DIRECT POLY(A) RNA NANOPORE SEQUENCING ON THE FRESHWATER DUCK MUSSEL ANODONTA ANATINA FOLLOWING EXPOSURE TO COPPER
A pilot study

Bachelor Degree Project in Bioscience
G2E 30 credits
Spring term 2019

Author
Erik Engström
a16erien@student.his.se

Supervisor
Mikael Ejdebäck
mikael.ejdeback@his.se

Co-supervisor
Magnus Fagerlind
magnus.fagerlind@his.se

Examiner
Maria Algerin
maria.algerin@his.se
Abstract

Aquatic ecotoxicology is the study of toxic chemicals and its effects on aquatic biological systems with the aim of minimising threats to human health and ensure self-sustained ecosystems. Freshwater bivalves are excellent sentinels for use in ecotoxilogical research due to their filter feeding properties, stationary lifestyle and inability to regulate body temperature. This project aimed to assess the feasibility and use of nanopore sequencing, a real-time single-molecule sequencing technology in comparative expression analysis by sequencing transcriptomic RNA from the freshwater mussel *Anodonta anatina* following exposure to copper. RNAs were extracted from 80 mg hepatopancreas tissue, followed by poly(A) RNA selection. Furthermore, the poly(A) RNA was used to construct a nanopore sequencing library. Sequencing a total amount of 560 ng poly(A) RNA over the course of two separate runs generated 239,448 reads, in which 75% of the reads were obtained during the first run (control) and 25% of the reads were obtained during the second run (case). The median read lengths ranged between 534-650 nucleotides, with a base call accuracy <90%. Due to the big differences in sequence data output between the two sequencing runs, the data was ineligible for comparative analysis. The findings conclude that nanopore sequencing is capable of generating longer read lengths when compared to other sequencing platforms. However, the technology is error-prone in terms of accurate base call identifications and relies on other platforms for error corrections. Future advances include *de novo* transcriptome assembly for efficient use of *Anodonta anatina* as a bioindicator in aquatic ecotoxicology.
<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>AChE</td>
<td>Acetylcholinesterase</td>
</tr>
<tr>
<td>BLASTn</td>
<td>Basic Local Alignment Search Tool nucleotide</td>
</tr>
<tr>
<td>BP</td>
<td>Base Pair</td>
</tr>
<tr>
<td>CAT</td>
<td>Catalase</td>
</tr>
<tr>
<td>CNS</td>
<td>Central Nervous System</td>
</tr>
<tr>
<td>cDNA</td>
<td>Complementary DNA</td>
</tr>
<tr>
<td>DNase</td>
<td>Deoxyribonuclease</td>
</tr>
<tr>
<td>GPx</td>
<td>Glutathione Peroxidase</td>
</tr>
<tr>
<td>GST</td>
<td>Glutathione S-Transferase</td>
</tr>
<tr>
<td>HSP</td>
<td>Heat Shock Protein</td>
</tr>
<tr>
<td>IQ</td>
<td>Integrity and Quality</td>
</tr>
<tr>
<td>mRNA</td>
<td>Messenger RNA</td>
</tr>
<tr>
<td>MT</td>
<td>Metallothionein</td>
</tr>
<tr>
<td>NCBI</td>
<td>National Centre for Biotechnology Information</td>
</tr>
<tr>
<td>NGS</td>
<td>Next Generation Sequencing</td>
</tr>
<tr>
<td>NRR</td>
<td>Neutral Red Retention</td>
</tr>
<tr>
<td>NT</td>
<td>Nucleotide</td>
</tr>
<tr>
<td>Poly(A)</td>
<td>Polyadenylated</td>
</tr>
<tr>
<td>PCR</td>
<td>Polymerase Chain Reaction</td>
</tr>
<tr>
<td>QC</td>
<td>Quality Control</td>
</tr>
<tr>
<td>qPCR</td>
<td>Quantitative Polymerase Chain Reaction</td>
</tr>
<tr>
<td>RIN</td>
<td>RNA Integrity Number</td>
</tr>
<tr>
<td>RNase</td>
<td>Ribonuclease</td>
</tr>
<tr>
<td>ROS</td>
<td>Reactive Oxygen Species</td>
</tr>
<tr>
<td>rRNA</td>
<td>Ribosomal RNA</td>
</tr>
<tr>
<td>RT</td>
<td>Reverse Transcriptase</td>
</tr>
<tr>
<td>SOD</td>
<td>Superoxide Dismutase</td>
</tr>
<tr>
<td>tRNA</td>
<td>Transfer RNA</td>
</tr>
<tr>
<td>UTR</td>
<td>Untranslated Region</td>
</tr>
</tbody>
</table>
# Table of contents

1. **Introduction** ............................................................................................................. 1
   1.1 The current state of copper-induced responses in aquatic invertebrates .................. 1
   1.2 The MinION nanopore technology ............................................................................ 3
   1.3 Aim and objective .................................................................................................... 5
   1.4 Ethical and environmental aspects .......................................................................... 5

2. **Materials and Methods** .......................................................................................... 9
   2.1 Experimental background ...................................................................................... 9
   2.2 RNA extraction ...................................................................................................... 6
   2.3 Polyadenylated (poly(A)) RNA enrichment and selection .................................... 10
   2.4 Preparation of poly(A) RNA sequencing library ................................................... 7
   2.5 Loading the MinION sequencing device and experimental software settings ........ 7
   2.6 Interpretation and processing of sequence output data ......................................... 8

3. **Results** .................................................................................................................. 8
   3.1 Quantity and quality of total hepatopancreas RNA ................................................. 8
   3.2 Recovery of Poly(A) RNA and sequencing library ................................................ 9
   3.3 RNA sequence data quality control and output ..................................................... 9
   3.4 Interpretation and analysis of sequence data .......................................................... 11

4. **Discussion** ............................................................................................................. 15
   4.1 Sample preparation ................................................................................................. 15
   4.2 Sequence data output ............................................................................................. 19
   4.3 Interpretation and analysis ..................................................................................... 21
   4.4 Conclusion .............................................................................................................. 24

Acknowledgements ......................................................................................................... 25

References ....................................................................................................................... 26
1. Introduction

The freshwater duck mussel *Anodonta anatina* (*A. anatina*), an aquatic invertebrate belonging to the phyla Mollusca, class Bivalvia and the family Unionidae is the most widely distributed bivalve in Sweden. The distribution of *A. anatina* goes all across Sweden except some parts of the inner mountain chains in the north. *A. anatina* are filter feeders and filters approximately up to 50 litres of water per day and this is believed to be a shared trait amongst large freshwater mussels. This subsequently emphasises the preference of eutrophic rather than oligotrophic habitats for large freshwater mussels. Due to the properties of filtering large quantities of water, mussels may continuously experience exposure to a large series of chemical compounds. This subsequently may lead to bioaccumulation of ecotoxilogical pollutants in gill and digestive specific tissue. By adding a stationary lifestyle and inability to regulate body temperature to the equation, these three variables make freshwater mussels such as *A. anatina* excellent sentinels for aquatic ecosystem health research (Negri et al., 2013).

Aquatic ecotoxicology is the study of toxic chemicals and its effects on aquatic biological systems. This field screens and evaluates the toxic effect of manufactured chemicals on organism, population and ecosystem levels. The purpose is to fulfil the following objectives: 1. Minimise threats to human health; 2. Ensure self-sustained ecosystems; and 3. Ensure that particular species do not decline (Depledge & Fossi, 1994). In contrast to direct measurements and screening of chemical levels in aquatic environments, the approach of understanding the effects of these chemicals on biological systems has been subjected to an increased level of interest over the past 30 years. In order to achieve this, particular responses in organisms caused by elevated levels of contaminants needs to be analysed. Examples of responses can be fluctuations in activity of specific enzymes, destabilisation of cellular membranes and changes in heart rate patterns. These responses are defined according to the scientific term of a so-called biomarker. The definition of a biomarker is “a biological response to a chemical or chemicals that gives a measure of exposure and sometimes, also, of toxic effect” (Peakall, 1994). A molecular biomarker is a molecule e.g. proteins that can be used in this regard, either through direct quantity measurements in biological samples or indirect through quantitative gene expression analysis.

The objective of this thesis is how *A. anatina* can be used as an indicator organism to reveal anthropogenic-coupled pollutants in freshwater ecosystems. The selected pollutant for this thesis will be copper (Cu$^{2+}$). It is a widespread contaminant across freshwater ecosystems and it is well understood how it interferes with biochemical, cellular and physiological pathways in invertebrates through its capability of binding to intracellular and membrane bound signalling receptors, thus causes interference with signal processing and gene expression levels (Hebel, Jones, & Depledge, 1997).

1.1 The current state of copper-induced responses in aquatic invertebrates

Copper is a trace element essential for cells to maintain its proper functions, however it is also one of the most toxic heavy metals. A study which investigated the toxicity of dissolved Cu, Zn, Ni and Cd to developing embryos of *Mytilus trossolus* (*M. trossolus*) showed that Cu is by far the most toxic metal (Nadella et al., 2009). One of the recurring traits of copper regarding its mode of action is its ability to catalyse the formation of reactive oxygen species (ROS) (Kayali & Tarhan, 2003). These oxygen species react and oxidises cellular components like proteins and
nucleic acids which leads to inactivation of enzymes, disruption of membranes and mutations which eventually causes cell death (Halliwell & Gutteridge, 1990; Manzl, Enrich, Ebner, Dallinger, & Krumsschnabel, 2004). There are several compounds involved in protecting cells from oxidative damage, a process where the oxidation and production of ROS is inhibited. Some of these defences involves enzymes like superoxide dismutase (SOD), catalase (CAT), glutathione peroxidases (GPx) and glutathione S-transferases (GST). Previous findings illustrate a decrease in SOD, CAT, GPx and GST messenger RNA (mRNA) levels 12 hours after exposure to 10 µg/L Cu^{2+} in the digestive gland of freshwater bivalve Corbicula fluminea (C. fluminea) (Bigot, Minguez, Giambérini, & Rodius, 2010). However, it has also been shown that activities of the antioxidant system do fluctuate with additional variables such as heat stress and life stages. Aquatic invertebrate embryos and larvae appears to be more sensitive to trace metals than adults (Boukadida, Cachot, Clérandeaux, Gourves, & Banni, 2017).

Previous studies on shore crabs Carcinus maenas (C. maenas) revealed changes in heart rate (stimulation of chemoreceptors) at 3 mg/L copper ion exposure (Depledge, 1984). A study on three different aquatic invertebrates Patella vulgata (P. vulgata), C. maenas and Mytilus edulis (M. edulis) which evaluated multiple biomarkers showed a significant decrease in heart rate for M. edulis at 38.5 µg/L Cu^{2+}. Further, a decrease in neutral red retention (NRR) time at 68.1 µg/L Cu^{2+} was obtained in M. edulis which implies reduced lysosomal stability (Viarengo, Burlando, & Bolognesi, 2002; Brown et al., 2004). Metallothioneins (MTs) are non-enzymatic proteins with high metal-binding capacity. The general consideration regarding MTs biological function consists of its role in homeostatic control of essential metals (Cu and Zn), thus acting as metal stores for enzymatic and metabolic demands (Brouwer, Winge, & Gray, 1989). It has been shown in multiple species (annelids, crustaceans, molluscs and fish) that MTs synthesis is induced upon exposure to contaminant trace metals (copper, zinc and cadmium) (Amiard, Amiard-Triquet, Barka, Pellerin, & Rainbow, 2006). These traits have prompted MTs into a core suite of biomarkers in aquatic biomonitoring programmes (Mathiessen, 2000). A study on A. anatina exposed to 10 µg/L Cu^{2+} for 14 days in laboratory conditions showed a significant increase of MTs in both gill and digestive tissue. However, the objective of this study was to evaluate in situ exposure history on molecular responses through comparison of mussels living in polluted versus unpolluted areas. Mussels exposed to long-term pollution were unable to activate MTs responses (Falfushynska, Gnatyshyna, & Stoliar, 2012). The declared conclusion of this was due to exhausting of their defence mechanisms as a result of persistent exposure to contaminants (Falfushynska et al., 2012). Induced genotoxic and cytotoxic responses in unionid mussels has also proved to be reliable indicators when evaluating elevated temperatures due to climate change (Falfushynska et al., 2014). Screening for relationships between pesticide levels, enzymatic activity and abiotic factors such as pH and temperature in Anodonta cygnea (A. cygnea) over the course of one year showed limitations to CAT and acetylcholinesterase (AChE) in their use as biomarkers, whereas GST activity decreased with increases of pesticide levels (Robillard, Beauchamp, & Laulier, 2003). A complementary DNA (cDNA) microarray assay on Mytilus galloprovincialis (M. galloprovincialis) exposed to heat stress (24 °C) and copper (40 µg/L) showed differentially expressed genes involved in translation, heat-shock protein and "microtubule-based movement' protein synthesis (Negri et al., 2013).

Heat-shock proteins (HSPs) are a family of proteins with highly considered protective properties for cells. HSPs serves a purpose of behaving as chaperones for other cellular proteins. Under external environmental stress that induces cellular protein damage like heavy metal exposure,
HSPs protects and restores damaged proteins to their original function. HSPs are a large superfamily of proteins with a broad range of representative HSP isoforms. Several studies emphasise the correlation between induced HSP levels following exposure to a variety of stressors like temperature, heavy metal exposure and pathogenic infections. A relative expression analysis of HSP70 (70 indicates the molecular mass in kDa) in *Mytilus coruscus* (*M. coruscus*) haemocytes showed a significant increase in mRNA levels after 5 days of exposure to 20 µg/L Cu²⁺ (Liu, He, Chi, & Shao, 2014). Similar results as of those mentioned previously were obtained for HSP90 (Liu, Wu, Xu, & He, 2016). Elevated levels of HSP70 has also been revealed in gill tissue of *M. edulis* as a result of exposure to copper (Radlowska & Pempkowiak, 2002), as well as increased concentrations of HSP60 in mantle tissue of *M. edulis* (Sanders, Martin, Nelson, Phelps, & Welch, 1991). Unlike *A. anatina*, the *Mytilus* genus are marine bivalve molluscs and there is an emerging difference between fresh- and seawater organisms regarding copper and its level of toxicity. Cations, such as Na⁺ and Ca²⁺ reduces copper toxicity in marine environments by increasing the competition between biological action sites. Absence of cations in freshwater ecosystems reduces competition of biological action sites, hence resulting in increased levels of copper toxicity (Brooks et al., 2007). Implementation of assays like the one conducted by Negri et al. (2013) on *A. anatina* have great potential in regards to aquatic biomonitoring. However, the functional transcriptomics of *A. anatina* is yet to be fully mapped and evaluated, thus adding constraints to this subject. Previous findings about molluscs emphasises their importance and role as indicator species in aquatic ecotoxicological research. This suggests that profound transcriptomics on *A. anatina* may perhaps provide increased knowledge about states and health in freshwater ecosystems.

This thesis project will be linked to 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. The essential objective will be based on preliminary findings by Ekelund Ugge, Jonsson and Berglund (n.d) where a quantitative stress gene expression assessment on *A. anatina* post-exposure to different concentrations of copper was performed. In total, six genes were subjected to analyses of differentiation in expression levels, including both gill and digestive tissue. Quantitative polymerase chain reaction (qPCR) assessment of MT, GST, HSP70, HSP90, CAT and SOD showed no significant change in expression levels (gill and digestive tissue). Long term goals and future needs mentioned by Ekelund Ugge et al. (n.d) suggests genome and/or transcriptome sequence analysis of *A. anatina* and further use of bioinformatic tools for biomarker panel development.

**1.2 The MinION nanopore technology**

Minion nanopore technology is a portable device that can be used for real-time sequencing of DNA and RNA molecules. The MinION device has proved to generate reads with longer read lengths when compared with next generation sequencing (NGS) technologies. The nanopore technology has the potential of producing nucleotide fragment reads up to 100 kilobases (kb) or more (Laver et al., 2015). Due to its capability of generating long reads, it increases the speed of species identification which further provides a quick and accurate technique for pathogen diagnostics. The portability also enables the minion device to be used in field studies, which further can contribute to the screening and surveillance of disease outbreaks. The nanopore technology is based on membrane bound pores. These kinds of pores are being present in all types of cells and works by transporting molecules in out of the intracellular environment, thus allowing the cells to maintain its proper function. With the nanopore technology, the pores
transport genetic material through its aperture. The nanopores are embedded into a consumable flowcell which is connected to the minion device. The flowcell is the site where the prepared DNA/RNA sample is added. Preparing sequencing strands involves mixing the DNA/RNA strands with copies of a processive enzyme which is also known as a "motor protein", forming a DNA/RNA-enzyme complex. The prepared samples are added to the flowcell and approaches the nanopores aperture. The DNA strand is then pulled through the pore one base at a time which is controlled by the motor protein. As the DNA/RNA strand moves through the pore, the combination of nucleotides and the strand that is being processed creates a disruption in the electrical current that exists through the nanopore. The processed signal of the electrical current can be used to determine the order of nucleotide bases on the strand that is being processed (Patel et al., 2018). An illustration of the technology is shown in Figure 1.

![Figure 1. A simplified illustration of nanopore sequencing technology. The motor protein processes the RNA strand through the aperture of the pore. The translocation of the RNA causes a change in the electrical current across the pore. The electrical signal produces a unique data point for different combinations of nucleotides. The electrical signals are further being processed through specific base call algorithms which translates the electrical current data point to a nucleotide sequence.](image)

The minion nanopore sequencing device can be plugged into any computer with a standard USB 3.0 port. The device generates sequence data by operating with the MinKNOW software tool. The MinKNOW software tool also allows multiple interface control and options for the minion device. The generated raw data from the minion sequencing device is saved as a FAST5-file which must undergo further analysis in order for the data to be converted into a nucleotide sequence. The conversion of a FAST5-file to a nucleotide sequence is a process termed base calling. There are multiple software tools available online that performs the base calling process including Metrichor (Metrichor, n.d), Nanocall (David, Dursi, Yao, Boutros, & Simpson, 2016), Albacore (Oxford Nanopore Technologies, n.d) and Chiron (Teng et al., 2018).
1.3 Aim and objective

The following objectives will be investigated during this thesis project: (1) Implement a method that can be used to obtain total RNAs from female *A. anatina* hepatopancreas; (2) Select and enrich poly(A) RNA that fulfils criteria's in regards to quantity and quality; (3) Perform direct RNA sequencing on native poly(A) RNA using the MinION Nanopore Technology (Oxford Nanopore Technologies); (4) Evaluate its performance; and (5) Compare RNA sequence output data between two conditions, including exposure to 100 µg/L Cu$^{2+}$ and 0 µg/L Cu$^{2+}$.

This project is considered as a pilot study because the analysed samples only contain one individual in each treatment group (n=1). Hence, no significant data will be obtained nor produced.

1.4 Ethical and environmental aspects

*A. anatina* are invertebrates which implies their lack of a central nervous system (CNS) and nociceptors, which is considered to be central aspects in regards to sensing pain and suffering. Another aspect related to pain, involves changes in behavioural patterns as a response to harmful stimuli, which is yet to be proved. The population of *A. anatina* in Sweden is considered to be viable, resident and reproductive (Artdatabanken, 2015). Collecting samples of *A. anatina* for research purposes requires a permission provided by the environmental protection agency or the county administrative board. This project will manage samples that already has been collected.

The results from this project could contribute to increased knowledge in regards to environmental biomonitoring by allowing rapid and efficient identification of harmful pollutions in freshwater ecosystems. Future impacts involve freshwater bivalves and their filter feeding properties. This is an important ecosystem service as they are considered as nature’s own water treatment plant. Hence, these services are crucial in regards to conservation biology and the perseverance of self-sustained ecosystems. The social aspects of this project emphasise public health, where prevention and protection of important freshwater resources are crucial in order to maintain the quality of municipal tap waters. Lastly, to put things in perspective, the objectives and future impacts of this project can be connected to the United Nations sustainable development goals. In total, 17 goals have been formulated and this project can be linked to goal 6: Clean water and sanitation, goal 14: Life below water and goal 15: Biodiversity and ecosystems (United Nations, 2015).

2. Materials and Methods

2.1 Experimental background

The samples that will be used for this thesis project consist of dissected hepatopancreas from *A. anatina*. The bivalves were sampled from Vinne å in the county of Skåne in the southern parts of Sweden. Over the course of one year, water quality measurements were conducted and the levels of copper did not exceed 0.7 µg/L Cu$^{2+}$. The mussels were acclimatised to laboratory conditions for 14 days before experimental treatments. The samples that will be used during this project consists of mussels treated in two separated freshwater tanks, including one case tank with a concentration of 100 µg/L Cu$^{2+}$ and one control tank. Exposure was interrupted after
96 h and hepatopancreas was immediately dissected and stored at -80 °C. The dissected tissue samples have been stored since October 2017. The RNA extraction of hepatopancreas started 2019-03-15 at the University of Skövde, approximately 18 months post dissection. In total, two individual samples of hepatopancreas will be analysed during this project, including one case (100 µg/L Cu²⁺) and one control sample (0 µg/L Cu²⁺) (n=1 in each group).

2.2 RNA extraction

Total RNAs were extracted using Norgens Total RNA Purification Kit (Norgen Biotek Corp, Cat #37500), a column-based extraction kit. A total amount of 80 mg female *A. anatina* hepatopancreas was crushed in liquid nitrogen and homogenised with a 23-gauge syringe needle in a mixture of 10 µl β-mercaptoethanol in 1 mL Buffer RL (Norgen Biotek Corp). The homogenate was passed through the needle ten times. Further, the homogenate was divided into eight separate RNase-free microcentrifuge tubes with a total amount of 10 mg tissue distributed in 600 µl of lysis buffer in each tube. This process was carried out in a cold room with a temperature of 4 °C. The lysate was centrifuged for 2 minutes at 13 400 x g in order to pellet the cell debris. The supernatant was separated from the pellet and an equal volume of 70 % ethanol was added to the lysate and vortexed for 10 seconds. Binding, washing and elution was carried out according to the manufacturer's instructions (Norgen Biotek Corp). The total RNAs was further cleaned-up and concentrated using Norgens RNA Clean-Up and Concentration Kit (Norgen Biotek Corp, Cat #43200). The eluate (eight reactions) from the previous step was pooled into one column supplied with the Norgen RNA Clean-Up and Concentration kit. Once the RNAs were tied to the column, the reaction was treated with DNase I. The enzyme mixture was prepared by adding 15 µl DNase I (Norgen Biotek Corp, Cat #25710) in 100 µl enzyme incubation buffer A (Norgen Biotek Corp, Cat #25710) and mixed by pipetting. The enzyme mixture (115 µl) was added to the column and centrifuged at 13 400 x g for 1 minute. The flowthrough was added back into the column and the reaction was incubated for 15 minutes at 28 °C. Washing and elution was performed according to the manufacturer's instructions (Norgen Biotek Corp). Several tissue input quantities were tested including 10 mg, 20 mg and 40 mg before deciding the final input quantity (80 mg). The quantity of the RNA was recorded by applying Qubit RNA HS Assay Kit (Invitrogen, Cat #Q32852). Confirmation of a successful removal of excessive gDNA was recorded with Qubit 1X dsDNA HS Assay Kit (Invitrogen, Cat #Q33230). An integrity and quality (IQ) control check of the RNA was recorded with Qubit RNA IQ Assay Kit (Invitrogen, Cat #Q33221). NanoDrop ND-1000 Spectrophotometer (NanoDrop Technologies, USA) (absorbance 260/280 and 260/230 nm) measurements were performed in order to confirm the presence or absence of contaminants. RNA was stored in -80 °C until use. This process was repeated for case and control.

2.3 Polyadenylated (poly(A)) RNA enrichment and selection

Poly(A) mRNA was isolated from total RNA using Dynabeads™ mRNA Purification Kit (Invitrogen, Cat #61006). First, the volume of the RNA eluate was adjusted to 100 µl with RNase free DEPC-treated water. The sample was further heated to 65 °C for 2 minutes and immediately placed on ice. The Dynabeads™ magnetic beads were resuspended by vortexing for 10 seconds. The oligo(dT)₂₅ (mg): RNA (µg) ratio was optimised according to the manufacturer’s instructions which suggested a use of 1 mg of oligo(dT)₂₅ per 75 µg of total RNAs. The magnetic beads were prepared together with a binding buffer provided with the kit according to the manufacturer’s instructions (Invitrogen). Total RNA extracted from the two samples (control 90 µg and case
76.5 µg) was separately added to the prepared bead suspensions, mixed thoroughly and incubated on a HulaMixer™ (ThermoFisher Scientific, Cat #15920D) for 5 minutes at room temperature. The mRNA-bead complex was washed and pelleted on a magnet twice. A 10 mM Tris-HCl, pH 7.5 solution was used to elute the mRNA and the sample was heated at 78 °C for 2 minutes in order to detach the beads from the mRNA. The sample was immediately placed on a magnet, pelleted and the eluate was transferred to a new RNase-free microcentrifuge tube, recorded the quantity with Qubit RNA HS Assay Kit (Invitrogen, Cat #Q32852) and stored the mRNA at -80 °C until use. This process was repeated for case and control.

2.4 Preparation of poly(A) RNA sequencing library

The poly(A) RNA sequencing library was prepared using the direct RNA sequencing kit (SQK-RNA002) (Oxford Nanopore Technologies). The first step included ligating a reverse transcriptase (RT)-adapter to the poly(A) RNA strand. Reagents were added to the reaction in following order; NEBNext Quick Ligation Reaction Buffer (New England Biolabs, Cat #B6058S), poly(A) RNA (486 ng from the control sample and 320 ng from the case sample), RNA CS (Oxford Nanopore Technologies), RT-adapter (RTA) (Oxford Nanopore Technologies) and T4 DNA ligase (New England Biolabs, Cat #M0202S). All reagents were added according to the instructions provided with the direct RNA sequencing kit (SQK-RNA002) in terms of reagent volumes (Oxford Nanopore Technologies). The reaction was incubated for 10 minutes at room temperature. This process was repeated for each sample (control and case) respectively. The optional reverse transcription (RT) step was skipped and the RT mastermix + enzyme mixture was replaced with 25 µl nuclease-free DEPC-treated water. Cleaning and removing reagent residues from the RT-adapter ligation reaction was done using Mag-Bind® TotalPure NGS (Omega Bio-Tek, Cat #M1378-00) according to the instructions in direct RNA sequencing kit (SQK-RNA002) (Oxford Nanopore Technologies). Further, a second RNA adapter was ligated to the poly(A)-RT adapter complex. Reagents were added to the reaction in the following order; Poly(A) RNA (486 ng from the control sample and 320 ng from the case sample), NEBNext Quick Ligation Reaction Buffer (New England Biolabs, Cat #B6058S), RNA adapter (RMX) (Oxford Nanopore Technologies), Nuclease-free water, T4 DNA ligase (New England Biolabs, Cat #M0202S) and incubated for 10 minutes at room temperature. All reagents were added according to the instructions provided with the direct RNA sequencing kit (SQK-RNA002) in terms of reagent volumes (Oxford Nanopore Technologies). This process was repeated for each sample (control and case) respectively. A second cleaning was performed with Mag-Bind® TotalPure NGS (Omega Bio-Tek, Cat #M1378-00). The quantity of the purified poly(A) RNA was recorded with Qubit RNA HS Assay Kit (Invitrogen, Cat #Q32852) with a recovery aim of 200 ng in 20 µl of eluate. Samples were kept on ice until use.

2.5 Loading the MinION sequencing device and experimental software settings

RNA sequencing on the MinION platform was performed using Oxford Nanopore Technologies R9.4 flowcell (FLO-MIN106) and the prepared poly(A) RNA library was loaded by carefully following the instructions in the direct RNA sequencing protocol (SQK-RNA002). In between each sequencing run, the R9.4 flowcell was washed using a flowcell wash kit (EXP-WSH002) (Oxford Nanopore Technologies). The flowcell was stored at 4 °C in between the two runs. The standard MinKNOW protocol script was used for both case and control, with one exception. The run time was changed to 18 hours for the control sample (NC_18Hr_sequencing_FLO-MIN106
2.6 Interpretation and processing of sequence output data

A quality control (QC) report was produced with PycoQC (version 2.2). The fastq-files were obtained using the automated base call algorithm in the MinKNOW software. The generated fastq-files were compressed using Winzip (version 23.0) (WinZip Computing). The control and case fastq-files were compressed separately. Once the files had been fused, an alignment was done using NanoPipe, a web-based tool for analysis of nanopore sequencing datasets (Shabardina et al., 2019). A genome assembly of the freshwater mussel *Venustaconcha ellipsiformis* (*V. ellipsiformis*) was used as reference (GenBank assembly accession: GCA_003401595.1). The reference *V. ellipsiformis* file was uploaded as the target sequence and the generated nanopore sequence data was uploaded as query. Two separate runs were conducted, including one for each respective dataset (control and case). The pipeline was done using default parameters. The pipeline generated several target consensus regions which were used to run a basic local alignment tool nucleotide (BLASTn) analysis available at the National Centre for Biotechnology Information (NCBI). The BLASTn parameters were set by default except selecting "somewhat similar sequences" (blastn) in the program selection tab.

Furthermore, a BLASTn analysis was done using eight randomly selected fastq-files from the control sequencing run in order to confirm RNA species identification. The BLASTn analysis was also used to compare the number of identities for each sequence alignment with the generated quality scores. The BLASTn parameters were set by default except using "somewhat similar sequences" (blastn) in the program selection tab and further narrowing down the organism search tab level to molluscs (taxid:6447).

3. Results

3.1 Quantity and quality of total hepatopancreas RNA

A total amount of 80 mg of female *A. anatina* hepatopancreas was used to obtain total RNAs through a two-step RNA extraction procedure. The first step included direct total RNA extraction from tissue followed up by pooling, clean-up and DNase I treatment in the second step. The subsequent extraction was done in order to concentrate the amount of total RNAs due to boundaries of the tissue extraction kit having a maximum tissue input amount of 10 mg/column. Further, IQ analysis on the total tissue RNA extracts without any additionally treatments recorded IQ values ≤5. Including DNase I treatment did increase IQ, which ascertains that an excessive amount of genomic DNA present in the sample causes interference with the IQ assay. Further, an 260/280 absorbance ratio closer to ~2 confirms the absence of genomic DNA, thus concluding a successful DNase I treatment. Table 1 summarises the overall quantity and quality controls that were conducted. The primary difference between the control and case sample is illustrated by the IQ, where 68% (IQ 6.8) of the total RNA is preserved in control whereas case having a 76% (IQ 7.6) recovery of preserved RNAs. The absorbance ratios are close to identical with a 0.01 difference at 260/280 nm and 260/230 nm respectively. In regards to quantity (µg), the control sample had a ~15% (90 µg control and 76.5 µg case) higher quantity percentage yield compared to the case sample (Table1).
Table 1. The samples include one control 0 µg/L Cu\(^{2+}\) and one case exposed to 100 µg/L Cu\(^{2+}\) for 96 h in aquatic tanks during laboratory-controlled conditions (n=1 in each group). Quantity (µg) of the total RNA extracted from female A. anatina hepatopancreas was recorded with Qubit RNA HS Assay Kit. Quality is illustrated with an IQ number which was obtained with Qubit RNA IQ Assay Kit. Absorbance values were obtained with NanoDrop ND-1000 Spectrophotometer at wavelength 260/280 nm and 260/230 nm respectively.

<table>
<thead>
<tr>
<th>Sample ID</th>
<th>Quantity (µg)</th>
<th>Quality (IQ)</th>
<th>Absorbance 260/280 (nm)</th>
<th>Absorbance 260/230 (nm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control</td>
<td>90.0</td>
<td>6.8</td>
<td>2.12</td>
<td>2.26</td>
</tr>
<tr>
<td>Case</td>
<td>76.5</td>
<td>7.6</td>
<td>2.13</td>
<td>2.25</td>
</tr>
</tbody>
</table>

3.2 Recovery of Poly(A) RNA and sequencing library

Selection and enrichment of poly(A) RNA was performed using Dynabeads oligo (dT)\(_{25}\). From a total amount of 90 µg total RNAs (control, Table 1), 486 ng poly(A) RNA was obtained, with a recovery percentage rate of 0.5%. Further, 320 ng poly(A) RNA was obtained from a total amount of 76.5 µg total RNAs (case, Table 1), with a recovery percentage rate of 0.4%. Recording the quality of the obtained poly(A) RNA with Qubit RNA IQ Assay Kit was not possible to perform due to the low volumes of eluate (10 µl). In this case, having a sample concentration ranging between 32-48 ng/µl would have required a total input volume of 20 µl according to the manufacturer's product user guide (Invitrogen, n.d) in order for the assay to quantify and produce reliable data.

The poly(A) RNA was prepared for sequencing by ligating a RT-adapter to the poly(A) tail followed up by ligating a second RNA sequencing adapter (RMX). In between each of the ligation steps, a magnetic bead-based cleaning process with Mag-Bind TotalPure NGS was conducted in order to remove any reagent residues. According to the manufacturer’s protocol, having 500 ng as starting material should result in a recovery of 200 ng after completing the preparations of the RNA sequence library, thus implying a 60% loss over the course of this procedure. The poly(A) RNA input for control was 486 ng and 320 ng for case respectively. After ligation and cleaning with Mag-Bind TotalPure NGS, the amount of recovered poly(A) RNA was 318 ng for control with a percentage loss of 35% and 242 ng was obtained for case with a percentage loss of 24%. Hence, the recovery of the poly(A) RNA sequence library was higher compared to the expected outcome. It was decided to use all of the obtained material from both samples in order maximise the sequence output rather than equalising the amount of input between the two groups.

3.3 RNA sequence data quality control and output

A quality control (QC) report which summarises the overall performance of the sequencing runs was produced with PycoQC (version 2.2). Table 2 shows the results from each respective run, including the number of reads, bases, read length, read quality, active channels and run duration. In total, the two sequencing runs produced 239,448 reads whereas 179,651 (75%) of the reads were obtained during the first run (control) and 59,797 (25%) of the reads were obtained in the second run (case). In regards to the number of nucleotides, a total of 208,362,167 bases were sequenced with 164,022,191 (79%) of the bases being obtained in the first sequencing run (control) and 44,339,976 (21%) of the bases were obtained during the second run (case). The median read lengths were 116 nucleotides (nts) longer in the first sequencing run (control)
compared to the second run (case). The median read quality exhibited a 0.12 decrease in the second run (case). According to the manufacturer’s product details, the MinION nanopore sequencing device has a maximum number of 512 channels available for sequencing (Oxford Nanopore Technologies, n.d). At the start of the first sequencing run (control), 505 (99%) of the channels were active, whilst at the start of the second run (case), only 284 (55%) of the channels were active. Figure 2a illustrates the channel activity over time for the first sequencing run (control), where the number of active channels is high during the first 2 hours of the experiment. However, the channel count showed a continuous decrease over time after the 2-hour timepoint. Figure 2b illustrates the channel activity over time for the second sequencing run (case) and it shows a substantial decrease in number of active channels at the start of the experiment when compared to figure 2a. However, figure 2b shows similar tendencies as figure 2a in regards to the reduction of active channels over time. The run time of the first sequencing run (control) was set to 18 hours. However, the run was completed after 15.33 hours. The run time of the second sequencing run (case) was set to 36 hours but was manually cancelled after 16.86 hours.

In summary, the overall performance of the two runs, respectively, illustrates that the first run (control) performed better over the second run (case) in terms of quality and sequence output.

Table 2. QC report showing the overall results of the sequencing run for each respective sample. Values generated and obtained by PycoQC (version 2.2).

<table>
<thead>
<tr>
<th>Sample ID</th>
<th>Reads</th>
<th>Bases</th>
<th>Read length (Median)</th>
<th>Read Quality (Median)</th>
<th>Active Channels</th>
<th>Run Duration (h)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control</td>
<td>179,651</td>
<td>164,022,191</td>
<td>650.00</td>
<td>9.13</td>
<td>505</td>
<td>15.33</td>
</tr>
<tr>
<td>Case</td>
<td>59,797</td>
<td>44,339,976</td>
<td>534.00</td>
<td>9.01</td>
<td>284</td>
<td>16.86</td>
</tr>
</tbody>
</table>
Figure 2. Nanopore channel activity over time including the first sequencing run (control) (A) and the second sequencing run (case) (B). Experiment time (h) on y-axis and channel ID on x-axis. (A) shows high channel activity (dark red areas) during the first 2 hours of the first sequencing run followed by a continuous decrease in channel activities after the 2-hour timepoint (yellow/white areas). (B) has low channel activities at the start of the sequencing run (white areas) which is continuously decreased over the course of the sequencing run, thus resulting in almost no channel activities at the end of the second run.

3.4 Interpretation and analysis of sequence data

The data was analysed using NanoPipe, a web-based tool for analysis of nanopore sequence data (Shabardina et al., 2019). Two runs were conducted, one for each data set produced from the two nanopore sequencing runs (control and case). A genome assembly of the freshwater mussel *V. ellipsoidalis* was used as reference due to absence of an available reference genome/transcriptome for *A. anatina*. With 179,651 reads present in the query from the first sequencing run (control), only 852 reads did map back to the reference genome which corresponds to 0.5% of the query reads generated from the first sequencing run (control).
Furthermore, 59,797 reads were present in the query from the second sequencing run (case) and 202 of the reads mapped back to the reference which corresponds to 0.3% of the total reads generated from the second sequencing run (case). Furthermore, out of the 852 mapped reads from the first sequencing run (control), 9 (1%) of the mapped reads targeted and produced an aligned consensus region. Out of the 202 mapped reads from the second sequencing run (case), 4 (2%) of the mapped reads targeted and generated an aligned consensus region (Table 3).

Table 3. Results from alignments generated by NanoPipe, including sample ID, reference species, median read length, number of query reads, number of mapped reads and number of targets identified.

<table>
<thead>
<tr>
<th>Sample ID</th>
<th>Reference</th>
<th>Read length (median)</th>
<th># Reads in query</th>
<th># Reads mapped</th>
<th># Targets identified</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control</td>
<td>V. ellipsiformis</td>
<td>650.00</td>
<td>179,651</td>
<td>852</td>
<td>9</td>
</tr>
<tr>
<td>Case</td>
<td>V. ellipsiformis</td>
<td>534.00</td>
<td>59,797</td>
<td>202</td>
<td>4</td>
</tr>
</tbody>
</table>

Table 4 shows the analysis of the nine targets listed in Table 3 which produced an aligned consensus region from the control sample sequence data within the *V. ellipsiformis* genome. Furthermore, Table 4 lists a number showing how many times each specific consensus region was found within the reference. The consensus region listed with the highest number of reads was mapped 134 times, whilst the consensus region with the lowest number of reads was mapped 13 times. Interestingly, the four consensus regions with the highest number of mapped reads in the first run (control) were identical to the four consensus regions found during the second run (case) (Table 3 & 4). These consensus regions also had a similar distribution in regards to the number of mapped reads, except QKMX01244513.1 which was mapped 24 times and QKMX01361759.1 which was mapped 20 times (Table 4). The length of the aligned consensus regions varied from 26 to 123 nucleotides with a percentage of identity matches ranging from 76% (70/92 nts) to 92% (24/26 nts) (Table 4). The BLASTn analysis of the nine consensus regions identified 2 out of 9 species within the Mollusca phyla, one of which had the highest number of reads mapped to the reference (134) (Table 4). Furthermore, 7 out of 9 consensus regions aligned to aquatic species in the BLASTn analysis. An mRNA molecule was found in 6 out of 9 cases, whereas 6 out of 6 mRNAs were identified as ribosomal proteins. Lastly, the consensus regions had an mRNA length coverage ranging from 10% (lowest) to 18% (highest) (Table 4).
Table 4. Identified target IDs and number of reads mapped by NanoPipe, followed by the BLASTn results of each respective target consensus region including species ID, # of identities to query target, mRNA name and length coverage of mRNA. Not available (n/a) meaning no mRNA molecule was targeted nor identified. Four of the targets were identified in both control and case sequence data (*) and the five remaining targets (no star) were identified in the control sequence data.

<table>
<thead>
<tr>
<th>Identified targets (ID)</th>
<th># Reads mapped</th>
<th>Species</th>
<th># Identities</th>
<th>mRNA</th>
<th>mRNA length coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>*QKMX01353604.1</td>
<td>134</td>
<td><em>Crassostrea gigas</em></td>
<td>70/92 (76%)</td>
<td>60S ribosomal protein</td>
<td>92/851 (11%)</td>
</tr>
<tr>
<td>*QKMX01361759.1</td>
<td>88</td>
<td><em>Penaeus vannamei</em></td>
<td>43/51 (84%)</td>
<td>40S ribosomal protein</td>
<td>51/491 (10%)</td>
</tr>
<tr>
<td>*QKMX01244513.1</td>
<td>71</td>
<td><em>Arenicola marina</em></td>
<td>93/123 (76%)</td>
<td>rpL19 ribosomal protein</td>
<td>123/678 (18%)</td>
</tr>
<tr>
<td>*QKMX01348912.1</td>
<td>46</td>
<td><em>Coptotermes formosanus</em></td>
<td>52/59 (88%)</td>
<td>CFSNI2820 ribosomal protein</td>
<td>59/564 (10%)</td>
</tr>
<tr>
<td>QKMX01074369.1</td>
<td>38</td>
<td><em>Haliotis discus</em></td>
<td>78/98 (80%)</td>
<td>L3 ribosomal protein</td>
<td>98/658 (15%)</td>
</tr>
<tr>
<td>QKMX01350935.1</td>
<td>29</td>
<td><em>Lateolabrax maculatus</em></td>
<td>29/32 (91%)</td>
<td>n/a</td>
<td>n/a</td>
</tr>
<tr>
<td>QKMX01297435.1</td>
<td>17</td>
<td><em>Talaromyces funiculosus</em></td>
<td>24/26 (92%)</td>
<td>n/a</td>
<td>n/a</td>
</tr>
<tr>
<td>QKMX01347782.1</td>
<td>16</td>
<td><em>Anabas testudineus</em></td>
<td>48/57 (84%)</td>
<td>n/a</td>
<td>n/a</td>
</tr>
<tr>
<td>QKMX01357269.1</td>
<td>13</td>
<td><em>Lingula anatina</em></td>
<td>63/74 (85%)</td>
<td>40S ribosomal protein</td>
<td>74/566 (13%)</td>
</tr>
</tbody>
</table>

Cells marked with * represents identified target consensus regions in both control and case.

A BLASTn analysis was conducted on eight randomly selected fastq-files from the control sequencing run in order to confirm RNA species identification as well as comparing the number of nucleotide identity matches with the generated quality score (9.13). Table 5 shows a summary of the BLASTn sequence alignments with the highest number of identities in each fastq-file respectively. The findings should be interpreted as observations rather than qualitative results, since the analysis was restricted to search for homologs within the Mollusca phylum (taxid:6447). Out of the eight respective BLASTn searches, every hit with the highest number of identities did correspond to species within the freshwater mussel family Unionidae. In this case, four freshwater mussel species were identified, where Cristaria plicata (*C. plicata*) shares its closest common ancestor with *A. anatina* (Wen et al., 2017). The number of identities varied.
from 83% (1174/1414 nts) to 89% (853/963 nts) (Table 5). Furthermore, 100% (8/8) of the alignments identified mRNA molecules including two heat shock protein isoforms (HSP70 and HSP90), β-actin, elongation factor 1-α, cathepsin B, catalase, Y-box protein and α-amylase. The coverage of the total mRNA lengths varied from 60% (1505/2482 nts) to 100% (1333/1339 nts) (Table 5).

Table 5. Alignments with the highest number of identities generated by BLASTn using eight randomly selected fastq-files from the control sequencing run. Listed with BLASTn ID, species, # identities, mRNA name and mRNA length coverage.

<table>
<thead>
<tr>
<th>BLASTn ID</th>
<th>Species</th>
<th># Identities</th>
<th>mRNA</th>
<th>mRNA length coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>FFD2RN4Y015</td>
<td>Sinanodonta woodiana</td>
<td>1995/2325 (84%)</td>
<td>Heat shock protein 70 (HSP70)</td>
<td>2325/2305 (100%)</td>
</tr>
<tr>
<td>FFD37154014</td>
<td>Sinanodonta woodiana</td>
<td>1156/1333 (87%)</td>
<td>β-actin</td>
<td>1333/1339 (100%)</td>
</tr>
<tr>
<td>FFD3XURE014</td>
<td>Fusconaia flava</td>
<td>853/963 (89%)</td>
<td>Elongation factor 1-α</td>
<td>963/1113 (87%)</td>
</tr>
<tr>
<td>FFD4V8H014</td>
<td>Cristaria plicata</td>
<td>1333/1536 (87%)</td>
<td>Cathepsin B</td>
<td>1536/1825 (84%)</td>
</tr>
<tr>
<td>FFD5EKXJ015</td>
<td>Cristaria plicata</td>
<td>1258/1505 (84%)</td>
<td>Catalase (CAT)</td>
<td>1505/2482 (60%)</td>
</tr>
<tr>
<td>FFD6USMJ01R</td>
<td>Cristaria plicata</td>
<td>1669/1960 (85%)</td>
<td>Heat shock protein 90 (HSP90)</td>
<td>1960/2674 (73%)</td>
</tr>
<tr>
<td>FFD7RBEM01R</td>
<td>Hyriopsis cumingii</td>
<td>1174/1414 (83%)</td>
<td>Y-box protein</td>
<td>1414/1427 (99%)</td>
</tr>
<tr>
<td>FFD1UHD6015</td>
<td>Hyriopsis cumingii</td>
<td>1416/1687 (84%)</td>
<td>α-amylase</td>
<td>1687/1701 (99%)</td>
</tr>
</tbody>
</table>
4. Discussion

Prior to all sequencing experiments, it is important to plan and specify its goals and further assess its feasibility in regards to budget and availability of methods. In this case, the goal was to perform RNA-sequencing on *A. anatina* exposed to copper with the intention of screening for differences in transcript abundances between two treatment groups. Further, the final goal was to identify suitable biomarkers based of the RNA-sequence data that potentially could be used to assess and track freshwater pollution related to copper. The objectives of this project were restricted to a budget of 10,000 SEK and use of a predetermined sequencing technology (nanopore sequencing). Currently, the absence of available data and research which utilises nanopore sequencing in environmental biomonitoring do add constraints to the discussion part of this assay. Instead, this section will mainly be focusing on discussing future advances and alternatives that can be used to increase the significance of the results.

In general, when comparing differences between treatment groups in RNA-sequence analyses, it is crucial to keep the variance between groups as low as possible. These variables may include feeding patterns, temperature, life stages and gender specificity (Wolf, 2013). This will further prevent possible obscuration's in the differences of interest. For this study, implications in the treatment variations appeared and involved the concentrations of copper in the treatment tanks. It was found out at a later point that the sand present in the tanks caused interference in the treatment conditions by absorbing a substantial amount of the copper ions (Ekelund Ugge et al., n.d). This did result in reduced concentrations of copper which may have caused alterations and thus affecting the desired outcome. Another crucial aspect of RNA-sequencing experiments involves the number of biological replicates. There is still an ongoing debate on the optimal number of biological replicates needed in RNA-sequence experiments. However, the inclusion of biological replicates does improve the estimation of the inter-individual variation. Currently, most RNA-sequencing experiments contain at least three biological replicates (ENCODE, 2016). Due to time and consumable constraints during this project, only one biological replicate was used in each treatment group (n=1).

4.1 Sample preparation

This project was provided with *A. anatina* hepatopancreas tissue from a study that was conducted in October 2017. From this date, the tissue had been stored at -80 °C until use in early March 2019. Storing tissue for later use, especially when it comes to RNA, requires the tissue to be stored in a proper way. Storing requires the tissues to be flash frozen in liquid nitrogen and optimally stored together with an RNA stabilisation reagent e.g. RNAprotect or RNAlater. This will allow the RNA to stabilise and maintain its integrity for later use. One important point to highlight as well involves the timepoints between treatment and RNA-sequencing. Usually, the majority of research on molecular tissue profiles are being conducted directly post treatment by dissecting fresh tissue, followed by immediate RNA extraction.

The first step in the sample preparations included total RNA extraction obtained from hepatopancreas. This step has proved to be the most challenging and crucial part in terms of sample preparations prior to sequencing. The reason for this is hepatopancreas and its properties. Hepatopancreas is rich in endogenous ribonucleases (RNases) (Beintema, Campagne, & Gruber, 1973), a type of enzyme that catalyses the degradation of RNA into smaller fragments. Hepatopancreas does also have autolytic properties which refers to the cells capability to self-
destruct through the action of its own enzymes (Beintema et al., 1973). The richness of RNases in hepatopancreatic tissues makes RNA extraction a challenge. RNases are very thermostable, and it is very hard to inhibit their activities. Not only are they present in the endogenous environment, they are also present in the surrounding environment. For instance, RNase 7, a member of the superfamily RNase A is secreted by human skin cells, serving as a defence against pathogens (Harder & Schröder, 2002). Due to the presence of RNases, RNA extractions require strictly sterile working areas treated with specified RNase decontamination reagents e.g. RNase AWAY. In order to inhibit RNase activity during extraction, β-mercaptoethanol was included in the lysis buffer for this experiment. For RNA extraction, the total RNA purification kit from Norgen was used. This kit is column-based which makes it very easy to use and follow. However, a downside with this kit is the capacity of input quantities. The quantities are restricted to a total amount of 10 mg tissue per column. Multiple input quantities were tested in order to find the most suitable number. The benchmark required 50 µg of total RNAs. This did require use of eight column reactions in total (see section 3.1) in order to obtain the desired amount of RNA. Managing eight reactions at a time does not only imply a risk of errors arising, it is also not optimal in terms of time efficiency. Therefore, it might be relevant to sought for alternative extraction techniques in order to increase the efficiency and further ensure that the techniques being used does allow the user to obtain high-quality intact RNAs. One of the most commonly used reagents for RNA extraction from tissue is TRlZol (Meng et al., 2013; Chen et al., 2015; Sun, Li, Liu, Lee, & Wang, 2016; Workman et al., 2018). This reagent enables the user to adapt quantities of tissue inputs in to a single reaction, which subsequently increases efficiency in terms of time savings. The method would had also excluded the additional pooling step conducted during this project, which further would have allowed treatment with DNase I on the total RNAs obtained directly from hepatopancreas. Previously, a study which aimed to optimise the use of TRlZol reagent in rat pancreas showed that the time period for collecting tissue on RNA extraction is crucial when it comes to maintaining the integrity of high-quality RNA (Li et al., 2009).

For quality control of the RNA samples, Qubit RNA IQ assay was used. This is a kit which produces an IQ number based of the ratios between intact vs. degraded RNAs present in the sample. The technique utilises two different reagent dyes, one that binds to large and highly structured RNA molecules and one that binds to small and degraded RNA molecules. Together, this produces a number between 1-10, with 10 corresponding to 100% of large and highly structured RNA and 1 corresponding to 10% of large and intact RNA molecules present in the sample. This kit is very time and cost effective. An integrity and quality number can be obtained within 15 minutes. However, quality control checks in RNA-sequencing experiments has proved to be crucial in regards to obtaining information about the quality of the RNA molecules. Even though, the IQ is cost and time effective, the information it provides is scarce. The general guidelines for assessing the quality of RNA involves using micro-capillary electrophoresis (Künstner et al., 2010; Laver et al., 2015; Seki et al., 2019). This technique allows the user to obtain profound information about the molecule of interest by showing the distribution of fragment sizes present in the sample. Micro-capillary electrophoresis can be conducted on a Bioanalyzer by Agilent with specified kits depending on what type of molecule is to be examined. It also produces an RNA integrity number (RIN) similar to the IQ. However, there appears to be a slight difference between RIN and IQ. An analysis conducted by the supplier of the RNA IQ assay showed differences in quality scores between RIN (#8.9) and IQ (#9.8) (ThermoFisher Scientific, n.d). The difference was due to the signal produced from 5s ribosomal RNA (rRNA) and transfer
RNA (tRNA) which exhibit tertiary structure and bind the large RNA dye in the IQ assay (ThermoFisher Scientific, n.d). Hence, the IQ scores generated in this project might exhibit a lower RIN. Quality assays with bioanalyzer were not performed due to lack of suitable kits for RNA analysis. Furthermore, there seem to be a correlation between quality scores and the amount of tissue being used for RNA extraction. The tissue inputs used during this project ranged between 10-80 mg. Interestingly, the IQ value did exhibit decreases with increases of tissue quantities. Using 10 mg of tissue did result in IQ-scores >9 after DNase I treatment, whereas the IQ-score decreased to 6.8-7.6 when 80 mg of tissue was used. These observations were confirmed by Li et al. (2009) showing that an increased amount of pancreatic tissues results in decreased integrity of high-quality RNAs. This issue is probably due to the β-mercaptoethanol not being able to completely destroy and inhibit the activities of RNases. A final point to emphasise involves DNase I and how it affects the quality score. A quality assay on the RNA samples without DNase I treatment generated quality scores ≤5. However, including DNase I treatment during the RNA extraction process did result in an increase of the quality scores. Thus, this concludes that an excessive amount of DNA present in the RNA sample causes major interference with the RNA IQ assay.

NanoDrop ND-1000 Spectrophotometer was used to confirm absence or presence of contaminants in the samples, e.g. proteins, carbohydrates or phenols. A 260/280 absorbance ratio ≥2 is generally accepted for RNA to be considered free from contaminants (Table 1). The 260/230 absorbance ratios for pure nucleic acid samples are commonly higher than the 260/280 values. Expected 260/230 values are usually in the range of 2.0-2.2 (Table 1) (ThermoFisher Scientific, n.d). 260/230 ratios <2 indicates presence of contaminants with absorbance at 230 nm, e.g. carbohydrates or phenols (ThermoFisher Scientific, n.d). This technology can also be used for sample quantification but does not discriminate between DNA and RNA, so instead Qubit RNA HS assay was used for sample quantification because this kit only quantifies RNA molecules.

Measuring and quantifying transcript levels in cells and tissue extracts by RNA-sequencing requires the user to select and enrich transcriptional RNA (mRNA). These molecules are gene coding products present in the intermediate step before generating a functional protein. This process is referred to as the "Central Dogma of Molecular Biology", by which information within a gene is transcribed and further translated to a protein (Crick, 1970). However, obtaining mRNA from total RNA can be difficult to achieve since the majority of RNA present in cells consists of noncoding RNA species (rRNA and tRNA). Gene coding RNAs is believed to cover approximately 1-5% of the cells total RNA content (Westermann, Gorski, & Vogel, 2012). In order to select and enrich mRNA from total RNA, unique traits within the mRNA molecule must be sought. In contrast to other RNA species, mRNAs are subjected to post-transcriptional modifications. This includes polyadenylation, a process in which multiple sequences of adenines are applied to the 3´ end of the mRNA molecule and this is commonly referred to as the poly(A) tail. These traits enable the user to select for mRNA with specialised oligo-dT magnetic beads. This technique revolves around an oligonucleotide sequence consisting of thymine attached to a magnetic bead. Further, it utilises the base pair process of DNAs in which adenine binds to thymine. By using a magnet, the user can pellet and select for pure mRNAs. Alternative methods for mRNA selection include rRNA depletion, a technique which removes rRNA from the total RNA sample. It relies on the hybridisation of specific probes to the rRNA molecule, which further can be removed from the sample (O'Neil, Glowatz, & Schlumpeberger, 2013). However, this
The mRNA enrichment was expected to have an average percentage recovery rate of ~1-5% (Westermann et al., 2012), and based on the total RNA quantities (90 µg control and 76.5 µg case), the mRNA quantities were expected to be within a range of 900-4 500 ng (1-5% of 90 µg control) and 765-3 825 ng (1-5% of 76.5 µg case) for each sample respectively (Table 1.). Unfortunately, the recovery only reached 0.5% (486 ng control) and 0.4% (320 ng case). The major concern for this involves the quality controls. Due to the low quantities of mRNAs, no quality controls were conducted since the RNA IQ assay required to high sample volumes for it to be able to record any quality scores. In this case, a quality check required 20 µl of sample for analysis and the total sample volumes were 10 µl. A reason for such low percentage recoveries of the mRNAs could have something to do with the quality scores listed in Table 1 for the total RNAs. By interpreting the quality scores, they tell that 32% (control) and 24% (case) of the total RNAs respectively is considered small and degraded based on the IQ technology (Table 1). In addition, this does not preclude the chance of the 1-5% of mRNAs present in total RNAs to be included within the percentages of low quality RNAs prior to mRNA selection. If this is the case, having degraded mRNAs might introduce biases towards the 3’ end of the mRNA strand as polyadenylation and oligo-dT selection occurs at the 3’ end. Thus, if degraded mRNAs enriched through 3’ end selection are to be sequenced, there will be an extensive overrepresentation of sequencing reads from the 3’ end over reads from the 5’ end of the mRNA strands (Künstner et al., 2010). This issue has previously been discussed by Garalde et al. (2018) which emphasise that degraded mRNAs may hinder proper detection of splice variants because the sample mRNA is no longer full length. One way to overcome the troubles of having degraded mRNAs described by Garalde et al. (2018) would be to implement a method that targets the 5’ end of eukaryotic mRNAs. Including such a step in the sample preparation steps prior to sequencing should further improve read-length distribution of direct poly(A) RNA reads (Garalde et al., 2018).

The sequencing library preparations were performed according to the direct RNA-sequencing protocol provided from Oxford Nanopore Technologies with two exceptions. First, the optional cDNA synthesis step was excluded. Secondly, Agencourt RNAClean XP beads was replaced with Mag-Bind TotalPure NGS. The required sample input prior to library construction was 500 ng of poly(A) RNA. In this case, these quantities were 486 ng (control) and 320 ng (case) respectively. Further, the protocol required an average fragment size of >500 base pairs (bp). The average fragment size of the poly(A) RNA was unknown since no micro-capillary electrophoresis was conducted. Synthesis of cDNA was excluded since cDNA strands are amplified by polymerase chain reaction (PCR) and this can introduce bias in sequencing coverage such as secondary strand synthesis, random formation of primer dimers and dropout of some RNA species which further can inflict relative transcript abundances. Thus, this reduces the complexity of the sequencing library (Kozarewa et al, 2009). The Agencourt RNAClean XP beads were replaced due to time and financial constraints. According to the manufacturer’s guidelines for sequencing library preparations, the recovery of quantities post library preparation was expected to be 40%, thus implying a 60% loss of the total input quantities. After the library had been constructed, the quantities were recorded and showed a percentage loss of 35% (486 ng to 318 ng control) and 24% (320 ng to 242 ng case) for each respective sample preparation. Thus, the recovery of the RNA-seq library did compensate for the lower input quantities prior to library construction. This do confirm that the Mag-Bind TotalPure NGS beads performs well in terms of
quantity recoveries. However, no quality checks were performed on the final RNA-sequencing libraries. Hence, this did exclude evaluation of its ability to remove sample residues. Furthermore, uncertainties remain regarding its compatibility with nanopore sequencing due to absence of publicly available research where alterations in the bead cleaning process has been conducted. Reasons why the protocol suggests to use Agencourt RNAClean XP beads has not been further investigated. One assumption could be that this revolves around business affairs rather than compatible product properties. However, this issue remains unresolved.

According to the general guidelines for RNA-sequence experiments provided from ENCODE, quality and quantity assays should be performed after: (1) Total RNA extraction; (2) Oligo-dT selection; and (3) Library construction (ENCODE, 2016). Quality assays are preferably conducted on a Bioanalyzer. Quantification was done after each of these steps during this experiment. However, the quality was only assessed after total RNA extraction. In addition, no quality assays were conducted on a bioanalyzer, thus this did result in loss of important information about fragment size distributions.

In summary, the sample preparation procedures prior to sequencing are crucial in terms of acquiring eligible data for analysis. In this section, the process of sample storage, RNA extraction, poly(A) selection and library construction has been evaluated. The findings emphasise key factors which needs to be improved such as securing the integrity of RNAs when storing samples for later use, obtaining profound information about fragment size distributions and lastly how time efficiencies and input quantities play a key role in terms of securing and obtaining high-quality RNAs for RNA-sequencing.

4.2 Sequence data output

Poly(A) RNA sequencing was performed using the MinION nanopore technology (Oxford Nanopore Technologies). The first run included the prepared control sample and the second run included the prepared case sample (Table 2). There was a substantial difference in sequencing output between the two runs, where the first run (control) had a higher read output when compared to the second run (case). The input quantities were 318 ng of poly(A) RNA in the first run (control) and 242 ng in the second run (case). The median read lengths were 650 nts in the first run (control) and 534 nts in the second run (case). These read lengths are longer when compared to other NGS platforms. The fragment size criteria for nanopore sequencing was >500 nts, which implies that a vast majority of the reads were close to the borderline of not being able to be detected by the nanopore sequencer. The median read quality ranged between 9.13 in the first run (control) to 9.01 in the second run (case) which corresponds to a base call accuracy <90%. According to Oxford Nanopore Technologies, the current standards for nanopore sequencing generates a quality score of 10 with a base call accuracy of 90%. The nanopore quality scores can be compared to Illumina HiSeq 2000 which generates quality scores ≥30, with a base call accuracy of 99.9% (Illumina, 2012). This proves that nanopore quality scores are error-prone in terms of accurate base call identification, which requires prerequisites and guidance for error corrections (Depledge et al., 2019; Goodwin et al., 2019). The number of active channels were 505 at the start of the first run (control) and declined to 284 at the start of the second run (case). According to Oxford Nanopore Technologies, the MinION flowcells has a total number of 512 active channels available for sequencing (Oxford Nanopore Technologies, n.d). The number of active channels were subjected to a substantial decline over the course of
the sequencing runs after the 2-hour timepoint and this observation was recurrent for both runs (see Figure 2). Reasons might be due to excess amounts of detergents or surfactants present in the samples causing interference with the nanopores by blocking its apertures and further inhibiting sequencing activities. Since no assays of the quality were recorded for the libraries, this issue remains unresolved.

The number of reads that were obtained during the two sequencing runs are low compared to other direct poly(A) RNA nanopore sequencing studies (Workman et al., 2018). The study conducted by Workman et al. (2018) generated 10,300,000 native poly(A) RNA reads using 30 flowcells in total. Thus, this implies that ~333,333 poly(A) RNA reads were generated from each flowcell (Workman et al., 2018), which can be compared to the 239,448 poly(A) RNA reads generated from one flowcell during this project. Reasons might be due to low poly(A) RNA quantities prior to sequencing or due to a vast majority of the fragments having a size <500 nts. Having degraded fragments present in the sample libraries reduces the number of read output since they are not being detected by the nanopores. The processing of small fragments produces background noises rather than eligible sequencing reads due to the base call algorithms not being able to translate the electrical current produced from fragments <500 nts. Furthermore, the issues related to read length distributions can be confirmed by comparing the median read lengths with the average size of a human mRNA which is 3392 bp (Piovesan, Caracausi, Antonaros, Pelleri, & Vitale, 2016). Interestingly, only half of the 3392 bp human mRNA, on average, is coding (Piovesan et al., 2016). This do imply that, on average, 1696 bp of the mRNA consist of untranslated regions (UTRs) at the 3’ and 5’ end of the mRNA. By assuming that the 3’ and 5’ UTRs are equal in length, the length of the 3’ UTR is 848 bp. Further, as mentioned earlier, oligo-dT selection enrich mRNAs by binding the poly(A) tail at the 3’ end of the mRNA. This might conclude that a vast majority of the read lengths produced during the sequencing runs consists of 3’ UTRs rather than protein coding regions.

Another aspect regarding sequencing output revolves around the big differences in read output between the two runs. It is very clear that performing two separate sequencing runs reduces the ability to compare sequence abundances between two treatments. These observations can be confirmed by previously published research showing that there is a substantial degradation of the flowcell following washing which is done in between two runs (Tyler et al., 2018). Further, Tyler et al. (2018) also showed that there is a carryover of DNA/RNA from previous runs and these findings were also confirmed by the supplier (Oxford Nanopore Technologies, n.d). One way to overcome this issue revolves around the concept of multiplexing. This allows the user to pool a large number of sequencing libraries and sequence them simultaneously in one single run. This is achieved by adding individual barcode sequences to each respective sample during library preparation so that each read from each individual sample can be identified and sorted prior to analysis (Illumina, n.d). This will further allow more efficient use of the flowcell (Oxford Nanopore Technologies, n.d).

A final concern to emphasise regarding poor sequencing qualities involves the time periods between receipt and use of the flowcell. The flowcell used during this thesis was received in early February and used in late April. According to the supplier, the recommendation is to use the flowcell within 3 months upon receipt and this require the flowcell to be stored at 4 °C. Thus, this means that the flowcell was used at the borderline of the expiration date. Noticeably, it appears to be a difference in performance of flowcells depending on whether they have been used early or later upon receipt. Flowcells stored for extended periods of time prior to use are
generally, of lower quality at the time of sequencing than those used immediately upon receipt (Tyler et al., 2018).

4.3 Interpretation and analysis

The comparative interpretation and analysis of the sequence output data using a bioinformatic approach is scarce due to limited time and knowledge in RNA-sequence analysis. Due to absence of a transcriptome assembly of the freshwater mussel *A. anatina*, a trial in aligning the RNA-sequence data was performed using a genome assembly of the freshwater mussel *V. ellipsiformis* (GenBank assembly accession: GCA_003401595.1) as reference. Aligning the sequence data was done using NanoPipe, a web-based tool for analysis of nanopore sequence datasets (Shabardina et al., 2019).

Prior to bioinformatic analysis of RNA-sequence data, it is important to conduct several quality assessments and pre-processes of the data in order to ensure a coherent analyses pipeline. This includes removing low-quality sequences, base-adaptors and overrepresented sequences (Han, Gao, Muegge, Zhang, & Zhou, 2015). Trimmomatic is a tool that can be used to enhance the quality of sequence data by removing adaptors and trimming low-quality bases (Bolger, Lohse, & Usadel, 2014). However, this tool is adapted for sequence datasets generated by the Illumina platform which makes it difficult to use when analysing nanopore datasets. Hence, no pre-processing of the nanopore data was conducted except using the automated base call algorithm provided in the MinKNOW software.

Only ~0.5% (852/179,651 reads control and 202/59,797 reads case) of the reads did align to the reference for each dataset respectively. This issue is probably due to the error-prone base call accuracies (<90%) causing interference with the algorithms during mapping. Thus, this does imply the need for profound pre-processing of the datasets as mentioned earlier. There may also be difficulties in detecting mRNA species abundances using a genome as reference due to a vast majority of expressed eukaryotic mRNAs being derived from alternative splice junctions (Johnson et al., 2003), in which different combinations of exons are being spliced together and non-coding intron regions are removed to produce different mRNA isoforms of a gene (Gilbert, 1978). Thus, the intron regions present in the genome reference could have affected the results by reducing the numbers and lengths of homologous regions as they are absent in the RNA-sequence data. Furthermore, only ~1% (9/852 control) and ~2% (4/202 case) of the mapped reads for each sample respectively did target and produce an aligned consensus region. The length of the targeted consensus regions varied from varied from 26 to 123 nucleotides (Table 4). The coverage of the median read length varied from 4% (26/650 nts) to 19% (123/650 nts) based of the median read length generated from the first sequencing run (control). The coverage of the median read length generated from the second sequencing run (case) varied from 10% (51/534 nts) to 23% (123/534 nts). Despite the smaller size of the consensus regions, the share of identities ranged from 76% (93/123 nts) to 92% (24/26 nts) (Table 4), which goes well in line with the generated quality scores (9.01 case and 9.13 control). However, the low number of mapped reads and targeted consensus regions makes the dataset ineligible for quantification of transcript abundances.

Interestingly, the four consensus target regions with the highest number of mapped reads in the first run (control) were identical to the four consensus regions found during the second run (case) (Table 3 & 4). Given that the two runs consist of two individuals as well as two different
treatment conditions, one would expect some inter-individual variations in the sequence data output. These issues could be linked to the previous section in the discussion part of this assay (section 4.2), which mentioned the implications of performing two sequencing runs separately. As demonstrated by Tyler et al. (2018), there is a carryover of DNA/RNA from previous sequencing runs. Hence, it cannot be ruled out that the sequence data output from the second run (case) is a result of re-sequenced carryovers of RNAs from the previous run (control).

Furthermore, the BLASTn analysis of the targeted consensus regions were scarce due to the short-read lengths and the low percentage of nucleotide identities (Table 4). This suggests that an increased read length is needed for correct species identifications as only 22% (2 out of 9) of the consensus regions identified species within the Mollusca phyla. Additionally, all of the identified mRNAs were ribosomal proteins, which are among the most highly conserved proteins across all life forms (Lecompte, Ripp, Thierry, Moras, & Poch, 2002). However, the low mRNA length coverage (10-18%) needs to be improved in order for it to be considered eligible and useful in mRNA species identification.

The BLASTn analysis of eight randomly selected fastq-files from the control sequencing run was conducted in order to confirm the generated quality score (9.13), Unionidae species identification, mRNA target identification and to observe mRNA transcript length coverage. The case sample will be covered in another thesis project. The top hits with the highest number of identities in each fastq-file respectively ranged from 83% (1174/1414 nts) to 89% (853/963 nts) number of identities which confirms an accurate quality score identification (<90%). The species target was narrowed down to search for Mollusca and all of the top hits targeted and identified species within the freshwater mussel Unionidae family. None of the hits identified A. anatina mRNA species, which is probably due to the scarce amount of transcriptomic profile data available for A. anatina. The closest relative of A. anatina was C. plicata, based of mitochondrial genomic data (Wen et al., 2017). Furthermore, the eight top hits did identify mRNA molecules, which supports a somewhat successful poly(A) RNA selection and sequencing. However, the data is not sufficient to determine and evaluate tissue-specific profiles based on the generated mRNAs. The total mRNA length coverage varied from 73% (1960/2674 nts) to 100% (1333/1339 nts), which does not confirm, but observes the potential ability of nanopore sequencing to detect and quantify full-length transcript isoforms. It has previously been demonstrated that nanopore sequencing can be used to identify full-length transcript isoforms derived from genes with complex splicing patterns (Bolisetty, Rajadinakaran, & Graveley, 2015). Hence, nanopore sequencing is slightly advantageous compared to other sequencing platforms when it comes to acquiring profound insights of the complexity of eukaryotic transcriptomes. Lastly, these findings are not definitive because the data was too weak to be processed with the BLASTn “highly similar sequences” (megablast) option.

Additionally, when a reference genome or transcriptome is absent or incomplete for the species of interest, the RNA-sequence dataset is recommended to be assembled de novo prior to analysis (Conesa et al., 2016). There are multiple packages available for de novo transcriptome assemblies, e.g. Oases (Schulz, Zerbino, Vingron, & Birney, 2012) and Trinity (Haas et al., 2013). Once the data has been assembled, the generated RNA-sequence reads can be mapped back to the novel constructed transcriptome reference using multiple sequence alignment tools, e.g. Bowtie (Langmead, Trapnell, Pop, & Salzberg, 2009). The reads can then be quantified based on the number of reads that map to each transcript sequence within the assembled transcriptome using tools such as HTSeq-count (Anders, Pyl, & Huber, 2015). Moreover, RNA-sequence
analyses pipelines are facilitated by basic knowledge in UNIX shell programming, Perl/Python scripting and the R programming environment (Wolf, 2013). Notably, it is also important for the transcriptome assemblies and transcript quantifications to be tissue specific, since genes are expressed based of necessities and functionality. This means that certain genes expressed in a particular tissue are not being expressed at all or at a lower level in other tissues because each tissue has its own function and properties. In this case, hepatopancreas was the targeted tissue, which function in the production of digestive enzymes and adsorption of food. Thus, this conclude that hepatopancreas is likely to have a higher and more specific expression of genes involved in metabolic regulation when compared to other tissues (Chen et al., 2015; Sun et al., 2016). Same goes for molecular responses as a result of exposure to environmental stressors, where changes in gene expression levels varies among specific tissue profiles.

Currently, MinION nanopore sequencing has primarily been utilised in genome assemblies of prokaryotic and eukaryotic genomes as it is able to sequence long DNA fragments (>10 kb) (Loman, Quick, & Simpson, 2015; Goodwin et al., 2019). In addition, it has also successfully been used in disease surveillance due to its real-time features (Quick et al., 2016). However, the use of nanopore RNA-sequencing in differential expression analysis is still disadvantageous compared to other established NGS technologies, e.g. Illumina, due to the error-prone base call accuracies where short-read NGS is superior to nanopore sequencing. Further, the sequencing throughput of the MinION flowcell is still not very stable with high fluctuations in the generation of raw sequence data (Ip et al., 2015). According to the ENCODE guidelines for RNA-sequence experiments, differential expression analysis usually requires a minimum of 20 million reads per sample library in order to detect big changes in expression levels (ENCODE, 2016). On the other hand, in order to detect small changes in expression levels, the number of biological replicates needs to be increased (Schurch et al., 2016). The number of required reads would be reduced, since nanopore sequencing generates longer reads. For instance, the Illumina platform provides a sequence library with a maximum read length of 300 bp. Based on this example, let’s assume nanopore sequencing provides a read length of 1000 bp, on average. This would approximately reduce the required read output by two thirds, thus it would require ~6.6 million reads to be generated with nanopore sequencing. Interestingly, a study which aimed to sequence human poly(A) RNA produced 10.3 million nanopore reads with a median read length of 771 bases. This, in turn, required a total use of 30 flowcells (Workman et al., 2018). This suggests that using one flowcell with the purpose of detecting changes in expression levels is insufficient. Furthermore, hybrid error correction is a recurrent method commonly used to overcome the issues with error-prone nanopore base call accuracies (Koren et al., 2012; Bayliss, Hunt, Yokoyama, Thorpe, & Feil, 2017; Goodwin et al., 2019; Depledge et al., 2019). This method combines long error-prone nanopore reads with low error rate short reads in order to detect and adjust incorrect base calls.

In summary, nanopore sequencing has not yet been fully optimised and standardised for use in RNA-sequencing and differential expression analyses when compared to other NGS platforms. Nanopore sequencing still relies on other technologies for error correction and suffers from high variations in the throughput of raw sequence data. This makes the technology very unpredictable, which in turn establishes a state of never knowing what to expect.
4.4 Conclusion

This study aimed to sought for molecular responses in *A. anatina* using nanopore sequencing, however over the course of the project, several implications appeared, and the direction of the project was changed multiple times. It ended up becoming a pilot study rather than an entitled scientific study in regards to the significance of the results. The discussion section of this thesis emphasises crucial aspects in the sample preparation steps prior to sequencing, nanopore sequence output implications and advances and lastly essential key features that needs to be taken under consideration during analysis and interpretation of RNA-sequence data. In summary, this project shows that nanopore sequencing is capable of producing longer single RNA-sequencing reads when compared to other NGS platforms. However, RNA-sequencing experiments are very fragile and thus requires a thorough and effective design during the sample preparation steps prior to sequencing. Future advances include a hybrid error corrected *de novo* transcriptome assembly for efficient use of *A. anatina* as a bioindicator in aquatic ecotoxicology.
Acknowledgements

I would like to thank my main supervisor Mikael Ejdebäck for his encouragements and input expertise over the course of this thesis project.

I would also like to thank my co-supervisor Magnus Fagerlind for his participation and help during the experimental part of this thesis.

Special thanks to my fellow comrade Johan Bergholm at the University of Skövde, this thesis would not have been possible to accomplish without your guidance and help during the experimental sessions.

I would further like to acknowledge Johan Baxter at the University of Skövde, as he voluntarily helped and shared his acquired knowledges about nanopore sequencing.

Lastly, I would like to thank my mother and my father for always supporting me through prosperity and adversity.

Thank you all.
References


