PhD in Aquatic Biosciences // No. 45 - 2022

Examining challenges in species-level taxonomy among *Calanus* copepods in the Northern seas using genome and transcriptome data

Apollo Marco Dalonos Lizano

FACULTY OF BIOSCIENCES AND AQUACULTURE

NORD University

www.nord.no

Examining challenges in species-level taxonomy among *Calanus* copepods in the Northern seas using genome and transcriptome data

Apollo Marco Dalonos Lizano

A thesis for the degree of
Philosophiae Doctor (PhD)

PhD in Aquatic Biosciences no. 45 (2022)
Faculty of Biosciences and Aquaculture

PhD in Aquatic Biosciences no. 45 (2022)

Apollo Marco Dalonos Lizano
Examining challenges in species-level taxonomy among *Calanus* copepods in the Northern seas using genome and transcriptome data

ISBN: 978-82-93165-44-6

Print: Trykkeriet NORD

Nord University
N-8049 Bodø
Tel: +47 75 51 72 00
www.nord.no

Preface

This thesis is submitted in fulfilment of the requirements for the degree of Philosophiae Doctor (PhD) at the Faculty of Biosciences and Aquaculture (FBA), Nord University, Bodø Norway. The presented original research was performed as part of the Stipendiat program. The studies carried out were financially supported by Nord University.

The project team consisted of the following members:

Apollo Marco D. Lizano, MSc: PhD candidate
Galice Hoarau, Professor, FBA Nord University: Main Supervisor
Marvin Choquet, Researcher, FBA Nord University / Department of Medical Chemistry and Microbiology, University of Uppsala, Sweden: Co-supervisor
Leslie Noble, Professor, FBA Nord University: Co-supervisor

Apollo Marco D. Lizano
Bodø, August 2022

I dedicate this thesis in the memory of my beloved Aunt, ***Dr. Florentina Lizano****, who always supported me throughout my academic career and taught me to believe that nothing is impossible in a world full of possibilities*

"Ang katalinuhan ay walang bansa. Ito ay nagbubunga sa lahat ng lugar.
Ito ay parang liwanag at hangin. Ito ay katangian ng lahat"
Dr. Jose Rizal

Acknowledgements

I would like to express my deepest and sincerest gratitude to all the people and institutions who helped me throughout the course of my Ph.D. studies.

To my thesis adviser, Galice, who opened the window of opportunities for me, all his countless efforts and sacrifices, and to his superb mentorship and enduring patience, to his trust, and for his motivation to help me finish this degree. Thank you for all the knowledge and wisdom that you have imparted on me. I really appreciate all the moments that you pushed me to improve. Rest assured that I will bring them as I continue my journey towards a career in science.

To my co-supervisor, Marvin, for all his patience and dedication, for all his advice, comments, and wisdom. Without him, I would not be able to finish this Ph.D. I really value all the lessons that you have taught me, especially with writing. Thank you also for believing in my abilities and for supporting me all the time. I have never felt alone in this journey because of you. I will always look up to you and hopefully follow your footsteps.

To the two outstanding lab engineers Irina and Martina, who are always there when you need them and who do not hesitate to help you all the time. Thank you for always having my back and for teaching me all the skills that I needed in the lab. You are the backbone of FBA laboratory, and I will strive to be like you, if ever I have the opportunity to work in a lab.

To Nord University Faculty of Biosciences and Aquaculture, for the moral and financial support from November 2018-October 2022. Thank you for making my Ph.D. so much easier and less complicated.

To all the past and present Ph.D. students from FBA who have been part of my journey. You know who you are. Even though we didn't interact as much, I know in some

ways that you have been very helpful by just listening to me during lunch time. Special shout out to my batch mates, Adnan, Saima, Rbbani, and Mads for their countless support and help whenever I have an emotional breakdown. To Le Qin and Joshua for always being so dependable and having my back all the time.

To all the friends that I have met in these past years who made me overcome living in Norway alone. To my Norwegian friends: William, Eric, Markus, Simon, Chris, Antonije Ayla, and Vilde. To my Erasmus friends: Diego, Pauline, Lotte, Antoine, Jeanne, and Elias. To my CS friends: Stelli, Leon, Steffan, Nills, Torsten, Anders, Oskar, and Magnus. To my Filipino Family here in Bodø: Tita Ruth, Tita Rowena, Uncle Odd, Christian, Daniel, Uncle Magne, Mariann, Madeleine, and to my pingpong buddies.

Most especially, to my family: to my Mother Susana, Father Angelito, sisters, Karen, Kristal, Kristine, and Katrina, brothers, Joker and Leo, and cousins, Francis, Marion, Peter and Nadine. For letting me pursue graduate school and allowing me to go where I wanted to be. For all the moral and financial supports, through the ups and downs.

Thank you all for being a ray of sunshine even on my darkest days (literally).

Table of Contents

List of figures and tables

Figures

Tables

List of abbreviations

HMW – High Molecular Weight

HTS - High Throughput Sequencing

IT IS – Integrated Taxonomic Information System

OBIS – Ocean biodiversity information system

cDNA – complementary deoxyribonucleic acid

mtDNA – mitochondrial deoxyribonucleic acid

mRNA – messenger ribonucleic acid

NGS – Next generation sequencing

POC – Particulate organic matter

RAD-seq – Restriction site-associated DNA sequencing

RGS – Reduced genome sequencing

RRS – Reduced-representation sequencing

SNPs – Single Nucleotide Polymorphisms

TCS – Target Capture Sequencing

WGS – Whole genome sequencing

WoRMS – World register of marine species

SLR – Synthetic long-read sequencing

List of papers

Paper I: **Lizano, A. M.,** Smolina, I., Choquet, M., Kopp, M., & Hoarau, G. (2022). Insights into the species evolution of *Calanus* copepods in the northern seas revealed by de novo transcriptome sequencing. *Ecology and Evolution*, *12*(2), e8606.

Paper II: Choquet, M., **Lizano, A.M.,** Ravinet M., Dhanasiri A.K.S., Hoarau G. (2022). Molecular characterization of species boundaries within the genus *Calanus* in the North Atlantic and Arctic oceans. (Manuscript)

Paper III: *Lizano, A.M., * Choquet, M., Smolina I., Bal Thijs, Dhanasiri A.K.S., Kopp M., Hoarau G. (2022). Genome-wide SNPs reveal three distinct genetic lineages within the cryptic *Calanus glacialis / marshallae* species complex. (Manuscript)

General Abstract

Copepods of the genus *Calanus* dominate the zooplankton biomass in the North Atlantic, Arctic, and Northern Pacific regions (Northern seas) and play a key role in marine ecosystems. *Calanus* species have been the focus of numerous ecological research, especially climate change related studies because of their sensitivity to environmental changes. Indeed, *Calanus* species are among the fastest organisms to respond to climate variations by shifting their distributions, potentially impacting the whole ecosystem. Therefore, these species are often used as indicators of climate change. Despite their popularity, species-level identification remains a challenge within the genus because of their morphological similarities and overlapping distributions. Molecular species identification using limited numbers of mitochondrial and nuclear genes have been very useful but can also have limited resolution to fully discriminate all the closely related species within the genus. Genome-wide molecular markers, such as single nucleotide polymorphism (SNPs) are powerful tools to address taxonomic issues compared to traditional molecular markers. However, generating genomic resources for *Calanus* has been challenging mainly because of their huge and complex genomes ranging between ~6 to 12 Gbps. Nonetheless, reduced-representation sequencing (RRS) methods offer a solution to these problems by sequencing only portions of the whole genome. Therefore, the overall goal of this thesis is to use RRS methods to generate genome-wide polymorphism data to address taxonomical challenges related to the species-level identification in the genus *Calanus* in the Northern seas (i.e., species boundaries, putative hybridization, and species complex).

Phylogenetic relationships within the genus *Calanus* have only been partially resolved. Fifteen new transcriptomes were generated and assembled for five *Calanus* species. Single copy orthologs from transcriptome data were used to reappraise, *Calanus* phylogeny and the resulting phylotranscriptome analysis resolved phylogenetic relationships and revealing a tree topology similar to a previously proposed phylogeny based on morphology. The functional annotation of protein families based on clusters

of orthologous genes (COG) and gene ontology (GO) annotations showed conserved and analogous patterns of protein functions across species and also identified protein-coding genes that are present in other Arthropods but are absent among all *Calanus*. Both genes are related to sphingolipid metabolism, which has not yet been explored in *Calanus*. Furthermore, the two new high-quality *de novo* transcriptomes generated (for *C. hyperboreus* and *C. marshallae*) contributed to the existing transcriptomic resources available for the genus *Calanus.*

The recent application of molecular tools has unveiled numerous areas of sympatry among *C. helgolandicus, C. finmarchicus, C. glacialis,* and *C. hyperboreus* in the North Atlantic and Arctic regions, raising questions about the porosity of species boundaries within the genus *Calanus*. Using single nucleotide polymorphism markers (SNPs) mined from both genome and transcriptome data, species boundaries were assessed for the four species of *Calanus*. Our results indicated the presence of four distinct genetic entities, with *C. finmarchicus* and *C. glacialis* genetically closer to each other. Multiple molecular analyses support the lack of recent genetic introgression among the four species included in the study.

The taxonomic status of *C. glacialis / C. marshallae* species complex has been questioned by taxonomists for decades, because of the extreme similarity on their morphology and mtDNA sequences. Here, genome-wide SNPs were used to characterize the level of genetic variations within the *C. glacialis / C. marshallae* species complex. Interestingly, specimens identified as *C. glacialis* and *C. marshallae* from the Northern seas fell into three evolutionary distinct units. Individuals of *C. glacialis* from the North Atlantic / Arctic locations diverged from Northern Pacific *C. glacialis,* representing two distinct genetic lineages. Both lineages showed further localized sub-structuring within the Arctic and North Atlantic regions. Meanwhile, individuals from the Puget Sound identified as *C. marshallae* formed a highly divergent cluster compared to these two evolutionary lineages, confirming the presence of three distinct lineages within the *C. glacialis / C. marshallae* species complex.

Overall, this thesis has substantially increased our understanding about the challenges related to species-level taxonomy in the genus *Calanus*. Here, the use of recent and advanced molecular tools resolved some of the taxonomic/evolutionary issues previously unanswered by morphology and the few existing mitochondrial and nuclear molecular markers. The application of genome and transcriptome data is crucial to generate the most accurate knowledge about hidden or cryptic diversity, species-complex, species boundaries, and putative hybridization within the genus *Calanus*. The new transcriptomic and genomic resources generated will open the way to new research studies and allow a better understanding of the future impact of climate change in each individual *Calanus* species, especially in the northern pelagic ecosystems.

Sammendrag på Norsk

Copepoder av slekten *Calanus* dominerer dyreplanktonbiomassen i de nordlige Atlanterhavet, arktiske og nordlige stillehavsregioner (nordlige hav) og spiller en nøkkelrolle i marine økosystemer. *Calanus*-arter har vært i fokus for en rekke økologiske forskning, spesielt klimaendringer relaterte studier på grunn av deres følsomhet for miljøendringer. *Calanus*-arter er faktisk blant de raskeste organismene som reagerer på klimavariasjoner ved å skifte distribusjon, noe som potensielt påvirker hele økosystemet. Derfor brukes disse artene ofte som indikatorer på klimaendringer. Til tross for deres popularitet, er identifikasjon på artsnivå fortsatt en utfordring innenfor slekten på grunn av deres morfologiske likheter og overlappende distribusjoner. Identifikasjon av molekylære arter ved bruk av begrenset antall mitokondrielle og kjernefysiske gener har vært veldig nyttig, men kan også ha begrenset oppløsning for fullt ut å diskriminere alle de nært beslektede artene i slekten. Genomomfattende molekylære markører, som enkeltnukleotidpolymorfisme (SNP) er kraftige verktøy for å løse taksonomiske problemer sammenlignet med tradisjonelle molekylære markører. Generering av genomiske ressurser for *Calanus* har imidlertid vært utfordrende hovedsakelig på grunn av deres enorme og komplekse genomene som varierer mellom ~6 til 12 Gbps. Ikke desto mindre tilbyr metoder for redusert representasjon sekvensering (RRS) en løsning på disse problemene ved å sekvensere bare deler av hele genomet. Derfor er det overordnede målet med denne oppgaven å bruke RRS-metoder for å generere genomomfattende polymorfismedata for å adressere taksonomiske utfordringer knyttet til artsnivåidentifikasjon i slekten *Calanus* i de nordlige hav (dvs. artsgrenser, antatt hybridisering og arter kompleks).

Fylogenetiske forhold innen slekten *Calanus* er bare delvis løst. Femten nye transkriptomer ble generert og satt sammen for fem *Calanus*-arter. Enkeltkopiortologer fra transkriptomdata ble brukt til å vurdere, *Calanus*-fylogeni og den resulterende phylotranskriptomanalysen løste fylogenetiske forhold og avslørte en tretopologi som ligner på en tidligere foreslått fylogeni basert på morfologi. Den funksjonelle merknaden

av proteinfamilier basert på klynger av ortologe gener (COG) og genontologi (GO) merknader viste konserverte og analoge mønstre av proteinfunksjoner på tvers av arter og identifiserte også proteinkodende gener som er tilstede i andre leddyr, men er fraværende blant alle *Calanus*. Begge gener er relatert til sfingolipidmetabolisme, som ennå ikke har blitt utforsket i *Calanus*. Videre bidro de to nye høykvalitets de novo-transkriptomene generert (for *C. hyperboreus* og *C. marshallae*) til de eksisterende transkriptomiske ressursene tilgjengelig for slekten *Calanus*.

Den nylige bruken av molekylære verktøy har avslørt mange områder med sympati blant *C. helgolandicus, C. finmarchicus, C. glacialis og C. hyperboreus* i de nord-atlantiske og arktiske områdene, og har reist spørsmål om porøsiteten til artsgrensene innenfor slekten Calanus. Ved å bruke enkeltnukleotidpolymorfismemarkører (SNP) utvunnet fra både genom- og transkriptomdata, ble artsgrenser vurdert for de fire artene av *Calanus*. Resultatene våre indikerte tilstedeværelsen av fire distinkte genetiske enheter, med C*. finmarchicus* og *C. glacialis* genetisk nærmere hverandre. Flere molekylære analyser støtter mangelen på nylig genetisk introgresjon blant de fire artene som er inkludert i studien.

Den taksonomiske statusen til *C. glacialis / C. marshallae*-artskomplekset har blitt stilt spørsmål ved av taksonomer i flere tiår, på grunn av den ekstreme likheten i deres morfologi og mtDNA-sekvenser. Her ble genomomfattende SNP-er brukt for å karakterisere nivået av genetiske variasjoner innenfor *C. glacialis / C. marshallae*-artskomplekset. Interessant nok falt prøver identifisert som *C. glacialis* og *C. marshallae* fra de nordlige hav i tre evolusjonært distinkte enheter. Individer av *C. glacialis* fra steder i Nord-Atlanteren/Arctic divergerte fra det nordlige Stillehavet *C. glacialis*, og representerte to forskjellige genetiske linjer. Begge avstamningene viste ytterligere lokalisert substrukturering innenfor de arktiske og nordatlantiske områdene. I mellomtiden dannet individer fra Puget Sound identifisert som *C. marshallae* en svært divergerende klynge sammenlignet med disse to evolusjonære linjene, noe som

bekrefter tilstedeværelsen av tre distinkte avstamninger innenfor *C. glacialis / C. marshallae*-artkomplekset.

Samlet sett har denne oppgaven betydelig økt vår forståelse av utfordringene knyttet til taksonomi på artsnivå i slekten *Calanus*. Her løste bruken av nyere og avanserte molekylære verktøy noen av de taksonomiske/evolusjonære problemene som tidligere ikke ble besvart av morfologi og de få eksisterende mitokondrielle og nukleære molekylære markørene. Anvendelsen av genom- og transkriptomdata er avgjørende for å generere den mest nøyaktige kunnskapen om skjult eller kryptisk mangfold, artskompleks, artsgrenser og antatt hybridisering innen slekten *Calanus*. De nye transkriptomiske og genomiske ressursene som genereres vil åpne veien for nye forskningsstudier og gi en bedre forståelse av den fremtidige virkningen av klimaendringer i hver enkelt *Calanus*-art, spesielt i de nordlige pelagiske økosystemene

1. Introduction

1.1 Ecological importance of marine zooplankton and their role in climate change

Marine zooplankton communities are comprised of a very diverse group of organisms including ~6,800 described species across 15 phyla (Wiebe et al., 2010; Bucklin et al., 2004; Bucklin et al., 2010). Due to their diversity, marine zooplankton communities perform a wide variety of ecosystem functions (Richardson 2008; Hébert et al., 2017), playing key roles in pelagic food webs, as crucial links between phytoplankton and higher trophic levels (Alcaraz & Calbet 2007). Indeed, as major grazers of the ocean, marine zooplankton enable the transfer of energy from primary producers (i.e., phytoplankton) to higher tropic levels in marine pelagic ecosystems (Lee et al., 1971; Lomartire et al., 2021). They are an important prey item for large predators such as fish, marine birds, marine mammals, and marine invertebrates (Arnkværn et al., 2005; Bonnet & Frid 2004; Cleary et al., 2017); but are also vital for the microbial community as a major source of nutrients (De Corte et al., 2018). Fecal pellets and carcasses of marine zooplankton are rich sources of organic carbon for detritus feeders (Lampert 1989; Steinberg et al., 2000). Furthermore, dead zooplankton slowly reach the deeper parts of ocean, thereby contributing to nutrients needed to sustain diverse benthic communities, such as sponges, echinoderms, anemones, crabs, fish, etc. (Ruhl and Smith 2004). Without zooplankton, the oceans would be devoid of important microorganisms thriving in the deep waters, large fish, and predator's dependent on them which entail, ecological, financial, and social values to the human communities.

Marine zooplankton communities also play a key role in the ocean carbon cycle. The ocean's ability to sequester carbon dioxide (CO_2) relies partially on the biological pump, a process where organic matter in the ocean descends from the surface layers to depth by means of advection, vertical mixing, or transportation by animals such as zooplankton (Turner 2015). CO_2 is absorbed into the ocean from the atmosphere and converted into particulate organic matter (POC) through photosynthesis by the phytoplankton

(Steinberg et al., 2000). The POC is processed by microbes, zooplankton, and their consumers into fecal pellets, organic aggregates (marine snow), and other forms of dissolved organic matter, which are then transported into deeper layers of the ocean by sinking and vertical migration of fish or zooplankton, usually ending up at the bottom of the ocean (i.e., mesopelagic, and bathypelagic zones) (Passow and Carlson 2012). This process continues and eventually reduces the amount of carbon in the atmosphere available for the carbon cycle (**Figure 1**).

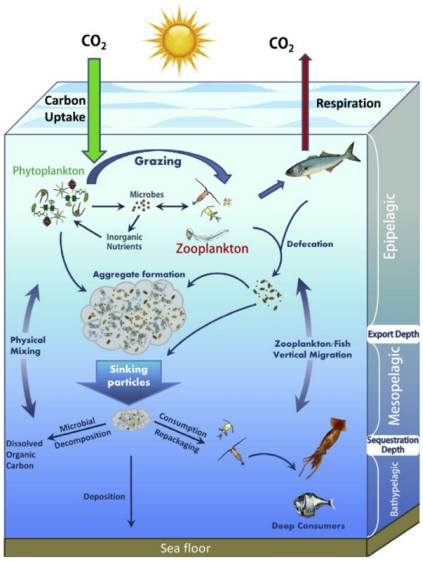


Figure 1. *Pathways of carbon export in the ocean involving marine zooplankton communities (Image adapted from Turner 2015).*

Marine zooplankton are also known as beacons of climate change due to their short life cycles, sensitivity to temperature changes, and scarce exploitation (Richardson 2008; Benedetti et al., 2018). Indeed, most zooplankton have short life spans (<1 year), so there can be a tight link between climate and population dynamics (Hays et al., 2005). Also, their physiological processes are highly sensitive to temperature changes, with metabolic rates doubling or tripling with every 10°C temperature rise (Mauchline 1998). Finally, most marine zooplankton species are not commercially exploited, so long-term

studies relating to response to environmental changes are not affected with trends of commercial exploitation (Richardson 2008). Marine zooplankton have been conspicuously affected by climate change, responding by profound spatial relocations, changes in size and abundance, shifts in their timing of occurrence, and/or extinction events (Garzke et al., 2015; Brun et al., 2019). For example, contrasting long-term trends in species abundance were found for two co-occurring congeneric copepods *Calanus finmarchicus* and *C. helgolandicus* in the North Sea and Skagerrak (south of Norway) (Falkenhaug et al., 2022). The analysis of a 26-year time-series (1994-2019) found a significant shift in dominant zooplankton species from a *C. finmarchicus* dominated to a *C. helgolandicus* dominated zooplankton community. The study also showed that the timing of the annual peak in abundance in both species had advanced by about one month, affecting the stability of predator populations feeding in the area, especially among some commercially important fish species such as cods, herrings, and sand eels (Munk and Nielsen 1994; Falkenhaug et al., 2022).

1.2. *Calanus,* the most important zooplankton genus in the Northern hemisphere?

Copepods (phylum: Arthropoda; class: Maxillopoda; subclass Copepoda) are zooplankton that consist of group of small crustaceans and can be found in almost all aquatic ecosystems (Eyun et al., 2017; Bernot et al., 2021; Walter & Boxshall 2022). There are approximately 15,000 described species of copepods with extremely diverse morphology and lifestyle traits (Walter & Boxshall 2022). Of these, 6,000 species are associated with marine metazoans, mostly as parasites (Kabata 1979; Huys & Boxshall 1991; Boxshall & Halsey 2004). The challenges in both morphological and molecular identification have limited our understanding of the copepod phylogeny, taxonomy, and systematics. Many copepodite samples were historically preserved in formalin, which limits the availability of specimens for combined morphological evaluation and molecular analysis (Bernot et al., 2021). Among the nine orders under the subclass Copepoda, the order Calanoida shows the highest number of species with sequence

data or phylogenetic information available (N=169/2,709, phylogenetic position shown in Figure 2). It is followed by order Harpacticoida (N=30/4,771 species), and Cyclopodia (N=141/4,500 species). These numbers indicate that many copepod taxa remain unexplored in terms of their phylogenetic and taxonomic relationships (Bernot et al., 2021).

The order Calanoida includes around 46 families containing ~1,800 described species, in both marine and fresh water (Mauchline 1988; Walter & Boxshall 2022). They are the most dominant marine zooplankton, making up to ~55% of known zooplanktonic organisms (Suarez-Morales 2015). In the North Atlantic, Arctic, and the Northern Pacific oceans (referred to as "Northern Seas" in the rest of this thesis), the genus *Calanus* is the most dominant genera in terms of biomass (Blaxter et al., 1988; Conover 1988; Fleminger & Hulsemann 1977; Kosobokova et al., 2011; Søreide et al., 2008). Aside from their importance as prey for fish, seabirds, and mammals, several *Calanus* species has been extensively used as model organisms for climate change studies (Reid et al., 2001). They are consequently among the most studied organisms, and there are numerous reasons why they are often used for ecological research. These includes their abundance and wide distributions in the higher latitudes, occupying key positions within their respective food webs (Longhurst 2007). Each species also has high affinity to water masses, making them good indicators of specific waters masses (Blachowiak-Samolyk 2008; Bonnet and Frid 2004; Conover 1998; Bucklin et al., 1995) and their relative short life cycle makes them an ideal species for observing physiological changes in the marine environment (Tarrant et al., 2019; Bailey et al., 2017; Payton et al., 2020).

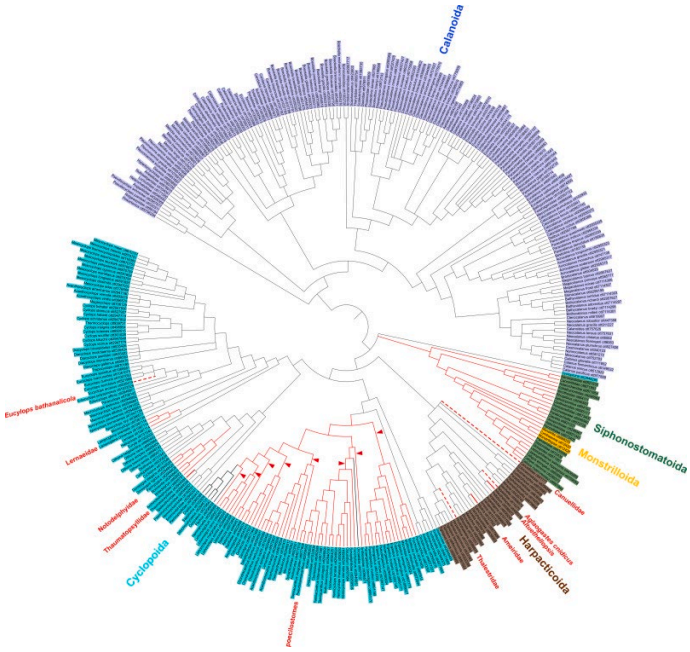


Figure 2. *Phylogeny of 365 copepod taxa from 31 published phylogenies. Colors represent the different taxonomic orders. Parasitic taxa are shown with red branches. Dotted red branches show taxa that contains free-living and parasitic species. (Image adapted from Bernot et al., 2021).*

1.2.1 Biology / Ecology and life history

There are currently 16 species and two subspecies in the genus *Calanus* with a valid taxonomic status according to WoRMs database (accessed March 2022), of which, only nine species have both available sequence data online and validated literature (ITIS, accessed March 2022). These nine species include: *C. finmarchicus* (Gunner, 1765), *C. helgolandicus* (Claus, 1863), *C. hyperboreus* Krøyer, 1838, *C. pacificus* Brodsky, 1948, *C. propinquus* Brady, 1883, *C. similimus* Giesbrecht 1902, *C. sinicus* Brodsky, 1962, *C. marshallae* Frost, 1974, *C. glacialis* Jaschnov, 1955. Of these nine species, seven dominate the Northern Seas. Four species are commonly found in the North Atlantic and Arctic Oceans, namely: *Calanus glacialis, C. helgolandicus, C. finmarchicus,* and *C. hyperboreus* (Conover 1988; Fleminger & Hulsemann 1977; Frost 1974). Three species

are mostly found in the Northern and subarctic Pacific basins: *Calanus pacificus, C. sinicus,* and *C. marshallae* (actual pictures of two *Calanus* species: *C. glacialis* and *C. finmarchicus* are shown in Figure 3).

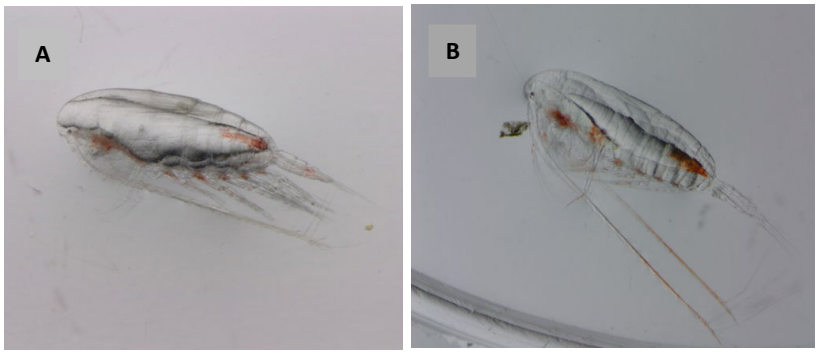


Figure 3*. Actual images of two Calanus species collected from Skjerstadfjord.* ***A.*** *CIV stage of C. glacialis* ***B.*** *Adult female of C. finmarchicus (Photos by Mads Schultz).*

Table 1. *List of taxonomically valid Calanus species throughout the years.*

	Brodsky, 1959	Bradford and Jilllet 1974	Bradford 1988	Worms database based on. Leach, 1819 (accessed March 2022, Boxshall)	Valid taxa under *Calanus* Leach 1819 (accessed from ITIS, March 2022)
	Genus *Calanus* Species	Genus: *Calanus* Species	Genus: *Calanus* Species	Genus: *Calanus* Species	Genus: *Calanus* Species
	C. finmarchicus	*C. finmarchicus*	*C. finmarchicus*	*C. aculeatus* Brady, 1918	*C. finmarchicus* (Gunner, 1765)
	C. helgolandicus	*C. helgolandicus*	*C. helgolandicus*	*C. agulhensis* De Decker, Kaczmaruk & Marska, 1991	*C. helgolandicus* (Claus, 1863)
	C. hyperboreus	*C. hyperboreus*	*C. hyperboreus*	*C. chilensis* Brodsky, 1959	*C. hyperboreus* Krøyer, 1838
	C. pacificus	*C. pacificus*	*C. pacificus*	*C. dorsalis* (Rafinesque, 1815)	*C. pacificus* Brodsky, 1948
	C. australis	*C. australis*	*C. australis*	*C. euxinus* Hulsemann, 1991	*C. pacificus californicus* Brodsky, 1965
	C. chilensis	*C. chilensis*	*C. chilensis*	*C. finmarchicus* (Gunnerus , 1770)	*C. pacificus pacificus* Brodsky 1948
	C. sinicus	*C. sinicus*	*C. sinicus*	*C. glacialis* Jaschnov, 1955	*C. propinquus* Brady, 1883
	C. cristatus	*C. similimus*	*C. similimus*	*C. helgolandicus* (Claus, 1863)	*C. similimus* Giesbrecht 1902
	C. plumchrus	*C. propinquus*	*C. propinquus*	*C. hyperboreus* Krøyer, 1838	*C. sinicus* Brodsky, 1962
	C. tonsius	*C. minor*	*C. glacialis*	*C. jashnovi* Hulsemann, 1994	*C. marshallae* Frost, 1974
		C. glacialis	*C. marshallae*	*C. marshallae* Frost, 1974	*C. glacialis* Jaschnov, 1955
				C. pacificus Brodsky, 1948	
				C. propinquus Brady, 1883	
				C. similimus Giesbrecht 1902	
				C. sinicus Brodsky, 1962	
				C. torticornis (Brady, 1918)	
Total No. of species under genus *Calanus*	10	11	11	16	9 species & 2 sub-genus

Calanus species are mostly broadcast spawners and follow six naupliar and five copepodite stages of development before molting and maturing to adulthood (Melle and Skjoldal 1998; Falk-Petersen et al., 2009). The success of each *Calanus* species highly depends on the accumulation of low-energy carbohydrates and in the conversion of high-energy wax ester lipids, which they store in their lipid sacs for future energy consumption (Lee et al., 2006; Vogedes et al., 2014). The size of this lipid storage is highly variable depending on the species, which allows them to survive for a longer period under starvation and with low availability of food resources (Wilson et al., 2016; Wang et al., 2017). Some *Calanus* species can undergo a short or long-term period of dormancy, called diapause, which is similar to hibernation (Aarflot et al., 2019), where they migrate down to deeper layers of the water column and lower their metabolic activity to the minimum (Hirche 1983; Hirche 1997).

Life cycle duration also varies between species. For instance, *C. finmarchicus* and *C. helgolandicus* have a one-year life cycle, *C. glacialis* have a 1-3 years life cycle, *C. hyperboreus* have a 1-5 years of life cycle (multi-year life cycle) (Vogedes et al., 2014; Falk-Petersen 2009). Meanwhile, *C. marshallae* have a lifetime of least 1 year (Frost 1974; Peterson 1979) similar with *C. sinicus and C. pacificus* (Pu et al., 2004; Runge et al., 1984).

1.2.2 Species identification challenges and advances

Traditionally, *Calanus* species have been identified using a set of morphological characters (prosome length, length and width of caudal ramus, shape, and width of cephalosome, etc.) (Brodsky 1988; Frost 1974), but several studies have shown that morphological characters are not always reliable to distinguish different *Calanus* species accurately, especially between *C. glacialis* and *C. finmarchicus* (e.g., Lindeque et al., 1999; Gabrielsen et al., 2012; Choquet et al., 2018). Given the challenges with morphological species identification, molecular methods have been developed to help identify *Calanus* species without ambiguities.

Molecular species identification in the genus *Calanus* started with allozymes (Sywula et al., 1994), followed by sequencing of mtDNA, using fragments of the COI and 16S rRNA genes (Bucklin et al., 1995). For rapid assessment and species identification, species-specific PCR assays have been developed (Hill et al., 2001; Stupnikova et al., 2013), together with restriction fragment length polymorphism to identify species (e.g., Lindeque et al., 1999; Lindeque et al., 2004), without resorting to sequencing. However, this single locus mtDNA based taxonomy can have limited resolution due to uniparental inheritance and relatively short sequence size (Hurst & Jiggins 2005; Sundt & Melle et al., 1998; Ashijan et al., 2017). As hybridization has been suspected in the genus (Parent et al., 2012), adequate co-dominant nuclear markers that are biparentally inherited, have been developed. Smolina et al. (2014) generated a panel of 12 insertion/deletion (InDels) markers, allowing for the identification of several *Calanus* species including: *C. hyperboreus, C. helgolandicus, C. glacialis,* and *C. finmarchicus.*

1.2.3 Species distribution

The use of molecular tools for species identification allowed redrawing the distributional ranges of *Calanus* species in the North Atlantic and Arctic Oceans, revealing much wider and more overlapping distributions than previously described (Choquet et al., 2017). *Calanus glacialis,* previously considered as an endemic Arctic shelf species, was thereby detected in several fjords along the Norwegian coast, as far south as 60°N where only *C. finmarchicus* was reported in the past (based on morphological identification). Meanwhile, the Arctic *C. hyperboreus*, which had previously been reported to occur in the Northern Norwegian Sea was recorded by molecular tools in large proportions along the Norwegian coast, and as far south as 58°N, in Oslofjord (Choquet et al., 2017). *C. helgolandicus* known as a pseudo-oceanic species was known to occur from the Mediterranean Sea up to the North Sea (58°N). However, recent genetic analyses identified individuals in several Norwegian fjords and in the Norwegian Sea as far north as 70°N. *C. finmarchicus* originally considered to be an indicator species of North Atlantic water masses are now confirmed to be found as far

north as 87°N and as far east in the Arctic at the eastern border of the Laptev Sea (78°N, 113°E) (Choquet et al., 2017).

In the Northern Pacific region, three species of *Calanus* are commonly found*: C. pacificus, C. sinicus,* and *C. marshallae*. Geographic distributions of these three species are based solely on morphological identification, thus lacking species distribution data based on molecular identification (Ocean Biodiversity Information System (OBIS) database, accessed June 2022). Figure 4 illustrates the current available distribution ranges of Pacific *Calanus* based on OBIS database. *Calanus marshallae* can be found from the Northern Pacific extending through the Northern Arctic region (4,191 occurrence records identified at the species-level: 15-147 meters). *C. pacificus* is broadly distributed across the Pacific with 3,623 documented observations, from 1906 to 2020. *C. sinicus*, meanwhile, have the lowest number of reported sightings with only 1,664 occurrences recorded. Most of these sightings are reported in the South China Sea area, including the Yellow Sea, which is located off the coast of Japan (Uye 2000; Pu et al., 2004; Yang et al., 2014).

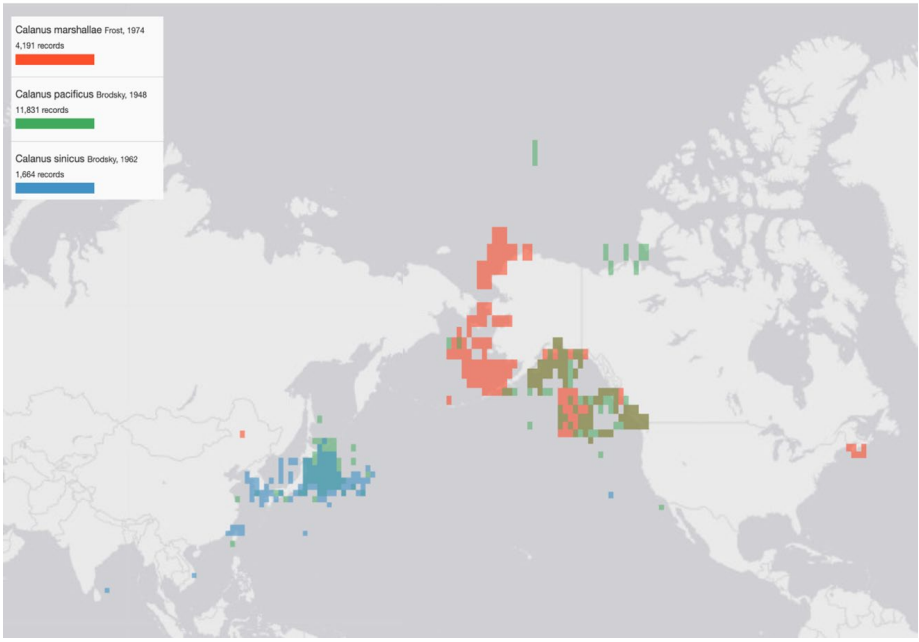


***Figure 4**. Species distribution map of the three Calanus species from the Northern Pacific region: C. marshallae, C. pacificus, and C. sinicus. The figure was created using OBIS distribution map creator based on reported observations.*

1.2.4 Unresolved taxonomical issues

Species of the copepod genus *Calanus* present several challenges for species-level taxonomy as they contain numerous groups of sibling species, which are difficult or nearly impossible to discriminate based on morphology alone (Bucklin et al., 1995; Lindeque et al., 1999). The genus also shows overlapping geographical distributions making discrimination of species a huge problem, especially in the regions of sympatry (Bucklin et al., 1995; Choquet et al., 2017; Choquet et al., 2018). These resulted in multiple questions and unresolved issues regarding the species-level taxonomy in the genus *Calanus.*

1.2.4.1 Phylogenetic relationships

Phylogenetic relationships in the genus *Calanus* have not yet been fully resolved. A phylogeny based on several morphological characters (i.e., relative size of accessory photoreceptor, caudal ramus, anal segment, genital pore, etc.) was proposed by Frost in 1974. The morphology-based identification identified two different groups and a separate *C. hyperboreus* clade. The *C. finmarchicus* group included two other species, *C. glacialis* and *C. marshallae*, while the *C. helgolandicus* group included six other species including *C. pacificus* and *C. sinicus*, and four other species not included in this study (Figure 5A). Molecular phylogenies later emerged based on two genetic markers 16S rRNA (Bucklin et al., 1995; Figure 5B) and 28S rRNA (Kozol et al., 2012; Figure 5C), both phylogenies showed lack of congruence. Although the 28S-based phylogeny seemed to agree with the morphology-based phylogeny on the clustering of species, some branches were not well supported. In contrast, the 16S-based phylogeny suggested a different grouping of species, with several species from *C. helgolandicus* group (identified by Frost 1974) not clustering together in a single clade and other clades only supported by low bootstrap values. Contrasting tree topologies on both molecular analyses can be due to the low resolution provided by a single molecular marker in both instances.

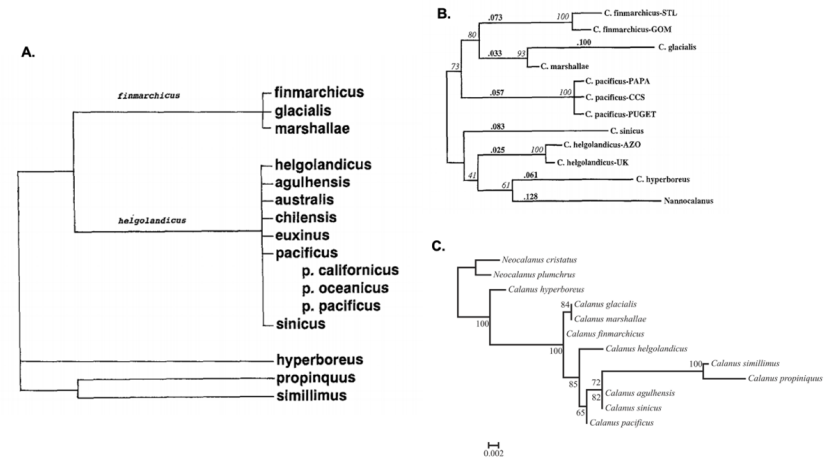


Figure 5. *The three currently available phylogenetic trees for the genus Calanus.* ***A.*** *based on morphology (Frost 1974) reported in Bucklin et al., 1995.* ***B.*** *based on the 16S rRNA gene (Bucklin et al., 1995).* ***C.*** *based on the 28S rRNA gene (Kozol et al., 2012).*

1.2.5.2 Species boundaries and putative hybridization

A large-scale species biogeography study performed by Choquet et al., (2017) using six InDels markers on more than 4,000 *Calanus* individuals collected from 83 locations across the North Atlantic and Arctic Oceans revealed broader and more overlapping distributions among the four *Calanus* species: *C. glacialis, C. finmarchicus, C. helgolandicus,* and *C. hyperboreus*. Numerous areas of sympatry were unveiled, raising questions on the potential for inter-specific hybridization. A particular emphasis was placed on *C. glacialis* and *C. finmarchicus* due to the earlier claim of putative hybridization between the two species (Parent et al., 2012), and because of their vast zone of sympatry identified in Choquet et al. (2017). After genotyping 684 individuals using 10 microsatellite loci developed for *C. finmarchicus*, Parent et al., 2012 reported the presence of hybrids in areas of sympatry in the East Canadian Arctic basin. Meanwhile, Choquet et al., 2017 analyzed thousands of individuals across the North Atlantic and Arctic Oceans. Notably, no hybrids were identified in this study using nuclear, co-dominant InDels markers. In addition, Choquet et al., 2020 studied the *Calanus* community specifically during the reproductive season in two newly uncovered subarctic areas of sympatry, where *C. glacialis* and *C. finmarchicus* co-occur in similar abundances. Genetic admixture tests of six InDels markers performed on all 1,126 individuals revealed no indications of hybridization, implying a strong reproductive isolation mechanism even in sympatric areas. However, as of today, there has still been no consensus agreement on the validity of the InDel markers to assess hybridization and to whether there is, or not, hybridization happening between these two *Calanus* species (Choquet et al. 2020; Parent et al. 2021; Choquet et al., 2021). Genome-wide molecular markers and wide geographic sampling are necessary to resolve this question of putative hybridization between *C. glacialis* and *C. finmarchicus*.

1.2.5.3 Cryptic species

Morphological similarity and overlapping geographical ranges among *Calanus* species have resulted in persistent problems in their identification resulting in species

complex– defined as a group of closely related organisms that are so similar in appearance that species boundaries between them are often unclear (De Queiroz 2007). Species complexes contain several "cryptic species", which are genetically distinct species, but are erroneously classified as a single species due to their superficial morphological similarities.

One example of cryptic species in the genus *Calanus* is the species complex comprising of *C. glacialis* and *C. marshallae*. Previously, only a single morphological character (relative size of photoreceptors, Frost 1974) was used to distinguish between these two recognized species. Both mitochondrial COI and 16S rRNA genes separated them by only one-base pair resulting to taxonomic experts questioning their validity as two separate taxa (Sundt & Melle 1998; Ashijan et al., 2003; Ashijan et al., 2017). Interestingly, the InDel markers developed from *C. finmarchicus* and *C. glacialis* transcriptome and genome data (Smolina et al., 2014), which have been shown to reliably discriminate between *C. finmarchicus*, *C. glacialis*, *C. hyperboreus* and *C. helgolandicus* (Smolina et al., 2014; Choquet et al., 2017), are not able to detect any difference between individuals morphologically identified as *C. glacialis* versus individuals morphologically identified as *C. marshallae* (Smolina, *personal communication)*. Another case of cryptic species is between *C. agulhensis* and *C. sinicus*. A multi-gene analyses showed the lack of genetic divergence between these two closely related taxa (Kozol et al., 2012). The large subunit of 28S revealed the absence of genetic variation between these two species. The lack of genetic distinction between *C. sinicus* and *C. agulhensis* raised questions of whether *C. agulhensis* warrants status of a distinct species. Lastly, there has also been reports of species complex existing between *C. helgolandicus* and *C. euxinus. Calanus euxinus* has recently been designated as a distinct species, although it is known to be closely related to *C. helgolandicus*. Both species are known to occur sympatrically in the Black Sea. Very subtle morphological differences distinguish these two species (Unal et al., 2006). Sequence variation using mtDNA COI resulted to only <0.5% sequence divergence, which is substantially low in comparison with other congeneric pairs. In general, the lack of resolution among these species-

complex, especially among closely related species indicate a clear need for more intensive and extensive ecological and genetic analyses.

1.2.6 Genomic resources

There are currently no published whole-genome data for the genus *Calanus* and only five mitochondrial genomes are found in public database (*C. glacialis, C. finmarchicu*s, *C. sinicus, C. hyperboreus, and C. similimus*) (NCBI genome database, accessed June 2022). However, whole-genome data for several other species of copepods have already been sequenced and assembled (Bucklin et al., 2018). These species include: *Acartia tonsa, Apocalypsi royi, Eurytemorra affinis, Oithona nana, Tigriopus californicus, T. japonicus,* and *T. kinsinjogensis* (Barreto et al., 2018; Eyun et al., 2017; Jørgensen et al., 2019; Choi et al., 2021; Madoui et al., 2017).

Advancements in sequencing technology now allows genome sequencing even for non-model species (da Fonseca et al., 2016; Singhal 2013; Ning et al., 2013; reviewed at Ellegren 2014). However, despite these technological advancements, generating genomic information in the genus *Calanus* still remains challenging (Choquet et al., 2019; reviewed in Bucklin et al., 2018). The main issues still precluding whole genome sequencing of *Calanus* are genome size and complexity, as well as the limited amount of high molecular weight DNA that can be extracted from a single individual. The reported haploid genome sizes of *Calanus* species vary and are estimated to be ca. 6.5 Gbp for *C. finmarchicus* and *C. pacificus,* ca. 8.5 Gbp for *C. sinicus,* ca. 11 Gbp in *C. helgolandicus* and *C. marshallae*, and 12.5 Gbp in *C. glacialis* to *C. hyperboreus* (McLaren 1988). These estimates are 6-10 folds larger than most mammalian genomes (Kapusta et al., 2017). In addition to the large genome size, characterized genomes of other copepod species appear to have low GC content and complex genome architecture with the presence of many repetitive sequences (Bron et al., 2011; Grishanin 2014; Tarrant et al., 2019), making them a challenge to assemble. The recent development in single molecule –sequencing technologies generating longer reads (e.g., PacBio and Oxford

Nanopore Technologies), may help to overcome the limitations of short NGS reads in reconstructing repetitive regions and complex genome architectures (Thompson & Milos 2011; Ameur 2019). Generating longer reads is essential for species with repeated genomes and would be ideal for *Calanus,* but these single molecule sequencing approaches still require large amount of high molecular weight (HMW) DNA, which is clearly a limiting factor for small body sized species such as most zooplankton.

Contrasting to the lack of a genome for the genus *Calanus* in online databases, several whole-transcriptome datasets are available, but only three species are represented so far. In total, seventeen independent RNA-seq datasets are currently available in the NCBI SRA database, of which, only three studies generated complete *de novo* transcriptome assemblies for the following species: *C. helgolandicus* (Asai et al., 2020), *C. finmarchicus* (Lenz et al., 2014; Tarrant et al., 2014), and *C. sinicus* (Ning et al., 2013; Yang et al., 2014).

1.3. High-throughput Sequencing (HTS) / Reduced-representation sequencing (RRS)

High-throughput sequencing technologies also known as next-generation sequencing (NGS) can generate huge amount of data more rapidly and cost-efficiently than ever before (Luikart et al., 2003; Schuster 2008; Metzker 2010). The rapid drop in sequencing cost and large volume of sequencing output makes HTS a powerful tool for phylogenomics, transcriptomics, population genomics, paleogenomics, etc. NGS now allows the sequencing of whole genome for many non-model species at a reasonable cost. Nonetheless, such approach remains challenging for certain taxa, in particular crustaceans, because of their complex genome architecture containing a large number of repetitive sequences, as mentioned earlier (Lenz et al., 2019; Bron et al., 2011). Furthermore, most evolutionary studies require many individuals to be sequenced making the use of whole genomes too costly and unpractical. Alternatively, reduced-representation sequencing (RRS) approaches can be used to generate genome-wide data and genotyping of large numbers of genetic polymorphisms (i.e., SNPs, inDels, SSR,

etc.) without sequencing the entire genome (Van Tassell et al., 2008; Good 2012). This is achieved by sequencing only homologous portions of the genome by either using restriction enzyme digestion (RAD-seq, Figure 6A), sequencing only expressed genes (transcriptome sequencing, Figure 6B), and/or by designing probes or microarrays to target specific regions of the genome (target-capture sequencing, Figure 6C). These approaches, which only look at the fraction of the genome (reduced genome sequencing) give us an opportunity to study the diversity and evolution of many non-model organisms, especially among marine zooplankton including the genus *Calanus.*

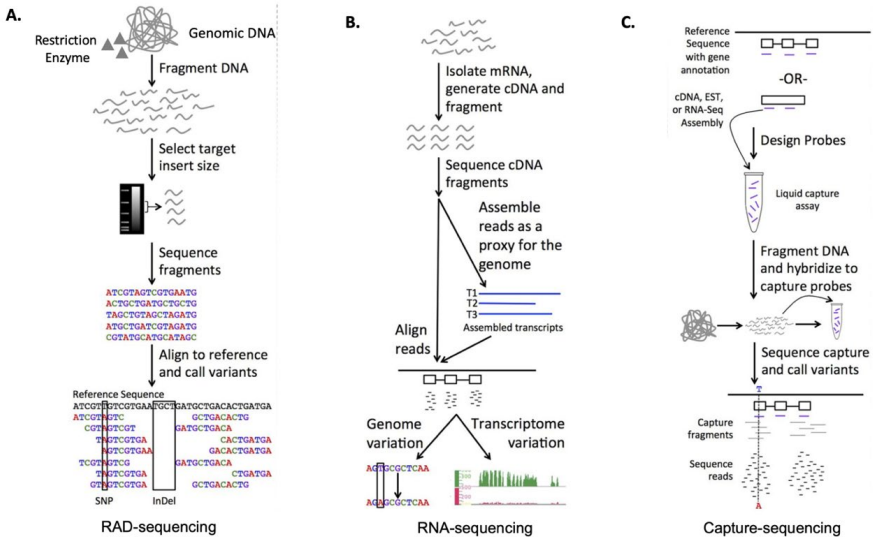


Figure 6. *Reduced-representation sequencing (RRS) strategies to generate genome-wide SNPs.* ***A.*** *Restriction site associated DNA sequencing* ***B.*** *RNA or transcriptome sequencing* ***C.*** *Target-capture sequencing. Figure adapted from Hirsch et al., 2014.*

1.3.1 Restriction Site-Associated DNA sequencing (RAD-seq)

Restriction site-associated DNA sequencing (Baird et al., 2008) was developed particularly for non-model organisms without a reference genome (Davey & Blaxter 2010) and the number of studies relying on this approach has increased exponentially in the recent years (Andrews et al., 2016; Catchen et al., 2017). RAD-seq uses one or several restriction enzymes to digest the genome, resulting in multiple fragments size

selected to generate genomic libraries consisting of all the genomic regions adjacent to restriction enzyme cut sites (Andrews et al., 2016; Peterson et al., 2012).

RAD-seq has been successfully used for some zooplankton species. For example, Blanco-Bercial & Bucklin (2016) used the 2b RAD-seq variant protocol to yield SNPs and investigate population structure patterns in the North Atlantic planktonic copepod *Centropyges typicus.* RAD-seq has also been tested for genus *Calanus,* but with limited success. Due to the high HMW DNA requirements of RAD-seq, several individuals had to be pooled together in a protocol of double-digest RAD-seq (ddRAD), which involves the use of two restriction enzymes. In the end, only 1,871 high-quality SNPs could be genotyped with sufficient coverage in *C. finmarchicus,* with only 343 SNPs in common among the different geographic sites included (Choquet et al., 2019). The somewhat low success of the RAD-seq approach in *Calanus finmarchicus* can be attributed to the combination of different factors such as the pooling of multiple individuals from the same locations limiting the use of sequencing data generated, or the suboptimal choice of restriction enzyme which could not be reliably guided by the simulation software (see explanation on limitations of RAD-seq in Choquet 2021). In summary, the number of SNPs identified for *Calanus finmarchicus* using a RAD-seq approach was not worth the sequencing cost and efforts (Choquet et al., 2019).

1.3.2 Transcriptome Sequencing (RNA-seq)

Transcriptome sequencing is another type of reduced-representation sequencing method which targets expressed genes (transcripts) of specific tissues or cells of a particular species by converting mRNA fragments to cDNA sequences (Hrdlickova et al., 2017; De Wit et al., 2012; Ozsolak et al., 2011). Transcriptome sequencing has become the basis of most gene expression studies (Finotello & Camillo 2015; Spies et al., 2019), transcriptional products, and transcription structure of genes (i.e., splicing patterns, post-transcriptional modification).

There has been a growing interest in utilizing RNA-seq approaches to answer evolutionary questions for non-model species because of the ease in assembling and analyzing transcriptome data compared to genomic data, and in the possibility of obtaining additional information from exonic regions of multiple genes (see e.g., Bi et al., 2012; Lenz et al., 2019; Lenz et al., 2021). Many zooplankton studies have embraced transcriptomics as an alternative to genomics in assessing biodiversity, gene expression, and for population studies (Roncalli et al., 2015; Pratlong et al., 2015; De Wit et al., 2012; Andrews & Luikart 2014). Nonetheless, despite their growing numbers, studies capitalizing on these advantages continue to be limited (Lenz et al., 2021). For the genus *Calanus*, various RNA-seq studies have investigated, for example, the classification of genes associated with developmental cycles from embryos to adult (Lenz et al., 2014) and genes contributing to molecular mechanisms during diapause, diapause termination, and starvation (Ohnishi et al., 2019; Skottene et al., 2019). In addition, studies have also looked at patterns of daily gene expression changes at different latitudes and sea-ice coverage (Payton et al., 2020), and the effects of ocean acidification on the regulation of gene expression (Bailey et al., 2017). Yet, there have been no studies utilizing RNA-seq-based data to mine genes to resolve the phylogeny of the genus *Calanus*, despite recent studies showing that phylotranscriptomes are as powerful as phylogenies derived from whole or partial genome, regardless of the tissue origin (Cheon et al., 2020; Zhao et al., 2021).

Lenz et al., 2021 have recently called to generate more transcriptome resources for marine zooplankton including *Calanus.* This call, under the Ocean Zooplankton Open 'Omics' project. (Ocean Zoop), aims to generate a collection of more comparable *de novo* transcriptomes for hundreds of zooplankton species, ultimately creating a new framework for ecological and physiological studies based on transcriptomic resources (Lenz et al., 2021). Defining species ecological niches, identifying optimal habitats, assessing adaptive capacity, and predicting changes in phenology are just few examples of how transcriptome data can transform our understanding about zooplankton ecology. Moreover, the sequencing of additional transcriptome will enable the

determination and identification of genetic variations from within and across species for several hundreds of species including the genus *Calanus*.

1.3.3 Target-capture sequencing (TCS or Capture-seq)

Target capture sequencing (TCS) or target capture enrichment is a method where only a specific and *a priori* selected subset of genes or regions of the genome are isolated and sequenced. It works by capturing genomic regions of interest by using PCR, microarray chips, or hybridization using target-specific biotinylated probes, which are later recovered by magnetic beads (Meyer & Kircher 2010; Rohland & Reich 2012). With the use of TCS, researchers can sequence up to 100s to millions of genomic regions giving the users the ability to choose which samples to process and sequence in parallel (Manthey et al., 2016; Jones & Good 2016; Andrews et al., 2016). Recent advancement in TCS technology includes genotyping in thousands (Campbell et al., 2015), anchored hybrid enrichment (Lemmon et al., 2012), and target-capture of ultra-conserved element, (Faircloth et al., 2012). Newer approaches also include the target capture of mitochondrial genomes (e.g., Li et al., 2015), overcoming the need for a reference genome to infer historical relationship between species (Jensen et al., 2021). The main difference with RAD-seq is that TCS requires a genomic reference to allow the design of specific capture probes, hence RAD-seq has often been considered more suitable for non-model species with no pre-existing reference. However, even RAD-seq presents challenges and limitations with non-model species, and in particular for species with large and complex genomes, such as *Calanus* (see perspective in Choquet 2021 Choquet et al., 2019). Alternatively, the reference of a sole transcriptome can be used in exome capture sequencing and represent a promising approach to generate large SNP datasets for population structure analyses (Oliveira et al., 2017).

For *Calanus*, no reference genome is available, but several transcriptomes exist. A first attempt at classic exome capture sequencing for *C. finmarchicus* did not lead to promising results in terms of number of SNPs yielded (see Choquet et al., 2019).

However, a fine-tuned protocol using the reference of transcripts mapped into genomic reads, including but not limited to gene regions (illustrated in Figure 7), led to high success for *C. finmarchicus* (Choquet et al., 2019), opening for the first time the possibilities to conduct population genomic studies in the genus. Indeed, the capture enrichment protocol based on 2,656 single-copy genes of *C. finmarchicus* yielded more than 100,000s high-quality SNPs, greatly surpassing the RAD-seq approach. Moreover, the set of capture probes developed by Choquet et al., 2019 originally developed for *C. finmarchicus* was also successfully applied to the congeneric species *C. glacialis*. The optimization of this same protocol (Figure 7) later led to success in other taxa such as the pelagic gastropod *Limacina bulimoides* (Choo et al., 2020), marine algae (Jüterbock et al., *in prep*) and sharks (Wagner et al., *unpublished*).

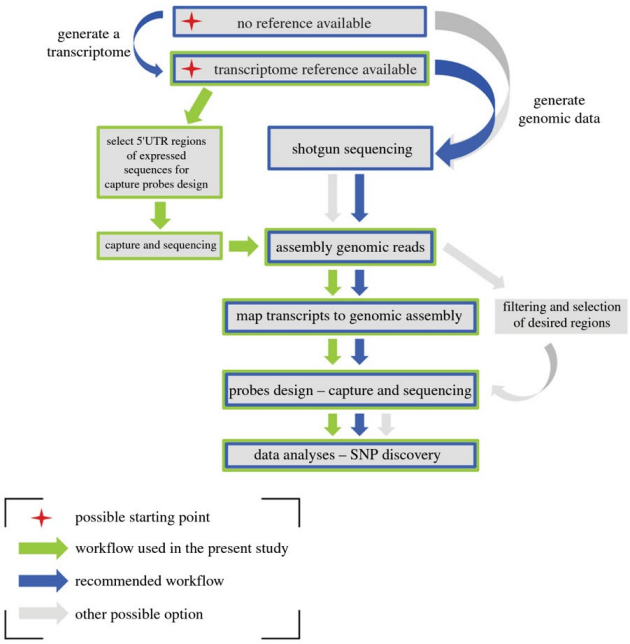


Figure 7. *Target capture workflows for SNP discovery in non-model species with large genomes proposed in Choquet et al. (2019). Figure extracted from Choquet et al. (2019). The original workflow shown by green arrow is the process followed by Choquet et al., 2019 to generate capture-seq data for C. finmarchicus. The optimized workflow in blue was used for pteropods (Choo et al., 2020), sharks (Wagner et al., in prep) and brown algae (Jüterbock et al., in prep).*

2. Knowledge gaps and aims of this thesis

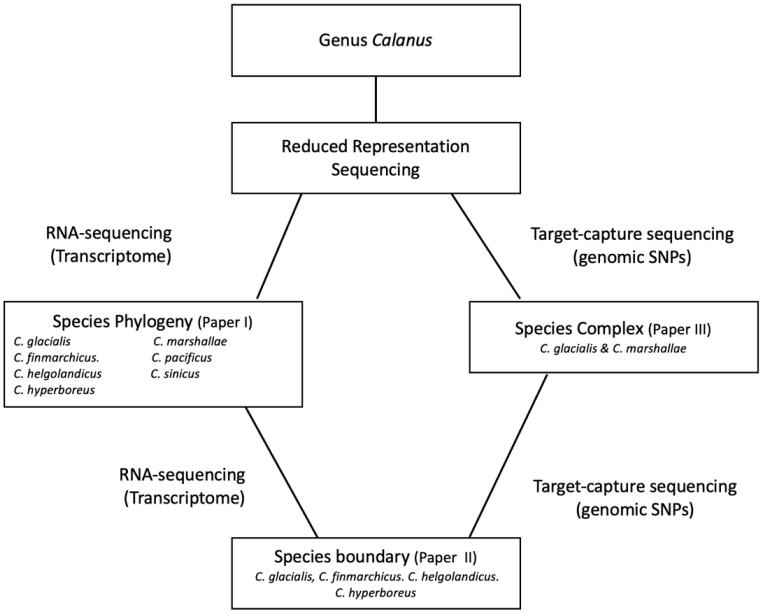
The overall goal of this thesis is to use both genomic and transcriptomic data to examine some of the issues related to the species-level taxonomy in the genus *Calanus* from the Northern seas (North Atlantic, Arctic, and Northern Pacific). This was accomplished through the following specific objectives:

1. Reconstruct the phylogeny and gain insight into the evolutionary relationships among *Calanus* species in the Northern seas based on genome-wide SNPs and transcriptome data (**Paper I**)

2. Address the long withstanding claims of putative hybridization between *C. glacialis* and *C. finmarchicus* species using genome-wide SNPs and transcriptome data (**Paper II**)

3. Characterize the genetic divergence and the taxonomic status between *C. glacialis* and *C. marshallae* species-complex using genome-wide SNPs (**Paper III**)

3. Conceptual framework



4. Main results

4.1 Species phylogeny and evolutionary relationships among the genus *Calanus* in the Northern Seas (Paper I)

RNA-sequencing was used to reconstruct the species phylogeny of seven species of the genus *Calanus* commonly found in the North Atlantic, Arctic, and Northern Pacific regions (Northern Seas). Results showed a fully resolved species tree with a topology concordant with the earlier tree topology based on morphology proposed by Frost 1974. Maximum-likelihood tree showed that the seven species were clustered into three well-supported clades (*C. sinicus*) (*C. helgolandicus + C. pacificus*); (*C. finmarchicus* (*C. glacialis+C. marshallae*); (*C. hyperboreus*). The monophyly of the *C. helgolandicus* group consisting of *C. helgolandicus, C. pacificus,* and *C. sinicus* was supported by high bootstrap values, in contrast with all the previously proposed molecular phylogenies. The highest duplication events were identified for the *C. helgolandicus* group followed by *C. hyperboreus,* and eventually in the *C. finmarchicus* group: *C. finmarchicus, C. glacialis,* and *C. marshallae*. Interestingly, the number of duplication events was not correlated with the genome size of the species. However, the data only showed gene duplication events among expressed genes, and therefore do not reflect other genomic regions. Only whole genome data will unequivocally confirm this result. Gene functional annotation across the seven species of the genus *Calanus* was also performed and showed conserved protein functions across the seven species, which is similar with other Arthropods. However, two gene orthologs involved in sphingolipid metabolism appear to be missing from the seven *Calanus* species but are present in other Arthropods.

4.2 Absence of hybridization between the sympatric Calanus species in the North Atlantic (Paper II)

Genome-wide and transcriptome derived SNPs were used to estimate genetic divergence and to examine species boundaries between the four *Calanus* species from the North Atlantic and Arctic regions, *C. glacialis, C. finmarchicus, C. helgolandicus,* and *C. hyperboreus*. The two independent SNP datasets congruently revealed four clear genetic clusters corresponding to the four *Calanus* species included in this study with no apparent genetic admixture. This study also showed clear genetic differentiation and lack of gene flow between *C. finmarchicus* and *C. glacialis* supporting the absence of introgression and hybridization among sympatric and allopatric populations of these two species.

4.3 Characterization of *C. glacialis* / *C. marshallae* species complex using genome-wide SNPs (Paper III)

Hundreds of gene orthologs derived from RNA-seq data showed high genetic divergence between *C. glacialis* from Norway and *C. marshallae* from Puget Sound. However, the lack of geographic coverage limits the interpretation of this dataset (**Paper I**). In the 3rd chapter of this thesis, analyses of genome-wide SNPs obtained by target capture sequencing from individuals sampled from nine locations (including Pacific, Arctic, and Atlantic) recovered three distinct genetic lineages. These three lineages included both the Arctic/North Atlantic, and Pacific populations of *C. glacialis* and the highly divergent *C. marshallae* lineage from Puget Sound. Lastly, this study also identified possible significant genetic structuring among the North Atlantic/Arctic populations of *C. glacialis*, which can be interesting for future population genomics study of this copepod taxa. The level of genetic differentiation identified using genome-wide SNPs markers demonstrated the importance of large genome-wide datasets in resolving species complex of closely related or sibling species.

5. General discussion

Species taxonomy of the genus *Calanus* have undergone significant revisions in the past, but the taxonomic status of some key species remains in question. In this PhD project, I investigated some of the taxonomic challenges in the genus *Calanus* that were unresolved by traditional morphology and molecular studies based on single or few mitochondrial or nuclear markers. Here, I used two reduced-representation sequencing methods: RNA sequencing (transcriptome) and target-capture sequencing (genome-wide SNPs) to generate powerful datasets of molecular markers, which resulted in an improvement of the species-level taxonomy of the genus *Calanus* and contributed to expanding the limited genomic and transcriptomic resources available for this group of marine organisms. First, I examined *de novo* transcriptomes of seven species of the genus *Calanus* from the Northern hemisphere and gathered some insights about their species evolution **(Paper I).** Then, I explored the species boundaries of four *Calanus* species in the North Atlantic and Artic Oceans, focusing mainly on the putative hybridization between *C. glacialis* and *C. finmarchicus* in both allopatric and sympatric locations **(Paper II).** Lastly, I examined the species complex of two closely related *Calanus* copepods, *C. glacialis* and *C. marshallae* using genome-wide SNP datasets based on a broader geographic range **(Paper III).**

5.1 Huge genomes? The promise of reduced-representation sequencing

By generating transcriptomic and genomic data for the genus *Calanus*, I was able to identify genetic differences and characterize genome-wide patterns of variation for several species of the genus *Calanus* found in the North Atlantic, Arctic, and Northern Pacific.

Phylotranscriptomics were a less costly, but effective alternative to phylogenomics for species with huge and complex genome architecture, such as *Calanus*. Given the relatively low cost of transcriptome and ease in data analysis compared to genome

sequencing, the phylotranscriptomic approach is expected to stimulate wider uses for challenging taxa especially among non-model species. Generating transcriptomic data for multiple species will not only give information that can be useful for phylogenetic reconstruction (**Paper I**), but it can also provide information which can be valuable for comparative analyses (**Paper I**) and for delimiting species boundaries (**Paper II**). Nonetheless, RNA-seq also faces many challenges especially from organisms with many repetitive sequences, unavailable reference genome, high sequence similarity between alternative spliced isoforms, and paralogous gene families (Tarrant et al., 2019). In addition, as RNAs are much less stable than DNA, sampling for RNA-seq requires more precautions. Despite these challenges, the current applications and information provided by RNA-seq make it a very valuable tool for determination of gene expression profiles, gene fusions, SNPs (**Paper II**), alternative splicing events, improvement of gene annotation (**Paper I**), and finding novel gene and transcripts (**Paper I**), expressed gene catalogs, gene orthologs (**Paper I)**.Overall, transcriptomics in general is highly relevant and offers robust information related to zooplankton ecology and evolution, provided that necessary experiments are established to validate gene expression data with traditional and non-traditional methods for ecological measurements (Lenz et al., 2021), and that the issue of scope, focusing mostly on a single species, is taken into consideration (Tarrant et al., 2014; Roncalli et al., 2016; Roncalli et al., 2019; Piccolin et al. 2020; Payton et al., 2021; Lenz et al., 2021).

DNA barcoding studies have unveiled an unforeseeable high number of morphologically cryptic species. However, if speciation has occurred relatively recently and rapidly, the use of single genetic markers, especially the exclusive use of mitochondrial markers, may have limited power in delimiting species, thereby, underestimating the true number of biological species (Moritz & Cicero 2004). The present thesis demonstrated the success and reliability of RRS (RNA-seq and TCS) in generating SNP datasets capable of addressing and examining species-level taxonomic problems in the genus *Calanus.* It also demonstrated the strength of TCS in identifying genome-wide variation/diversity in the focal species *C. finmarchicus,* but also in other

closely related species in the genus *Calanus* (*C. helgolandicus, C. hyperboreus, C. glacialis* and *C. marshallae)*. These genome-wide SNPs are also suitable for future population genomics studies (Choquet et al., *unpublished*; Lizano et al., *unpublished*). However, it requires initial investments, careful optimization, and fine-tuning to deliver high-quality marker density with little cost. Despite the success demonstrated by RRS methods, these approaches can still be limited to fully understand speciation events for the genus *Calanus* (i.e., whole genome duplication, selection, mutation, etc.).

Target-capture sequencing in *Calanus* is useful for both inter-specific (**Paper II**) and intra-specific population studies (*C. finmarchicus*; Choquet et al., *unpublished*, *C. glacialis*; Lizano et al., *unpublished*), but deeper understanding about speciation and demographic history in *Calanus* will require whole genomes. In recent years, the number of marine zooplankton species sequenced for whole genome studies has been growing, including species from Cnidaria to the Urochordata (Bucklin et al., 2018; Bucklin et al., 2021). There are currently 13 genomes available for the order Copepoda (NCBI genome database, accessed June 27, 2022). None of these are from the genus *Calanus*. One of the significant criteria considered by researchers in sequencing the whole genome is the estimated genome size. Notably, most of the genomes of marine zooplankton uploaded in public databases have a size of <1GB. This is mainly because of the sequencing efforts needed and the total cost of generating such output. Single molecule sequencing (Oxford Nanopore Technology, Pacific Biosciences) offers a jump in sequencing technology allowing much longer reads, higher throughput, and improved quantitative accuracy (Thompson & M 2011). Longer reads are crucial for resolving complex and repetitive genomes, but still a challenge for *Calanus* given the limited amount of HMW DNA that can be obtained from a single individual. Interestingly, synthetic long-read sequencing (SLR) coupled with Tell-seq library preparation method offers a promise of producing long sequencing reads of around 10kb with significantly shorter library preparation time and only requiring small quantity of DNA input (as low as 0.1 ng), suitable for WGS (Chen et al., *unpublished*; Universal Sequencing Technology Corporation). The method was only tested and validated for genomes comparable to

human genome size (~3.5 Gbp). How useful this new method will be for *Calanus* species genomes remains to be explored.

5.2. Revisiting morphology according to genome-wide species identification

Traditionally, *Calanus* species have been identified based on morphology alone (Frost 1974; Conover, 1988; Bucklin et al., 1995). However, examination of all morphological characters requires performing demanding preparations on each specimen and is therefore seldom applied during routine zooplankton samples analyses. In the North Atlantic/Arctic oceans, body size and geographical location are commonly used for rapid sorting between species, but this has been proven to be inconsistent and unreliable (Lindeque et al., 1999; Choquet et al., 2017; Choquet et al., 2018). The rise of molecular markers for species ID in the genus *Calanus* has shown the problem caused by morphological-based species ID to differentiate between species. A recent reappraisal of *Calanus* species distribution in the North Atlantic/Arctic Oceans using molecular technique coupled with large scale sampling showed that misidentification is widespread leading to erroneous conclusions about *Calanus* distribution and biogeography (Choquet et al., 2017).

It is thus very interesting that morphological-based (Frost 1974) and transcriptome-based phylogenies (**Paper I**) showed concordant topology, suggesting that some of the morphological characters selected by Frost 1974 may reflect the evolutionary history of the genus, especially those that are related with secondary sexual structures (i.e., relative size of accessory photoreceptor, caudal ramus, anal segment, and genital pore – see Frost 1974). These morphological characters need to be revisited in the light of molecular species ID as potential candidates for investigations of morphological transitions between species and their putative roles as barriers for inter-species copulation (Bucklin et al., 1995). Furthermore, proper taxonomic analyses are needed for pre-identified individuals using genome-wide markers to find useful morphological landmarks that could be used routinely for species ID and are consistent with molecular species identification

5.3 Unresolved taxonomic issues among *Calanus* spp. in the Northern Seas

This thesis identified clear genetic clustering of individuals per species and the lack of putative hybridization between species in the genus *Calanus* is consistent in both genomic and transcriptomic datasets (**Paper I**, **Paper II, and Paper III**). Delimiting species boundaries is important for understanding evolutionary mechanisms and processes. For instance, knowing the species boundary will help define limits within or across, which evolutionary process operates and at what level (Barton & Gale 1993). Over- or under-resolving species boundaries can confound studies aimed at understanding higher functional system such as population or ecosystem level processes.

A species can be defined as two individuals interbreeding via natural reproduction resulting to viable offspring (Wright 1940; Mayr 1942; Dobzhansky & Pavan 1950). However, the 'species concept' does not have a simple definition because of the several alternative contemporary species concepts that obscure a single general concept of a species (i.e., isolation (Mayr 1942; Dobzhansky & Pavan 1950, recognition Paterson & Vrba 1985; Lambert 1995; ecological, Van Valen 1976; Andersson 1990; evolutionary, Simpson 1951; Wiley 1978; Mayden 1997, de Quiroz 2007). De Quiroz (1998; 2007) proposed a unified species concept mostly based on these alternative species concepts. The contemporary unified species concept can be defined as separately evolving metapopulation lineages, where metapopulations are populations made up of connected subpopulations (de Quiroz 1998). The 'species concept' has been used as a fundamental unit of diversity for many analyses in biology, ecology, macroevolution, and conservation biology (Brown et al., 1996; Blackburn and Gaston 1998; Barraclough & Nee 2001). Erroneous species identification can confound conservation measures and strategies leading to over or under exploitation of the natural resources especially in the marine waters. In Norway, small scale fishery of harvesting *C. finmarchicus* started in 2013, with a yearly quota of up to 1000 tons. The most efficient method of harvesting *C. finmarchicus* is using a pelagic trawl with the largest catch during summer/spring months when it forms loose aggregations in the upper water column (0-50m, from 62°N up to 24°E). Depth, timing, and range plays an

important role for determining the optimum measures in harvesting *C. finmarchicus* in Norway. For instance, it was suggested that harvesting of *C. finmarchicus* should be between 62°N up to 24°E (quota is based on NEZ and Jan Mayen zone) with up to 1000 m depth to generate up to 165,000 tons harvest (Långard, 2016 Norwegian management plan for harvesting *Calanus finmarchicus*). Molecular species identification of *Calanus* within this geographic range showed overlapping distribution across four species (*C. glacialis, C. helgolandicus, C. finmarchicus, and C. hyperboreus*). Future management plans should take into consideration the existence of multiple species within a geographic area where *Calanus* fisheries are planned. Selective harvesting based on species could have a negative impact on fish populations due to reduced food availability during seasonal feeding, which can have indirect effect on genetic diversity (when several species of *Calanus* becomes a bycatch). Species-level taxonomy plays an important role to determine the accurate number of species existing in the locality, which can be the basis of conservation strategies for managing biodiversity to avoid exploitation of other non-targeted species.

Molecular studies have shown in the recent years the high prevalence of cryptic species in marine zooplankton, specifically among copepods (reviewed in van der Sprong 2021). However, turning DNA-based discoveries into robustly and formally named taxonomic entities remains challenging (Satler et al., 2013). Consequently, the literature includes a large number of undescribed and unnamed cryptic or pseudo-cryptic species for which identification from one study to another is challenging or nearly impossible to identify (Bucklin et al., 2016, van der Sprong 2021). With large uncertainties about the taxonomy of many marine zooplankton, current management suffers from the lack of expertise and knowledge to properly address these issues related to taxonomy, especially on cryptic species. For instance, while recent oceanographic efforts such as Tara Oceans (Pesant et al., 2015) and Malaspina (Duarte 2015) expeditions have generated staggering wealth of novel observational data on plankton distribution and diversity (Chust et al., 2017), these same data have revealed the extent of our limited knowledge with proper taxonomy and the challenges brought about by cryptic species

in marine zooplankton. Data coming from these expeditions showed that a large fraction of plankton recorded and sampled cannot be assigned to known taxonomic groups (de Vargas et al., 2015; Chust et al., 2017). Furthermore, the effects of climate change in *C. glacialis*, specifically the atlantification of the Arctic is currently being monitored by various programs (Choquet, *personal communications*, Daase & Søreide 2021; Ershova et al., 2021; Hop et al., 2021). The two possible populations or genetic entities identified for *C. glacialis* should be taken into account, as this may result in wrong interpretation of data and possible confusion, which can affect our understanding of the changing climate in the Northern seas. These examples highlight profoundly the need to examine species-level taxonomy in marine zooplankton groups using more advanced molecular techniques such as methods demonstrated in this PhD thesis.

6. Conclusions: Present day challenges & future perspectives

Reduced representation sequencing (RRS) approach using RNA-seq and Capture-seq was successful in examining species-level taxonomy questions in marine zooplankton with large and complex genomes, specifically in the genus *Calanus.*

With almost all phylogenetics/population genetics investigations, more samples are better. This is also the case for the present thesis. For example, including all the *Calanus* species in the phylotranscriptomic analysis would be very informative. Regarding the *C. glacialis/ C. marshallae* complex, a much finer-scale sampling will be required to further understand the distribution and dynamics of the three genetic entities identified.

Despite the success of RRS, developing target capture probes and performing TCS remains tedious and relatively costly especially for 100s or 1000s of samples. Developing practical methods for routine, high throughput, and accurate species-level identification is necessary moving forward. Single nucleotide polymorphism chip set containing useful SNPs to differentiate across several species (derived from TCS) are now being used for the genus *Limacina* (Choo et al., 2020; Choo et al., *unpublished*). This method can also be developed for the genus *Calanus* based on the genome-wide and transcriptome SNPs that we have generated in this study, to increase the number of individuals that can be analyzed without comprising the efficiency and accuracy of species identification.

The genus *Calanus* is gaining popularity as a commercial species for their potential source of Omega-3 fatty acid for both human and aquaculture demands (Pedersen et al., 2014). Examining the species-level taxonomy and resolving taxonomical issues in the genus *Calanus,* such as species boundaries and species complex will help for their proper management and for our understanding of their fate under climate change.

7. References

Aarflot, J. M., Aksnes, D. L., Opdal, A. F., Skjoldal, H. R., & Fiksen, Ø. (2019). Caught in broad daylight: Topographic constraints of zooplankton depth distributions. *Limnology and Oceanography*, 64(3), 849-859.

Alcaraz, M., & Calbet, A. (2007). Large zooplankton: its role in pelagic food webs.

Ameur, A., Kloosterman, W. P., & Hestand, M. S. (2019). Single-molecule sequencing: towards clinical applications. *Trends in biotechnology*, *37*(1), 72-85.

Andersson, L. (1990). The driving force: species concepts and ecology. *Taxon*, *39*(3), 375-382.

Andrews, K. R., Good, J. M., Miller, M. R., Luikart, G., & Hohenlohe, P. A. (2016). Harnessing the power of RADseq for ecological and evolutionary genomics. *Nature Reviews Genetics*, *17*(2), 81-92.

Andrews, K. R., & Luikart, G. (2014). Recent novel approaches for population genomics data analysis.

Arnkværn, G., Daase, M., & Eiane, K. (2005). Dynamics of coexisting *Calanus finmarchicus*, *Calanus glacialis* and *Calanus hyperboreus* populations in a high-Arctic fjord. *Polar Biology*, **28**(7), 528–538. https://doi.org/10.1007/s00300-005-0715-8

Asai, S., Sanges, R., Lauritano, C., Lindeque, P. K., Esposito, F., Ianora, A., & Carotenuto, Y. (2020). De novo transcriptome assembly and gene expression profiling of the copepod Calanus helgolandicus feeding on the PUA-producing diatom Skeletonema marinoi. *Marine drugs*, *18*(8), 392.

Ashjian, Carin J., Robert G. Campbell, Harold E. Welch, Mari Butler, and Donna Van Keuren. "Annual cycle in abundance, distribution, and size in relation to hydrography of important copepod species in the western Arctic Ocean." *Deep Sea Research Part I: Oceanographic Research Papers* 50, no. 10-11 (2003): 1235-1261.

Ashjian, C. J., Campbell, R. G., Gelfman, C., Alatalo, P., & Elliott, S. M. (2017). Mesozooplankton abundance and distribution in association with hydrography on Hanna Shoal, NE Chukchi Sea, during August 2012 and 2013. *Deep Sea Research Part II: Topical Studies in Oceanography*, *144*, 21-36.

Bailey, A., De Wit, P., Thor, P., Browman, H. I., Bjelland, R., Shema, S., & Hop, H. (2017). Regulation of gene expression is associated with tolerance of the Arctic copepod *Calanus glacialis* to CO_2-acidified sea water. *Ecology and Evolution*, **7**(18), 7145–7160

Baird, N. A., Etter, P. D., Atwood, T. S., Currey, M. C., Shiver, A. L., Lewis, Z. A., ... & Johnson, E. A. (2008). Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PloS one*, *3*(10), e3376.

Barraclough, T. G., & Nee, S. (2001). Phylogenetics and speciation. *Trends in ecology & evolution*, *16*(7), 391-399.

Barreto, F. S., Watson, E. T., Lima, T. G., Willett, C. S., Edmands, S., Li, W., & Burton, R. S. (2018). Genomic signatures of mitonuclear coevolution across populations of Tigriopus californicus. *Nature ecology & evolution*, *2*(8), 1250-1257.

Barton, N. H., & Gale, K. S. (1993). Genetic analysis of hybrid zones. *Hybrid zones and the evolutionary process*, 13-45.

Benedetti, F., Guilhaumon, F., Adloff, F., & Ayata, S. D. (2018). Investigating uncertainties in zooplankton composition shifts under climate change scenarios in the Mediterranean Sea. *Ecography*, *41*(2), 345-360

Bernot, J. P., Boxshall, G. A., & Crandall, K. A. (2021). A synthesis tree of the Copepoda: integrating phylogenetic and taxonomic data reveals multiple origins of parasitism. *PeerJ*, *9*, e12034.

Bi, K., Vanderpool, D., Singhal, S., Linderoth, T., Moritz, C., & Good, J. M. (2012). Transcriptome-based exon capture enables highly cost-effective comparative genomic data collection at moderate evolutionary scales. *BMC Genomics*, **13**(1), 403.

Blackburn, T. M., & Gaston, K. J. (1998). Some methodological issues in macroecology. *The American Naturalist*, *151*(1), 68-83.

Blachowiak-Samolyk, K., Kwasniewski, S., Hop, H., & Falk-Petersen, S. (2008). Magnitude of mesozooplankton variability: a case study from the Marginal Ice Zone of the Barents Sea in spring. *Journal of Plankton Research*, *30*(3), 311-323.

Blanco-Bercial, L., & Bucklin, A. (2016). New view of population genetics of zooplankton: RAD-seq analysis reveals population structure of the North Atlantic planktonic copepod Centropages typicus. *Molecular ecology*, *25*(7), 1566-1580.

Blaxter, J. H., Douglas, B., Tyler, P. A., & Mauchline, J. (1998). *The biology of calanoid copepods*. Academic Press.

Bonnet, D., & Frid, C. (2004). Seven copepod species considered as indicators of water-mass influence and changes: Results from a Northumberland coastal station. *ICES Journal of Marine Science*, **61**(4), 485–491. https://doi.org/10.1016/j.icesjms.2004.03.005

Boxshall & Halsey (2004) Boxshall GA, Halsey SH. *An introduction to copepod diversity.* The Ray Society; London: 2004. p. 966.

Bradford, J. M. (1988). Review of the taxonomy of the Calanidae (Copepoda) and the limits to the genus Calanus. *Hydrobiologia*, *167*(1), 73-81.

Bradford, J. M., & Jillett, J. B. (1974). A revision of generic definitions in the Calanidae (Copepoda, Calanoida). *Crustaceana*, 5-16.

Brodsky, K. A., 1959. On the phylogenetic relationship of certain species of *Calanus* (Copepoda) from the northern and southern hemispheres. Zool. J., Acad. of Sci. URSS., 10: 1537–1553.

Bron, J. E., Frisch, D., Goetze, E., Johnson, S. C., Lee, C. E., & Wyngaard, G. A. (2011). Observing copepods through a genomic lens. *Frontiers in zoology*, *8*(1), 1-15.

Brown, J. H., Stevens, G. C., & Kaufman, D. M. (1996). The geographic range: size, shape, boundaries, and internal structure. *Annual review of ecology and systematics*, 597-623.

Brun, P., Stamieszkin, K., Visser, A. W., Licandro, P., Payne, M. R., & Kiørboe, T. (2019). Climate change has altered zooplankton-fuelled carbon export in the North Atlantic. *Nature ecology & evolution*, *3*(3), 416-423.

Bucklin, A., Frost, B. W., & Kocher, T. D. (1995). Molecular systematics of six Calanus and three Metridia species (Calanoida: Copepoda).

Bucklin, A., Nishida, S., & Schiel, S. (2004). Census of marine zooplankton: A new global survey of Marine Biodiversity. *Global Change Newsletter*, *60*, 8-10.

Bucklin, A., Nishida, S., Schnack-Schiel, S., Wiebe, P. H., Lindsay, D., Machida, R. J., & Copley, N. J. (2010). A census of zooplankton of the global ocean. *Life in the World's Oceans: Diversity, Distribution, and Abundance, edited by: McIntyre, A*, 247-265.

Bucklin, A., Steinke, D., & Blanco-Bercial, L. (2011). DNA barcoding of marine metazoa. *Annual review of marine science*, *3*, 471-508.

Bucklin, A., Lindeque, P. K., Rodriguez-Ezpeleta, N., Albaina, A., & Lehtiniemi, M. (2016). Metabarcoding of marine zooplankton: prospects, progress and pitfalls. *Journal of Plankton Research*, *38*(3), 393-400.

Bucklin, A., DiVito, K. R., Smolina, I., Choquet, M., Questel, J. M., Hoarau, G., & O'Neill, R. J. (2018). Population genomics of marine zooplankton. In *Population genomics: Marine organisms* (pp. 61-102).

Bucklin, A., Peijnenburg, K. T., Kosobokova, K. N., O'Brien, T. D., Blanco-Bercial, L., Cornils, A., ... & Weydmann-Zwolicka, A. (2021). Toward a global reference database of COI barcodes for marine zooplankton. *Marine Biology*, *168*(6), 1-26.

Campbell, N. R., Harmon, S. A., & Narum, S. R. (2015). Genotyping-in-Thousands by sequencing (GT-seq): A cost effective SNP genotyping method based on custom amplicon sequencing. *Molecular ecology resources*, *15*(4), 855-867.

Catchen, J. M., Hohenlohe, P. A., Bernatchez, L., Funk, W. C., Andrews, K. R., & Allendorf, F. W. (2017). Unbroken: RADseq remains a powerful tool for understanding the genetics of adaptation in natural populations. *Molecular ecology resources*, *17*(3), 362-365.

Cheon, S., Zhang, J., & Park, C. (2020). Is phylotranscriptomics as reliable as phylogenomics? *Molecular biology and evolution*, *37*(12), 3672-3683.

Choi, B. S., Kim, D. H., Kim, M. S., Park, J. C., Lee, Y. H., Kim, H. J., ... & Lee, J. S. (2021). The genome of the European estuarine calanoid copepod Eurytemora affinis: Potential use in molecular ecotoxicology. *Marine Pollution Bulletin*, *166*, 112190.

Choo, L. Q., Bal, T. M., Choquet, M., Smolina, I., Ramos-Silva, P., Marlétaz, F., ... & Peijnenburg, K. T. (2020). Novel genomic resources for shelled pteropods: a draft genome and target capture probes for Limacina bulimoides, tested for cross-species relevance. *BMC genomics*, *21*(1), 1-14.

Choquet, M., Hatlebakk, M., Dhanasiri, A. K., Kosobokova, K., Smolina, I., Søreide, J. E., ... & Hoarau, G. (2017). Genetics redraws pelagic biogeography of Calanus. *Biology letters*, *13*(12), 20170588.

Choquet, M., Kosobokova, K., Kwaśniewski, S., Hatlebakk, M., Dhanasiri, A. K., Melle, W., ... & Hoarau, G. (2018). Can morphology reliably distinguish between the copepods Calanus finmarchicus and C. glacialis, or is DNA the only way? *Limnology and Oceanography: Methods*, *16*(4), 237-252.

Choquet, M., Smolina, I., Dhanasiri, A. K., Blanco-Bercial, L., Kopp, M., Jueterbock, A., ... & Hoarau, G. (2019). Towards population genomics in non-model species with large genomes: a case study of the marine zooplankton Calanus finmarchicus. *Royal Society open science*, *6*(2), 180608.

Choquet, M. (2021). Customize and get the most out of your reduced-representation sequencing experiment with the new simulation software RADinitio. *Molecular ecology resources*, *21*(2), 351-354.

Chust, G., Vogt, M., Benedetti, F., Nakov, T., Villeger, S., Aubert, A., ... & Ayata, S. D. (2017). Mare Incognitum: A glimpse into future plankton diversity and ecology research. *Frontiers in Marine Science*, *4*, 68.

Cleary, A. C., Søreide, J. E., Freese, D., Niehoff, B., & Gabrielsen, T. M. (2017). Feeding by *Calanus glacialis* in a high arctic fjord: potential seasonal importance of alternative prey. *ICES Journal of Marine Science*, **74**(7), 1937–1946. https://doi.org/10.1093/icesjms/fsx106.

Conover, R. J. (1988). Comparative life histories in the genera Calanus and Neocalanus in high latitudes of the northern hemisphere. *Hydrobiologia*, *167*(1), 127-142.

Daase, M., & Søreide, J. E. (2021). Seasonal variability in non-consumptive mortality of Arctic zooplankton. *Journal of plankton research*, *43*(4), 565-585.

Dam, H. G. (2013). Evolutionary adaptation of marine zooplankton to global change. *Annual Review of Marine Science*, *5*, 349-370.

Davey, J. W., & Blaxter, M. L. (2010). RADSeq: next-generation population genetics. *Briefings in functional genomics*, *9*(5-6), 416-423.

De Corte, D., Srivastava, A., Koski, M., Garcia, J. A. L., Takaki, Y., Yokokawa, T., ... & Herndl, G. J. (2018). Metagenomic insights into zooplankton-associated bacterial communities. *Environmental microbiology*, *20*(2), 492-505.

De Vargas, C., Audic, S., Henry, N., Decelle, J., Mahé, F., Logares, R., ... & Velayoudon, D. (2015). Eukaryotic plankton diversity in the sunlit ocean. *Science*, *348*(6237), 1261605.

De Wit, P., Pespeni, M. H., Ladner, J. T., Barshis, D. J., Seneca, F., Jaris, H., ... & Palumbi, S. R. (2012). The simple fool's guide to population genomics via RNA-Seq: an introduction to high-throughput sequencing data analysis. *Molecular ecology resources*, *12*(6), 1058-1067.

De Queiroz, K. (1998). The general lineage concept of species, species criteria, and the process of speciation. *Endless forms: species and speciation*.

De Queiroz, K. (2007). Species concepts and species delimitation. *Systematic biology*, *56*(6), 879-886.

Dobzhansky, T., & Pavan, C. (1950). Local and seasonal variations in relative frequencies of species of Drosophila in Brazil. *The Journal of Animal Ecology*, 1-14.

Duarte, C. M. (2015). Seafaring in the 21st century: the Malaspina 2010 circumnavigation expedition.

Ellegren, H. (2014). Genome sequencing and population genomics in non-model organisms. *Trends in ecology & evolution*, *29*(1), 51-63.

Ershova, E. A., Kosobokova, K. N., Banas, N. S., Ellingsen, I., Niehoff, B., Hildebrandt, N., & Hirche, H. J. (2021). Sea ice decline drives biogeographical shifts of key Calanus species in the central Arctic Ocean. *Global Change Biology*, *27*(10), 2128-2143.

Eyun, S. I. (2017). Phylogenomic analysis of Copepoda (Arthropoda, Crustacea) reveals unexpected similarities with earlier proposed morphological phylogenies. *BMC evolutionary biology*, *17*(1), 1-12.

Faircloth, B. C., McCormack, J. E., Crawford, N. G., Harvey, M. G., Brumfield, R. T., & Glenn, T. C. (2012). Ultraconserved elements anchor thousands of genetic markers spanning multiple evolutionary timescales. *Systematic biology*, *61*(5), 717-726.

Falk-Petersen, S., Mayzaud, P., Kattner, G., & Sargent, J. R. (2009). Lipids and life strategy of Arctic Calanus. *Marine Biology Research*, *5*(1), 18-39.

Falkenhaug, T., Broms, C., Bagøien, E., & Nikolioudakis, N. (2022). Temporal variability of co-occurring Calanus finmarchicus and C. helgolandicus in Skagerrak. *Frontiers in Marine Science*, 631.

Finotello, F., & Di Camillo, B. (2015). Measuring differential gene expression with RNA-seq: challenges and strategies for data analysis. *Briefings in functional genomics*, *14*(2), 130-142.

Fleminger, A., & Hulsemann, K. (1977). Geographical range and taxonomic divergence in North Atlantic Calanus (C. helgolandicus, C. finmarchicus and C. glacialis). *Marine Biology*, *40*(3), 233-248.

da Fonseca, R. R., Albrechtsen, A., Themudo, G. E., Ramos-Madrigal, J., Sibbesen, J. A., Maretty, L., ... & Pereira, R. J. (2016). Next-generation biology: sequencing and data analysis approaches for non-model organisms. *Marine genomics*, *30*, 3-13.

Frost, B. W. (1974). *Calanus marshallae*, a new species of calanoid copepod closely allied to the sibling species *C. finmarchicus* and *C. glacialis*. *Marine Biology*, **26**(1), 77-99. https://doi.org/10.1007/BF00389089

Gabrielsen, Tove M., Benjamin Merkel, J. E. Søreide, Emma Johansson-Karlsson, Allison Bailey, D. Vogedes, H. Nygård, Øystein Varpe, and Jørgen Berge. "Potential misidentifications of two climate indicator species of the marine arctic ecosystem: Calanus glacialis and C. finmarchicus." *Polar biology* 35, no. 11 (2012): 1621-1628.

Garzke, J., Ismar, S. M., & Sommer, U. (2015). Climate change affects low trophic level marine consumers: warming decreases copepod size and abundance. *Oecologia*, *177*(3), 849-860.

Good, J. M. (2012). Reduced representation methods for subgenomic enrichment and next-generation sequencing. *Molecular methods for evolutionary genetics*, 85-103.

Grishanin, A. (2014). Chromatin diminution in Copepoda (Crustacea): pattern, biological role and evolutionary aspects. *Comparative cytogenetics*, *8*(1), 1.

Hays, G. C., Richardson, A. J., & Robinson, C. (2005). Climate change and marine plankton. *Trends in ecology & evolution*, *20*(6), 337-344.

Hébert, M. P., Beisner, B. E., & Maranger, R. (2017). Linking zooplankton communities to ecosystem functioning: toward an effect-trait framework. *Journal of Plankton Research*, *39*(1), 3-12.

Hill, R., Allen, L., & Bucklin, A. (2001). Multiplexed species-specific PCR protocol to discriminate four N. Atlantic Calanus species, with an mtCOI gene tree for ten Calanus species. *Marine Biology*, *139*(2), 279-287.

Hirche, H. J. (1997). Life cycle of the copepod Calanus hyperboreus in the Greenland Sea. *Marine Biology*, *128*(4), 607-618.

Hirche, H. J. (1983). Overwintering of Calanus finmarchicus and Calanus helgolandicus. *Marine ecology progress series. Oldendorf*, *11*(3), 281-290.

Hirsch, C. D., Evans, J., Buell, C. R., & Hirsch, C. N. (2014). Reduced representation approaches to interrogate genome diversity in large repetitive plant genomes. *Briefings in functional genomics*, *13*(4), 257-267.

Hop, H., Wold, A., Meyer, A., Bailey, A., Hatlebakk, M., Kwasniewski, S., ... & Søreide, J. E. (2021). Winter-spring development of the zooplankton community below sea ice in the Arctic Ocean. *Frontiers in Marine Science*, *8*, 609480.

Hosmani, P. S., Flores-Gonzalez, M., van de Geest, H., Maumus, F., Bakker, L. V., Schijlen, E., ... & Saha, S. (2019). An improved de novo assembly and annotation of the tomato reference genome using single-molecule sequencing, Hi-C proximity ligation and optical maps. *BioRxiv*, 767764.

Hrdlickova, R., Toloue, M., & Tian, B. (2017). RNA-Seq methods for transcriptome analysis. *Wiley Interdisciplinary Reviews: RNA*, *8*(1), e1364.

Hurst, G. D., & Jiggins, F. M. (2005). Problems with mitochondrial DNA as a marker in population, phylogeographic and phylogenetic studies: the effects of inherited symbionts. *Proceedings of the Royal Society B: Biological Sciences*, *272*(1572), 1525-1534.

Huys & Boxshall (1991) Huys R, Boxshall GA. *Copepod evolution.* The Ray Society; London: 1991. p. 468.

Integrated Taxonomic Information System (ITIS), Retrieved March 2022. www.itis.gov, CC0, https://doi.org/10.5066/F7KH0KBK

Jensen, M. R., Sigsgaard, E. E., Liu, S., Manica, A., Bach, S. S., Hansen, M. M., ... & Thomsen, P. F. (2021). Genome-scale target capture of mitochondrial and nuclear environmental DNA from water samples. *Molecular ecology resources*, *21*(3), 690-702.

Jørgensen, T. S., Petersen, B., Petersen, H. C. B., Browne, P. D., Prost, S., Stillman, J. H., ... & Hansen, B. W. (2019). The genome and mRNA transcriptome of the cosmopolitan calanoid copepod acartia tonsa dana improve the understanding of copepod genome size evolution. *Genome biology and evolution*, *11*(5), 1440-1450.

Jones, M. R., & Good, J. M. (2016). Targeted capture in evolutionary and ecological genomics. *Molecular ecology*, *25*(1), 185-202.

Kabata, Z. (1979). Parasitic copepoda of Australian fishes, XII. family Lernanthropidae. *Crustaceana*, 198-213.

Kapusta, A., Suh, A., & Feschotte, C. (2017). Dynamics of genome size evolution in birds and mammals. *Proceedings of the National Academy of Sciences*, *114*(8), E1460-E1469.

Kosobokova, K. N., Hopcroft, R. R., & Hirche, H. J. (2011). Patterns of zooplankton diversity through the depths of the Arctic's central basins. *Marine Biodiversity*, *41*(1), 29-50.

Kozol, R., Blanco-Bercial, L., & Bucklin, A. (2012). Multi-gene analysis reveals a lack of genetic divergence between Calanus agulhensis and C. sinicus (Copepoda; Calanoida). *PLoS One*, *7*(10), e45710.

Lambert, D. (1995). GUEST EDITORIAL: DAVID LAMBERT. *New Zealand Journal of Ecology*, *19*(2), 93-96.

Lampert, W. (1989). The adaptive significance of diel vertical migration of zooplankton. *Funct. Ecol.* 3, 21–27. doi: 10.2307/2389671.

Lee, R. F., Hirota, J., & Barnett, A. M. (1971, December). Distribution and importance of wax esters in marine copepods and other zooplankton. In *Deep Sea Research and Oceanographic Abstracts* (Vol. 18, No. 12, pp. 1147-1165).

Lee, R. F., Hagen, W., & Kattner, G. (2006). Lipid storage in marine zooplankton. *Marine Ecology Progress Series*, *307*, 273-306.

Lemmon, A. R., Emme, S. A., & Lemmon, E. M. (2012). Anchored hybrid enrichment for massively high throughput phylogenomics. *Systematic biology*, *61*(5), 727-744.

Lenz, P. H., Unal, E., Hassett, R. P., Smith, C. M., Bucklin, A., Christie, A. E., & Towle, D. W. (2012). Functional genomics resources for the North Atlantic copepod, Calanus finmarchicus: EST database and physiological microarray. *Comparative Biochemistry and Physiology Part D: Genomics and Proteomics*, *7*(2), 110-123.

Lenz, P. H., Roncalli, V., Hassett, R. P., Wu, L. S., Cieslak, M. C., Hartline, D. K., & Christie, A. E. (2014). De novo assembly of a transcriptome for Calanus finmarchicus (Crustacea, Copepoda)–the dominant zooplankter of the North Atlantic Ocean. *PloS one*, *9*(2), e88589.

Lenz, P. H., Lieberman, B., Cieslak, M. C., Roncalli, V., & Hartline, D. K. (2021). Transcriptomics and metatranscriptomics in zooplankton: wave of the future? *Journal of Plankton Research*, *43*(1), 3-9.

Li, C., Corrigan, S., Yang, L., Straube, N., Harris, M., Hofreiter, M., ... & Naylor, G. J. (2015). DNA capture reveals transoceanic gene flow in endangered river sharks. *Proceedings of the National Academy of Sciences*, *112*(43), 13302-13307.

Lindeque, P. K., Harris, R. P., Jones, M. B., & Smerdon, G. R. (1999). Simple molecular method to distinguish the identity of Calanus species (Copepoda: Calanoida) at any developmental stage. *Marine Biology*, *133*(1), 91-96.

Lindeque, P. K., Harris, R. P., Jones, M. B., & Smerdon, G. R. (2004). Distribution of Calanus spp. as determined using a genetic identification system. *Scientia Marina*, *68*(S1), 121-128.

Lomartire, S., Marques, J. C., & Gonçalves, A. M. (2021). The key role of zooplankton in ecosystem services: A perspective of interaction between zooplankton and fish recruitment. *Ecological Indicators*, *129*, 107867.

Longhurst, Alan R. (2007). *Ecological Geography of the Sea || The Indian Ocean., (), 275–325.* doi:10.1016/B978-012455521-1/50011-5.

Luikart, G., England, P. R., Tallmon, D., Jordan, S., & Taberlet, P. (2003). The power and promise of population genomics: From genotyping to genome typing. *Nature Reviews Genetics*, **4**(12), 981–994.

Madoui, M. A., Poulain, J., Sugier, K., Wessner, M., Noel, B., Berline, L., ... & Wincker, P. (2017). New insights into global biogeography, population structure and natural selection from the genome of the epipelagic copepod Oithona. *Molecular ecology*, *26*(17), 4467-44.

Manthey, J. D., Campillo, L. C., Burns, K. J., & Moyle, R. G. (2016). Comparison of target-capture and restriction-site associated DNA sequencing for phylogenomics: a test in cardinalid tanagers (Aves, Genus: Piranga). *Systematic biology*, *65*(4), 640-650.

Mauchline, J. (1998). *Adv. Mar. Biol. 33: The biology of calanoid copepods*.

Mayden, R. L. (1997). A hierarchy of species concepts: the denouement in the saga of the species problem.

Mayr, E. (1988). The why and how of species. *Biology and Philosophy*, *3*(4), 431-441.

Melle, W., & Skjoldal, H. R. (1998). Reproduction and development of Calanus finmarchicus, C. glacialis and C. hyperboreus in the Barents Sea. *Marine Ecology Progress Series*, *169*, 211-228.

McLaren, I. A., Sevigny, J. M., & Corkett, C. J. (1988). Body sizes, development rates, and genome sizes among Calanus species. In *Biology of copepods* (pp. 275-284). Springer, Dordrecht.

Metzker, M. L. (2010). Sequencing technologies the next generation. *Nature Reviews Genetics*, **11**(1), 31–46.

Meyer, M., & Kircher, M. (2010). Illumina sequencing library preparation for highly multiplexed target capture and sequencing. *Cold Spring Harbor Protocols*, *2010*(6), pdb-prot5448.

Munk, P., & Nielsen, T. G. (1994). Trophodynamics of the plankton community at Dogger Bank: predatory impact by larval fish. *Journal of Plankton Research*, *16*(9), 1225-1245.

National Center for Biotechnology Information. [June 2022]. https://www.ncbi.nlm.nih.gov/genome/.

Ning, J., Wang, M., Li, C., & Sun, S. (2013). Transcriptome sequencing and de novo analysis of the copepod Calanus sinicus using 454 GS FLX. *PloS one*, *8*(5), e63741.

Ocean Biogeographic Information System. [June 2022] www.iobis.org.

Ohnishi, T., Hirai, J., Shimode, S., & Tsuda, A. (2019). Identification of molecular markers associated with starvation in female *Calanus sinicus* (Copepoda: Calanoida). *Marine Ecology Progress Series*, *614*, 51-65.

Oliveira, M. C., Repetti, S. I., Iha, C., Jackson, C. J., Díaz-Tapia, P., Lubiana, K. M. F., ... & Verbruggen, H. (2018). High-throughput sequencing for algal systematics. *European Journal of Phycology*, *53*(3), 256-272.

Ozsolak, F., & Milos, P. M. (2011). RNA sequencing: advances, challenges and opportunities. *Nature reviews genetics*, *12*(2), 87-98.

Parent, G. J., Plourde, S., & Turgeon, J. (2012). Natural hybridization between Calanus finmarchicus and C. glacialis (Copepoda) in the Arctic and Northwest Atlantic. *Limnology and oceanography*, *57*(4), 1057-1066.

Passow, U., & Carlson, C. A. (2012). The biological pump in a high CO2 world. *Marine Ecology Progress Series*, *470*, 249-271.

Paterson, H. E., & Vrba, E. S. (1985). The recognition concept of species. *The Units of Evolution: Essays on the Nature of Species*, 139-58.

Payton, L., Noirot, C., Hoede, C., Hüppe, L., Last, K., Wilcockson, D., Ershova, E. A., Valière, S., & Meyer, B. (2020). Daily transcriptomes of the copepod *Calanus finmarchicus* during the summer solstice at high Arctic latitudes. *Scientific Data*, **7**(1), 1–9. https://doi.org/10.1038/s41597-020-00751-4.

Pesant, S., Not, F., Picheral, M., Kandels-Lewis, S., Le Bescot, N., Gorsky, G., ... & Searson, S. (2015). Open science resources for the discovery and analysis of Tara Oceans data. *Scientific data*, *2*(1), 1-16.

Peterson, W. T. (1979). Life history and ecology of Calanus marshallae Frost in the Oregon upwelling zone.

Peterson, B. K., Weber, J. N., Kay, E. H., Fisher, H. S., & Hoekstra, H. E. (2012). Double digest RADseq: an inexpensive method for de novo SNP discovery and genotyping in model and non-model species. *PloS one*, *7*(5), e37135.

Pettersson, M. E., Rochus, C. M., Han, F., Chen, J., Hill, J., Wallerman, O., ... & Andersson, L. (2019). A chromosome-level assembly of the Atlantic herring genome—detection of a supergene and other signals of selection. *Genome research*, *29*(11), 1919-1928.

Pratlong, M., Haguenauer, A., Chabrol, O., Klopp, C., Pontarotti, P., & Aurelle, D. (2015). The red coral (Corallium rubrum) transcriptome: a new resource for population genetics and local adaptation studies. *Molecular ecology resources*, *15*(5), 1205-1215.

Provan, J. I. M., Beatty, G. E., Maggs, C. A., & Savidge, G. (2007). Expressed sequence tag-derived microsatellites for the cool-water marine copepod *Calanus finmarchicus*. *Molecular ecology notes*, *7*(6), 1369-1371.

Pu, X. M., Sun, S., Yang, B., Zhang, G. T., & Zhang, F. (2004). Life history strategies of Calanus sinicus in the southern Yellow Sea in summer. *Journal of Plankton Research*, *26*(9), 1059-1068.

Reid, P. C., Holliday, N. P., & Smyth, T. J. (2001). Pulses in the eastern margin current and warmer water off the northwest European shelf linked to North Sea ecosystem changes. *Marine Ecology Progress Series*, *215*, 283-287.

Richardson, A. J. (2008). In hot water: zooplankton and climate change. *ICES Journal of Marine Science*, *65*(3), 279-295.

Rohland, N., & Reich, D. (2012). Cost-effective, high-throughput DNA sequencing libraries for multiplexed target capture. *Genome research*, *22*(5), 939-946.

Roncalli, V., Cieslak, M. C., Passamaneck, Y., Christie, A. E., & Lenz, P. H. (2015). Glutathione S-transferase (GST) gene diversity in the crustacean *Calanus finmarchicus*–contributors to cellular detoxification. *PloS one*, *10*(5), e0123322.

Roncalli, V., Turner, J. T., Kulis, D., Anderson, D. M., & Lenz, P. H. (2016). The effect of the toxic dinoflagellate Alexandrium fundyense on the fitness of the calanoid copepod *Calanus finmarchicus*. *Harmful Algae*, *51*, 56-66.

Roncalli, V., Cieslak, M. C., Germano, M., Hopcroft, R. R., & Lenz, P. H. (2019). Regional heterogeneity impacts gene expression in the subarctic zooplankter Neocalanus flemingeri in the northern Gulf of Alaska. *Communications biology*, *2*(1), 1-13.

Ruhl, H. A., & Smith Jr, K. L. (2004). Shifts in deep-sea community structure linked to climate and food supply. *Science*, *305*(5683), 513-515.

Runge, J. A. (1984). Egg production of the marine, planktonic copepod, Calanus pacificus Brodsky: laboratory observations. *Journal of Experimental Marine Biology and Ecology*, *74*(1), 53-66.

Satler, J. D., Carstens, B. C., & Hedin, M. (2013). Multilocus species delimitation in a complex of morphologically conserved trapdoor spiders (Mygalomorphae, Antrodiaetidae, Aliatypus). *Systematic biology*, *62*(6), 805-823.

Schuster, S. C. (2008). Next-generation sequencing transforms today's biology. *Nature Methods*, **5**(1), 16–18.

Singhal, S. (2013). De novo transcriptomic analyses for non-model organisms: An evaluation of methods across a multi-species data set. *Molecular Ecology Resources*, *13*(3), 403-416.

Spies, D., Renz, P. F., Beyer, T. A., & Ciaudo, C. (2019). Comparative analysis of differential gene expression tools for RNA sequencing time course data. *Briefings in bioinformatics*, *20*(1), 288-298.

Søreide, J. E., Falk-Petersen, S., Hegseth, E. N., Hop, H., Carroll, M. L., Hobson, K. A., & Blachowiak-Samolyk, K. (2008). Seasonal feeding strategies of Calanus in the high-Arctic Svalbard region. *Deep Sea Research Part II: Topical Studies in Oceanography*, *55*(20-21), 2225-2244.

Simpson, G. G. (1951). The species concept. *Evolution*, *5*(4), 285-298.

Smolina, I., Kollias, S., Poortvliet, M., Nielsen, T. G., Lindeque, P., Castellani, C., ... & Hoarau, G. (2014). Genome-and transcriptome-assisted development of nuclear insertion/deletion markers for C alanus species (C opepoda: C alanoida) identification. *Molecular ecology resources*, *14*(5), 1072-1079.

Skottene, E., Tarrant, A. M., Olsen, A. J., Altin, D., Østensen, M. A., Hansen, B. H., ... & Olsen, R. E. (2019). The β-oxidation pathway is downregulated during diapause termination in Calanus copepods. *Scientific reports*, *9*(1), 1-13.

Steinberg, D. K., Carlson, C. A., Bates, N. R., Goldthwait, S. A., Madin, L. P., and Michaels, A. F. (2000). Zooplankton vertical migration and the active transport of dissolved organic and inorganic carbon in the Sargasso Sea. *Deep Sea Res. Part I* 47, 137–158. doi: 10.1016/s0967-0637(99)000527

Stupnikova, A. N., Kulagin, D. N., Neretina, T. V., & Mugue, N. S. (2013). Mitochondrial DNA polymorphism (CO1) of three dominant copepod species in the South Atlantic. *Oceanology*, *53*(4), 451-459.

Suárez-Morales, Eduardo (2015). *Thorp and Covich's Freshwater Invertebrates || Class Maxillopoda. 709–755.* doi:10.1016/B978-0-12-385026-3.00029-2.

Sundt, R. C., & Melle, W. (1998). Atlantic observation of Calanus marshallae (Copepoda; Calanoida). *Marine Ecology Progress Series*, *166*, 207-210.

Sywula, T., Namiotko, T., Sell, J., Witkowski, A., & Zajaczkowski, M. (1994). Crustacean species new to Spitsbergen with notes on the polymorphism and the subfossil preservation of Cytherissa lacustris (GO Sars). *Polar Research*, *13*(2), 233-235.

Tarrant, A. M., Baumgartner, M. F., Hansen, B. H., Altin, D., Nordtug, T., & Olsen, A. J. (2014). Transcriptional profiling of reproductive development, lipid storage and molting throughout the last juvenile stage of the marine copepod Calanus finmarchicus. *Frontiers in zoology*, *11*(1), 1-15.

Tarrant, A. M., Nilsson, B., & Hansen, B. W. (2019). Molecular physiology of copepods—From biomarkers to transcriptomes and back again. *Comparative Biochemistry and Physiology - Part D: Genomics and Proteomics*, **30**, 230–247.

Thompson, J. F., & Milos, P. M. (2011). The properties and applications of single-molecule DNA sequencing. *Genome biology*, *12*(2), 1-10.

Toonen, R. J., Puritz, J. B., Forsman, Z. H., Whitney, J. L., Fernandez-Silva, I., Andrews, K. R., & Bird, C. E. (2013). ezRAD: a simplified method for genomic genotyping in non-model organisms. *PeerJ*, *1*, e203.

Turner, J. T. (2015). Zooplankton fecal pellets, marine snow, phytodetritus and the ocean's biological pump. *Progress in Oceanography*, *130*, 205-248.

Unal, E., Frost, B. W., Armbrust, V., & Kideys, A. E. (2006). Phylogeography of Calanus helgolandicus and the Black Sea copepod Calanus euxinus, with notes on Pseudocalanus elongatus (Copepoda, Calanoida). *Deep Sea Research Part II: Topical Studies in Oceanography*, *53*(17-19), 1961-1975.

Uye, S. (2000). Why does Calanus sinicus prosper in the shelf ecosystem of the Northwest Pacific Ocean? *ICES Journal of Marine Science*, *57*(6), 1850-1855.

Van der Sprong, J. (2021). What do we know about cryptic diversity in marine zooplankton? (Unpublished Master's thesis). Universitiet van Amsterdam, Amsterdam Netherlands.

Van Tassell, C. P., Smith, T. P., Matukumalli, L. K., Taylor, J. F., Schnabel, R. D., Lawley, C. T., ... & Sonstegard, T. S. (2008). SNP discovery and allele frequency estimation by deep sequencing of reduced representation libraries. *Nature methods*, *5*(3), 247-252.

Van Valen, L. (1976). Ecological species, multispecies, and oaks. *Taxon*, 233-239.

Vogedes, D., Eiane, K., Båtnes, A. S., & Berge, J. (2014). Variability in Calanus spp. abundance on fine-to mesoscales in an Arctic fjord: implications for little auk feeding. *Marine Biology Research*, *10*(5), 437-448.

Voznesensky, M., Lenz, P. H., Spanings-Pierrot, C., & Towle, D. W. (2004). Genomic approaches to detecting thermal stress in Calanus finmarchicus (Copepoda: Calanoida). *Journal of Experimental Marine Biology and Ecology*, *311*(1), 37-46.

Walter & Boxshall (2022). [07 June 2022]. http://www.marinespecies.org/copepoda.

Wang, B., Xie, X., Liu, S., Wang, X., Pang, H., & Liu, Y. (2017). Development and characterization of novel microsatellite markers for the Common Pheasant (Phasianus colchicus) using RADseq. *Avian Research*, *8*(1), 1-7.

Wang, Y., Li, C., Liu, M., & Jin, X. (2017). Lipids and fatty acids in Calanus sinicus during oversummering in the southern Yellow Sea. *Chinese Journal of Oceanology and Limnology*, *35*(4), 874-882.

Wiebe, P. H., Schnack-Schiel, S. B., & Nishida, S. (2010). Introduction to species diversity of marine zooplankton. *Deep Sea Research Part II: Topical Studies in Oceanography*, *57*(24-26), 2061-2063.

Wilson, R. J., Banas, N. S., Heath, M. R., & Speirs, D. C. (2016). Projected impacts of 21st century climate change on diapause in *Calanus finmarchicus*. *Global change biology*, *22*(10), 3332-3340.

World Register of Marine Species (2022). Available from https://www.marinespecies.org at VLIZ. Accessed 2022-07-25. doi:10.14284/170.

Wiley, E. O. (1978). The evolutionary species concept reconsidered. *Systematic zoology*, *27*(1), 17-26.

Wright, S. (1940). Breeding structure of populations in relation to speciation. *The American Naturalist*, *74*(752), 232-248.

Yang, Q., Sun, F., Yang, Z., & Li, H. (2014). Comprehensive transcriptome study to develop molecular resources of the copepod Calanus sinicus for their potential ecological applications. *BioMed Research International*, *2014.*

Zhao, Y., Zhang, R., Jiang, K. W., Qi, J., Hu, Y., Guo, J., ... & Ma, H. (2021). Nuclear phylotranscriptomics and phylogenomics support numerous polyploidization events and hypotheses for the evolution of rhizobial nitrogen-fixing symbiosis in Fabaceae. *Molecular Plant*, *14*(5), 748-773.

Paper I

Received: 27 September 2021 | Revised: 14 January 2022 | Accepted: 19 January 2022

DOI: 10.1002/ece3.8606

RESEARCH ARTICLE

Ecology and Evolution Open Access WILEY

Insights into the species evolution of *Calanus* copepods in the northern seas revealed by *de novo* transcriptome sequencing

Apollo Marco Lizano[1] | Irina Smolina[1] | Marvin Choquet[1,2] | Martina Kopp[1] | Galice Hoarau[1]

[1]Faculty of Biosciences and Aquaculture, Nord University, Bodø, Norway

[2]Department of Medical Biochemistry and Microbiology, Uppsala University, Uppsala, Sweden

Correspondence
Apollo Marco Lizano, Faculty of Biosciences and Aquaculture, Nord University, Bodø, Norway.
Email: apollo.m.lizano@nord.no

Funding information
Nord universitet

Abstract

Copepods of the zooplankton genus *Calanus* play a key role in marine ecosystems in the northern seas. Although being among the most studied organisms on Earth, due to their ecological importance, genomic resources for *Calanus* spp. remain scarce, mostly due to their large genome size (from 6 to 12 Gbps). As an alternative to whole-genome sequencing in *Calanus* spp., we sequenced and de novo assembled transcriptomes of five *Calanus* species: *Calanus glacialis*, *C. hyperboreus*, *C. marshallae*, *C. pacificus*, and *C. helgolandicus*. Functional assignment of protein families based on clusters of orthologous genes (COG) and gene ontology (GO) annotations showed analogous patterns of protein functions across species. Phylogenetic analyses using maximum likelihood (ML) of 191 protein-coding genes mined from RNA-seq data fully resolved evolutionary relationships among seven *Calanus* species investigated (five species sequenced for this study and two species with published datasets), with gene and site concordance factors showing that 109 out of 191 protein-coding genes support a separation between three groups: the *C. finmarchicus* group (including *C. finmarchicus*, *C. glacialis*, and *C. marshallae*), the *C. helgolandicus* group (including *C. helgolandicus*, *C. sinicus*, and *C. pacificus*) and the monophyletic *C. hyperboreus* group. The tree topology obtained in ML analyses was similar to a previously proposed phylogeny based on morphological criteria and cleared certain ambiguities from past studies on evolutionary relationships among *Calanus* species.

KEYWORDS
Calanus, concordance factor, de novo transcriptome, phylotranscriptomics, RNA-seq

TAXONOMY CLASSIFICATION
Ecological genetics

1 | INTRODUCTION

Recent developments of next-generation sequencing (NGS) technologies for nucleotide sequencing have revolutionized the field of molecular biology (Metzker, 2010; Schuster, 2008) by allowing the generation of massive amounts of data more rapidly and cost-efficiently than ever before (Luikart et al., 2003). Nonetheless, generating whole-genome data can still be challenging for many zooplankton groups, due to their typically small body size yielding only small amounts of DNA, usually not enough for whole-genome

sequencing, exacerbated by their often large and complex genome architecture characterized by the presence of many repetitive sequences (Bucklin et al., 2018; Tarrant et al., 2019). Copepod species of the marine zooplankton genus *Calanus*, although morphologically very similar (Choquet et al., 2018; Fleminger & Hulsemann, 1977; Frost, 1974), have large genomes that differ greatly in size (from nearly 6 Gbps in *C. finmarchicus* to 12 Gbps in *C. hyperboreus*, McLaren et al., 1988). *Calanus* species play a key role in energy transfer in marine food webs, both as primary consumers and as prey for fish, seabirds, and other marine predators (Arnkværn et al., 2005; Bonnet & Frid, 2004; Cleary et al., 2017). Despite their ecological importance, genomic resources currently available for *Calanus* spp. remain limited, which has led to poor understanding of phylogenetic relationships within the *Calanus* genus so far.

Before development of genetic tools, a phylogeny based on the analysis of morphological characteristics (i.e., relative size of accessory photoreceptor, caudal ramus, anal segment, and genital pore—see Frost, 1974) was proposed for species of the genus *Calanus* (Frost, 1974 and reported in Bucklin et al., 1995). The morphology-based phylogeny identified two distinct groups: the *C. finmarchicus* group, including *C. finmarchicus*, *C. glacialis*, and *C. marshallae*; and the *C. helgolandicus* group, including *C. helgolandicus*, *C. pacificus*, *C. sinicus*, and four other species not investigated in the present study. *Calanus hyperboreus* was considered as a separate clade, distinct from the *C. finmarchicus* and *C. helgolandicus* groups. Later, new phylogenies emerged from analyses of the two genetic markers 16S rRNA (Bucklin et al., 1995) and 28S rRNA (Kozol et al., 2012), but showed a lack of congruence. Although the 28S-based phylogeny proposed by Kozol et al. (2012) seemed to agree with the morphology-based phylogeny on the clustering of species, some branches were not well supported. In contrast, the 16S-based phylogeny suggested a different grouping of species, with several species from the "*C. helgolandicus* group" identified by Frost (1974) not clustering together (i.e., *C. pacificus* separated from *C. sinicus* and *C. helgolandicus*; Bucklin et al., 1995). The discrepancy observed between the two different molecular phylogenies may be explained by the potentially limited resolution of using only a single molecular marker. The use of NGS approaches to obtain larger numbers of molecular markers from genome-wide data can overcome this problem and provide more powerful datasets, needed to accurately characterize species relationships (Leaché & Oaks, 2017).

There has been a growing interest in utilizing RNA-seq approaches to answer evolutionary questions for non-model species because of the ease in assembling and analyzing transcriptome data compared to genomic data and of the possibility to obtain additional information from exonic regions of multiple genes (see e.g., Bi et al., 2012; Tarrant et al., 2019). Transcriptomics studies are on the rise in non-model marine organisms (Eldem et al., 2017; Marlétaz et al., 2019; Pai et al., 2018; Tarrant et al., 2019; Ungaro et al., 2017), but are still too limited to understand zooplankton species ecology and evolution (Lenz et al., 2021). For the genus *Calanus*, various RNA-seq studies have investigated, for example, the classification of genes associated with developmental cycles from embryos to adult (Lenz et al., 2014) and genes contributing to molecular mechanisms during diapause, diapause termination, and starvation (Ohnishi et al., 2019; Skottene et al., 2019). In addition, studies have also looked at patterns of daily gene expression changes at different latitudes and sea-ice coverage (Payton et al., 2020) and the effects of ocean acidification on the regulation of gene expression (Bailey et al., 2017). Yet, there have been no studies utilizing RNA-seq-based data to mine genes to resolve the phylogeny of the genus *Calanus*. Recent studies have validated the use of transcriptomes in phylogenetic analyses, showing virtually identical results with phylogenies derived from whole or partial genome, regardless of the tissue origin and whether the same tissue was used across species (Cheon et al., 2020; Zhao et al., 2021). Multiple studies have already used transcriptomics for phylogenetics of various marine organisms, including dinoflagellates (Annenkova et al., 2018), pteropods (Peijnenburg et al., 2020), bivalves (Li et al., 2020), and crustaceans (Gan et al., 2020).

For the genus *Calanus*, there are currently 17 independent RNA-seq datasets available in the NCBI SRA database (Appendix Table S1, accessed November 2020) representing five *Calanus* species (*C. helgolandicus*, *C. finmarchicus*, *C. sinicus*, *C. pacificus*, and *C. glacialis*), of which only three species have a complete transcriptome assembly: *C. helgolandicus* (Asai et al., 2020), *C. finmarchicus* (Lenz et al., 2014; Tarrant et al., 2014); and *C. sinicus* (Ning et al., 2013; Yang et al., 2014). For *Calanus* species living in the northern seas (covered by the North Atlantic, Arctic, and North Pacific oceans), where they dominate the zooplankton biomass, RNA-seq studies have targeted mostly four species (i.e., *Calanus helgolandicus*, *C. finmarchicus*, *C. sinicus*, and *C. glacialis*), while other species have been mostly ignored (*C. pacificus*, *C. marshallae*, and *C. hyperboreus*).

Our objective is to contribute and improve the currently available transcriptomic resources for *Calanus* spp. and explore the suitability of de novo transcriptome data to infer evolutionary relationships within the genus *Calanus*. To achieve this, we sequenced, assembled, and annotated de novo transcriptomes of two *Calanus* species for the first time (*C. hyperboreus* and *C. marshallae*) and of three species with limited transcriptomic data available (*C. glacialis*, *C. helgolandicus*, and *C. pacificus*). We also investigated hundreds of single-copy orthologs present among the seven species of *Calanus* derived from RNA-seq data and updated the phylogeny of the genus *Calanus*.

2 | MATERIALS AND METHODS

2.1 | Specimen collection and molecular species identification

Three individual copepodites of five species of *Calanus* (*C. helgolandicus*, *C. pacificus*, *C. glacialis*, *C. marshallae*, and *C. hyperboreus*) were sourced from various collaborators (see details in Table 1). These specimens originated from zooplankton samples collected between March 2018 and June 2019 at different sites across the North Atlantic, the North Pacific, and the Arctic Oceans (Figure 1), from different depth ranges (Table 1) by vertically towing a plankton

TABLE 1 Sampling information for *Calanus* species from the North Atlantic, Arctic, and North Pacific Oceans used in this study

Species	Individual ID	Date of collection	Sampling site	Coordinates		Sampling depth (m)	Developmental stage	Collector or study
				Lat	Lon			
Calanus glacialis	Cgla_007	06/2019	Skjerstadfjord	67°14′N	14°44′E	300–500	CV	M. Krogstad
	Cgla_010							
	Cgla_011							
Calanus hyperboreus	Chype_012	09/2018	West Greenland Sea	74°34′N	11°18′W	0–350	Adult female	E. Friis Møller
	Chype_021							
	Chype_030							
Calanus marshallae	Cmar_005	03/2018	Main basin of Puget Sound	47°40′N	122°28′W	0–140	CV	A. Bucklin & B. Frost
	Cmar_007							
	Cmar_008							
Calanus pacificus	Cpac_006	03/2018	Main basin of Puget Sound	47°40′N	122°28′W	0–140	CV	A. Bucklin & B. Frost
	Cpac_007							
	Cpac_008							
Calanus helgolandicus	Chelg_003	04/2019	Stonehaven - north-east Scotland	56°57′N	02°07′W	0–48	CV	L. Noble
	Chelg_007							
	Chelg_008							
Calanus finmarchicus	Cfin_SRR1153468	07/2011	Mount Desert Rock, Gulf of Maine	44°2′N	68°3′W	Not specified	CV	Lenz et al. (2014)
	Cfin_SRR1141107	05/2012	NTNU/SINTEF Sealab facility Trondheim, Norway	Not specified	Not specified	70		Tarrant et al. (2014)
	Cfin_SRR1141110			Not specified	Not specified			
Calanus sinicus	Csin_DRR144876	10/2015	Off the coast of Japan along the Kuroshio Current	34°00′N	138°00′E	0–100	Adult female	Ohnishi et al. (2019)
	Csin_DRR144878							
	Csin_SRP032493	05/2013	Yellow Sea	38°45′N	121°45′E	Not specified	Adult copepod unspecified sex	Yang et al. (2014)
Acartia tonsa	Atonsa_Nilsson	09/2016	Øresund Denmark	56°N	12°E	Culture	Adult	Nilsson et al. (2018)
Eurytemora affinis	Eurytemora_affinis	NA	Bred at WHOI for 1 year	NA	NA	Culture	Adult female	Almada & Tarrant (2016)

Note: Three individuals were used for each species. For *C. finmarchicus* and *C. sinicus*, previously published data were used, individual ID contains the reference number for sequences downloaded from the NCBI SRA database.

net (WP3, Juday or multinet). For each sample, *Calanus* spp. individuals were pre-sorted from the rest of the zooplankton and subsequently preserved at −20°C in RNAlater. Individuals of *C. pacificus* and *C. marshallae* were morphologically identified as such by the taxonomist B. Frost. Molecular species identification was performed to confirm the species identity of *C. helgolandicus*, *C. glacialis*, and *C. hyperboreus* using six InDel (Insertion-Deletion) molecular markers (Smolina et al., 2014), with DNA extracted from antennules separately, following the optimized protocol from Choquet et al. (2017).

2.2 | RNA extraction, library preparation, and sequencing

Total RNA was extracted from the 15 pre-identified (using antennules DNA or morphology) *Calanus* individuals using the Tri(Qia)zol from Qiagen RNeasy mini kit with minor modifications to the manufacturer's protocol and was concentrated using the RNA Clean & Concentrator™ kit (Zymo Research). RNA quality was assessed using an Agilent Bioanalyzer 2100 (Agilent Technologies) and showed integrity numbers (RIN values) between 9.6 and 10.0, indicating high quality of extracted RNA. Aside from RIN values, we did not observe evidence of DNA contamination and limited smearing in small size range.

Individual RNA libraries were prepared using the NEBNext® Ultra™ II Directional RNA Library Prep Kit for Illumina® (New England Biolabs) following the manufacturer's protocol. Libraries were quantified using an Agilent Bioanalyzer 2100 (Agilent Technologies) and pooled in equimolar concentrations before sequencing on an Illumina NextSeq 500 platform with a 2 × 150 bp high-output NextSeq 500/550 v. 2.5. kit.

2.3 | De novo transcriptome assembly and transcript filtering

Raw sequencing reads, generated from the 15 individuals, were demultiplexed using bcl2fastq v. 2.2 (https://support.illumina.com/sequencing/sequencing_software/bcl2fastq-conversion-software.html) and quality checked using FastQC v. 0.11.5 (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Low-quality reads (see Appendix S1) and adapter sequences were trimmed using cutadapt v. 1.18 (Martin, 2011).

RNA-seq reads from transcriptomes of two additional *Calanus* species (*C. finmarchicus* and *C. sinicus*) with three individuals per species, and two closely related taxa (*Acartia tonsa* and *Eurytemora affinis*) with one individual per species, were downloaded from the NCBI SRA database and included in the subsequent analyses (see Table 1 for more information). We chose to sequence new individuals for both *C. glacialis* and *C. helgolandicus* as most of the transcriptome assemblies available online are either incomplete or limited

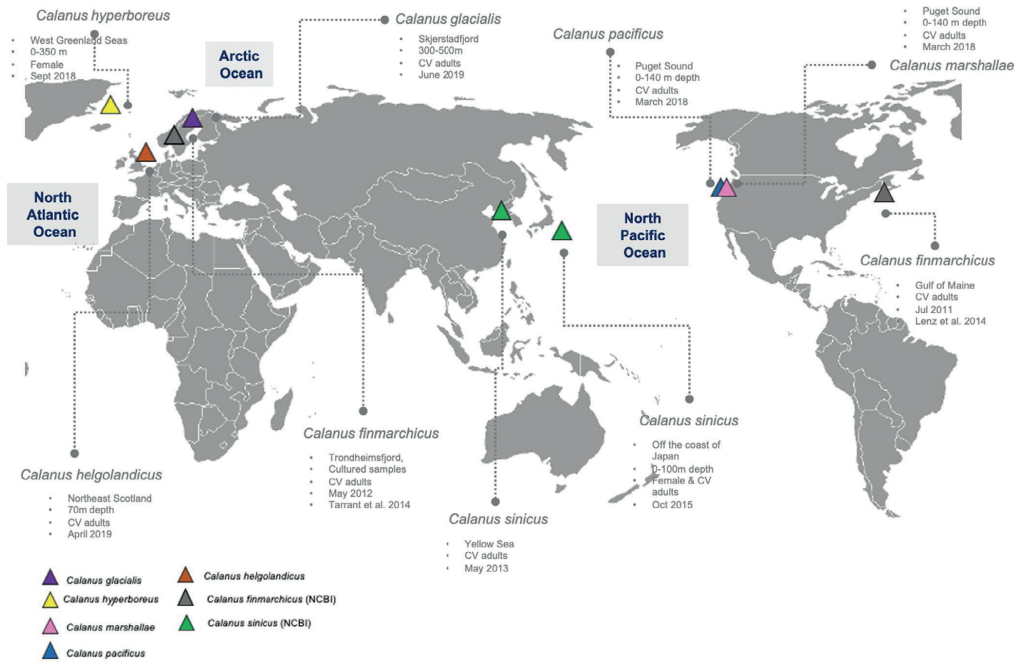


FIGURE 1 Sampling locations for the seven species of *Calanus* analyzed in this study

to certain species-specific stages. In order to compare across species, we aimed for only CV or adult females. We selected *Acartia tonsa* and *E. affinis* as outgroups since they are among the closest extant species to *Calanus* spp. within the order Calanoida to have complete transcriptomes available online (Tarrant et al., 2019), and importantly because both species have been used as outgroups in a previous study where divergence time was estimated (Eyun, 2017).

Individual de novo transcriptome assemblies were performed for all individuals of seven species using Trinity v. 2.9.1. (Grabherr et al., 2011). Assembly statistics were computed using the Perl script TrinityStats.pl contained in the Trinity software package. Based on the lengths of assembled transcriptome contigs, we computed for N50 based on the single longest isoform per gene (Nx50) and ExN50 statistics, which are limited to the topmost highly expressed genes. N50 values can often be exaggerated due to Trinity program generating too many transcript isoforms. To attenuate the probability of including false and redundant transcripts, contigs were filtered in five steps: (1) cross-species contamination of contigs in assembled transcriptome was removed using CroCo v. 1.1 (Simion et al., 2018); (2) weakly expressed transcript isoforms or isoforms which are not expressed as much as other isoforms were removed and only the most highly expressed isoform per gene were retained using Trinity perl scripts *align_and_estimate_abundance.pl* and *filter_low_expr_transcripts.pl* with the "--highest_iso_only" parameter included; (3) redundant transcripts with ≥95% identity were filtered out using cd-hit-est v. 4.7 (Fu et al., 2012; Li & Godzik, 2006); (4) misassembled or incomplete contigs were filtered out based on read mapping metrics using TransRate v. 1.0.3 (Smith-Unna et al., 2016); and lastly, (5) only transcripts containing open-reading frames (ORFs) with a length of at least 100 amino acids were retained using Transdecoder v. 5.5 (Haas et al., 2013, Figure S1). Then, transcriptome assembly completeness was assessed using BUSCO v. 4.0.2 (*Benchmarking Universal Single-Copy Orthologs*, Seppey et al., 2019) to obtain the overview of all single-copy, duplicated, and missing orthologs represented in the arthropod dataset (arthropoda_odb10). Downstream analyses were performed using the resulting 23 filtered transcriptomes.

2.4 | Identification of coding regions and functional annotation

Candidate coding regions within the transcriptome assemblies were identified using TransDecoder v. 5.5 (Haas et al., 2013). Functional annotation was performed using eggNOG-mapper v. 2.0 (Huerta-cepas et al., 2017) based on fast orthology assignment using precomputed eggNOG v. 5.0 (Huerta-Cepas et al., 2019) clusters and phylogenies. Protein families were assigned to known functional class using gene ontology (GO) terms and the database of clusters of orthologous genes (COG, Galperin et al., 2021). Two COG groups "Cell motility, N" and "Nuclear Structure, Y" showed very low protein counts across the seven *Calanus* species, and the reason behind it was investigated in detail using eggNOG database for Arthropoda (6656 single-copy orthologs, downloaded on 25.04.2021). Proteins in the Arthropoda database for N and Y categories that were not identified by eggNOG-mapper in *C. hyperboreus* (species with the lowest number of obtained matches to N and Y category) were manually searched using HHMER v3.1b2 (Johnson et al., 2010). The created hidden Markov models were used to mine predicted ORFs from *C. hyperboreus*. Best significant hits (with the lowest *E*-value) were searched against NCBI nonredundant protein database using BLASTp to further confirm their identity and relatedness to COG functional categories. In addition, annotated transcripts were also classified to three GOSlim functional categories (biological process, cellular component, and molecular function) using the webserver "PANTHER Classification System" (Mi et al., 2019, 2021) with *Drosophila melanogaster* chosen as the reference organism.

2.5 | Ortholog identification and phylogenetic analyses

Orthofinder v. 2.3.1 (Emms & Kelly, 2019) was used to prepare a dataset containing only single-copy orthologs for phylogenetic inference using all 23 transcriptomes, while for estimation of gene duplication events along the phylogeny only representative individuals per species were used. A phylogenetic tree was constructed by aligning protein-coding sequences of single-copy orthologs using clustalomega v. 1.2.4 (Sievers & Higgins, 2014). A customized Python script (*convert.py*, https://github.com/mmatschiner/tutorials/tree/master/ml_species_tree_inference) was used to remove sequences containing missing information from all the alignments and to translate each sequence alignment into a Nexus format, needed for IQ-tree. Maximum likelihood (ML) phylogenetic trees were generated for each single-copy orthogroup using both the bootstrap method (with 1000 replicates) and the maximum-likelihood method with branch lengths calculation using IQ-tree ver. 1.6.1 (Nguyen et al., 2015). Substitution models were not specified, allowing IQ-tree to choose the best-fitting model for each orthogroup. A single species tree with maximum number of quartets shared among gene trees based on ML was inferred among individual species trees using ASTRAL v. 5.7.3 (Zhang et al., 2018). Gene and site concordance factors (sCF) were also computed to determine which branches show concordant and disconcordant genes and to calculate site variances to the reference ML tree.

3 | RESULTS

3.1 | Sequencing

We sequenced 15 individual transcriptomes, with three individuals for each of five *Calanus* species (*C. helgolandicus*, *C. pacificus*, *C. glacialis*, *C. marshallae*, and *C. hyperboreus*). The mean sequencing output per individual was 47.6 million reads (ranging from 10.8 to

102 million, Table S2). All raw reads were uploaded to NCBI SRA database under BioProject PRJNA744376 and all de novo transcriptome assemblies generated in this project has been deposited at DDBJ/EMBL/GenBank TSA database and DRYAD server (https://doi.org/10.5061/dryad.n8pk0p2ww, see Table S4).

3.2 | *De novo* transcriptome assembly and quality assessment

De novo transcriptome assembly using Trinity was performed for each *Calanus* individual sequenced in addition to six datasets publicly available, generating a total of 21 transcriptome assemblies. Quality metrics for each assembly are presented in Table 2. Mapping of quality trimmed reads to their corresponding de novo transcriptome assemblies generated alignment rates from 95.62% to 99.13% with a mean alignment rate of 96.69%. Based on 1013 conserved arthropod orthologs, our BUSCO analysis identified 93.57% (mean among three individuals) complete single-copy and complete duplicated BUSCO's in *C. glacialis*, (91.33%) in *C. hyperboreus*, (88.43%) in *C. marshallae*, (91.80%) in *C. pacificus*, and (92.10%) in *C. helgolandicus*. These parameters indicated that the 15 de novo transcriptomes were well assembled and relatively complete (Table 2; Figure S2). Moreover, we investigated orthologs identified by BUSCO as missing and found eight orthologs that are common among all the 21 *Calanus* transcriptomes (Table S3). However, further manual BLAST of these proteins against the translated transcriptome of *C. hyperboreus* revealed query hits >50% for six proteins, and the absence of significant hits for the two other proteins: 3-ketodihydrosphingosine reductase (*Plutella xylostella*, XP_037969682.1) and serine palmitoyltransferase 1 (*Plutella xylostella*, XP_037971886.1), both reportedly involved in sphingolipid metabolism.

We tested for cross-species contamination among the 15 new transcriptomes that were sequenced on the same flow-cell, and none was detected. After filtering out weakly expressed isoforms, a total of 332,489 transcripts (representing 66.44% of all the generated transcripts) were retained for all individuals of *C. glacialis*, 58.56% for *C. hyperboreus*, 66.99% for *C. marshallae*, 58.37% for *C. pacificus*, 58.94% for *C. helgolandicus*, 44.73% for *C. finmarchicus*, and 43.75% for *C. sinicus*. The final number of peptides with ORF meeting the minimum criteria set by Transdecoder-v.5.5 (Haas et al., 2013) ranged from 33,135 peptide sequences for *C. finmarchicus_SRR1141110* to 72,916 for *C. pacificus_008*. The mean number of remaining peptides among the seven species was 53,265 (Table 2).

3.3 | Functional classification of protein families

Functional annotation of orthologous protein families was based on GO terms and COG databases implemented in eggNOG-mapper v.2.0 (Huerta-Cepas et al., 2017). Functional classification of protein-coding sequences based on COG yielded different numbers of protein queries from one species to another. For *A. tonsa*, we were able to functionally assign 5947 protein queries, 10,925 *for C. marshallae*, 12,435 for *C. hyperboreus*, 16,057 for *C. finmarchicus*, 16,308 for *C. glacialis*, 16,971 for *C. pacificus*, 17,349 for *C. helgolandicus*, and 17,706 for *C. sinicus*. The orthologous protein families were subdivided into 25 COG classifications (Figure 2a). Among them, the category "Unknown Function, S" represented the largest group with a cumulative query hit comprising of 24.62% of the total protein assignments for all the *Calanus* species. It was followed by "Signal transduction mechanisms, T" (12.48%), "Post-translational modification, protein turn-over, and chaperon, O" (10.27%). Groups with the lowest protein count were linked to functions related to "Cell motility, N" (0.05%), "Nuclear Structure, Y" (0.04%), and "General function prediction only, R" (0,0%). In-depth look into N and Y categories using Arthropoda eggNOG database showed that this database only contains 15 and 10 proteins in N and Y categories respectively, thus explaining general low numbers of these categories in *Calanus* transcriptomes (from 5 to 10 proteins per species in N category and from 5 to 8 proteins per species in Y category). Furthermore, manual search with HHMER reduced the number of unfound proteins from the Arthropoda database in *C. hyperboreus* from 10 to 5 in N category, and from 5 to 1 in Y category.

In total, the number of transcripts with GO and KEGG annotations were, respectively, 4161 and 4447 for *A. tonsa*; 10,869 and 12,106 for *C. finmarchicus*; 10,837 and 12,197 for *C. glacialis*; 7701 and 8305 for *C. marshallae*; 11,027 and 11,908 for *C. helgolandicus*; 11,231 and 12,313 for *C. pacificus*; 12,154 and 13,271 for *C. sinicus*; and 8575 and 9220 for *C. hyperboreus*. Overall, the percentage of annotated transcripts to major GO categories was similar, if not comparable among all examined species (Figure 2b). Within the GO category Biological Process (BP), the largest proportion of transcripts were assigned to cellular process (~33%), metabolic process (~24%), and biological regulation (~15%). Within the GO category Molecular Function (MF), binding and catalytic activities were the largest terms with ~40% and ~34%, respectively, of all transcripts with GO term hits. For Cellular Component (CC), within GO_Slim category in Panther database, transcripts were assigned to only three groups with cellular anatomical entity (~42%) and intracellular (~37%) as the largest groups (Figure S3).

3.4 | Ortholog identification and phylogenetic analyses

On average, Orthofinder assigned 32,446 (91%) genes to orthogroups in each *Calanus* species, while 11,610 (84%) genes were assigned to orthogroups in the two outgroup species, meaning that taxon sampling was sufficient. The number of genes in species-specific orthogroups ranged from 123 in *C. marshallae* to 898 in *C. helgolandicus* with a mean of 498 genes (1.3%). Estimated

TABLE 2 Transcriptome assembly statistics for *Calanus* spp. and two outgroup species *Acartia tonsa* and *Eurytemora affinis* downloaded from NCBI SRA database showing the total no. of assembled bases, total no. of genes, total no. of transcripts, %GC content, % alignment, no. of retained transcripts, no. of peptides (ORF ≥ 100 aa), and BUSCO results

Species	Individual ID	Total no of assembled bases	Total no genes	Total no of transcripts	%GC	% alignment	No of retained transcripts	No of peptides (ORF ≥ 100 aa)	BUSCO
Calanus glacialis	Cgla_007	80,786,787	107,689	191,809	42.61	98.10	58,057	55,050	C: 93.7% [S: 53.0%, D: 40.7%], F: 1.2%, M: 5.1%
	Cgla_010	52,916,092	107,265	191,130	42.63	95.64	82,924	69,473	C: 93.8% [S: 51.8%, D: 42.0%], F: 0.8%, M: 5.4%
	Cgla_011	83,871,692	117,560	208,862	42.63	95.62	41,072	72,199	C: 93.2% [S: 50.1%, D: 43.1%], F: 1.4%, M: 5.4%
Calanus hyperboreus	Chype_012	64,862,613	89,686	154,261	43.86	99.13	57,073	43,074	C: 92.0% [S: 48.7%, D: 43.3%], F: 1.8%, M: 6.2%
	Chype_021	44,640,167	57,792	98,478	44.80	96.87	55,106	67,106	C: 89.6% [S: 53.0%, D: 36.6%], F: 2.5%, M: 7.9%
	Chype_030	69,174,829	97,334	165,321	43.13	96.05	79,215	64,262	C: 92.4% [S: 51.3%, D: 41.1%], F: 1.6%, M: 6.0%
Calanus marshallae	Cmar_005	43,950,718	58,868	86,851	45.35	96.61	30,505	34,572	C: 88.8% [S: 59.5%, D: 29.3%], F: 1.8%, M: 9.4%
	Cmar_007	53,405,794	70,397	108,982	44.96	97.07	42,388	46,829	C: 89.8% [S: 56.4%, D: 33.4%], F: 2.2%, M: 8.0%
	Cmar_008	30,405,070	39,378	55,916	45,89	95.86	48,111	53,770	C: 86.7% [S: 58.7%, D: 28.0%], F: 2.1%, M: 11.2%
Calanus pacificus	Cpac_006	57,260,966	79,107	133,448	45.60	96.53	57,755	60,389	C: 90.7% [S: 38.2%, D: 52.5%], F: 2.0%, M: 7.3%
	Cpac_007	65,331,309	85,103	153,092	45.30	96.68	48,704	56,214	C: 93.5% [S: 40.4%, D: 53.1%], F: 1.8%, M: 4.7%
	Cpac_008	62,163,761	87,403	144,545	45.31	96.34	76,121	72,916	C: 91.2% [S: 38.4%, D: 52.8%], F: 2.5%, M: 6.3%
Calanus helgolandicus	Chelg_003	82,334,519	110,120	199,181	44.70	97.17	62,642	67,106	C: 93.6% [S: 39.8%, D: 53.8%], F: 1.2%, M: 5.2%
	Chelg_007	61,489,417	83,960	137,554	45.23	96.31	54,695	60,542	C: 90.6% [S: 41.6%, D: 49.0%], F: 2.2%, M: 7.2%
	Chelg_008	45,658,464	61,171	96,333	45.86	96.34	44,779	49,783	C: 89.3% [S: 45.3%, D: 44.0%], F: 2.9%, M: 7.8%
Calanus finmarchicus	Cfin_SRR1141107	29,971,015	43,252	75,504	46,28	88.89	53,751	51,970	C: 81.3% [S: 42.2%, D: 39.1%], F: 8.1%, M: 10.6%
	Cfin_SRR1141110	39,691,175	53,703	113,701	45,39	88.73	25,298	33,135	C: 86.7% [S: 27.8%, D: 58.9%], F: 4.7%, M: 8.6%
	Cfin_SRR1153468	62,399,753	90,151	229,051	44,22	96.20	34,899	43,427	C: 90.2% [S: 54.5%, D: 35.7%], F: 4.3%, M: 5.5%
Calanus sinicus	Csin_SRP032493	61,756,777	102,986	235,405	46,44	94.98	67,651	61,031	C: 92.1% [S: 20.2%, D: 71.9%], F: 2.1%, M: 5.8%
	Csin_DRR144876	63,116,211	113,366	282,710	46,23	94.98	77,949	63,851	C: 91.8% [S: 11.5%, D: 80.3%], F: 2.5%, M: 5.7%
	Csin_DRR144878	58,403,986	106,792	263,104	46,16	94.67	69,934	58,166	C: 90.4% [S: 10.4%, D: 80.0%], F: 2.6%, M: 7.0%
Acartia tonsa	Atonsa_Nilsson	118,203,047	48,149	114,717	37,9	98.79	31,986	16,174	C: 56.8% [S: 56.2%, D: 0.6%], F: 2.8%, M: 40.4%
Eureytemora affinis	Eaffinis_Almada	181,412,865	90,855	170,681	38,61	95.23	57,397	24,056	C: 71.1% [S: 69.4%, D: 1.7%], F: 2.7%, M: 26.2%

Note: BUSCO assessment was based on arthropoda database (*odb_10* containing 1103 orthologs). C = complete, S = single, D = duplicated, F = fragmented, and M = missing no. of orthologs.

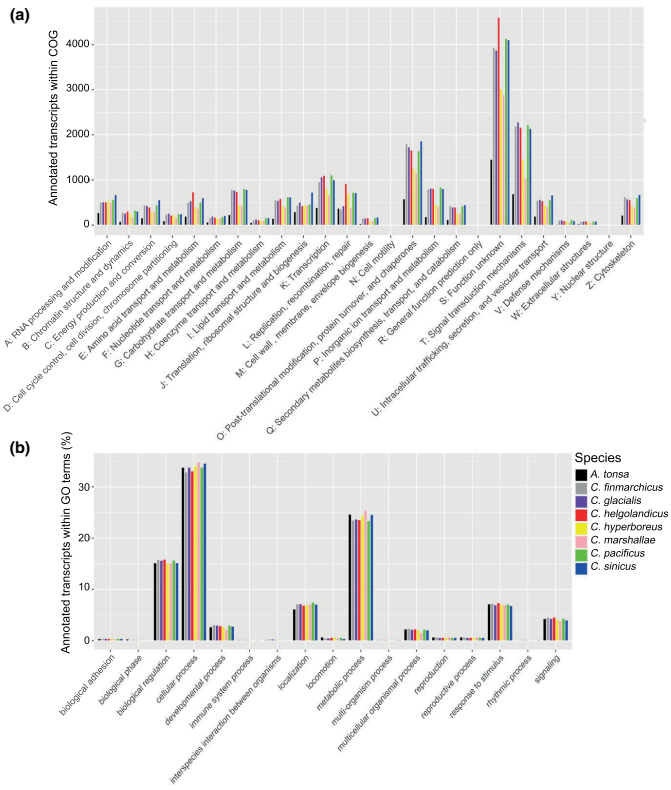


FIGURE 2 (a) Functional annotation of protein-coding sequences based on Clusters of Orthologous Groups (COG) database. (b) Gene Ontology (GO) annotation representing biological process for seven species of *Calanus* and one outgroup taxon *Acartia tonsa*

duplication events inferred by Orthofinder showed that the largest number of duplication events happened prior to the emergence of the genus *Calanus* and after the split of *C. hyperboreus* (Figure 3). The *C. helgolandicus* group had the most duplication events with the most duplication (4970) inferred for *C. helgolandicus* itself. The lowest number of duplications was observed in *C. marshallae* (988). We also observed that species with the highest number of duplication events are also the species with the highest number of peptides and sequencing output (Tables 2 and S2). Notably, a strong positive correlation is detected between the number of final peptides in each species and the number of estimated duplications (Pearson's $r = .82$, p value = .023).

Orthofinder identified 191 single-copy protein coding orthologous genes across seven species of *Calanus* and the two outgroup taxa *A. tonsa* and *E. affinis*. The generated ML trees for each gene ortholog showed that the seven species were clustered into three well-supported clades based on bootstrap support values (*C. sinicus* (*C. helgolandicus* + *C. pacificus*); (*C. finmarchicus* (*C. glacialis* + *C. marshallae*); (*C. hyperboreus*). Moreover, *C. sinicus* shared a recent common ancestor with the monophyletic species *C. helgolandicus* and *C. pacificus*. *C. finmarchicus* shared a recent common ancestor with *C. glacialis* and *C. marshallae*, and there was a consistent split between *C. hyperboreus* and the two other monophyletic groups (*C. helgolandicus* and *C. finmarchicus* groups, Figure 4a).

Gene and site concordance analyses revealed branches that show a concordant and disconcordant gene and site variations within our reference ML tree (Figure 4b). The split between *C. hyperboreus* group and the two other groups (*C. finmarchicus* and *C. helgolandicus* group) showed a 100% maximum-likelihood value with a gCF of 100% and sCF of 96.3%. This means that all 191 protein-coding genes support this grouping, with most of the amino acid sites informative for this ML branch. Meanwhile, a gCF value of 57.3% or 109 out of 191 single-copy orthologous genes support the split between *C. finmarchicus* and *C. helgolandicus* groups with most of the sites (67.8%) informative of the branch topology. 36.7% or ~70 single-copy orthologs support the separation of *C. sinicus* from *C. helgolandicus* and *C. pacificus*. A third of all single-copy protein-coding genes (30.9% gCF) indicate that *C. helgolandicus* and *C. pacificus* are sister species and about 54.5% of the genes supported the split of *C. finmarchicus* with *C. glacialis* and *C. marshallae* with 56.2% of these protein-coding genes containing informative sites. Furthermore, the scatter plot of gCF and sCF values for all branches shows that the majority of single-copy orthologs used to reconstruct the species tree contained informative gene and site information.

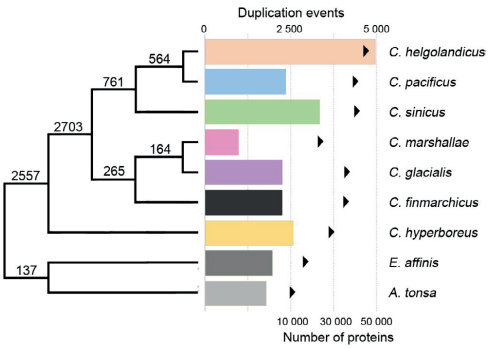


FIGURE 3 Inferred number of gene duplication events along *Calanus* species tree. Numbers on each branch are duplication events of each respective branch that are retained in all descendant species. Bar plots represent the number of gene duplication events for each species. Black arrows indicate number of proteins per species used for the inference

4 | DISCUSSION

4.1 | 15 new *de novo* transcriptome assemblies for *Calanus* spp.

With the aim of contributing to and improving the existing transcriptomic resources available for the genus *Calanus*, we analyzed the transcriptomes of five species of *Calanus* together with two outgroups and two *Calanus* species mined from NCBI database. Overall, we were able to contribute to the existing online high-throughput sequencing database with 15 new transcriptomes, including two species never sequenced before (*C. hyperboreus* and *C. marshallae*). The quality of our transcriptome assemblies was similar, if not slightly higher, than previously assembled de novo transcriptomes for *Calanus* spp. according to several metrics. For instance, the mean percentage of alignment among the individuals generated from the present study was 96.69%, while the mean percentage alignment among the six de novo transcriptome assemblies that we downloaded from NCBI database is 93.08%. Our results satisfy the currently accepted criteria that a good Trinity transcriptome assembly should have a high percentage alignment or that most of the reads should map back to the assembly. This is also true with the Nx50 value where the mean Nx50 from our study is 1118.8, while the mean Nx50 among database samples is 1007.83. Furthermore, we also computed for the ExN50 (i.e., N50 based only on the topmost highly expressed genes), considered to be one of the most appropriate metrics to assess transcriptome data quality (Haas et al., 2013). Our ExN50 ranged between 2257 and 2854 with a mean value of 2438 (Figure S4). Lastly, to complement the technical metrics of N50 statistics, we used BUSCO to indirectly assess our assembly completeness using a specific set of near-universal single-copy orthologs based on the BUSCO Arthropoda database. Our results show a relatively complete assembly among the 15 individuals with a mean BUSCO completeness value of 90.55%. Both the summary statistics and BUSCO measurements showed similar or higher values compared to previously published studies related to de novo transcriptome assemblies in *Calanus* spp. and other copepods (Berger et al., 2021; Lenz et al., 2014; Tarrant et al., 2014, 2019; Yang et al., 2014). Although the transcriptomes presented here are of high quality, they remain partially incomplete because they only represent one or two specific developmental stages of a species in a snapshot of natural conditions and are likely lacking stage-specific, condition- or stress-specific transcripts. More sequencing efforts are needed to further improve transcriptomic resources for the genus *Calanus*.

The recent Ocean ZOOP initiative (Ocean Zooplankton Open 'Omics Project) has called for multispecies high-quality de novo transcriptomes for zooplankton species spanning diverse taxa from across the world's oceans, to generate a new framework for evolutionary, ecological, and physiological studies (Lenz et al., 2021). Our results on multiple *Calanus* species will thus contribute to such most needed project. In addition, here we demonstrated how the use of whole transcriptome data can help resolve evolutionary relationships among closely related zooplankton species.

4.2 | Conserved protein functions across species

Functional assignment based on COG and GO term annotations showed a conserved pattern of protein functions across different *Calanus* species. In particular, we noticed that relatively few protein families encode for functions relating to nuclear structure compared to other eukaryotic species (i.e., *Arabidopsis thaliana*, *Caenorhabditis elegans*, *Drosophila melanogaster*, *Homo sapiens*, and *Saccharomyces cerivisae* Tatusov et al., 2003). This COG pattern is also distinct from other distantly related marine organisms such as the fish *Coilia nasus* (Du et al., 2014); the diatom *Skeletonema costatum* (Zhang et al., 2016); and the crab *Eriochier sinensis* (Li et al., 2013). The results of our COG annotation indicate a lack of protein families in functional categories N (Cell motility) and Y (Nuclear structure). However, this may be linked to their small numbers in the database compared to protein families in other COG categories, and slightly lower efficiency of automated annotation compared to manual, but not necessarily to the absence of these proteins in *Calanus* transcriptomes. We also found that 24.3% of the protein families among seven *Calanus* species do not have a known function and that no proteins were functionally assigned to the general function predictions (R) category. Our results are almost similar with the COG annotation performed by Yang et al., 2014 for *C. sinicus* except that we were able to assign more protein families with functions related to extracellular structures (393–677 protein queries vs. <100, for *C. sinicus*, Yang et al., 2014). In general, the total number of protein families assigned to known functions (based on COG annotation) was much lower (~6383 out of 43,417 protein queries assigned) in the study of Yang et al. (2014), due to the limited number of reference genomes present in the COG database back in 2014. GO annotation also showed that sequenced and assembled transcriptomes of *Calanus*

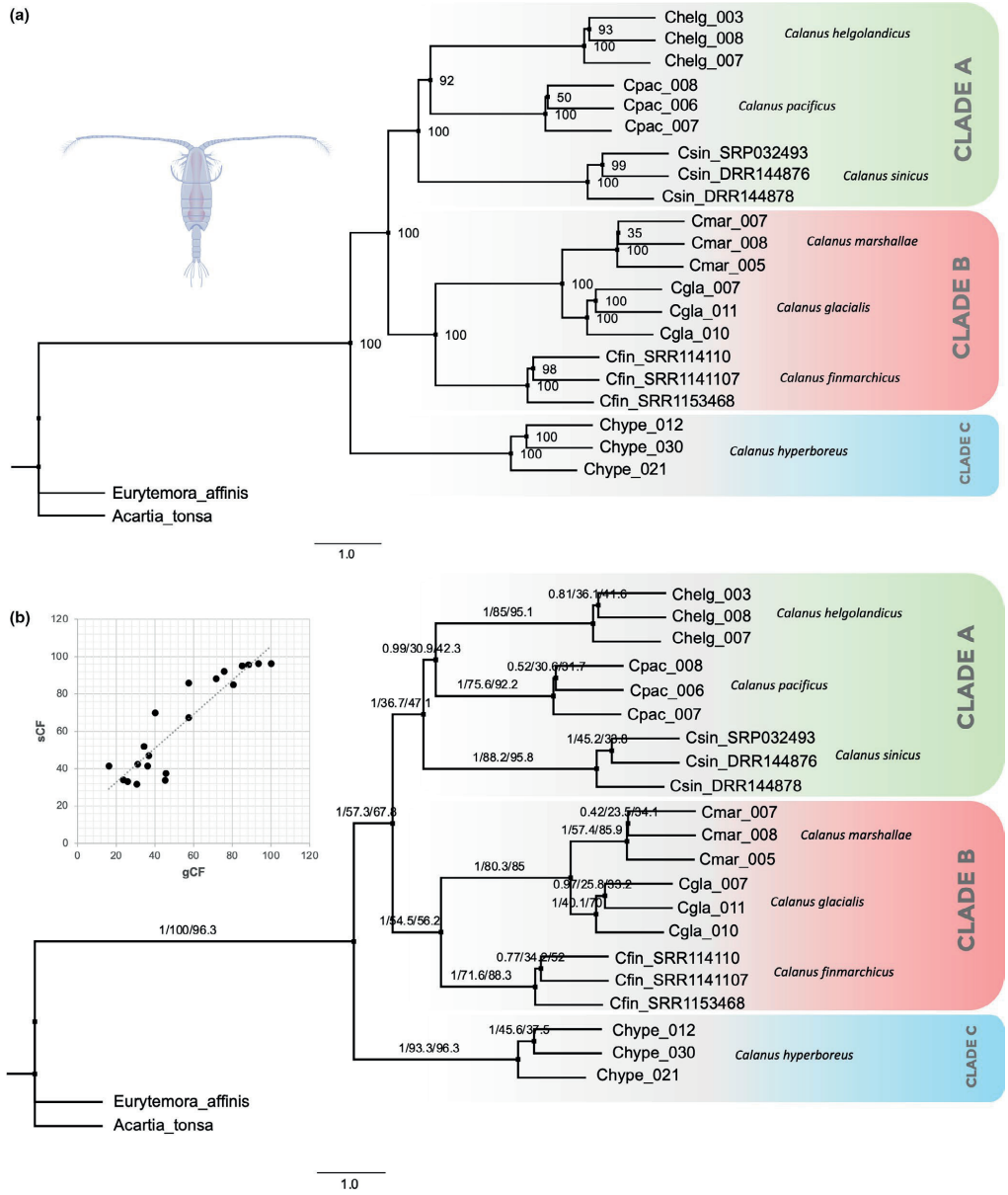


FIGURE 4 (a) Maximum-likelihood (ML) phylogenetic tree of seven *Calanus* species and two outgroup taxa *Acartia tonsa* and *Eurytemora affinis* based on 191 single-copy orthologs derived from transcriptomes. ML tree showing bootstrap support values that are at the maximum on the majority of nodes (ML bootstrap = 100) except *C. helgolandicus* & *C. pacificus* (bootstrap = 92%). (b) Corresponding ML tree for the *Calanus* spp. dataset including two outgroup taxa. Numbers on each branch represent maximum-likelihood support value, gCF, and sCF. The inset shows the scatterplot of gCF and sCF values

species have patterns of annotated transcripts similar to previously sequenced and here re-assembled *C. finmarchicus* (Tarrant et al., 2014), *C. sinicus* (Yang et al., 2014), and other recently investigated copepods, for example, *Labidocera madurae* (Roncalli et al., 2017) and *Rhincalanus gigas* (Lauritano et al., 2020).

4.3 | Orthologs and duplication events

According to Orthofinder, most protein-coding genes are conserved among the seven *Calanus* species with only 1.3% of the genes being species-specific. This percentage of species-specific genes is a bit higher than what was reported for five species of bats in the recent study of (Moreno-Santillán et al., 2019) and for five species of lizards in (Maldonado et al., 2020). The number of duplication events are expected to be high for *Calanus* copepods due to their large genomes (McLaren et al., 1988) and extended gene families (e.g., Lenz et al., 2014). Surprisingly, the largest number of duplication events was inferred in *C. helgolandicus* and other species of the *C. helgolandicus* group (*C. pacificus* and *C. sinicus*). It was only intermediate in *C. hyperboreus* and *C. glacialis*, despite them having the largest genome size estimations (McLaren et al., 1988). The lowest number of duplication events was found in *C. marshallae*, which has a genome size estimate comparable to that of *C. helgolandicus*. The strong positive correlation between duplication events and the number of predicted peptides used as input suggests that these results could be biased by sequencing effort and completeness of species transcriptomes. However, the positive correlation between the number of peptides and duplication events does not shed light on this question and only sequenced genomes of *Calanus* species will unequivocally resolve this issue.

4.4 | Fully resolved phylogenetic relationships for seven *Calanus* species

In the present study, we were able to reconstruct the phylogenetic relationships of seven species of *Calanus* based on 191 single-copy protein-coding orthologs from RNA-seq data. The results of our maximum-likelihood tree indicated that these seven species cluster into three well-supported clades. Our phylogenetic analysis also showed a fully resolved *C. helgolandicus* group (*C. sinicus* as a sister clade to the sister species *C. helgolandicus* and *C. pacificus*, clade A) in comparison with the earlier study of Bucklin et al. (1995) using 16S rRNA, wherein several species of the *C. helgolandicus* group (*C. helgolandicus*, *C. pacificus*, and *C. sinicus*) did not group together well, largely because of the ambiguous position of *C. pacificus*. The relationships within the *C. helgolandicus* group relied on a single molecular marker and were only loosely supported by the bootstraps (Bucklin et al., 1995). Moreover, our phylogenetic results slightly disagree with the phylogenetic tree from Kozol et al. (2012) based on 658 bp region of the 28S rRNA gene, where the authors found that *C. helgolandicus* is a sister species to *C. pacificus* and *C. sinicus*, while our ML tree suggests that *C. helgolandicus* and *C. pacificus* are more related than *C. pacificus* and *C. sinicus* are. Our tree topology was supported by high bootstrap values (bootstrap = 100, *C. sinicus* (*C. pacificus* + *C. helgolandicus*); and bootstrap = 92, *C. pacificus* + *C. helgolandicus*), with more individuals per species and higher number of genes (~60 to 70 single-copy orthologs supporting these branches) compared to earlier studies. Meanwhile, species relationships in both Clade B and Clade C are consistent with all the previously proposed *Calanus* phylogenies (Bucklin et al., 1995; Frost, 1974; Kozol et al., 2012).

The topology of our ML tree was more concordant with the topology found by Frost (1974) and by Fleminger (reported in Bucklin et al., 1995) based on the examination of several morphological characters. Our study demonstrates the power that can be obtained with large molecular datasets to fully resolve phylogenies, in comparison with using only a single genetic marker. The concordance between morphology-based and well supported/confirmed molecular phylogenies is particularly interesting in the case of *Calanus* spp., as the strong morphological similarity between species within the genus has challenged the work of taxonomists for decades, especially for species within Clade B. Recently, a thorough assessment of morphological characters considered to be diagnostic for species discrimination between *C. finmarchicus* and *C. glacialis* was made using genetic tools and revealed that most if not all morphological characters were unreliable, depending on geographical location (Choquet et al., 2018). To reconstruct the *Calanus* phylogeny, Frost (1974) examined several morphological characters including: the fifth pair of swimming legs, the relative size of accessory photoreceptor, the length of caudal ramus, the size of genital pore, and the shape of the ventral surface of the genital segment. Some of these taxonomic characters have been re-investigated recently together with some genetic information (mostly between *C. glacialis* and *C. finmarchicus* species). No evident species-specific patterns were observed, and results are shown to be species independent (e.g., 5th pair of swimming legs Choquet et al., 2018; secondary sexual structures—K. Kosobokova, *personal communication*) or significantly variable depending on geography. However, the combined analysis of multiple morphological traits, as performed in the study of Frost (1974) from individuals sampled in regions where species-specific morphological differences may be more distinct (*see* Choquet et al., 2018) could have contributed to a more similar phylogenetic tree between the phylogeny based on morphology and the phylogeny based on new transcriptome datasets (this study).

The taxonomic status of *C. marshallae* has been in question due to the extreme similarity of mtCOI and mt16S sequences with that of *C. glacialis* reported in GenBank. Minimal differences in barcode sequences have raised doubts among experts about the current taxonomic status of *C. marshallae* species. Ashjian et al. (2017) reported a single base pair difference between species in the COI mitochondrial gene across 1500 specimens of *C. glacialis*/*C. marshallae* analyzed in their study. Their genetic analyses revealed three genetically differentiated groups (Ashjian et al., 2017): *C. marshallae* from Puget Sound (identified by B. Frost), the Arctic *C. glacialis*

collected at SHEBA ice camp in Canadian basin (Ashjian et al., 2003), and *C. glacialis* samples from the Bering Sea. Here, our data suggest a strong genetic differentiation between *C. glacialis* (collected from Skjerstadfjord) and *C. marshallae* (from Puget Sound c/o B. Frost) and may indicate that they are two distinct species, notably because of the high bootstrap support value and concordance factor observed in the branch separating these taxa (ML bootstrap = 100%; gCF = 80.3%; sCF = 85%), relative to the other species present in our phylogenetic analysis. However, we could not investigate further the taxonomic status of the three genetically differentiated clusters reported by analyses of mtCOI (Ashjian et al., 2017) since we did not include samples of *C. glacialis* from the Bering Sea. More individuals identified as *C. marshallae* and *C. glacialis* from these three localities must be analyzed to further assess the actual taxonomic status of these three genetic clusters. Combining comprehensive morphological examinations together with analyses of large numbers of genome-wide molecular markers such as single nucleotide polymorphisms (SNPs), following the genome-reduced representation protocol developed by Choquet et al. (2019), will allow testing for reproductive isolation among these three genetic entities.

Lastly, to get more insights on our ML tree, we quantified genealogical concordance in our phylogenetic dataset. Gene concordance factor (gCF) is defined as the percentage of decisive gene trees supporting a branch, while sCF is the percentage of decisive alignment sites supporting a branch in a reference tree (Minh et al., 2020). These concepts are important for phylogenomics datasets because genes that are from different chromosomes or from distant regions of the genome tend to show different levels of resolution or phylogenetic signals (see Minh et al., 2020; Rota et al., 2021). The results of our concordance estimations fully supported the split between *C. hyperboreus* and the two other *Calanus* groups: *C. finmarchicus* and *C. helgolandicus* (gCF = 100%; sCF = 96.3%). We also found that 109/191 single-copy protein-coding genes supported the separation between *C. finmarchicus* and *C. helgolandicus* groups. Although we obtained the maximum possible bootstrap support values for each branch of our ML tree, we still observed low gCF and sCF values for branches splitting the three species in the *C. helgolandicus* group and in the shallower splits in general. High bootstrap values, but somewhat low concordance factors between lineages are common and have also been observed in several studies using gene orthologs derived from transcriptomes (i.e., in frogs, Chan et al., 2020; in spiders, Kallal et al., 2020; in butterflies Rota et al., 2021; in ants van Elst et al., 2021). Low concordance values do not mean that the phylogenetic tree is unresolved, but rather gives us further insights on how related or congruent the genes are in resolving the species phylogeny (Minh et al., 2020). In addition, the relatively low concordance factors in the shallower branches of our phylogenetic tree may be attributed to conflicting signals among the 191 single-copy orthologs used to reconstruct our phylogenetic tree (Minh et al., 2020). Unfortunately, this issue of concordance between multiple genomic markers represents an obstacle for the calculation of divergence time within the genus *Calanus* and needs to be resolved by either clustering genes with similar phylogenetic signals or by using a different set of genomic markers from transcriptome datasets such as SNPs. Nevertheless, our study provides the most recent, well-resolved and multigenic phylogenetic analysis for copepod species of the genus *Calanus* in the northern seas.

5 | CONCLUSION

The use of RNA-sequencing enabled us to contribute and improve the existing transcriptome database for the genus *Calanus* as well as build a baseline information for future comparative transcriptomics in evolutionary and eco-physiological contexts. Our study is the first attempt to utilize phylotranscriptomics to resolve species relationships among the *Calanus* species living in the North Atlantic, North Pacific, and Arctic Oceans, and in copepods in general. This resulted in the reconstruction of a much-improved phylogenetic tree and clarification of certain ambiguities within the *Calanus* genus. Moreover, the phylogenetic tree inferred in this study showed the potential of using concordance factor to look at underlying variations in phylogenomics data beyond the limitations of bootstrapping method. As phylotranscriptomic analyses are getting more accessible and popular, more robust and streamlined, improvements in the ease of analyses and development of a consensus in interpretation of data shall be expected in the near future.

ACKNOWLEDGMENTS

We would like to acknowledge Nord University, which funded this research under the Phd funding no. 224000-138. We are very grateful to all the partners who provided us with samples, including Morten Krogstad, Eva Friis Møller, Ann Bucklin, Bruce Frost, and Leslie Noble. For their valuable inputs in bioinformatics, we would also like to thank Adnan Gora, Golam Rbbani, Le Qin Choo, Joost Raeymaekers, Lars Martin Jakt, and Alexander Jüterbock.

CONFLICT OF INTEREST

The authors declare they have no conflict of interest.

AUTHOR CONTRIBUTION

Apollo Marco Dalonos Lizano: Conceptualization (equal); Data curation (equal); Formal analysis (equal); Investigation (equal); Methodology (equal); Software (equal); Validation (equal); Visualization (equal); Writing – original draft (equal); Writing – review & editing (equal). **Irina Smolina:** Conceptualization (equal); Formal analysis (equal); Methodology (equal); Software (equal); Visualization (equal); Writing – review & editing (equal). **Marvin Choquet:** Conceptualization (equal); Project administration (equal); Supervision (equal); Writing – review & editing (equal). **Martina Kopp:** Methodology (equal); Project administration (equal); Resources (equal). **Galice Hoarau:** Conceptualization (equal); Funding acquisition (equal); Project administration (equal); Supervision (equal); Writing – original draft (equal); Writing – review & editing (equal).

DATA AVAILABILITY STATEMENT

RNA-seq reads used for assembling of 15 transcriptomes are available at NCBI SRA database under BioProject PRJNA744376 and 15 individual de novo transcriptomes generated on this project has been deposited at DDBJ/EMBL/GenBank TSA database and DRYAD server (https://doi.org/10.5061/dryad.n8pk0p2ww).

ORCID

Apollo Marco Lizano https://orcid.org/0000-0002-8216-3078
Irina Smolina https://orcid.org/0000-0002-0205-7663
Marvin Choquet https://orcid.org/0000-0001-6719-2332

REFERENCES

Annenkova, N. V., Ahrén, D., Logares, R., Kremp, A., & Rengefors, K. (2018). Delineating closely related dinoflagellate lineages using phylotranscriptomics. *Journal of Phycology*, *54*(4), 571–576. https://doi.org/10.1111/jpy.12748

Arnkværn, G., Daase, M., & Eiane, K. (2005). Dynamics of coexisting *Calanus finmarchicus*, *Calanus glacialis* and *Calanus hyperboreus* populations in a high-Arctic fjord. *Polar Biology*, *28*(7), 528–538. https://doi.org/10.1007/s00300-005-0715-8

Asai, S., Sanges, R., Lauritano, C., Lindeque, P. K., Esposito, F., Ianora, A., & Carotenuto, Y. (2020). De novo transcriptome assembly and gene expression profiling of the copepod Calanus helgolandicus feeding on the PUA-producing diatom Skeletonema marinoi. *Marine Drugs*, *18*(8), 392.

Ashjian, C. J., Campbell, R. G., Gelfman, C., Alatalo, P., & Elliott, S. M. (2017). Mesozooplankton abundance and distribution in association with hydrography on Hanna Shoal, NE Chukchi Sea, during August 2012 and 2013. *Deep Sea Research Part II: Topical Studies in Oceanography*, *144*, 21–36. https://doi.org/10.1016/j.dsr2.2017.08.012

Ashjian, C. J., Campbell, R. G., Welch, H. E., Butler, M., & Van Keuren, D. (2003). Annual cycle in abundance, distribution, and size in relation to hydrography of important copepod species in the western Arctic Ocean. *Deep Sea Research Part I: Oceanographic Research Papers*, *50*(10–11), 1235–1261. https://doi.org/10.1016/S0967-0637(03)00129-8

Bailey, A., De Wit, P., Thor, P., Browman, H. I., Bjelland, R., Shema, S., & Hop, H. (2017). Regulation of gene expression is associated with tolerance of the Arctic copepod *Calanus glacialis* to CO_2-acidified sea water. *Ecology and Evolution*, *7*(18), 7145–7160.

Berger, C. A., Steinberg, D. K., Copley, N. J., & Tarrant, A. M. (2021). *De novo* transcriptome assembly of the Southern Ocean copepod *Rhincalanus gigas* sheds light on developmental changes in gene expression. *Marine Genomics*, *58*, 100835. https://doi.org/10.1016/j.margen.2021.100835

Bi, K., Vanderpool, D., Singhal, S., Linderoth, T., Moritz, C., & Good, J. M. (2012). Transcriptome-based exon capture enables highly cost-effective comparative genomic data collection at moderate evolutionary scales. *BMC Genomics*, *13*(1), 403. https://doi.org/10.1186/1471-2164-13-403

Bonnet, D., & Frid, C. (2004). Seven copepod species considered as indicators of water-mass influence and changes: Results from a Northumberland coastal station. *ICES Journal of Marine Science*, *61*(4), 485–491. https://doi.org/10.1016/j.icesjms.2004.03.005

Bucklin, A., DiVito, K. R., Smolina, I., Choquet, M., Questel, J. M., Hoarau, G., & O'Neill, R. J. (2018). Population genomics of marine zooplankton. In M. F. Oleksiak & O. P. Rajora (Eds.), *Population genomics: Marine organisms* (pp. 61–102). Springer.

Bucklin, A., Frost, B. W., & Kocher, T. D. (1995). Molecular systematics of six *Calanus* and three *Metridia* species (Calanoida: Copepoda). *Marine Biology*, *121*(4), 655–664. https://doi.org/10.1007/BF00349301

Chan, K. O., Hutter, C. R., Wood, P. L. Jr., Grismer, L. L., & Brown, R. M. (2020). Larger, unfiltered datasets are more effective at resolving phylogenetic conflict: Introns, exons, and UCEs resolve ambiguities in Golden-backed frogs (Anura: Ranidae; genus Hylarana). *Molecular Phylogenetics and Evolution*, *151*, 106899. https://doi.org/10.1016/j.ympev.2020.106899

Cheon, S., Zhang, J., & Park, C. (2020). Is phylotranscriptomics as reliable as phylogenomics? *Molecular Biology and Evolution*, *37*(12), 3672–3683. https://doi.org/10.1093/molbev/msaa181

Choquet, M., Hatlebakk, M., Dhanasiri, A. K., Kosobokova, K., Smolina, I., Søreide, J. E., Svensen, C., Melle, W., Kwaśniewski, S., Eiane, K., Daase, M., Tverberg, V., Skreslet, S., Bucklin, A., & Hoarau, G. (2017). Genetics redraws pelagic biogeography of *Calanus*. *Biology Letters*, *13*(12), 20170588.

Choquet, M., Kosobokova, K., Kwaśniewski, S., Hatlebakk, M., Dhanasiri, A. K. S., Melle, W., Daase, M., Svensen, C., Søreide, J. E., & Hoarau, G. (2018). Can morphology reliably distinguish between the copepods *Calanus finmarchicus* and *C. glacialis*, or is DNA the only way? *Limnology and Oceanography: Methods*, *16*(4), 237–252. https://doi.org/10.1002/lom3.10240

Choquet, M., Smolina, I., Dhanasiri, A. K., Blanco-Bercial, L., Kopp, M., Jueterbock, A., Sundaram, A. Y., & Hoarau, G. (2019). Towards population genomics in non-model species with large genomes: a case study of the marine zooplankton Calanus finmarchicus. *Royal Society open science*, *6*(2), 180608.

Cleary, A. C., Søreide, J. E., Freese, D., Niehoff, B., & Gabrielsen, T. M. (2017). Feeding by *Calanus glacialis* in a high arctic fjord: potential seasonal importance of alternative prey. *ICES Journal of Marine Science*, *74*(7), 1937–1946. https://doi.org/10.1093/icesjms/fsx106

Du, F., Xu, G., Nie, Z., Xu, P., & Gu, R. (2014). Transcriptome analysis gene expression in the liver of *Coilia nasus* during the stress response. *BMC Genomics*, *15*(1), 1–11. https://doi.org/10.1186/1471-2164-15-558

Eldem, V., Zararsiz, G., Taşçi, T., Duru, I. P., Bakir, Y., & Erkan, M. (2017). Transcriptome analysis for non-model organism: Current status and best-practices. *Applications of RNA-Seq and Omics Strategies-From Microorganisms to Human Health*, *1*(2), 1–19.

Emms, D. M., & Kelly, S. (2019). OrthoFinder: Phylogenetic orthology inference for comparative genomics. *Genome Biology*, *20*(1), 1–14.

Eyun, S. I. (2017). Phylogenomic analysis of Copepoda (Arthropoda, Crustacea) reveals unexpected similarities with earlier proposed morphological phylogenies. *BMC Evolutionary Biology*, *17*(1), 1–12. https://doi.org/10.1186/s12862-017-0883-5

Fleminger, A., & Hulsemann, K. (1977). Geographical range and taxonomic divergence in North Atlantic *Calanus* (*C. helgolandicus*, *C. finmarchicus* and *C. glacialis*). *Marine Biology*, *40*(3), 233–248. https://doi.org/10.1007/BF00390879

Frost, B. W. (1974). *Calanus marshallae*, a new species of calanoid copepod closely allied to the sibling species *C. finmarchicus* and *C. glacialis*. *Marine Biology*, *26*(1), 77–99. https://doi.org/10.1007/BF00389089

Fu, L., Niu, B., Zhu, Z., Wu, S., & Li, W. (2012). CD-HIT: Accelerated for clustering the next-generation sequencing data. *Bioinformatics*, *28*(23), 3150–3152. https://doi.org/10.1093/bioinformatics/bts565

Galperin, M. Y., Wolf, Y. I., Makarova, K. S., Vera Alvarez, R., Landsman, D., & Koonin, E. V. (2021). COG database update: Focus on microbial diversity, model organisms, and widespread pathogens. *Nucleic Acids Research*, *49*(D1), D274–D281. https://doi.org/10.1093/nar/gkaa1018

Gan, Z., Yuan, J., Liu, X., Dong, D., Li, F., & Li, X. (2020). Comparative transcriptomic analysis of deep-and shallow-water barnacle species (Cirripedia, Poecilasmatidae) provides insights into deep-sea adaptation of sessile crustaceans. *BMC Genomics*, *21*(1), 1–13. https://doi.org/10.1186/s12864-020-6642-9

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., Chen, Z., Mauceli, E., Hacohen, N., Gnirke, A., Rhind, N., di Palma, F., Birren, B. W., Nusbaum, C., Lindblad-Toh, K., … Regev, A. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology*, *29*(7), 644–652. https://doi.org/10.1038/nbt.1883

Haas, B. J., Papanicolaou, A., Yassour, M., Grabherr, M., Philip, D., Bowden, J., Couger, M. B., Eccles, D., Li, B., Lieber, M., MacManes, M. D., Ott, M., Orvis, J., Pochet, N., Strozzi, F., Weeks, N., Westerman, R., William, T., Dewey, C. N., & Regev, A. (2013). *De novo* transcript sequence recostruction from RNA-Seq: Reference generation and analysis with Trinity. *Nature Protocols*, *8*, 1494–1512. https://doi.org/10.1038/nprot.2013.084

Huerta-cepas, J., Forslund, K., Coelho, L. P., Szklarczyk, D., Jensen, L. J., Mering, C. V., & Bork, P. (2017). Fast genome-wide functional annotation through orthology assignment by eggNOG-mapper. *Molecular Biology and Evolution*, *34*(8), 2115–2122. https://doi.org/10.1093/molbev/msx148

Huerta-Cepas, J., Szklarczyk, D., Heller, D., Hernández-Plaza, A., Forslund, S. K., Cook, H., Mende, D. R., Letunic, I., Rattei, T., Jensen, L. J., von Mering, C., & Bork, P. (2019). EggNOG 5.0: A hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Research*, *47*(D1), D309–D314. https://doi.org/10.1093/nar/gky1085

Johnson, L. S., Eddy, S. R., & Portugaly, E. (2010). Hidden Markov model speed heuristic and iterative HMM search procedure. *BMC Bioinformatics*, *11*(1), 1–8. https://doi.org/10.1186/1471-2105-11-431

Kallal, R. J., Dimitrov, D., Arnedo, M. A., Giribet, G., & Hormiga, G. (2020). Monophyly, taxon sampling, and the nature of ranks in the classification of orb-weaving spiders (Araneae: Araneoidea). *Systematic Biology*, *69*(2), 401–411.

Kozol, R., Blanco-Bercial, L., & Bucklin, A. (2012). Multi-gene analysis reveals a lack of genetic divergence between *Calanus agulhensis* and *C. sinicus* (Copepoda; Calanoida). *PLoS One*, *7*(10), e45710. https://doi.org/10.1371/journal.pone.0045710

Lauritano, C., Roncalli, V., Ambrosino, L., Cieslak, M. C., & Ianora, A. (2020). First de novo transcriptome of the copepod Rhincalanus gigas from Antarctic waters. *Biology*, *9*(11), 410. https://doi.org/10.3390/biology9110410

Leaché, A. D., & Oaks, J. R. (2017). The utility of single nucleotide polymorphism (SNP) data in phylogenetics. *Annual Review of Ecology, Evolution, and Systematics*, *48*, 69–84. https://doi.org/10.1146/annurev-ecolsys-110316-022645

Lenz, P. H., Lieberman, B., Cieslak, M. C., Roncalli, V., & Hartline, D. K. (2021). Transcriptomics and metatranscriptomics in zooplankton: wave of the future? *Journal of Plankton Research*, *43*(1), 3–9. https://doi.org/10.1093/plankt/fbaa058

Lenz, P. H., Roncalli, V., Hassett, R. P., Wu, L. S., Cieslak, M. C., Hartline, D. K., & Christie, A. E. (2014). *De novo* assembly of a transcriptome for *Calanus finmarchicus* (crustacea, copepoda) – The dominant zooplankter of the North Atlantic Ocean. *PLoS One*, *9*(2), e88589. https://doi.org/10.1371/journal.pone.0088589

Li, J., Lemer, S., Kirkendale, L., Bieler, R., Cavanaugh, C., & Giribet, G. (2020). Shedding light: A phylotranscriptomic perspective illuminates the origin of photosymbiosis in marine bivalves. *BMC Evolutionary Biology*, *20*(1), 1–15. https://doi.org/10.1186/s12862-020-01614-7

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. https://doi.org/10.1093/bioinformatics/btl158

Li, X., Cui, Z., Liu, Y., Song, C., & Shi, G. (2013). Transcriptome analysis and discovery of genes involved in immune pathways from hepatopancreas of microbial challenged mitten crab *Eriocheir sinensis*. *PLoS One*, *8*(7), 1–15. https://doi.org/10.1371/journal.pone.0068233

Luikart, G., England, P. R., Tallmon, D., Jordan, S., & Taberlet, P. (2003). The power and promise of population genomics: From genotyping to genome typing. *Nature Reviews Genetics*, *4*(12), 981–994. https://doi.org/10.1038/nrg1226

Maldonado, J. A., Firneno, T. J. Jr., Roelke, C. E., Rains, N. D., Mwgiri, J., & Fujita, M. K. (2020). Transcriptome sequencing reveals signatures of positive selection in the Spot-Tailed Earless Lizard. *PLoS One*, *15*(6), e0234504. https://doi.org/10.1371/journal.pone.0234504

Marlétaz, F., Peijnenburg, K. T. C. A., Goto, T., Satoh, N., & Rokhsar, D. S. (2019). A new spiralian phylogeny places the enigmatic arrow worms among gnathiferans. *Current Biology*, *29*(2), 312–318.e3. https://doi.org/10.1016/j.cub.2018.11.042

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*, *17*(1), 10–12. https://doi.org/10.14806/ej.17.1.200

McLaren, I. A., Sevigny, J. M., & Corkett, C. J. (1988). Body sizes, development rates, and genome sizes among *Calanus* species. *Hydrobiologia*, *167-168*(1), 275–284. https://doi.org/10.1007/BF00026315

Metzker, M. L. (2010). Sequencing technologies the next generation. *Nature Reviews Genetics*, *11*(1), 31–46. https://doi.org/10.1038/nrg2626

Mi, H., Ebert, D., Muruganujan, A., Mills, C., Albou, L. P., Mushayamaha, T., & Thomas, P. D. (2021). PANTHER version 16: A revised family classification, tree-based classification tool, enhancer regions and extensive API. *Nucleic Acids Research*, *49*(D1), D394–D403. https://doi.org/10.1093/nar/gkaa1106

Mi, H., Muruganujan, A., Ebert, D., Huang, X., & Thomas, P. D. (2019). PANTHER version 14: More genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools. *Nucleic Acids Research*, *47*(D1), D419–D426. https://doi.org/10.1093/nar/gky1038

Minh, B. Q., Hahn, M. W., & Lanfear, R. (2020). New methods to calculate concordance factors for phylogenomic datasets. *Molecular Biology and Evolution*, *37*(9), 2727–2733. https://doi.org/10.1093/molbev/msaa106

Moreno-Santillán, D. D., Machain-Williams, C., Hernández-Montes, G., & Ortega, J. (2019). De novo transcriptome assembly and functional annotation in five species of bats. *Scientific Reports*, *9*(1), 1–12. https://doi.org/10.1038/s41598-019-42560-9

Nguyen, L. T., Schmidt, H. A., Von Haeseler, A., & Minh, B. Q. (2015). IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular Biology and Evolution*, *32*(1), 268–274. https://doi.org/10.1093/molbev/msu300

Ning, J., Wang, M., Li, C., & Sun, S. (2013). Transcriptome sequencing and *de novo* analysis of the copepod *Calanus sinicus* using 454 GS FLX. *PLoS One*, *8*(5), e63741. https://doi.org/10.1371/journal.pone.0063741

Ohnishi, T., Hirai, J., Shimode, S., & Tsuda, A. (2019). Identification of molecular markers associated with starvation in female *Calanus sinicus* (Copepoda: Calanoida). *Marine Ecology Progress Series*, *614*, 51–65. https://doi.org/10.3354/meps12904

Pai, T. W., Li, K. H., Yang, C. H., Hu, C. H., Lin, H. J., Wang, W. D., & Chen, Y. R. (2018). Multiple model species selection for transcriptomics analysis of non-model organisms. *BMC Bioinformatics*, *19*(9), 53–66. https://doi.org/10.1186/s12859-018-2278-z

Payton, L., Noirot, C., Hoede, C., Hüppe, L., Last, K., Wilcockson, D., Ershova, E. A., Valière, S., & Meyer, B. (2020). Daily transcriptomes of the copepod *Calanus finmarchicus* during the summer solstice at high Arctic latitudes. *Scientific Data*, *7*(1), 1–9. https://doi.org/10.1038/s41597-020-00751-4

Peijnenburg, K. T., Janssen, A. W., Wall-Palmer, D., Goetze, E., Maas, A. E., Todd, J. A., & Marlétaz, F. (2020). The origin and diversification of pteropods precede past perturbations in the Earth's carbon cycle. *Proceedings of the National Academy of Sciences*, *117*(41), 25609–25617. https://doi.org/10.1073/pnas.1920918117

Roncalli, V., Christie, A. E., Sommer, S. A., Cieslak, M. C., Hartline, D. K., & Lenz, P. H. (2017). A deep transcriptomic resource for the copepod

crustacean *Labidocera madurae*: A potential indicator species for assessing near shore ecosystem health. *PLoS One*, *12*(10), e0186794. https://doi.org/10.1371/journal.pone.0186794

Rota, J., Twort, V., Chiocchio, A., Peña, C., Wheat, C. W., Kaila, L., & Wahlberg, N. (2021). The unresolved phylogenomic tree of butterflies and moths (Lepidoptera): Assessing the potential causes and consequences. *bioRxiv*. https://www.hindawi.com/journals/bmri/2014/493825/#copyright

Schuster, S. C. (2008). Next-generation sequencing transforms today's biology. *Nature Methods*, *5*(1), 16–18. https://doi.org/10.1038/nmeth1156

Seppey, M., Manni, M., & Zdobnov, E. M. (2019). BUSCO: Assessing genome assembly and annotation completeness. *Methods in Molecular Biology*, *1962*, 227–245.

Sievers, F., & Higgins, D. G. (2014). Clustal omega, accurate alignment of very large numbers of sequences. *Methods in Molecular Biology*, *1079*, 105–116. https://doi.org/10.1007/978-1-62703-646-7_6

Simion, P., Belkhir, K., François, C., Veyssier, J., Rink, J. C., Manuel, M., Philippe, H., & Telford, M. J. (2018). A software tool 'CroCo' detects pervasive cross-species contamination in next generation sequencing data. *BMC Biology*, *16*(1), 1–9. https://doi.org/10.1186/s12915-018-0486-7

Skottene, E., Tarrant, A. M., Olsen, A. J., Altin, D., Østensen, M.-A., Hansen, B. H., Choquet, M., Jenssen, B. M., & Olsen, R. E. (2019). The β-oxidation pathway is downregulated during diapause termination in *Calanus* copepods. *Scientific Reports*, *9*(1), 1–13. https://doi.org/10.1038/s41598-019-53032-5

Smith-Unna, R., Boursnell, C., Patro, R., Hibberd, J. M., & Kelly, S. (2016). TransRate: reference-free quality assessment of *de novo* transcriptome assemblies. *Genome Research*, *26*(8), 1134–1144. https://doi.org/10.1101/gr.196469.115

Smolina, I., Kollias, S., Poortvliet, M., Nielsen, T. G., Lindeque, P., Castellani, C., Møller, E. F., Blanco-Bercial, L., & Hoarau, G. (2014). Genome- and transcriptome-assisted development of nuclear insertion/deletion markers for *Calanus* species (Copepoda: Calanoida) identification. *Molecular Ecology Resources*, *14*(5), 1072–1079. https://doi.org/10.1111/1755-0998.12241

Tarrant, A. M., Baumgartner, M. F., Hansen, B. H., Altin, D., Nordtug, T., & Olsen, A. J. (2014). Transcriptional profiling of reproductive development, lipid storage and molting throughout the last juvenile stage of the marine copepod *Calanus finmarchicus*. *Frontiers in Zoology*, *11*(1), 1–15. https://doi.org/10.1186/s12983-014-0091-8

Tarrant, A. M., Nilsson, B., & Hansen, B. W. (2019). Molecular physiology of copepods—From biomarkers to transcriptomes and back again. *Comparative Biochemistry and Physiology - Part D: Genomics and Proteomics*, *30*, 230–247. https://doi.org/10.1016/j.cbd.2019.03.005

Tatusov, R. L., Fedorova, N. D., Jackson, J. D., Jacobs, A. R., Kiryutin, B., Koonin, E. V., Krylov, D. M., Mazumder, R., Mekhedov, S. L., Nikolskaya, A. N., Rao, B. S., Smirnov, S., Sverdlov, A. V., Vasudevan, S., Wolf, Y. I., Yin, J. J., & Natale, D. A. (2003). The COG database: An updated version includes eukaryotes. *BMC Bioinformatics*, *4*(4), 1–14.

Ungaro, A., Pech, N., Martin, J. F., Scott McCairns, R. J., Mévy, J. P., Chappaz, R., & Gilles, A. (2017). Challenges and advances for transcriptome assembly in non-model species. *PLoS One*, *12*(9), 1–21. https://doi.org/10.1371/JOURNAL.PONE.0185020

van Elst, T., Eriksson, T. H., Gadau, J., Johnson, R. A., Rabeling, C., Taylor, J. E., & Borowiec, M. L. (2021). Comprehensive phylogeny of Myrmecocystus honey ants highlights cryptic diversity and infers evolution during aridification of the American Southwest. *Molecular Phylogenetics and Evolution*, *155*, 107036.

Yang, Q., Sun, F., Yang, Z., & Li, H. (2014). Comprehensive transcriptome study to develop molecular resources of the copepod calanus sinicus for their potential ecological applications. *BioMed Research International*. https://doi.org/10.1155/2014/493825

Zhang, C., Rabiee, M., Sayyari, E., & Mirarab, S. (2018). ASTRAL-III: Polynomial time species tree reconstruction from partially resolved gene trees. *BMC Bioinformatics*, *19*(Suppl. 6), 15–30. https://doi.org/10.1186/s12859-018-2129-y

Zhang, S.-F., Yuan, C.-J., Chen, Y., Chen, X.-H., Li, D.-X., Liu, J.-L., Lin, L., & Wang, D.-Z. (2016). Comparative transcriptomic analysis reveals novel insights into the adaptive response of *Skeletonema costatum* to changing ambient phosphorus. *Frontiers in Microbiology*, *7*, 1–15. https://doi.org/10.3389/fmicb.2016.01476

Zhao, Y., Zhang, R., Jiang, K.-W., Qi, J. I., Hu, Y. I., Guo, J., Zhu, R., Zhang, T., Egan, A. N., Yi, T.-S., Huang, C.-H., & Ma, H. (2021). Nuclear phylotranscriptomics and phylogenomics support numerous polyploidization events and hypotheses for the evolution of rhizobial nitrogen-fixing symbiosis in Fabaceae. *Molecular Plant*, *14*(5), 748–773. https://doi.org/10.1016/j.molp.2021.02.006

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: Lizano, A. M., Smolina, I., Choquet, M., Kopp, M., & Hoarau, G. (2022). Insights into the species evolution of *Calanus* copepods in the northern seas revealed by *de novo* transcriptome sequencing. *Ecology and Evolution*, 12, e8606. https://doi.org/10.1002/ece3.8606

List of previously published theses for PhD in Aquaculture / PhD in Aquatic Biosciences, Nord University

No. 1 (2011)
PhD in Aquaculture
Chris André Johnsen
Flesh quality and growth of farmed Atlantic salmon (*Salmo salar* L.) in relation to feed, feeding, smolt type and season
ISBN: 978-82-93165-00-2

No. 2 (2012)
PhD in Aquaculture
Jareeporn Ruangsri
Characterization of antimicrobial peptides in Atlantic cod
ISBN: 978-82-93165-01-9

No. 3 (2012)
PhD in Aquaculture
Muhammad Naveed Yousaf
Characterization of the cardiac pacemaker and pathological responses to cardiac diseases in Atlantic salmon (*Salmo salar* L.)
ISBN: 978-82-93165-02-6

No. 4 (2012)
PhD in Aquaculture
Carlos Frederico Ceccon Lanes
Comparative Studies on the quality of eggs and larvae from broodstocks of farmed and wild Atlantic cod
ISBN: 978-82-93165-03-3

No. 5 (2012)
PhD in Aquaculture
Arvind Sundaram
Understanding the specificity of the innate immune response in teleosts: Characterisation and differential expression of teleost-specific Toll-like receptors and microRNAs
ISBN: 978-82-93165-04-0

No. 6 (2012)
PhD in Aquaculture
Teshome Tilahun Bizuayehu
Characterization of microRNA during early ontogeny and sexual development of Atlantic halibut (*Hippoglossus hippoglossus* L.)
ISBN: 978-82-93165-05-7

No. 7 (2013)
PhD in Aquaculture
Binoy Rajan
Proteomic characterization of Atlantic cod skin mucosa – Emphasis on innate immunity and lectins
ISBN: 978-82-93165-06-04

No. 8 (2013)
PhD in Aquaculture
Anusha Krishanthi Shyamali Dhanasiri
Transport related stress in zebrafish: physiological responses and bioremediation
ISBN: 978-82-93165-07-1

No. 9 (2013)
PhD in Aquaculture
Martin Haugmo Iversen
Stress and its impact on animal welfare during commercial production of Atlantic salmon (*Salmo salar* L.)
ISBN: 978-82-93165-08-8

No. 10 (2013)
PhD in Aquatic Biosciences
Alexander Jüterbock
Climate change impact on the seaweed *Fucus serratus*, a key foundational species on North Atlantic rocky shores
ISBN: 978-82-93165-09-5

No. 11 (2014)
PhD in Aquatic Biosciences
Amod Kulkarni
Responses in the gut of black tiger shrimp *Penaeus monodon* to oral vaccine candidates against white spot disease
ISBN: 978-82-93165-10-1

No. 12 (2014)
PhD in Aquatic Biosciences
Carlo C. Lazado
Molecular basis of daily rhythmicity in fast skeletal muscle of Atlantic cod *(Gadus morhua)*
ISBN: 978-82-93165-11-8

No. 13 (2014)
PhD in Aquaculture
Joanna Babiak
Induced masculinization of Atlantic halibut (*Hippoglossus hippoglossus* L.): towards the goal of all-female production
ISBN: 978-82-93165-12-5

No. 14 (2015)
PhD in Aquaculture
Cecilia Campos Vargas
Production of triploid Atlantic cod: A comparative study of muscle growth dynamics and gut morphology
ISBN: 978-82-93165-13-2

No. 15 (2015)
PhD in Aquatic Biosciences
Irina Smolina
Calanus in the North Atlantic: species identification, stress response, and population genetic structure
ISBN: 978-82-93165-14-9

No. 16 (2016)
PhD in Aquatic Biosciences
Lokesh Jeppinamogeru
Microbiota of Atlantic salmon (*Salmo salar L.*), during their early and adult life
ISBN: 978-82-93165-15-6

No. 17 (2017)
PhD in Aquatic Biosciences
Christopher Edward Presslauer
Comparative and functional analysis of microRNAs during zebrafish gonadal development
ISBN: 978-82-93165-16-3

No. 18 (2017)
PhD in Aquatic Biosciences
Marc Jürgen Silberberger
Spatial scales of benthic ecosystems in the sub-Arctic Lofoten-Vesterålen region
ISBN: 978-82-93165-17-0

No. 19 (2017)
PhD in Aquatic Biosciences
Marvin Choquet
Combining ecological and molecular approaches to redefine the baseline knowledge of the genus Calanus in the North Atlantic and the Arctic Oceans
ISBN: 978-82-93165-18-7

No. 20 (2017)
PhD in Aquatic Biosciences
Torvald B. Egeland
Reproduction in Arctic charr – timing and the need for speed
ISBN: 978-82-93165-19-4

No. 21 (2017)
PhD in Aquatic Biosciences
Marina Espinasse
Interannual variability in key zooplankton species in the North-East Atlantic: an analysis based on abundance and phenology
ISBN: 978-82-93165-20-0

No. 22 (2018)
PhD in Aquatic Biosciences
Kanchana Bandara
Diel and seasonal vertical migrations of high-latitude zooplankton: knowledge gaps and a high-resolution bridge
ISBN: 978-82-93165-21-7

No. 23 (2018)
PhD in Aquatic Biosciences
Deepti Manjari Patel
Characterization of skin immune and stress factors of lumpfish, *Cyclopterus lumpus*
ISBN: 978-82-93165-21-7

No. 24 (2018)
PhD in Aquatic Biosciences
Prabhugouda Siriyappagouder
The intestinal mycobiota of zebrafish – community profiling and exploration of the impact of yeast exposure early in life
ISBN: 978-82-93165-23-1

No. 25 (2018)
PhD in Aquatic Biosciences
Tor Erik Jørgensen
Molecular and evolutionary characterization of the Atlantic cod mitochondrial genome
ISBN: 978-82-93165-24-8

No. 26 (2018)
PhD in Aquatic Biosciences
Yangyang Gong
Microalgae as feed ingredients for Atlantic salmon
ISBN: 978-82-93165-25-5

No. 27 (2018)
PhD in Aquatic Biosciences
Ove Nicolaisen
Approaches to optimize marine larvae production
ISBN: 978-82-93165-26-2

No. 28 (2019)
PhD in Aquatic Biosciences
Qirui Zhang
The effect of embryonic incubation temperature on the immune response of larval and adult zebrafish (*Danio rerio*)
ISBN: 978-82-93165-27-9

No. 29 (2019)
PhD in Aquatic Biosciences
Andrea Bozman
The structuring effects of light on the deep-water scyphozoan *Periphylla periphylla*
ISBN: 978-82-93165-28-6

No. 30 (2019)
PhD in Aquatic Biosciences
Helene Rønquist Knutsen
Growth and development of juvenile spotted wolffish (*Anarhichas minor*) fed microalgae incorporated diets
ISBN: 978-82-93165-29-3

No. 31 (2019)
PhD in Aquatic Biosciences
Shruti Gupta
Feed additives elicit changes in the structure of the intestinal bacterial community of Atlantic salmon
ISBN: 978-82-93165-30-9

No. 32 (2019)
PhD in Aquatic Biosciences
Peter Simon Claus Schulze
Phototrophic microalgal cultivation in cold and light-limited environments
ISBN: 978-82-93165-31-6

No. 33 (2019)
PhD in Aquatic Biosciences
Maja Karoline Viddal Hatlebakk
New insights into Calanus glacialis and C. finmarchicus distribution, life histories and physiology in high-latitude seas
ISBN: 978-82-93165-32-3

No. 34 (2019)
PhD in Aquatic Biosciences
Arseny Dubin
Exploration of an anglerfish genome
ISBN: 978-82-93165-33-0

No. 35 (2020)
PhD in Aquatic Biosciences
Florence Chandima Perera Willora Arachchilage
The potential of plant ingredients in diets of juvenile lumpfish (*Cyclopterus lumpus*)
ISBN: 978-82-93165-35-4

No. 36 (2020)
PhD in Aquatic Biosciences
Ioannis Konstantinidis
DNA hydroxymethylation and improved growth of Nile tilapia (*Oreochromis niloticus*) during domestication
ISBN: 978-82-93165-36-1

No. 37 (2021)
PhD in Aquatic Biosciences
Youngjin Park
Transcriptomic and cellular studies on the intestine of Atlantic salmon
Discovering intestinal macrophages using omic tools
ISBN: 978-82-93165-34-7

No. 38 (2021)
PhD in Aquatic Biosciences
Purushothaman Kathiresan
Proteomics of early embryonic development of zebrafish (Danio rerio)
ISBN: 978-82-93165-37-8

No. 39 (2021)
PhD in Aquatic Biosciences
Valentin Kokarev
Macrobenthic communities of sub-Arctic deep fjords: composition, spatial patterns and community assembly
ISBN: 978-82-93165-38-5

No. 40 (2021)
PhD in Aquatic Biosciences
Aurélien Delaval
Population genomics of a critically endangered data-deficient elasmobranch, the blue skate *Dipturus batis*
ISBN: 978-82-93165-39-2

No. 41 (2021)
PhD in Aquatic Biosciences
Isabel Sofía Abihssira García
Environmental impact of microplastics in relation to Atlantic salmon farming
ISBN: 978-82-93165-40-8

No. 42 (2022)
PhD in Aquatic Biosciences
Yousri Abdelmutalab Ahmed Abdelhafiz
Insights into the bacterial communities of Nile tilapia – core members and intergenerational transfer
ISBN: 978-82-93165-41-5

No. 43 (2022)
PhD in Aquatic Biosciences
Fredrik Ribsskog Staven
Interaction studies on lumpfish exposed to Atlantic salmon: behavioural observations and the underlying physiological and neurobiological mechanisms
ISBN: 978-82-93165-42-2

No. 44 (2022)
PhD in Aquatic Biosciences
Solveig Lysfjord Sørensen
Influence of feed ingredients and additives on mucosal health with focus on the intestine of Atlantic salmon (*Salmo salar*)
ISBN: 978-82-93165-43-9

The copepod genus *Calanus* or "Raudåte" by Norwegian name is an ecologically important marine species in the Northern Seas (North Atlantic, Arctic, and Northern Pacific). *Calanus* species play a key role in energy transfer in marine food webs, both as primary consumers and as prey for fish, seabirds, and other marine predators. Commercially, they have been a great source of Omega-3 fatty acid. Despite their ecological and commercial importance, species-level taxonomy in the genus *Calanus* remains challenging because of their identical morphology, overlapping distribution, huge genome sizes, and limited genomic and transcriptomics resources. Species-level identification in the genus *Calanus* is very important for their proper conservation and management. The current fisheries regulations in Norway only focuses on a single species, *C. finmarchicus*. However, recent genetic analyses using few molecular markers have shown overlapping distribution for at least several species across the Northern Seas, including the Norwegian fjords. This thesis addressed some of these species-level taxonomy problems present in the genus *Calanus* using genomics and transcriptomics approaches, which aides in the proper understanding of their ecology and evolution. Results of these studies may also have implications in the current fisheries regulations implemented in the country.

NORD University

ISBN: 978-82-93165-44-6
Trykk: Trykkeriet, Nord universitet
www.nord.no