UPPSALA
UNIVERSITET

UPTEC X 18 016
Examensarbete 30 hp
Juni 2018

Development of a phylogenomic framework for the krill

Arusjak Gevorgyan

UPPSALA UNIVERSITET

Teknisk- naturvetenskaplig fakultet
UTH-enheten

Besöksadress:
Ångströmlaboratoriet
Lägerhyddsvägen 1
Hus 4, Plan 0

Postadress:
Box 536
751 21 Uppsala

Telefon:
018 – 471 30 03

Telefax:
018 – 471 30 00

Hemsida:
http://www.teknat.uu.se/student

Abstract

Development of a phylogenomic framework for the krill

Arusjak Gevorgyan

Over the last few decades, many krill stocks have declined in size and number, likely as a consequence of global climate change (Siegel 2016). A major risk factor is the increased level of carbon dioxide (CO2) in the ocean. A collapse of the krill population has the potential to cause disruption of the ocean ecosystem, as krill are the main connection between primary producers such as phytoplankton and larger animals (Murphy et al. 2012). The aim of this project is to produce the first phylogenomic framework with help of powerful comparative bioinformatics and phylogenomic methods in order to find and analyse the genes that help krill adapt to its environment. Problem with these studies is that we still do not have access to a reference genome sequence of any krill species. To strengthen and increase trust in our studies two different pipelines were performed, each with different Orthology Assessment Toolkits (OATs), Orthograph and UPhO, in order to establish orthology relationships between transcripts/genes. Since UPhO produces well-supported trees where the majority of the gene trees match the species tree, it is recommended as the proper OATs for generating a robust molecular phylogeny of krill. The second aim with his project was to estimate the level of positive selection in E. superba in order to lay a foundation about level of selection acting on protein-coding sequences in krill. As expected, the level of selection was quite high in E. superba, which indicates that krill are adapted to the changing environment by positive selection rather than natural genetic drift.

Handledare: Andreas Wallberg
Ämnesgranskare: Fabien Burki
Examinator: Jan Andersson
ISSN: 1401-2138, UPTEC X 18 016

Sammanfattning

Det senaste årtiondet har växthusgasen koldioxids (CO_2) ökning i atmosfären orsakat obalans och oro hos havsmiljön. En av arterna som påverkas av denna förändring är krill. Havsförsurningen medför att krillägg inte kläcks och minskas drastisk, vilket har stor påverkan på marina ekosystemet då krill är länken mellan primärproducenter som växtplankton och större djur så som vithajar. För att hantera dessa extrema förändringar som orsakas av den ökande koldioxidhalten i atmosfären är det jätteviktig att dyka in i den molekylära nivån hos krill.

Syftet med detta projekt var att konstruera fylogenetiskt träd med hjälp av transkriptomdata för att öka vår uppfattning om krills evolution. Eftersom målet var att producera den mest robusta molekylära fylogenin av krill, två metoder användes med två olika sätt att gruppera ortologa sekvenser. Orthograph (Petersen *et al.* 2017), som var grafbaserad jämförde och klustrade sekvenser för att identifiera ortologa grupper medans UPhO (Ballesteros et al., 2017), en trädbaserad metod, utöver klustring även använde fylogenetiska metoder.

Eftersom de två olika metoderna gav två olika stöd för artträdet, antyder det att i praktiken är det viktigt att överväga metoderna noggrant för att studera krills evolution. Då UPhO genererade genträd med mycket bättre stöd för artträdet jämfört med Orthograph, ansåg den vara den mest pålitliga metoden för att skapa robust molekylärfylogeni av krill.

I den andra delen av projektet analyserades graden av positiv selektion hos *Euphausia superba* för att få en bild av hur krill adapteras till miljöförändringarna som förekommer i havet. Med hjälp av polymorfism data (SNP) inom *E. superba* och divergensdata mellan *E. superba* och *E. crystallorophias* kunde graden av positiv selektion i de ortologa generna hos *E. superba* uppskattas. Då det är vanlig att arter med stor biomassa utsetts för selektion snarare än genetisk drift och krill är känd att ha stor biomassa, selektionsvärdet för *E. superba* förväntades att vara rätt så högt. Precis som det var förväntad var selektionsvärdet för *E. superba* högt ($\alpha = 0.69$) vilket indikerar att krill adapteras till miljöförändringar via selektion.

Table of contents

Abbreviations

D	Fixed differences
MSA	Multiple sequence alignments
N	Non- synonymous mutations
N_e	Effective population size
OAT	Orthology Assessment Toolkit
OG	Orthology group
P	Polymorphisms
S	Synonymous mutations
SNP	Single Nucleotide Polymorphism

1 Background

1.1 Environmental impact on krill

Among the most important animals for the ocean ecosystem are krill. These crustacean animals comprise 85 species that belong to the Euphausiidae family, have large population sizes and biomass, and lifespans of up to six years. Krill are keystone species for the oceans. They feed on phytoplankton and zooplankton from the surface of the ocean and are important food for marine mammals, birds and fish (*Jarman 2001*). Some krill, such as the Antarctic krill Euphausia superba, have remarkable plasticity and can grow in summer when access to food is good and shrink during winter when food is limited (*Siegel 2016*), making them well adapted to seasonal fluctuations in resources. However, over the last few decades, many krill stocks have declined in size and number, likely as a consequence of global climate change (*Siegel 2016*). A major risk factor is the increased level of carbon dioxide (CO_2) in the ocean. As CO_2 increases, carbonate (CO_3^{-2}) in the ocean decreases and the amount of H^+ increases, which in turn reduces the pH value in the ocean. As a consequence, krill eggs do not hatch at the expected rate (*Kawaguchi et al. 2013*). A collapse of the krill population has the potential to cause disruption of the ocean ecosystem, as krill are the main connection between primary producers such as phytoplankton and larger animals (*Murphy et al. 2013*).

1.2 The importance of genomic studies on krill

Development of genomic resources is important for helping us understand how krill are genetically adapted to the marine environment and may respond to climate change (*Stapley et al. 2010*). Several different studies have been carried out in order to expand this our insights into environmental genomics and physiology in krill. In one molecular study, researchers have closely investigated the adaptation that takes place in Antarctic krill metabolic pathways due to increasing level of CO_2 in the environment (*Meyer et al. 2015*). A second group has performed laboratory experiments and investigated how sensitive krill eggs are against ocean acidification. Since krill eggs sink 700 -1000 m from the ocean's surface (where carbon dioxide partial pressure is high) in order to hatch, they experience the changes of the CO_2 much more rigorously, which affects the krill eggs by decreasing hatching rate (*Kawaguchi et al. 2013*). Not only krill eggs are affected by climate change but also matured krill, since growth rate among them decrease due to increasing temperature in the ocean (*Atkinson et al. 2006*). These genetic and physiological studies have provided the first glimpses into how krill responds to changes to the environment and can cope with climate change in the ocean in the short term. However, we still have almost no insight into the genetic adaptation in krill and how they may respond and adapt at the genetic level in the longer term.

1.3 Potential and limitations of krill data

One of the main challenges of studying krill evolution and adaptation is that we still do not have access to a reference genome sequence of any krill species. This is due to the extremely large size of krill genomes, which range from 12 to 48 Gbp among different species (up to 16 times larger than the human genome) and has therefore not been possible to sequence and assemble (*Jeffrey et al. 2011; Meyer et al. 2015*). One hypothesis about large genome size in krill is that gene duplications have provided new genetic material for natural selection to act on, and produce various novel and adaptive variants and functions (*Coulbone et al. 2011; Rothet et al. 2007*). One such potential mechanism could be increased physiological and developmental plasticity due to a vast a number of unique non-coding transcripts in krill (*Mayer et al. 2015*). Due to the difficulty of assembling full krill genomes, several research groups have instead analysed transcriptome data and studied the expression of different kinds of genes to learn more about environmental adaptation in these species. In one article (*Clark et al. 2011*), they looked at chaperone genes to better understand how krill cope with stress. In a different study (*De Pittà et al. 2013*), scientists analysed circadian clock genes and clock-controlled genes to understand the biological rhythm of krill. Both of these articles used transcriptome data targeting mostly a small subset of candidate genes and functions to understand how krill are adapted to its environment. But what else can full transcriptome datasets tell us about krill evolution? How does the choice of different methods used to study evolution in krill (e.g. for inferring gene/transcript orthology) affect the results? By answering these questions, we will learn more about the evolutionary history of krill, better understand the impact and biases of methods on the evolutionary inferences and find the most reliable methods and data. Transcriptomes are now well-established sources of data in phylogenomic and population genomic analyses, but have not yet been extensively analysed in krill (*Galtier 2016*). My project, therefore, is important to the research community, as it will help identify practical and powerful methods and produce some of the first robust information about molecular evolution, genetic adaptation and phylogenetic interrelationships in krill.

1.4 Potential and limitations to study evolution of krill and similar species

In this project, I have performed comparative genomic analyses of krill in order to increase understanding of krill evolution and adaptation, and help predict how they may adapt to the changing environment in the Antarctic ocean. By using transcriptome data from five krill species and five crustacean outgroups, spanning many thousands of genes, I created molecular phylogenies that shows the relationship between krill and other crustacean outgroups. Establishing orthology and correctly cluster genes from different species for comparative analyses is a central task in phylogenomics. However, because genome sequences are not available for any krill, it is difficult to know with certainty which transcripts are orthologues (derived from the same ancestral gene in all species) and suitable for phylogenetic analyses, or paralogues (derived from gene duplications that may be species-specific) that could potentially mislead inferences if mistakenly included (*Koonin 2005*).

2 Aim

2.1 Phylogenomic framework for krill

The aim of this thesis is to produce the first phylogenomic framework for krill and determine the extent to which genes in the Antarctic krill *E. superba* evolve due to positive selection on adaptive mutations. Such variants may help krill adapt to its environment. To this end, I applied powerful comparative bioinformatics and phylogenomic methods to transcriptome data to first produce a well-supported molecular phylogeny of krill. The data spans five different krill species and five outgroup species. The outgroups are model-crustaceans, for which the genomes have already been sequenced. To be able to create a reliable molecular phylogeny, suitable orthology groups (OGs) were identified and selected from all the transcripts and genes from the ten species. Two pipelines using different Orthology Assessment Toolkits (OATs) were compared in order to establish and compare orthology relationships between transcripts/genes. Pipeline 1 used the software Orthograph (*Petersen et al. 2017*), which uses clustering for identifying transcript orthologues between species and for Pipeline 2 used UPhO (*Ballesteros et al. 2017*) that use phylogenetic methods in addition to clustering. The Pipeline that retained the most data and produces well-supported trees was considered to be the main Pipeline for generating a robust molecular phylogeny (or species-tree) of the krill. These analyses have allowed me to infer relationships between species and resolve ancient events in krill evolution.

2.2 Investigation of adaptive evolution on krill

The second aim of this thesis was to perform an extended population genomic study where recent selective forces that influence gene evolution in the Antarctic krill *E. superba* has been assessed and analysed. Within-species polymorphism data (SNPs) in *E. superba* and divergence data between *E. superba* and *E. crystallorophias* for orthologous genes was used to estimate the degree of recent neutral and positive selection on coding sequences in *E. superba*. This will lay the foundation for studying how selection acts on protein-coding sequences in krill and help us understand how krill are adapted to different environments and may respond to climate change at the genetic level.

3 Theory

3.1 The importance of species evolutionary history

If you have sequenced the whole genomes or transcriptomes of different taxonomically distant species, you can reconstruct the evolutionary history of each gene in species with the help of phylogenomics by comparing sequences within and between species (*Koonin 2005*). But why is the evolutionary history of species so intriguing? For krill, this information will help us to better understand how krill have diversified, adapted and evolved in response to the environmental changes and selection pressures that they have been exposed to in the past. This will help us better understand how krill can adapt to changing environments also in the future.

3.2 Significance of orthology and paralogy

To understand and work with phylogenomics it is essential to distinguish between orthologous genes and paralogous genes. Genes that are orthologous are derived from speciation events while those that are paralogous are a result of gene duplication (*Koonin 2005).* Gradual changes in orthologous sequences and the addition or deletion of paralogues are believed to be among the primary events that cause the genomes to evolve differently between species (*Heijden 2007*). Considering this statement, one should be careful clustering all the genes when trying to construct robust and reliable phylogenetic trees in order to get information about a species evolutionary history (*Zmesek et al. 2001*).

For this particular study, the aim has been to group all genes that are orthologous together, since these are comparatively reliable predictors of interrelationships following speciation. Genes that are derived from gene duplication (paralogous genes), on the other hand, are less reliable for tracing speciation events and may also have undergone rapid sequence evolution and functional diversification that could mislead phylogenetic inferences (*Gabaldón et al. 2013*).

3.2.1 The occurrence of paralogous genes

New functions within species can be created with help of paralogous genes. During evolution, when gene duplication occurs, some of these paralogous genes manage to escape extinction or removal due to purifying selection, and later evolve and create a new function within that specific species (*Koonin 2005*).

3.3 Issues with grouping

Due to genome complexity (i.e. the size, structure and repetitive nature of genomes) and incongruent histories of individual genes, inferring the evolutionary history of a group of species is not easy. The difficulties here are caused by issues such as gene loss, horizontal gene transfer, gene rearrangements and divergent evolutionary rates or sequence composition between species (*Koonin 2005*). This makes it difficult to group orthologous genes together while avoiding including paralogous genes.

3.3.1 Way to solve the problem with grouping

To deal with this problem and find the right orthologous genes, there exist different kinds of orthology prediction methods. One is based on the sequence similarity and is called graph based (*Trachana et al. 2011*). It calculates similarity during genome comparison of the species and groups those with the highest score to OG (orthology group). The second method, besides clustering, are also using tree reconciliation where it maps all gene trees together with a species tree, this is called a tree-based method (*Heijden 2007*). One of the main differences between these two approaches is that the tree-based method is computationally demanding compared with the graph-based method (*Trachana et al. 2011*).

The issue with the graph-based method is the importance of high coverage in the genomes. Genomes that are compared with each other, should be reasonably similar to each other in order to group right OGs. This can be achieved by choosing the closest relative as an outgroup (*Koonin 2005*). Moreover, varying evolutionary rates or gene losses following gene duplication can also make it difficult to correctly create OGs. This problem is less severing in tree-based methods. The problem here is instead the choice of the tree root, which requires choosing the correct outgroup. This is difficult when you have a large set of data since it's not always the case that the outgroup is present in all the gene families, which means that these gene family trees will not have the correct root that is important for this approach (*Heijden 2007*). In this project, multiple outgroups were used to deal with this issue.

3.4 Transcriptome data

3.4.1 Advantages and disadvantages of transcriptome data

Because krill are non-model organisms with a large genome size, we still do not have a reference genome in order to better understand and get representative example about the set of genes in krill. Since we lack a reference genome, transcriptome data is preferable as it is cost effective, and does not dependent on a reference genome. Things to consider when using transcriptome data is that you risk missing some of the genes since not all genes are expressed simultaneously. To tackle this problem and get transcriptomes with the highest coverage, you should collect samples from several tissues. Another issue is that the risk is higher to mistakenly identify false positive orthologous transcripts when gene duplication is present in the organism (*Wen et al. 2013*).

3.5 Adaptive evolution

3.5.1 The cause of adaptation in species

Environmental changes such as global warming, have affected many organisms negatively. Some affected species may escape their habitats and migrate to more favourable environments. The changes to the distribution and abundance of species due to climate change have been particularly drastic in the oceans (*Poloczanska et al. 2013*). However, not all species have the ability to escape these changes. Instead, they may have to quickly adapt at the genetic level to the changes in order to survive (*Peck et al. 2014*), which requires adaptive variants to be present in populations, or appear before they go extinct. Adaptive evolution studies dive into the depths of the organism to analyse these genetic adjustments. By looking at genetic diversity researchers can find out how krill are adapting to the changing environment.

3.5.2 The way to discover adaptation in species

With help of different approaches, using population genetic theory, it is possible to estimate how efficiently natural selection operates in a species (*Booker et al. 2017*). This can be done by analysing adaptive molecular evolution., from studying the ratio of non-synonymous mutations (those that change amino-acids in proteins; so-called N mutations) and synonymous mutations (those that do not change the amino acids; S) (*Papot et al. 2016*). McDonald-Kreitman test uses this principle. It compares nucleotide divergence (fixed differences; D) between two species against nucleotide diversity (polymorphisms; P) within species. When there is extensive adaptive evolution in a species, the ratio of D_N/D_S should be greater than the ratio of P_N/P_S. Because adaptive non-synonymous variants are expected to quickly reach fixation, and non-adaptive variants to be selected against and removed from populations, these variants are expected to contribute less to the species' polymorphism than what selectively neutral synonymous variants do, which decreases the value of P_N/P_S. Instead, the ratio of D_N/D_S is expected to be high, since the fixation probability in a species is greater than neutral evolution, which indicates adaptive evolution (*Galtier 2016*). In the case of neutrality, the ratio of D_N/D_S and P_N/P_S would be equal since both of synonymous mutation and non-synonymous mutations would evolve neutrally.

3.5.3 Estimation of the nucleotide substitutions fraction

The ratio of D_N/D_S and P_N/P_S can be used in order to estimate the fraction of non-synonymous variants that has been driven to fixation by positive selection (α; see equation 1). α values close to one, indicate very strong positive selection while values close to zero are a sign of neutral evolution (*Haller et al. 2017*). Different species have different α values, since the efficacy of selection differs between species. It is typically believed that species with large effective population size (N_e) have high α, while species with relatively small N_e, (for example humans) have low α. This is because populations with large N_e are less affected of genetic drift, which increases the chance that rare beneficial mutations become fixed. Krill have enormous population sizes and combined biomass (*Siegel 2016*), and are assumed to have high N_e. However, the effective population size has not been estimated in any krill so far, and the efficacy of positive selection (e.g. α) is unknown.

$$\alpha = 1 - \frac{D_S \times P_N}{D_N \times P_S} \quad (1)$$

4 Metod and implementation

4.1 Data

The transcript data was taken from 5 different krill species with different length. These are, 1) *Euphausia superba* (n=133,965 transcripts) (*Sales et al. 2017*), 2) *E. crystallorophias* (n=42,362) (*Toullec et al. 2013*), 3) *Meganyctiphanes norvegica* (n=405,497) (*Blanco-Bercial et al. 2017*), 4) *Thysanoessa inermis* (n=340,890) (*Huenerlage et al. 2016*) and 5) *T. raschii* (n=161,028). These five transcripts of krill are the only ones that are available in the present. The first four transcripts were taken from different published studies while the fifth was shared with the project supervisor by krill researcher Jean-Yves Toullec at Station Biologique de Roscoff, France.

To infer OGs, five different crustaceans was used as outgroups since those species whole genome has already been sequenced. Those species were 1) *Daphnia pulex*; NCBI txid=6669; n=30,590 genes, 2) *D. magna*; txid=35525; n=29,127 genes), 3) *Lepeophtheirus salmonis*; txid=72036; n=13,844, 4) *Eurytemora affinis*; txid=88015; n=29,783 and 5) *Hyalella azteca*; txid=294128; n=12,906. All these genes were downloaded from OrthoDB Hierarchical Catalog of Orthologs.

The transcriptomes for analysing adaptation in the Antarctic krill had been produced from 48 specimens of *E. superba* collected from five different Antarctic stocks. The data was kindly provided by Prof Bettina Meyer at the Alfred Wegener Institute, Germany.

4.2 OAT execution and Post OG-assessment processing

Before aligning sequence data, we have to find all the relationships (orthologous, not paralogous) between our genes from our species in order to get a robust phylogenetic tree. By using different Orthology Assessment Toolkits (OATs), it is possible to collect and cluster putatively orthologous genes that are similar in our different species, based on different criteria. Here, I have developed two Pipelines and used and evaluated two different OATs: 1) Orthograph, which is suitable for a large dataset (*Nichio et al. 2017*); and 2) UPhO. Both are said to be relatively tolerant towards missing sequences and appropriate to use for clustering transcriptome data. The output from Orthograph was available at the project start and my task was to evaluate it in detail. For the OAT in Pipeline 2, I run UPhO myself and compared the results from all two Pipelines in subsequent steps.

4.2.1 Orthograph

Orthograph is a graph-based approach. It uses pairwise similarity between the data (transcripts) and reference sequences in order to find the right cluster of orthologous genes. Reference sequences can be found from different databases such as OrthoDB, eggNOG, OrthoMCL etc. (*Trachana et al. 2011*). In our case, OrthoDB was used. Reference proteins (crustacean genes in our case) were used as sequence templates to infer orthologous groups (OGs). To be able to search for candidate orthology (transcript sequences) in the targeted compound library, profile Hidden Markov Model (pHMMs) were first created with multiple sequence alignments (MSA) of precomputed OGs. This profile is later used to BLAST search to get information about candidate orthology (*Petersen et al. 2017*). Only a single best transcript was kept per krill species and OG.

4.2.2 Unrooted Phylogenetic Orthology (UPhO)

UPhO is an unrooted tree-based approach. This approach does not require precomputed OGs to be able to produce orthologies for the species. It starts with grouping sequences into homologous genes (gene families) with help of explicit sequence similarity threshold such as BLAST. In order to do this, it uses All versus All BLAST to create a database containing pairwise blast scores. In the next step, files containing homologous clustered sequences are aligned, regions containing gaps are trimmed and cleaned, in order to use it for phylogenetic inference, which is performed by RAxML. The final step is orthology assessment where trees containing gene families are evaluated to find the right orthologous groups (*Ballesteros et al. 2017*).

4.3 Polishing and evaluation of alignments

To get good quality and comparable data, the output files with orthologues from both OATs containing orthologous sequences were re-aligned. Afterwards, the alignment was to polish all output files with help of trimming program. This was done to get rid of unreliable sites in all files. Lastly, quality scores were calculated to estimate the quality of the alignment.

4.3.1 MAFFT (Alignment)

MAFFT is a multiple sequence alignment (MSA) and can compare many homologous sequences simultaneously. This approach starts with calculating the pairwise distance between species in order to create distance matrix for construction of the guide tree with help of clustering method called Unweighted Pair Group Method with Arithmetic Mean (UPGMA) (*Katoh et al. 2008*). Iteration parameter was chosen to perform 1000 iteration, where it calculates the distance matrixes for each iteration. To be able to run this analysis for all the gene files, a bash script was created.

4.3.2 Gblocks (Trimming tool)

One of the disruptions in the phylogenetic analysis is caused by gene regions that are not evolving continuously. This can cause noisy and poorly-aligned regions with gaps in the alignment file. In order to deal with this issue, trimming programs are used. The purpose of Gblocks is to scan the alignment and only maintain regions that are relatively conserved. It begins with evaluating and classifying blocks that are conserved and blocks that is not. Those that lack a certain level of conservation and contains gaps are rejected since they are ambiguous for the phylogenetic analysis (*Castresana 2000*). Gblocks was running with the sequence type chosen to protein and allowed to retain half of the gap positions. A bash script was created to run this analysis for all the gene files.

4.3.3 T-Coffee (Quality estimation of the alignment)

When alignment and trimming of the data/genes was completed, I estimated alignment qualities with the aim to identify genes that were well-aligned or poorly aligned. T-Coffee (*Notredame et al. 2000*), was used to calculate this alignment score. T-Coffee start with generating libraries containing pairwise alignment with the source from global and local alignment, which will be used as help file to guide MSA. It is precisely this that is beneficial with T-Coffee, the ability to combine global and local alignment. No specific parameter was chosen for the T-Coffee run but a bash script was created in order to run this analysis for all the gene files.

4.4 Polishing of the data

All OGs/alignment files were not equally good: some did not contain sequences from all ten species and others were poorly aligned overall. Before I could continue to create my gene- and the species tree, I wrote a Perl script to select those genes that had alignment quality 80 % and above. After that step, the genes that also contained eight to ten species and more than 80% of their sequence after trimming with Gblocks were kept.

Since data from Pipeline 1 could sometimes contain multiple sequences from some of the outgroups, a subtree function in the program TreeKO was used to remove those extra sequences. This function prunes paralogues scattered across the tree, while trying to retain the orthologues that are consistent with the species tree.

By using Figtree, redundancy was detected also in gene trees generated from UPhO. This problem was managed with a script in pearl where paralogues sequences was removed. Since the redundant data was in-paralogs (duplicates at the tips of the trees), this issue could be used without using a tool in TreeKO.

4.5 Phylogenetic inference

After polishing, all the gene trees were created (one per orthology group alignment) before the concatenation and were used to create the species tree.

4.5.1 RAxML (Making of maximum likelihood (ML) gene- and specie trees)

To construct gene and species tree RAxML was used since it is a popular program that can handle large dataset and construct phylogenetic trees using maximum likelihood (*Stamatakis 2014*). The parameters that were chosen to create gene trees were –p 12345 that help to debug the program by using a seed that generates reproducible random numbers for the parsimony inferences[1]. The second parameter was –m PROTGAMMAUTO that choose a substitution model, specific for amino acid sequences. With help of AUTO, the best substitution matrix/model with the best score was chosen automatically by the software. The gamma parameter was used to incorporate among-site rate variations.

To construct the species tree, all gene alignments were concatenated to a big file and used as an input to create the species tree. Three additional parameters were used for the creation of species tree beside –p and –m. –f a that allows us to select an algorithm for rapid bootstrapping. This search ML tree with the best score. –x 12345 that work as –p but is for the rapid bootstrapping case. And lastly -#100, which allow us to choose the number of replicates/run, in this case, 100 replicates.

4.6 Concatenate and removal of additional sequences

4.6.1 TreeKO (Reconstruction of pruned trees, st)

Orthograph generated duplicates in the data, which has to be removed before the creation of the final gene- and species trees. This can be done with help of a software called TreeKO. Those gene trees that contain redundant data are selected and nodes including duplicates are used as a “cutting” points for the pruning procedure. Gene trees with nodes that contain duplicates are splitting and pruning until only one copy of each species in the tree (subtree) remains (*Marcet-Houben et al. 2011*).

4.7 Tree-evaluation

In this step, species trees and gene trees from both pipelines were compared with each other in order to estimate the most accurate tree and pipeline. Here, two tree-metrics were used to

[1] https://sco.h-its.org/exelixis/resource/download/NewManual.pdf 6/5/18

evaluate the distance between gene and species trees. The best species tree and orthology assessments were chosen for the next step.

4.8 Compute distance between gene and species tree

4.8.1 TreeKO (Tree comparison, tc)

With help of tree comparison algorithm, built in TreeKO we could calculate the topological differences between the species tree and all the gene trees. Speciation distance compares all the gene trees against the species tree in order to find similarity between them without taking into account duplications and gene losses. If both trees, which are compared with each other have same evolutionary history, the value of the speciation distance is going to be 0. Since speciation distance is not considering gene losses and duplications, it is important to take in account that trees with speciation distance 0 are not necessarily going to be identical in their composition, but only share the same history of speciation (*Marcet-Houben et al. 2011).*

4.8.2 RF (Robinson–Foulds) distance

When you are comparing phylogenetic trees to each other, you can calculate the dissimilarity between them by comparing clusters of descendant leaves in those trees, this is the RF distance (*Asano et al. 2012*). If two trees are identical the RF distance between those trees are going to be zero. In this thesis, all gene trees were compared with the specie tree in order to calculate the dissimilarity among them. The speciation distance and RF distance are expected to be relatively highly correlated.

4.8.3 Phylome support

Phylome support is defined by looking at the percentage of trees in a complete set of gene phylogenies (phylome) that support a specific topological arrangement defined by two daughter nodes. The advantage of this support compared with bootstrap support is that it takes into consideration arrangements between multiple partitions and not only single partition (*Marcet-Houben et al. 2009*).

4.9 Statistical test

Now when we have distance values from both OATs, it is time to decide which method is most accurate and for which OATs the gene trees are most congruent with the species tree. To establish that a statistical test was used. This test helps us to determine if the data is enough to draw conclusions about precision of the methods.

4.9.1 Boot-ci

Because our data was not normally distributed, I used the non-parametric bootstrap method to test if the distance parameters (speciation distance and RF) were significantly different between the two pipelines. The Boot-ci function in MATLAB was the best chose for the statistical test. For every bootstrap replicate, the average value was calculated. At the end, values are sorted from lowest to highest and the upper/lower 2.5% of the data represents the 95% confidence interval. It computes the 95% confidence interval from all the chosen 2000 samples by sampling with replacement. The output you get is the values of the higher and lower bounds of the confidence interval[2].

4.10 Adaptive evolution in the Antarctic krill E. Superba

For this analysis, I used the original set of OGs (2653 genes) detected in Pipeline 1 but the alignment files covered only *E. superba* and *E. crystallorophias*. An independently produced dataset with Single Nucleotide Polymorphisms (SNPs) (14879 SNPs) for *E. superba* was used (see Data above). The method distinguishes between ancestral and derived alleles at SNPs in *E. superba* by comparing against *E. crystallorophias* (the outgroup for these analyses). Only OGs with at least one SNP where one of the two alleles were the same between the two species were included in this analysis. With this data, I estimated the number of synonymous and non-synonymous alleles segregating at different frequencies and formatted input parameters for each gene to estimate the degree of positive selection in *E. superba* (*Booker et al. 2017*).

4.11 Evaluate polymorphism

4.11.1 AsymptoticMK

This is a web-based tool that helps you to estimate and understand if nucleotide substitution has driven to fixation by positive selection or not, α. In order to run this server, the calculated values of divergence and polymorphism within and between *E. superba* and *E. crystallorophias* was given together with a file (was developed using Perl) containing the estimated frequency of non-synonymous and synonymous substitution ratio. The α values close to one indicted that the substitution in *E. superba* was due to positive selection while dose α values close to zero indicated for neutral evolution (*Haller et al. 2017*).

5 Result

5.1 Analysis of the species tree

A maximum likelihood phylogeny, representing the species tree was created from 1562 orthologues groups and 541484 amino acids, generated from Pipeline 1 (see Figure 1). I validated the interrelationships inferred in this species tree by comparing the tree against the taxonomic hierarchy in the NCBI taxonomy database, which contains phylogenetic lineages

[2] https://www.mathworks.com/help/stats/bootci.html 6/5/18

from different organisms with molecular data. The result was that the species tree indeed coheres with NCBI hierarchy, indicating that groupings are consistent with the currently held hypotheses about crustacean interrelationships.

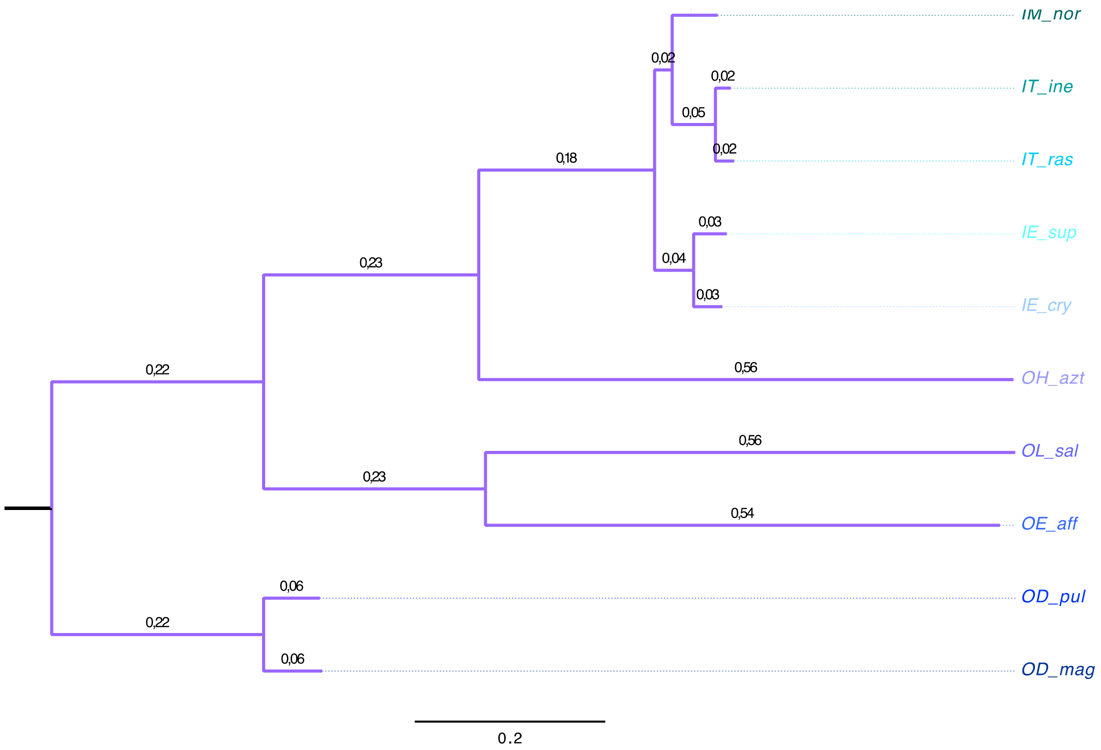


Figure 1: Molecular maximum likelihood phylogeny of five different krill species (light blue), IM_nor (Meganyctiphanes norvegica), IT_ine (Thysanoessa inermis), IT_ras (Thysanoessa raschii), IE_sup (Euphausia superba), IE_cry (Euphausia crystallorophias) and five different crustacean species (dark blue), OH_azt (Hyalella azteca), OL_sal (Lepeophtheirus salmonis), OE_alf (Eurytemora affinis), OD_pul (Daphnia pulex) and OD_mag (Daphnia magna). Numbers indicate the branch length.

All nodes were 100% supported by the bootstrap analysis, but was unclear how many genes supported each node in the species tree. To increase our confidence about the interrelationships inferred in the species tree, the species tree was redrawn with phylome support for the nodes. Phylome support values show us a measure of how well a certain combination is supported in the species tree, see Figure 2 for Pipeline 1 and Figure 3 for Pipline2. Branches containing value one show us those species whose combination of gene tree supports their grouping with 100%.

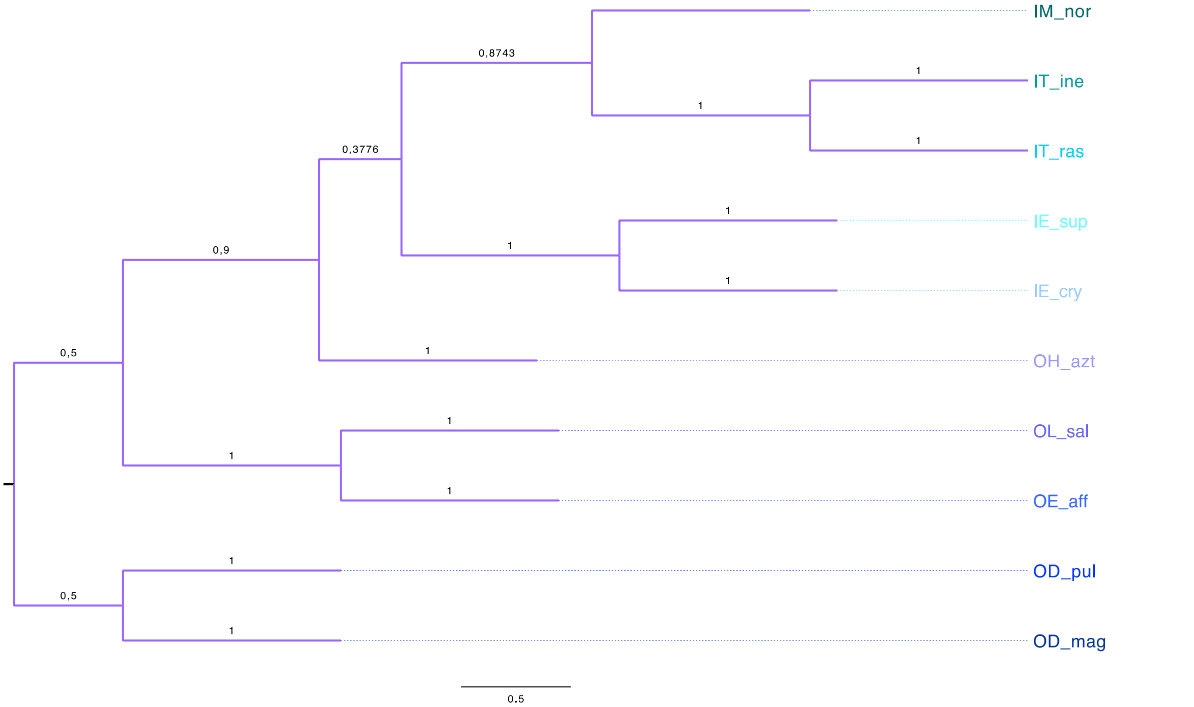


Figure 2: Molecular maximum likelihood phylogeny of five different krill species (light blue) and five different crustacean species (dark blue). Branch values on the tree shows us the phylome support for genes generated with Orthograph.

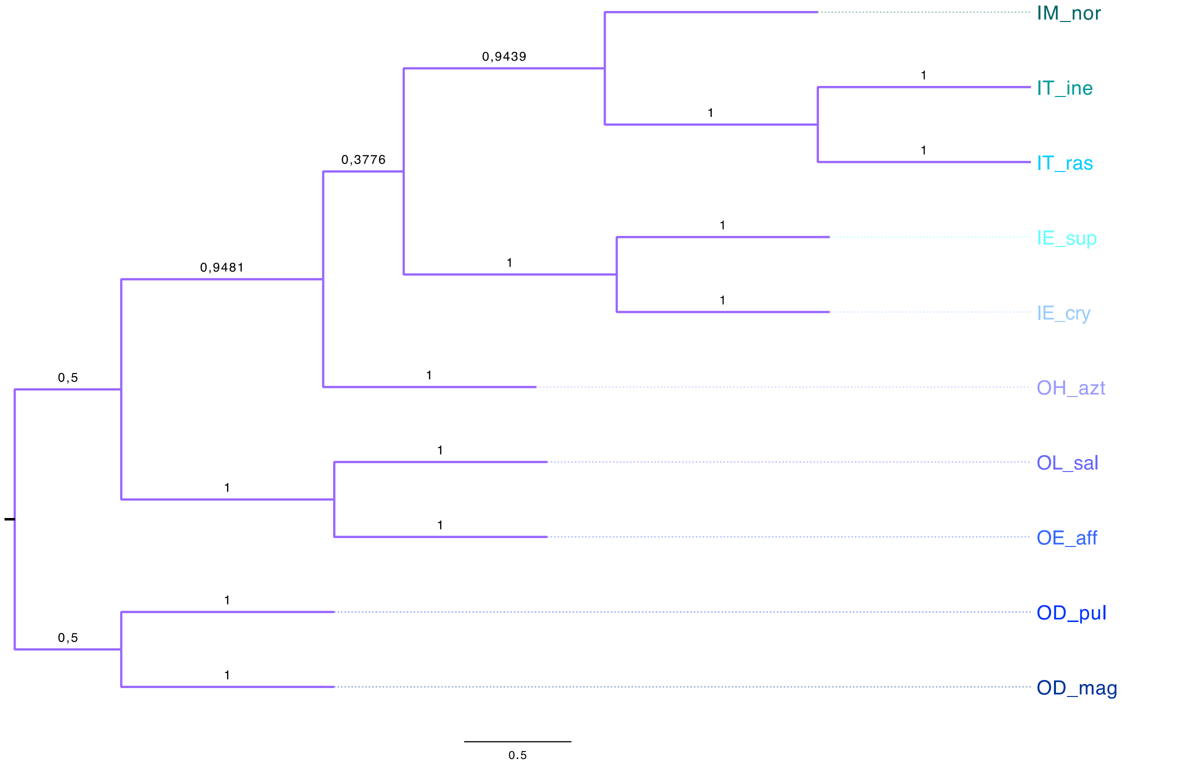
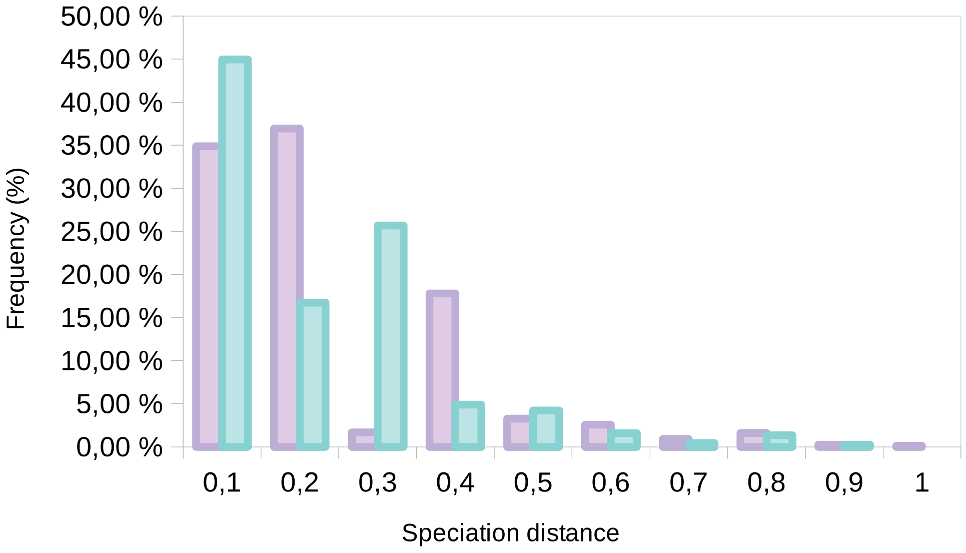


Figure 3: Molecular maximum likelihood phylogeny of five different krill species (light blue) and five different crustacean species (dark blue). Branch values on the tree shows us the phylome support for genes generated with UPhO.

To test how well-supported the species tree are, I used the tree comparison (tc) algorithm in TreeKO and Robinson-Foulds symmetric difference in ete3 toolkit. These tools compared all the 1562 gene trees from Pipeline 1 (Orthograph) and 903 gene trees from Pipeline 2 (UPhO) against the species tree. It calculates the topological distance and clade differences that separate a gene tree from the species tree. A small distance indicates that a gene tree is similar to the species tree and that the gene sequence has an evolutionary history that matches the order of speciation suggested by the species tree. If most gene trees have short distances, the species tree can be considered well-supported by the data and reliable. To visualize the distribution of these two distances, each pipeline was plotted as a histogram. Figure 4 shows the full distributing of speciation and RF distances for both Pipelines.



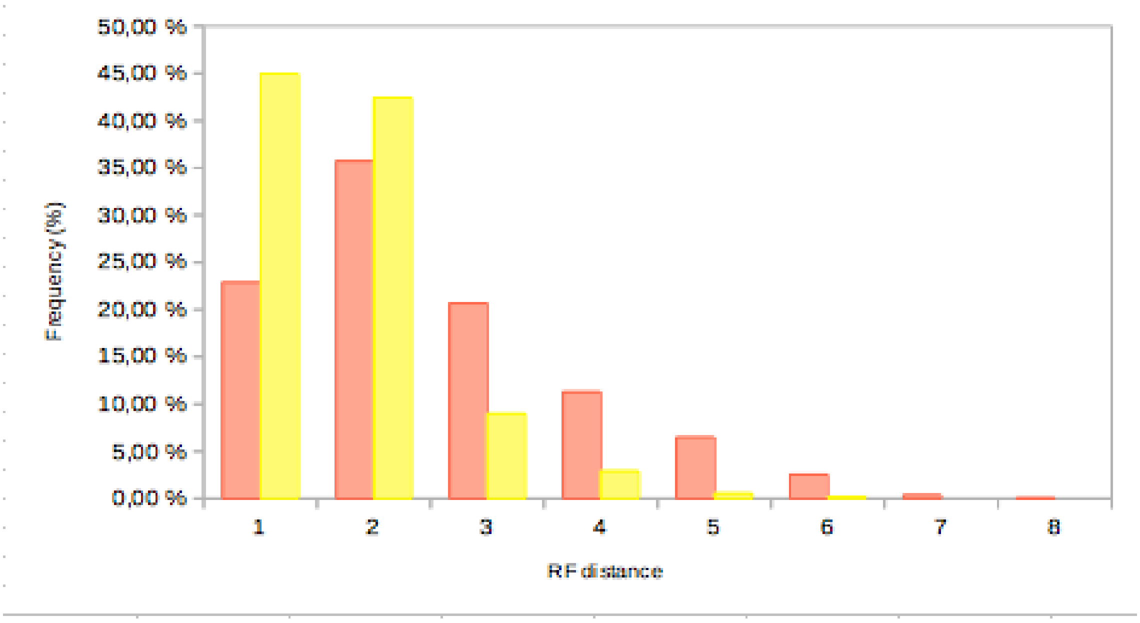


Figure 4: Histogram for Pipline 1 and Pipline 2. The collection of the distribution of bins with different frequency values of the speciation- and RF distances both for pipeline 1 (purple, orange) and for pipeline 2 (light blue, yellow).

An overview of the two orthology Pipelines is available in Table 1. The number of remaining genes and amino acids positions was calculated before and after trimming to make a conclusion about which method preserved the highest number of genes and amino acid positions.

Table 1: Containing values generated before and after trimming with Gblocks

	PIPELINE 1	**PIPELINE 2**
Nr. of genes from start	*3010*	*1942*
Nr. of genes at end	*1562*	*903*
Nr. of aa before trim	*964237*	*1150992*
Nr. of aa after trim	*541484*	*368959*
% of remaining aa	*56*	*32*

To be able to evaluate which of these two methods has the minimum mean of speciation distance, see Figure 5 and RF distance, see Figure 6, I used bootstrapping to randomly resample the data 2.000 times and generate 95% confidence intervals. The calculated values of the mean for the speciation distance was 0.1904 for pipeline 1 and 0,1607 for the pipeline 2. The calculated mean for the RF distance was 2.0384 for pipeline 1 and 1.4463 for the pipeline 2. Since the lower bound for the Pipeline 1(Speciation distance, 0.1812; RF distance, 1.936) and upper bound for the Pipeline 2 (Speciation distance, 0.1731; RF distance, 1.5532) are not joined, the statistics are significantly different.

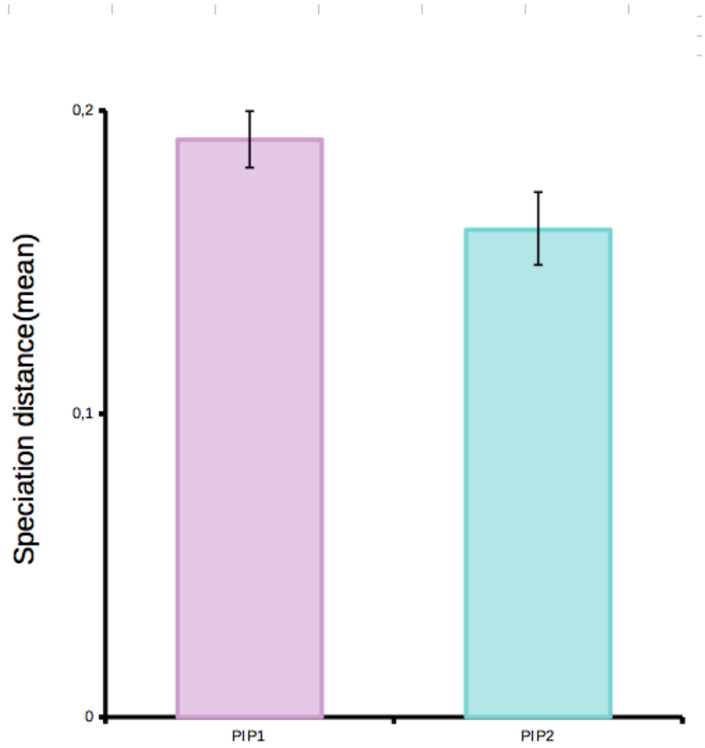


Figure 5: Histogram with confidence interval calculated with boot-ci statistical test, containing mean values of the speciation distance for the pipeline 1(purple) and for pipeline 2(light blue).

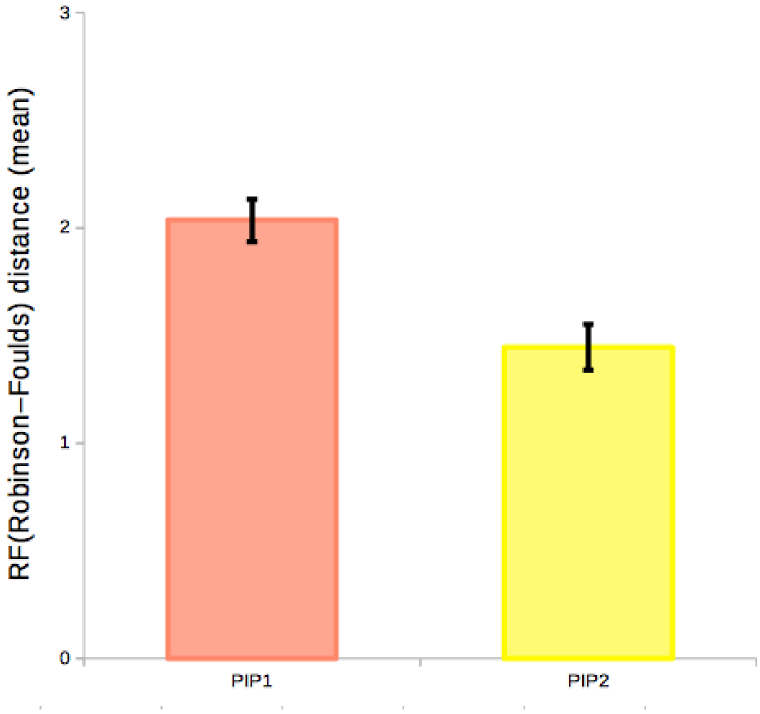


Figure 6: Histogram with confidence interval calculated with boot-ci statistical test, containing mean values of the RF distance for the pipeline 1(orange) and for pipeline 2(yellow).

5.2 Importance of distance calculation

I next tested for a correlation between speciation distances and the proportion of amino acid positions removed from gene alignments during trimming with Gblocks. The aim was to learn if genes with noisy alignments tended to have higher speciation distances. Such alignments could potentially indicate poor automated selection of orthologous sequences that do not represent the evolutionary history of the species, and could perhaps be identified already at an early step in the pipeline. The R^2 was calculated using a linear regression model in MATLAB, to get information about how strongly the parameters were correlated. We can see from Figure 7 that most of the data points are widely spread in the correlation graph and therefore the R-square gives us value near zero ($R^2 = 0.0042$, or 0.42%). This indicates that the number of removed positions does not help us to predict the fit of our gene trees against the species tree and further studies need to be done such as calculation of speciation- and RF distances.

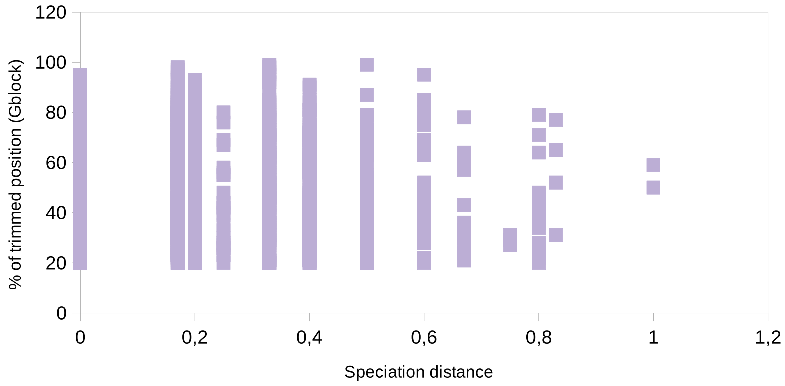


Figure 7: Correlation graph with data (purple) from 1,562 genes. x-axis is speciation distance data and y-axis is percentage of sequences that was trimmed with Gblocks. The large distribution indicates that no or little correlation between the proportion of trimmed sites and the speciation distance.

5.3 Analysis of positive selection in *E. superba*

The analysis of adaptation through positive selection on proteins in the Antarctic krill spanned 2653 genes and 14879 SNPs. Estimated frequencies of synonymous and non-synonymous alleles, which I generated with help of a Perl script with help of data given from my supervisor, were used in asymptoticMK as an input together with dN (non-synonymous substitutions between species) = 69270.7 and dS (synonymous substitutions between species) = 162671.2, which had been inferred by the supervisor using PAML. The cut-off values for allele frequencies that the web-server asymptoticMK required for the data were set to 0.1 and 0.9. The output graph received from asymptoticMK, see Figure 8, showing the distribution of the allele frequency α (x). The dotted grey line in the graph shows the level of selection in our data, which was $\alpha = 0.69534$. Whit this α value the level of selection in *E. superba* was determined.

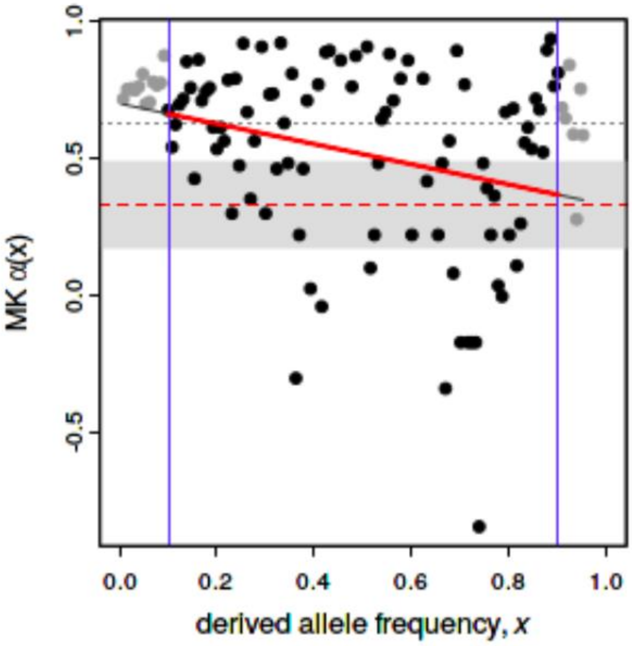


Figure 8: Graph received from asymptoticMK web-server. The black points show the normalized binned polymorphism frequencies, blue lines show the chosen cut-off values. Red line is the best fit with regard to cut off value. Grey dotted lines are the estimated α value and red dotted line show us the asymptotic fit.

6 Discussion

A species tree of krill has been produced and analysed using different phylogenomic methods and techniques. I can now with confidence talk about its robustness. One of these methods was to compare my tree with the taxonomic hierarchy in the NCBI taxonomy database. This indicated that grouping in my species trees indeed was consistent with the currently held hypotheses about crustacean interrelationships, and from the tree that I have created, see Figure 1, we can see that krill genera is monophyletic (descended from a single ancestor). The species tree was also compared with a study done with ribosomal data in order to construct the maximum likelihood tree (*Jarman 2001*). Even in this case, grouping in my species tree matched with the tree constructed performed with ribosomal data. In addition, an analysis was performed with help of a tool in TreeKO that calculates the phylome support in the species tree, see Figure 2 and Figure 3. I could see that all phylome support values was 1 in those branches where latest speciation event has happened, due to no possible alternative groupings.

This is the reason why phylome support values should be analysed for the internal nodes only. Since for both trees, the phylome support values overall had good support indicated that high percentage of the gene trees matched the species tree for that specific combination and increased the theory about the robustness of the way species are grouped in the species tree.

Some of the branches earlier in the evolution have phylome support values below one, even as low as 0.3586. It may be due to that construction of phylogenetic trees for lineages that have large variation in their evolutionary rate, which may cause LBA (long branch attraction) (*Felsenstein 1978*) or be due to ILS (incomplete lineage-sorting) of alleles. *Meganyctiphanes norvegica* may be the reason of instability in the species tree since it jumps around more than the other species. This may be due to the fact that other krill species belong to two common families. *Thysanoessa inermis* and *Thysanoessa raschii* belong to the Thysanoessa family while *E. superba* and *Euphausia crystallorophias* belong to the Euphausia family, causing similarity in sequences composition. While *M. norvegica* belongs to the Meganyctiphanes family and has a more "deviating" sequence composition compared with the other species.

My task was not only to create a robust phylogenetic tree but also evaluate which of the OATs produced the highest amount of orthology groups and created gene trees similar to the species tree. To estimate this, the tool in ete3 toolkit and TreeKO was used in order to calculate the distance between all gene trees compared against the species tree. The distribution of specific distance measures suggested that UPhO gave me distributions leaning towards smaller distance values (Figures 4). In order to strengthen my conclusion a statistical test was performed, see Figure 5 and Figure 6. The result of the statistical test showed us that the mean of both speciation and RF distance were lower for genes generated from UPhO. This result supports the conclusion that gene trees generated with UPhO, give us a well support species tree since they have shorter distance compared with gene trees given by Orthograph. The confidence intervals given from the statistical test showed that these two methods are indeed significantly separated from each other, meaning that they should give us a result different from each other. This interpretation could be done since the lower boundary from the confidence interval from Pipeline 1 did not overlap with it higher interval generated from Pipeline 2 distance values.

By comparing genetic variation in *E. superba* to divergence against *E. crystallorophias*, I estimated the proportion of amino-acid substitutions that had been fixed in *E. superba* due to natural selection (α), rather than neutral genetic drift. The α value was estimated to be 0.69 using the method implemented in asymptoticMK. This value is much higher than for example in mammals where the estimated α values is 0.38 but similar to other arthropods α value (α = 0.65; Galtier 2016). I could see that even in this study the α of these species are relatively high, which gives little confidence for the result. Since krill have a large biomass and species with large N_e are expected to have higher α, this value is believed to be in line with the current hypothesis. However, one should take into account that this value changes with different cut-off values.

7 Conclusion

I have here generated and evaluated the first genome-scale orthology-sets and phylogeny of krill. The different methods used here to analyse my species tree, suggests that it is indeed robust. However, it does only include five krill species, which we had data from. In the future when more transcriptome data are available for different krill species, the topology of my species tree may change and hopefully gives us a much better picture of krill evolution.

Since we did not get correlation for the data (Figure 7), it is important to use a proper phylogenetic test (e.g. speciation distance or RF) to see how well a single gene-alignment can represent the species tree.

Since the two different methods that predicted our OGs gave different level of support of the species tree, it indicates that in practice it is important consider the methods carefully for studying evolution of krill. In our case UPhO gave more reliable result, because the gene tress gave us much better support for the species tree, compared with Orthograph. 17.7% better support according to speciation distance data and 40.9% according to RF distance data. The only problem occurring in UPhO was the in-paralogs but with help of a Perl script, it was relatively easy to get read of this redundancy. The sequence-redundancy caused by Orthograph was more problematic, and required a lot of time and work in order to fix. In the end, the better performance by UPhO was actually expected since it is a tree-based approach that not only uses sequence comparison but also tree reconciliation. When we get access to more transcriptome data of krill in order to learn more about the evolutionary history of these species, UPhO are recommended as the chosen OATs because of its impact and biases on the evolutionary inferences.

As for the SNP data taken from *E. superba*, although the generated α value seems to match the current hypotheses that species with large Ne tend to have higher alpha values, I would still recommend performing this analysis with significantly more genes and SNPs for more robust results and uses sophisticated methods.

8 References

Asano, Tetsuo, Jesper Jansson, Kunihiko Sadakane, Ryuhei Uehara, och Gabriel Valiente. ”Faster Computation of the Robinson–Foulds Distance between Phylogenetic Networks”. *Information Sciences* 197 (augusti 2012): 77–90. https://doi.org/10.1016/j.ins.2012.01.038.

Atkinson, Angus, Rachael S. Shreeve, Andrew G. Hirst, Peter Rothery, Geraint A. Tarling, David W. Pond, Rebecca E. Korb, Eugene J. Murphy, och Jonathon L. Watkins. ”Natural Growth Rates in Antarctic Krill (Euphausia Superba): II. Predictive Models Based on Food, Temperature, Body Length, Sex, and Maturity Stage”. *Limnology and Oceanography* 51, nr 2 (u.å.): 973–87. https://doi.org/10.4319/lo.2006.51.2.0973.

Ballesteros, Jesús A., och Gustavo Hormiga. ”A New Orthology Assessment Method for Phylogenomic Data: Unrooted Phylogenetic Orthology”. *Molecular Biology and Evolution* 33, nr 8 (2016): 2117–34. https://doi.org/10.1093/molbev/msw069.

Blanco-Bercial, Leocadio, och Amy E. Maas. ”A Transcriptomic Resource for the Northern Krill Meganyctiphanes Norvegica Based on a Short-Term Temperature Exposure Experiment”. *Marine Genomics* 38 (april 2018): 25–32. https://doi.org/10.1016/j.margen.2017.05.013.

Booker, Tom R., Benjamin C. Jackson, och Peter D. Keightley. ”Detecting positive selection in the genome”. *BMC Biology* 15 (30 oktober 2017): 98. https://doi.org/10.1186/s12915-017-0434-y.

Castresana, J. ”Selection of Conserved Blocks from Multiple Alignments for Their Use in Phylogenetic Analysis”. *Molecular Biology and Evolution* 17, nr 4 (april 2000): 540–52. https://doi.org/10.1093/oxfordjournals.molbev.a026334.

Clark, Melody S., Michael A. S. Thorne, Jean-Yves Toullec, Yan Meng, Le Luo Guan, Lloyd S. Peck, och Stephen Moore. ”Antarctic Krill 454 Pyrosequencing Reveals Chaperone and Stress Transcriptome”. *PloS One* 6, nr 1 (06 januari 2011): e15919. https://doi.org/10.1371/journal.pone.0015919.

Colbourne, John K., Michael E. Pfrender, Donald Gilbert, W. Kelley Thomas, Abraham Tucker, Todd H. Oakley, Shinichi Tokishita, m.fl. ”The Ecoresponsive Genome of Daphnia Pulex”. *Science (New York, N.Y.)* 331, nr 6017 (04 februari 2011): 555–61. https://doi.org/10.1126/science.1197761.

De Pittà, Cristiano, Alberto Biscontin, Alessandro Albiero, Gabriele Sales, Caterina Millino, Gabriella M. Mazzotta, Cristiano Bertolucci, och Rodolfo Costa. ”The Antarctic Krill Euphausia Superba Shows Diurnal Cycles of Transcription under Natural Conditions”. Redigerad av Nicholas S. Foulkes. *PLoS ONE* 8, nr 7 (17 juli 2013): e68652. https://doi.org/10.1371/journal.pone.0068652.

Emms, David M., och Steven Kelly. ”OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy”. *Genome Biology* 16 (06 augusti 2015): 157. https://doi.org/10.1186/s13059-015-0721-2.

Felsenstein, Joseph. ”Cases in which Parsimony or Compatibility Methods Will be Positively Misleading”. *Systematic Zoology* 27, nr 4 (1978): 401–10. https://doi.org/10.2307/2412923.

Gabaldón, Toni, och Eugene V. Koonin. ”Functional and Evolutionary Implications of Gene Orthology”. *Nature Reviews Genetics* 14, nr 5 (maj 2013): 360–66. https://doi.org/10.1038/nrg3456.

Galtier, Nicolas. ”Adaptive Protein Evolution in Animals and the Effective Population Size Hypothesis”. *PLOS Genetics* 12, nr 1 (januari 2016): e1005774. https://doi.org/10.1371/journal.pgen.1005774.

Haller, Benjamin C., och Philipp W. Messer. ”AsymptoticMK: A Web-Based Tool for the Asymptotic McDonald–Kreitman Test”. *G3&amp;#58; Genes|Genomes|Genetics* 7, nr 5 (maj 2017): 1569–75. https://doi.org/10.1534/g3.117.039693.

Heijden, René TJM van der, Berend Snel, Vera van Noort, och Martijn A. Huynen. ”Orthology prediction at scalable resolution by phylogenetic tree analysis”. *BMC Bioinformatics* 8 (08 mars 2007): 83. https://doi.org/10.1186/1471-2105-8-83.

Huenerlage, Kim, Kévin Cascella, Erwan Corre, Lola Toomey, Chi-Ying Lee, Friedrich Buchholz, och Jean-Yves Toullec. ”Responses of the arcto-boreal krill species Thysanoessa inermis to variations in water temperature: coupling Hsp70 isoform expressions with metabolism”. *Cell Stress & Chaperones* 21, nr 6 (november 2016): 969–81. https://doi.org/10.1007/s12192-016-0720-6.

Jarman, S N. ”The Evolutionary History of Krill Inferred from Nuclear Large Subunit RDNA Sequence Analysis”, nr. 14 (november 2001) .

Jeffery, Nicholas W. ”The First Genome Size Estimates for Six Species of Krill (Malacostraca, Euphausiidae): Large Genomes at the North and South Poles”. *Polar Biology* 35, nr 6 (01 juni 2012): 959–62. https://doi.org/10.1007/s00300-011-1137-4.

Katoh, Kazutaka, och Hiroyuki Toh. ”Recent Developments in the MAFFT Multiple Sequence Alignment Program”. *Briefings in Bioinformatics* 9, nr 4 (juli 2008): 286–98. https://doi.org/10.1093/bib/bbn013.

Kawaguchi, S., A. Ishida, R. King, B. Raymond, N. Waller, A. Constable, S. Nicol, M. Wakita, och A. Ishimatsu. ”Risk Maps for Antarctic Krill under Projected Southern Ocean Acidification”. *Nature Climate Change* 3, nr 9 (september 2013): 843–47. https://doi.org/10.1038/nclimate1937.

Koonin, Eugene V. ”Orthologs, Paralogs, and Evolutionary Genomics”. *Annual Review of Genetics* 39 (2005): 309–38. https://doi.org/10.1146/annurev.genet.39.073003.114725.

Marcet-Houben, Marina, och Toni Gabaldón. ”The Tree versus the Forest: The Fungal Tree of Life and the Topological Diversity within the Yeast Phylome”. Redigerad av Christophe d’Enfert. *PLoS ONE* 4, nr 2 (03 februari 2009): e4357. https://doi.org/10.1371/journal.pone.0004357.

Marcet-Houben, Marina, och Toni Gabaldón. ”TreeKO: A Duplication-Aware Algorithm for the Comparison of Phylogenetic Trees”. *Nucleic Acids Research* 39, nr 10 (maj 2011): e66–e66. https://doi.org/10.1093/nar/gkr087.

Meyer, Bettina, Lutz Auerswald, Volker Siegel, Susanne Spahić, Carsten Pape, Bettina A. Fach, Mathias Teschke, Andreas L. Lopata, och Veronica Fuentes. ”Seasonal variation in body composition, metabolic activity, feeding, and growth of adult krill Euphausia superba in the Lazarev Sea”. *Marine Ecology Progress Series* 398 (2010): 1–18. http://dx.doi.org/10.3354/meps08371.

Meyer, Bettina. ”The Overwintering of Antarctic Krill, <Emphasis Type="Italic">Euphausia Superba,</Emphasis> from an Ecophysiological Perspective”. *Polar Biology* 35, nr 1 (01 januari 2012): 15–37. https://doi.org/10.1007/s00300-011-1120-0.

Meyer, B., P. Martini, A. Biscontin, C. De Pittà, C. Romualdi, M. Teschke, S. Frickenhaus, m.fl. ”Pyrosequencing and de Novo Assembly of Antarctic Krill (Euphausia Superba) Transcriptome to Study the Adaptability of Krill to Climate-Induced Environmental Changes”. *Molecular Ecology Resources* 15, nr 6 (november 2015): 1460–71. https://doi.org/10.1111/1755-0998.12408.

Murphy, E. J., E. E. Hofmann, J. L. Watkins, N. M. Johnston, A. Piñones, T. Ballerini, S. L. Hill, m.fl. ”Comparison of the structure and function of Southern Ocean regional ecosystems: The Antarctic Peninsula and South Georgia”. *Journal of Marine Systems*, Large-scale regional comparisons of marine biogeochemistry and ecosystem processes - research approaches and results, 109–110 (01 januari 2013): 22–42. https://doi.org/10.1016/j.jmarsys.2012.03.011.

Nichio, Bruno T. L., Jeroniza Nunes Marchaukoski, och Roberto Tadeu Raittz. ”New Tools in Orthology Analysis: A Brief Review of Promising Perspectives”. *Frontiers in Genetics* 8 (31 oktober 2017). https://doi.org/10.3389/fgene.2017.00165.

Notredame, C., D. G. Higgins, och J. Heringa. ”T-Coffee: A Novel Method for Fast and Accurate Multiple Sequence Alignment”. *Journal of Molecular Biology* 302, nr 1 (08 september 2000): 205–17. https://doi.org/10.1006/jmbi.2000.4042.

Papot, Claire, Kévin Cascella, Jean-Yves Toullec, och Didier Jollivet. ”Divergent Ecological Histories of Two Sister Antarctic Krill Species Led to Contrasted Patterns of Genetic Diversity in Their Heat-Shock Protein (Hsp70) Arsenal”. *Ecology and Evolution* 6, nr 5 (mars 2016): 1555–75. https://doi.org/10.1002/ece3.1989.

Peck, Lloyd S., Simon A. Morley, Joëlle Richard, och Melody S. Clark. ”Acclimation and Thermal Tolerance in Antarctic Marine Ectotherms”. *Journal of Experimental Biology* 217, nr

1 (01 januari 2014): 16–22. https://doi.org/10.1242/jeb.089946.

Petersen, Malte, Karen Meusemann, Alexander Donath, Daniel Dowling, Shanlin Liu, Ralph S. Peters, Lars Podsiadlowski, m.fl. ”Orthograph: a versatile tool for mapping coding nucleotide sequences to clusters of orthologous genes”. *BMC Bioinformatics* 18 (16 februari 2017): 111. https://doi.org/10.1186/s12859-017-1529-8.

Poloczanska, Elvira S., Christopher J. Brown, William J. Sydeman, Wolfgang Kiessling, David S. Schoeman, Pippa J. Moore, Keith Brander, m.fl. ”Global Imprint of Climate Change on Marine Life”. *Nature Climate Change* 3, nr 10 (oktober 2013): 919–25. https://doi.org/10.1038/nclimate1958.

Roth, Christian, Shruti Rastogi, Lars Arvestad, Katharina Dittmar, Sara Light, Diana Ekman, och David A. Liberles. ”Evolution after Gene Duplication: Models, Mechanisms, Sequences, Systems, and Organisms”. *Journal of Experimental Zoology. Part B, Molecular and Developmental Evolution* 308, nr 1 (15 januari 2007): 58–73. https://doi.org/10.1002/jez.b.21124.

Sales, Gabriele, Bruce E. Deagle, Enrica Calura, Paolo Martini, Alberto Biscontin, Cristiano De Pittà, So Kawaguchi, m.fl. ”KrillDB: A de Novo Transcriptome Database for the Antarctic Krill (Euphausia Superba)”. Redigerad av Cristiano Bertolucci. *PLOS ONE* 12, nr 2 (10 februari 2017): e0171908. https://doi.org/10.1371/journal.pone.0171908.

Siegel, Volker. *Biology and Ecology of Antarctic Krill*. New York, NY: Springer Berlin Heidelberg, 2016.

Stamatakis, Alexandros. ”RAxML Version 8: A Tool for Phylogenetic Analysis and Post-Analysis of Large Phylogenies”. *Bioinformatics (Oxford, England)* 30, nr 9 (01 maj 2014): 1312–13. https://doi.org/10.1093/bioinformatics/btu033.

Stapley, Jessica, Julia Reger, Philine G. D. Feulner, Carole Smadja, Juan Galindo, Robert Ekblom, Clair Bennison, Alexander D. Ball, Andrew P. Beckerman, och Jon Slate. ”Adaptation Genomics: The next Generation”. *Trends in Ecology & Evolution* 25, nr 12 (december 2010): 705–12. https://doi.org/10.1016/j.tree.2010.09.002.

Toullec, Jean-Yves, Erwan Corre, Benoît Bernay, Michael A. S. Thorne, Kévin Cascella, Céline Ollivaux, Joël Henry, och Melody S. Clark. ”Transcriptome and Peptidome Characterisation of the Main Neuropeptides and Peptidic Hormones of a Euphausiid: The Ice Krill, Euphausia Crystallorophias”. Redigerad av Frederique Lisacek. *PLoS ONE* 8, nr 8 (21 augusti 2013): e71609. https://doi.org/10.1371/journal.pone.0071609.

Trachana, Kalliopi, Tomas A. Larsson, Sean Powell, Wei-Hua Chen, Tobias Doerks, Jean Muller, och Peer Bork. ”Orthology Prediction Methods: A Quality Assessment Using Curated Protein Families”. *BioEssays: News and Reviews in Molecular, Cellular and Developmental Biology* 33, nr 10 (oktober 2011): 769–80. https://doi.org/10.1002/bies.201100062.

Wen, Jun, Zhiqiang Xiong, Ze-Long Nie, Likai Mao, Yabing Zhu, Xian-Zhao Kan, Stefanie M. Ickert-Bond, Jean Gerrath, Elizabeth A. Zimmer, och Xiao-Dong Fang. ”Transcriptome

Sequences Resolve Deep Relationships of the Grape Family”. Redigerad av Hector Candela. *PLoS ONE* 8, nr 9 (17 september 2013): e74394. https://doi.org/10.1371/journal.pone.0074394.

Zmasek, Christian M., och Sean R. Eddy. ”A Simple Algorithm to Infer Gene Duplication and Speciation Events on a Gene Tree”. *Bioinformatics* 17, nr 9 (01 september 2001): 821–28. https://doi.org/10.1093/bioinformatics/17.9.821.