UNIVERSITÀ
DEGLI STUDI
DI PADOVA

HEAD OFFICE: UNIVERSITÀ DEGLI STUDI DI PADOVA

Department of Biology

Ph.D. COURSE IN: Biosciences
CURRICULUM: Genetics, Genomics and Bioinformatics
SERIES 34°

Integrated transcriptome reconstruction and annotation of three krill species: new insights on their life cycle and environmental adaptation

Coordinator: Prof. Ildikò Szabò
Supervisor: Prof. Chiara Romualdi
Co-Supervisor: Prof. Gabriele Sales

Ph.D. student: Ilenia Urso

SUMMARY **3**

INTRODUCTION **1**

KRILL SPECIES AND THEIR CRUCIAL ROLE 1

Euphausia superba *4*

Meganyctiphanes norvegica *6*

Thysanoessa inermis *7*

STATE OF THE ART AND EVALUATION OF THE SEQUENCING APPROACHES 8

RESEARCH OBJECTIVES 10

KRILL SAMPLING AND SEQUENCING 13

Euphausia superba *13*

Meganyctiphanes norvegica *14*

Thysanoessa inermis *16*

TRANSCRIPTOME ASSEMBLY STRATEGY **17**

DE NOVO ASSEMBLY TEST 17

QUALITY ASSESSMENT 19

KRILL DE NOVO TRANSCRIPTOME ASSEMBLY **22**

META-ASSEMBLY STRATEGY **27**

FUNCTIONAL ANNOTATION **33**

Euphausia superba *33*

Meganyctiphanes norvegica *34*

Thysanoessa inermis *34*

EXPRESSION ATLAS **36**

DIFFERENTIAL EXPRESSION ANALYSIS **38**

Euphausia superba *39*

Meganyctiphanes norvegica *42*

Thysanoessa inermis *45*

IDENTIFICATION OF MICRORNA PRECURSORS **47**

Euphausia superba *48*

Meganyctiphanes norvegica *49*

Thysanoessa inermis *50*

CROSS SPECIES COMPARISON **52**

KRILLDB^2 WEB INTERFACE **56**

CONCLUSIONS AND FUTURE PERSPECTIVES **60**

ACKNOWLEDGMENTS **63**

REFERENCES **64**

Summary

Euphausia superba, *Meganyctiphanes norvegica* and *Thysanoessa inermis* are three krill species. The first one is typical of the Antarctic region while the other two are common in the North Atlantic.

In their specific environment they all play crucial roles in the food web and are characterized by a great abundance in terms of biomass. Their ecologically fundamental role makes it crucial to increase the knowledge of their life cycle, adaptation strategies and responses to environmental changes and human activities.

The aim of this dissertation is to create and annotate a *de novo* transcriptome assembly for the three krill species, to explore their gene expression levels across different developmental stages and their responses to different environmental conditions, allowing for a comparative and physiological study. The present study can be subdivided into three main parts: i) a preliminary investigation focusing on the efficiency and quality of already existing strategies for *de novo* transcriptome assembly of non-model organisms; ii) a second part consisting in the identification and application of the best transcriptome assembly strategy; iii) a final step during which I analyze all the results and collect them in the first comparative analysis of the species under study.

At first, I performed a separate transcriptome reconstruction using five different *de novo* assembly programs: Trinity, BinPacker, rnaSPAdes, TransABySS and IDBA-tran. A combination of two filtering steps was applied to the newly reconstructed transcriptomes to discard artifacts and improve the assembly quality. First, I estimated the abundances of all the transcripts reconstructed by each assembler retaining only those transcripts showing an expression level of at least 1 transcript per million (TPM) within each of the three experimental conditions. In a second step, I considered the results of all assemblers jointly: I ran the “cd-hit-est” program to cluster similar sequences and to produce a set of non-

redundant representative transcripts. In order to group resulting transcripts into units corresponding to genes I ran the “EvidentialGene” pipeline, followed by another round of analyses to identify redundant or mis-assembled sequences still appearing in the transcriptome. Through the combination of all these filters I produced a new transcriptome for each krill species, retaining alternative and paralog transcripts with sufficient level of uniqueness in their sequence.

The quality assessment of the reconstructed transcriptomes confirmed the reliability of the strategy applied. Furthermore, the availability of these new references suggested the possibility to perform a detailed investigation of differential expression patterns, which highlighted the presence of genes that were already described in literature together with new ones likely involved in crucial steps of the species life cycle. The comparison between differential expression results across the different krill species highlighted the presence of genes involved in the same functions and processes that appear to be differentially expressed for each krill. Together with these analyses, an orthology inference was performed, allowing the identification of transcripts showing an orthology relationship with genes from other species. By cross-referencing all these results, I identified a series of sequences showing an orthology relationship across all the three krill species, suggesting that those transcripts derived from a common ancestor.

The new results produced for *E. superba*, then, allowed me to extend the KrillDB website, now named KrillDB^2, which provides the most complete source of information about the krill transcriptome and will offer a reliable starting point development of novel ecological studies.

In conclusion, this work highlighted the importance of a precise transcriptome reconstruction to maximize the potential to describe the expression profile and transcriptional phenotypes of a species in the most accurate way; moreover, the results produced in terms of computational approaches have improved our possibilities in the field of transcriptomic studies and represent a starting point for a deeper and more accurate fine-tuning of the available procedure. From a genetic and ecological point of view, this work represents a basis for future functional studies on krill responses to environmental changes and the underlying molecular mechanisms.

Introduction

In this section I will collect the background information about the krill species under study, specifically in terms of geographical distribution, life cycle and available genetic resources. In addition, a summary of the most common and referenced techniques found in literature regarding the transcriptome reconstruction in non-model organisms will be evaluated, in order to give an overview of the state of the art, which represented the starting point for the optimization of the more general procedure. Therefore, I will explain the research objectives of this project and describe the samples collection used in this work.

Krill species and their crucial role

Krill are small decapod crustaceans which can be found throughout the oceans. These organisms are considered key species in the food web, especially in the Southern Ocean, representing a crucial link between apex predators and primary producers. In fact, on the one hand they are the major grazers of primary production, repacking vast amounts of primary production into their own body by grazing micro-size phytoplankton and, on the other hand, they also represent the main planktonic diet of marine mammals. This is the reason why the word “krill” usually refers to “whale food”, although its actual meaning is “small fry of fish”. It has been estimated that the overall consumption of krill by marine mammals is about 10–20 million tons/year in the North Pacific, 15–25 million tons/year in the North Atlantic, and 125–250 million tons/year in the Southern Hemisphere, where the largest part is consumed in the Southern Ocean (Hewitt & Lipsky, 2018). Specifically, in the Antarctic region the term “krill” usually refers to a single species, the Antarctic krill *Euphausia superba*. Together with the other two species under study, the Northern krill *Meganyctiphanes norvegica* and *Thysanoessa inermis* they all belong to the same taxonomic order, the Euphausiacea, and the Euphausiidae family, that collects also other krill species (*Euphausia* *pacifica*, *Thysanoessa raschii*, *Thysanoessa spinifera*). In particular, Euphausiidis term comprises a total of 86 species in 10 genera (Janine Cuzin-Roudy, in Advances in Marine Biology, 2010). These species occur in all oceans worldwide and are mainly epipelagic or may distribute in the water column at a depth between 180 m and 900 m. They must continually swim to maintain their position, aggregating into swarms that may have a diameter length from one meter to tens of meters and may extend horizontally tens of meters to several thousand meters (Hewitt & Lipsky, 2018). A common characteristic relies on krill capability to perform diel vertical migrations along the water column, usually moving towards the surface at night providing food for surface predators and living near the bottom during the day.

As proof that the role of these species in their environment is fundamental to ensure the marine ecosystem conservation, it has been observed that a decrease in the Southern West Atlantic region in krill abundance directly produced a parallel decrease in the number of krill dependent predators (Trivelpiece et al., 2011). There are different factors that affect krill distribution and abundance, but most of the pressure is addressed to climate change and fisheries.

Climate change represents a global issue that impact oceans worldwide. Changes in climate can fundamentally alter many properties of the oceans, in particular the increase in greenhouse gases emissions, which trap more energy from the sun, causes an increase of sea surface temperatures and rising sea level. It is true that the oceans can reduce climate change by storing large amounts of carbon dioxide (CO_2), however it contextually produces an increasing level of dissolved carbon into sea water, changing its chemistry and making it more acidic.

Ocean acidification, represented by a decrease of sea water pH due to high levels of CO_2 dissolving into it, and consequently climate change is some of the main concerns of the scientific community about future sustainability of marine ecosystems and organisms. Increased ocean acidity causes a series of problems to marine communities. A common reported effect regards, for example, corals and shellfish difficulties building their skeletons and shells. All these information make it clear how such environmental changes can in turn alter the biodiversity and productivity of ocean ecosystems.

Regarding krill species, an increasing attention has been developed during the last years for the effects of a series of stressors on Antarctic krill. During the last quarter of the 20th century Antarctic krill abundance has experienced a decline caused by a decrease of the sea ice coverage due to increasing temperatures (Atkinson et al., 2004). This greater and most documented impact on *Euphausia superba* ecology and abundance is mainly due to the fact the Southern Ocean are affected by ocean acidification more than other marine environments because of the higher solubilities of CO_2 in cold waters and because of the upwelling of deep sea water with high concentration of CO_2 (Sabine et al., 2004).

Recently, it has been documented how growth and Antarctic krill population dynamics seem to be influenced by consistent changes in climate and ocean temperature (Murphy et al., 2017; Veytia et al., 2020), which led to shifts in growth habitat timing. It is known that krill life cycle is strictly synchronized with seasonal cycles in Southern Ocean and winter sea-ice cover (Ducklow et al., 2006; Flores et al. 2012), therefore the observed early end of the growth season in ecologically important regions such as Antarctic Peninsula have already altered the krill population dynamics.

On the other hand, as described by Riquelme-Bugueño et al. (2020) some euphausiidis are capable to adapt and live in association with such extreme condition (high CO_2 level, minimum oxygen concentration), which represents an interesting and important signal to be analyzed in detail in order to understand how to overcome future changes caused by these increasing stressors.

Together with environmental changes, krill are also affected by human activities, such as commercial fishery. Due to the strong impact of climate changes on krill populations and ecology the need of an accurate resource and conservation management has increased over the years. In Southern Ocean the Convention for the Conservation of Antarctic Marine Living Resources (CCAMLR), established in 1982, specifically aims at guaranteeing the conservation of Antarctic marine organisms and, at the same time, allow for the rational use of marine living resources. The fishery of Antarctic krill has been considered as one of the world's last under-exploited fisheries (Garcia and Rosenberg 2010) for several years, and the reasons why are probably related to its high costs and the difficulty of obtaining commercially useful product from it. However, over the years, new technologies have been developed, together with the possibility of an increased usage of derived krill products. Although a small part of these catches is intended for human consumption and uses (e.g., medical applications), a lot of krill products are used in aquaculture applications. Over the years, the CCAMLR has established a series of rules in order to manage krill catches and exploitation. Today, Antarctic krill may be taken in Southern Ocean only in four established sub-areas and divisions. The total allowable catch for the southwest Atlantic is currently about 5.6 million tons annually but distributed across the different areas the actual annual catch is around 0.3% of the unexploited biomass of krill.

All these information demonstrate how crucial is the study of these species and how strong is the need of a deep knowledge of the dynamics – and the molecular mechanisms underlying - of these species within their environment in order to prevent them from dramatic declines and ecological impacts.

Euphausia superba

Antarctic krill *Euphausia superba* represents a widespread distributed crustacean species of the Southern Ocean and one of the world's most abundant species with a total biomass between 100 and 500 million tons (Nicol & Endo, 1997). The average lifespan of Antarctic krill is 5-7 years. (Everson, 2000; Nicol, 2006).

E. superba, as all Euphausiidis species, is characterized by a life cycle were several larval stages follow each other; the main and common larval stages are represented by nauplius, metanauplius, calyptopus, and furcilia. As all crustaceans, one of the major parts of krill life history is the moult cycle, also called "ecdysis", consisting in the shedding of the exoskeleton for the purpose of growth. In particular, the moulting process characterizes the development from each larval stage to another, sometimes even occurring within each larval stage, and it continues also throughout adult life (Buchholz CM, Pehlemann FW & Sprang RR, 1989). This process of shrinking and growing involves the synthesis of new cuticle, which is a complex matrix with different layers characterized by α-chitin microfibrils forming a protein matrix. In general, all euphausiidis moult cycle follows the same steps series: at first, a postmoult step, characterized by a soft cuticle; an intermoult stage, with a hard cuticle; the early premoult, followed by late premoult; finally, the ecdysis stage, when krill lose the old cuticle (Rosato et al., 2010).

The furcilia stage is the last larval stage, followed by a small and immature juvenile. The juveniles develop their gonads during Antarctic spring/summer and begin to spawn at two years of age (Everson, 2000; Kawaguchi, et al., 2011).

As a common characteristic in krill species, during life cycle there are several spawning events occurring. In particular, Antarctic krill females produce eggs periodically and they are laid in deep waters usually between December and March. Similarly to other krill species, including the ones considered in this work, their spawning and moulting processes are influenced by water temperatures; specifically, warmer temperature appear to increase these events (Cuzin-Roudy, 2000; Kawaguchi, et al., 2011; Knox, 1994).

Due to its crucial ecological role in the Antarctic ecosystem during the years several studies have been carried out in order to characterize krill distribution (Marr 1962; Siegel 2005;

Atkinson et al. 2008; Hofmann & Murphy 2004), population dynamics and structuring (Valentine & Ayala 1976; Bortolotto et al. 2011) and above all understand its complex genetics (Zane et al. 1998; Goodall-Copestake et al. 2010; Batta-Lona et al. 2011; Bortolotto et al. 2011). Among them, a consistent part is represented by DNA-based studies, specifically focusing on mtDNA variation; however, the information available about krill genetics still remain relatively modest. The difficulty in reconstructing this kind of resources relies on the extraordinarily large Antarctic krill genome size (Jeffery 2011), which is more than 15 times larger than human genome. It has been suggested that the large genome size could be related to the rapid accumulation of transposable elements and repetitive DNA, as the sequencing of a small portion revealed no protein coding sequences, but instead only novel non-coding regions (Jarman et al., 1999).

Meganyctiphanes norvegica

Northern krill *Meganyctiphanes norvegica* (M. Sars, 1857) is an Euphusiidis krill species common in North Atlantic and in the Mediterranean Sea, found also in the North Sea and in the Skagerrak. According to information found in literature, it seems that from both morphological and molecular studies the Northern krill represents a distinct Euphausiacea lineage, more related to Thysanoessa rather than to Thysanopoda. It is considered one of the most wide-spread euphausiid species in the northern hemisphere and this large range of distribution exposes these organisms to a large variety of different bathymetric, hydrographic, seasonal, and trophic conditions. However, besides food availability, temperature is the major abiotic factor that directly and indirectly dominates the physiological ecology of this euphausiid. In general, krill occur in waters from 2°C to 15°C (Einarsson 1945; Lindley 1982), experiencing both low and high temperatures which affect in different ways life cycle and metabolic processes. Although the wide range of distribution and the differences between the geographical areas in terms of food availability, temperature and light exposure, the developmental process and the main steps of reproduction are the same (Cuzin-Roudy and Buchholz, 1999).

Similarly to *E. superba*, it has been demonstrated that Northern krill release eggs seasonally in multiple spawning events (Couzin-Roudy, 2000). Available data show that during the reproductive season male individuals continuously produce spermatozoa packed into spermatophores. These spermatozoa are then transferred to the females, where ovarian cycles functionally follow the moult cycles with spawning occurring in the early premoult step and vitellogenesis between the spawning periods. Interestingly, it has been demonstrated that higher temperatures reduce the duration of moult and ovarian cycles (Cuzin-Roudy et al., 1999; Cuzin-Roudy and Buchholz, 1999; Albessard et al., 2001; Tarling and Cuzin-Roudy, 2003).

Regarding the available resources for *M. norvegica*, in 2011 a study carried out by Jeffery (2011) aimed at estimating the genome size of this species, resulting in a haploid genome size of about 18 Gb (Gigabases) with a haploid chromosome number of 19.

Thysanoessa inermis

Among all the three species under study, *Thysanoessa inermis* is the one with less information available in the literature. This species remains poorly studied, both in terms of genomic and transcriptomic data, therefore few is known about its seasonal distribution and adaptation, as well as the expression pattern under its capability to adapt to the environmental changes. It is known that in the sub-Arctic area, together with *Meganyctiphanes norvegica*, *T. inermis* represent the most abundant krill species (Einarsson, 1945). Spawning occurs in the spring, at the end of the first and second years, and the young stages are found near the surface during the summer. Growth in both years is rapid during the summer months with little or no increase in size during the winter. The growth rate of the female surpasses that of the males in the second year (Kulka, D W & Corey S, 1978).

As observed in the Gulf of Alaska by Pinchuk & Hopcroft in 2006, gravid females can be found only during April and May, when the spawning actually occurs at the same time of the spring bloom of large diatoms, an evidence already described in other works carried out in the subarctic Atlantic and Pacific (Kulka and Corey, 1978, Hanamura et al., 1989; Timofeev, 1996).

Literature lacks information about the genomic or transcriptomic resources for this species, although the genome size estimates for *Thysanoessa* sp. appear to be about 13 Gbp (Jeffery 2011). Therefore, few or no progresses have been done to our knowledge about the molecular mechanisms driving its responses to seasonal changes as well as the expression profiles describing its development and life cycle.

State of the art and evaluation of the sequencing approaches

As mentioned in the previous paragraphs, the genetic resources and knowledge about the molecular mechanisms underlying the complex dynamics of these species are relatively modest. Although *E. superba* and *M. norvegica* are the most well-studied euphausiidis species from an ecological and economical point of view, little is still known about them in terms of genetics. We can identify some main reasons why it is difficult to produce such kind of resources. First, the large genomes, particularly the one from *E. superba,* could make a genome sequencing and assembly project unfeasible and challenging, both in terms of costs and bioinformatic applications. In addition, although DNA represents a mirror of all cellular processes and its sequencing gives us the genetic profile of an organism, it is with RNA sequencing that we can have information about the sequences that are actively expressed in the cells, tissues, or organism at a specific time or in different conditions (Wang, Gerstein,

& Snyder, 2009). This gives us the advantage to reconstruct the transcriptome of an organism, which means producing the complete set of RNA molecules (in other terms, all the genes' readouts) in a cell for a specific developmental stage, a specific physiological condition or derived from a specific experimental treatment. The result, in terms of information obtained and analyses that we can perform, is that we can quantify the changing expression levels of each transcript during development and under different conditions. Therefore, this kind of approach allows us to understand the molecular mechanisms that drive specific capabilities of the species to adapt to different environmental conditions and changes, which is important in the perspective of an accurate management of their stocks and conservation.

With the advances in high-throughput RNA-sequencing techniques - different Antarctic krill transcriptomes have been developed (De Pittà et al., 2008; Seear et al., 2010; Clark et al. 2011; De Pittà et al. 2013; Meyer et al., 2015; Martins et al., 2015). However, it is with the KrillDB project (Sales et al., 2017) that a detailed and advanced genetic resource was produced and made available to the community as an organized database, the KrillDB website. Here, I expanded the amount of transcriptomic information and expression data available for Antarctic krill, thus updating KrillDB to a new and more complete version.

As for *E. superba*, the large genome size of *M. norvegica* makes a project of RNA-sequencing much more advantageous. However, literature is still very poor in this context and today only one *M. norvegica* transcriptome resource has been produced (Blanco-Bercial & Maas, 2018). In this work the authors reconstructed a total of 405,497 transcripts, with 319,012 corresponding genes. The poor genomic resources available for this and other related species, together with the use of one computational method for the reconstruction, probably led to the functional annotation of a limited number of sequences (16%). Here, I produced a more accurate and complete set of Northern krill putative transcripts, enhancing the power to identify processes or functions in which they may be involved and improving the interpretability of expression patterns.

Another explanation for the lack of genetic resources and the few transcriptomic data produced is that, specifically in the context of *E. superba* studies, collecting samples from the Southern Ocean can be difficult, especially in some regions such as Eastern Antarctica which is more than 8000 km far from the Antarctic circle.

Research objectives

This gap in the knowledge of such ecologically important species makes it difficult to identify the best strategies in terms of assessment and management of their stocks and in order to preserve their abundances from decline. We have already seen that, among the major threats, krill species abundance and conservation are affected by ocean acidification, but also by increasing commercial fishing. For these reasons the CCAMLR is constantly monitoring krill catches to minimize the impact on the ecosystem. However, a deeper knowledge of krill genetics and population dynamics would significantly help the management and conservation efforts.

Working with non-model organisms we do not have at our disposal a suitable reference genome. As already mentioned, RNA-seq thus becomes a powerful tool to reconstruct and quantify the whole transcriptome. Nowadays, RNA-seq uses Next Generation Sequencing (NGS) techniques, which are massive parallel sequencing procedures where several sequencing reactions are parallelized in order to generate hundreds of megabases to gigabases of nucleotide sequence reads in a single instrument run. A typical RNA-Seq experiment starts with the isolation of RNA, which is then converted into complementary DNA (cDNA); then, the sequencing library is prepared and is finally sequenced on an NGS platform (Kukurba & Montgomery, 2015). The most commonly used techniques are the

Illumina systems, and the sequence data produced are the so called "*reads*", which are raw sequences that may consist of multiple segments. The reads represent the starting point for the transcriptome reconstruction, which helps us assign these raw sequences to genomic features. If the genome is unknown, as in this case, the technique used consists in a "*de novo* transcriptome assembly" reconstruction. A series of algorithms have been produced and improved over the years, through the development of new optimized software and the identification of strategies that have helped minimize the number of artifacts and errors included in the reconstructed transcriptome.

However, no assembly tool is optimal for all RNA-Seq datasets, mainly because there are a series of different parameters and characteristics that must be taken into account, such as the species of interest or the sequencing protocols. Over the years, one of the most useful applications has become the combination of multiple assemblies through a step called "meta-assembly" (Lu, Zeng & Shi, 2013). In particular, it is with EvidentialGene (Gilbert, 2013), that an accurate pipeline for the management of multiple transcripts set into a biologically best set of mRNAs was introduced. The need for such a strategy derives from the evidence that assemblers all have different biases which can cause them to miss some real sequences or, on the other hand to erroneously assemble ("misassemble") sequences by forming chimeras. In particular, recent benchmarks such as Hölzer & Marz (2019) have shown that, while reconstructing the transcriptome of a species, no single approach is uniformly superior: the quality of each result is influenced by a number of factors, both technical (k-mer size, strategy for duplicate resolution) and biological (genome size, presence of contaminants).

The goal of this approach is to overcome the different disadvantages of certain assemblers and to combine their advantages in a comprehensive *de novo* transcriptome assembly. In particular, EvidentialGene clusters fragments according to their similarity in terms of sequence and then uses the presence and the length of putative coding sequences (CDS) to select the best fragment for a specific transcript among all the available assemblies. It is easily predictable, at this point, that as we increase the number of assemblies used, it will contextually introduce a lot of redundancy. This is a crucial issue: as demonstrated by the results reported in this work, there are still some limitations to the whole procedure, especially when working with huge amount of data and non-model organisms with significantly large genomes.

For all these reasons, the aim of my work is comprised of four main parts:

1) The identification of the most accurate transcriptome reconstruction procedure by implementing the use of already existing methods with intermediate steps aiming at reducing the mis-assemblies and redundant sequences.
2) The application of the new pipeline on three krill species and the demonstration of the reliability of the newly reconstructed transcriptomes in terms of quality metrics.
3) The quantification of the power of the new procedure and of the filtering approaches to produce a reference that can enhance our possibility to identify important transcriptional phenotypes that have not been previously described.
4) The collection of all the information and data obtained in a unique and complete resource, easily available and accessible by all scientific community.

In the perspective of producing new annotation data, I therefore considered the possibility to introduce an improvement of the online available resources for *Euphausia superba*. As previously mentioned, an accurate Antarctic krill transcriptome reconstruction was already performed by Sales et al. (2017), using two larval stages. The data produced have been collected in KrillDB, which is a web-graphical interface with annotation results coming from the *de novo* reconstruction of krill transcriptome, assembling more than 360 million Illumina sequence reads. Here, I used it as a starting point to re-design the website by updating the number and variety of information available and increasing the user search possibilities.

Krill sampling and sequencing

Euphausia superba

This study aims at covering the entire developmental process of Antarctic krill. Therefore, I used samples coming from different developmental stages to cover the entire *E. superba* transcriptome, from larval to adult specimens. Specifically, adults included both male and female specimens, as well as summer and winter individuals and they also came from 3 different geographical regions: Lazarev Sea (62°S -66°S), South Georgia (54°S), and Bransfield Strait/South Orkney (60°S-63°S).

Specifically, a part of the Antarctic krill samples was the same used in a previous work (Höring et al., 2021) and they were obtained from five different expeditions and from a Norwegian fishing vessel. Then snap-frozen Antarctic krill samples stored at −80 °C were transferred to the Alfred-Wegener-Institute (Bremerhaven) for molecular analysis.

Total RNA from these samples was individually extracted from frozen krill heads with the RNeasy Midi Kit (QIAGEN, Hilden, Germany). The RNA-Seq of each individual sample was carried out from IGA Technology Services (Udine, Italy). cDNA libraries were constructed with 1–2 µg of total RNA by using the TruSeq Stranded mRNA Sample Prep kit (Illumina, San Diego, CA, USA). Sequencing was carried out on paired-end mode (2 × 100 bp) by using HiSeq 2500 (Illumina) with a targeted sequencing depth of about 80 million reads *per* sample. The paired-end sequencing implies that the reading starts at one read following its direction until the specified read length, and then another round of reading starts from the opposite end of the fragment. This kind of sequencing, compared to the single-end technique, improves the ability to identify the relative positions of various reads in the genome/transcriptome, which enhance the possibility to detect rearrangements and repetitive sequence elements, gene fusions and novel transcripts.

Larval samples came from a different experiment carried out in April 2011, where adult individuals were collected from the Indian Ocean sector of the Southern Ocean (64°09’ S, 100°460 E); then they were maintained in the Australian Antarctic Division’s marine research acquarium, were they matured in laboratory. Specifically, gravid females spawned between December 2011 and February 2012. Finally, batches of larvae obtained were reared through different stages of development over several months. Total RNA was extracted using the ZR-Duet™ RNA miniPrep procedure (Zymo Research Corporation) and samples were sent to GeneWorks (Australia) for sequencing, using a TruSeq RNA sample prep kit (Illumina) and for each sample one lane of paired-end reads (2x100 bp) was produced in a Genome Analyzer IIx sequencer (Illumina).

In total, more than 6 billion of reads were produced from all the experiments considered, which represents an unprecedented number of samples. Such a huge number of input data represented one of the challenges of this work, specifically in terms of management from a computational point of view (e.g., time and memory usage).

Meganyctiphanes norvegica

Norther krill dataset included 36 samples provided by AWI Institute and University of Bremerhaven and the ICBM (Institute for Chemistry and Biology of Marine Environment, Oldenburg). Samples were characterized by male and female specimens, collected at two different time points (May and November) and representing three different tissue types (sex, carapax and head). In addition, 12 samples were downloaded from NCBI BioProject PRJNA324094, coming from a previous physiological and gene expression study on *M. norvegica* (Blanco-Bercial & Maas, 2018).

AWI dataset included specimens sampled at Gullmarsfjord in cooperation with the Sven Lovén Centre of Marine Research (Kristineberg, Sweden). Sampling was conducted during a 24-hour sampling campaign during spring (May) and winter (November/December) 2017. It was conducted in the center of the fjord (58°19 `17.7 N, 11°32`68.7 E) using an Isaacs Kidd Midwater Trawl (IKMT). Immediately after catch, individuals were transferred to bins containing filtered fjord water. Total catch was recorded, and subsamples were taken for the determination of standard length [mm], sex as well as maturity stage. Individual krill were then immediately stored in RNAlater™ or snap-frozen and stored at -80°C for further analysis. RNA was extracted from the head, the carapace and sexual organs, respectively. Individuals either stored in RNAlater™ (Thermo Fisher Scientific, USA) or snap frozen directly after catch were used. Before dissection, snap-frozen samples were thawed in RNAlater™- ICE (Thermo Fisher Scientific, USA) at -20°C for at least 24h. Dissection was performed using a binocular (Leica MZ125) and a cooled, RNAse-free petri dish. Stomach, intestines as well as eyeballs were carefully removed. After dissection, tissues were directly transferred into Precellys® tubes (1.4 mm and 2.8 mm ceramic mixed beads) containing TRIzol® Reagent. Tissues were then homogenized using a Precellys ® 24 homogenizer (Bertin Technologies, France). The homogenate was transferred into 2.0 ml RNase-free Eppendorf® tubes (Eppendorf, Hamburg, Germany) and chloroform (Sigma-Aldrich, USA) was added in a ratio of Trizol:chloroform = 5:1. Samples were centrifuged at 12,000 g for 15 min at 4°C for phase separation. A fixed volume of the upper aqueous phase was transferred to a fresh 1.5 ml RNase-free Eppendorf® tube. RNA was then extracted using the Direct-zol™RNA Miniprep Plus Kit (Zymo Research, USA) according to the manufacturers protocol. RNA quantity and quality of each sample were determined using a Nanodrop 2000 (Thermo Fisher Scientific, USA) and the Agilent Bioanalyzer 2100 (Agilent Technology). RNA was finally sequenced on a HiSeq2500 platform, producing 80 million paired-end reads 125bp, for a total of 36 samples (n=6 per season; 3 males/3 females each; 3 tissues each).

Thysanoessa inermis

A total of 36 *T. inermis* samples coming from AWI/ICBM cruises were used to *de novo* assemble the transcriptome. Both male and female individuals were included into the dataset, sampled at two different time points (April and July) and, as for *M. norvegica*, coming from three different tissue types (sex, carapax, head). Moreover, a small dataset of 4 samples coming from an experiment of deep sequencing (100 million reads/sample) was added to increase the potential of an accurate transcriptome reconstruction. The samples for deep sequencing were kindly provided by Flores, H. et al; sampled in June 2015 during Polarstern cruise PS92 at 81° 17.79' N 17° 6.17' E using a MRMT. RNA was extracted from the head, the carapace and sexual organs, using the same procedure applied to *M. norvegica* samples. For deep sequencing of *T. inermis*, whole-animal RNA was extracted from 4 individuals (2 males/2 females). RNA was then sequenced on a HiSeq2500 platform, resulting in 40 million paired-end reads 125bp; 36 samples (n=6 per season; 3 males/3 females each; 3 tissues) and 100 million paired-end reads 125bp for deep sequencing.

Transcriptome assembly strategy

De novo assembly test

As a starting point, seven different *de novo* assemblers were selected, among the most used and referenced found in literature:

1) Trinity (Grabherr et al. 2011). It is one of the most used, documented and powerful *de novo* assembly software, and it consists of three software components (*Inchworm, Chrysalis,* and *Butterfly*) applied sequentially. Inchworm assembles reads into the unique sequences of transcripts, using a greedy k-mer based approach for fast and efficient transcript assembly, recovering only a single (best) representative for a set of alternative variants that share k-mers (due to alternative splicing, gene duplication, or allelic variation). K-mers represent substring of length k contained within the sequence. Then, Chrysalis clusters related sequences that correspond to portions of alternatively spliced transcripts or unique portions of paralogous genes. Based on these results, a de Bruijn graph for each cluster of related sequences is constructed, each reflecting the complexity of overlaps between variants. De Bruijn graph are in fact kind of graphs used in bioinformatics to assemble sequencing reads, representing overlap between sequences. Finally, Butterfly analyzes the paths taken by reads and read pairings in the context of the corresponding de Bruijn graph and reports all plausible transcript sequences, considering both alternatively spliced isoforms and transcripts derived from paralogous genes.
2) Oases-Velvet (Schulz et al. 2012); it is a well-known *de novo* assembler for next generation sequencing short reads. It builds on two algorithms: Velvet, which is a *de novo* assembler based on de Bruijn graphs, as Trinity; Oases, a software consisting of independent assemblies, which vary according to the *k*-mer length. In each of the

assemblies, the reads are used to build a de Bruijn graph, which is then simplified for errors, organized into a scaffold, divided into loci and finally analyzed to extract transcript assemblies or the so-called *transfrags*. Finally, the individual k-mer assemblies are merged into a final assembly.

3) BinPacker (Liu et al. 2016), which is an update of the previous Bridger software and it is based on a splicing graph construction instead of de *Bruijn* graphs. This approach differs from those previously described as it integrates coverage information into the procedure. In particular, only splicing junctions are involved in the assembling procedure and through the use of a rigorous mathematical model, called the minimum path cover, the software searches for a minimal set of paths (transcripts) that are supported by the provided RNA-seq reads and can explain all the observed splicing events of the created graph.
4) rnaSPAdes (Bushmanova et al. 2019), works using a single k-mer size which is automatically tuned depending on the read length. Specifically, it constructs a de Bruijn graph from short reads, then the graph is simplified removing erroneous edges and producing a so-called assembly graph. An alignment of reads to the assembly graph is run and, finally, an isoform reconstruction procedure is performed.
5) TransABySS (Zhao et al. 2011), also based on de Bruijn graph for the transcriptome reconstruction, computing assemblies of substrings using various stringencies and then merging all the separate assemblies into contigs.
6) IDBA-tran (Peng et al. 2013). It also uses multiple de Brujin graphs constructed using distinct k-mer lengths to handle transcripts with different expression levels. It then employs a probabilistic progressive approach to removes erroneous k-mers based on local thresholds. At each run, the algorithm uses the output of the last run rather than raw reads.

7) Drap (Cabau et al. 2017). This software represents a pipeline that wraps two other assemblers, Trinity and Oases, in order to improve their results. It comprises a set of three command-line tools respectively called runDrap, runMeta and runAssessment. runDrap performs the assembly including compaction and correction; runAssessment compares different contig sets and gathers the results in a global report, while runMeta merges and compacts different contigs sets.

The selected assemblers were tested using a small test dataset of strand-specific RNA-seq reads from *Schizosaccharmomyces pombe*, provided by Haas et al. (2013). Strand specificity is an important parameter in RNA-seq experiments because it adds information on the originating strand, in other terms it identifies the strand of the sequenced transcript. On the contrary, in unstranded libraries the sequences of both sense strand and the antisense strand of the original mRNA are obtained without knowing which strand of the cDNA corresponds to the original mRNA. Therefore, when working with unstranded libraries it is not clear the original strand a read derived from, and this makes it impossible to deconvolve the signals of sense and antisense transcripts harbored by the same genomic locus. Some of the above-mentioned software gives the possibility to specify the type of sequencing protocol used (stranded or unstranded) and exploit this knowledge to improve reconstruction accuracy.

The aim of this pre-processing step was to compare the performances of each software, assess which one produced the most reasonable result and identify the best setup optimization for each *de novo* assembler (both in terms of time and memory/CPUs usage).

Quality assessment

After using the seven algorithms described earlier, I selected a series of independent quality measures to assess the reliability of their reconstructions. Specifically, the quality of a transcriptome can be evaluated using both "internal" measures, which relies on the RNA-seq input data used to perform the *de novo* assembly, and “external” measures, which are based on the use of known protein or nucleotide databases, such as sequences coming from other species and organisms. I used as internal measures a series of basic statistics, such as the number of total transcripts, the %GC content, the average fragment length, the total number of bases and the N50 value; in addition, I also considered the evaluation of the mapping rate of each transcriptome by mapping back the reads on the assembled transcripts. The latter is especially important as it gives an idea of how well the input data is represented by the reconstructed transcriptome. Regarding the external measures, I used the representation of full-length reconstructed protein-coding genes, by searching the assembled transcripts against a database of known protein sequences (i.e., SwissProt, https://www.uniprot.org) in order to examine the number of assembled transcripts that appear to be full-length or nearly full-length. In addition, I ran a BUSCO analysis, which provides a measure of transcriptome completeness based on evolutionarily-informed expectations of gene content from near-universal single-copy orthologs (Simão. et al. 2015). BUSCO (Benchmarking Universal Single-Copy Orthologs) is a tool based on the concept of single-copy orthologs that should be highly conserved among the closely related species. In particular, we refer to single-copy orthologs as the set of genes that have remained in single copy (without duplications or losses occurring) since the last common ancestor at a specific taxonomical range. It relies on OrthoDB, a catalogue of orthologues (https://www.orthodb.org/) to identify complete, duplicated, fragmented and missing genes. The percentage of “complete” results actually represents the number of known single-copy orthologs expected to be in all the species belonging to the selected taxonomic clade that are included in the reconstructed transcriptome. The fraction of “fragmented” results corresponds to the amount of those single-copy orthologs that have been partially reconstructed. Those that have not been reconstructed at all are included in the “missing” category.

On the basis of the results (**Figure 1**; **Table 1**) produced by these quality control analyses, I decided to discard Oases-Velvet and Drap assemblers. The basic statistics shows a very low number of transcripts reconstructed (Drap) and very low median contig value (Oases-Velvet). Together with the results of the analysis of full-length reconstructed transcripts (**Figure 1**), these numbers confirmed that the two algorithms produced less reliable and accurate transcriptome reconstructions with respect to the other software I tested.

	# Total transctits	%GC	Median Contig Length	Average Contig	# Total bases	N50
Trinity	9231	38.03	738	1020.58	9421944	1595
BinPacker	6540	38.05	1188	1386.09	9065012	1838
IDBA-tran	6553	37.87	1057	1347.1	8827551	2007
rnaSPAdes	6959	38.44	556	978.75	6811150	1659
Oases-Velvet	6993	37.78	1075	1349.05	9432539	2175
Drap	3372	38.54	1140	1322.23	4458566	1638
Trans-ABySS	5707	38.08	1277	1466.32	8368277	1922

Table 1. Measure of basic statistics of each transcriptome produced using *S. pombe* dataset (Haas et al., 2013), testing the seven assemblers selected.

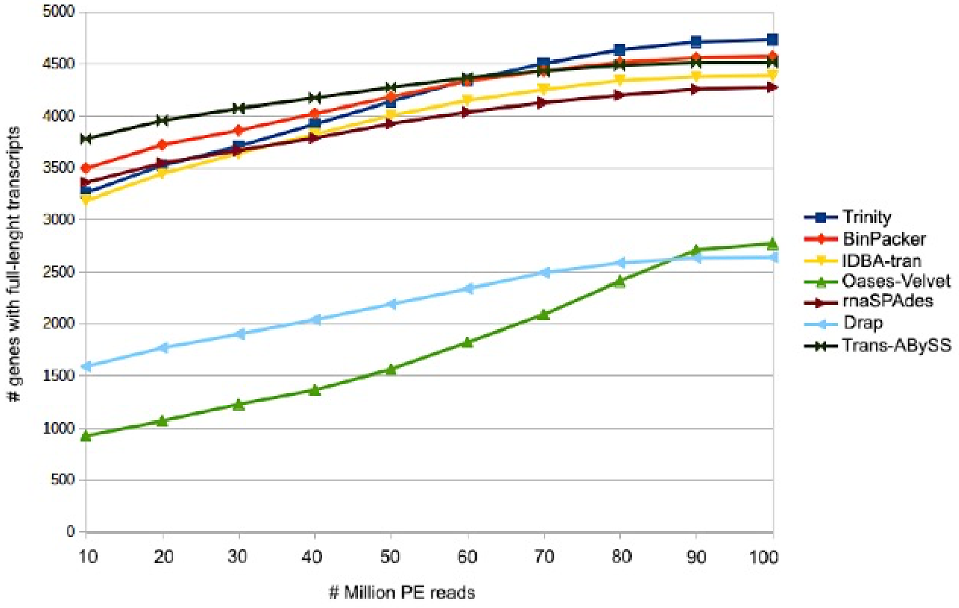


Figure 1. Number of full-length transcripts vs. number of input RNA-Seq fragments, performed separately for each of the assemblies produced from S. pombe dataset (Haas et al., 2013).

Krill *de novo* transcriptome assembly

Once selected the best *de novo* assembly algorithms, I focused on the krill transcriptome reconstruction.

I have summarized all the steps of the assembly reconstruction strategy, annotation process and downstream analyses in **Figure 2**.

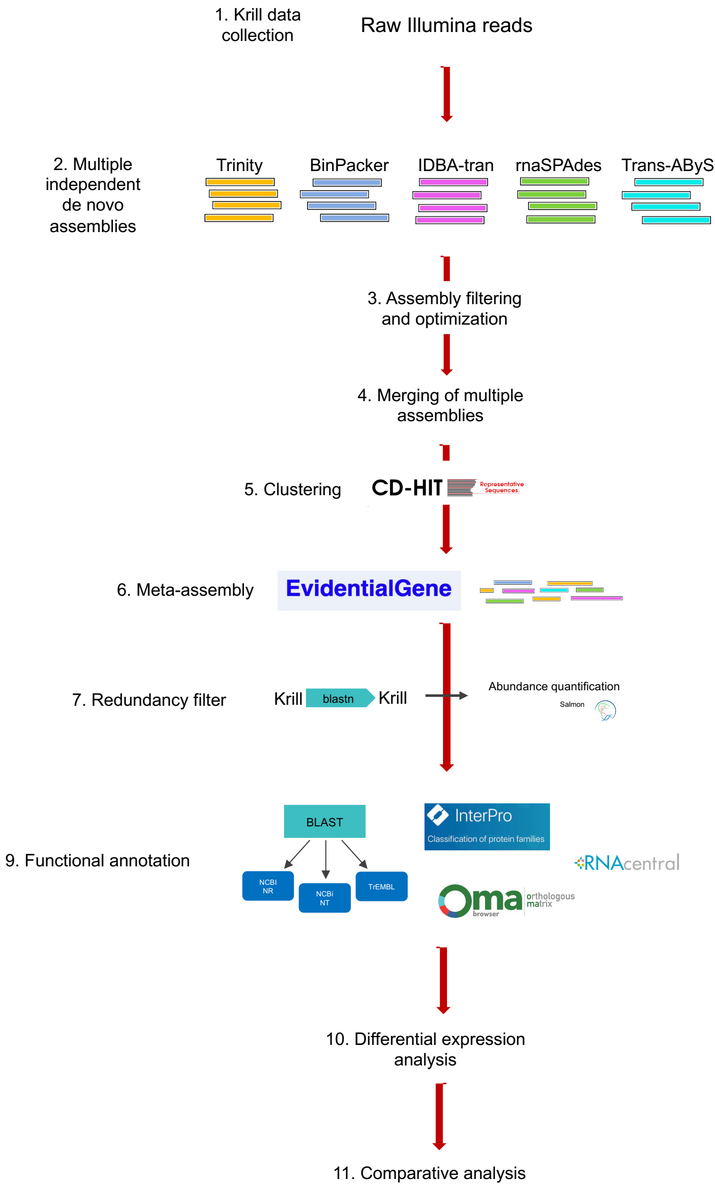
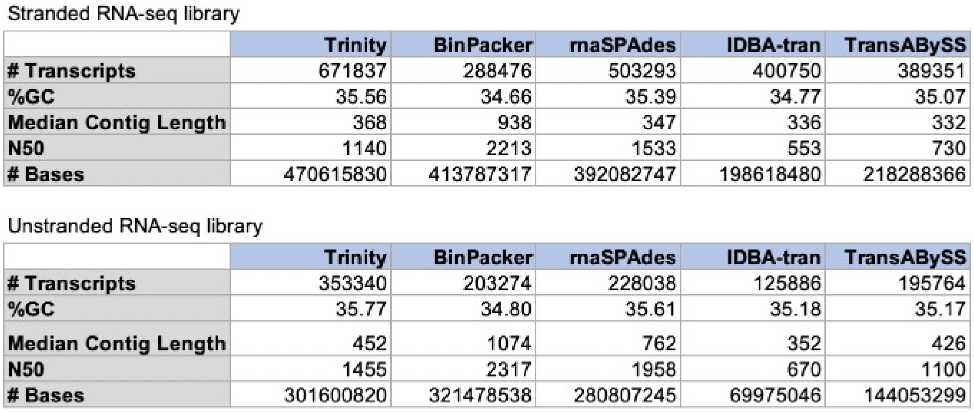


Figure 2. Workflow of the assembly process, annotation, database re-design and downstream analyses.

I first performed a separate transcriptome reconstruction with each of the tools listed above. I evaluated their respective advantages measuring their quality through the computing of the total number of transcripts, their GC content, median contig length, total number of bases and N50 value (**Table 2**, **Table 3**, **Table 4**).

The raw sequencing data I used for the assemblies was obtained from different experiments and included both stranded and unstranded libraries (*E. superba*), different experimental designs (*M. norvegica*) and different sequencing methods (*T. inermis*). As mixing different types of libraries or experiments in a single assembly is not well supported, I decided to run, for each species, each software twice: I thus generated a total of ten different *de novo* assemblies for each species.



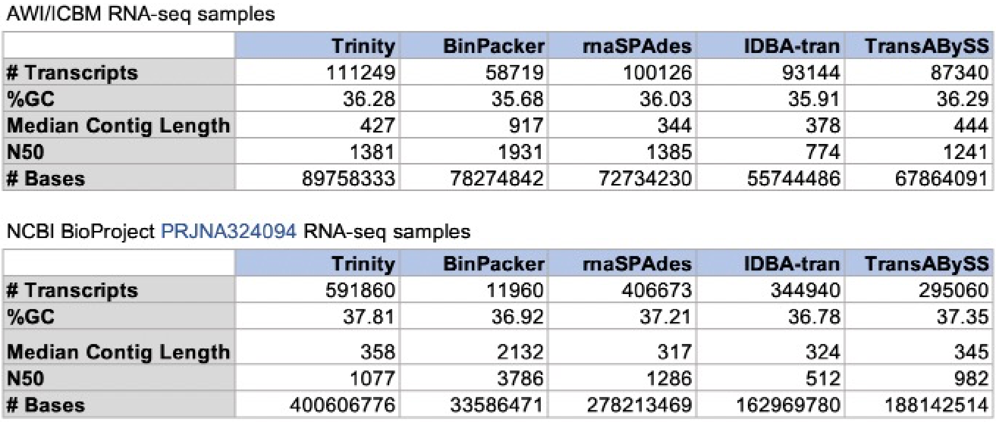
Stranded RNA-seq library

	Trinity	BinPacker	rnaSPAdes	IDBA-tran	TransABySS
# Transcripts	671837	288476	503293	400750	389351
%GC	35.56	34.66	35.39	34.77	35.07
Median Contig Length	368	938	347	336	332
N50	1140	2213	1533	553	730
# Bases	470615830	413787317	392082747	198618480	218288366

Unstranded RNA-seq library

	Trinity	BinPacker	rnaSPAdes	IDBA-tran	TransABySS
# Transcripts	353340	203274	228038	125886	195764
%GC	35.77	34.80	35.61	35.18	35.17
Median Contig Length	452	1074	762	352	426
N50	1455	2317	1958	670	1100
# Bases	301600820	321478538	280807245	69975046	144053299

Table 2. Quality assessment of *Euphausia superba* raw assemblies produced by each of the five assemblers used. *De novo* assembly with the five software was performed independently on samples coming from different RNA-seq libraries.



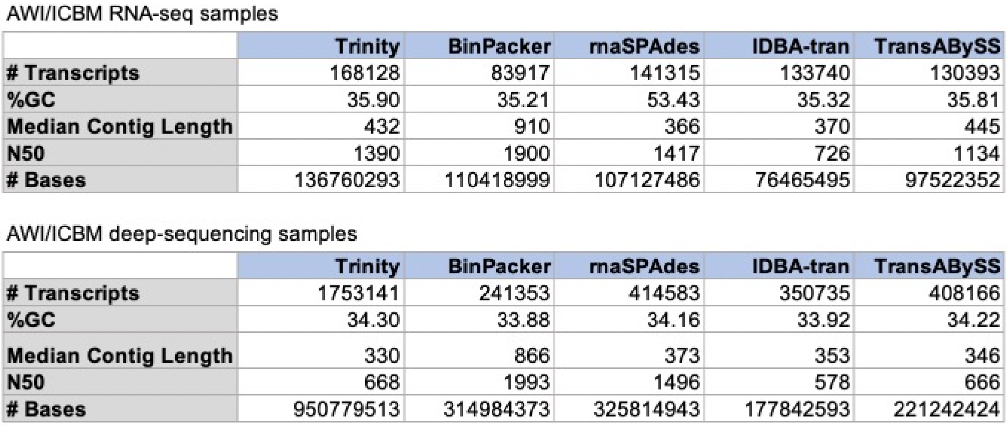
AWI/ICBM RNA-seq samples

	Trinity	BinPacker	rnaSPAdes	IDBA-tran	TransABySS
# Transcripts	111249	58719	100126	93144	87340
%GC	36.28	35.68	36.03	35.91	36.29
Median Contig Length	427	917	344	378	444
N50	1381	1931	1385	774	1241
# Bases	89758333	78274842	72734230	55744486	67864091

NCBI BioProject PRJNA324094 RNA-seq samples

	Trinity	BinPacker	rnaSPAdes	IDBA-tran	TransABySS
# Transcripts	591860	11960	406673	344940	295060
%GC	37.81	36.92	37.21	36.78	37.35
Median Contig Length	358	2132	317	324	345
N50	1077	3786	1286	512	982
# Bases	400606776	33586471	278213469	162969780	188142514

Table 3. Quality assessment of *Meganyctiphanes norvegica* raw assemblies produced by each of the five assemblers used. *De novo* assembly with the five software was performed independently on samples coming from different RNA-seq experiments.



AWI/ICBM RNA-seq samples

	Trinity	BinPacker	rnaSPAdes	IDBA-tran	TransABySS
# Transcripts	168128	83917	141315	133740	130393
%GC	35.90	35.21	53.43	35.32	35.81
Median Contig Length	432	910	366	370	445
N50	1390	1900	1417	726	1134
# Bases	136760293	110418999	107127486	76465495	97522352

AWI/ICBM deep-sequencing samples

	Trinity	BinPacker	rnaSPAdes	IDBA-tran	TransABySS
# Transcripts	1753141	241353	414583	350735	408166
%GC	34.30	33.88	34.16	33.92	34.22
Median Contig Length	330	866	373	353	346
N50	668	1993	1496	578	666
# Bases	950779513	314984373	325814943	177842593	221242424

Table 4. Quality assessment of *Thysanoessa inermis* raw assemblies produced by each of the five assemblers used. *De novo* assembly with the five software was performed independently on samples coming from different sequencing methods.

Then, a combination of two filtering steps was applied to the newly reconstructed transcriptomes in order to discard artifacts and improve the assembly quality.

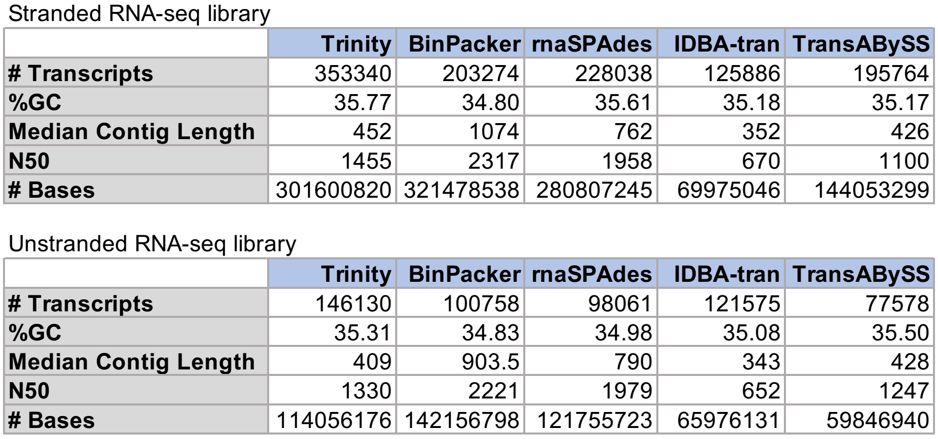
First, I estimated the abundances of all the transcripts reconstructed by each assembler using the Salmon software (version 1.4.0, Patro, Duggal, Love, Irizarry & Kingsford, 2017), which is a popular method for quantifying transcript abundance from RNA–seq reads.

Samples were grouped according to the main experimental conditions:

- *Euphausia superba*: (1) sex, with female and male levels; (2) geographical area, covering Bransfield Strait-South Orkney, South Georgia and Lazarev Sea; and (3) season, with summer and winter levels.
- *Meganyctiphanes norvegica*: (1) sex, with female and male levels; (2) time point, with samples coming from May and November; (3) tissue type, characterized by sex, carapax and head samples.
- *Thysanoessa inermis*: (1) sex, including both male and female individuals; (2) time point, with July and April levels; (3) tissue type, with sex, carapax and head levels.

Abundance estimates were imported in the R statistical software using the tximport package (Love et al. 2017) and I implemented a filter to keep only those transcripts showing an expression level of at least 1 transcript per million (TPM) within each of the three experimental conditions.

The same evaluation of quality in terms of basic internal statistics was performed on the filtered transcriptome (**Table 5, Table 6**, **Table 7**).



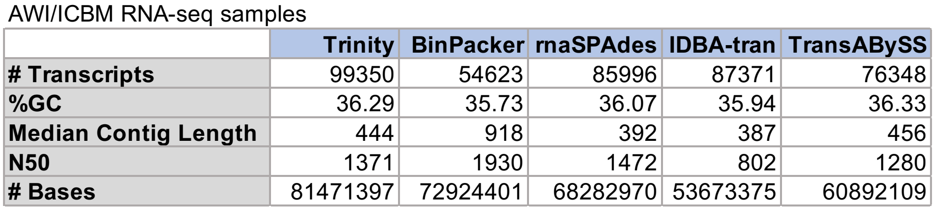
Stranded RNA-seq library

	Trinity	BinPacker	rnaSPAdes	IDBA-tran	TransABySS
# Transcripts	353340	203274	228038	125886	195764
%GC	35.77	34.80	35.61	35.18	35.17
Median Contig Length	452	1074	762	352	426
N50	1455	2317	1958	670	1100
# Bases	301600820	321478538	280807245	69975046	144053299

Unstranded RNA-seq library

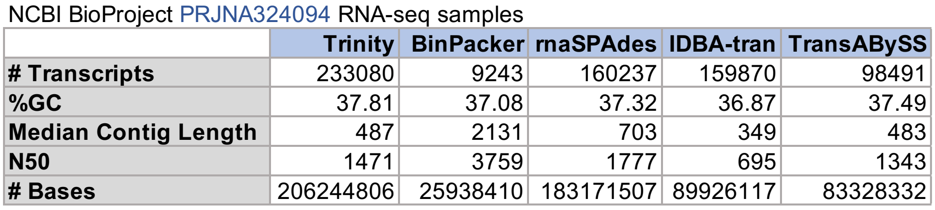
	Trinity	BinPacker	rnaSPAdes	IDBA-tran	TransABySS
# Transcripts	146130	100758	98061	121575	77578
%GC	35.31	34.83	34.98	35.08	35.50
Median Contig Length	409	903.5	790	343	428
N50	1330	2221	1979	652	1247
# Bases	114056176	142156798	121755723	65976131	59846940

Table 5. Quality assessment of *Euphausia superba* raw assemblies after low-abundance-filter.



AWI/ICBM RNA-seq samples

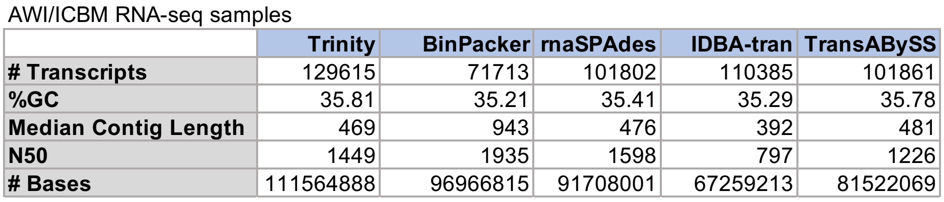
	Trinity	BinPacker	rnaSPAdes	IDBA-tran	TransABySS
# Transcripts	99350	54623	85996	87371	76348
%GC	36.29	35.73	36.07	35.94	36.33
Median Contig Length	444	918	392	387	456
N50	1371	1930	1472	802	1280
# Bases	81471397	72924401	68282970	53673375	60892109



NCBI BioProject PRJNA324094 RNA-seq samples

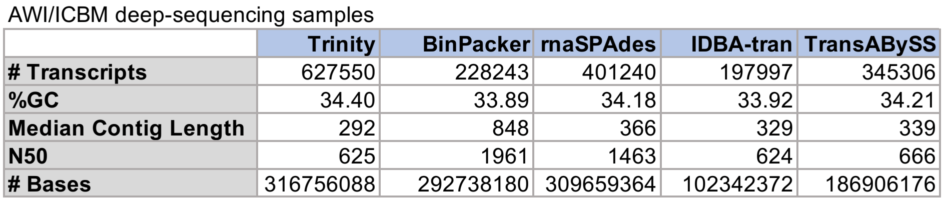
	Trinity	BinPacker	rnaSPAdes	IDBA-tran	TransABySS
# Transcripts	233080	9243	160237	159870	98491
%GC	37.81	37.08	37.32	36.87	37.49
Median Contig Length	487	2131	703	349	483
N50	1471	3759	1777	695	1343
# Bases	206244806	25938410	183171507	89926117	83328332

Table 6. Quality assessment of *Meganyctiphanes norvegica* raw assemblies after low-abundance-filter.



AWI/ICBM RNA-seq samples

	Trinity	BinPacker	rnaSPAdes	IDBA-tran	TransABySS
# Transcripts	129615	71713	101802	110385	101861
%GC	35.81	35.21	35.41	35.29	35.78
Median Contig Length	469	943	476	392	481
N50	1449	1935	1598	797	1226
# Bases	111564888	96966815	91708001	67259213	81522069



AWI/ICBM deep-sequencing samples

	Trinity	BinPacker	rnaSPAdes	IDBA-tran	TransABySS
# Transcripts	627550	228243	401240	197997	345306
%GC	34.40	33.89	34.18	33.92	34.21
Median Contig Length	292	848	366	329	339
N50	625	1961	1463	624	666
# Bases	316756088	292738180	309659364	102342372	186906176

Table 7. Quality assessment of *Thysanoessa inermis* raw assemblies after low-abundance-filter.

Meta-assembly strategy

In a second step, I considered the results of all assemblers jointly, and I ran the "cd-hit-est" program (Li & Godzik, 2006) in order to cluster similar sequences and to produce a set of non-redundant representative transcripts. Specifically, I collapsed all sequences sharing 95% or more of their content.

First attempts and tests in order to find the best combination of steps and to optimize the transcriptome reconstruction procedure were applied on the Antarctic krill data. At this point, I managed to reduce the total number of transcripts from 1,650,404 to 551,110.

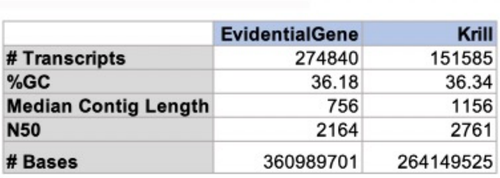
The procedure described above was designed to identify near-duplicate sequences deriving from different software, but likely corresponding to the same biological transcript. As a further refinement, I was also interested in grouping resulting transcripts into units corresponding to genes. To this end, I used the EvidentialGene pipeline (Gilbert 2013, 2019). I applied the "*tr2aacds*" tool which clusters transcripts and classifies them to identify the most likely coding sequence representing each gene. The software subdivides sequences into different categories, including primary transcript with alternates (*main*), primary without alternates (*noclass*), alternates with high and medium alignment to primary (*althi1*, *althi*, *altmid*) and partial (*part*) incomplete transcripts. A "coding potential" flag is also added, separating coding from non-coding sequences (see section "KrillDB^2 Web Interface"). The meta-assembly thus obtained consisted in 274,840 putative transcripts, subdivided into 173,549 genes.

As these figures remained unrealistically high, I performed another round of analyses to identify redundant or mis-assembled sequences still appearing in the Antarctic krill transcriptome. Here I used a combination of BLAST searches against known protein and nucleotide databases (NCBI NR, NCBI NT, UniProtKB/TREMBL) and information deriving from full-length, experimentally validated transcripts from a previous study (Biscontin et al., 2017). Results confirmed that the newly reconstructed transcriptome fully represented krill RNAs, but the large amount of input reads, together with the number of independent *de novo* assemblers, likely led to an inflation in the number of alternative splicing variants being reconstructed. Moreover, transcript alignments against BUSCO genes (Simão et al., 2015) and the doubletime, cry1, shaggy and vrille full-length transcripts from Biscontin et al. (2017) highlighted the fact that multiple fragments of the same gene were incorrectly assembled as separate transfrags. To remove these artifacts, first I aligned all transcript sequences in the dataset against each other using the *blastn* tool. I discarded all sequences already included in a longer transcript for more than the 90% of their length. This filter helped me remove 78,731 redundant sequences (29% of transcripts, overall). Then, I ran a new abundance quantification using Salmon, discarding all transcripts with an average abundance below 0.1 TPM.

The combination of all the filters discussed above allowed me to reduce the number of Antarctic krill transcripts to 151,585 and, correspondingly, that of genes to 85,905. I tested the power of the strategy applied by comparing the quality of the EvidentialGene transcriptome, both in terms of internal basic statistics and completeness according to the number of single-copy orthologs represented searching our sequences among all expected orthologs from Arthropoda phylum (**Table 8**, **Figure 3**). These results highlighted an increase of the assembly quality due to the filters applied. In this study I observed that, although a consistent number of sequences was removed through each step of assembly, merging and filtering, I did not find any decline in the quality described by the basic statistics of the reconstructed transcripts.

In particular, the results of the analyses comparing the EvidentialGene reconstructed transcriptome to the redundancy-filtered assembly in terms of completeness according to single-copy orthologs showed a minimal reduction of complete matches to the BUSCO profile for *E. superba*.

Specifically, a total number of 274,840 *E. superba* transcripts was produced by the EvidentialGene pipeline. Then, the redundancy filter helped me remove 45% of the total transcripts reconstructed, thus increasing the median contig length from 756 to 1,156 and N50 value from 2,164 to 2,761. Transcriptome completeness in terms of amount of complete single-copy orthologs reconstructed in krill transcriptome showed that the loss in the amount of sequences produced a little decrease in this value (from 95.3% to 93.2%). In addition, the redundancy filter barely affected the average mapping rate value, reducing it from 89% (EvidentialGene only) to 88% (full filtering).



	EvidentialGene	Krill
# Transcripts	274840	151585
%GC	36.18	36.34
Median Contig Length	756	1156
N50	2164	2761
# Bases	360989701	264149525

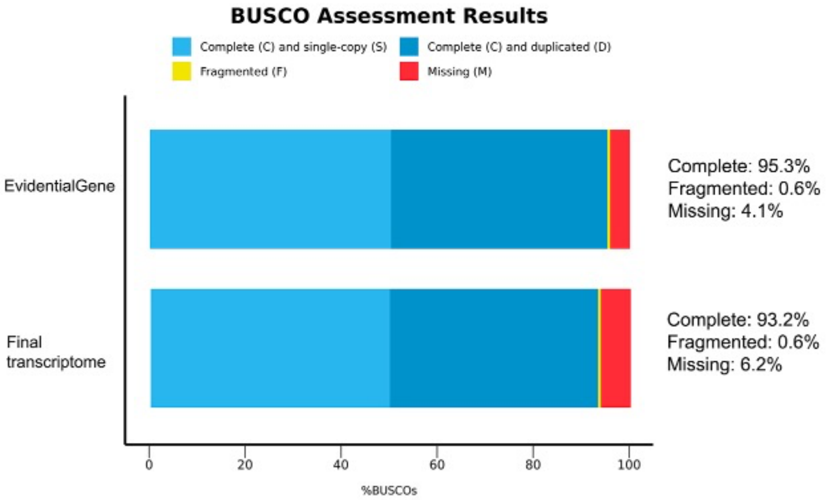
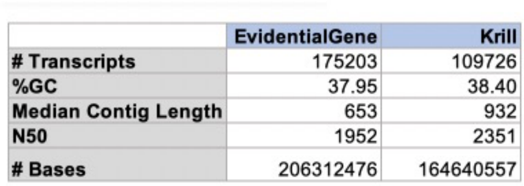


Table 8. Table (up) shows a comparison between the EvidentialGene assembly and the final transcriptome of Antarctic krill in terms of basic statistics. **Figure 3**. BUSCO assessment results (bottom) on EvidentialGene transcriptome and krill transcriptome after last filter: the EvidentialGene transcriptome was characterized by 95.3% Complete sequences (50.2% single-copy, 45.1% duplicated), 0.6% Fragmented and 4.1% Missing sequences. The same analysis on the final krill transcriptome reconstruction produced 93.2% Complete transcripts (49.8% Single-copy, 43.4% Duplicated), 0.6% Fragmented and 6.2% Missing sequences.

The same approach was then applied on the datasets from the other two krill species. In particular, the sequence of assembly, filtering and meta-assembly steps reduced the total number of Northern krill transcripts/genes from 175,203/123564 (EvidentialGene) to 109,726 transcripts and 73,624 corresponding genes. Although the 37% of the transcripts was removed, both median contig length (from 653 to 932) and N50 value (from 1,952 to 2,351) increased; moreover, a little increase was registered also in the number of complete results from BUSCO analysis, moving from 94.4% to 94.6% (**Table 9**, **Figure 4**). The average mapping rate of the final transcriptome was around the 86%, the same as for EvidentialGene only.



	EvidentialGene	Krill
# Transcripts	175203	109726
%GC	37.95	38.40
Median Contig Length	653	932
N50	1952	2351
# Bases	206312476	164640557

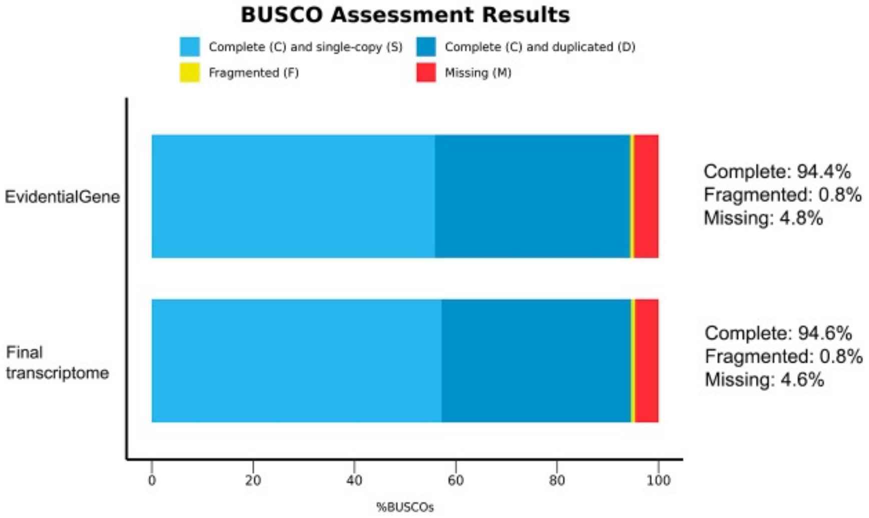
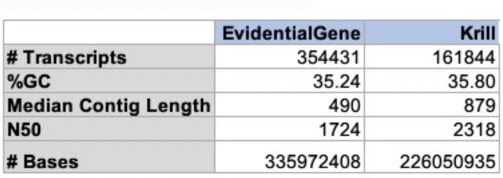


Table 9. Table (up) shows a comparison between the EvidentialGene assembly and the final transcriptome of Antarctic krill in terms of basic statistics. **Figure 4**. BUSCO assessment results (bottom) on EvidentialGene transcriptome and krill transcriptome after last filter: the EvidentialGene transcriptome was characterized by 94.4% Complete sequences (55.9% single-copy, 38.5% duplicated), 0.8% Fragmented and 4.8% Missing sequences. The same analysis on the final krill transcriptome reconstruction produced 94.6% Complete transcripts (57.2% Single-copy, 37.4% Duplicated), 0.8% Fragmented and 4.6% Missing sequences.

The *Thysanoessa inermis* transcriptome, as reconstructed by EvidentialGene, consisted of 354,431 total sequences; the redundancy filter removed the 54% of those transcripts, thus resulting in a final transcriptome with 161,844 sequences. Both median contig length and N50 value followed the same trend as for *E. superba* and *M. norvegica*, increasing the first from 490 to 879 and the second from 1,724 to 2,318. BUSCO results showed a little decrease of the number of complete single-copy orthologs present in the transcriptome, from 91.8% to 91.5% (**Table 10**, **Figure 5**); of the reads that contributed to the EvidentialGene meta-assembly 86% mapped back, but after the redundancy filter the average mapping rate increased to 89%.



	EvidentialGene	Krill
# Transcripts	354431	161844
%GC	35.24	35.80
Median Contig Length	490	879
N50	1724	2318
# Bases	335972408	226050935

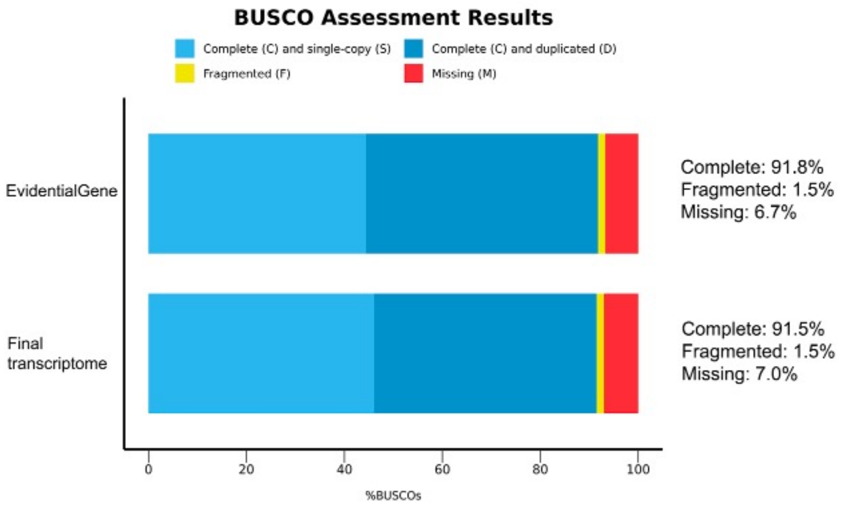


Table 10. Table (up) shows a comparison between the EvidentialGene assembly and the final transcriptome of Antarctic krill in terms of basic statistics. **Figure 5**. BUSCO assessment results (bottom) on EvidentialGene transcriptome and krill transcriptome after last filter: the EvidentialGene transcriptome was characterized by 91.8% Complete sequences (44.4% single-copy, 47.4% duplicated), 1.5% Fragmented and 6.7% Missing sequences. The same analysis on the final krill transcriptome reconstruction produced 91.5% Complete transcripts (46.1% Single-copy, 45.4% Duplicated), 1.5% Fragmented and 7.0% Missing sequences.

As demonstrated by all these data, the approach I developed retained alternative and paralog transcripts with sufficient level of uniqueness in their sequence, as confirmed by the fact that although I removed almost 45%, 37% and 54% of the initially assembled transcripts of *E. superba*, *M. norvegica* and *T. inermis* respectively, the drop barely affected the average mapping rate value, reducing it from about 89% (EvidentialGene only) to about 88% (full filtering).

Functional annotation

At this point, it was fundamental to understand whether the reconstructed transcripts could match with some known protein or nucleotide sequences and thus be linked to specific functions or processes already described in other species. To this aim, the assembled fragments were aligned against the NCBI NR (non-redundant) UniProtKB/TrEMBL protein databases and against the NCBI NT nucleotide collection (data downloaded on 22/04/2021). I also ran InterproScan (version 5.51-85.0) in order to search for known functional domains and predict protein family membership. **Figure 6** shows the relative percentage of each GO category represented within each krill transcriptome.

In addition, all krill transcripts were searched against RNAcentral database (https://rnacentral.org/; https://doi.org/10.1093/nar/gkw1008) in order to identify and characterize the fraction of non-coding RNAs within the reconstructed sequences. Contextually, an orthology inference was ran using Orthologous MAtrix (OMA) standalone package (https://omabrowser.org/standalone/; Altenhoff et al., 2019) which uses information available on public OMA browser, a database with a complete catalog of orthologous genes among more than 2300 genomes across the tree of life. This analysis helped me identify, on the basis of protein sequences, those krill transcripts showing orthology relationship with genes from other species and set of genes belonging to a single common ancestral gene at a specific taxonomic range of interest (Altenhoff et al., 2013).

Euphausia superba

Results from the functional annotation analyses showed that 63,633 contigs matched at least one protein from the NR collection, corresponding to about 42% of the total krill transcriptome, while 62,249 contigs found a match among TREMBL protein sequences (41% of the total). A search against the NT nucleotide sequences produced 22,071 krill matches (15% of the total). To classify transcripts by putative function, I performed a GO assignment. Specifically, 2,612 GO terms (corresponding to 13068 genes) were assigned: 1,128 of those (corresponding to 1178 genes) represented molecular functions; 1,099 terms (corresponding to 6991 genes) were linked to biological processes; 385 terms (corresponding to 4303 genes) represented cellular components.

Meganyctiphanes norvegica

Among the total Northern krill transcriptome 47,275 sequences matched at least one protein from NR database, representing about 43% of the total. 16,136 contigs were annotated among NT nucleotide sequenced, corresponding to about 15% of the total, while 46,548 transcripts found a match among TREMBL protein sequences (42% of the total). As for E. superba, a few transcripts matched with a non-coding RNA from RNAcentral database (2% of the total). The search of putative protein domains produced a total of 76,533 matches (70% of the total), among them 28,341 did not find any other match against the other nucleotide/protein databases. The GO analysis showed that a total of 2,665 GO terms could be assigned: 1,145 of them represented a molecular function, 392 corresponded to a cellular component while 1,128 were related to a biological process.

Thysanoessa inermis

The annotation of the newly reconstructed transcriptome of *T. inermis* resulted in a total of 58,392 contigs finding at least a match with an entry from NR collection, corresponding to 36% of the total transcripts, while 19,264 sequences were annotated against NT nucleotide database (12% of the total). The alignment against TREMBL protein collection produced 57,278 matches (35% of the total), while 2,421 transcripts were annotated against RNAcentral database, corresponding to 1% of the total. InterproScan analysis showed that 105,371 sequences found at least a match with a known protein domain (65% of the total) and, among them, 43% of the sequences did not find annotation with any of the previously mentioned databases. I assigned 2,600 GO terms to the assembled transcripts, divided into 1,117 terms corresponding to a molecular function, 387 related to a cellular component and 1,096 representing a biological process.

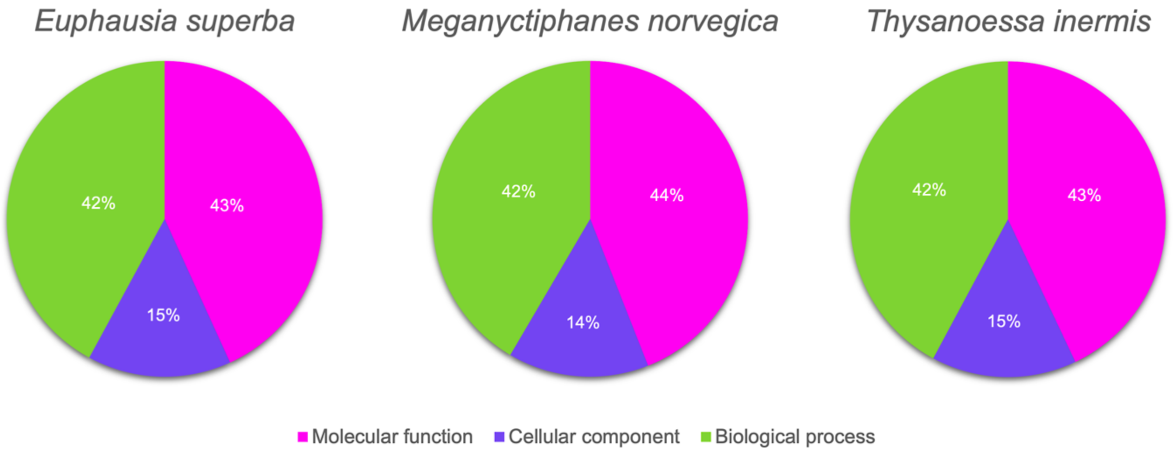


Figure 6. Representation (%) of each Gene Ontology category among the GO terms assigned to the assembled transcripts for each species.

Expression Atlas

With the aim to implement KrillDB website with a new section collecting the expression data produced from all the newly available sequences, I estimated the abundances of all the transcripts in the final assembly using Salmon over a wide range of RNA-seq datasets, including:

- Larval krill at two different stages of development exposed to different CO_2 conditions, coming from Sales et al. (2017)
- Adult krill (48 samples) coming from different geographical areas (Bransfield Strait, Lazarev Sea, South Georgia, South Orkney) and different seasons (summer and winter), divided into male and female specimens (Höring et al., 2021)
- Adult krill divided into male and female specimens (Suter et al., 2019)
- Adult krill exposed to three different temperatures – Low Temperature, Mid temperature, High Temperature (data downloaded from NCBI BioProject: PRJNA640244)

Overall, the experimental design included six factors: geographical area, season, developmental stage, pCO_2 exposure condition, sex and temperature. Abundances and raw counts were imported using R (version 4.0.5) and the package *tximport* (version 1.18.0). Batch effect removal was performed using the *removeBatchEffect* function implemented in the *limma* package (version 3.46.0). The resulting count matrix of transcripts (rows) across samples (columns) was then converted to the transcripts per million (TPM) scale. Finally, results were summarized to the gene level using the *isoformToGeneExp* function (IsoformSwitchAnalyzeR version 1.12.0 package). The expression levels for each experimental condition are displayed in KrillDB^2 as a barplot, as part of the webpage for each gene or transcript (see section “KrillDB^2 Web Interface”).

Differential expression analysis

Transcript-level abundances and estimated counts were summarized at the gene-level using the package tximport. Resulting counts were normalized to remove unwanted variation by means of the RUVg method (Risso & Course, 2015). Normalization is needed anytime we want to compare expression across samples or genes as it is essential to ensure that observed differences in expression are truly due to differential expression and not technical artifacts (e.g., differences in sequencing depth, library preparation). RUVg method is specifically based on the main assumption that one can identify a set of negative control genes, which are non-differentially expressed genes (e.g., genes whose expression is not influenced by the biological covariates of interest). Therefore, the procedure performs a factor analysis of the read counts based on the set of control genes selected to estimate the amount of unwanted variation.

Here I performed a preliminary between-sample normalization (EDASeq package, version 2.24.0) to adjust for sequencing depth. Then, following the workflow outlined in the RUVseq vignette, I defined the set of empirical negative control genes with an FDR level larger than 0.8. My design matrix for the model included all the independent factors (season, area and sex for *E. superba* and sex, time point and tissue type for *M. norvegica* and *T. inermis*) and, in addition, the interaction between all the independent factors. This kind of model (e.g. the interaction model) is useful in identifying the effect of the combined conditions, understanding whether the differential expression is due to the effect of the main factors or can be addressed to the combined effects of those factors. In other words, it would help us understand whether the change in expression between males and females is the same for winter season as it is for summer season. I selected the 30,000 less differentially expressed genes as negative controls and applied the RUVg method to estimate k=2 factors of unwanted variation. I then included those factors in the design matrix for the final differential expression analysis, performed using the GLM method implemented by the edgeR software (version 3.32.1). All p-values were corrected using the Benjamini-Hochberg method.

Euphausia superba

The availability of a new assembly of the krill transcriptome, reconstructed collecting the largest amount of experimental data available thus far, suggested the possibility of performing a detailed investigation of differential expression patterns. In particular, I decided to reanalyze the dataset from (Höring et al., 2021) in order to assess the possibility of identifying differentially expressed genes which were not detected in the original study due to the use of an older reference (Meyer et al., 2015).

In total 1,741 genes were differentially expressed (DEG) among experimental conditions, corresponding to around 2% of the total reconstructed genes. In the previous work by Höring et al. (2021) the same samples were quantified against a total of 58,581 contigs (Meyer et al., 2015) producing a total of 1,654 DEGs. **Table 11** summarizes all the contrasts that were performed, each one with the number of differentially expressed up and down regulated genes.

1,195 DEGs were identified in the comparison between summer and winter specimens: 1,078 were up-regulated and 117 down-regulated. 396 of such DEGs had some form of functional annotation. Comparison between males and females produced 14 DEGs (7 up-regulated and 7 down-regulated); none had some form of annotation deriving from sequence homology. In general, these results are in accordance with the discussion by Höring et al. (2021), which found that seasonal differences are predominant in comparison to regional ones.

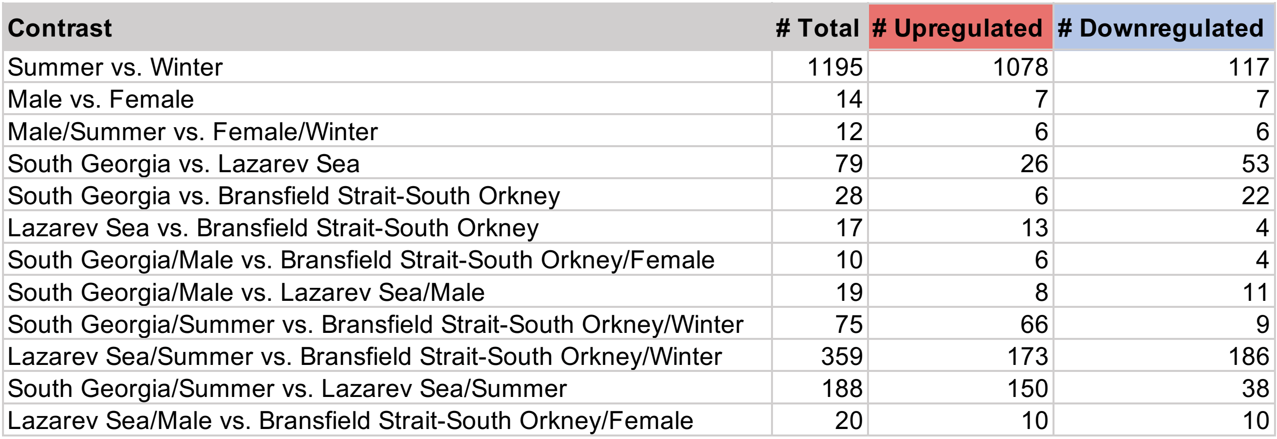
I selected a series of genes among seasonal DEGs according to what has been already described in literature. Höring et al. (2021) previously identified and described 35 relevant

DEGs involved in seasonal physiology and behavior: I recovered the same gene signature in our own analysis by comparing summer to winter samples. The majority of these DEGs appear to be involved in the development of cuticles (*chitine synthase*, *carbohydrate sulfotransferase 11*), lipid metabolism (*fatty acid synthase 2*, *enoyl-CoA ligase*), reproduction (*vitellogenin*, *hematopoietic prostaglandin D synthase*), metabolism of different hormones (*type 1 iodothyronine deiodinase*) and in the circadian clock (*cryptochrome*). Our results also include DEGs that were found to be involved in the moult cycle of krill in other studies (Seear et al., 2010). Specifically, I identified a larger group of genes involved in the different stages of cuticle developmental process (*peritrophin-A domain*, *calcified cuticle protein*, *glycosyltransferase 9-domain containing protein*, *collagen alpha 1*, *glutamine-fructose 6 phosphate*), including proteins such as cuticle protein-3,6,19.8, early cuticle protein, pupal cuticle protein, endocuticle structural glycoprotein, chitinase-3 and chitinase-4, the latter representing a group of chitinase which have been shown to be expressed predominantly in gut tissue during larval and/or adult stages in other arthropods and are proposed to be involved in digestion of chitin-containing substrates (Khajuria, Buschman, Chen, Muthukrishnan & Zhu, 2010). Finally, in addition to *trypsin* and *crustin 4* (immune-related gene, essential in early pre-moult stage when krill still have a soft cuticle to protect them from pathogen attack, as seen by Seear et al., 2010) I also identified crustin-1,2,3,5 and 7. All the reported genes were up-regulated in summer, the period in which growth take place and krill moult regularly.

Cuticle development genes were also identified as differentially expressed in the analysis of the interaction of multiple factors, in particular between male samples coming from South Georgia and female specimens coming from the area of Bransfield Strait-South Orkney (considered as a unique area since they are placed at similar latitudes). Strikingly, I also identified a pro-resilin gene, whose role in many insects consists in providing efficient energy storage, being upregulated in South Georgia male specimens.

A number of relevant DEGs were found among specific interactions of regional and seasonal changes. In the comparison between krill samples in South Georgia in summer and individuals sampled in Bransfield Strait-South Orkney in winter I found genes, up-regulated in summer in South Georgia, that are related to reproductive activities, such as *doublesex and mab-3 related transcription factor*. The latter, in particular, is a transcription factor crucial for sex determination and sexual differentiation which has already described in other arthropods (Yi Jia et al., 2018) where the male-specific isoform represses its targets. Since no differentially expressed gene related to reproduction was found by Höring et al. (2021) in the same comparisons, this suggests that the new krill transcriptome significantly improves the interpretability of expression studies and the characterizations of krill samples.

Finally, comparison between male individuals from Lazarev Sea and female specimens from Bransfield Strait-South Orkney showed additional DEGs involved in reproduction, such as *ovochymase 2*, usually highly expressed in female adults or eggs, *serine protease* and a *trypsin-like gene*. In particular, trypsin-like genes are usually thought to be digestive serine proteases, but previous works suggested that they can play other roles (Bao et al., 2014); many trypsins show female or male-specific expression patterns and have been found exclusively expressed in males, as in our analysis, suggesting that they play a role in the reproductive processes.



Contrast	# Total	# Upregulated	# Downregulated
Summer vs. Winter	1195	1078	117
Male vs. Female	14	7	7
Male/Summer vs. Female/Winter	12	6	6
South Georgia vs. Lazarev Sea	79	26	53
South Georgia vs. Bransfield Strait-South Orkney	28	6	22
Lazarev Sea vs. Bransfield Strait-South Orkney	17	13	4
South Georgia/Male vs. Bransfield Strait-South Orkney/Female	10	6	4
South Georgia/Male vs. Lazarev Sea/Male	19	8	11
South Georgia/Summer vs. Bransfield Strait-South Orkney/Winter	75	66	9
Lazarev Sea/Summer vs. Bransfield Strait-South Orkney/Winter	359	173	186
South Georgia/Summer vs. Lazarev Sea/Summer	188	150	38
Lazarev Sea/Male vs. Bransfield Strait-South Orkney/Female	20	10	10

Table 11. Number of up and downregulated genes among the total of DEGs identified within each comparison considered in *E. superba* differential expression analysis.

Meganyctiphanes norvegica

Differential expression analysis produced a total of 19,242 Northern krill differentially expressed genes, corresponding to about the 26% of the total number of genes. A summary of the total number of DEGs related to each comparison is reported in **Table 12**.

The comparison between male and female individuals captured a total of 12,115 DEGs (5,841 up-regulated and 6,274 down-regulated), while the amount of differential expression that could be addressed to changes between different time points was represented by 427 genes (200 up-regulated and 227 down-regulated). A consistent part of DEGs was distributed among comparisons between different tissue types: 4,173 between sexual tissue and carapax (2,667 up-regulated, 1,506 down-regulated), 15,543 between head tissue and carapax (7,534 up-regulated and 8,009 down-regulated) and 5,092 DEGs between sexual and head tissues (3,311 up-regulated, 1,781 down-regulated).

Changes between sexual tissue and head tissue highlighted the presence of differentially expressed genes annotated as *opsin 1*, *opsin 2*, *opsin 3*, and also the *opsin 4* from *Euphausia superba*; in addition, I identified an *onychopsin*, a *peropsin*, a *compound eye opsin*, *visual pigment-like receptor* and a *peropsin isoform*, all up-regulated in samples from head tissue.

The availability of samples coming from different tissue types likely enhanced the possibility to observe as differentially expressed some genes involved in specific tissue differentiation, as well as processes or reaction taking place in specific tissues. For example, the comparison between head tissue and carapax and the one between sexual and head tissues highlighted the presence of a series of neuropeptides F (*neuropeptide F1*, *neuropeptide F2*, *neuropeptide F3, preproneuropeptide F I, preproneuropeptide F II*) and *neuropeptide Y*, all up-regulated in samples coming from head tissue; these neuropeptides have already been described in other arthropods species as multi-functional neuropeptides playing important roles in feeding, metabolism, reproduction, as well as circadian rhythms and stress responses (Cui HY & Zhao ZW, 2020).

Despite the results of differential expression analysis for Antarctic krill, where few genes were found DE due to sex differentiation, here I identified a series of DEGs in comparison between male and female individuals. Among them, *crustacyanin A1*, *crustacyanin C1* and *crustacyanin C3* are all genes involved in body color variation, binding carotenoid astaxanthin which is responsible of the color of crustacean carapax. Together with them, I also observed genes related to cuticle development (*crustin like*, *arthrodial cuticle proteins, cuticle protein*), but also an *ecdysis triggering hormone receptor* (ETH), described in other arthropod species as a master hormone in regulating the ecdysis process; this process is characterized by the molting of an external skeleton for the purpose of growth or change in shape (Žitňan D et al, 2010). In addition, a gene involved in post-mating interactions (*ejaculatory bulb specific protein 3*) was found to be differentially expressed, up-regulated in male samples, as well as *androgen-dependent TFPI-regulating protein isoform* and *hematopoietic prostaglandin d synthase*, the latter usually involved in smooth muscle contraction and relaxation.

Although a great number of genes were found to be differentially expressed among the comparisons between independent factors, a part of them were also differentially expressed in the interaction between the main factors. In particular, the comparison between head tissue coming from male samples and carapax from female specimens produced a series of genes related to important functions and process of krill life cycle. Among them, *sex-determination protein fruitless like*, up-regulated in male head tissue, was already described in *Drosophila melanogaster*, where controls the development of the male specific abdominal muscle of Lawrence and seems to play a role in male courtship behavior and sexual orientation (Taylor BJ et al., 2001). Similarly, other genes involved in sex-determination and oogenesis in *D. melanogaster* were found to be differentially expressed in this comparison, up-regulated in male head tissue too (*sex lethal variant 1*, *sex lethal variant 2*, *sex lethal variant 6*). Other DEGs up regulated in male head tissue have been identified, such as *cryptochrome-2-like isoform*, *putative cryptochrome-1-like isoform*, a series of *heat shock proteins, small androgen receptor interacting protein 1*, *male specific lethal 3*, *doublesex and mab-3 related transcription factor* already found as differentially expressed in *E. superba* data and two neuropetides, *neuropeptide FNPF* and *pro-neuropeptide Y*. Moreover, a *retinal rod rhodopsin-sensitive cGMP 3'-5'* was identified as up-regulated in male head tissue, a gene which has been already described as involved in the amplification of visual signal. This comparison also revealed the presence of other DEGs involved in reproduction and molt cycle, which are up-regulated in female carapax samples, such as cuticular proteins, procollagen-proline 4-dioxygenase, ovochymase-2 and vitelline membrane outer layer protein 1-like; the latter represents a basic protein present in the outer layer of the vitelline membrane of eggs, playing an essential role in separating the yolk from the egg white.

Other genes involved in the process of oogenesis (*vitellogenin*) were differentially expressed in the comparison between male samples from November and female individuals from May.

Contrast	# Total	Upregulated	Downregulated
November vs May	427	200	227
Male vs Female	12115	4841	6274
Male/November vs Female/May	144	77	67
Head vs Carapax	15543	7534	8009
Sex vs Carapax	4173	2667	1506
Sex vs Head	5092	3311	1781
Male/Head vs Female/Carapax	5122	3275	1847
Male/Sex vs Female/Carapax	326	13	313
November/Head vs May/Carapax	24	15	9
November/Sex vs May/Carapax	39	27	12
Male/November/Head vs Female/May/Carapax	9	6	3
Male/November/Sex vs Female/May/Carapax	15	5	10

Table 12. Number of up and downregulated genes among the total of DEGs identified within each comparison considered in *M. norvegica* differential expression analysis.

Thysanoessa inermis

The comparisons that could be performed starting from *T. inermis* dataset and experimental design are reported in **Table 13**, with the number of up- and down-regulated genes.

I identified a total of 9,892 DEGs, corresponding to 6% of the total reconstructed genes. The distribution of these genes among the different comparisons follows a similar signature to the one observed for *M. norvegica*. Therefore, most of the differential expression seems to be addressed to differences between males and females, as well as differences between different tissue types and specific interaction between these main factors.

Looking at the functions and processes in which the identified DEGs can be involved, I observed that most of them were annotated similarly to the DEGs detected in *M. norvegica* differential expression analysis. However, in addition to the genes previously described comparison between head tissue samples from July and carapax samples from April produced a DEG annotated as *glia derived nexin*, which seems to be up-regulated during neuronal differentiation (here up-regulated in head samples coming from July)

Contrast	# Total	Upregulated	Downregulated
Juvy vs April	328	98	234
Female vs Male	5614	3881	1733
Female/July vs Male/April	224	185	39
Head vs Carapax	3706	1358	2348
Sex vs Carapax	884	132	752
Sex vs Head	2262	880	1382
Female/Head vs Male/Carapax	1533	182	1351
Female/Sex vs Male/Carapax	29	23	6
July/Head vs April/Carapax	57	48	9
July/Sex vs April/Carapax	38	29	9
Female/July/Head vs Male/April/Carapax	12	4	8
Female/July/Sex vs Male/April/Carapax	61	10	51

Table 13. Number of up and downregulated genes among the total of DEGs identified within each comparison considered in *T. inermis* differential expression analysis.

Identification of microRNA precursors

Together with coding sequences, over the years the development of new approaches in molecular biology and new computational methods have increased our possibility to also understand the importance of the non-coding RNAs fraction. In particular, the most studied of these non-coding RNAs are represented by microRNAs (miRNA). The first miRNA was identified in 1993 by Lee et al., describing its importance in the regulation of the timing of development in *Caenorhabditis elegans*. miRNAs are small non-coding RNAs, with an average 22 nucleotides in length. They are usually transcribed from DNA sequences into primary miRNAs (pri-miRNAs), then processed into precursor miRNAs (pre-miRNAs) with a typical hairpin shape and finally mature miRNAs. Specifically, miRNA genes are transcribed by RNA polymerase II (Pol II) in the nucleus, while, once processed, the pre-miRNAs are exported from the nucleus into the cytoplasm to be transformed into miRNAs. These precursors have been widely studied, in fact there are a series of studies that have shown how some miRNA families are widely conserved across different lineages (Zhang et al. 2006; Axtell and Bartel 2005). Such evidence suggests that using specific computational methods for the identification of homology it can be possible to discovery mature sequences of conserved miRNA families. Most of the approaches focus on the identification of pre-miRNAs instead of mature miRNAs, due to the difficulty in experimentally identify the lowly expressed miRNAs or the miRNAs that are expressed in the specific tissues or in the developmental stage. Moreover, the characteristic stem-loops of the pre-miRNAs have been found to be well conserved in both vertebrate and invertebrate animals.

The importance of miRNAs is related to the fact that they act as post-transcriptional and translational regulators of gene expression interacting with a target mRNA. Usually, they determine the expression suppression of a target mRNA, by the interaction with its 3′ UTR (Ha & Kim, 2014), but it has been found that miRNAs can also interact with other regions, such as 5' UTR or coding sequences (O'Brien, Hayder, Zayed & Peng, 2018). Although their key role in the regulation of gene expression and, therefore, in many important biological processes, such as development or cell differentiation, there are still no information about miRNAs in krill species.

For these reasons, I also investigated the possibility that the new transcriptomes included sequences corresponding to miRNA precursors.

To this aim, I ran the HHMMiR software (Kadri, Hinman & Benos, 2009), which combines structural and sequence information to train a Hierarchical Hidden Markov Model (HHMM) for the identification of miRNA genes. In particular, it predicts the miRNAs using a template for the structure of a typical pre-miRNA hairpin from publicly available data. This template is characterized by the terminal loop, the extension (area between the terminal loop and the miRNA duplex), the miRNA duplex and the pri-miRNA extension. Then, the software uses the HHMM model over this template. The output is represented by a list of input transcripts sequences identified as a putative miRNA, with a score indicating the degree of similarity.

Once identified the hairpins, I also performed a *blastn* search of all our assembled transcripts against the collection of miRBase (http://www.mirbase.org/) mature sequences. Results from these two analyses were combined: I collected all transcripts with a HHMMiR score below or equal to 0.71 and an alignment to a known mature microRNA with at most two mismatches. I then used the QuickGO tool (https://www.ebi.ac.uk/QuickGO/) to identify any potential association among the putatively identified microRNAs and GO categories.

Euphausia superba

In total I identified 261 Antarctic krill transcripts with sequence homology to 644 known microRNAs from other species. 306 sequences were linked to at least one GO term, matching 54 krill transcripts. Among them, I identified 5 putative microRNAs involved with changes in cellular metabolism (*age-dependent general metabolic decline* - GO:0001321, GO:0001323), as well as changes in the state or activity of cells (*age-dependent response to oxidative stress* - GO:0001306, GO:0001322, GO:0001324), 35 microRNAs involved in interleukin activity and production regulation. Moreover, I found 26 putative microRNAs likely involved in ecdysteroidogenesis (specifically GO:0042768), a process resulting in the production of ecdysteroids, moulting and sex hormones found in many arthropods. In addition, I found a microRNA involved in fused antrum stage (GO:0048165) which appears to be related, in other species, to the oogenesis process. I also identified 27 microRNAs related to rhombomere morphogenesis, formation and development (GO:0021661, GO:0021663, GO:0021570). These functions are well described as involved in the development of portions of the central nervous system in vertebrates, but they are also homologous to a portion of the arthropod brain. Moreover, 26 krill sequences showed high similarity with 2 mature microRNA related to the formation of tectum (GO:0043676), which represents in arthropods and, specifically, crustaceans, the part of the brain acting as visual center.

Meganyctiphanes norvegica

The results produced by the analysis of Antarctic krill transcriptome in search of putative microRNA precursors suggested the possibility to repeat the same procedure also on the other two krill species data, trying - in a second step - to highlight possible similar patterns in terms of microRNAs included in the reconstructed transcriptomes.

The microRNA precursors identification performed on Northern krill data produced a total of 265 transcripts matching 212 microRNAs from other species. Once annotated using the QuickGO tool in search of GO terms related to the identified microRNAs, I observed 49 sequences matching at least a GO term, corresponding to 95 Northern krill transcripts. As for Antarctic krill, 2 microRNAs likely involved in metabolism changes were detected (*age-dependent general metabolic decline*, *age-related resistance and others*), while 85 Northern krill transcripts showed similarity with 5 microRNAs related to interleukin binding and production. Same Antarctic krill microRNAs responsible of the formation and development of portions of the central nervous system can be described for *M. norvegica*; however here a greater number of transcripts (83) found a match with 4 microRNAs annotated as *rhombomere 1 development*, *rhombomere 1 formation*, *rhombomere 2*, *rhombomere 4* (GO:0021567, GO:0021568, GO:0021663, GO:0021570).

Interestingly, a transcript matched a microRNA corresponding to *mature follicle stage* (GO:0048166), representing for most mammals but also insects a step occurring in ovulation process (Knapp E & Sun J, 2017). Moreover, a microRNA involved in pre-ovulatory period as a prerequisite for ovulation was identified (*prostaglandin-E2 9-reductase activity*, GO:0050221).

A total of 81 transcripts matched with 2 mature microRNAs involved in the formation of tectum, the same described for *E. superba* (GO:0043676).

Thysanoessa inermis

The same analysis carried out on *T. inermis* transcripts collection produced a total of 227 sequences matching 620 microRNAs from other species; among them, 417 mature microRNAs could be annotated with at least a GO term, corresponding to 44 *T. inermis* transcripts. The signature was almost the same reported for E. superba and M. norvegica; in particular, a transcript matched a microRNA related to changes in metabolic processes, as observed for the other two krill species, while here 18 transcripts showed similarity with 138 microRNAs involved in interleukin activity (*interleukin-10 binding*, *interleukin-10 mediated signaling pathway*, *interleukin-2 binding and production*, *interleukin-4 binding and production, interleukin-7,8,9*). A transcript matched with a total of 34 mature microRNAs annotated as *fused antrum stage* (GO:0048165) already reported for Antarctic krill and representing a stage of the oogenesis. Tectum microRNAs were identified also in *T. inermis* transcriptome, in particular 14 transcripts matched with 4 of such microRNAs (GO:0043676).

As seen in *M. norvegica* microRNA precursors identification, a transcript matched with 3 microRNAs from mature follicle stage, while another transcript showed similarity with the same prostaglandin receptor previously described. In addition, I identified a sequence corresponding to a microRNA likely involved in steroid biosynthesis and lipid metabolism (steroid 9-alpha-monooxygenase activity, GO:0050292).

Cross species comparison

A powerful way of characterizing newly sequenced and assembled transcripts is to compare them with evolutionarily related genes, specifically with orthologs in other species (corresponding sequences across multiple species). Therefore, I decided to perform an orthology inference procedure for each krill species. This analysis would help not only the identification of krill transcripts with an orthology relationship with sequences from other organisms but would also allow for a comparison among the three krill species, in order to find patterns, sequences and transcriptional profiles conserved or unique among them.

To this aim, I uses the OMA software in order to identify orthologous genes among the sequences of all krill species under study together with other 42 species across Arthropoda phylum and, as outgroups, 4 species from Chordata, 2 from Nematoda, a species from Cnidaria and another one from Porifera (https://omabrowser.org/standalone/; Altenhoff et al., 2019). Results were then filtered using OMA identifier mapping lists, extracting, were available, the annotation of OMA genes from UniProt, EntrezGene and Ensembl IDS. OMA performs a parallel comparison of all transcript sequences computing a series of all-against-all sequence alignments followed by pairwise maximum likelihood distance estimation for all pairs with a statistically significant similarity. Then, the orthology inference is performed, producing a series of results: lists of pairwise orthologs (one-to-one, one-to-many, many-to-one, many-to-many orthology); transcript groups, which are sets of genes for which all pairs are inferred to be orthologous; and Hierarchical Orthologous Groups (HOGs). The latter are groups defined for every internal node of the species tree, so that each HOG contains the genes that are inferred to have descended from a common ancestral gene among the species attached to that internal node.

The all-against-all sequence alignment is the most time-consuming step of the algorithm, in this case made more challenging by the number of species processed and the amount of data produced for each krill species. However, the parallelization of the process helped me optimize this step. Overall, I managed to obtain the full list of results in 3 days by using a computing cluster equipped with 256 cores.

I identified 20,478 *E. superba* sequences showing an orthology relationship with an OMA gene, while 17,431 and 16,037 orthologs were found for *M. norvegica* and *T. inermis* respectively. In a second moment I have crossed OMA results in search of groups of orthologs that included different combinations of the three krill species, such as: i) *M. norvegica* and *T. inermis*, ii) *M. norvegica* and *E. superba*, iii) *T. inermis* and *E. superba*, iv) *M. norvegica*, *T. inermis* and *E. superba.*

As shown in **Figure 7**, a total of 8,341 sequences from each krill species showed an orthology relationship, suggesting that they would likely derive from the same common ancestor. The two arctic species, *T. inermis* and *M. norvegica*, shared a total of 2,295 sequences classified as orthologs by OMA analysis; 3,245 and 4,470 sequences were grouped as orthologs across *T. inermis* and *E. superba* and *M. norvegica* and *E. superba*, respectively.

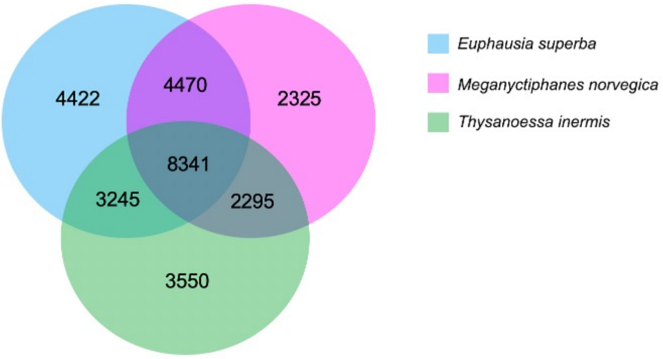


Figure 7. Number of genes from the different krill species belonging to the same OMA orthologous group.

Furthermore, a comparison of the results in terms of differentially expressed genes was contextually performed. The three experimental designs shared two main variables: sex and season/time point. I compared the number of differentially expressed genes among different sexes, seasons or time points and the interaction between these main factors across all the species, finding that *M. norvegica* and *T. inermis* almost showed the same signature in terms of number of DEGs among the different contrasts performed, while a different trend was registered in *E. superba*. Specifically, most of the differential expression in Antarctic krill was due to seasonal changes, while changes in sex were predominant in the other two species (**Figure 8**). In addition, as *T. inermis* and *M. norvegica* shared the same experimental design, I compared the number of DEGs among all the contrasts performed across the two krill species, which highlighted a highly similar trend.

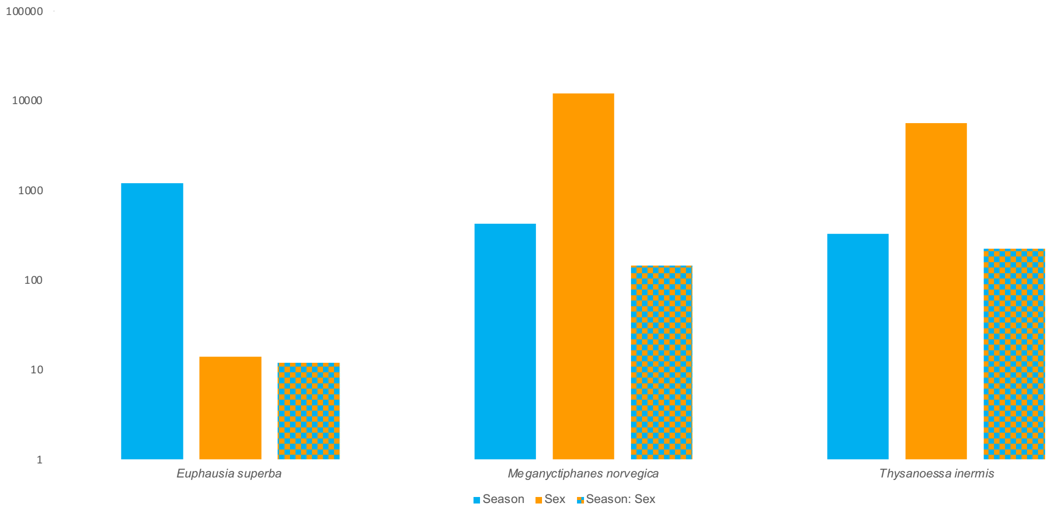
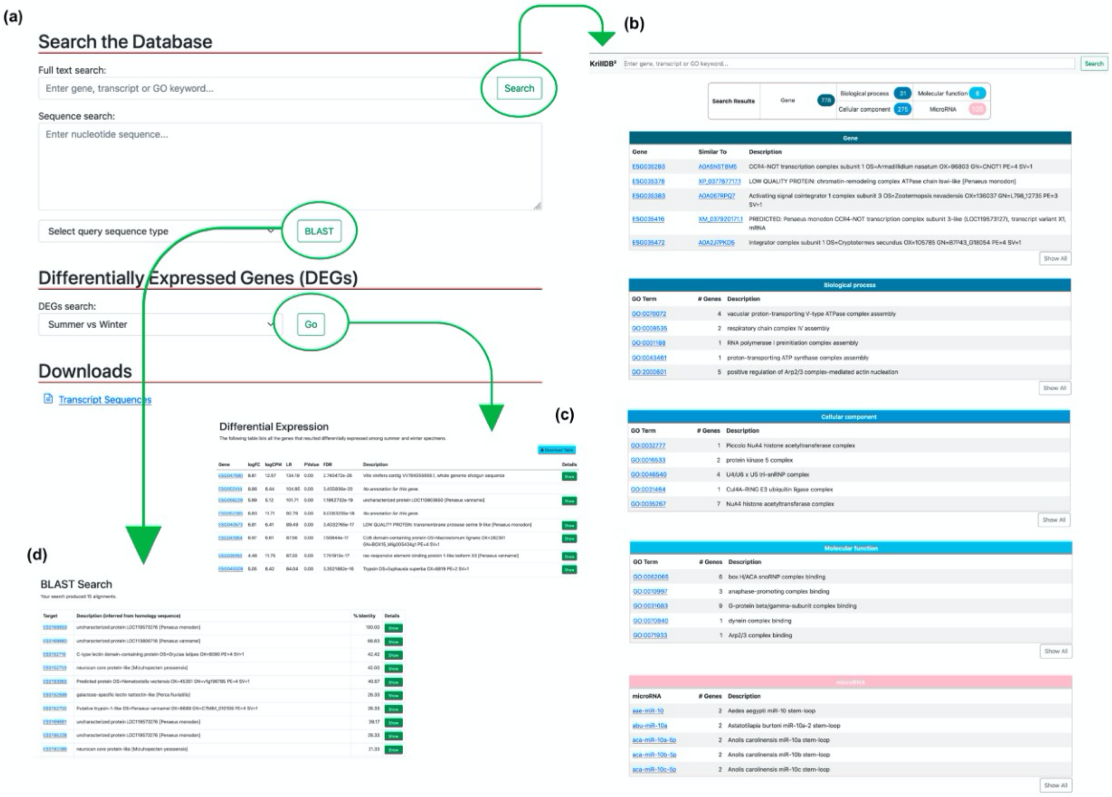


Figure 8. Number of DEGs among the comparable contrasts across *E. superba*, *M. norvegica*, *T. inermis*.

Together with these observations, I decided to cross the differentially expressed genes tables from all krill species, I identified 3,640 DEGs from Northern krill, 2,163 from *T. inermis* and 450 from Antarctic krill. The majority of these genes were involved in the krill moult cycle (*chitinase, crustin, collagenolytic serine, cartilage oligomeric matrix precursor, cuticle protein, trypsin, carbohydrate sulfotransferase, peritrophin* ect.), embryogenesis, sexual differentiation, reproductive process (*serine protease, ovochymase, vitellogenin, prostaglandin, apolipocrustacein, blastula protease, androgen-induced gene, CUB-serine protease*) but also related to the immune system (*C-type lectine, antimicrobial peptide*) and lipid metabolism (*fatty acid synthase 2, enoyl-CoA ligase*).

KrillDB^2 Web Interface

I re-designed the KrillDB website interface to include the new version of the Antarctic krill transcriptome assembly. **Figures 9**, **10** and **11** collect images taken from the main new sections of the updated version of KrillDB. The integrated full-text search engine allows the user to search for a transcript ID, gene ID, GO term, a microRNA ID or any other free-form query. In particular, results of full-text searches are now organized into a number of separate tables, each representing a different data source or biological aspect (**Figure 9b**).



Search the Database

Full text search: Enter gene, transcript or GO keyword...

Sequence search: Enter nucleotide sequence...

Select query sequence type

Differentially Expressed Genes (DEGs)

DEGs search: Summer vs Winter

Downloads

Transcript Sequences

KrillDB² Enter gene, transcript or GO keyword...

Search Results	Gene 778	Biological process 31	Molecular function 6
		Cellular component 275	MicroRNA 105

Gene

Gene	Similar To	Description
ESG035293	A0A5N5T8M5	CCR4-NOT transcription complex subunit 1 OS=Armadillidium nasatum OX=96803 GN=CNOT1 PE=4 SV=1
ESG035378	XP_037787717.1	LOW QUALITY PROTEIN: chromatin-remodeling complex ATPase chain Iswi-like [Penaeus monodon]
ESG035383	A0A067RPQ7	Activating signal cointegrator 1 complex subunit 3 OS=Zootermopsis nevadensis OX=136037 GN=L798_12735 PE=3 SV=1
ESG035416	XM_037920171.1	PREDICTED: Penaeus monodon CCR4-NOT transcription complex subunit 3-like (LOC119573127), transcript variant X1, mRNA
ESG035472	A0A2J7PKD5	Integrator complex subunit 1 OS=Cryptotermes secundus OX=105785 GN=B7P43_G18054 PE=4 SV=1

Biological process

GO Term	# Genes	Description
GO:0070072	4	vacuolar proton-transporting V-type ATPase complex assembly
GO:0008535	2	respiratory chain complex IV assembly
GO:0001188	1	RNA polymerase I preinitiation complex assembly
GO:0043461	1	proton-transporting ATP synthase complex assembly
GO:2000601	5	positive regulation of Arp2/3 complex-mediated actin nucleation

Cellular component

GO Term	# Genes	Description
GO:0032777	1	Piccolo NuA4 histone acetyltransferase complex
GO:0016533	2	protein kinase 5 complex
GO:0046540	4	U4/U6 x U5 tri-snRNP complex
GO:0031464	1	Cul4A-RING E3 ubiquitin ligase complex
GO:0035267	7	NuA4 histone acetyltransferase complex

Molecular function

GO Term	# Genes	Description
GO:0062065	6	box H/ACA snoRNP complex binding
GO:0010997	3	anaphase-promoting complex binding
GO:0031683	9	G-protein beta/gamma-subunit complex binding
GO:0070840	1	dynein complex binding
GO:0071933	1	Arp2/3 complex binding

microRNA

microRNA	# Genes	Description
aae-miR-10	2	Aedes aegypti miR-10 stem-loop
abu-miR-10a	2	Astatotilapia burtoni miR-10a-2 stem-loop
aca-miR-10a-5p	2	Anolis carolinensis miR-10a stem-loop
aca-miR-10b-5p	2	Anolis carolinensis miR-10b stem-loop
aca-miR-10c-5p	2	Anolis carolinensis miR-10c stem-loop

Differential Expression

The following table lists all the genes that resulted differentially expressed among summer and winter specimens.

Gene	logFC	logCPM	LR	PValue	FDR	Description	Details
ESG047580	8.81	12.57	134.19	0.00	2.740472e-26	Vitis vinifera contig VV78X055658.1, whole genome shotgun sequence	Show
ESG055134	8.86	8.44	104.95	0.00	3.455836e-20	*No annotation for this gene.*	
ESG056229	5.99	5.12	101.71	0.00	1.1862732e-19	uncharacterized protein LOC113803650 [Penaeus vannamei]	Show
ESG052385	6.83	11.71	92.79	0.00	8.0263255e-18	*No annotation for this gene.*	
ESG040573	6.81	6.41	89.49	0.00	3.4032765e-17	LOW QUALITY PROTEIN: transmembrane protease serine 9-like [Penaeus monodon]	Show
ESG041954	6.97	6.61	87.56	0.00	7.50844e-17	CUB domain-containing protein OS=Macrostomum lignano OX=282301 GN=BOX15_Mlig005434g1 PE=4 SV=1	Show
ESG036993	4.46	11.75	87.20	0.00	7.741913e-17	ras-responsive element-binding protein 1-like isoform X3 [Penaeus vannamei]	Show
ESG045328	5.05	8.42	84.04	0.00	3.3521882e-16	Trypsin OS=Euphausia superba OX=6819 PE=2 SV=1	Show

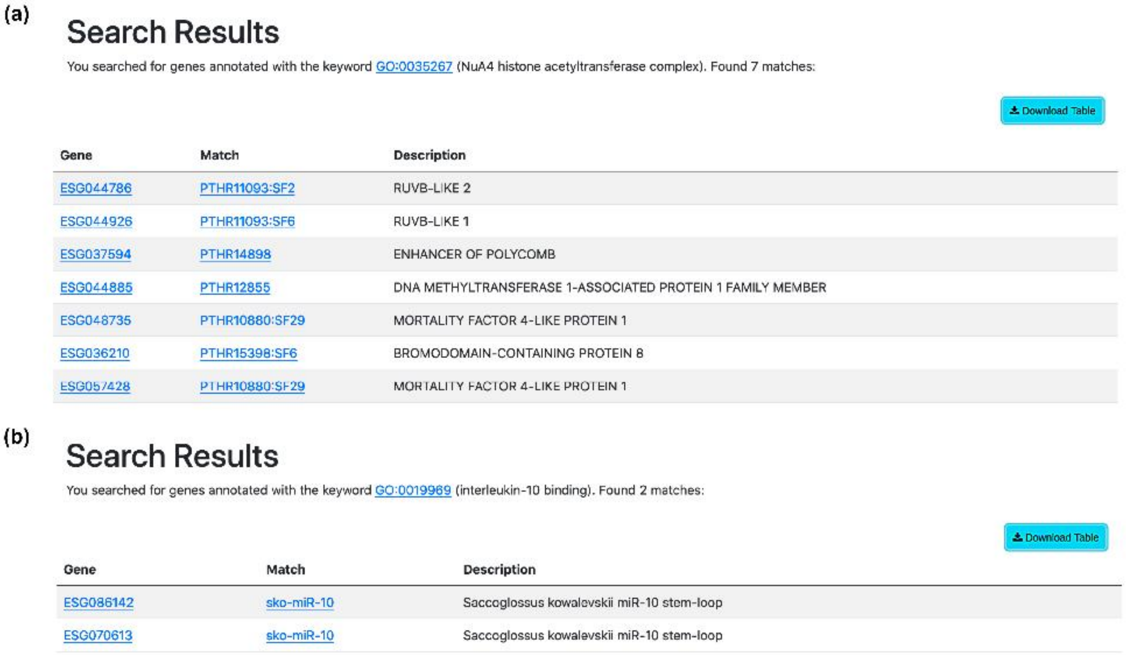
BLAST Search

Your search produced 15 alignments.

Target	Description (inferred from homology sequence)	% Identity	Details
ESS169859	uncharacterized protein LOC119573276 [Penaeus monodon]	100.00	Show
ESS169860	uncharacterized protein LOC113806718 [Penaeus vannamei]	68.83	Show
ESS152719	C-type lectin domain-containing protein OS=Oryzias latipes OX=8090 PE=4 SV=1	42.42	Show
ESS152703	neurocan core protein-like [Mizuhopecten yessoensis]	42.00	Show
ESS193365	Predicted protein OS=Nematostella vectensis OX=45351 GN=v1g196785 PE=4 SV=1	40.67	Show
ESS152699	galactose-specific lectin nattectin-like [Perca fluviatilis]	28.33	Show
ESS152700	Putative trypsin-1-like OS=Penaeus vannamei OX=6689 GN=C7M84_010109 PE=4 SV=1	28.33	Show
ESS169861	uncharacterized protein LOC119573276 [Penaeus monodon]	39.17	Show
ESS195728	uncharacterized protein LOC119573276 [Penaeus monodon]	28.33	Show
ESS192386	neurocan core protein-like [Mizuhopecten yessoensis]	21.33	Show

Figure 9. New search engine of KrillDB^2: the homepage sections **(a)** with an example of the results of a fulltext-search **(b)**, of a blast-search **(d)** and the tables listing differentially expressed genes related to the contrast selected **(c)**.

Results of GO term searches are summarized in a table reporting the related genes with corresponding domain (**Figure 10a**) or microRNA (**Figure 10b**) match and associated description.



Search Results

You searched for genes annotated with the keyword GO:0035267 (NuA4 histone acetyltransferase complex). Found 7 matches:

Gene	Match	Description
ESG044786	PTHR11093:SF2	RUVB-LIKE 2
ESG044926	PTHR11093:SF6	RUVB-LIKE 1
ESG037594	PTHR14898	ENHANCER OF POLYCOMB
ESG044885	PTHR12855	DNA METHYLTRANSFERASE 1-ASSOCIATED PROTEIN 1 FAMILY MEMBER
ESG048735	PTHR10880:SF29	MORTALITY FACTOR 4-LIKE PROTEIN 1
ESG036210	PTHR15398:SF6	BROMODOMAIN-CONTAINING PROTEIN 8
ESG057428	PTHR10880:SF29	MORTALITY FACTOR 4-LIKE PROTEIN 1

Search Results

You searched for genes annotated with the keyword GO:0019969 (interleukin-10 binding). Found 2 matches:

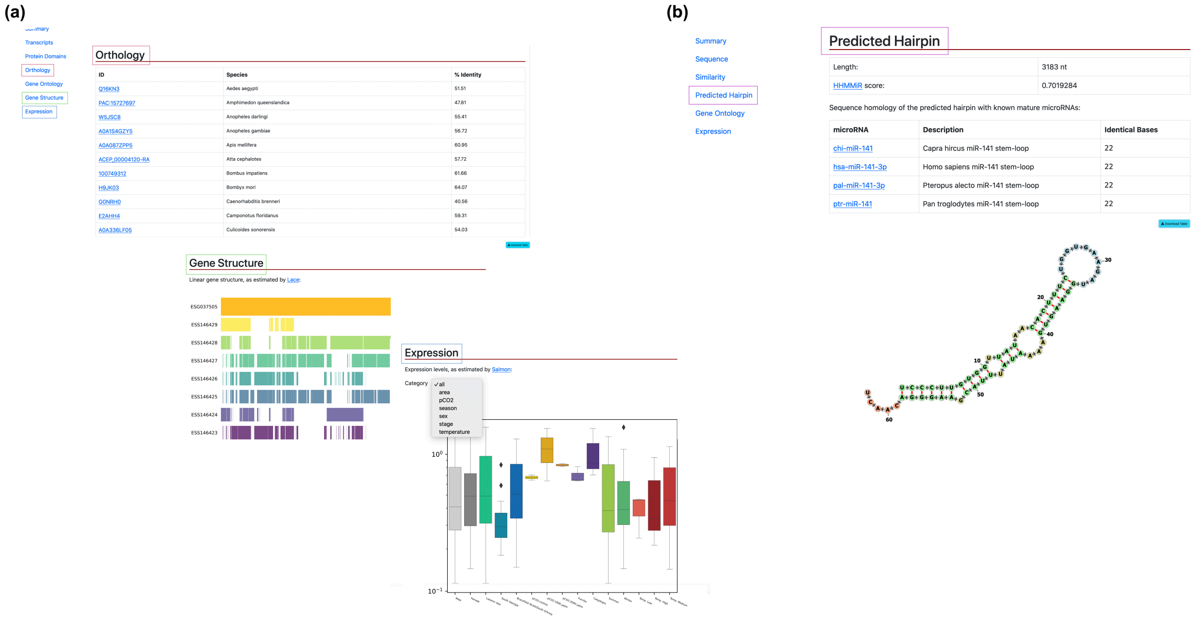
Gene	Match	Description
ESG086142	sko-miR-10	Saccoglossus kowalevskii miR-10 stem-loop
ESG070613	sko-miR-10	Saccoglossus kowalevskii miR-10 stem-loop

Figure 10. Table summarizing results of a GO term search: an example of a GO term associated to genes matching different protein domains **(a)** and a GO term associated to genes matching a known mature microRNA **(b)**.

Both gene and transcript-centric pages have been extended with two new sections: “Orthology" and “Expression” (**Figure 10a**). The Orthology section summarizes the list of orthologous sequences coming from OMA analysis, each one with the species it belongs to and the identity score.

The “Expression” section shows a barplot, implemented using the Seaborn Python library (version 0.11.1), representing abundances estimates obtained from Salmon (Patro et al., 2017). An additional section, called “Gene Structure” (**Figure 11a**), was added to the gene page on the basis of the results coming from the SuperTranscript analysis. Specifically, I modified the STViewer.py Python script (from Lace), optimizing and adapting it to our own data and database structure, in order to produce a visualization of each gene with its transcripts. Since Lace relies on the construction of a single directed splice graph and it is not able to compute it for complex clusters with more than 30 splicing variants, this section is available for a selection of genes only.

The new KrillDB^2 release includes completely updated transcript and gene identifiers. However, the user searching for a retired ID is automatically redirected to the page describing the newest definition of the appropriate transcript or gene.



Orthology

ID	Species	% Identity
Q16KN3	Aedes aegypti	51.51
PAC:15727697	Amphimedon queenslandica	47.81
W5JSC8	Anopheles darlingi	55.41
A0A1S4GZY5	Anopheles gambiae	56.72
A0A087ZPP5	Apis mellifera	60.95
ACEP_00004120-RA	Atta cephalotes	57.72
100749312	Bombus impatiens	61.66
H9JK03	Bombyx mori	64.07
G0NRH0	Caenorhabditis brenneri	40.56
E2AHH4	Camponotus floridanus	59.31
A0A336LF05	Culicoides sonorensis	54.03

Gene Structure

Linear gene structure, as estimated by Lace:

Expression

Expression levels, as estimated by Salmon:

Predicted Hairpin

Length:	3183 nt
HHMMiR score:	0.7019284

Sequence homology of the predicted hairpin with known mature microRNAs:

microRNA	Description	Identical Bases
chi-miR-141	Capra hircus miR-141 stem-loop	22
hsa-miR-141-3p	Homo sapiens miR-141 stem-loop	22
pal-miR-141-3p	Pteropus alecto miR-141 stem-loop	22
ptr-miR-141	Pan troglodytes miR-141 stem-loop	22

Figure 11. Additional sections in gene and transcript pages. The new sections in the gene-centric page show a table listing the orthologous sequences with their belonging species and the identity score, a visualization of the gene structure as estimated by Lace software and a boxplot coming from Expression Atlas analyses **(a)**. Both Orthology and Expression section are integrated also in the transcript-centric page. When a transcript is annotated as a putative microRNA, a “Predicted Hairpin” section displays a visualization of the hairpin predicted secondary structure and tables showing the alignment length, the HHMMiR score and the list of mature microRNAs matching **(b)**.

The KrillDB^2 homepage now includes two additional sections (**Figure 9a**): one is represented by the possibility to perform a BLAST search. Any nucleotide or protein sequence (query) can be aligned against krill sequences stored in the database; results are summarized in a table containing information about the krill transcripts (target) that matched with the user’s query, and the e-value corresponding to the alignment. The other new section, called “Differentially Expressed Genes”, allows the user to browse all the tables listing the genes that were found to be differentially expressed among the conditions I have described above. A drop-down menu gives access to the different comparisons; DEG tables (**Figure 9c**) list for each gene its log fold-change, p- and FDR values as estimated by edgeR. Moreover, each gene is linked to a functional description (if available) inferred from sequence homology searches.

Information about krill transcripts that showed homology with an annotated microRNA is available in the section “Predicted Hairpin” (**Figure 11b**). It contains a summary table with details about the hairpin length and the similarity score (as estimated by HHMMiR), followed by full listing of all the corresponding mature microRNAs (including links to their miRBase page). In addition, an image displaying the predicted secondary structure of the hairpin is included (computed by the “fornac” visualization software from the ViennaRNA suite).

Conclusions and future perspectives

The availability of a large amount of public RNA-seq data capturing Antarctic krill transcripts allowed me to re-assemble its transcriptome and to significantly extend its annotation. In particular, the huge and unprecedented amount of data used has given the possibility to fine tuning the more general approach for transcriptome reconstruction, which could be tested and applied also to other two Euphausiidis species. I have now covered the entire developmental process of *E. superba* and included in the analysis individuals belonging to different seasons and affected by different environmental conditions, while producing the first accurate transcripts collection and comprehensive annotation data for both *M. norvegica* and *T. inermis*, two krill species that have been poorly investigated until now.

The differential expression analyses performed on Antarctic krill dataset showed that the multiple-assemblies combination together with the filtering steps developed allowed not only to preserve the gene signature already described in previous works, but also to add new information and sequences likely involved in important steps of the krill developmental process and responses to environmental changes, with further improvements and investigations, will help the interpretation of the life and evolution of this key species.

Moreover, the expression profiles described for *E. superba* have been combined with those from *M. norvegica* and *T. inermis*, highlighting the presence of a set of genes related to the same processes and functions that are differentially expressed across all the three krill species. Most of these shared genes were found to be differentially expressed due to changes among sex and tissue types (or the interaction of these factors) for both *M. norvegica* and *T. inermis*, while the same genes were differentially expressed due to seasonal changes in *E. superba* samples. Such differences must be thoroughly investigated in order to understand whether they can be addressed to real specific and unique patterns of each species likely related to their significant different latitude, or they can be the result of experimental designs that are not comparable. In fact, *E. superba* dataset lacked information on tissue type, therefore it could not be included in the model matrix design and estimate its weight in differential expression. Consequently, a consistent part of this differential expression was captured by seasonal changes, already found to be predominant in previous works. However, any hypothesis must be tested with more detailed analyses, likely using experimental designs that can be compared.

All the differential expression analyses performed showed the simultaneous presence of genes involved in different steps of the krill moulting cycle and in the reproductive process and sexual maturation that appear to be differentially expressed in same comparisons; this evidence is in accordance with what was already observed in krill (Buchholz, Watkins, Priddle, Morris & Ricketts, 1996) and other krill species (Tarling & Cuzin-Roudy, 2003). In particular, there are already evidences of a strong relation between the krill moulting process and its growth and sexual maturation during the year, which support and confirm the reliability of our results in terms of genes involved in such krill life cycle steps.

The new transcriptomes also provided the possibility of an analysis of the putative microRNAs reconstructed by the approach described. Across all the species, the identified non-coding RNAs were similar in terms of functions and processes in which they appear to be involved. Although this is just a preliminary analysis, the results described already hint at a role of microRNAs in defining the adaptive capabilities of these species to their environment. This represents a promising starting point for the study of non-coding RNAs in krill and in other close species.

Finally, here I describe an update to KrillDB, now renamed KrillDB^2 (available at the address https://krilldb2.bio.unipd.it/), specifically focusing on two aspects: the improvement of the quality and breadth of the Antarctic krill transcriptome sequences previously reconstructed, thanks to the addition of an unprecedented amount of RNA-sequencing data; and, correspondingly, an increase in the amount of annotation information associated to each transcript, which is made available through interactive graphs, images and downloadable files. KrillDB^2 now provides the most complete source of information about the Antarctic krill transcriptome and will offer a reliable starting point development of novel ecological studies. All the new sections have been designed in order to be easily accessible and all tables can be downloaded.

As a future perspective, the availability of the same set of information and results from *Meganyctiphanes norvegica* and *Thysanoessa inermis* suggests the possibility of a greater implementation in the next future of the KrillDB^2 website, with the aim to produce a “multi-krill species” database. This will give the possibility to the scientific community to have a collection of krill sequences, differential expression analysis results, annotation information, orthologs, non-coding RNAs, comparable data from same experimental conditions as well as roles and expression of the same genes across the different species, all in a unique resource.

Acknowledgments

My position was supported by the Helmholtz Vitual Institute "PolarTime": Biological timing in a changing marine environment - clocks and rhythms in polar pelagic organisms (VH-VI-500), headed by Bettina Meyer. I would also like to acknowledge the CAPRI initiative (Calcolo ad Alte Prestazioni per la Ricerca e l'Innovazione", University of Padova Strategic Research Infrastructure Grant 2017) for the technical support and the HPC resources I have used for the analyses.

References

Albessard, E., Mayzaud, P., & Cuzin-Roudy, J. (2001). Variation of lipid classes among organs of the northern krill Meganyctiphanes norvegica, with respect to reproduction. *Comparative Biochemistry and Physiology Part A: Molecular & Integrative Physiology*, *129*(2-3), 373-390. doi: 10.1016/S1095-6433(00)00355-X

Altenhoff, A. M., Gil, M., Gonnet, G. H., & Dessimoz, C. (2013). Inferring hierarchical orthologous groups from orthologous gene pairs. PloS one, 8(1), e53786. doi: 10.1371/journal.pone.0053786

Altenhoff, A. M., Levy, J., Zarowiecki, M., Tomiczek, B., Vesztrocy, A. W., Dalquen, D. A., ... & Dessimoz, C. (2019). OMA standalone: orthology inference among public and custom genomes and transcriptomes. Genome research, 29(7), 1152-1163. doi: 10.1101/gr.243212.118

Anand, A., Villella, A., Ryner, L. C., Carlo, T., Goodwin, S. F., Song, H. J., ... & Taylor, B. J. (2001). Molecular genetic dissection of the sex-specific and vital functions of the Drosophila melanogaster sex determination gene fruitless. *Genetics*, *158*(4), 1569-1595. doi: 10.1093/genetics/158.4.1569

Atkinson, A., Siegel, V., Pakhomov, E. A., Rothery, P., Loeb, V., Ross, R. M., ... & Fleming, A. H. Atkinson, A., Siegel, V., Pakhomov, E. A., Rothery, P., Loeb, V., Ross, R. M., ... & Fleming, A. H. (2008). Oceanic circumpolar habitats of Antarctic krill. Marine Ecology Progress Series, 362, 1-23., 2008. doi: org/10.3354/meps07498

Bao, Y. Y., Qin, X., Yu, B., Chen, L. B., Wang, Z. C., & Zhang, C. X. (2014). Genomic insights into the serine protease gene family and expression profile analysis in the planthopper, Nilaparvata lugens. BMC genomics, 15(1), 1-17. doi: 10.1186/1471-2164-15-507

Batta-Lona, P. G., Bucklin, A., Wiebe, P. H., Patarnello, T., & Copley, N. J. (2011). Population genetic variation of the Southern Ocean krill, Euphausia superba, in the Western Antarctic Peninsula region based on mitochondrial single nucleotide polymorphisms (SNPs). Deep Sea Research Part II: Topical Studies in Oceanography, 58(13-16), 1652-1661. doi: 10.1016/j.dsr2.2010.11.017

Beckmann, H., Hering, L., Henze, M. J., Kelber, A., Stevenson, P. A., & Mayer, G. (2015). Spectral sensitivity in Onychophora (velvet worms) revealed by electroretinograms, phototactic behaviour and opsin gene expression. Journal of Experimental Biology, 218(6), 915-922. doi: 10.1242/jeb.116780

Biscontin, A., Frigato, E., Sales, G., Mazzotta, G. M., Teschke, M., De Pittà, C., ... & Bertolucci, C. (2016). The opsin repertoire of the Antarctic krill Euphausia superba. Marine genomics, 29, 61-68. doi: 10.1016/j.margen.2016.04.010

Bortolotto, E., Bucklin, A., Mezzavilla, M., Zane, L., & Patarnello, T. (2011). Gone with the currents: lack of genetic differentiation at the circum-continental scale in the Antarctic krill Euphausia superba. BMC genetics, 12(1), 1-18. doi: 10.1186/1471-2156-12-32

Buchholz, C. M., Pehlemann, F. W., & Sprang, R. R. (1989). The cuticle of krill (Euphausia superba) in comparison to that of other crustaceans. *Pesq Antárt Bras*, *1*, 103-111.

Buchholz, F., Watkins, J. L., Priddle, J., Morris, D. J., & Ricketts, C. (1996). Moult in relation to some aspects of reproduction and growth in swarms of Antarctic krill, Euphausia superba. Marine Biology, 127(2), 201-208. doi: 10.1007/BF00942104

Bushmanova, E., Antipov, D., Lapidus, A., & Prjibelski, A. D. (2019). rnaSPAdes: a de novo transcriptome assembler and its application to RNA-Seq data. GigaScience, 8(9), giz100. doi: 10.1093/gigascience/giz100

Clark, M. S., Thorne, M. A., Toullec, J. Y., Meng, Y., Peck, L. S., & Moore, S. (2011). Antarctic krill 454 pyrosequencing reveals chaperone and stress transcriptome. PLos one, 6(1), e15919. doi: 10.1371/journal.pone.0015919

Colbourne, J. K., Pfrender, M. E., Gilbert, D., Thomas, W. K., Tucker, A., Oakley, T. H., et al. (2011). The ecoresponsive genome of *Daphnia pulex*. Science 331, 555–561. doi: 10.1126/science.1197761

Cui, H. Y., & Zhao, Z. W. (2020). Structure and function of neuropeptide F in insects. *Journal of Integrative Agriculture*, *19*(6), 1429-1438. doi: 10.1016/S2095-3119(19)62804-2

Cuzin-Roudy, J., & Buchholz, F. (1999). Ovarian development and spawning in relation to the moult cycle in Northern krill, Meganyctiphanes norvegica (Crustacea: Euphausiacea), along a climatic gradient. *Marine Biology*, *133*(2), 267-281. doi: 10.1007/s002270050466

Cuzin-Roudy, J. (2000). Seasonal reproduction, multiple spawning, and fecundity in northern krill, Meganyctiphanes norvegica, and Antarctic krill, Euphausia superba. *Canadian Journal of Fisheries and Aquatic Sciences*, *57*(S3), 6-15. doi: doi.org/10.1139/f00-165

Davidson, N. M., Hawkins, A. D., & Oshlack, A. (2017). SuperTranscripts: a data driven reference for analysis and visualization of transcriptomes. Genome biology, 18(1), 1-10. doi: 10.5281/zenodo.830594

De Pittà, C., Bertolucci, C., Mazzotta, G. M., Bernante, F., Rizzo, G., De Nardi, B., ... & Costa, R. (2008). Systematic sequencing of mRNA from the Antarctic krill (Euphausia superba) and first tissue specific transcriptional signature. BMC genomics, 9(1), 1-14. doi: 10.1186/1471-2164-9-45

De Pittà, C., Biscontin, A., Albiero, A., Sales, G., Millino, C., Mazzotta, G. M., ... & Costa, R. (2013). The Antarctic krill Euphausia superba shows diurnal cycles of transcription under natural conditions. PLoS One, 8(7), e68652. doi: 10.1371/journal.pone.0068652

DeLeo, D. M., & Bracken-Grissom, H. D. (2020). Illuminating the impact of diel vertical migration on visual gene expression in deep-sea shrimp. Molecular Ecology, 29(18), 3494-3510. doi: 10.1111/mec.15570

Ducklow, H. W., Baker, K., Martinson, D. G., Quetin, L. B., Ross, R. M., Smith, R. C., ... & Fraser, W. (2007). Marine pelagic ecosystems: the west Antarctic Peninsula. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *362*(1477), 67-94. doi: 10.1098/rstb.2006.1955

Eriksson, B. J., Fredman, D., Steiner, G., and Schmid, A. (2013). Characterisation and localisation of the opsin protein repertoire in the brain and retinas of a spider and an onychophoran. BMC Evol. Biol. 13:186. doi: 10.1186/1471-2148-13-186

Everson, I. 2000. *Krill: Biology, Ecology and Fisheries*. Oxford: Blackwell Science Ltd.

Flores, H., Van Franeker, J. A., Siegel, V., Haraldsson, M., Strass, V., Meesters, E. H., ... & Wolff, W. J. (2012). The association of Antarctic krill Euphausia superba with the under-ice habitat. *PloS one*, *7*(2), e31775. doi: 10.1371/journal.pone.0031775

Futahashi, R., Kawahara-Miki, R., Kinoshita, M., Yoshitake, K., Yajima, S., Arikawa, K., et al. (2015). Extraordinary diversity of visual opsin genes in dragonflies. Proc. Natl. Acad. Sci. U.S.A. 112, E1247–E1256. doi: 10.1073/pnas.1424670112

Gilbert, D. G. (2019). Longest protein, longest transcript or most expression, for accurate gene reconstruction of transcriptomes? bioRxiv, 829184. doi: 10.1101/829184

Goodall-Copestake, W. P., Perez-Espona, S., Clark, M. S., Murphy, E. J., Seear, P. J., & Tarling, G. A. (2010). Swarms of diversity at the gene cox1 in Antarctic krill. Heredity, 104(5), 513-518. doi: 10.1038/hdy.2009.188

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., ... & Regev, A. (2011). Trinity: reconstructing a full-length transcriptome without a genome from RNA-Seq data. Nature biotechnology, 29(7), 644. doi: 10.1038%2Fnbt.1883

Ha, M., & Kim, V. N. (2014). Regulation of microRNA biogenesis. *Nature reviews Molecular cell biology*, *15*(8), 509-524. doi: 10.1038/nrm3838

Hanamura, Y., Kotori, M. and Hamaoka, S. (1989) Daytime surface swarms of the euphausiid Thysanoessa inermis off the west coast of Hokkaido, northern Japan. Mar. Biol. 102:369–376. doi: 10.1007/BF00428489

Henze, M. J., & Oakley, T. H. (2015). The dynamic evolutionary history of pancrustacean eyes and opsins. Integrative and comparative biology, 55(5), 830-842. doi: 10.1093/icb/icv100

Hering, L, Henze, M.J., Kohler, M., Kelber, A., Bleidorn, C., Leschke, M., Nickel, B., Meyer, M., Kircher, M., Sunnucks, P., Mayer, G. (2012) Opsins in onychophora (velvet worms) suggest a single origin and subsequent diversification of visual pigments in arthropods. Molecular Biology & Evolution, 29(11), 3451-3458. doi: 10.1093/molbev/mss148

Hering, L., and Mayer, G. (2014). Analysis of the opsin repertoire in the tardigrade Hypsibius dujardini provides insights into the evolution of opsin genes in Panarthropoda. Genome Biol. Evol. 6, 2380–2391. doi: 10.1093/gbe/evu193

Hofmann, E. E., & Murphy, E. J. (2004). Advection, krill, and Antarctic marine ecosystems. Antarctic Science, 16(4). doi: 10.1017/s0954102004002275

Hölzer, M., & Marz, M. (2019). De novo transcriptome assembly: A comprehensive cross-species comparison of short-read RNA-Seq assemblers. GigaScience, 8(5), giz039. doi: 10.1093/gigascience/giz039

Höring, F., Biscontin, A., Harms, L., Sales, G., Reiss, C. S., De Pittà, C., & Meyer, B. (2021). Seasonal gene expression profiling of Antarctic krill in three different latitudinal regions. Marine Genomics, 56, 100806. doi: 10.1016/j.margen.2020.100806

Jeffery, N. W. (2012). The first genome size estimates for six species of krill (Malacostraca, Euphausiidae): large genomes at the north and south poles. Polar Biology, 35(6), 959-962. doi: 10.1007/s00300-011-1137-4

Jia, L. Y., Chen, L., Keller, L., Wang, J., Xiao, J. H., & Huang, D. W. (2018). Doublesex evolution is correlated with social complexity in ants. Genome biology and evolution, 10(12), 3230-3242. doi: 10.1093/gbe/evy250

Kadri, S., Hinman, V., & Benos, P. V. (2009). HHMMiR: efficient de novo prediction of microRNAs using hierarchical hidden Markov models. BMC bioinformatics, 10(1), 1-12. doi: 10.1186/1471-2105-10-S1-S35

Kawaguchi, S., Kilpatrick, R., Roberts, L., King, R. A., & Nicol, S. (2011). Ocean-bottom krill sex. *Journal of plankton research*, *33*(7), 1134-1138. doi: 10.1093/plankt/fbr006

Khajuria, C., Buschman, L. L., Chen, M. S., Muthukrishnan, S., & Zhu, K. Y. (2010). A gut-specific chitinase gene essential for regulation of chitin content of peritrophic matrix and growth of Ostrinia nubilalis larvae. Insect biochemistry and molecular biology, 40(8), 621-629. doi: 10.1016/j.ibmb.2010.06.003

Knapp, E., & Sun, J. (2017). Steroid signaling in mature follicles is important for Drosophila ovulation. *Proceedings of the National Academy of Sciences*, *114*(4), 699-704. doi: 10.1073/pnas.1614383114

Kukurba, K. R., & Montgomery, S. B. (2015). RNA sequencing and analysis. *Cold Spring Harbor Protocols*, *2015*(11), pdb-top084970.

Kulka, D. W., & Corey, S. (1978). The life history of Thysanoessa inermis (Krøyer) in the Bay of Fundy. *Canadian Journal of Zoology*, *56*(3), 492-506. doi: 10.1139/z78-069

Li, W., & Godzik, A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics, 22(13), 1658-1659. doi: 10.1093/bioinformatics/btl158

Liu, J., Li, G., Chang, Z., Yu, T., Liu, B., McMullen, R., ... & Huang, X. (2016). BinPacker: packing-based de novo transcriptome assembly from RNA-seq data. PLoS computational biology, 12(2), e1004772. doi: 10.1371/journal.pcbi.1004772

Lu, B., Zeng, Z., & Shi, T. (2013). Comparative study of de novo assembly and genome-guided assembly strategies for transcriptome reconstruction based on RNA-Seq. *Science China Life Sciences*, *56*(2), 143-155. doi: 10.1007/s11427-013-4442-z

Martins, M. J. F., Lago-Leston, A., Anjos, A., Duarte, C. M., Agusti, S., Serrão, E. A., & Pearson, G. A. (2015). A transcriptome resource for Antarctic krill (Euphausia superba Dana) exposed to short-term stress. Marine genomics, 23, 45-47. doi: 10.1016/j.margen.2015.04.008

Meyer, B., Martini, P., Biscontin, A., De Pittà, C., Romualdi, C., Teschke, M., ... & Kawaguchi, S. (2015). Pyrosequencing and de novo assembly of Antarctic krill (E uphausia superba) transcriptome to study the adaptability of krill to climate-induced environmental changes. Molecular ecology resources, 15(6), 1460-1471. doi: 10.1111/1755-0998.12408

Murphy, E. J., Thorpe, S. E., Tarling, G. A., Watkins, J. L., Fielding, S., & Underwood, P. (2017). Restricted regions of enhanced growth of Antarctic krill in the circumpolar Southern Ocean. *Scientific reports*, *7*(1), 1-14. doi: 10.1038/s41598-017-07205-9

Nicol, S., & Endo, Y. (1997). Krill fisheries of the world (No. 367). Food & Agriculture Org..

O'Brien, J., Hayder, H., Zayed, Y., & Peng, C. (2018). Overview of microRNA biogenesis, mechanisms of actions, and circulation. *Frontiers in endocrinology*, *9*, 402. doi: 10.3389/fendo.2018.00402

Patro, R., Duggal, G., Love, M. I., Irizarry, R. A., & Kingsford, C. (2017). Salmon provides fast and bias-aware quantification of transcript expression. Nature methods, 14(4), 417-419. doi: 10.1038/nmeth.4197

Peng, Y., Leung, H. C., Yiu, S. M., Lv, M. J., Zhu, X. G., & Chin, F. Y. (2013). IDBA-tran: a more robust de novo de Bruijn graph assembler for transcriptomes with uneven expression levels. Bioinformatics, 29(13), i326-i334. doi: 10.1093/bioinformatics/btt219

Pinchuk, A. I., & Hopcroft, R. R. (2006). Egg production and early development of Thysanoessa inermis and Euphausia pacifica (Crustacea: Euphausiacea) in the northern Gulf of Alaska. *Journal of Experimental Marine Biology and Ecology*, *332*(2), 206-215. doi: 10.1016/j.jembe.2005.11.019

Raven J, Caldeira K, Elderfield H, Hoegh-Guldberg O, Liss PS, Riebesell U., Sheperd, J, Turley, C & Watson, A. (2005). Ocean Acidification due to Increasing Atmospheric Carbon Dioxide. Royal Society Policy Document.

Risso, D., & Course, I. B. S. (2015). RNA-seq Normalization and Batch Effect Removal.

Roller, L., Žitňanová, I., Dai, L., Šimo, L., Park, Y., Satake, H., ... & Žitňan, D. (2010). Ecdysis triggering hormone signaling in arthropods. *Peptides*, *31*(3), 429-441. doi: 10.1016/j.peptides.2009.11.022

Sabine, C. L., Feely, R. A., Gruber, N., Key, R. M., Lee, K., Bullister, J. L., ... & Rios, A. F. (2004). The oceanic sink for anthropogenic CO2. *science*, *305*(5682), 367-371. doi: 10.1126/science.1097403

Sales, G., Deagle, B. E., Calura, E., Martini, P., Biscontin, A., De Pittà, C., ... & Jarman, S. (2017). KrillDB: A de novo transcriptome database for the Antarctic krill (Euphausia superba). PLoS One, 12(2), e0171908. doi: 10.1371/journal.pone.0171908

Seear, P. J., Goodall-Copestake, W. P., Fleming, A. H., Rosato, E., & Tarling, G. A. (2012). Seasonal and spatial influences on gene expression in Antarctic krill Euphausia superba. Marine Ecology Progress Series, 467, 61-75. doi: 10.3354/meps09947

Seear, P. J., Tarling, G. A., Burns, G., Goodall-Copestake, W. P., Gaten, E., Özkaya, Ö., & Rosato, E. (2010). Differential gene expression during the moult cycle of Antarctic krill (Euphausia superba). BMC genomics, 11(1), 1-13. doi: 10.1186/1471-2164-11-582

Siegel, V. (2005). Distribution and population dynamics of Euphausia superba: summary of recent findings. Polar Biology, 29(1), 1-22. doi: 10.1007/s00300-005-0058-5

Siegel, V. (Ed.). (2016). Biology and ecology of Antarctic krill. Cham, Switzerland: Springer.

Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., & Zdobnov, E. M. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics, 31(19), 3210-3212. doi: 10.1093/bioinformatics/btv351

Suter, L., Polanowski, A. M., King, R., Romualdi, C., Sales, G., Kawaguchi, S., ... & Deagle, B. E. (2019). Sex identification from distinctive gene expression patterns in Antarctic krill (Euphausia superba). Polar Biology, 42(12), 2205-2217. doi: 10.1007/s00300-019-02592-3

Tarling, G. A., & Cuzin-Roudy, J. (2003). Synchronization in the molting and spawning activity of northern krill (Meganyctiphanes norvegica) and its effect on recruitment. Limnology and Oceanography, 48(5), 2020-2033. doi: 10.4319/lo.2003.48.5.2020

Timofeev, S. F. (1996). Ontogenetic ecology of euphausiid crustaceans (Crustacea, Euphausiacea) of the northern seas. Nauka.

Trivelpiece, W. Z., Hinke, J. T., Miller, A. K., Reiss, C. S., Trivelpiece, S. G., & Watters, G. M. (2011). Variability in krill biomass links harvesting and climate warming to penguin population changes in Antarctica. *Proceedings of the National Academy of Sciences*, *108*(18), 7625-7628. doi: 10.1073/pnas.1016560108

Valentine, J. W., & Ayala, F. J. (1976). Genetic variability in krill. Proceedings of the National Academy of Sciences, 73(2), 658-660. doi: 10.1073/pnas.73.2.658.

Veytia, D., Corney, S., Meiners, K. M., Kawaguchi, S., Murphy, E. J., & Bestley, S. (2020). Circumpolar projections of Antarctic krill growth potential. *Nature Climate Change*, *10*(6), 568-575.

Zane, L., Ostellari, L., Maccatrozzo, L., Bargelloni, L., Battaglia, B., & Patarnello, T. (1998). Molecular evidence for genetic subdivision of Antarctic krill (Euphausia superba Dana) populations. Proceedings of the Royal Society of London. Series B: Biological Sciences, 265(1413), 2387-2391. doi: 10.1098/rspb.1998.0588

Zhao, Q. Y., Wang, Y., Kong, Y. M., Luo, D., Li, X., & Hao, P. (2011, December). Optimizing de novo transcriptome assembly from short-read RNA-Seq data: a comparative study. In BMC bioinformatics (Vol. 12, No. 14, pp. 1-12). BioMed Central. doi: 10.1186/1471-2105-12-S14-S2

Wang, Z., Gerstein, M., & Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nature reviews genetics*, *10*(1), 57-63. doi: 10.1038/nrg2484