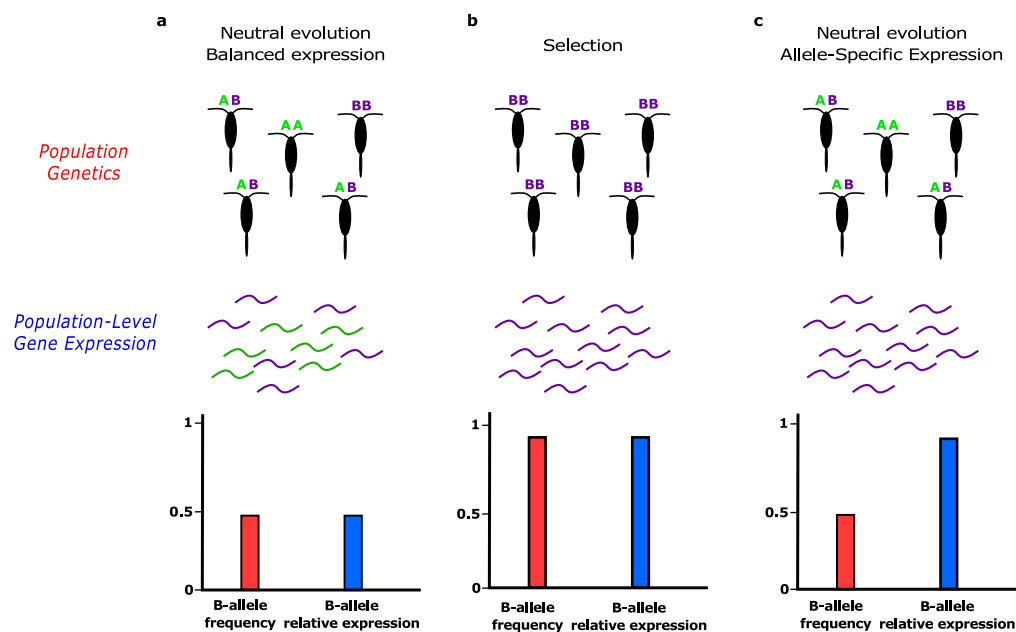
626

627

27

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.
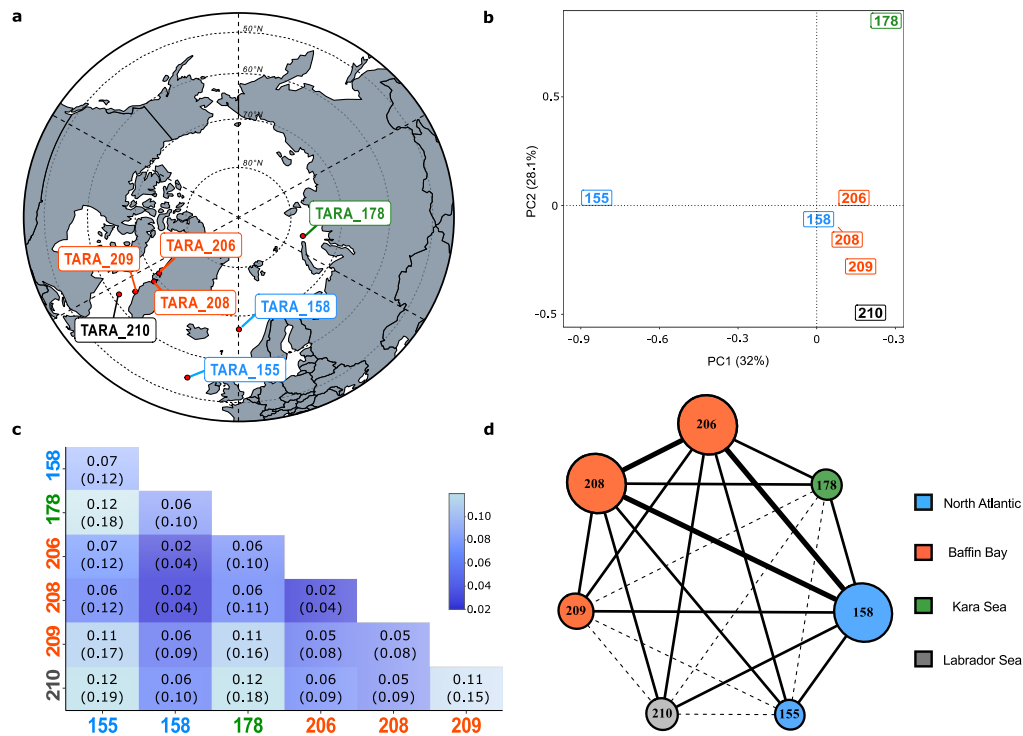
Figure 2: Genomic differentiation of *O. similis* populations from Arctic Seas. **a,** Geographic locations of the seven *Tara* Oceans sampling sites: Northern Atlantic (blue), Kara Sea (green), Baffin Bay (orange) and Labrador Sea (grey). **b,** Principal Component Analysis (PCA) computed by *pcadapt* 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).
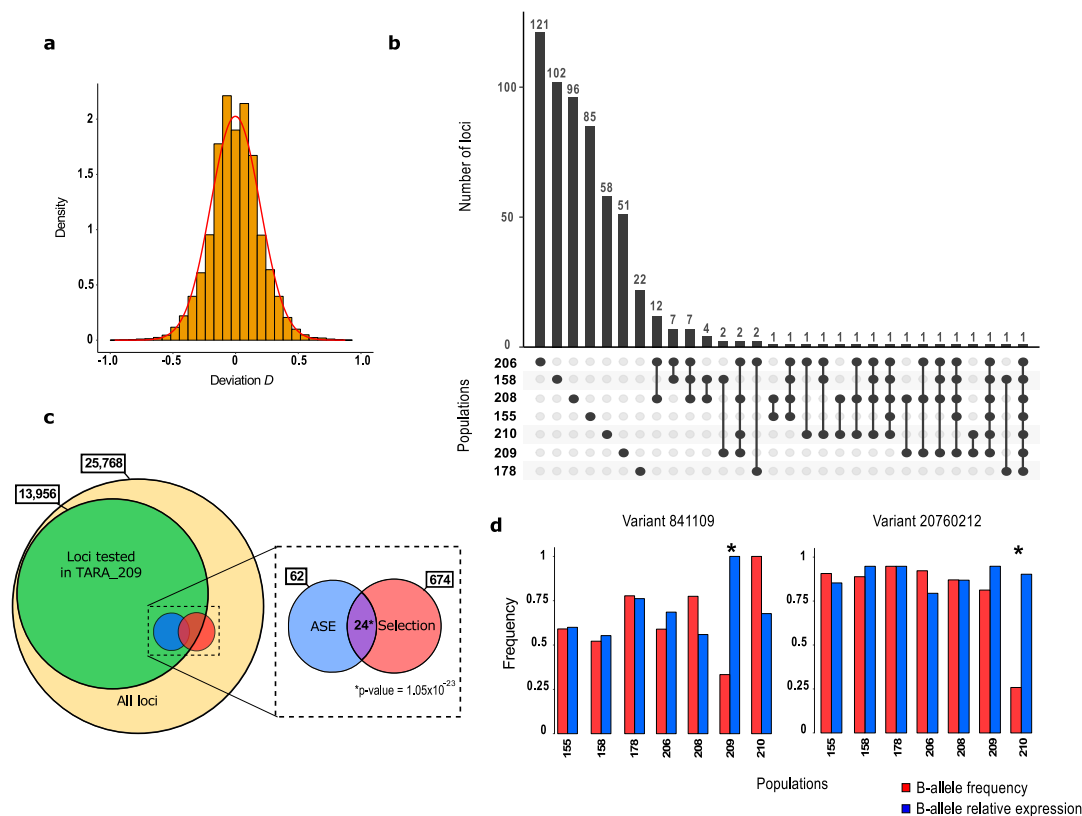
Figure 3: Population allele-specific expression detection and link with natural selection. **a,**The deviation *D* distribution in TARA_209. The red line corresponds to the Gaussian distribution estimated from the data. **b,** Upset plot of the ASE detection in the seven populations. Each bar of the 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.
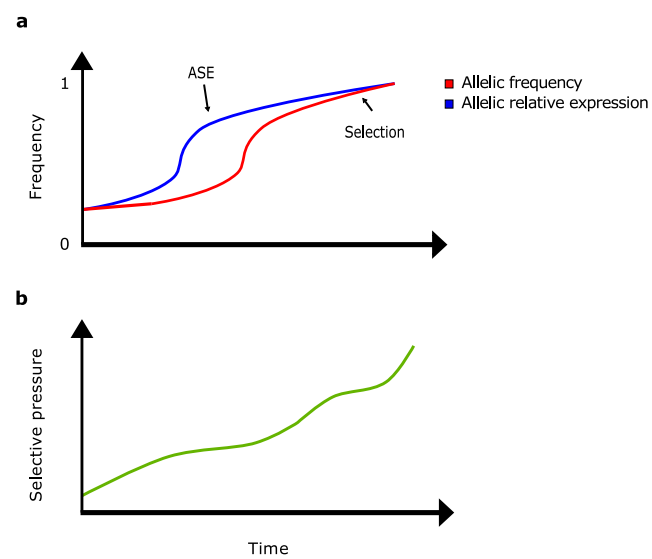
Figure 4: From allele-specific expression to natural selection. **a,** Evolution of allele frequency and allele relative expression over time. **b,** Evolution of selective pressure over time