Degree project

UNIVERSITY OF SKÖVDE

Sequencing the Genome of *Anodonta anatina* to Support Future Research on Pollution Impacts

Master Degree Project in Bioscience A1E

Second Cycle 30 credits

Year : 2026

Students:

Muhammed Shanoob Kannamthodi Pariyapurath

Abhijith Puthenveettil Janesh

Supervisor: Magnus Fagerlind

Examiner: Sanja Jurcevic

Abstract

Freshwater ecosystems are increasingly threatened by anthropogenic pollution, habitat degradation and other environmental stressors, creating a need for sensitive biological approaches to assess their effects. Freshwater mussels are valuable bioindicator organisms because their sedentary lifestyle and filter-feeding behaviour allow them to accumulate environmental contaminants and respond to pollution at physiological, biochemical and molecular levels. *Anodonta anatina* is a widespread European freshwater mussel that has previously demonstrated molecular and biochemical responses to environmental contaminants. However, limited genomic resources restrict the investigation of these responses at a genome-wide level. This project aimed to develop a genomic resource for *A. anatina* using Oxford Nanopore long-read sequencing and de novo genome assembly. High-molecular-weight genomic DNA was extracted using two extraction approaches, the Monarch HMW DNA Extraction Kit and the QIAGEN MagAttract HMW DNA Kit. DNA concentration and purity were assessed using fluorometric, spectrophotometric and Pulsed-Field Gel Electrophoresis analyses before library preparation and sequencing. Sequencing reads were assessed using NanoPlot and assembled using Flye, with assembly statistics evaluated using QUAST. The selected QIAGEN and Monarch samples generated draft assemblies of 2.06 Gb and 1.99 Gb, respectively, with N50 values of 102,364 bp and 51,996 bp and 79,055 and 79,491 contigs, respectively. The QIAGEN and Monarch datasets were generated from four and two sequencing runs, respectively; therefore, differences in sequencing yield and assembly metrics cannot be attributed solely to extraction method. Genome completeness was not independently assessed. The assemblies may support future transcriptomic, comparative genomic and population genomic studies of *A. anatina*, including research on environmental responses and conservation.

Popular Scientific Summary

Freshwater mussels live in rivers, lakes and other freshwater environments, where they filter water to obtain food. During this process, they may also take in pollutants from their surroundings. Because they cannot easily move away from polluted areas, they are useful for studying how pollution affects aquatic life. Understanding these effects at the genetic level requires information about the mussel's genome, which contains the instructions that help an organism develop and function.

This study focused on the freshwater mussel *Anodonta anatina* and aimed to produce a draft genome to support future research into pollution impacts. DNA, the material carrying genetic information, was extracted using two different methods. The amount and purity of the DNA were measured. A laboratory technique called pulsed-field gel electrophoresis was also used. This technique separates DNA fragments according to their size, allowing researchers to check whether the DNA remains in large, intact pieces. The best sample from each extraction method was selected for further analysis.

The selected DNA samples were analysed using Oxford Nanopore technology, a method that reads DNA by measuring changes in electrical signals as DNA passes through tiny protein pores. This approach can produce long DNA sequences, which can help researchers piece together a genome. The two datasets generated approximately 39.42 billion and 20.18 billion DNA bases, respectively. One dataset contained more sequence data overall, while the other had longer DNA reads on average.

Computer programs were then used to assemble the DNA sequences, much like putting together overlapping pieces of a puzzle. The QIAGEN method produced a more continuous genome assembly than the Monarch method. This means the assembled DNA was arranged into longer connected sections, rather than being divided into as many shorter pieces. Both methods produced draft genome assemblies of approximately two billion DNA bases. However, the study could not determine how much of the mussel's complete genome was represented.

These draft genomes may provide a starting point for future research into how freshwater mussels respond to pollution. By combining genome information with studies of which genes become more or less active after exposure to pollutants, researchers may identify genes and biological processes involved in environmental stress responses. This could improve our understanding of pollution's effects on freshwater ecosystems and help researchers use mussels to assess environmental health.

Contents

Abstract .....

Popular Scientific Summary .....

List of Abbreviations .....

Introduction 1
- Freshwater ecosystems and environmental pollution 1
- Freshwater mussels as bioindicators 1
- *Anodonta anatina* as a model for environmental monitoring 2
- From targeted biomarkers to genomics for environmental research 2
- Genome complexity and long-read sequencing 3
- High-molecular-weight DNA and genome assembly 4
- Knowledge gap and rationale 4

Aim and Objectives 5

Materials and Methods 5
- Collection and Dissection of *Anodonta anatina* 5
- High Molecular Weight DNA extraction 6
  - DNA extraction using the Monarch HMW DNA Kit 6
  - DNA extraction using the QIAGEN MagAttract HMW DNA Kit 6
- DNA Quality Assessment 6
  - DNA concentration measurement using Qubit fluorometry 7
  - DNA purity assessment using NanoDrop spectrophotometry 7
  - Pulsed-Field Gel Electrophoresis (PFGE) 7
- Library Preparation 7
- Oxford Nanopore Sequencing 8
- Bioinformatic Analysis 8
  - Read Quality Assessment 8
  - Genome Assembly 9
  - Assembly Quality Assessment 9

Results 9
- DNA extraction and quality assessment 9
- PFGE Analysis of Genomic DNA Integrity 10
- Oxford Nanopore library preparation and sequencing 11
- Read-quality assessment using NanoPlot 11
- De novo genome assembly using Flye 13
- Genome-assembly quality assessment using QUAST 13
- Overall comparison of sequencing and assembly results 14

Discussion 14
- DNA quality and suitability for long-read sequencing 15
- Sequencing output and read quality 15
- Read-length distribution and implications for assembly 16
- De novo assembly and comparison of the two assemblies 16
- Relevance to the biological aim of the study 17
- Methodological limitations 17
- Conclusion 18

Ethical aspects, gender equality perspective, and impact on the society 18

Future perspective 19

AI Declaration 20
References 20
Appendix 1 24
Appendix 2 24
Appendix 3 24
Appendix 4 24

List of Abbreviations

bp	Base pair(s)
BSA	Bovine serum albumin
BUSCO	Benchmarking Universal Single-Copy Orthologs
DNA	Deoxyribonucleic acid
EB	Elution Buffer
FCF	Flow Cell Flush
FCT	Flow Cell Tether
Gb	Giga base(s)
GC	Guanine–cytosine
HAC	High accuracy (base calling mode)
HMW	High-molecular-weight
kb	Kilobase(s)
LA	Ligation Adapter
LFB	Long Fragment Buffer
ONT	Oxford Nanopore Technologies
PFGE	Pulsed-field gel electrophoresis
QC	Quality control
QUAST	Quality Assessment Tool for Genome Assemblies
SB	Sequencing Buffer
TBE	Tris-borate-EDTA

Introduction

Freshwater ecosystems and environmental pollution

Freshwater ecosystems, including rivers, streams, lakes and wetlands, support a large proportion of global biodiversity and provide important ecosystem services such as water purification, nutrient cycling and habitat provision (Dudgeon et al., 2006; Vörösmarty et al., 2010). However, they are increasingly threatened by pollution, habitat degradation, invasive species and climate change (Reid et al., 2019; Tickner et al., 2020).

Pollution from agriculture, industry and urban activities introduces contaminants such as heavy metals, pesticides and pharmaceuticals into freshwater environments (Schwarzenbach et al., 2006; aus der Beek et al., 2016). These contaminants can cause physiological, biochemical and molecular responses in aquatic organisms, making biological monitoring useful alongside chemical measurements (Regoli & Giuliani, 2014). Freshwater mussels are valuable organisms for environmental research because their filter-feeding lifestyle and limited mobility expose them to local environmental conditions (Lopes-Lima et al., 2018). However, studying pollution responses at the genetic level requires reliable genomic information. A draft genome of *Anodonta anatina* could therefore provide a foundation for future studies investigating genes and biological pathways involved in responses to environmental stress and pollution (Ekelund Ugge et al., 2020).

Freshwater mussels as bioindicators

Freshwater mussels belonging to the order Unionida are important components of freshwater ecosystems and are among the most threatened aquatic animal groups. Through filtration, nutrient cycling and interactions with sediments, they influence ecosystem functioning and contribute to habitat complexity (Vaughn & Hoellein, 2018). Their ecological importance and widespread population declines have increased interest in understanding their biology, genetics and responses to environmental change (Lopes-Lima et al., 2018). Genomic resources can support studies of genetic variation, adaptation and the molecular mechanisms underlying responses to environmental stress (Gerdol et al., 2020).

The conservation status of freshwater mussels further increases the importance of understanding their responses to environmental stress. European freshwater mussel populations have experienced substantial declines because of pollution, habitat modification, hydrological changes, invasive species and changes in host-fish communities (Lopes-Lima et al., 2017; Lopes-Lima et al., 2018). Their complex reproductive biology can further increase their vulnerability because many unionid species require suitable fish hosts during their larval stage (Haag, 2012). Understanding how freshwater mussels respond to environmental stress at the molecular level is therefore relevant both to ecological monitoring and conservation.

The use of bivalves as biological indicators is well established. Their capacity to accumulate contaminants has been demonstrated for numerous metals and organic pollutants, while physiological and biochemical responses can provide additional evidence of environmental stress (Goldberg, 1975; Rainbow, 1995).

Anodonta anatina as a model for environmental monitoring

The duck mussel, *Anodonta anatina* (Linnaeus, 1758), is a widely distributed European freshwater mussel belonging to the family Unionidae. Its distribution and ecological characteristics make it a relevant organism for investigating environmental contamination. Previous studies have demonstrated that *A. anatina* can accumulate metals in its tissues and exhibit biochemical and transcriptional responses following contaminant exposure (Ekelund Ugge et al., 2020; Ekelund Ugge et al., 2022).

Early research on *Anodonta anatina* from the River Thames demonstrated the accumulation of several metals, including zinc, nickel, lead, cadmium, copper, and mercury, and highlighted the potential of this species as a bioindicator of aquatic pollution (Manly & George, 1977). Subsequent experimental studies have examined molecular responses to trace metals including copper, zinc and cadmium and identified changes in metallothionein-related responses and other cellular stress mechanisms (Falfushynska et al., 2013). These findings demonstrate that *A. anatina* can respond biologically to environmental metal exposure rather than simply accumulating contaminants.

More recent studies have investigated transcriptional and biochemical responses to environmentally relevant concentrations of copper. Ekelund Ugge et al. (2020) reported changes in biomarkers associated with antioxidant defence, xenobiotic responses and cellular stress, including catalase (cat), glutathione S-transferase (gst), heat-shock proteins (hsp70 and hsp90), metallothionein (mt), superoxide dismutase (sod) and acetylcholinesterase. Responses to industrial wastewater have also been observed, with changes detected in selected transcriptional and biochemical biomarkers in tissues of *A. anatina* (Ekelund Ugge et al., 2022). These targeted studies provide evidence that *A. anatina* responds to environmental contaminants at the molecular level, but they focus on selected biomarkers rather than the genome. The draft genome generated in the present study provides a broader genomic resource that can support future studies investigating these and other pollution-responsive genes and pathways at a genome-wide level.

From targeted biomarkers to genomics for environmental research

Although targeted biomarker studies provide important information about contaminant exposure, they are limited by the number of genes or pathways selected for investigation. Environmental contaminants can affect multiple cellular processes simultaneously, including oxidative stress, xenobiotic metabolism, DNA repair, apoptosis, immune regulation and energy metabolism. Consequently, investigating only a small number of biomarkers may provide an incomplete picture of the biological response (Ekelund Ugge et al., 2022).

The development of next-generation sequencing technologies has made genome-scale investigations increasingly accessible for non-model organisms. Genome sequencing can provide a reference framework for gene annotation, comparative genomics and the investigation of genetic variation (Ekblom & Galindo, 2011; Ellegren, 2014). A reference genome provides an assembled representation of an organism's genomic sequence, supporting the identification and annotation of genes and other genomic features. When a suitable reference genome is available, transcriptomic data can be analysed in relation to the genomic sequence, facilitating the identification and interpretation of expressed genes (Hornett & Wheat, 2012).

For *Anodonta anatina*, an improved reference genome would therefore provide an important foundation for future environmental genomics research. Once genes and genomic regions have

been identified and annotated, transcriptomic experiments could investigate genome-wide changes in gene expression following exposure to metals, pesticides, pharmaceuticals and other contaminants. This could enable the identification of genes and molecular pathways associated with oxidative stress, metal detoxification, xenobiotic metabolism, DNA repair, immune responses, apoptosis and energy metabolism. Such genome-wide approaches may provide a more comprehensive understanding of contaminant-induced biological responses than studies focusing on a limited number of targeted biomarkers (Balbi et al., 2021; Ekelund Ugge et al., 2022).

Genomic resources are increasingly being developed for freshwater mussels (Smith et al. (2018) generated a draft genome for *Venustaconcha ellipsiformis*, demonstrating the feasibility of genomic analysis in unionid mussels. Improved assemblies for other species, including *Margaritifera margaritifera*, have further demonstrated the value of long-read sequencing for freshwater mussel genomics (Gomes-dos-Santos et al., 2023). A draft genome of *A. anatina* is also available and was generated using PacBio HiFi sequencing and assembly with hifiasm. The assembly has a size of approximately 2.46 Gb distributed across 2,008 scaffolds, with a scaffold N50 of 2.7 Mb and a Benchmarking Universal Single-Copy Orthologs (BUSCO) completeness score of 83.5% (Faust et al., 2025). Although this represents an important genomic resource for the species, additional genomic data and assembly approaches may further improve genome characterisation and support future genomic applications. Together, these studies demonstrate the growing importance of genomic resources for understanding freshwater mussel biology and provide the scientific context for generating improved genomic data for *A. anatina*.

In addition to environmental research, genomic resources may contribute to conservation studies. Population genomic analyses can provide information about genetic diversity, population connectivity and demographic history, while comparisons among populations may identify genetic variation associated with different environmental conditions (Ellegren, 2014). Such information could complement ecological monitoring and support conservation strategies for freshwater mussels, many of which are experiencing substantial population declines (Lopes-Lima et al., 2017; Tickner et al., 2020).

Genome complexity and long-read sequencing

Genome assembly can be challenging when genomes are large, repetitive or highly heterozygous. Short-read sequencing technologies provide highly accurate sequences but may be insufficiently long to span repetitive genomic regions, potentially resulting in fragmented assemblies (Logsdon et al., 2020). The recently reported *A. anatina* draft genome also illustrates these challenges, as the assembly was highly fragmented and the available sequencing data had relatively short read lengths and limited genome coverage (Faust et al., 2025). These characteristics highlight the importance of obtaining sufficiently long reads and adequate sequencing coverage when generating genomic resources for *A. anatina.*

Long-read sequencing technologies provide an effective approach for overcoming some of these limitations. Platforms such as Pacific Biosciences and Oxford Nanopore Technologies generate reads that can span repetitive regions and structural variants that may be difficult to resolve using short-read data (Jain et al., 2018; Logsdon et al., 2020). Longer reads can therefore improve contiguity and facilitate more accurate reconstruction of complex genomic regions.

Oxford Nanopore Technologies (ONT) is well suited to de novo genome sequencing because it generates long reads and provides data in real time. In this study, sequencing was performed

using an R10.4.1 (FLO-MIN114) flow cell, with basecalling performed through MinKNOW. The ability to obtain long reads depends strongly on DNA integrity; therefore, high-molecular-weight (HMW) DNA is important for long-read sequencing and genome assembly (Jain et al., 2018; Logsdon et al., 2020).

High-molecular-weight DNA and genome assembly

DNA fragmentation during extraction, pipetting or library preparation can reduce sequencing read length and limit the ability of long-read assemblers to resolve repetitive regions. Optimisation of HMW DNA extraction is therefore an important component of long-read genome projects (Jones et al., 2021; Trigodet et al., 2022). In the present study, the Monarch HMW DNA Extraction Kit and QIAGEN MagAttract HMW DNA Kit are considered for extraction of genomic DNA from *A. anatina*. Comparing extraction approaches can help identify a protocol capable of producing DNA with suitable concentration, purity and molecular integrity.

Following DNA extraction and quality assessment, Oxford Nanopore Technologies (ONT) sequencing generates long reads suitable for *de novo* genome assembly. *De novo* assembly reconstructs a genome directly from sequencing reads without relying on an existing reference genome and is therefore particularly suitable for non-model species with limited genomic resources. Several assemblers have been developed for long-read sequencing data, including Canu and Flye (Koren et al., 2017; Kolmogorov et al., 2019). In this project, Flye was selected in advance for genome assembly rather than being compared experimentally with alternative assemblers. The choice was based on its specific design for assembling long and error-prone reads using a repeat-graph approach, which is advantageous for complex genomes containing repetitive regions (Kolmogorov et al., 2019). Flye also supports Oxford Nanopore sequencing data and provides an automated workflow for generating *de novo* assemblies. Therefore, Flye was considered appropriate for generating a draft genome assembly of the freshwater mussel from the ONT sequencing reads. Genome assemblies must subsequently be evaluated using multiple quality metrics. QUAST provides information about assembly size, number of contigs, N50, L50, longest contig and GC content (Gurevich et al., 2013).

Knowledge gap and rationale

Despite increasing availability of genome assemblies for unionid mussels, genomic resources remain unevenly distributed among freshwater mussel species. *A. anatina* has already been used extensively in environmental biomonitoring and has demonstrated measurable biochemical and molecular responses to contaminants (Falfushynska et al., 2013; Ekelund Ugge et al., 2020; Ekelund Ugge et al., 2022). However, although a draft genome assembly is now available for *A. anatina*, further genomic data and improved assembly approaches may help extend these targeted studies towards genome-wide investigations (Faust et al., 2025).

Although a draft genome assembly for *A. anatina* is available, its characteristics and level of fragmentation may limit some applications requiring a highly contiguous genomic resource. Additional long-read sequencing data and alternative assembly approaches may therefore complement existing genomic resources and support future transcriptome analysis, comparative genomics and investigations of molecular pathways involved in environmental responses (Faust et al., 2025).

The present study focuses on generating additional long-read genomic data for *Anodonta anatina* through the evaluation of HMW DNA extraction methods, assessment of DNA integrity, Oxford Nanopore long-read sequencing and de novo genome assembly. By comparing HMW DNA

extraction methods and applying a workflow combining DNA quality assessment, long-read sequencing and systematic assembly evaluation, this study aims to assess an alternative approach for generating genomic data for the species. The resulting genomic data and draft assembly may support future investigations of how freshwater mussels respond to environmental pollution at the molecular level.

Aim and Objectives

The aim of this study was to generate and evaluate a draft genome assembly of the freshwater mussel *Anodonta anatina* using Oxford Nanopore long-read sequencing technology. The study aimed to generate additional long-read genomic data for the species and evaluate a workflow based on high molecular weight DNA extraction, long-read sequencing and *de novo* genome assembly. The resulting genomic data and draft assembly may complement existing genomic resources for *A. anatina* and support future genomic and environmental research.

To achieve this aim, the following objectives were established:

- To extract high-molecular-weight genomic DNA from *Anodonta anatina* using the Monarch HMW DNA Extraction Kit and the QIAGEN MagAttract HMW DNA Kit, and to evaluate the quality and sequencing performance of the analysed DNA samples generated by the two extraction workflows.
- To compare the quality, purity, concentration, and integrity of DNA obtained from both extraction methods using Qubit fluorometry, NanoDrop spectrophotometry, and Pulsed-Field Gel Electrophoresis (PFGE).
- To prepare a long-read sequencing library using the Oxford Nanopore Ligation Sequencing Kit (SQK-LSK114) and perform genome sequencing using the Oxford Nanopore MinION platform.
- To assess sequencing read characteristics using NanoPlot, perform *de novo* genome assembly using Flye, and evaluate assembly contiguity and basic assembly statistics using QUAST.

Materials and Methods

Collection and Dissection of *Anodonta anatina*

A total of eight *Anodonta anatina* specimens were collected from Lake Axvall, Sweden, during spring 2026. Collection and species identification were performed based on shell morphology, including shell shape, size, hinge structure, and umbo characteristics, using standard freshwater mussel identification keys with the assistance of Professor Annie Jonsson (University of Skövde). Following collection, the mussels were individually maintained in freshwater at 4 °C and dissected within three days to minimise deterioration. Before dissection, the shells were gently rinsed with water to remove surface debris and opened using sterile dissection instruments. Foot and gill tissues were aseptically excised using sterile forceps and scissors, transferred to sterile 1.5 mL microcentrifuge tubes, and kept on ice until DNA extraction. Twelve tissue samples were obtained from the eight specimens, with Mussels 7 and 8 each contributing two additional tissue samples. Six samples (three gill and three foot) were processed using the Monarch HMW DNA Extraction Kit (New England Biolabs), and six (three gill and three foot) using the QIAGEN MagAttract HMW DNA Kit (QIAGEN).

High Molecular Weight DNA extraction

DNA extraction using the Monarch HMW DNA Kit

High-molecular-weight genomic DNA was extracted from gill and foot tissue using the Monarch HMW DNA Extraction Kit for Tissue (New England Biolabs) according to the manufacturer's protocol, with DNA handling performed carefully to preserve DNA integrity for subsequent Oxford Nanopore sequencing (New England Biolabs).

Fresh gill tissue (19 mg) and foot tissue (20 mg) were transferred separately to sterile microcentrifuge tubes containing the master mix of 600 µl HMW gDNA lysis buffer and 20 µl proteinase K. The tissue was mechanically homogenized using a rotor–stator homogenizer at the lowest instrument setting for 10 seconds, until no visible tissue fragments remained, following the manufacturer's recommendations (New England Biolabs). The sample was incubated at 56 °C for 45 minutes with shaking at 600 rpm in a thermoshaker. Tissue lysis was monitored visually during incubation, and the samples were considered lysed when no visible tissue fragments remained. Following lysis,10 µl of RNase A was added to each sample was incubated at 56°C for 10 minutes to remove contaminating RNA from the genomic DNA preparation.

Throughout the extraction procedure, careful handling was maintained to minimise mechanical shearing of genomic DNA. Samples were mixed by gentle inversion rather than vortexing, and pipetting was performed slowly and carefully. These precautions helped preserve long DNA fragments required for long-read sequencing. Following DNA purification, genomic DNA was eluted using 100 µl Monarch DNA Elution Buffer II. The eluted DNA was carefully transferred into sterile microcentrifuge tubes and stored at -18°C.

DNA extraction using the QIAGEN MagAttract HMW DNA Kit

Genomic DNA was also extracted using the QIAGEN MagAttract HMW DNA Kit (QIAGEN) following the manufacturer's recommended protocol (QIAGEN). This magnetic bead-based extraction method was included to compare DNA yield, purity, and fragment integrity with the Monarch extraction method.

Fresh gill tissue (22 mg) and foot tissue (20 mg) were lysed separately in 220 µL Buffer ATL and 20 µL Proteinase K by incubation at 56°C for 16 hours. A 200 µL aliquot of each lysate was transferred to a new tube and mixed with 4 µL RNase A, 150 µL Buffer AL, 280 µL Buffer MB, and 40 µL MagAttract Suspension G. After magnetic bead binding, the DNA was washed twice with 700 µL Buffer MW1, twice with 700 µL Buffer PE, and twice with 700 µL distilled water before elution in 100–200 µL Buffer AE. To minimise DNA fragmentation, vigorous mixing and unnecessary pipetting were avoided throughout the procedure. Purified genomic DNA was stored at 4°C until further analysis.

For each extraction kit, six tissue samples (three gill and three foot) from eight mussels were analysed. Corresponding tissues from the same individuals were processed using both workflows, with Mussels 7 and 8 providing additional samples. No technical extraction replicates were performed. Samples were ranked by Qubit concentration, A260/A280 and A260/A230 ratios, and PFGE integrity before selecting one gill extract per workflow.

DNA Quality Assessment

The quality of the extracted genomic DNA was evaluated using complementary analytical methods to determine DNA concentration, purity, and fragment integrity before library preparation.

DNA concentration measurement using Qubit fluorometry

DNA concentration was measured using a Qubit 4 Fluorometer (Thermo Fisher Scientific) and the Qubit dsDNA High Sensitivity (HS) Assay Kit (Thermo Fisher Scientific) according to the manufacturer's instructions.

DNA purity assessment using NanoDrop spectrophotometry

DNA purity was evaluated using a NanoDrop™ spectrophotometer (Thermo Fisher Scientific). Prior to measurement, the instrument was blanked using nuclease-free water. A 1 µL aliquot of each DNA sample was analysed to determine absorbance. The A260/A280 and A260/A230 absorbance ratios were recorded for each DNA sample and used to assess DNA purity.

Pulsed-Field Gel Electrophoresis (PFGE)

The integrity and molecular size distribution of the extracted genomic DNA were assessed using pulsed-field gel electrophoresis (PFGE). A 1% pulsed-field-certified agarose gel was prepared by dissolving 1.5 g of pulsed-field-certified agarose in 135 mL of ultrapure water, followed by the addition of 15 mL of 5× TBE buffer to obtain a final volume of 150 mL containing 0.5× TBE. The gel was prepared following the recommendations for PFGE using the CHEF-DR III Pulsed Field Electrophoresis System (Bio-Rad Laboratories).

An aliquot of 4 mL molten agarose was maintained at 60 °C for sealing the wells, while the remaining gel was cast and allowed to solidify. A Lambda DNA ladder (New England Biolabs) was loaded into the first and fourteenth wells as a high-molecular-weight DNA size marker. The gel was run in a CHEF-DR III Pulsed Field Electrophoresis System containing recirculated 0.5× TBE buffer maintained at 14 °C. Electrophoresis was performed for 22 h at 6 V/cm with a switch-time ramp of 50–90 s and an included angle of 120°. The gel was then stained with GelRed (New England Biolabs) at a final concentration of 1×, visualised using a gel documentation system, and the DNA migration pattern was compared with the Lambda DNA ladder to assess DNA integrity and fragment size distribution.

A volume of 8 µL of each DNA sample was mixed with 2 µL of 5× DNA loading dye (QIAGEN) by gentle pipetting. The mixtures were carefully loaded into the designated wells of the agarose gel. A Lambda DNA ladder was loaded into the first and fourteenth wells of the gel as a high-molecular-weight DNA size marker to enable estimation of DNA fragment size.

Library Preparation

The highest-quality genomic DNA sample obtained from the gill tissue of a different individual mussel for each DNA extraction method was selected for library preparation. Libraries were prepared using the Oxford Nanopore Ligation Sequencing Kit V14 (SQK-LSK114) according to the manufacturers protocol. For library preparation, 1 µg of genomic DNA was used as input, corresponding to 8.9 µl of QIAGEN-extracted DNA (113 ng/µl) or 9.0 µl of Monarch-extracted DNA (111 ng/µl). Four libraries were prepared using the same selected QIAGEN DNA extract, while two libraries were prepared using the same selected Monarch DNA extract. For each kit, two sequencing runs were performed on the same flow cell, with the QIAGEN dataset comprising four runs and the Monarch dataset comprising two runs. For the DNA repair and end-preparation step, 47 µl of DNA was combined with 1 µl DNA CS, 7 µl NEBNext FFPE DNA Repair Buffer v2, 2 µL NEBNext FFPE DNA Repair Mix, and 3 µl Ultra II End-prep Enzyme Mix, resulting in a final reaction volume of 60 µl.

The reaction was incubated in a thermal cycler at 20°C for 5 minutes, followed by 65°C for 5 minutes, and then cooled to 4°C until further processing. The repaired DNA was purified using 60 µL AMPure XP beads, washed twice with freshly prepared 80% ethanol, air-dried for approximately 30 seconds, and eluted in 61 µL nuclease-free water.

Adapter ligation was performed by mixing 60 µL repaired DNA, 5 µL Ligation Adapter (LA), 25 µL Ligation Buffer (LNB), and 10 µL Salt-T4 DNA Ligase, resulting in a final reaction volume of 100 µL, followed by incubation at room temperature for 10 minutes. The ligated DNA library was purified using 40 µL AMPure XP beads and washed twice with 250 µL Long Fragment Buffer (LFB). The purified library was eluted in 15 µL Elution Buffer (EB), quantified using a Qubit fluorometer, and 12 µL was used for preparation of the final sequencing library.

Oxford Nanopore Sequencing

Sequencing was performed on an Oxford Nanopore MinION (R10.4.1 flow cell, FLO-MIN114) and operated and monitored using MinKNOW software (Oxford Nanopore Technologies) to monitor flow cell status, pore activity, yield, and data acquisition in real time. Prior to loading, flow cells were primed according to the manufacturer's instructions. To prepare the priming mix (per flow cell), combine in a suitable tube at room temperature and mix by pipetting: Flow Cell Flush (FCF) 1,170 µL, Bovine Serum Albumin (BSA, 50 mg/mL) 5 µL, and Flow Cell Tether (FCT) 30 µL (total 1,205 µL).

The final sequencing library was prepared by mixing 12 µL DNA library, 37.5 µL Sequencing Buffer (SB), and 25.5 µL Library Beads (LIB), resulting in a total loading volume of 75 µL. The library was mixed gently by pipetting immediately before loading and carefully loaded dropwise through the SpotON sample port onto an R10.4.1 (FLO-MIN114) MinION flow cell. Sequencing was initiated using MinKNOW software according to the manufacturer's instructions.

Sequencing was performed using MinKNOW software version 25.09.16, which monitored flow cell performance and data acquisition throughout the sequencing run. Basecalling was performed using the dna_r10.4.1_e8.2_400bps_hac basecalling model (version 5.2.0) in high-accuracy (HAC) mode to generate FASTQ files for subsequent bioinformatic analysis.

Bioinformatic Analysis

Read Quality Assessment

Following Oxford Nanopore sequencing, reads were basecalled and converted to FASTQ format using MinKNOW. Reads with quality scores ≥Q10 were classified as passed, and only passed reads were retained. No additional filtering, trimming, or read selection was performed before NanoPlot or Flye analysis. FASTQ files from four QIAGEN and two Monarch sequencing runs were merged separately using command-line processing. Because sequencing effort differed, read yield and assembly metrics were interpreted accordingly rather than as direct extraction-method comparisons. The merged datasets were analysed using NanoPlot (De Coster et al., 2018) for read-quality assessment and visualisation. Commands for FASTQ merging and NanoPlot analysis are provided in Appendices 1 and 2, respectively.

NanoPlot (version 1.46.2) was used to generate summary statistics and graphical outputs, including the total number of reads, total bases sequenced, read length distribution, mean read length, maximum read length, N50 read length, and read quality scores. These metrics were used to evaluate the overall sequencing quality and confirm the suitability of the long-read datasets for downstream de novo genome assembly (Amarasinghe et al., 2020).

Genome Assembly

De novo genome assembly was performed using Flye (version 2.9.6-b1802) (Kolmogorov et al., 2019). The merged FASTQ files from the QIAGEN and Monarch samples were used as input, and separate assemblies were generated for each dataset. A genome-size estimate of 2 Gb was specified in Flye as an approximate value based on previously available information for *Anodonta anatina*. A recently published *A. anatina* draft assembly was approximately 2.8 Gb in assembled sequence length (Faust et al., 2025). The resulting assemblies were subsequently assessed for basic assembly statistics and contiguity using QUAST. All commands used for Flye are provided in Appendix 3.

Assembly Quality Assessment

The de novo genome assemblies were evaluated separately using QUAST (Quality Assessment Tool for Genome Assemblies; version 5.3.0) with identical parameters (Gurevich et al., 2013). Flye-generated contigs from the QIAGEN and Monarch datasets were analysed for number of contigs, total assembly length, largest contig, GC content, N50, L50, and ambiguous bases (Ns). These metrics were used to describe assembly size and contiguity. As no suitable reference genome was used, QUAST results were not interpreted as measures of assembly accuracy or biological completeness. Differences in sequencing effort between the QIAGEN (four runs) and Monarch (two runs) datasets were considered when comparing assembly statistics. The exact QUAST command and parameters are provided in Appendix 4.

Results

DNA extraction and quality assessment

The DNA quality results for the *A. anatina* samples extracted using the Monarch HMW DNA Extraction Kit and QIAGEN MagAttract HMW DNA Kit are summarised in Table 1. The A260/A280 ratios ranged from 1.70 to 1.93 for Monarch and from 1.71 to 1.97 for QIAGEN, while A260/A230 ratios ranged from 0.71 to 1.52 and 0.45 to 2.88, respectively. The highest-quality samples selected for subsequent library preparation were Monarch Gill-1 and QIAGEN Gill-1, with concentrations of 111 and 113 ng/µL, respectively. The quantitative results are summarised in Table 1.

Table 1. DNA quality metrics obtained with the Monarch HMW DNA Extraction Kit and the QIAGEN MagAttract HMW DNA Kit.

Kits	Tissue	Qubit (ng/µL)	A260/A280	A260/A230
Monarch	Gill-1	111	1.93	1.20
Monarch	Gill-2	62.3	1.86	1.00
Monarch	Gill-3	59	1.85	0.71
Monarch	Foot-1	31.4	1.91	1.06
Monarch	Foot-2	46.5	1.80	1.42
Monarch	Foot-3	37.4	1.70	1.52
Qiagen	Gill -1	113	1.89	2.30
Qiagen	Gill -2	59.1	1.97	2.88
Qiagen	Gill -3	38.6	1.71	0.45

Qiagen	Foot -1	37.9	1.80	1.06
Qiagen	Foot- 2	40.6	1.93	1.00
Qiagen	Foot-3	49.4	1.77	1.00

To obtain approximately 1 µg of DNA for library preparation, 9.0 µL of the Monarch extract and 8.9 µL of the QIAGEN extract were used. The selected Monarch extract had an A260/A280 ratio of 1.93 and an A260/A230 ratio of 1.20, while the corresponding values for the QIAGEN extract were 1.89 and 2.30, respectively.

PFGE Analysis of Genomic DNA Integrity

Pulsed-field gel electrophoresis (PFGE) was performed to evaluate the integrity and molecular size distribution of genomic DNA from 12 *Anodonta anatina* tissue samples. The DNA molecular-weight ladder was loaded in the first lane, followed by the genomic DNA samples. The 12 samples were assessed based on the intensity and appearance of high-molecular-weight DNA bands, together with their DNA concentration and purity measurements. The Monarch-extracted gill DNA in lane 3 and the QIAGEN-extracted gill DNA in lane 6 showed comparatively better DNA quality and were selected for further analysis (Figure 1). Both selected samples showed detectable high-molecular-weight DNA, although the DNA bands appeared diffuse rather than sharp and discrete, with some smearing across the gel.

The remaining DNA samples showed varying degrees of DNA migration and high-molecular-weight DNA signals, with differences in signal intensity and distribution between lanes. Some samples displayed weaker or more diffuse signals. The observed migration patterns indicated variation in DNA fragment distribution among the samples. Overall, PFGE showed the presence of HMW DNA in the extracts. Based on the PFGE results together with the DNA concentration and purity measurements, the Monarch gill sample (lane 3) and QIAGEN gill sample (lane 6) were selected for subsequent Oxford Nanopore library preparation and sequencing.

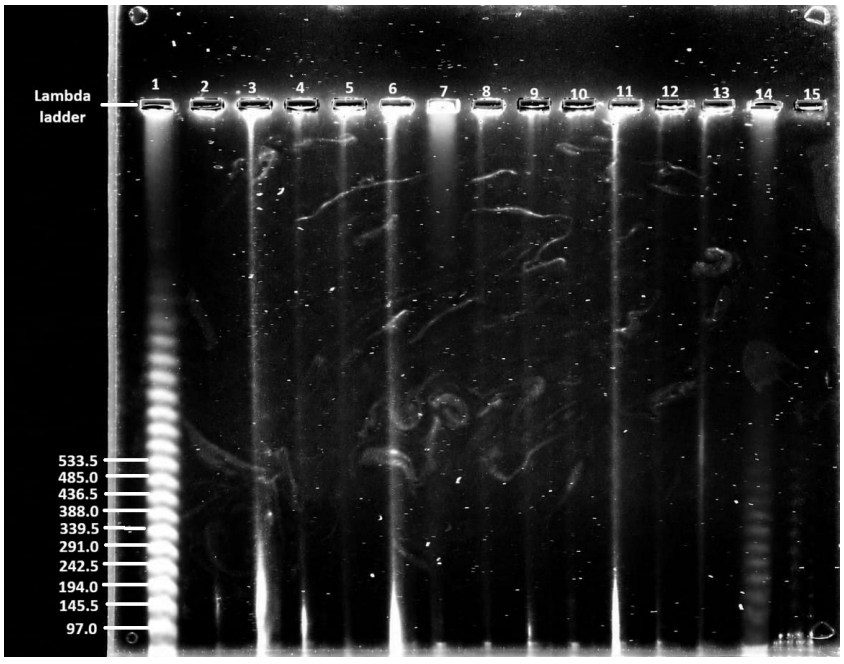


Figure 1. Pulsed-field gel electrophoresis (PFGE) analysis of genomic DNA extracted from *Anodonta anatina* gill and foot tissues using the Monarch and QIAGEN extraction methods. Lanes 1 and 14: DNA

molecular-weight ladder; lanes 2–4: Monarch gill tissue samples; lanes 5–7: QIAGEN gill tissue samples; lanes 8–10: Monarch foot tissue samples; lanes 11–13: QIAGEN foot tissue samples.

Oxford Nanopore library preparation and sequencing

Libraries were prepared from 1 µg of each selected DNA extract using the Oxford Nanopore Ligation Sequencing Kit V14 (SQK-LSK114). The input volumes were 9.0 µL of the Monarch extract (111 ng/µL) and 8.9 µL of the QIAGEN extract (113 ng/µL). The prepared libraries were sequenced using Oxford Nanopore MinION flow cells. The QIAGEN sample was sequenced in four runs, while the Monarch sample was sequenced in two runs. The resulting FASTQ files were used for subsequent read-quality assessment and genome assembly.

Read-quality assessment using NanoPlot

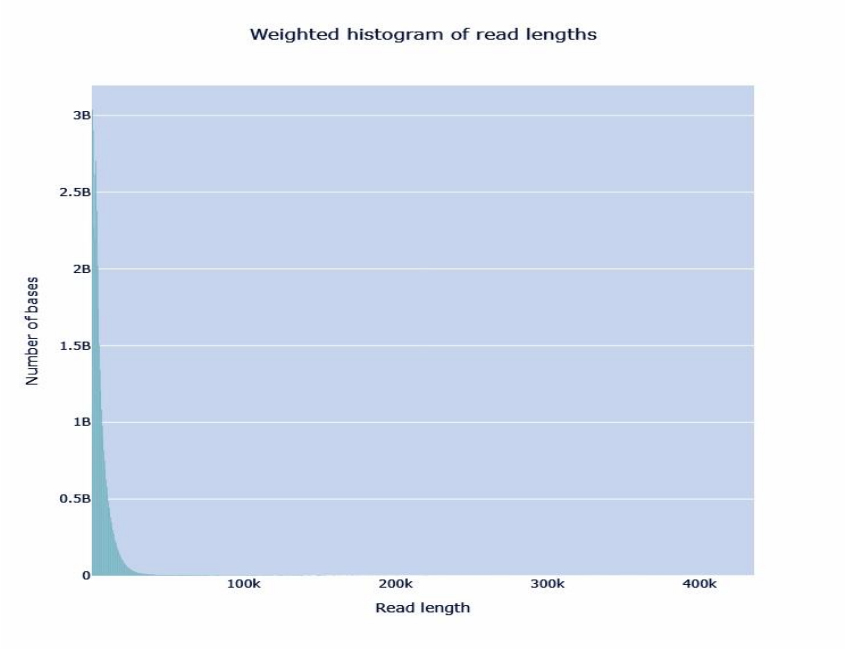
FASTQ files generated by the MinKNOW basecalling workflow were merged separately for each extraction approach and analysed using NanoPlot. The combined QIAGEN dataset contained 23,404,683 reads and 39.42 Gb of sequence data, whereas the combined Monarch dataset contained 5,267,181 reads and 20.18 Gb. The mean read lengths were 1.68 kb for QIAGEN and 3.83 kb for Monarch, with median read lengths of 0.62 and 2.94 kb, respectively. The longest reads were 435.72 kb for QIAGEN and 685.87 kb for Monarch. Read-length N50 values were 3.90 kb and 5.98 kb, respectively. Mean read-quality scores were Q14.7 for QIAGEN and Q15.5 for Monarch (Table 2).

Table 2. Oxford Nanopore sequencing statistics generated from DNA extracted with the two kits.

Parameter	Monarch Kit	QIAGEN Kit
Total reads	5 267 181	23 404 683
Total bases (Gb)	20.18	39.42
Mean read length (kb)	3.83	1.68
Median read length (kb)	2.94	0.62
Longest read (kb)	685.87	435.72
Read-length N50 (kb)	5.98	3.90
Mean read quality	15.5	14.7

For the QIAGEN dataset, the read-length distribution was concentrated below 5 kb, with peaks around 200–500 bp and a broader distribution around 2–3 kb (Figure 2). Regarding read quality, 96.5% of reads exceeded Q10, 71.4% exceeded Q15, 21.0% exceeded Q20, and 0.1% exceeded Q25.

A. QIAGEN



B. Monarch

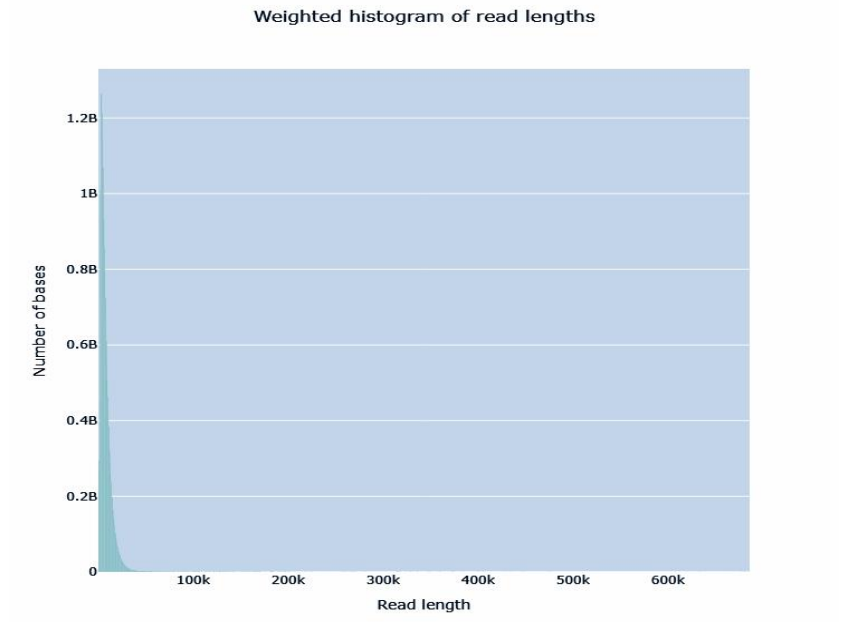


Figure 2. Weighted read-length distributions of Oxford Nanopore sequencing reads from (A) the Monarch dataset and (B) the QIAGEN dataset. The x-axis represents read length in base pairs (bp), and the y-axis represents the number of bases contributed by reads within each read-length bin. The shaded distributions show the weighted read-length histograms, in which reads are weighted by length to represent their contribution to the total sequenced bases.

For the Monarch dataset, the read-length distribution was also concentrated below 5 kb, with a more pronounced distribution around 2–5 kb (Figure 2). Regarding read quality, 96.5% of reads exceeded Q10, 71.4% exceeded Q15, 21.0% exceeded Q20, and 0.1% exceeded Q25. The length-versus-quality plot did not show a clear relationship between read length and quality score in either dataset.

De novo genome assembly using Flye

The sequencing datasets generated from the QIAGEN and Monarch DNA samples were assembled into draft genome sequences using Flye. The resulting assemblies were obtained as FASTA files and subsequently assessed using QUAST. The Monarch assembly had a total length of 1.99 Gb, while the QIAGEN assembly had a total length of 2.06 Gb, compared with the 2 Gb genome-size estimate specified during Flye assembly. The assembly characteristics, including assembly size, number of contigs, contig length distribution, and assembly contiguity, are presented in the subsequent QUAST analysis.

Genome-assembly quality assessment using QUAST

The two Flye assemblies were further evaluated using QUAST to obtain standardised assembly statistics, including contiguity and compositional metrics, as presented in Table 3. The total assembly lengths were similar, measuring 1.99 Gb for the Monarch assembly and 2.06 Gb for the QIAGEN assembly. These metrics showed greater contiguity in the QIAGEN assembly compared with the Monarch assembly, although the assemblies were generated from datasets with different sequencing effort. The N50 was 102,364 bp for QIAGEN compared with 51,996 bp for Monarch, while the largest contig was 916,719 bp for QIAGEN and 475,548 bp for Monarch. Although the total numbers of contigs were similar between the QIAGEN (79,055) and Monarch (79,491) assemblies, the QIAGEN assembly had lower L50 and L90 values. The L50 values were 5,468 for QIAGEN and 10,731 for Monarch, while the L90 values were 26,458 and 43,820, respectively. These metrics indicate that the QIAGEN assembly contained a greater proportion of longer contigs than the Monarch assembly. Both assemblies had total lengths of approximately 2 Gb, with the QIAGEN assembly being slightly larger than the Monarch assembly.

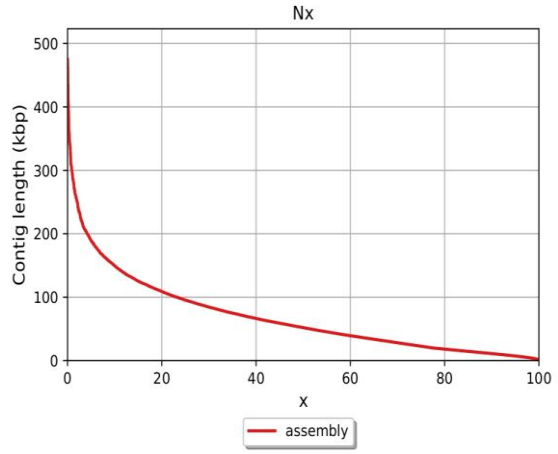
GC content was similar between the two assemblies, at 34.04% for QIAGEN and 34.51% for Monarch. Both assemblies contained 0.00 Ns per 100 kbp, indicating that no ambiguous bases were reported within the assembled contigs. Nx and cumulative-length plots showed higher contiguity metrics for the QIAGEN assembly across the genome fractions examined.

Table 3. QUAST assembly-quality statistics.

Parameter	Monarch Kit	QIAGEN Kit
Total length (Gb)	1.99	2.06
Number of contigs	79 491	79 055
Largest contig (bp)	475 548	916 719
N50 (bp)	51 996	102 364
L50	10 731	5 468
L90	43 820	26 458
GC content (%)	34.51	34.04
Ns per 100 kbp	0.00	0.00

The assembly metrics obtained from Flye and QUAST should be interpreted according to the specific metrics reported by each analysis. In the present study, the QUAST results were used for the standardised comparison of the two assemblies. Based on these metrics, the QIAGEN assembly showed greater contiguity than the Monarch assembly (Figure 3). Both assemblies had total lengths of approximately 2 Gb, indicating similar assembly sizes for the analysed datasets of *A. anatina*.

A. Monarch



B. QIAGEN

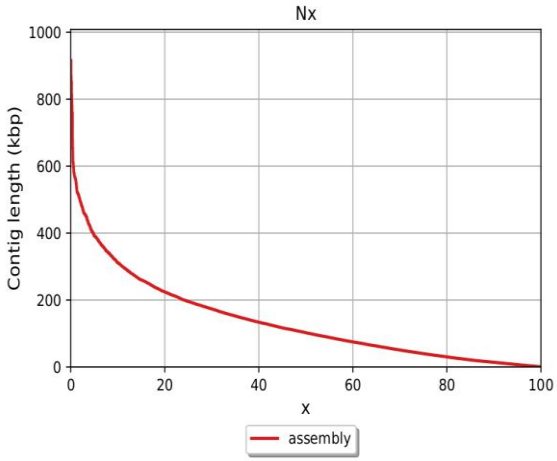


Figure 3. Nx plots of the draft *Anodonta anatina* genome assemblies generated from the (A) Monarch and (B) QIAGEN sequencing datasets. Nx represents the contig length for which contigs of that length or longer account for x% of the total assembly length. The x-axis represents the percentage of the assembly (x, 0–100%), and the y-axis represents contig length in kilobase pairs (kbp). The red curve shows the Nx contig-length distribution across the assembly.

Overall comparison of sequencing and assembly results

The selected Monarch and QIAGEN DNA samples from *A. anatina* gill tissue supported long-read sequencing and de novo genome assembly. The QIAGEN library generated a higher total sequencing yield, whereas the Monarch library produced longer average and median reads and a higher read-length N50.

Discussion

The aim of this study was to generate a draft genome assembly of the freshwater mussel *Anodonta anatina* using Oxford Nanopore long-read sequencing and to evaluate its potential as a basis for future research into pollution-related responses. Overall, the study successfully progressed from high-molecular-weight DNA extraction through sequencing, read-quality assessment, de novo assembly and assembly evaluation. Both DNA extraction approaches produced DNA suitable for sequencing, and both datasets generated draft assemblies of approximately 2 Gb. The results therefore demonstrate the feasibility of generating preliminary genomic assemblies for *A. anatina* using a relatively accessible long-read sequencing workflow. However, the assembly metrics also show that the resulting genomes should be considered draft assemblies rather than complete reference-quality genomes. As genome completeness, contamination, sequence accuracy and annotation were not independently assessed, the suitability of these assemblies as genomic resources remains uncertain. Additional sequencing, assembly refinement, accuracy and

completeness assessment, and genome annotation will therefore be required before these assemblies can reliably support comprehensive studies of pollution-related gene expression and other downstream applications.

DNA quality and suitability for long-read sequencing

The selected Monarch and QIAGEN extracts produced genomic DNA at comparable concentrations (approximately 110 ng/µL), with A260/A280 ratios indicating relatively pure DNA. The QIAGEN extract showed a higher A260/A230 ratio than the Monarch extract, suggesting lower contaminant carry-over in that sample. PFGE showed high-molecular-weight DNA in both extracts, although diffuse signals and some smearing indicated partial DNA degradation. Nevertheless, both selected extracts were suitable for Oxford Nanopore library preparation and sequencing. The successful sequencing of the Monarch extract confirms the suitability of that sample for sequencing, but the effect of its lower A260/A230 ratio on sequencing performance cannot be determined from the available results (Logsdon et al., 2020).

Sequencing output and read quality

The sequencing results demonstrated clear differences between the two datasets. The QIAGEN dataset generated 39.42 Gb from 23,404,683 reads, whereas the Monarch dataset generated 20.18 Gb from 5,267,181 reads. Thus, QIAGEN produced approximately twice as much sequence data. However, the Monarch dataset contained substantially longer reads, with a mean read length of 3.83 kb compared with 1.68 kb for QIAGEN, and a read-length N50 of 5.98 kb compared with 3.90 kb. The longest Monarch read was also longer, reaching 685.87 kb compared with 435.72 kb for QIAGEN.

These results demonstrate an important characteristic of long-read sequencing: total sequence yield and read length are different aspects of sequencing performance. A dataset can contain more bases without necessarily containing longer individual molecules. Long reads are advantageous because they can span repetitive regions and improve the continuity of de novo assemblies (Amarasinghe et al., 2020; Jain et al., 2018).

The higher QIAGEN sequence yield may have contributed to the greater contiguity of its assembly by providing higher effective sequence coverage and more overlapping reads. However, this interpretation cannot be separated completely from the difference in sequencing effort between the two datasets. The mean read quality scores were similar between the datasets, at Q15.5 for Monarch and Q14.7 for QIAGEN. Therefore, the major difference between the sequencing datasets was not read quality but the balance between sequence yield and read length. NanoPlot was appropriate for examining these characteristics because it provides visualisation and summary statistics specifically designed for long-read sequencing datasets (De Coster et al., 2018).

An important experimental consideration is that the datasets were generated from different numbers of sequencing runs. The QIAGEN dataset contained four sequencing runs, whereas the Monarch dataset contained two. Therefore, the higher total QIAGEN yield cannot be attributed solely to extraction chemistry because the datasets were generated using different amounts of sequencing effort. The comparison should consequently be interpreted as a comparison of the combined datasets obtained for each extraction method rather than as a direct single-run comparison.

Read-length distribution and implications for assembly

NanoPlot revealed that both datasets contained a large proportion of relatively short reads, despite the presence of exceptionally long individual reads. This is important because the longest read alone does not determine assembly quality. A genome assembler requires enough informative reads distributed across the genome. The Monarch dataset had a stronger mid-length component around 2–5 kb, whereas the QIAGEN dataset contributed substantially more total sequence.

The presence of many short reads may have limited the overall assembly contiguity. Long-read sequencing is most effective for resolving repetitive and structurally complex regions when a substantial proportion of the reads extend across those regions. Although both datasets contained reads extending hundreds of kilobases, most of the sequence was derived from much shorter molecules. This likely contributed to the relatively fragmented assemblies obtained. Previous studies have shown that long-read sequencing can substantially improve assembly continuity, but the benefit depends on read length, coverage and genome complexity (Jain et al., 2018; Amarasinghe et al., 2020).

Genomic characteristics common among freshwater mussels, including high heterozygosity and repetitive elements, can present additional challenges for genome assembly (Smith et al., 2018). Such characteristics make it difficult to reconstruct a genome solely from moderate-length sequencing reads. Consequently, the sequencing results should be interpreted as a useful first genomic dataset rather than evidence that the complete genomic structure of *A. anatina* has been resolved.

De novo assembly and comparison of the two assemblies

Flye successfully generated draft assemblies from both sequencing datasets using a repeat-graph approach suitable for long, error-prone reads (Kolmogorov et al., 2019). The resulting assemblies were 1.99 Gb for Monarch and 2.06 Gb for QIAGEN. Although these sizes indicate comparable amounts of assembled sequence, they do not necessarily reflect genome completeness because duplicated regions, haplotigs, and missing regions can affect assembly size. The assemblies were smaller than the published *A. anatina* draft genome assembly of approximately 2.46 Gb generated using PacBio HiFi sequencing and hifiasm (Faust et al., 2025). Therefore, additional assessment of assembly quality and completeness is required.

QUAST showed higher contiguity for QIAGEN than Monarch. The QIAGEN assembly had an N50 of 102,364 bp and a largest contig of 916,719 bp, compared with 51,996 bp and 475,548 bp for Monarch. Both assemblies were substantially more fragmented than the published *A. anatina* assembly, which consisted of 2,008 scaffolds with a scaffold N50 of 2.7 Mb (Faust et al., 2025). Although scaffold and contig metrics are not directly comparable, the comparison indicates that considerable assembly improvement remains possible.

This result is particularly interesting because Monarch generated longer sequencing reads. It demonstrates that longer individual reads do not automatically produce a more contiguous genome assembly. In this experiment, the substantially greater amount of QIAGEN sequence data may have provided higher effective coverage and improved the ability of Flye to connect overlapping genomic regions. The results therefore suggest that the higher sequencing yield and greater assembly contiguity observed in the QIAGEN-derived dataset may be related, although this cannot be attributed solely to sequencing yield or coverage because several experimental factors were confounded.

The GC contents were very similar between the assemblies, at 34.51% for Monarch and 34.04% for QIAGEN. Both assemblies also contained no reported ambiguous bases. These similarities suggest that the two assemblies had broadly comparable nucleotide composition and that the differences in assembly structure were not associated with a major difference in overall GC composition.

Relevance to the biological aim of the study

The generation of approximately 2 Gb assemblies represents a useful first step towards establishing genomic resources for *A. anatina*. Freshwater mussels are ecologically important but remain comparatively underrepresented in genomic research, despite their importance as bioindicators and their sensitivity to environmental stress. European freshwater mussels face multiple threats, including pollution, habitat degradation, invasive species and climate-related pressures (Lopes-Lima et al., 2017; Reid et al., 2019).

The biological relevance of developing a genome for *A. anatina* is supported by previous research showing that this species responds at molecular and biochemical levels to environmental contaminants. For example, Ekelund Ugge et al. (2020) demonstrated transcriptional and biochemical responses of *A. anatina* to environmentally relevant copper exposure, including changes in biomarkers associated with cellular stress. Other studies have similarly demonstrated molecular responses of freshwater mussels to pollution and environmental stress (Falfushynska et al., 2014). A genome assembly can provide an important foundation for interpreting such responses at a broader genomic scale.

Recent freshwater mussel genome projects have produced substantially more contiguous assemblies, such as *Unio delphinus* (N50 ≈ 10 Mb) and *Sinosolenaia oleivora* (N50 ≈ 20.36 Mb) (Gomes-dos-Santos et al., 2023; Ma et al., 2024). In comparison, the present QIAGEN and Monarch assemblies had N50 values of 102,364 bp and 51,996 bp, respectively. Although a draft *A. anatina* genome already exists, this study provides an independent assembly and demonstrates that QIAGEN extraction produced the more contiguous assembly, supporting its use in future genomic studies of this species.

Methodological limitations

Several limitations should be considered when interpreting the results. First, the comparison between extraction methods was not completely balanced because the QIAGEN and Monarch datasets were generated from four and two sequencing runs, respectively. In addition, only selected samples from each extraction method were sequenced, limiting conclusions regarding differences between the kits. The higher QIAGEN yield therefore cannot be attributed solely to extraction chemistry. Differences in flow-cell performance, pore availability, and sequencing duration may also have influenced sequencing yield. A controlled comparison would require equivalent sequencing effort, including the same number of runs and comparable flow-cell conditions.

Second, although the assemblies were approximately 2 Gb, assembly size alone cannot establish genome completeness. QUAST provided useful contiguity and composition metrics. An independent gene-content completeness assessment, such as BUSCO, was not performed. In addition, structural accuracy was not assessed against a reference genome. QUAST can evaluate assemblies without a reference, but reference-free metrics cannot establish whether all expected genes or genomic regions are present.

Third, the assemblies remained highly fragmented, with approximately 79,000 contigs in each assembly. Although the QIAGEN assembly had substantially better N50 and largest-contig values, its contig count indicates that considerable fragmentation remains. The available sequencing coverage may also have contributed to the fragmented assemblies, as insufficient coverage can reduce assembly contiguity and complicate the resolution of repetitive genomic regions. Additional genome polishing, contamination screening, repeat annotation, and independent completeness assessment would therefore be required before considering the assemblies high-quality draft genome assemblies.

Finally, the study focused primarily on sequencing and assembly. No gene prediction, functional annotation, repeat annotation, or pollution-exposure experiment was performed. Consequently, the present results cannot directly identify genes or pathways involved in pollutant responses. In addition, the absence of these downstream analyses limits the biological interpretation and immediate applicability of the assemblies. The genomic resource generated here should therefore be regarded as a preliminary foundation for future investigations, and further assembly refinement and annotation will be required before the assemblies can be reliably used for comprehensive studies of pollution-related gene expression and functional genomics.

Conclusion

Overall, this study generated draft genome assemblies of *Anodonta anatina* using Oxford Nanopore long-read sequencing. DNA extracted using the Monarch and QIAGEN workflows was suitable for sequencing, although the analysed datasets differed in sequencing yield and read-length characteristics. The combined QIAGEN-derived dataset generated more sequence data, whereas the Monarch-derived dataset contained longer reads on average. Despite the longer reads obtained from the Monarch-derived dataset, the QIAGEN-derived dataset produced the more contiguous assembly according to QUAST, with an N50 of 102.4 kb and the largest contig of 916.7 kb.

These findings demonstrate that genome assembly performance depends on multiple factors, including DNA integrity, read length, sequencing yield and effective coverage. However, because only selected samples from each workflow were sequenced and the sequencing effort differed between datasets, the observed differences should be interpreted cautiously. The resulting assemblies remain fragmented and should therefore be considered draft genome assemblies rather than complete reference genomes. Nevertheless, the approximately 2 Gb assemblies demonstrate the feasibility of generating genomic data for *A. anatina* and provide a basis for future assembly improvement, annotation and validation. Additional work will be required before the assemblies can reliably support downstream biological studies, including investigations of pollution-related responses.

Ethical aspects, gender equality perspective, and impact on the society

Ethical concerns regarding the necessity of using living organisms in research and the availability of alternatives should be considered in this project, publicly available genomic data were considered as a potential alternative; however, the available genomic data for *Anodonta anatina* did not provide the dataset required for the objectives of this study. Eight *A. anatina* specimens were used and handled carefully to minimise unnecessary stress and biological material use. After collection, mussels were stored at 4 °C until processing. Shells were carefully opened before dissection, after which the mussels were dead. No behavioural or pain-related experiments were performed. Based on the procedures used, no ethical approval was required from Länsstyrelsen or Jordbruksverket.

From a gender equality perspective, the research activities and opportunities associated with this project should be conducted according to principles of equal treatment and equal access, regardless of gender. The sex of the sampled mussels was not investigated and was therefore not considered in sample selection, experimental procedures, data analysis, or interpretation of the results.

Freshwater pollution and climate change are major challenges affecting aquatic ecosystems and biodiversity. The draft genome assembly generated in this project may provide a resource for future research investigating how *A. anatina* responds to environmental contaminants. In particular, the genomic resource may support future studies of gene expression, identification of pollution-related biomarkers, and investigation of biological pathways involved in environmental stress. The generated genomic information may also contribute to conservation research and improve understanding of freshwater mussel biology. However, direct applications for pollution monitoring or conservation would require further research beyond the present genome assembly.

Future perspective

The draft genome assemblies generated in this study provide a foundation for future genomic research on *Anodonta anatina*. Although the QIAGEN assembly showed higher contiguity according to QUAST, both assemblies remained fragmented, indicating that additional sequencing data and assembly refinement are needed. The higher sequencing yield of the QIAGEN dataset may have contributed to its improved assembly performance. Increasing the proportion of long and ultra-long Oxford Nanopore reads through optimised DNA extraction and handling could further improve contig continuity. High-accuracy long-read technologies, such as PacBio HiFi sequencing, may also complement Nanopore-based assemblies and support the development of a more complete genomic resource (Wenger et al., 2019).

Further assessment using BUSCO would provide a biologically meaningful estimate of conserved gene-content completeness by examining the presence of conserved single-copy genes, complementing the technical assembly metrics obtained from QUAST (Manni et al., 2021). Assembly polishing should be considered a future refinement step to potentially reduce sequencing errors and improve draft genome accuracy, with its effectiveness evaluated using appropriate accuracy assessment methods.

A further methodological improvement would be chromosome-level scaffolding using Hi-C sequencing. Prior to scaffolding, contamination screening, assessment of potential redundant haplotypes, and assembly refinement should be performed. Hi-C data can be used to order and orient contigs into chromosome-scale assemblies (Yamaguchi et al., 2021). Once improved, the assembly should be annotated to identify genes, repetitive elements, and other genomic features. A sufficiently complete and annotated genome would support future transcriptomic studies of gene expression and pollution-related responses.

Finally, the improved genome could be combined with transcriptomic experiments involving mussels exposed to environmentally relevant pollutants. Comparing exposed and control groups could identify pollution-responsive genes and biological pathways involved in stress responses. Such work would extend the present genomic resource toward its intended application in freshwater ecotoxicology, biomarker discovery and conservation research.

AI Declaration

AI-based tools, including ChatGPT (OpenAI) and Grammarly, were used during the preparation of this thesis to support language editing, grammar correction, sentence restructuring, and readability. ChatGPT was additionally used to assist with data-analysis commands and command-line troubleshooting. AI tools were not used to generate, modify, or interpret experimental data. All experimental work, data analysis, interpretation of results, and scientific conclusions were carried out independently by the authors.

References

Amarasinghe, S. L., Su, S., Dong, X., Zappia, L., Ritchie, M. E., & Gouil, Q. (2020). Opportunities and challenges in long-read sequencing data analysis. *Genome Biology*, 21, 30. https://doi.org/10.1186/s13059-020-1935-5

aus der Beek, T., Weber, F.-A., Bergmann, A., Hickmann, S., Ebert, I., Hein, A., & Küster, A. (2016). Pharmaceuticals in the environment—Global occurrences and perspectives. *Environmental Toxicology and Chemistry, 35*(4), 823–835. https://doi.org/10.1002/etc.3339

Balbi, T., Auguste, M., Ciacci, C., & Canesi, L. (2021). Immunological responses of marine bivalves to contaminant exposure: Contribution of the -omics approach. *Frontiers in Immunology,* 12, Article 618726. https://doi.org/10.3389/fimmu.2021.618726

De Coster, W., D'Hert, S., Schultz, D. T., Cruts, M., & Van Broeckhoven, C. (2018). NanoPack: Visualizing and processing long-read sequencing data. *Bioinformatics*, 34(15), 2666–2669. https://doi.org/10.1093/bioinformatics/bty149

Dudgeon, D., Arthington, A. H., Gessner, M. O., Kawabata, Z.-I., Knowler, D. J., Lévêque, C., Naiman, R. J., Prieur-Richard, A.-H., Soto, D., Stiassny, M. L. J., & Sullivan, C. A. (2006). Freshwater biodiversity: Importance, threats, status and conservation challenges. *Biological Reviews*, 81(2), 163–182. https://doi.org/10.1017/S1464793105006950

Ekblom, R., & Galindo, J. (2011). Applications of next generation sequencing in molecular ecology of non-model organisms. *Heredity, 107*(1), 1–15. https://doi.org/10.1038/hdy.2010.152

Ekelund Ugge, G. M. O., Jonsson, A., Olsson, B., Sjöback, R., & Berglund, O. (2020). Transcriptional and biochemical biomarker responses in a freshwater mussel (*Anodonta anatina*) under environmentally relevant Cu exposure. *Environmental Science and Pollution Research, 27*(9), 9999–10010. https://doi.org/10.1007/s11356-020-07660-4

Ekelund Ugge, G. M., Jonsson, A., & Berglund, O. (2022). Molecular biomarker responses in the freshwater mussel *Anodonta anatina* exposed to an industrial wastewater effluent. *Environmental Science and Pollution Research, 29*(2), 2158–2170. https://doi.org/10.1007/s11356-021-15633-4

Ellegren, H. (2014). Genome sequencing and population genomics in non-model organisms. *Trends in Ecology & Evolution, 29*(1), 51–63. https://doi.org/10.1016/j.tree.2013.09.008

Falfushynska, H. I., Gnatyshyna, L. L., & Stoliar, O. B. (2013). Effect of in situ exposure history on the molecular responses of freshwater bivalve *Anodonta anatina* (Unionidae) to trace metals.

Ecotoxicology and Environmental Safety, 89, 73–83. https://doi.org/10.1016/j.ecoenv.2012.11.024

Falfushynska, H. I., Gnatyshyna, L. L., Osadchuk, O. Y., Farkas, A., Vehovszky, A., Carpenter, D. O., Gyori, J., & Stoliar, O. B. (2014). Diversity of the molecular responses to separate wastewater effluents in freshwater mussels. *Comparative Biochemistry and Physiology Part C: Toxicology & Pharmacology*, 164, 51–58. https://doi.org/10.1016/j.cbpc.2014.04.007

Falfushynska, H. I., Gnatyshyna, L. L., Yurchak, L., Ivanina, A., Stoliar, O., & Sokolova, I. M. (2014). Habitat pollution and thermal regime modify molecular stress responses to elevated temperature in freshwater mussels (*Anodonta anatina*: Unionidae). *Science of the Total Environment,* 500–501, 339–350. https://doi.org/10.1016/j.scitotenv.2014.08.112

Faust, E., Conrads, J., Giulio, M., Ciofi, C., Natali, C., Feulner, P. G. D., & Weber, A. A.-T. (2025). Beyond genetic indicators: How reproductive mode and hybridisation challenge freshwater mussel conservation. *Molecular Ecology, 34*(21), e70066. https://doi.org/10.1111/mec.70066

Gerdol, M., Moreira, R., Cruz, F., Gómez-Garrido, J., Vlasova, A., Rosani, U., Venier, P., Naranjo-Ortiz, M. A., Murgarella, M., Greco, S., Balseiro, P., Corvelo, A., Frias, L., Gut, M., Gabaldón, T., Pallavicini, A., Canchaya, C., Novoa, B., Alioto, T. S., Posada, D., & Figueras, A. (2020). Massive gene presence-absence variation shapes an open pan-genome in the Mediterranean mussel. *Genome Biology, 21*, 275. https://doi.org/10.1186/s13059-020-02180-3

Goldberg, E. D. (1975). The mussel watch—A first step in global marine monitoring. *Marine Pollution Bulletin, 6*(7), 111. https://doi.org/10.1016/0025-326X(75)90271-4

Gomes-dos-Santos, A., Lopes-Lima, M., Machado, M. A., Teixeira, A., Castro, L. F. C., & Froufe, E. (2023). PacBio Hi-Fi genome assembly of the Iberian dolphin freshwater mussel *Unio delphinus* Spengler, 1793. *Scientific Data, 10*, 340. https://doi.org/10.1038/s41597-023-02251-7

Gurevich, A., Saveliev, V., Vyahhi, N., & Tesler, G. (2013). QUAST: Quality assessment tool for genome assemblies. *Bioinformatics*, 29(8), 1072–1075. https://doi.org/10.1093/bioinformatics/btt086

Haag, W. R. (2012). *North American freshwater mussels: Natural history, ecology, and conservation*. Cambridge University Press. https://doi.org/10.1017/CBO9781139048217

Hornett, E. A., & Wheat, C. W. (2012). Quantitative RNA-Seq analysis in non-model species: Assessing transcriptome assemblies as a scaffold and the utility of evolutionary divergent genomic reference species. *BMC Genomics*, 13, Article 361. https://doi.org/10.1186/1471-2164-13-361

Jain, M., Koren, S., Miga, K. H., Quick, J., Rand, A. C., Sasani, T. A., Tyson, J. R., Beggs, A. D., Dilthey, A. T., Fiddes, I. T., Malla, S., Marriott, H., Nieto, T., O'Grady, J., Olsen, H. E., Pedersen, B. S., Rhie, A., Richardson, H., Quinlan, A. R., ... Loose, M. (2018). Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nature Biotechnology*, 36, 338–345. https://doi.org/10.1038/nbt.4060

Jones, A., Torkel, C., Stanley, D., Nasim, J., Borevitz, J., & Schwessinger, B. (2021). High-molecular weight DNA extraction, clean-up and size selection for long-read sequencing. *PLOS ONE*, 16(7), Article e0253830. https://doi.org/10.1371/journal.pone.0253830

Kolmogorov, M., Yuan, J., Lin, Y., & Pevzner, P. A. (2019). Assembly of long, error-prone reads using repeat graphs. *Nature Biotechnology*, 37, 540–546. https://doi.org/10.1038/s41587-019-0072-8

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. https://doi.org/10.1101/gr.215087.116

Linnaeus, C., *Systema Naturae per Regna Tria Naturae*, Secundum Classes, Ordines, Genera, Species, cum Characteribus, Differentiis, Synonymis, Locis, 10th ed., Vol. 1 (Laurentii Salvii, Holmiae, 1758), p. 706.

Logsdon, G. A., Vollger, M. R., & Eichler, E. E. (2020). Long-read human genome sequencing and its applications. *Nature Reviews Genetics*, 21, 597–614. https://doi.org/10.1038/s41576-020-0236-x

Lopes-Lima, M., Sousa, R., Geist, J., Aldridge, D. C., Araujo, R., Bergengren, J., Bespalaya, Y., Bódis, E., Burlakova, L., Van Damme, D., Douda, K., Froufe, E., Georgiev, D., Gumpinger, C., Karatayev, A., Kebapçı, Ü., Killeen, I., Lajtner, J., Larsen, B. M., Lauceri, R., ... Zieritz, A. (2017). Conservation status of freshwater mussels in Europe: State of the art and future challenges. *Biological Reviews, 92*(1), 572–607. https://doi.org/10.1111/brv.12244

Lopes-Lima, M., Sousa, R., Geist, J., Aldridge, D. C., Araujo, R., Bergengren, J., Bespalaya, Y., Bódis, E., Burlakova, L., Van Damme, D., Douda, K., Froufe, E., Georgiev, D., Gumpinger, C., Karatayev, A., Kebapçı, Ü., Killeen, I., Lajtner, J., Larsen, B. M., ... Zieritz, A. (2018). Conservation status of freshwater mussels in the European Union: An overview. *Biological Conservation, 221*, 115–124. https://doi.org/10.1016/j.biocon.2018.02.033

Ma, X., Jin, W., Chen, W., Liu, Q., Jiang, H., Zhou, Y., Xu, P., Wen, H., ... Xu, D. (2024). Chromosome-level genome assembly of the freshwater mussel Sinosolenaia oleivora (Heude, 1877). *Scientific Data*, 11, 606. https://doi.org/10.1038/s41597-024-03451-5

Manly, R., & George, W. O. (1977). The occurrence of some heavy metals in populations of the freshwater mussel *Anodonta anatina* (L.) from the River Thames. *Environmental Pollution, 14*(2), 139–154. https://doi.org/10.1016/0013-9327(77)90106-9

Manni, M., Berkeley, M. R., Seppey, M., & Zdobnov, E. M. (2021). BUSCO: Assessing genomic data quality and beyond. *Current Protocols*, 1(12), e323. https://doi.org/10.1002/cpz1.323

Manni, M., Berkeley, M. R., Seppey, M., Simão, F. A., & Zdobnov, E. M. (2021). BUSCO update: Novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. *Molecular Biology and Evolution*, 38(10), 4647–4654. https://doi.org/10.1093/molbev/msab199

Rainbow, P. S. (1995). Biomonitoring of heavy metal availability in the marine environment. *Marine Pollution Bulletin, 31*(4–12), 183–192. https://doi.org/10.1016/0025-326X(95)00116-5

Regoli, F., & Giuliani, M. E. (2014). Oxidative pathways of chemical toxicity and oxidative stress biomarkers in marine organisms. *Marine Environmental Research, 93*, 106–117. https://doi.org/10.1016/j.marenvres.2013.07.006

Reid, A. J., Carlson, A. K., Creed, I. F., Eliason, E. J., Gell, P. A., Johnson, P. T. J., Kidd, K. A., MacCormack, T. J., Olden, J. D., Ormerod, S. J., Smol, J. P., Taylor, W. W., Tockner, K., Vermaire, J. C., Dudgeon, D., & Cooke, S. J. (2019). Emerging threats and persistent conservation challenges for freshwater biodiversity. *Biological Reviews*, 94(3), 849–873. https://doi.org/10.1111/brv.12480

Schwarzenbach, R. P., Escher, B. I., Fenner, K., Hofstetter, T. B., Johnson, C. A., von Gunten, U., & Wehrli, B. (2006). The challenge of micropollutants in aquatic systems. *Science, 313*(5790), 1072–1077. https://doi.org/10.1126/science.1127291

Smith, C. H., Pfeiffer, J. M., Johnson, N. A., & Moulton, M. J. (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. https://doi.org/10.1093/gbe/evy117

Tickner, D., Opperman, J. J., Abell, R., Acreman, M., Arthington, A. H., Bunn, S. E., Cooke, S. J., Dalton, J., Darwall, W., Edwards, G., Harrison, I., Hughes, K., Jones, T., Leclère, D., Lynch, A. J., Leonard, P., McClain, M. E., Muruven, D., Olden, J. D., Ormerod, S. J., Robinson, J., Tharme, R. E., Thieme, M., Tockner, K., Wright, M., & Young, L. (2020). Bending the curve of global freshwater biodiversity loss: An emergency recovery plan. *BioScience, 70*(4), 330–342. https://doi.org/10.1093/biosci/biaa002

Trigodet, F., Lolans, K., Fogarty, E., & Chang, S. L. (2022). High molecular weight DNA extraction strategies for long-read sequencing of complex metagenomes. *Molecular Ecology Resources,* 22(5), 1786–1802. https://doi.org/10.1111/1755-0998.13588

Vaughn, C. C., & Hoellein, T. J. (2018). Bivalve impacts in freshwater and marine ecosystems. *Annual Review of Ecology, Evolution, and Systematics, 49*, 183–208. https://doi.org/10.1146/annurev-ecolsys-110617-062703

Vörösmarty, C. J., McIntyre, P. B., Gessner, M. O., Dudgeon, D., Prusevich, A., Green, P., Glidden, S., Bunn, S. E., Sullivan, C. A., Liermann, C. R., & Davies, P. M. (2010). Global threats to human water security and river biodiversity. *Nature*, 467, 555–561. https://doi.org/10.1038/nature09440

Wenger, A. M., Peluso, P., Rowell, W. J., Chang, P.-C., Hall, R. J., Concepcion, G. T., Ebler, J., Fungtammasan, A., Kolesnikov, A., Olson, N. D., ... Hunkapiller, M. W. (2019). Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nature Biotechnology, 37*, 1155–1162. https://doi.org/10.1038/s41587-019-0217-9

Yamaguchi, K., Kadota, M., Nishimura, O., Ohishi, Y., Naito, Y., & Kuraku, S. (2021). Technical considerations in Hi-C scaffolding and evaluation of chromosome-scale genome assemblies. *Molecular Ecology*, 30(23), 5923–5934. https://doi.org/10.1111/mec.16146

Appendix 1

cat Abhijith_and_Shanoob/Qiagen/All_passed_fastq_files/*.fastq.gz > Merged_fastq.gz

cat Abhijith_and_Shanoob/Monarch/All_passed_fastq_files/*.fastq.gz > Merged_fastq.gz

Appendix 2

NanoPlot --fastq Abhijith_and_Shanoob/Qiagen/Merged_fastq/Merged_fastq.gz --outdir Abhijith_and_Shanoob/Qiagen/NanoPlot_output

NanoPlot --fastq Abhijith_and_Shanoob/Monarch/Merged_fastq/Merged_fastq.gz --outdir Abhijith_and_Shanoob/Monarch/NanoPlot_output

Appendix 3

flye --nano-hq ~/Abhijith_and_Shanoob/Qiagen/Merged_fastq/Merged_fastq.fastq.gz --out-dir ~/Abhijith_and_Shanoob/Qiagen/Flye_output --genome-size 2g --threads 16

flye --nano-hq ~/Abhijith_and_Shanoob/Monarch/Merged_fastq/Merged_fastq.fastq.gz --out-dir ~/Abhijith_and_Shanoob/Monarch/Flye_output --genome-size 2g --threads 16

Appendix 4

quast ~/Abhijith_and_Shanoob/Qiagen/Flye_output/assembly.fasta -o ~/Abhijith_and_Shanoob/Qiagen/Quast_output

quast ~/Abhijith_and_Shanoob/Monarch/Flye_output/assembly.fasta -o ~/Abhijith_and_Shanoob/Monarch/Quast_output