



## TRANSCRIPTOMIC ANALYSIS ON FRESHWATER MUSSELS FOR IDENTIFICATION OF POTENTIAL BIOMARKERS TO MONITOR WATER ECOSYSTEMS

Master's thesis project in  
BioScience (30 credits)

### **Author**

Md Mehedi Hasan Joyon  
[a19mdmjo@student.his.se](mailto:a19mdmjo@student.his.se)

### **Name of supervisor**

Mikael Ejdebäck (Ph.D.)  
System Biology Research Center, University of Skövde  
[mikael.ejdeback@his.se](mailto:mikael.ejdeback@his.se)

### **Name of co supervisor**

John Baxter  
Lecturer in Bioscience  
[John.baxter@his.se](mailto:John.baxter@his.se)

### **Name of Examiner**

Anna-Karin Pernestig  
Senior Lecturer in Bioscience  
[anna-karin.pernestig@his.se](mailto:anna-karin.pernestig@his.se)

## Abstract

Stress-specific expression of cellular proteins in responses to exogenous exposure and resulting physiological alteration provides important insight into the field of ecological research. Due to its habitat, feeding, lifestyle and physiologic properties, mussel has become an important indicative measure of aquatic environment pollution in order to assess effect of these pollution in aquatic life. In order to minimize the threats imposed on the aquatic ecosystem and advancement of sustainable lifestyle for human, recent ecological studies are more concern about monitoring different bioindicative properties. In this study, two widely distributed freshwater bivalve mussel species *Anodonta anatina* and *Unio tumidus* was used to conduct comparative study on the transcriptome of these species in order to identify and quantify the expressed transcripts on both species and investigate their biomarker properties in mussels for monitoring heavy metal or toxic exposure. mRNA was isolated and converted to cDNA through reverse transcription PCR. Quality and quantity assessments of purity, fragment size and concentration was performed. Each cDNA sample was barcoded and amplified for cDNA library preparation and nanopore sequencing. Basic bioinformatics tools were used to identify the transcripts for transcriptomic analysis. The findings shows some common mitochondrial and ribosomal transcripts along with a wide range of conserved and abundant transcript variants in mussels with important biomarker properties. Some of the transcripts exhibits expression in multiple samples suggesting characteristic bioindicator properties. Also in this study, a pipeline for transcriptomic analysis was generated and critical steps in the procedure were identified and discussed.

## Popular scientific summery

Water pollution is a common issue all across the world. A recent study showed that millions of peoples does not have adequate water supply, proper sanitizations or fresh water usage facility and still the condition is getting worse. Consumption of such polluted water brings enormous health hazards to the consumers. Water in river, ocean, lake and ponds are normally polluted through exposure of sudden hazardous substance including heavy metal, dangerous chemical, plastic and polybags, industrial sewage, household wastes, marine dumping, accidental oil leakages and many more. These toxic substances easily dissolves and contaminate water. Water contamination also impose a great threat for the aquatic life and ecosystem. Heavy pollutant contamination causes sterility, number of diseases, fatality and sometime extinction of marine species. Various experiments has been performed in order to minimize the effect of water pollution such as chemical approaches and bioremediation along with promoting public awareness. Two basic approaches are available to determine the presence of pollutant in an aquatic system. The first of which is physio-chemical approach such as chemical oxygen demand (COD), usage of chemical kits, water nutrient and pH measurement. The second approach Includes using organism to test presence the pollutants in an aquatic system commonly called as bioindicators.

Mussels are unique organism with high taxonomical richness and high diversity. Out of total 840 species diversity, there are seven different species of mussels that has been discovered in many freshwaters such as rivers, lakes, ponds in Sweden. Interestingly, mussels has an ability to filter approximately 50 liters of water during a day. Due to this filtering ability, they are exposed to a number of pollutant which accumulates in various tissues inside mussels. A number of studies has already been conducted where mussel samples has been used as bioindicator of pollutant exposure such as heavy metals as copper, zinc, cadmium and also other pollutant such as microplastics, chemical products and many more. These studies shows significant expression of physiologically important genes of mussel species in response to stress and pollutant exposure. A recent study at University of Skövde used a species of mussel as sample for monitoring the effect of copper exposure and biomarker investigation. Analysis of such gene expression in response to pollutant exposure reveals potential biomarker properties. The ultimate aim of these studies were to detect the water quality and effect of such pollutant on the transcriptome profile of the mussels. Developing multi-biomarker panel may also aid in the detection of a variety of contaminant at a time.

The aim of this study was to compare the transcriptome of two mussels sample *Anodonta anatina* and *Unio tumidus* for identification of potential biomarkers to monitor water ecosystems. Transcriptomic studies requires sequencing of coding RNA moieties of the species and gene expression analysis. Identification of various transcripts and their annotation analysis can provide important insights about the conserved characteristics of mussels, response to external stresses or exposures, functions and properties. Comparative transcriptomic analysis of two mussel species shows common transcript variants found in both mussel species and their physiological role along with transcripts that are only found in one species. These commonly expressed transcripts can further be investigated to find their role as indicators of freshwater pollution. Such information from the comparative transcriptomic analysis of two freshwater mussel species may aid in future ecological studies to detect marker genes for the assessment of water pollution. Outcomes from the experimental pipeline and approaches followed in this study may help better understanding of the necessary steps required to improvise quality and quantity of expected output.

## Abbreviations

A-1	<i>Anodonta anatina</i> sample 1
A-2	<i>Anodonta anatina</i> sample 2
A-3	<i>Anodonta anatina</i> sample 3
A-4	<i>Anodonta anatina</i> sample 4
A-5	<i>Anodonta anatina</i> sample 5
A-6	<i>Anodonta anatina</i> sample 6
A-7	<i>Anodonta anatina</i> sample 7
BLASTn	Basic Local Alignment Search Tool nucleotide
bp	Base Pair
CaM	Calmodulin
cDNA	Complementary DNA
COX	Cytochrome C oxidase
EF $\alpha$ -1	Elongation factor alpha-1
GO	Gene ontology
HSP	Heat Shock Protein
kb	Kilo base
mRNA	Messenger RNA
NCBI	National Centre for Biotechnology Information
ONT	Oxford Nanopore Technology
PCR	Polymerase Chain Reaction
Poly(A)	Polyadenylated
QC	Quality Control
qPCR	Quantitative Polymerase Chain Reaction
RNase	Ribonuclease
RQN	RNA quality number
rRNA	Ribosomal RNA
RT	Reverse Transcriptase
tRNA	Transfer RNA
U-1	<i>Unio tumidus</i> sample 1
U-2	<i>Unio tumidus</i> sample 2

U-3	<i>Unio tumidus</i> sample 3
U-4	<i>Unio tumidus</i> sample 4
U-5	<i>Unio tumidus</i> sample 5

# Table of Contents

1. Introduction .....	1
1.1 Bioindicators and biomarkers .....	2
1.2 mRNA extraction technology .....	3
1.3 Nanopore sequencing technology .....	3
1.4 Aim .....	4
2. Materials and Methods .....	5
2.1 Sample preparation .....	5
2.2 Sequencing Library Preparation .....	6
2.3 Processing of sequencing output data and bioinformatics analysis .....	7
3. Results .....	7
3.1 Sample preparation .....	7
3.2 Quality and quantity of extracted mRNA .....	8
3.3 Quality and Quantity assessment of sequence library .....	10
3.4 Quality control of mRNA sequence data .....	11
3.5 Data analysis and interpretations .....	11
4.1 Discussion .....	12
4.1 Sample preparation .....	12
4.2 Sequence data output .....	15
4.3 Interpretation and analysis .....	16
5. Ethical aspects and impacts on society .....	18
6. Conclusion .....	18
7. Future perspectives .....	18
8. Acknowledgments .....	19
9. References .....	20
10. Appendix .....	25

## 1. Introduction

Mussels are bivalve aquatic organism belonging to the phylum mollusk and mostly located in freshwater or saltwater all across the globe. A number of 840 diverse species has been known to be found and increased diversity has been recorded on the southern countries (Graf & Cummings, 2007). Their distribution and diversity in Sweden is relatively abundant and widespread all across the nation. Unionidae is the largest family type among all and also found to be rich in species diversity (Graf et al., 2007). Among 674 species diversity of Unionidae family worldwide, *Anodonta anatina* commonly known as duck mussels is one of the species that is found in abundance in many freshwater in Sweden. Similarly, *Unio tumidus* commonly known as swollen river mussel, also belong to the family Unionidae which is also found in many freshwater rivers and lakes in Sweden (Annie, Ann & Mats, 2013). Both of these species normally feed on freshwater plankton and filters approximately 50 liters water per day. During water filtration, they are exposed to a certain number of heavy metals (Chandurvelan, Marsden, Glover & Gaw, 2015) pollutants such as lipophilic pollutants and polychlorinated biphenyl pollutants (Tanabe, Tatsukawa & Phillips 1987), microplastic (Li et al., 2019) and chemical products e.g. tributyltin (Salazar & Salazar, 1996), diclofenac (Cunha, Pena & Fernandes, 2017). Exposure to exogenous pollutants or metal compounds causes accumulation of these compound in tissues such as hepatopancreas. Such exposure causes physiological stresses which in turn triggers expression of transcripts in mussels (Woo, Jeon, Kim & Yum, 2011) (Hüning et al., 2013). These physiological response and differential expression of specific stress related transcripts makes both of these species a potential candidate for investigation of bioindicator properties (Asif, Malik & Chaudhry, 2018) and ecological research (Hoellein, Zarnoch, Bruesewitz & DeMartini, 2017).

Mussels have unique physiology with certain organs conducting major physiological responses as shown in Figure 1 and a relatively large number of mRNA expression has been reported from these certain tissues, more specifically hepatopancreas and gills, due to environmental stress or exposure (Franco et al., 2006 and Moore, Viarengo, Donkin & Hawkins, 2007). This provides certain insights to understand tissue specific responses in monitoring the effects of certain stimuli and tissue specific feedback expressions by the organism. However, the study shows a relatively higher number of mRNA expressions reported in digestive gland i.e. hepatopancreas in respect to the other organs upon conduction of metabolic responses (Liu et al., 2014). For instance, monitoring the hepatopancreatic tissue specific expression can also provide important information about some characteristic bioindicator properties.

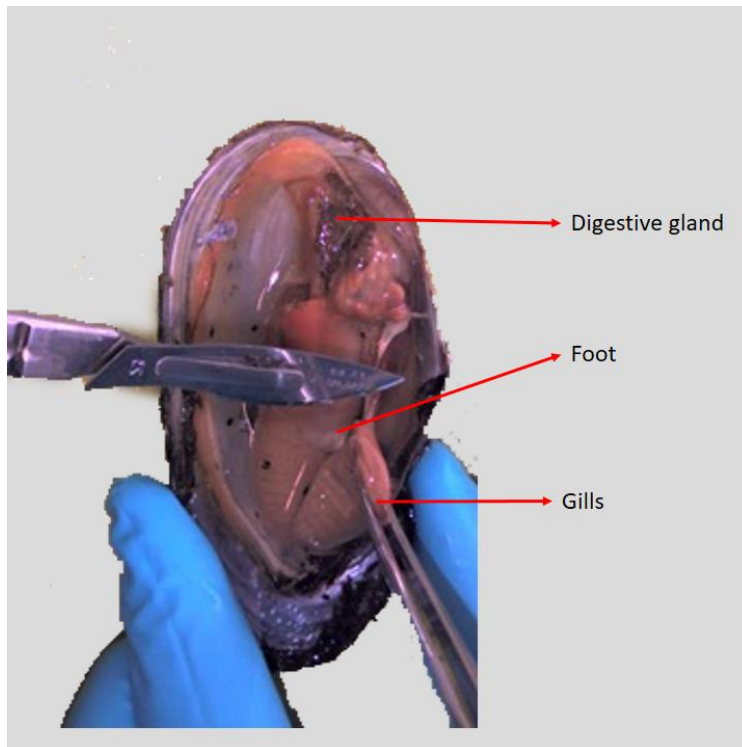


Figure 1. Anatomy of *Anodonta anatina* mussel. The blackish region in the upper right corner is the digestive gland where hepatopancreas is located. The pink muscular region attached to the right and left valves are the gills. And at the end of each gills, white small region is the foots of the mussels. The picture was taken with a camera device during dissection of one of the mussel samples of *Anodonta anatina*.

The figure 1 shows certain tissues that were found rich in mRNA concentration due to external stress (Franco et al., 2006) (Moore et al., 2007) and their tissue distribution in mussel. This indicates important anatomical target areas to extract mRNA for transcriptomic studies in mussel.

### 1.1. Bioindicators and biomarkers

In ecological studies, a bioindicator can be refer to as indicators of ecological health. Living organisms that has been used to monitor the changes in ecological environment can also be defined as bioindicators (Parmar, Rawtani & Agrawal, 2016). Mussels are generally assessed for bioindicator properties for four important reasons, the first being its sensitivity to pollution and its bioaccumulation properties to assess water quality. Secondly, the larvae of mussels lives on the aquatic fishes as parasite indicating abundance of fishes in following freshwater environment. Thirdly, the diversity and taxonomic richness of mussels influences with the geo-ecological features of the freshwater and affect taxonomic composition of other organisms living on that ecological environment. Finally, it helps in habituation for other aquatic organism imposing impact on planktonic ecosystem (Aldridge, Fayle & Jackson, 2007). Various ecological changes produce different stresses in aquatic organism which eventually trigger several signaling pathways causing elevated level of gene expressions enabling a chance to detect the quantitative measure of the changes. Investigation of these expressed gene type and comparative analysis of these certain elevated level of expressed genes with the closely related species makes it possible to predict genes that either directly or passively related to the ecological changes or their lifecycle. Such genes can also be predicted as potential biomarkers by monitoring their properties, ontology and characteristics of the gene products (Hoffmann & Willi, 2008).

Studies on mussels as bioindicators has been conducted over the time for their bioindicator properties (Lepoutre et al., 2020), monitoring bioaccumulation of different pollutants (Sohail, Khan, Chaudhry & Qureshi, 2016), bioremediation and effect of pollutant exposure (Ugge, Jonsson, Olsson, Sjöback, & Berglund, 2020). Also some studies has been conducted on family level composition where *Anodonta anatina* and *Unio tumidus* species were involved (Bolotov et al., 2020). A number of studies on these species has been performed for bioindicator property investigation within these recent period of time in the field of ecological research and diversity (Zieritz et al, 2018). Different approaches and sequencing methods has also been applied in this recent time period to make sufficient data available for advanced research in this fields (Prié et al., 2020). Many gene expression analysis was also performed with or without external stress provided on the blue mussel i.e. *Mytilus edulis* and *Mytilus galloprovincialis* samples (Moore et al., 2007). Modern sequencing method mostly uses either Illumina or Nanopore sequencing technology.

This thesis project was a part of the project “Waterassess Multi-biomarker panel for environmental impact assessment of wastewater effluents”, a cooperative research project between Lund University and University of Skövde funded by the Knowledge foundation. Major focus of this thesis project was to perform sequencing experiment of mRNA from hepatopancreas tissue of two mussel species *Anodonta anatina* and *Unio tumidus* for identification and quantification of transcripts expressed in natural condition for comparative transcriptome analysis. Advanced bioinformatics tools will be used in future projects to perform mRNA expression analysis of the transcripts that has been identified in this thesis project and will be compared to those with exogenous pollutant or metal ion exposure for biomarker property analysis.

### **1.2. mRNA extraction technology**

The cell contains different kinds of RNA, all combined are called total RNA content of the cell. Extracting mRNA requires capturing only those RNA species that can be coded to functional protein moiety. However, the extraction of only the coding mRNA is very crucial as majority of RNA species i.e. rRNA and tRNA in total RNA content of the cell are non-coding. The coding mRNA constituent of only 1-5% whereas the rest of the 95% is the non-coding RNA out of the total RNA content of the cell (Mattick & Gagen, 2001). It makes extraction of mRNA more challenging. During post transcriptional modification, primary transcripts undergoes 3' polyadenylation where a number of adenine bases is added by the end of the primary transcript. That makes it a long poly(A) chain. The poly(A) sequence of mRNA can be targeted to extract coding mRNA out of total RNA using oligo- DT magnetic beads. In this technique, a chain of oligonucleotide thymidine attached to a magnetic bead and forms complementarily pairs with the poly(A) tail of the mRNA. Further, a magnet is used to pull the mRNA bound with magnetic beads out of solution as pellet and thus washed to be extracted as purified poly(A) mRNA. This method is only applicable to eukaryotic cells. Prokaryotic RNA does usually not contain a polyA sequence. In such cases rRNA removal techniques are used to degrade the rRNA and tRNA out of total RNA contents and finally remaining mRNA is extracted using commercially available kits (Peano et al., 2013).

### **1.3. Nanopore sequencing technology**

Nanopore sequencing technology (ONT) uses small and pocket size device i.e. a MinION device to sequence DNA and RNA molecules. It provides longer reads compared to other next generation

sequencing technologies about 10 kb and sometimes even more. It is capable of quick and accurate sequencing of these long fragment molecules (Lu, Giordano & Ning, 2016). The feature of quick sequencing allows monitoring and diagnosis of pathogens and faster identification of new species (Hoenen et al., 2016). Producing faster sequencing may also aid in dealing with disease outbreak, producing genome sequences in real time, investigate properties and also identification of target molecules to minimize the outbreak (Jain, Olsen, Paten & Akeson, 2016). The nanopores are basically nano-scale holes in an electrically resistant membrane which is embedded into a small size flow cell that is connected the MinION sequencing device. These pores contain transporter molecules that allows biological molecules to pass through the pores and these pores are well stabilized by array of micro scaffolds. Ionic current passes through these pores which create electrical voltage across the membrane. The sample, which can be either DNA or RNA are first loaded into the flow cell. Flow cell prepares DNA/RNA- enzyme complex with the transporter protein and thus directed inwards the small pores. In the nanopore, fragments passes one nucleotide base at a time across the pore causing temporary changes in the electrical voltage as described in figure 2. This change in electrical voltage is specific for the unique nucleotide base and is recorded through the device (Feng, Zhang, Ying, Wang & Du, 2015).

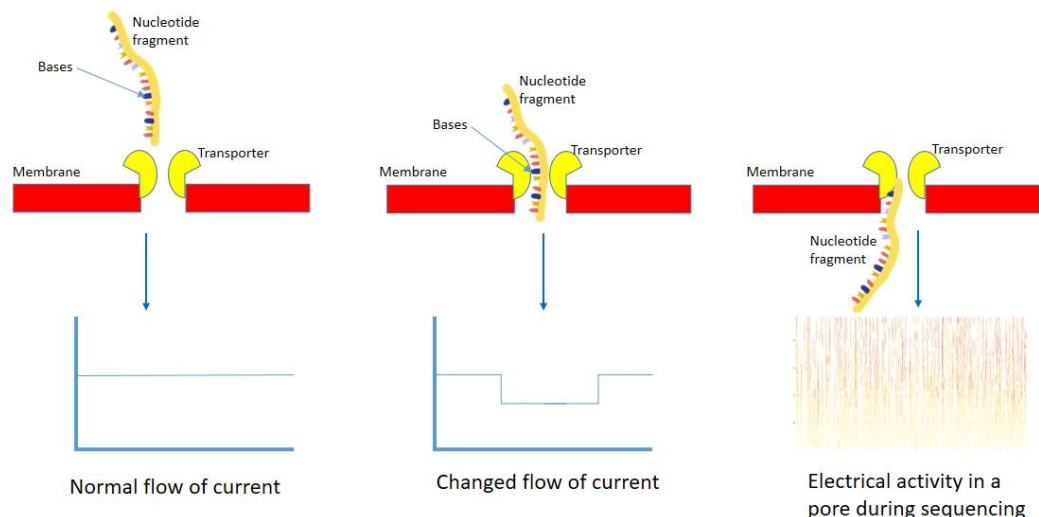


Figure 2. Figure showing the process of nanopore sequencing technology. The membrane contains transporter protein, has continuous flow of electric voltage over the pore (A). When nucleotide fragment comes to the opening of the pore, first nucleotide binds to the transporter protein and it allows the nucleotide pass through the pore causing a unique change in electric current which corresponds to that specific base (B). The cycle is repeated for the rest of nucleotide bases in the fragment. After passing the whole fragment through the channel, it produces a sequential wave of electrical signals indicating the sequence of the fragment (C).

#### 1.4. Aim

The aim of this thesis was to compare expression of gene in two freshwater mussel species i.e. *Anodonta anatina* and *Unio tumidus* in order to identify potential biomarkers for monitoring of water ecosystems.

Major objectives of this project was, 1) Generation of sequence data from two mussel species mentioned using available mRNA isolation method and nanopore sequencing technology on MinION device. 2) Performing alignment of the sample originated sequenced data originated with the sequences provided in publicly available database NCBI for transcript identification. 3) Performing GO analysis of each gene product of identified transcripts to reveal their association

with various conserved biological processes and response to external stresses which may indicate characteristic biomarker properties.

## 2. Materials and Methods

### 2.1. Sample preparation

The mussel samples were collected from the freshwater lake Näsbadet at Skärvalången, located near the Skövde city that belongs to Västra Götaland county of Sweden at 6<sup>th</sup> March 2020. During the collection of mussel samples the pH of the lake was 6.85 and the lake water temperature was 2 °C. Twenty mussel samples were collected in total where ten of them belongs to *Anodonta anatina* species and rest of the ten samples were from *Unio tumidus*. Seven samples from each species were dissected within three hours the same day. Hepatopancreas tissues were collected from each samples followed by immediate storage of the tissues in a solution of RNALater (ThermoFisher) separately in different tube at 4°C for further use.

Extraction of mRNA was performed using DynaBeads® mRNA DIRECT™ Kit (Thermo Fisher Scientific). Two methods for tissue homogenization were compared i.e. TissueLyser LT (Qiagen) and homogenization with syringe and needle. Initially, 20 mg of the hepatopancreas tissue from *Unio tumidus* sample was taken in two separate tubes with 1250 µl of lysis buffer from DynaBeads® mRNA DIRECT™ Kit (Thermo Fisher Scientific). Then the solid tissue in lysis buffer was either grinded in liquid nitrogen using a 21 gauge needle and syringe or homogenized with a TissueLyser LT (Qiagen) to make sample lysate. Different spinning frequency (30 Hertz and 50 Hertz) and run time (40 seconds and 60 seconds) were compared for the optimization of the extraction using the TissueLyser LT. Finally, a 50 hertz frequency and 40 seconds runtime was used for the final extractions. However, amount of tissue was increased to 50 mg for the five samples of *Unio tumidus* and seven samples of *Anodonta anatina*.

For extraction of mRNA, DynaBeads® Oligo(DT) pellet collected from 250 µl DynaBeads® Oligo(DT) (Thermo Fisher Scientific) was suspended in each sample lysate. The mixtures were incubated in a rotary shaker for 10 minutes at room temperature to allow hybridization of poly(A) tail of the mRNA with the Oligo(DT) on the beads. After incubation, the mixtures were placed in a vial magnet for 5 minutes and the pellet of mRNA/beads were collected followed by discarding the supernatant. The sample was washed several times as per the protocol provided by the DynaBeads® mRNA DIRECT™ Kit (Thermo Fisher Scientific) and 100 µl of 1x Reverse Transcriptase buffer from High-Capacity cDNA Reverse Transcription Kit (Thermo Fisher Scientific) was used as last wash. mRNA was eluted in 25 µl of 10mM Tris\_HCl pH 7.5 at 80 °C in a RNase-free microcentrifuge tube.

Quantity of the extracted mRNA was assessed on a Qubit 4 Fluorometer (Invitrogen) using Qubit™ RNA HS Assay Kit (Invitrogen) according to the protocol provided with the kit. Quality of extracted mRNA was assessed using Nanodrop 2000 spectrophotometer (Thermo Fisher Scientific). The absorbance of 260/280 nm was recorded to check the purity of the mRNA solution. Fragment Analyzer™ Automated CE System (Agilent) was used to assess the fragment length and quality using DNF-472 High Sensitivity RNA Analysis Kit, 15 nt (Agilent) according to the protocol provided with the kit.

## 2.2. Sequencing Library Preparation and sequencing

For construction of cDNA from mRNA, RT-PCR was performed according to the Protocol (ONT) provided with PCR-cDNA Barcoding kit (SQK-PCB 109). Only alternative was instead of using only Maxima H Reverse Transcriptase enzyme (200 U/ $\mu$ l concentration) from Maxima H Minus Reverse Transcriptase kit (Thermofisher) for all the samples, this enzyme was only used for each samples of *Anodonta anatina* species whereas Multiscribe™ Reverse Transcriptase enzyme (50 U/ $\mu$ l concentration) from High-Capacity cDNA Reverse Transcription Kit (Thermofisher) was used for each samples of *Unio Tumidus* species. However, same RT buffer recommended by PCR-cDNA Barcoding kit (SQK-PCB 109) protocol was used for both enzymes. This final mixture was incubated in RT-PCR program cycle where Reverse Transcription and strand switching step was performed in 42 °C for 90 minutes and Heat inactivation was performed in 85 °C for 5 minutes in CFX384™ Real-Time PCR Detection System (Bio-Rad, USA).

cDNA sequencing library was prepared according to the instruction provided with PCR-cDNA Barcoding kit (SQK-PCB 109). A PCR with a set of total 12 different barcode primers were used for the 12 individual samples, five samples from *Unio tumidus* and seven sample from *Anodonta anatina* species in order to identify samples after pooling. A reaction mixture for each sample containing 5  $\mu$ l of Reverse-transcribed RNA sample, 1.5  $\mu$ l of 400 nM Barcode primers (BP01-BP12) from PCR-cDNA Barcoding kit (SQK-PCB 109) each for one individual sample, 0.5  $\mu$ l of 2 U/ $\mu$ l PCRBIO High Fidelity polymerase from PCRBIO HiFi Polymerase kit (PCRbio systems), 10  $\mu$ l of 1x PCRBIO High Fidelity buffer containing MgCl<sub>2</sub>, dNTPs, enhancers and stabilizers from PCRBIO HiFi Polymerase kit (PCRbio systems) and 33  $\mu$ l Nuclease-free water was prepared and subjected to PCR (Table 1).

Table 1. Table describing programming cycle for amplification PCR.

Step	Temperature	Number of cycle	Time
<b>Initial denaturation</b>	95 °C		30 seconds
<b>Denaturation</b>	95 °C	18 cycles	15 seconds
<b>Annealing</b>	62 °C		15 seconds
<b>Extension</b>	65 °C		50 seconds
<b>Final extension</b>	65 °C		6 minutes
<b>Hold</b>	4 °C		

For purification of desired barcoded cDNA from the PCR reaction mixture, AMPure XP beads (Beckman Coulter Life Sciences) was added to each reaction mixture according to protocol provided for cDNA Barcoding kit (SQK-PCB 109). To each reaction tube 1  $\mu$ l 20 U/ $\mu$ l Exonuclease 1 (New England Biolabs) was added and incubated on HulaMixer and finally eluted in 12  $\mu$ l elution buffer (EB) from cDNA Barcoding kit (SQK-PCB 109) on a magnet as per instructions. After elution, quantity measurement of each eluted sample was carried out on Qubit 4 Fluorometer (Invitrogen) using Qubit™ dsDNA HS Assay Kit (Thermofisher). Subsequently, 1  $\mu$ l from the each sample was pooled together in one reaction tube and the final volume of the pooled cDNA library was 12  $\mu$ l. For addition of adapter, 1  $\mu$ l of Rapid Adapter (RAP) from cDNA Barcoding kit (SQK-PCB 109) was added.

Sequencing of the cDNA on MinION device was performed in R9.4.1 Flow cell (FLO-MIN 106D). Before loading, the flow cell was washed with Flow Cell Wash Kit (EXP-WSH 003A) according to the protocol (ONT). For priming and loading the MinION Spot on flow cell, instructions provided

with PCR-cDNA Barcoding kit (SQK-OCA-59) was followed. The standard MinKNOW protocol script was used for the sequencing. The run time of the MinION device was set to 24 hours in 190 voltage without base-calling and the quality score cut off was set to 7.

### **2.3. Processing of sequencing output data and bioinformatics analysis**

Fastq files were generated from sequencing barcodes using automated base call algorithm Guppy software using version 19.12.5. Quality of the sequencing was assessed using PycoQC software (Version 2.2).

Non-sequence data was removed from the fastq files using “Regex” function of Python software but no trimming of barcoding sequence or adapter sequence was performed, since the sequence data was only used for basic local alignment search tool nucleotide (BLASTn) available at the National Centre for Biotechnology Information (NCBI). Also no assembly of generated sequence was conducted. A total of eight fastq-files were chosen randomly from each barcodes where each barcodes represent a sample from a specific mussel species. For randomization of picking fastq files, different time period of output file generation and descending file size was taken into consideration and 97 individual BLASTn search was performed. During BLASTn search, maximum target sequence parameter was set to 10,000 for each search as the matrix score was too low below 10000 hits and also it takes a huge time for loading. Also the expected value in BLASTn was 1 for each search. The rest of the BLASTn parameters were kept default. The BLASTn results were filtered by “Mollusk (taxid:6447)” which displayed the result containing the species belongs to mollusk phylum. Not all the search produced alignment results and thus were excluded. In case of Barcode sample 1 which belongs to *Unio tumidus* species, no result was originated as each file contained single and short fragment data. In this case, all the fastq-files for this barcode samples were processed to create sequence only data and then all the data were put together in one file, which finally used to perform one extra BLASTn search. For analysis of gene ontology QuickGO (version 2020-08-10) and InterPro (version 81.0) servers were used. In order to study information about the transcript, secondary database, transcripts from closely related species, protein activity and proteins active site UniProt was used. In case where no information were found within the species specification, a different nearby species were used for Gene Ontology (GO) analysis.

## **3. Results**

### **3.1. Optimization of mRNA extraction conditions**

Selection of an appropriate extraction method and optimization of extraction conditions was required for achieving relatively pure mRNA with desired concentration. An optimization was first performed between two commonly used tissue homogenization methods i.e. TissueLyser LT (Qiagen) and homogenization with syringe and needle using liquid nitrogen in order to select the suitable method that can used for the mRNA extraction with high yield. After tissue grinding, mRNA extraction was performed twice followed by quality assessment by Nanodrop 2000 spectrophotometer (ThermoFisher) at the absorbance of 260/280 ratio and quantity assessment by Qubit 4 fluorometer (Invitrogen). Qubit assay showed grinding with syringe method yielded 1.7 ng/ $\mu$ l mRNA while 2.6 ng/ $\mu$ l mRNA was extracted from grinding with TissueLyser LT in average. Purity measured by Nanodrop (260/280 ratio) for mRNA solution extracted with needle-syringe method was 2.99 and for TissueLyser LT it was 1.86 in average. As the TissueLyser LT

gave the better quantity and quality it was selected for the remaining extractions. The results from the optimization are retrieved by using the qubit and nanodrop as shown in Appendix A.

After selecting TissueLyser LT tissue homogenization method, it was required to optimize the spinning frequency of TissueLyser and run time.

### 3.2. Quality and quantity of extracted mRNA

An amount of 50 mg of hepatopancreas tissue for each of the 12 different samples were used for mRNA extraction according to the optimized procedure mentioned above and also according to the tissue volume suggested by the protocol provided with DynaBeads® mRNA DIRECT™ Kit (Thermo Fisher Scientific) and further assessed for quality and quantity measurement. The overall assessment was designed into three part measurement (Table 2), a) Quantity estimation with Qubit 4 Fluorometer to measure the concentration of extracted mRNA in the sample b) Quality of purity of the extracted mRNA with Nanodrop 2000 spectrophotometer at 260/280 absorbance ratio, c) Quality of fragments present in extracted mRNA with Fragment Analyzer™ Automated CE System (Agilent). The sample extraction and quality and quantity measurement was only performed one during the overall experiment.

Table 2. Values for RNA concentration, absorbance ratio and RNA quantity number (RQN) is given for each sample in each row accordingly.

<b>Sample no</b>	<b>Concentration of RNA (ng/μl)</b>	<b>of (260/280) absorbance ratio</b>	<b>RQN</b>
<b>U-1</b>	4.60	2.80	1.1
<b>U-2</b>	4.00	1.70	3.8
<b>U-3</b>	3.20	1.84	1.1
<b>U-4</b>	9.40	2.09	9.4
<b>U-5</b>	8.40	1.91	8.9
<b>A-1</b>	35.60	2.10	4.2
<b>A-2</b>	30.20	2.20	1.9
<b>A-3</b>	4.20	2.25	1.4
<b>A-4</b>	2.60	1.42	3.0
<b>A-5</b>	51.00	2.20	7.3
<b>A-6</b>	4.00	1.92	2.5
<b>A-7</b>	16.1	1.89	7.5

The result of the quantity assessment shows values between 2.6 ng/μl to 51 in ng/μl. In purity assessment, six of the sample resulted a 280/ 260 absorbance ratio higher than 2.0. The RQN score mentioned in table 2 ranges from 1.1 to 9.4 as an indicative measure for RNA fragmentation. The graphical demonstration of gel image (figure 3) and fragmentation curve (figure 4) reported from fragmentation quality assessment is used in this study in order to provide a clear picture that what these values signifies. Also an elaborated explanation for these assessments is mentioned in the discussion part of this report.

Figure 3 shows a number of strong bands for sample A-2, A-3 and A-6 visible in lower band (<500 nucleotide) region. Also a similarity in relatively low RQN score is observable for all three samples mentioned. A common type of medium light bands are noticeable for sample U-2, A-1 and A-4 with a medium RQN score. The sample U-4, U-5, A-5 and A-7 exhibit light bands that is slightly visible in the lower region with a strong band in the 4000 nucleotide region and these sample carries

relatively high RQN score. But no bands observed for the sample U-1 and U-3 both carrying lowest RQN score. For this reason, 3 different fragment curve representing different level of RQN score is given on figure 4 where A) has shown a lots of short RNA fragments ranging from 100 to 1800 nucleotides and also carries lower RQN score. B) A number of short fragment ranging from 100 to 1800 with a larger density in 4000 nucleotide region with a medium RQN score. C) Very few number of short fragment around 100 but a high abundance of peaks in higher nucleotide region ranging from 3800 to 6500 nucleotide region with higher RQN score. In order to analyze the integrity of extracted RNA the library or cDNA was run on a fragment analyzer. High RQN indicating low fragmentation and low RQN indicating high fragmentation.

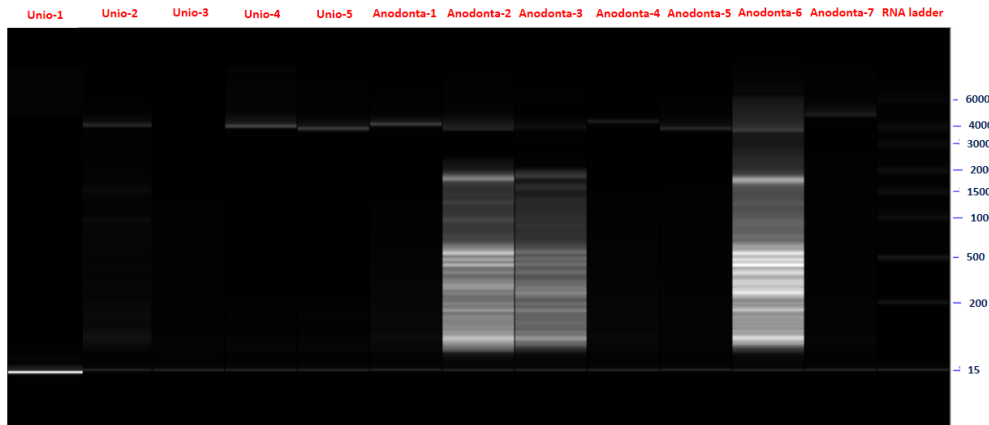
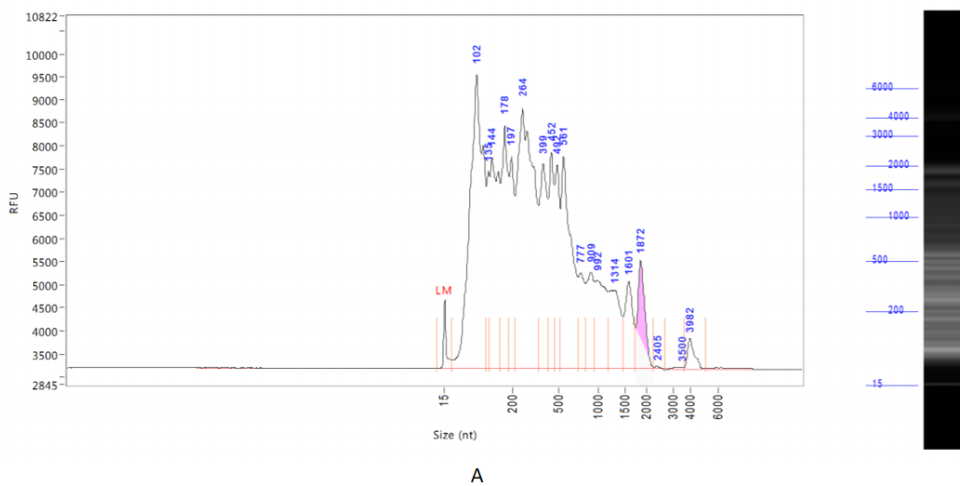


Figure 3. The gel image generated by ProSize Data Analysis Software after running all the samples in fragment analyzer. The lane in red in the upper row represents sample wells ranges from well 1 to well 12 while well 13 indicating the RNA ladder. The column showing blue lanes represent band in base pairs originated from RNA ladder.



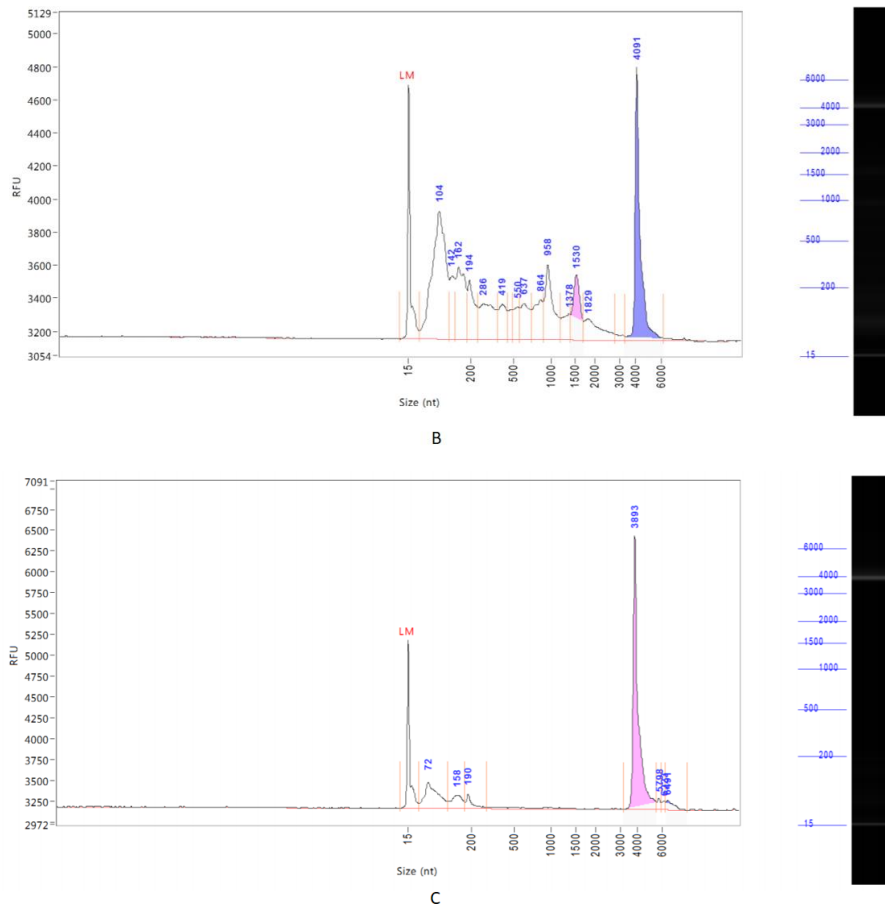


Figure 4. Fragmentation curve of three different sample containing three different RNA quality score is shown in the figure where A) is the sample anodonta-2 with RQN value 1.9 B) is Unio-2 with RQN score 3.8 and C) is Unio-5 with RQN score 8.9 respectively. In the following figures, size in length of nucleotide fragment is given on x-axis and relative fluorescence unit (RFU) is given on y-axis. LM corresponds to lower marker that is 15 nucleotides long. The numerics in blue are the length of fragments detected during fragment analysis.

### 3.3. Quality and Quantity assessment of sequence library

It is important to quantify the concentration of each eluted cDNA barcoded sample to record their final concentration for further understanding of its effect on sequence output quality. Also qualitative purity of each sample ensures that samples does not contain any contamination. Quality assesment with Nanodrop at absorbance ratio 260/280 ratio showed that all the samples in the mixture had shown a value with in the range of 1.8 to 2.0. In case of quantity assessment with Qubit difference in sample concentration was observed (table 3).

Table 3: Values of concentration written as “Conc.” for each sample is given in ng/μl measurement scale with 280/260 absorbance ratio as purity index. Here sample title U and A represents the sample from *Unio tumidus* and *Anodonta anatina* species respectively.

Sample	U-1	U-2	U-3	U-4	U-5	A-1	A-2	A-3	A-4	A-5	A-6	A-7
<b>Conc.</b>	2.00	1.68	2.00	1.30	1.71	1.02	0.46	0.22	0.50	0.82	1.53	0.80
<b>Purity</b>	2.00	1.98	1.87	1.92	1.95	1.88	1.80	2.00	1.87	1.98	2.00	1.85

Concentration of cDNA higher than 1 ng/μl was observed for eight samples where a very low concentration of cDNA was observed for five other samples. Samples containing low cDNA concentration were mostly from *Anodonta anatina* species.

### 3.4. Quality control of mRNA sequence data

Sequencing output and performance data was assessed qualitatively with quality control assessment software PycoQC (Version 2.2). It produces an output that provides information about the number of reads generated during a sequencing experiment, quality of sequencing, number of pass reads, average length of fragment present, channel activity during sequencing reaction and number of reads per barcodes. Table 4 shows a short summary of overall sequencing run with given number of reads and bases detected, average read length and quality, number of active channel and run duration and finally the number of barcodes detected. Figure 5 shows the overall summary in channel activity.

Table 4. The sequence run summary generated by quality control assessment software PycoQC (Version 2.2) is shown in this table.

Reads	Bases	Read length (Median)	Read Quality (Median)	Active Channels	Run Duration (h)	Barcodes
3,771,813	733,237,130	173.00	8.3	486	24 h	13

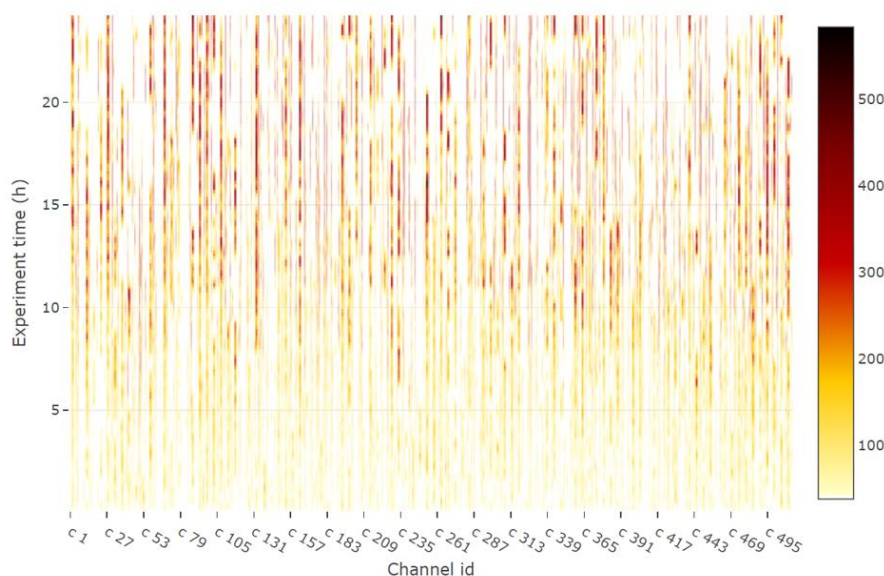


Figure 5. Channel activity during sequencing run is shown with an increasing time period. The overall summary of channel activity is plotted in a graphical representation where channel ID is given in x-axis and experiment time in hour is given on y-axis. The channel activity of each channel ID during the run period has shown in colors ranging from white to deep dark. Deep dark color indicating the high activity while white and yellow indicating poor channel activity.

### 3.5. Data analysis and interpretations

Sequencing with MinION device generates hundreds of fastq files for each barcode samples (Appendice B). After basecalling with MinKNOW software, total 12 barcode files containing fastq files of the corresponding sample and one unclassified samples containing fastq files whose sampling group was uncertain was generated. Due to lack of a complete reference genome for both species it was not possible to conduct alignment which might directly denote the corresponding gene or gene product. Instead, BLASTn (Zhang, Schwartz, Wagner, & Miller, 2000) was used to identify the transcripts present in the sample (Figure 6). BLASTn results showed high sequence identity above 90 percent for some of the transcripts from some of the closely related

species. The species that shares highest sequence identities for most of the BLASTn searches are mostly freshwater mussels from Unionidae family or other nearby species belonging to mollusk phylum. Species with high sequence identities to samples from both *Unio tumidus* and *Anodonta anatina* species are *Sinanodonta woodiana*, *Cristaria plicata* and *Hyriopsis cumingii* but the highest identity was observed for *Sinanodonta woodiana*. Although, few BLASTn search did not showed any alignment result but still a number of transcripts were identified from both of the species samples. These identified transcripts were mostly similar for same species samples but not so common for the other sampling species. Identified transcripts from *Unio tumidus* were shown in Appendix C and transcripts from *Anodonta anatina* were shown in appendix D.

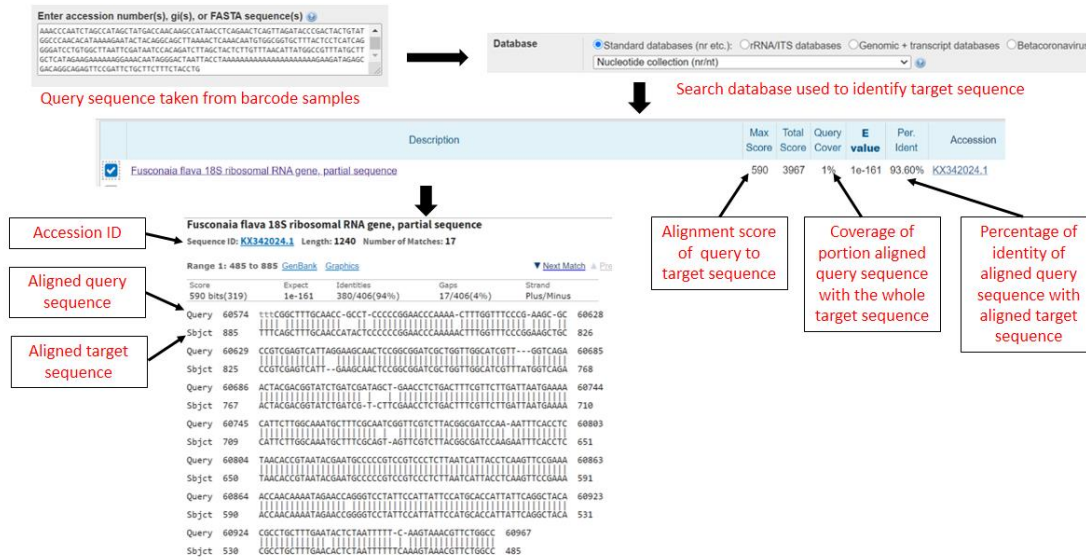


Figure 6: Figure showing the output generated for one of the query sequence file from barcode sample 12 (A-7). In the figure, query sequence was used for search into nucleotide database and then the generated result is shown with arrows.

## 4. Discussion

The main objective of this project was to perform mRNA sequencing experiment on two freshwater mussel species *Anodonta anatina* and *Unio tumidus* taken from their natural habitat which allows studying the common and uncommon transcripts between the species and their biomarker properties. The analyzed data originated from this project are discussed below emphasizing on discussion of the methods that has been used for this project with biomarker property investigation of the identified transcripts.

### 4.1. Sample preparation

A good RNA and mRNA quantity and quality was critical for downstream steps in the experimental procedure and the outcome of the experiment. Hepatopancreas tissue was used for RNA extraction in this project but this tissue itself contains a large quantity of endogenous ribonucleases (RNases) (Beintema,, Campagne & Gruber, 1973). This enzyme catalyzes the degradation of cellular RNA into smaller fragments and makes it difficult for extracting intact fragments of mRNA from the tissue. RNase are thermostable due to presence of its disulphide bridges and it is very difficult to inhibit its activity. Hepatopancreas also have cellular enzymes that causes self-destruction of cellular molecules through its autolytic properties and causes fragmentation and destruction of cellular RNA (Beintema et al, 1973). RNase 7, a ribonuclease that

belongs to the superfamily RNase A, were found on healthy human skin and participates in innate cutaneous defense and exhibits antimicrobial activity (Rademacher, Simanski & Harder, 2016) can also contaminate samples through human interaction and cause RNA degradation. In order to inhibit RNase activity, an area free from RNase's was required which was brought by cleaning the working area with RNase AWAY™ (Thermo Fisher Scientific), an RNase decontamination reagent. However, result in this study still shows certain fragmentation for some samples (Figure 3 and 4).

Before selecting the suitable extraction kit for mRNA extraction, it was important to optimize the tissue homogenization method and amount of tissue for extraction. In this case, TissueLyser LT showed relatively higher yield with higher purity of extracted mRNA than Syringe-needle method. Also during optimization of the spinning frequency and run time of TissueLyser LT, 40 Hz spinning frequency for 50 seconds showed relatively higher yield with higher purity (Appendix A). Finally 50 mg of tissue samples were used for extraction of mRNA. Below that amount of tissue sample the yield was too low for further experiments.

DynaBeads® mRNA DIRECT™ Kit (Thermo Fisher Scientific) is used for mRNA isolation because the reagents in the kit are maintained RNase free with RNase inhibitory agents. Also the protocol for this kit is very flexible to scale up or down to the sample size. This kit allows rapid isolation of pure polyadenylated mRNA and has a high expected yield (Thermo Fisher Scientific, 2012). The expected yield for 50 mg of liver tissue is 400 µg of total RNA and 1-5% of the total RNA is the expected mRNA concentration according to the instruction provided with the kit. A recent study also uses this kit for mRNA extraction (Engström, 2019) and thus was followed in this project.

Qubit™ RNA HS Assay Kit (Invitrogen) used in quantity assessment allows highly sensitive, fast and easy detection of RNA. It uses very small sample size of 1 µl and produce highly accurate result for measurement of RNA concentration present in the solution (Invitrogen, 2015). Qubit assessment on extracted mRNA showed that the concentration of mRNA ranged from 3.20 ng/ µl to 51.00 ng/ µl. The concentration required for the next step i.e. RT-PCR was 1 ng/µl (ONT, 2019). That means the concentration present in each sample was higher than the required concentration. However, the difference in concentration possibly arose from presence of less mRNA in the tissue sample or due to fragmentation that caused loosing 3' poly(A) tail (Vermeulen et al., 2011).

Nanodrop 2000 spectrophotometer was used for the assessment where absorbance of extracted mRNA was at 260/280 absorbance ratio indicating the purity of the mRNA present in sampling solution. The value of 260/280 absorbance ratio higher than 2.0 is considered pure for RNA solution. Six of the samples including U-1, U-4, A-1, A-2, A-3 and A-5 shows similar absorbance ratio and can be considered pure. But samples U-2, U-3, U-5, A-4, A-6 and A-7 shows absorbance ratio below 2.0. This might be possible due to presence of contaminants like protein, phenol, carbohydrates or other contaminants (Wilfinger, Mackey & Chomczynski, 1997). The kit used for the mRNA extraction does not contain any reagents with such contaminants but the incorporation of these contaminants possibly arose from external environment, tissue lysate or association of other tissue specific contaminants (Peirson & Butler, 2007)

DNF-472 High Sensitivity RNA Analysis Kit, 15 nt (Agilent) used for the fragment quality assessment uses a very small volume of sample size which is 2 µl and also have a high accuracy (Agilent, 2020). Fragment analyzer produces an electrophoresis graph from which the RQN number can be calculated. RQN rates the RNA degradation quality from scale 1 to 10 where 1 considered as highly degraded fragments and 10 corresponds to highly intact mRNA fragment (Schroeder et al., 2006). The electrophoresis gel image shows visual interpretation of

electrophoresis run and is used in this report to explain the RNA fragmentation quality. Four different types of bands is observable from the figure 3 as mentioned earlier, the first type with strong bands and low RQN score is a clear indication of high RNA fragmentation for corresponding samples. The second types being moderately light brands and medium RQN score is an indication of medium fragmentation. Which means, the corresponding samples are neither fully fragmented nor completely intact. The third type with light band and high RQN score indicates highly intact RNA fragment. The last type with no bands is a clear indication of highest fragmentation in RNA. These statements are supported by the fragmentation curve shown in figure 4 where three different samples were presented with three different RQN and lower to highest RQN indicating lowest to highest fragmentation depending on the abundance of peaks observable for fragment sizes mentioned in result section. This difference in fragmentation quality possibly arisen from the RNase contamination while working with samples. The volume of tissue used also has an impact on RNA integrity as the higher amount used shows lower the integrity of RNA (Li et al, 2009). Finally repeated freezing and thawing of frozen sample tissue also causes RNA degradation which can also be the case in here as the extracted samples were kept in freezer below 4°C for storage and were also used while required for further steps (Florell et al, 2001). Fragmentation in some of the samples can cause generation of low quality sequencing output as explained further in the discussion below.

No major relationship has been observed for fragment quality with sample concentration and purity. For example, sample A-1, A-2, A-5 and A-7 shows relatively higher concentration than other samples where only first three samples has purity higher than 2.0. RQN score for each sample mentioned is different. Only similarity has been observed between last two samples mentioned. A similar study also measured RNA integrity number with different RNA concentration using different quality and quantity assesment methods but no relationship was observed between RNA concentration and RNA integrity (Wong & Pang, 2013).

According to a study by Ugge et al. (2020) the expected amount of total RNA in 50 mg of mussel tissue sample is 56.25 µg (90 µg of total RNA was estimated in following study from 80 mg of sample). This means the expected concentration of mRNA per sample should be within 562.5 ng to 2810 ng in this study. Most of the sample shows a concentration ranges from 0.3% to 4.5 % of the expected total RNA quantity. However, the possible reason for few sample with lower concentration might lies with mRNA extraction method. As Dynabead's poly(A) based extraction method only extracts those mRNA that contains fragments containing poly(A) tail, it is possible that those mRNA that are fragmented and does not contain poly(A) tail at the 3' end were not extracted out of total RNA content (Garalde et al., 2018).

After extraction, the next step of the project included sequencing library preparation with PCR-cDNA Barcoding kit (SQK-PCB 109) that allows sample input as low as 1 ng and simultaneously sequence 12 samples. Two different RT enzymes were used for RT-PCR as mentioned earlier due to unavailability of enough volume of enzyme required for all of the samples. The protocol (ONT) for PCR-cDNA Barcoding kit suggested Maxima H Minus Reverse Transcriptase enzyme but the volume present in Maxima H Minus Reverse Transcriptase kit met the requirement only for samples of *Anodonta anatina* species. However, PCR amplification can introduce bias in sequencing coverage. The types of biases includes, secondary structure and primer dimer formation, loss of specific RNA species, sample loss due to target specificity, duplication of sequences (Van Dijk, Jaszczyszyn & Thermes, 2014). Possible effect of PCR bias was assumed in

this study where low concentration of library cDNA sample was observed (table 3). Such biases can be avoided by using cluster amplification step instead of PCR step (Kozarewa et al., 2009).

Quality assessment with Nanodrop at 280/260 ratio for cDNA samples ensured the purity of the samples as all the samples showed a value within the range of 1.8 to 2.0 (Table 3) and absorbance ratio within this range is considered pure for DNA samples (Wilfinger et al., 1997). Quantity measurement with Qubit showed two different types of concentration variance, the first one being below 1 ng/ $\mu$ l and second one is higher than 1 ng/ $\mu$ l (table 3). All the samples of *Unio tumidus* species and two samples from *Anodonta anatina* showed a concentration higher than 1 ng/ $\mu$ l. Possible reason for this difference may arise from sample loss due to PCR bias or usage of different polymerase enzyme as two different enzymatic unit/ $\mu$ l concentration for two enzymes has been used. In general, cDNA concentration of 1 ng/ $\mu$ l shows better sequencing output (Oxford Nanopore Technologies, 2019). Due to low concentration of cDNA for few samples lower than 1 ng/ $\mu$ l, cDNA integrity assessment on cDNA library samples was required. Only one attempt in quality assessment with fragment analyzer was performed with dsDNA 915 Reagent Kit (35-5000bp) (Aligent) on cDNA library samples but no result was observed due to presence of contamination in dilution marker. This problem also caused loss of cDNA library samples. After one attempt, no further quality assessment was possible to perform with fragment analyzer on cDNA library samples due to presence of lower volume of the samples than required. Thus, this caused loss of important data regarding fragment size and quality for cDNA library samples.

#### 4.2. Sequence data output

The sequencing of polyA mRNA was performed using a MinION and nanopore sequencing technology (ONT). An average of 943 fastq files (Appendix B) with total number of 3,771,813 number of sequence reads were generated (Table 4). This number of reads generated by the run is much lower than sequence reads from other similar studies, where a total number of 9,900,000 reads were generated by poly(A) mRNA sequencing with Oxford nanopore technology (Workman et al., 2019). However, similar sequence read result cannot be expected from a study conducted on human or other mammalian animals but there should be 7,000,000 to 12,000,000 sequence reads in a single run per flow cell with PCR-cDNA sequencing kit (ONT, 2019). Also, the number of reads generated in this study are also very lower than average sequence reads generated in general from those that uses illumina sequencing technology for sequencing (Mizrachi, Hefer, Ranik, Joubert & Myburg, 2010). This might caused from having a highly fragmented sequences or lower fragment size that is undetectable by the nanopore. These shorter fragments mostly creates background noise in sequence output rather than meaningful sequencing data which is usually not detectable by the basecalling algorithm.

The average read length for this sequencing experiment was 173 base pairs (Table 5). But the fragment lengths produced by nanopore sequencing can exceed 10,000 base pairs (Lu et al., 2016). Some studies showed relatively shorter median mRNA fragments of 600-1200 base pairs with nanopore sequencing (Boldogkői, Moldován, Szűcs & Tombácz, 2018). The length of fragment length might depends on the organism and experiment type. A study by Ugge et al. (2020) showed a medium fragment size of 650 base pairs. On the other hand, illumina sequencing technology generates much shorter fragment sizes than nanopore sequencing technology (Mizrachi et al., 2010). Problem with shorter fragment size than expected possibly arose from improper sample storage or its duration and repeated freezing and thawing of frozen mRNA samples numerous time. The sample preparation started at the beginning of April 2020 and sequence run was performed in 11<sup>th</sup> May 2020. Longer shortage period has effect on nucleic acid

integrity through biolysis (Srinivasan, Sedmak & Jewell, 2002). Also repeated freezing and thawing of tissue and sample produces bad quality sample (Botling et al., 2009).

Quality assessment on sequencing output showed an average sequencing quality score of 8.3 (Table 5). It corresponds to base call accuracy less than 90%. Basically, a read quality of 8.3 means there is a probability to receive one error nucleotide base read out of every 8.3 nucleotide reads higher than the cutoff value i.e. sequencing quality score 8 (Tyler et al, 2018). But the current standard for Oxford nanopore technologies is nanopore sequencing generating a quality score of 10 and has 90% accuracy of base call. Quality score produced by nanopore technology is very different from Illumina sequencing technology. Illumina HiSeq 2000 in average produces a quality score greater than 30 and also has a base call accuracy of 99.9% (Illumina, 2012). This shows that nanopore sequencing is much more error-prone than Illumina sequencing when it comes to base calling accuracy. Proper guidance for error correction is required (Goodwin, 2015). Also an average of 943 fastq file generation was observed after basecalling (Appendix B).

During five hour of sequence run, the channel activity was very low indicating low sequence read during this period (Figure 5). After 10 hours of runtime, the channel activity increased gradually till next 14 hours. This indicates that highest number of sequence read was received during this time period. According to product information of R9.4.1 Flow cell (FLO-MIN 106D) each flow cell should contain 512 active channels. But in this case only 486 channels were active. The possible reason might be loss of channel activity during washing step or blockage.

### 4.3. Interpretation and analysis

Due to lack of reference genome from *Anadonta anatina* and *Unio tumidus* in publicly available databases, no trail alignment with reference genome was possible for detection target transcripts. In case, NCBI BLASTn server was used to identify sequenced transcripts.

Output of BLASTn showed the alignment region of target sequence with the query sequence and the measurement parameters of the alignment (Figure 6). Expected threshold, alignment score and sequence identity is the most important factors for identification of homology for a genetic sequence (Pearson, 2013). A rule of thumb implies an alignment score higher than 50 is always statistically significant and is sufficient as statistically significant for a database with 7 million entries thus considered as a minimum threshold. Also the same rule predicts that the sequence identity higher than 30% and E value less than 0.01 is also always significantly higher also are minimum thresholds (Pearson, 2013). In this project, minimum alignment score was set to 70, E value was 1 but all transcripts reported had an e value less than 0.01 and finally all the identified transcript had sequence identity higher than 30%. However, the cutoff for sequence identity was found assumed different i.e. 90% in another study (Unneberg, Wennborg & Larsson, 2003). Also a study used 95% sequence coverage in order to accept transcripts fully reliable (Roberts, Pimentel, Trapnell & Pachter, 2011). In this study, the sequence output was not merged and highly fragmented, so in this case sequence identity, Expected threshold and alignment score was considered for more reliability over sequence coverage.

Comparison between the transcriptome generated from mussel species *Anodonta anatina* and *Unio tumidus* shows common mitochondrial transcripts such as COX I, II and III, NADH dehydrogenase subunit 4 that are part of in aerobic respiratory electron transport chain (GO:00019646) and ATP synthase F0 subunit 6 that is a part of ATP synthesis process (GO:0015986) and combinedly aid in proton pumping process across the mitochondrial membrane (Michel, Behr, Harrenga & Kannt, 1998). However these transcripts are highly conserved and common in most

living organism. Ribosomal subunits such as 12S, 16S and 28S rRNA found were also common for both species. But, presence of these rRNA transcripts was unexpected as the extraction method was specified only for mRNA species. The possible reason for this might be due to contamination with rRNA during mRNA purification step. A study also supports the statement that the optimization of reagent buffer and strictly controlled handling during purification is necessary to ensure higher purity of mRNA in sequence library sample (Wang, Wang, Zang, Sun & Yang, 2018).

Gonadotropin releasing hormone/ corazonin is the last common transcript found in both species and this protein is a part of reproduction system (GO:0007275). It is also found involved in central functions including feeding, locomotion, heart control, and reproduction (Fodor et al., 2020). This indicates the mussel samples that were going through stresses of reproductive phase.

Sigma-class glutathione S-transferase, Pi- glutathione S transferase are transcripts that give rise to protein binding (GO:0005515) protein that plays role to protect cells from oxidative stress and exogenous toxic compounds (Li, Yang, Huang & Li, 2015) These proteins were identified in different species in *Anodonta anatina* and their presence indicates those mussels were experiencing episodes of oxidative stress or toxic exposure. The concept of investigating glutathione S-transferases can further be explored for assessment of toxic exposure on aquatic environment.

Methallothionein is a type of transcript codes for ion binding protein that has been identified from different samples of *Anodonta anatina* species. Methallothionein is associated with metal ion binding (GO:0046872). Association of this protein was reported for protection against metal exposure and increase metal binding (Roesijadi et al., 1994). Association of this protein was also reported as indicator of heavy metal exposure in recent studies (Le, Zimmermann & Sures 2016). Presence of methallothionein indicates exposure of used mussel samples to heavy metal.

The transcript for Heat shock protein 70 produces a multifunctional regulatory protein. Major role of this highly conserved protein is protein folding chaperone (GO:0044183) activity to translocate protein molecules within cell. It also function as providing protein protection and restoration of damaged protein. Increased expression of HSP70 has been reported due to cellular and physiological stress and heavy metal exposure (Kiang & Tsokos, 1998). This transcript was identified from a sample of *Anodonta anatina* species. Elevated level of this transcript indicates cellular stress or metal exposure to the mussel samples. Thus this protein can further be investigated for bioindicator properties to monitor extracellular stress on aquatic life.

Some other transcripts that are only identified in *Anodonta anatina* indicates some important biomarker properties. Some of them found involved in biomineralization process (Calmodulin, Ferritin 1 and Ferritin 2) as shown in appendix E. Each of these transcripts are very important indicative of biomarker characteristics.

In summary, comparative transcriptome of *Anodonta anatina* and *Unio tumidus* species shows five common mitochondrial and three ribosomal transcripts along with some non-common transcript variants with biomarker characteristics which responds to exogenous compounds and stress. The analysis and interpretations focuses ultimately to investigate biomarker properties in mussels for monitoring heavy metal or toxic exposure. However, due to low quality sequencing output and high fragmentation of sequence library caused loss of important transcriptomic data. Error corrected base calling data for better transcriptomic analysis is needed.

## 5. Ethical aspects and impact on the society

The 3R principle (reduction, replacement and refinement) was considered in case of ethical aspects for this project, (Sneddon, Halsey & Bury, 2017). The reduction principle concerns with the concept of using as few animals as possible for the research purpose. Although frozen mussel samples from previous experiments in Högskolan i Skövde were available for a possible transcriptomic analysis but due to prolong storage time, the genetic material of the mussel samples were degraded and found difficult to extract mRNA from rendering tissue. In order to overcome this problem, fresh samples from freshwater were required and frozen samples had to be avoided for transcriptomic and sequencing analysis. In this project, minimum number of animal samples i.e. five mussels for *Unio tumidus* and seven for *Anodonta anatina* were collected for ensuring enough data for conducting transcriptomic research and check for individual variation.

Replacement principle concerns with the replacement of live animals with other techniques that do not require live animals. However, in this study, it is impossible to replace the mussels taken from their natural environment with a non-animal option.

Refinement principle states that pains should be eliminated from the animal samples during experiment. Nervous system and sense organ of species *Anodonta anatina* and *Unio tumidus* are capable of responding to diverse stimulus but they lack spinal cord and brain. Due to lack of central nervous system and nociceptors i.e. receptors that are responsible for sensing pain, they are not capable of feeling pain.

Considering the global situation of water pollution with toxic substances and ecosystem damage, use of mussels can also be motivated as they have been proven as potential bioindicators. The two species used, *Anodonta anatina* and *Unio tumidus*, are common in Sweden and considered in the Least Concern-category of the International Union for Conservation of Nature (IUCN) list. Using appropriate species may open a window to solve this global situation.

## 6. Conclusion

The major aim of this project was to study the comparative transcriptome of two mussel species *Anodonta anatina* and *Unio tumidus*. Because of complications during sample preparation and sequencing, the quality of sequence output was not satisfactory. Quality of mRNA fragments was moderately fragmented, sequence library had lower cDNA concentration and quality of sequence data had relatively low quality score. Comparison of transcripts identified from sequence output showed common mitochondrial transcripts and some uncommon transcripts with biomarker characteristics. Analysis on sequence output also revealed degree of conservation in identified transcripts between two species. However, unavailability of data and research in the same field from shared species makes it much challenging for monitoring bioindicator properties. For future investigation, more planned and careful method selection for sample preparation is required along with sample handling. In order to improve the accuracy of base calling, error correction method is recommended. Better quality sequencing and advanced bioinformatics analysis may give a better resolution on investigation of bioindicator characteristics in freshwater mussels.

## 7. Future perspectives

The pipelines and methods used in this project provide important insights into the areas to improve for better quality and quantity of output. In addition, outcome from this project suggests usage of better extraction method to avoid loss of mRNA concentration and quality. Also

this project emphasize on performing better base calling algorithm with higher accuracy and sequence quality followed by using error correction methods. Most importantly, comparative transcriptomic analysis revealed mitochondrial transcripts in both species along with other transcripts with possible biomarker properties. Information about these transcript that were investigated and predicted as potential biomarker can aid in future research to further validate the result and application.

## **8. Acknowledgments**

Firstly I would like to thank my supervisor Mikael Ejdebäck, Ph.D for his instructions, help, and inspiration.

Secondly I also want to thank my co supervisor John Baxter, for his continuous guidance and help. This thesis project would not be possible to finish without his help during laboratory experiment and analysis.

Then I would like to thank Nazmul Islam from University of Skövde (Masters in Bioinformatics) for his help during bioinformatics analysis and python scripting.

I had to participate the overall thesis project during global corona pandemic situation and country-wise lockdown. So special acknowledgement also goes to my dearest friend Helena Garcia from University of Skövde (Bachelor in Bioscience) and Sukhneet Kaur (Bachelor in Molecular BioDesign) from University of Skövde for being my best support during this time.

I would also like to thank my program coordinator Maria Algerin (University of Skövde) and my mentor Imtiaj Hasan, Ph.D (Associate professor in Biochemistry and Molecular Biology department, University of Rajshahi) for their mentorship, suggestions and help.

Finally I would like to thank my parents for being the best support for me. Their inspiration, love and confidence gave me the energy to fight through the situation and complete this project.

Deepest gratitude to all.

## 9. References

- Agilent (2020), Agilent DNF-472 HS RNA (15 nt) Kit Quick Guide for Fragment Analyzer Systems. Retrieved 2020 April 24 from <https://www.agilent.com/cs/library/usermanuals/public/quick-guide-dnf-472-hs-rna-kit-SD-AT000132.pdf>
- Aldridge, D. C., Fayle, T. M., & Jackson, N. (2007). Freshwater mussel abundance predicts biodiversity in UK lowland rivers. *Aquatic Conservation: Marine and Freshwater Ecosystems*, 17(6), 554-564.
- Annie, J., Ann, B., & Mats, R. (2013). Spatial distribution and age structure of the freshwater unionid mussels *Anodonta anatina* and *Unio tumidus*: implications for environmental monitoring. *Hydrobiologia*, 711(1), 61-70.
- Arvaniti, M., Jensen, M. M., Soni, N., Wang, H., Klein, A. B., Thiriet, N., ... & Kohlmeier, K. A. (2016). Functional interaction between Lypd6 and nicotinic acetylcholine receptors. *Journal of neurochemistry*, 138(6), 806-820.
- Arvaniti, M., Polli, F. S., Kohlmeier, K. A., Thomsen, M. S., & Andreasen, J. T. (2018). Loss of Lypd6 leads to reduced anxiety-like behaviour and enhanced responses to nicotine. *Progress in Neuro-Psychopharmacology and Biological Psychiatry*, 82, 86-94.
- Asif, N., Malik, M., & Chaudhry, F. N. (2018). A review of on environmental pollution bioindicators. *Pollution*, 4(1), 111-118.
- Beintema, J. J., Campagne, R. N., & Gruber, M. (1973). Rat pancreatic ribonuclease I. Isolation and properties. *Biochimica et Biophysica Acta (BBA)-Protein Structure*, 310(1), 148-160.
- Boldogkői, Z., Moldován, N., Szűcs, A., & Tombácz, D. (2018). Transcriptome-wide analysis of a baculovirus using nanopore sequencing. *Scientific data*, 5, 180276.
- Bolotov, I. N., Kondakov, A. V., Konopleva, E. S., Vikhrev, I. V., Aksenova, O. V., Aksenov, A. S., ... & Gofarov, M. Y. (2020). Integrative taxonomy, biogeography and conservation of freshwater mussels (Unionidae) in Russia. *Scientific reports*, 10(1), 1-20.
- Botling, J., Edlund, K., Segersten, U., Tahmasebpoor, S., Engström, M., Sundström, M., ... & Micke, P. (2009). Impact of thawing on RNA integrity and gene expression analysis in fresh frozen tissue. *Diagnostic Molecular Pathology*, 18(1), 44-52.
- Chandurvelan, R., Marsden, I. D., Glover, C. N., & Gaw, S. (2015). Assessment of a mussel as a metal bioindicator of coastal contamination: relationships between metal bioaccumulation and multiple biomarker responses. *Science of the Total Environment*, 511, 663-675
- Cunha, S. C., Pena, A., & Fernandes, J. O. (2017). Mussels as bioindicators of diclofenac contamination in coastal environments. *Environmental Pollution*, 225, 354-360.
- Engström, E. (2019). Direct poly (A) RNA nanopore sequencing on the freshwater duck mussel *Anodonta anatina* following exposure to copper: A pilot study.
- Feng, Y., Zhang, Y., Ying, C., Wang, D., & Du, C. (2015). Nanopore-based fourth-generation DNA sequencing technology. *Genomics, proteomics & bioinformatics*, 13(1), 4-16.

- Florell, S. R., Coffin, C. M., Holden, J. A., Zimmermann, J. W., Gerwels, J. W., Summers, B. K., ... & Leachman, S. A. (2001). Preservation of RNA for functional genomic studies: a multidisciplinary tumor bank protocol. *Modern pathology*, 14(2), 116-128.
- Fodor, I., Zrinyi, Z., Horvath, R., Urban, P., Herczeg, R., Buki, G., ... & Pirger, Z. (2020). Identification, presence, and possible multifunctional regulatory role of invertebrate gonadotropin-releasing hormone/corazonin molecule in the great pond snail (*Lymnaea stagnalis*). *BioRxiv*.
- Franco, J. L., Trivella, D. B., Trevisan, R., Dinslaken, D. F., Marques, M. R., Bainy, A. C., & Dafre, A. L. (2006). Antioxidant status and stress proteins in the gills of the brown mussel *Perna perna* exposed to zinc. *Chemico-Biological Interactions*, 160(3), 232-240.
- Galante, Y. M., & Hatefi, Y. (1979). Purification and molecular and enzymic properties of mitochondrial NADH dehydrogenase. *Archives of Biochemistry and Biophysics*, 192(2), 559-568.
- Garalde, D. R., Snell, E. A., Jachimowicz, D., Sipos, B., Lloyd, J. H., Bruce, M., Pantic, N., Admassu, T., James, P., Warland, A., Jordan, M., Ciccone, J., Serra, S., Keenan, J., Martin, S., McNeill, L., Wallace, E. J., Jayasinghe, L., Wright, C., Blasco, J., Young, S., Bocklebank, D., Juul, S., Clarke, J., Heron, A. J. & Turner, D. J. (2018). Highly parallel direct RNA sequencing on an array of nanopores. *Nature Methods*, 15(3), pp. 201-206. <https://doi.org/10.1038/nmeth.4577>
- Goodwin, S., Gurtowski, J., Ethe-Sayers, S., Deshpande, P., Schatz, M. C., & McCombie, W. R. (2015). Oxford Nanopore sequencing, hybrid error correction, and de novo assembly of a eukaryotic genome. *Genome research*, 25(11), 1750-1756.
- Graf, D. L., & Cummings, K. S. (2007). Review of the systematics and global diversity of freshwater mussel species (Bivalvia: *Unionoida*). *Journal of Molluscan Studies*, 73(4), 291-314.
- Hitchcock-DeGregori, S. E., & Barua, B. (2017). Tropomyosin structure, function, and interactions: a dynamic regulator. In *Fibrous proteins: Structures and mechanisms* (pp. 253-284). Springer, Cham.
- Hoellein, T. J., Zarnoch, C. B., Bruesewitz, D. A., & DeMartini, J. (2017). Contributions of freshwater mussels (*Unionidae*) to nutrient cycling in an urban river: filtration, recycling, storage, and removal. *Biogeochemistry*, 135(3), 307-324.
- Hoffmann, A. A., & Willi, Y. (2008). Detecting genetic responses to environmental change. *Nature Reviews Genetics*, 9(6), 421-432.
- Hüning, A. K., Melzner, F., Thomsen, J., Gutowska, M. A., Krämer, L., Frickenhaus, S., ... & Lucassen, M. (2013). Impacts of seawater acidification on mantle gene expression patterns of the Baltic Sea blue mussel: implications for shell formation and energy metabolism. *Marine Biology*, 160(8), 1845-1861.
- Jain, M., Olsen, H. E., Paten, B., & Akeson, M. (2016). The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. *Genome biology*, 17(1), 239.
- Kiang, J. G., & Tsokos, G. C. (1998). Heat shock protein 70 kDa: molecular biology, biochemistry, and physiology. *Pharmacology & therapeutics*, 80(2), 183-201.
- Le, T. Y., Zimmermann, S., & Sures, B. (2016). How does the metallothionein induction in bivalves meet the criteria for biomarkers of metal exposure?. *Environmental Pollution*, 212, 257-268.

- Lepoutre, A., Hervieux, J., Faassen, E. J., Zweekers, A. J., Lurling, M., Geffard, A., & Lance, E. (2020). Usability of the bivalves *Dreissena polymorpha* and *Anodonta anatina* for a biosurvey of the neurotoxin BMAA in freshwater ecosystems. *Environmental Pollution*, 259, 113885.
- Li, D., Ren, W., Wang, X., Wang, F., Gao, Y., Ning, Q., ... & Lu, S. (2009). A modified method using TRIzol® reagent and liquid nitrogen produces high-quality RNA from rat pancreas. *Applied biochemistry and biotechnology*, 158(2), 253-261.
- Li, J., Lusher, A. L., Rotchell, J. M., Deudero, S., Turra, A., Bråte, I. L. N., ... & Shi, H. (2019). Using mussel as a global bioindicator of coastal microplastic pollution. *Environmental pollution*, 244, 522-533.
- Liu, X., Ji, C., Zhao, J., Wang, Q., Li, F., & Wu, H. (2014). Metabolic profiling of the tissue-specific responses in mussel *Mytilus galloprovincialis* towards *Vibrio harveyi* challenge. *Fish & shellfish immunology*, 39(2), 372-377.
- Lu, H., Giordano, F., & Ning, Z. (2016). Oxford Nanopore MinION sequencing and genome assembly. *Genomics, proteomics & bioinformatics*, 14(5), 265-279.
- Mattick, J. S., & Gagen, M. J. (2001). The evolution of controlled multitasked gene networks: the role of introns and other noncoding RNAs in the development of complex organisms. *Molecular biology and evolution*, 18(9), 1611-1630.
- Michel, H., Behr, J., Harrenga, A., & Kannt, A. (1998). Cytochrome c oxidase: structure and spectroscopy. *Annual review of biophysics and biomolecular structure*, 27(1), 329-356.
- Mizrachi, E., Hefer, C. A., Ranik, M., Joubert, F., & Myburg, A. A. (2010). De novo assembled expressed gene catalog of a fast-growing Eucalyptus tree produced by Illumina mRNA-Seq. *BMC genomics*, 11(1), 1-12.
- Moore, M. N., Viarengo, A., Donkin, P., & Hawkins, A. J. (2007). Autophagic and lysosomal reactions to stress in the hepatopancreas of blue mussels. *Aquatic Toxicology*, 84(1), 80-91.
- Oxford Nanopore Technology (2019), PCR-cDNA Barcoding (SQK-PCB109). Retrieved 2020 May 18 from [https://store.nanoporetech.com/eu/media/wysiwyg/pdfs/SQK-PCB109/MSDS\\_-\\_SQK-PCB109.pdf](https://store.nanoporetech.com/eu/media/wysiwyg/pdfs/SQK-PCB109/MSDS_-_SQK-PCB109.pdf)
- Parmar, T. K., Rawtani, D., & Agrawal, Y. K. (2016). Bioindicators: the natural indicator of environmental pollution. *Frontiers in life science*, 9(2), 110-118.
- Peano, C., Pietrelli, A., Consolandi, C., Rossi, E., Petiti, L., Tagliabue, L., ... & Landini, P. (2013). An efficient rRNA removal method for RNA sequencing in GC-rich bacteria. *Microbial informatics and experimentation*, 3(1), 1-11.
- Pearson, W. R. (2013). An introduction to sequence similarity ("homology") searching. *Current protocols in bioinformatics*, 42(1), 3-1.
- Peirson, S. N., & Butler, J. N. (2007). RNA extraction from mammalian tissues. In *Circadian Rhythms* (pp. 315-327). Humana Press.
- Prié, V., Valentini, A., Lopes-Lima, M., Froufe, E., Rocle, M., Poulet, N., ... & Dejean, T. (2020). Environmental DNA metabarcoding for freshwater bivalves biodiversity assessment: methods and results for the Western Palearctic (European sub-region). *Hydrobiologia*, 1-20.

- Rademacher, F., Simanski, M., & Harder, J. (2016). RNase 7 in cutaneous defense. *International journal of molecular sciences*, 17(4), 560.
- Roberts, A., Pimentel, H., Trapnell, C., & Pachter, L. (2011). Identification of novel transcripts in annotated genomes using RNA-Seq. *Bioinformatics*, 27(17), 2325-2329.
- Roesijadi, G. (1994). Metallothionein induction as a measure of response to metal exposure in aquatic animals. *Environmental Health Perspectives*, 102(suppl 12), 91-95.
- Salazar, M. H., & Salazar, S. M. (1996). Mussels as bioindicators: effects of TBT on survival, bioaccumulation, and growth under natural conditions. In *Organotin* (pp. 305-330). Springer, Dordrecht.
- Schroeder, A., Mueller, O., Stocker, S., Salowsky, R., Leiber, M., Gassmann, M., ... & Ragg, T. (2006). The RIN: an RNA integrity number for assigning integrity values to RNA measurements. *BMC molecular biology*, 7(1), 1-14.
- Sneddon, L. U., Halsey, L. G., & Bury, N. R. (2017). Considering aspects of the 3Rs principles within experimental animal biology. *Journal of Experimental Biology*, 220(17), 3007-3016.
- Sohail, M., Khan, M. N., Chaudhry, A. S., & Qureshi, N. A. (2016). Bioaccumulation of heavy metals and analysis of mineral element alongside proximate composition in foot, gills and mantle of freshwater mussels (*Anodonta anatina*). *Rendiconti Lincei*, 27(4), 687-696.
- Srinivasan, M., Sedmak, D., & Jewell, S. (2002). Effect of fixatives and tissue processing on the content and integrity of nucleic acids. *The American journal of pathology*, 161(6), 1961-1971.
- Tanabe, S., Tatsukawa, R., & Phillips, D. J. (1987). Mussels as bioindicators of PCB pollution: A case study on uptake and release of PCB isomers and congeners in green-lipped mussels (*Perna viridis*) in Hong Kong waters. *Environmental Pollution*, 47(1), 41-62.
- Theil, E. C. (2004). Iron, ferritin, and nutrition. *Annu. Rev. Nutr.*, 24, 327-343.
- ThermoFisher Scientific (2012), Dynabeads® mRNA DIRECT™ Kit: For the isolation of pure mRNA directly from crude samples. Retrieved 2020 April 04 from [https://www.thermofisher.com/document-connect/dynabeads\\_mRNA\\_direct\\_man.pdf](https://www.thermofisher.com/document-connect/dynabeads_mRNA_direct_man.pdf)
- ThermoFisher Scientific (2015), Qubit® RNA HS Assay Kits: For use with the Qubit® Fluorometer. Retrieved 2020 April 08 from [https://www.thermofisher.com/document-connect/Qubit\\_RNA\\_HS\\_Assay\\_UG.pdf](https://www.thermofisher.com/document-connect/Qubit_RNA_HS_Assay_UG.pdf)
- Tyler, A. D., Mataseje, L., Urfano, C. J., Schmidt, L., Antonation, K. S., Mulvey, M. R., & Corbett, C. R. (2018). Evaluation of Oxford Nanopore's MinION sequencing device for microbial whole genome sequencing applications. *Scientific reports*, 8(1), 1-12.
- Ugge, G. M. O. E., Jonsson, A., Olsson, B., Sjöback, R., & Berglund, O. (2020). Transcriptional and biochemical biomarker responses in a freshwater mussel (*Anodonta anatina*) under environmentally relevant Cu exposure. *Environmental Science and Pollution Research*, 1-12.
- Unneberg, P., Wennborg, A., & Larsson, M. (2003). Transcript identification by analysis of short sequence tags—influence of tag length, restriction site and transcript database. *Nucleic acids research*, 31(8), 2217-2226.

- Van Dijk, E. L., Jaszczyszyn, Y., & Thermes, C. (2014). Library preparation methods for next-generation sequencing: tone down the bias. *Experimental cell research*, 322(1), 12-20.
- Wang, L., Wang, Y., Zang, D., Sun, Z., & Yang, C. (2018). Optimization of Poplar mRNA purification for transcriptome library construction. *Acta biochimica et biophysica Sinica*, 50(2), 224-226.
- Vermeulen, J., De Preter, K., Lefever, S., Nuytens, J., De Vloed, F., Derveaux, S., ... & Vandesompele, J. (2011). Measurable impact of RNA quality on gene expression results from quantitative PCR. *Nucleic acids research*, 39(9), e63-e63.
- Wilfinger, W. W., Mackey, K., & Chomczynski, P. (1997). 260/280 and 260/230 ratios NanoDrop® ND-1000 and ND-8000 8-sample spectrophotometers. *BioTechniques*, 22, 474-481.
- Wong, K. S., & Pang, H. M. (2013). Simplifying HT RNA Quality & Quantity Analysis: Automated CE System Designed to Improve Rapid Assessment. *Genetic Engineering & Biotechnology News*, 33(2), 17-17.
- Woo, S., Jeon, H. Y., Kim, S. R., & Yum, S. (2011). Differentially displayed genes with oxygen depletion stress and transcriptional responses in the marine mussel, *Mytilus galloprovincialis*. *Comparative Biochemistry and Physiology Part D: Genomics and Proteomics*, 6(4), 348-356.
- Workman, R. E., Tang, A. D., Tang, P. S., Jain, M., Tyson, J. R., Razaghi, R., ... & Sadowski, N. (2019). Nanopore native RNA sequencing of a human poly (A) transcriptome. *Nature Methods*, 16(12), 1297-1305.
- Zieritz, A., Bogan, A. E., Froufe, E., Klishko, O., Kondo, T., Kovitvadhi, U., ... & Sousa, R. (2018). Diversity, biogeography and conservation of freshwater mussels (Bivalvia: *Unionida*) in East and Southeast Asia. *Hydrobiologia*, 810(1), 29-44.
- Özhan, G., Sezgin, E., Wehner, D., Pfister, A. S., Kühl, S. J., Kagermeier-Schenk, B., ... & Weidinger, G. (2013). Lypd6 enhances Wnt/ $\beta$ -catenin signaling by promoting Lrp6 phosphorylation in raft plasma membrane domains. *Developmental cell*, 26(4), 331-345.
- Zhang, Z., Schwartz, S., Wagner, L., & Miller, W. (2000). A greedy algorithm for aligning DNA sequences. *Journal of Computational biology*, 7(1-2), 203-214.

## 10. Appendix

### A. Optimization of spinning frequency and Runtime in TussieLyser LT

Optimization of the spinning frequency and runtime in TissueLyser LT tissue homogenization method.

Assay	30 Hz 40 Sec.	50 Hz 40 Sec.	30 Hz 60 Sec.	50 Hz 60 Sec.
<b>Concentration (ng/ <math>\mu</math>l)</b>	4.1	4.4	3.1	3.2
<b>(260/280) absorbance ratio</b>	1.47	2.00	2.34	2.47

### B. Number of Fastq files generated per sample.

Number of fastq files generated per sample after base calling.

Sample	U-1	U-2	U-3	U-4	U-5	A-1	A-2	A-3	A-4	A-5	A-6	A-7
<b>Number of file</b>	298	943	943	943	943	943	943	943	943	943	943	943

### C. Transcripts identified from *Unio tumidus* species

Transcripts identified from *Unio tumidus* species. Result shown here only from species belonging to phylum mollusk (exception *Coturnix Japonica* and *parasutterella* sp.) and has a alignment score with corresponding target transcript higher than 70, expected value close to 1. Percentage identity of aligned query sequence with aligned target sequence is given with the coverage of the query sequence with the target transcript sequence according to each target transcript in each row along with species source and its accession ID. Barcode samples represent the source of query sequence.

Target transcript	Target identity	Species	Accession or protein ID of target	mRNA ID of query cover	Barcode samples
<b>12S Ribosomal RNA</b>	84/85 (99%)	<i>Anodonta anatina</i>	GU584015.1	85/570 (15%)	U-1, U-2
<b>16S Ribosomal RNA</b>	34/35 (97%)	<i>Parasutterella</i> sp.	MN135768.1	35/4114 (1%)	U-1, U-2, U-4, U-5
<b>28s Ribosomal RNA</b>	80/85 (94%)	<i>Aculamprotula polysticta</i>	MK687417.1	85/791 (11%)	U-2
<b>Cytochrome oxidase subunit I (Mitochondrial)</b>	<b>c</b> 84/98 (86%)	<i>Lamprotula scripta</i>	AND82411.1	98/1529 (6.4)	U-2
<b>cytochrome oxidase subunit II (Mitochondrial)</b>	<b>c</b> 263/286(92%)	<i>Unio tumidus</i>	AQM37837.1	286/681 (42%)	U-2
<b>cytochrome oxidase subunit III (Mitochondrial)</b>	<b>c</b> 195/242(81%)	<i>Lasmigona compressa</i>	ADL62635.1	242/880 (28%)	U-2

<b>ATP synthase subunit (Mitochondrial)</b>	<b>F06</b>	173/209(83%)	<i>Utterbackia imbecillis</i>	ADL62617.1	209/702 (30%)	U-2
<b>NADH dehydrogenase subunit (Mitochondrial)</b>	<b>4</b>	100/105(95%)	<i>Unio tumidus</i>	AQM37831.1	105/1347 (8%)	U-2
<b>*Gonadotropin releasing hormone/ corazonin</b>		31/32 (97%)	<i>Lymnaea stagnalis</i>	QIH29241.1	32/360 (9%)	U-3, U-4
<b>*Domain containing 6 (LYPD6)</b>		30/31 (97%)	<i>Coturnix Japonica</i>	XP_015724597.1	31/510 (6%)	U-5

#### D. Transcripts identified from *Anodonta anatina* species.

Identified transcripts from *Anodonta anatina* sample. Result shown here were taken from species only belongs to phylum mollusk and has a alignment score with corresponding target transcript higher than 70, expected value close to 1. Percentage identity of aligned query sequence with aligned target sequence is given with the coverage of the query sequence with the target transcript sequence according to each target transcript in each row along with species source and its accession ID. Barcode samples represent the source of query sequence.

<b>Transcript</b>	<b>Target identity</b>	<b>Species</b>	<b>Accession ID</b>	<b>mRNA query cover</b>	<b>Barcodes</b>
<b>60s acidic ribosomal protein P1- like</b>	61/72 (85%)	<i>Pomacea canaliculata</i>	XP025090121.1	72/118 (61%)	A-1
<b>*Gonadotropin releasing hormone/ corazonin</b>	32/33 (97%)	<i>Lymnaea stagnalis</i>	QIH29241.1	33/360 (9%)	A-1, A-3
<b>NADH dehydrogenase subunit 6 (Mitochondrial)</b>	289/311 (93%)	<i>Anodonta cygnea</i>	AVI15553.1	311/489 (64%)	A-1
<b>Cytochrome Oxidase subunit II (Mitochondrial)</b>	378/411(92%)	<i>Anodonta anatina</i>	AGS18010.1	411/681 (60%)	A-1, A-2, A-3, A-4, A-7
<b>18S Ribosomal RNA</b>	558/621 (90%)	<i>Fusconaia flava</i>	KX342024.1	621/1240 (50%)	A-1, A-2, A-5, A-7
<b>*Sigma-class glutathione transferase</b>	247/269 (92%)	<i>Sinanodonta woodiana</i>	AQW43003.1	269/612 (44%)	A-1, A-3
<b>*Ferritin 1</b>	189/208 (91%)	<i>Sinanodonta woodiana</i>	ADZ04888.1	208/525 (40%)	A-1, A-4
<b>*Ferritin 2</b>	189/208 (90%)	<i>Sinanodonta woodiana</i>	AEK27025.1	208/525 (40%)	A-1, A-4

<b>NADH dehydrogenase subunit 4 (Mitochondrial)</b>	112/116 (97%)	<i>Anodonta cygnea</i>	AVI15554.1	116/1347 (9%)	A-1, A-4
<b>12S Ribosomal RNA</b>	502/592 (85%)	<i>Anodonta anatina</i>	GU584012.1	592/1347 (44%)	A-1, A-4, A-5, A-7
<b>16S Ribosomal RNA</b>	182/192 (95%)	<i>Anodonta anatina</i>	MF781083.1	192/1310 (15%)	A-1, A-4, A-5, A-6
<b>*Pi- glutathione S transferase</b>	282/307 (92%)	<i>Cristaria plicata</i>	ADM88875.1	307/618 (50%)	A-2
<b>Trypsin- like protein</b>	340/393 (87%)	<i>Hyriopsis cumingii</i>	AEB70966.1	393/864 (45%)	A-2
<b>*Calmodulin (CaM)</b>	321/345 (95%)	<i>Hyriopsis cumingii</i>	ACI22622.1	345/450 (77%)	A-2
<b>Alpha- amylase</b>	370/467 (79%)	<i>Hyriopsis cumingii</i>	AGW45296.1	467/1572 (30%)	A-2
<b>Phage lysozyme 1</b>	131/144 (91%)	<i>Cristaria plicata</i>	ALL27411.1	144/468 (31%)	A-2
<b>Cytochrome C Oxidase subunit I (Mitochondrial)</b>	335/354 (95%)	<i>Anodonta anatina</i>	AGS18009.1	354/1542 (23%)	A-2, A-3, A-4
<b>Cytochrome C Oxidase subunit III (Mitochondrial)</b>	321/337 (95%)	<i>Anodonta anatina</i>	AGS17980.1	337/780 (43%)	A-2, A-3, A-4
<b>ATP synthase FO subunit 6 (Mitochondrial)</b>	354/373 (95%)	<i>Anodonta anatina</i>	AGS18007.1	373/780 (48%)	A-2, A-4
<b>28S Ribosomal RNA</b>	240/278 (86%)	<i>Anodonta cygnea</i>	AM779650.1	278/1487 (19%)	A-2, A-5
<b>Tropomyosin</b>	70/85 (82%)	<i>Saccostrea glomerata</i>	AVD53650.1	85/855 (10%)	A-3
<b>NADH dehydrogenase subunit 2 (Mitochondrial)</b>	254/309 (92%)	<i>Anodonta anatina</i>	AGS18013.1	309/966 (32%)	A-3, A-4
<b>*Methallothionein</b>	302/361 (84%)	<i>Unio tumidus</i>	ABP01350.1	361/413 (87%)	A-4
<b>*Heat shock protein 70 mRNA</b>	350/388 (90%)	<i>Sinanodonta woodiana</i>	AMR60410.1	388/1974 (20%)	A-7
<b>Beta-actin</b>	498/593 (84%)	<i>Fusconaia flava</i>	ANY58936.1	593/1048 (57%)	A-7
<b>Elongation factor 1- alpha</b>	279/309 (90%)	<i>Fusconaia flava</i>	ANY58937.1	309/1113 (28%)	A-7
<b>Y-box protein</b>	285/300 (95%)	<i>Hyriopsis cumingii</i>	AIT55908.1	300/690 (43%)	A-7

E. Gene ontological information of the identified transcripts.

Molecular functions of each transcripts identified from both species are given here along with the biological processes they are involved in. Here N/a means not applicable as not all the transcripts are involved in biological process and are thus can be called multifunctional protein.

<b>Transcript</b>	<b>Biological process</b>	<b>Molecular Function</b>
<b>ATP synthase F0 subunit 6 (Mitochondrial)</b>	ATP synthesis coupled proton transport (GO:0015986)	Proton transmembrane transporter activity (GO:0015078)
<b>Gonadotropin releasing hormone/ corazonin</b>	multicellular organism development (GO:0007275)	Hormone activity (GO:0005179)
<b>Domain containing 6 (LYPD6)</b>	Wnt/beta-catenin signaling (Özhan et al., 2013)	Acetylcholine receptor regulator activity (GO:0030548)
<b>60s acidic ribosomal protein P1- like</b>	Translational elongation (GO:0006414)	Structural constituent of ribosome (GO:0003735)
<b>Sigma-class glutathionine S-transferase, Pi-glutathione S transferase</b>	N/a	Protein binding (GO:0005515)
<b>Ferritin 1, Ferritin 2</b>	Iron ion transport (GO:0006826)	Ferric iron binding (GO:0008199)
<b>Trypsin- like protein</b>	Protein metabolic process (GO:0019538)	Serine-type endopeptidase activity (GO:0004252)
<b>Calmodulin (CaM)</b>	Calcium-mediated signaling (GO:0019722)	Calcium ion binding (GO:0005509)
<b>Alpha- amylase</b>	Carbohydrate metabolic process (GO:0005975)	Alpha-amylase activity (releasing maltohexaose) (GO:0103025)
<b>Phage lysozyme 1</b>	peptidoglycan catabolic process (GO:0009253)	Lysozyme activity (GO:0003796)
<b>Tropomyosin</b>	N/a	Regulation of muscle contraction (IPR000533)
<b>Methallothionein</b>	N/a	Metal ion binding (GO:0046872)
<b>Heat shock protein 70</b>	N/a	protein folding chaperone (GO:0044183)
<b>Beta-actin</b>	N/a	Formation of filaments of cytoskeleton (IPR020902)
<b>Elongation factor 1- alpha</b>	Translational elongation (GO:0006414)	Translation elongation factor activity (GO:0003746)
<b>Y-box protein</b>	N/a	Nucleic acid binding (GO:0003676)