Master Degree Project

UNIVERSITY OF SKÖVDE

GENOMIC DNA SEQUENCING OF FRESHWATER MUSSEL USING THE MINION

Master Degree Project in Bioscience
One year Level, 60 credits

Sanam Zeb
A18sanze@student.his.se

Supervisor: Mikael Ejdebäck
Mikael.ejdeback@his.se

Examiner: Anna-Karin Pernestig
Anna-karin.pernestig@his.se

Abstract

Freshwater mussels (Unionida) belong to phylum Mollusca and live in freshwater habitats, such as lakes and rivers. Freshwater mussels have high capacity for water purification and play an important role in calcium recycling. There is not much information about the freshwater mussel genome due to lack of genomic sequences in the database, till now only four species have been sequenced and the only Swedish one is *Margaritifera margaritifera*. This study aim was to use nanopore sequencing technology to sequence the genomic DNA of a freshwater mussel. The data about the genomic sequence is helpful in identification of their species and give a better understanding towards the genomics and transcriptomics, it also could help in the development of multi-biomarker panels for an early assessment of water pollution. In this study the DNA used was extracted from the foot tissues, and different tissue homogenization methods were tested to find the best approach. The genomic DNA was sequenced by using Oxford nanopore MinION device, and the reads were assembled and polished using multiple software through bioinformatic analysis. The number of reads from sequencing the DNA covered only 13.5x of the estimated genome size of the freshwater mussel, while the required coverage for a complete genome assembly is 20x-25x or higher. Due to low coverage and fragmentation, a partial sequence of the genomic DNA was obtained. This indicates that nanopore sequencing could be used, but additional sequencing runs are needed to get enough coverage to assemble a complete genomic DNA of the freshwater mussel.

Popular scientific summary

Water pollution is a serious environmental issue because human health can be affected by consuming, entering, or washing in polluted water. Water is sometimes referred to as the universal solvent, as it dissolves more substances than any other liquid. However, this ability means that water is easily prone to pollution. The main causes of water pollution are industrial waste, sewage, mining activity, marine dumping, accidental oil leakage, dangerous chemical, plastic and polybags etc. Water contamination has great threat for the aquatic life and ecosystem. Heavy pollutant contamination causes sterility, a number of diseases, fatality and sometime extinction of marine species. Early detection of water pollution will help in preventing further increase in water pollution. Two types of methods are currently used to detect the water pollution. The first type are the chemical methods, which are based on the tests to check the presences of a specific chemical contaminant in water, such a magnesium or zinc. The second type are the biological methods that rely on using the organisms that are characterized by particular susceptibility to contaminants to check if there is any contaminant in water that is affecting these organisms. Examples of these organisms are the mussels.

Freshwater mussels live in freshwater habitats, lakes and rivers. Freshwater mussels have high capacity for water purification in rivers and lakes. They also play an important role in calcium recycling. Therefore, freshwater mussels are important component of ecosystem and can be used as a biological indicator for water pollution. Freshwater mussels can remove a variety of contaminating materials from the water column, including sediment, organic matter, bacteria, and phytoplankton, which makes them extremely interesting from a biological perspective. And in ecological monitoring, mussels have been used previously to indicate pollutants in water, such as copper, cadmium, zinc, and other contaminants. A multi-biomarker panel is a method that use biological tests to determine the existence of multiple markers in a living organism. The biomarkers can be detected in the organism at all conditions but their levels or activities changes upon pollution. Different approaches and methods for sequencing the DNA have been developed and applied in recent time to make sufficient genomic data available for advanced research and this data will be also very helpful in development of multi-biomarker panels.

In Europe, sixteen species of unionid freshwater mussels are found, out of which seven are found in Sweden. These species can be found in many rivers and lakes around Sweden. The genome of many Unionid mussel species has not been sequenced yet. This study aim is to find the composition and order of the genomic DNA of freshwater mussel, which will be helpful in identification of their species and give a better understanding towards the genomics and transcriptomics. By knowing the genomic sequence, we can understand the transcriptomic changes caused by pollution. Then, these transcriptional changes can be studied to determine if they can be used as a biomarker for water pollution. Due to the challenges that were faced during this project the aim was partially achieved. The obtained part of genomic DNA sequence could be used to study the possibility of using freshwater mussels to develop a multi-biomarker panel. However, the whole genomic DNA sequence is still required to make sure that it is possible. The current study also included a considerable effort to understand the optimal methods that can be used to obtain the correct composition and order of the DNA sequence of freshwater mussels.

Table of Contents

Abbreviations

Bp	Base pair
BR	Broad range
CNS	Central nervous system
EPA	Environment protection agency
EU	European Union
HS	High sensitivity
ITS	Internal transcribed spacer
IUCN	International Union for conservation of nature
Mbp	Mega base pairs
NGS	Next generation sequencing
ONT	Oxford nanopore technology
PCR	Polymerase chain reaction
RFLP	Restriction fragment length polymorphism
SFB	Short fragment buffer
WGS	Whole-genome shotgun sequencing

Introduction

Freshwater mussels (Unionida) belong to phylum Mollusca, and live in freshwater habitats such as lakes and rivers (Rosenburg, 2014). They have high capacity for water purification (Vaughn, 2018) and they also play an important role in calcium recycling (Green et al., 1985). Freshwater mussels play important role in maintaining the ecosystems of the environment in the surrounding areas by regulating water purification by a process called bioremediation (Vaughn, 2018). Bioremediation is a process where the waste in the polluted environment is neutralized or removed by an organism (Prince, 2000). According to the environment protection agency (EPA), the definition of bioremediation is "a treatment that utilizes naturally occurring organisms to break-down a dangerous substance into a less hazardous or non-toxic substance". Freshwater mussels are important components of ecosystems and are used as a biological indicator for water pollution. Unionids remove a variety of materials from the water column, including sediment, organic matter, bacteria, and phytoplankton, which makes them extremely interesting from a biological perspective (Strayer et al., 2004).

An inventory of the global and regional diversity of the unionoida has been performed, to know how many species of freshwater mussels are there and how they are distributed (Daniel & Kevin, 2007).

In Europe, sixteen species of unionid freshwater mussels are found, out of which seven are found in Sweden. The most famous in Sweden are the freshwater pearl mussel (*Margaritifera margaritifera*) and the thick-shelled river mussel (*Unio crassus*). Both species are categorized as endangered on the international union for conservation of nature (IUCN) and the Swedish Red list of species (Osterling et al., 2012). Among the different species of freshwater mussels, the *Anodontia anatina* which belongs to family Unionidae consider to be safer and ecologically relevant model organism for field and laboratory. This specie is widely distributed in Sweden and other parts of Europe (Lopes-Lime, 2014). Research on *Anodontia anatina* has mostly focused on morphology, phylogeny, reproduction and seasonal behavior (Aldridge 1999; Jonsson et al., 2013; Lurman et al., 2014).

The identification of mussel species is a complicated task. Adult mussels are often identified by their shell morphology, which can be difficult as shell shape is variable and influenced by environmental factors, it is similar between close-related species and also affected by interspecies junction and intraspecies variability (Watters 1994; Zieritz et al., 2012). In addition, morphological approaches cannot easily be applied on the small mussel larvae. A complete molecular identification key based on the polymerase chain reaction (PCR) and restriction fragment length polymorphism (RFLP) principles was recently developed for all indigenous North and Central European unionoid species (Zieritz et al., 2012). This information can be helpful in identification of mussel species. There is no much information about the freshwater mussel genome and only four species can be found in the genomic databases and the only Swedish one is *Margaritifera margaritifera* freshwater pearl shell mussel.

DNA sequencing is the process of determining the nucleic acid sequence and the order of nucleotides in DNA and it includes the method that is used to determine the order of the four bases: adenine, guanine, cytosine, and thymine. The first DNA sequencing methods were obtained in the early 1970s included the Maxam-Gilbert method, discovered by and named for American molecular biologists Allan M. Maxam and Walter Gilbert, and the Sanger method or dideoxy method discovered by English biochemist Frederick Sanger (Sanger et al., 1977). The new DNA sequencing methods has greatly accelerated biological and medical research and discovery. The rapid speed of sequencing achieved with modern DNA sequencing technology has been involved in the sequencing of complete DNA sequences, or genomes, of numerous types and species of life, including the human genome and other complete DNA sequences of many animal, plant, and microbial species. It allows the researchers to identify changes in the genes that are associated with the diseases and abnormal condition. Sequencing also allows for the identification and diagnosis of viral infections and drug resistance testing, which is very useful in designing

medicines (Fei & Ng, 2019). Sequencing data can be used to determine different areas of the DNA; for example, which areas of the DNA contain genes and which areas provide regulatory instruction. In addition, most importantly sequence data can highlight changes in a gene that may cause inherited diseases (França et al., 2002).

Modern sequencing methods have been developed that lower the cost and increase sequencing accuracy and precision such as Next generation sequencing (NGS) and the Nanopore sequencing (ku & Roukos, 2013).

Despite all the scientific achievements that science have achieved in the past 50 years, sequencing a whole genome of a eukaryotic organism remains difficult task. Sequencing the whole genome requires breaking the DNA into small fragments, then sequencing the fragments and assembling the fragments into a long consensus (Jung et al., 2019). The significant steps involved in the genomic DNA sequencing (Figure 1). And the brief summary of the steps is discussed below.

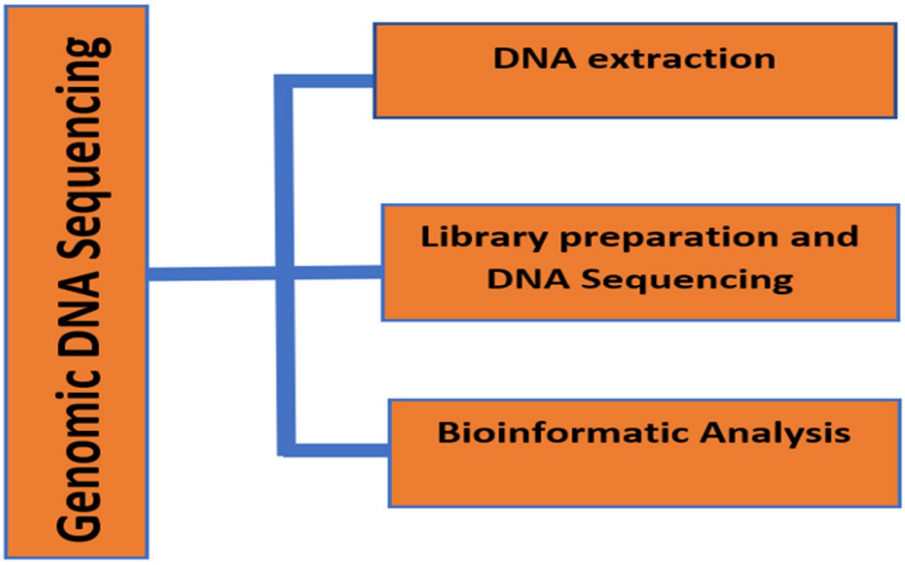

Figure 1. Representing the main steps of genomic DNA sequencing. This first step is the DNA extraction, second step is the library preparation for sequencing and last step is the bioinformatic analysis.

DNA extraction

DNA extraction (Figure 1) is a procedure used to isolate DNA from the nucleus of cell. It can be defined as the process of separating the DNA from protein and other cell debris (Elkins, 2013). The process of DNA extraction requires careful handling of the sample to prevent contamination (Elkins, 2013). In eukaryotic organisms there are three main steps of DNA extraction; cell lysis, DNA purification and the DNA precipitation. In the cell lysis step, disruption of the outer boundary or cell membrane is done to release the inter-cellular materials such as DNA etc. It can be done either by a chemical method using a cell lysis buffer or by a physical method such as tissue homogenizer. DNA purification is a method of extracting and purifying DNA from other cellular components. And it can be done both physically and chemically. The last step of DNA extraction is the DNA precipitation, which can be done either by isopropanol or ethanol (Kelly & Elkins,2013).

Library preparation

Library preparation (Figure 1) is the first step of NGS. It allows DNA or RNA to adhere to the sequencing flow cell and allows the sample to be identified. Library preparation is done by attaching an artificial DNA segment (adapters) to both ends of the DNA fragments (Gansauge & Meyer, 2013). These adapters help to allow for priming of the sequencing reaction and it also enables amplification of the DNA library by PCR. By amplifying all the fragments that were

transformed into library molecules, the library is successfully immortalized. Once the library preparation is done, DNA is conditioned and is ready to be load into the sequencing device (Mikheyev & Tin, 2014). The Library preparation generally takes about half a day, depending on the protocol and the sequencing device being used (Mikheyev & Tin, 2014).

Nanopore sequencing was introduced by David Deamer at the University of California and by George Church and Daniel Branton both at Harvard University. Nanopore sequencing technology was developed at the beginning of the 1990s, many academic laboratories reached series of achievements toward making a functional nanopore sequencing platform (Deamer et al., 2010). Multiple companies have proposed nanopore-based sequencing methods, but only MinION-based sequencing has been effectively used by independent genomics laboratories around the world (Jain et al., 2016). According to Oxford nanopore technology (ONT), two types of DNA adapters are there that allow for DNA priming and DNA amplification through the PCR. The first type of adapters, called 1D adapters which enable the template DNA strand and the complementary DNA strand to be sequenced as individual strands and result in a very high data yield (reads) from a wide variety of sample types. The second type of adapters called 1D2 adapters, are special type of DNA adapters that increase the chance of the complement strand to immediately follow the template strand during sequencing. The adapters using 1D2 analysis methods, have low rate errors in sequencing and produces a higher accuracy reads. So, the choice of sequencing adapters depends on the purpose of the sequencing experiment. If the sequencing experiment requires a high data yield such as whole-genome sequencing, then 1D sequencing adapters can be used. Whereas, if the experiment requires a high base accuracy and a low rate of errors then 1D2 sequencing adapters can be used.

Bioinformatic Analysis

The bioinformatic processing (Figure 1) is most complex and time-consuming part of genomic DNA sequencing. The first step in bioinformatic processing after obtaining the raw reads using the Oxford nanopore sequencing device is base-calling, which is the computational process of translating raw electrical signals to a nucleotide sequence (Wick et al., 2019). For Base-calling five different computer software's can be used. After base-calling, the next step is the genome assembly, which is the process of merging small sequenced fragments into larger fragments (contigs) using an assembly software such as Canu or Falcon (Foxman, 2012). The next step is contigs polishing, it is the process of aligning the consensus to the base-called data to find any possible errors, and later fix these errors using the original reads (Zimin & Aleksey, 2019). The output contigs obtained from Canu software and the polished sequence obtained from Racon and Medaka software will be assessed for their quality by Quast. And Quast can evaluate and compare genome assemblies both with a reference genome as well as without a reference genome (Gurevich et al., 2020).

Multi-biomarker panel

Knowledge about the genomic sequence of the freshwater mussels could be helpful in identification of their species and give a better understanding towards the genomics, transcriptomics and tracking the evolution. It is considered that pollution cause a change in several metabolites during metabolism in an organism. The knowledge about the genomic sequence is also helpful in the development of multi-biomarker panels, which are consider to be powerful tools in the early assessment of water pollution. The biomarkers can be detected in the organism at all conditions but their levels or activities changes upon the pollution. Hence by knowing the genome it becomes easy to understand the transcriptomic changes caused by pollution.

Aim

The genome of many mussel species has not been sequenced yet. The aim of this study was to determine the genomic DNA sequence of a freshwater mussels, which could be very helpful in identification of their species and give a better understanding towards the genomics and transcriptomics and in the development of multi-biomarker panel for early assessment of water pollution.

By knowing the genomic sequence it's easy to understand the transcriptomic changes caused by pollution. Then, these transcriptional changes can be studied to determine if they can be used as a biomarker for water pollution. In this study, the aim was to sequence the genomic DNA of one freshwater mussel. Mussels will be collected from a local source then will be dissected and DNA will be extracted by using different methods extracted to select the one with long DNA fragment size, sequencing will be done using Minion sequencing device followed by bioinformatic analysis.

Materials and Methods

Sample preparation

Freshwater mussels were collected on 21st March, 2021 from Vingsjön lake near Axvall, Sweden. Seven samples of freshwater mussels were collected from the lake and stored in a bucket on ice. A scalpel was used to cut the anterior and posterior valves of the mussel. The shell of the mussels was opened, and the mantle was removed by using a scalpel. The foot, pancreas and gills of freshwater mussels were excised at the university of Skövde, by a scalpel and stored in the freezer at -80°C until DNA extraction.

Two different kits were compared for DNA extraction. The blood and cell culture DNA midi kit (Qiagen) was used with two different homogenization methods, liquid nitrogen and the TissueLyser (Qiagen) to determine which homogenization method gives longer DNA fragment sizes. And the NucleoBond HMW DNA kit (Macherey-Nagel) was used instead with two different homogenization methods. In the first homogenization method, 300mg of gill tissue was ground with liquid nitrogen and a mortar. In the second method, 300mg of gill tissue was homogenized using enzymatic lysis NucleoBond HMW DNA kit (Macherey-Nagel). The final DNA sample of a freshwater mussel was then extracted, 300mg of the foot tissue was homogenized by using enzymatic lysis NucleoBond HMW DNA kit (Macherey-Nagel) and a NucleoBond HMW Column (including a filter) with a plastic washer arranged on a 50 ml tube was used for the extraction.

The DNA concentration was measured using a Qubit 4.0 fluorometer, and the dsDNA High-sensitivity (HS) assay kit (Invitrogen) and the dsDNA Broad-range (BR) assay kit (Invitrogen) protocol was followed. The purity of the DNA was tested using Nanodrop spectrophotometer (Saveen Werner). The length distribution of the fragments was analyzed using a 0.8% Agarose Gel electrophoresis technique by following the instructions in the addgene protocol.

Species Identification

For the identification of species of a freshwater mussel, a PCR along with a fragment analyzer was performed. For PCR reaction the protocol of Phusion high-fidelity PCR kit (New England Biolabs) was followed. The Invitrogen custom primers (Thermo Fisher Scientific), internal transcribed spacer (ITS), ITS-1-F (forward) with a sequence (5' to 3') AAG ACT GGG TTG CGG AGG and ITS-1-R (reverse) with a sequence (5' to 3') GAG TGA TCC ACC GCT TAG A were used. In total six reactions were performed each having a total volume of 50µl. For fragment analysis the length of the fragments was analyzed by using a Fragment Analyzer (Advanced analytical) following the instructions in the DNF-915 dsDNA reagent kit protocol.

Sequencing

The extracted DNA was then prepared for sequencing by following genomic DNA by ligation (SQK-LSK109) protocol (ONT). For Library preparation 2.5 µg of the extracted DNA was transferred into a 1.5 ml Eppendorf DNA LoBind tube, and the volume was adjusted to 49 µL with nuclease-free water was used for library preparation according to the instruction in the protocol. After finishing the preparation step, one microliter of the eluted sample was quantified by using a Qubit fluorometer (Invitrogen) by using a dsDNA High-sensitivity assay kit (Invitrogen) protocol. After preparation, adapter ligation and clean up were done by following genomic DNA by ligation (SQK-LSK109) protocol (ONT). Short Fragments buffer (SFB) was used to retain DNA fragments of all sizes. After the library preparation and adapter ligation steps, one microliter of eluted DNA sample was quantified by using a Qubit fluorometer (Invitrogen) by employing dsDNA HS assay kit (Invitrogen) protocol. The prepared library was stored on ice until loaded into the flow cells.

The library was loaded on a new R9.4.1 Minion flow cell, and the number of pores was tested before and after the sequencing run. In total, 2.5 µg of the final library preparation were loaded into the R9.4.1 flow cell by following the instructions of the genomic DNA by ligation (SQK-LSK109) protocol (ONT). The flow cell was attached to a laptop that contains MinKNOW version 19.12.5 software. Sequencing was started by using MinKNOW software. The duration of the sequencing run was decided and set to 24 hours, and the output format was set to fast5 files. The base-calling option was turned off during this stage in order not to devastate computer memory capacity.

Bioinformatic processing

The generated fast5 files from running the DNA by using MinKNOW software were base-called using Guppy software (Wick et al., 2020). The software was used with the default settings, and the GPU version of base-calling was used. Guppy has two running options (fast or accurate), and the accurate method was used in this experiment. The generated fastq files were separated into a pass or fail folder by using a quality score of seven. The results generated by the Guppy Base-calling software were analyzed with quality control assessment software PycoQC (Leger & Leonardi, 2019). Default settings were used for the software, and a base-calling summary file was used as an input. The obtained results from base-calling and quality control were saved on an external hard drive for further analysis. The passed fastq reads from base-calling were corrected and trimmed, and then assembled by using Canu software (Koren & Walenz, 2020). Canu software was run by using seven threads and using default parameters for assembly, while the genome size was set to 1.6Mbp, as the estimated reference genome size.

The contigs obtained by Canu software (Koren & Walenz, 2020) were aligned to the Bivalvia, Unionoida (*Venustaconcha ellipsiformis*) reference genome by using minimap2 software (Li, 2018). The software was run using the default parameters to map the reads to a reference genome, and an output mapping file in sam format was obtained.

The obtained contigs from Canu software (Koren & Walenz, 2020) were polished against the mapping file using Racon software (Vaser et al, 2020). The software was run by using the default parameters, and seven threads were used. The polished contigs obtained from Racon software (Vaser et al., 2020) were polished again to increase the accuracy of the consensus sequences using Medaka software (ONT, 2018). The software was run using default settings. The trimmed reads file obtained from Canu software was aligned to the polished contigs obtained by Racon software to increase the polishing accuracy of the contigs. Finally, the output contigs obtained from Canu software (Koren & Walenz, 2020) and the polished sequences obtained from Racon software (Vaser et al., 2020) and Medaka software (ONT, 2018) were assessed for their quality by using Quast software (Gurevich et al., 2020).

Results

Sample preparation

Tissue homogenization

Two different tissue homogenization methods were first tested to determine which method resulted in less DNA fragmentation and gave higher concentration and purity of the DNA (Table 1).

Table 1. DNA Extraction

	Enzymatic Lysis	Liquid Nitrogen
DNA Concentration (ng/µL) *	225.69	28.74
A260/230*	2.43	2.23
A260/280*	1.87	1.81
DNA Concentration (ng/µL) **	183	14.8

*Nanodrop **Qubit.

The fragment size of the extracted DNA homogenized by both methods was analyzed using 0.8% Agarose gel electrophoresis technique. There was the smear formation in extracted DNA homogenized by enzymatic lysis method and fragmentation with multiple bands were observed in the extracted DNA homogenized with liquid nitrogen (not shown).

The final DNA sample to be used for sequencing was extracted by using the enzymatic lysis method NucleoBond HMW DNA kit (Macherey-Nagel). The result obtained from analyzing the concentration and the purity of the genomic DNA using Nanodrop shows a DNA concentration of 446 ng/µL. The absorbance at 260/230 nm was 2.42, while the absorbance at 260/280 nm was 1.85. The concentration that was obtained by using Qubit dsDNA BR assay kit was 458 ng/µL.

The DNA fragment size was analyzed using 0.8% Agarose gel electrophoresis technique. The fragment size of both concentrated and diluted sample was analyzed. The result shown in (Figure 2).

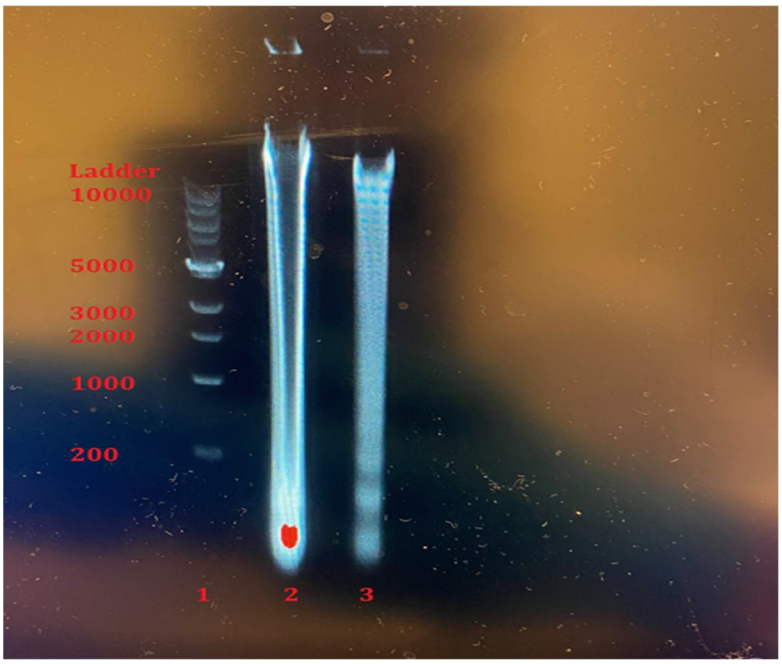

Figure 2. Gel electrophoresis result, the bands observed on the top of the gel is the DNA stuck in the wells (not entering the gel). Then in lane 1 of the gel is the 1kb DNA ladder and then in lane 2 is the concentrated sample which is definitely overloaded and in lane 3 is the diluted sample which is a fragmented smear.

Species Identification

The PCR along with a fragment analyzer was used for identification of freshwater mussel species. A complete molecular identification key based on the PCR and RFLP principles recently developed for all indigenous North and Central European unionoid species (Zieritz et al. 2012) was used along with different GenBank entries as a reference guide. The observed result on the fragment analyzer show a fragment size of 478 as shown (Figure 3).

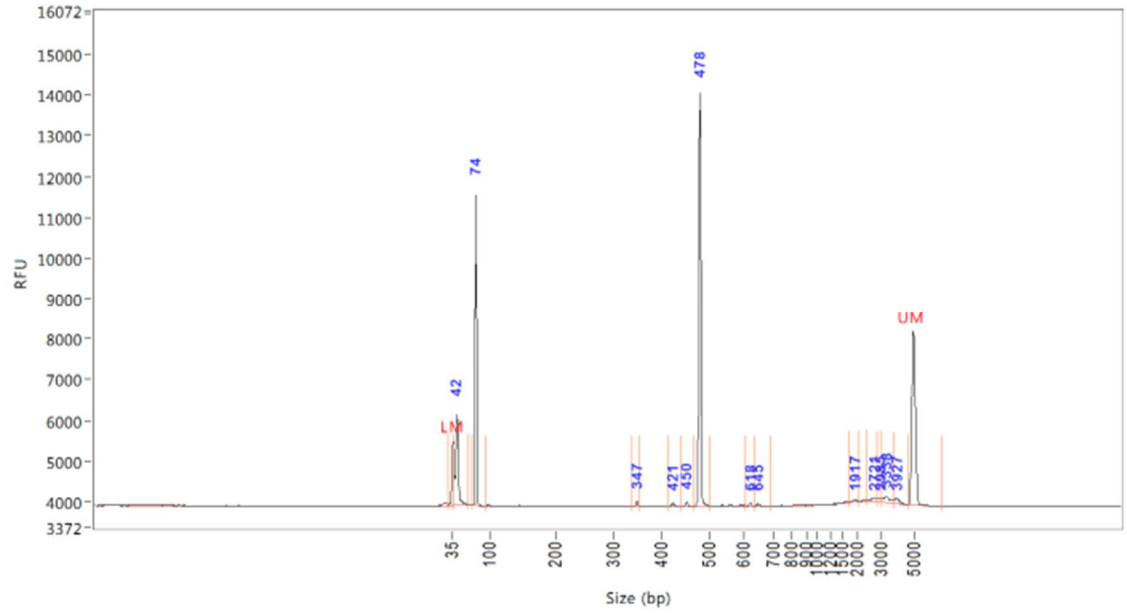

Figure 3. Represents the fragment analyzer result for the identification of freshwater Mussels Unionida using PCR and fragment analyzer. The x-axis shows the size of each fragment in bp, and relative fluorescence units (RFU) are shown on the y-axis corresponding to the amount of DNA for each fragment. Where (LM) represents the lower marker, and (UM) represents the upper marker used for calibration.

Sequencing and Base-calling

The number of pores available for sequencing R9.4.1 flow cell prior for running the sequencing run was 1022 pores. The duration of the sequencing run was decided and set to 24 hours and the output format was set to fast5 files. Base-calling the data using Guppy software (Wick et al., 2020) with the GPU accurate version of the software indicated that 192k reads had been base-called with 132k bases and a median read quality of 11.69. Analyzing the reads quality scores by using PycoQC software (Leger & Leonardi, 2019) indicates that more than 90% of the reads have passed the quality score of seven, and the reads median quality score is 11.69 as shown in (Figure 4). The median read length of the passed reads were 450bp.

Figure 4. Result from analyzing the base-called reads quality scores by using pycoQC version 2.5.2 software. The x-axis represents the read quality scores and y-axis represents the density of reads.

Bioinformatic processing

Assembly

After running Canu software (Koren & Walenz, 2020), the input sequences, generated 23660 reads with a total of 21.6Mbp bases with 13.5x of the total length of the genome size, which is a low coverage as 20x coverage or higher is required for assembly (Koren et al., 2017). The result from running Canu software to correct, trim, and assemble the reads assessed by using Quast software shows a total of 60 contigs with largest contig size of 16745 bp and a total length of 203 234 (Table 2).

Polishing

Polishing the contigs using Racon software (Vaser et al., 2020) software and Medaka software (ONT, 2018) has been assessed by using Quast software (Gurevich et al., 2020). Polishing the contigs by using Racon software (Vaser et al., 2020) shows that a total of 22 contigs has been reserved after the polishing step with the largest contig size of 5431bp. Polishing the contigs furthermore by using Medaka software (ONT, 2018) shows a total of 19 contigs that have been reserved after the polishing step with the largest contig size of 5441 bp. The result in (Table 2) shows the number of contigs, largest contig and total length obtained after running Canu software (Koren & Walenz, 2020) to assemble the data and the result obtained after the polishing steps from Racon software (Vaser et al., 2020) and the Medaka software (ONT, 2018).

Table 2. Represent the summarized result obtained from Quast by assessing the output contigs obtained from Canu software and the polished sequence obtained from Racon and Medaka software.

Software	Canu	Racon	Medaka
Number of contigs (>=0 bp)	60	22	19
Largest contigs (bp)	16745	5431	5441
Total length (bp)	203234	49685	47686

Discussion

Nanopore sequencing technology was developed at the beginning of 1990s when academic laboratories reached a series of milestones towards developing a functional nanopore sequencing platform (Branton et al., 2008). Nanopore sequencing is a third-generation sequencing technology that has two key advantages over second-generation technologies as it gives longer reads and the ability to perform real-time sequence analysis. As compare to other sequencing methods the MinION is significant for genomic DNA sequencing because of its small size, low cost, simple library prep, and portability (Greninger et al., 2015). The MinION platform allows for real-time analysis because individual DNA strands are translocated through the nanopore, allowing decisions to be made during the sequencing run, this is an advantage of nanopore especially in clinical use (Jain et al., 2016). Sequencing adapters are ligated to both ends of the genomic DNA fragments these adapters allow long and continuous sequencing of both strands of duplex molecule by covalently binding one strand to the other strand (Jain et al., 2016). As the DNA passes through the pore, the sensor will detect the change in an ionic current caused by differences in nucleotides that pass through the pore, which will give an accurate reading of the genomic DNA sequence loaded into the machine (Jain et al., 2016).

Homogenization and DNA extraction

For DNA extraction, different tissue homogenization methods can be used, depending on the type of tissue. In this study, foot tissue of freshwater mussel was used for DNA extraction. In order to get the good quantity and quality of DNA for genomic sequencing two different extraction kits were used along with different tissue homogenization method to select the procedure with least DNA fragmentation. A study performed by (Smith et al., 2011) to test three different tissue homogenization methods suggests that TissueLyser method did not cause over fragmentation of the genomic DNA. While in this study the result obtained by using the blood and cell culture midi kit (Qiagen), shows that over fragmentation has occurred. This problem can be avoided in the future by shortening the run time and the power of TissueLyser machine (Smith et al., 2011) or by using another tissue homogenization method that results in less mechanical shearing of the tissues.

To make the DNA ready for sequencing the quantity and quality check is important. As for every nanopore sequencing run the main starting material is the quality of the DNA during the library preparation step (Schalamun et al., 2019). The quality of DNA can be measured by using Nanodrop absorbance ratios at 260 nm and 280 nm. A dsDNA absorbance ratio at 260/230 between 2.0 and 2.2 and 260/280 between 1.8 and 2.0 is accepted as pure (Desjardins & Conklin, 2010). The 260/230 ratio for the DNA that was extracted by using enzymatic lysis and the liquid nitrogen methods (Table 1) were higher than the optimal level for this ratio, which show the presence of unwanted organic compounds. While the 260/280 ratio, falls within the optimal ratio interval (Boesenberg et al., 2012). The concentration of the DNA obtained by using enzymatic lysis method is 225.691 ng/µL, while the concentration obtained by using liquid nitrogen is 28.74 ng/µL (Table 1). This indicates that the concentration of DNA obtained from the Liquid nitrogen method is significantly lower than the enzymatic lysis method. So, for final DNA extraction enzymatic lysis method was used and gave a higher concentration of 446.194 ng/µL.

The fragment size of the extracted DNA homogenized by enzymatic lysis was analyzed by 0.8 % Agarose gel electrophoresis technique. As the sample was highly concentrated so it was diluted in the resuspension buffer HE. And the gel electrophoresis results a fragmented smear (Figure 2, lane 3).

The ONT evaluates that the optimal DNA size for MinION device should be within the range of 200bp to 8000bp (Lopez et al., 2019). Sequencing fragments with a size below 200 bp are consider to be useless, according to ONT because event detection and base-calling will not be possible at this range of size (Lopez et al., 2019). Which can affect the output of the sequencing run because

the MinION device will be busy in sequencing the short fragments instead of the longer fragments, which will lower the number of reads that will get base-called and pass the quality score test.

To summarize, the best approach for tissue homogenization would be pure DNA with long DNA fragment size. The result in (Table 1) indicates that enzymatic lysis approach gave higher concentration and purity than liquid nitrogen homogenization method. The fragment size of the extracted DNA homogenized by both methods was measured by 0.8 % Agarose gel electrophoresis technique. And for the final DNA extraction the enzymatic lysis approach was selected and the genome of freshwater mussels was sequenced by using Oxford MinION technology.

Species Identification

The adult mussels are often identified by their shell morphology and color of the tissues, which can be difficult as shell shape is very variable and influenced by environmental factors, it is similar between close-related species and also affected by interspecies junction and intraspecies variability (Watters 1994; Zieritz et al., 2012). In addition, morphological approaches cannot easily be applied on the small mussel larvae. Molecular genetic methods such as RFLP has been commonly used to develop molecular identification keys for freshwater mussel species found in United States (Kneeland & Rhymer 2007) and Europe (Gerke & Tiedmann 2001).

A previous study by (Zieritz et al., 2012) on complete molecular identification key based on the PCR and RFLP principles developed for all indigenous North and Central European unionoid species, along with the GenBank entries were used as a reference guide for specie identification (Figure 3), suggests that the specie could be *Anodontia anatine* but for correct identification sequencing is required.

Sequencing

Different approaches and methods for sequencing the DNA have been developed and applied in recent time to make sufficient genomic data available for advanced research (Prié et al., 2020). Using the MinION device for sequencing the DNA, generated 23660 reads with a total of 21.6Mbp bases with a coverage of 13.5x. As in this case the coverage is not enough for assembly. In the flow cell the number of active pores were 1022 before starting the sequencing run. The estimated number of pores in the flow cell after the sequencing run was 812. Whereas, the minimum number of pores suggested by ONT to start a sequencing run is 800 pores (McIntyre et al., 2016). This indicates that the low coverage obtained during the sequencing run is not due to a problem in the loaded DNA library. Hence, to sequence the reads that cover 20x or higher of the expected genome size of a freshwater mussel, suggests that the duration of the sequencing run should be increased to obtain enough coverage to assemble the whole genomic DNA of a freshwater mussel by using MinION flow cell. A high number of reads will result in an increased coverage for the sequenced genome, which gives a more accurate genomic DNA assembly (Jung et al., 2019).

Bioinformatic processing

Base-calling is a process of translating electronic raw signals of a sequencer into base sequences, it is of great importance to the sequencing platforms produced by Oxford nanopore Technologies (Wick et al., 2019). The result obtained from base-calling the reads by using Guppy software (Wick et al., 2020), was analyzed with quality control assessment software PycoQC which indicates that more than 90% of the reads have passed the quality score test (Figure 4).

Canu is a software used to assemble long reads from either PacBio or Oxford Nanopore, which have higher error rates than short reads from Illumina. And compared to Celera Assembler, this tool runs much faster and implemented some new overlapping and assembly algorithms such as adaptive overlapping strategy and sparse assembly graph construction. It can also provide output in graphical fragment assembly (GFA) format (Koren et al., 2017). The minimum coverage that Canu can perform an assembly on is 20x of the genome size of the sequenced organism (Koren et al., 2017). The result obtained by sequencing the genomic DNA of freshwater mussels for 24 hours,

indicates a coverage of 13.5x, this low coverage is a sign that correction of reads did not work well. This coverage is not enough for Canu to assemble a highly contiguous *de novo* genome of the freshwater mussel. In future, problem of low coverage can be avoided by increasing duration of the sequencing run from 24 hours to 36 or 48 hours based on the data output, to obtain more reads or doing multiple sequencing runs to obtained enough coverage to cover at least 20x of the genome size (Xu et al., 2017).

Contig polishing is the process of aligning the reads obtained by sequencing to the contigs obtained by the assembly to increase the quality and the accuracy of the contigs. Polishing the contigs using Racon software (Vaser et al., 2020) software and Medaka software (ONT, 2018), were assessed by using Quast software (Gurevich et al., 2020) determine that the Racon and Medaka assembly have much fewer and shorter number of contigs as compared to Canu (Table 2).

Most large-scale genome sequencing experiments nowadays use the whole-genome shotgun sequencing (WGS) strategy on large-scale genome sequencing. This approach is based on cutting the DNA into multiple small fragments that differ in size and can be sequenced on both ends to obtain multiple fragments size (Pop, 2004). To assemble these fragments these into longer sequences (contigs) by using any of the multiple assembly software available such as Canu or Falcon (Giordano, 2017).

The third-generation long reads sequencing such as PacBio and Nanopore methods, have great benefits over second-generation Illumina sequencing in *de novo* assembly studies (Wang et al., 2020). A study by (Renaut et al., 2018), used a hybrid *de novo* assembly and annotation of the genome of a freshwater mussel, *Venustaconcha ellipsiformis*. The genome described here was obtained by hybrid sequencing technologies using long read–low coverage and short read–high coverage, which offer an affordable strategy with the advantage of assembling repeated regions of the genome (for which short reads are ineffective) and circumventing the relatively higher error rate of long reads (Koren et al., 2012; Miller et al., 2017).

Conclusion

In this study, the genomic sequence of a freshwater mussel, was partially obtained but the species could not be identified. The data that was obtained can be used to analyze the genomics of species and aid analysis of future transcriptomic data obtained from the mussel species. And to solve the problem of DNA fragmentation, more focus should be on pure DNA with long DNA fragment size during the sample preparation. The sequencing and base-calling were performed properly. The obtained amount of reads and coverage was too low to assemble the whole genome of a freshwater mussel. The duration of time for the sequencing run should be increased to get more reads that cover at least 20x of the genome size of a freshwater mussel. A high number of reads will result in an increased coverage for the sequenced genome, which gives a more accurate genomic DNA assembly.

Ethical aspects, gender perspectives, and impact on the society

As the European Union (EU), directive puts emphasis on species-specific education and on the implementation of the three R's in every aspect of use and care of laboratory animal. In this experimental study the three R's ethical guiding principles, the replacement, refinement and reduction were followed. The replacement which refers to replace the experiments with other alternative methods if possible is unfortunately impossible due to the lack of published sequencing read for freshwater mussel. The refinement method refers to avoid the pain and discomfort to animals. This method is less relevant as the organism lacks Central nervous system (CNS), means freshwater mussels do not have the sense of pain and sufferings. The reduction refers to reduce the number of animals used in the experiment. This method was followed where the lowest number of individual (one is enough for sequencing) being used to give the expected outcome result. Freshwater mussels are important component of ecosystem and can be used as a biological indicator for water pollution. The aim of this experiment was to sequence the genomic DNA of a freshwater mussel, that can be used to develop a multi-biomarker panel to identify water pollution. By having knowledge about the genomic sequence, we can understand the transcriptomic changes caused by pollution. Then, these transcriptional changes can be studied to determine if they can be used as a biomarker for the early assessment of water pollution. The genomic DNA of freshwater mussels can also be used to develop a fast and accurate method to identify the species of freshwater mussels and give a better understanding towards the genomics including evolution and transcriptomics.

Future perspectives

The main purpose of this research project was partially accomplished. The sequencing result indicates that only 13.5x of the genome coverage was obtained and this coverage is not enough to assemble a sequence that covers a whole genome. Which means, more sequencing runs are required to obtain sufficient reads that can cover at least 20x-25x of the genome size. According to a study by (Minei et al., 2018) established a pipeline of a *de novo* assembly of middle-sized eukaryotic genomes at a low cost and with high quality using long and short reads. Using this approach will increase the base accuracy of the sequence and the reads coverage needed to assemble the genomic DNA sequence of a freshwater mussel.

Study by (Daniel et al., 2014) on genetic damage induced by water pollutants in the freshwater fish, indicates that some aquatic populations are at risk of exposure to genotoxic pollutants in water. Detection on the genetic damage can make a tool available for monitoring and identifying the genotoxicity of pollutants in water ecosystem. Freshwater mussels are important component of ecosystem and can be used as a biological indicator for water pollution. Freshwater mussels can remove a variety of contaminating materials from the water column, which makes them extremely interesting from a biological perspective.

After the data of genomic sequence of a freshwater mussel is obtained, further studies are also required on the gene expression of the specie to validate the possibility of finding a transcriptional biomarker for water pollution. These biomarkers can be detected in the organisms at all conditions but their levels or activities changes upon the pollution.

Acknowledgments

I would like to be thankful to the School of Bioscience at the University of Skövde, Sweden for providing all the necessary equipments and resources for making this thesis project to be completed effectively on time. I would like to give a big thanks to my supervisor Mikael Ejdebäck for his support, advice and valuable time throughout the project. He was the one who read my revisions and helped make some sense of the confusion. Then I would like to thanks my co-supervisor John Baxter for his help in laboratory work and in the bioinformatic analysis part. Extending my thanks to the Associate Professor Annie Jonsson for her help in collecting sample for the thesis project. And Finally, thanks to my husband for his support and help, the way he handles my little munchkin throughout the project and making my way easy to manage my work space. And special thanks to my parents for their tremendous support and hope they had given to me without that hope, this thesis has not been possible. In short, thanks to my family and friends who endured this long process with me, always offering support and love.

References

Aldridge, D. C. (1999). The morphology, growth and reproduction of Unionoida (Bivalvia) in a Fenland waterway, *Journal of Molluscan Study* 65: 47–60.

Boesenberg-Smith, K. A., Pessarakli, M. M., & Wolk, D. M. (2012). Assessment of DNA yield and purity: an overlooked detail of PCR troubleshooting. *Clinical Microbiology* Newsletter, 34(1), 1-6.

Branton, D., Deamer, D. W., Marziali, A., Bayley, H., Benner, S. A., Butler, T., Di Ventra, M., Garaj, S., Hibbs, A., Huang, X., Jovanovich, S. B., Krstic, P. S., Lindsay, S., Ling, X. S., Mastrangelo, C. H., Meller, A., Oliver, J. S., Pershin, Y. V., Ramsey, J. M., Riehn, R., ... Schloss, J. A. (2008). The potential and challenges of nanopore sequencing. *Nature biotechnology*, *26*(10), 1146–1153. Retrieved from https://doi.org/10.1038/nbt.1495

Daniel, B., Daniel, R. M., Marcelo, P. d. B., & Luciano, B. d. S. (2014). Genetic damage induced by water pollutants in the freshwater fish *Hyphessobrycon luetkenii* (Characidae) in a reservoir of the Canela National Forest, Brazil, *Journal of Freshwater Ecology*, 29:2, 295-299. doi: 10.1080/02705060.2013.879539

Daniel, L., Graf, Kevin, S., Cummings. (2007). Review of the systematics and global diversity of freshwater mussel species (Bivalvia: Unionoida), *Journal of Molluscan Studies*, Volume 73, Issue 4, Pages 291–314. Retrieved from https://doi.org/10.1093/mollus/eym029

Deamer, D., Akeson, M., & Branton, D. (2016). Three decades of nanopore sequencing. *Nature biotechnology*, 34(5), 518.

Desjardins, P., & Conklin, D. (2010). NanoDrop microvolume quantitation of nucleic acids. JoVE (*Journal of Visualized Experiments*), (45), e2565.

Elkins, K. M. (2012). *Forensic DNA biology*: *a laboratory manual*. Academic Press.

Foxman, B. (2010). *Molecular tools and infectious disease epidemiology*. Academic Press.

França, L., Carrilho, E., & Kist, T. (2002). A review of DNA sequencing techniques. *Quarterly Reviews of Biophysics*, 35(2), 169-200. doi:10.1017/S0033583502003797

Gansauge, M. T., & Meyer, M. (2013). Single-stranded DNA library preparation for the sequencing of ancient or damaged DNA. *Nature protocols*, *8*(4), 737–748. Retrieved from https://doi.org/10.1038/nprot.2013.038

Gerke, N., Tiedemann, R. (2001). A PCR-based molecular identification key to the glochidia of European freshwater mussels (Unionidae). *Conservation Genetics*, 2, 287–289.

Giordano, F., Aigrain, L., Quail, M. A., Coupland, P., Bonfield, J. K., Davies, R. M., & Yue, J. X. (2017). De novo yeast genome assemblies from MinION, PacBio and MiSeq platforms. *Scientific reports*, 7(1), 1-10.

Green, R. H., S. M. Singh & R. C. Bailey, 1985. Bivalve molluscs as response systems for modelling spatial and temporal environmental patterns. The Science of the Total Environment 46: 147–169.

Gurevich, A., Saveliev, V., Vyahhi, N., & Tesler, G. (2013). QUAST: quality assessment tool for genome assemblies. *Bioinformatics*, *29*(8), 1072-1075.

Gurevich, A., Saveliev, V., Vyahhi, N. & Tesler, G. (2020). QUAST: quality assessment tool for genome assemblies (version 5.1.0) [Computer software]. Retrieved from https://github.com/ablab/quast

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. Retrieved from https://doi.org/10.1186/s13059-016-1103-0

Jonsson, A., Bertilsson, A., Rydgård, M. (2013). Spatial distribution and age structure of the freshwater unionid mussels *Anodonta anatina* and *Unio tumidus*: implications for environmental monitoring. *Hydrobiologia* 711:61–70.

Jung, H., Winefield, C., Bombarely, A., Prentis, P., & Waterhouse, P. (2019). Tools and strategies for long-read sequencing and de novo assembly of plant genomes. *Trends in plant science*.

Kneeland, S. C., Rhymer, J. M. (2007). A molecular identification key for freshwater mussel glochidia encysted on naturally parasitized fish hosts in Maine, USA. *Journal of Molluscan Studies*, 73, 279-282.

Koren, S., Schatz, M. C., Walenz, B. P., Martin, J., Howard, J. T., Ganapathy, G., ... & Phillippy, A. M. (2012). Hybrid error correction and de novo assembly of single-molecule sequencing reads. *Nature biotechnology*, *30*(7), 693-700.

Koren, S., & Walenz, B. (2020). Marbl/Canu (Version 2.0) [Computer software]. Retrieved from https://github.com/marbl/Canu

Koren, S., Walenz, B. P., Berlin, K., Miller, J. R., Bergman, N. H., & Phillippy, A. M. (2017). Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome research,* 27(5), 722-736.

Ku, C. S., & Roukos, D. H. (2013). From next-generation sequencing to nanopore sequencing technology: paving the way to personalized genomic medicine. *Expert review of medical devices*, *10*(1), 1–6. Retrieved from https://doi.org/10.1586/erd.12.63

Leger, A., Leonardi, T. (2019). pycoQC, interactive quality control for Oxford Nanopore Sequencing (Version 3.0) [Computer software]. Retrieved from https://github.com/a-slide/pycoQC. Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, 34:3094- 3100. doi:10.1093/bioinformatics/bty191

Li H. (2018). Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics (Oxford, England)*, *34*(18), 3094–3100. Retrieved from https://doi.org/10.1093/bioinformatics/bty191

Lopes-Lima, M. (2014). *Anodonta anatina*. The IUCN Red List of Threatened Species 2014: e. T155667A21400363. Retrieved from https://doi.org/10.2305/IUCN.UK.2014-1.RLTS.T155667A21400363.en

Lopez, R., Chen, Y. J., Dumas Ang, S., Yekhanin, S., Makarychev, K., Racz, M. Z., Seelig, G., Strauss, K., & Ceze, L. (2019). DNA assembly for nanopore data storage readout. *Nature communications*, *10*(1), 2933. Retrieved from https://doi.org/10.1038/s41467-019-10978-4

Lurman, G., Walter, J., Hoppeler, H. H. (2014). Seasonal changes in the behavior and respiration physiology of the freshwater duck mussel, *Anodonta anatina*. *Journal of Experimental Biology* 217:235–243.

McIntyre, A. B., Rizzardi, L., Angela, M. Y., Alexander, N., Rosen, G. L., Botkin, D. J., & Burton, A. S. (2016). Nanopore sequencing in microgravity. npj *Microgravity*, 2(1), 1-9.

ONT. (2018). Medaka Sequence correction provided by ONT research (Version 1.0.1) [Computer software]. Retrieved from https://github.com/ONT/medaka

Osterling, M., Zulsdorff, V., & Schneider, L. (2012). Host fish species of freshwater mussels in seven Swedish river systems (The thick-shelled river mussel brings back life to rivers. Serious No. 46).

Retrieved from the website of UCforlife: http://www.ucforlife.se/wp-content/uploads/2012/12/7.2.2_TECHNICAL-REPORT_C1a.pdf

Pop, M., Phillippy, A., Delcher, A. L., & Salzberg, S. L. (2004). Comparative genome assembly. *Briefings in bioinformatics*, 5(3), 237-248.

Prince, R. C. (2000). Bioremediation. *Kirk-Othmer Encyclopedia of Chemical Technology*. Retrieved from https://doi.org/10.1002/0471238961.0209151816180914.a01

Prié, V., Valentini, A., Lopes-Lima, M., Froufe, E., Rocle, M., Poulet, N., & Dejean, T. (2020). Environmental DNA metabarcoding for freshwater bivalve biodiversity assessment: methods and results for the Western Palearctic (European sub-region). *Hydrobiologia,* 1-20.

Renaut, S., Guerra, D., Hoeh, W. R., Stewart, D. T., Bogan, A. E., Ghiselli, F., Milani, L., Passamonti, M., & Breton, S. (2018). Genome Survey of the Freshwater Mussel Venustaconcha ellipsiformis (Bivalvia: Unionida) Using a Hybrid De Novo Assembly Approach. *Genome biology and evolution*, *10*(7), 1637–1646. Retrieved from https://doi.org/10.1093/gbe/evy117

Rosenberg, G. (2014). A new critical estimate of named species-level diversity of the recent Mollusca. *American Malacological Bulletin*, 32(2), 308-322.

Sanger, F., Nicklen, S., & Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proceedings of the national academy of sciences*, 74(12), 5463-5467.

Schalamun, M., Nagar, R., Kainer, D., Beavan, E., Eccles, D., Rathjen, J. P., & Schwessinger, B. (2019). Harnessing the MinION: An example of how to establish long-read sequencing in a laboratory using challenging plant tissue from Eucalyptus pauciflora. *Molecular ecology resources*, 19(1), 77-89.

Smith, B., Li, N., Andersen, A. S., Slotved, H. C., & Krogfelt, K. A. (2011). Optimizing bacterial DNA extraction from faecal samples: comparison of three methods. *The Open microbiology journal*, 5, 14.

Strayer, D.L., Downing, J.A., Haag, W.R., King, T.L., Layzer, J.B., Newton, T.J., & Nichols, J. (2004). Changing perspectives on pearly mussels, North America's most imperilled animals. *Bioscience*, 54: 429 –439.

Vaughn, C. C. (2018). Ecosystem services provided by freshwater mussels. *Hydrobiologia*, 810(1), 15-27.

Wang, H., Liu, B., Zhang, Y., Jiang, F., Ren, Y., Yin, L., & Fan, W. (2020). Estimation of genome size using k-mer frequencies from corrected long reads. arXiv preprint arXiv:2003.11817.

Watters, G. T. (1994). Form and function of unionoidean shell sculpture and shape (Bivalvia). *American Malacological Bulletin*, 11, 1–20.

Wick, R. R., Judd, L. M., & Holt, K. E. (2020). Guppy Local accelerated base calling for Nanopore data (Version 3.6) [Computer software]. Retrieved from https://community.ONT.com/downloads

Xu, C., Wu, K., Zhang, J. G., Shen, H., ... & Deng, H. W. (2017). Low-, high-coverage, and two-stage DNA sequencing in the design of the genetic association study. *Genetic epidemiology*, 41(3), 187-197.

Zieritz A, Gum B, Kuehn, Geist J (2012) Identifying freshwater mussels (Unionoida) and parasitic glochidia larvae from host fish gills: a molecular key to the North and Central European species. *Ecology and Evolution*, 2, 740-50.

Zimin, A. V., & Salzberg, S. L. (2020). The genome polishing tool POLCA makes fast and accurate corrections in genome assemblies. *PLoS computational biology*, *16*(6), e1007981. Retrieved from https://doi.org/10.1371/journal.pcbi.1007981