Molecular and anatomical study of the epipelagic copepod Oithona nana (Copepoda; Cyclopoida)

Kevin Sugier

université PARIS-SACLAY

Thèse de doctorat

NMT : 2019SACLE045

Molecular and anatomical study of the epipelagic copepod *Oithona nana* (*Crustacea; Cyclopoida*).

Thèse de doctorat de l'Université Paris-Saclay
préparée à l'Université Evry Val d'Essonne et au CEA-Genoscope.

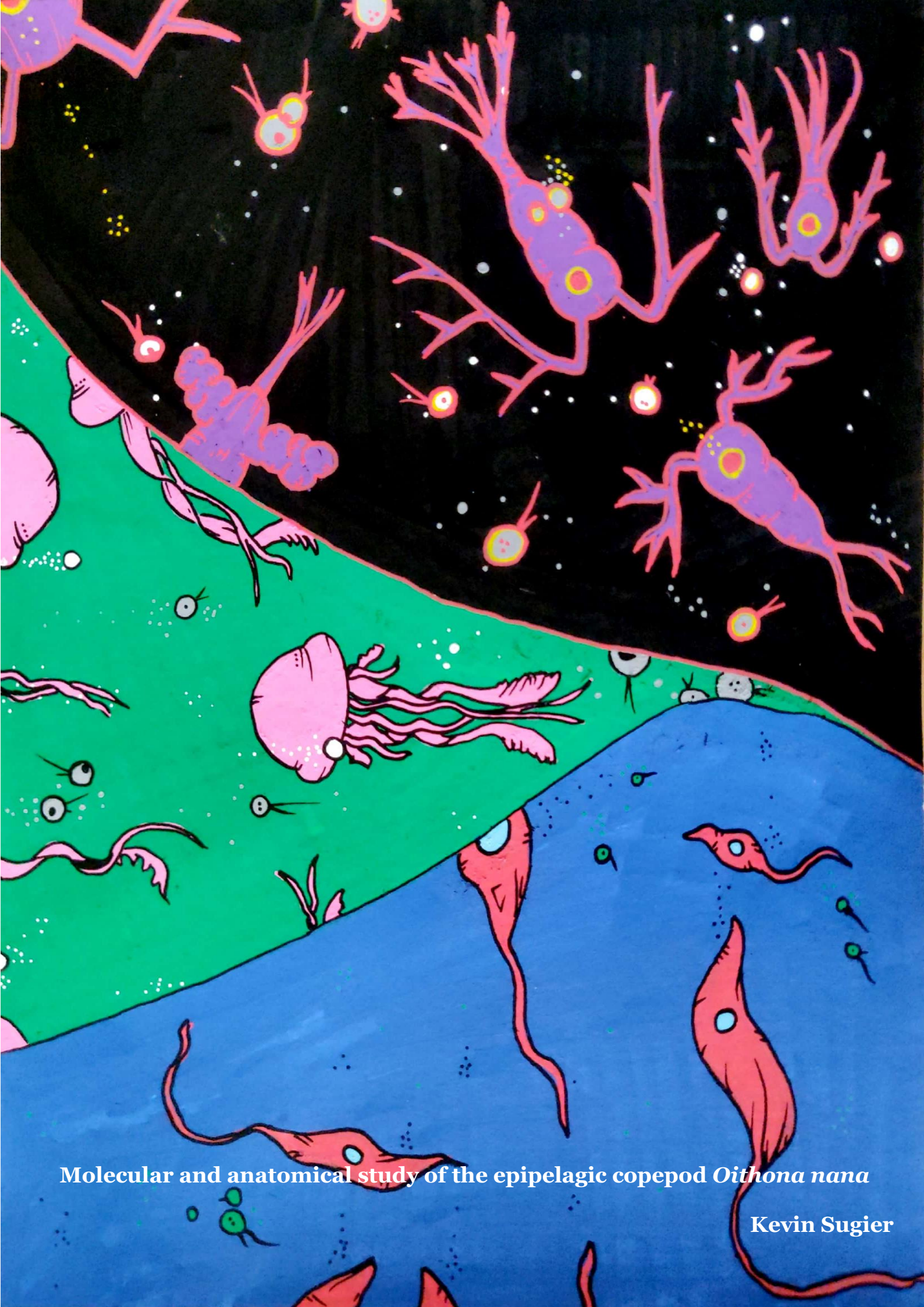
Ecole doctorale n° 577 Structure et dynamiques des systèmes vivants (SDSV)
Spécialité de doctorat : Sciences de la Vie et de la Santé

université PARIS-SACLAY | ÉCOLE DOCTORALE Structure et dynamique des systèmes vivants (SDSV)

Thèse présentée et soutenue à Evry, le 12/12/2019, par
Kevin Sugier

Composition du Jury :

Céline Boulangé-Lecomte Professeur, Université Le Havre Normandie (UMR-I 02)	Rapporteur
Christine Coustau DR CNRS, Institut Sophia Agrobiotech	Rapporteur
Richard Cordaux DR CNRS, Université de Poitiers (UMR 7267)	Examinateur
Elodie Fleury PhD, Ifremer Brest (UMR 6539)	Examinateur
Abdelghani Sghir Professeur, Paris-Saclay (UMR 8030)	Président
Mohammed-Amin Madoui Chercheur CEA, CEA/Genoscope (UMR 8030)	Directeur de thèse
Jean-Louis Jamet Professeur, Université de Toulon (UMR 110)	Directeur de thèse
Patrick Wincker DR CEA, CEA/Genoscope (UMR 8030)	Directeur de thèse



Molecular and anatomical study of the epipelagic copepod *Oithona nana*

Kevin Sugier

« Nothing in Biology makes sense except in the light of Evolution »

Theodosius Dobzhansky, 1964

Acknowledgements

I sincerely thank all the people who have supported me for the past three years: my family, my friends, my colleagues and my collaborators.
You have all my gratitude.

Table of Contents

List of Figures

Introduction

Chapter 1: *Oithona* as you never saw it

Article 1

Chapter 2: The *O. nana* genome shows an explosion of LNR domain targeted by natural selection

Article 2

Article 3

Chapter 3: LDPGs are involved in auto-proteolysis and neurogenesis modulation in *O. nana* males

Article 4

Chapter 4: Nervous system specific genes are targeted by population-scale ASE and natural selection in Arctic *O. similis* population

Article 5

List of Tables

Chapter 2: The *O. nana* genome shows an explosion of LNR domain targeted by natural selection

Article 2

Article 3

Chapter 4: Nervous system specific genes are targeted by population-scale ASE and natural selection in Arctic *O. similis* population

Article 5:

List of Appendix

Table of Abbreviation

AO:	Atlantic Ocean
ASE:	Allele-specific expression
BAF:	B-allele frequency
CI to CVI:	1st to 5th Nauplius stages (larvae)
$C_6H_{12}O_6$:	Glucose
Cal.:	Calanus
CO_2 :	Carbon dioxide
Cycl.:	Cyclopoid
DAPI:	Diaminido-2-phenylindole
DE:	Differential expression
DNA:	Deoxyribonucleic Acid
ECM:	Extracellular Matrix
GABA:	Gamma-aminobutyric Acid
Gb:	Giga base pairs (billion)
H_2O:	Water
H^+:	Hydrogen ion
Harp.:	Harpacticoid
HCO^{3-}:	Bicarbonate
IAGH:	Insulin-like Androgenic Gland Hormone
IGF:	Insulin-like Growth Factor
IGFBP:	Insulin-like Growth Factor Binding Protein
LDP:	LNR domain-containing protein
LDPG:	LNR domain-containing protein-coding gene
LNR:	Lin-12 Notch repeat
Ma:	Megaannum (million years)
Mb:	Mega base pairs (million)
MGV:	Marine genomic variant
MS:	Mediterranean Sea
Mya:	Million years ago
NI to NVI:	1st to 5th Copepodite stages (juvenile)
NAO:	North Atlantic Ocean
O_2:	Dioxygen
OTU:	Operational taxonomic unit
Pappa:	Pappalysin
PCA:	Principal component analysis
PM:	Peritrophic membrane
PPI:	Protein-protein interaction
RNA:	Ribonucleic acid
SNP:	Single-nucleotide polymorphism
TF:	Transcription factor
UTR:	Untranslated region
WCGNA:	Weighted correlation network analysis
WGA-FITC:	Wheat germ agglutinin-fluorescein isothiocyanate
Y2H:	Yeast two-hybrid

Copepod anatomy (p. 15)

R :	Rostrum
A1:	Antennule
A2:	Antennae
La:	Labrum
Md:	Mandibule
Mx1:	Maxillule
Mx:	Maxillae
Mxp:	Maxilliped
P1 to P5:	Legs pairs
Th1 to Th5:	Thoracic somite
ip:	Intercoxal plate
cx:	Coxa
Bsp:	Basipod
Exp1 to Exp3:	Exopodites
Enp1 to Enp3:	Endopodites
RS:	Receptaculum Seminis
Gs:	Genital somite
Ur1 to Ur5:	Urosomites
Fu:	Furcal rami
Me:	External (marginal) furcal seta
Sd:	Dorsal furcal seta
Te:	Terminal external furcal seta
Tme:	Terminal median external seta
Tmi:	Terminal median internal seta
Ti:	Terminal internal seta

Eukaryota (domain)

...Opisthokonta

.....**Metazoa** (kingdom)

.....Eumetazoa

.....Bilateria

.....Protostomia

.....Ecdysozoa

.....Panarthropoda

.....**Arthropoda** (phylum)

.....Mandibulata

.....Pancrustacea

.....**Crustacea** (subphylum)

.....**Multicrustacea** (superclass)

.....**Hexanauplia** (class)

.....**Copepoda** (subclass)

FIG. 1: **Taxonomic rank of Copepoda** (after WoRMS and NCBI).

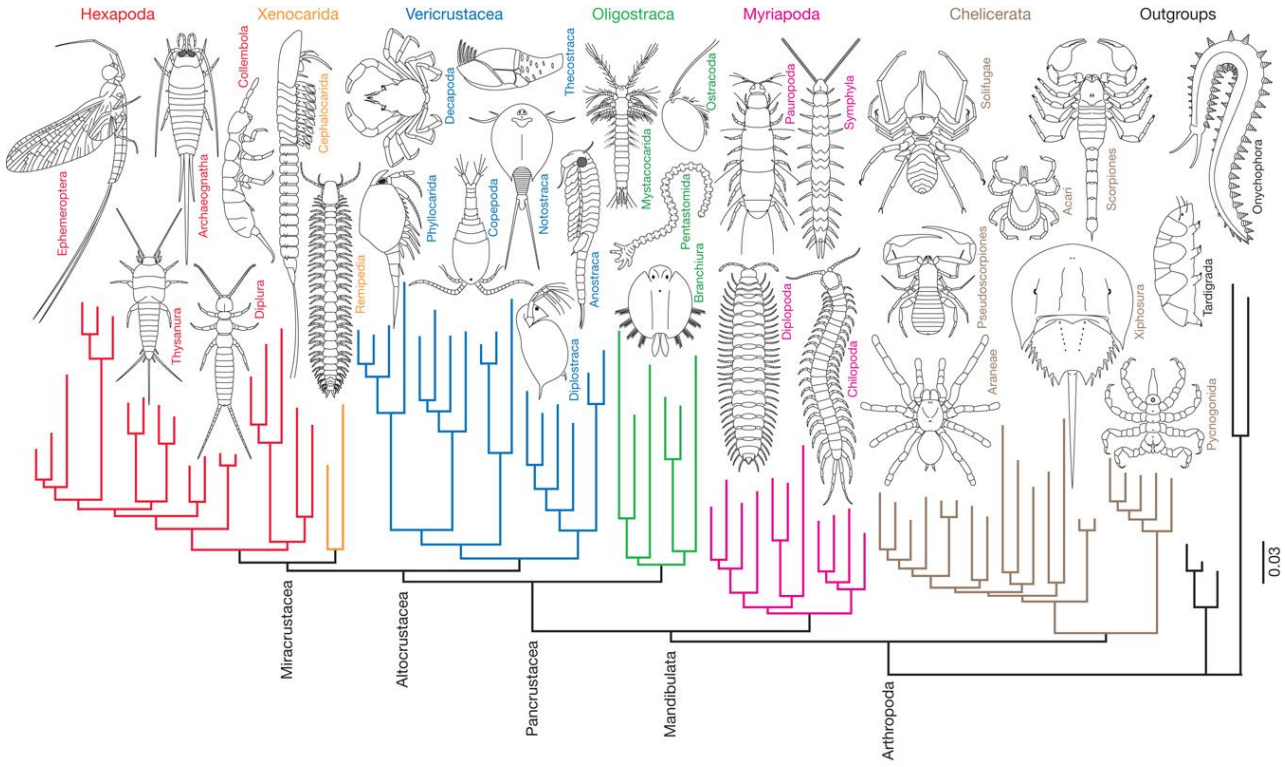


FIG. 2: Phylogenetic tree of Arthropods obtained by likelihood (from Regier et al., 2010)

Introduction

1.1. Copepods fit the world

Copepods are small aquatic animals. Their name comes from the ancient Greek "*kôpê-podos*", literally "handle of foot", meaning ream-shaped foot; and was introduced by the French zoologist Henri Milne Ewards in 1840 (WoRMS 2019a; Milne Edwards 1840). The name *Copepoda* constitutes a subclass of Crustacean and belongs to the Arthropoda phylum (figure 1).

The divergence between copepods and other arthropod taxa is estimated molecularly between 395 and 495 million years ago (Mya) (Eyun 2017; Sanders and Lee 2010). Copepods fossils are scarce and little diversified (few orders are represented by fossils) (Huys et al. 2016), the oldest and not ambiguous fossils date from the late Carboniferous age (around 300 Mya) (Selden et al. 2010), but some copepodologists declared they found a copepod-like mandible fossil from Cambrian (more than 500 Mya) (Harvey and Pedder 2013). *Copepoda* is sister-phylum of *Thecostraca* (*e.g.* barnacle), and *Eumalacostraca* (*e.g.* crabs, shrimp), and together forms the *Multicrustacea* superclass (Eyun 2017; Regier et al. 2010) (figure 2).

To date, eight orders are recognised using taxonomic and molecular data, but there is no consensus among copepodologists, and three other orders can be mentioned in the literature. Among the eight orders, four gather 99% of known families: *Calanoida*, *Harpacticoida*, *Siphonostomatoida* and *Cyclopoida* (which includes the former *Poecilostomatoida* order). The *Copepoda* subclass is divided into two superorders by the German zoologist Wilhelm Giesbrecht: *Gymnoplea* (*Calanoida* order) and *Podoplea* that includes the seven other orders (see Morphology/Anatomy section). Copepods contain more than 14,000 species (WoRMS, 2019), dozens of new species are described each year, and more than 20,000 are expected after discoveries from the future marine sediment explorations. This high species diversity leads to an explosion of forms and sizes occupying various ecological niches.

The copepods living in or near the sediments belong to the benthos, from the ancient Greek "*bathos"* meaning depth. Benthic copepods are at the origin of their present diversity: they probably started to colonise the water column between 400 Mya and 450 Mya, and the pelagic waters between 250 Mya and 300 Mya. (Bradford-Grieve 2004; Selden et al. 2010).

The pelagic copepods belong to the broader concept of zooplankton, coming from the ancient Greek "*zoion - plagktos"* meaning animal and wandering, respectively. The German physiologist Victor Hensen was the first to give this name to describe the small marine animals that could not swim against the current. However, planktonic copepods can migrate vertically. One-third of the known copepod species are parasites with a broad host spectrum including sponges, mammals, reptiles and fishes (Selden et al. 2010).

Copepods are described as the most abundant animals on Earth, followed by Insects and Nematodes (Humes 1994), and are estimated to probably outnumber all other combined metazoans. Dr Geoff Boxshall and Dr Rony Huys estimated that, with the presumption of a unique copepod per litre of ocean water, more than $1.347x10^{21}$ copepods are present on Earth. This abundance is also illustrated by the presence of copepods in all aquatic environments, salinity ranges, temperature regimes, a wide range of pH and a broad latitude spectrum: from freshwater to hypersaline seawaters, from sub-zero glacier to hot water-spring, from deep-sea trenches (down to -10 km) to mountain (up to +5km) (Kiørboe 2011; Selden et al. 2010; Huys and Boxshall 1991). Some copepods were also observed in other atypical environments like abandoned tyres, pineapples, bromeliads, puddles, trees mosses and holes and water tanks (Huys and Boxshall 1991). Anecdotally, the presence of this crustacean in drinkable waters was a big issue in New York Jewish (crustaceans are not kosher) and vegetarian communities (BBC News 2004). Copepods have adapted their reproduction, feeding and laying eggs strategies to all these environments which explain their cosmopolitan distribution.

The majority of the copepods has a size between 500µm and 5mm. To date, the smallest known copepods measure a few micrometres and are fish gill parasites of the smallest vertebrate, *Paedocypris progenetica;* and the largest known copepods (*Pennella sp.*) are up to 50cm with egg sacs, and are baleen whales (*Balaenoptera musculus*) parasites (Selden et al. 2010). Copepods are also described, related to their size, as the world's strongest and fastest animals. Indeed, copepods are ten to thirty times strongest than any other measured metazoan, and can move up to half a meter (around 500 times their body length) in less than one second (Kiørboe *et al.*, 2010; Technical University of Denmark, 2010).

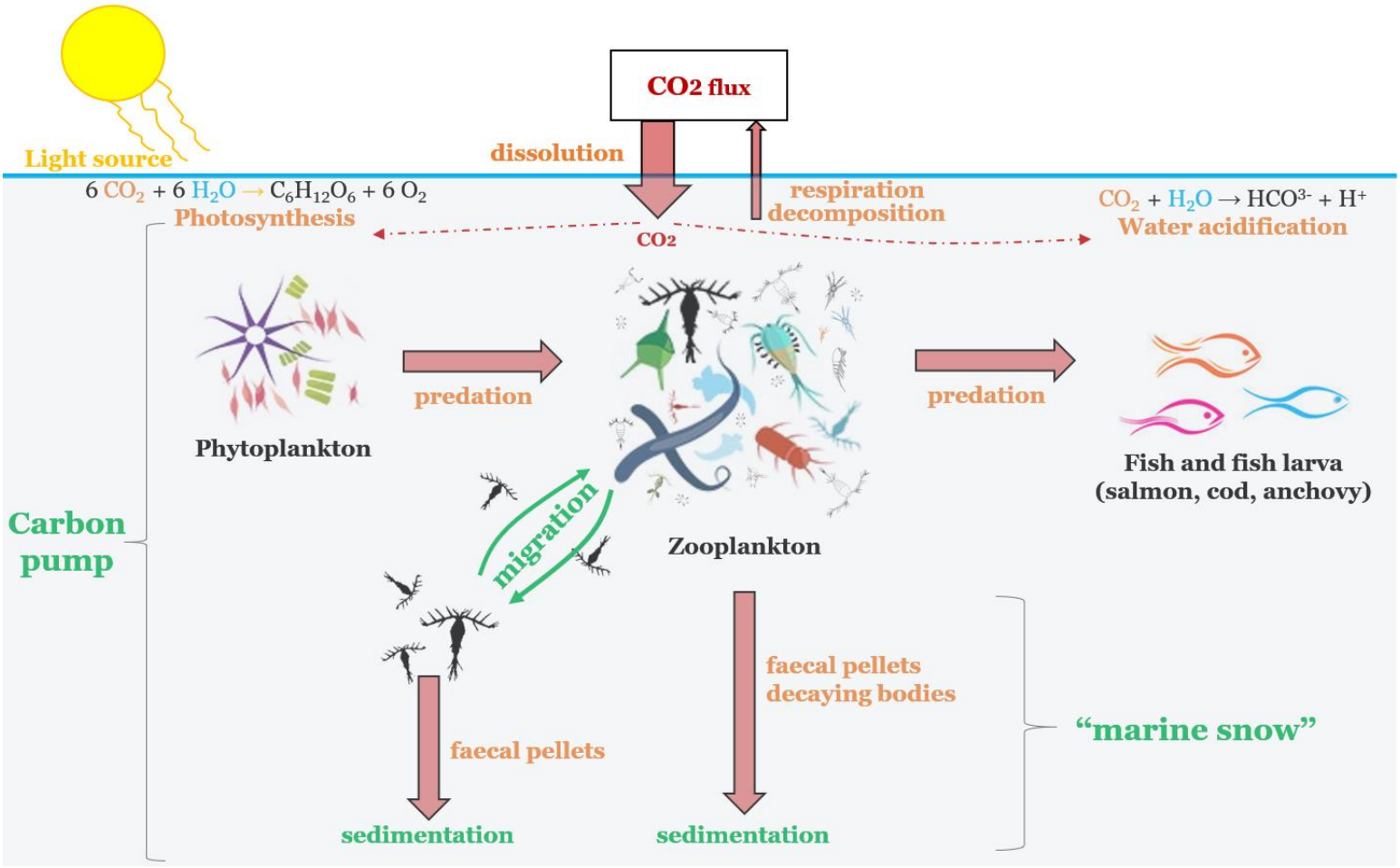


FIG. 3: Copepods are the foundation of aquatic metazoan life, and essential to marine biogeochemical cycles.

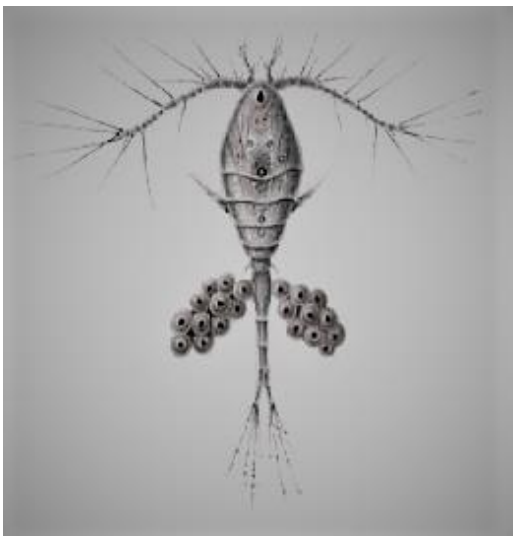


FIG. 4: First drawing of *Oithona nana* (from Giesbrecht, 1892).

Some copepods can also be human disease vectors. Some freshwater cyclops carry the Guinea worm larva (*Dracuncula medinensis*), the causative agent of the debilitating disease dracunculiasis (Huys and Boxshall 1991). The abundant copepods *Eurytemora affinis* and *Acartia tonsa* are also known to carry the *Vibrio cholerae* bacteria, responsible for cholera (Rawlings, Ruiz, and Colwell 2007).

1.2. Copepods feed the world

Copepods represent the majority of the zooplankton: in pelagic zones, up to 88% in terms of biomass, and up to 99% in terms of numerical density (Thompson, Dinofrio, and Alder 2013). They consume microorganisms, phytoplankton and small zooplankton, and are eaten by the fish larva of species mostly consumed by humans (cod, salmon and anchovies). The disappearance or sudden decline in copepod abundance may have a severe economic impact for Northern countries and a health impact for countries where fish is the primary source of protein. In the North Sea, since the mid- ’80s, an increase in water temperature was observed (Beaugrand et al. 2003). This temperature increase led to a migration to the colder Northern waters of the copepod *Calanus finmarchicus,* the most consumed by cod larvae, which is gradually replaced by *Calanus helgolandicus*. This latter having a different reproductive phenology, the species replacement resulted in a fall in cod larval recruitment, and thus, a fall in coastal fish biomass (Beaugrand et al. 2003).

Copepods are the most significant marine carbon sink and active transports of nitrogen and phosphorous into the deep oceans through two phenomena (figure 3). (i) By respiration and excretion, especially for species that migrate to 200 meters depth and below (Kobari et al. 2008). In the cold marine environment, several calanoids species at the last juvenile or adult stages accumulate lipid resources in oil sacs, migrate in deep waters and enter in diapause, a hibernation-like period of several months where the individuals stop their development (growth and sexual differentiation) and reduce their physiological activities to the minimum (Baumgartner and Tarrant 2017). (ii) By the “marine snow”, which comprises decaying bodies and faecal pellets sedimentation (Boyd et al. 2019). In North Atlantic, along the year, copepods constitute 10% to 25% of the biological carbon pump (Jónasdóttir et al. 2015).

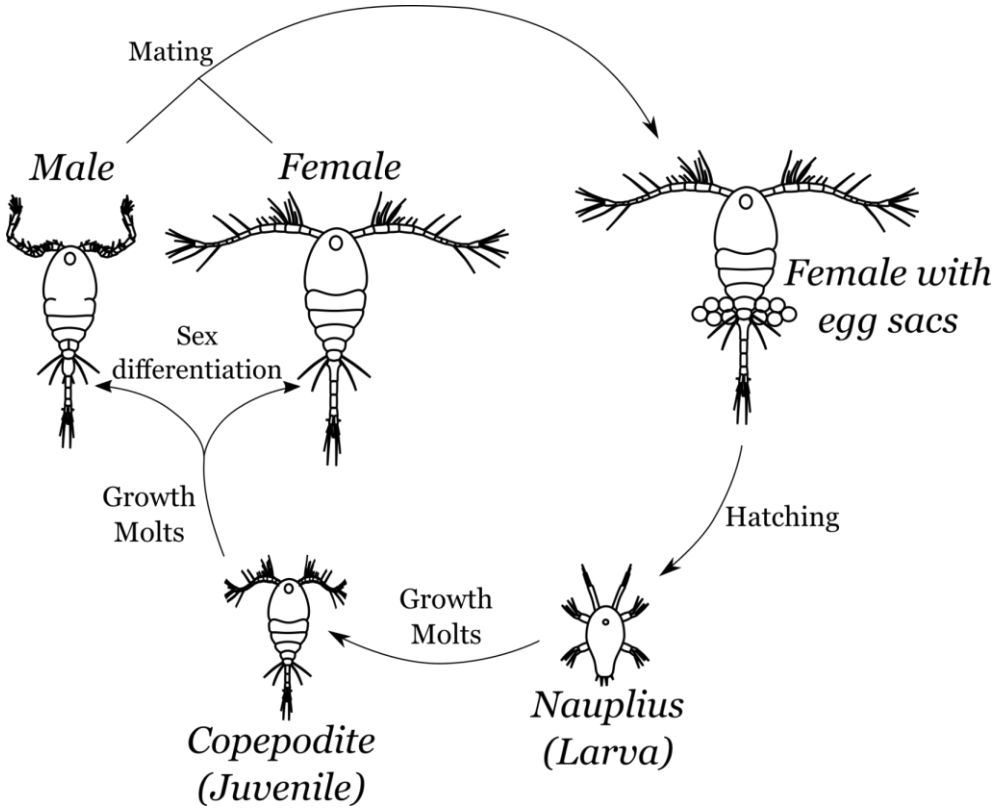


FIG. 5: The simplified life cycle of free-living cyclopoid (Sugier et al., 2019).

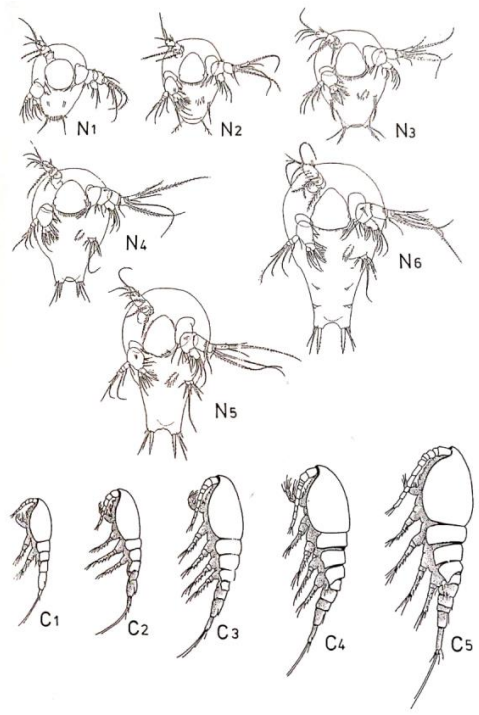


FIG. 6: Developmental stage of free-living *Cyclopoida*: nauplii (larva) N1 to N6 and copepodids (juvenile) C1 to C5 (from Dussart and Defaye, 2001).

With actual climate changes, a dormancy period reduction was observed which have a significant impact on biogeochemical cycles, particularly the carbon one (Baumgartner and Tarrant 2017).

Through these same processes, copepods are also essential for the life of marine bacteria: free-living bacteria can feed on copepod carcasses and faecal pellets, while other communities live on the surface or in the gut of the copepods (Tang 2005; Shoemaker, Locey, and Lennon 2017). Community differences are observed between bacteria attached to copepods and free-living, but exchanges exist (De Corte et al. 2014). Thus, copepods provide organic substances and/or are the hosts of bacterioplankton.

Copepods are also used by humans for fish farming but also as a source of fatty acid oil for cosmetics and food (Pedersen, Vang, and Olsen 2014); some Asian and Scandinavian populations directly consume copepods they harvest (*e.g.* in Laos in Kottelat 2007 study). Like insects, copepods may become a source of protein and fatty acid to respond to the decline of fish stocks and the ecological problems induced by conventional farming.

Copepods are also used as pest and diseases controllers. *Mesocyclops aspericornis* is described as the most effective species in Polynesia, Australia and parts of Asia to kill *Aedes aegypt*, the vector of the Dengue haemorrhagic fever. Each individual may kill up to 40 *Aedes* larva by day (Brown, Kay, and Hendrikz 1991). *Mesocyclops* was introduced in some water tanks of villages in Vietnam touched by the dengue. One year after the introduction of the copepod, *A. aegypt* disappeared (Marten 2001).

1.3. *Oithona*: Virgin of the wave

The *Oithona* genus was described for the first time by the Scottish zoologist William Baird in 1843 in the journal 'The Zoologist' (Baird 1843; WoRMS 2019b). He was investigating "insects" responsible for the luminescence of the sea and making the first *Oithona* descriptions by observing *O. plumifera* and *O. splendens* (Zamora Terol 2013; Baird 1843). The genus name comes from the James Macpherson book "The Poems of Ossian", and is composed of "*oi-thóna*", meaning respectively virgin and wave in Gaelic (Zamora Terol 2013). To date, 48 species of *Oithona* are listed

(WoRMS 2019b), including *O. nana* Giesbrecht 1892 (figure 4), *O. frigida,* Giesbrecht 1902, *O. atlantica* Farran 1908 and *O. davisae* Ferrari and Orsi 1984.

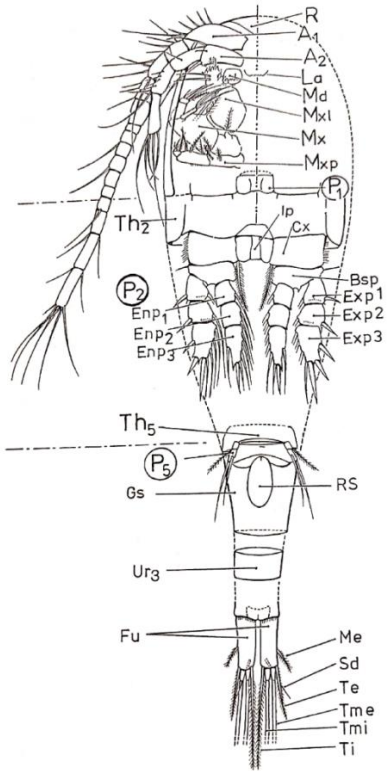


FIG. 7: Drawing of the ventral view a cyclopoid copepod (from Dussart and Defaye, 2001)**.** Legend abbreviation on page 8.

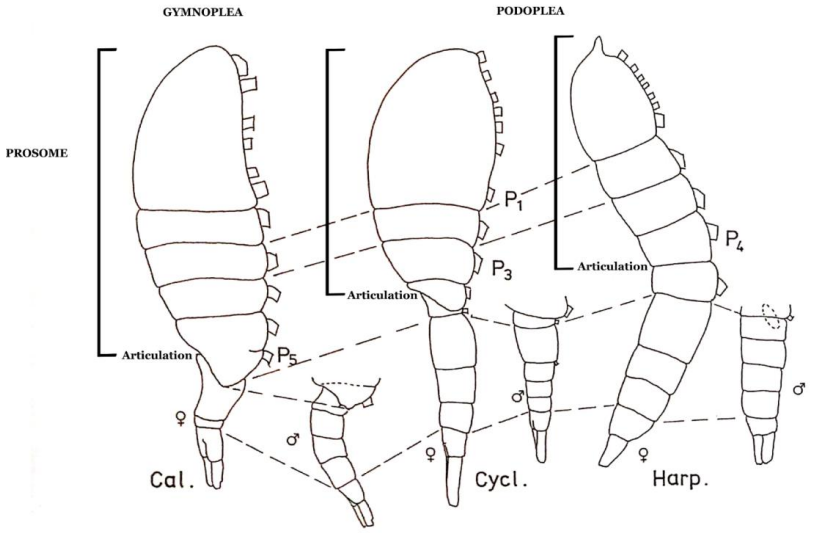


FIG. 8: Difference of morphological structure between the two copepod super-orders, *Gymnoplea* and *Podoplea* (after Dussart and Defaye, 2001)**.**

Until the end of the XXth century, the conventional net used for sampling zooplankton had a 200µm mesh (Gallienne and Robins 2001). Because of the use of large-mesh nets, the small zooplankton was underestimated. Without this sampling bias, *Oithona* genus is described as the most abundant marine copepod (Gallienne and Robins 2001), and present in all marine environments (Nishida 1985). Because of this high abundance and cosmopolitan biogeography, *Oithona* is one of the key genera for marine ecology.

The life cycle of copepods is divided into two phases (figure 5). The first larval phase, called nauplius, is composed of six stages (from NI to NVI) (figure 6). During this phase, the larvae will acquire new appendages by an alternation of molten and growth steps. This allows reaching a second phase (juvenile) called copepodite. This phase is also composed of six stages (CI to CVI) during which the juveniles, by molten and growth steps alternation, will increase their size, develop their segments and undergo a sexual maturation (figure 6). The definitive adult form is called copepodite VI.

In adults, despite the multitude of species occupying various ecological niches, a vast majority of ovoid forms are observed; except for host-specialised parasitic species. The number or shape of appendages is characteristic of the species and the sex of the individuals. To identify the different *Oithona* species, the morphological descriptions of Nishida (1985) and Rose (1933) are currently used. At the anatomical level, the descriptions of Huys and Boxshall (1991), Dussart and Defaye (2001) and Mironova and Pasternak (2017) are currently used.

The free-living adult body is divided into two parts: the prosome (the head and the thorax) and the urosome (the abdomen) (figure 7). These elements are composed of segments of different sizes, not overlapping, and more or less fused according to the species. On the segments of the prosome are observed various appendages (antennas, antennules, mandibles, maxilla, maxillipeds, swimming legs) that have sensory, movement, grip or food capture role (figure 7). Depending on the ecological niche and the feeding behaviour, the appendixes can have differences in the number of segments, spine and setae (Huys and Boxshall 1991). The genital orifices are located on the first segment of the abdomen and the anal orifice on the last segments. From morphological observations, Wilhelm Giesbrecht established two super-orders: *Gymnoplea* (“*gymno-*“ meaning naked) because the urosome is lacking appendixes, contrary to

FIG. 9: *Cyclopoida* copepods mating. (from 호프콤파니, 2014)**.** The male (smallest individual) uses antennas to grip the 4th female pair of legs and transferred it spermatophores with its modified P5.

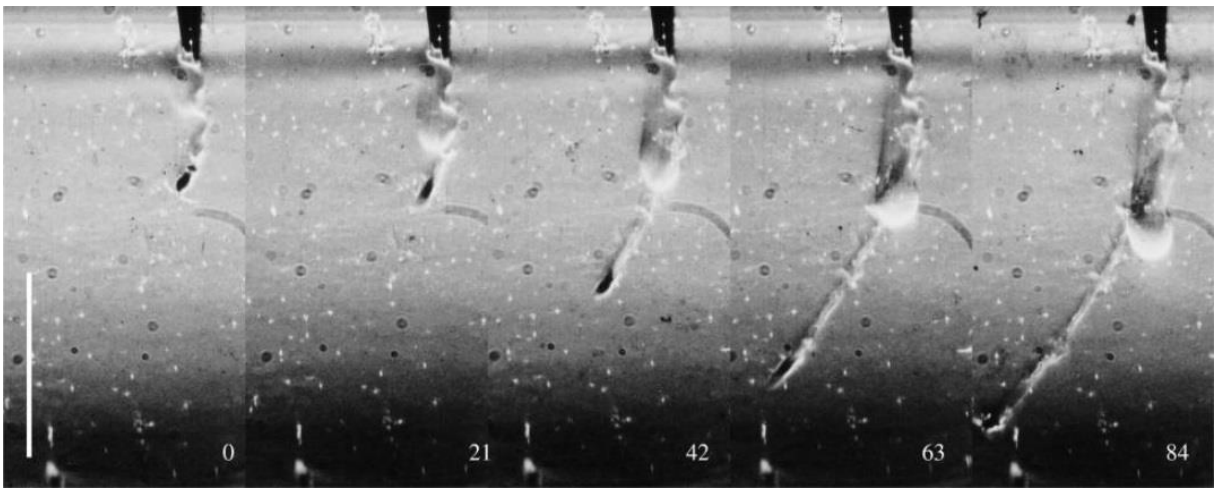


FIG. 10: *Cyclopoida* copepod escaping from a disturbance (from Strickler and Balázsi, 2007). Numbers represent time in milliseconds. Scale bar = 1cm.

Podoplea ("*pod-*" meaning foot) where the fifth leg pair (P5) is located on the first urosome segment (figure 8).

In the majority of free-living copepods, some of these appendixes have gender-dependent characteristics. The male antennas gain geniculations that allow it to grip the female P4 during mating (figure 9). In males, the P5 is also modified to permit the transfer of the spermatophore to the genital orifice of the captured female. To date, none cyclomorphosis was observed in copepods. All copepods possess a sexual reproduction between two individuals of the opposite sex (male and female), except in some harpacticoid species with a parthenogenesis capacity. The quasi-motionless *Oithona* female is detected by males thanks to the trail created after motion (figure 10). At the copepod scale, water has a high viscosity, and so a wake is visible following the copepod movement (Strickler and Balázsi 2007). Moreover, *Oithona* females can be detected by their pheromone that has not yet been chemically identified. These pheromones are detected by a specialised organ: the aesthetasc (figure 11). Generally, males have two aesthetascs by urosome segments (Huys and Boxshall 1991). When the female is detected, the male tries to capture it. The pursuit can end by the escape of the female, or its capture. During the mating, the male clings, using its articulated antennae, and transfers its two spermatophores into the seminal orifices by using its P5 (figure 12). The sperm is stored in the spermathecae that allow fertilisation of the female's eggs. With a single mating, the female can fertilise all the eggs produced in its lifespan; for this reason, the male has a preference for 'virgin' females (Heuschele and Kiørboe 2012). The eggs are carried in bags by the female until they hatch (figures 5 & 6).

Oithona is described as an ambush feeder, more precisely as a small active ambush feeder (Benedetti, Gasparini, and Ayata 2015), meaning that it uses a sit-and-wait strategy. The long antennae pairs (length depending on the species) have a sensory function due to their large number of setae (Strickler and Bal 1973). This permits to detect prey movements (hydromechanical perception), to jump on it and to capture it using its mouth appendixes in a few milliseconds (figure 13). This successful strategy limits the predation risk by remaining motionless, especially for the *Oithona* females with eggs. Moreover, the antennae can help during a powerful escape jump (Borazjani et al. 2010), and it is one of the main argument to the ecological success of copepods (Kiørboe et al. 2010).

FIG. 11: Copepod chemosensor organ (aesthetasc) by electron microscope (Huys and Boxshall 1991)

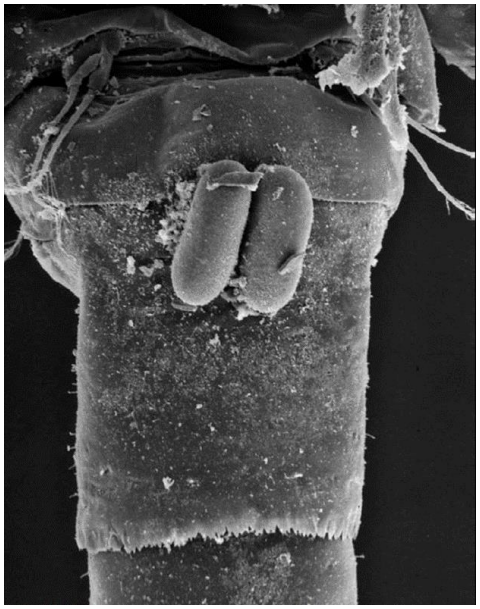


FIG. 12: **Ventral view of an *Oithona* female genital somite after one mating by electron microscope** (Huys and Boxshall 1991).

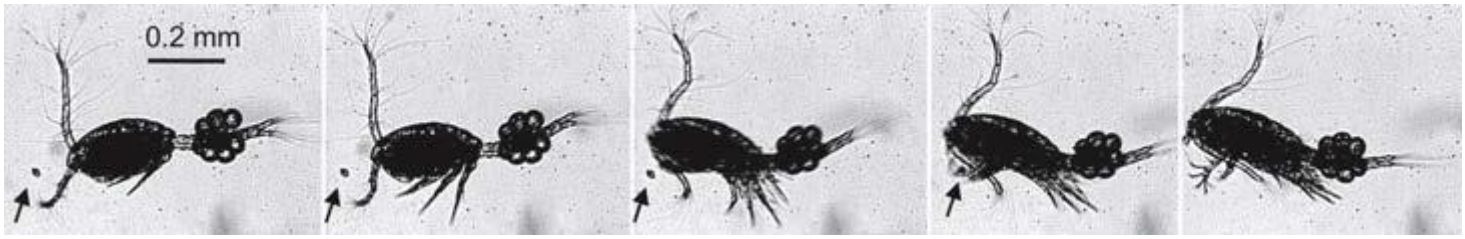


FIG. 13: *Oithona* ambush-feeder strategy (from Kiørboe et al., 2009). Each frame is separated by 0.05ms.

Because of its cosmopolitan distribution, ecological success, hyper abundance, essential role in marine trophic web and biogeochemical cycles and the absence of molecular information, we chose the *Oithona* genus as a potential model for small zooplankton. *Oithona nana* seems to be present in most of the coastal waters (Razouls et al. 2019; Nishida 1985), notably in the polluted ones and possesses a small genome. For all these reasons, I used *Oithona* as the model organism of my thesis.

1.4. Thesis aim

This thesis aims to fill the gaps between the behaviour ecology and biology of the copepod *Oithona nana* using anatomical analysis and molecular data. In the first chapter, I analysed the structure of the digestive and reproductive system. This study revealed their chitin constitution and confirmed the copepod importance in the marine "biological carbon pump". In the second chapter, I highlighted the explosion of a gene family, called LNR domain-containing protein-coding genes (LDPGs), and highlighted their importance in an evolutionary point of view using *Tara* Oceans metagenomic data. The third chapter confirms, at a molecularly level, the difference between the male behaviour compared to the other developmental stages. The transcriptomes of the five *O. nana* developmental stages were built, and the genes specific to each phase were identified. Using expression and protein-protein interaction analysis, I have tried to characterise the functional role of LDPGs. They were likely to play a role in modulating the auto-proteolysis and neurogenesis in male and thus they may participate to the sacrificial behaviour of the male during search mating-partner, which seems to be the first case in iteroparous animals. The last chapter proves the link between population-scale allele-specific expression (ASE) and natural selection in *O. similis* population from the Arctic Seas. In these populations, genes involved in the nervous system also appeared to be under population-scale evolution processes.

Chapter 1:

Oithona as you never saw it

Initially, I wanted to test by silencing the function of some gene candidates. I hypothesised a role of these genes in the development of the male and its reproductive system. I needed the wild type *O. nana* anatomy as a control to compare and identify the gene-candidate role. This microscopy analysis was done because of the lack of precise images of the *Oithona* morphology/anatomy in the literature.

The developed protocol made it possible by fluorescence labelling of the chitinous structures. The fact that the digestive and reproductive systems were entirely covered by chitin was unexpected. Also, an undescribed chitinous structure in the last segment of the legs was detected in some individuals.

In this study, my role was to apply the protocol developed by Vacherie B. and Madoui M-A on *Oithona* individuals, to make microscopic observations and to take the pictures. I also had to analyse the pictures, write the article and make the figures, with the help of my co-authors.

Article source

Sugier, Kevin, Benoit Vacherie, Astrid Cornils, Patrick Wincker, Jean-Louis Jamet, and Mohammed-Amin Madoui. 2018. 'Chitin Distribution in the *Oithona* Digestive and Reproductive Systems Revealed by Fluorescence Microscopy'. PeerJ 6 (May): e4685. https://doi.org/10.7717/peerj.4685.

A French abstract is available in appendix 1 (page 152)

Chitin distribution in the *Oithona* digestive and reproductive systems revealed by fluorescence microscopy

Kevin Sugier[1], Benoit Vacherie[2], Astrid Cornils[3], Patrick Wincker[1], Jean-Louis Jamet[4] and Mohammed-Amin Madoui[1]

[1] Génomique Métabolique, Genoscope, Institut François Jacob, CEA, CNRS, Univ Evry, Université Paris-Saclay, Evry, France
[2] Commissariat à l'Energie Atomique (CEA), Institut François Jacob, Genoscope, Evry, France
[3] Alfred-Wegener-Institut Helmholtz-Zentrum für Polar- und Meeresforschung, Polar Biological Oceanography, Bremerhaven, Germany
[4] Université de Toulon, Aix Marseille Université, CNRS/INSU, IRD, MIO UM 110, Mediterranean Institute of Oceanography, La Garde, France

ABSTRACT

Among copepods, which are the most abundant animals on Earth, the genus *Oithona* is described as one of the most numerous and plays a major role in the marine food chain and biogeochemical cycles, particularly through the excretion of chitin-coated fecal pellets. Despite the morphology of several *Oithona* species is well known, knowledge of its internal anatomy and chitin distribution is still limited. To answer this problem, *Oithona nana* and *O. similis* individuals were stained by Wheat Germ Agglutinin-Fluorescein IsoThioCyanate (WGA-FITC) and DiAmidino-2-PhenylIndole (DAPI) for fluorescence microscopy observations. The image analyses allowed a new description of the organization and chitin content of the digestive and reproductive systems of *Oithona* male and female. Chitin microfibrils were found all along the digestive system from the stomach to the hindgut with a higher concentration at the peritrophic membrane of the anterior midgut. Several midgut shrinkages were observed and proposed to be involved in faecal pellet shaping and motion. Amorphous chitin structures were also found to be a major component of the ducts and seminal vesicles and receptacles. The rapid staining protocol we proposed allowed a new insight into the *Oithona* internal anatomy and highlighted the role of chitin in the digestion and reproduction. This method could be applied to a wide range of copepods in order to perform comparative anatomy analyses.

Subjects Marine Biology, Aquatic and Marine Chemistry, Biogeochemistry
Keywords Chitin, Microscopy, Biology marine, Anatomy, Copepod, *Oithona*

INTRODUCTION

Copepods are the most abundant animals on Earth ahead of insects and nematodes (*Humes, 1994*) and inhabit all aquatic niches: groundwater, vernal ponds, glaciers, lakes, rivers and oceans (*Huys & Boxshall, 1991*). Among marine copepods, *Oithona* has been described as the most important marine planktonic genus in terms of abundance (*Gallienne & Robins, 2001*). A recent study, based on the *Tara* Oceans metagenomic data,

Submitted 4 December 2017
Accepted 10 April 2018
Published 14 May 2018

Corresponding author
Kevin Sugier,
ksugier@genoscope.cns.fr

Academic editor
Robert Toonen

Additional Information and Declarations can be found on page 11

DOI **10.7717/peerj.4685**

has shown the global distribution of *Oithona* in coastal and open ocean waters (*Madoui et al., 2017*), which highlighted its key role as a major secondary producer of the marine food chain (*Beaugrand et al., 2003*; *Zamora-Terol et al., 2014*). The important contribution of copepods in the biological carbon pump has also been demonstrated (*Jonasdottir et al., 2015*), in particular through the excretion of faecal pellets (*Steinberg & Landry, 2017*) that sink, provide organic and inorganic compounds to microplankton (*Steinberg, Goldthwait & Hansell, 2002*; *Valdés et al., 2017*), and deposit on the sediments where they could remain as fossils for several thousand years (*Bathmann et al., 1987*; *Haberyan, 1985*). The biochemical analysis of the copepod faecal pellets has revealed a high amount of chitin (*Kirchner, 1995*), a β-1-4-*N*-acetylglucosamine polymer, the most abundant biopolymer in nature after celluloses (*Kirchner, 1995*), and mostly known in copepods as a component of the exoskeleton. Besides the role of copepods in the carbon pump, the abundance of chitin in the faecal pellets also points out the implication of copepods in the global nitrogen cycle (*Frangoulis, Christou & Hecq, 2004*).

Morphological traits of more than forty *Oithona* species are well known (*Razouls et al., 2005–2018*), especially the structure of the antennules, the oral appendages, the swimming legs and the caudal rami (*Nishida, 1985*). However, such morphological traits are only accessible through finical dissections under the microscope that need expertise and are time-consuming. Recently, molecular tools have proven their usefulness in species identification (*Cornils, Wend-Heckmann & Held, 2017*; *Madoui et al., 2017*).

The detailed external anatomy of copepods has been analysed through Congo red fluorescence (*Michels & Büntzow, 2010*) and through electron microscopy that allowed the species identification of copepods and the characterization of their external structures (*Chang, 2013*; *Cuoc et al., 1997*; *Marques et al., 2017*). Using an electron microscope, *Oithona nana* Giesbrecht, 1892 female sexual orifices with attached male spermatophores were observable (*Huys & Boxshall, 1991*). Diagrams of marine and freshwater cyclopoids, which provide the structures of the reproductive and digestive systems (*Borradaile & Potts, 1935*; *Dussart & Defaye, 2001*; *Kellogg, 1902*) were available. Some studies proposed methods to observe the reproductive system of aldehyde-preserved copepods by direct light microscopy observation of individuals (*Eisfeld & Niehoff, 2007*; *Niehoff, 2003*; *Niehoff & Hirche, 1996*; *Tande & Hopkins, 1981*), by staining of gonad with borax carmine (*Tande & Gronvik, 1983*; *Tande & Hopkins, 1981*), with fluorescent polyunsaturated aldehydes (PUAs) probes (*Wolfram, Nejstgaard & Pohnert, 2014*), or with Fast Green (*Batchelder, 1986*). The internal anatomy of *O. similis* Claus, 1863 has been recently described using phase contrast microscopy and provided the first insight into the organization of the *Oithona* female reproductive system (*Mironova & Pasternak, 2017*). In the *Wolfram, Nejstgaard & Pohnert (2014)* study, some pictures of the calanoid *Acartia tonsa* obtained using fluorescent PUA probes, also allowed to determine the anatomy of the digestive system. Other studies (*Bautista & Harris, 1992*; *Debes, Eliasen & Gaard, 2008*) used the chlorophyll fluorescence to determine the ingestion rates and the gut contents, but without providing a clear structure of the digestive organs. Electron microscopy revealed that chitin microfibrils are present in the anterior and posterior midgut peritrophic membrane (PM) of free-living and in the

posterior PM of parasitic copepods (*Yoshikoshi & Kô, 1988*), but no *Oithona* species have been included in the study.

For a better understanding of the ecological success of *Oithona*, a detailed knowledge of its internal anatomy is crucial. Fluorescence microscopy based on a double staining coupling Wheat Germ Agglutinin-Fluorescein IsoThioCyanate (WGA-FITC) and Diamidino-2-phenylIndole (DAPI) were used to elucidate the internal anatomy. DAPI is a blue fluorescent protein which has an affinity to two nucleoids: adenine and thyrosin (*Lin, Comings & Alfi, 1977*). This staining is widely used to detect DNA in eukaryotes, prokaryotes and some viruses, without tissue-specificity. FITC is a green fluorescent protein that can be conjugated with a wheat lectin that has an affinity and specificity to *N*-acetyl-β-D-glucosamine (*Allen, Neuberger & Sharon, 1973*). WGA-FITC staining is widely used for chitin detection by fluorescence, in a liquid medium containing lysed cells or directly on whole organisms (*El Gueddari et al., 2002*; *Farnesi et al., 2015*; *Fones, Mardon & Gurr, 2016*; *Godoy, Fernandes & Martins, 2015*). On copepods, WGA-FITC was used only once; but after dissolution of the soft tissues which did not allow the investigation of the internal anatomy (*Mravec et al., 2014*). In the present study, we used WGA-FITC and DAPI staining to provide a new insight into the internal anatomy and chitin content of *O. nana* and *O. similis* with a focus on their digestive and reproductive systems.

MATERIAL AND METHODS

Biological materials samples

Oithona nana and *O. similis* specimens were sampled at two locations of the Toulon harbour, France, at the East of the little harbour of Toulon (Lat 43°06′52.1″N and Long 05°55′42.7″E) and the North of the great harbour of Toulon (Lat 43°06′02.3″N and Long 05°56′53.4″E). Sampling took place in November 2016, January, March and June 2017. The samples were collected from the upper water layers (0–10 m) using zooplankton nets with a mesh of 90 and 200 µm. Samples were preserved in 70% ethanol and stored at –4 °C. In the samples, individuals of the four different development stages were observable (nauplii, copepodites and adults of both sexes), but the large majority were female adults.

Individual staining

This protocol was adapted from *Farnesi et al. (2015)*. After gently mixing the ethanol preserved samples (about 20 reversals), 100 µL were sampled in a 1.5 mL tube. After 2 min, the ethanol was removed, and 100 µL of phosphate buffered saline (PBS) at 1× and 10 µL of WGA-FITC at 2 mg mL^{-1} (L4895 SIGMA, Lectin from Triticum vulgaris (wheat) FITC conjugate, lyophilized powder; Sigma-Aldrich, St. Louis, MO, USA) were added for chitin staining. After mixing, the sample was incubated for 30 min protected from light before supernatant removing. To stain the DNA, dual staining with DAPI can be performed by adding, 100 µL of PBS at 1× and 10 µL of DAPI (D9542 SIGMA, DAPI for nucleic acid staining; Sigma-Aldrich, St. Louis, MO, USA) at 10×. The microscopy observations were done directly after mixing. This protocol can also be used on living individuals from a seawater sample; in this case, sodium chloride at 39 g L^{-1} has to be added to the PBS solution.

Microscopy

The stained individuals were placed between slide and coverslip and observed under a reflected fluorescence microscope Olympus BX43. WGA-FITC was excited with the 460/495 nm line from a 100 W mercury lamp with an interference excitation filter (BP460), and collected with a 505 nm dichroic mirror (DM505) and a 510 nm interference barrier filter (BA510IF). DAPI fluorescence was excited with the 340/390 nm line from a 100 W mercury lamp with an interference excitation filter (BP340), and collected with a 410 nm dichroic mirror (DM410) and a 420 nm interference barrier filter (BA420IF). Selected *Oithona* individuals were photographed with a 16-megapixel camera using the ToupView software (v.3.7). For each individual, three photographs were taken: one in polarized light, one with the WGA-FITC fluorescence and one with the DAPI fluorescence. Some colour adjustments were made with the ImageJ software (*Schneider, Rasband & Eliceiri, 2012*).

RESULTS

Oithona morphology with WGA-FITC microscopy

The *Oithona* chitin was labelled with WGA-FITC directly on the individuals and observed by fluorescence microscopy. The setae and spines of the exopod segments of the five leg pairs could be identified and counted on *O. nana* (Fig. 1A). These first results revealed the chitinous structure of the setae and the spines, and could provide a rapid method for taxonomical identification. However, because of the individuals and setae position on the plate, we were not able to identify and count the setae of all tested individuals. Chitinous elliptic or spherical structures of unknown function and larger than 6 μm (Fig. 1A) were also visible in the exopods of the swimming legs. These globular structures were observed in both sexes of *O. nana* (Figs. 1A and 2A–2E), but only in female individuals of *O. similis* (Figs. 2F and 2G). They may also be smaller (Fig. 2F), or absent (Figs. 3A and 4C) in other individuals. These structures can also be present in other exopod segments (Fig. 2A). Another tubular structure, in the distal part of the exopods three of the right third leg, right and left fourth legs and right and left fifth legs were noticeable (Fig. 1A). In other *Oithona* individuals, these tubular structures appear to be attached to the globular structure (Figs. 2B, 2D and 2G).

Chitin distribution in the *Oithona* digestive system

Chitin was detected all along the digestive system, from the stomach to the hindgut of the nauplius (Fig. 1B) and adults (Figs. 1C–1E) of the two species. The exoskeleton chitin was also stained by the WGA-FITC, which allowed a clear identification of the stomach in the prosome, of the midgut in the prosome and in the urosome and the hindgut in the urosome. Along the digestive system, the chitin had a microfibrillar structure aligned along the antero-posterior axis with regions showing higher microfibrils density, especially the anterior midgut and some stomach areas (Figs. 1C–1F). Some individuals contained in their anterior and posterior midgut one or several elliptical faecal pellets completely engulfed by chitin (Figs. 1E and 4B). However, no faecal pellets were found in the nauplius. In the anterior and posterior midgut, we observed several shrinkages

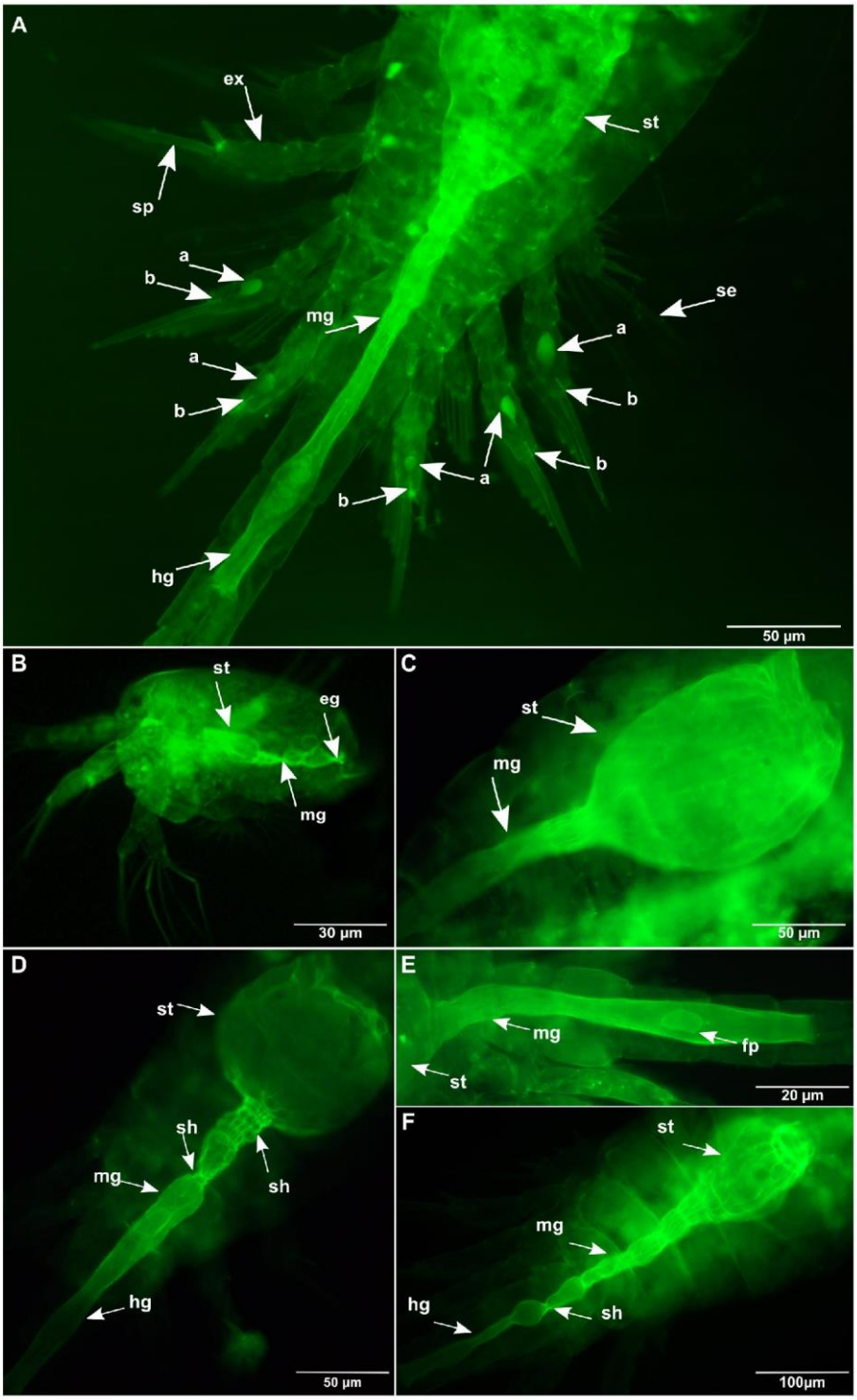


Figure 1 *Oithona* appendages morphology and digestive system by WGA-FITC fluorescence microscopy. (A) Dorsal view of the *O. nana* female swimming appendages. (B) Lateral view of the *Oithona* nauplius digestive system. (C) Lateral view of the *O. nana* female stomach. (D) Dorsal view of the *O. nana* female stomach. (E) Lateral view of an *O. nana* male gut. (F) Dorsal view of an *O. similis* female adult stomach. st, stomach; mg, midgut; hg, hindgut; sh, shrinkage; ex, exopod; se, seta; sp, spine; fp, faecal pellet; a, globular structure; b, tubular structure. Full-size DOI: 10.7717/peerj.4685/fig-1

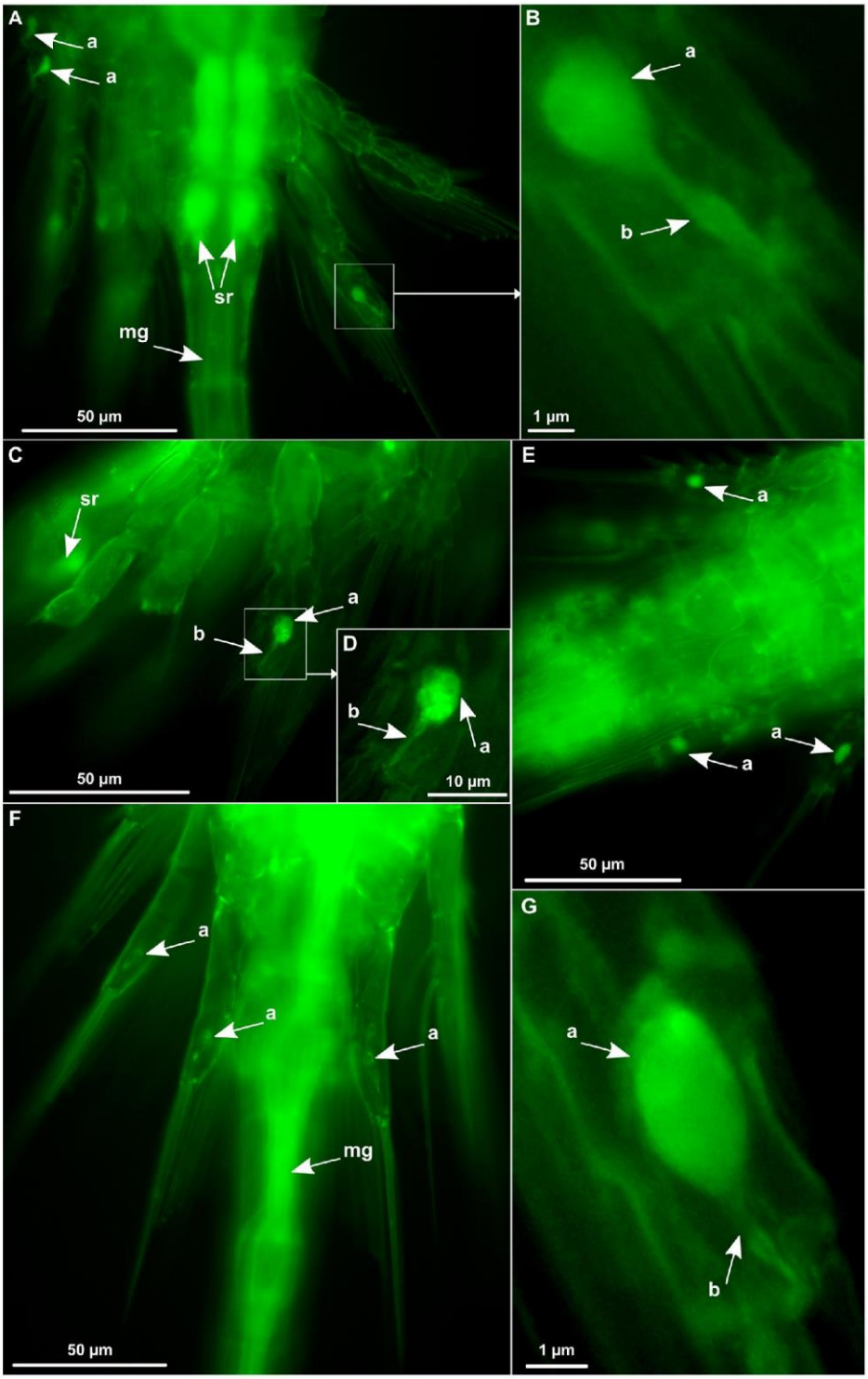


Figure 2 ***Oithona*** **globular and tubular chitinous structures in the swimming appendages by WGA-FITC fluorescence microscopy.** (A) Dorsal view of *O. nana* female urosome/prosome junction. (B) Zoom on the right P4 exopod of the same *O. nana* individual. (C) Lateral view of *O. nana* female swimming appendages. (D) Zoom on the right P3 exopod of the same *O. nana* individual (E) Ventral view of *O. nana* male abdomen. (F) Dorsal view of *O. similis* female urosome/prosome junction. (G) Zoom on the right P5 exopod of a female *O. similis*. sr, seminal receptacle; mg, midgut; a, globular structure; b, tubular structure. Full-size DOI: 10.7717/peerj.4685/fig-2

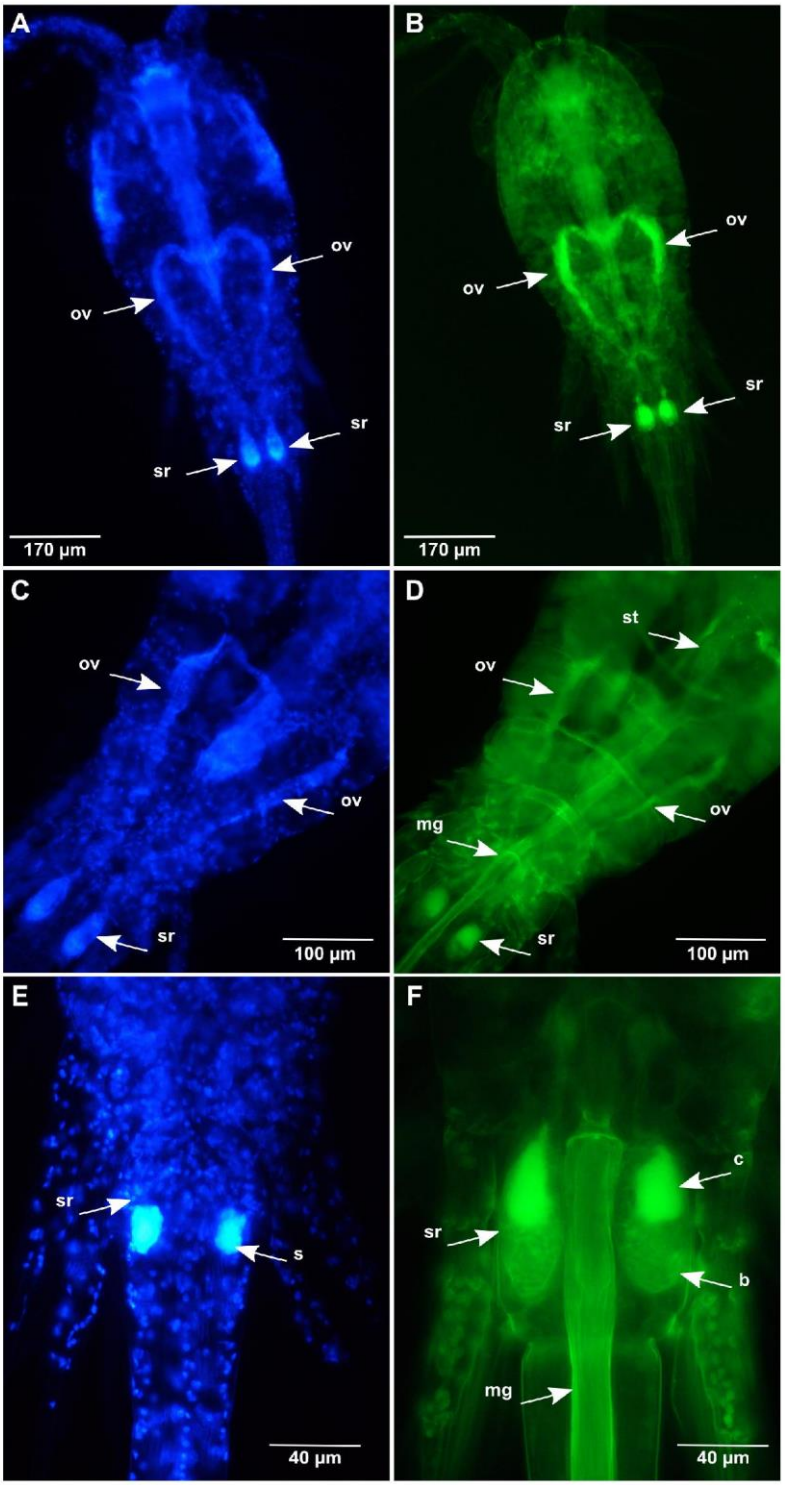


Figure 3 ***Oithona*** **female reproductive system by DAPI and WGA-FITC fluorescence microscopy.** (A and B) Dorsal view of the *O. nana* female reproductive system (DAPI staining on the left and WGA-FITC staining on the right). (C and D) Dorsal view of the *O. similis* female reproductive system (DAPI staining on the left and WGA-FITC staining on the right). (E and F) Dorsal view of the *O. nana* female double sexual somite (DAPI staining on the left and WGA-FITC staining on the right). mg, midgut; sr, seminal receptacle; s, semen; ov, oviduct; hg, hindgut; b, diffuse chitin region; c, chitin rich region.

Full-size DOI: 10.7717/peerj.4685/fig-3

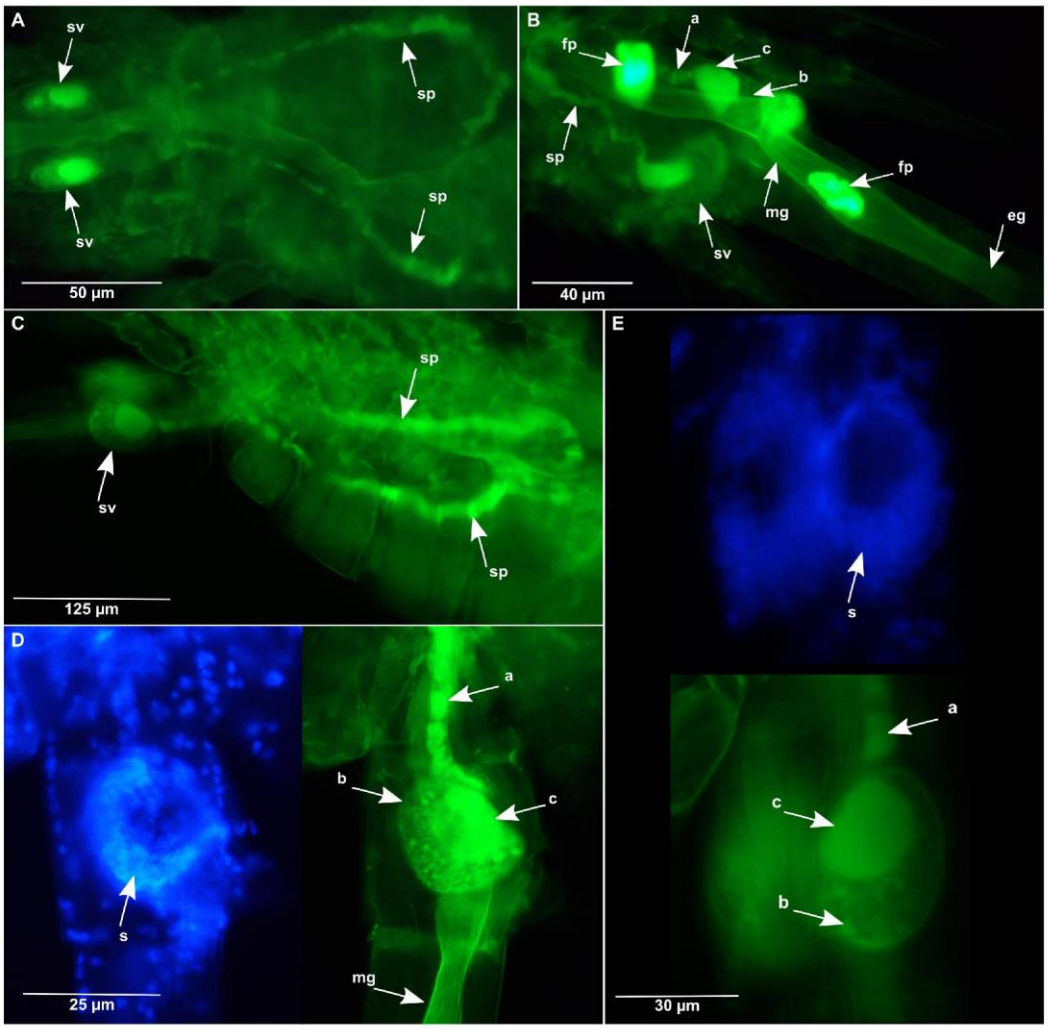


Figure 4 *Oithona* male reproductive system by DAPI and WGA-FITC fluorescence microscopy. (A) Dorsal view of the *O. nana* male reproductive system. (B) Dorso-lateral view of the *O. nana* seminal vesicle. (C) Lateral view of the male *O. similis* reproductive system. (D) Lateral view of the *O. nana* male double sexual somite. (E) Lateral view of the *O. similis* male double sexual somite. mg, midgut; fp, faecal pellet; sv, seminal vesicle; sp, spermiduct; a, heterogeneous chitin; b, diffuse chitin; c, chitin rich region.
Full-size DOI: 10.7717/peerj.4685/fig-4

at different interval distances corresponding to midgut contractions. In certain cases, several shrinkages (up to four) were separated by less than 5 µm (Fig. 1D), while other individuals showed more distant shrinkages (Fig. 1F).

Chitin distribution in the *Oithona* reproductive system

The DAPI and WGA-FITC stainings on *Oithona* females allowed the identification of the ovaries and the oviducts that presented a heart shape in the middle of the prosome (Figs. 3A and 3B) as previously described by Mironova and Pasternak. The oviducts start from each lateral side of the gonads to the seminal receptacle in the genital double somite (the first two segments of the urosome). Comparing to the microfibrillar structure of the chitin found in the digestive system, the chitin staining in the reproductive

system was mainly amorphous. Besides, its distribution was discontinuous along the ducts, altering chitin-rich and poor areas (Figs. 3C and 3D).

In females, we distinguished two parts forming the seminal receptacle (Figs. 3E and 3F). The first part was chitin-rich and located in the anterior region of the receptacle. The chitin distribution between the anterior receptacle and the oviduct was discontinuous. The second part was located in the posterior receptacle and contained less and sparser chitin, presenting a mix of microfibrillar and amorphous structures. Thanks to DAPI staining, in some females the presence of the DNA rich material in the posterior region of the seminal receptacle was observed and was likely to be male semen.

In males, the chitin staining allowed the identification of the spermiducts, which presented the same chitin pattern observed in the oviducts (Figs. 4A and 4C). The spermiducts probably start from each side of the male gonads (not visible on the pictures) to the seminal vesicles, in the sexual somite (Fig. 4B). As for the female seminal receptacle, the male seminal vesicle can be divided into two parts (Figs. 4D and 4E). The first part of the vesicles is chitin-rich, located in the anterior region of the vesicle. The distribution of the chitin from this upper part of the vesicle to the spermiduct was not continuous. The second part, located in the posterior region of the vesicle, was observed by DAPI staining and was likely to be filled by DNA-rich male semen.

DISCUSSION

Comparing to previous staining methods used to observe the digestive and reproductive systems of copepods (*Batchelder, 1986*; *Eisfeld & Niehoff, 2007*; *Mironova & Pasternak, 2017*; *Niehoff, 2003*; *Niehoff & Hirche, 1996*; *Tande & Gronvik, 1983*; *Tande & Hopkins, 1981*; *Wolfram, Nejstgaard & Pohnert, 2014*), the protocol proposed here allows a clear insight into the chitin distribution in these systems. Moreover, this protocol is simple and rapid, taking a few minutes of manipulation, 30 min of incubation, and can be used on living, but also on alcohol-preserved copepods. The main limit of our method remains in the short-time staining of the WGA-FITC: a picture must be taken, a few minutes after fluorescence excitation to save any microscopic observation without loss of quality. Furthermore, for the reproductive system, the DAPI staining allows only the observation of the gonad structure; while the Mironova and Pasternak protocol allows a better identification of the oocytes.

The use of WGA-FITC revealed chitinous spherical structures in the exopods of the swimming legs in *O. nana* males and females and in *O. similis* females, which were not observed in previous studies. The absence of these structures in *O. similis* male individuals may be a bias due to their low presence in our samples. Despite, luminescence is not conspicuous in *Oithona*, these structures could be luminous glands (*Herring, 1988*). The green staining revealed also a chitinous tubular structure in the exopods penultimate segment of the swimming legs. These structures resemble the 'Crusalis organ,' an osmoregulatory structure that was described by *Johnson et al. (2014)* from the coastal/estuarine copepod *Eurytemora affinis*. Since the globular and the tubular structures seemed attached, we suggest they form only one organ involved either in bioluminescence, osmoregulation or both.

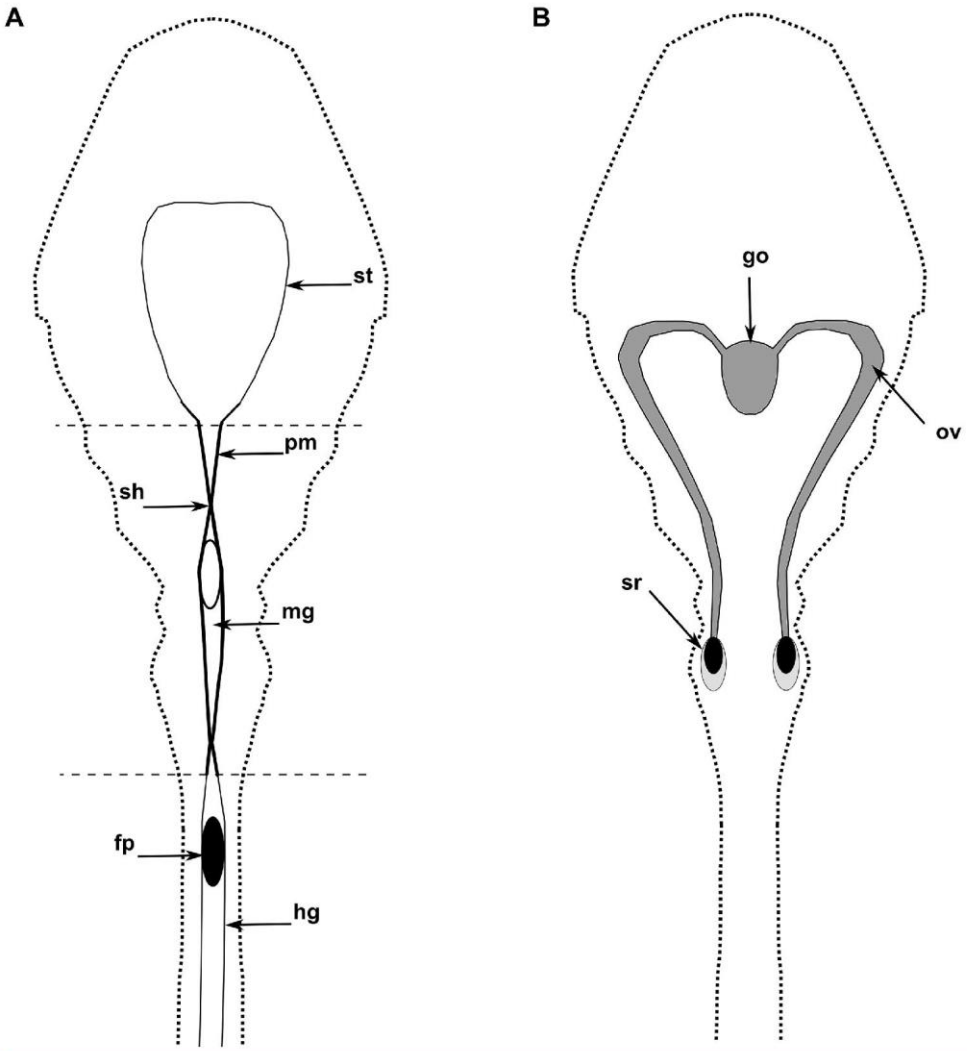


Figure 5 The diagram of the internal anatomy of a female *O. nana*. (A) Diagram of the dorsal view of the digestive system. (B) Diagram of the dorsal view of the reproductive system. st, stomach; mg, midgut; pm, peritrophic membrane; go, gonads; sh, shrinkage; fp, faecal pellet; hg, hindgut; sr, seminal receptacle; ov, oviduct. Thick black zones correspond to chitin rich areas. Dark grey zones correspond to heterogeneous chitin area. Light grey zones correspond to amorphous chitin areas.

Full-size DOI: 10.7717/peerj.4685/fig-5

The WGA-FITC staining allowed also the identification of the chitin distribution in the *Oithona* organs, which provides a high-quality view of the external and internal anatomy and pointed out the major role of chitin in the *Oithona* digestion and reproduction. According to insect studies, the distribution of chitin in the digestive system is limited to the midgut (*Hegedus et al., 2009*; *Terra, 2001*). The same chitin distribution was observed in decapods (*Martin et al., 2006*; *Wang et al., 2012*). In both *Oithona* species, we detected chitin throughout the digestive system, which distinguishes it from insects and decapods, but seems consistent with observations in other free-living cyclopoids made by *Yoshikoshi & Kô (1988)*.

In the PM of some insects (*Hegedus et al., 2009*; *Kelkenberg et al., 2015*; *Lehane, 1997*), chitin plays a role in protection (chemical, mechanical and against viruses, bacteria and pathogens) and digestion (*Terra, 2001*). As the synthesis of chitin has a significant metabolic cost for the organism, we hypothesized that, like the insects and decapods PM, the formation of a chitin coat around faecal pellets help to protect against toxins and pathogens that were not degraded during digestion.

In copepods, no evidence of midgut contraction has previously been described although the phenomenon has been suggested at several instances (*Gauld, 1957*). We suppose that the midgut shrinkages observed in this study could play a key role in the formation and motion of the faecal pellets to the anus. However, we observed intestine shrinkages without the presence of faecal pellets, and vice versa. As proposed by Yoshikoshi and Kô for other copepods, we also suggest that, in *Oithona*, the formation of chitin coat around the faecal pellets can be produced by engulfing digested food in chitin microfibrils present in the PM of the anterior midgut (Fig. 5; *Yoshikoshi & Kô, 1988*).

The presence of chitin along the oviduct and spermiduct walls validates the cuticular appearance of the ducts described by *Cuoc et al. (1997)*. In all *Oithona* males, we observed a pair of spermiducts, while in *Calanus finmarchicus* one of the two spermiducts disappeared during the male differentiation (*Tande & Hopkins, 1981*). The bipartite structure of the seminal receptacles and vesicles found in *O. nana* and *O. similis* males and females were very similar. In males, we hypothesized that the chitinous structure of the vesicle plays a role in the holding of the spermatophores during their formation. Likewise, in the females, this structure would play a role in the holding of the ovisac but also in the opening and closing of the oviduct to release oocytes in the seminal receptacle.

CONCLUSION

With this study, we adapted and tested a simple and rapid chitin-staining protocol that can help to the taxonomic identification of copepods, and enable new studies on copepod comparative anatomy at a larger scale. The application of the method to *Oithona* extended the knowledge of the structure of its digestive and reproductive systems. Considering the important role of copepods in the carbon and nitrogen sequestration through chitin synthesis, more efforts should be undergone to better understand the molecular and physiological mechanisms involved in faecal pellets formation.

ACKNOWLEDGEMENTS

We thank Julie Poulain for initiating the *Oithona* genome project and Dr. Leocadio Blanco-Bercial for helpful comments on the manuscript.

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

This work was supported by the Commissariat à l'Energie Atomique et aux Energies Alternatives, the French Ministry of Research and OCEANOMICS (ANR-11-BTBR-0008).

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Grant Disclosures

The following grant information was disclosed by the authors:
French Ministry of Research and OCEANOMICS: ANR-11-BTBR-0008.

Competing Interests

The authors declare that they have no competing interests.

Author Contributions

- Kevin Sugier performed the experiments, analysed the data, contributed reagents/materials/analysis tools, prepared figures and/or tables, approved the final draft.
- Benoit Vacherie conceived and designed the experiments, performed the experiments, contributed reagents/materials/analysis tools, approved the final draft.
- Astrid Cornils authored or reviewed drafts of the paper, approved the final draft.
- Patrick Wincker authored or reviewed drafts of the paper, approved the final draft.
- Jean-Louis Jamet contributed reagents/materials/analysis tools, authored or reviewed drafts of the paper, approved the final draft.
- Mohammed-Amin Madoui conceived and designed the experiments, performed the experiments, analysed the data, contributed reagents/materials/analysis tools, prepared figures and/or tables, authored or reviewed drafts of the paper, approved the final draft.

Data Availability

The following information was supplied regarding data availability:

The raw data are provided in the article as photomicrographs.

REFERENCES

Allen AK, Neuberger A, Sharon N. 1973. The purification, composition and specificity of wheat-germ agglutinin. *Biochemical Journal* **131(1)**:155–162 DOI 10.1042/bj1310155.

Batchelder HP. 1986. A staining technique for determining copepod gonad maturation—application to Metridia-Pacifica from the northeast Pacific-Ocean. *Journal of Crustacean Biology* **6(2)**:227–231 DOI 10.2307/1547983.

Bathmann UV, Noji TT, Voss M, Peinert R. 1987. Copepod fecal pellets—abundance, sedimentation and content at a permanent station in the Norwegian Sea in May/June 1986. *Marine Ecology Progress Series* **38**:45–51 DOI 10.3354/Meps038045.

Bautista B, Harris RP. 1992. Copepod gut contents, ingestion rates and grazing impact on phytoplankton in relation to size structure of zooplankton and phytoplankton during a spring bloom. *Marine Ecology Progress Series* **82**:41–50 DOI 10.3354/Meps082041.

Beaugrand G, Brander KM, Alistair Lindley J, Souissi S, Reid PC. 2003. Plankton effect on cod recruitment in the North Sea. *Nature* **426(6967)**:661–664 DOI 10.1038/nature02164.

Borradaile LA, Potts FA. 1935. Class Copepoda. In: *The Invertebrata: A Manual for the Use of Students*. New York: The MacMillan Company, 370–376.

Chang C. 2013. A New Species of Cletocamptus Copepoda (Harpacticoida, Canthocamptidae) from Salt Marshes in Korea. *Animal Systematics, Evolution and Diversity* 29(**3**):227–237 DOI 10.5635/ased.2013.29.3.227.

Cornils A, Wend-Heckmann B, Held C. 2017. Global phylogeography of *Oithona similis* s.l. (Crustacea, Copepoda, Oithonidae)—a cosmopolitan plankton species or a complex of cryptic lineages? *Molecular Phylogenetics and Evolution* **107**:473–485 DOI 10.1016/j.ympev.2016.12.019.

Cuoc C, Defaye D, Brunet M, Notonier R, Mazza J. 1997. Female genital structures of Metridinidae (Copepoda: Calanoida). *Marine Biology* **129(4)**:651–665 DOI 10.1007/s002270050208.

Debes H, Eliasen K, Gaard E. 2008. Seasonal variability in copepod ingestion and egg production on the Faroe shelf. *Hydrobiologia* **600(1)**:247–265 DOI 10.1007/s10750-007-9238-3.

Dussart BH, Defaye D. 2001. *Introduction to the Copepoda*. Leiden: Backhuys.

Eisfeld SM, Niehoff B. 2007. Gonad morphology, oocyte development and spawning cycle of the calanoid copepod *Acartia clausi*. *Helgoland Marine Research* **61(3)**:193–201 DOI 10.1007/s10152-007-0066-7.

El Gueddari NE, Rauchhaus U, Moerschbacher BM, Deising HB. 2002. Developmentally regulated conversion of surface-exposed chitin to chitosan in cell walls of plant pathogenic fungi. *New Phytologist* **156(1)**:103–112 DOI 10.1046/j.1469-8137.2002.00487.x.

Farnesi LC, Menna-Barreto RF, Martins AJ, Valle D, Rezende GL. 2015. Physical features and chitin content of eggs from the mosquito vectors *Aedes aegypti*, *Anopheles aquasalis* and *Culex quinquefasciatus*: connection with distinct levels of resistance to desiccation. *Journal of Insect Physiology* **83**:43–52 DOI 10.1016/j.jinsphys.2015.10.006.

Fones HN, Mardon C, Gurr SJ. 2016. A role for the asexual spores in infection of *Fraxinus excelsior* by the ash-dieback fungus *Hymenoscyphus fraxineus*. *Scientific Reports* **6(1)**:34638 DOI 10.1038/srep34638.

Frangoulis C, Christou ED, Hecq JH. 2004. Comparison of marine copepod outfluxes: nature, rate, fate and role in the carbon and nitrogen cycles. *Advances in Marine Biology* **47**:253–309 DOI 10.1016/S0065-2881(04)47004-7.

Gallienne CP, Robins DB. 2001. Is Oithona the most important copepod in the world's oceans? *Journal of Plankton Research* **23(12)**:1421–1432 DOI 10.1093/plankt/23.12.1421.

Gauld DT. 1957. A peritrophic membrane in calanoid copepods. *Nature* **179(4554)**:325–326 DOI 10.1038/179325a0.

Godoy RS, Fernandes KM, Martins GF. 2015. Midgut of the non-hematophagous mosquito *Toxorhynchites theobaldi* (Diptera, Culicidae). *Scientific Reports* **5(1)**:15836 DOI 10.1038/srep15836.

Haberyan KA. 1985. The role of copepod fecal pellets in the deposition of diatoms in Lake Tanganyika. *Limnology and Oceanography* **30(5)**:1010–1023 DOI 10.4319/lo.1985.30.5.1010.

Hegedus D, Erlandson M, Gillott C, Toprak U. 2009. New insights into peritrophic matrix synthesis, architecture, and function. *Annual Review of Entomology* **54(1)**:285–302 DOI 10.1146/annurev.ento.54.110807.090559.

Herring PJ. 1988. Copepod luminescence. *Hydrobiologia* **167–168(1)**:183–195 DOI 10.1007/Bf00026304.

Humes AG. 1994. How many copepods? *Hydrobiologia* **292–293(1)**:1–7 DOI 10.1007/BF00229916.

Huys R, Boxshall GA. 1991. *Copepod Evolution*. London: The Ray Society.

Johnson KE, Perreau L, Charmantier G, Charmantier-Daures M, Lee CE. 2014. Without gills: localization of osmoregulatory function in the copepod *Eurytemora affinis*. *Physiological and Biochemical Zoology* **87(2)**:310–324 DOI 10.1086/674319.

Jonasdottir SH, Visser AW, Richardson K, Heath MR. 2015. Seasonal copepod lipid pump promotes carbon sequestration in the deep North Atlantic. *Proceedings of the National Academy of Sciences of the United States of America* **112(39)**:12122–12126 DOI 10.1073/pnas.1512110112.

Kelkenberg M, Odman-Naresh J, Muthukrishnan S, Merzendorfer H. 2015. Chitin is a necessary component to maintain the barrier function of the peritrophic matrix in the insect midgut. *Insect Biochemistry and Molecular Biology* **56**:21–28 DOI 10.1016/j.ibmb.2014.11.005.

Kellogg VL. 1902. Branch Arthropoda: the crustaceans, centipeds, insects, and spiders. In: *Elementary Zoology*. New York: Henry Holt & Company, 148.

Kirchner M. 1995. Microbial colonization of copepod body surfaces and chitin degradation in the sea. *Helgoländer Meeresuntersuchungen* **49(1–4)**:201–212 DOI 10.1007/Bf02368350.

Lehane MJ. 1997. Peritrophic matrix structure and function. *Annual Review of Entomology* **42(1)**:525–550 DOI 10.1146/annurev.ento.42.1.525.

Lin MS, Comings DE, Alfi OS. 1977. Optical studies of the interaction of 4′-6′-diamidino-2-phenylindole with DNA and metaphase chromosomes. *Chromosoma* **60(1)**:15–25 DOI 10.1007/bf00330407.

Madoui MA, Poulain J, Sugier K, Wessner M, Noel B, Berline L, Labadie K, Cornils A, Blanco-Bercial L, Stemmann L, Jamet JL, Wincker P. 2017. New insights into global biogeography, population structure and natural selection from the genome of the epipelagic copepod Oithona. *Molecular Ecology* **26(17)**:4467–4482 DOI 10.1111/mec.14214.

Marques TM, Clebsh L, Córdova L, Boeger WA. 2017. Ergasilus turkayi n. sp. (Copepoda, Cyclopoida, Ergasilidae): a gill parasite of Serrasalmus hollandi Jégu, 2003 (Characiformes, Serrasalmidae) from the Paragùa River, Bolivia. *Nauplius* **25**:e2017020 DOI 10.1590/2358-2936e2017020.

Martin GG, Simcox R, Nguyen A, Chilingaryan A. 2006. Peritrophic membrane of the penaeid shrimp *Sicyonia ingentis*: structure, formation, and permeability. *Biological Bulletin* **211(3)**:275–285 DOI 10.2307/4134549.

Michels J, Büntzow M. 2010. Assessment of Congo red as a fluorescence marker for the exoskeleton of small crustaceans and the cuticle of polychaetes. *Journal of Microscopy* **238(2)**:95–101 DOI 10.1111/j.1365-2818.2009.03360.x.

Mironova E, Pasternak A. 2017. Female gonad morphology of small copepods *Oithona similis* and *Microsetella norvegica*. *Polar Biology* **40(3)**:685–696 DOI 10.1007/s00300-016-1993-z.

Mravec J, Kracun SK, Rydahl MG, Westereng B, Miart F, Clausen MH, Fangel JU, Daugaard M, Van Cutsem P, De Fine Licht HH, Hofte H, Malinovsky FG, Domozych DS, Willats WG. 2014. Tracking developmentally regulated post-synthetic processing of homogalacturonan and chitin using reciprocal oligosaccharide probes. *Development* **141(24)**:4841–4850 DOI 10.1242/dev.113365.

Niehoff B. 2003. Gonad morphology and oocyte development in Pseudocalanus spp. in relation to spawning activity. *Marine Biology* **143(4)**:759–768 DOI 10.1007/s00227-003-1034-7.

Niehoff B, Hirche HJ. 1996. Oogenesis and gonad maturation in the copepod *Calanus finmarchicus* and the prediction of egg production from preserved samples. *Polar Biology* **16(8)**:601–612 DOI 10.1007/s003000050095.

Nishida S. 1985. *Taxonomy and Distribution of the Family Oithonidae (Copepoda, Cyclopoida) in the Pacific and Indian Oceans*. Tokyo: Ocean Research Institute, University of Tokyo.

Razouls C, De Bovée F, Kouwenberg J, Desreumaux N. 2005–2018. Diversity and geographic distribution of marine planktonic copepods. *Available at https://copepodes.obs-banyuls.fr/*.

Schneider CA, Rasband WS, Eliceiri KW. 2012. NIH Image to ImageJ: 25 years of image analysis. *Nature Methods* **9(7)**:671–675 DOI 10.1038/nmeth.2089.

Steinberg DK, Goldthwait SA, Hansell DA. 2002. Zooplankton vertical migration and the active transport of dissolved organic and inorganic nitrogen in the Sargasso Sea. *Deep-Sea Research Part I: Oceanographic Research Papers* **49(8)**:1445–1461 DOI 10.1016/S0967-0637(02)00037-7.

Steinberg DK, Landry MR. 2017. Zooplankton and the ocean carbon cycle. *Annual Review of Marine Science* **9(1)**:413–444 DOI 10.1146/annurev-marine-010814-015924.

Tande KS, Gronvik S. 1983. Ecological investigations on the zooplankton community of Balsfjorden, northern Norway—sex-ratio and gonad maturation cycle in the copepod Metridia-longa (Lubbock). *Journal of Experimental Marine Biology and Ecology* **71(1)**:43–54 DOI 10.1016/0022-0981(83)90103-X.

Tande KS, Hopkins CCE. 1981. Ecological investigations of the zooplankton community of Balsfjorden, northern Norway—the genital system in *Calanus finmarchicus* and the role of gonad development in overwintering strategy. *Marine Biology* **63(2)**:159–164 DOI 10.1007/Bf00406824.

Terra WR. 2001. The origin and functions of the insect peritrophic membrane and peritrophic gel. *Archives of Insect Biochemistry and Physiology* **47(2)**:47–61 DOI 10.1002/arch.1036.

Valdés VP, Fernandez C, Molina V, Escribano R, Joux F. 2017. Dissolved compounds excreted by copepods reshape the active marine bacterioplankton community composition. *Frontiers in Marine Science* **4**:343 DOI 10.3389/fmars.2017.00343.

Wang L, Li F, Wang B, Xiang J. 2012. Structure and partial protein profiles of the peritrophic membrane (PM) from the gut of the shrimp *Litopenaeus vannamei*. *Fish & Shellfish Immunology* **33(6)**:1285–1291 DOI 10.1016/j.fsi.2012.09.014.

Wolfram S, Nejstgaard JC, Pohnert G. 2014. Accumulation of polyunsaturated aldehydes in the gonads of the copepod *Acartia tonsa* revealed by tailored fluorescent probes. *PLOS ONE* **9(11)**:e112522 DOI 10.1371/journal.pone.0112522.

Yoshikoshi K, Kô Y. 1988. Structure and function of the peritrophic membranes of copepods. *Nippon Suisan Gakkaishi* **54(7)**:1077–1082 DOI 10.2331/suisan.54.1077.

Zamora-Terol S, Kjellerup S, Swalethorp R, Saiz E, Nielsen TG. 2014. Population dynamics and production of the small copepod Oithona spp. in a subarctic fjord of West Greenland. *Polar Biology* **37(7)**:953–965 DOI 10.1007/s00300-014-1493-y.

Chapter 2:

The *Oithona nana* genome shows an explosion of LNR domain targeted by natural selection

We wanted to construct the genome of three small zooplanktons living in different temperature ranges; and to determine their biogeography, their population genetic structure and also to identify loci under selection using the *Tara* Oceans metagenomics data. We chose the cosmopolitan copepod genus: *Oithona.* Because of the small amount of DNA by individuals, and the large genome size of some species (more than five gigabases), we focused further only on the *Oithona nana* genome.

This study provided the first genome of a Cyclopoid copepod; and the third genome of copepods. I performed comparative genomic analyses and observed the structural differences between Cylopoids, Calanoids and Harpacticoids. Comparative analysis at a functional level showed the explosion of Lin-12 Notch Repeat (LNR) protein domain in the *O. nana* deduced proteome. Using the metagenomic data of the *Tara* Oceans expeditions, the *O. nana* genetic structure in the Mediterranean Sea was established, and genes under selection were detected. One of the genes, possessing only LNR domains, was detected under selection and was expressed only in males.

In this study, my role was to perform the comparative genomic analysis. I also had to write the material and method of my analysis and make the first figure and the first appendix table, with the help of my co-authors.

Article source

Madoui, Mohammed-Amin, Julie Poulain, Kevin Sugier, Marc Wessner, Benjamin Noel, Leo Berline, Karine Labadie, et al. 2017. 'New Insights into Global Biogeography, Population Structure and Natural Selection from the Genome of the Epipelagic Copepod *Oithona*'. Molecular Ecology 26 (17): 4467–82. https://doi.org/10.1111/mec.14214.

A French abstract is available in appendix 2 (page 153)

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

DOI: 10.1111/mec.14214

ORIGINAL ARTICLE

WILEY MOLECULAR ECOLOGY

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

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

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

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

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

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

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

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

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

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

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

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

Abstract

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

KEYWORDS
genome, Mediterranean Sea, phylogeography, selection

1 | INTRODUCTION

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

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

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

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

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

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

2 | MATERIALS AND METHODS

2.1 | Genome sequencing

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

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

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

2.2 | Genome assembly

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

kept in the final assembly. The gene completeness of the assembly was estimated with CEGMA 2.4.010312 (Parra, Bradnam, & Korf, 2007).

2.3 | mRNA sequencing and assembly

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

2.4 | Genome annotation

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

2.5 | Comparative genomic analysis of copepods

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

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

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

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

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

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

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

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

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

3 | RESULTS

3.1 | Genome assembly and annotation of *Oithona nana*

The *O. nana* somatic genome was sequenced and assembled in 7,375 contigs with a N50 of 39.6 kb that were linked in 4,626 scaffolds with a N50 of ~400 kb (Appendix S9). After gap closing, only 3.7% of undetermined bases remained. The final genome sequence contains 89% of the 458 conserved core eukaryotic proteins used by CEGMA with an average of 1.2 copies per gene. Thus, the sequencing strategy (i.e., pooled individuals) and the assembly based on dipSpades successfully merged the different haplotypes present in the initial genomic DNA into a single representative haplotype suitable for gene annotation. Low-complexity and repetitive elements represented 3.6% of the genome sequence and were masked

prior to the genome annotation. The genome annotation procedure applied with GMOVE predicted 15,359 genes with 23% monoexonic genes. The resulting proteome was scanned to identify conserved protein domains and 72.3% of the proteins harboured at least one conserved domain present in the INTERPRO database.

3.2 | Comparative analysis of the *Oithona nana* genome

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

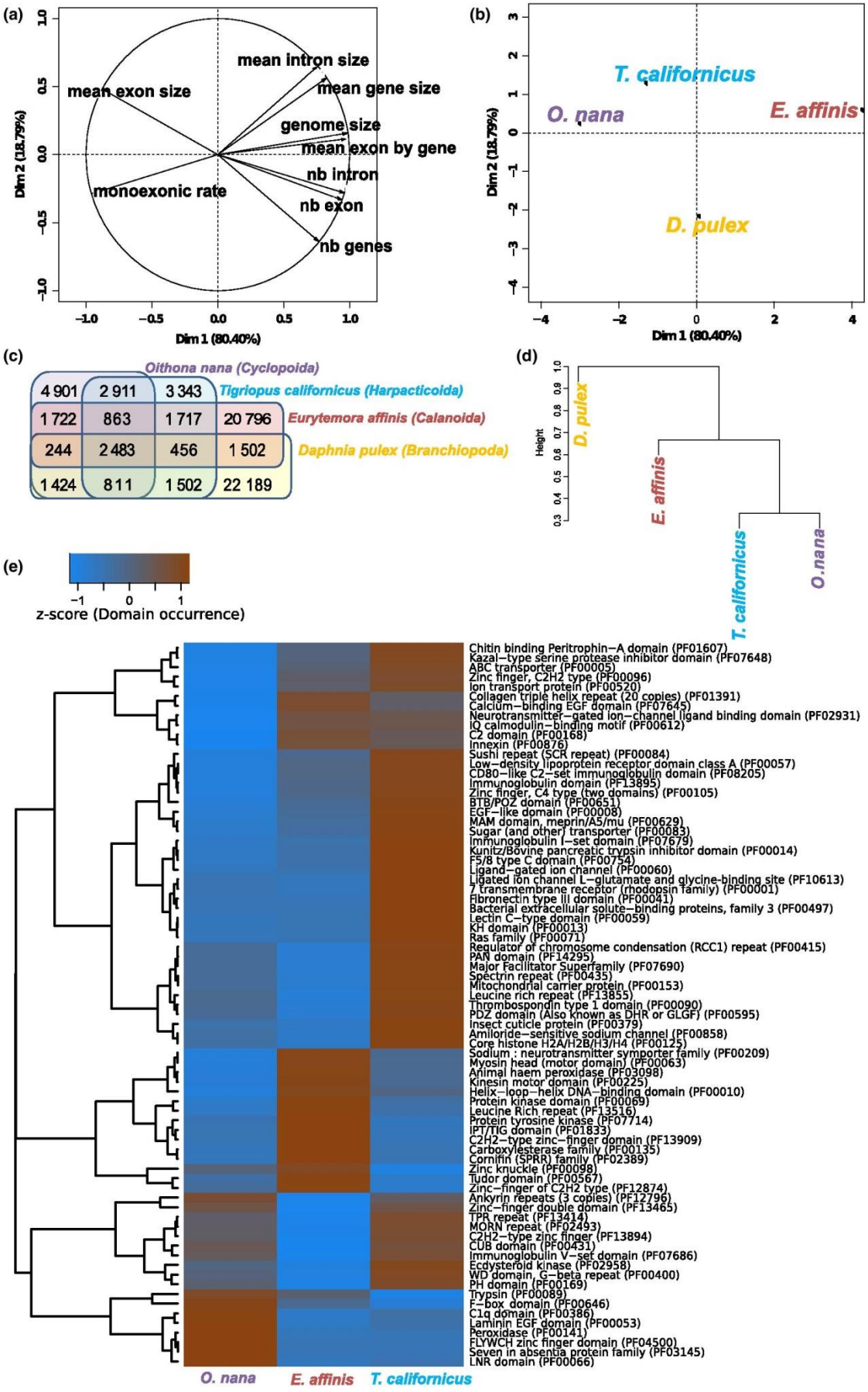
3.3 | Functional protein domains

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

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

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

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


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

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

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

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

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

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

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

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

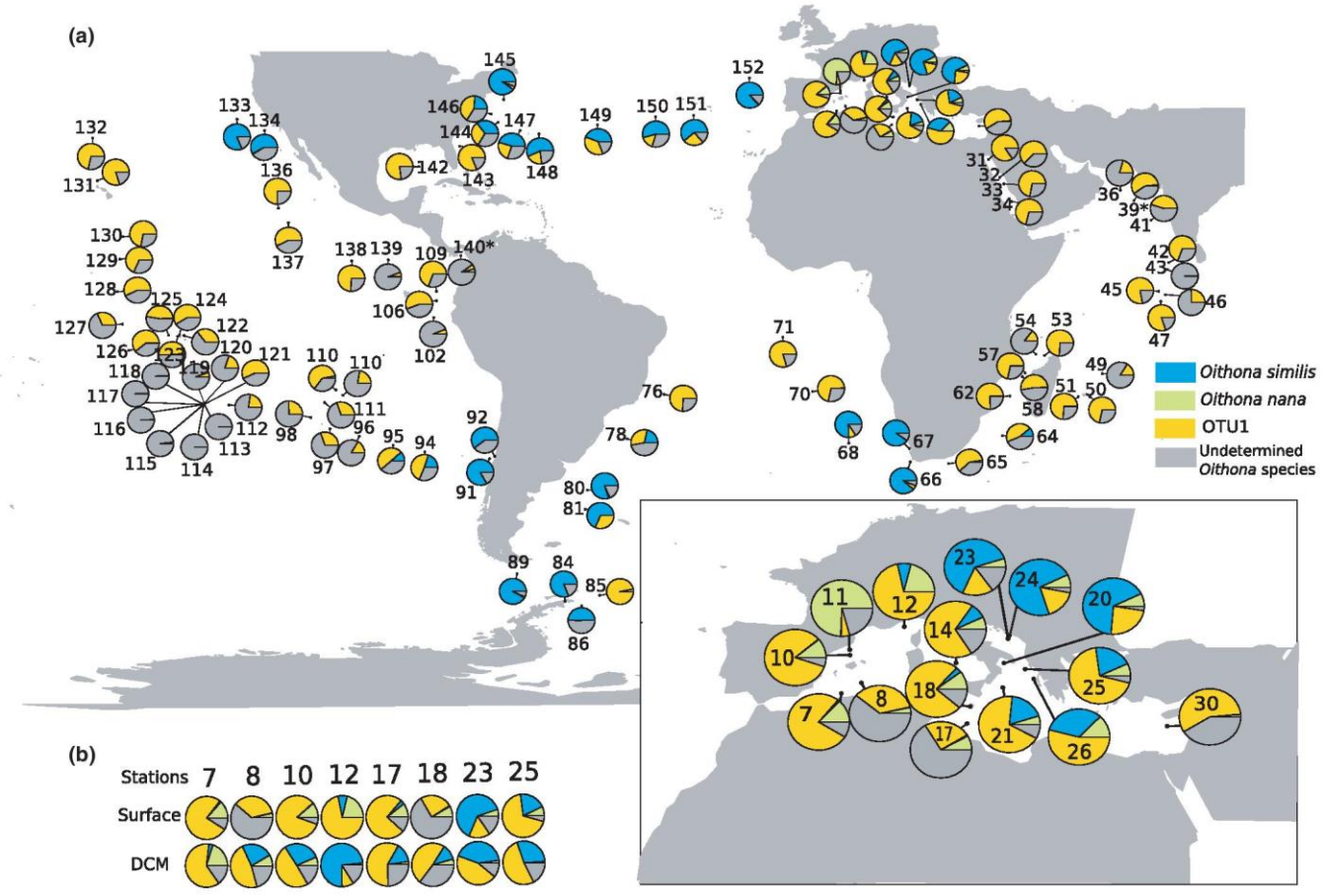


FIGURE 2 Biogeography of *Oithona* species. (a) Global biogeography of the *Oithona* species using the *Tara* Oceans metagenomic data and 28S sequences. The proportion of each species is represented by a pie chart; asterisks correspond to *Oithona nana* presence with a very low abundance. The MS region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

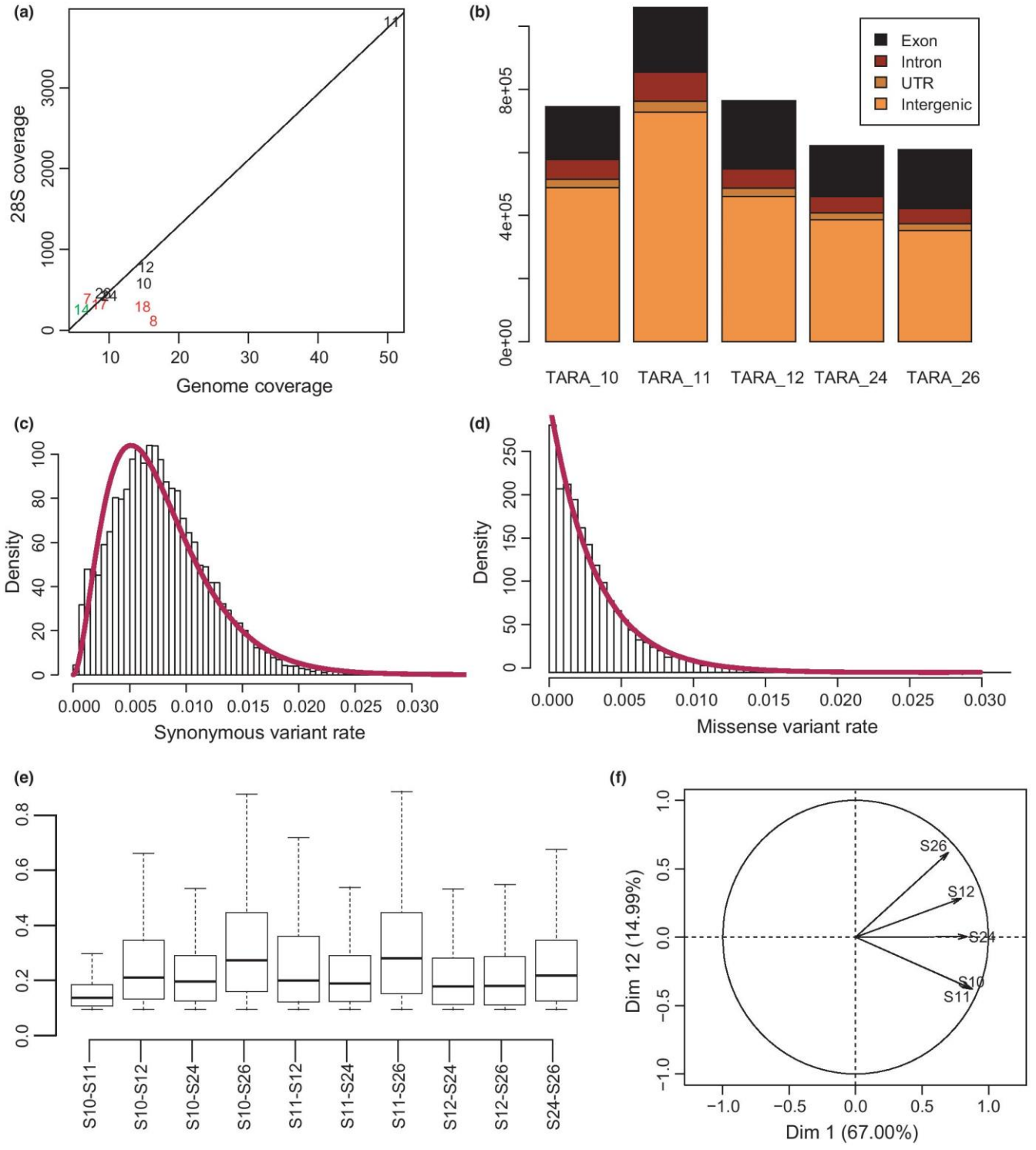


FIGURE 3 Genomic variants in *Oithona nana* populations of the Mediterranean Sea. (a) Correlation between the 28S and the genomic coverage. Station numbers are plotted; numbers in red correspond to stations discarded due to a bimodal distribution of the identity percentage. Stations in green correspond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

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

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

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

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

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

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

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

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

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

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

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

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

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

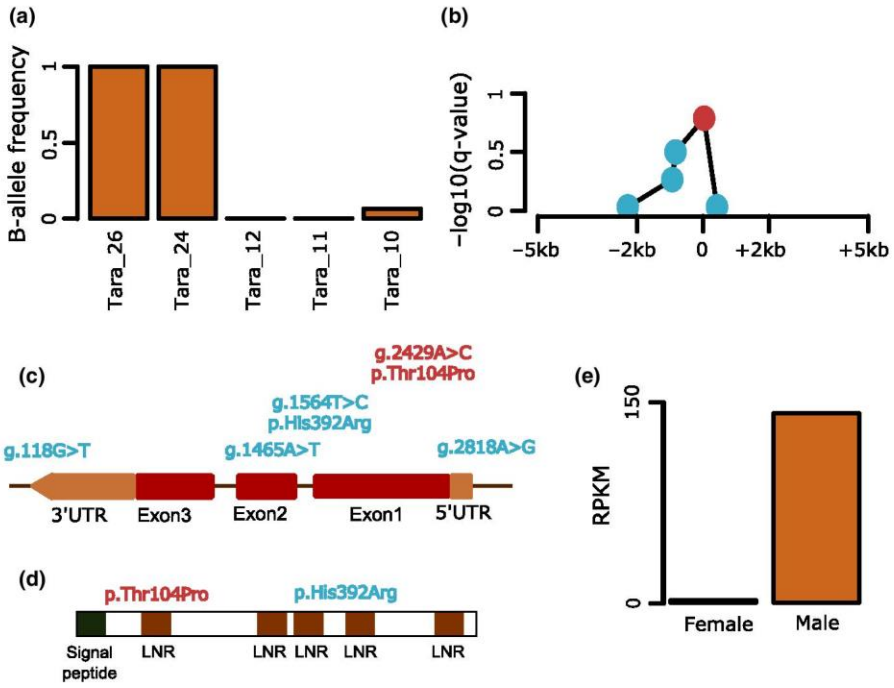


FIGURE 4 Positive selection of variant in the GSONAT00014698001 gene. (a) B-allele frequency value of the loci under selection in the northern stations. (b) Manhattan plot in a 2-kb window around the loci under selection. The dot in red is considered under selection (q-value < 0.2). (c) GSONAT00014698001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

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

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

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

BAF, B-allele frequency.

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

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

4 | DISCUSSION

4.1 | Towards the building of new copepod reference genomes

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

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

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

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

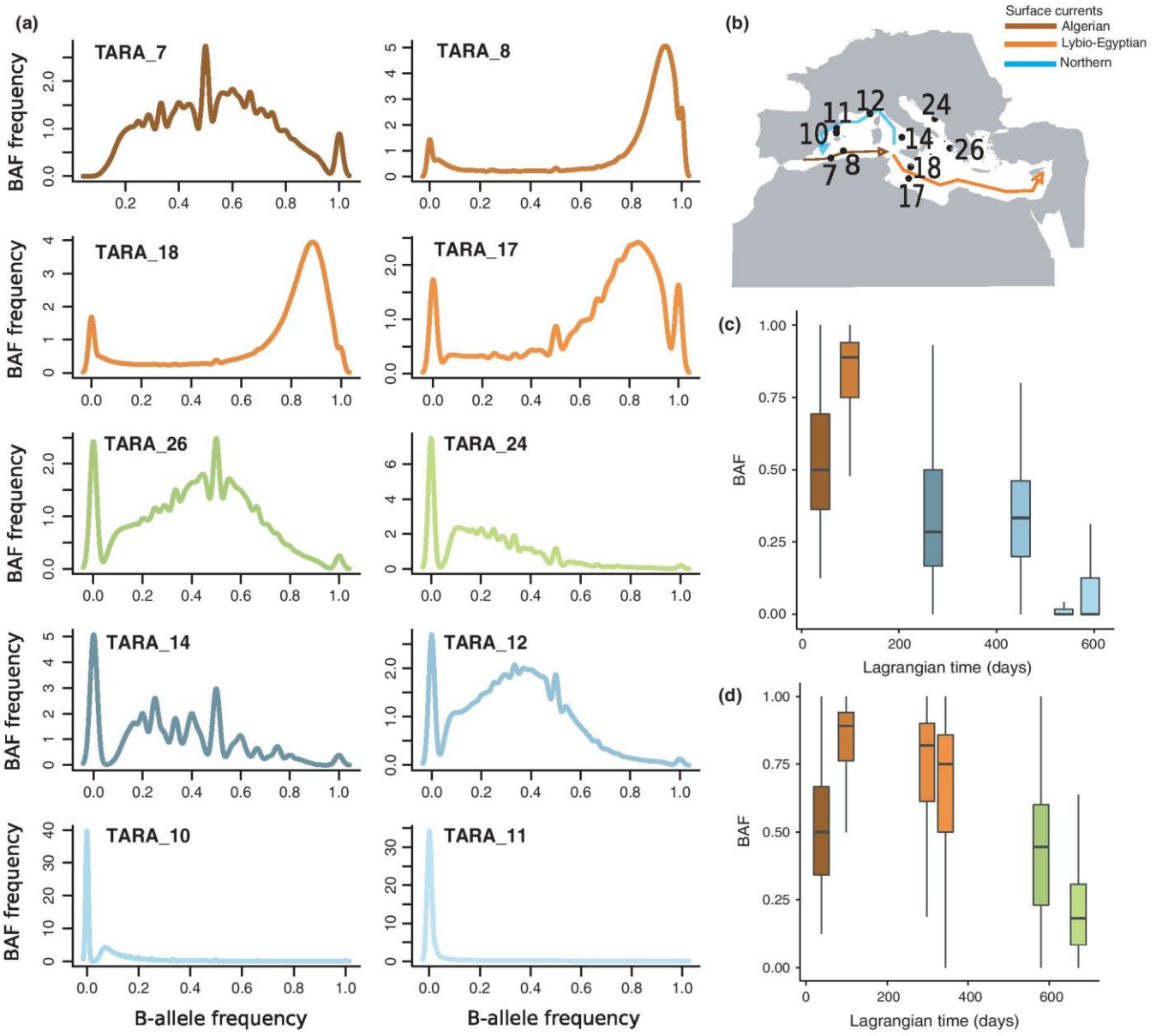


FIGURE 5 *Oithona* population genomics in the Mediterranean Sea. (a) B-allele frequency (BAF) distribution of *Oithona* variants in the southern stations (TARA_7, 8, 17 and 18), northwestern stations (TARA_10, 11, 12 and 14) and northeastern stations (TARA_24 and 26). (b) Dominant Mediterranean surface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

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

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

Until recently, we lacked a global genetic data set to study the distribution of *Oithona* species across the world. Taking advantage of the *Tara* Oceans metagenomic effort, we extracted reads from the 28S DNA marker from the fraction size compatible with the *Oithona* presence. The analysis of Oithonidae 28S sequences showed that 28S resolution varied depending on the species and that OTU1 contained at least two species. The mitochondrial lineages found recently for *O. similis* (Cornils et al., 2017) were also not visible in 28S. Unfortunately, this does not allow a comprehensive

biogeography of *Oithona* species by this method but provides still some new aspects of the *Oithona* biogeography. The predominance of *O. similis* in temperate and polar waters illustrates its ecological relevance as a small-sized zooplankton in these oceanic areas. Its absence in tropical regions also confirms previous distribution assumptions (Nishida, 1985). Thus, *O. similis* is a strong candidate for future genomic efforts. *Oithona nana* was identified only in a few *Tara* Oceans samples and mostly in the MS. Despite its worldwide distribution (Razouls et al., 2016), *O. nana* is limited to coastal waters and, unfortunately, the *Tara* Oceans project did not focus on such environments. Several samples were identified as containing undefined *Oithona* spp.; the lack of resolution was partially due to the general lack of references for most of the 44 *Oithona* species. More efforts have to be undertaken for generating new references for under-represented species in the public databases.

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

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

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

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

5 | CONCLUSIONS

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

ACKNOWLEDGEMENTS

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

DATA ACCESSIBILITY

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

AUTHOR CONTRIBUTIONS

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

L.B.B., A.C. and P.W. contributed to the manuscript. M.-A.M. and P.W. supervised the study.

REFERENCES

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

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

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

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

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

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

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

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

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

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

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

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

Cingolani, P., Platts, A., Wang le, L., Coon, M., Nguyen, T., Wang, L., … Ruden, D. M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly (Austin)*, *6*, 80–92.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

SUPPORTING INFORMATION

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

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

DiscoSNP ++ is a bioinformatics tool that was designed to detect SNP and indels directly from raw Illumina reads and without reference. By using this tool on *Tara* metagenomic data, this study provides a database of several million variants detected in the Atlantic Ocean and the Mediterranean Sea. Now, it is sufficient to align only the identified variants on reference genomes, without realigning all the *Tara* data. To test the method, the alignment was performed on the genome of *Oithona nana*.

Twenty-six functionally-annotated genes were detected under selection, whose five were LDPGs: three were LNR-only genes (whose one was detected in the previous study), one was associated with a Kelch domain, and the last one was associated with a metalloprotease domain.

In this study, I performed functional analyses of loci detected under selection. I also had to write the material and method of my analysis and make a supplementary table, with the help of my co-authors.

Article source

Arif, Majda, Jérémy Gauthier, Kevin Sugier, Daniele Iudicone, Olivier Jaillon, Patrick Wincker, Pierre Peterlongo, and Mohammed-Amin Mohammed-Amin Mohammed-Amin Madoui. 2019. 'Discovering Millions of Plankton Genomic Markers from the Atlantic Ocean and the Mediterranean Sea'. Molecular Ecology Resources 19 (2): 0–3. https://doi.org/10.1111/1755-0998.12985.

A French abstract is available in appendix 5 (page 156)

Received: 20 March 2018 | Revised: 16 November 2018 | Accepted: 19 November 2018

DOI: 10.1111/1755-0998.12985

RESOURCE ARTICLE

WILEY MOLECULAR ECOLOGY RESOURCES

Discovering millions of plankton genomic markers from the Atlantic Ocean and the Mediterranean Sea

Majda Arif[1] | Jérémy Gauthier[2] | Kevin Sugier[1] | Daniele Iudicone[3] | Olivier Jaillon[1] | Patrick Wincker[1] | Pierre Peterlongo[2] | Mohammed-Amin Madoui[1]

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

[2]Univ Rennes, CNRS, Inria, IRISA–UMR 6074, Rennes, France

[3]Stazione Zoologica Anton Dohrn, Naples, Italy

Correspondence
Mohammed-Amin Madoui, Génomique Métabolique, Genoscope, Institut François Jacob, CEA, CNRS, Univ Evry, Université Paris-Saclay, Evry, France.
Email: amadoui@genoscope.cns.fr

Funding information
Agence Nationale de la Recherche, Grant/Award Number: ANR-10-INBS-09-08, ANR-11- BTBR-0008, ANR-14-CE23-0001

Abstract

Comparison of the molecular diversity in all plankton populations present in geographically distant water columns may allow for a holistic view of the connectivity, isolation and adaptation of organisms in the marine environment. In this context, a large-scale detection and analysis of genomic variants directly in metagenomic data appeared as a powerful strategy for the identification of genetic structures and genes under natural selection in plankton. Here, we used DISCOSNP++, a reference-free variant caller, to produce genetic variants from large-scale metagenomic data and assessed its accuracy on the copepod *Oithona nana* in terms of variant calling, allele frequency estimation and population genomic statistics by comparing it to the state-of-the-art method. DISCOSNP ++ produces variants leading to similar conclusions regarding the genetic structure and identification of loci under natural selection. DISCOSNP++ was then applied to 120 metagenomic samples from four size fractions, including prokaryotes, protists and zooplankton sampled from 39 TARA Oceans sampling stations located in the Atlantic Ocean and the Mediterranean Sea to produce a new set of marine genomic markers containing more than 19 million of variants. This new genomic resource can be used by the community to relocate these markers on their plankton genomes or transcriptomes of interest. This resource will be updated with new marine expeditions and the increase of metagenomic data (availability: http://bioinformatique.rennes.inria.fr/taravariants/).

1 | INTRODUCTION

The identification of population connectivity, isolation and adaptation is of great interest to understand the current and future ecological responses of plankton communities to environmental variations such as the rise of water temperature and acidity (Freer, Partridge, Tarling, Collins, & Genner, 2018; Pelejero, Calvo, & Hoegh-Guldberg, 2010), especially in a climate change context (Beaugrand, Brander, Alistair Lindley, Souissi, & Reid, 2003; Beaugrand, Reid, Ibanez, Lindley, & Edwards, 2002). To understand the impact of these changes on living organisms, the study of plankton populations at the molecular level is a valuable option since it allows us not only to characterize genetic structures but also to determine which genes and biological functions are under natural selection (Avise, 2004; Peijnenburg & Goetze, 2013). Previous studies performed on plankton were based mostly on a few molecular markers, such as ribosomal DNA or mitochondrial genes (Blanco-Bercial, Cornils, Copley, & Bucklin, 2014; Cepeda, Blanco-Bercial, Bucklin, Beron, & Vinas, 2012). An alternative capture-based approach based on RAD-seq has also been proposed (Blanco-Bercial & Bucklin, 2016). These approaches permitted the construction of population genetic structures using only a subset of the whole genomic variability. Furthermore, as the loci under selection represent only a very small fraction of a genome, the lack of resolution of these methods does not allow a comprehensive view of the natural selection occurring on plankton. To be

able to capture the entire genomic variability of these organisms, whole-genome sequencing of individuals could be the ideal strategy. However, due to the small size of certain major zooplankters and their large genome size (Wyngaard & Rasch, 2000; Wyngaard, Rasch, Manning, Gasser, & Domangue, 2005), the current DNA extraction methods applied on a single individual do not permit us to retrieve a sufficient amount of genomic DNA that captures the whole-genome complexity and that is needed to build genomic DNA libraries (without random genomic amplification) usable for high-throughput sequencing.

Recently, the use of metagenomic data has been proposed to identify natural selection in prokaryotes (Costea et al., 2017; Delmont et al., 2017; Schloissnig et al., 2013). A similar approach has also been applied to the widespread marine copepod *Oithona* (Madoui et al., 2017) to establish a population genomic analysis at the whole-genome level. The methods used in these studies were all based on metagenomic reads mapping to reference genomes, followed by several filtering steps based on the nucleic identity cut-off and depth of sequencing coverage prior to the variant calling step. This allowed the detection of polymorphic loci and the estimation of allele frequencies in each sample that were followed by a wide range of analyses to characterize the nucleic variations and to identify selection using population genetic metrics such as F_{ST} (Wright, 1951), LK (Lewontin & Krakauer, 1973) and FLK (Bonhomme et al., 2010). In these previous studies, the arbitrary nucleic identity cut-off was used to decrease the amount of false-positive variants that can be generated by the alignment of metagenomic reads provided by a closely related species that can be present in the sample. Although the use of such a filter is justified, reads harbouring more variation (<97% identity) but belonging to the studied organism are de facto discarded. Moreover, the time and computational resources needed for metagenomic read alignments increase with the number of reference genomes included in the analysis. Finally, methods based on read alignments suffer from bias due to the incompleteness and imperfectness of reference genome sequences unless reference genomes are exhaustively and correctly assembled, which is rarely the case.

To bypass these problems, the use of an alignment-free variant calling method could be a solution. Therefore, in the present study, we used DISCOSNP++ (Peterlongo, Riou, & Drezen, 2017; Uricaru et al., 2015), a reference-free variant caller, and compared its performance to the one obtained with bwa/samtools/bcftools (BSB; Li et al., 2009) first using simulated data and then the TARA Oceans metagenomic data (Karsenti et al., 2011; Pesant et al., 2015) on the *O. nana* reference genome as a case study to determine DISCOSNP++ accuracy for variant calling, allele frequency estimation and downstream population genomic analysis. Then, we applied DISCOSNP++ to TARA Oceans metagenomic data from the Atlantic Ocean (AO) and the Mediterranean Sea (MS) to provide a new genomic resource that contains more than 19 million marine genomic variants (MGVs) that can be used as is or directly mapped on plankton genomes and transcriptomes of interest for population genomic analysis using the provided DISCOSNP++ module.

2 | MATERIAL AND METHODS

2.1 | Metagenomic data and genome reference

To compare DISCOSNP++ to BSB, we used metagenomics reads from the MS collected by the TARA Oceans expedition (Alberti et al., 2017) that correspond to the 20–180 µm fraction size from the surface (≤20 m) water layers of Mediterranean stations TARA_8, 10, 11, 12, 24 and 26 (Supporting information Appendix S1; Pesant et al., 2015). Only the metagenomic data from the two stations TARA_8 and 11 were used to compare the performances of DISCOSNP++ versus BSB for variant calling and B-allele frequency accuracy. Data from the five stations TARA_10, 11, 12, 24 and 26 were used to compare the two approaches in order to perform population genomic analysis. The *O. nana* genome was downloaded from NCBI (Accession no.: GCA 900157175.1). To build the marine genomic variants sets (MGVs), we used TARA Oceans metagenomic reads generated from samples corresponding to four size fractions (0.8–5, 5–20, 20–180 and 180–2,000 µm) collected from stations located in the AO and MS (Supporting information Appendix S1).

2.2 | The BSB pipeline

The *bwa mem* (Li & Durbin, 2009) command was used to align the metagenomics reads on the *O. nana* genome with a 17 bp seed, and alignments were stored in one sorted BAM file per station. To avoid spurious read alignments, Dust was applied with default parameters to discard reads with low complexity. The reads with an identity under 97% with the *O. nana* genome were discarded. For the variant (in this study, we will systematically use the term "variant" to refer to a single nucleotide polymorphism) calling step, we used the *samtools mpileup* and *bcftools call -m* commands (Li et al., 2009) with default parameters. Loci with a maximum of two alleles were kept. Only positions with a vertical coverage between the median coverage ± two *SD* were kept with a minimum of 4× coverage (Supporting information Appendix S2).

2.3 | DISCOSNP++ method overview

DISCOSNP++ was originally designed for genomic data analysis; however, the core of the program also applies to cases of metagenomic data. The tool is based on the analysis of the *de Bruijn* Graph (dBG). In the genome assembly context (Pevzner, Tang, & Tesler, 2004), a dBG is a graph in which nodes are words of length k (k-mers), and each edge connects two k-mers that share a $k-1$ overlap. For assembling purposes, the dBG is constructed from k-mers of a read set, and contigs are obtained by finding paths in this graph. In practice, k-mers are counted and those having an unexpected low abundance are removed as they are considered to contain sequencing errors. The dBG is constructed with the remaining k-mers. Basically, in a dBG, a bubble denotes a path in the graph which diverges into two distinct paths before they reunite. Any couple of distinct sequences that exists in the data, starting and finishing with the same

k-nucleotides, generates a bubble in the dBG. In particular, small indels and SNPs generate such a topological pattern. The DISCOSNP++ algorithm detects bubbles whose couple of paths is of equal length (generated by substitutions in the data) and bubbles whose couple of paths has a difference of length $\leq D$ (generated by insertion or deletion of size at most *D*). The detection of bubbles in the dBG can be performed through different methods corresponding to different stringencies: parameter *–b* 0 or 1, with *–b* 0 providing high precision, and lower recall and conversely (Uricaru et al., 2015 for more details). In a second step, raw reads are mapped back on the sequence of these paths. This step provides a way to remove non-coherent sequences (Myers, 2005) and to supply read coverage per variant and per input read set, whatever the number of input read set(s). This allows the simultaneous analysis of large metagenomic data sets. When a reference genome is available, sequence variants can be mapped to it. Thus, mapped predicted variants have a genomic position, provided in a VCF file.

DISCOSNP++ was run using the default parameters but avoiding indels (*–D* 0) for the methods comparison, and additionally using *–k* 51 to build the MGVs. Using a large *k* value (here *k* = 51) decreases the method sensitivity, increases the precision and, by simplifying the *de Bruijn* graph, allows faster computing and lower memory use on very large data sets. Depending on the situation, DISCOSNP++ was run using *–b* 0 (default) or *–b* 1, and this parameter is specified and motivated in the text. For the BSB pipeline, biallelic loci were kept and only positions with a vertical coverage between the median coverage ± two *SD* were kept with a minimum of 4× coverage (Supporting information Appendix S2).

2.4 | Comparison of the variant calling methods on simulated data

We simulated a first population of 20 *O. nana* genomes having 99% identity with the *O. nana* reference genome and a second population of 20 genomes of "Oithona2" based on a new reference having 95% identity to the *O. nana* genome and with a 99% identity within the population. SNPs were simulated to reproduce their natural distribution along the genome (Supporting information Appendix S3). We generated 100× of Illumina reads on each population and created 20 read data sets by mixing the two populations in different proportions and each data set contained a total of 30× of simulated Illumina reads. We applied the two approaches to these simulated data sets using the *O. nana* genome as a reference for reads mapping and variant calling for BSB and variant relocating for DISCOSNP++. The variants found by the two methods (DISCOSNP++ was run in relaxed mode only) were compared to the simulated ones (methodology presented in Figure 1). Considering *O. nana* as the organism of interest, the signal-to-noise ratio was calculated, that is, the ratio of the number of *O. nana* variants over the number of Oithona2 variants.

2.5 | Comparison of the variant calling methods on real data

The two methods were compared on their performance to identify intra-species variants present in the *O. nana* genome. The TARA_8 sample was known to contain an abundant species closely related to *O. nana* with a median identity percentage of 95% and very few *O.*

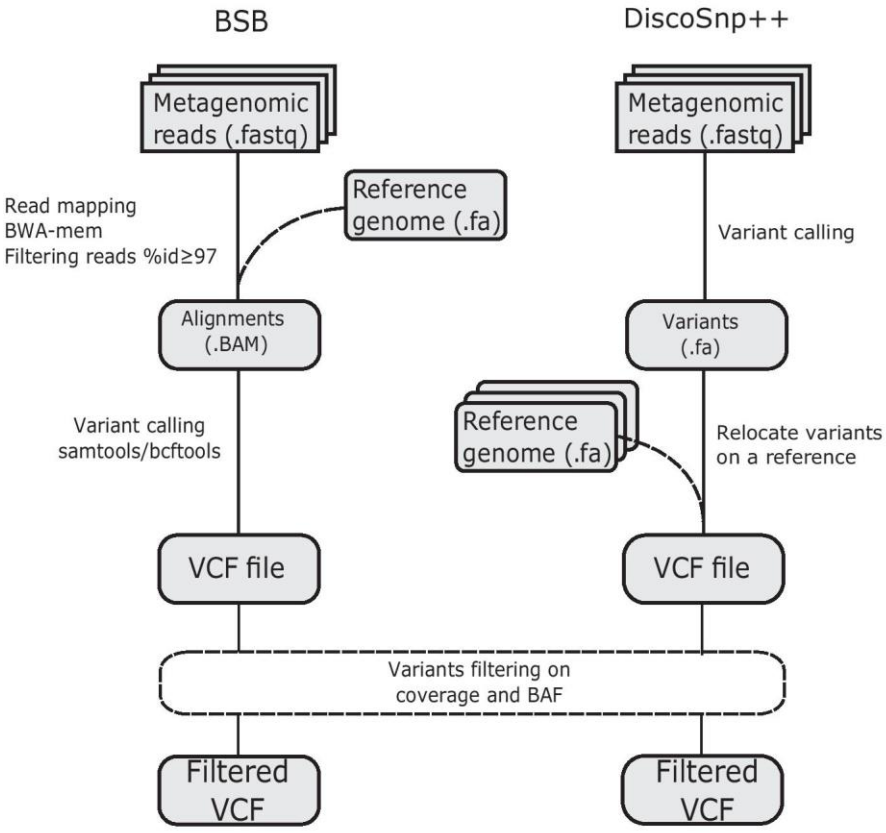


FIGURE 1 Workflow for BSB and DISCOSNP++ methods comparison

nana (<10% of total *Oithona* based on the 28S relative abundance; Madoui et al., 2017). The variants predicted from this sample by any method and remapped on the *O. nana* genome can be considered as enriched in inter-species variants. The TARA_11 sample was known to contain a large majority of *O. nana* (>60% of total *Oithona*) with other *Oithona* species that are not closely related to *O. nana* (here *Oithona similis* and *Oithona atlantica*). The variants predicted from this sample by any method and remapped on the *O. nana* genome can be considered enriched in intra-species variants. The stations TARA_8 and 11 were used to compare the two approaches (methodology presented Figure 1), in terms of variant calling, allele frequency accuracy and population genomics statistics (DISCOSNP++ was run in relaxed and stringent mode). To evaluate the possible biases on the coverage of biallelic loci that could be introduced by the variant calling methods, the read depth of the biallelic loci was fitted to a negative binomial distribution and the expected skewness of the distribution was calculated and compared to the observed one. The significance of the method's impact on the coverage skewness was tested by Wilcoxon signed-rank tests.

2.6 | Population genomics analysis

The B-allele frequencies (BAFs), also named alternative allele frequencies compared to a haploid reference genome, were calculated from the VCF files generated by the two methods (DISCOSNP++ in relaxed and stringent mode). Only loci with at least a BAF ≥0.05 in one population were selected. To identify populations having the same genomic variant pattern, a PCA was performed based on the BAF of the five populations. To measure the genetic differentiation between the populations, we used the F_{ST} (Wright's fixation index): $F_{ST} = V(p)/E(p)(1-E(p))$, with p being a set of BAFs observed in n populations at the same biallelic locus, E the mean and V the variance. For each locus, the F_{ST} was calculated between each population (pairwise F_{ST}). The median pairwise F_{ST} was then used to estimate the genetic differentiation between each population.

To evaluate the use of DISCOSNP++ to identify population differentiation, we calculated the median pairwise F_{ST} using four sets of BAFs; set 1: BAFs inferred from BSB variants found in common with DISCOSNP++; set 2: BAFs inferred from DISCOSNP++ variants found in common with BSB; set 3: BAFs inferred from all BSB variants; and set 4: BAFs inferred from all DISCOSNP++ variants. For the four sets, we used the DISCOSNP++ variants called using the –*b* 0.

2.7 | Detection of loci under selection

To detect loci under selection, we calculated the Lewontin–Krakauer (LK) statistic, which is an improvement of the F_{ST} that can be used for testing the neutrality of polymorphic genes, $LK = (n\text{-}1)F_{ST}/E(F_{ST})$. To be able to detect loci under selection, the LK distribution must follow a chi-square distribution χ^2 $(n-1)$ with n being the number of different populations. The fitting between the theoretical χ^2 distribution and the observed LK distribution obtained from BSB and DISCOSNP++ (with –*b* 0 and –*b* 1 options) was observed to validate the neutral model, that is, the majority of the biallelic loci are not under selection (Supporting information Appendix S4). The FLK statistics (Bonhomme et al., 2010) were also calculated; this metric is an extension of the LK test that uses a kinship matrix of the populations based on the BAF to correct genetic distance biases due to population structure. The FLK statistics were also tested for the neutral model. The first hundred loci having the highest LK or FLK values higher than expected (with a p-value ≤ 0.05) were considered to be under selection. The annotation of the variants and their possible effect on protein structure was performed with SNPEFF (Cingolani et al., 2012).

3 | RESULTS

3.1 | Variant calling

The BSB and DISCOSNP++ pipelines were compared for variant detection (methodology presented in Figure 1) using simulated data representing an admixture of *O. nana* and a closely related species in different proportions. Here, we considered *O. nana* as the organism of interest, and its variants were considered as true positives and the variants of Oithona2 as false positives. BSB found more true positives than DISCOSNP++ in all admixtures especially for low *O. nana* content (between 5% and 50% of *O. nana*; Figure 2a, Supporting information Appendix S5). BSB also identified more false positives than DISCOSNP++ especially for admixtures with *O. nana* lower than 90%. Based on these simulations, DISCOSNP++ was less sensitive in any admixture but more specific than BSB when dealing with an admixture of two closely related species. However, the signal-to-noise ratio was higher for BSB, especially for admixtures with more than 90% of *O. nana* (Figure 2b, Supporting information Appendix S5).

The BSB and DISCOSNP++ pipelines were also compared using TARA Oceans metagenomic data from the stations TARA_8 and 11 and the *O. nana* genome. Compared to DISCOSNP++ in stringent mode, BSB found approximately 14 times more intra-species variants (TARA_11) and eight times more variants enriched in inter-species variants (TARA_8; Figure 3). Compared to DISCOSNP++ in relaxed mode (–*b* 1), BSB found 3.5 times more intra-species variants, and 1.3 times more variants enriched in inter-species variants. On real metagenomic data, the results provided the same trend given by the comparison on the simulated data, showing that DISCOSNP++ is less sensitive but more specific for intra-species variant detection in a population admixture, even in relaxed mode. The effect of the variant calling methods on the skewness of the depth of coverage distribution was not significant but still present (p-value = 0.06, Wilcoxon signed-rank test; Supporting information Appendix S2b), and the skewness obtained using the –*b* 0 option of DISCOSNP++ was closer to the expected one (Supporting information Appendix S2c).

3.2 | Allele frequency accuracy

The BAFs obtained for variants found by the two calling methods were compared (Figure 4), and we observed a strong correlation

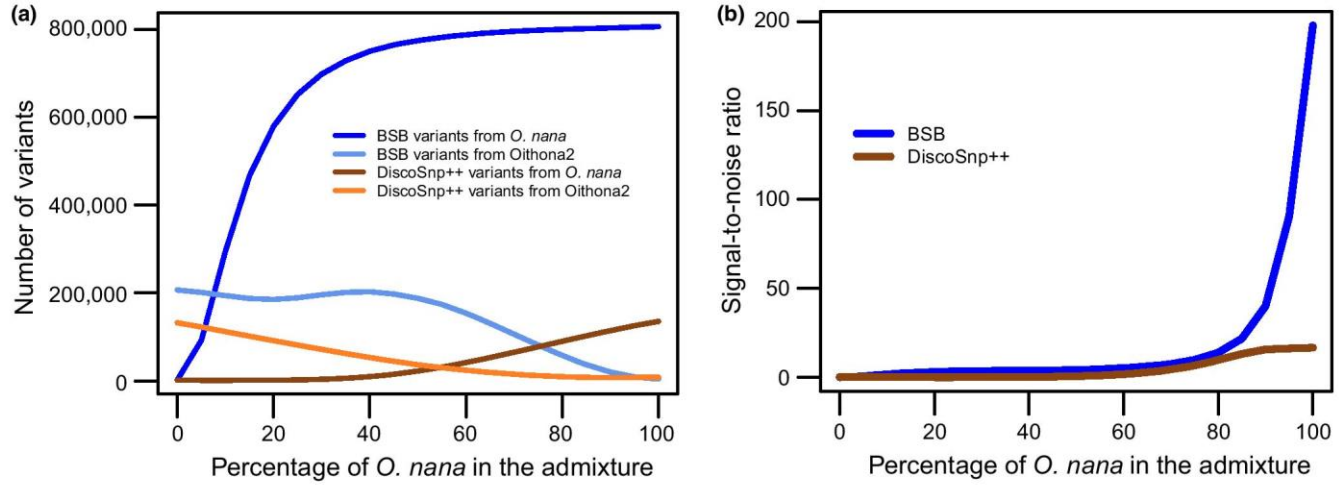


FIGURE 2 Comparison of variant calling between DISCOSNP++ and BSB on simulated data. (a) Variants recall for increasing proportion of *Oithona nana* in the admixture. (b) Methods efficiency for increasing proportion of *O. nana* in the admixture [Colour figure can be viewed at wileyonlinelibrary.com]

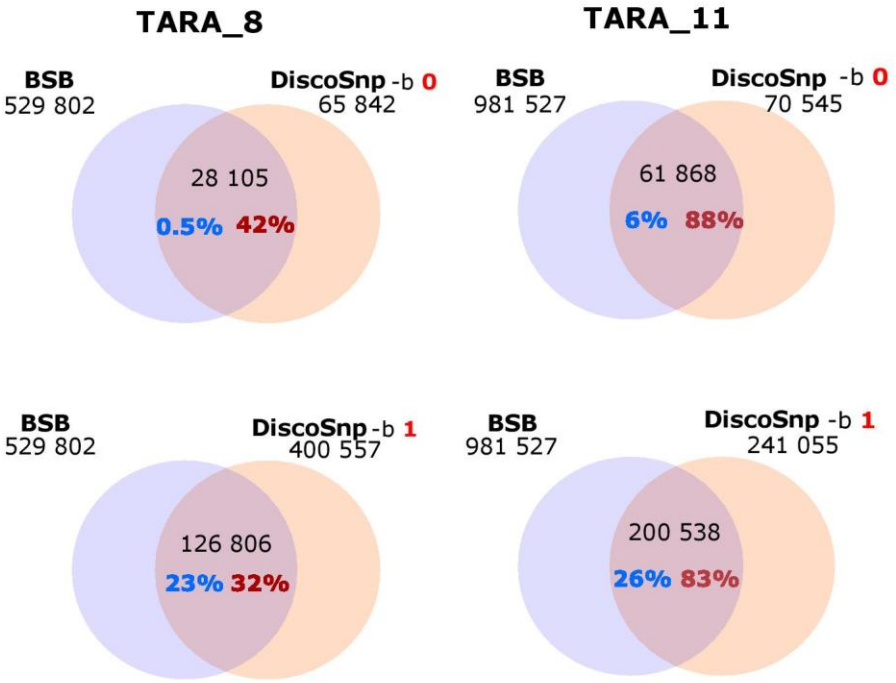


FIGURE 3 Comparison of variant calling between DISCOSNP++ and BSB on TARA Oceans metagenomic data. Variants found by each method are written under the method name. Variants found in common with DISCOSNP++ and BSB in populations from TARA_8 and 11. DISCOSNP++ parameters –b 0 (stringent mode) and –b 1 (relaxed relaxed) were tested. The percentages correspond to the fraction of variants found by the two methods (in blue for BSB and red for DISCOSNP++) [Colour figure can be viewed at wileyonlinelibrary.com]

between the two methods in *O. nana* populations from TARA_8 and 11 ($R^2 \geq 0.95$). However, we found that 7.5% of the variants had a higher BAF difference than expected between the two methods (i.e., with a BAF difference higher/lower than the median difference plus/minus two *SD*, Supporting information Appendix S6a). For variants having a higher BAF with DISCOSNP++ (6.3% of the total variants found in common), we explained the difference by the identity cut-off of 97% used in the BSB pipeline (Supporting information Appendix S7). The variants presenting a strong BAF deviation between the two methods were annotated based on their genomic location (i.e., intronic, exonic, UTR and intergenic) and compared to (a) the genomic location of the variants presenting no significant BAF differences, and (b) a random distribution of the variants on the genome. Significant differences (p-value < 0.001, chi-square test) were found, with an increase of biallelic loci having higher BAFs with DISCOSNP++ located in the non-coding regions of the genome (Supporting information Appendix S6b,c). This result suggests that DISCOSNP++ can recruit more reads than BSB in non-coding regions of the genome. These regions are indeed expected to contain more polymorphisms than coding regions within populations. Therefore, filters that are applied in BSB tend to discard reads that should be aligned at a reduced similarity threshold. Consequently, DISCOSNP++ seems to provide a better estimation of the allele frequency in more variable regions of the genome compared to BSB applied with a 97% identity cut-off.

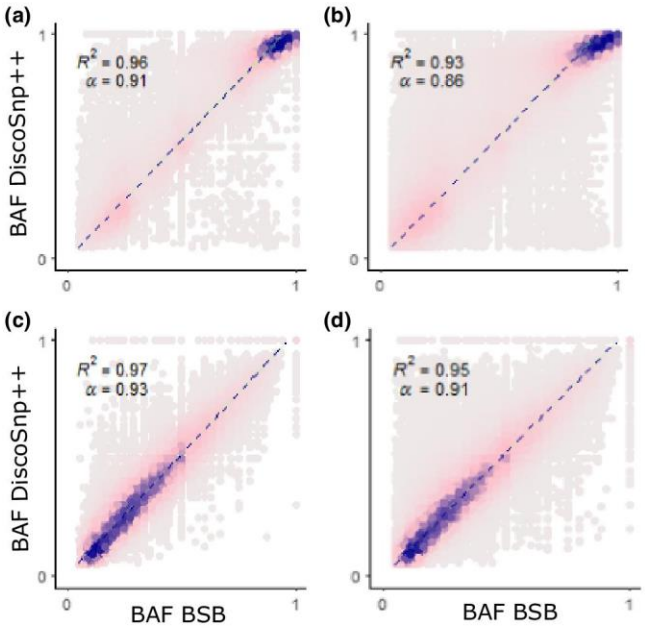


FIGURE 4 B-Allele frequency correlation between DISCOSNP++ and BSB. The *x*-axis of the scatter-plots corresponds to BAFs obtained with BSB, and the *y*-axis corresponds to BAFs obtained with DISCOSNP++. (a) *y*-axis is BAFs computed from TARA_8 with DISCOSNP++ –b 0. (b) *y*-axis is BAFs computed from TARA_8 with DISCOSNP++ option –b 1. (c) *y*-axis is BAFs computed from TARA_11 with DISCOSNP++ –b 0. (d) *y*-axis is BAFs computed from TARA_11 with DISCOSNP++ –b 1 [Colour figure can be viewed at wileyonlinelibrary.com]

TABLE 1 Median pairwise F_{ST} between *Oithona nana* populations obtained from the four BAFs sets. Set 1: BAFs inferred from BSB for all variants found in common with DISCOSNP++; set 2: BAFs inferred from DISCOSNP++ for variants found in common with BSB; set 3: all BAFs inferred from BSB; set 4: all BAFs inferred from DISCOSNP++

	Median pairwise F_{ST}				
Populations	Set 1	Set 2	Set 3	Set 4	*SD*
TARA_10 versus TARA_11	0.074	0.074	0.065	0.075	0.0046
TARA_10 versus TARA_12	0.077	0.084	0.065	0.086	0.0096
TARA_10 versus TARA_24	0.096	0.096	0.077	0.099	0.01
TARA_10 versus TARA_26	0.109	0.125	0.096	0.133	0.016
TARA_11 versus TARA_12	0.077	0.099	0.071	0.1	0.0149
TARA_11 versus TARA_24	0.100	0.124	0.089	0.128	0.0189
TARA_11 versus TARA_26	0.114	0.142	0.099	0.143	0.0216
TARA_12 versus TARA_24	0.096	0.1	0.077	0.105	0.0121
TARA_12 versus TARA_26	0.096	0.1	0.077	0.111	0.0141
TARA_24 versus TARA_26	0.105	0.116	0.095	0.125	0.013

3.3 | Population genomic analysis

Five *O. nana* populations from sampling stations (TARA_10, 11, 12, 24 and 26) were clustered by PCA based on their BAFs (Figure 5a,b). For the two methods, the clustering showed a similar grouping of the populations by geographic location, separating the ones from the Western MS (WMS) from the ones of the Eastern MS (EMS). We estimated the genetic differentiation between the *O. nana* populations by calculating the pairwise F_{ST} using four different sets of variants and related BAFs (Supporting information Appendix S6) and compared the median F_{ST} values to evaluate any biases that could be introduced by DISCOSNP++ ran in stringent mode (Table 1). Using only variants detected by the two methods (i.e., using BAF from set 1 and set 2), we found no F_{ST} difference over 0.024 and the average difference between the median pairwise F_{ST} was 0.012 ± 0.01 (Figure 5c). A higher difference was observed for high F_{ST} values. We found a negligible difference between pairwise F_{ST} computed by DISCOSNP++ in relaxed mode versus stringent mode (mean = 0.003, *SD* = 0.008). For the selected variants, the two methods allowed the identification of the same genetic pattern between the five *O. nana* populations. The genetic distance observed using all DISCOSNP++ or all BSB variants also produced a similar genetic pattern with an absence of genetic structure within the WMS and a weak differentiation between the two MS basins and within the

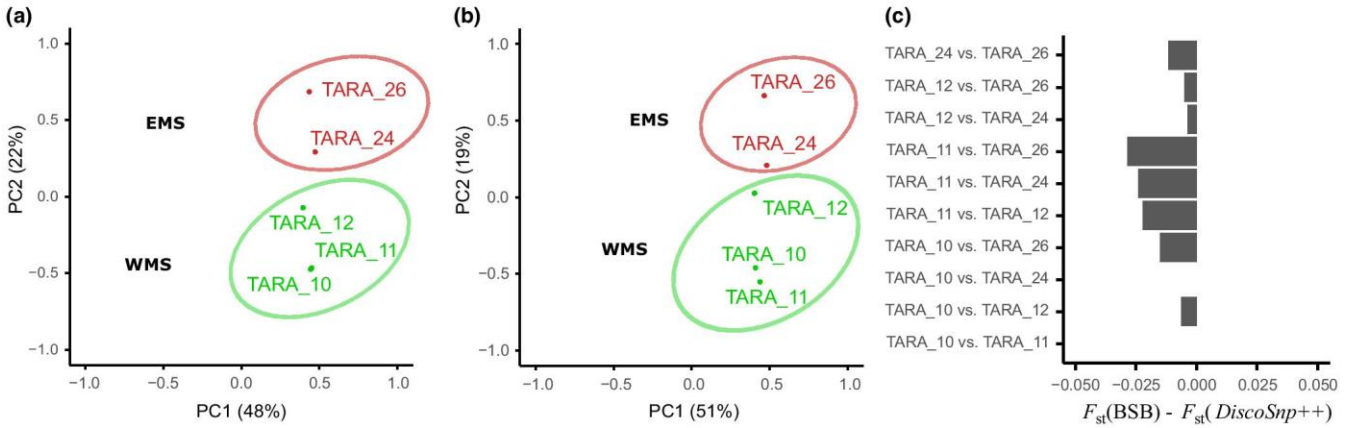


FIGURE 5 *Oithona nana* genetic structure in the Mediterranean Sea obtained with DISCOSNP++ and BSB. (a) PCA on five *O. nana* populations on the MS based on BAFs obtained with BSB. (b) PCA on five *O. nana* populations on the MS based on BAFs obtained with DISCOSNP++. (c) Differences of the median pairwise F_{ST} between BSB and DISCOSNP++ [Colour figure can be viewed at wileyonlinelibrary.com]

EMS (Table 1). Compared to the previously published results (Madoui et al., 2017), there was a lower genetic distance between the population of TARA_26 and the four other populations. This difference can be due to the more stringent filtering on reads coverage used in the current study (see Materials and Methods) to consider valid variants compared to the previous study where biallelic loci with a read coverage up to 80× were kept. The current coverage filters may have discarded reads provided by repeated regions or a closely related species possibly present in the TARA_26 sample.

3.4 | Detection of loci under natural selection

To identify loci under natural selection, the LK and FLK statistics were computed from the BAFs of sets 1 and 2. For each variant set and statistics, the hundred loci with the highest LK and FLK were compared to estimate the congruence between the two variant calling methods (Figure 6). We found more loci in common with LK than FLK and by using the *–b* 0 option of DISCOSNP++ suggesting a more accurate detection of loci under selection as being more stringent in the variant calling of DISCOSNP++.

The functional annotation of the 79 variants under natural selection (detected by DISCOSNP++ *–b* 0 and using the LK outliers) that were found in common with BSB (Supporting information Appendix S7) showed 16 non-synonymous variants and 14 synonymous variants. Compared to the previous study (Madoui et al., 2017), we found four new Lin12 Notch Repeat (LNR) domain-coding genes. These domain-coding genes are of particular interest in *O. nana* where they were found to be over-abundant compared to other metazoans and one of them detected under positive selection was male-specific based on expression data (Madoui et al., 2017).

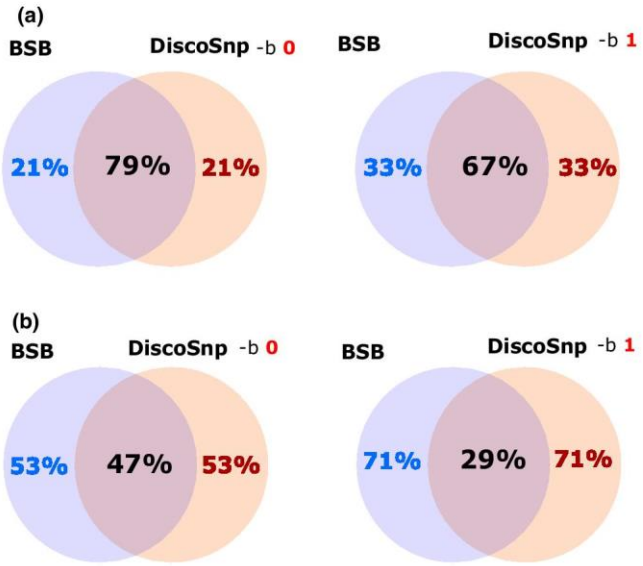


FIGURE 6 Loci under natural selection found in common between DISCOSNP++ and BSB. (a) Loci in common using the 100 LK highest values (p-value < 0.001, chi-square test). (b) Loci in common using the 100 highest FLK values (p-value < 0.001, chi-square test) [Colour figure can be viewed at wileyonlinelibrary.com]

Among the four new LNR domain-coding genes found to be under selection, one (GSONAT00015400001) codes a metallopeptidase domain protein, another (GSONAT00015380001) codes an LNR protein associated with a Kelch domain, and two others (GSONAT00013822001, GSONAT00015410001) code only LNR domain proteins without association to other known domains. These new results reinforce the highly evolutionary potential of LNR domain-containing proteins and their importance in the *O. nana* biology.

3.5 | Plankton genomic variant resources from the TARA Oceans metagenomic data

We produced the new set of MGVs by running DISCOSNP++ in relaxed mode (to optimize the number of MGVs) on more than 40 billion metagenomic 100 bp reads from 39 TARA stations located in the AO and the MS (Figure 7 and Supporting information Appendix S8). These MGVs correspond to genomic variants (SNVs and indels) found from natural populations of prokaryotic, protist and animal plankton that were sampled during the 3-year expedition of TARA. For the four different size fractions, we generated more than nineteen million MGVs (Table 2). The amount of input data was relatively similar among all size fractions (~11–12 × 10^9 of 100 bp reads) but the computation time globally increased with the size fraction and all had the same very low memory usage (~100 Gb). The amount of MGVs found in the different fraction sizes was at the same scale (5.2–6.2 × 10^6 variants) except for fraction 5–20 μm that presented half the MGVs and had the lowest computation time. This may be because of less genomic complexity in this fraction size, as shown previously (Carradec et al., 2018). The MGVs can be downloaded and directly used by the scientific community in order to perform new analyses of genomic diversity on any organism of interest as demonstrated on *O. nana*.

4 | DISCUSSION

Like any reference-based variant detection method, DISCOSNP++ limitations are mainly due to genomic approximate repeats. Reads from approximate repeats and, in the metagenomic framework, reads from similar inter-species genomic regions contain the same signal as those from regions containing intra-species variants. As shown by the results from this study, those imperfect predictions are not an insurmountable limitation for population genomics analysis where alignment-based and reference-free-based approaches provide similar conclusions in terms of population differentiation and overlapping results in terms of natural selection. Moreover, DISCOSNP++ is an order of magnitude faster and uses fewer resources (Peterlongo et al., 2017; Uricaru et al., 2015). In the case of the admixture of two closely related species, neither the alignment-based nor the reference-free-based approach allows the removal of inter-species variants which reduce the number of populations that can be integrated into the population genomic analysis focused on a single species.

The MGVs detected de novo with DISCOSNP++ from TARA Oceans data can be downloaded from http://bioinformatique.rennes.inria.f

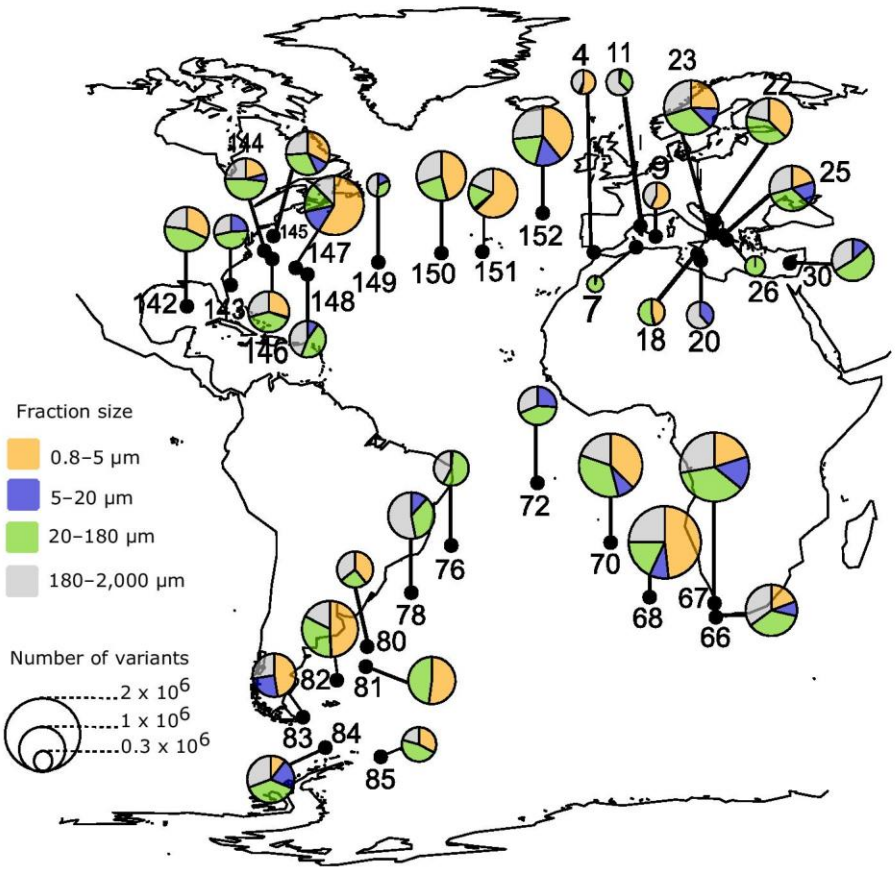


FIGURE 7 Geographic and size fraction distribution of MGVs [Colour figure can be viewed at wileyonlinelibrary.com]

TABLE 2 Marine genomic variants produced by DISCOSNP++ on TARA Oceans metagenomic data from the Atlantic Ocean and the Mediterranean Sea

Fraction size (µm)	Number of stations	Number of reads used	Number of variants	Computation time (hours)	Max memory used (Gb)
0.8–5	25	11.3×10^9	5.5×10^6	64	107
5–20	27	11.8×10^9	2.3×10^6	60	107
20–180	31	11.2×10^9	5.2×10^6	105	110
180–2,000	31	11.2×10^9	6.2×10^6	124	120

r/taravariants/and used directly on any genome or transcriptome provided by the users to create VCF files without computation of the variant calling. This can be done by running only the final DISCOSNP++ step "*run_VCF_creator.sh*" that can be done on a laptop computer. This allows the community to avoid (a) the systematic downloading of the whole TARA Oceans metagenomic data set that needs investment in large infrastructure for data storage and backup and (b) the alignment of the reads to their genomes and transcriptomes of interest that needs investment in computational power. As demonstrated in this study, the MGVs allow an accurate analysis of the molecular diversity of the plankton present in the AO and MS that were captured during the TARA Oceans expedition. In addition to the lack of reference sequences for plankton, depending on the genome size and abundance of the studied plankton in the TARA Oceans samples, the use of the MGVs collection may have some limits. Analyses focusing on small-size genomes (<100 Mb) and abundant protists such as green algae are more likely to provide interesting results compared to those focusing on copepods with large-size genomes (>1 Gb).

The increasing number of large collections of marine plankton samples and their related metagenomic data set forces a rethinking of the way population genomics can be performed. This can push the community towards the use of a universal genomic resource of variants that can be updated with the accumulation of newly released metagenomic data. From this perspective, the use of DISCOSNP++ offers a great advantage by providing a uniform method to generate community shared markers that store all the information needed to perform robust downstream population genetic analyses of plankton.

ACKNOWLEDGEMENTS

We wish to thank the individuals 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) and HYDROGEN (ANR-14-CE23-0001). This is TARA Oceans contribution number 83.

AUTHORS' CONTRIBUTION

M.A., K.S., P.P., J.G. and M.A.M. performed the analyses. O.J., D.L., P.P. and M.A.M. designed the study. M.A.M. wrote the manuscript, and all authors accepted its final version.

DATA AVAILABILITY

The metagenomic data from TARA Oceans are available at ENA (Supporting information Appendix S1). The *O. nana* genome sequence and annotation are available at ENA with the study Accession no. PRJEB18938. The MGVs files and their corresponding tutorial are available at http://bioinformatique.rennes.inria.fr/taravariants/.

ORCID

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

REFERENCES

Alberti, A., Poulain, J., Engelen, S., Labadie, K., Romac, S., Ferrera, I., … Wincker, P. (2017). Viral to metazoan marine plankton nucleotide sequences from the Tara Oceans expedition. *Scientific Data*, *4*, 170093.

Avise, J. C. (2004). *Molecular markers, natural history and evolution*, 2nd ed. Sunderland, MA: Sinauer.

Beaugrand, G., Brander, K. M., Alistair Lindley, J., Souissi, S., & Reid, P. C. (2003). Plankton effect on cod recruitment in the North Sea. *Nature*, *426*, 661–664. https://doi.org/10.1038/nature02164

Beaugrand, G., Reid, P. C., Ibanez, F., Lindley, J. A., & Edwards, M. (2002). Reorganization of North Atlantic marine copepod biodiversity and climate. *Science*, *296*, 1692–1694. https://doi.org/10.1126/science.1071329

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

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

Bonhomme, M., Chevalet, C., Servin, B., Boitard, S., Abdallah, J., Blott, S., & SanCristobal, M. (2010). Detecting selection in population trees: The Lewontin and Krakauer test extended. *Genetics*, *186*, 241–262. https://doi.org/10.1534/genetics.110.117275

Carradec, Q., Pelletier, E., Da Silva, C., Alberti, A., Seeleuthner, Y., Blanc-Mathieu, R., … Wincker, P. (2018). A global ocean atlas of eukaryotic genes. *Nature Communications*, *9*, 373. https://doi.org/10.1038/s41467-017-02342-1

Cepeda, G. D., Blanco-Bercial, L., Bucklin, A., Beron, C. M., & Vinas, M. D. (2012). Molecular systematic of three species of Oithona (Copepoda, Cyclopoida) from the Atlantic Ocean: Comparative analysis using 28S rDNA. *PLoS ONE*, *7*, e35861. https://doi.org/10.1371/journal.pone.0035861

Cingolani, P., Platts, A., Wang, L. L., Coon, M., Nguyen, T., Wang, L., … Ruden, D. M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly (Austin)*, *6*, 80–92. https://doi.org/10.4161/fly.19695

Costea, P. I., Munch, R., Coelho, L. P., Paoli, L., Sunagawa, S., & Bork, P. (2017). metaSNV: A tool for metagenomic strain level analysis. *PLoS ONE*, *12*, e0182392. https://doi.org/10.1371/journal.pone.0182392

Delmont, O., Kiefl, E., Kilinc, O., Esen, O. C., Uysal, I., Rappe, M. S., … Eren, A. M. (2017). *The global biogeography of amino acid variants within a single SAR11 population is governed by natural selection.*

Freer, J. J., Partridge, J. C., Tarling, G. A., Collins, M. A., & Genner, M. J. (2018). Predicting ecological responses in a changing ocean: The effects of future climate uncertainty. *Marine Biology*, *165*, 7. https://doi.org/10.1007/s00227-017-3239-1

Karsenti, E., Acinas, S. G., Bork, P., Bowler, C., De Vargas, C., Raes, J., … Wincker, P. (2011). A holistic approach to marine eco-systems biology. *PLoS Biology*, *9*, e1001177. https://doi.org/10.1371/journal.pbio.1001177

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

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, *25*, 1754–1760. https://doi.org/10.1093/bioinformatics/btp324

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., … Durbin, R. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics*, *25*, 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Madoui, M.-A., Poulain, J., Sugier, K., Wessner, M., Noel, B., Berline, L., … Wincker, P. (2017). New insights into global biogeography, population structure and natural selection from the genome of the epipelagic copepod Oithona. *Molecular Ecology*, *26*, 4467–4482.

Myers, E. W. (2005). The fragment assembly string graph. *Bioinformatics*, *21*(Suppl 2), ii79–85. https://doi.org/10.1093/bioinformatics/bti1114

Peijnenburg, K. T., & Goetze, E. (2013). High evolutionary potential of marine zooplankton. *Ecology and Evolution*, *3*, 2765–2781. https://doi.org/10.1002/ece3.644

Pelejero, C., Calvo, E., & Hoegh-Guldberg, O. (2010). Paleo-perspectives on ocean acidification. *Trends in Ecology & Evolution*, *25*, 332–344. https://doi.org/10.1016/j.tree.2010.02.002

Pesant, S., Not, F., Picheral, M., Kandels-Lewis, S., Le Bescot, N., Gorsky, G., … Wincker, P. (2015). Open science resources for the discovery and analysis of Tara Oceans data. *Scientific Data*, *2*, 150023. https://doi.org/10.1038/sdata.2015.23

Peterlongo, P., Riou, C., Drezen, E., & Lemaitre, C. (2017). *DiscoSnp++: De novo detection of small variants from raw unassembled read set(s). bioRxiv*. 209965.

Pevzner, P. A., Tang, H., & Tesler, G. (2004). De novo repeat classification and fragment assembly. *Genome Research*, *14*, 1786–1796. https://doi.org/10.1101/gr.2395204

Schloissnig, S., Arumugam, M., Sunagawa, S., Mitreva, M., Tap, J., Zhu, A., … Bork, P. (2013). Genomic variation landscape of the human gut microbiome. *Nature*, *493*, 45–50. https://doi.org/10.1038/nature11711

Uricaru, R., Rizk, G., Lacroix, V., Quillery, E., Plantard, O., Chikhi, R., … Peterlongo, P. (2015). Reference-free detection of isolated SNPs. *Nucleic Acids Research*, *43*, e11. https://doi.org/10.1093/nar/gku1187

Wright, S. (1951). The genetical structure of populations. *Annals of Eugenics*, *15*, 323–354. https://doi.org/10.1111/j.1469-1809.1949.tb02451.x

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

Wyngaard, G. A., Rasch, E. M., Manning, N. M., Gasser, K., & Domangue, K. (2005). The relationship between genome size, development rate, and body size in copepods. *Hydrobiologia*, *532*, 123–137. https://doi.org/10.1007/s10750-004-9521-5

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Arif M, Gauthier J, Sugier K, et al. Discovering millions of plankton genomic markers from the Atlantic Ocean and the Mediterranean Sea. *Mol Ecol Resour*. 2019;19:526–535. https://doi.org/10.1111/1755-0998.12985

Chapter 3:

LDPGs are involved in auto-proteolysis and neurogenesis modulation in *O. nana* males

This study was done to try to answer to two problematics: (i) can we explain the strongly biased sex-ratio (9:1) toward females in the Little Bay of Toulon at the molecular level? And (ii) can we identify developmental stage-specific genes and can we infer a function to these genes? To address these questions, we performed a differential expression (DE) and protein-protein interaction (PPI) analyses.

The DE analysis showed a large number of male-specific up-regulated genes (445), with enrichment in LDPGs, proteolysis genes and genes participating in neurogenesis and nervous system functioning. The PPI analysis detected the formation of LDP complexes and interactions with proteases, ECM and neurogenesis-related proteins. All these results seem to show, at the transcriptomic scale, a sacrificial behaviour in *O. nana* males.

In this study, my role was to do the analysis (sex-ratio, DE, functional annotation, phylogenetic, PPI), to write the article and make the figures, with the help of my co-authors. I was also widely involved in the wet lab works: I collected samples in the Toulon Little Bay, I performed the mRNA extraction and quality analyses of the *O. nana* egg sacs, I sorted more than 1,000 *O. nana* male individuals for the PPI, and I applied the Yeast two-hybrid protocol, developed by Alberti A. and her team, on two LDPG candidates.

Article source

Kevin Sugier, Romuald Laso-Jadart, Soheib Kerbache, Jos Kafer, Majda Arif, Laurie Bertrand, Karine Labadie, Nathalie Martins, Celine Orvain, Emmanuelle Petit, Julie Poulain, Patrick Wincker, Jean-Louis Jamet, Adriana Alberti, and Mohammed-Amin Madoui. 2019. 'Proteolysis and neurogenesis modulated by LNR domain proteins explosion support male differentiation in the crustacean *Oithona nana*. BioRxiv, Octobre, 818179

A French abstract is available in appendix 7 (page 159)

Proteolysis and neurogenesis modulated by LNR domain proteins explosion support male differentiation in the crustacean *Oithona nana*

Kevin Sugier[1], Romuald Laso-Jadart[1], Soheib Kerbache[1], Jos Kafer[4], Majda Arif[1], Laurie Bertrand[2], Karine Labadie[2], Nathalie Martins[2], Celine Orvain[2], Emmanuelle Petit[2], Julie Poulain[1], Patrick Wincker[1], Jean-Louis Jamet[3], Adriana Alberti[2] and Mohammed-Amin Madoui[1]

1. Génomique Métabolique, Genoscope, Institut François Jacob, CEA, CNRS, Univ Evry, Université Paris-Saclay, Evry, France
2. Commissariat à l'Energie Atomique (CEA), Institut François Jacob, Genoscope, Evry, France
3. Université de Toulon, Aix-Marseille Université, CNRS/INSU/IRD, Mediterranean Institute of Oceanography MIO UMR 7294, CS 60584, 83041 Toulon cedex 9, France
4. Laboratoire de Biométrie et Biologie Evolutive, Université Lyon 1, CNRS UMR 5558, France

Abstract

Copepods are the most numerous animals and play an essential role in the marine trophic web and biogeochemical cycles. The genus *Oithona* is described as having the highest numerical density, as the most cosmopolite copepod and iteroparous. The *Oithona* male paradox obliges it to alternate feeding (immobile) and mating (mobile) phases. As the molecular basis of this trade-off is unknown, we investigated this sexual dimorphism at the molecular level by integrating genomic, transcriptomic and protein-protein interaction analyses.

While a ZW sex-determination system was predicted in *O. nana*, a fifteen-year time-series in the Toulon Little Bay showed a biased sex ratio toward females (male / female ratio < 0.15± 0.11) highlighting a higher mortality in male. Here, the transcriptomic analysis of the five different developmental stages showed enrichment of Lin12-Notch Repeat (LNR) domains-containing proteins coding genes (LDPGs) in male transcripts. The male also showed enrichment in differentially expressed transcripts involved in proteolysis, nervous system development, synapse assembly and functioning and also amino acid conversion to glutamate. Moreover, several male down-regulated genes were involved in the increase of food uptake and digestion. The formation of LDP complexes was detected by yeast two-hybrid, with interactions involving proteases, extracellular matrix proteins and neurogenesis related proteins.

Together, these results suggest that the *O. nana* male hypermotility is sustained by LDP-modulated proteolysis allowing the releases and conversions of amino acid into the excitatory neurotransmitter glutamate. This process could permit new axons and dendrites formation suggesting a sexual nervous system dimorphism. This could support the hypothesis of a sacrificial behaviour in males at the metabolic level.

Introduction

Copepods are small planktonic crustacean forming the most abundant metazoan subclass on Earth and occupy all ecological aquatic niches (Huys and Boxshall 1991; Kiørboe 2011). Among them, the genus *Oithona* is described as having the highest numerical density (Gallienne and Robins 2001), cosmopolite (Nishida 1985) and playing a key role of secondary producer in the marine food web and biogeochemical cycles (Steinberg and Landry 2017). Because of its importance, *Oithona* phylogeography, ecology, behaviour, life cycle, anatomy and genomics are studied (Cornils, Wend-Heckmann, and Held 2017; Dvoretsky and Dvoretsky 2009; Kiørboe 2007; Madoui et al. 2017; Mironova and Pasternak 2017; Paffenhöfer 1993; Sugier et al. 2018; Zamora-Terol et al. 2014; Zamora Terol 2013).

Oithona is an ambush feeder: to feed, the individual remains static, jumps on preys that come on its range, and captures them with its buccal appendages (Kiørboe 2007). While females are mostly feeding and thus static, males actively seek females for mating. The male mating success increases by being motile and non-feeding, but its searching activity is limited by the energy resources previously stored. Theoretically, to maximise mating success, males have to alternate feeding and female searching periods which constitutes a paradox in the *Oithona* male behaviour (Kiørboe 2007).

In the Little Bay of Toulon, *O. nana* is the dominant zooplankton throughout the year without significant seasonal variation suggesting a continuous reproduction (Richard and Jamet 2001) as observed in others *Oithona* populations (Temperoni et al. 2011). Under laboratory condition with two sexes incubated separately, *O. nana* males have a mean lifetime of 25 days and 42 days for females. However, the lifespan *in situ* is unknown as well as its reproduction rate. Nonetheless, in the case of female saturation, *O. davisae* males have a reproduction rate (0.9 females male^{-1} day^{-1}) depending on the production of spermatophores that are transferred during a mating (Kiørboe 2007).

A biased sex-ratio toward females (male/female ratio<0.22) was observed in *O. nana* population of the Toulon Little Bay (Richard and Jamet 2001). Several causes could explain this observation. One factor evoked was the higher male exposure to predators due to its higher motility; this behaviour was assimilated to a "risky" behaviour (Hirst et al. 2010; Kiørboe 2006). However, we can propose other possibilities such as environmental sex determination (ESD) that has already been observed in other copepods (Voordouw and Anholt 2002) but also energy resource depletion as a consequence of male energy consumption during mate search. A risky behaviour implies environmental factors decreasing fitness, like encountering a predator. In contrast, a sacrificial behaviour is deterministic and thus only driven by genetic factors, for example, forcing the consumption of all energy resources to find a partner or die while trying. The sacrifice of males is observed in semelparous animals (single reproductive cycle in a lifetime) like insects or arachnids and is associated to female fitness increase. This behaviour has never been described in iteroparous animals like copepods (multiple reproductive cycles in a lifetime) (Hairston and Bohonak 1998).

Recently, the *O. nana* genome was sequenced, and its comparison to other genomes showed an explosion of Lin-12 Notch Repeat (LNR) domains-containing proteins coding genes (LDPGs) (Madoui et al. 2017). Among the 75 LDPGs present in the genome, five were found under natural selection in Mediterranean Sea populations, including notably one-point mutation generating an amino-acid change within the LNR domain of a male-specific protein (Arif et al. 2019; Madoui et al. 2017). This provided a first evidence of *O. nana* molecular differences between sexes at the transcriptional level and a potential gene repertoire of interest.

To further investigate the molecular basis of *O. nana* sexual differentiation, we propose in this study a multi-approach analysis including, (i) *in situ* sex ratio determination in time series (ii) sexual system determination by sex-specific polymorphism analysis; (iii) *in silico* analysis of

84 the structure and evolution of the LDPGs, (iv) sex-specific gene expression through RNA-seq
85 analysis and (v) LDPs interaction protein network by yeast two-hybrid.

86 Material and methods

87 Sex-ratio report in the Toulon Little Bay

88 *Oithona nana* specimens were sampled at the East of the Toulon Little Bay, France (Lat 43°
89 06' 52.1" N and Long 05° 55' 42.7" E). The samples were collected from the upper water layer
90 (0-10m) using zooplankton nets with a mesh of 90µm and 200µm. Samples were preserved in
91 5% formaldehyde. The monitoring of *O. nana* in Toulon Little Bay was performed from 2002
92 to 2017. Individuals of both sexes were identified and counted under the stereomicroscope.

93 Biological materials and RNA-seq experiments

94 For the RNA-seq experiment, the plankton sampling took place in November 2015 and
95 November 2016 using the same collecting method than previously described. The samples were
96 preserved in 70% ethanol and stored at -20°C. The *Oithona nana* were isolated under the
97 stereomicroscope. We selected individuals from five different development stages: five pairs
98 of egg-sac, four nauplii (larvae), four copepodites (juveniles), four female adults and four male
99 adults. All individuals were isolated from the November 2015 sample, except for eggs. Each
100 individual was transferred alone and crushed with a tissue grinder (Axygen) into a 1.5 mL tube
101 (Eppendorf). Total mRNAs were extracted with the NucleoSpin RNA XS kit (Macherey-Nagel)
102 following the manufacturer instructions, then quantified on Qubit 2.0 with the RNA HS Assay
103 kit (ThermoFisher Scientific) and quality assessed on Bioanalyzer 2100 with the RNA 6000
104 Pico Assay kit (Agilent). cDNAs were constructed using the SMARTer v4 Ultra low Input
105 RNA kit (Takara). After cDNA shearing by Covaris E210 instrument, Illumina libraries were
106 constructed using the NEBNext Ultra II kit (New England Biolabs) and sequenced on Illumina

HiSeq2500. A minimum of $9.7e^6$ reads pairs was produced from each individual (Supplementary Notes S1).

Sex-determination system identification by RNA-seq

RNA-seq reads of both sexes (four females and four males) were alignment against predicted cDNA. Reads having an alignment length ≤ 80% and identity cut-off ≤ 97% were removed. The variant calling step was performed with the '*samtools mpileup*' and '*bcftools call*' commands, with default parameters (Li et al. 2009) and only bi-allelic sites were kept.

To identify the most likely sexual system in *O. nana*, we used *SD-pop* (Käfer, Lartillot, Picard & Marais, *in prep*). Just like its predecessor *SEX-DETector* (Muyle et al. 2016), *SD-pop* calculates the likelihood of three sexual models (absence of sex chromosomes, XY system or ZW system) which can be compared using their BIC (Bayesian Information Criterion). The difference with *SEX-DETector* is that *SD-pop* is based on population genetics (i.e. Hardy-Weinberg equilibrium for autosomal genes, and different equilibria for sex-linked genes) instead of mendelian transmission from parents to offspring, and thus can be used without the requirement of obtaining a controlled cross.

The number of individuals used (four for each sex) is close to the lower limit for the use of *SD-pop,* where the robustness of the method is weakening. To test whether the model preferred by *SD-pop* could have been preferred purely by chance, we permuted the sex of the individuals, with the constraint of keeping four females and four males ((8!/(4!*4!)-1=69 permuted datasets). As the XY model is strictly equivalent to the ZW model with the sexes of all individuals changed, two *SD-pop* models (no sex chromosomes, ZW) were run on all possible permutations of the data, and the BIC of each model was calculated. The genes inferred as sex-linked based on their posterior probability (>0.8) were manually annotated.

***Arthropoda* phylogenetic tree**

131 The ribosomal 18S sequences from seven arthropods including five copepods (*O. nana,*
132 *Lepeophtheirus salmonis, Tigriopus californicus, Eurytemora affinis, Calanus glacialis,*
133 *Daphnia pulex and Drosophila melanogaster*) were downloaded from NCBI. The sequences
134 were aligned with MAFFT (Katoh and Standley 2013) using default parameters. The nucleotide
135 blocks conserved in the seven species were selected by Gblock on Seaview (Gouy, Guindon,
136 and Gascuel 2010) and manually curated. The Maximum-Likelihood phylogenetic tree was
137 constructed using PhyML 3.0 with General Time Reversible (GTR) model and branch support
138 computed by approximate likelihood ratio test (aLRT) method (Guindon et al. 2010).

139 **Genes annotation**

140 The functional annotation of the genes was updated from the previous genome annotation
141 (Madoui et al. 2017) using InterProScan v5.8-49.0 (Jones et al. 2014), BlastKOALA v2.1
142 (Kanehisa, Sato, and Morishima 2016) and by alignment on NCBI non-redundant protein
143 database using Diamond *(Buchfink, Xie, and Huso*n 2015). Furthermore, a list of *O. nana* genes
144 under natural selection in the Mediterranean Sea was added based on previous population
145 genomic analysis (Arif et al. 2019). We further considered the annotation provided by either (i)
146 Pfam (Finn et al. 2014) or SMART (Letunic and Bork 2018) protein domains, (ii) GO terms
147 (molecular function, biological process or cellular component) (Ashburner et al. 2000) (iii)
148 KEGG pathways (Kanehisa et al. 2012) and (iv) presence of locus under natural selection.
149 These four gene features were used to identify specific enrichment in a given set of genes using
150 a hypergeometric test that estimates the significance of the intersection between a specific gene
151 list and one of the four global annotation lists.

152 **HMM search for LDPGs identification**

153 From the InterProScan annotation of the *O. nana* proteome, 25 LNR domain sequences were
154 detected (*p-value* $\leq$10-6), extracted and aligned with MAFFT using default parameters (Katoh
155 and Standley 2013). A Hidden Markov Model (HMM) was generated from the aligned

sequences using the '*hmmbuild*' function of the HMMER tool version 3.1b1 (Eddy 2011). The *O. nana* proteome was scanned by '*hmmsearch*' using the LNR HMM profile. Detected domains were considered as canonical LNR domains for *E-value*, *c-E-value* and *i- E- value* $<10^{-6}$ and containing at least six cysteines or considered as LNR-like domains for *E-value, c-Evalue* and *i-Evalue* between 10^{-6} and 10^{-1} and containing at least four cysteines. A weblogo (Crooks et al. 2004) was generated to represent the conserved residues for the three LNRs of Notch protein, and the LNRs and LNR-like detected by HMM. The LNR and LNR-like domain-containing proteins constitute the LDPs final set used further. Deep-Loc (online execution) (Almagro Armenteros et al. 2017) was used to determinate the LDPs cellular localisation. To detect signal peptides and membrane protein topology, we used the online services of SignalP 5.0 (Almagro Armenteros et al. 2019) and TOPCONS (Tsirigos et al. 2015), respectively.

Phylogeny tree of *O. nana* LNR domains

The *O. nana* nucleotide sequences of the LNR and LNR-like domains were aligned using MAFFT with default parameters. The Maximum-Likelihood phylogenetic tree was constructed using PhyML.3.0 with a model designed by the online execution of Smart Model Selection v.1.8.1 (Lefort, Longueville, and Gascuel 2017) and with branch supports computed by the aLRT method. The GTR model was used with an estimated discrete Gamma distribution (-a= 1.418) and a proportion of fixed invariable sites (-I = 0.25). The tree was visualised using MEGA-X (Kumar et al. 2018).

Differential expression analysis

RNA-seq reads from the 20 libraries were mapped, independently, against the *O. nana* virtual cDNA, with '*bwa-mem*' (v. 0.7.15-r1140) using default parameters (Li 2013) and read counts were extracted from the 20 BAM files with samtools (v. 1.4) (Li et al. 2009). Each reads set was validated by pairwise MA-plot to ensure a global representation of the *O. nana*

transcriptome in each sample (Supplementary Notes S2). One nauplius sampled showing a biased read count distribution was discarded. Read counts from valid replicates were used as input data for the DESeq R package (Anders and Huber 2010) to identify differential gene expression between the five development stages through pairwise comparisons of each developmental stages. Genes having a Benjamini-Hochberg corrected p-value ≤ 0.05 in one of the pairwise comparisons were considered significantly differentially expressed. To identify stage-specific genes among these genes, we selected those being twice more expressed based on the normalised read count mean ($\log_2$(foldChange) > 1) in one development stage comparing to the four others. Up-regulated stage-specific genes were represented by a heatmap. The same method was used to determined down-regulated stage-specific genes (with log2(foldChange)<-1).

Protein-protein interaction assays by yeast two-hybrid screening

Yeast two-hybrid experiments were performed using Matchmaker Gold Yeast Two-Hybrid System (Takara). The coding sequences of interest were first cloned into the entry vector pDONR/Zeo (ThermoFisher) and the correct ORF sequence verified by Sanger sequencing. To this aim, LDPGs were PCR-amplified with Gateway-compatible primers (Supplementary Notes S6) using cDNAs of pooled male individuals as template. In the case of secreted proteins, the amplified ORF lacked the signal peptide. Then, the cloned ORFs were reamplified by a two step-PCR protocol allowing the creation of a recombination cassette containing the ORF flanked by 40-nucleotide tails homologous to the ends of the pGBKT7 bait vector at the cloning site. Linearised bait vectors and ORF cassettes were co-transformed in Y2HGold yeast strain, and ORF cloning was obtained by homologous recombination directly in yeast. Y2H screening for potential interacting partners of the baits was performed against a cDNA library obtained from a pool of total mRNA of 100 *O. nana* male individuals constructed into the pGAD-AD prey vector.

206 Before screening, the self-activity of each bait clone was tested by mating with the Y187 strain
207 harbouring an empty pGADT7-AD vector and then plating on SD/-His/-Leu/-Trp/ medium
208 supplemented with 0, 1, 3, 5 or 10 mM 3-amino 1,2,4-triazole (3-AT). Each bait clone was then
209 mated with the prey library containing approximately 4 x 10^6 individual clones and plated on
210 low-stringency agar plates (SD/-Trp/-Leu/-His/) supplemented with the optimal concentration
211 of 3-AT based on the results of the self-activity test. To decrease the false positive rate, after
212 five days of growth at 30°C, isolated colonies were spotted on high-stringency agar plates (SD/-
213 Leu/-Trp/-Ade/-His) supplemented with 3-AT and allowed to grow another five days. Colony
214 PCR on positive clones on this high stringency medium was performed with primers flanking
215 the cDNA insert on the pGAD-AD vector, and PCR products were directly Sanger sequenced.

216 ## Results

217 ### *Oithona nana* female-biased sex-ratio

218 Between 2002 and 2017, 186 samples were collected in the Toulon Little Bay (figure 1.a), from
219 which *O. nana* female and male adults were isolated. (figure 1.b). Across the fifteen years of
220 observations, we noted a minimum male/female ratio in February (0.11), maxima in September,
221 October and November (0.17) and a mean sex-ratio of 0.15 ± 0.11 all over the years (figure
222 1.c). This monitoring showed a relative stability along the year (ANOVA, P=0.87) but strongly
223 biased sex-ratio toward females.

224 ### Male homogamety

225 To identify the most likely cause of this sex-ratio bias between environment sex determination
226 (ESD) and higher male mortality, we used SD-pop on four individual transcriptomes of both
227 sexes to determine the *O. nana* sexual system. According to SD-pop, the ZW model was
228 preferred (lowest BIC) for *O. nana*. This result is unlikely to be due to chance, as for none of
229 the runs on the 69 datasets for which the sex was permuted, the ZW model had the lowest BIC.

Eleven genes had a posterior probability of being sex-linked in *O. nana* greater than 0.8. None of the SNPs in these genes showed the typical pattern of a fixed ZW SNP, i.e. the four females heterozygote and the four males homozygote (although, for some SNPs for which one individual was not genotyped, all four females were heterozygote, and all three genotyped males homozygote), indicating that the recombination suppression between the gametologs is recent, and that no or few mutations have gone to fixation independently in both gametolog copies. Annotation of these eleven genes shows that only one has homologs in other metazoans, ATP5H, that codes a subunit of mitochondrial ATP synthase (Supplementary Notes S8). Like in *Drosophila*, this gene is located in the nucleus (Liao et al. 2006).

LNR domains burst in the *O. nana* proteome

To identify LDPs, we developed a HMM dedicated to *O. nana* LNR identification based on 31 conserved amino-acid residues. In the *O. nana* proteome, 178 LNR and LNR-like domains were detected and coded by 75 LDPGs, while a maximum of eight domains coded by six LDPGs was detected in the four other copepods (figure 2.b). Among the 178 *O. nana* domains, 22 were canonical LNR and 156 LNR-like domains (figure 2.c). By comparing the structure of Notch LNRs and LNR-like, we observed the loss of two cysteines (figure 2.c) in the LNR-like domains. Among the 75 LDPs, we identified nine different protein structure patterns (figure 2.d), including notably 47 LNR-only proteins, 12 trypsin-associated LDPs and eight metallopeptidase-associated LDPs. Overall, LDPs were predicted to contain a maximum of 5 LNR domains and 13 LNR-like domains.

Forty-nine LDPs were predicted as secreted (eLDP), six membranous (mLDPs) and twenty intracellular (iLDPs) (Supplementary Notes S3). Among the iLDPs, two were associated with proteolytic domains, three associated with sugar-protein or protein-protein interaction domains (PAN/Appel, Lectin and Ankyrin) and 13 (65%) were LNR-only proteins. Among the eLDP, 18 (37%) contained proteolytic domains corresponding to a significant enrichment of

proteolysis in eLDPs (*p-value*=2.13e-17); other eLDPs corresponded to LNR-only proteins (63%). The mLDPs were constituted of one Notch protein, two proteins with LNR domains associated with lectin or thrombospondin domains respectively, and three LNR-only proteins.

In the LNR and LNR-like domains phylogenetic tree bases on nucleic sequences (figure 2e), only 17% of the nodes had a support over 90%. Twenty-seven branch splits corresponded to tandem duplications involving 15 LDPGs, including Notch and a cluster of five trypsin-associated LDPGs codings three eLDPs and two iLDPs.

Oithona nana male gene expression.

Among the 15,399 genes predicted on the *O. nana* reference genome, 1,233 (~8%) were significantly differentially expressed in at least one of the five developmental stages. Among them, 619 genes were specifically up-regulated in one stage, with 53 genes up-regulated in eggs, 19 in nauplii, 75 in copepodids, 27 in adult females and 445 in adult males (figure 3.a). The male up-regulated genes were categorised based on their functional annotation (figure 3.b).

Up-regulation of LNR-coding and proteolytic genes in adult male

The 1,233 differentially expressed genes contained 27 LDPGs (36% of total LDPGs) (figure 3.c). Over these 27 genes, 18 were specifically up-regulated in adult males producing a significant and robust enrichment of LDPGs in the adult males transcriptomes (fold > 8; *p-value*=2.95e-12) (figure 3.c). Among the 445 male-specific genes, 27 are predicted to play a role in proteolysis including 16 trypsins with three trypsin-associated LDPGs and showing significant enrichment of trypsin coding genes in males (*p-value*=1.73e-05), three metalloproteinases and five proteases inhibitors.

Up-regulation of nervous system associated genes in male adult

Forty-eight up-regulated genes in males are predicted functions in the nervous system (Supplementary Notes S4). These included 36 genes related to neuropeptides and hormones,

through their metabolism (10 genes) with seven enzymes involved in neuropeptide maturation and one allatostatin, through their transport and release (9 genes), and through neuropeptide or hormone receptors (17 genes), seven of which are FMRFamide receptors. Six genes are predicted involved in the neuron polarisation, four in the axonal and dendrites organisation and growth guidance (including homologs to B4GAT1, futsch-like and zig-like genes), two in the development and maintenance of sensory and motor neurons (IMPL2, and DYF-5) and one in synapse formation (SYG-2).

Up-regulation of amino-acid conversion into neurotransmitters in male adults

We observed ten up-regulated genes in males predicted to play a role in amino acid metabolism (figure 3.b). This includes five enzymes that convert directly lysine, tyrosine and glutamine into glutamate through the activity of one α-aminoadipic semialdehyde synthase (AASS), one tyrosine aminotransferase (TyrAT) and three glutaminases, respectively. Three other enzymes might play a role in the formation of pyruvate: one alanine dehydrogenase (AlaDH), one serine dehydrogenase (SDH), and indirectly one phosphoglycerate mutase (PGM). Furthermore, two other enzymes might be involved in the formation of glycine, one sarcosine dehydrogenase (SARDH) and one betaine-homocysteine methyltransferase (BHMT) (figure 3.d).

Food uptake regulation in male adult

Three genes, which had predicted functions in food uptake regulation, showed specific patterns in male. These included the increase of the allatostatin-coding gene expression, a neuropeptide known in arthropods to reduce food uptake, but also three male under-expressed genes, a crustacean cardioactive peptide (CCAP), a neuropeptide that triggers digestive enzymes activation and the two bursicon protein subunits. These latter two hormones are known to be involved in intestinal and metabolic homeostasis.

Protein-protein interaction involving LDPs and IGFBP

303 In order to further characterise the function of LDPGs, we studied potential proteins interactions
304 by Yeast two-hybrid (Y2H) analysis (Supplementary Notes S5). To this aim, we selected eleven
305 genes: seven male-overexpressed LDPGs, and four potential IGFBPs (Supplementary Notes
306 S6).

307 We performed Y2H analysis by two different approaches (Supplementary Notes S5): the first
308 was a matrix-based screen where potential binary interactions within candidate proteins were
309 tested one-to-one. The second approach aimed to identify potential interactors in the entire
310 *O. nana* proteome by a random library screen. This more time-consuming screening was
311 applied only to a subset of four genes (two LDPGs and two IGFBP) used as baits against a Y2H
312 library constructed from *O. nana* cDNAs.

313 Together, these two approaches allowed the reconstruction of a protein network containing 17
314 proteins including two LDPs and one IGFBP used as baits (figure 4.a), and 14 interacting
315 partners of which six have an ortholog in other metazoans and five have no ortholog but at least
316 one detected InterProScan domain (figure 4.b).

317 On_LDP1, an extracellular trypsin-containing LDP, formed a homodimer and interacted with
318 a trypsin, two extracellular matrix (ECM) proteins and also an insulin-like growth factor
319 binding protein (On_IGFBP7) that contains a trypsin inhibitor kazal domain. Based on its
320 phylogeny, this protein is homolog to IGFBP7, also present in vertebrates (Supplementary
321 Notes S7). On_IGFBP7 formed a homodimer and interacted with three other proteins: one
322 spondin-1 like protein (On_Spon1-like) containing a kazal domain, one thrombospondin
323 domain-containing protein and one vitellogenin 2-like protein (On_Vtg2). On_LDP2 is coded
324 by a gene up-regulated in male (figure 4.c), detected under selection (Arif et al. 2019) and
325 interacted with nine proteins: one vitellogenin 2-like protein (same interactant as On_IGFBP7),
326 three uncharacterized proteins, one thrombospondin domain-containing protein (different than
327 the On_IGFBP7 partner); one secretogranin V-like protein, one wnt-like protein, one laminin 1

subunit β and one a furin-like protein. No PPI with IGF was detected, and no homolog of insulin-like androgenic gland hormone (Ventura, Rosen, and Sagi 2011) was found in the *O. nana* proteome.

Discussion

High mortality rate of *O. nana* males.

Over 15 years of sampling, we observed a stable and strongly female-biased sex-ratio (~1:9) in the Toulon Little Bay. A similar observation was done in another *O. nana* population (Temperoni et al. 2011) and in other 132 *Oithonidae* populations (Kiørboe 2006). Two main causes could lead to these observations: a higher mortality of males or an environment-induced sex determination. As we showed that the *O. nana* sexual system is likely to be ZW, which would conduct to an expected 1:1 sex-ratio, the higher mortality rate of males seems more likely to explain our observations. These results are in accordance with the previously described risky behaviour of males, that is frequently in motion to find females and thus more vulnerable to predators than immobile females (Hirst et al. 2010).

LDPs driven proteolysis in *O. nana* males.

The explosion of LDPGs in the *O. nana* genome is unique in metazoan and is associated with the formation of new protein structures containing notably proteolytic domains. Owing to the LNR domain shortness (~40 amino acids) and the substantial polymorphism within the LNR domain sequences, the deep branches of the tree are weakly supported, and the evolutionary scenario of the domain burst remains hard to determine. However, the duplications of genes located in different scaffolds suggest post-duplication chromosomal rearrangements. Two previous studies on *O. nana* population genomics (Madoui et al. 2017; Arif et al. 2019) identified five LDPGs under natural selection with point mutations within an LNR domain. These results reinforce the idea of an ongoing evolution of these domains, forming new

352 structures and thus allowing the emergence of new functions, especially in *O. nana* males
353 according to the expression pattern of the LDPGs.

354 In metazoa, LNR domains are known to be involved in extracellular PPI (Boldt and Conover
355 2007) and cleavage site accessibility modulation (Sanchez-Irizarry et al. 2004). In *O. nana*,
356 iLDPs are the most abundant type of LDPs. Their associations with other protein-binding
357 domains like Kelsh, Ankyrin, PAN/Apple and thrombospondin repeat support a role of iLDPs
358 in intracellular PPI. Half of the eLDPs are LNR-only and the other half is associated with
359 peptidases (trypsin and metalloproteinase). From the PPI network, we showed that two eLDPs
360 (On_LDP1 and On_LPD2) might interact with different types of extracellular proteins involved
361 notably in tissue structure, energy storage and extracellular proteolysis. On the other hand,
362 transcriptomic analyses showed an enrichment of trypsins in male adults (figure 3.c). The
363 upregulation of allatostatin (Hergarden, Tayler, and Anderson 2012) and the downregulation of
364 CCAP (Žitňan and Daubnerová 2016) and bursicon (Scopelliti et al. 2019) allow the male to
365 reduce its food uptake. Taken together, this information supports a self-digestion of
366 extracellular proteins in males driven by eLDPs and trypsin complexes that could act on
367 proteolysis specificity modulation and/or on targeting/protecting specific extracellular proteins.
368 This autolysis allows the release of amino acid and energy not supplied by feeding. Thus, the
369 male adult autolysis could permit an expansion of mating period to increase its chances of
370 mating (Heuschele and Kiørboe 2012) without motionless feeding phases. On the other hand,
371 the deleterious effect of the autolysis on the organism supports a molecular-scale sacrificial
372 behaviour in *O. nana* male.

373 **Neurotransmitter biosynthesis and nervous system development in *O. nana* male**

374 From gene expression profile in *O. nana* male, we highlighted the direct and indirect conversion
375 of four amino acids to glutamate, an excitatory neurotransmitter in arthropods, and the
376 production of proteins composing the neurotransmitter vesicle transport which is consistent

with the hyperactivity of the males during mate search. From the upregulation of neuronal developmental genes in male adult, and especially syg-2 and zig-8 normally expressed during the larval phase (Shen, Fetter, and Bargmann 2004), we infer an ongoing formation of new axons and/or dendrites and synapses in the male motor and/or sensory neurons. These results suggest a sexual dimorphism of the *O. nana* nervous system, as recently demonstrated in *Caenorhabditis elegans* (Cook et al. 2019).

Moreover, On_LDPG2, a male over-expressed gene and under natural selection in Mediterranean Sea populations, has been shown to interacts in yeast with two proteins involved in the nervous system development (On_Wnt and On_Lamβ1), notably in axon guidance (Zou 2004; Randlett et al. 2011). So, through its interactions, On_LDP2 may modulate neurogenesis in males and participate to the sexual dimorphism of the *O. nana* nervous system.

Conclusion

In the Toulon Little Bay, *O. nana* presents a strong biased sex-ratio toward female while a ZW sexual determination system is favoured, which supports a higher mortality rate in male that can be explained by the sacrificial behaviour of males due to non-feeding, high motility and autolysis. The explosion of LDPGs in the *O. nana* genome seems to play an important role in the male-specific neurogenesis and autolysis. However, more investigation should be undergone to identify which part of the nervous system is developing and which tissues are lysed. To our knowledge, sacrificial behaviour was only observed in semelparous animals. Thus, the sacrificial behaviour supported at the molecular level in *O. nana* by autolysis may represent the first example of sacrificial behaviour in semelparous animals. Subtle trade-offs and molecular control mechanisms for the autolysis of specific tissues may occur to adjust the autolysis to the lifespan of the male allowing it to reproduce several times. The low mortality rate of the motionless females and the male sacrificial behaviour could be one of the main factors of the ecological success of *O. nana (Razouls et al. 2019)*.

Acknowledgements

We acknowledge the support of the Genoscope-CEA, France Génomique (ANR-10-INBS-09) and the French Ministry of Research.

Data availability

The *O. nana* RNA-seq data are available at ENA (*Supplementary Notes* S1)

Authors' contribution

KS, JP and JLJ collected the samples; JLJ generated the sex-ratio data; KS, KL, MAM, EP and JP generated the molecular data; JK performed sexual system analysis; KS, AA, LB, SK, NM, and CO performed the yeast two-hybrid analysis; AA designed the yeast two-hybrid method; KS and MAM performed the analyses; KS, AA and MAM wrote to the manuscript; MAM supervised the study.

Bibliography

Almagro Armenteros, José Juan, Casper Kaae Sønderby, Søren Kaae Sønderby, Henrik Nielsen, and Ole Winther. 2017. 'DeepLoc: Prediction of Protein Subcellular Localization Using Deep Learning'. Edited by John Hancock. *Bioinformatics* 33 (21): 3387–95. https://doi.org/10.1093/bioinformatics/btx431.

Almagro Armenteros, José Juan, Konstantinos D. Tsirigos, Casper Kaae Sønderby, Thomas Nordahl Petersen, Ole Winther, Søren Brunak, Gunnar von Heijne, and Henrik Nielsen. 2019. 'SignalP 5.0 Improves Signal Peptide Predictions Using Deep Neural Networks'. *Nature Biotechnology* 37 (4): 420–23. https://doi.org/10.1038/s41587-019-0036-z.

Anders, Simon, and Wolfgang Huber. 2010. 'Differential Expression Analysis for Sequence Count Data'. *Genome Biology* 11 (10): R106. https://doi.org/10.1186/gb-2010-11-10-r106.

Arif, Majda, Jérémy Gauthier, Kevin Sugier, Daniele Iudicone, Olivier Jaillon, Patrick Wincker, Pierre Peterlongo, and Mohammed-Amin Mohammed-Amin Mohammed-Amin Madoui. 2019. 'Discovering Millions of Plankton Genomic Markers from the Atlantic Ocean and the Mediterranean Sea'. *Molecular Ecology Resources* 19 (2): 0–3. https://doi.org/10.1111/1755-0998.12985.

Ashburner, M, C A Ball, J A Blake, D Botstein, H Butler, J M Cherry, A P Davis, et al. 2000. 'Gene Ontology: Tool for the Unification of Biology. The Gene Ontology Consortium.' *Nature Genetics* 25 (1): 25–29. https://doi.org/10.1038/75556.

Boldt, Henning B., and Cheryl A. Conover. 2007. 'Pregnancy-Associated Plasma Protein-A (PAPP-A): A Local Regulator of IGF Bioavailability through Cleavage of IGFBPs'. *Growth Hormone & IGF Research* 17 (1): 10–18. https://doi.org/10.1016/j.ghir.2006.11.003.

435 Buchfink, Benjamin, Chao Xie, and Daniel H Huson. 2015. 'Fast and Sensitive Protein Alignment Using
436 DIAMOND'. *Nature Methods* 12 (1): 59–60. https://doi.org/10.1038/nmeth.3176.

437 Cook, Steven J., Travis A. Jarrell, Christopher A. Brittin, Yi Wang, Adam E. Bloniarz, Maksim A.
438 Yakovlev, Ken C. Q. Nguyen, et al. 2019. 'Whole-Animal Connectomes of Both Caenorhabditis
439 Elegans Sexes'. *Nature* 571 (7763): 63–71. https://doi.org/10.1038/s41586-019-1352-7.

440 Cornils, Astrid, Britta Wend-Heckmann, and Christoph Held. 2017. 'Global Phylogeography of Oithona
441 Similis s.l. (Crustacea, Copepoda, Oithonidae) – A Cosmopolitan Plankton Species or a Complex
442 of Cryptic Lineages?' *Molecular Phylogenetics and Evolution* 107 (February): 473–85.
443 https://doi.org/10.1016/j.ympev.2016.12.019.

444 Crooks, Gavin E, Gary Hon, John-Marc Chandonia, and Steven E Brenner. 2004. 'WebLogo: A
445 Sequence Logo Generator'. *Genome Research* 14: 1188–90. https://doi.org/10.1101/gr.849004.

446 Dvoretsky, V. G., and A. G. Dvoretsky. 2009. 'Life Cycle of Oithona Similis (Copepoda: Cyclopoida)
447 in Kola Bay (Barents Sea)'. *Marine Biology* 156 (7): 1433–46. https://doi.org/10.1007/s00227-
448 009-1183-4.

449 Eddy, Sean R. 2011. 'Accelerated Profile HMM Searches'. Edited by William R. Pearson. *PLoS*
450 *Computational Biology* 7 (10): e1002195. https://doi.org/10.1371/journal.pcbi.1002195.

451 Finn, Robert D, Alex Bateman, Jody Clements, Penelope Coggill, Ruth Y Eberhardt, Sean R Eddy,
452 Andreas Heger, et al. 2014. 'Pfam: The Protein Families Database'. *Nucleic Acids Res.* 42.
453 https://doi.org/10.1093/nar/gkt1223.

454 Gallienne, C. P., and D. B. Robins. 2001. 'Is Oithona the Most Important Copepod in the World's
455 Oceans?' *Journal of Plankton Research* 23 (12): 1421–32.
456 https://doi.org/10.1093/plankt/23.12.1421.

457 Gouy, M., S. Guindon, and O. Gascuel. 2010. 'SeaView Version 4: A Multiplatform Graphical User
458 Interface for Sequence Alignment and Phylogenetic Tree Building'. *Molecular Biology and*
459 *Evolution* 27 (2): 221–24. https://doi.org/10.1093/molbev/msp259.

460 Guindon, Stéphane, Jean-François Dufayard, Vincent Lefort, Maria Anisimova, Wim Hordijk, and
461 Olivier Gascuel. 2010. 'New Algorithms and Methods to Estimate Maximum-Likelihood
462 Phylogenies: Assessing the Performance of PhyML 3.0'. *Systematic Biology* 59 (3): 307–21.
463 https://doi.org/10.1093/sysbio/syq010.

464 Hairston, Nelson G), and Andrew J Bohonak. 1998. *Copepod Reproductive Strategies: Life-History*
465 *Theory, Phylogenetic Pattern and Invasion of Inland Waters. Journal of Marine Systems*. Vol. 15.
466 Elsevier. https://doi.org/10.1016/S0924-7963(97)00046-8.

467 Hergarden, Anne Christina, Timothy D Tayler, and David J Anderson. 2012. 'Allatostatin-A Neurons
468 Inhibit Feeding Behavior in Adult Drosophila.' *Proceedings of the National Academy of Sciences*
469 *of the United States of America* 109 (10): 3967–72. https://doi.org/10.1073/pnas.1200778109.

470 Heuschele, Jan, and Thomas Kiørboe. 2012. 'The Smell of Virgins: Mating Status of Females Affects
471 Male Swimming Behaviour in Oithona Davisae'. *Journal of Plankton Research* 34 (11): 929–35.
472 https://doi.org/10.1093/plankt/fbs054.

473 Hirst, A. G., D. Bonnet, D. V.P. P. Conway, and T. Kiørboe. 2010. 'Does Predation Control Adult Sex
474 Ratios and Longevities in Marine Pelagic Copepods?' *Limnology and Oceanography* 55 (5): 2193–
475 2206. https://doi.org/10.4319/lo.2010.55.5.2193.

476 Huys, Rony., and Geoffrey Allan. Boxshall. 1991. *Copepod Evolution*. Ray Society.
477 http://www.raysociety.org.uk/publications/zoology/copepod-evolution-r-huys-and-g-a-boxshall/.

478 Jones, Philip, David Binns, Hsin-Yu Chang, Matthew Fraser, Weizhong Li, Craig McAnulla, Hamish
479 McWilliam, et al. 2014. 'InterProScan 5: Genome-Scale Protein Function Classification.'
480 *Bioinformatics (Oxford, England)* 30 (9): 1236–40. https://doi.org/10.1093/bioinformatics/btu031.

481 Kanehisa, Minoru, Susumu Goto, Yoko Sato, Miho Furumichi, and Mao Tanabe. 2012. 'KEGG for

Integration and Interpretation of Large-Scale Molecular Data Sets.' *Nucleic Acids Research* 40 (Database issue): D109-14. https://doi.org/10.1093/nar/gkr988.

Kanehisa, Minoru, Yoko Sato, and Kanae Morishima. 2016. 'BlastKOALA and GhostKOALA: KEGG Tools for Functional Characterization of Genome and Metagenome Sequences'. *Journal of Molecular Biology* 428 (4): 726–31. https://doi.org/10.1016/J.JMB.2015.11.006.

Katoh, Kazutaka, and Daron M Standley. 2013. 'MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability.' *Molecular Biology and Evolution* 30 (4): 772–80. https://doi.org/10.1093/molbev/mst010.

Kiørboe, Thomas. 2006. 'Sex, Sex-Ratios, and the Dynamics of Pelagic Copepod Populations'. *Oecologia* 148 (1): 40–50. https://doi.org/10.1007/s00442-005-0346-3.

———. 2007. 'Mate Finding, Mating, and Population Dynamics in a Planktonic Copepod Oithona Davisae: There Are Too Few Males'. *Limnology and Oceanography* 52 (4): 1511–22. https://doi.org/10.4319/lo.2007.52.4.1511.

———. 2011. 'What Makes Pelagic Copepods so Successful?' *Journal of Plankton Research* 33 (5): 677–85. https://doi.org/10.1093/plankt/fbq159.

Letunic, Ivica, and Peer Bork. 2018. '20 Years of the SMART Protein Domain Annotation Resource'. *Nucleic Acids Research* 46 (D1): D493–96. https://doi.org/10.1093/nar/gkx922.

Li, Heng. 2013. 'Aligning Sequence Reads, Clone Sequences and Assembly Contigs with BWA-MEM', March. http://arxiv.org/abs/1303.3997.

Li, Heng, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, and 1000 Genome Project Data Processing 1000 Genome Project Data Processing Subgroup. 2009. 'The Sequence Alignment/Map Format and SAMtools.' *Bioinformatics (Oxford, England)* 25 (16): 2078–79. https://doi.org/10.1093/bioinformatics/btp352.

Liao, T. S.Vivian, Gerald B. Call, Preeta Guptan, Albert Cespedes, Jamie Marshall, Kevin Yackle, Edward Owusu-Ansah, et al. 2006. 'An Efficient Genetic Screen in Drosophila to Identify Nuclear-Encoded Genes with Mitochondrial Function'. *Genetics* 174 (1): 525–33. https://doi.org/10.1534/genetics.106.061705.

Madoui, Mohammed-Amin, Julie Poulain, Kevin Sugier, Marc Wessner, Benjamin Noel, Leo Berline, Karine Labadie, et al. 2017. 'New Insights into Global Biogeography, Population Structure and Natural Selection from the Genome of the Epipelagic Copepod *Oithona*'. *Molecular Ecology* 26 (17): 4467–82. https://doi.org/10.1111/mec.14214.

Mironova, Ekaterina, and Anna Pasternak. 2017. 'Female Gonad Morphology of Small Copepods Oithona Similis and Microsetella Norvegica'. *Polar Biology* 40 (3): 685–96. https://doi.org/10.1007/s00300-016-1993-z.

Muyle, Aline, Jos Käfer, Niklaus Zemp, Sylvain Mousset, Franck Picard, and Gabriel A.B. Marais. 2016. 'Sex-Detector: A Probabilistic Approach to Study Sex Chromosomes in Non-Model Organisms'. *Genome Biology and Evolution* 8 (8): 2530–43. https://doi.org/10.1093/gbe/evw172.

Nishida, Shuhei. 1985. 'Taxonomy and Distribution of the Family Oithonidae (Copepoda, Cyclopoida) in the Pacific and Indian'. *Bull. Ocean Res. Inst.* 20: 1–167.

Paffenhöfer, Gustav-Adolf. 1993. 'On the Ecology of Marine Cyclopoid Copepods (Crustacea, Copepoda)'. *Journal of Plankton Research* 15 (1): 37–55. https://doi.org/10.1093/plankt/15.1.37.

Randlett, Owen, Lucia Poggi, Flavio R Zolessi, and William A Harris. 2011. 'The Oriented Emergence of Axons from Retinal Ganglion Cells Is Directed by Laminin Contact in Vivo.' *Neuron* 70 (2): 266–80. https://doi.org/10.1016/j.neuron.2011.03.013.

Razouls, C., F. de Bovée, J. Kouwenberg, and N. Desreumaux. 2019. 'Diversité et Répartition Géographique Chez Les Copépodes Planctoniques Marins'. 2019. https://copepodes.obs-

banyuls.fr/.

Richard, Simone, and Jean-louis Jamet. 2001. 'An Unusual Distribution of Oithona Nana GIESBRECHT (1892) (Crustacea : Cyclopoida) in a Bay : The Case of Toulon Bay (France, Miditerranean Sea)'. *Journal of Coastal Research*, no. April 2015.

Sanchez-Irizarry, C., A. C. Carpenter, A. P. Weng, W. S. Pear, J. C. Aster, and S. C. Blacklow. 2004. 'Notch Subunit Heterodimerization and Prevention of Ligand-Independent Proteolytic Activation Depend, Respectively, on a Novel Domain and the LNR Repeats'. *Molecular and Cellular Biology* 24 (21): 9265–73. https://doi.org/10.1128/MCB.24.21.9265-9273.2004.

Scopelliti, Alessandro, Christin Bauer, Yachuan Yu, Tong Zhang, Björn Kruspig, Daniel J. Murphy, Marcos Vidal, Oliver D.K. Maddocks, and Julia B. Cordero. 2019. 'A Neuronal Relay Mediates a Nutrient Responsive Gut/Fat Body Axis Regulating Energy Homeostasis in Adult Drosophila'. *Cell Metabolism* 29 (2): 269-284.e10. https://doi.org/10.1016/J.CMET.2018.09.021.

Shen, Kang, Richard D Fetter, and Cornelia I Bargmann. 2004. 'Synaptic Specificity Is Generated by the Synaptic Guidepost Protein SYG-2 and Its Receptor, SYG-1'. *Cell* 116 (6): 869–81. https://doi.org/10.1016/S0092-8674(04)00251-X.

Steinberg, Deborah K., and Michael R. Landry. 2017. 'Zooplankton and the Ocean Carbon Cycle'. *Annual Review of Marine Science* 9 (1): 413–44. https://doi.org/10.1146/annurev-marine-010814-015924.

Sugier, Kevin, Benoit Vacherie, Astrid Cornils, Patrick Wincker, Jean-Louis Jamet, and Mohammed-Amin Madoui. 2018. 'Chitin Distribution in the Oithona Digestive and Reproductive Systems Revealed by Fluorescence Microscopy'. *PeerJ* 6 (May): e4685. https://doi.org/10.7717/peerj.4685.

Temperoni, B., M. D. Vinas, N. Diovisalvi, and R. Negri. 2011. 'Seasonal Production of Oithona Nana Giesbrecht, 1893 (Copepoda: Cyclopoida) in Temperate Coastal Waters off Argentina'. *Journal of Plankton Research* 33 (5): 729–40. https://doi.org/10.1093/plankt/fbq141.

Tsirigos, Konstantinos D, Christoph Peters, Nanjiang Shu, Lukas Käll, and Arne Elofsson. 2015. 'The TOPCONS Web Server for Consensus Prediction of Membrane Protein Topology and Signal Peptides.' *Nucleic Acids Research* 43 (W1): W401-7. https://doi.org/10.1093/nar/gkv485.

Ventura, Tomer, Ohad Rosen, and Amir Sagi. 2011. 'From the Discovery of the Crustacean Androgenic Gland to the Insulin-like Hormone in Six Decades'. *General and Comparative Endocrinology* 173 (3): 381–88. https://doi.org/10.1016/j.ygcen.2011.05.018.

Voordouw, M. J., and B. R. Anholt. 2002. 'Environmental Sex Determination in a Splash Pool Copepod'. *Biological Journal of the Linnean Society* 76 (4): 511–20. https://doi.org/10.1046/j.1095-8312.2002.00087.x.

Zamora-Terol, Sara, Sanne Kjellerup, Rasmus Swalethorp, Enric Saiz, and Torkel Gissel Nielsen. 2014. 'Population Dynamics and Production of the Small Copepod Oithona Spp. in a Subarctic Fjord of West Greenland'. *Polar Biology* 37 (7): 953–65. https://doi.org/10.1007/s00300-014-1493-y.

Zamora Terol, Sara. 2013. 'Ecology of the Marine Copepod Genus Oithona'. *TDX (Tesis Doctorals En Xarxa)*. Universitat Politècnica de Catalunya. https://upcommons.upc.edu/handle/2117/95142.

Žitňan, Dušan, and Ivana Daubnerová. 2016. 'Crustacean Cardioactive Peptide'. In *Handbook of Hormones*, 442-e69-2. Academic Press. https://doi.org/10.1016/B978-0-12-801028-0.00069-6.

Zou, Yimin. 2004. 'Wnt Signaling in Axon Guidance.' *Trends in Neurosciences* 27 (9): 528–32. https://doi.org/10.1016/j.tins.2004.06.015.

Figure Legends

Figure1: **Life cycle and sex-ratio of the copepod *Oithona nana* in the Toulon Little Bay. a.** Sampling station map in the Toulon Little Bay**. b**. The life cycle of *O. nana*. **c**. Sex ratio of *O. nana* time series in the Toulon Little Bay from 2002 to 2017. Black circles represent the mean by month. The blue line represents the fifteen-years mean (0.15).

Figure 2: **Lin-12 Notch Repeat (LNR) protein domain burst and high divergence with new domain associations in the *Oithona nana* proteome**. **a**. Phylogeny of five copepod species and two other arthropod species based on 18S ribosomal sequences. The numbers at internal branches show the aLRT branch support. The scale bar represents the nucleotide substitution rate. **b**. LNR domain occurrences in seven Arthropoda proteomes detected by HMM. In front of each bar corresponds to the number of detected genes. **c**. Consensus sequences of the *O. nana* Notch LNR, LNR and LNR-like domains generated by WebLogo. The asterisks represent the conserved sites. **d**. Schemata of the *O. nana* LNR and LNR-like proteins structure. Numbers under each domain represent the possible occurrence range. The barplot represents the occurrence of the nine structures. **e**. Phylogenetic tree of the *O. nana* LNR and LNR-like domains. Bold branches have aLRT support $\geq$ 0.90. The red circles represent tandem duplication.

Figure 3: **Differential expression analysis of the *Oithona nana* transcriptomes**. **a**. Heatmap of the 1,233 significantly differentially expressed genes in at least one of the five developmental stages. **b**. Functional annotation distribution of 445 genes explicitly overexpressed in male adults. **c**. Heatmap of the 27 significantly differentially expressed LDPGs and their protein domains composition. **d**. Amino acid conversion to neurotransmitters in *O. nana* male. Over-expressed enzymes in males are in red, amino acids in blue and neurotransmitter amino acids in green.

Figure 4: **Protein-Protein Interaction of LNR-containing proteins in the *O. nana* male proteome**. **a**. Structure and expression of the PPI candidates. The red arrows represent the PCR primers. **b**. PPI network of LDPGs obtained by Yeast two-hybrid assays. Lamβ 1: Laminin subunit beta 1; Spon1-like: Spondin1-like; Vtg2: Vitellogenin2; SgIV: Secretogranin IV; thbs: thrombospondins; unk: unknow. **c**. RPKM normalised expression for the five developmental stages. Only the five statistically differentially expressed genes are shown. From left to right: egg (e), larva (l), juvenile (j), female adult (f), male adult (m).

Supplementary Notes S1: **Transcriptomic data.** RNA-seq quality metrics of the 20 samples. One nauplii sample was discarded after MA plots pairwise comparisons analysis (see Supplementary Notes 2).

Supplementary Notes S2: **Transcriptomes quality**. MA plots pairwise comparisons between the five developmental stages. One nauplius sampled showing a biased read count distribution was discarded.

Supplementary Notes S3: **Structure and localisation of the *Oithona nana* LDPs**. *e, i* and *m* correspond to extracellular, intracellular and membranous respectively

Supplementary Notes S4: **Functional annotation of *O. nana* genes over-expressed in male**.

Supplementary Notes S5: **Experimental design of the protein interaction (PPI) analysis**.

Supplementary Notes S6: **Primers used for PCR amplification**. The first table contains the primers used for bait amplification; the second contains the prey ones.

Supplementary Notes S7: **Phylogenetic tree of On_IGFBP7**. The numbers at internal branches show the bootstrap branch support (100). We used NCBI data from *Mus muculus* (mouse), *Danio rerio* (danre) and *Bos taurus* (bovin).

Supplementary Notes S8. **Gene annotation of the sex-determination system associated genes of *Oithona nana***

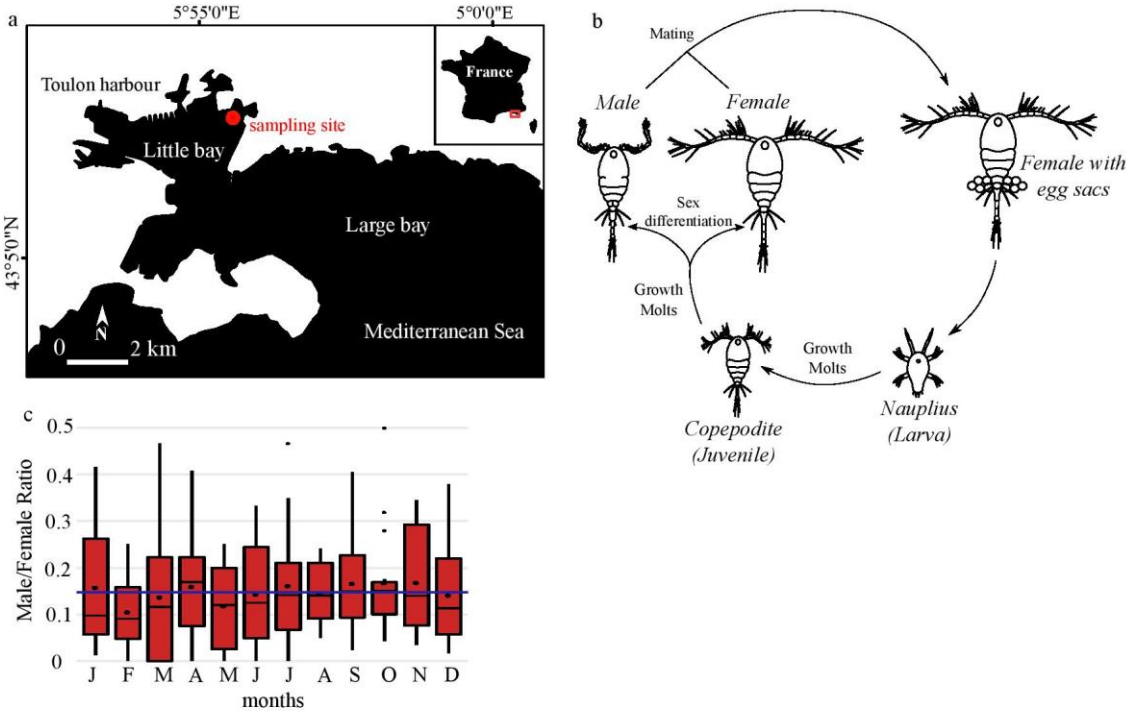


Figure1: **Life cycle and sex-ratio of the copepod *Oithona nana* in the Toulon Little Bay. a.** Sampling station map in the Toulon Little Bay**. b**. The life cycle of *O. nana*. **c**. Sex ratio of *O. nana* time series in the Toulon Little Bay from 2002 to 2017. Black circles represent the mean by month. The blue line represents the fifteen-years mean (0.15).

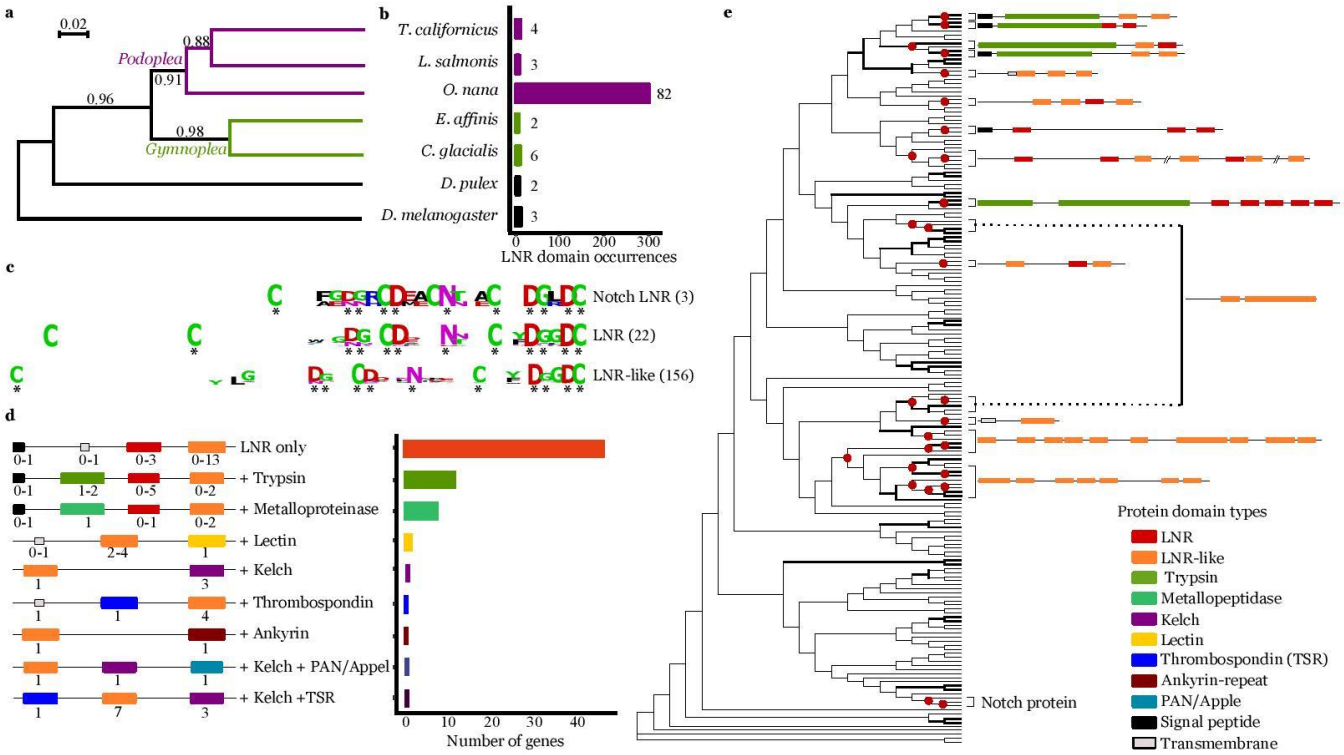


Figure 2: **Lin-12 Notch Repeat (LNR) protein domain burst and high divergence with new domain associations in the *Oithona nana* proteome. a**. Phylogeny of five copepod species and two other arthropod species based on 18S ribosomal sequences. The numbers at internal branches show the aLRT branch support. The scale bar represents the nucleotide substitution rate. **b**. LNR domain occurrences in seven Arthropoda proteomes detected by HMM. In front of each bar corresponds to the number of detected genes. **c.** Consensus sequences of the *O. nana* Notch LNR, LNR and LNR-like domains generated by WebLogo. The asterisks represent the conserved sites. **d.** Schemata of the *O. nana* LNR and LNR-like proteins structure. Numbers under each domain represent the possible occurrence range. The barplot represents the occurrence of the nine structures. **e.** Phylogenetic tree of the *O. nana* LNR and LNR-like domains. Bold branches have aLRT support ≥ 0.90. The red circles represent tandem duplication.

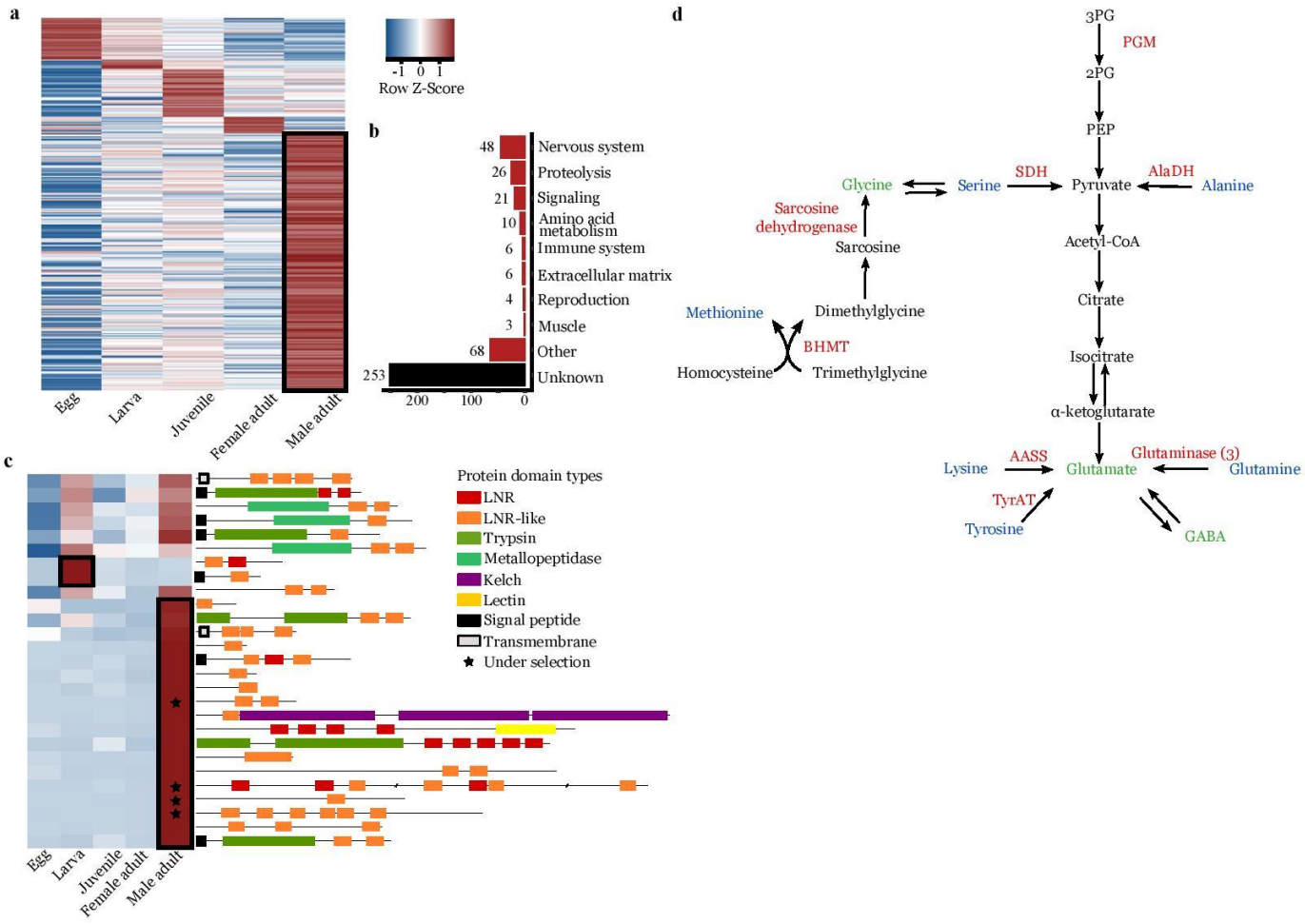


Figure 3: **Differential expression analysis of the *Oithona nana* transcriptomes**. **a**. Heatmap of the 1,233 significantly differentially expressed genes in at least one of the five developmental stages. **b**. Functional annotation distribution of 445 genes explicitly overexpressed in male adults. **c**. Heatmap of the 27 significantly differentially expressed LDPGs and their protein domains composition. **d**. Amino acid conversion to neurotransmitters in *O. nana* male. Over-expressed enzymes in males are in red, amino acids in blue and neurotransmitter amino acids in green.

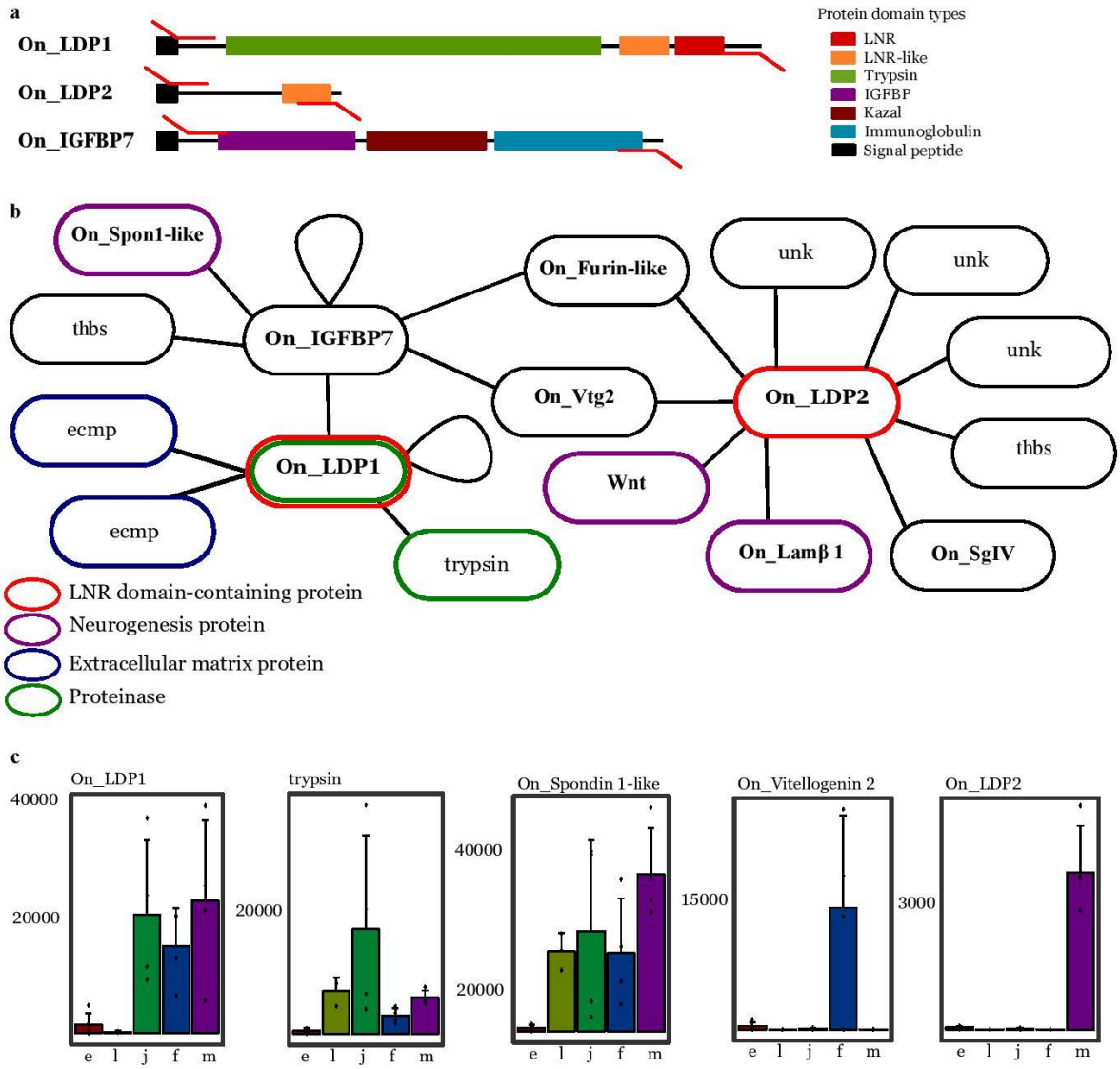


Figure 4: **Protein-Protein Interaction of LNR -containing proteins in the *O. nana* male proteome**. **a**. Structure and expression of the PPI candidates. The red arrows represent the PCR primers. **b**. PPI network of LDPGs obtained by Yeast double-hybrid assays. Lamβ 1: Laminin subunit beta 1; Spon1-like: Spondin1-like; Vtg2: Vitellogenin2; SgIV: Secretogranin IV; thbs: thrombospondins; unk: unknow. **c**. RPKM normalised expression for the five developmental stages. Only the five statistically differentially expressed genes are shown. From lef t to right: egg (e), larva (l), juvenile (j), female adult (f), male adult (m).

Chapter 4:

Nervous system specific genes are targeted by population-scale ASE and natural selection in Arctic *O. similis* population

This study tries to establish a link between population-scale allele-specific expression (ASE) and natural selection. To test this hypothesis, we used *Tara* Oceans metagenomic and metatranscriptomic data and also the *Oithona similis* transcriptomes data.

This investigation provides the first analysis of ASE at a population level. The analysis of seven populations of the copepod *O. similis* showed a significant amount of genes targeted by both population-scale ASE and natural selection. Among these genes, eight were involved in the nervous system (glutamate metabolism, glycine and/or GABA receptors, visual perception). As a GABA receptor was also found under selection in *O. nana* in the Mediterranean Sea in the previous studies of the laboratory (Madoui et al. 2017; Arif et al. 2019), these new results are coherent and highlight the *Oithona* nervous system as a main target of the natural selection.

In this study, I performed all the wet lab work except for the *Tara* Oceans data. I collected the samples in the large Toulon Bay, I sorted the of *O. similis* individuals at differential development states, and I performed the mRNA extraction and quality analyses.

Article source

Laso-Jadart, Romuald, Kevin Sugier, Emmanuelle Petit, Karine Labadie, Pierre Peterlongo, Christophe Ambroise, Patrick Wincker, Jean-Louis Jamet, and Mohammed-Amin Madoui. 2019. “Linking Allele-Specific Expression And Natural Selection In Wild Populations”. BioRxiv, April, 599076. https://doi.org/10.1101/599076.

Laso-Jadart R., Sugier K., Petit E., Labadie K., Peterlongo P., Ambroise C., Wincker P., Jamet J.-L. and Madoui M.-A., “Linking Allele-Specific Expression And Natural Selection In Wild Populations”, *Ecology and Evolution*, 2020

A French abstract is available in appendix 11 (page 164)

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

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 Arctic[30–33]. Copepods, and particularly *Oithona*, are small crustaceans forming the most abundant metazoan on Earth, reflecting strong adaptive capacities to environmental fluctuations[34–36].They play a key ecological role in biogeochemical cycles and in the marine trophic food chain[37]; therefore copepods constitute an ideal model to study wild population evolution[38–41]

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 expedition[42] 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[43] using default parameters and transcripts were clustered using cd-hit v4.6.1[44] (Supplementary Table 1).To ensure the classification of the sampled individuals, each ribosomal read set were detected with SortMeRNA [45] and mapped with bwa v0.7.15using default parameters [46] 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[43] 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 [40] using DIAMOND v0.9.22[48]. Protein domain annotation was performed on the final ORF predictions with Interproscan v5.17.56.0[49] 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

database, restricted to Arthropoda (taxid: 6656), with DIAMOND v0.9.22 (Supplementary Fig. 1).

Variant calling using *Tara* Oceans metagenomic and metatranscriptomic data

We used metagenomic and metatranscriptomic reads generated from samples of the size fraction 20–180 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 [50].

The reference-free variant caller *DiscoSNP++* [51,52] 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 [53,54] 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 σ 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 [55] 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 [56,57].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 [58]. We retained the median pairwise-F_{ST} as a measure of the genomic differentiation between each population. The *pcadapt* R package v4.0.2 [59] 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 [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 ≠ {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.

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 [62]. 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[63], 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).

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).

198 **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,

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).

226 # Discussion

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

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.

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

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].

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

572 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

625 **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	-

626

627

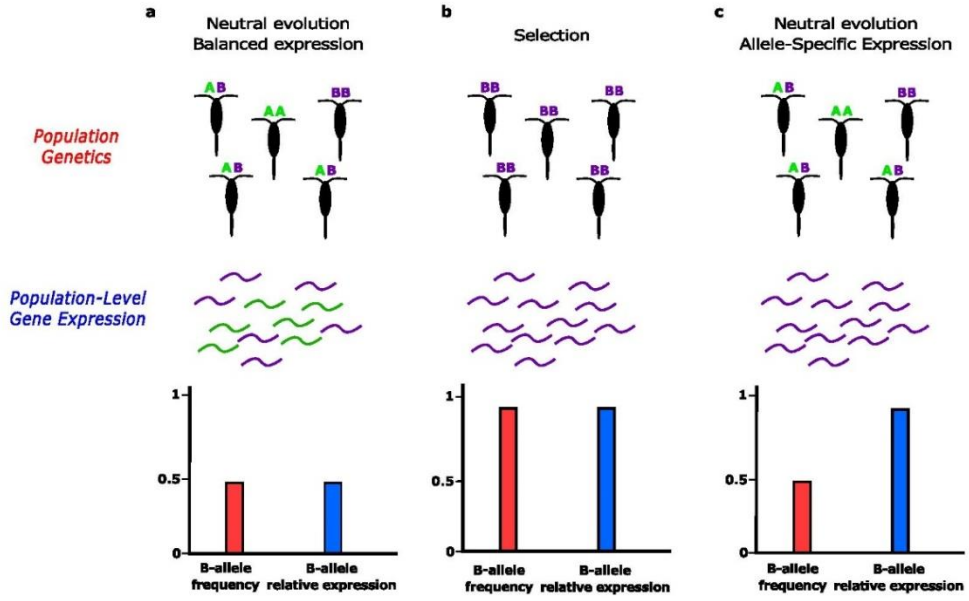


Figure 1: Population genomic and transcriptomic profiles of a biallelic locus in a case of **a,** Neutral evolution and balanced expression; **b,** Selection in favor of the B-allele; **c,** ASE in favor of the B-allele.

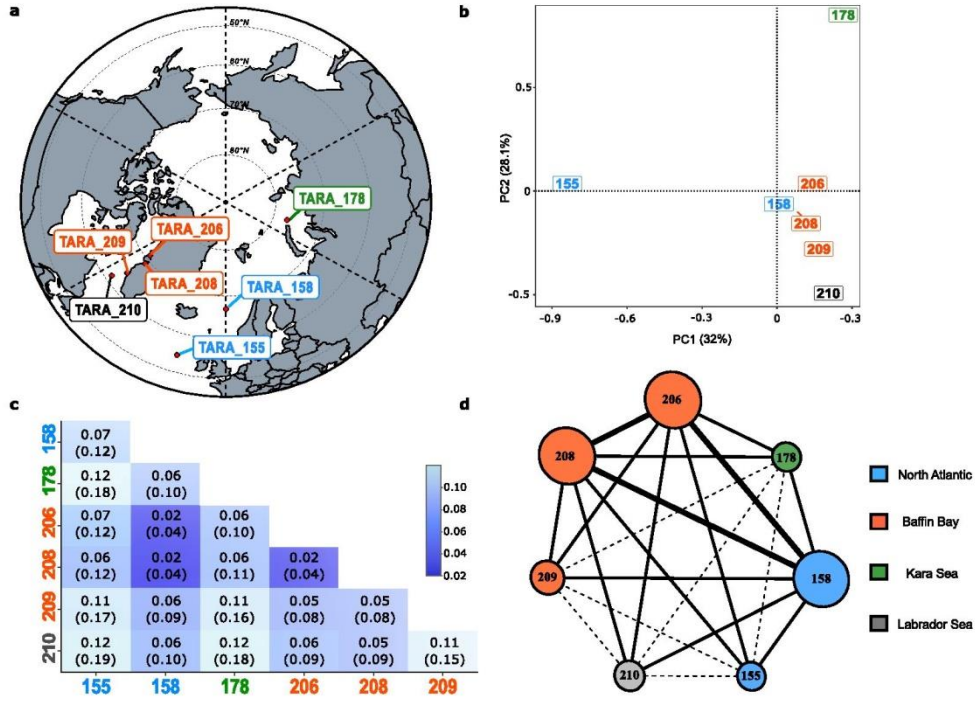


Figure 2: Genomic differentiation of *O. similis* populations from Arctic Seas. **a,** Geographic locations of the seven *Tara* Oceans sampling sites: Northern Atlantic (blue), Kara Sea (green), Baffin Bay (orange) and Labrador Sea (grey). **b,** Principal Component Analysis (PCA) computed by *pcadapt* 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).

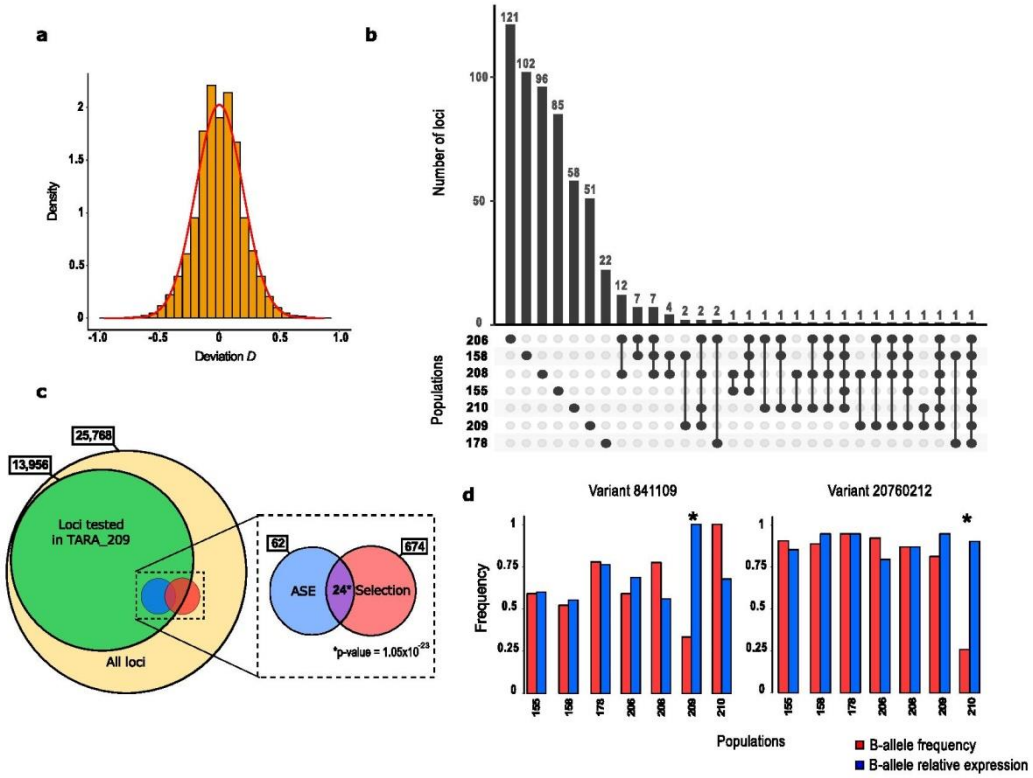


Figure 3: Population allele-specific expression detection and link with natural selection. **a,**The deviation *D* distribution in TARA_209. The red line corresponds to the Gaussian distribution estimated from the data. **b,** Upset plot of the ASE detection in the seven populations. Each bar of the 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.

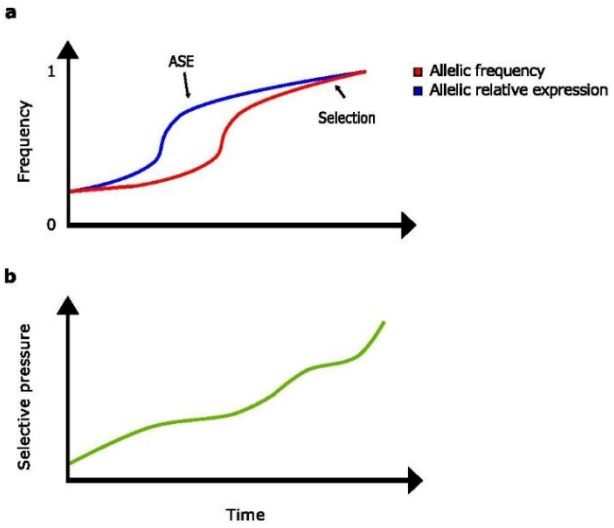


Figure 4: From allele-specific expression to natural selection. **a,** Evolution of allele frequency and allele relative expression over time. **b,** Evolution of selective pressure over time

Discussion

The purpose of this thesis was to better understand, from a molecular and anatomical point of view, the biology of the cosmopolitan and abundant *Oithona* genus to better explain its ecological success and behaviour and to finally propose it as a model for the study of small zooplankton. For this, we analysed its anatomy (chapter 1) and constructed the genome of *Otihona nana* (chapter 2). From the *Tara* data, we carried out a population genomic analysis and identified the loci under selection (chapter 2), identified an explosion of the LNR domain proteins in *Oithona* (chapter 2), constructed transcriptomes at different development stages, and identified genes specific to each stage (chapter 3), inferred the role of LDPGs in males, in particular, in its sacrificial behaviour to find a reproductive partner (chapter 3). Moreover, we observed a link between population-scale ASE and natural selection and highlighted the importance of the nervous system and neurotransmitters in the evolution of *O. similis* Arctic populations (Chapter 4). In this last part of my thesis manuscript, I will synthesise the results, but also highlight some of the failures as well as the experiences that we wish to perform in the near future.

Oithona molecular data

To build Illumina libraries, the amount of DNA extracted from a single *O. nana* individual was not sufficient (<5ng) and it was necessary to pool several individuals (more than 2,000) to finally build its genome. Therefore, the genome assembly we proposed is a mosaic of individuals from the same population of the Little Bay of Toulon. The annotated genome has a size of 85 Mb and is about 90% complete, with more than 15,000 genes. It is the first available genome for the *Oithona* genus, but also the first for the *Cyclopoida* order.

On the other hand, the extraction of mRNA from a single environmental individual retrieves enough mRNA for Illumina libraries construction. Through my thesis, transcriptomes from individuals at different stages of development (egg, larva, juvenile, adults male and female) are now also available. Since the genes specific of different stages were identified, it is now possible to use these genes to perform molecular monitoring of the population dynamic and thus to estimate the proportion

of *O. nana* individuals at each different stage in environmental samples. Moreover, as *O. nana* is known to be particularly robust and present in polluted waters (Richard and Jamet 2001); this species can be a biomarker of the poor quality of the environment when its population increases to the detriment of other species.

Our initial purpose was also to build the genome of *O. similis* and *O. atlantica* and make a comparative genomic analysis between Oithona from different ecosystems (coastal, cold and tropical waters). However, the *O. similis* genome is much larger than expected: between three and five gigabases, against 85Mb for *O. nana*. During the first DNA extraction tests performed on eggs, *O. nana* genomic DNA was extracted, but the genome construction was not possible because of egg genome size (>1Gb based on k-mer spectrum analysis). This genome size difference between *O. nana* germinal and somatic cells has already been observed in other cyclopoid copepods like freshwater Cyclops and can be explained by a phenomenon called chromatin reduction (Grishanin 2014). An *O. similis* genome draft assembly was built, according to its size (>3Gb), it seems to have lost or not acquired the chromatin reduction ability. The first analysis of this draft showed a dramatic incompleteness (less than 50%) and a high fragmentation which make it unusable for gene prediction.

However, as for *O. nana*, I constructed *O. similis* transcriptomes at different developmental stages. Unfortunately, I could not isolate female carrying egg bags from the Toulon Large Bay. Given our analyses, we can even say that I could not isolate female *O. similis*. This is problematic because it is the most accessible stage to identify. On six individuals I isolated for mRNA extraction and identified as *O. similis* female, none were *bona fide O. similis* (based on posterior molecular analysis of the transcriptomes, see Appendix 12). This can be explained by one of these two reasons: either I am a terrible taxonomist having misidentified the species-specific characters, or the criteria to designate *O. similis* correspond to several species at the molecular level with very few differences at the phenotypic level (binocular lens). To answer this question, I should ask other taxonomists to identify female *O. similis* individuals, and then I could build a reliable female transcriptome.

From the *Tara* metagenomic data and the available copepods 28S data, we were able to confirm molecularly the presence of the genus *Oithona* in all *Tara* Oceans expedition sampling stations (Appendix 4). We only observed *O; nana* in coastal sampling stations from the Northern Mediterranean Sea. While we observed *O. similis*

in several sampling stations located in the cold waters of the Atlantic, Pacific, Arctic and Antarctic oceans. In tropical waters, we observed either *O. atlantica, O. plumifera* or *O. frigida* that we have grouped in a single operational taxonomic unit (OTU) because of their molecular similarity (>98% average nucleic identify on the variable part of the 28S). Other unknown species of *Oithona* were detected but remain undefined because of lack of molecular data. For this reason, a molecular markers atlas obtained from individuals well identified by taxonomists must be constructed.

Characterisation of LDPGs structure and function

By performing a comparative analysis with other available copepod genomes, we observed an explosion of the Lin-12 Notch Repeat (LNR) domain in the *O. nana* deduced proteome (Appendix 8). It seems that this explosion is also visible in *O. similis* transcriptomes, with less amplitude (about twenty LDPGs, data not shown). Since the genome is incomplete, we were not able to do an exhaustive analysis, and so I will not discuss this point further.

However, in the *O. nana* LDPGs, we observed a large diversity of structures with new domain associations not described in the literature, and a strong evolution of the sequences coding the domain. In the Mediterranean Sea, 7% of LDPGs are under selection in the *O. nana* populations, including three single-nucleotide mutations within LNR domains. We also observed 24% of the 75 LDPGs over-expressed in *O. nana* males from the Toulon Little Bay. However, in the literature, only three proteins possessing an LNR domain are described: Notch, from which it takes its name, Stealth protein and Pappalysin.

Notch protein contains three LNR domains in tandem, playing a role in cleavage site protection in the 'normal' configuration and so negatively regulates the Notch pathway (Sanchez-Irizarry et al. 2004). Stealth protein is present in some eukaryotes and few prokaryotes (Sperisen et al. 2005). The protein seems to play a role in the innate immune system, but the role of its LNR domain is still unknown. Pappalysin (Pappa) contains three LNR domains: two in N-terminal, and one in C-terminal (Monget and Oxvig 2016). By mean of this LNR, Pappa can form a homodimer, and then activate its metallopeptidase domain. In humans, Pappa can cleave IGFPB 4 and 5 (Insulin-like Growth Factor Binding Protein), to positively regulate the amount of

free insulin-like growth factor (IGF) (Monget and Oxvig 2016). In some crustaceans, IAG (Insulin-like Androgenic Gland) is the hormone responsible for differentiation into male and is regulated by an IGFBP (Rosen et al. 2013; Ventura and Sagi 2012). Our goal at the beginning of my thesis was to find an IAG in *O. nana* proteome and to characterise the LDPs by detecting their protein partners by PPI with the *a priori* that one of the LDPs carrying a protease domain could interact with *O. nana* IGF. Initially, I also planned to identify the genes that are co-expressed using the WCGNA method, to determine the LDPGs functions through gene-silencing technics and to build the evolutionary tree of all the *O nana* LNR domains by phylogeny.

The Yeast two-hybrid (Y2H) experiment set up (Appendix 10), the false-positive verifications and the interpretation were laborious, but we finally obtained robust results (chapter 3). We detected the formation of LDP-containing protein complexes as well as interactions with proteases, extracellular matrix (ECM) proteins and proteins known to be involved in neurogenesis.

After a complete analysis of the genes co-expression by the WGCNA tool, we obtained 179 modules comprising at least 15 genes. Among them, 33 were significantly associated to a developmental stage, from which eleven were specific to males. One of these male modules, called M3, is composed of 443 genes and is enriched in LDPGs, under selection genes, and trypsin. Another, called M1, was composed of 1,199 genes that were enriched in neuroactive receptors and ion transporters. I searched for regulatory sequences in the upstream regions of the M3 module genes and found a conserved sequence of nine nucleotides, possibly regulating nine genes including one LDPG and a protease inhibitor; but no transcription factor associated with this sequence has been found in the Insect transcription factor binding profile databases (Khan et al. 2018). This analysis was not discussed in chapter 3 but will probably be resumed.

For the gene silencing set up, we started from a protocol developed for the harpacticoid *Tigriopus californicus* (Barreto, Schoville, and Burton 2015), a copepod model for ecotoxicology studies, and we adapted it for *O. nana*. According to the protocol, we used the electroporation method to insert the dsRNA. Due to both our technical inability to keep an *O. nana* culture for more than two weeks and the distance between the sampling area and the Evry laboratory, we had intense difficulties developing a proper protocol. We also had a lot of loss during each electrical pulse. In

the best experimentations, we observed more than one-third of the living individuals after the ten-pulse sequence and two dead individuals whose labelled dsRNA penetrated the chitinous exoskeleton, but we did not manage to get living individuals with labelled dsRNA inside. This experiment will probably be set up directly in the laboratory of University of Toulon, France where we regularly capture *O. nana*.

For the LNR domains evolutionary tree, because of their small size and the high divergence of the domain in *O. nana*, we used the nucleotide sequences instead of the protein one to keep more information. We tried to perform a phylogeny using the protein sequences, but we did not observe a strong signal. Based on nucleic sequences, the phylogenetic tree had only 17% of the nodes with a strong support, and 27 branch splits corresponding to tandem duplication involving 15 LDPGs (20% of the total LDPGs), including Notch.

Role of LDPGs

From the differential expression and Y2H results, our main hypotheses for the LDPGs function is the modulation, in males, (i) of the autolysis and (ii) of the neurogenesis. The autolysis or “auto-digestion” may allow the male to search for a female for extended time periods without requiring food intake by providing the necessary energy and amino acids to its metabolism for glutamate synthesis and nervous system development. The neurogenesis in male may participate to their high motility and ability to catch females. To confirm these two hypotheses, we are currently testing fluorescence microscopy protocols for the nervous system and connective tissue specific labelling on *Oithona* males and females. According to our hypotheses, we expect males to present a more developed nervous system with more axons, dendrites and synapses, and also a tissue loss notably for the connective tissue. This microscopy analysis is currently still in progress. If we succeed in making the observations, the experiment will be eventually added to the article of chapter 3, currently in preprint. If we do not succeed in the protocol set up, the study will be submitted in its current state. If our analyses do not confirm our hypotheses, we will have to change the discussion of the article before further submission.

Annexes

Appendix 1: French abstract of "Chitin Distribution in the *Oithona* Digestive and Reproductive Systems Revealed by Fluorescence Microscopy."

Les copépodes sont les animaux les plus abondants sur Terre, et parmi eux, le genre *Oithona* est décrit comme l'un des plus nombreux. Ils jouent un rôle majeur dans le réseau trophique et les cycles biogéochimiques marins, particulièrement grâce à l'excrétion de granulés fécaux enrobés de chitine. Alors que la morphologie de nombreuses espèces d*'Oithona* est bien connue, la connaissance de l'anatomie du copépode et de sa constitution en chitine est toujours limitée. Pour répondre à ce problème, des individus d'*Oithona nana* et d'*O. similis* ont été colorés par Wheat Germ Agglutinin-Fluorescein IsoThioCynata (WGA-FITC) et DiAmidino-2-PhenylIndole (DAPI) pour des observations au microscope à épifluorescence.

L'analyse d'image permet une nouvelle description de l'organisation et de la constitution en chitine des systèmes digestifs et reproducteurs des mâles et femelles *Oithona*. Des microfibrilles de chitine ont été observées tout le long du système digestif : de l'estomac jusqu'à l'intestin, avec une plus forte concentration au niveau de la membrane péritrophique de la partie supérieure de l'intestin. De nombreux rétrécissements de l'intestin ont été observés, pouvant être impliqués dans la formation et le transport des granulés fécaux. Des amas chitineux ont été observés comme composantes majeures des canaux, vésicules et réceptacles séminaux. Des structures sphériques et chitineuses ont également été observées à l'extrémité des pattes de certains individus et n'avaient jamais été décrites auparavant. La fonction de cet organe reste inconnue, mais il pourrait être impliqué dans la bioluminescence ou dans l'osmorégulation.

Le protocole de coloration rapide proposé dans cette étude permet une nouvelle vision de l'anatomie d'*Oithona*, et met en lumière le rôle de la chitine dans la digestion et la reproduction. Cette méthode pourrait être appliquée à une large palette de copépodes dans le but d'effectuer des identifications taxonomiques ainsi que des analyses comparatives de leurs anatomies.

Appendix 2: **French abstract of "New insights into global biogeography, population structure and natural selection from the genome of the epipelagic copepod *Oithona nana*"**

Dans la zone épipélagique des océans, le genre *Oithona* est décrit comme l'un des plus abondants et des plus répandus, jouant un rôle essentiel dans le réseau trophique. Malgré son importance écologique, peu de connaissances existent sur la génomique d'*Oithona* et des cyclopoïdes. Par conséquent, nous avons séquencé, assemblé, et annoté le génome d'*Oithona nana*.

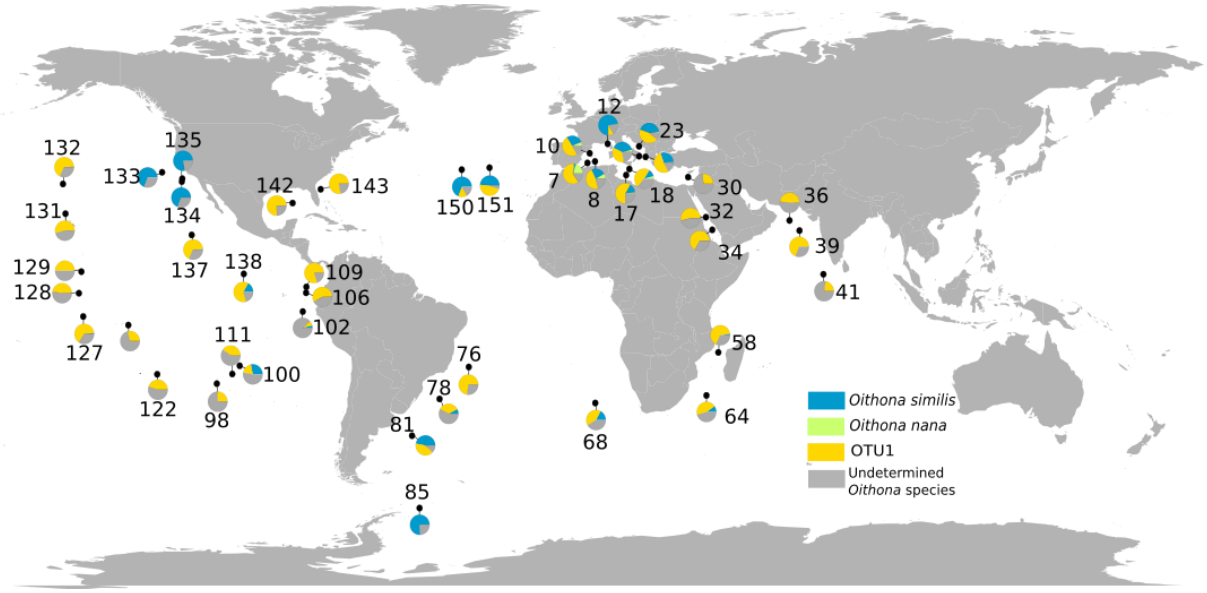
L'analyse comparative avec les autres génomes de copépodes disponibles met en lumière l'explosion de gènes liés à la réponse au stress, à la différenciation cellulaire et au développement, dont de nombreux gènes codant pour le domaine protéique Lin-12 Notch repeat (LNR). La biogéographie d'*Oithona*, reposant sur les séquences 28S et les lectures métagénomiques des expéditions Tara Oceans, montre la présence de *O. nana* principalement en mer Méditerranée (MS) et confirme la présence d'*O. similis* dans les eaux froides et tempérées (non-tropicales). L'analyse génétique de la population d'*O. nana* dans le nord de la mer Méditerranée, intégrant les données métatranscriptomiques de Tara Oceans et le génome de *O. nana*, permet d'identifier la structure génétique des populations des différents bassins méditerranéens. De plus, 20 loci ont été détectés sous sélection positive, dont quatre mutations faux-sens et huit mutations synonymes, avec des motifs de balayage sélectif. Une de ces mutations faux-sens fut localisée au sein d'un gène spécifique aux mâles, dans une région codant pour un domaine protéique LNR. La variation de la fréquence allélique, en prenant en compte les courants de la MS, montre la présence d'un cline génomique entre *O. nana* et une autre espèce indéfinie d'*Oithona*, probablement importée par les eaux de l'océan Atlantique.

Cette étude fournit une nouvelle approche et de nouveaux résultats (structures et pression naturelle) sur la génomique des populations abondantes de zooplancton grâce à l'intégration de données métagénomiques et océanographiques.

Appendix 3 : **Crustacean genome and genes metrics** (from Madoui et al. 2017)

	D.pulex v.3	***E. affinis v.0.5.3***	***O. nana* v.1.0**	***T. californicus v.1.0***
annotator	ab initio + manual	Maker 2	Gmove	Maker 2.3.1
source	NCBI (gff)	NCBI (gff)	Genoscope (gff)	NCBI (gff)
genome size (Mpb)	197	494	85	178
# genes	30 616	29 783	15 359	14 086
gene size (mean)	2 127	8 339	2 477	4 760
gene size (median)	1 289	5 853	1 444	2 856
gene size (maximal)	114 500	113 400	77 560	98 780
gene size (minimal)	60	96	105	180
# monoexonics genes	5 037	853	3 120	2 046
% monoexonic	16.45	2.86	20.31	14.53
# exon	144 569	188 898	50 809	66 651
mean exon/gene	4.72	6.34	3.31	4.73
median exon/gene	4	5	3	4
exon size (mean)	227	192	401	391
exon size (median)	154	91	261	225
# max exon/gene	83	96	37	65
sum exons sizes	32 758 251	36 354 301	20 351 900	26 086 403
exon size (maximal)	12 780	15 360	15 940	18 110
exon size (minimal)	1	1	1	3
# intron	113 953	159 115	35 450	52 565
mean intron/gene	3.72	5.34	2.31	3.73
median intron/gene	3	4	2	3
intron size (mean)	284	1 332	489	779
intron size (median)	76	666	106	140
# max intron/gene	82	95	36	64
sum introns sizes	32 373 539	212 010 761	17 322 537	40 965 296
intron size (maximal)	48 490	38 270	49 920	60 510
intron size (minimal)	1	4	9	1

Appendix 4: **Global biogeography of *Oithona* species in the DCM water layer** (from Madoui et al. 2017)



Appendix 5: **French abstract of "Discovering millions of plankton genomic markers from the Atlantic Ocean and the Mediterranean Sea."**

La comparaison de la diversité moléculaire de toutes les populations planctoniques présentes dans des colonnes d'eau géographiquement éloignées pourrait permettre une vue holistique de la connectivité, de l'isolation et de l'adaptation des organismes dans un environnement marin. Dans ce contexte, une détection à large échelle et l'analyse des mutations génomiques, directement dans les données métagénomiques du plancton, apparaît comme une stratégie pertinente pour l'identification des structures génétiques et la détection des gènes sous pression de sélection naturelle.

Dans cette étude, nous avons utilisé DiscoSNP++, un détecteur de variance sans utilisation de référence, pour détecter les mutations dans les données métagénomiques de grande échelle. Nous avons également estimé sa précision chez le copépode *Oithona nana*, en termes de détection de variations, d'estimation de fréquence allélique et de génomique des populations, en le comparant aux autres méthodes de l'état de l'art. DiscoSNP++ détecte des variations qui mènent à des conclusions similaires, pour la structure génétique et l'identification de loci sous pression de sélection naturelle, aux autres méthodes. DiscoSNP++ fut appliqué sur 120 échantillons métagénomiques de quatre fractions de taille, dont des procaryotes, des protistes et du zooplanctons échantillonnés dans 39 stations *Tara* Oceans situées en océan Atlantique et en mer Méditerranée, produisant ainsi un nouveau lot de marqueurs génomiques marins contenant plus de 19 millions de mutations.

Cette nouvelle ressource génomique peut être utilisée par la communauté scientifique pour relocaliser ces marqueurs sur les génomes ou transcriptomes des espèces planctoniques d'intérêt. Cette ressource sera mise à jour avec de nouvelles expéditions, et l'augmentation des données métagénomiques.

Appendix 6: **Annotation of loci detected under natural selection** (from Arif et al. 2019)**.**

Scaffold	Position	Ref	Alt	Gene model	Variant	AA modification	Annotation
1	218542	C	T		Upstream gene		
1	467134	C	A		Upstream gene		
2	966381	G	C	GSONAT00000516001	NA		Serine-pyruvate aminotransferase
8	891235	T	C		NA		None predicted
8	891330	T	A		NA		None predicted
9	971651	A	T	GSONAT00015360001	NA		Pickpocket protein like
10	237360	T	C	GSONAT00002382001	Synonymous	p.Leu2027Leu	Dynein heavy chain
10	247395	A	G	GSONAT00002383001	Upstream gene		None predicted
10	247465	G	A	GSONAT00002383001	Synonymous	p.Phe19Phe	None predicted
10	247594	A	C	GSONAT00002383001	5 prime UTR		None predicted
10	247669	G	T	GSONAT00002383001	5 prime UTR		None predicted
15	277847	A	T	GSONAT00003212001	Downstream gene		
23	372498	T	C		Upstream gene		
25	231999	G	A	GSONAT00005206001	5 prime UTR		Fermitin family
51	153302	C	T	GSONAT00008031001	synonymous	p.Ser351Ser	Secretin-like
67	27053	T	C		Upstream gene		
68	10871	T	C		Upstream gene		
75	294967	G	A	GSONAT00009904001	Synonymous	p.Cys213Cys	PAN PAN/Appel domain
75	295105	G	A	GSONAT00009904001	Synonymous	p.Phe167Phe	PAN PAN/Appel domain
75	295475	T	C	GSONAT00009904001	Missense	p.Thr64Ala	PAN PAN/Appel domain
75	304202	C	T	GSONAT00009907001	Synonymous	p.Pro11Pro	ARL14 effector protein
75	304522	T	C	GSONAT00009907001	5 prime UTR		ARL14 effector protein
75	304956	T	G	GSONAT00009908001	Synonymous	p.Ala196Ala	None predicted
76	166168	T	C	GSONAT00009935001	Synonymous	p.Thr536Thr	Kelch-type beta propeller domain
76	166248	A	G	GSONAT00009935001	Missense	p.Glu563Gly	Kelch-type beta propeller domain
76	166428	A	G	GSONAT00009935001	Missense	p.Lys623Arg	Kelch-type beta propeller domain
86	141392	C	T	GSONAT00010501001	Intron		RasGRF2
86	141429	A	G	GSONAT00010501001	Intron		RasGRF2
86	142467	T	A	GSONAT00010501001	Intron		RasGRF2
88	109848	T	G		Upstream gene		
94	105665	G	A	GSONAT00015420001	NA		TNF-like domain superfamily
94	109199	C	T		Upstream gene		
94	113922	T	C	GSONAT00010837001	Missense	p.Val11Ala	Sugar transporter-like
102	143819	A	T	GSONAT00011159001	NA		Arylsulfatase
102	159878	G	T	GSONAT00011161001	Missense	p.Gln382Lys	FMRFamide receptor
103	208056	C	T	GSONAT00011184001	Synonymous	p.Leu412Leu	None predicted
103	210242	C	T		Downstream gene		
120	27573	A	G	GSONAT00011726001	Synonymous	p.Gly48Gly	None predicted
126	151167	G	A	GSONAT00011915001	Synonymous	p.Thr87Thr	None predicted
126	151204	A	G	GSONAT00011915001	Splice region&intron		None predicted
131	32856	G	T		Upstream gene		
140	42203	A	G	GSONAT00012258001	3 prime UTR		Innexin

169	20585	G	A	GSONAT00012792001	3 prime UTR		Glutamic rich SH3 binding domain
175	29781	T	G		Upstream gene		
196	842	T	A	GSONAT00013101001	Missense	p.Lys112Met	None predicted
212	24957	G	A	GSONAT00015370001	NA		
212	45266	A	T	GSONAT00013238001	Synonymous	p.Ala493Ala	Peroxidase
262	14087	C	T	GSONAT00013467001	Synonymous	p.Ser60Ser	None predicted
360	660	G	A	GSONAT00015450001	NA		Uncharacterized protein
408	5071	G	A	GSONAT00015430001	Intron		None predicted
408	6331	C	A	GSONAT00015430001	Intron		None predicted
541	2367	C	T	GSONAT00013822001	Missense	p.Glu1091Lys	LNR domain
556	3426	C	T	GSONAT00015380001	NA		LNR domain/Kelch-type domain
556	3591	C	T	GSONAT00015380001	NA		LNR domain/Kelch-type domain
1090	1757	C	G	GSONAT00014235001	Intron		None predicted
1239	1585	C	T		Intergenic		
1239	606	T	C	GSONAT00015400001	NA		LNR domain/Metalopeptidase
1365	1534	C	G	GSONAT00014370001	Missense	p.Glu699Asp	Laminin subunit
1365	2873	C	A	GSONAT00014370001	Missense	p.Asp337Tyr	Laminin subunit
1429	1120	A	G		Upstream gene		
1604	1098	G	A	GSONAT00014466001	Synonymous	p.Ile487Ile	FAD/NAD(P)-binding domain
1807	2980	G	C	GSONAT00015410001	NA		LNR domain
1819	1461	C	G	GSONAT00014573001	Missense	p.Pro171Ala	Kelch-type beta propeller domain
1819	2618	C	T	GSONAT00014573001	Splice region&intron		Kelch-type beta propeller domain
1819	2868	T	C	GSONAT00014573001	Missense	p.Tyr483His	Kelch-type beta propeller domain
1886	3092	A	T		Downstream gene		
2017	1371	G	A		Intergenic		
2017	1736	G	T		Intergenic		
2023	2729	A	G		Upstream gene		
2066	3045	T	A		Intergenic		
2085	1564	T	C	GSONAT00014698001	Missense	p.His392Arg	LNR domain
2085	1603	C	G	GSONAT00014698001	Missense	p.Trp379Ser	LNR domain
2085	2429	T	G	GSONAT00014698001	Missense	p.Thr104Pro	LNR domain
2487	2406	T	G	GSONAT00015390001	NA		Kelch-type beta propeller domain
3137	1880	G	T	GSONAT00015041001	Missense	p.Phe109Leu	Gamma-glutamyltranspeptidase
3137	1912	T	C	GSONAT00015041001	Missense	p.Lys99Glu	Gamma-glutamyltranspeptidase
3250	232	C	T		Intergenic		
3397	1439	T	C		Intergenic		
3651	2020	G	A		Intergenic		

Appendix 7: **French abstract of "Proteolysis and neurogenesis modulated by LNR domain proteins explosion support male differentiation in the crustacean *Oithona nana*"**

Les copépodes sont les animaux les plus abondants sur Terre, et jouent un rôle essentiel dans le réseau trophique et les cycles biogéochimiques marins. Le genre *Oithona* est décrit comme ayant la plus grande densité en termes de nombre d'individus, comme le copépode le plus cosmopolite et itéropare. Le paradoxe du mâle *Oithona* l'oblige à alterner entre des phases d'alimentation (immobile) et des phases de recherche de partenaires de reproduction (mobile). Comme les mécanismes moléculaires de ce compromis sont inconnus, nous avons étudié, à une échelle moléculaire, ce dimorphisme sexuel, en nous fondant sur des analyses génomiques, transcriptomiques et d'interactions protéine-protéine.

Un système ZW de détermination sexuelle est prédit chez *O. nana*. Une série temporelle de quinze ans dans la petite rade de Toulon a montré un sex-*ratio* biaisé en faveur des femelles (rapport mâle/femelle <0,15 ± 0,11), mettant en évidence une mortalité plus élevée chez les mâles. L'analyse d'expression différentielle aux cinq stades de développement montre un enrichissement en gène codant pour des protéines contenant des domaines LNR (LDPG) chez le mâle. Le mâle montre également un enrichissement en transcrits impliqués dans la protéolyse, la formation de nouveaux axones et dendrites, l'assemblage et le fonctionnement de la synapse ainsi que la conversion d'acides aminés en glutamate, un neurotransmetteur excitateur. De plus, plusieurs gènes, spécifiquement régulés négativement chez le mâle, sont impliqués dans l'augmentation de la prise alimentaire et la modulation de la digestion. La formation de complexes de LDP a été détectée par la technique du double hybride, avec des interactions impliquant des protéases, des protéines de la matrice extracellulaire et des protéines associées à la neurogenèse.

L'ensemble de ces résultats semble montrer que l'hypermobilité du mâle serait possible grâce à une autolyse modulée par les LDPs qui libèrent des acides aminés, permettant leur conversion en glutamate et ainsi, le développement du système nerveux. Cet effet délétère renforce l'hypothèse, à l'échelle moléculaire, d'un comportement sacrificiel chez le mâle *O. nana*.

Appendix 8: **Structure and localisation of the *Oithona nana* LDPs.**

e, i and *m* correspond to extracellular, intracellular and membranous respectively.

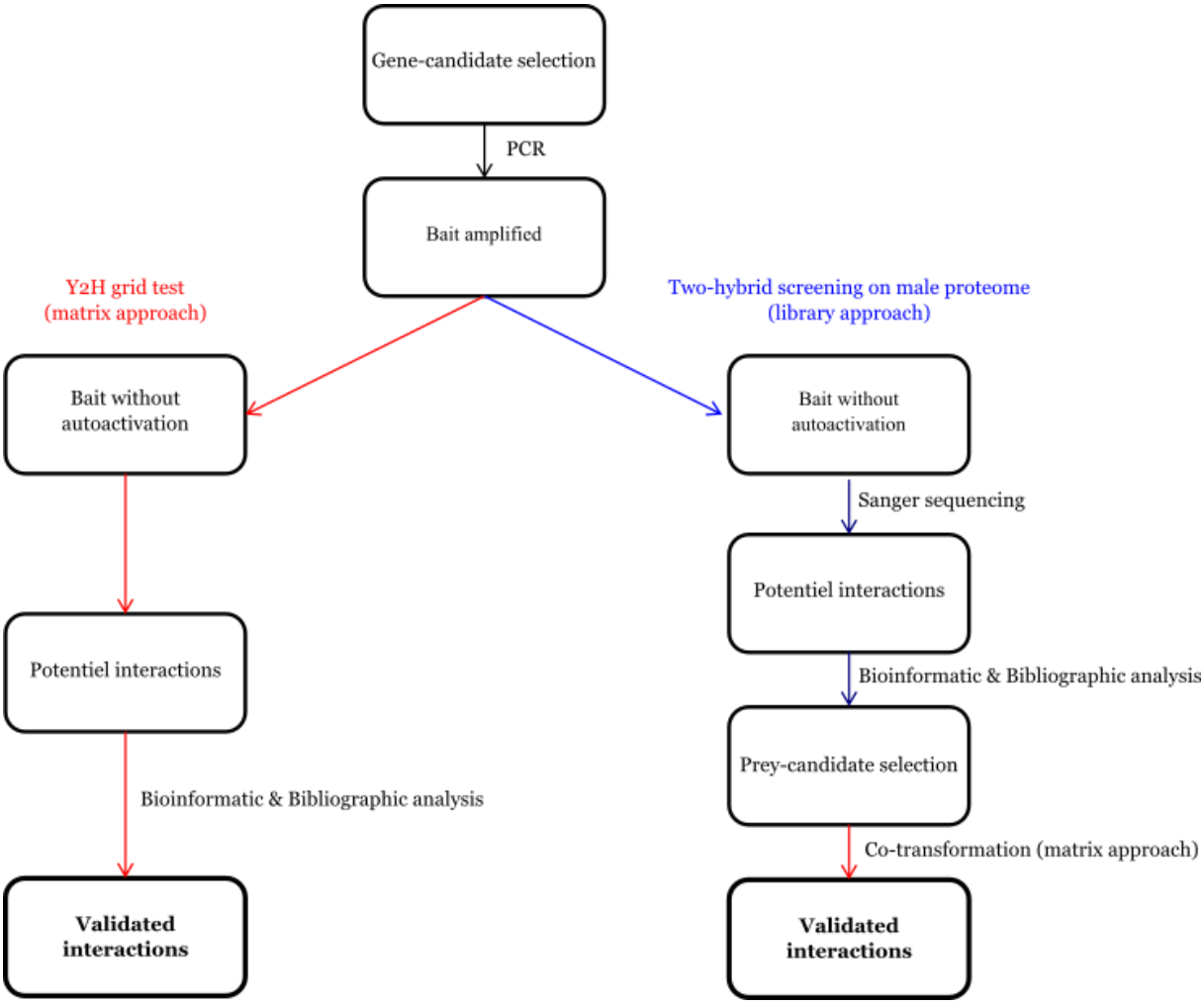
Protein name	Protein size (aa)	#Total LNR domain (#canonical LNR - #LNR-like)	Other InterPro domains	Localisation
100:211670..212713	328	3 (1-2)	-	e
11.480.1:721370..723090	531	1 (0-1)	Metallo-peptidase family M12	e
12.169.2:773247..776065	808	5 (5-0)	Trypsin	e
12.175.1	386	1 (0-1)	Trypsin	e
12.182.1	445	2 (0-2)	Trypsin	e
12.185.1	357	1 (0-1)	Trypsin	e
123:42215..42819	181	2 (0-2)	-	e
1239:190..1043	228	2 (0-2)	-	e
14.373.1:814017..816142	226	2 (0-2)	-	e
1531:53..2073	222	3 (0-3)	-	e
1646:48..560	147	1 (0-1)	-	e
174:49279..49938	219	3 (0-3)	-	e
1781:78..483	113	1 (1-0)	Metallo-peptidase family M12B	e
18.220.1:802020..803354	420	1 (0-1)	Trypsin	e
188.202.1:42571..49247	547	2 (2-0)	-	e
19.334.1	365	4 (1-3)	-	e
2.84.1	195	2 (1-1)	Trypsin	e
20.537.1	353	2 (1-1)	-	e
2085.199.1	687	6 (0-6)	-	e
21.77.1:327349..328866	458	2 (1-1)	Trypsin	e
2201:1341..2883	426	3 (0-3)	-	e
On_LDP1	378	2 (1-1)	Trypsin	e
27:531099..531903	248	2 (0-2)	-	e
32.270.1:20364..22049	561	2 (0-2)	-	e
3256:80..502	140	1 (0-1)	-	e
37:34427..34747	107	1 (0-1)	-	e
3704.33.1	266	1 (0-1)	Metallo-peptidase family M12	e
3765:1004..2202	383	1 (0-1)	-	e
38:505765..507559	438	5 (0-5)	-	e
3888:228..1050	202	2 (1-1)	-	e
4.222.1	525	2 (0-2)	Metallo-peptidase family M12	e
4.25.1	90	1 (0-1)	-	e
41:85441..86498	276	2 (0-2)	-	e
42.291.1	122	1 (0-1)	-	e
45.412.1	262	1 (0-1)	Trypsin	e
45.424.1	139	1 (0-1)	-	e
4546:128..817	229	3 (0-3)	-	e
48.267.1:6982..8492	461	2 (0-2)	Trypsin	e
48.344.1	525	2 (1-1)	Metallo-peptidase family M12	e
54.120.2:255915..258850	940	7 (0-7)	-	e
55:11070..12199	357	4 (0-4)	-	e
57.114.1	159	2 (1-1)	-	e
6.168.1:380139..381062	227	1 (0-1)	Metallo-peptidase family M12B	e
68.40.1	493	1 (0-1)	Metallo-peptidase family M12	e
821:3497..4457	291	6 (0-6)	-	e
On_LDP2	116	1 (0-1)	-	e
92.33.1:248003..248641	213	2 (0-2)	-	e
98.272.1	74	1 (1-0)	-	e
99.319.1:158136..159228	344	1 (0-1)	Trypsin	e
12.180.1	489	2 (0-2)	Trypsin	i
126.234.1	823	2 (0-2)	-	i

14:255249..255707	252	1 (0-1)	-	i
15.286.1	207	2 (0-2)	-	i
1807:1535..3342	517	1 (0-1)	-	i
188:49497..50375	177	1 (1-0)	-	i
1921.111.1	137	1 (0-1)	-	i
2.15.1:36291..38976	767	13 (0-13)	-	i
2.380.1:1294082..1295071	329	2 (0-2)	-	i
20.536.3	1083	1 (0-1)	Kelch motif	i
2184:300..2200	310	1 (0-1)	Ankyrin repeat-containing domain superfamily	i
2277:24..1778	518	8 (0-8)	-	i
23.148.2	460	2 (0-2)	Metallo-peptidase family M12	i
243.135.1	316	2 (0-2)	-	i
284.43.1	132	1 (0-1)	-	i
468:255..2448	602	1 (0-1)	PAN/Apple domain ; Kelch motif	i
541.168.1	1564	7 (3-4)	-	i
556:85..5023	1494	6 (0-6)	TSP1 ; Kelch-type_b-tropeller	i
60:181372..182580	402	1 (0-1)	-	i
94.274.1:231104..233771	867	4 (0-4)	Lectin	i
147.226.2	266	3 (0-3)	-	m
36:371592..372773	336	1 (0-1)	-	m
On_Notch	2147	3(3-0)	Notch protein	m
59:239420..240555	358	2 (0-2)	-	m
76.32.1:162363..164050	435	4 (0-4)	Thrombospondin type-1 (TSP1) repeat superfamily	m
96.130.1:322..1331	262	2 (0-2)	Lectin	m

Appendix 9: **Functional annotation of *O. nana* genes over-expressed in male.**

Protein functional groups	Protein	Number of male-specific genes
Neuropeptide and Hormone metabolism	Allatostatin precursor protein	1
	Carboxypeptidase E	1
	Furin-like protease	1
	Neuroendocrine convertase 2	1
	Peptidyl-alpha-hydroxyglycine alpha-amidating lyase	1
	Peptidylglycine alpha-hydroxylating monooxygenase-like	1
	ITG-like peptide	1
	Sulfoacetaldehyde acetyltransferase	1
	Total	***8***
Neuropeptide and hormone transport and release	Bestrophin 1	1
	Excitatory amino acid transporter 1	1
	Multiple C2 and transmembrane domain-containing protein	1
	Sodium- and chloride-dependent GABA transporter 1-like	1
	Synaptic vesicular amine transporter	1
	Synaptobrevin	1
	Synaptotagmin	1
	Synaptosomal-associated protein 25-like	1
	Neuronal calcium sensor-like	1
	Total	***9***
Neuropeptide and hormone receptors	FMRFamide receptor	6
	Muscarinic acetylcholine receptor	1
	GABA receptor subunit	2
	Glycine receptor subunit	2
	Ionotropic glutamate receptor subunit	2
	Capa receptor	1
	Octopamine receptor subunit	1
	Corticotropin hormone receptor	1
	Acetylcholine receptor subunit	1
	Total	***17***
Neuron polarization Current propagation and ion channel	Kv channel-interacting protein 1-like	2
	Potassium voltage-gated channel, Shab-related subfamily	2
	Transient receptor potential protein	1
	Two pore potassium channel protein sup-9	1
	Total	***6***
Neuron development Synapse assembly	Beta-1,4-glucuronyltransferase 1	1
	Microtubule-associated protein futsch-like	1
	Serine/threonine-protein kinase dyf-5 ou ICK	1
	Neural/ectodermal development factor IMP L2-like	1
	zwei Ig domain protein	2
	copine-like	1
	synaptogenesis syg-2	1
	Total	***8***
Total		***48***

Appendix 10: **Experimental design of the protein-protein interaction (PPI) analysis.**



Appendix 11: **French abstract of "Linking Allele-Specific Expression and Natural Selection in Wild Populations."**

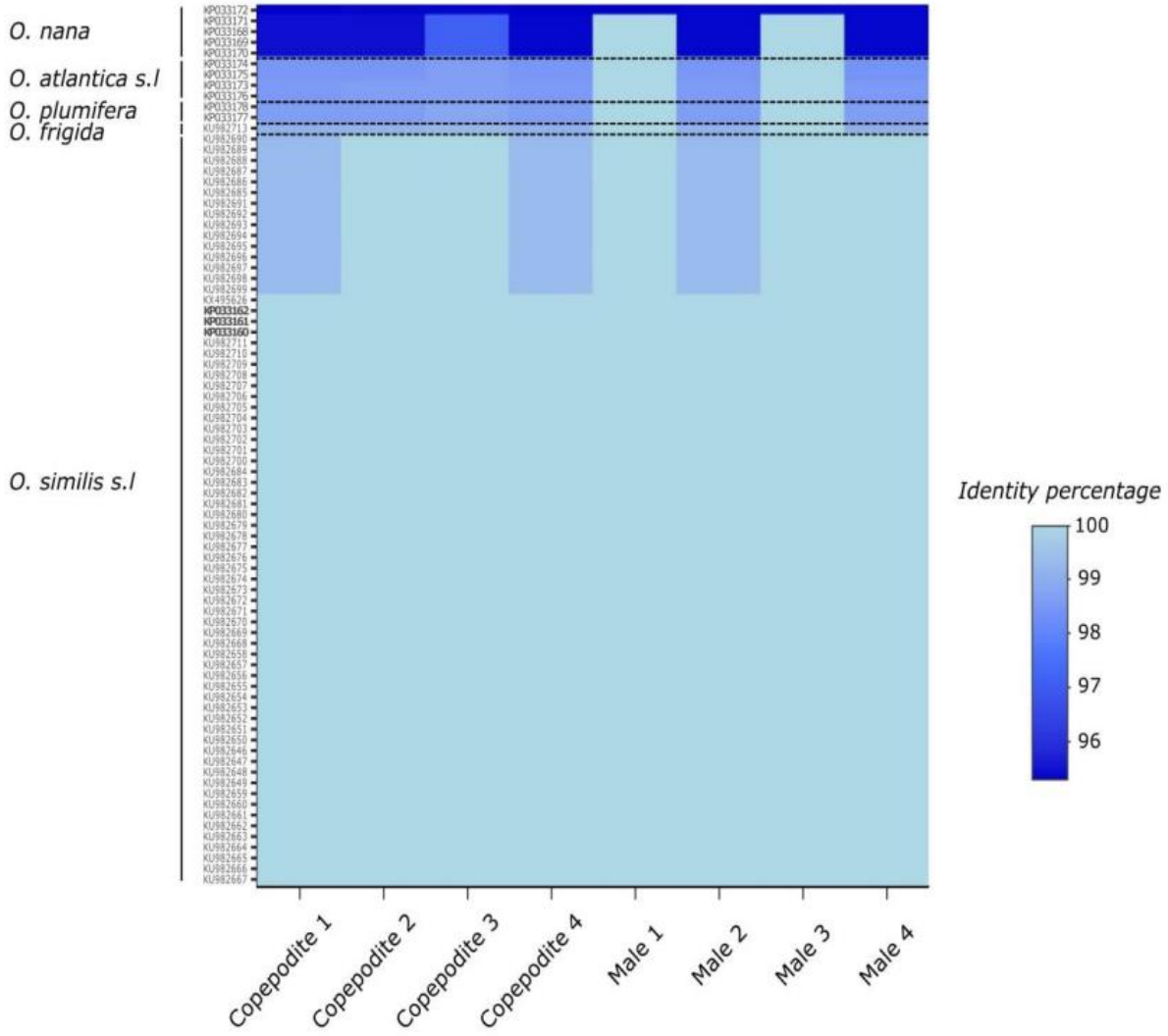
Les mécanismes de l'expression allèle-spécifique (ASE) sont grandement étudiés, aux niveaux cellulaire, tissulaire ou de l'organisme entier. Toutefois, les ASE au niveau d'une population et leur impact évolutif n'ont toujours pas été observés.

Dans cette étude, nous faisons l'hypothèse d'un potentiel lien entre ASE et sélection naturelle chez le copépode cosmopolite *Oithona similis*. Nous avons combiné les données métagénomiques et métatranscriptomiques issues de sept populations sauvages d'*O. similis* capturées durant les expéditions *Tara Oceans*. Nous avons détecté 587 variants nucléotidiques uniques (SNVs) sous ASE, et un nombre significatif (152 SNVs) sous ASE dans au moins une population, et sous sélection naturelle dans toutes les populations. Parmi les gènes sous ASE et pression de sélection, certains jouent un rôle dans le métabolisme du glutamate, ou dans les récepteurs de glycine et/ou GABA.

Ceci constitue une première preuve que la sélection naturelle et l'expression allèle-spécifique ciblent plus de loci en commun que par pur hasard ; ce qui soulève de nouvelles questions quant à la nature des liens évolutifs entre ces deux mécanismes.

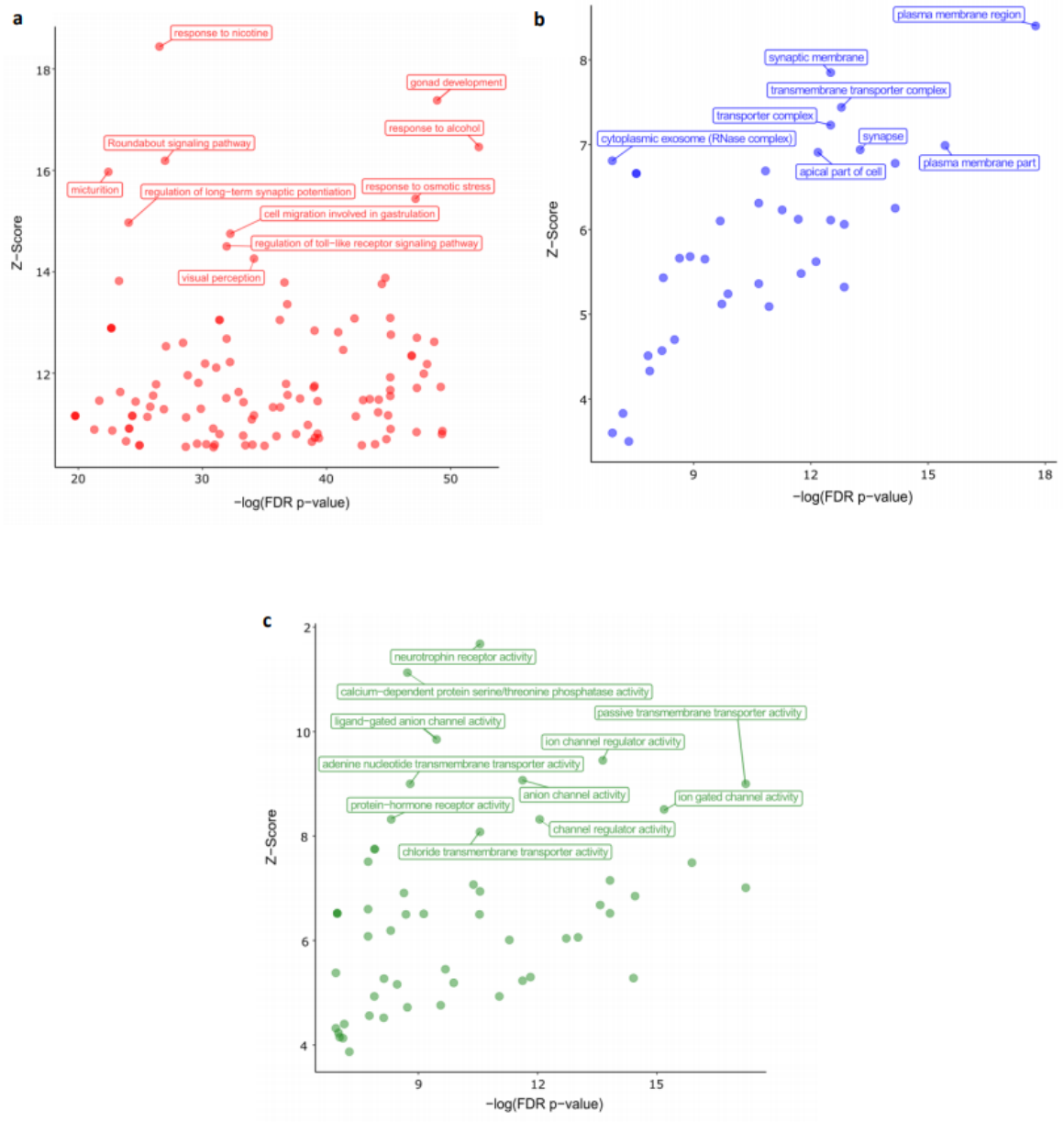
Appendix 12: **Validation of taxonomic assignation** (from Laso-Jadart et al. 2019).

In rows are represented the 82 accession numbers of *Oithona* species 28S sequences. In bold, type localities of *O. similis* as described in Cornils, Wend-Heckmann, and Held 2017. In columns are represented ribosomal read sets of the eight individuals.



Appendix 13: **Functional analysis of *Oithona similis* transcripts targeted by ASE and selection**(from Laso-Jadart et al. 2019).

The axes correspond to statistical metrics computed by dcGO Enrichment. Are highlighted the most significantly enriched terms in **a,** Biological Process GO-terms, **b,** Cellular Component GO-terms, **c,** Molecular Functions GO-terms.



Appendix 14: **Remerciments infinis à :**

----- In order of appearance -----

La meilleure famille au monde

L'éducation nationale tout particulièrement Mr Espitalier, Mr Marotte, Pr Terzian et Dr Berta.

Anastasia G., Jade S. et Manon S.

Amin M.

Marc W., Marion Gmove D., Benjamin N. et toute le RDBioSeq

Julie P. et Benoit V.

Manue P. Maria D. Muriel R. Eric

Karine L. et toute l'équipe séquençage

Janaina R., Alexey V., Sarah F. Jade L., Soheib K., Majda A. et tout le LAGE

David R., Guillaume G, David V., Laura B. et tout le LabGem

Patrick W.

Jean-Louis et Dominique J.

Adriana A. (l'une de mes quatre directrices !)

Céline O., Laurie B., Nathalie M. et toute l'équipe Interactome

Catherine S. Catherine C. et Nancy D.

Peggy P.

Dr Cornils et Dr Blanco-Bercial

Corinne C., Odette B., Emilie, Thomas et toute l'équipe Nanopore

Romuald L.-J. et Julie L.-H.

Magali B., Sophia B., Ivan D. et tous les gens du 1er

Claude S., Frank et toute l'équipe Système

Les derniers (mais non les moindres) : , Chloe, Oriane, Ombeline, Thomas B., Marion T., Nachida T., Aimeric B., Adrien J., Céline, Christophe P., Nastasia, Christine P., Adelme B., Remi P., Mathieu D., Johan, Mylène, Jonathan, Léo, Artem, Nöelie, Yohann et tous les autres kamarades du GenoPub.

Beaucoup d'animaux ont été maltraité durant cette thèse : excusez-nous !

Les bandes originales de la thèse sont disponibles sur Spotify à TheseOithona et MrSardou.

Appendix 15: Résumé en français de la thèse.

Introduction

Les copépodes sont de petits animaux aquatiques. Leur nom vient du grec ancien « kope -podos », littéralement « manche - pied », ce qui signifie pied en forme de rame. Il a été introduit par le zoologiste français Henri Milne Ewards en 1840 (WoRMS 2019a ; Milne Edwards 1840). Le nom *Copepoda* constitue une sous-classe de crustacés et appartient à l'embranchement des *Arthropoda*.

La divergence entre les copépodes et les autres taxons d'arthropodes est estimée, moléculairement, entre 395 et 495 millions d'années (Eyun 2017 ; Sanders et Lee 2010). Les fossiles de copépodes sont rares et peu diversifiées : peu d'ordres sont représentés (Huys et al. 2016). Le plus ancien fossile de copépode, non ambiguë, date du Carbonifère (environ -300 Ma) (Selden et al., 2010). Certains copépodologues ont déclaré avoir trouvé un fossile d'une mandibulaire de copépode datant du Cambrien (plus de 500 Ma avant notre ère) (Harvey et Pedder 2013). La sous-classe Copepoda est un clade-sœur avec les *Thecostraca* (*e.g.* balanes) et *Eumalacostraca* (*e.g.* crabes, crevettes) forment ensemble la superclasse *Multicrustacea* (Eyun 2017; Regier et al. 2010).

À ce jour, huit ordres sont reconnus en utilisant les données taxonomiques et moléculaires, mais il n'y a pas de consensus entre copepodologues, et trois autres ordres peuvent être mentionnées dans la littérature. Parmi les huit ordres, quatre regroupent 99% des familles connues : *Calanoida*, *Harpacticoida*, *Siphonostomatoida* et *Cyclopoida* (qui comprend l'ancien ordre *Poecilostomatoida*). La sous-classe *Copepoda* est divisée en deux superordres par le zoologiste allemand Wilhelm Giesbrecht : *Gymnoplea* (comportant les calanoïdes) et *Podoplea* (qui comprend les sept autres ordres). A ce jour, un peu plus de 14 000 espèces de copépodes sont répertoriés (WoRMS , 2019) et des dizaines de nouvelles espèces sont décrites chaque année. Avec l'exploration des sédiments marins, le total du nombre d'espèce serait d'un peu plus de 20 000. Cette forte diversité d'espèce est également accompagnée d'une explosion de formes et de tailles et de niches écologiques occupées.

Les copépodes vivant dans ou à proximité des sédiments appartiennent au benthos, du grec ancien « bathos » qui signifie profondeur. Les copépodes benthiques sont à l'origine de la diversité actuelle : ils ont probablement commencé à coloniser les colonnes d'eau entre -400 Ma et -450 Ma, et les eaux pélagiques entre -250 Ma et -300 Ma. (Bradford-Grieve 2004 ; Selden et al. 2010).

Les copépodes pélagiques appartiennent au zooplancton, venant du grec ancien « zoion – plagktos » signifiant respectivement animaux et errant. Le physiologiste allemand Victor Hensen a été le premier à donner ce nom à la description « des petits animaux marins qui ne peuvent pas nager à contre-courant ». Les copépodes planctoniques peuvent cependant migrer verticalement, en profondeur. Un tiers des espèces de copépodes connues sont des parasites ayant un large spectre d'hôtes, notamment des éponges, des mammifères, des reptiles et des poissons (Selden et al. 2010).

Les copépodes sont décrits comme les animaux les plus abondants sur Terre, suivis par les insectes et les nématodes (Humes 1994). Il est estimé qu'ils sont probablement plus nombreux que tous les autres métazoaires combinés. Dr Geoff Boxshall et Dr Rony Huys estimé que, avec la présomption d'un copépode unique par litre d'eau de mer, plus de $1.347x10^{21}$ copépodes sont présents sur la Terre. Cette abondance est également illustrée par la présence de copépodes dans tous les milieux aquatiques, dans un grand spectre de salinité, température, pH, latitude : de l'eau douce aux eaux de mer hypersalines, du glacier à la source d'eau chaude, des fausses océaniques (à 10 km de profondeur) aux montagne (jusqu'à 5 km d'altitude) (Kiørboe 2011 ; Selden et al 2010 ; Huys et Boxshall 1991). Certains copépodes ont également été observés dans des environnements atypiques comme des pneus abandonnés, des ananas, des broméliacées, des flaques d'eau, des mousses d'arbres et des réservoirs d'eau (Huys et Boxshall 1991). Pour l'anecdote, la présence de ce crustacé dans les eaux potables était un gros problème dans les communautés juives (les crustacés ne sont pas casher) et végétariennes de New York (BBC News 2004). Les copépodes ont adapté leurs stratégies de reproduction et d'alimentation à tous ces environnements, ce qui expliquent leur distribution cosmopolite.

La majorité des copépodes ont une taille comprise entre 500 µm et 5 mm. À ce jour, les plus petits copépodes connus mesurent quelques micromètres et sont des parasites des branchies du plus petit vertébré, *Paedocypris progenetica*. Les plus

grands copépodes connus (*Pennella sp.*) font jusqu'à 50 cm avec les sacs d'œufs et sont des parasites des baleines à fanons (*Balaenoptera musculus*) (Selden et al. 2010). Les copépodes sont également décrits, par rapport à leur taille, comme les plus forts et les plus rapide animaux sur Terre. En effet, les copépodes sont dix à trente fois plus forts que tout autre métazoaire mesuré et peuvent se déplacer jusqu'à un demi-mètre (environ 500 fois la longueur de leur corps) en moins d'une seconde (Kiørboe et al., 2010 ; Université technique du Danemark, 2010).

Certains copépodes peuvent également être des vecteurs de maladies humaines. Certains cyclopes d'eau douce portent la larve du ver de Guinée (*Dracuncula medinensis*), l'agent responsable de la dracunculose, une maladie débilitante (Huys et Boxshall 1991). Les copépodes abondants *Eurytemora affinis* et *Acartia tonsa* sont également connus pour être porteurs de la bactérie *Vibrio cholerae*, responsable du choléra (Rawlings, Ruiz et Colwell 2007).

Les copépodes constituent la majorité du zooplancton : dans les zones pélagiques, jusqu'à 88% en termes de biomasse et jusqu'à 99% en termes de nombre d'individus (Thompson, Dinofrio et Alder 2013). Ils consomment des microorganismes, du phytoplancton et zooplancton de petite taille et sont eux-mêmes consommés par les niveaux trophiques supérieurs, dont les larves d'espèces de poisson consommées par les humains (le cabillaud, le saumon et l'anchois). La baisse soudaine de la biomasse ou la disparition de copépodes peut avoir un impact économique pour les pays du Nord et un impact sanitaire important pour les pays où le poisson est la principale source de protéines. En mer du Nord, depuis le milieu des années 80, une augmentation de la température de l'eau a été observée (Beaugrand et al. 2003). Cette augmentation de température a entraîné une migration, vers les eaux plus froides du nord, de *Calanus finmarchicus* le copépode le plus consommé par les larves de cabillaud et qui est progressivement remplacé par *Calanus helgolandicus*. Ce dernier ayant une autre phénologie de reproduction, ce remplacement d'espèce a donné lieu à une baisse du recrutement des larves morue et donc une baisse importante de la biomasse (Beaugrand et al. 2003).

Les copépodes sont le puits de carbone marin le plus important et les transporteurs actifs d'azote et de phosphore vers les profondeurs à travers deux phénomènes. (i) Par respiration et excrétion, en particulier pour les espèces qui migrent à 200 mètres de profondeur (Kobari et al. 2008). Dans les milieux marins

froids, plusieurs espèces de calanoïdes aux derniers stades juvénile ou adulte accumulent des ressources lipidiques dans leurs sacs d'huile, migrent dans les eaux profondes et entrent en diapause, une période d'hibernation de plusieurs mois où les individus arrêtent leur développement (croissance et différenciation sexuelle) et réduisent au minimum leurs activités physiologiques (Baumgartner et Tarrant 2017). (ii) Par la « neige marine » qui désigne la chute et sédimentation des corps en décomposition et des pelotes fécales (Boyd et al. 2019). Dans l'Atlantique nord, les copépodes constituent de 10% à 25% de la pompe à carbone biologique (Jónasdóttir et al. 2015). A cause du dérèglement climatique, une réduction de la période de diapause a été observé, ce qui impacte significativement les cycles biogéochimiques, en particulier celui du carbone (Baumgartner et Tarrant 2017)

Grâce à ces mêmes processus, les copépodes sont également essentiels à la vie des bactéries marines : certaines se nourrissent des carcasses et des pelotes fécales de copépodes, tandis que d'autres communautés vivent directement à la surface ou dans l'intestin des copépodes (Tang 2005). Des différences de communauté sont observées entre les bactéries attachées aux copépodes et celles vivant librement, mais des échanges existent (De Corte et al. 2014). Ainsi, les copépodes fournissent des substances organiques et/ou sont les hôtes du bactérioplancton.

Les copépodes sont utilisés par l'humain pour la pisciculture mais également comme source d'acide gras pour les cosmétiques et les aliments (Pedersen, Vang et Olsen 2014). Certaines populations asiatiques et scandinaves consomment directement des copépodes qu'elles récoltent (un exemple au Laos dans l'étude de Kottelat 2007). Comme les insectes, les copépodes peuvent devenir une source de protéines et d'acides gras pour répondre au déclin des stocks de poissons et aux problèmes écologiques induits par l'élevage conventionnel.

Les copépodes sont également utilisés comme contrôleurs des ravageurs et des maladies. *Mesocyclops aspericornis* est décrit comme l'espèce la plus efficace en Polynésie, en Australie et dans certaines régions d'Asie pour tuer *Aedes aegypt*, le vecteur de Dengue hémorragique. Chaque individu peut tuer jusqu'à 40 larves d'*Aedes* par jour (Brown, Kay et Hendrikz 1991). Des mésocyclopes ont été introduits dans certains réservoirs d'eau de villages du Vietnam touchés par la dengue. Un an après l'introduction du copépode, *A. aegypt* a disparu (Marten 2001).

Oithona : vierge des vagues

Le genre *Oithona* a été décrit pour la première fois par le zoologiste écossais William Baird en 1843 dans la revue « The Zoologist » (Baird 1843; WoRMS 2019b). Il enquêtait sur les « insectes » marins responsables de la luminescence de l'eau et faisait les premières descriptions d'*Oithona* en observant *O. plumifera* et *O. splendens* (Zamora Terol 2013 ; Baird 1843). Le nom vient du livre « Les Poèmes d'Ossian » de James Macpherson et se compose de « oi - thona », ce qui signifie respectivement vierge et vague en gaélique (Zamora Terol 2013). À ce jour, 48 espèces d'*Oithona* sont répertoriées (WoRMS 2019b) dont *O. nana* Giesbrecht 1892, *O. frigida*, Giesbrecht 1902, *O. atlantica* Farran 1908 et *O. davisae* Ferrari et Orsi 1984.

Jusqu'à la fin du XX^{e} siècle, le filet conventionnel utilisé pour l'échantillonnage du zooplancton avait un maillage de 200 µm (Gallienne et Robins 2001). En raison de l'utilisation de ce filet à grande maille, la fraction de petite taille du zooplancton était sous-estimée. Sans ce biais d'échantillonnage, le genre *Oithona* est décrit comme le copépode marin le plus abondant (Gallienne et Robins 2001) et présent dans tous les environnements marins (Nishida 1985). En raison de cette forte abondance et sa biogéographie cosmopolite, *Oithona* est l'un des genres clé de l'écologie marine.

Le cycle de vie des copépodes non parasite est divisé en deux phases. Une première phase larvaire, appelée Nauplius, est composée de six étapes (du NI au NVI). Au cours de cette phase, les larves vont acquérir de nouveaux appendices par une alternance d'étapes de croissance et mue. Cela permet d'atteindre la deuxième phase (juvénile) appelée Copépodite. Cette phase est également composée de six étapes (CI à CVI) au cours desquelles les juvéniles, par alternance de croissance et de mue, vont augmenter leur taille, développer leurs segments et subir une maturation sexuelle. La forme adulte définitive est appelée Copépodite VI.

Malgré la multitude d'espèces occupant diverses niches écologiques, une grande majorité des copépodes adultes observés sont de formes ovoïdes, sauf pour les espèces parasitaires spécialisées pour l'hôte. Le nombre ou la forme des appendices est caractéristique de l'espèce et du sexe des individus. Pour identifier les différentes

espèces d'*Oithona*, les descriptions morphologiques de Nishida (1985) et Rose (1933) sont actuellement utilisés. Au niveau anatomique, les descriptions de Huys et Boxshall (1991), Dussart et Defaye (2001) et Mironova et Pasternak (2017) sont actuellement utilisés.

Le corps d'un copépode adulte non parasite est divisé en deux : le prosome (la tête et le thorax) et l'urosome (l'abdomen). Ils sont composés de segments de tailles différentes, ne se chevauchant pas et plus ou moins fusionnés selon les espèces. Sur les segments du prosome sont observés divers appendices (antennes, antennules, mandibules, maxillaire, maxillipèdes, pattes pour la nage) qui ont un rôle sensoriel, de moteur, d'adhérence ou de détection des proies. Selon la niche écologique, le comportement et la stratégie alimentaire, les appendices peuvent avoir une différence dans le nombre de segments, d'épines ou de soies (Huys et Boxshall 1991). L'orifice génital est situé sur le premier segment de l'abdomen et l'orifice anal sur le dernier segment. Par observation morphologique, Wilhelm Giesbrecht a établi deux superordres au sein des copépodes : Gymnoplea (« gymno- » signifiant nu) dû à l'absence d'appendices sur l'urosome, contrairement aux Podoplea (« pod- » signifiant pied) pour lesquels la cinquième paire de pattes (P5) est située sur le premier segment de l'urosome.

Chez la majorité des copépodes libres, certains appendices présentent des caractéristiques dépendantes du sexe. Pendant la différenciation en mâle, les antennes acquièrent des géniculations. Ceci permet au mâle de saisir la P4 de la femelle lors de l'accouplement. La P5 est également modifié pour permettre le transfert du spermatophore vers l'orifice génital de la femelle capturée. À ce jour, aucun cyclomorphosis n'a était observé.

Les copépodes possèdent donc une reproduction sexuelle entre deux individus de sexe opposé (mâle et femelle), sauf chez certaines espèces d'harpacticoïdes avec une capacité de parthénogenèse. La femelle *Oithona*, quasi immobile, est détectée par les mâles grâce à la traînée créée par leur mouvement. A l'échelle d'un copépode, l'eau est d'une haute viscosité : leur mouvement forme donc une trainée (Strickler et Balázsi 2007). De plus, les femelles *Oithona* relargue constamment des phéromones qui n'ont pas encore été identifiée chimiquement. Ces phéromones sont détectées par un organe spécialisé : l'aesthétasc. En général, les mâles possèdent deux aesthétascs par segment d'urosome (Huys et Boxshall 1991). Lorsque la femelle est détectée, le mâle essaie de

la capturer. La poursuite peut se terminer par l'évasion ou la capture de la femelle. Au cours de l'accouplement, le mâle agrippe la femelle à l'aide de son antenne articulée et transfère ses deux spermatophores sur l'orifice séminal en utilisant sa P5. La femelle stock ce sperme dans sa spermathèque, ce qui permet la fécondation de ces œufs. Avec un seul accouplement, la femelle peut féconder tous les œufs produits au cours de sa vie. Pour cette raison, le mâle a une préférence pour les femmes « vierges » (Heuschele et Kiørboe 2012). Les œufs sont transportés dans des sacs par la femelle jusqu'à leur éclosion.

Les *Oithona* chassent à l'affût (ambush feeders) (Benedetti, Gasparini et Ayata 2015), ce qui signifie qu'ils utilisent une stratégie d'attente, sans mouvement, que la proie soit à portée. Les longues paires d'antennes (variable selon les espèces) ont une fonction sensorielle en raison de leur grand nombre de soies (Strickler et Bal 1973). Cela permet de détecter les mouvements des proies (perception hydromécanique), de sauter dessus et de les capturer/manger à l'aide des appendices buccaux en quelques millisecondes. Cette stratégie limite également le risque d'être détecter par les prédateurs, en particulier pour les femelles *Oithona* avec des œufs. De plus, les antennes peuvent aider lors d'un puissant saut d'évasion (Borazjani et al. 2010), et c'est l'un des principaux arguments du succès écologique des copépodes (Kiørboe et al. 2010).

En raison de sa distribution cosmopolite, de son succès écologique, de son hyper-abondance, de son rôle essentiel dans le réseau trophique marin et les cycles biogéochimiques et de l'absence d'informations moléculaires, nous avons choisi le genre *Oithona* comme modèle potentiel pour le petit zooplancton. *Oithona* nana semble être présente dans la plupart des eaux côtières (Razouls et al.2019 ; Nishida 1985), notamment dans les eaux polluées et possède un petit génome. Pour toutes ces raisons, j'ai utilisé *Oithona* comme l'organisme modèle de ma thèse.

Objectifs de thèse

Cette thèse vise à combler les lacunes entre l'écologie comportementale et la biologie du copépode *Oithona* nana à l'aide d'analyses anatomiques et de données moléculaires. Dans le premier chapitre, j'analyse la structure du système digestif et de reproduction. Cette étude révèle leur constitution en chitine et confirment l'importance

des copépodes dans la pompe à carbone biologique marine. Dans le deuxième chapitre, je souligne l'explosion d'une famille de gènes, appelé gènes codant pour des protéines contenant des domaines Lin12 Notch repeat (LNR)(LDPGs) et souligne leur importance d'un point de vue évolutif en utilisant les données métagénomiques *Tara* Oceans. Le troisième chapitre confirment, à un niveau moléculaire, la différence entre le comportement des mâles par rapport aux autres stades de développement. Les transcriptomes des cinq stades de développement d'*O. nana* ont été construits et les gènes spécifiques à chaque stade identifié. En utilisant l'expression et l'analyse d'interaction protéine-protéine, j'ai essayé de caractériser le rôle fonctionnel des LDPG. Ils sont susceptibles de jouer un rôle dans la modulation de l'auto-protéolyse et de la neurogenèse chez le mâle et ainsi ils pourraient participer au comportement sacrificiel du mâle lors de la recherche de partenaire, ce qui semble être le premier cas observé chez un animal itéropare. Le dernier chapitre prouve le lien entre l'expression allèle spécifique (ASE) et la sélection naturelle, à l'échelle de la population, dans les populations *O. similis* des eaux arctiques. Dans ces populations, les gènes impliqués dans le système nerveux semblent également être soumis à des processus d'évolution à l'échelle de la population.

Données moléculaires d'*Oithona*

Pour construire des bibliothèques Illumina, la quantité d'ADN extrait d'un seul individu *O. nana* n'était pas suffisante (<5ng). Il a fallu regrouper plusieurs individus (plus de 2 000) pour finalement construire son génome. Par conséquent, l'assemblage du génome que nous avons proposé est une mosaïque d'individus de la même population de la petite rade de Toulon. Le génome annoté a une taille de 85Mb et est à environ 90% complet, avec plus de 15 000 gènes. C'est le premier génome disponible pour le genre *Oithona*, le premier pour l'ordre des cycloïdes et le troisième pour la sous-classe des copépodes.

D'autre part, l'extraction de l'ARN à partir d'un seul individu environnemental était suffisante pour la construction des bibliothèques Illumina. Grâce à ma thèse, des transcriptomes d'individus à différents stades de développement (œuf, larve, juvénile, mâle et femelle adulte) sont désormais également disponibles. Les gènes spécifiques des différents stages étant identifiés, il est maintenant possible d'utiliser ces gènes

comme marqueurs pour effectuer le contrôle moléculaire de la dynamique des populations d'*Oithona* et d'estimer la proportion d'individus aux différents stades dans les échantillons environnementaux. De plus, *O. nana* étant connu pour être particulièrement robuste et présent dans les eaux polluées (Richard et Jamet 2001) ; cette espèce peut être un biomarqueur de la mauvaise qualité de l'environnement lorsque sa population augmente au détriment d'autres espèces.

Notre objectif initial était également de construire le génome de *O. similis* et *O. atlantica* et de faire une analyse génomique comparative entre *Oithona* de trois différents écosystèmes (eaux côtières, froides et tropicales). Cependant, le génome de *O. similis* est beaucoup plus important que prévu : entre trois et cinq gigabases (Gb), contre 85 Mb pour *O. nana (*entre 35 fois et 60 fois plus grand*)*. Durant les premières extractions d'ADN d'*O.* nana, l'extraction s'effectuait sur les œufs, mais la construction du génome n'a pas été possible en raison de sa grande taille (> 1 Go d'après l'analyse du spectre de k- mer). Cette différence de taille entre cellules somatiques et germinales a déjà été observé chez d'autres cyclopoïdes d'eaux douces (*Cyclops)* et peut être expliqué par un phénomène appelé réduction chromatique (Grishanin 2014). Une ébauche d'assemblage du génome de *O. similis* a été construite. L'analyse du spectre de k-mer semble indiqué une taille supérieure à 3 Gb : l'espèce semble avoir perdu ou n'a pas acquis (contrairement à *O. nana*) la capacité de réduction de la chromatine. La première analyse de cet assemblage a montré une incomplétude importante (moins de 50%) et une fragmentation élevée, ce qui le rendent inutilisable pour la prédiction des gènes.

Cependant, j'ai construit des transcriptomes *O. similis* à quatre différents stades de développement. Malheureusement, je n'ai pas pu isoler de femelles portant des sacs à œufs dans les échantillons de la grande rade de Toulon. Au vu de nos analyses, on peut également dire que je n'ai pas pu isoler de femelle *O. similis*. C'est problématique car normalement c'est le stade le plus facile à identifier. Sur six individus que j'ai isolés pour l'extraction d'ARN et identifiés comme des femelles, aucune n'appartenait à l'espèce *O. similis* (basé sur l'analyse moléculaire des transcriptomes). Cela peut s'expliquer par l'une de ces deux raisons : soit je suis un mauvais taxonomiste ayant mal identifié les caractères spécifiques à l'espèce ou bien les critères pour désigner *O. similis* correspondent, au niveau moléculaire, à plusieurs espèces avec très peu de différences au niveau phénotypique. Pour répondre à cette question, je devrais

demander à d'autres taxonomistes d'identifier des femelles *O. similis*, puis de construire leur transcriptome.

À partir des données métagénomiques de *Tara* et des données de séquence 28S disponibles dans les banques publiques, nous avons pu confirmer la présence moléculaire du genre *Oithona* dans toutes les stations d'échantillonnage de l'expédition *Tara* Oceans. Nous avons observé *O. nana* dans plusieurs stations côtières du nord de la Méditerranée. Nous avons observé *O. similis* dans toutes les stations d'échantillonnage situées dans les eaux froides des océans Atlantique, Pacifique, Arctique et Antarctique. Dans les eaux tropicales, nous avons observé soit *O. atlantica, O. plumifera* ou *O. frigida* que nous avons regroupés en une seule unité taxonomique opérationnelle (OTU) en raison de leur similitude moléculaire (>98% d'identité nucléique moyenne sur la partie variable du 28S). D'autres espèces inconnues d'*Oithona* ont été détectées mais restent non identifiées en raison du manque de données moléculaires. Pour cette raison, un atlas de marqueurs moléculaires obtenu auprès d'individus bien identifiés par les taxonomistes doit être construit.

Caractérisation structurelle et fonctionnelle des LDPGs

En effectuant une analyse comparative avec d'autres génomes de copépodes disponibles, nous avons observé une explosion du domaine Lin-12 Notch Repeat (LNR) dans le protéome d'*O. nana*. Il semble que cette explosion soit également visible dans les transcriptomes *O. similis* avec moins d'amplitude (une vingtaine de LDPGs). Le génome étant incomplet, nous n'avons pas pu faire une analyse exhaustive et donc je ne vais pas discuter de ce point.

Cependant, dans les LDPGs d'*O. nana*, nous observons une grande diversité de structures avec de nouvelles associations de domaines, non décrites dans la littérature, ainsi qu'une forte divergence des séquences codant le domaine LNR. En mer Méditerranée, 7% des LDPGs sont sous pression de sélection dans les populations d'*O. nana*, dont trois mutations mono-nucléotidiques (SNPs) au sein des domaines LNR. Nous avons également observé 24% des 75 LDPGs surexprimés chez le mâle *O. nana* de la petite rade de Toulon. Cependant, dans la littérature, seules trois protéines possédant un domaine LNR sont décrites : Notch, dont elle tire son nom, la protéine Stealth et la Pappalysin.

La protéine Notch contient trois domaines LNR en tandem, jouant un rôle dans la protection d'un site de clivage dans la configuration « normale » et ainsi régule négativement la voie Notch (Sanchez-Irizarry et al., 2004). La protéine Stealth est présente chez certains eucaryotes et quelques procaryotes (Sperisen et al. 2005). Elle semble jouer un rôle dans le système immunitaire inné, mais le rôle des domaine LNR est encore inconnu. La pappalysine (Pappa) contient trois domaines LNR : deux en N-terminal et un en C-terminal (Monget et Oxvig 2016). Grace à ces domaines LNR, la Pappa peut former un homodimère (deux Pappa en tête-bêche) puis activer son domaine métallo-peptidase. Chez l'humain, Pappa peut cliver l'IGFPB 4 et 5, une protéine qui peut se lier à un facteur de croissance analogue à l'insuline (IGF). Ce clivage a pour conséquence de réguler positivement la quantité d'IGF libre (Monget et Oxvig 2016). Chez certains crustacés, l'IAG (Insulin-like Androgenic Gland) est l'hormone responsable de la différenciation en mâle et est régulée par une IGFBP (Rosen et al.2013 ; Ventura et Sagi 2012). Notre objectif au début de ma thèse était de trouver une IAG dans le protéome d'*O. nana* et de caractériser les LDP en détectant leurs partenaires protéiques par PPI avec un a priori qu'un des LDP portant un domaine protéase pourrait interagir avec une IGF. Nous avons également essayé d'identifier les gènes coexprimés en utilisant la méthode WCGNA, de caractériser la fonction des LDPGs par la technique d'extinction de gène (Gene Silencing) et également de construire l'arbre évolutif de tous les domaines LNR détectés chez *O. nana* par phylogénie.

La mise en place de la technique du double hybrides (Y2H) ainsi que les vérifications de faux positives et l'interprétation ont été laborieuses, mais nous avons finalement obtenu des résultats robustes. Nous avons détecté la formation de complexes protéiques contenant des LDP ainsi que des interactions avec des protéases, des protéines de la matrice extracellulaire (ECM) et des protéines connues pour être impliquées dans la neurogenèse.

Après une analyse complète de coexpression des gènes, nous avons obtenu 179 modules comprenant au moins 15 gènes. Parmi eux, 33 étaient significativement associés à un stade de développement, dont onze étaient spécifiques aux mâles. L'un de ces modules mâles, appelé M3, est composé de 443 gènes et est enrichi en LDPG, en gènes sous sélection, et en trypsine. Un autre, appelé M1, est composé de 1 199 gènes enrichis en récepteurs neuroactifs et transporteurs d'ions. J'ai cherché des

séquences régulatrices dans les régions en amont des gènes du module M3 et ai trouvé une séquence conservée de neuf nucléotides, régulant potentiellement neuf gènes comprenant un LDPG et un inhibiteur de protéase ; mais aucun facteur de transcription associé à cette séquence n'a été trouvée dans les bases de données.

Pour la mise en place du gene silencing, nous sommes partis d'un protocole développé chez l'harpacticoïde *Tigriopus californicus* (Barreto, Schoville et Burton 2015), un organisme modèle pour les études d'écotoxicologie, et nous l'avons adapté pour *O. nana*. Suivant le protocole, nous avons utilisés la technique d'électroporation pour insérer l'ARNds . En raison de notre incapacité technique à conserver une culture de *O. nana* pendant plus de deux semaines et de la distance entre la zone d'échantillonnage et le laboratoire d'Evry, nous avons eu d'intenses difficultés à développer un protocole approprié. Dans les meilleures expérimentations, nous avons observé plus du tiers des individus vivants après une séries de de dix impulsions et deux individus morts dont l'ARNds marqué a pénétré l'exosquelette chitineux, mais nous n'avons pas réussi à faire entrer à l'intérieur des individus vivants de l'ARNds marqué. Cette expérience sera probablement mise en place directement dans le laboratoire de Toulon où nous régulièrement capturer des individus *O. nana*.

Pour l'arbre évolutif des domaines LNR chez *O. nana*, en raison de leur petite taille et de leur forte divergence, nous avons utilisé les séquences nucléotidiques au lieu de la protéine pour conserver le plus d'informations possible. Nous avons essayé d'effectuer une phylogénie en utilisant les séquences protéiques, mais nous n'avons pas observé de signal fort. Sur la base des séquences nucléiques, l'arbre phylogénétique n'avait que 17% des nœuds avec un support solide, et 27 divisions de branche correspondant à la duplication en tandem impliquant 15 LDPGs (20% du total des LDPGs), y compris Notch.

Rôles des LDPGs

D'après les résultats d'expression différentielles et du Y2H, nos principales hypothèses pour la fonction des LDPGs sont la modulation, chez le mâle,(i) de l'autolyses et (ii) de la neurogenèse. L'autolyse ou « autodigestion » peut permettre au mâle de rechercher une femelle pendant de longues périodes sans nécessiter de prise alimentaire en fournissant l'énergie et les acides aminés nécessaires à son métabolisme

pour la synthèse du glutamate et le développement du système nerveux. La neurogenèse chez les mâles peut participer à leur motilité élevée et à leur capacité à capturer des femelles. Pour confirmer ces deux hypothèses, nous sommes en train de tester des protocoles de microscopie à fluorescences pour observer le système nerveux et le tissu conjonctif chez des individus *Oithona* mâles et femelles. Selon nos hypothèses, nous nous attendons à ce que les mâles présentent un système nerveux plus développé avec plus d'axones, de dendrites et de synapses, ainsi qu'une perte de tissu notamment pour le tissu conjonctif. Cette analyse microscopique est actuellement en cours.

Bibliography

Arif, Majda, Jérémy Gauthier, Kevin Sugier, Daniele Iudicone, Olivier Jaillon, Patrick Wincker, Pierre Peterlongo, and Mohammed-Amin Mohammed-Amin Mohammed-Amin Madoui. 2019. "Discovering Millions of Plankton Genomic Markers from the Atlantic Ocean and the Mediterranean Sea." *Molecular Ecology Resources* 19 (2): 0–3. https://doi.org/10.1111/1755-0998.12985.

Baird, W. 1843. "Notes on British Entomostraca." *The Zoologist (Newman)*, 1:193-197. http://marinespecies.org/aphia.php?p=sourcedetails&id=75847.

Barreto, Felipe S., Sean D. Schoville, and Ronald S. Burton. 2015. "Reverse Genetics in the Tide Pool: Knock-down of Target Gene Expression via RNA Interference in the Copepod *Tigriopus Californicus*." *Molecular Ecology Resources* 15 (4): 868–79. https://doi.org/10.1111/1755-0998.12359.

Baumgartner, Mark F., and Ann M. Tarrant. 2017. "The Physiology and Ecology of Diapause in Marine Copepods." *Annual Review of Marine Science* 9 (1): 387–411. https://doi.org/10.1146/annurev-marine-010816-060505.

BBC News. 2004. "Alarm over 'non-Kosher' NY Water." BBC News Online. 2004. http://news.bbc.co.uk/2/hi/africa/3774005.stm.

Beaugrand, Grégory, Keith M. Brander, J. Alistair Lindley, Sami Souissi, and Philip C. Reid. 2003. "Plankton Effect on Cod Recruitment in the North Sea." *Nature* 426 (6967): 661–64. https://doi.org/10.1038/nature02164.

Benedetti, Fabio, Stéphane Gasparini, and Sakina Dorothée Ayata. 2015. "Identifying Copepod Functional Groups from Species Functional Traits." *Journal of Plankton Research* 38 (1): 159–66. https://doi.org/10.1093/plankt/fbv096.

Borazjani, Iman, Fotis Sotiropoulos, Edwin Malkiel, and Joseph Katz. 2010. "On the Role of Copepod Antennae in the Production of Hydrodynamic Force during Hopping." *The Journal of Experimental Biology* 213 (Pt 17): 3019–35. https://doi.org/10.1242/jeb.043588.

Boyd, Philip W., Hervé Claustre, Marina Levy, David A. Siegel, and Thomas Weber. 2019. "Multi-Faceted Particle Pumps Drive Carbon Sequestration in the Ocean." *Nature* 568 (7752): 327–35. https://doi.org/10.1038/s41586-019-1098-2.

Bradford-Grieve, Janet M. 2004. "Deep-Sea Benthopelagic Calanoid Copepods and Their Colonization of the Near-Bottom Environment." *Zoological Studies* 43 (2): 276–91. http://www.sinica.edu.tw/zool/zoolstud/43.2/276.pdf.

Brown, Michael D., Brian H. Kay, and Joan K. Hendrikz. 1991. "Evaluation of Australian Mesocyclops (Cyclopoida: Cyclopidae) for Mosquito Control." *Journal of Medical Entomology* 28 (5): 618–23. https://doi.org/10.1093/jmedent/28.5.618.

Cornils, Astrid, Britta Wend-Heckmann, and Christoph Held. 2017. "Global Phylogeography of Oithona Similis s.l. (Crustacea, Copepoda, Oithonidae) – A Cosmopolitan Plankton Species or a Complex of Cryptic Lineages?" *Molecular*

Phylogenetics and Evolution 107 (February): 473–85. https://doi.org/10.1016/j.ympev.2016.12.019.

Corte, D De, I Lekunberri, E Sintes, JAL Garcia, S Gonzales, and GJ Herndl. 2014. "Linkage between Copepods and Bacteria in the North Atlantic Ocean." *Aquatic Microbial Ecology* 72 (3): 215–25. https://doi.org/10.3354/ame01696.

Dussart, B. H., and D. Defaye. 2001. *Copepoda. Introduction to the Copepoda. Guides to the Identification of the Microinvertebrates of the Continental Waters of the World.* Edited by H. J. F. Dumont. 2nd ed. Vol. 16. Leiden: Backhuys Publishers.

Eyun, Seong Il. 2017. "Phylogenomic Analysis of Copepoda (Arthropoda, Crustacea) Reveals Unexpected Similarities with Earlier Proposed Morphological Phylogenies." *BMC Evolutionary Biology* 17 (1): 1–12. https://doi.org/10.1186/s12862-017-0883-5.

Ferrari, Frank D., and James Orsi. 1984. "Oithona Davisae, New Species, and Limnoithona Sinensis (Burckhardt, 1912) (Copepoda: Oithonidae) From the Sacramento-San Joaquin Estuary, California." *Journal of Crustacean Biology* 4 (1): 106–26. https://doi.org/10.2307/1547900.

Gallienne, C. P., and D. B. Robins. 2001. "Is Oithona the Most Important Copepod in the World's Oceans?" *Journal of Plankton Research* 23 (12): 1421–32. https://doi.org/10.1093/plankt/23.12.1421.

Giesbrecht, Wilhelm. 1892. *Systematik Und Faunistik Der Pelagischen Copepoden Des Golfes von Neapel Und Der Angrenzenden Meeres-Abschnitte.* Berlin,: R. Friedlander & Sohn,. https://www.biodiversitylibrary.org/item/15504.

Grishanin, Andrey. 2014. "Chromatin Diminution in Copepoda (Crustacea): Pattern, Biological Role and Evolutionary Aspects." *Comparative Cytogenetics* 8 (1): 1–10. https://doi.org/10.3897/CompCytogen.v8i1.5913.

Harvey, T. H. P., and B. E. Pedder. 2013. "Copepod Mandible Palynomorphs From the Nolichucky Shale (Cambrian, Tennessee): Implications for the Taphonomy and Recovery of Small Carbonaceous Fossils." *Palaios* 28 (5): 278–84. https://doi.org/10.2110/palo.2012.p12-124r.

Heuschele, Jan, and Thomas Kiørboe. 2012. "The Smell of Virgins: Mating Status of Females Affects Male Swimming Behaviour in Oithona Davisae." *Journal of Plankton Research* 34 (11): 929–35. https://doi.org/10.1093/plankt/fbs054.

Humes, Arthur G. 1994. "How Many Copepods?" *Hydrobiologia* 292–293 (1): 1–7. https://doi.org/10.1007/BF00229916.

Huys, Rony., and Geoffrey Allan. Boxshall. 1991. *Copepod Evolution*. Ray Society. http://www.raysociety.org.uk/publications/zoology/copepod-evolution-r-huys-and-g-a-boxshall/.

Huys, Rony, Eduardo Suárez-Morales, Mariá De Lourdes Serrano-Sánchez, Elena Centeno-Garciá, and Francisco J. Vega. 2016. "Early Miocene Amber Inclusions from Mexico Reveal Antiquity of Mangrove-Associated Copepods." *Scientific Reports* 6 (September): 1–12. https://doi.org/10.1038/srep34872.

Jónasdóttir, Sigrún Huld, André W Visser, Katherine Richardson, and Michael R Heath. 2015. "Seasonal Copepod Lipid Pump Promotes Carbon Sequestration in the Deep North Atlantic." *Proceedings of the National Academy of Sciences of the United States of America* 112 (39): 12122–26. https://doi.org/10.1073/pnas.1512110112.

Khan, Aziz, Oriol Fornes, Arnaud Stigliani, Marius Gheorghe, Jaime A Castro-Mondragon, Robin Van Der Lee, Adrien Bessy, et al. 2018. "JASPAR 2018: Update of the Open-Access Database of Transcription Factor Binding Profiles and Its Web Framework." *Nucleic Acids Research* 46. https://doi.org/10.1093/nar/gkx1126.

Kiørboe, Thomas. 2011. "What Makes Pelagic Copepods so Successful?" *Journal of Plankton Research* 33 (5): 677–85. https://doi.org/10.1093/plankt/fbq159.

Kiørboe, Thomas, Anders Andersen, Vincent J Langlois, and Hans H Jakobsen. 2010. "Unsteady Motion: Escape Jumps in Planktonic Copepods, Their Kinematics and Energetics." *Journal of the Royal Society, Interface* 7 (52): 1591–1602. https://doi.org/10.1098/rsif.2010.0176.

Kiørboe, Thomas, Anders Andersen, Vincent J Langlois, Hans Henrik Jakobsen, and Tomas Bohr. 2009. "Mechanisms and Feasibility of Prey Capture in Ambush-Feeding Zooplankton." *Proceedings of the National Academy of Sciences of the United States of America* 106 (30): 12394–99. https://doi.org/10.1073/pnas.0903350106.

Kobari, Toru, Deborah K. Steinberg, Ai Ueda, Atsushi Tsuda, Mary W. Silver, and Minoru Kitamura. 2008. "Impacts of Ontogenetically Migrating Copepods on Downward Carbon Flux in the Western Subarctic Pacific Ocean." *Deep Sea Research Part II: Topical Studies in Oceanography* 55 (14–15): 1648–60. https://doi.org/10.1016/J.DSR2.2008.04.016.

Kottelat, Maurice. 2007. "A Freshwater Diaptomid Copepod Harvasted for Human Consumption in Central Laos." *The Raffles Bulletin of Zoology* 16: 355–57.

Laso-Jadart, Romuald, Kevin Sugier, Emmanuelle Petit, Karine Labadie, Pierre Peterlongo, Christophe Ambroise, Patrick Wincker, Jean-Louis Jamet, and Mohammed-Amin Madoui. 2019. "Linking Allele-Specific Expression And Natural Selection In Wild Populations." *BioRxiv*, April, 599076. https://doi.org/10.1101/599076.

Madoui, Mohammed-Amin, Julie Poulain, Kevin Sugier, Marc Wessner, Benjamin Noel, Leo Berline, Karine Labadie, et al. 2017. "New Insights into Global Biogeography, Population Structure and Natural Selection from the Genome of the Epipelagic Copepod *Oithona*." *Molecular Ecology* 26 (17): 4467–82. https://doi.org/10.1111/mec.14214.

Marten, Gerald G. 2001. *Human Ecology : Basic Concepts for Sustainable Development*. Earthscan Publications.

Milne Edwards, H. 1840. *Histoire Naturelle Des Crustacés : Comprenant l'anatomie, La Physiologie et La Classification de Ces Animaux.* Paris: Librairie encyclopédique de Roret,. https://www.biodiversitylibrary.org/item/54604.

Mironova, Ekaterina, and Anna Pasternak. 2017. "Female Gonad Morphology of Small

Copepods Oithona Similis and Microsetella Norvegica." *Polar Biology* 40 (3): 685–96. https://doi.org/10.1007/s00300-016-1993-z.

Monget, Philippe, and Claus Oxvig. 2016. "PAPP-A and the IGF System." *Annales d'Endocrinologie* 77 (2): 90–96. https://doi.org/10.1016/j.ando.2016.04.015.

Nishida, Shuhei. 1985. "Taxonomy and Distribution of the Family Oithonidae (Copepoda, Cyclopoida) in the Pacific and Indian." *Bull. Ocean Res. Inst.* 20: 1–167.

Pedersen, Alice Marie, Birthe Vang, and Ragnar L. Olsen. 2014. "Oil from Calanus Finmarchicus —Composition and Possible Use: A Review." *Journal of Aquatic Food Product Technology* 23 (6): 633–46. https://doi.org/10.1080/10498850.2012.741662.

Rawlings, Tonya K, Gregory M Ruiz, and Rita R Colwell. 2007. "Association of Vibrio Cholerae O1 El Tor and O139 Bengal with the Copepods Acartia Tonsa and Eurytemora Affinis." *Applied and Environmental Microbiology* 73 (24): 7926–33. https://doi.org/10.1128/AEM.01238-07.

Razouls, C., F. de Bovée, J. Kouwenberg, and N. Desreumaux. 2019. "Diversité et Répartition Géographique Chez Les Copépodes Planctoniques Marins." 2019. https://copepodes.obs-banyuls.fr/.

Regier, Jerome C., Jeffrey W. Shultz, Andreas Zwick, April Hussey, Bernard Ball, Regina Wetzer, Joel W. Martin, and Clifford W. Cunningham. 2010. "Arthropod Relationships Revealed by Phylogenomic Analysis of Nuclear Protein-Coding Sequences." *Nature* 463 (7284): 1079–83. https://doi.org/10.1038/nature08742.

Richard, Simone, and Jean-louis Jamet. 2001. "An Unusual Distribution of Oithona Nana GIESBRECHT (1892) (Crustacea : Cyclopoida) in a Bay : The Case of Toulon Bay (France, Miditerranean Sea)." *Journal of Coastal Research*, no. April 2015.

Rose, M. 1933. *Faune de France - Copépodes Pélagiques*. Lechevalier. Paris. www.faunedefrance.org.

Rosen, Ohad, Simy Weil, Rivka Manor, Ziv Roth, Isam Khalaila, and Amir Sagi. 2013. "A Crayfish Insulin-like-Binding Protein: Another Piece in the Androgenic Gland Insulin-like Hormone Puzzle Is Revealed." *Journal of Biological Chemistry* 288 (31): 22289–98. https://doi.org/10.1074/jbc.M113.484279.

Sanchez-Irizarry, C., A. C. Carpenter, A. P. Weng, W. S. Pear, J. C. Aster, and S. C. Blacklow. 2004. "Notch Subunit Heterodimerization and Prevention of Ligand-Independent Proteolytic Activation Depend, Respectively, on a Novel Domain and the LNR Repeats." *Molecular and Cellular Biology* 24 (21): 9265–73. https://doi.org/10.1128/MCB.24.21.9265-9273.2004.

Sanders, Kate L., and Michael S.Y. Lee. 2010. "Arthropod Molecular Divergence Times and the Cambrian Origin of Pentastomids." *Systematics and Biodiversity* 8 (1): 63–74. https://doi.org/10.1080/14772000903562012.

Selden, Paul A., Rony Huys, Michael H. Stephenson, Alan P. Heward, and Paul N.

Taylor. 2010. “Crustaceans from Bitumen Clast in Carboniferous Glacial Diamictite Extend Fossil Record of Copepods.” *Nature Communications* 1 (5): 1–6. https://doi.org/10.1038/ncomms1049.

Shoemaker, William R., Kenneth J. Locey, and Jay T. Lennon. 2017. “A Macroecological Theory of Microbial Biodiversity.” *Nature Ecology and Evolution* 1 (5). https://doi.org/10.1038/s41559-017-0107.

Sperisen, Peter, Christoph D Schmid, Philipp Bucher, and Olav Zilian. 2005. “Stealth Proteins: In Silico Identification of a Novel Protein Family Rendering Bacterial Pathogens Invisible to Host Immune Defense.” *PLoS Computational Biology* 1 (6): e63. https://doi.org/10.1371/journal.pcbi.0010063.

Strickler, J. Rudi, and Arya K. Bal. 1973. “Setae of the First Antennae of the Copepod Cyclops Scutifer (Sars): Their Structure and Importance.” *Proceedings of the National Academy of Sciences of the United States of America* 70 (9): 2656. https://doi.org/10.1073/PNAS.70.9.2656.

Strickler, J Rudi, and Gábor Balázsi. 2007. “Planktonic Copepods Reacting Selectively to Hydrodynamic Disturbances.” *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 362 (1487): 1947–58. https://doi.org/10.1098/rstb.2007.2080.

Sugier, Kevin, Benoit Vacherie, Astrid Cornils, Patrick Wincker, Jean-Louis Jamet, and Mohammed-Amin Madoui. 2018. “Chitin Distribution in the Oithona Digestive and Reproductive Systems Revealed by Fluorescence Microscopy.” *PeerJ* 6 (May): e4685. https://doi.org/10.7717/peerj.4685.

Tang, KW. 2005. “Copepods as Microbial Hotspots in the Ocean: Effects of Host Feeding Activities on Attached Bacteria.” *Aquatic Microbial Ecology* 38 (1): 31–40. https://doi.org/10.3354/ame038031.

Technical University of Denmark (DTU). 2010. “What Makes World’s Strongest Animal -- the Tiny Copepod -- so Successful? -- ScienceDaily.” 2010. https://www.sciencedaily.com/releases/2010/05/100512172444.htm.

Thompson, Gustavo A., Estela O. Dinofrio, and Viviana A. Alder. 2013. “Structure, Abundance and Biomass Size Spectra of Copepods and Other Zooplankton Communities in Upper Waters of the Southwestern Atlantic Ocean during Summer.” *Journal of Plankton Research* 35 (3): 610–29. https://doi.org/10.1093/plankt/fbt014.

Ventura, Tomer, and Amir Sagi. 2012. “The Insulin-like Androgenic Gland Hormone in Crustaceans: From a Single Gene Silencing to a Wide Array of Sexual Manipulation-Based Biotechnologies.” *Biotechnology Advances* 30 (6): 1543–50. https://doi.org/10.1016/j.biotechadv.2012.04.008.

WoRMS. 2019a. “Copepoda.” 2019. http://www.marinespecies.org/aphia.php?p=taxdetails&id=1080.

———. 2019b. “Oithona Baird.” 2019. http://www.marinespecies.org/aphia.php?p=taxdetails&id=106485.

Zamora Terol, Sara. 2013. “Ecology of the Marine Copepod Genus Oithona.” *TDX*

(Tesis Doctorals En Xarxa). Universitat Politècnica de Catalunya. https://upcommons.upc.edu/handle/2117/95142.

호프콤파니. 2014. “수조와 연못 속 작은 생명들 (Hope Aqua Art).” 2014. https://wongj81.blog.me/220105785244.

université PARIS-SACLAY | ÉCOLE DOCTORALE Structure et dynamique des systèmes vivants (SDSV)

Titre : Etude moléculaire et anatomique du copépode épipélagique *Oithona nana* (*Copepoda*; *Cyclopoida*)

Mots clés : Copépode ; *Oithona* ; Génome ; Transcriptomes ; Anatomie

Résumé :

Les copépodes sont les animaux les plus nombreux sur Terre, devant les insectes, et jouent un rôle essentiel dans le réseau trophique et les cycles biogéochimiques marins. Dans la zone épipélagique, *Oithona* est considéré comme un des genres les plus abondants et les plus répandus. Malgré l'importance écologique des copépodes, peu d'informations sur leur anatomie et peu de données moléculaires sont disponibles, particulièrement pour *Oithona*. Nous avons adapté un protocole qui permet d'observer les systèmes digestifs et reproducteurs d'*Oithona,* mais également de déterminer, pour la première fois, la distribution en chitine au sein de ces systèmes. Nous avons aussi rendu disponible le génome de *Oithona nana* ainsi que les transcriptomes aux différents stades de développement.

L'analyse de génomique comparative entre les trois différents génomes de copépodes disponibles nous a permis d'observer une explosion du domaine protéique Lin-12 Notch Repeat (LNR) dans le génome de *O. nana*. En utilisant les données métagénomique de Tara, nous avons identifié la structure de la population d'*O. nana* en mer Méditerranéen, ainsi que détecté les loci sous sélection naturelle. Parmi eux, cinq sont des gènes codant pour une protéine contenant des domaines LNR (nommé LDPGs), dont trois gènes contenant des mutations au sein même d'un domaine LNR.

Bien qu'un système ZW de détermination sexuelle soit prédit chez *O. nana*, une série temporelle de quinze ans dans la petite rade de Toulon montre un sexe-ratio biaisé en faveurs des femelles (ratio mâle/femelle < 0.15± 0.11), mettant en évidence une plus forte mortalité chez les mâles. Ceci peut être expliqué par le « paradoxe du mâle » : les mâles *Oithona* doivent alterner entre des phases d'alimentation (immobile) et des phases de recherche de partenaire pour s'accoupler (mobile). Comme les mécanismes moléculaires de ce compromis sont inconnus, nous avons effectué une analyse d'expression différentielle. Vingt pourcents des LDPGs sont surexprimés chez le mâle. Les males montrent également un enrichissement en transcrits impliqués dans la protéolyse, la formation de nouveaux axones et dendrites, la formation et le fonctionnement des synapses, ainsi que la conversion d'acides aminées en glutamate, un neurotransmetteur excitateur. De plus, plusieurs gènes négativement régulés chez le mâle sont impliqués dans l'augmentation de la prise alimentaire et l'activation de la digestion. La formation de complexe protéique contenant des domaine LNR a été détectée par double hybride, ainsi que des interactions impliquant des protéases, des protéines de la matrice extracellulaire et des protéines liées à la neurogenèse.

L'ensemble de ces résultats montrent que la recherche de partenaire chez les mâles *O. nana* est soutenue par une autolyse modulée par des LDP, permettant la libération d'acides aminés convertis en glutamate, et permettant le développement du système nerveux. Ceci soutient un comportement sacrificiel à une échelle moléculaire chez le mâle *O. nana* et pourrait constituer le premier cas de comportement sacrificiel chez un animal itéropare.

université PARIS-SACLAY | ÉCOLE DOCTORALE Structure et dynamique des systèmes vivants (SDSV)

Title: Molecular and anatomical study of the epipelagic copepod *Oithona nana* (*Copepoda*; *Cyclopoida*)

Keywords: Copepod; *Oithona;* Genome; Transcriptomes; Anatomy

Abstract :

Copepods are the most numerous animals on Earth and play an essential role in the marine trophic web and biogeochemical cycles. In the epipelagic ocean, the genus *Oithona* is considered as one of the most abundant and widespread copepods. Despite the ecological importance, few internal anatomy and molecular data are available for copepods in general, and for *Oithona* in particular. We updated a protocol permitting to observe the digestif and reproductive systems of *Oithona*, but also to determine, for the first time, the chitin distribution inside these systems. We also made available the *Oithona nana* genome and transcriptomes at the different development stages.

After comparative genomic analysis against other available copepod genomes, we observe an explosion of the Lin-12 Notch Repeat (LNR) protein domain in the *O. nana* genome. Using the *Tara* metagenomic data, we identified the population structure of *O. nana* in the Mediterranean Sea, but also detected loci under natural selection. Among them, five were LNR domain-containing protein-coding genes (LDPGs) whose three single-nucleotide mutations within LNR domain.

While a ZW sex-determination system was predicted in *O. nana*, a fifteen-year time-series in the Toulon Little Bay showed a biased sex ratio toward females (male / female ratio < 0.15± 0.11) highlighting higher mortality in males. This can be explained by the "*Oithona* male paradox": males have to alternate between feeding (immobile) and partner research (mobile) phases. As the molecular basis of this trade-off is unknown, we make a differential expression analysis. Twenty-four per cent of LDPGs were up-regulated in males. The male also showed enrichment in transcripts involved in proteolysis, the formation of new axons and dendrites, synapse assembly and functioning and even amino acid conversion to glutamate (an excitatory neurotransmitter). Moreover, several male down-regulated genes were involved in the increase of food uptake and digestion. The formation of LDP complexes was detected by yeast two-hybrid, with interactions involving proteases, extracellular matrix proteins and neurogenesis related proteins.

Together, these results showed that the mating partner search of *O. nana* male is sustained by LDP-modulated autolysis that releases amino acid, convert them to the glutamate and permit a nervous system development. This could support a molecular-scale sacrificial behaviour in *O. nana* male.