Molekularne osnove adaptacija jednakonožnih rakova roda Proasellus na špiljski okoliš

Marković, Eva

University of Zagreb

Faculty of Science

Department of Biology

Eva Marković

The molecular basis of adaptations in isopod crustaceans of the *Proasellus* genus to cave environments

Master thesis

Zagreb, 2024.

H. Bilan

Sveučilište u Zagrebu

Prirodoslovno-matematički fakultet

Biološki odsjek

Eva Marković

Molekularne osnove adaptacija jednakonožnih rakova roda *Proasellus* na špiljski okoliš

Diplomski rad

Zagreb, 2024.

Ovaj rad je izrađen u Laboratoriju za molekularnu genetiku na Zavodu za molekularnu biologiju Instituta "Ruđer Bošković" u Zagrebu, pod mentorstvom dr. sc. Helene Bilandžije, te komentorstvom izv. prof. dr. sc Anamarije Štambuk. Rad je predan na ocjenu Biološkom odsjeku Prirodoslovno-matematičkog fakulteta Sveučilišta u Zagrebu radi stjecanja zvanja magistre molekularne biologije.

TEMELJNA DOKUMENTACIJSKA KARTICA

Sveučilište u Zagrebu
Prirodoslovno-matematički fakultet
Biološki odsjek

Diplomski rad

Molekularne osnove adaptacija jednakonožnih rakova roda *Proasellus* na špiljski okoliš

Eva Marković
Rooseveltov trg 6, 10000 Zagreb, Hrvatska

Život u špiljama može biti stresan. Nedostatak svjetlosti i limitirana dostupnost hrane usmjeravaju evoluciju špiljskih vrsta, selektirajući za prilagodbe nužne za preživljavanje u takvom okolišu. Bilo da se radi o gubitku pigmentacije, redukciji ili potpunom gubitku očiju, sporijem metabolizmu ili odgovoru na hipoksiju izazvanu niskim razinama kisika u špiljskim vodama, sve prilagodbe počinju na molekularnoj razini. Kako bi se došlo do dubljeg uvida u ove prilagodbe, provedeno je istraživanje komparativne transkriptomike, uspoređujući dvije špiljske (*Proasellus anophtalmus* i *Proasellus hercegovinensis*) i dvije vanjske vrste (*Proasellus coxalis* i *Proasellus karamani*) slatkovodnih jednakonožnih rakova. U ovom istraživanju sklopljeno je četiri transkiptoma kutikula ovih jednakonožnih rakova. Određena je genska ortologija, te je provedena analiza diferencijalne ekspresije na zajedničkim ortolozima. Rezultati istraživanja ukazali su na visoke razine ekspresije gena *C15orf48, attractin* i *tramtrac* u obje špiljske vrste, sugerirajući na zajedničku konvergentnu strategiju prilagodbe za špiljske vrste. Ovaj pronalazak ukazuje na prilagodbu na hipoksiju, očuvanje i skladištenje energije, imunološki odgovor te regresiju očiju, rasvijetljavajući načine na koje se špiljske vrste prilagođavaju kako bi opstajale u zahtjevnom špiljskom okolišu.

Ključne riječi: *de novo* sklapanje transkriptoma, komparativna transkriptomika, analiza diferencijalne ekspresije

(97 stranica, 31 slika, 16 tablica, 86 literaturnih navoda, jezik izvornika: engleski)
Rad je pohranjen u Središnjoj biološkoj knjižnici

Mentor: dr. sc. Helena Bilandžija
Komentor: izv. Prof. Anamaria Štambuk

Ocjenitelji:

Izv. prof. dr. sc. Anamaria Štambuk
Izv. prof. dr. sc. Duje Lisičić
Izv. prof. dr. sc. Sandra Hudina

Rad prihvaćen: 5. 9. 2024.

BASIC DOCUMENTATION CARD

University of Zagreb
Faculty of Science
Department of Biology Master thesis

The molecular basis of adaptations in isopod crustaceans of the *Proasellus* genus to cave environments

Eva Marković
Rooseveltov trg 6, 10000 Zagreb, Croatia

Life in a cave can be stressful. An absence of light and limited food availability drives the evolution of cave species, selecting adaptations crucial for survival in such environments. Whether it's the loss of pigmentation, reduction or loss of eyes, a slower metabolism due to food scarcity, or a hypoxic response due to oxygen-poor cave waters, all adaptations originate at a molecular level. To gain deeper insights into these adaptations, a comparative transcriptomics study was conducted, comparing two cave-dwelling (*Proasellus anophtalmus* and *Proasellus hercegovinensis*) and two surface-dwelling species of freshwater isopods (*Proasellus coxalis* and *Proasellus karamani*). In this study, four high-quality *de novo* transcriptomes were assembled from the cuticular tissue of the isopods. Gene orthology was inferred, and a differential expression analysis was performed on the common orthologues, a first of its kind for the species. The study revealed high expression levels of the genes *C15orf48, attractin* and *tramtrac* in both cave species suggesting a shared, convergent adaptive strategy among the cave-dwellers. Indicating an adaptation to hypoxia, energy conservation and storage, immune response, and eye regression, highlighting the intricate ways these species have evolved to thrive in the demanding cave environment.

Keywords: *de novo* transcriptome assembly, comparative transcriptomics, differential expression analysis
(97 pages, 31 figures, 16 tables, 86 references, original in: English)
Thesis is deposited in Central Biological Library.

Mentor: Dr. sc. Helena Bilandžija
Co-mentor: Assoc. Prof. Anamaria Štambuk

Reviewers:
Asst. Prof. Anamaria Štambuk, PhD
Assoc. Prof. Duje Lisičić, PhD
Assoc. Prof. Sandra Hudina, PhD

Thesis accepted: 5. 9. 2024.

Contents

Abbreviations

bp – Base pairs

BW – Bead Wash

cDNA – Complementary deoxyribonucleic acid

DNA – Deoxyribonucleic acid

DSP – Displacement Stop Primers

E1 – Enzyme Mix 1

E2 – Enzyme Mix 2

EB – Elution Buffer

ERCC – External RNA Controls Consortium

FPKM – Fragments Per Kilobase of transcript per Million mapped reads

Gb – Giga base pairs

GO – Gene Ontology

HYB – RNA Hybridization Buffer

ID – Identifier

L2FC – Log2 fold change

LM – Ligation Mix

LO – Ligation Oligo

M – Million

mRNA – Messenger ribonucleic acid

NGS – Next Generation Sequencing

ORF – Open reading frame

PA – *Proasellus anophtalmus*

PB – Purification Beads

PC – *Proasellus coxalis*; or Principal component

PCA – Principal Component Analysis

PCR – Polymerase Chain Reaction

PH – *Proasellus hercegovinensis*

PK – *Proasellus karamani*

PS – Purification Solution

qPCR – Quantitative real-time PCR

RNA – Ribonucleic acid

RNA-seq – RNA sequencing

rpm – Revolutions per minute

ROS – Reactive oxygen species

rRNA – Ribosome ribonucleic acid

RTM – Reverse Transcription Mix

SIRV – Spike-In RNA Variants

TPM – Transcripts per Million mapped reads

UDI – Unique Dual Index

UMI – Unique Molecular Identifiers

UV – Ultra-violet

WB – Wash Buffer

1. Introduction

1.1. Life in the cave environment

Constant darkness, lack of food, and high humidity, all characteristics of caves, make them a rather stressful environment. These and other conditions observed in caves are usually very constant with few variations but can nevertheless be quite extreme (Howarth & Moldovan, 2018). Despite these challenges, numerous organisms have successfully inhabited caves and developed various adaptations that allow them to thrive in such an environment (Krishnan and Rohner, 2017).

For example, air and water temperatures in caves are usually the same (Camacho, 1992) and very stable, defying seasonal variations (Cigna, 2002), with overall humidity reaching up to 100% (Lauritzen, 2018). On the other hand, despite the temperature stability, cave water pH levels that mostly fall between 7 and 9 (Lauritzen, 2018) can sometimes drop significantly due to dissolved substances (White, 1997), acidifying the environment. Moreover, toxic metal ions can be present in cave water (Macalady et al., 2007), creating additional environmental pressure on the inhabiting organisms.

However, the biggest challenge for a cave organism is the lack of light, which dictates almost all life aspects. No primary producers can live in a lightless environment, meaning only detritivores and predators can survive in caves (Howarth & Moldovan, 2018). By eliminating photosynthesizing organisms from the food web, the overall nutrient amount becomes extremely low, making food sources scarce (Howarth, 1993). Furthermore, the sense of vision becomes useless in the dark, making it harder to find the little food that is present. Finding a partner in the darkness becomes a further challenge, especially considering the low population density (Howarth & Moldovan, 2018).

1.2. Adaptations to life in caves

Only a small number of species are capable of colonizing the cave environment due to its particular conditions. These organisms are typically highly specialized for this environment, exibiting a distinct set of adaptations (Francis G. Howarth et al., 2018) which can be unified under the term troglomorphism. These adaptations result from convergent evolution and can be categorized into four major groups: morphological, physiological, behavioral, and other specialized adaptations (Derkarabetian et al., 2010). Each of these can be further classified as either constructive (e.g., hypertrophy of mechanosensory organs) or regressive (e.g., loss of vision) (Fišer, 2019), with the outcomes primarily influenced by the selective pressures of darkness (Derkarabetian et al., 2010; Culver et al., 2010) and food scarcity (Culver et al., 2015).

Morphological adaptations, which are the most easily observed, tend to follow two main pathways. One involves the reduction of characteristics that are useless in a dark environment, such as the loss of pigment, or albinism, which is commonly seen in stygobionts like the cave-dwelling *Proasellus* isopods, as well as many other cave species, both invertebrate and vertebrate (Bilandžija et al., 2012; M. Protas et al., 2012).

Functions such as attracting a mate, camouflage, and UV protection—all primary roles of body pigment (Ducrest et al., 2008; Sugumaran et al., 2016; True, 2003)—become irrelevant in a dark cave, leading to relaxed selection and subsequent loss of the pigment synthesis pathways (Bilandžija et al., 2012). In the case of melanin, a convergent pattern is observed—synthesis often halts at the first step (Bilandžija et al., 2012; McCauley et al., 2004; M. E. Protas et al., 2006). This may occur because it is economically efficient for the organism, prevents the accumulation of toxic intermediary compounds (Graham et al., 1978), or allows the buildup of other compounds that are beneficial in different pathways (Bilandžija et al., 2012). A pleiotropic trade-off of this kind has been observed in the cave-dwelling *Astyanax mexicanus*, where a mutation in the *oca2* gene, responsible for the first step in melanin synthesis, leads to the accumulation of L-tyrosine, which in turn allows for an increase in catecholamine synthesis (Audus et al., 1986; Fernstrom et al., 2007; M. E. Protas et al., 2006). This can potentially lead to a higher feeding efficiency (Stricker and Zigmond, 1984) as well as enhance the alertness in the species, resulting in less sleep (Duboué et al., 2011) and improving the chances of finding food (Bilandžija et al., 2013). On the other hand, arthropods rely on melanin for their immune response, including pathogen defence and wound healing (Ashida et al., 1995; Söderhäll et al., 1998), suggesting that albinism could negatively impact certain cave dwellers. Most tested arthropods have retained the ability to produce melanin in response to injury (Bilandžija et al., 2017), indicating precise regulation of the synthesis pathway rather than a complete blockage. However, this immune response was not observed in the *Proasellus* species, suggesting a lack of melanin-based immune defence (Bilandžija et al., 2017).

The loss of vision, or the complete absence of eyes, is another common morphological adaptation observed in cave dwellers. Extensive research has focused on cavefish, particularly *A. mexicanus*. Since eyesight is useless in a lightless environment, such regression is unsurprising. The loss of vision and eyes can be attributed to the accumulation of mutations in eye development genes due to the absence of selective pressure, as an energy-conserving measure or to protect against eye damage (Krishnan et al., 2017; Moran et al., 2015). However, a pleiotropic function has been noted where the loss of eyes in *A. mexicanus* is due to the regulation of *ssh* (Yamamoto et al., 2004) and *pax6* (Jeffery et al., 1998) genes. In turn, this results in elevated *ssh* levels in the mouth and pharynx, leading to the development of a shovel-like jaw that aids in food collection in cave sediments (Yamamoto et al., 2009). This suggests that the reduction of certain morphological traits often has pleiotropic effects, optimizing other traits or pathways more useful in the cave environment, rather than merely conserving energy or occurring spontaneously. In contrast to regression, constructive morphological adaptations such as enhanced senses of touch, smell (Langecker, 2000), and taste (Yamamoto et al., 2009), along with elongated antennae to accommodate more receptors (Moldovan et al., 2004), may compensate for the loss of vision and facilitate in navigating the cave environment.

Most physiological adaptations are driven by food scarcity, forcing organisms to scavenge whatever they find (Howarth & Moldovan, 2018). In addition to a broader diet, these animals typically have the ability to consume large quantities of food at once, survive longer periods without food, store more fat, and maintain a slower metabolism (Frédéric Hervant et al., 2002), often coupled with reduced movement and overall activity (Hüppop, 1985). Furthermore, because cave waters are often hypoxic (Malard & Hervant, 2012), a slower metabolism helps these organisms adapt to low-oxygen environments (Hervant et al., 1997, 1998, 2002).

Behavioural changes are also evident in cave animals. Most of these changes align with physiological adaptations, aiming to conserve energy and increase the efficiency of food search by moving more slowly (Moldovan and Paredes Bartolome, 1998/1999; Kuštor and Novak, 1980). Other adaptations include intraspecies communication shifts, relying on pheromones rather than sight (Cazals and Juberthie-Jupeau, 1983; Juberthie-Jupeau and Cazals, 1984; Moldovan and Juberthie, 1994), a reduced response to predation (Kowalko, 2019), and the loss of circadian rhythms due to constant darkness and the absence of seasonal changes (Howarth & Moldovan, 2018).

1.3. The *Proasellus* genus

Four species from the *Proasellus* genus are at the centre of this research, intending to shed light on the evolution of the cave species and the adaptations that make life in caves possible. The genus belongs to the Asellidae family, a group of isopod crustaceans that inhabit the freshwater environment. Two of the four species inhabit surface waters, *Proasellus coxalis* and *Proasellus karamani* (*Figure 1A*). Both species have developed eyes and body pigmentation, giving them a brownish colour (Wouters & Vercauteren, 2009; Henry et al., 1986). *P. karamani* is distributed on the Balkan peninsula, from Bosnia and Hercegovina to Macedonia (Sket, 1967), while *P. coxalis* has a much broader distribution from the North of Europe to the North of Africa. On the other hand, *Proasellus anophtalmus* (*Figure 1B*) and *Proasellus hercegovinensis* are cave dwellers. Both are depigmented and have no eyes. *P. anophtalmus* inhabits the Dinaric Karst and is the smallest of the four species, with a body up to 4.5 mm. While *P. hercegovinensis* is double the size (up to 10 mm), it inhabits only a small area of Popovo polje in Hercegovina (Henry et al., 1986; Karaman, 1955).

These four closely related species might help elucidate how cave animals evolve since a comparison can be made between two surface and two cave-dwelling species. Unique features discovered on the molecular basis might point towards the direction of selection and evolution, with common features between *P. anophtalmus* and *P. hercegovinensis* potentially highlighting the adaptations arising from the cave colonization.

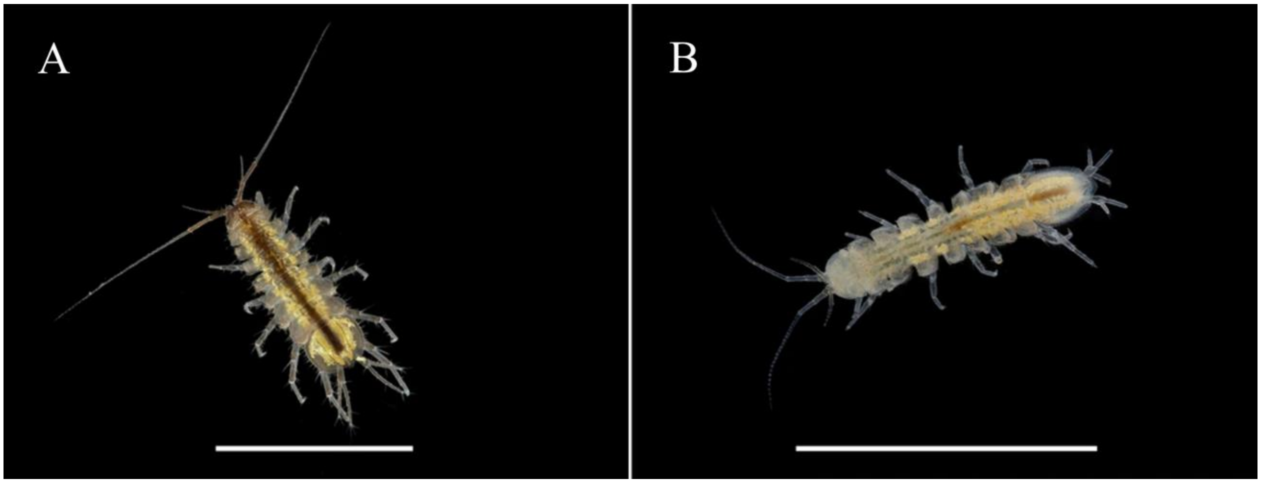


Figure 1 Two representatives of the Proasellus genus: (A) Proasellus karamani (surface-dwelling), and (B) Proasellus anophtalmus (cave-dwelling). Scale bars are 5 mm. Modified from Jovović et al., 2024. Photo credit: Tin Rožman

1.4. Next Generation RNA sequencing enabling a comparative transcriptomics study

Nucleotide sequencing has come a long way from the First-Generation Maxam-Gilber chemical degradation technique (Heather et al., 2016) and revolutionary Sanger's method based on chain termination using dideoxynucleotides which are able to sequence only short nucleotide fragments (Sanger et al., 1977). Today, entire genomes can be sequenced at once, utilizing Next-Generation sequencing (NGS) techniques. Illumina, a short-read Second-Generation sequencing technology is one such method. Along with other NGS methods, it has revolutionised omics techniques, including transcriptomics utilized in this research. A complete set of RNA molecules in an organism – the transcriptome, can be sequenced at once and subsequently studied. This allows for not only the entire transcriptome assemblies, but also for a thorough analysis of gene expression patterns in an organism (or condition) and inferring differentially expressed genes between two or more biological samples (Satam et al., 2023).

With over 10000 isopod species described, only a few of those have their genomes sequenced, and only two studies have been done on transcriptomes of *P. coxalis*, *P. karamani*, *P. hercegovinensis* and *P. anophtalmus* (Jovović et al., 2024). A comparative transcriptomics study can give valuable insights by providing sets of up- or down-regulated genes when comparing these four species. In order to do so, whole transcriptomes of each species have to be sequenced and assembled.

To obtain RNA-seq data, sequencing libraries need to be constructed first. These libraries contain cDNA fragments, complementary to the mRNA extracted from a sample, and adapters attached to the cDNA which allow for attaching and sequencing the cDNA on a sequencing machine. To prepare the libraries, total RNA is extracted from the sample, rRNA is removed, and mRNA fragmented. Reverse transcription is conducted using random priming to synthesize cDNA, RNA is degraded, and a second strand (identical to the mRNA) is synthesized. Then, adapters are added to the ends of the fragment which allow for library amplification with PCR. Library amplification ensures multiple copies of each fragment, increasing the chances for each fragment to be sequenced, and also adds adapters needed for sequencing. In this research, Lexogen’s RNA-seq library preparation protocol was followed which slightly differs from classical methods, primarily due to the elimination of the RNA fragmentation step. In this case, short cDNA first strands are generated from random primers which carry the partial adapter sequences. These partial adapter sequences then allow for the complementary adapter to hybridize, and a fragment is synthesized between two pairs of hybridized primers. In this case the first cDNA strand already carries partial adapters on both its 5’ and 3’ ends, preserving the strandedness of the library, which is lost with classical methods (*Figure 2*) (*RNA LEXICON | Lexogen*, n.d.).

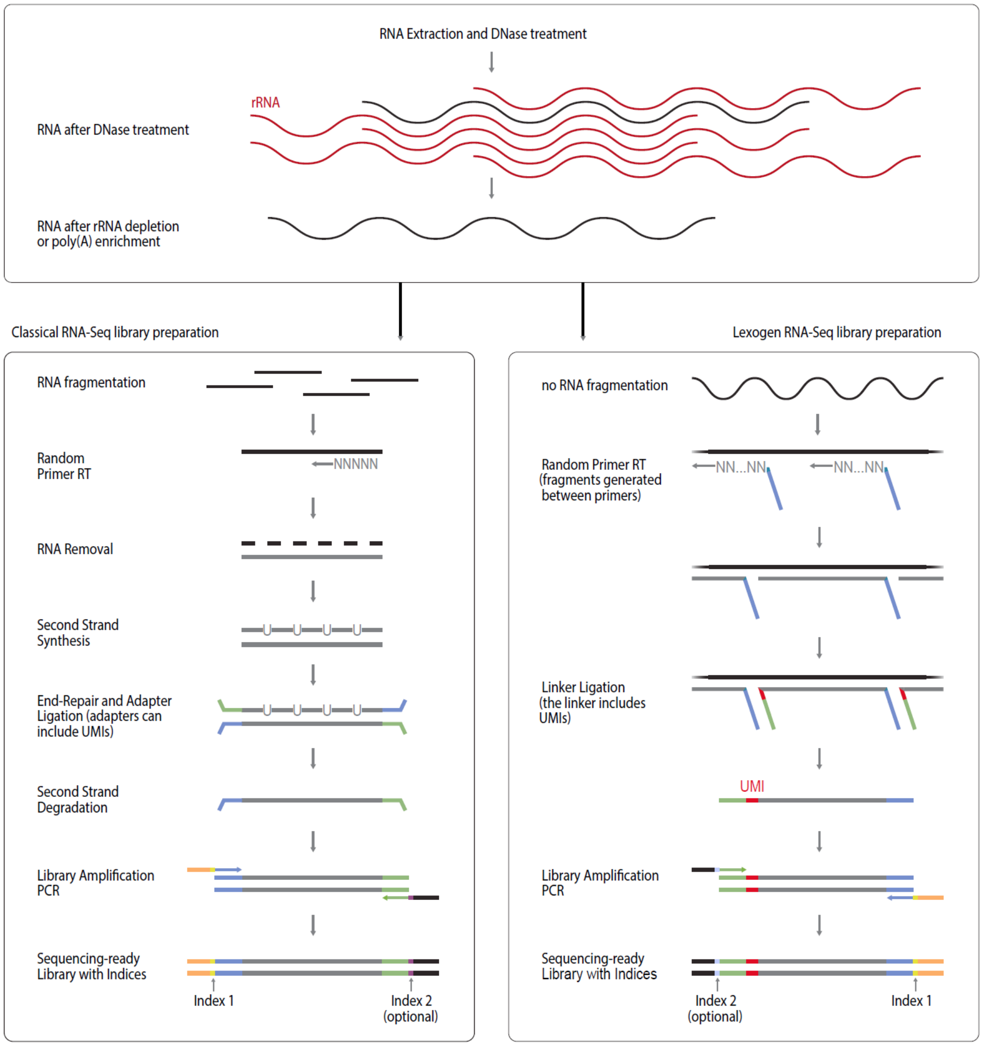

RNA Extraction and DNase treatment
RNA after DNase treatment
rRNA
RNA after rRNA depletion or poly(A) enrichment
Classical RNA-Seq library preparation
Lexogen RNA-Seq library preparation
RNA fragmentation
Random Primer RT
NNNNN
RNA Removal
Second Strand Synthesis
End-Repair and Adapter Ligation (adapters can include UMIs)
Second Strand Degradation
Library Amplification PCR
Sequencing-ready Library with Indices
Index 1
Index 2 (optional)
no RNA fragmentation
Random Primer RT (fragments generated between primers)
NN...NN
NN...NN
Linker Ligation (the linker includes UMIs)
UMI
Library Amplification PCR
Sequencing-ready Library with Indices
Index 2 (optional)
Index 1

Figure 2 A comparison of the classical RNA-Seq library preparation method, and Lexogen's method. Modified from RNA LEXICON (Lexogen).

The structure of the final sequencing-ready cDNA fragment is presented in *Figure 3*. The double stranded insert sequence consists of Read 1 (first, forward strand) and Read 2 (second, reverse strand) flanked by two adapters necessary for Illumina sequencing (P5 and P7 adapters). Both adapters contain parts of sequences necessary for binding to the complementary sequences on the Illumina sequencing flow cells. These are the outer regions coloured black and orange, respectively. Inner regions of the adapter (green and blue) are binding sites for the sequencing primers, which are used for the sequencing process. Furthermore, each adapter can carry an additional index (i5 and i7 respectively), which are short sequences shown in light blue and yellow colours. In case of Lexogen's library preparation protocol, these indices are named Unique Dual Indices (UDIs), and are necessary for library multiplexing. Multiplexing allows for multiple samples to be sequences on one flow cell, with UDIs being specific for each sample allowing for a demultiplexing process (*RNA LEXICON* | *Lexogen*, n.d.).

Lexogen's Illumina compatible adapters carry an additional tag, called Unique Molecular Identifier (UMI). These short 12 nucleotide sequences are placed at the 3' end of the P5 adapter, placing themselves between the Read 1 Sequencing Primer and Read 1. After the sequencing process is over, the beginning of each Read 1 carries the tag. Their purpose is to eliminate PCR duplicates that arise during library amplification due to the exponential nature of a PCR reaction and preferential amplification. There is enough variety in UMI sequences (4^{10} different sequences) that the probability of two identical cDNA fragment carrying the same UMI sequence is really low (*RNA LEXICON* | *Lexogen*, n.d.).

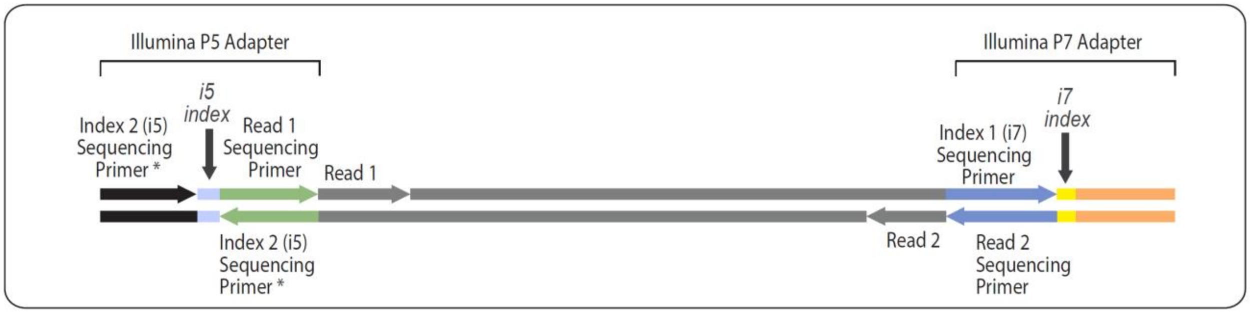


Figure 3 Completed Illumina cDNA library, with P5 and P7 adapter sequences. Modified from RNA LEXICON (Lexogen).

This kind of library is then ready to be sequenced on an Illumina platform, utilizing the sequencing by synthesis method.

Another benefit of the Lexogen's library preparation protocol is the fact that it enables the use of spike-in controls. These are artificial sequences that can be added to the samples before library construction is done, and are used to asses quantification accuracy and transcript coverage. Since these sequenced are of known length and composition, the expected sequencing results can be defined and then used to asses the quality of the entire sequencing process. Additionally, they can be used to compare the results between different biological samples and replicates, allowing for outlier detection (*RNA LEXICON* | *Lexogen*, n.d.).

A full pipeline of this comparative transcriptomics study is presented in *Figure 4*. It outlines the main steps of the process, from tissue isolation to bioinformatic analysis.

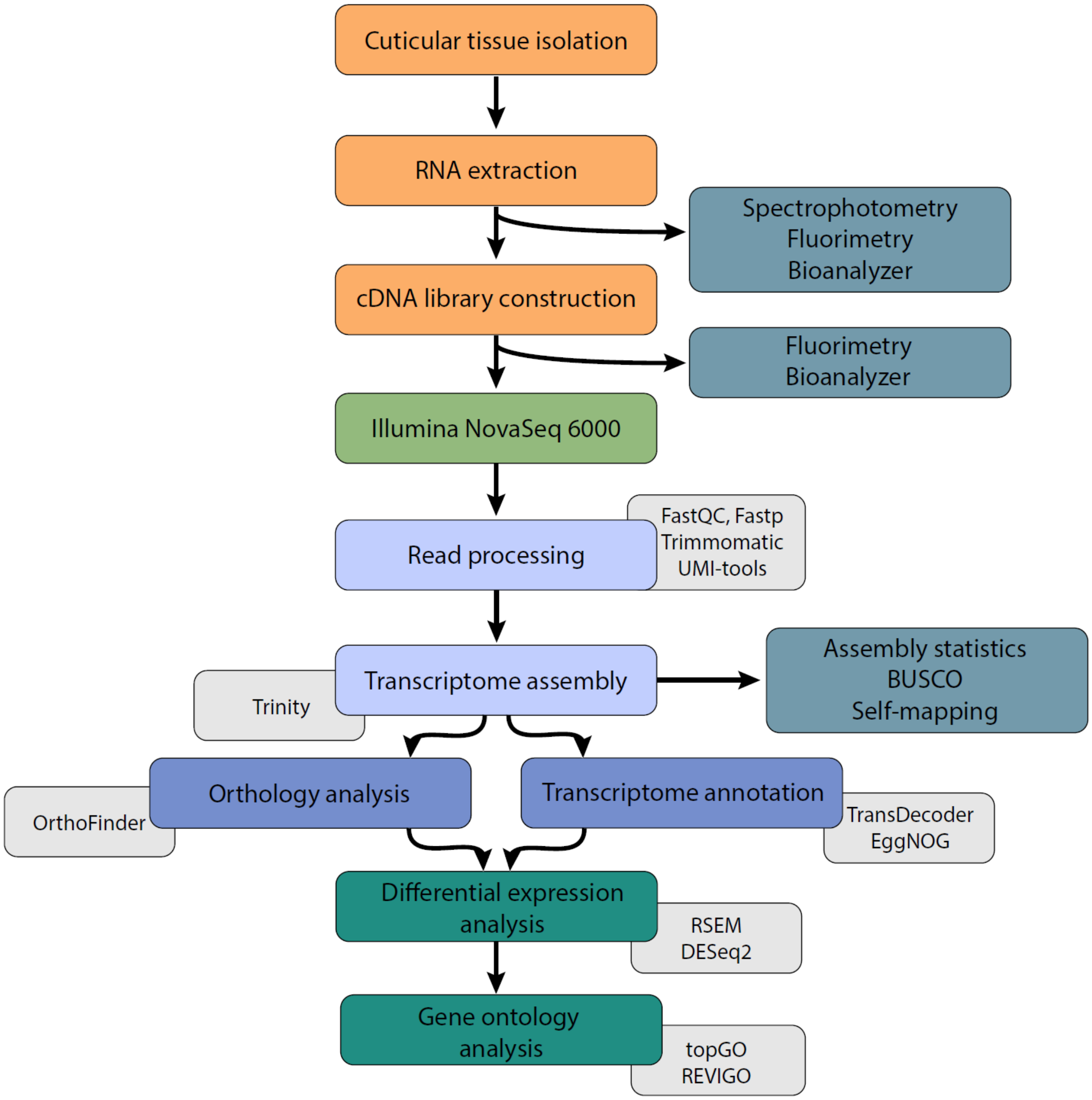


Figure 4 Comparative transcriptomics pipeline used in this study with the main steps of the process highlighted in coloured text boxes. A full process is shown, from sample preparation to bioinformatic analysis. Grey boxes adjacent to some of the steps indicate the main tools used in each of those steps, while the blue-grey boxes indicate main quality control steps done in the process.

2. Research goals

The research goal of this Master's thesis is to gain insights into the molecular mechanisms underlying the convergent and divergent troglomorphic adaptations of isopod crustaceans belonging to the *Proasellus* genus, based on transcriptomic data. The specific goals of this research are outlined as follows:

1. Inferring the transcript orthology between the four *Proasellus* species.
2. Conducting a differential gene expression analysis of the four species by creating pairwise comparisons between each two species, and clustering the samples.
3. Extracting common up- and down-regulated genes of each cave-dwelling species to determine the divergent adaptations.
4. Determining the commonly up- and down-regulated genes between the two cave-dwelling species to gain insights into the convergent adaptations.

3. Materials and methods

3.1. Specimen collection and cuticle isolation

Specimens of *P. coxalis*, *P. karamani*, *P. hercegovinensis*, and *P. anophtalmus* were collected during field research conducted in 2020, 2021, and 2022 (*Table 1*). Individuals were manually collected and transferred to the laboratory, where they were held in stable conditions according to the standard laboratory procedures (SLP) of the Laboratory for Molecular Genetics, Ruđer Bošković Institute (Lukić et al., 2024).

Table 1: Locations and dates of specimen collection during field research conducted from 2020 to 2022.

Species	Location	Date
P. coxalis	Vransko (lake)	30th Oct 2020
P. karamani	Ključ (stream)	22nd Jun 2021
P. hercegovinensis	Bjelušica (cave)	15th Sep 2021 4th Nov 2022
P. anophtalmus	Močiljska (cave)	4th Nov 2022

Between 25 and 40 healthy specimens of each species were randomly selected and transferred to a new plastic container, where they were subjected to starvation for one week prior to RNA isolation. I isolated the cuticles of each specimen in an RNase-free environment using histological needles and tweezers. First, I placed a live sample in a small Petri dish filled with 1 mL of RNA*later*™ Stabilization Solution (*Thermo Fisher Scientific*) and immediately decapitated the individual using histological tools. Then I proceeded to separate the cuticle from the bodies of the specimen (excluding appendages), placed them into molecular grade 1.5 mL Eppendorf tubes filled with 200 µL of RNA*later*™ Stabilization Solution and stored them at 4 °C until the next step (RNA isolation). I obtained a total of 32 *P. karamani*, 37 *P. coxalis*, 35 *P. anophtalmus* and 25 *P. hercegovinensis* cuticles for downstream use.

3.2. RNA isolation

After obtaining the cuticles, I proceeded with RNA isolation using the SPLIT RNA Extraction Kit (*Lexogen*) following the kit's protocol with slight modifications. The protocol consists of three main steps: sample homogenization, phenol/chloroform extraction, and column-based purification. The homogenization step is performed using a highly chaotropic isolation buffer, enabling solubilization and RNase inhibition. First, the samples are centrifuged on a bench-top centrifuge at 12000 g and 25 °C for one minute. Then, I placed the samples on ice for the rest of the homogenization process. I replaced RNA*later*™ Stabilization Solution by 400 µL of ice-cold Isolation Buffer, and added four 2.5 mm metal beads per tube. I placed the tubes on a homogenizer at speed level 5 for 20 seconds and then immediately back on ice for another 20 seconds. The process is repeated three times for 10 seconds with 20-second pauses on ice in between, finishing with a 3-minute incubation on ice. Afterward, the samples are centrifuged at 12000 g and 4 °C for one minute. Finally, I transferred the supernatant to a clean molecular grade Eppendorf tube and then centrifuged again at maximum speed and 4 °C for 3 minutes.

The next step of the protocol enables RNA extraction from the homogenized tissue, utilizing the phenol/chloroform extraction method. Firstly, I transferred the samples to Phase Lock Gel™ tubes that I centrifuged beforehand (12000 g, 18 °C, 1 minute). This method of RNA extraction is highly specific, partitioning proteins and DNA in the organic phase and leaving only RNA in the aqueous phase, with the gel from the tubes separating these two phases. I added a volume of 400 µL of a phenol solution (saturated with 0.1 M citrate buffer, pH 4.3, BioReagent, *Sigma Aldrich*) to the samples and mixed by inverting the tubes five times. Next, I added 150 µL of Acidic Buffer and mixed it by resuspending it ten times. Subsequently, I added 400 µL of chloroform (*Kemika*) and mixed it by swiftly and thoroughly inverting the tubes for 15 seconds. The samples are then incubated at room temperature for 2 minutes, followed by two subsequent centrifugations at 12000 g and 18 °C for 2 minutes each. The upper aqueous phase, partitioned above the gel, can then be transferred to a new molecular grade 2 mL Eppendorf tube by carefully decanting.

Finally, I purified the extracted RNA by a column-based method. From this moment on, all RNA samples were handled inside the UV-hood. I measured the volume of each sample (aqueous phase), added isopropanol at 1.75× of its volume, and mixed by vortexing for 10 seconds. A maximum of 800 µL is loaded into the purification column and placed in the collection tube. I then centrifuged the samples at 12000 g and 18 °C for 30 seconds. I discarded the flowthrough and repeated the process until the entire sample had been loaded into the purification column. After discarding the last flowthrough, I washed the sample by applying 500 µL of Wash Buffer (WB) to the column and centrifuged at 12000 g and 18 °C for 30 seconds. The flowthrough is discarded again, and the step is repeated twice more. Lastly, the sample is centrifuged again at 12000 g and 18 °C for one minute to spin-dry. I transferred the column to a new molecular grade 1.5 mL Eppendorf tube, where the RNA is eluted from the column with 15 µL of Elution Buffer (EB), prewarmed to 70 °C before being centrifuged at 12000 g and 18 °C for 1.5 minutes. I re-eluted the sample and centrifuged it once again. Then, I proceeded to the DNase treatment immediately.

3.3. DNase treatment

To make sure there is no left-over DNA in the RNA samples, I treated all the samples with a TURBO™ DNase (2 U/µL, Invitrogen™ by *Thermo Fisher Scientific*), which cleaves double-stranded DNA nonspecifically (*Thermo Fisher Scientific*). First, I added 30 µL of Nuclease-free water to each sample and prepared a mastermix consisting of 5 µL of 10X TURBO™ DNase Buffer (Invitrogen™ by *Thermo Fisher Scientific*) and 1 µL of TURBO™ DNase (2 U/µL) per sample. Then I added 6 µL of the mastermix to each sample and mixed by resuspending. The samples were then incubated at 30 °C for 30 minutes and mixed by resuspending at 15-minute incubation time. Once the incubation was over, the samples needed to precipitate, which I achieved by adding 5 µL of Lipoprotein(a) (LPA, 5 µg/µL, *Alfa Aesar*, *Ru-Ve*), 5.5 µL (1/10 of the total volume) of 3M NaOAc (3 M, pH 5.2, *Thermo Fisher*, *Ru-Ve*) and 151.3 µL (2.5× volume) of ice-cold absolute ethanol (*Kemika*) to each sample. I mixed everything by resuspending and precipitated overnight at -20 °C. After the precipitation, I centrifuged the samples at maximum speed and 4 °C for 20 minutes. Then, I removed the supernatant by pipetting and washed each sample three times with 500 µL of ice-cold 80% ethanol (prepared from 100% ethanol, *Kemika*) with gentle resuspending whilst washing. All ethanol had to be removed, so the samples were air-dried for 5 to 10 minutes with the tube caps opened. Lastly, I added 10 or 15 µL of nuclease-free water to the samples, which were then incubated at 60 °C for 2 minutes and mixed by resuspending before storing the finished samples at -20 °C. I determined the volume of nuclease-free empirically, using 15 µL for larger species (*P. coxalis*, *P. karamani*, and *P. hercegovinensis*) and 10 µL for the smaller *P. anophtalmus* to make sure the obtained RNA concentrations wouldn't be too low.

3.4. RNA quality control

A set of quality control checks were done on all isolated RNA samples to determine their usability for downstream applications. Quality checks included inquiring into the effectiveness of the DNase treatment, RNA concentration, sample purity, and RNA integrity.

3.4.1. DNase treatment effectiveness (PCR)

To check the success of the DNase treatment, I performed a PCR (polymerase chain reaction) to detect any left-over DNA traces in the RNA samples. For the PCR, I used primers for a 16S rRNA marker, which amplify 500 bp of the 16S rRNA gene. Primer sequences are stated in *Table 2*. To perform the PCR reaction, I used the GoTaq® G2 Green (or Colorless) Master Mix (*Promega*), a DNA polymerase in a ready-to-use master mix, following the reaction conditions stated in *Table 3*.

Table 2: Primer sequences used to amplify the 500 bp of the 16S rRNA standard marker gene.

Primer	Strand	Sequence
16Sbr-L	F	5'- CGCCTGTTTATCAAAAACAT - 3'
16S_Stena_R1	R	5'- CGTGGAAGTTTAATAGTCGAACAGAC - 3'

Table 3: PCR reaction conditions.

Step	Temperature	Time	Cycle number
Initial denaturation	95 °C	2:00	
Denaturation	95 °C	0:45	35×
Annealing	52 °C	0:45	
Extension	72 °C	0:45	
Final extension	72 °C	5:00	
Hold	4 °C	∞	

The PCR results, and thus the effectiveness of the DNase treatment, were determined by agarose gel electrophoresis. I prepared 50- or 90-mL gels with 1.5% agarose (*Sigma Aldrich*), staining the gel with the MIDORI Green Advance DNA Stain (0.05 µL per mL of gel; *NIPPON Genetics*). I loaded 5 µL of the PCR amplicons to the finished agarose gels, together with 1 µL of 6×DNA Dye if GoTaq® G2 Colorless Master Mix was used for the PCR. I conducted the electrophoresis at 120 V for 20 minutes using Mini-Sub Cell GT Cell or Wide Mini-Sub Cell GT Cell (*Bio Rad*) horizontal electrophoretic chambers and PowerPac™ HC High-Current Power Supply (*Bio Rad*) direct current power supply. Upon finishing the electrophoresis, I visualised the gels.

3.4.2. Spectrophotometry and fluorimetry

For the initial quantification of RNA obtained from the cuticles, I utilized the DS-11 Series Spectrophotometer/Fluorometer (*DeNovix*). The spectrophotometer reading provided information about the mass concentrations of the samples and the values of A_{260}/A_{280} and A_{260}/A_{230} ratios which I used to assess the purity/potential contaminations of the samples. Additionally, to achieve more accurate and precise RNA quantification results, I measured the RNA concentrations using fluorimetry, with a DeNovix RNA Assay (*DeNovix*) kit and the same, DS-11 Series Spectrophotometer/Fluorometer (*DeNovix*) instrument.

3.4.3. RNA gel electrophoresis and microcapillary electrophoresis

To assess the integrity of the RNA samples, I used methods of agarose gel electrophoresis and microcapillary electrophoresis. Since at least 200 ng of RNA is needed to successfully visualise it on the agarose gel, I used this method for samples with higher yields only (C > 50 ng/µL). I prepared 50- or 90-mL gels with 1.5% agarose, staining the gel with the MIDORI Green Advance DNA Stain (0.05 µL per mL of gel; *NIPPON Genetics*). I calculated the volumes of samples needed so that they contained 200 ng of RNA, diluting the samples with water if the volume to be loaded was less than 2 µL. Then, I mixed in the RNA Gel Loading Dye (*Thermo Fisher Scientific*) in a 1:1 ratio and loaded the samples onto the gel. Electrophoresis was conducted at 90 V for 40 minutes using Mini-Sub Cell GT Cell or Wide Mini-Sub Cell GT Cell (*Bio-Rad*) horizontal electrophoretic chambers and PowerPac™ HC High-Current Power Supply (*Bio-Rad*) direct current power supply. Samples with lower yields were assessed with the 2100 Bioanalyzer (*Agligent*) microcapillary electrophoresis instrument by the Laboratory for advanced genomics at the Ruđer Bošković Institute, using the Eukaryote Total RNA Nano assay. The Bioanalyzer system allows for the visualisation of RNA molecules based on the size distribution. It provides information about RNA integrity, quantification, and purity of the samples (*Sample Quality Control, Electrophoresis, Bioanalyzer | Agilent*, n.d.).

3.5. Equimolar sample pooling and spiking

Prior to library preparation, I performed the equimolar pooling of samples, whereby individual RNA samples of the same species were combined into groups of five, referred to as a "sample pool" hereafter. To execute sample pooling, I selected 20 individual cuticle RNA samples from *P. hercegovinensis* and 25 samples for each of the other species (*P. karamani*, *P. coxalis,* and *P. anophtalmus*) which showed no signs of contamination or degradation (confirmed by PCR and RNA electrophoresis/Bioanalyzer results, respectively), and favourable concentrations (at least 5 ng/µL for *P. anophtalmus*, and at least 20 ng/µL for every other species). This resulted in a total of five equimolar sample pools for *P. coxalis*, *P. karamani,* and *P. anophtalmus*, and four equimolar sample pools for *P. hercegovinensis*. The samples of *P. coxalis*, *P. karamani*, and *P. hercegovinensis* were pooled to a final concentration of 20 ng/µL and a total volume of

40 μL, meaning the total mass of RNA in a single sample pool was 800 ng. Accordingly, the mass of RNA from a single sample was 160 ng. On the other hand, the RNA yields from *P. anophtalmus* were pooled to the final concentrations of 10 ng/μL, and total volumes of 20 μL, with the total RNA mass of a sample pool being 200 ng, due to lower RNA yields. For the purpose of comparison amongst the RNA sequencing reads obtained from different sample pools and general library preparation workflow control, I added Spike-In RNA Variants (SIRV-Set 3, *Lexogen*), consisting of SIRV and ERCC sequences, to each sample pool. Two equations (1, 2) were used to calculate the volume of spike-in controls needed for the sample pools:

$$m_{SIRV} = F_{SIRV} \times F_{target\ RNA} \times m_{RNA\ input} \tag{1}$$

Where m_{SIRV} is the total mass of the spike-ins to be added to each sample pool, F_{SIRV} the fraction of desired SIRV reads, $F_{target\ RNA}$ the fraction of the RNA targeted in the experiment and $m_{RNA\ input}$ the mass of RNA input per sample pool. To calculate the volume of the SIRVs needed for each sample pool, the total mass of the SIRVs from equation (1) is divided by the concentration of the SIRVs at the prepared dilution (C_{SIRV}):

$$V_{SIRV} = \frac{m_{SIRV}}{c_{SIRV}} \tag{2}$$

The volumes of SIRVs added to each sample pool were 1.32 μL.

3.6. cDNA library construction

To carry out RNA sequencing cDNA libraries were constructed and amplified to concentrations suitable for the next-generation sequencing technology. The lowest RNA mass used was 47.7 ng, while the highest was 1571.3 ng.

3.6.1. Poly(A) selection

To extract mRNA molecules from the total isolated cell RNA, I used the Poly(A) RNA Selection Kit V1.5 (*Lexogen*) according to the manufacturer's instructions with some slight modifications. This method is based on extraction using magnetic beads with dT oligos attached. Polyadenylated RNA hybridizes to the oligos and is extracted using a magnet. The protocol is split into four steps: magnetic bead preparation, RNA denaturation, poly(A) RNA hybridization, and RNA elution. First, I prepared the magnetic beads by transferring 2 μL of resuspended beads to a new 1.5 mL DNA LoBind Tube (*Eppendorf AG*) per sample and placed the tubes on the magnet for 5 minutes. While still on the magnet, I removed the supernatant and then lifted the tubes off the magnet, adding 75 μL of the Bead Wash Buffer (BW) and washed by resuspending. I then put the tubes with the beads back on the magnet until the supernatant became clear (up to 5 minutes), removed the supernatant, and repeated the washing step. Lastly, after removing the clear supernatant, I added 10 – 20 μL of RNA Hybridization Buffer (HYB) to each sample and mixed by resuspending (I adjusted the volume of HYB based on the volumes of individual pools so that the volumes would be equal). With this, the beads were prepared for hybridization. For mRNA (polyadenylated RNA) to hybridize, it first had to be denatured. This was done by denaturing 10 – 20 µL of the pooled samples (depending on the volumes of the individual sample pools) at 60 °C for 1 minute on the thermocycler and then cooling and holding at 25 °C. Immediately after I finished denaturing the RNA, I added the entire volume of RNA to the 10 µL of the previously prepared magnetic tubes and put the tubes on the thermomixer at 25 °C and 1250 rpm for a 20-minute incubation. Then, I transferred the tubes to the magnetic rack until the beads were collected and the supernatant was clear (up to 5 minutes). I removed the supernatant and then removed the tubes from the magnetic rack, adding 100 µL of BW and resuspending to wash the beads. Then, I placed the tubes on a thermomixer at 25 °C and 1250 rpm for 5 minutes. After the incubation finished, I placed the tubes on the magnetic rack once again and let the beads collect for up to 5 minutes, removing and discarding the supernatant after collection. I repeated the washing step and continued with the final part of the protocol – RNA elution. To elute the hybridized RNA, I removed the supernatant and added 12 µL of the nuclease-free water, resuspending the beads. I put the tubes on the thermomixer at 70 °C for 1 minute, then immediately on the magnetic rack to collect the beads, for up to 5 minutes. Finally, I transferred 10 µL of the sample to new PCR tubes (0.5 mL AXYGEN, *Corning Incorporated*), avoiding the transfer of magnetic beads. With this, the mRNA was extracted and prepared for cDNA library generation, to which I proceeded immediately.

3.6.2. Library generation

The mRNA has to be reversely transcribed into cDNA to be sequenced. For this purpose, I used the CORALL Total RNA-Seq Library Generation Module and Purification Module from the CORALL RNA Seq with UDIs Kit (*Lexogen*), according to the kit's instructions with minor modifications. The protocol is split into two main steps – reverse transcription and linker oligo ligation, with each of these steps followed by purification. The first step in generating a cDNA library is reverse transcription which is done by utilizing Displacement Stop Primers (DSP). To each 10 µL RNA sample, I added 18 µL of Reverse Transcription Mix (RTM) and 1 µL of DSP. After mixing and spinning down, samples were incubated at 94 °C for 3 minutes, and then at 16 °C for 15 minutes on a thermocycler. Subsequently, I added 1 µL of Enzyme Mix 1, mixed and spun down. Another incubation cycle proceeded (10 minutes at 25 °C, 40 minutes at 37 °C, 10 minutes at 42 °C, cool to 25 °C and 1-minute hold at 25 °C). Immediately after the reverse transcription reaction, the samples had to be purified. I prepared a mastermix of Purification Beads (PB) and Bead Diluent (BD) in a 9 µL:29 µL ratio. I added a total of 38 µL of the mastermix to each sample. After mixing, samples were incubated for 5 minutes at room temperature. I moved the samples to the magnet and waited approximately 2 minutes before removing the clear supernatant. At this point, the library fragments should have been attached to the magnets. Then, I washed the beads (while still on the magnet) with 120 µL of 80 % ethanol, each time incubating for 30 seconds. I removed ethanol thoroughly, carefully removing any leftover drops, before letting the beads dry for approximately 7 minutes. Then, I added 20 µL of Elution

Buffer (EB) to the beads, and let them incubate for 2 minutes at room temperature. Finally, I placed the samples back on the magnet and transferred the supernatant to a new PCR tube after ensuring it was completely clear. At this point, the libraries should consist of small cDNA fragments with partial adapter sequences at their 5' ends.

The next step in generating a library is Linker Oligo Ligation, which adds partial Illumina-compatible adapters at the 3' ends of the first strand cDNA fragments. I added a mastermix of 36 µL Ligation Mix (LM), 1 µL dithiothreitol (DTT), 1 µL of Ligation Oligo (LO), and 2 µL of Enzyme Mix 2 (E2) to each sample. After mixing and spinning down, samples were incubated at 37 °C for 30 minutes. Again, another purification step proceeded. I added 9 µL of PB and 50 µL of BD to each sample, before incubating for another 5 minutes at room temperature. I transferred the samples to the magnetic rack and removed the supernatant after it became completely clear. Then, I added EB, removed the samples from the magnet, and left the samples to incubate for 2 minutes before adding Purification Solution (PS) and incubating again for 5 minutes at room temperature, to reprecipitate the libraries. I transferred the samples back to the magnet and removed the supernatant when it became clear. I proceeded with the ethanol wash in the same manner as in the previous purification. Finally, I eluted the libraries in 20 µL of EB, thoroughly mixed them, incubated them for 2 minutes at room temperature, and put them back to the magnet for 5 minutes, before transferring 17 µL to a new tube and proceeding to the library generation step.

3.6.3. Library amplification

To generate enough material for sequencing, the libraries need to be amplified using PCR. Additionally, the PCR step adds complete adapter sequences along with UDIs (Unique Dual Indices) required for multiplexing. But, before proceeding to the amplification step, I performed qPCR to precisely determine the optimal number of cycles for the PCR reaction, to prevent over- or under-cycling of the samples. For the qPCR reaction, I added 2 µL of Elution Buffer to the 17 µL of samples from the previous step. Then, I combined 1.7 µL of the sample, 7 µL of PCR Mix, 5 µL of P7 Primer (5' CAAGCAGAAGACGGCATA CGAGAT 3'), 1 µL of Enzyme Mix, 1.2 µL of 2.5x SYBR Green dye, and 14.1 µL of EB. I then performed the reaction program stated in *Table 4*.

Table 4 qPCR reaction conditions.

Step	Temperature	Time	Cycle number
Initial denaturation	98 °C	0:30	
Denaturation	98 °C	0:10	35×
Annealing	65 °C	0:20	
Extension	72 °C	0:30	
Final extension	72 °C	1:00	
Hold	10 °C	∞	

I calculated the number of cycles needed for the Endpoint PCR reaction by doing the following: 1) determining the maximum value of fluorescence (at the plateau phase of the qPCR), 2) calculating 50% of this maximum, 3) determining which reaction cycle reaches the 50% fluorescence value, 4) and finally subtracting 3 from the cycle number to get the optimal number of cycles for each sample.

After determining the number of cycles needed for the final PCR amplification step, I prepared a mastermix containing 7 µL of Dual PCR Mix and 1 µL of Enzyme Mix for each sample, added it to each library, and added 10 µL of Unique Dual Index Primer pair (UDI) to each. Each sample got unique UDIs that allows multiplexing all libraries together. The reaction program is stated in *Table 5*.

Table 5 Endpoint PCR reaction conditions.

Step	Temperature	Time	Cycle number
Initial denaturation	98 °C	0:30	
Denaturation	98 °C	0:10	14-17×
Annealing	65 °C	0:20	
Extension	72 °C	0:30	
Final extension	72 °C	1:00	
Hold	10 °C	∞	

Lastly, I performed a purification to complete the library generation process. Again, I added 31.5 µL Purification Beads to each reaction, collected the beads on the magnet, and removed the supernatant when it became clear. I removed the samples from the magnet, added 30 µL of EB, and incubated them at room temperature for 2 minutes. Then, I added 30 µL of Purification Solution for the libraries to reprecipitate, and the libraries were incubated again for 5 minutes at room temperature before placing them on the magnet and discarding the clear supernatant. I repeated the ethanol wash exactly like in the previous purifications, added 20 µL of EB, placed the samples back onto the magnet, and finally transferred 17 µL of fully finished libraries to new tubes. This concluded the library generation process.

3.7. Library quality control, pooling, and next generation sequencing

Similarly to RNA quality control, I estimated the quality of the cDNA libraries by measuring their concentration using fluorimetry with a DeNovix dsDNA High Sensitivity Assay (*DeNovix*) kit on the DS-11 Series Spectrophotometer/Fluorometer (*DeNovix*) instrument. Furthermore, to get insights into the integrity of the libraries, as well as their average fragment sizes, libraries were measured with the 2100

Bioanalyzer (*Agligent*) microcapillary electrophoresis instrument by the Laboratory for advanced genomics at the Ruđer Bošković Institute, using the High Sensitivity DNA Assay.

After determining that libraries were of good quality, I proceeded to pool them all together in an equimolar manner so that 25 fmol of each library ended up in the final pool. The library pool was sequenced by Novogene on the Illumina NovaSeq 6000 Sequencing System (*Illumina*).

3.8. Quality control and processing of raw reads

To assess the quality of the raw reads, I used the FastQC tool (version 0.12.1, Andrews, 2010) by specifying the forward and reverse paired-end reads for each sample. I then concatenated the output files into a single comprehensive report using MultiQC (version 1.14, Ewels et al., 2016).

Since Unique Molecular Identifiers (UMIs) were used during Library Generation (section 3.6.2) and are located at the beginning of the forward read, they had to be excluded from the read sequence. To accomplish this, I used the extract option in UMI-tools (version 1.1.5, Smith et al., 2017), which extracts the UMI sequence from the read and places it in the header of the sequence (the command is specified in the Supplement *9.1.1 UMI-tools command*).

To further ensure the quality of the reads, I trimmed adapter sequences and poly(G) sequences with the sequence trimming tool fastp (version 0.23.2, Chen et al., 2018). Poly(G) nucleotide sequences commonly occur at the read ends when sequencing strategies based on two-colour chemistry are employed, like Illumina NovaSeq used in this experiment. Trimming poly(G) sequences by fastp is essential for obtaining the best read quality. Furthermore, adapter sequences can sometimes be found at the ends of reads if sequencing proceeded beyond the read. Fastp automatically detects and trims adapters by overlapping each read pair (Chen et al., 2018). The command used for the fastp tool is specified in Supplement *9.1.2. fastp command*.

Since I added spike-in controls during library preparation, the SIRV and ERCC sequences were mixed with the endogenous mRNA sequences. For downstream *de novo* transcriptome assembly, the SIRV and ERCC reads needed to be filtered out. To achieve this, I used the STAR (version 2.7.10b) mapping tool (Dobin et al., 2013). (The command used for the STAR tool is specified in Supplement *9.1.3. STAR command*).

First, I mapped the processed reads (from the fastp output) to the SIRVome (a FASTA file containing all SIRV sequences) and saved the unmapped reads as an output. Next, I mapped these unmapped reads (which should now contain only ERCC sequences and endogenous mRNA) to the ERCC multi-FASTA file (a FASTA file containing all ERCC sequences) and again saved the unmapped reads as an output.

The final unmapped reads file should contain only endogenous mRNA sequences, which can then be used for the *de novo* genome assembly (section 3.9.).

3.9. *De novo* transcriptome assembly, quality assessment, and annotation

To assemble the transcriptomes of the four *Proasellus* species, I used Trinity (Trinity-v2.13.2), a tool specialized in *de novo* transcriptome assembly (Grabherr et al., 2011), using only RNA-seq reads (prepared as described in section 3.9.), and no reference (the command alongside the explanation is in the Supplement *9.1.4. Trinity command*).

After obtaining the assemblies, I assessed their quality with BUSCO (version 5.5.0, (Manni, Berkeley, Seppey, & Zdobnov, 2021; Manni, Berkeley, Seppey, Simão, et al., 2021)), specifying transcriptome-containing file and the *Arthropoda* database, which is the most taxonomically relevant database offered (Supplement *9.1.5 BUSCO command*). Following the quality assessment, I proceeded with annotating the obtained transcriptomes. Prior to the annotation itself, I used the TransDecoder tool (version 5.7.1, (Haas, BJ.)) to find the longest open reading frames (ORFs) which then yields the most probable coding sequences (CDS) of the transcripts by using the TransDecoder.LongOrfs option. Then, I used the TransDecoder.Predict option to obtain predicted peptide sequences by translating the nucleic coding sequences (command specified in Supplement *9.1.6. TransDecoder options*). Peptide sequences should result in better alignments when comparing them to database sequences since the amino-acid code is more conserved (Bininda-Emonds, 2005). Therefore, using them instead of nucleic acid sequences should provide better annotations for the transcriptome.

Finally, I used these peptide sequences to create annotations using the EggNOG-mapper tool (version 2.1.10, Cantalapiedra et al., 2021) which creates functional annotations based on gene orthology (command specified in Supplement *9.1.7. EggNOG-mapper command*). The mapping is done against the EggNOG database (version 5.0.2, Huerta-Cepas et al., 2019) with the DIAMOND sequence aligner (version 2.1.8., Buchfink et al., 2021).

3.10. Read mapping and calculating expression

The next step in the analysis was to obtain gene expression levels by mapping the reads back to the assembled transcriptomes and counting the number of reads mapped to each transcript. This process can be achieved with the RSEM tool (Li et al., 2011), but since I used UMIs for deduplication purposes, a lot of preprocessing had to be done beforehand (the entire pipeline can be found in the Supplement *9.1.8. Read processing, mapping, and expression calculation pipeline*).

First, I trimmed the reads with Trimmomatic (version 0.39, Bolger et al., 2014a) with the same parameters as the ones used in Trinity (see Supplement *9.1.4.*) to ensure that I used the same reads of the same quality. Next, I mapped the trimmed reads to the corresponding transcriptomes using bowtie2 (Langmead et al., 2012), following the exact parameters that the RSEM tool uses for mapping if bowtie2 is specified, for consistency. After the mapping process, I converted the obtained SAM file to a BAM file for easier manipulation using the SAMtools view command (version 1.19, Danecek et al., 2021). The BAM file had to then be sorted, which I again achieved using SAMtools (Supplement *9.1.8. sort command*).

Finally, I could deduplicate the sequences using the UMI-tools (Smith et al., 2017) dedup function based on UMI sequences that were extracted to the header (as explained in section 3.8.), also discarding any chimeric or unpaired reads. Again, I sorted the BAM file since it was unsorted during UMI deduplication using SAMtools (Supplement *9.1.8. sort command #2*).

The last step before calculating the gene expression was preparing the BAM file for RSEM. Redundant reads need to be removed because otherwise, they prevent RSEM from working. I achieved this with a dedicated script included in UMI-tools: prepare-for-rsem (Supplement *9.1.8. UMI-tools prepare-for-rsem command*). Lastly, I calculated the expression levels using the RSEM rsem-calculate-expression function (version 1.3.1), specifying alignments since I already had BAM files (mapped reads).

3.11. Assessing the quality of the RNA-seq workflow with spike-in controls

To further examine the quality of the RNA sequencing workflow, I assessed the spike-in controls added during library preparation. For consistency and comparability with the sample reads, I followed the same procedure of read mapping and calculating expression, matching the exact parameters described in section 3.10. It consisted of mapping the trimmed reads to the SIRVome (a fasta file containing all SIRV sequences) and to the ERCC transcripts (a fasta file containing all ERCC sequences). These mapped reads went through the same sorting, deduplicating, and expression calculating steps. Finally, to assess these spike-in controls, I utilized the SIRVsuite tool (command in Supplement *9.1.9. SIRVsuite command*).

3.12. Finding orthologues with OrthoFinder

To achieve the cross-species differential gene expression comparison, orthology has to be inferred in order to compare the orthologues genes. For this purpose, I utilized OrthoFinder (version 2.5.4, Emms et al., 2019), a tool for comparative genomics, under the default settings, with DIAMOND sequence aligner (version 2.1.8., Buchfink et al., 2021).

The next step was to associate the transcripts of all four species to their respective orthogroups inferred with OrthoFinder. I decided to use the hierarchical orthogroups output file since it should be the most accurate as it uses rooted gene (transcript) trees instead of gene (transcript) similarity (Emms et al., 2019). Firstly, I filtered out any orthogroup that didn’t have at least one representative transcript for each species. This step ensures only comparing transcripts (genes) present in all four species (Supplement *9.1.10. Python script 1*).

Then, I extracted the orthogrups from this file, associating them to the corresponding transcripts in the RSEM counts output file that was to be used for the differential gene expression analysis (Supplement *9.1.11. Python script 2*). After adding the orthogroups to all the corresponding transcripts in the RSEM files, I filtered out any transcripts from any of the new files that didn't have an associated orthogroup (these transcripts should be the ones excluded in the first step after removing orthogroups that didn't have representative transcripts of all four species; Supplement *9.1.12. Python script 3*).

Some of the orthogroups had multiple transcripts associated within one species. In such cases, following the method of Stern et al., (2018), I summed up the expected counts of all those transcripts under one orthogroup (Supplement *9.1.13. Python script 4*). Since the RSEM output file also contains information about the length and effective length of each transcript, in occurrences where I had to group multiple transcripts, I only left the length value of the longest transcript and the corresponding TPM and FPKM values, removing this information for the shorter transcripts. After modifying the RSEM files in such a way, the orthogroup IDs act as gene IDs, and their differential expression can be inferred downstream.

3.13. Differential gene expression analysis with DESeq2

To perform the differential gene expression analysis, I employed R (v4.3.2, R Core Team, 2024), utilizing the DESeq2 package (v1.42.0, Love et al., 2014) within R-Studio (v2023.9.1.494, Posit team, 2024), creating pairwise comparisons of differentially expressed genes between each two species. I also implemented the DESeq2 package for sample clustering, including hierarchical clustering to produce heatmaps, and PCA analysis for a broader comparison of all four species. To obtain clustering results, I implemented the variance-stabilizing transformation (VST) data transformation method. Full scripts used can be found in the Supplement (*9.1.14. DESeq2 analysis: A pairwise comparison example script; 9.1.15. DESeq2 analysis: sample clustering*). After identifying the differentially expressed genes, I analyzed the overlap of up-regulated genes between *P. anophtalmus* and *P. hercegovinensis*. I performed a similar analysis for the down-regulated genes. Additionally, I compiled separate lists of genes that were exclusively up-regulated or down-regulated in either *P. anophtalmus* or *P. hercegovinensis*.

3.14. Gene ontology analysis with topGO and REVIGO

After inferring differentially expressed genes with DESeq2, I performed a gene ontology (GO) analysis with the topGO (Alexa & Rahnenfuhrer, 2024) package in R (v1.42.0, Love et al., 2014) within R-Studio (v2023.9.1.494, Posit team, 2024), analysing all three categories of GO terms of up- and down-regulated genes: biological processes, molecular functions and cellular components based on GO IDs obtained by EggNOG annotation. I utilized the 'classic' algorithm and Fisher's method to obtain p-values of GO terms, analysing only GO terms with a p-value of 0.05 or lower. I inferred GO terms of genes up-regulated solely in *P. anophtalmus* utilizing the intersections of pairwise comparisons of *P. anophtalmus* and every other species. I did the same for genes up-regulated in *P. hercegoviensis.* To visualise and reduce the gene ontology terms, I utilised the REVIGO tool (version 1.8.1, Supek et al., 2011). Full scripts can be found in the Supplement (*9.1.16. topGO analysis*).

4. Results

Additional RNA and cDNA library quality control results can be found in the Supplement (*9.2.1. RNA and cDNA library quality*), and the spike-in quality control results can be found in the Supplement (*9.2.3 Spike-in quality control – individual sample quality*).

4.1. Sequence processing statistics and transcriptome assembly statistics

The NovaSeq 6000 Illumina sequencing of the cDNA libraries yielded between 17.4 and 24.6 million raw reads per library, or between 2.52 and 3.53 Gb, with at least 92.1% of those bases having a quality score of Q30 or above. Duplication levels of raw, unprocessed reads ranged from 16.1% to a maximum of 39.7%, depending on the sample, with only four samples having such high duplication levels. The GC content ranged from 38% to 41%, being relatively consistent within the species (*Table 6)*. None of the samples had a significant number of unassigned basses within the reads. In contrast, all of the samples had some degree of leading bases present at the beginning of the reads, and the majority (18 out of 19) samples had overrepresented sequences as a result of the adapter content that was detected in all samples (Supplement *Figure S3*). Removal of the adapters after trimming the reads with the fastp tool can be seen in Supplement *Figure S4* where base contents are more uniform. Any remaining low-quality reads or parts of reads were successfully removed with further trimming and filtering with Trimmomatic, dropping no more than 1.9% of reads of any sample.

After assembling the reads into transcripts, the entire transcriptomes consisted of more than 100000 transcripts, ranging from around 280 bp to over 36000 bp. The mean transcript length is around 1100 bp for all four species. The GC content also appears uniform across species, being 36% in the assemblies. Overall mapping rates of processed reads back to the assembled transcriptomes were high, the lowest being 87.6%. Additionally, all four transcriptomes showed high levels of completeness, having more than 89% of complete BUSCOs and less than 5% of missing BUSCOs.

Table 6 Sequence processing statistics and mapping rates for each sample, transcriptome assembly statistics and transcriptome completeness for each species.

Species	*Proasellus anophtalmus* (PA)					*Proasellus coxalis* (PC)					*Proasellus hercegovinensis* (PH)				*Proasellus karamani* (PK)				
Sample	1	2	3	4	5	1	2	3	4	5	1	2	3	4	1	2	3	4	5
Sequence processing																			
Raw reads (M)	18	17.6	19.2	18.4	19.4	18.4	23	17.4	24.6	18.2	21.2	19.6	20.6	18.8	20	18	21.2	18.2	18.8
Raw bases (Gb)	2.58	2.52	2.75	2.65	2.80	2.65	3.31	2.52	3.53	2.06	3.06	2.84	2.97	2.71	2.88	2.88	2.60	3.05	2.63
Q30 %	92.7	92.4	92.6	92.0	92.1	93.4	93.5	93.4	93.0	95.0	93.0	92.2	92.8	92.9	92.1	92.1	92.1	92.7	93.4
Clean reads (M)	17.8	17.4	19	18.2	19.2	18.2	22.8	17.4	24.2	18	21	19.4	20.4	18.6	19.8	17.8	20.8	18	18.6
GC content (%)	41	39.5	41	41	41	40.5	40	40	40	40	39	40	40	40	40	41	38	39.5	39
Assembly statistics																			
Transcripts (#)	124866					144241					120069				161524				
Min. length (bp)	289					279					280				282				
Max. length (bp)	31547					36818					23040				33109				
Mean length (bp)	1107.69					1160.77					1089.9				1072.8				
Median length (bp)	697					731					688				666				
N90	469					489					461				450				
N70	939					1004					927				902				
N50	1597					1699					1570				1546				
N30	2602					2731					2532				2556				
GC content (%)	36%					36%					36%				36%				
L50	23576					27216					22918				30164				
L90	87933					100563					84756				114206				
Transcriptome completeness																			
Complete (%)	95.46%					90.33%					89.63%				93.39%				
Fragmented (%)	2.86%					5.33%					6.61%				4.05%				
Missing (%)	1.68%					4.34%					3.75%				2.57%				
Mapping																			
Overall mapping (%)	91.0	89.3	91.6	91.7	91.5	92.0	90.8	91.2	90.9	90.9	89.6	89.3	89.4	89.5	90.9	90.1	87.6	89.4	89.8

4.2. Gene orthology

A total of 80597 orthogroups were found using OrthoFinder when analysing 550700 sequences from all four transcriptomes. Out of these orthogroups, 40768 (50.6%) were species-specific, amounting to 31.5% of all sequences. Of all orthogroups, 9120 (11.3%) were present in all four species and therefore used for downstream analysis. Overall, 87.8% of sequences were assigned to orthogroups, leaving 12.2% of sequences unassigned. An average orthogroup consisted of six sequences, while a median orthogroup size was four sequences. The G50 value for assigned sequences is 8, indicating that 50% of orthogroups contain at least eight sequences. Meanwhile, the O50 is 16159, meaning that 50% of all assigned sequences are encompassed in 16159 orthogroups. Among all four species, *P. anophtalmus* had the highest percentage of sequences assigned to orthgroups, and *P. karamani* had the least, with close to 20% of all sequences unassigned to orthogroups. On the other hand, *P. karamani* had the highest percentage of sequences in species-specific orthogroups, while *P. hercegovinensis* had the least (*Table 7*).

Table 7 OrthoFinder gene orthology statistics for all four Proasellus species.

Species	*Proasellus anophtalmus*	*Proasellus coxalis*	*Proasellus hercegovinensis*	*Proasellus karamani*
Percentage of sequences in orthogroups	91.80%	89.60%	91.50%	80.20%
Percentage of unassigned sequences	8.20%	10.40%	8.50%	19.80%
Percentage of sequences in species-specific orthogroups	18.60%	38.10%	15.30%	47.70%

When comparing the overlaps in orthogroups among all four species, it is evident that *P. anophtalmus* and *P. hercegovinensis* share the largest number of orthogroups, nearly doubling every other overlap (*Table 8*). On the other hand, *P. karamani* shares the smallest amount of orthogroups with other species, aligning with the highest percentage of species specific orthogroups.

Table 8 Overlap of orthogroups in pairwise comparisons between each two species.

Species	*P. coxalis*	*P. hercegovinensis*	*P. karamani*
P. coxalis			
P. hercegovinensis	16727		
P. karamani	12479	15523	
P. anophtalmus	17267	30017	15410

Furthermore, OrthoFinder produced a species tree inferred from all analysed orthogroups, (*Figure 5*). It is evident that based on orthogroups, or rather transcriptome sequences, cave species (*P. anophtalmus* and *P. hercegovinensis*) are grouped together. Both of these species branch from the same root. *P. coxalis* has the longest branch, while *P. karamani* is the root of the tree.

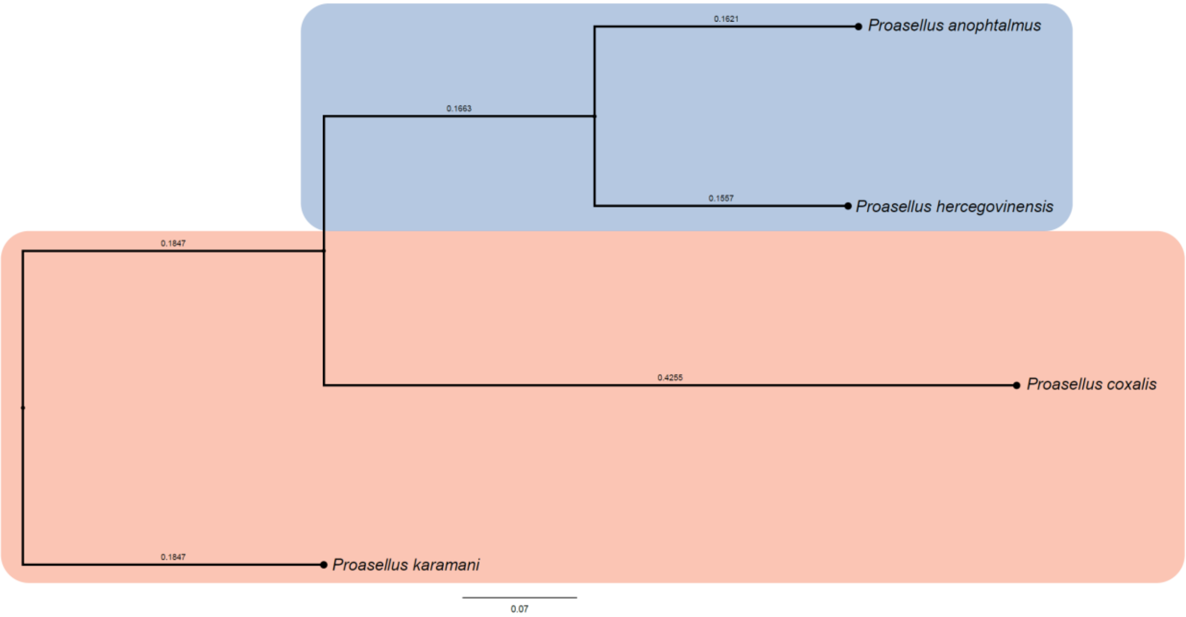


Figure 5 Species tree inferred from all analysed orthogroups. Tree was produced by OrthoFinder.

4.3. Differentially expressed genes and gene ontology

Comparing each species to all other species resulted in a total of six pairwise comparisons of differentially expressed genes with DESeq2. It has to be noted that, in this case, a gene is approximated to an orthogroup since one orthogroup encompasses multiple similar transcripts, likely transcribed from the same gene. Each comparison was visualised with a MA plot, showing the distribution of differentially expressed genes – plotting the log2 fold change (L2FC) value of a gene against its mean of normalized counts. All six comparisons resulted in very similar MA plots, an example of which is shown in *Figure 6.* The MA plot shows the comparison of *P. coxalis* and *P. anophtalmus* where dots on the plot represent differentially expressed genes – dots above the middle line (with an L2FC value above zero) are genes up-regulated in *P. coxalis* (or down-regulated in *P. anophtalmus*). The dots bellow the middle line (with L2FC below zero) are genes up-regulated in *P. anophtalmus* (or down-regulated in *P. coxalis*). If a dot is blue, the result is statistically significant, with a p-value of 0.05 or lower (95% significance). Grey dots are statistically unsupported genes. Even though the lower range of gene counts has more unsupported genes, the statistically significant genes span the entire range of count numbers. The L2FC values span from -11.6 to 13.4, with the lowest changes in expression being around -0.4 to 0.4.

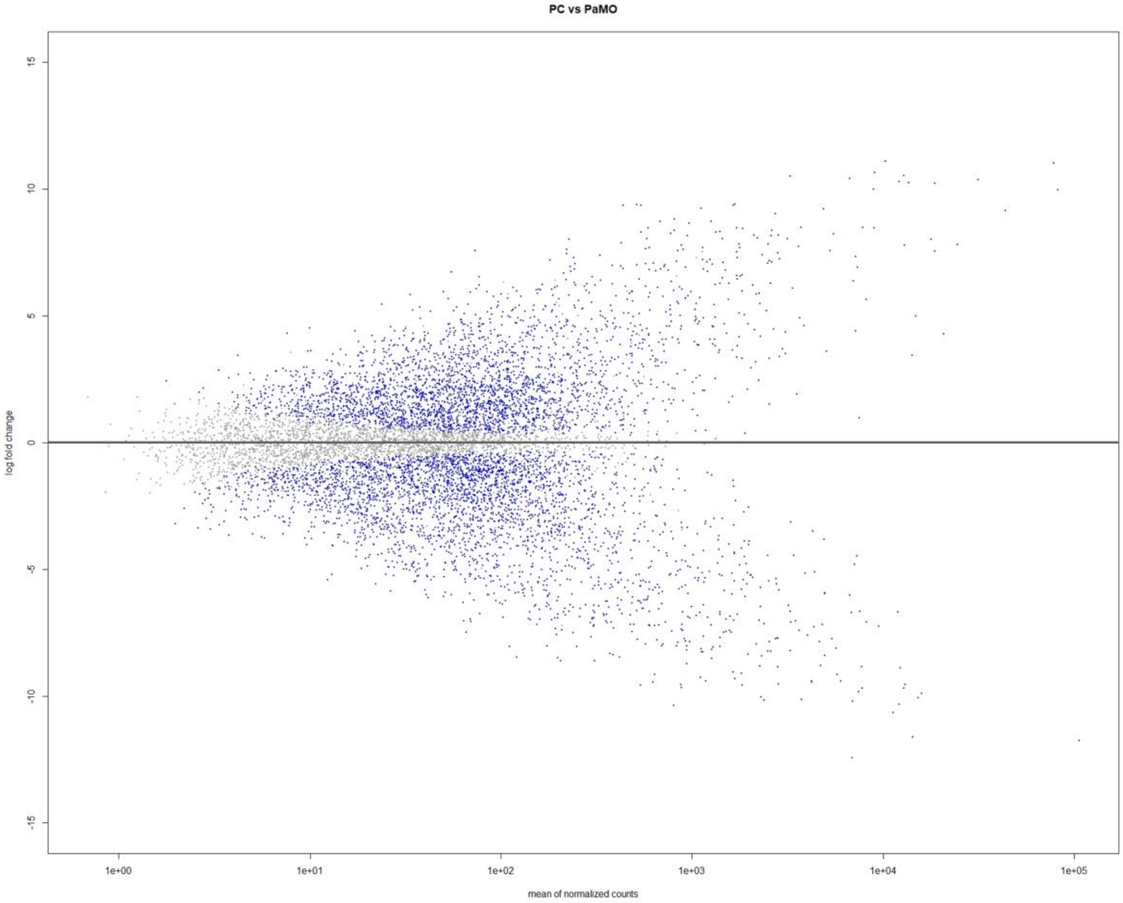


Figure 6 MA plot of differentially expressed genes between P. coxalis and P. anophtalmus. Positive log2 fold change values represent genes up-regulated in P. coxalis, while negative values represent genes up-regulated in P. anophtalmus. Blue dots represent statistically significant genes. The plot was produced with DESeq2.

4.3.1. Sample clustering

Apart from pairwise comparisons, sample clustering was performed in an all-versus-all manner. A heatmap displaying the top 200 differentially expressed genes with the highest L2FC values is shown in *Figure 7*. Red tones represent the highest L2FC values, while the lowest values are represented with blue tones. Each species has its pattern on the heatmap, with *P. karamani* and *P. anophtalmus* having higher L2FC values than the other two species.

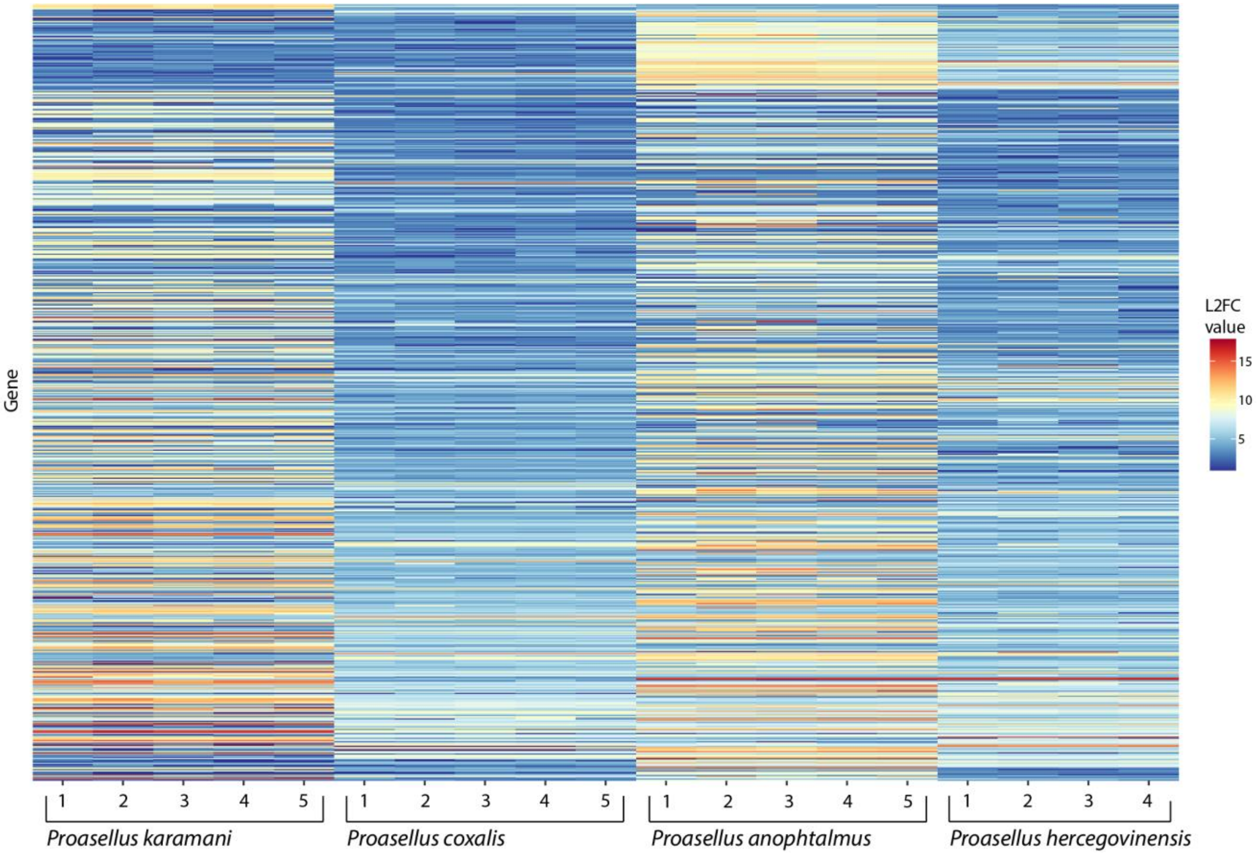


Figure 7 Heatmap displaying the top 200 differentially expressed genes with the highest L2FC values across all samples. Each row denotes a single differentially expressed gene while the plot is split into four major sections, one for each species. The numbers in each of the samples represent the replicate samples of each species. A legend is depicted on the right side, showing a coloured scale for the L2FC values, with red being the highest. The plot was produced with DESeq2.

A sample-to-sample distance plot shown in *Figure 8* highlights the uniformity of the samples withing a species (dark blue colour), and also a disparity between species (light blue colour). Additionally, the tree surrounding the distance heatmap implies the same relatedness between species as the species tree produced by OrthoFinder in *Figure 5*.

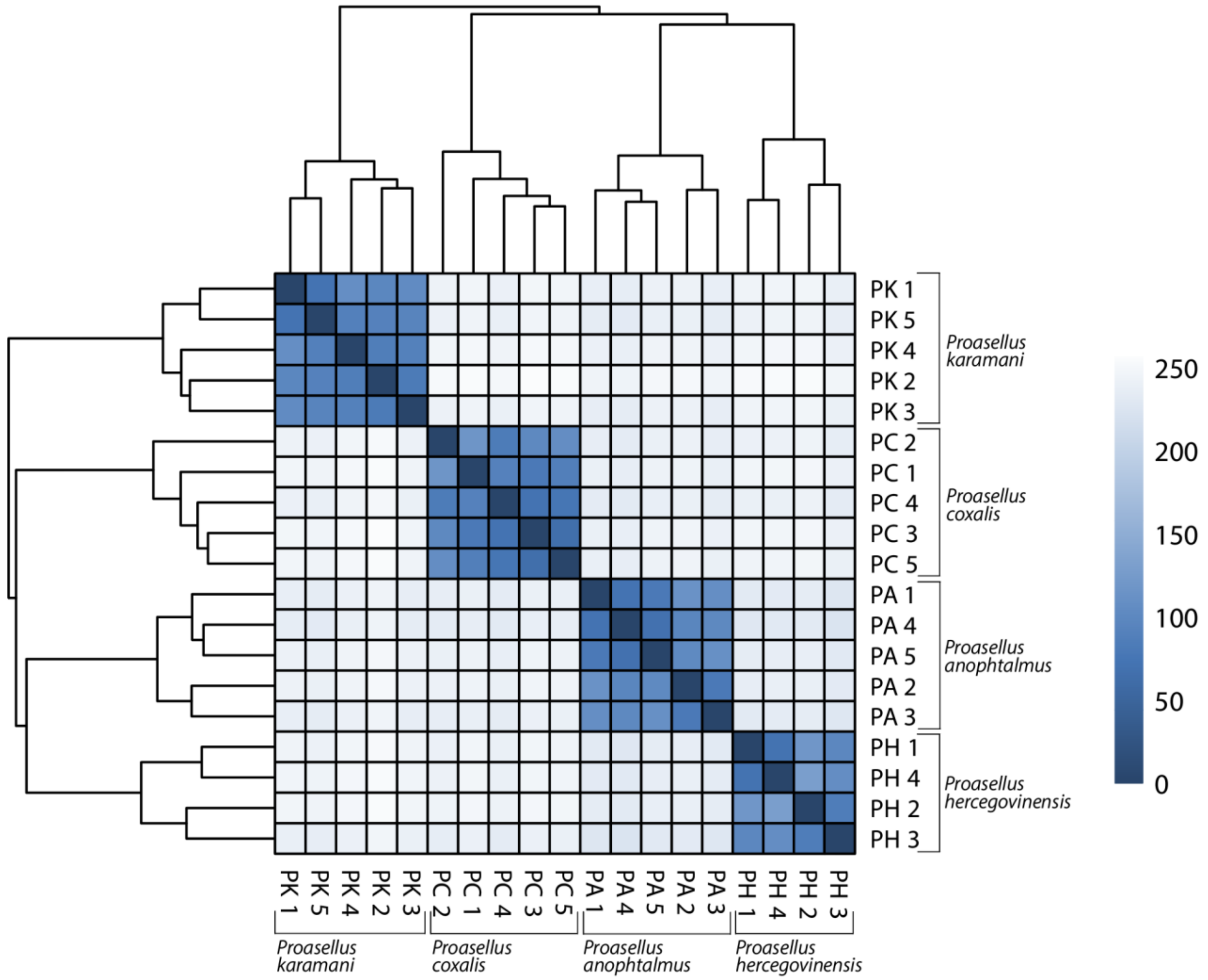


Figure 8 Sample-to-sample distance plot displaying distances between each sample with dark blue indicating more similarity, and light blue indicating the most distance. Each individual sample is denoted by a symbol indicating the species, and a number. The plot was produced by DESeq2.

Following the sample-to-sample distance plot, the PCA plot in *Figure 9* confirms a strong sample grouping within each species and a discrepancy among all four species. The plot displays the two first principal components: PC1 holding 38% of all variance, and PC2, holding 34% of all variance. The two components hold a nearly identical percentage of variance, implying the two components are crucial for species differentiation. The two cave species (*P. anophtalmus* and *P. hercegovinensis*) separate due to the PC2 component, while there is virtually no separation in the PC1 component. The two surface species (P. coxalis and P. karamani) display an opposite trend, differentiating according to the PC1 component. Accordingly, the cave species and surface species differentiate themselves with both of the principal components.

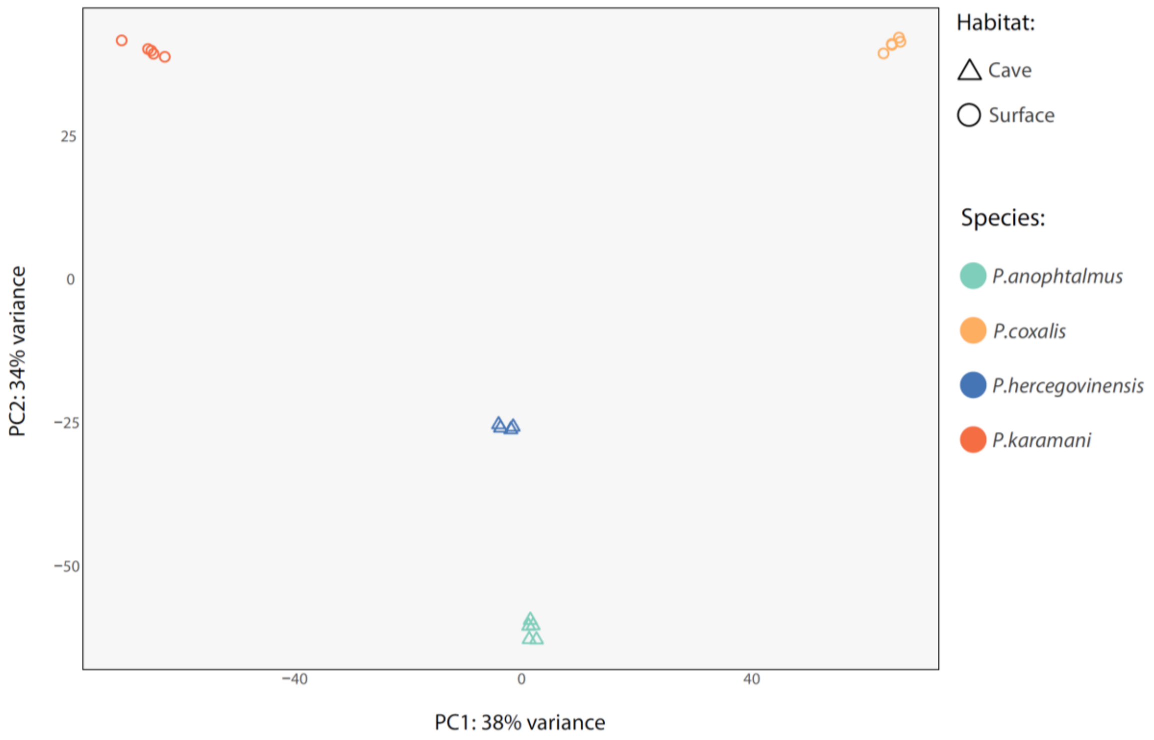


Figure 9 PCA plot showing the first two principal components (PC1 and PC2) displaying the clustering and the separation of the Proasellus species. Habitat is represented by shapes, and species by colours. The plot was produced with DESeq2.

4.3.2. Differentially expressed genes – statistics

The numbers of statistically significant differentially expressed genes between each two species are listed in *Table 9*. The numbers exceed 5000 genes in each case, with the number being the highest between *P. anophtalmus* and *P. karamani* and the lowest being between *P. anophtalmus* and *P. hercegovinensis*. The total number of genes compared was 9120 (the number of orthogroups present in all four species), indicating that 61% to 65% of all analysed genes are differentially expressed.

Table 9 Numbers of differentially expressed genes in pairwise comparisons between each two species.

Species	*Proasellus anophtalmus*	*Proasellus coxalis*	*Proasellus hercegovinensis*
Proasellus anophtalmus			
Proasellus coxalis	5539		
Proasellus hercegovinensis	5335	5539	
Proasellus karamani	5912	5849	5550

4.3.3. Differentially expressed genes of *P. anophtalmus*

A total of 134 annotated genes are uniquely up-regulated in *P. anophtalmus* compared to all other species. All of these 134 genes have an L2FC value of at least one in at least one of the three comparisons and a p-value of 0.05 or less in all instances. A list containing the top 50 up-regulated genes can be found in Supplement *9.2.4. (Table S2).*

Based on the GO terms (*Figure 10*), most of the up-regulated genes are associated with biological processes of anatomical morphogenesis (e.g., organ growth, adipose tissue development, mechanoreceptor differentiation), detection of chemical stimulus involved in sensory perception (e.g., adaptive immune response, circadian sleep/wake cycle, response to oxidative stress), and positive regulation of cellular component biogenesis (e.g., negative regulation of Notch signaling pathway). Two other major groups are the chondroitin sulfate proteoglycan metabolic process and cellular anatomical entity morphogenesis.

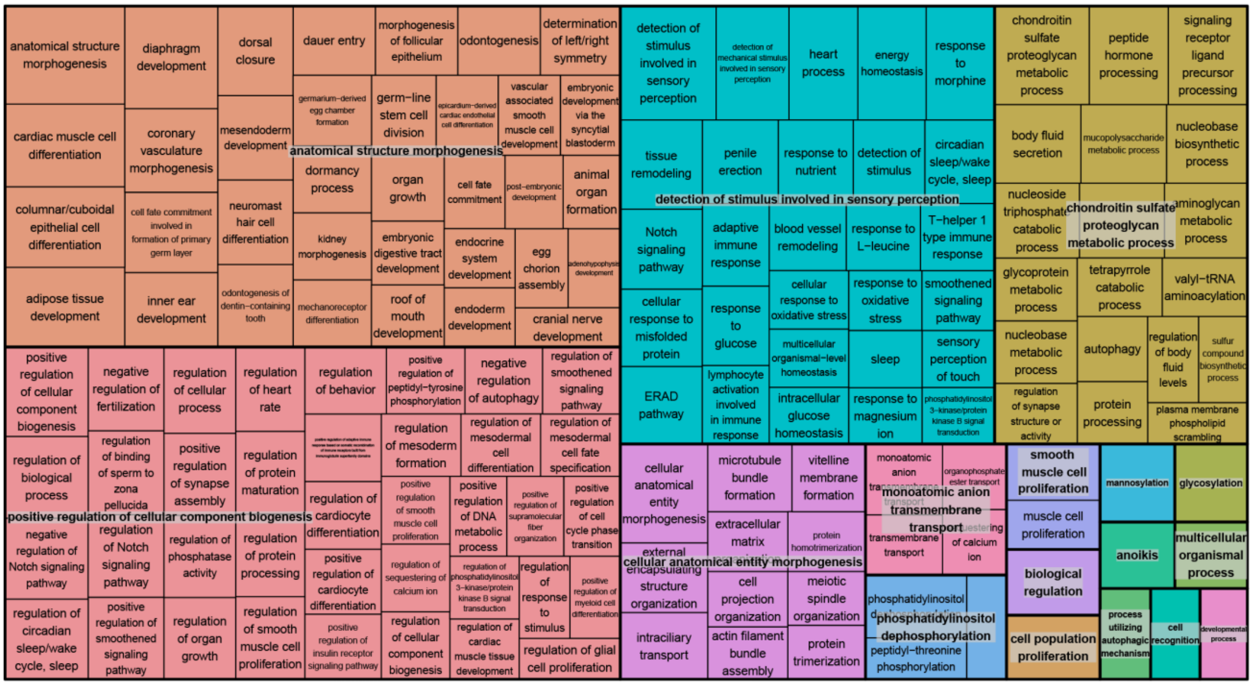


Figure 10 Tree map of gene ontology terms of up-regulated genes in P. anophtalmus associated with biological processes inferred by topGO. Tree map made with REVIGO.

When it comes to the cellular localisation of the up-regulated genes, it is evident that most of them localise in the plasma membrane and the cell periphery, but also as parts of various complexes such as the ubiquitin ligase complex (*Figure 11*).

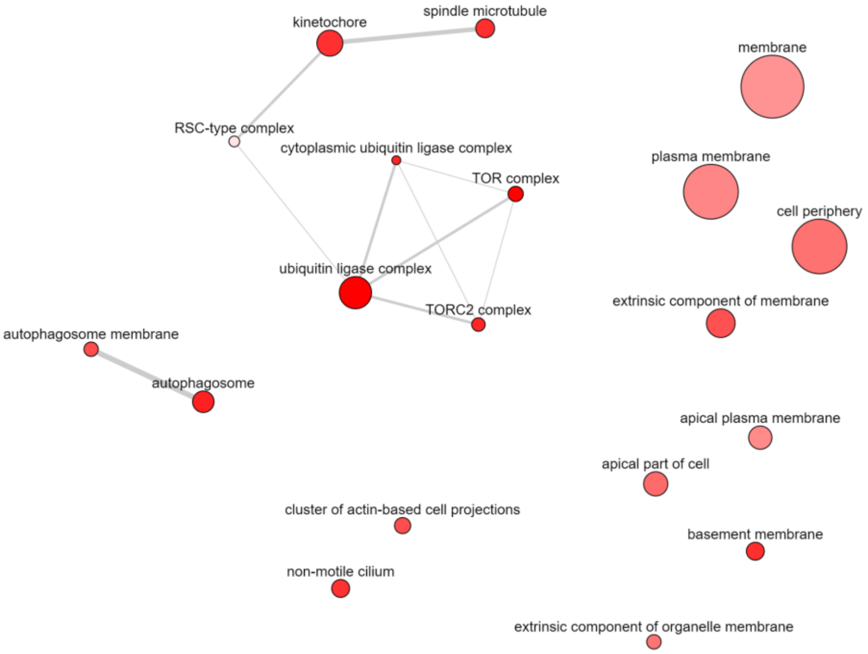


Figure 11 Graph of gene ontology cellular compartments terms of up-regulated genes in P. anophtalmus inferred by topGO. Tree map made with REVIGO.

The molecular functions of these genes range from glycosiltransferase activity and peptide transmembrane transporter activity to phosphoric ester hydrolase activity (*Figure 12*).

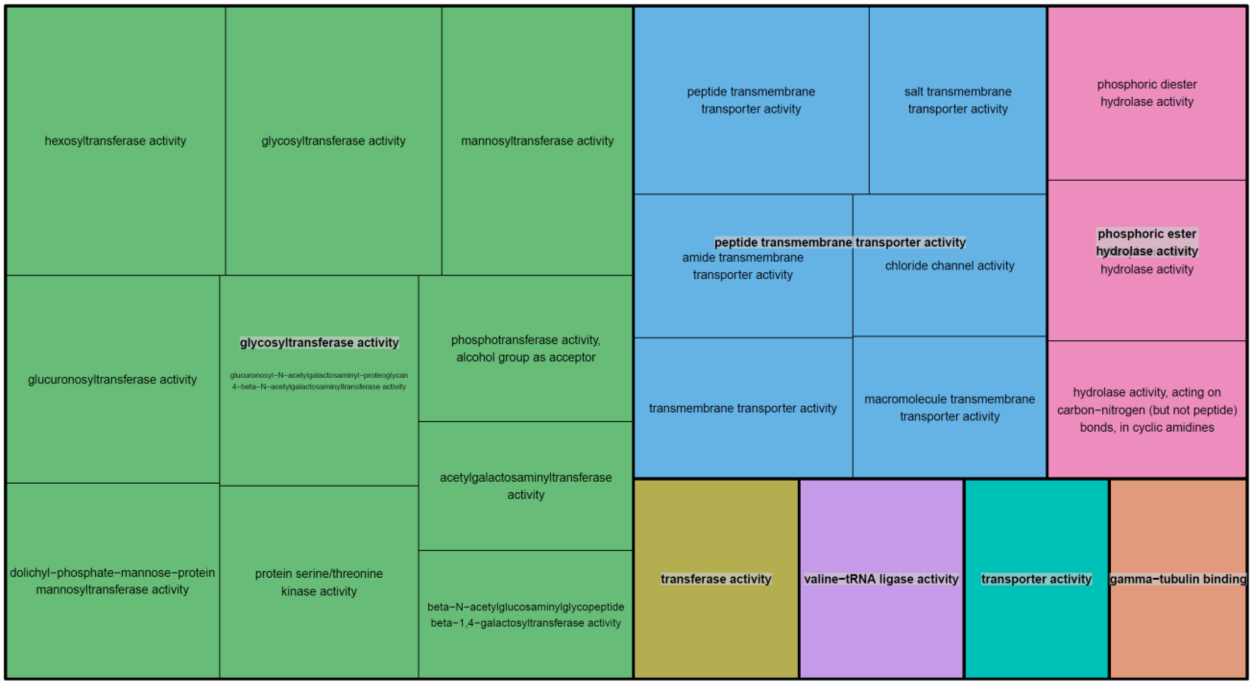


Figure 12 Tree map of gene ontology terms of up-regulated genes in P. anophtalmus associated with molecular functions inferred by topGO. Tree map made with REVIGO.

On the other hand, there are 643 annotated genes that are uniquely down-regulated in *P. anophtalmus* compared to all other species. All of these 643 genes have a L2FC value of at least one in at least one of the three comparisons, and a p-value of 0.05 or less in all instances. A list of genes with the highest expression levels is in the Supplement *9.2.4. (Table S3)*.

Biological processes related GO terms of down-regulated genes are shown in *Figure 13*, with major groups being translation (e.g., mRNA processing, RNA splicing, translational initiation), ribosome biogenesis and negative regulation of macromolecule biosynthetic process (e.g. regulation of amide metabolic process, cellular response to cAMP).

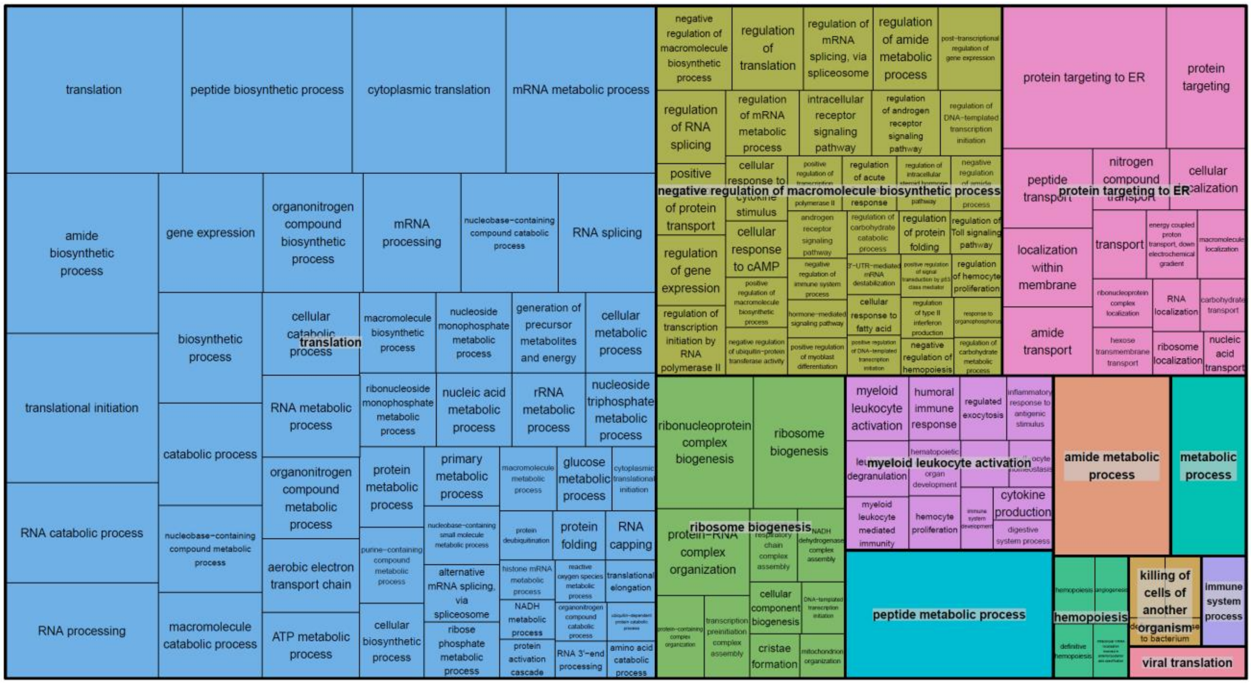


Figure 13 Tree map of gene ontology terms of down-regulated genes of P. anophtalmus associated with biological processes inferred by topGO. Tree map made with REVIGO.

A network of cellular compartments associated with the down-regulated genes is far more complex than the one for up-regulated genes, as seen in *Figure 14.* Down-regulated genes localise in various organelles like the mitochondrion or the nucleus, in the cytoplasm, vesicles, ribosomes, and many cellular complexes. Their molecular functions seem primarily associated with RNA and protein binding, peptidase regulation, transmembrane transport activity, or ribosome-associated (*Figure 15*).

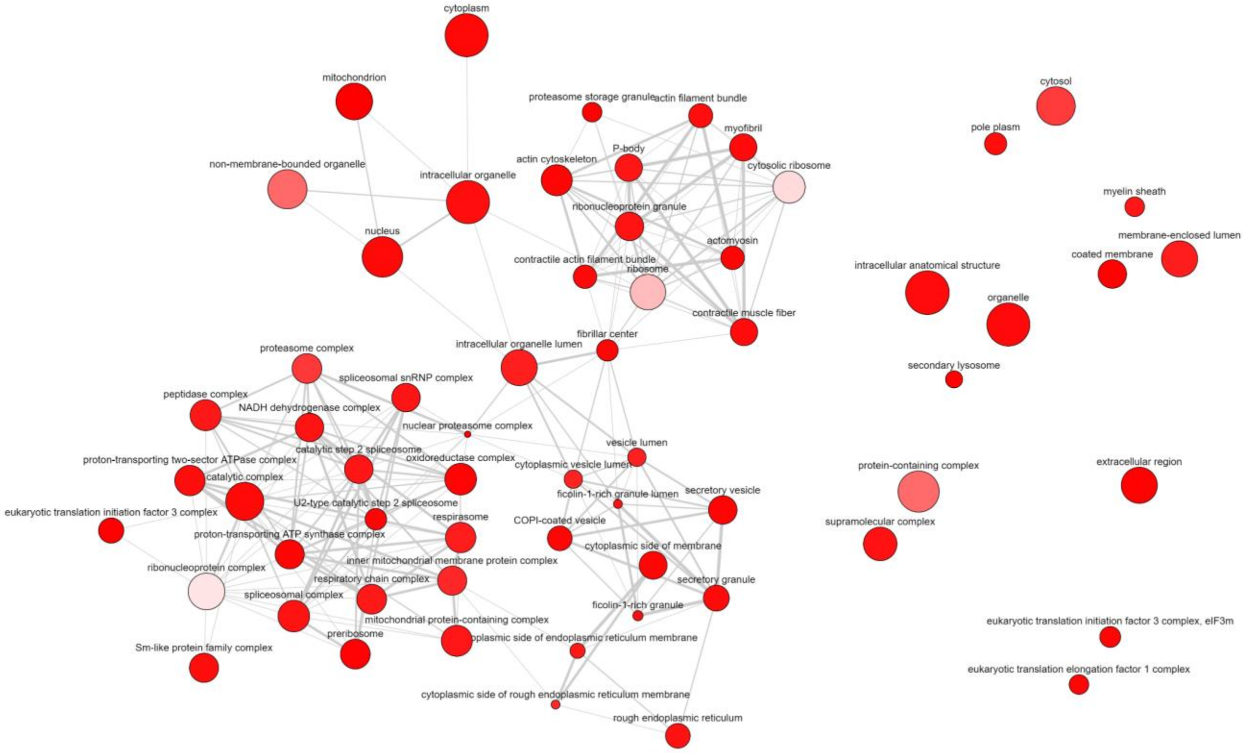

cytoplasm
mitochondrion
proteasome storage granule
actin filament bundle
myofibril
P-body
actin cytoskeleton
non-membrane-bounded organelle
intracellular organelle
cytosolic ribosome
ribonucleoprotein granule
nucleus
actomyosin
contractile actin filament bundle
ribosome
contractile muscle fiber
fibrillar center
intracellular organelle lumen
proteasome complex
spliceosomal snRNP complex
peptidase complex
NADH dehydrogenase complex
nuclear proteasome complex
catalytic step 2 spliceosome
oxidoreductase complex
vesicle lumen
proton-transporting two-sector ATPase complex
cytoplasmic vesicle lumen
catalytic complex
ficolin-1-rich granule lumen
secretory vesicle
U2-type catalytic step 2 spliceosome
eukaryotic translation initiation factor 3 complex
respirasome
COPI-coated vesicle
proton-transporting ATP synthase complex
cytoplasmic side of membrane
inner mitochondrial membrane protein complex
ribonucleoprotein complex
respiratory chain complex
secretory granule
spliceosomal complex
ficolin-1-rich granule
mitochondrial protein-containing complex
preribosome
oplasmic side of endoplasmic reticulum membrane
Sm-like protein family complex
cytoplasmic side of rough endoplasmic reticulum membrane
rough endoplasmic reticulum
cytosol
pole plasm
myelin sheath
membrane-enclosed lumen
coated membrane
intracellular anatomical structure
organelle
secondary lysosome
protein-containing complex
extracellular region
supramolecular complex
eukaryotic translation initiation factor 3 complex, eIF3m
eukaryotic translation elongation factor 1 complex

Figure 14 Graph of gene ontology cellular compartments terms of down-regulated genes in P. anophtalmus inferred by topGO. Tree map made with REVIGO.

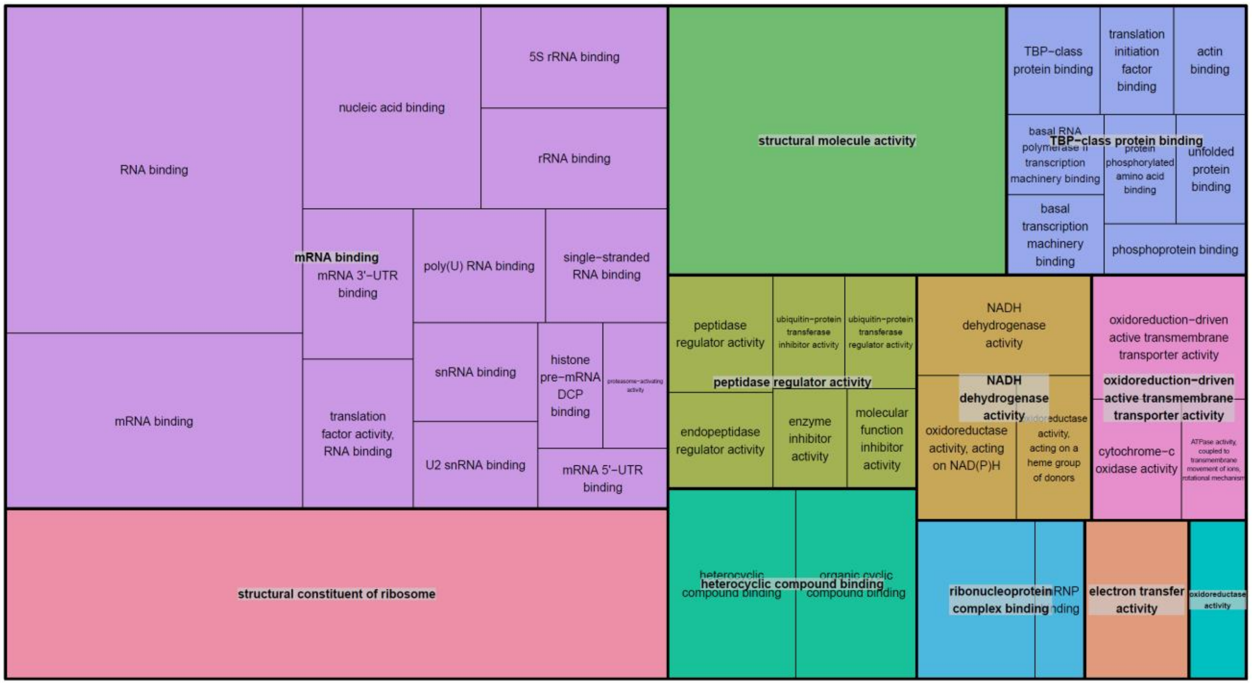


Figure 5 Tree map of gene ontology terms of down-regulated genes of P. anophtalmus associated with molecular functions inferred by topGO. Tree map made with REVIGO.

4.3.4. Differentially expressed genes of *P. hercegovinensis*

There is a total of 128 annotated genes that are up-regulated in *P. hercegovinensis* in comparison to the other three species. All of these genes have a L2FC of at least one, in at least one of the comparisons against other three species, and p-values of 0.05 or less. Genes with the highest L2FC values (top 50) are listed in Supplement *9.2.5. (Table S4)*.

A look into the GO terms of the 128 up-regulated genes puts them in four major categories based on the biological processes they are associated with (*Figure 16*). The categories are behavioural response to cocaine (e.g., habituation, behaviour), regulation of cellular response to hypoxia, myosin filament assembly and mitochondrial translation (e.g., mitochondrial gene expression).

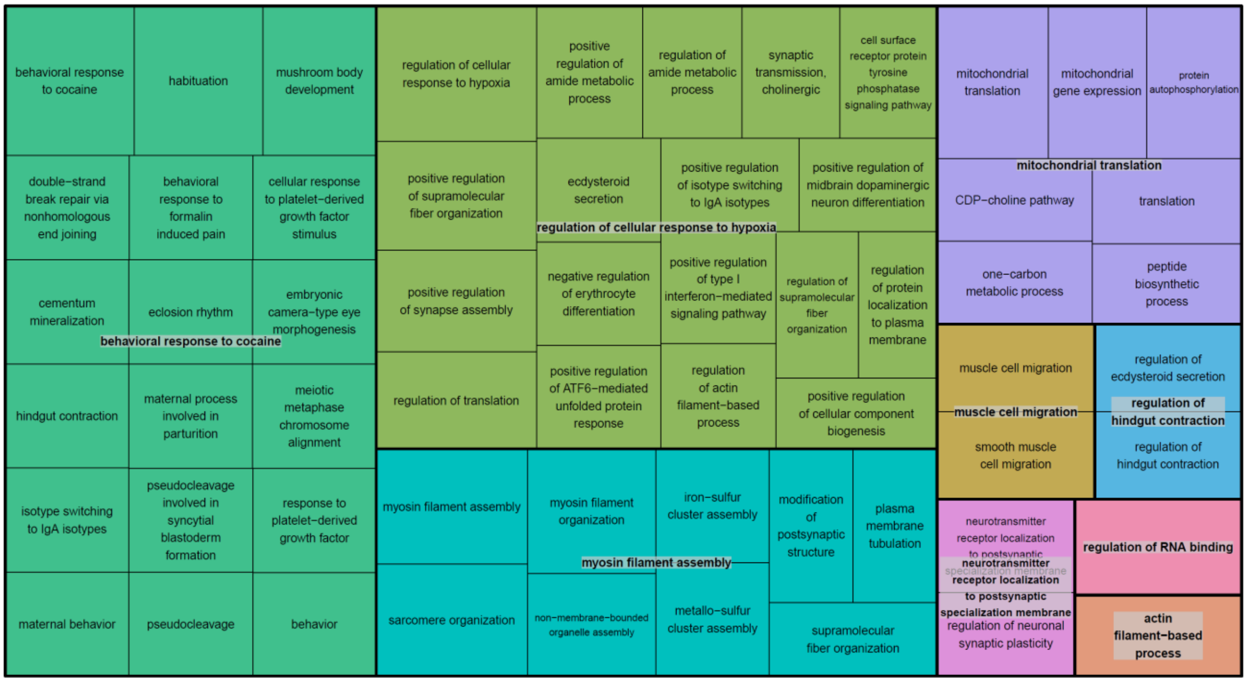


Figure 6 Tree map of gene ontology terms of up-regulated genes of P. hercegovinensis associated with biological processes inferred by topGO. Tree map made with REVIGO.

The activity of these genes is localised in supramolecular complexes, actin cytoskeleton, organellar ribosomes and other cellular compartments shown in *Figure 17.*

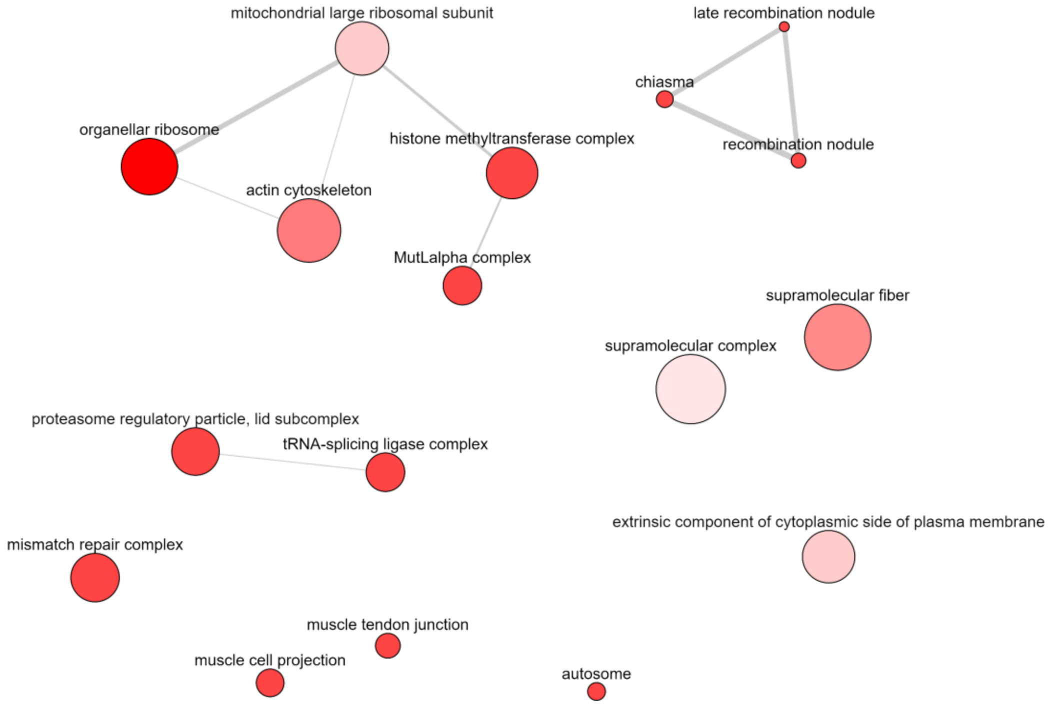

mitochondrial large ribosomal subunit
late recombination nodule
chiasma
organellar ribosome
histone methyltransferase complex
recombination nodule
actin cytoskeleton
MutLalpha complex
supramolecular fiber
supramolecular complex
proteasome regulatory particle, lid subcomplex
tRNA-splicing ligase complex
extrinsic component of cytoplasmic side of plasma membrane
mismatch repair complex
muscle tendon junction
muscle cell projection
autosome

Figure 7 Graph of gene ontology cellular compartments terms of up-regulated genes in P. hercegovinensis inferred by topGO. Tree map made with REVIGO.

Molecular functions of the up-regulated genes are separated in seven groups, the two major being heat shock protein binding and pre-mRNA binding (*Figure 18*).

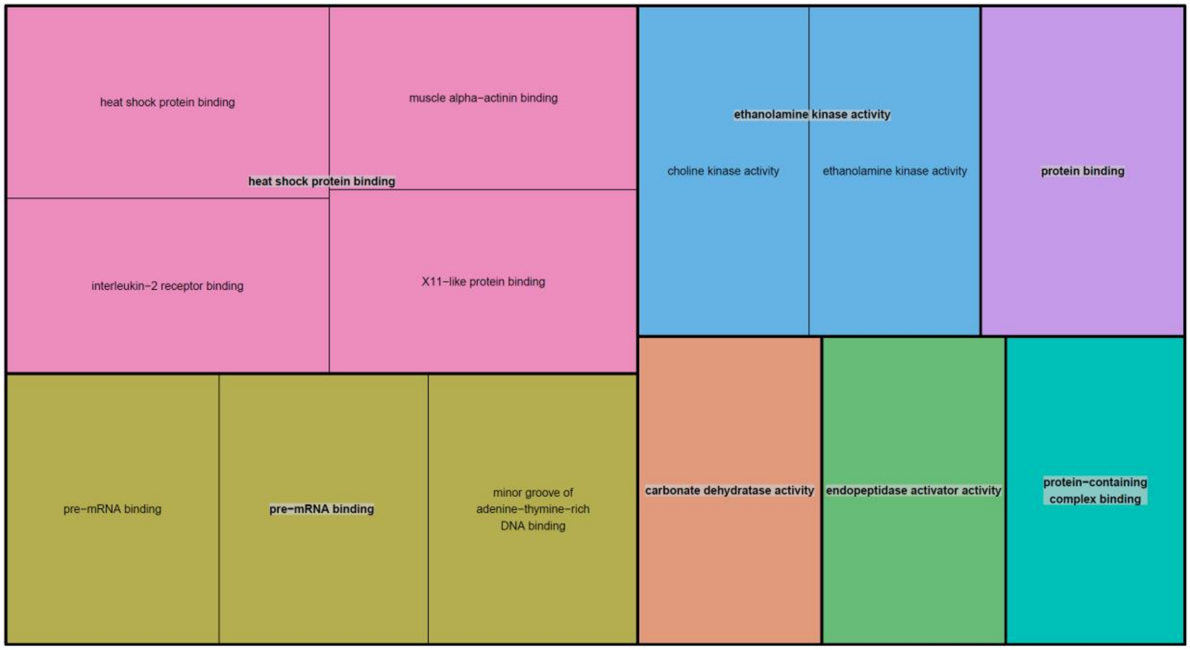


Figure 18 Tree map of gene ontology terms of up-regulated genes of P. hercegovinensis associated with molecular functions inferred by topGO. Tree map made with REVIGO.

A total of 93 annotated genes are down-regulated in *P. hercegovinensis* in comparison to the other three species. All of these genes have a L2FC of at least one, in at least one of the comparisons against other three species, and p-values of 0.05 or less. Genes with the highest L2FC values (top 50) are listed in Supplement *9.2.5 (Table S5).*

Most of these genes are associated with the biological processes of positive regulation of synapse maturation, transcription and tissue migration. Various terms are united under the GO term of positive regulation of synapse maturation, like the regulation of stem cell division, or positive regulation of pigment cell differentiation (*Figure 19*).

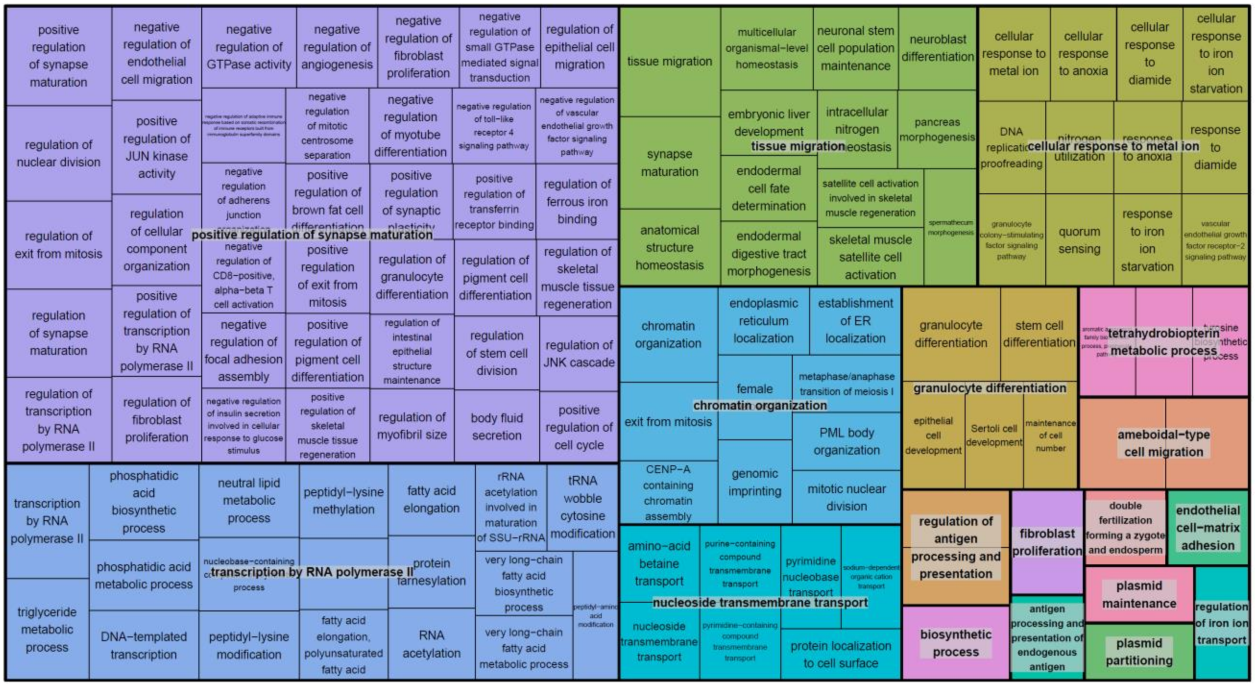


Figure 8 Tree map of gene ontology terms of down-regulated genes of P. hercegovinensis associated with biological processes inferred by topGO. Tree map made with REVIGO.

According to the cellular compartment GO terms, the down-regulated genes primarily localise in organelles, and chromosomes, but also in various complexes like the protein-DNA complex or transcription repression complex (*Figure 20*). The molecular functions of these genes are associated with histone binding, nucleic acid binding and acyltransferase activity (*Figure 21*).

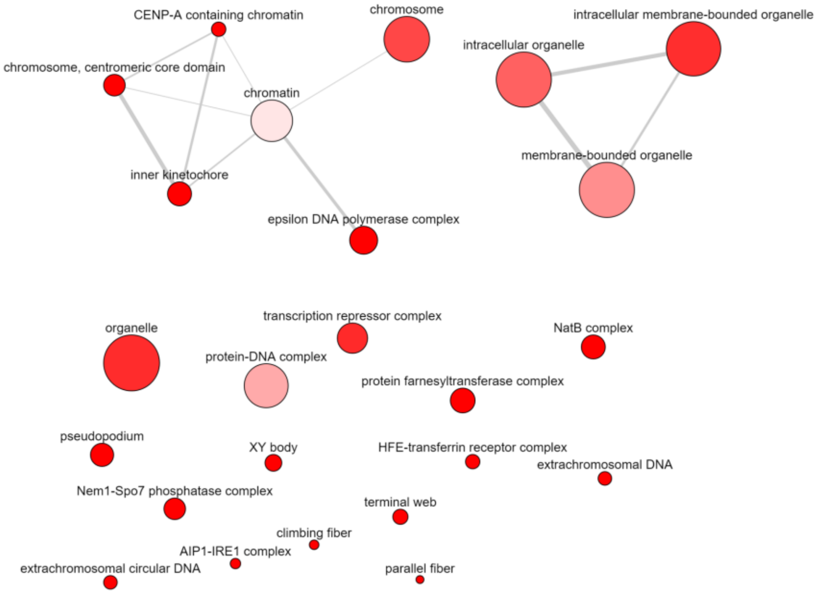

CENP-A containing chromatin
chromosome
intracellular membrane-bounded organelle
intracellular organelle
chromosome, centromeric core domain
chromatin
membrane-bounded organelle
inner kinetochore
epsilon DNA polymerase complex
organelle
transcription repressor complex
NatB complex
protein-DNA complex
protein farnesyltransferase complex
pseudopodium
XY body
HFE-transferrin receptor complex
extrachromosomal DNA
Nem1-Spo7 phosphatase complex
terminal web
climbing fiber
AIP1-IRE1 complex
extrachromosomal circular DNA
parallel fiber

Figure 20 Graph of gene ontology cellular compartments terms of down-regulated genes in P. hercegovinensis inferred by topGO. Tree map made with REVIGO.

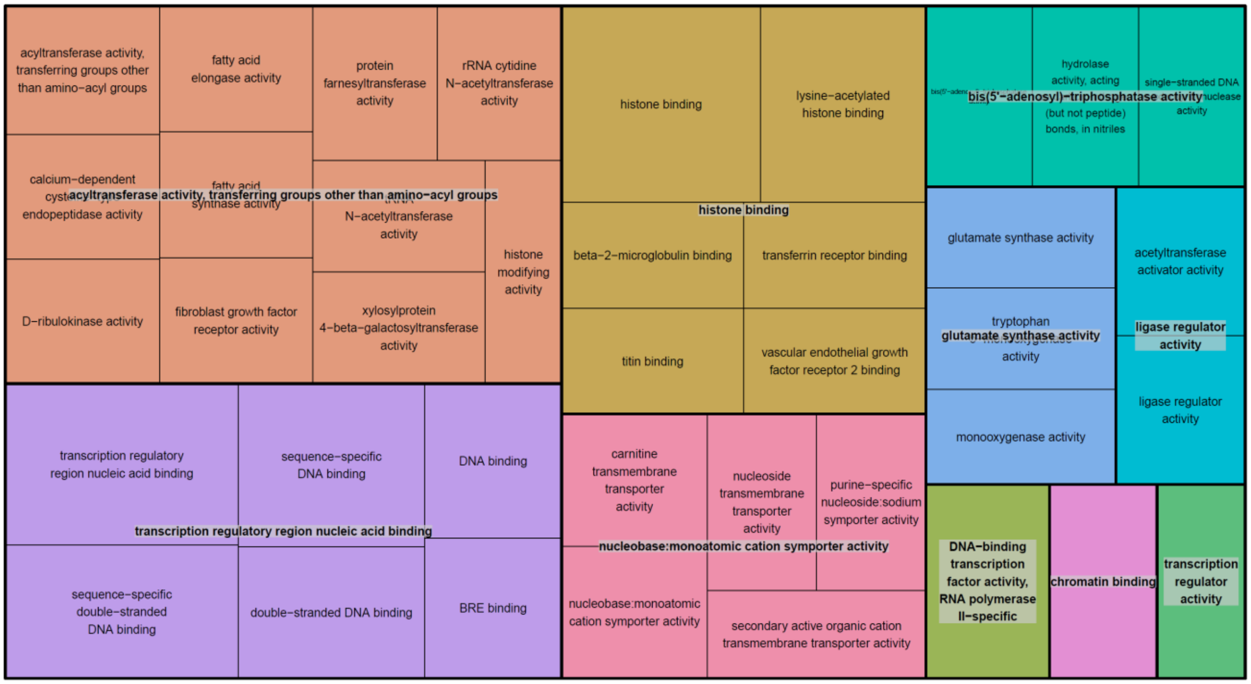


Figure 21 Tree map of gene ontology terms of down-regulated genes of P. hercegovinensis associated with molecular functions inferred by topGO. Tree map made with REVIGO.

4.3.5. Common gene expression patterns of *P. anophtalmus* and *P. hercegovinesis*

In total, 12 common genes are up-regulated in both cave species, *P. anophtalmus* and *P. hercegovinensis* when each is compared to both outside species (*P. coxalis* and *P. karamani*), with seven of them having a L2FC value above one in at least three out of four pairwise comparisons. Those seven genes are listed in *Table 10*. Some genes reach extremely high expression levels, like *C15orf48* that has 362 times higher expression levels in *P. hercegovinensis* compared to *P. karamani*. On the other hand, there are far more down-regulated genes which meet the criteria, a total of 77, which are listed in the Supplement *9.2.6.* (*Table S6)*.

Table 10 Commonly up-regulated genes of P. anophtalmus and P. hercegovinensis. This list contains an overlap of genes that are up-regulated both in P. anophtalmus and P. hercegovinensis in four pairwise comparisons (P. anophtalmus (PA) vs P. coxalis (PC), P. anophtalmus vs P. karamani (PK), P. hercegovinensis (PH) vs P. coxalis (PC), and P. hercegovinensis vs P. karamani (PK)). A sum of L2FC values determined the order of the list. A preferred name for the gene, a brief description of the function, and PFAM domains are also listed.

Gene ID/HOG ID	L2FC PA vs PC	L2FC PH vs PC	L2FC PA vs PK	L2FC PH vs PK	Preferred name	Description	PFAMs
N0.HOG0034425	3.462	2.161	2.437	8.509	C15orf48	Proton transmembrane transport	B12D
N0.HOG0009928	5.324	2.608	1.504	0.730	ATRN	Kelch motif	CUB, EGF_2, Kelch_1, Kelch_3, Kelch_4, Kelch_5, Kelch_6, Laminin_EGF, Lectin_C, PSI
N0.HOG0006722	2.103	1.017	1.889	4.579	IPO9	Ran GTPase binding	IBN_N, Xpo1
N0.HOG0006566	1.669	1.025	2.921	4.287	DESI2	PPPDE putative peptidase domain	Peptidase_C97
N0.HOG0024952	1.702	2.466	4.031	1.799	TTK	Broad-Complex, Tramtrack and Bric a brac	BTB, HTH_psq
N0.HOG0023553	3.090	1.688	1.862	2.365	BMP2	Transforming growth factor-beta (TGF-beta) family	TGF_beta, TGFb_propeptide
N0.HOG0016371	1.104	2.556	1.518	2.693	YARS2	Tyrosyl-tRNA synthetase	tRNA-synt_1b

Gene ontology terms associated with biological processes point out to several major functions of the down-regulated genes (*Figure 22*). They are enrolled in processes of NADH dehydrogenase complex assembly, many catabolic processes (e.g., glycoprotein catabolic process, insulin catabolic process, protein catabolic process) and general metabolic processes. Furthermore, there is a down-regulation of respiratory electron chain associated genes, as well as vision-related genes (post-embryonic eye development, eye pigment precursor transport, phototransduction). Many of these genes localise in the mitochondrion and its complexes, but also in Golgi-associated vesicles and membranes like the melanosome membrane (*Figure 23*).

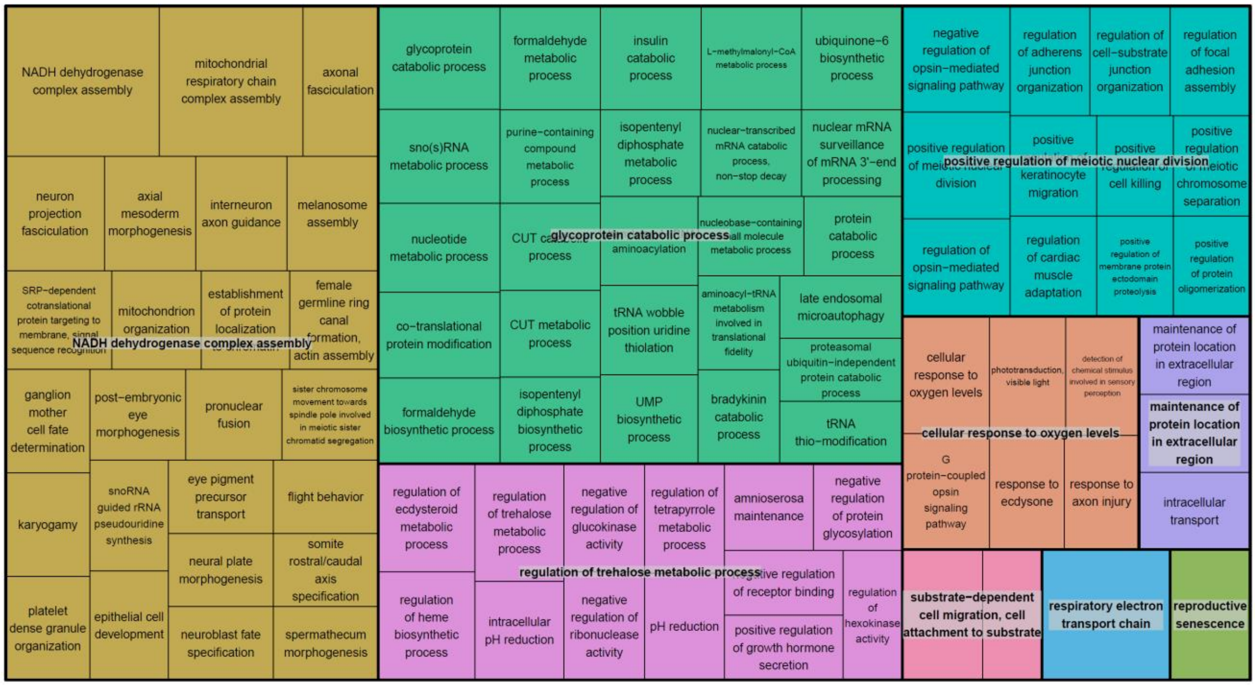


Figure 22 Tree map of gene ontology terms of common down-regulated genes of P. anophtalmus and P. hercegovinensis associated with biological processes inferred by topGO. Tree map made with REVIGO.

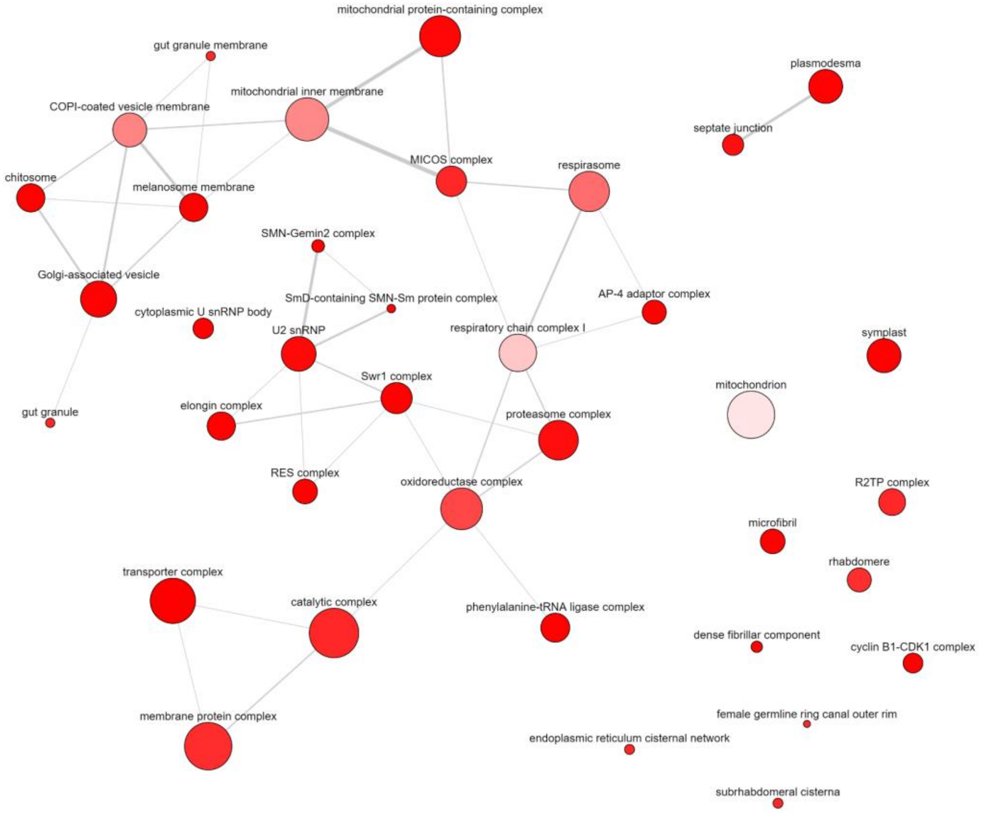


Figure 23 Graph of gene ontology cellular compartments terms of common down-regulated genes in P. anophtalmus and P. hercegovinensis inferred by topGO. Tree map made with REVIGO.

Molecular functions of the commonly down-regulated genes include nuclear receptor activity and associated functions, 7S RNA binding, chromatin binding, electron transfer activity, hormone binding and many others (*Figure 24*).

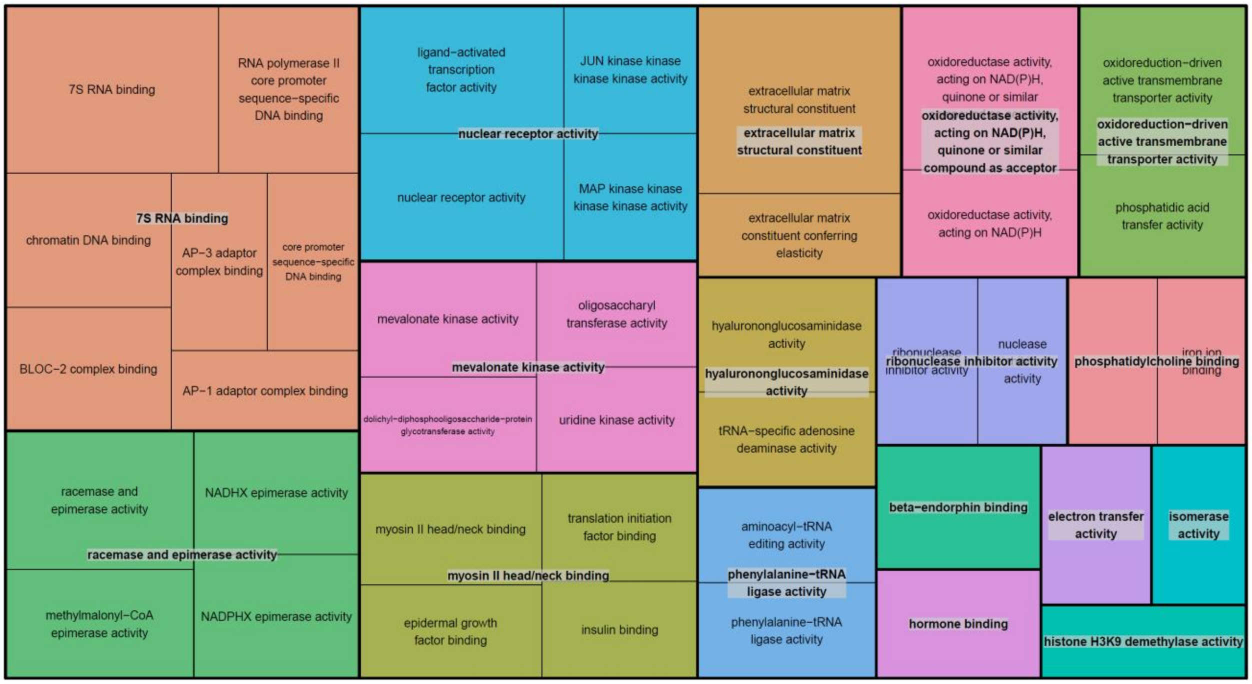


Figure 24 Tree map of gene ontology terms of common down-regulated genes of P. anophtalmus and P. hercegovinensis associated with molecular functions inferred by topGO. Tree map made with REVIGO.

5. Discussion

5.1. RNA-seq library construction and sequencing was successful regardless of small cuticle tissue amounts

Despite RNA yields from some individuals being as much as four times lower than the recommended amount of 200 ng (Wang et al., 2019) for RNA sequencing experiments, careful equimolar sample pooling of high-quality samples ensured a successful library construction. All libraries were of expected size and quality, thus proving that RNA extraction and subsequent cDNA library construction can be done from cuticular tissue of *Proasellus* isopods. Consequently, the quality of raw reads was very high as well, with over 92% of bases of each sample reaching the Phred score Q30 or above with less than 2% of reads being dropped after trimming and filtering. Sequence duplication levels were also relatively low, suggesting a higher complexity of cDNA libraries and sequenced reads.

Spike-in controls, artificial sequences added before advancing to the library generation process, signal a high uniformity of the samples and no bias towards a species. Even though there were some deviations from the expected concentrations of certain SIRV isoforms, the patterns of deviations were relatively uniform across all samples. This could suggest that during library preparation or sequencing some RNA sequences were underrepresented or entirely missing. Nevertheless, the high Pearson correlation coefficients observed in the ERCC correlation analysis suggest no apparent bias towards sequences based on their abundance. Even the ERCC sequences with very low concentrations were captured in the library preparation and sequencing processes. Thereafter, these controls prove to be a valuable asset when analysing RNA-seq data. This is especially the case in comparative transcriptomics when mRNA from different species needs to be compared. The uniformity of the results obtained by the spike-in analysis confirms the accuracy and reproducibility of the implemented methods and the comparability of the results between different species. There is, of course, room for improvement of the used methods, to try and capture more of the SIRV isoforms in their expected concentrations.

5.2. Cuticles yield transcriptomes of high completeness

Even though transcriptomes were assembled *de novo* solely from the cuticles of the individuals, according to BUSCO results, their completeness is very high compared to the Arthropoda database. Assembly statistics are uniform across all four transcriptomes, from median transcript length to N50 statistics. Maximum, minimum, and mean transcript lengths all match the values of whole-body transcriptomes of the same species assembled by Jovović et al. (2024). The N50 values particularly indicate the quality of the assembly. In all four transcriptomes, the N50 values were larger than the mean transcript lengths and also larger than the mean transcript lengths of independently assembled whole-body transcriptomes (Jovović et al., 2024), indicating a good assembly quality (Clarke et al., 2013). Additionally, high mapping rates of 89% and above are considered excellent (Clarke et al., 2013), and further confirm the good quality of the assembled transcriptomes, indicating low amounts of misassembled transcripts.

A large majority (87.8%) of the transcripts were assigned to orthogroups, but only a small amount (11.3%) were present in all four species and therefore useable for gene expression comparison. At the same time, with G50 of orthogroups being 8, this suggests that the majority of orthogroups might consist of intraspecies paralogues rather than interspecies orthologues. Considering the fact that transcriptomes were used in the analysis, the increased number of sequences per species could stem from multiple isoforms a certain gene can have, which were captured by sequencing, rather than being actual paralogues. Further analysis should be done with tools like PIC-Me, to differentiate isoforms and paralogues (Oh et al., 2021).

In addition, it is to be noted that *P. coxalis* and *P. karamani* have double the amounts of species-specific orthogroups compared to *P. anophtalmus* and P. *hercegovinensis* which also have the largest number of common orthogroups. This finding could indicate a larger evolutionary divergence of *P. coxalis* and P. *karamani* to all other compared species. Or, at the same time, allude to a close relatedness of *P. anophtalmus* and *P. hercegovinensis.*

The same kind of relationship between species is portrayed in the species tree. While the connection between *P. anophtalmus* and *P. hercegovinensis* was also confirmed in the species tree obtained by Jovović et al. (2024), the root of the *Asselidae* family in this study appears to be *P. coxalis* instead of *P. karamani*. This difference in species trees could arise from the fact that only cuticular mRNA was used in this experiment, compared to the whole-body mRNA, or because of the fact that the species tree in Jovović et al. (2024) was constructed with 11 instead of four species.

5.3. Adaptations to the cave environment

Arguably the best way to isolate genes responsible for adapting to the cave environment is by looking at overlaps between up-regulated genes of *P. anophtalmus* and *P. hercegovinensis* in their pairwise comparisons to the two surface species, and doing the same with the down-regulated genes. This should indicate genes which are up- or down-regulated as a result of living in a cave environment, rather than species-specific differential expression. There is a total of only seven such up-regulated genes, with consistently high L2FC values (one and above) in at least three out of four pairwise comparisons.

On the contrary, 77 down-regulated genes meet the same criteria, exactly 11 times more. This might not be that surprising, considering the fact that a lot of outside influences are eliminated in the cave environment, and the overall metabolism is slower (Hervant et al., 2002), lowering the demands for many gene products.

One such differentially expressed gene, with particularly high expression levels, is C15orf48, a mitochondrial protein with a function in modulating cytochrome *c* oxidase in complex IV of the electron transport chain. It has also been shown that it induced stress-independent autophagy, and regulated oxidative stress (Takakura et al., 2024). Reactive oxygen species (ROS) which induce oxidative stress, have been shown to be released from complex III in the electron transport chain during hypoxia, as a signal to trigger a response for the condition (Guzy et al., 2006). This could potentially explain the up-regulation of *C15orf48* which could serve as one of the mechanisms to mitigate the hypoxia-induced oxidative stress. Up-regulated genes of *P. hercegovinensis* show GO terms associated with response to hypoxia, giving additional confirmation to the observation. This could indicate a potential mechanism that cave-dwelling isopods use to deal with the lack of oxygen in subterranean waters.

Another interesting find is the up-regulation of *IPO9* gene that encodes the nuclear Importin-9. It has been found that Importin-9 functions as a storage chaperone for histones (H2A and H2B), it escorts them to the nucleus, but also sequesters them from DNA, hinting at a transcription regulation mediator function (Padavannil et al., 2019). Furthermore, Importin-9 has a role in proteasome import, which has been shown in *Drosophilla* where it has also been found to regulate chromosome segregation (Palacios et al., 2021). The maintenance of nuclear actin levels needed for transcriptional activity regulation is also mediated by Importin-9, which has been found to transport actin into the nucleus (Dopie et al., 2012). All of these findings suggest an important role of Importin-9, mediating transcriptional activity through histones, actin, and protein degradation. Its up-regulation could suggest a convergent mechanism of transcription regulation among the cave-dwelling *Proasellus* species. With emphasis on a potentially broad influence of Importin-9, it could be suggested that an up-regulation of one such gene can influence the activity of multiple pathways by mediating their expression levels.

5.3.1. *Attractin* – a “jack of all trades” in a cave environment?

Attractin, encoded by the *Atrn* gene is a widely expressed gene in vertebrates (Gunn et al., 1999). The protein contains a CUB domain, and multiple EGF and Plexin domains (He et al., 2001). While this protein is a single-transmembrane-domain glycoprotein (Gunn et al., 1999), a secreted isoform has been detected in humans with a regulatory role during an inflammatory reaction (Tang et al., 2000). Homologs of this protein have been found in invertebrates as well, suggesting an evolutionary conservation of the sequence and a pleiotropic role of the protein (He et al., 2001). As such, attractin has been found to serve multiple functions. Mutations in the *Atrn* gene led to a reduced body mass and adiposity as well as an increase in locomotor activity in mice carrying homozygous mutations (Gunn et al., 2001). If such an effect is present in invertebrates, or more specifically, cave isopods, it could have devastating consequences for an organism trying to navigate a nutrient-poor environment. On the contrary, an increase in the expression levels of an *Atrn* homolog in both *P. anophtalmnus* and *P. hercegovinensis* could suggest the opposite, less locomotor activity and better nutrient preservation and storage, all highly valuable in a scarce cave environment.

Furthermore, attractin seems to have a role in the central nervous system ensuring normal myelination (Kuramoto et al., 2001) while some findings suggest that attractin has a protective role against environmental toxins and helps prevent neurodegeneration (Paz et al., 2007). Again, if such role is present in the *Proasellus* attractin homologues, this gene's up-regulation would prove beneficial in a cave environment, especially focusing on toxin protection.

Additionally, *Atrn* appears to be expressed in the hair folicle melanocytes, while its expression is low in the non-pigmented cells. Mice with mutations in the *Atrn* gene appear to synthesize only eumelanin and no pheomelanin (He et al., 2001). Isopods use ommochrome pigments for their body pigmentation, however there are structures in their bodies which are melanized such as mouth parts (Jovović et al., unpublished). If a pigmentation-related function of *Atrn* is present in the cave isopods, it could provide additional insights into pigment rearrangements upon entering a cave environment.

Presuming the conservation of attractin functions in a *Proasellus* homolog, it is clear to see why it's up-regulation would be beneficial in a cave environment. With such a pleiotropic function in pigmentation, energy conservation, immune response and toxin protection, up-regulation of this single gene could have a highly beneficial role.

5.3.2. Focusing on the optimal sensory imputs with *Tramtrac*

Tramtrac is a transcription factor involved in a variety of biological processes together with other BTB-ZF transcription factors in the group (Kelly et al., 2006). BTB-ZF transcription factors act either as transcriptional activators or repressors, and are conserved across eukaryotes (Siggs et al., 2012). While PLZF, a human BTB-ZF acts as a tumour suppressor, these transcription factors have a different role in *Drosophila*, such as neurogenesis, metamorphosis and development of ovaries, by controlling cell proliferation and differentiation (Simon et al., 2019). Specifically, Tramtrac (*Ttk*) regulates the cell fate of cells in the peripheral nervous system by promoting them in non-neural development (Guo et al., 1995). It impacts cell proliferation and development in photoreceptors, intestinal stem cells and tracheal cells of *Drosophilla* (Simon et al., 2019) by working as a repressor (Brown et al., 1991). Two proteins are encoded by the *ttk* gene, Ttk69 and Ttk88 (Read et al., 1992). In this case Ttk69 is of more interest, as it has a role in cell cycle regulation. This, in turn, regulates the mitosis in the eye disc morphogenetic furrow of *Drosophila*. It has been shown that up-regulation of Ttk69 causes a complete stop of mitosis in the eye disc furrow (Baonza et al., 2002). Because of its highly conserved function (Siggs et al., 2012), it could be proposed that a similar effect is in place in *P. anophtalmus* and *P. hercegovinensis*, where its expression levels are 3 to 16 times as high compared to the surface-dwelling species. Restricting eye development can be advantageous in a cave environment where no light is present and vision holds no importance. This could conserve resources and energy, which is crucial in the nutrient-poor caves.

Much like attractin, tramtrac's up-regulation seems to harbour multiple benefits for the cave-dwelling isopods. Research has shown that a loss-of-function in the *ttk* gene transforms sensory cells into neurons, in the mechanosensory organs (Guo et al., 1995). Accordingly, Ttk69 determines the fate of progenitor cells, directing them to a non-neural fate during the mechanosensory bristles formation in *Drosophila* (Simon et al., 2019). It can be argued that the up-regulation of *Ttk* homolog in cave-dwelling *Proasellus* would promote more progenitor cells to become sensory cells. Since vision is useless in caves, animals have to rely on other senses to navigate their environment, implying that more developed and sensitive mechanosensory organs could indeed be beneficial. Again, up-regulation of a single gene indicates multiple benefits for an organism in a cave environment.

5.3.3. Down-regulation to slow down and conserve energy

Commonly down-regulated genes could suggest lower demands for certain metabolic processes or functions. For example, genes involved in mitochondrial function and therefore related to energy production e.g., *NDUFS8* (Wang et al., 2022), *NDUFA9* (Stroud et al., 2013), and *NDUFB11* (Amate-García et al., 2023) suggest lower demands for energy production due to the low-nutrient environment. Observed GO terms of down-regulated genes related to many metabolic processes and electron transport chain suggest the same. There are also three unannotated genes that seem to be cuticle-related with some of the lowest expression levels observed. Cuticular changes are a troglomorphic trait, with thinning observed in terrestrial cave isopods (Vittori et al., 2017). Furthermore, downregulation of *CDH23* is likely reflecting the reduced reliance on vision (Takahashi et al., 2016a) in the perpetual cave darkness, which is again supported by the vision-related GO terms. Additionally, slower growth rates of cave organisms could be connected to down-regulation of genes regulating cell cycle and division, like *CDK1* (Adhikari et al., 2012), and *CCDC86* (Stamatiou et al., 2023). All these genes point to a slower-paced life in a cave environment, where nutrients are poor, and vision unnecessary.

5.3.4. Divergent adaptation strategies of *Proasellus anophtalmus* and *Proasellus hercegovinensis*

Even though both *P. anophtalmus* and *P. hercegovinensis* are cave-dwellers which, according to the species tree, seem to be the closest relatives out of all four species analysed, these species show very different gene expression patterns. There is a total of 5335 genes differentially expressed between them, 134 genes uniquely up-regulated in *P. anophtalmus* against all other species, and 128 in *P. hercegovinensis.* Albeit being different species, and different expression patterns are to be expected, some of the mentioned genes could potentially allude to unique adaptations or strategies these species use in order to live in a cave environment. The GO terms analysed point to different gene groups, focusing on morphogenesis in *P. anophtalmus* and behaviour in *P. hercegovinensis.* Observations like this are more difficult to interpret, since it’s hard to connect these species-specific independent responses with a cave lifestyle, or with any specific environmental features of their respective caves, which are largely unknown. Further examination and analysis are needed to gain more insights into why certain genes show different expression patterns, and whether it harbours any benefits in the cave environment, or is a mere coincidence, compensation, or just a species-specific feature.

Nevertheless, it's interesting to note that some of the up-regulated terms in *P. anophtalmus* include adipose tissue development, mechanoreceptor differentiation, adaptive immune response, circadian sleep/wake cycle, and response to oxidative stress. Adipose tissue development indicates energy conservancy important to the cave environment, and mechanoreceptor differentiation confirms the switch from vision to other senses. In comparison, down-regulated terms are associated with splicing, mRNA processing, ribosome biogenesis and translational initiation. All of these indicate to lower (post)transcriptional activity and protein synthesis, pointing to a general metabolism slow-down as another mean of energy conservation. While the implications of up- and down-regulated terms of *P. hercegovinensis* are harder to determine, it has to be noted that a down-regulation of genes associated with positive regulation of pigment cell differentiation is present, potentially alluding to a mechanism of pigment loss in the cave-dweller.

5.3.5. A problem of pairwise comparisons

The entirety of this research is based on pairwise comparisons between each two species. While it is clear this approach produced a lot of results and provided insights into phenotype evolution, it can lead to unsupported conclusions. The problem of the approach is not including evolutionary relationships information in the analysis and can lead to statistical problems (Dunn et al., 2018). After examining multiple studies, Dunn and coworkers noticed that the results obtained by pairwise comparisons reflected evolutionary relationships between species, rather than supported a certain evolutionary process. They argue that more traits are shared between more closely related species, since they have a more recent common ancestor, which stands against the assumption of independence needed for statistical methods. A better approach for cross-species comparison could be the Expression Variance and Evolution (Eve) model which can analyse quantitative traits between species as well as within species. It is an orthology based, and phylogenetic ANOVA based model which deals with expression variance in a phylogenetic context (Rohlfs et al., 2015).

Since *P. anophtalmus* and *P. hercegovinensis* seem to be closely related, at least in the context of this research, the commonly up-regulated and down-regulated genes could be a consequence of their close relationship rather than an evolutionary process leading to adaptation. On the other hand, there are also 5335 differentially expressed genes between the two cave species which is not that much different than their comparisons with the two surface species (ranging from 5539 to 5912). Also, the observed L2FC values should indicate the importance of the up-regulation to such a high level, and should not be overlooked. The findings of this research show strong indications of benefits of up-regulated attractin and tramtrac genes both of which seem to have high expression levels in the two cave species. The analysis should be reproduced utilizing the Eve model in the context of phylogeny of the species. For this, a better phylogenetic tree is required, since the one obtained in this research, and the one obtained by Jovović et al., (2024) show discrepancies. Additionally, such an analysis would greatly benefit from complete genome assemblies of all four species which would ensure more precise mapping and gene counts, since reads would be mapped to genes instead of transcripts with many isoforms. Complete high-quality genomes would also help determine more precise ortholog and paralog gene relationships, especially considering that the orthology in this research was determined solely based on cuticular transcriptomes, which may have not captured the entire transcript (gene) diversity of the species.

6. Conclusion

The cave environment provides a unique system for tracking the course of evolution, from its "beginning" in surface dwellers to its "end" in highly adapted cave dwellers. The *Proasellus* genus fits well within that framework, which is why the two cave species and the two surface species were compared in this research.

Gene orthology analysis indicates that the two cave species, *P. anophtalmus* and *P. hercegovinensis* are the closest relatives among the four species, sharing the largest number of common orthogroups. Similarly, the PCA plot inferred from the differential gene expression analysis shows a clear distinction between the surface and cave dwellers, with smaller, yet noticeable, differences within each group. Although the patterns of differentially expressed genes vary between species, with hundreds of genes uniquely up- or down-regulated in each cave species, some common trends emerge among the cave dwellers. For instance, the up-regulation of *C15orf48*, a gene potentially involved in the hypoxic response to low oxygen environments, attractin, a gene regulating metabolism, energy conservation, toxin protection, and pigmentation, and tramtrac, a gene that may facilitate the transition from vision to mechanosensory reliance by halting eye development and promoting mechanosensory organ growth, are notable examples. Even though there is much more to be explored, and these findings need to be experimentally confirmed with a research focus on the functions of these genes in the *Proasellus* genus, these observations hint at a direction of convergent evolution and solutions for life in caves. Moreover, this study underscores the molecular basis of adaptation and neatly showcases how a change in expression levels of a certain gene pushes the adaptation in a direction beneficial for survival in a given environment. Interestingly, attractin and tramtrac seem to be pleiotropic, impacting more than one organismal function, pointing to an economic solution to multiple problems by an up-regulation of a single gene.

Even if these observations are not proven in the future, this research will still hold value by providing four new transcriptome assemblies and proving the cuticular tissue of isopods is suitable for RNA sequencing. It also emphasized the importance of using UMI sequences and spike-in controls in RNA sequencing experiments. Furthermore, the newly assembled transcriptomes gave more information about the phylogeny of the *Proasellus* genus and enabled a successful inference of orthologues between the four species. And lastly, this research gave a general insight into a comparative transcriptomics method, whether it is proven to be a good one, or not.

7. References

Adhikari, D., Zheng, W., Shen, Y., Gorre, N., Ning, Y., Halet, G., Kaldis, P., & Liu, K. (2012). Cdk1, but not Cdk2, is the sole Cdk that is essential and sufficient to drive resumption of meiosis in mouse oocytes. *Human Molecular Genetics*, *21*(11), 2476–2484. doi: 10.1093/HMG/DDS061

Alexa A, R. J. (2024). *Alexa A, Rahnenfuhrer J . topGO: Enrichment Analysis for Gene Ontology. R package version 2.56.0.*

Amate-García, G., Ballesta-Martínez, M. J., Serrano-Lorenzo, P., Garrido-Moraga, R., González-Quintana, A., Blázquez, A., Rubio, J. C., García-Consuegra, I., Arenas, J., Ugalde, C., Morán, M., Guillén-Navarro, E., & Martín, M. A. (2023). A Novel Mutation Associated with Neonatal Lethal Cardiomyopathy Leads to an Alternative Transcript Expression in the X-Linked Complex I NDUFB11 Gene. *International Journal of Molecular Sciences*, *24*(2). doi: 10.3390/IJMS24021743

Ashida, M., & Brey, P. T. (1995). Role of the integument in insect defense: pro-phenol oxidase cascade in the cuticular matrix. *Proceedings of the National Academy of Sciences of the United States of America*, *92*(23), 10698. doi: 10.1073/PNAS.92.23.10698

Audus, K. L., & Borchardt, R. T. (1986). Characteristics of the large neutral amino acid transport system of bovine brain microvessel endothelial cell monolayers. *Journal of Neurochemistry*, *47*(2), 484–488. doi: 10.1111/J.1471-4159.1986.TB04527.X

Baonza, A., Murawsky, C. M., Travers, A. A., & Freeman, M. (2002). Pointed and Tramtrack69 establish an EGFR-dependent transcriptional switch to regulate mitosis. *Nature Cell Biology 2002 4:12*, *4*(12), 976–980. doi: 10.1038/ncb887

Bilandžija, H., Ćetković, H., & Jeffery, W. R. (2012). Evolution of albinism in cave planthoppers by a convergent defect in the first step of melanin biosynthesis. *Evolution & Development*, *14*(2), 196. doi: 10.1111/J.1525-142X.2012.00535.X

Bilandžija, H., Laslo, M., Porter, M. L., & Fong, D. W. (2017). Melanization in response to wounding is ancestral in arthropods and conserved in albino cave species. *Scientific Reports 2017 7:1*, *7*(1), 1–11. doi: 10.1038/s41598-017-17471-2

Bilandžija, H., Ma, L., Parkhurst, A., & Jeffery, W. R. (2013). A potential benefit of albinism in Astyanax cavefish: downregulation of the oca2 gene increases tyrosine and catecholamine levels as an alternative to melanin synthesis. *PloS One*, *8*(11). doi: 10.1371/JOURNAL.PONE.0080823

Bininda-Emonds, O. R. P. (2005). transAlign: using amino acids to facilitate the multiple alignment of protein-coding DNA sequences. *BMC Bioinformatics*, *6*, 156. doi: 10.1186/1471-2105-6-156

Bolger, A. M., Lohse, M., & Usadel, B. (2014a). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, *30*(15), 2114. doi: 10.1093/BIOINFORMATICS/BTU170

Bolger, A. M., Lohse, M., & Usadel, B. (2014b). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, *30*(15), 2114. doi: 10.1093/BIOINFORMATICS/BTU170

Brown, J. L., Sonoda2, S., Ueda"', H., Scott214, M. P., & Wu1', C. (1991). Repression of the Drosophila fushi tarazu (ftz) segmentation gene. *The EMBO Journal*, *10*(3), 665–674. doi: 10.1002/J.1460-2075.1991.TB07995.X

Buchfink, B., Reuter, K., & Drost, H. G. (2021). Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nature Methods 2021 18:4*, *18*(4), 366–368. doi: 10.1038/s41592-021-01101-x

Cantalapiedra, C. P., Herņandez-Plaza, A., Letunic, I., Bork, P., & Huerta-Cepas, J. (2021). eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale. *Molecular Biology and Evolution*, *38*(12), 5825–5829. doi: 10.1093/MOLBEV/MSAB293

Cigna, A. A. (2002). Modern Trend in Cave Monitoring. *Acta Carsologica*, *31*(1). doi: 10.3986/AC.V31I1.402

Clarke, K., Yang, Y., Marsh, R., Xie, L. L., & Zhang, K. K. (2013). Comparative analysis of de novo transcriptome assembly. *Science China Life Sciences*, *56*(2), 156–162. doi: 10.1007/S11427-013-4444-X/METRICS

Culver, D. C., & Pipan, T. (2015). Shifting Paradigms of the Evolution of Cave Life. *Acta Carsologica*, *44*(3), 415–425. doi: 10.3986/AC.V44I3.1688

Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., Whitwham, A., Keane, T., McCarthy, S. A., & Davies, R. M. (2021). Twelve years of SAMtools and BCFtools. *GigaScience*, *10*(2), 1–4. doi: 10.1093/GIGASCIENCE/GIAB008

Deleo, D. M., Pérez-Moreno, J. L., Vázquez-Miranda, H., & Bracken-Grissom, H. D. (2018). RNA profile diversity across arthropoda: guidelines, methodological artifacts, and expected outcomes. *Biology Methods & Protocols*, *3*(1). doi: 10.1093/BIOMETHODS/BPY012

Derkarabetian, S., Steinmann, D. B., & Hedin, M. (2010). Repeated and Time-Correlated Morphological Convergence in Cave-Dwelling Harvestmen (Opiliones, Laniatores) from Montane Western North America. *PLOS ONE*, *5*(5), e10388. doi: 10.1371/JOURNAL.PONE.0010388

Dopie, J., Skarp, K. P., Rajakylä, E. K., Tanhuanpää, K., & Vartiainen, M. K. (2012). Active maintenance of nuclear actin by importin 9 supports transcription. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(9), E544–E552. doi: 10.1073/PNAS.1118880109/SUPPL_FILE/PNAS.1118880109_SI.PDF

Duboué, E. R., Keene, A. C., & Borowsky, R. L. (2011). Evolutionary convergence on sleep loss in cavefish populations. *Current Biology : CB*, *21*(8), 671–676. doi: 10.1016/J.CUB.2011.03.020

Ducrest, A. L., Keller, L., & Roulin, A. (2008). Pleiotropy in the melanocortin system, coloration and behavioural syndromes. *Trends in Ecology & Evolution*, *23*(9), 502–510. doi: 10.1016/J.TREE.2008.06.001

Dunn, C. W., Zapata, F., Munro, C., Siebert, S., & Hejnol, A. (2018). Pairwise comparisons across species are problematic when analyzing functional genomic data. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(3), E409–E417. doi: 10.1073/PNAS.1707515115/SUPPL_FILE/PNAS.1707515115.SD01.PDF

Emms, D. M., & Kelly, S. (2019). OrthoFinder: Phylogenetic orthology inference for comparative genomics. *Genome Biology*, *20*(1), 1–14. doi: 10.1186/S13059-019-1832-Y/FIGURES/5

Fernstrom, J. D., & Fernstrom, M. H. (2007). Tyrosine, phenylalanine, and catecholamine synthesis and function in the brain. *The Journal of Nutrition*, *137*(6 Suppl 1). doi: 10.1093/JN/137.6.1539S

Graham, D. G., Tiffany, S. M., & Vogel, F. S. (1978). The toxicity of melanin precursors. *The Journal of Investigative Dermatology*, *70*(2), 113–116. doi: 10.1111/1523-1747.EP12541249

Gunn, T. M., Inui, T., Kitada, K., Ito, S., Wakamatsu, K., He, L., Bouley, D. M., Serikawa, T., & Barsh, G. S. (2001). Molecular and Phenotypic Analysis of Attractin Mutant Mice. *Genetics*, *158*(4), 1683–1695. doi: 10.1093/GENETICS/158.4.1683

Gunn, Teresa M., Miller, K. A., He, L., Hyman, R. W., Davis, R. W., Azarani, A., Schlossman, S. F., Duke-Cohan, J. S., & Barsh, G. S. (1999). The mouse mahogany locus encodes a transmembrane form of human attractin. *Nature 1999 398:6723*, *398*(6723), 152–156. doi: 10.1038/18217

Guo, M., Bier, E., Yeh Jan, L., & Nung Jan, Y. (1995). tramtrack Acts Downstream of numb to Specify Distinct Daughter Cell Fates during Asymmetric Cell Divisions in the Drosophila PNS. In Neuron (Vol. 14).

Guzy, R. D., & Schumacker, P. T. (2006). Oxygen sensing by mitochondria at complex III: the paradox of increased reactive oxygen species during hypoxia. *Experimental Physiology*, *91*(5), 807–819. doi: 10.1113/EXPPHYSIOL.2006.033506

Haas, BJ. (n.d.). *TransDecoder*.

He, L., Gunn, T. M., Bouley, D. M., Lu, X. Y., Watson, S. J., Schlossman, S. F., Duke-Cohan, J. S., & Barsh, G. S. (2001). A biochemical function for attractin in agouti-induced pigmentation and obesity. *Nature Genetics 2001 27:1*, *27*(1), 40–47. doi: 10.1038/83741

Heather, J. M., & Chain, B. (2016). The sequence of sequencers: The history of sequencing DNA. *Genomics*, *107*(1), 1. doi: 10.1016/J.YGENO.2015.11.003

Hervant, F., Mathieu, J., & Messana, G. (1998). Oxygen Consumption and Ventilation in Declining Oxygen Tension and Posthypoxic Recovery in Epigean and Hypogean Crustaceans. *Journal of Crustacean Biology*, *18*(4), 717–727. doi: 10.1163/193724098X00593

Hervant, Frédéric, Mathieu, J., Barré, H., Simon, K., & Pinon, C. (1997). Comparative study on the behavioral, ventilatory, and respiratory responses of hypogean and epigean crustaceans to long-term starvation and subsequent feeding. *Comparative Biochemistry and Physiology Part A: Physiology*, *118*(4), 1277–1283. doi: 10.1016/S0300-9629(97)00047-9

Hervant, Frédéric, & Renault, D. (2002). Long-term fasting and realimentation in hypogean and epigean isopods: a proposed adaptive strategy for groundwater organisms. *Journal of Experimental Biology*, *205*(14), 2079–2087. doi: 10.1242/JEB.205.14.2079

Howarth, F. G. (1993). High-stress subterranean habitats and evolutionary change in cave-inhabiting arthropods. *The American Naturalist*, *142 Suppl 1*(Suppl.). doi: 10.1086/285523

Howarth, Francis G., & Moldovan, O. T. (2018). *The Ecological Classification of Cave Animals and Their Adaptations*. 41–67. doi: 10.1007/978-3-319-98852-8_4

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. doi: 10.1093/NAR/GKY1085

Jeffery, W. R., & Martasian, D. P. (1998). Evolution of Eye Regression in the Cavefish Astyanax: Apoptosis and the Pax-6 Gene. *Integrative and Comparative Biology*, *38*(4), 685–696. doi: 10.1093/ICB/38.4.685

Jovović, L., Bedek, J., Malard, F., & Bilandžija, H. (2024). De novo transcriptomes of cave and surface isopod crustaceans: insights from 11 species across three suborders. *Scientific Data 2024 11:1*, *11*(1), 1–11. doi: 10.1038/s41597-024-03393-y

Kelly, K. F., & Daniel, J. M. (2006). POZ for effect - POZ-ZF transcription factors in cancer and development. *Trends in Cell Biology*, *16*(11), 578–587. doi: 10.1016/j.tcb.2006.09.003

Krishnan, J., & Rohner, N. (2017). Cavefish and the basis for eye loss. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *372*(1713). doi: 10.1098/RSTB.2015.0487

Kuramoto, T., Kitada, K., Inui, T., Sasaki, Y., Ito, K., Hase, T., Kawagachi, S., Ogawa, Y., Nakao, K., Barsh, G. S., Nagao, M., Ushijima, T., & Serikawa, T. (2001). Attractin/mahogany/zitter plays a critical role in myelination of the central nervous system. *Proceedings of the National Academy of Sciences of the United States of America*, *98*(2), 559–564. doi: 10.1073/PNAS.98.2.559/ASSET/FE3A5685-AFBD-4CEC-B3A8-313CDF851055/ASSETS/GRAPHIC/PQ0214663005.JPEG

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods 2012 9:4*, *9*(4), 357–359. doi: 10.1038/nmeth.1923

Li, B., & Dewey, C. N. (2011). RSEM: Accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*, *12*(1), 1–16. doi: 10.1186/1471-2105-12-323/TABLES/6

Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, *15*(12), 1–21. doi: 10.1186/S13059-014-0550-8/FIGURES/9

Lukic, M., Jovovic, L., Bedek, J., Grgic, M., Kuharic, N., Rozman, T., Cupic, I., Weck, B., Fong, D., & Bilandzija, H. (2024). A practical guide for the husbandry of cave and surface invertebrates as the first step in establishing new model organisms. *PLOS ONE*, *19*(4), e0300962. doi: 10.1371/JOURNAL.PONE.0300962

Macalady, J. L., Jones, D. S., & Lyon, E. H. (2007). Extremely acidic, pendulous cave wall biofilms from the Frasassi cave system, Italy. *Environmental Microbiology*, *9*(6), 1402–1414. doi: 10.1111/J.1462-2920.2007.01256.X

Manni, M., Berkeley, M. R., Seppey, M., Simão, F. A., & Zdobnov, E. M. (2021). BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes. *Molecular Biology and Evolution*, *38*(10), 4647–4654. doi: 10.1093/MOLBEV/MSAB199

Manni, M., Berkeley, M. R., Seppey, M., & Zdobnov, E. M. (2021). BUSCO: Assessing Genomic Data Quality and Beyond. *Current Protocols*, *1*(12), e323. doi: 10.1002/CPZ1.323

McCauley, D. W., Hixon, E., & Jeffery, W. R. (2004). Evolution of pigment cell regression in the cavefish Astyanax: a late step in melanogenesis. *Evolution & Development*, *6*(4), 209–218. doi: 10.1111/J.1525-142X.2004.04026.X

Moran, D., Softley, R., & Warrant, E. J. (2015). The energetic cost of vision and the evolution of eyeless Mexican cavefish. *Science Advances*, *1*(8). doi: 10.1126/SCIADV.1500363

Oh, J., Lee, S. G., & Park, C. (2021). PIC-Me: paralogs and isoforms classifier based on machine-learning approaches. *BMC Bioinformatics*, *22*(Suppl 11). doi: 10.1186/S12859-021-04229-X

Padavannil, A., Sarkar, P., Kim, S. J., Cagatay, T., Jiou, J., Brautigam, C. A., Tomchick, D. R., Sali, A., D'Arcy, S., & Chook, Y. M. (2019). Importin-9 wraps around the H2A-H2B core to act as nuclear importer and histone chaperone. *ELife*, *8*. doi: 10.7554/ELIFE.43630

Palacios, V., Kimble, G. C., Tootle, T. L., & Buszczak, M. (2021). Importin-9 regulates chromosome segregation and packaging in Drosophila germ cells. *Journal of Cell Science*, *134*(7). doi: 10.1242/JCS.258391/237786

Paz, J., Yao, H., Lim, H. S., Lu, X. Y., & Zhang, W. (2007). The neuroprotective role of attractin in neurodegeneration. *Neurobiology of Aging*, *28*(9), 1446–1456. doi: 10.1016/J.NEUROBIOLAGING.2006.06.014

Protas, M. E., Hersey, C., Kochanek, D., Zhou, Y., Wilkens, H., Jeffery, W. R., Zon, L. I., Borowsky, R., & Tabin, C. J. (2006). Genetic analysis of cavefish reveals molecular convergence in the evolution of albinism. *Nature Genetics*, *38*(1), 107–111. doi: 10.1038/NG1700

Protas, M., & Jeffery, W. R. (2012). Evolution and development in cave animals: from fish to crustaceans. *Wiley Interdisciplinary Reviews. Developmental Biology*, *1*(6), 823–845. doi: 10.1002/WDEV.61

Read, D., & Manley, J. L. (1992). Alternatively spliced transcripts of the Drosophila tramtrack gene encode zinc finger proteins with distinct DNA binding specificities. *The EMBO Journal*, *11*(3), 1035–1044. doi: 10.1002/J.1460-2075.1992.TB05142.X

RNA LEXICON | Lexogen. (n.d.). Retrieved from https://www.lexogen.com/rna-lexicon/

Rohlfs, R. V., & Nielsen, R. (2015). Phylogenetic ANOVA: The Expression Variance and Evolution Model for Quantitative Trait Evolution. *Systematic Biology*, *64*(5), 695–708. doi: 10.1093/SYSBIO/SYV042

Sample Quality Control, Electrophoresis, Bioanalyzer | Agilent. (n.d.). Retrieved from https://www.agilent.com/en/product/automated-electrophoresis/bioanalyzer-systems/bioanalyzer-instrument/2100-bioanalyzer-instrument-228250

Sanger, F., Nicklen, S., & Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America*, *74*(12), 5463–5467. doi: 10.1073/PNAS.74.12.5463

Satam, H., Joshi, K., Mangrolia, U., Waghoo, S., Zaidi, G., Rawool, S., Thakare, R. P., Banday, S., Mishra, A. K., Das, G., & Malonia, S. K. (2023). Next-Generation Sequencing Technology: Current Trends and Advancements. *Biology 2023, Vol. 12, Page 997*, *12*(7), 997. doi: 10.3390/BIOLOGY12070997

Siggs, O. M., & Beutler, B. (2012). The BTB-ZF transcription factors. *Cell Cycle*, *11*(18), 3358–3369. doi: 10.4161/CC.21277

Simon, F., Ramat, A., Louvet-Vallée, S., Lacoste, J., Burg, A., Audibert, A., & Gho, M. (2019). Shaping of Drosophila Neural Cell Lineages Through Coordination of Cell Proliferation and Cell Fate by the BTB-ZF Transcription Factor Tramtrack-69. *Genetics*, *212*(3), 773–788. doi: 10.1534/GENETICS.119.302234

Smith, T., Heger, A., & Sudbery, I. (2017). UMI-tools: Modeling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy. *Genome Research*, *27*(3), 491–499. doi: 10.1101/GR.209601.116/-/DC1

Söderhäll, K., & Cerenius, L. (1998). Role of the prophenoloxidase-activating system in invertebrate immunity. *Current Opinion in Immunology*, *10*(1), 23–28. doi: 10.1016/S0952-7915(98)80026-5

Stamatiou, K., Chmielewska, A., Ohta, S., Earnshaw, W. C., & Vagnarelli, P. (2023). CCDC86 is a novel Ki-67-interacting protein important for cell division. *Journal of Cell Science*, *136*(2). doi: 10.1242/JCS.260391

Stern, D. B., & Crandall, K. A. (2018). The Evolution of Gene Expression Underlying Vision Loss in Cave Animals. *Molecular Biology and Evolution*, *35*(8), 2005–2014. doi: 10.1093/MOLBEV/MSY106

Stroud, D. A., Formosa, L. E., Wijeyeratne, X. W., Nguyen, T. N., & Ryan, M. T. (2013). Gene knockout using transcription activator-like effector nucleases (TALENs) reveals that human ndufa9 protein is essential for stabilizing the junction between membrane and matrix arms of complex i. *Journal of Biological Chemistry*, *288*(3), 1685–1690. doi: 10.1074/jbc.C112.436766

Sugumaran, M., & Barek, H. (2016). Critical Analysis of the Melanogenic Pathway in Insects and Higher Animals. *International Journal of Molecular Sciences*, *17*(10). doi: 10.3390/IJMS17101753

Supek, F., Bošnjak, M., Škunca, N., & Šmuc, T. (2011). REVIGO Summarizes and Visualizes Long Lists of Gene Ontology Terms. *PLOS ONE*, *6*(7), e21800. doi: 10.1371/JOURNAL.PONE.0021800

Takahashi, S., Mui, V. J., Rosenberg, S. K., Homma, K., Cheatham, M. A., & Zheng, J. (2016). Cadherin 23-C Regulates Microtubule Networks by Modifying CAMSAP3's Function. *Scientific Reports 2016 6:1*, *6*(1), 1–13. doi: 10.1038/srep28706

Takakura, Y., Machida, M., Terada, N., Katsumi, Y., Kawamura, S., Horie, K., Miyauchi, M., Ishikawa, T., Akiyama, N., Seki, T., Miyao, T., Hayama, M., Endo, R., Ishii, H., Maruyama, Y., Hagiwara, N., Kobayashi, T. J., Yamaguchi, N., Takano, H., … Yamaguchi, N. (2024). Mitochondrial protein C15ORF48 is a stress-independent inducer of autophagy that regulates oxidative stress and autoimmunity. *Nature Communications*, *15*(1). doi: 10.1038/S41467-024-45206-1

Tang, W., Gunn, T. M., McLaughlin, D. F., Barsh, G. S., Schlossman, S. F., & Duke-Cohan, J. S. (2000). Secreted and membrane attractin result from alternative splicing of the human ATRN gene. *Proceedings of the National Academy of Sciences of the United States of America*, *97*(11), 6025–6030. doi: 10.1073/PNAS.110139897/ASSET/3ED33837-772F-4433-BC9B-77A2B146456B/ASSETS/GRAPHIC/PQ1101398005.JPEG

True, J. R. (2003). Insect melanism: the molecules matter. *Trends in Ecology & Evolution*, *18*(12), 640–647. doi: 10.1016/J.TREE.2003.09.006

Vittori, M., Tušek-Žnidarič, M., & Štrus, J. (2017). Exoskeletal cuticle of cavernicolous and epigean terrestrial isopods: A review and perspectives. *Arthropod Structure & Development*, *46*(1), 96–107. doi: 10.1016/J.ASD.2016.08.002

Wang, J., Rieder, S. A., Wu, J., Hayes, S., Halpin, R. A., de los Reyes, M., Shrestha, Y., Kolbeck, R., & Raja, R. (2019). Evaluation of ultra-low input RNA sequencing for the study of human T cell transcriptome. *Scientific Reports 2019 9:1*, *9*(1), 1–13. doi: 10.1038/s41598-019-44902-z

Wang, S., Kang, Y., Wang, R., Deng, J., Yu, Y., Yu, J., & Wang, J. (2022). Emerging Roles of NDUFS8 Located in Mitochondrial Complex I in Different Diseases. *Molecules*, *27*(24). doi: 10.3390/MOLECULES27248754

White, W. B. (1997). Thermodynamic equilibrium kinetics, activation barriers, and reaction mechanisms for chemical reactions in Karst Terrains. *Environmental Geology*, *30*(1–2), 46–58. doi: 10.1007/S002540050131/METRICS

Yamamoto, Y., Byerly, M. S., Jackman, W. R., & Jeffery, W. R. (2009). Pleiotropic functions of embryonic sonic hedgehog expression link jaw and taste bud amplification with eye loss during cavefish evolution. *Developmental Biology*, *330*(1), 200–211. doi: 10.1016/J.YDBIO.2009.03.003

Yamamoto, Y., Stock, D. W., & Jeffery, W. R. (2004). Hedgehog signalling controls eye degeneration in blind cavefish. *Nature 2004 431:7010*, *431*(7010), 844–847. doi: 10.1038/nature02864

8. Resume

Eva Marković was born on September 22nd, 1999, in Koprivnica, Croatia. She completed her elementary education and attended the Fran Galović Gymnasium in her hometown. In 2018, she began her studies at the Faculty of Science, University of Zagreb, where she earned a summa cum laude bachelor's degree in Molecular Biology in 2021. Her thesis was titled 'The Dark Side of Evolution – How Cave Animals Evolve'.

Continuing her academic journey, Eva pursued a master's degree in Molecular Biology at the same institution. Throughout her studies, she gained research experience through internships in various labs. These include the Bilandžija Group at the Laboratory of Molecular Genetics at the Ruđer Bošković Institute in Zagreb, the Department of Animal Physiology at the Faculty of Science, University of Zagreb, and the Multicellgenome Group at the Institut de Biologia Evolutiva in Barcelona.

In addition to her academic pursuits, Eva actively participated in several science popularization projects within and outside the Faculty. She also contributed to *In Vivo*, a science magazine produced by biology students. From 2019 to 2021, she was a recipient of the STEM Scholarship, and in the 2022/2023 academic year, she was awarded the University of Zagreb Scholarship for Excellence.

9. Supplement

9.1. Supplementary methods

9.1.1. UMI-tools command

Command used for UMI extraction:

```
umi_tools extract -I forward_reads.fq.gz \
    --bc-pattern=NNNNNNNNNNNN \
    --read2-in=reverse_reads.fq.gz \
    --stdout=sample_name_F_umi-out.fq.gz \
    --read2-out=sample_name_R_umi-out.fq.gz
```

The `--bc-pattern` option specifies the UMI sequence, which, for Lexogen's CORALL RNA-Seq libraries, are random 12-nucleotide sequences (denoted by 12 Ns).

9.1.2. fastp command

Command used for read trimming:

```
fastp -i sample_name_F_umi-out.fq.gz \
     -I sample_name_R_umi-out.fq.gz \
     -o sample_name_trimmed_umi-out_F.fq.gz \
     -O sample_name_trimmed_umi-out_R.fq.gz \
     -c --trim_poly_g
```

9.1.3. STAR command (mapping SIRV and ERCC reads)

Command used for SIRV and ERCC read mapping:

```
STAR --runThreadN 16 \
   --genomeDir /path/to/genome_dir/ \
   --readFilesIn sample_name_trimmed_umi-out_F.fq.gz sample_name_trimmed_umi-out_R.fq.gz \
   --readFilesCommand zcat \
   --outFileNamePrefix sample_name \
   --outReadsUnmapped Fastx
```

9.1.4. Trinity command and explanation

Command used for transcriptome assembly:

```
Trinity --grid_exec sbatch \
      --grid_node_CPU 60 \
      --grid_node_max_memory 8G \
      --seqType fq \
      --NO_SEQTK \
      --max_memory 470G \
      --CPU 60 \
      --left sample_1_F.fq, sample_2_F.fq,  sample_3_F.fq, sample_4_F.fq, sample_5_F.fq \
      --right sample_1_R.fq, sample_2_R.fq, sample_3_R.fq, sample_4_R.fq, sample_5_R.fq \
      --SS_lib_type FR \
      --trimmomatic \
      --quality_trimming_params "SLIDINGWINDOW:4:15 LEADING:10 TRAILING:10 MINLEN:25" \
      --output /trinity_output_dir/ \
      --bflyHeapSpaceMax 6G \
      --bflyCPU 60 \
      --min_kmer_cov 2 \
      --min_contig_length 300
```

Where the first three options specify the job submission on a computing cluster, `--seqType fq` specifies input file type as fastq, `--NO_SEQTK` helps resolve the issue of Trinity not recognizing read type as paired-end, `--max_memory` to specify the maximum memory that Trinity can use, `--CPU` specifies the number of CPUs Trinity can use, `--left` is to specify the forward read files (where I listed all replicates of the species), `--right` is to specify the reverse read files, `--SS_lib_type` stands for strand-specific library type. FR indicates that the first read in the pair is the sense strand (sequenced as forward) while the second read in the pair is the anti-sense strand (sequenced as reverse), which is the case with the CORALL RNA-Seq libraries. To ensure only the best quality reads got used for the assembly, I invoked Trimmomatic (`--trimmomatic`), a stand-alone trimming tool (Bolger et al., 2014b) incorporated into Trinity, and specified the trimming parameters with the `--quality_trimming_params` option. The `SLIDINGWINDOW:4:15` option scans the entire read length by sliding a window that is four bases wide and cutting where (if) the average per-base phred33 quality score drops below 15. The `LEADING:10` option removes the leading bases (bases at the beginning of the read) if the phred33 quality score is below 10, and the `TRAILING` option does the same with the trailing bases (bases at the end of a read). The `MINLEN:25` option removes the reads shorter than 25 bases since Trinity would discard them anyway. The `--bflyHeapSpaceMax` and `--bflyCPU` settings help limit the memory usage of one of Trinity's processes to no more than 80% of the job memory capacity, which is crucial to prevent the job from crashing. Finally, I set the `--min_kmer_cov` to 2, which prevents very lowly expressed transcripts from assembling, and `--min_contig_length` to 300, which prevents transcripts shorter than 300 bp from assembling. These two options help reduce the number of temporary files created by Trinity, which can cause the job to crash due to exceeding the disk quota.

9.1.5. BUSCO command

Command used for transcriptome assessment wit BUSCO:

```
busco --in Proasselus_x_transcriptome.Trinity.fasta
     --mode transcriptome
     --lineage_dataset arthropoda_odb10
```

9.1.6. TransDecoder commands

TransDecoder.LongOrfs option:

```
TransDecoder.LongOrfs -t Proasselus_x_transcriptome.Trinity.fasta
```

TransDecoder.Predict option:

```
TransDecoder.Predict -t Proasselus_x_transcriptome.Trinity.fasta
```

9.1.7. EggNOG-mapper command

Command used for annotating the transcriptomes with EggNOG:

```
emapper.py -I PX_transcriptome.Trinity.fasta.transdecoder.pep \
        --cpu 64 \
        --output P.x \
        --output_dir ./eggnog/
```

9.1.8. Read processing, mapping, and expression calculation pipeline

Trimmomatic command:

```
java -jar trimmomatic-0.39.jar PE -threads 32 \
$forward $reverse \
${dir_name}_trimmed_forward_paired.fq.gz ${dir_name}_trimmed_forward_unpaired.fq.gz \
${dir_name}_trimmed_reverse_paired.fq.gz ${dir_name}_trimmed_reverse_unpaired.fq.gz \
SLIDINGWINDOW:4:15 LEADING:10 TRAILING:10 MINLEN:25 \
1> Trimmomatic_stdout.txt \
2> Trimmomatic_stderr.txt
```

bowtie2 command:

```
bowtie2 -q \
     --phred33 \
     --sensitive \
     --dpad 0 \
     --gbar 99999999 \
     --mp 1,1 \
     --np 1 \
     --score-min L,0,-0.1 \
     -I 1 \
     -X 1000 \
     --no-mixed \
     --no-discordant \
     --norc \
     -p 30 \
     -k 200 \
     -x Proasselus_x_transcriptome.Trinity.fasta \
     -1 PX_CUT_X_trimmed_forward_paired.fq.gz \
     -2 PX_CUT_X_trimmed_reverse_paired.fq.gz \
     -p 30 -S PX_CUT_X_mapped.sam
```

SAMtools view command:

```
samtools view -bT Proasselus_x_transcriptome.Trinity.fasta PX_CUT_X_mapped.sam \
            -o PX_CUT_X_mapped.bam
```

SAMtools sort command:

```
samtools sort -o PX_CUT_X_mapped_sorted.bam \
            -O BAM \
            --threads 30 \
            --write-index \
            --reference Proasselus_x_transcriptome.Trinity.fasta \
            PX_CUT_X_mapped.bam
```

UMI-tools dedup command:

```
umi_tools dedup -I PX_CUT_X_mapped_sorted.bam \
              --stdout= PX_CUT_X_mapped_sorted_dedup.bam \
              --paired \
              --chimeric-pairs=discard \
              --unpaired-reads=discard
```

SAMtools sort command #2:

```
samtools sort -n \
            -o PX_CUT_X_mapped_sorted_dedup_sorted.bam \
            -O BAM \
            --threads 30 \
            --reference Proasselus_x_transcriptome.Trinity.fasta \
            PX_CUT_X_mapped_sorted_dedup.bam
```

UMI-tools prepare-for-rsem command:

```
umi_tools prepare-for-rsem -I PX_CUT_X_mapped_sorted_dedup_sorted.bam \
                         --stdout= PX_CUT_X_ready_for_rsem.bam
```

RSEM calculate-expression command:

```
rsem-calculate-expression --paired-end \
                        --alignments \
                        -p 30 \
                        PX_CUT_X_ready_for_rsem.bam \
                        Proasselus_x_transcriptome.Trinity.fasta \
                        PX_CUT_X
```

9.1.9. SIRVsuite command

Command used to invoke SIRVsuite for spike-in control assessment:

```
SIRVsuite -i sample_metadata.csv \
       -o ./SIRV_output \
       --SIRV-concentration \
       --ERCC-correlation \
       --experiment-name Proasellus
```

9.1.10. Python script 1: Filtering orthogroups

```
import pandas as pd
# Load the file into a DataFrame
file_path = 'N0.tsv'
df = pd.read_csv(file_path, sep='\t')

# Filter the rows where all species columns are not empty
filtered_df = df.dropna(subset=['PC_transcriptome.Trinity', 'PH_transcriptome.Trinity',
'PK_transcriptome.Trinity', 'PaMO_transcriptome.Trinity'])

# Save the filtered DataFrame to a new file
filtered_file_path = 'filtered_orthogroups.csv'
filtered_df.to_csv(filtered_file_path, index=False)

filtered_file_path
```

9.1.11. Python script 2: Extracting orthogroups and adding them to corresponding transcripts

```
import pandas as pd

# Load the RSEM gene counts file
rsem_file = 'PX_CUT_X.genes.results'
rsem_df = pd.read_csv(rsem_file, sep='\t')

# Load the orthogroups file
orthogroups_file = 'filtered_orthogroups()_modified_wo_isoforms.txt'
orthogroups_df = pd.read_csv(orthogroups_file, sep='\t')

# Create a dictionary to hold the mapping of gene_id to HOG_IDs
gene_to_hogs = {}
for index, row in orthogroups_df.iterrows():
    hog_id = row['HOG']
    genes = row['PX_transcriptome.Trinity']
    if pd.notna(genes):
        gene_list = genes.split(', ')
        for gene in gene_list:
            if gene not in gene_to_hogs:
                gene_to_hogs[gene] = set()
            gene_to_hogs[gene].add(hog_id)

# Create a new dataframe to hold the expanded rows
expanded_rows = []

# Iterate through the RSEM dataframe and create new rows with HOG_IDs
for index, row in rsem_df.iterrows():
    gene_id = row['gene_id']
    if gene_id in gene_to_hogs:
        for hog_id in gene_to_hogs[gene_id]:
            new_row = row.copy()
            new_row['HOG'] = hog_id
            expanded_rows.append(new_row)
    else:
        new_row = row.copy()
        new_row['HOG'] = None
        expanded_rows.append(new_row)

# Create the expanded dataframe
```

```python
expanded_df = pd.DataFrame(expanded_rows)

# Save the updated RSEM file
output_file = 'orthogroups_PK_CUT_5.genes.results'
expanded_df.to_csv(output_file, sep='\t', index=False)

print(f"Updated RSEM file saved as {output_file}")
```

9.1.12. Python script 3: Filtering out transcripts with no orthogroups associated

```python
import os
import pandas as pd

def filter_files_in_directory(directory):
    for filename in os.listdir(directory):
        if filename.startswith("orthogroups_"):
            file_path = os.path.join(directory, filename)
            df = pd.read_csv(file_path, delimiter='\t')
            # Filter out rows without a HOG_ID
            filtered_df = df[df['HOG'].notnull()]

            # Remove the 'transcript_id(s)' column
            filtered_df = filtered_df.drop(columns=['transcript_id(s)'])

            # Save the filtered DataFrame to a new file
            output_filename = f"filtered_{filename}"
            output_path = os.path.join(directory, output_filename)
            filtered_df.to_csv(output_path, sep='\t', index=False)

# Define the working directory
working_directory = os.getcwd()

# Run the filtering function
filter_files_in_directory(working_directory)
```

9.1.13. Python script 4: Summing up count of transcripts belonging to the same orthogroup

```python
import os
import pandas as pd

def process_file(file_path):
    df = pd.read_csv(file_path, delimiter='\t')  # Assuming the file is tab-delimited

    print("Original DataFrame:")
    print(df.head())

    # Sort by HOG
    df = df.sort_values(by='HOG')

    # Separate DataFrame for single and multiple gene_ids
    single_gene_id_df = df.groupby('HOG').filter(lambda x: len(x) == 1)
    multiple_gene_id_df = df.groupby('HOG').filter(lambda x: len(x) > 1)

    print("Single gene_id DataFrame:")
    print(single_gene_id_df.head())

    print("Multiple gene_id DataFrame:")
    print(multiple_gene_id_df.head())

    # Process single gene_id_df
    single_gene_id_df = single_gene_id_df[['HOG', 'gene_id', 'expected_count', 'length', 'effective_length', 'TPM', 'FPKM']]

    # Group by HOG and perform the required aggregations for multiple gene_id_df
    grouped = multiple_gene_id_df.groupby('HOG').agg({
        'gene_id': lambda x: ','.join(x),
        'expected_count': 'sum',
        'length': 'max',
        'effective_length': 'max'
    }).reset_index()

    print("Grouped DataFrame after aggregation:")
    print(grouped.head())

    # Function to get the TPM and FPKM associated with the max effective_length
    def get_tpm_fpkm_for_max_effective_length(HOG_group):
        max_length_row = HOG_group.loc[HOG_group['effective_length'].idxmax()]
        return pd.Series([max_length_row['TPM'], max_length_row['FPKM']], index=['TPM', 'FPKM'])

    # Apply the function to get the TPM and FPKM values
    tpm_fpkm = multiple_gene_id_df.groupby('HOG').apply(get_tpm_fpkm_for_max_effective_length).reset_index()

    print("TPM and FPKM DataFrame:")
    print(tpm_fpkm.head())

    # Merge the TPM and FPKM values back to the grouped dataframe
    grouped = grouped.merge(tpm_fpkm, on='HOG', how='left')

    print("Grouped DataFrame after merging TPM and FPKM:")
    print(grouped.head())

    # Concatenate the single and multiple gene_id DataFrames
    final_df = pd.concat([single_gene_id_df, grouped], ignore_index=True)

    print("Final DataFrame:")
    print(final_df.head())

    # Save the final DataFrame to a new file
    output_filename = f"merged_{os.path.basename(file_path)}"
    output_path = os.path.join(os.path.dirname(file_path), output_filename)
    final_df.to_csv(output_path, sep='\t', index=False)

def process_files_in_directory(directory):
    for filename in os.listdir(directory):
        if filename.startswith("filtered_orthogroups_"):
            file_path = os.path.join(directory, filename)
            process_file(file_path)

# Define the working directory as the current directory
working_directory = os.getcwd()

# Run the processing function
process_files_in_directory(working_directory)
```

9.1.14. DESeq2 analysis: A pairwise comparison example script

```r
#--------------------------------------------------------------------------------
#INSTALL PACKAGES IF NEEDED:

#if (!require("BiocManager", quietly = TRUE))
  #install.packages("BiocManager")
#BiocManager::install("DESeq2")
#BiocManager::install("tximport")
#BiocManager::install("tximportData")
#install.packages("readr")

#Load all required packages
library ("DESeq2")
library("tximport")
library("readr")
library("tximportData")

#--------------------------------------------------------------------------------
#PERFORM DIFFERENTIAL GENE EXPRESSION ANALYSIS
#--------------------------------------------------------------------------------
#Set working directory to the directory containing sample files to analyse
#Note: have a specific directory containing .genes.results files of the samples you're analyzing
dir <- getwd() #Set your working directory under the "dir" value
list.files(dir)

#Manually create a .csv file containing sample names and all metadata of your experiment (cell type, generation, 02 levels...)
#Load the .csv file and assign it to the "samples" value
samples <- read.csv(file.path(dir, "samples_PC_PaMO.csv"), header = TRUE, sep = ";")
samples

#Specify the path to the files using the appropriate columns of samples, and read in a table that links transcripts to genes for this dataset
files <- file.path(dir, paste0(samples$SAMPLE,".genes.results"))
names(files) <- paste0(samples$SAMPLE)

#Construct the txi object
txi.rsem <- tximport(files, type = "rsem", txIn = FALSE, txOut = FALSE)
head(txi.rsem$counts)


#Construct a DESeqDataSet from the txi object
ddsTxi <- DESeqDataSetFromTximport(txi.rsem,
                                   colData = samples,
                                   design = ~ SPECIES)

#In case this error appears after running the previos line of code:
#Error in DESeqDataSetFromTximport(txi.rsem, colData = samples, design = ~SPECIES) : all(lengths > 0) is not TRUE
#Run the following
txi.rsem$length[txi.rsem$length == 0] <- 1
#Then re-run ddsTxi <- DESeqDataSetFromTximport ...

#Run the DESeq2 analysis of the data
dds <- DESeq(ddsTxi)
sizeFactors(dds)
```

```
colSums(counts(dds))
colSums(counts(dds, normalized=T))
res <- results(dds)
#Print the results
res

#Check the number of up- and down-regulated genes for padj < 0.1
summary(res)
#Check the number of padj < 0.1
sum(res$padj < 0.1, na.rm=TRUE)
#Create a dataset with genes for which the padj < 0.05
res05 <- results(dds, alpha=0.05)
#Check the number of up- and down-regulated genes for padj < 0.05
summary(res05)
#Check the number of padj < 0.05
sum(res05$padj < 0.05, na.rm=TRUE)
#Order the results by p-value
resOrdered <- res05[order(res05$pvalue),]

#Output differentially expressed genes (padj < 0.1)
write.csv(as.data.frame(resOrdered),
          file="PC_vs_PaMO.csv")

#Create a subset of the differentially expresed genes containing only the ones with pajd < 0.05
resSig <- subset(resOrdered, padj < 0.05)
resSig
#Output differentially expressed genes (padj < 0.05)
write.csv(as.data.frame(resSig),
          file="PC_vs_PaMO_padj_005.csv")

#Plot the log fold change against the mean of normalized counts
#Adjust ylim as needed, write the plot title under main =
plotMA(res, ylim=c(-15,15), main = "PC vs PaMO")
```

9.1.15. DESeq2 analysis: Sample clustering

```
#-------------------------------------------------------------------------------
#INSTALL PACKAGES IF NEEDED:

#if (!require("BiocManager", quietly = TRUE))
#install.packages("BiocManager")
#BiocManager::install("DESeq2")
#BiocManager::install("tximport")
#BiocManager::install("tximportData")
#install.packages("readr")

#Load all required packages
library ("DESeq2")
library("tximport")
library("readr")
library("tximportData")

#-------------------------------------------------------------------------------
#PERFORM DIFFERENTIAL GENE EXPRESSION ANALYSIS
#-------------------------------------------------------------------------------
#Set working directory to the directory containing sample files to analyse
#Note: have a specific directory containing .genes.results files of the samples
```

```r
you're analyzing
dir <- getwd() #Set your working directory under the "dir" value
list.files(dir)

#Create a .csv file containing sample names and all metadata (e.g. type, generation, replicate...)
#Load the .csv file and assign it to the "samples" value
samples <- read.csv(file.path(dir, "samples_all.csv"), header = TRUE, sep = ";")
samples

#Specify the path to the files using the appropriate columns of samples, and read in a table that links transcripts to genes for this dataset
files <- file.path(dir, paste0(samples$SAMPLE,".genes.results"))
names(files) <- paste0(samples$SAMPLE)

#Construct the txi object
txi.rsem <- tximport(files, type = "rsem", txIn = FALSE, txOut = FALSE)
head(txi.rsem$counts)

#Use this in case of error:
#Error in DESeqDataSetFromTximport(txi.rsem, colData = samples, design = ~SPECIES) : all(lengths > 0) is not TRUE
#if it appears after running the ddsTxi <- DESeqDataSetFromTximport
txi.rsem$length[txi.rsem$length == 0] <- 1

#Construct a DESeqDataSet from the txi object
#Under the "design" variable, include experimental design variables relevant to the gene expression comparison between samples
ddsTxi <- DESeqDataSetFromTximport(txi.rsem,
                                   colData = samples,
                                   design = ~ SPECIES)

#Run the DESeq2 analysis of the data
dds <- DESeq(ddsTxi)
sizeFactors(dds)
colSums(counts(dds))
colSums(counts(dds, normalized=T))
res <- results(dds)
res

#Plot the log fold change against the mean of normalized counts
#Adjust ylim as needed, write the plot title under main =
plotMA(res, ylim=c(-10,22), main = "Hypoxia Ad and Agg")

#-------------------------------------------------------------------------------
#COUNT DATA TRANSFORMATION
#-------------------------------------------------------------------------------
#There are 3 types of data transformation available: ntd, vsd and rld
#Note that vsd and rld work better than ntd, but keep in mind that rld is slow compared to vsd

#BiocManager::install("vsn")
library("vsn")

#ntd transformation: This gives log2(n + 1)
ntd <- normTransform(dds)
#Plot the standard deviation agains the mean of normalized counts
meanSdPlot(assay(ntd))

```

```r
#vsd transformation
vsd <- vst(dds, blind=FALSE) #blind = TRUE when you want to omit any experimental design variables
head(assay(vsd), 3)
#Plot the standard deviation agains the mean of normalized counts
meanSdPlot(assay(vsd))

#rld transformation
rld <- rlog(dds, blind=FALSE) #blind = TRUE when you want to omit any experimental design variables
#Plot the standard deviation agains the mean of normalized counts
meanSdPlot(assay(rld))

#-------------------------------------------------------------------------------
#SAMPLE CLUSTERING
#-------------------------------------------------------------------------------
library("pheatmap")
library("RColorBrewer")
#...............................................................................
#Create heatmaps with nts, vsd and rld transformed data

pheatmap(deseq2vsd, color=heatmap_colors, cluster_rows=TRUE, show_rownames=FALSE,
         cluster_cols=TRUE, annotation_col=df, fontsize_col = fontsize, fontsize_row = 5, main = "all samples")

#Create a heatmap using vsd transformed data
pheatmap(assay(vsd)[select,], cluster_rows=FALSE, show_rownames=FALSE,
         cluster_cols=FALSE, annotation_col=df, fontsize_col = fontsize, main = "All samples")

#Create a heatmap using rld transformed data
pheatmap(assay(rld)[select,], cluster_rows=FALSE, show_rownames=FALSE,
         cluster_cols=FALSE, annotation_col=df, fontsize_col = fontsize, main = "Hypoxia Ad and Agg, rld transform")

#...............................................................................
#Create sample distances heatmaps with nts, vsd and rld transformed data

#Calculate sample distances with ntd transformed data
sampleDists <- dist(t(assay(ntd)))

#Create a sample distance matrix
sampleDistMatrix <- as.matrix(sampleDists)
rownames(sampleDistMatrix) <- paste(samples$SAMPLE)
colnames(sampleDistMatrix) <- paste(samples$SAMPLE)
colors <- colorRampPalette((brewer.pal(9, "YlGnBu")))(255)
colors <- colorRampPalette((brewer.pal(9, "GnBu")))(255)
colors <- (colorRampPalette(c("#29456a","#375d8f","#4575b4","#6790c5","#8cabd3","#b1c6e1", "#d6e1ef", "#fbfcfd"),bias=1, space=c("rgb","Lab"))(255))
#colors <- (colorRampPalette(c( "#f8fce0","#f1f9c1", "#edf8b1","#a4dbce", "#7fcdbb", "#2c7fb8", "#22638f"), bias=1, space=c("rgb","Lab"))(255))

fontsize <- 5 #Adjust as needed
#Plot the heatmap of sample to sample distances
pheatmap(sampleDistMatrix,
         clustering_distance_rows=sampleDists,
         clustering_distance_cols=sampleDists,
         col=colors,
         fontsize_row = fontsize,
         fontsize_col = fontsize,
         border_color = "black",
         cellwidth=6,
         cellheight=6,
         main = "All samples sample distance, ntd")

#Calculate sample distances with vsd transformed data
sampleDists <- dist(t(assay(vsd)))

#Create a sample distance matrix
sampleDistMatrix <- as.matrix(sampleDists)
rownames(sampleDistMatrix) <- paste(samples$SAMPLE)
colnames(sampleDistMatrix) <- paste(samples$SAMPLE)
colors <- colorRampPalette((brewer.pal(9, "YlGnBu")))(255)
colors <- colorRampPalette((brewer.pal(9, "GnBu")))(255)
colors <-
(colorRampPalette(c("#29456a","#375d8f","#4575b4","#6790c5","#8cabd3","#b1c6e1",
"#d6e1ef", "#fbfcfd"),bias=1, space=c("rgb","Lab"))(255))
colors <- (colorRampPalette(c( "#f8fce0","#f1f9c1", "#edf8b1","#a4dbce",
"#7fcdbb", "#2c7fb8", "#22638f"), bias=1, space=c("rgb","Lab"))(255))

fontsize <- 5 #Adjust as needed
#Plot the heatmap of sample to sample distances
pheatmap(sampleDistMatrix,
         clustering_distance_rows=sampleDists,
         clustering_distance_cols=sampleDists,
         col=colors,
         fontsize_row = fontsize,
         fontsize_col = fontsize,
         border_color = "black",
         cellwidth=6,
         cellheight=6,
         main = "All samples sample distance, vsd")

#Calculate sample distances with rld transformed data
sampleDists <- dist(t(assay(rld)))

#Create a sample distance matrix
sampleDistMatrix <- as.matrix(sampleDists)
rownames(sampleDistMatrix) <- paste(samples$SAMPLE)
colnames(sampleDistMatrix) <- paste(samples$SAMPLE)
colors <- colorRampPalette((brewer.pal(9, "YlGnBu")))(255)
colors <- colorRampPalette((brewer.pal(9, "GnBu")))(255)
colors <-
(colorRampPalette(c("#29456a","#375d8f","#4575b4","#6790c5","#8cabd3","#b1c6e1",
"#d6e1ef", "#fbfcfd"),bias=1, space=c("rgb","Lab"))(255))
colors <- (colorRampPalette(c( "#f8fce0","#f1f9c1", "#edf8b1","#a4dbce",
"#7fcdbb", "#2c7fb8", "#22638f"), bias=1, space=c("rgb","Lab"))(255))

fontsize <- 5 #Adjust as needed
#Plot the heatmap of sample to sample distances
pheatmap(sampleDistMatrix,
         clustering_distance_rows=sampleDists,
         clustering_distance_cols=sampleDists,
         col=colors,
         fontsize_row = fontsize,
         fontsize_col = fontsize,
         border_color = "black",
          cellwidth=6,
          cellheight=6,
          main = "All samples sample distance, rld")

#.............................................................................
#Perform PCA analysis with ntd, vsd and rld transformed data

plotPCA(ntd, intgroup==c("SPECIES", "CAVE")) #under intgroup variable, input experimental design variables relevant for PCA clustering

plotPCA(vsd, intgroup=c("SPECIES", "CAVE"))

plotPCA(rld, intgroup=c("SPECIES", "CAVE"))176.
```

9.1.16. topGO analysis

```
# if (!requireNamespace("BiocManager", quietly=TRUE))
#   + install.packages("BiocManager")
# BiocManager::install("topGO")

library(topGO)

all_pval <- read.csv(file="all_pval_PaMO_&_PH.csv",header=TRUE, sep=",")
all_pval_mat <- as.vector(all_pval$padj)
names(all_pval_mat) <- as.character(all_pval$GeneID)

down <- read.csv(file="PaMO-up(vsPK)-l2fc2.csv",header=TRUE)
up <- read.csv(file="PaMO_&_PH_up.csv",header=TRUE)
gid_down <- as.character(down$GeneID)
gid_up <- as.character(up$GeneID)
gid_all <- names(all_pval_mat)
PA_down <- factor(as.integer(gid_all %in% gid_down))
PA_up <-  factor(as.integer(gid_all %in% gid_up))
names(PA_down) <- names(all_pval_mat)
names(PA_up) <- names(all_pval_mat)

geneID2GO <- readMappings(file="Proasellus_transcript2GOID.txt")

up_topGO_BP <- new("topGOdata", ontology="BP", allGenes=PA_up,annot=annFUN.gene2GO,gene2GO=geneID2GO)

resultFisher_up_topGO_BP <- runTest(up_topGO_BP,algorithm="classic",statistic="fisher")
#resultFisher_up_topGO_BP_elim <- runTest(up_topGO_BP,algorithm="elim",statistic="fisher")
#resultFisher_up_topGO_BP_weight <- runTest(up_topGO_BP,algorithm="weight",statistic="fisher")
#resultFisher_up_topGO_BP_weight01 <- runTest(up_topGO_BP,algorithm="weight01",statistic="fisher")
#resultFisher_up_topGO_BP_lea <- runTest(up_topGO_BP,algorithm="lea",statistic="fisher")
#resultFisher_up_topGO_BP_pc <- runTest(up_topGO_BP,algorithm="parentchild",statistic="fisher")

allRes_Fisher_up_topGO_BP <- GenTable(up_topGO_BP,classic=resultFisher_up_topGO_BP,orderBy="classic",ranksOf="classic",topNodes=length(usedGO(up_topGO_BP)),numChar=1000)

```

```
write.csv(allRes_Fisher_up_topGO_BP,file="PaMO_&_PH_UP_topGO_BP.csv", row.names=FALSE)

up_topGO_MF <- new("topGOdata",ontology="MF",allGenes=PA_up,annot=annFUN.gene2GO,gene2GO=geneID2GO)
resultFisher_up_topGO_MF <- runTest(up_topGO_MF,algorithm="classic",statistic="fisher")
#resultFisher_up_topGO_MF_elim <- runTest(up_topGO_MF,algorithm="elim",statistic="fisher")
#resultFisher_up_topGO_MF_weight <- runTest(up_topGO_MF,algorithm="weight",statistic="fisher")
#resultFisher_up_topGO_MF_weight01 <- runTest(up_topGO_MF,algorithm="weight01",statistic="fisher")
#resultFisher_up_topGO_MF_lea <- runTest(up_topGO_MF,algorithm="lea",statistic="fisher")
#resultFisher_up_topGO_MF_pc <- runTest(up_topGO_MF,algorithm="parentchild",statistic="fisher")

allRes_Fisher_up_topGO_MF <- GenTable(up_topGO_MF,classic=resultFisher_up_topGO_MF,orderBy="classic",ranksOf="classic",topNodes=length(usedGO(up_topGO_MF)),numChar=1000)

write.csv(allRes_Fisher_up_topGO_MF,file="PaMO_&_PH_UP_topGO_MF.csv", row.names=FALSE)

up_topGO_CC <- new("topGOdata",ontology="CC",allGenes=PA_up,annot=annFUN.gene2GO,gene2GO=geneID2GO)
resultFisher_up_topGO_CC <- runTest(up_topGO_CC,algorithm="classic",statistic="fisher")
#resultFisher_up_topGO_CC_elim <- runTest(up_topGO_CC,algorithm="elim",statistic="fisher")
#resultFisher_up_topGO_CC_weight <- runTest(up_topGO_CC,algorithm="weight",statistic="fisher")
#resultFisher_up_topGO_CC_weight01 <- runTest(up_topGO_CC,algorithm="weight01",statistic="fisher")
#resultFisher_up_topGO_CC_lea <- runTest(up_topGO_CC,algorithm="lea",statistic="fisher")
#resultFisher_up_topGO_CC_pc <- runTest(up_topGO_CC,algorithm="parentchild",statistic="fisher")

allRes_Fisher_up_topGO_CC <- GenTable(up_topGO_CC,classic=resultFisher_up_topGO_CC,orderBy="classic",ranksOf="classic",topNodes=length(usedGO(up_topGO_CC)),numChar=1000)

write.csv(allRes_Fisher_up_topGO_CC,file="PaMO_&_PH_UP_topGO_CC.csv", row.names=FALSE)

down_topGO_BP <- new("topGOdata", ontology="BP", allGenes=PA_down,annot=annFUN.gene2GO,gene2GO=geneID2GO)

resultFisher_down_topGO_BP <- runTest(down_topGO_BP,algorithm="classic",statistic="fisher")
#resultFisher_down_topGO_BP_elim <- runTest(down_topGO_BP,algorithm="elim",statistic="fisher")
#resultFisher_down_topGO_BP_weight <- runTest(down_topGO_BP,algorithm="weight",statistic="fisher")
#resultFisher_down_topGO_BP_weight01 <- runTest(down_topGO_BP,algorithm="weight01",statistic="fisher")
#resultFisher_down_topGO_BP_lea <-
```

```
runTest(down_topGO_BP,algorithm="lea",statistic="fisher")
#resultFisher_down_topGO_BP_pc <- runTest(down_topGO_BP,algorithm="parentchild",statistic="fisher")

allRes_Fisher_down_topGO_BP <- GenTable(down_topGO_BP,classic=resultFisher_down_topGO_BP,orderBy="classic",ranksOf="classic",topNodes=length(usedGO(down_topGO_BP)),numChar=1000)

write.csv(allRes_Fisher_down_topGO_BP,file=" PaMO_&_PH_DOWN_topGO_BP.csv", row.names=FALSE)

down_topGO_MF <- new("topGOdata",ontology="MF",allGenes=PA_down,annot=annFUN.gene2GO,gene2GO=geneID2GO)
resultFisher_down_topGO_MF <- runTest(down_topGO_MF,algorithm="classic",statistic="fisher")
#resultFisher_down_topGO_MF_elim <- runTest(down_topGO_MF,algorithm="elim",statistic="fisher")
#resultFisher_down_topGO_MF_weight <- runTest(down_topGO_MF,algorithm="weight",statistic="fisher")
#resultFisher_down_topGO_MF_weight01 <- runTest(down_topGO_MF,algorithm="weight01",statistic="fisher")
#resultFisher_down_topGO_MF_lea <- runTest(down_topGO_MF,algorithm="lea",statistic="fisher")
#resultFisher_down_topGO_MF_pc <- runTest(down_topGO_MF,algorithm="parentchild",statistic="fisher")

allRes_Fisher_down_topGO_MF <- GenTable(down_topGO_MF,classic=resultFisher_down_topGO_MF,orderBy="classic",ranksOf="classic",topNodes=length(usedGO(down_topGO_MF)),numChar=1000)

write.csv(allRes_Fisher_down_topGO_MF,file=" PaMO_&_PH_DOWN_topGO_MF.csv", row.names=FALSE)

down_topGO_CC <- new("topGOdata",ontology="CC",allGenes=PA_down,annot=annFUN.gene2GO,gene2GO=geneID2GO)
resultFisher_down_topGO_CC <- runTest(down_topGO_CC,algorithm="classic",statistic="fisher")
#resultFisher_down_topGO_CC_elim <- runTest(down_topGO_CC,algorithm="elim",statistic="fisher")
#resultFisher_down_topGO_CC_weight <- runTest(down_topGO_CC,algorithm="weight",statistic="fisher")
#resultFisher_down_topGO_CC_weight01 <- runTest(down_topGO_CC,algorithm="weight01",statistic="fisher")
#resultFisher_down_topGO_CC_lea <- runTest(down_topGO_CC,algorithm="lea",statistic="fisher")
#resultFisher_down_topGO_CC_pc <- runTest(down_topGO_CC,algorithm="parentchild",statistic="fisher")

allRes_Fisher_down_topGO_CC <- GenTable(down_topGO_CC,classic=resultFisher_down_topGO_CC,orderBy="classic",ranksOf="classic",topNodes=length(usedGO(down_topGO_CC)),numChar=1000)

write.csv(allRes_Fisher_down_topGO_CC,file=" PaMO_&_PH_DOWN_topGO_CC.csv", row.names=FALSE)
```

9.2. Supplementary results

9.2.1. RNA and cDNA library quality

The average mass of total RNA isolated from the cuticles was 661.5 ng, giving the mean mass concentration of 60.26 ng/µL. The total isolated RNA was of high quality, shown by the values of A_{260}/A_{280} and A_{260}/A_{230} ratios which confirmed the purity of the RNA. Furthermore, the integrity of the RNA was confirmed with microcapillary electrophoresis, which showed three distinct peaks belonging to rRNA in the samples, as it can be seen in *Figure S1* (*P. coxalis* sample pool 1), which is characteristic to isopods (Deleo et al., 2018).

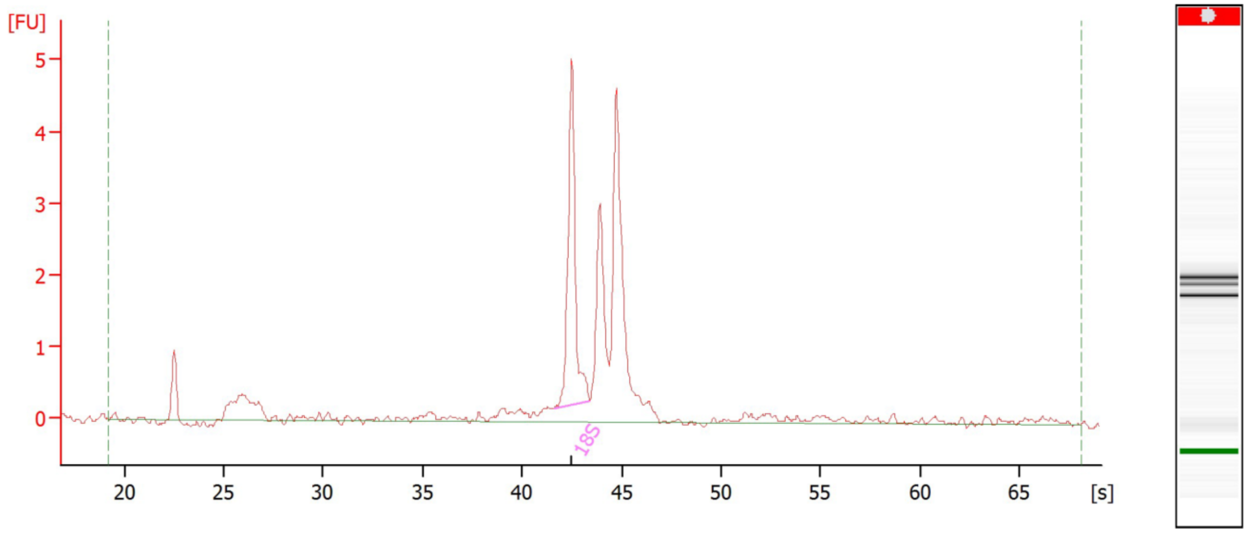


Figure S1 Distinct rRNA peaks of Proasellus coxalis visualised by microcapillary electrophoresis with Bioanalyser

The average cDNA library size (size of the cDNA fragments) obtained from the isolated RNA was 326 bp, and the average molarity of the libraries was 12.82 nM. All libraries showed high integrity, no overcycling, and no degradation in the microcapillary electrophoresis results, as it can be seen in the example in *Figure S2*.

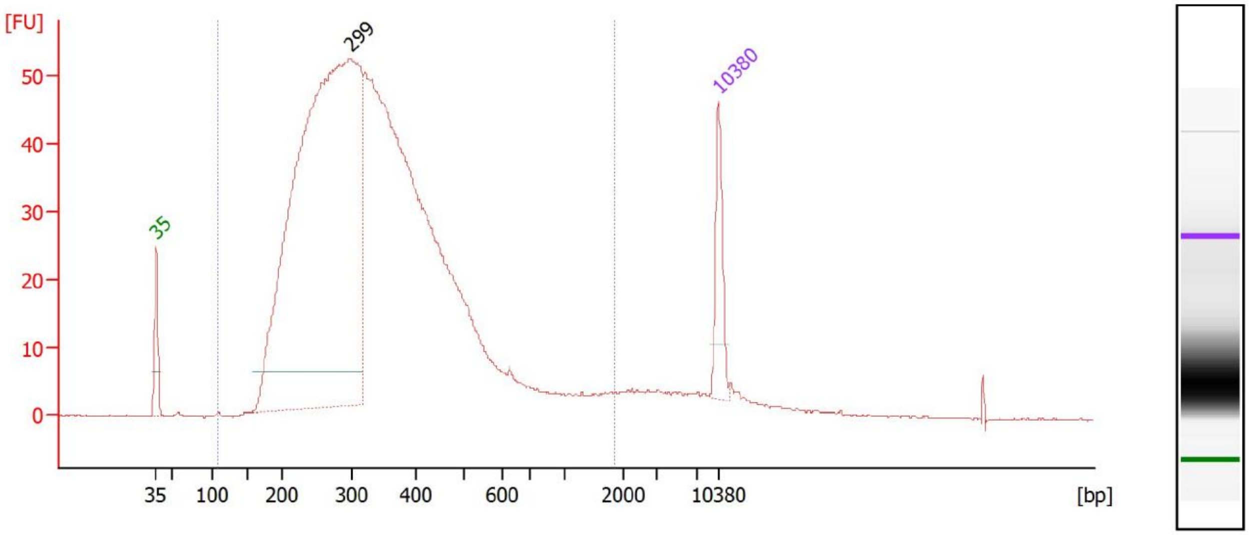


Figure S2 Library size distribution visualized by microcapillary electrophoresis with Bioanalyser. Green and purple lines represent the ladder, with the ladder fragment sizes (in bp) denoted in corresponding colours. Average library fragment size is 299 bp.

9.2.2. Sequence processing: fastp trimming and filtering results

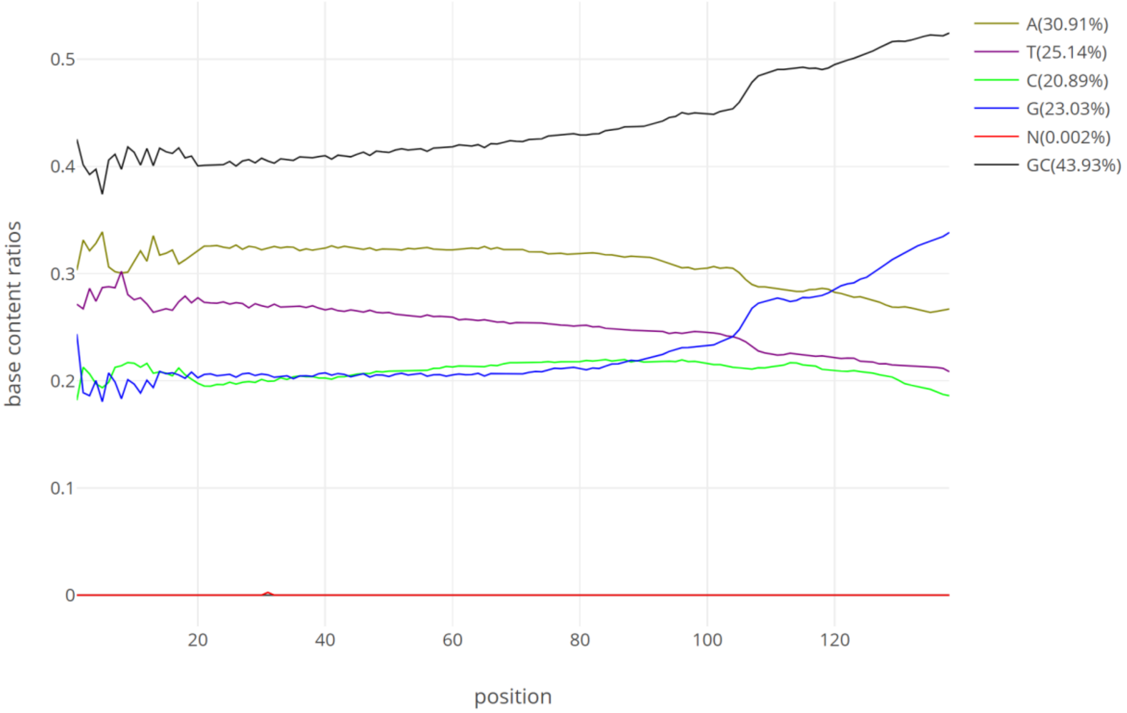


Figure S3 Base content ratios corresponding to each position in the reads before trimming and filtering. Graph inferred by fastp.

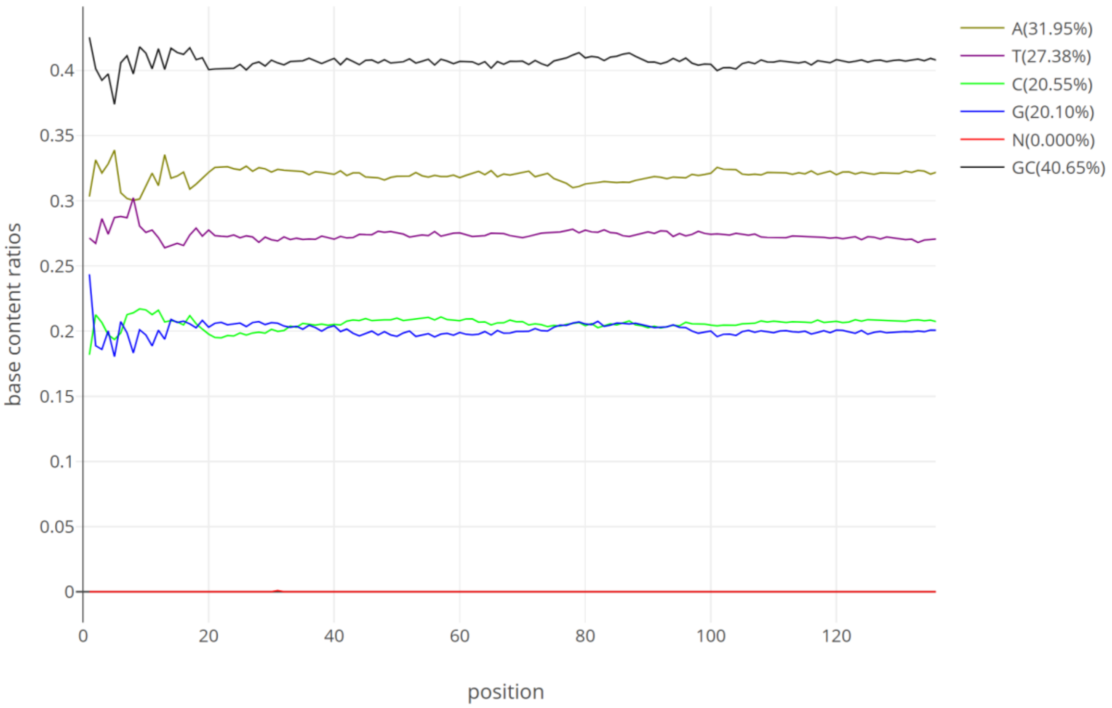


Figure S4 Base content ratios corresponding to each position in the reads after trimming and filtering with fastp. Graph inferred by fastp.

9.2.3. Spike-in quality control – individual sample quality

Additional sample preparation, library preparation, and sequencing quality was assessed with spike-in internal controls. *Figure S5* shows relative concentrations of all SIRV isoforms, denoted by log2 fold changes (L2FC) compared to expected concentrations. Blue tones indicate concentrations up to two (or more) times lower than expected (L2FC = -1), while red tones indicate concentrations up to two (or more) times higher than expected (L2FC = 1). SIRV set 1 isoforms (SIRV1, first block) show consistent concentrations across all samples. The only exception is the SIRV102 isoform having low concentrations in four out of five *P. karamani* samples, and two out of five *P. anophtalmus* samples. Contrastingly, the SIRV2 set (second block) shows more deviations from expected concentrations. The first three isoforms of the set (SIRV201, SIRV202, SIRV203) have expected concentrations, but the remaining three isoforms (SIRV204, SIRV205, SIRV206) show low concentrations (around L2FC of -1, or more). Nevertheless, the lower concentrations are consistent across all four species and samples, indicating the error is systemic, and no bias towards a species is present. The SIRV3 set is similar to the preceding one, with the majority of isoforms showing no shifts from the expected concentrations, with SIRV305's, SIRV308's, and SIRV311's L2FC values indicating lower concentrations. In the case of SIRV311, it is undetected in nine out of 19 samples. Again, these shifts are present in all samples regardless of the species, not implying any bias. On the other hand, the SIRV4 set of isoforms indicates three isoforms (SIRV403, SIRV404, SIRV405) with concentrations above expected. Two of the isoforms (SIRV406, SIRV409) show concentrations slightly below the expected values. Much like the other SIRV sets, this one indicates no apparent bias towards any species, further confirming the comparability between species and a consistency of the experimental procedures. However, the changes in concentrations among the SIRV5 set of isoforms appear to be less consistent among the species and samples. While the SIRV501 isoform shows consistently higher concentrations across all samples, and the SIRV512 isoforms shows consistently lower concentrations across all samples, other isoform, with the exclusion of SIRV504 and SIRV508, are less consistent. Still, no trend is apparent that would show a particular bias toward a species or a set of samples. The next set, SIRV6, paints a similar picture, with SIRV618 being particularly inconsistent, with concentrations being too low in some samples, and too high in others. The SIRV617 shows very low concentrations across all samples, with it being undetected in two cases. The last set, SIRV7, is fairly consistent across all samples and species, with the majority of isoforms showing slightly elevated concentrations. Overall, even though a lot of isoforms deviate from the expected concentration values, the changes of concentrations are mostly consistent, indicating no apparent bias towards a sample or species. Analysing all isoforms from all of the sets together, the only sample showing a pattern that slightly deviates from the rest in some instances is the PK_CUT_3 sample, belonging to the *P. karamani* species.

Furthermore, *Figure S6* illustrates the relative concentration distribution of all SIRV isoforms for each individual sample. Relative concentration distributions of isoforms seem uniform across all samples. The distribution deviates more in the lower concentration range which is a trend observed in all samples. This is especially evident in PC_CUT_3, PH_CUT_3, PC_CUT_1, PC_CUT_3, and PC_CUT_5.

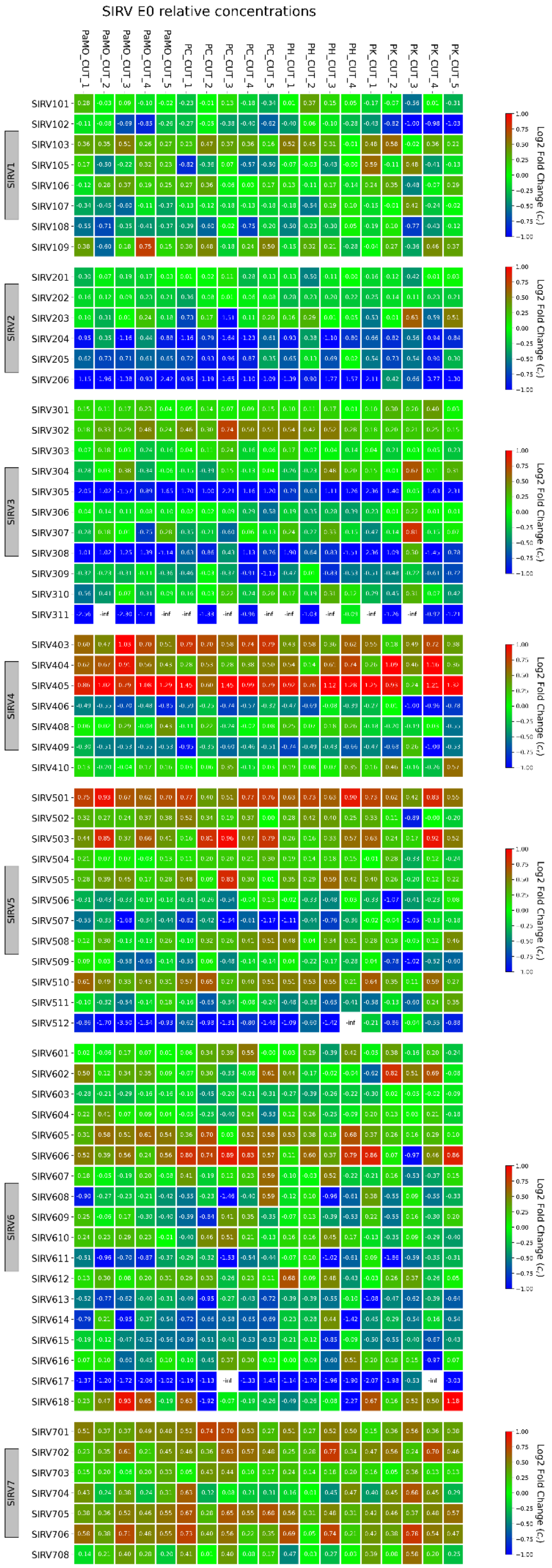


Figure S5 Relative concentrations of SIRV isoforms across samples, denoted on a log2 fold change scale with blue showing low, and red showing high concentration values.

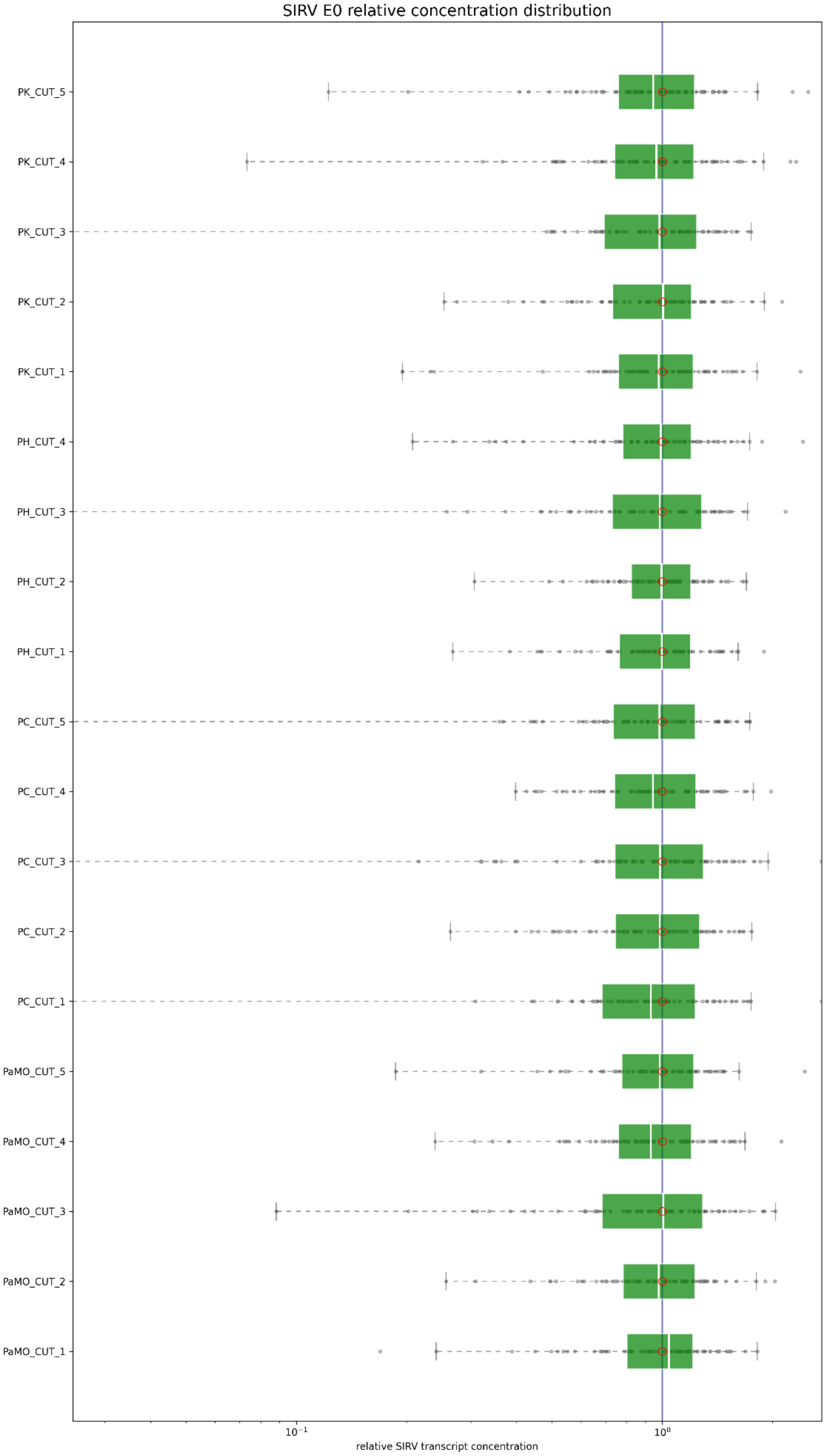


Figure S6 Relative concentration distributions of SIRV isoforms across samples.

The ERCC controls were used to validate the observed amounts of the control sequences compared to the expected values. A plot showing the correlation between the observed (measured) and expected (theoretical) concentrations can be observed in *Figure S7*. The Pearson correlation coefficient (R) in this example is 0.951 which indicates a high correlation of the two values. There are no major deviations from the theoretical concentrations at any concentration range, even at the lowest range of concentrations around 10^{-4}.

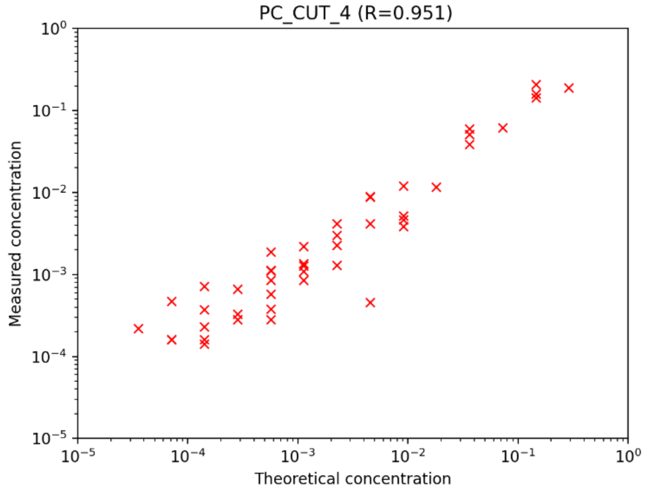


Figure S7 Correlation of measured and theoretical concentrations of ERCC control sequences

Other samples show similar results to the plot shown in *Figure S7*. A table listing all Pearson correlation coefficients is presented below (*Table S1*).

Table S1 Pearson correlation coefficients reflecting the correlation of measured and theoretical concentrations of ERCC control sequences, across all samples

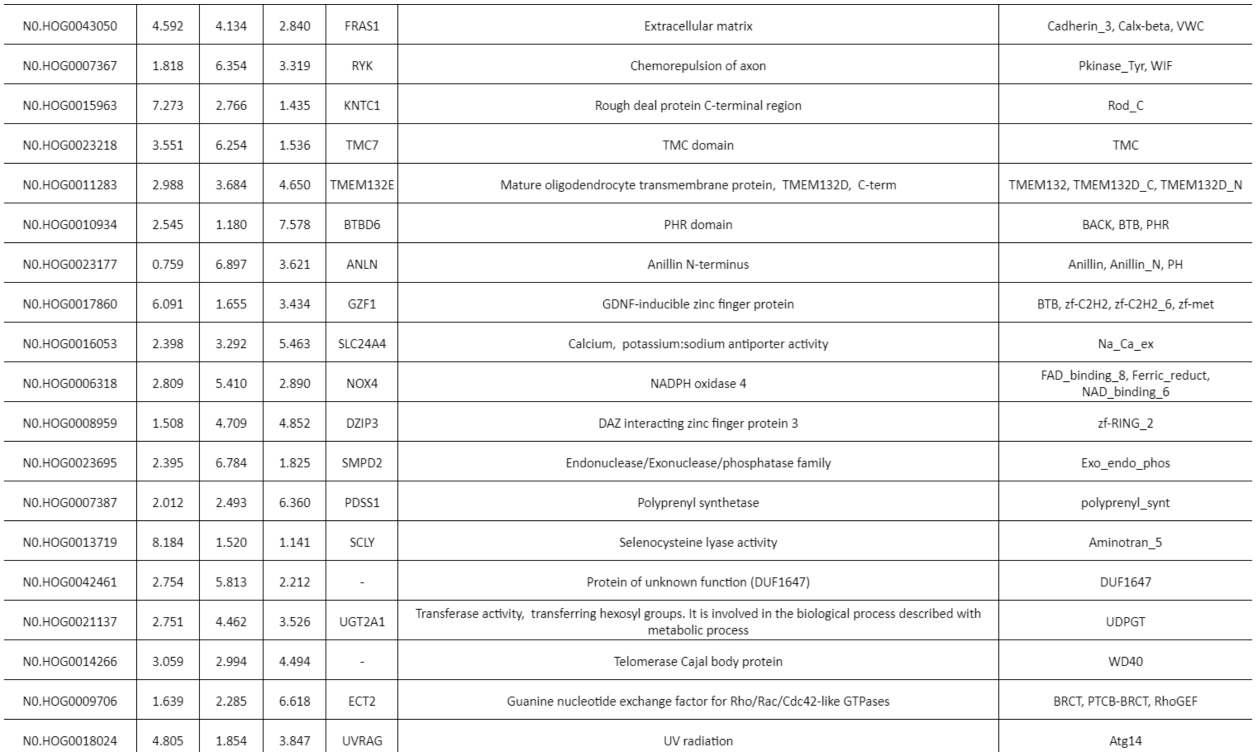
Sample	ERCC Pearson R
PaMO_CUT_1	0.910384
PaMO_CUT_2	0.950981
PaMO_CUT_3	0.932471
PaMO_CUT_4	0.943329
PaMO_CUT_5	0.937743
PC_CUT_1	0.946059
PC_CUT_2	0.907055
PC_CUT_3	0.901259
PC_CUT_4	0.950966
PC_CUT_5	0.927524
PH_CUT_1	0.932833
PH_CUT_2	0.948654
PH_CUT_3	0.906645
PH_CUT_4	0.920024
PK_CUT_1	0.911687
PK_CUT_2	0.927877
PK_CUT_3	0.913337
PK_CUT_4	0.889292
PK_CUT_5	0.957303

9.2.4. Differentially expressed genes of *P. anophtalmus*

Table S2 The top 50 up-regulated genes of P. anophtalmus. This list contains an overlap of genes that are up-regulated in P. anophtalmus in all three pairwise comparisons (P. anophtalmus (PA) vs P. hercegovinensis (PH), P. anophtalmus vs P. coxalis (PC), and P. anophtalmus vs P. karamani (PK)). A sum of L2FC values determined the order of the list. A preferred name for the gene, a brief description of the function, and PFAM domains are also listed.

Gene ID/HOG ID	L2FC vs PH	L2FC vs PC	L2FC vs PK	Preferred name	Description	PFAMs
N0.HOG0015942	10.102	4.543	7.163	-	Transcription regulatory region sequence-specific DNA binding	BACK, BTB, Kelch_1, bZIP_Maf, zf-C2H2
N0.HOG0033985	6.055	8.835	4.625	CLPTM1L	Cleft lip and palate associated transmembrane protein 1, like	CLPTM1
N0.HOG0004603	9.945	5.016	4.405	ANKIB1	Ankyrin repeats (3 copies)	Ank, Ank_5, IBR, zf-RING_UBOX
N0.HOG0034759	5.881	8.005	2.375	CSGALNACT1	Acetylgalactosaminyltransferase activity	CHGN
N0.HOG0015836	8.394	2.030	5.769	MROH1	Maestro heat-like repeat-containing protein family member	HEAT
N0.HOG0000518	2.407	2.643	11.121	-	Reverse transcriptase (RNA-dependent DNA polymerase)	RVT_1
N0.HOG0042863	8.672	4.308	1.905	-	Immunoglobulin V-set domain	I-set, Ig_3, V-set, ig
N0.HOG0007048	3.759	9.407	1.717	cac	Voltage-gated calcium channel activity. It is involved in the biological process described with calcium ion transmembrane transport	Ca_chan_IQ, GPHH, Ion_trans
N0.HOG0042836	8.217	4.772	1.417	ANO7	Dimerisation domain of Ca+-activated chloride-channel, anoctamin	Anoct_dimer, Anoctamin
N0.HOG0042518	5.641	6.665	2.015	-	-	-
N0.HOG0015990	3.398	6.006	4.660	ASTL	Astacin (Peptidase family M12A)	Astacin, CUB
N0.HOG0017459	7.969	3.976	1.896	PMT2	to Saccharomyces cerevisiae PMT2 (YAL023C) and PMT3 (YOR321W)	MIR, PMT, PMT_4TMC
N0.HOG0021480	3.124	4.203	6.441	SEL1L2	Sel-1 suppressor of lin-12-like 2 (C. elegans)	Sel1
N0.HOG0024219	6.739	4.096	2.664	SLC35C2	Triose-phosphate Transporter family	Collagen, TPT
N0.HOG0010492	5.921	5.436	2.118	STK36	Kinase 36	HEAT_2, Pkinase
N0.HOG0042927	8.753	2.641	1.976	PIK3CA	Hypomethylation of CpG island	PI3K_C2, PI3K_p85B, PI3K_rbd, PI3Ka, PI3_PI4_kinase
N0.HOG0005131	1.373	7.395	4.541	-	MYND finger	SET, zf-MYND
N0.HOG0042791	8.959	2.402	1.941	IFT88	Regulation of autophagosome assembly	TPR_1, TPR_12, TPR_16, TPR_2, TPR_6, TPR_7, TPR_8
N0.HOG0015249	6.326	4.849	2.035	CLCA1	Chloride channel	CLCA, VWA, VWA_2
N0.HOG0042799	8.083	1.791	3.213	KIAA0513	Myotubularin protein	SBF2
N0.HOG0017670	4.905	3.829	4.303	MYO3A	Belongs to the TRAFAC class myosin-kinesin ATPase superfamily. Myosin family	IQ, Myosin_head, Pkinase
N0.HOG0014354	1.271	2.550	9.010	FRS2	Phosphotyrosine-binding domain (IRS1-like)	IRS
N0.HOG0004575	2.652	4.256	5.822	GDA	Guanine deaminase	Amidohydro_1
N0.HOG0043155	5.800	4.223	2.591	RTEL1	ATP-dependent DNA helicase implicated in DNA repair and the maintenance of genomic stability. Acts as an anti-recombinase to counteract toxic recombination and limit crossover during meiosis. Regulates meiotic recombination and crossover homeostasis by physically dissociating strand invasion events and thereby promotes noncrossover repair by meiotic synthesis dependent strand annealing (SDSA) as well as disassembly of D loop recombination intermediates	DEAD_2, Helicase_C_2
N0.HOG0000744	5.724	5.347	1.416	-	Belongs to the helicase family	Helitron_like_N, Herpes_Helicase, PIF1
N0.HOG0003108	5.016	3.531	3.669	INTS1	Protein of unknown function (DUF3677)	DUF3677
N0.HOG0005335	2.527	2.963	6.566	NEDD1	WD40 repeats	ANAPC4_WD40, WD40
N0.HOG0042918	5.511	3.442	3.069	CASD1	10 TM Acyl Transferase domain found in Cas1p	Cas1_AcylT
N0.HOG0000837	8.198	2.760	1.016	-	Domain of unknown function (DUF4598)	DUF4598
N0.HOG0015886	1.072	3.879	6.731	ABCC10	It is involved in the biological process described with transmembrane transport	ABC_membrane, ABC_tran
N0.HOG0014077	4.802	3.551	3.224	-	-	-

Table S2 continued. (The top 50 up-regulated genes of P. anophtalmus. This list contains an overlap of genes that are up-regulated in P. anophtalmus in all three pairwise comparisons (P. anophtalmus (PA) vs P. hercegovinensis (PH), P. anophtalmus vs P. coxalis (PC), and P. anophtalmus vs P. karamani (PK)). A sum of L2FC values determined the order of the list. A preferred name for the gene, a brief description of the function, and PFAM domains are also listed.)

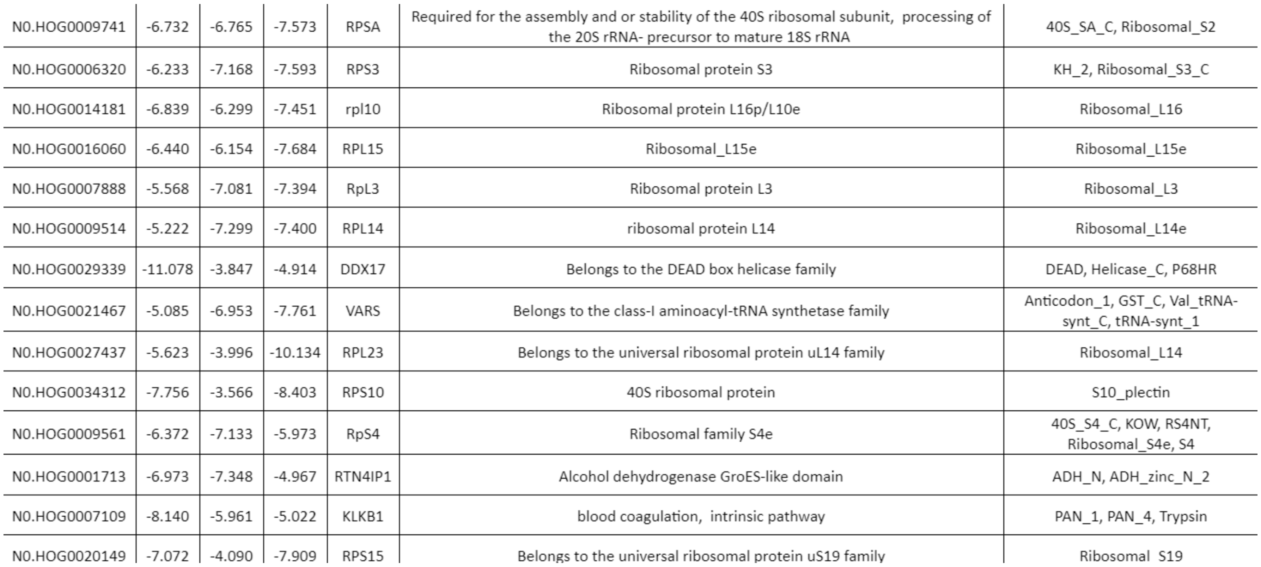


N0.HOG0043050	4.592	4.134	2.840	FRAS1	Extracellular matrix	Cadherin_3, Calx-beta, VWC
N0.HOG0007367	1.818	6.354	3.319	RYK	Chemorepulsion of axon	Pkinase_Tyr, WIF
N0.HOG0015963	7.273	2.766	1.435	KNTC1	Rough deal protein C-terminal region	Rod_C
N0.HOG0023218	3.551	6.254	1.536	TMC7	TMC domain	TMC
N0.HOG0011283	2.988	3.684	4.650	TMEM132E	Mature oligodendrocyte transmembrane protein, TMEM132D, C-term	TMEM132, TMEM132D_C, TMEM132D_N
N0.HOG0010934	2.545	1.180	7.578	BTBD6	PHR domain	BACK, BTB, PHR
N0.HOG0023177	0.759	6.897	3.621	ANLN	Anillin N-terminus	Anillin, Anillin_N, PH
N0.HOG0017860	6.091	1.655	3.434	GZF1	GDNF-inducible zinc finger protein	BTB, zf-C2H2, zf-C2H2_6, zf-met
N0.HOG0016053	2.398	3.292	5.463	SLC24A4	Calcium, potassium:sodium antiporter activity	Na_Ca_ex
N0.HOG0006318	2.809	5.410	2.890	NOX4	NADPH oxidase 4	FAD_binding_8, Ferric_reduct, NAD_binding_6
N0.HOG0008959	1.508	4.709	4.852	DZIP3	DAZ interacting zinc finger protein 3	zf-RING_2
N0.HOG0023695	2.395	6.784	1.825	SMPD2	Endonuclease/Exonuclease/phosphatase family	Exo_endo_phos
N0.HOG0007387	2.012	2.493	6.360	PDSS1	Polyprenyl synthetase	polyprenyl_synt
N0.HOG0013719	8.184	1.520	1.141	SCLY	Selenocysteine lyase activity	Aminotran_5
N0.HOG0042461	2.754	5.813	2.212	-	Protein of unknown function (DUF1647)	DUF1647
N0.HOG0021137	2.751	4.462	3.526	UGT2A1	Transferase activity, transferring hexosyl groups. It is involved in the biological process described with metabolic process	UDPGT
N0.HOG0014266	3.059	2.994	4.494	-	Telomerase Cajal body protein	WD40
N0.HOG0009706	1.639	2.285	6.618	ECT2	Guanine nucleotide exchange factor for Rho/Rac/Cdc42-like GTPases	BRCT, PTCB-BRCT, RhoGEF
N0.HOG0018024	4.805	1.854	3.847	UVRAG	UV radiation	Atg14

Table S3 The top 50 down-regulated genes of P. anophtalmus. This list contains an overlap of genes that are down-regulated in P. anophtalmus in all three pairwise comparisons (P. anophtalmus (PA) vs P. hercegovinensis (PH), P. anophtalmus vs P. coxalis (PC), and P. anophtalmus vs P. karamani (PK)). A sum of L2FC values determined the order of the list. A preferred name for the gene, a brief description of the function, and PFAM domains are also listed.

Gene ID/HOG ID	L2FC vs PH	L2FC vs PC	L2FC vs PK	Preferred name	Description	PFAMs
N0.HOG0008826	-10.829	-9.148	-11.685	Mlc1	mesoderm development	-
N0.HOG0033792	-10.344	-7.701	-10.580	RpS18	Belongs to the universal ribosomal protein uS13 family	Ribosomal_S13
N0.HOG0018492	-8.525	-8.828	-9.353	RPL12	Ribosomal protein L11/L12	Ribosomal_L11, Ribosomal_L11_N
N0.HOG0007868	-10.054	-7.709	-8.213	RpL5	Ribosomal large subunit proteins 60S L5, and 50S L18	Ribosomal_L18_c, Ribosomal_L5e
N0.HOG0023297	-7.953	-9.366	-7.906	RPL27	structural constituent of ribosome	KOW, Ribosomal_L27e
N0.HOG0040417	-8.862	-9.384	-6.686	RPS20	RNA binding. It is involved in the biological process described with translation	Ribosomal_S10
N0.HOG0004018	-8.593	-7.811	-8.524	Hsc70-4	Heat shock 70 kDa protein cognate	HSP70
N0.HOG0019979	-7.794	-8.055	-8.990	rpl35	60S ribosomal protein L35	Ribosomal_L29
N0.HOG0009332	-8.762	-9.415	-6.502	Mlc2	Myosin regulatory light chain	EF-hand_6, EF-hand_8
N0.HOG0028587	-8.532	-8.319	-6.964	RPL27A	ribosomal protein	Ribosomal_L27A
N0.HOG0012884	-9.716	-7.228	-6.826	RPS2	Belongs to the universal ribosomal protein uS5 family	Ribosomal_S5, Ribosomal_S5_C
N0.HOG0017834	-7.019	-7.627	-9.063	RPLP1	Belongs to the eukaryotic ribosomal protein P1 P2 family	Ribosomal_60s
N0.HOG0000567	-11.986	-4.289	-7.315	EEF1A1	This protein promotes the GTP-dependent binding of aminoacyl-tRNA to the A-site of ribosomes during protein biosynthesis	GTP_EFTU, GTP_EFTU_D2, GTP_EFTU_D3
N0.HOG0024894	-6.338	-8.027	-9.173	RPS17	Ribosomal protein S17	Ribosomal_S17e
N0.HOG0029255	-6.707	-8.373	-8.455	RPS26	Belongs to the eukaryotic ribosomal protein eS26 family	Ribosomal_S26e
N0.HOG0003123	-7.950	-5.001	-10.527	TNNI1	Troponin	Troponin, Troponin-I_N
N0.HOG0004314	-5.477	-8.377	-9.595	NME2	protein histidine kinase activity	NDK
N0.HOG0014330	-10.309	-7.056	-5.837	FABP6	lipid binding	Lipocalin_7
N0.HOG0025008	-7.725	-8.260	-7.149	rps15a	Ribosomal protein S8	Ribosomal_S8
N0.HOG0013905	-7.654	-8.078	-7.297	RPL7	Ribosomal L30 N-terminal domain	Ribosomal_L30, Ribosomal_L30_N
N0.HOG0015918	-8.316	-6.105	-8.587	TPT1	tumor protein	TCTP
N0.HOG0024365	-7.198	-9.248	-6.527	RPL18A	Belongs to the eukaryotic ribosomal protein eL20 family	Ribosomal_L18A
N0.HOG0024030	-6.025	-8.338	-8.235	RPS13	Belongs to the universal ribosomal protein uS15 family	Ribosomal_S13_N, Ribosomal_S15
N0.HOG0029293	-10.383	-6.842	-5.198	LYZ	Alpha-lactalbumin / lysozyme C	Lys
N0.HOG0015916	-8.299	-6.356	-7.665	FABP5	Lipocalin / cytosolic fatty-acid binding protein family	Lipocalin
N0.HOG0004395	-8.206	-7.132	-6.878	RPS12	Belongs to the eukaryotic ribosomal protein eS12 family	Ribosomal_L7Ae
N0.HOG0007092	-7.865	-6.565	-7.481	RpL8	Ribosomal Proteins L2, C-terminal domain	Ribosomal_L2, Ribosomal_L2_C
N0.HOG0000821	-5.966	-7.802	-7.787	fln	-	-
N0.HOG0007729	-8.210	-6.804	-6.538	RPS23	Ribosomal protein S12/S23	Ribosom_S12_S23
N0.HOG0003453	-0.793	-9.983	-10.774	CNOT10	CCR4-NOT transcription complex subunit	TPR_8
N0.HOG0012820	-8.502	-6.164	-6.827	CTRC	Serine-type endopeptidase activity. It is involved in the biological process described with proteolysis	Trypsin
N0.HOG0000661	-11.249	-3.452	-6.655	unc-15	Myosin tail	Myosin_tail_1
N0.HOG0021117	-5.797	-8.164	-7.379	RPS25	Ribosomal protein S25	Ribosomal_S25
N0.HOG0007400	-7.618	-5.870	-7.832	RPL32	Ribosomal_L32e	Ribosomal_L32e
N0.HOG0007143	-6.517	-7.500	-7.068	RPS8	Belongs to the eukaryotic ribosomal protein eS8 family	Ribosomal_S8e
N0.HOG0014331	-6.307	-6.611	-8.162	EEF1B2	Eukaryotic translation elongation factor 1 beta 2	EF-1_beta_acid, EF1_GNE

Table S3 continued. (The top 50 down-regulated genes of P. anophtalmus. This list contains an overlap of genes that are down-regulated in P. anophtalmus in all three pairwise comparisons (P. anophtalmus (PA) vs P. hercegovinensis (PH), P. anophtalmus vs P. coxalis (PC), and P. anophtalmus vs P. karamani (PK)). A sum of L2FC values determined the order of the list. A preferred name for the gene, a brief description of the function, and PFAM domains are also listed.)



N0.HOG0009741	-6.732	-6.765	-7.573	RPSA	Required for the assembly and or stability of the 40S ribosomal subunit, processing of the 20S rRNA- precursor to mature 18S rRNA	40S_SA_C, Ribosomal_S2
N0.HOG0006320	-6.233	-7.168	-7.593	RPS3	Ribosomal protein S3	KH_2, Ribosomal_S3_C
N0.HOG0014181	-6.839	-6.299	-7.451	rpl10	Ribosomal protein L16p/L10e	Ribosomal_L16
N0.HOG0016060	-6.440	-6.154	-7.684	RPL15	Ribosomal_L15e	Ribosomal_L15e
N0.HOG0007888	-5.568	-7.081	-7.394	RpL3	Ribosomal protein L3	Ribosomal_L3
N0.HOG0009514	-5.222	-7.299	-7.400	RPL14	ribosomal protein L14	Ribosomal_L14e
N0.HOG0029339	-11.078	-3.847	-4.914	DDX17	Belongs to the DEAD box helicase family	DEAD, Helicase_C, P68HR
N0.HOG0021467	-5.085	-6.953	-7.761	VARS	Belongs to the class-I aminoacyl-tRNA synthetase family	Anticodon_1, GST_C, Val_tRNA-synt_C, tRNA-synt_1
N0.HOG0027437	-5.623	-3.996	-10.134	RPL23	Belongs to the universal ribosomal protein uL14 family	Ribosomal_L14
N0.HOG0034312	-7.756	-3.566	-8.403	RPS10	40S ribosomal protein	S10_plectin
N0.HOG0009561	-6.372	-7.133	-5.973	RpS4	Ribosomal family S4e	40S_S4_C, KOW, RS4NT, Ribosomal_S4e, S4
N0.HOG0001713	-6.973	-7.348	-4.967	RTN4IP1	Alcohol dehydrogenase GroES-like domain	ADH_N, ADH_zinc_N_2
N0.HOG0007109	-8.140	-5.961	-5.022	KLKB1	blood coagulation, intrinsic pathway	PAN_1, PAN_4, Trypsin
N0.HOG0020149	-7.072	-4.090	-7.909	RPS15	Belongs to the universal ribosomal protein uS19 family	Ribosomal_S19

9.2.5. Differentially expressed genes of *P. hercegovinensis*

Table S4 The top 50 up-regulated genes of P. hercegovinensis. This list contains an overlap of genes that are up-regulated in P. anophtalmus in all three pairwise comparisons (P. hercegovinensis (PH) vs P. anophtalmus (PA), P. hercegovinensis vs P. coxalis (PC), and P. hercegovinensis vs P. karamani (PK)). A sum of L2FC values determined the order of the list. A preferred name for the gene, a brief description of the function, and PFAM domains are also listed.

Gene ID/HOG ID	L2FC vs PA	L2FC vs PC	L2FC vs PK	Preferred name	Description	PFAMs
N0.HOG0000661	11.249	3.841	1.567	unc-15	Myosin tail	Myosin_tail_1
N0.HOG0000628	5.385	6.908	3.044	-	Calcium ion binding	EGF, Ldl_recept_b, hEGF
N0.HOG0034419	4.656	6.213	3.637	LSM14A	Homolog A	FDF, LSM14
N0.HOG0008310	3.220	4.826	6.392	Jafrac1	activity. It is involved in the biological process described with oxidation-reduction process	1-cysPrx_C, AhpC-TSA
N0.HOG0014294	1.109	1.885	9.873	NEDD9	Domain of unknown function (DUF3513)	DUF3513, SH3_1, SH3_9, Serine_rich
N0.HOG0000582	6.381	1.572	4.855	HNRNPLL	Nucleotide binding. It is involved in the biological process described with mRNA processing	RRM_1, RRM_5
N0.HOG0010058	7.912	2.104	2.688	-	Transcriptional regulator	Transcrip_reg
N0.HOG0028285	2.346	5.131	4.484	ZBTB8OS	Zinc finger and BTB domain containing 8 opposite strand	Archease
N0.HOG0008286	2.246	2.304	7.143	MRPL22	Belongs to the universal ribosomal protein uL22 family	Ribosomal_L22
N0.HOG0020505	3.114	3.122	5.196	-	carbohydrate binding	F5_F8_type_C, Laminin_G_3, Lectin_C, Methyltransf_FA, PAN_1, SEA
N0.HOG0002315	5.615	3.423	2.204	RAVER2	RNA recognition motif	RRM_1
N0.HOG0000433	2.949	5.052	2.824	DMXL1	WD40 repeats	Rav1p_C, WD40
N0.HOG0027846	1.432	1.453	7.863	-	-	-
N0.HOG0021236	1.444	4.764	4.389	PPAN	Brix	7tm_1, Brix
N0.HOG0013963	3.143	2.444	4.983	-	Transporter activity. It is involved in the biological process described with transport	CRAL_TRIO
N0.HOG0002707	2.212	2.307	5.960	SYAP1	Synapse-associated protein	BSD
N0.HOG0011937	1.614	0.772	7.874	EGLN1	2OG-Fe(II) oxygenase superfamily	2OG-FeII_Oxy_3, zf-MYND
N0.HOG0007553	3.983	3.915	2.303	LDB3	Zinc-binding domain present in Lin-11, Isl-1, Mec-3.	DUF4749, LIM, PDZ
N0.HOG0014133	3.656	3.028	3.348	PSMD14	Lys63-specific deubiquitinase activity	JAB, MitMem_reg
N0.HOG0024607	4.274	3.015	2.465	TpnC4	Troponin C	EF-hand_1, EF-hand_6, EF-hand_7
N0.HOG0008527	1.283	3.064	5.120	SGCG	Sarcoglycan complex subunit protein	Sarcoglycan_1
N0.HOG0009985	2.291	4.402	2.611	DIRAS1	GTP binding. It is involved in the biological process described with	Ras
N0.HOG0015958	4.296	3.910	1.030	CHCHD2	Regulation of cellular response to hypoxia.	CHCH
N0.HOG0020759	2.702	4.839	1.678	SF3B6	Pfam:RRM_6	RRM_1
N0.HOG0029250	2.791	3.659	2.688	UTP14A	Utp14 protein	Utp14
N0.HOG0017451	3.437	2.540	2.946	-	O-methyltransferase	Methyltransf_24, Methyltransf_3
N0.HOG0016042	1.864	3.526	3.521	-	NADH dehydrogenase (ubiquinone) activity. It is involved in the biological process described with ATP synthesis coupled electron transport	CI-B14_5a
N0.HOG0014184	1.079	3.708	4.108	UBTD2	Ubiquitin-binding domain	UBD, ubiquitin
N0.HOG0015576	2.276	4.468	2.088	SSR3	It is involved in the biological process described with cotranslational protein targeting to membrane	TRAP-gamma
N0.HOG0005725	3.062	0.803	4.727	CLSTN2	Calcium ion binding. It is involved in the biological process described with homophilic cell adhesion via plasma membrane adhesion molecules	Cadherin, Laminin_G_3
N0.HOG0004143	0.880	1.444	6.225	SETD3	Belongs to the class V-like SAM-binding methyltransferase superfamily. Histone-lysine methyltransferase family. SETD3 subfamily	Rubis-subs-bind, SET

Table S4 continued. (The top 50 up-regulated genes of P. hercegovinensis. This list contains an overlap of genes that are up-regulated in P. anophtalmus in all three pairwise comparisons (P. hercegovinensis (PH) vs P. anophtalmus (PA), P. hercegovinensis vs P. coxalis (PC), and P. hercegovinensis vs P. karamani (PK)). A sum of L2FC values determined the order of the list. A preferred name for the gene, a brief description of the function, and PFAM domains are also listed.)

N0.HOG0000926	1.472	1.548	5.465	pgbd4	Transposase IS4	DDE_Tnp_1_7, Tnp_zf-ribbon_2
N0.HOG0016469	3.333	2.520	2.501	ACTR3B	Belongs to the actin family	Actin
N0.HOG0003153	1.019	4.837	2.375	Nle	NLE (NUC135) domain	NLE, WD40
N0.HOG0002278	1.867	1.640	4.717	NIT2	Nitrilase family, member 2	CN_hydrolase
N0.HOG0009716	3.202	3.305	1.600	RWDD1	positive regulation of androgen receptor activity	DFRP_C, RWD
N0.HOG0016263	1.412	3.244	3.414	-	BCL (B-Cell lymphoma); contains BH1, BH2 regions	Bcl-2
N0.HOG0008650	3.066	2.430	2.556	AAMP	WD domain, G-beta repeat	WD40
N0.HOG0013876	2.132	2.475	3.435	MLH1	Heterodimerizes with Pms2 to form MutL alpha, a component of the post-replicative DNA mismatch repair system (MMR).	DNA_mis_repair, HATPase_c_3, Mlh1_C
N0.HOG0008470	1.983	2.978	3.013	EAF1	RNA polymerase II transcription elongation factor	EAF
N0.HOG0008087	3.435	2.400	2.101	Car15	Eukaryotic-type carbonic anhydrase	Carb_anhydrase
N0.HOG0015321	3.984	3.056	0.872	ATP5D	proton-transporting ATPase activity, rotational mechanism. It is involved in the biological process described with ATP synthesis coupled proton transport	ATP-synt_DE_N
N0.HOG0023416	1.459	4.029	2.310	UGCG	Transferase activity, transferring glycosyl groups	Glyco_transf_21
N0.HOG0014311	1.440	3.204	3.150	RBM42	RNA recognition motif	RRM_1
N0.HOG0006810	1.231	4.010	2.495	RDH13	Retinol dehydrogenase 13	adh_short
N0.HOG0018653	2.786	3.812	1.123	-	Nucleoplasmin/nucleophosmin domain	Nucleoplasmin
N0.HOG0013575	1.636	0.751	5.253	TWISTNB	Transcription by RNA polymerase I	SHS2_Rpb7-N
N0.HOG0035172	4.037	2.488	0.955	CSNK1D	Casein kinase I isoform delta	Pkinase
N0.HOG0013751	2.136	1.318	3.984	UHMK1	U2AF homology motif (UHM) kinase 1	Pkinase, RRM_1
N0.HOG0004486	1.713	2.566	3.107	FBXL6	F-box-like	F-box-like, LRR_6
N0.HOG0009479	4.383	1.895	1.038	HYOU1	Belongs to the heat shock protein 70 family	HSP70

Table S5 The top 50 down-regulated genes of P. hercegovinensis. This list contains an overlap of genes that are down-regulated in P. anophtalmus in all three pairwise comparisons (P. hercegovinensis (PH) vs P. anophtalmus (PA), P. hercegovinensis vs P. coxalis (PC), and P. hercegovinensis vs P. karamani (PK)). A sum of L2FC values determined the order of the list. A preferred name for the gene, a brief description of the function, and PFAM domains are also listed.

Gene ID/HOG ID	L2FC PH vs PA	L2FC PH vs PC	L2FC PH vs PK	Preferred name	Description	PFAMs
N0.HOG0004277	-3.093	-11.006	-2.328	CPSF3L	Beta-lactamase superfamily domain	Beta-Casp, Lactamase_B_6, RMMBL
N0.HOG0042743	-8.486	-5.531	-1.387	CELF2	nucleic acid binding	RRM_1
N0.HOG0016423	-9.317	-4.545	-0.924	ATAD2	ATPase family associated with various cellular activities (AAA)	AAA, Bromodomain
N0.HOG0000831	-4.696	-4.034	-5.965	PAH	Biopterin-dependent aromatic amino acid hydroxylase	ACT, Biopterin_H
N0.HOG0010128	-1.547	-3.833	-9.031	RPL4	Structural constituent of ribosome. It is involved in the biological process described with	Ribos_L4_asso_C, Ribosomal_L4
N0.HOG0024202	-2.919	-6.402	-4.939	-	RNA splicing	RRM_1
N0.HOG0015360	-5.672	-6.870	-0.964	CDC20	anaphase-promoting complex binding	ANAPC4_WD40, WD40
N0.HOG0014192	-8.153	-1.657	-1.620	Hsc70-4	Hsp70 protein	HSP70
N0.HOG0024373	-2.435	-3.991	-4.951	-	Zinc finger protein	zf-C2H2, zf-C2H2_11, zf-C2H2_4
N0.HOG0043002	-7.382	-2.193	-1.396	POLE	DUF1744	DNA_pol_B, DNA_pol_B_exo1, DUF1744
N0.HOG0018664	-3.041	-6.364	-1.485	-	Got1/Sft2-like family	Got1
N0.HOG0007877	-7.019	-2.088	-1.485	NIT1	Hydrolase activity, acting on carbon-nitrogen (but not peptide) bonds. It is involved in the biological process described with nitrogen compound metabolic process	CN_hydrolase, HIT
N0.HOG0028217	-2.056	-4.485	-3.436	AGPAT6	transferase activity, transferring acyl groups. It is involved in the biological process described with metabolic process	Acyltransferase
N0.HOG0021260	-3.941	-3.658	-2.038	CDH23	Calcium ion binding. It is involved in the biological process described with homophilic cell adhesion via plasma membrane adhesion molecules	Cadherin
N0.HOG0018516	-5.375	-1.831	-2.166	JAGN1	Jagunal, ER re-organisation during oogenesis	Jagunal
N0.HOG0002512	-3.280	-4.938	-0.941	ARID4B	RNA binding activity-knot of a chromodomain	ARID, RBB1NT, Tudor-knot
N0.HOG0009302	-1.470	-3.638	-3.933	MRPL3	Ribosomal protein L3	Ribosomal_L3
N0.HOG0018075	-4.426	-2.637	-1.949	ADAMTS9	It is involved in the biological process described with proteolysis	ADAM_spacer1, GON, Pep_M12B_propep, Reprolysin, TSP_1
N0.HOG0014010	-1.343	-6.338	-1.281	-	MYND finger	SET, zf-MYND
N0.HOG0006183	-3.947	-3.435	-1.574	-	Elongation of very long chain fatty acids protein	ELO
N0.HOG0023714	-2.599	-4.299	-1.990	SLC39A11	ZIP Zinc transporter	Zip
N0.HOG0007900	-0.831	-5.949	-2.043	MED23	Mediator complex subunit 23	Med23
N0.HOG0016209	-4.268	-2.811	-1.354	SLC22A5	carnitine transmembrane transporter activity	MFS_1, Sugar_tr
N0.HOG0009422	-5.284	-1.751	-1.339	BRWD3	bromo domain	Bromodomain, WD40
N0.HOG0009774	-3.757	-1.156	-3.413	-	Zinc ion binding	Bromodomain, PHD, PWWP
N0.HOG0013510	-2.525	-4.177	-1.386	-	serine-type endopeptidase activity. It is involved in the biological process described with proteolysis	Trypsin
N0.HOG0007278	-4.338	-3.052	-0.660	B4GALT7	N-terminal region of glycosyl transferase group 7	Glyco_transf_7C, Glyco_transf_7N
N0.HOG0008671	-5.573	-1.188	-1.237	DAXX	Death domain-associated protein 6	Daxx
N0.HOG0010851	-2.645	-3.146	-2.129	uzip	Protein of unknown function (DUF3421)	DUF3421
N0.HOG0011289	-4.323	-1.704	-1.805	FGGY	FGGY family of carbohydrate kinases, C-terminal domain	FGGY_C, FGGY_N

Table S5 continued (The top 50 down-regulated genes of P. hercegovinensis. This list contains an overlap of genes that are down-regulated in P. anophtalmus in all three pairwise comparisons (P. hercegovinensis (PH) vs P. anophtalmus (PA), P. hercegovinensis vs P. coxalis (PC), and P. hercegovinensis vs P. karamani (PK)). A sum of L2FC values determined the order of the list. A preferred name for the gene, a brief description of the function, and PFAM domains are also listed.)

N0.HOG0000612	-0.651	-5.661	-1.510	NFIB	Recognizes and binds the palindromic sequence 5'- TTGGCNNNNNGCCAA-3' present in viral and cellular promoters and in the origin of replication of adenovirus type 2. These proteins are individually capable of activating transcription and replication	CTF_NFI, MH1, NfI_DNAbd_pre-N
N0.HOG0010778	-2.584	-2.622	-2.562	LARP6	Domain in the RNA-binding Lupus La protein; unknown function	La, SUZ-C
N0.HOG0001395	-2.384	-1.238	-3.854	Hr39	Ligand binding domain of hormone receptors	Hormone_recep, zf-C4
N0.HOG0042944	-4.416	-1.943	-1.050	LIPI	Lipase, member I	Lipase
N0.HOG0001840	-1.520	-4.605	-1.222	QPCTL	Glutaminyl-peptide cyclotransferase-like	Peptidase_M28
N0.HOG0013866	-0.967	-3.497	-2.852	-	Immunoglobulin C-2 Type	I-set, Ig_3, V-set
N0.HOG0008456	-2.900	-2.876	-1.495	CAPN1	calpain_III	Calpain_III, Calpain_u2, EF-hand_5, EF-hand_8, Peptidase_C2
N0.HOG0012609	-3.338	-2.269	-1.624	-	C2H2-type zinc finger	zf-C2H2
N0.HOG0001132	-1.462	-2.808	-2.939	IKBKE	Protein tyrosine kinase	Pkinase
N0.HOG0015518	-2.151	-3.467	-1.526	WDR59	response to amino acid starvation	WD40, Zn_ribbon_17
N0.HOG0002638	-2.623	-1.345	-3.105	CREG2	Cellular repressor of	Pyrid_oxidase_2
N0.HOG0007653	-0.874	-5.083	-0.954	RAG1	V(D)J recombination-activating protein 1	RAG1, RAG1_imp_bd, zf-C3HC4, zf-C3HC4_2, zf-RAG1
N0.HOG0004555	-4.179	-0.887	-1.845	FAM43A	Phosphotyrosine interaction domain (PTB/PID)	PID_2
N0.HOG0012947	-3.021	-1.779	-1.841	SLC28A3	Nucleoside sodium symporter activity. It is involved in the biological process described with transport	Gate, Nucleos_tra2_C, Nucleos_tra2_N
N0.HOG0013993	-2.746	-1.821	-1.998	MECOM	nucleic acid binding	zf-C2H2, zf-C2H2_6
N0.HOG0005610	-1.495	-4.128	-0.874	MFNG	Fringe-like	Fringe
N0.HOG0004428	-1.418	-3.283	-1.695	METTL13	methyltransferase like 13	Methyltransf_11, Methyltransf_25, Methyltransf_31, Spermine_synth
N0.HOG0007053	-0.810	-4.451	-1.077	FAAH2	Carbon-nitrogen ligase activity, with glutamine as amido-N-donor	Amidase
N0.HOG0011056	-0.923	-2.076	-3.252	Ppib	PPIases accelerate the folding of proteins. It catalyzes the cis-trans isomerization of proline imidic peptide bonds in oligopeptides	Pro_isomerase
N0.HOG0008896	-1.427	-3.403	-1.397	UBA5	Ubiquitin-like modifier-activating enzyme 5	ThiF

9.2.6. Common gene expression patterns of *P. anophtalmus* and *P. hercegovinesis*

Table S6 Commonly down-regulated genes of P. anophtalmus and P. hercegovinensis. This list contains an overlap of genes that are down-regulated both in P. anophtalmus and P. hercegovinensis in four pairwise comparisons (P. anophtalmus (PA) vs P. coxalis (PC), P. anophtalmus vs P. karamani (PK), P. hercegovinensis (PH) vs P. coxalis (PC), and P. hercegovinensis vs P. karamani (PK)). A sum of L2FC values determined the order of the list. A preferred name for the gene, a brief description of the function, and PFAM domains are also listed.

Gene ID/HOG ID	L2FC PA vs PC	L2FC PH vs PC	L2FC PA vs PK	L2FC PH vs PK	Preferred name	Description	PFAMs
N0.HOG0028738	-0.078	-10.272	-4.313	-6.252	-	Structural constituent of cuticle	Chitin_bind_4
N0.HOG0024600	-1.259	-3.581	-7.066	-3.981	-	Insect cuticle protein	Chitin_bind_4
N0.HOG0005562	-3.295	-0.612	-4.980	-6.887	MCEE	methylmalonyl-CoA epimerase, mitochondrial	Glyoxalase_4
N0.HOG0024777	-5.059	-1.349	-4.131	-4.909	TBL2	WD domain, G-beta repeat	WD40
N0.HOG0007344	-3.335	-4.283	-4.867	-2.950	-	HIT domain	HIT
N0.HOG0016290	-3.783	-4.157	-3.949	-3.081	NDUFS8	4Fe-4S dicluster domain	Fer4
N0.HOG0007847	-3.517	-1.631	-5.070	-4.744	E75	It is involved in the biological process described with binding.	Hormone_recep, zf-C4
N0.HOG0015480	-4.562	-2.367	-5.570	-2.305	-	Insect cuticle protein	Chitin_bind_4
N0.HOG0016111	-1.308	-7.982	-2.815	-1.660	-	PMP-22/EMP/MP20/Claudin family	PMP22_Claudin
N0.HOG0011056	-2.854	-2.076	-5.410	-3.252	Ppib	PPIases accelerate the folding of proteins. It catalyzes the cis-trans isomerization of proline imidic peptide bonds in oligopeptides	Pro_isomerase
N0.HOG0035152	-3.008	-2.575	-4.241	-3.687	GAR1	Required for ribosome biogenesis. Part of a complex which catalyzes pseudouridylation of rRNA. This involves the isomerization of uridine such that the ribose is subsequently attached to C5, instead of the normal N1	Gar1
N0.HOG0020582	-4.577	-1.839	-3.258	-3.728	TMEM14C	Transmembrane proteins 14C	Tmemb_14
N0.HOG0023512	-5.217	-1.492	-1.556	-4.890	IMMT	Component of the MICOS complex, a large protein complex of the mitochondrial inner membrane that plays crucial roles in the maintenance of crista junctions, inner membrane architecture, and formation of contact sites to the outer membrane	Mitofilin
N0.HOG0013592	-5.179	-2.026	-4.203	-1.594	-	Acidic leucine-rich nuclear phosphoprotein 32-related protein	LRR_4, LRR_9
N0.HOG0000910	-1.825	-7.119	-0.965	-3.066	NID1	Calcium ion binding. It is involved in the biological process described with cell-matrix adhesion	EGF, EGF_3, EGF_CA, FXa_inhibition, G2F, Ldl_recept_b, NIDO, Thyroglobulin_1, cEGF
N0.HOG0012900	-2.323	-3.272	-4.980	-2.034	WDR74	WD repeat-containing protein	WD40
N0.HOG0017480	-3.072	-1.384	-3.796	-4.276	PIN4	Peptidyl-prolyl cis-trans isomerase	Rotamase_3
N0.HOG0012973	-2.353	-1.656	-3.568	-4.556	CCDC86	Coiled-coil domain-containing protein	Cgr1
N0.HOG0020147	-1.096	-1.203	-4.434	-5.032	PSMA7	threonine-type endopeptidase activity	Proteasome, Proteasome_A_N
N0.HOG0024914	-0.036	-3.968	-3.257	-4.500	STT3B	Oligosaccharyl transferase activity. It is involved in the biological process described with protein glycosylation	STT3
N0.HOG0014429	-3.709	-2.433	-3.648	-1.754	NDUFB11	ESSS subunit of NADH:ubiquinone oxidoreductase (complex I)	ESSS
N0.HOG0041464	-3.125	-2.113	-3.189	-3.092	EPB41L4B	cytoskeletal protein binding	FA, FERM_C, FERM_M, FERM_N
N0.HOG0024370	-1.406	-3.656	-3.364	-3.064	MRPL55	Mitochondrial ribosomal protein L55	Mitoc_L55
N0.HOG0018014	-1.527	-2.341	-4.867	-2.704	MIDN	Midnolin	ubiquitin
N0.HOG0009365	-1.785	-2.196	-2.298	-4.563	RAB32	GTP binding. It is involved in the biological process described with protein transport	Ras
N0.HOG0005699	-1.047	-6.483	-1.670	-1.420	-	NADPHX epimerase activity	YjeF_N
N0.HOG0018259	-1.792	-0.729	-3.924	-4.001	Jtb	Jumping translocation breakpoint protein (JTB)	JTB
N0.HOG0024436	-0.739	-1.857	-3.575	-4.207	msn	It is involved in the biological process described with protein phosphorylation	CNH, Pkinase
N0.HOG0006202	-0.308	-6.104	-1.652	-2.204	GADD45GIP1	Growth arrest and DNA-damage-inducible proteins-interacting protein 1	CR6_interact

Table S6 continued. (Commonly down-regulated genes of P. anophtalmus and P. hercegovinensis. This list contains an overlap of genes that are down-regulated both in P. anophtalmus and P. hercegovinensis in four pairwise comparisons (P. anophtalmus (PA) vs P. coxalis (PC), P. anophtalmus vs P. karamani (PK), P. hercegovinensis (PH) vs P. coxalis (PC), and P. hercegovinensis vs P. karamani (PK)). A sum of L2FC values determined the order of the list. A preferred name for the gene, a brief description of the function, and PFAM domains are also listed.)

N0.HOG0011397	-2.184	-1.444	-2.865	-3.639	-	Chitin-binding domain type 2	CBM_14
N0.HOG0000746	-3.519	-1.961	-2.121	-2.419	cher	actin binding	CH, Filamin
N0.HOG0004426	-2.064	-3.507	-2.284	-2.156	NDUFA9	Coenzyme binding	3Beta_HSD, Epimerase, NAD_binding_10, NmrA
N0.HOG0024751	-0.300	-5.336	-1.738	-2.386	SF3B3	Splicing factor 3b, subunit 3	CPSF_A, MMS1_N
N0.HOG0018067	-4.250	-2.954	-0.842	-1.458	TCEB1	Belongs to the SKP1 family	Skp1_POZ
N0.HOG0029511	-1.461	-3.950	-2.020	-1.954	PLCB4	Protein of unknown function (DUF1154)	C2, DUF1154, EF-hand_like, PI-PLC-X, PI-PLC-Y
N0.HOG0000909	-1.709	-1.238	-5.210	-1.077	SSRP1	Component of the FACT complex, a general chromatin factor that acts to reorganize nucleosomes. The FACT complex is involved in multiple processes that require DNA as a template such as mRNA elongation, DNA replication and DNA repair.	HMG_box, POB3_N, Rtt106, SSrecog
N0.HOG0004737	-2.349	-1.134	-1.910	-3.415	PITPNM2	LNS2	DDHD, IP_trans
N0.HOG0008249	-0.131	-1.155	-3.488	-3.944	TNXB	Fibronectin type 3 domain	EGF_2, Fibrinogen_C, fn1, fn3
N0.HOG0012756	-1.503	-3.567	-1.399	-2.054	CDK1	RNA polymerase II CTD heptapeptide repeat kinase activity	Pkinase
N0.HOG0011380	-1.480	-2.104	-2.142	-2.694	UCK1	Phosphoribulokinase / Uridine kinase family	PRK
N0.HOG0015917	-2.617	-0.670	-2.242	-2.860	SRP19	SRP-dependent cotranslational protein targeting to membrane, signal sequence recognition	SRP19
N0.HOG0004313	-1.024	-0.575	-4.711	-2.026	-	Male enhanced antigen 1 (MEA1)	MEA1
N0.HOG0014016	-2.608	-2.437	-1.675	-1.572	ITFG1	integrin alpha FG-GAP repeat containing 1	VCBS
N0.HOG0012344	-1.193	-2.114	-2.817	-1.915	IDE	Belongs to the peptidase M16 family	Peptidase_M16, Peptidase_M16_C, Peptidase_M16_M
N0.HOG0007848	-2.303	-1.024	-1.372	-3.126	NDUFA10	Accessory subunit of the mitochondrial membrane respiratory chain NADH dehydrogenase (Complex I), that is believed not to be involved in catalysis. Complex I functions in the transfer of electrons from NADH to the respiratory chain. The immediate electron acceptor for the enzyme is believed to be ubiquinone	dNK
N0.HOG0028170	-1.647	-3.334	-1.113	-1.674	VPS53	Vacuolar protein sorting-associated protein 53 homolog	Vps53_N
N0.HOG0003091	-1.093	-3.693	-1.065	-1.882	RUFY1	Metal ion binding	FYVE, RUN
N0.HOG0008762	-0.688	-3.199	-1.713	-2.014	CTU2	Plays a central role in 2-thiolation of mcm(5)S(2)U at tRNA wobble positions of tRNA(Lys), tRNA(Glu) and tRNA(Gln). May act by forming a heterodimer with NCS6 CTU1 that ligates sulfur from thiocarboxylated URM1 onto the uridine of tRNAs at wobble position	CTU2
N0.HOG0001395	-0.522	-1.238	-1.978	-3.854	Hr39	Ligand binding domain of hormone receptors	Hormone_recep, zf-C4
N0.HOG0004966	-1.814	-0.879	-3.438	-1.444	ABCE1	Possible Fer4-like domain in RNase L inhibitor, RLI	ABC_tran, Fer4, RLI
N0.HOG0020543	-1.356	-1.445	-0.905	-3.835	LMO7	myosin II head/neck binding	CH, DUF4757, LIM, PDZ
N0.HOG0006872	-1.974	-1.198	-2.604	-1.446	YIPF5	Yip1 domain	Yip1
N0.HOG0005141	-1.012	-3.137	-2.399	-0.642	-	AP-4 complex subunit	Adaptin_N
N0.HOG0010826	-1.957	-1.478	-2.295	-1.300	-	Broad-Complex, Tramtrack and Bric a brac	BTB, zf-C2H2, zf-C2H2_4, zf-met
N0.HOG0021260	-0.001	-3.658	-1.157	-2.038	CDH23	Calcium ion binding. It is involved in the biological process described with homophilic cell adhesion via plasma membrane adhesion molecules	Cadherin
N0.HOG0005226	-1.423	-1.921	-2.380	-1.129	MVK	It is involved in the biological process described with isoprenoid biosynthetic process	GHMP_kinases_C, GHMP_kinases_N
N0.HOG0007206	-0.701	-3.610	-1.435	-1.062	PECR	Peroxisomal trans-2-enoyl-CoA	adh_short, adh_short_C2
N0.HOG0004377	-1.308	-1.742	-1.769	-1.958	MGEA5	beta-N-acetylglucosaminidase	NAGidase

Table S6 continued. (Commonly down-regulated genes of P. anophtalmus and P. hercegovinensis. This list contains an overlap of genes that are down-regulated both in P. anophtalmus and P. hercegovinensis in four pairwise comparisons (P. anophtalmus (PA) vs P. coxalis (PC), P. anophtalmus vs P. karamani (PK), P. hercegovinensis (PH) vs P. coxalis (PC), and P. hercegovinensis vs P. karamani (PK)). A sum of L2FC values determined the order of the list. A preferred name for the gene, a brief description of the function, and PFAM domains are also listed.)

N0.HOG0003576	-1.827	-0.659	-2.858	-1.402	RUVBL2	Proposed core component of the chromatin remodeling Ino80 complex which is involved in transcriptional regulation, DNA replication and probably DNA repair	TIP49
N0.HOG0012560	-0.778	-1.055	-1.703	-3.193	PSMD8	It is involved in the biological process described with proteolysis	CSN8_PSD8_EIF3K
N0.HOG0004866	-0.021	-1.173	-1.538	-3.993	FAM173B	positive regulation of sensory perception of pain	-
N0.HOG0002135	-1.720	-1.748	-1.782	-1.414	FBN1	Complement Clr-like EGF-like	EGF_CA, FXa_inhibition, Sushi, TB, hEGF
N0.HOG0013739	-0.528	-1.640	-1.385	-3.101	NRD1	Belongs to the peptidase M16 family	Peptidase_M16, Peptidase_M16_C, Peptidase_M16_M
N0.HOG0004428	-0.635	-3.283	-1.026	-1.695	METTL13	methyltransferase like 13	Methyltransf_11, Methyltransf_25, Methyltransf_31, Spermine_synth
N0.HOG0014192	-1.233	-1.657	-2.118	-1.620	Hsc70-4	Hsp70 protein	HSP70
N0.HOG0041221	-1.559	-2.795	-1.041	-1.060	H15	Sequence-specific DNA binding transcription factor activity. It is involved in the biological process described with regulation of transcription, DNA-templated	T-box
N0.HOG0001157	-1.167	-0.858	-1.962	-2.162	COPE	The coatomer is a cytosolic protein complex that binds to dilysine motifs and reversibly associates with Golgi non- clathrin-coated vesicles, which further mediate biosynthetic protein transport from the ER, via the Golgi up to the trans Golgi network. The coatomer complex is required for budding from Golgi membranes, and is essential for the retrograde Golgi-to-ER transport of dilysine-tagged proteins	Coatomer_E
N0.HOG0002235	-1.640	-1.116	-1.056	-2.131	FARSA	phenylalanyl-tRNA synthetase, alpha subunit	HTH_11, tRNA-synt_2d
N0.HOG0008486	-2.230	-1.352	-1.427	-0.836	RBMX2	nucleic acid binding	RRM_1
N0.HOG0009667	-1.885	-1.955	-1.157	-0.836	TIMMDC1	Tim17/Tim22/Tim23/Pmp24 family	Tim17
N0.HOG0001163	-1.697	-0.399	-1.944	-1.596	KDM3B	A domain family that is part of the cupin metalloenzyme superfamily.	JmjC
N0.HOG0007812	-0.080	-3.204	-1.032	-1.319	SCCPDH	oxidoreductase activity	Sacchrp_dh_NADP
N0.HOG0001985	-1.469	-0.991	-2.003	-1.055	TMEM199	Endoplasmic reticulum-based factor for assembly of V-ATPase	Vma12
N0.HOG0004298	-0.148	-1.107	-1.847	-2.093	-	Adenosine-deaminase (editase) domain	A_deamin
N0.HOG0008085	-0.433	-1.359	-1.242	-1.772	EXOSC2	Exosome complex component RRP4	ECR1_N, KH_6
N0.HOG0008379	-0.460	-2.003	-1.250	-1.084	-	aldo-keto reductase family 1, member	Aldo_ket_red
N0.HOG0003032	-1.013	-1.486	-0.867	-1.216	GEMIN2	It is involved in the biological process described with spliceosomal snRNP assembly	SIP1