BUREAU OF RECLAMATION

# Sequencing of the quagga mussel (*Dreissena rostriformis bugensis*) genome as a tool for biocontrol

**Science and Technology Program**
**Research and Development Office**
**Final Report No. ST-2020-1866-01**

**U.S. Department of the Interior**

September 23, 2020

# REPORT DOCUMENTATION PAGE

*Form Approved*
*OMB No. 0704-0188*



| 1. REPORT DATE *(DD-MM-YYYY)* | 2. REPORT TYPE | 3. DATES COVERED *(From - To)* |
|---|---|---|
| 23-09-2020 | Research | 2018-2020 |

**4. TITLE AND SUBTITLE**
Sequencing of the quagga mussel genome as a tool for biocontrol

**5a. CONTRACT NUMBER**
X1866 / RY.15412018.ZQ31866

**5b. GRANT NUMBER**

**5c. PROGRAM ELEMENT NUMBER**
1541 (S&T)

**6. AUTHOR(S)**
Yale Passamaneck, Ph.D. (Bureau of Reclamation)
Kevin Kocot, Ph.D. (University of Alabama)

**5d. PROJECT NUMBER**
Final Report ST-2020-1866-01

**5e. TASK NUMBER**
N/A

**5f. WORK UNIT NUMBER**
N/A

**7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)**
Technical Service Center
Bureau of Reclamation
U.S. Department of the Interior
Denver Federal Center PO Box 25007, Denver, CO 80225-0007

Department of Biological Sciences and Alabama Museum of Natural History
University of Alabama
Tuscaloosa, Alabama, 35487, USA

**8. PERFORMING ORGANIZATION REPORT NUMBER**

**9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)**
Science and Technology Program
Research and Development Office
Bureau of Reclamation
U.S. Department of the Interior
Denver Federal Center
PO Box 25007, Denver, CO 80225-0007

**10. SPONSOR/MONITOR'S ACRONYM(S)**
Reclamation

**11. SPONSOR/MONITOR'S REPORT NUMBER(S)**
Final Report ST-2020-1866-01

**12. DISTRIBUTION/AVAILABILITY STATEMENT**
Final Report may be downloaded from https://www.usbr.gov/research/projects/index.html

**13. SUPPLEMENTARY NOTES**

**14. ABSTRACT**
The quagga mussel is an invasive freshwater bivalve which poses a significant risk to Reclamation facilities and operations. The current project undertook sequencing and assembly of the quagga mussel genome to provide a resource for efforts to understand mussel biology and to develop genetic biocontrols. A chromosome-scale reference genome assembly was developed during the project. Analyses of the assembled genome show that it displays high levels of both contiguity and completeness.

**15. SUBJECT TERMS**
Quagga mussels, genomics, genetic biocontrol

| 16. SECURITY CLASSIFICATION OF: | | | 17. LIMITATION OF ABSTRACT | 18. NUMBER OF PAGES | 19a. NAME OF RESPONSIBLE PERSON |
|---|---|---|---|---|---|
| a. REPORT | b. ABSTRACT | THIS PAGE | | | 19b. TELEPHONE NUMBER *(Include area code)* |
| U | U | U | | | |

# Mission Statements

The Department of the Interior (DOI) conserves and manages the Nation's natural resources and cultural heritage for the benefit and enjoyment of the American people, provides scientific and other information about natural resources and natural hazards to address societal challenges and create opportunities for the American people, and honors the Nation's trust responsibilities or special commitments to American Indians, Alaska Natives, and affiliated island communities to help them prosper.

The mission of the Bureau of Reclamation is to manage, develop, and protect water and related resources in an environmentally and economically sound manner in the interest of the American public.

# Disclaimer

Information in this report may not be used for advertising or promotional purposes. The data and findings should not be construed as an endorsement of any product or firm by the Bureau of Reclamation, Department of Interior, or Federal Government. The products evaluated in the report were evaluated for purposes specific to the Bureau of Reclamation mission. Reclamation gives no warranties or guarantees, expressed or implied, for the products evaluated in this report, including merchantability or fitness for a particular purpose.

# Acknowledgements

The Science and Technology Program, Bureau of Reclamation, sponsored this research. Dr Michael McCarthy provided valuable discussion on genomic DNA extraction methods. The University of Alabama High-Performance Computing (UAHPC) administrators provided assistance with software installation and maintenance.

# Sequencing of the quagga mussel (*Dreissena rostriformis bugensis*) genome as a tool for biocontrol

**Final Report No. ST-2020-1866-01**

*prepared by*

**Technical Service Center**
**Yale Passamaneck, Ph.D., Biologist**

**University of Alabama**
**Kevin Kocot, Ph.D., Assistant Professor**

# Peer Review

## Bureau of Reclamation
## Research and Development Office
## Science and Technology Program

Final Report ST-2020-1866-01

**Sequencing of the quagga mussel (*Dreissena rostriformis bugensis*) genome as a tool for biocontrol**

YALE PASSAMANECK

Digitally signed by YALE PASSAMANECK
Date: 2020.09.28 15:58:11 -06'00'

**Prepared by: Yale J. Passamaneck, Ph.D.**
**Biologist, Ecological Research Laboratory, Hydraulic Investigations and Laboratory Services, 86-68560, Technical Service Center, Bureau of Reclamation**

RHEANNAN QUATTLEBAUM

Digitally signed by RHEANNAN QUATTLEBAUM
Date: 2020.09.29 09:46:18 -06'00'

**Peer Review by: Rheannan A Quattlebaum**
**Biologist, Ecological Research Laboratory, Hydraulic Investigations and Laboratory Services, 86-68560, Technical Service Center, Bureau of Reclamation**

# Acronyms and Abbreviations

| | |
|---|---|
| ASCII | American Standard Code for Information Exchange |
| CDS | Coding sequence |
| CTAB | Cetyl trimethylammonium bromide |
| DNA | Deoxyribonucleic acid |
| HMM | Hidden Markov Model |
| mRNA | messenger RNA |
| NGS | Next-generation sequencing |
| PacBio | Pacific Biosciences |
| PE | Paired-end |
| Reclamation | Bureau of Reclamation |
| RNA | ribonucleic acid |
| SMRT | Single Molecule Real Time |
| RIN | RNA integrity number |
| RNASeq | RNA sequencing |
| rRNA | Ribosomal RNA |
| UTR | Untranslated region |

# Measurements

| | |
|---|---|
| bp | base pairs |
| Mbp | Mega base pairs |
| nt | nucleotides |
| µg | microgram |

# Contents

# Executive Summary

The quagga mussel (*Dreissena rostriformis bugensis*) is a non-native freshwater bivalve that has become established in several Reclamation reservoirs. Quagga mussels can cause significant risks for Reclamation operations due to the settlement and growth of mussels on hard substrates, or the entrainment of shell debris in waters carried through Reclamation facilities. Potential impacts include interruption to power generation and water delivery, as well as damage to equipment. Quagga mussels can dramatically alter the ecology of reservoirs and other waters where they occur, as their filter feeding removes large amounts of algae and bacteria from the water column, depriving native species of nourishment and creating conditions for benthic algal blooms though increased light penetrance. Mussel infestations can also impact recreational use of Reclamation waters, as efforts to prevent the spread of mussels can limit or completely prevent access by the public.

In the 30 years since quagga mussels were first detected in the United States, significant effort has gone into controlling their populations and their impacts. Although several tools have shown promise for controlling their impact on infrastructure and operations, there currently exist no methods for their control or eradication in large open-water habitats such as Lake Mead.

Analysis of the quagga mussel genome holds the potential to facilitate development of new approaches to control. For example, methods for genetic biocontrol have developed rapidly in the last few years and have the potential to provide technologies that could be self-perpetuating and scalable, allowing for control or eradication of populations even in large reservoirs.

The current project developed a high-quality, chromosome-scale assembly of the quagga mussel genome as a resource for the development of biocontrols and investigations of quagga mussel biology. Two methods of high-throughput DNA sequencing were used to maximize both completeness and the accuracy of genome sequencing. Multiple different bioinformatic approaches were tested to optimize the genome assembly. Scaffolding of the assembly was complemented using Hi-C chromosomal conformation analysis, which facilitated reconstruction of chromosome-scale pseudomolecules. RNA sequencing and transcript assembly was also conducted to allow identification of genes of interest and to facilitate ongoing efforts to annotate the genome.

# Introduction

## Project Background

Quagga mussels (*Dreissena rostriformis bugenesis*) are freshwater bivalves whose native range is the Dneiper River drainage in Ukraine. Due to unintentional translocation, they became established in the Great Lakes by 1989, following the congener zebra mussel (*Dreissena polymorpha*), which was discovered in the same waters a year earlier. Despite efforts to keep these invasive species out of the Western United States, quagga mussels were discovered in Lake Mead, upstream of Hoover Dam. Shortly thereafter quagga mussels were also found in Lake Mohave upstream of Davis Dam and in Lake Havasu upstream of Parker Dam. In subsequent years quagga mussel have become established in additional reservoirs outside of the Lower Colorado River system, particularly in Lake Powell upstream of Glen Canyon Dam, and in several reservoirs of the Salt River Project in Arizona. Additionally, early detection monitoring has found evidence for introductions of quagga mussels in numerous other Reclamation reservoirs, suggesting that quagga mussels continue to be new waterbodies, and the risk of additional infestations persists.

At locations where quagga mussels have become established, they have the potential to cause significant operational, ecological, recreational, and economic impacts. Because of their shells and their propensity for settlement and growth on hard substrates, quagga mussels pose numerous risks to Reclamation facilities. Fouling by settled mussels can cause increased labor and operational costs as piping, trashracks, and other equipment needs to be monitored and cleaned with increased regularity. Settled mussels and shell debris have the potential to occlude pipes and to be carried into equipment. Such events have the potential to damage and block equipment, including fire suppression systems, and could disruptions in water delivery and power generation. Within reservoirs, quagga mussels can have significant ecological impacts as their filter feeding on algae and bacteria can deprive other organisms of nutrition. Native and commercially important fish species may be negatively impacted by these mussel-induced changes to the food web. In addition, dreissenid mussels have been linked to increases in harmful algal blooms, due to their selective feeding and impact on nutrient cycling. The need to control the spread of quagga mussels can also impact public recreation, as some locations have chosen to limit or block boating activity to prevent the spread or introduction of mussels.

A major impediment to the management of quagga mussels is the absence of methods for controlling populations, particularly in open water. Numerous control treatments have been tested, and several, including ultraviolet light and potash, have shown promise, particularly for use in closed-pipe systems. However, none of the available treatments have shown promise for scaling to large open-water systems, such as most infested reservoirs, due to technical and economic limitations.

In the last few years, there has been significant interest in the use of genetic and genomic approaches to control invasive species. So called genetic biocontrol relies on using an organisms own biology, or changes to its genome, to effect changes in the population with the goal of control or eradication.

This work has been accelerated by several technical and theoretical advances, in particular the advent of the CRISPR/Cas9 system. CRISPR/Cas9 facilitates the generation of targeted genomic modifications, and the system has been shown to be widely applicable in a wide variety of organisms. In the case of quagga mussels, genetic biocontrol could be used to target reproduction, growth, or other aspects of their biology that could lower their fitness and lead to decreases in populations. Before any such efforts could be pursued, a detailed understanding of the organism's genome is required. High-quality genomic data is foundational to understanding what parts of the genome might be targeted for genetic biocontrol and how they might be targeted.

Prior to the advent of this project relatively little genetic data was publicly available. A limited number of sequences were available in the National Center for Biotechnology Information database. However, genome-wide data required to investigate quagga mussel biology and develop genetic biocontrols was lacking. The current project set out to collect genomic sequence data for the quagga mussel and to develop a high-quality assembly to serve as a resource for future investigations. During the course of the project a parallel effort to sequence the genome of a quagga mussel collected in Europe was published (Calcino et al., 2019). This second genome will serve as an important point of comparison for the data from the current project.

# Methods and Results

## Sample collection

Adult quagga mussels were collected from the Davis Dam forebay. Trash racks hung from the side of the forebay provide a substrate for mussel settlement and growth. Adult mussels were collected and returned to Reclamation's Lower Colorado Area Office in Boulder City, NV, where they were maintained in aerated fresh water. Individuals were dissected under a stereo microscope, with sterile dishes, forceps, and scalpels used for each individual. Tissue samples including foot, gills, and gonads were placed in individual screw-top cryotubes. Samples were flash frozen in liquid nitrogen and held on dry ice for shipment to Reclamation's Ecological Research Laboratory in Denver CO, where they were stored in a -80°C freezer.

## DNA extraction and isolation

Requirements for sequencing quagga mussel genomic DNA included large amounts of DNA (over 50 μg), high molecular weight DNA, and high purity. To optimize DNA extraction and isolation, multiple methods were tested, including several commercial kits and traditional CTAB buffer extraction with phenol:chloroform isolation. The best results for all three criteria were achieved with the Qiagen Blood & Cell Culture Kit (QIAGEN Inc., USA), using Genomic Tip 100/G. The kit relies on gravity filtration through Genomic Tip, rather than centrifugation as used in most other genomic DNA isolation kit protocols. Gravity filtration decreases shearing of the DNA, but also makes the filters prone to clogging. The issue of clogging with quagga mussel tissues was significant, likely due to the presence of polysaccharides, particularly in testes tissue, that are not degraded in the

Proteinase K-based tissue digestion. Use of a plunger was tested to force samples through the Genomic Tip column with gentle positive pressure, but this was found to decrease yield and potentially increase shearing. The most effective method found was to pre-filter the digested sample through a nylon screen filter basket with a 100 μm pore size (Corning Sterile Cell Strainer – manufacturer product number 431752) prior to binding the sample to the Genomic Tip filter. This approach greatly improved flow-through from the Genomic Tip column and resulted in increased genomic DNA yield.

DNA yield was measured with a Qubit 4 fluorometer (Thermo Fisher Scientific, USA) using the Qubit dsDNA BR Assay Kit to label samples. DNA purity was determined by measurement of light absorption values at 260 nm and 280 nm using a Cary-60 spectrophotometer with a 1.0mm microcell cap for measurement of small volumes. The ration of these two values, referred to as the 260/280 ratio, is expected to be near 1.8 for pure DNA. Values lower than this may indicate the presence of protein or other contaminants.

## DNA sequencing

Two sequencing technologies, Pacific Biosciences (PacBio) Sequel Single Molecule Real Time (SMRT) and Illumina HiSeq paired-end (PE), were used to collect genomic DNA sequence data for the quagga mussel. Both of these technologies are so-called high-throughput or "next-generation" (NGS) DNA sequencing technologies, wherein sequencing is massively parallelized and thousands or millions of DNA fragments may be analyzed simultaneously. Such technologies have made it economical to collect genome-scale data, as compared to the traditional Sanger method where each DNA fragment had to be prepared and analyzed in a separate reaction. PacBio Sequel SMRT and Illumina HiSeq PE sequencing each have distinct advantages and disadvantages. PacBio Sequel SMRT can produce very long individual reads (mean read lengths over 10,000 base pairs (bp) from high quality genomic DNA preparations). However, the error rate of individual base calls may be relatively high. In contrast, Illumina HiSeq sequencing produces shorter reads, typically 100 or 150 bp in length (depending on the strategy used), but with low error rates.

Both technologies facilitate the modern approach to genome sequencing which relies on "shotgun" sequencing. In shotgun sequencing, the genome is fragmented and then randomly sequenced in a parallelized manner. Reconstruction of the genome sequence then relies on *de novo* assembly of the resultant fragments (Figure 1). This assembly step is computationally intensive, but acquisition of the sequence data itself is orders of magnitude less expensive per base read than is the traditional Sanger method. Two of the main challenges in *de novo* sequence assembly are to place DNA sequence read fragments in the correct relative position, and to maximize the length of the assembled regions of genome sequence, called contigs. Both goals are confounded by the fact that animal genomes often contain large numbers of repetitive elements, regions of repeated DNA sequences that may make up to as much of 70% of the genome. Assembly of a highly complex and repetitive genome from short Illumina HiSeq reads makes it inherently difficult to correctly place these repetitive elements, and resultant genome assemblies tend to be highly fragmented (often with tens of thousands of fragments depending on the genome size). The long read-lengths from PacBio Sequel SMRT helps to overcome this challenge as individual reads may span repetitive elements, facilitating assembly of longer contigs. The lower error rate of the shorter Illumina HiSeq reads can be used to improve base call accuracy by proofreading the assembly. Combining data from two different sequencing

technologies into the assembly is referred to as hybrid assembly, a term which encompasses a variety of different methodologies and algorithims.

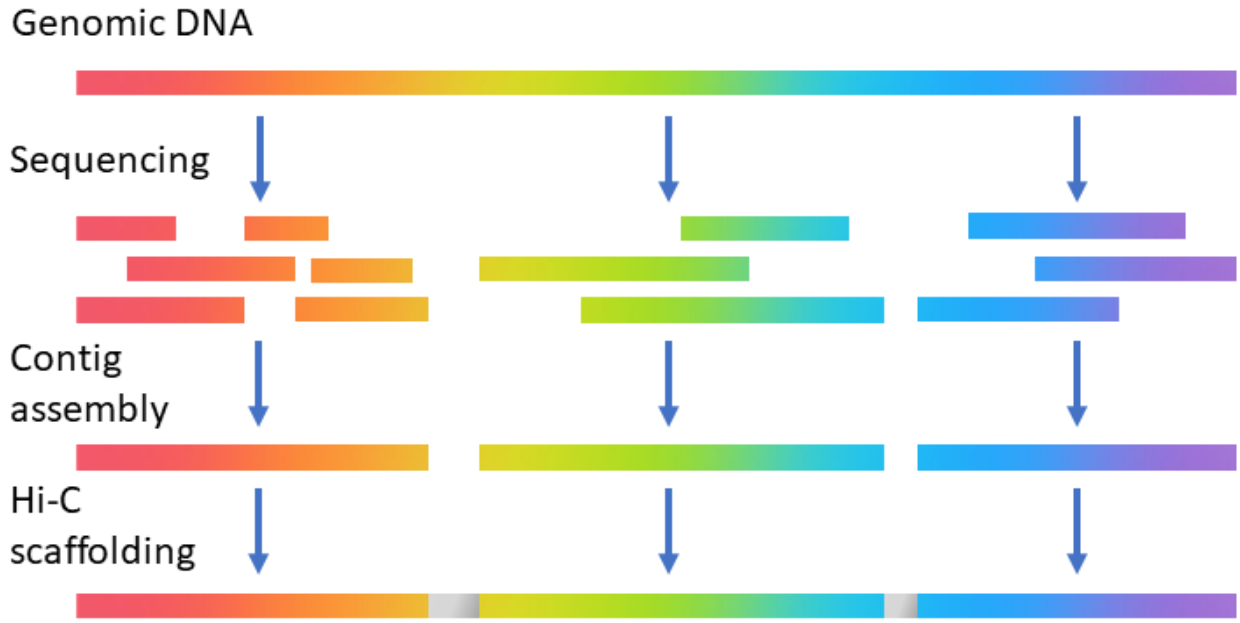


Figure 1 Schema of genome sequence and assembly process

Shotgun sequencing and *de novo* genome assembly rely on oversampling of the genome. The higher the coverage of sequencing of the genome, the more overlapping fragments that are available for accurate reconstruction of any given region. The practical and technical limitations on this maxim are the increased cost of high-coverage sequencing, and the computational resources required to perform assemblies on massive quantities of sequencing data. At some point additional sequencing provides diminishing returns, however there is no hard and fast rule for this, as the optimal sequencing depth is dependent upon the size and complexity of the genome, the goals of the project, the assembly software used, and the computational resources available. For the present study a goal of 100x coverage from both sequencing technologies was chosen as a target that would maximize the data available to develop a high-quality assembly while keeping costs within budget.

A key piece of data for designing a genome sequencing and assembly project is knowledge of the genome size. An accurate estimate of genome size is critical to selecting the appropriate amount of sequencing and to properly evaluate the results of assembly efforts. Genome sizes of animals can range widely, from as small as 15 mega base pairs (Mbp) for some parasitic nematodes (Slyusarev et al., 2020) up to an estimated 130,000 Mbp for the marbled lungfish (Pedersen, 1971); (www.genomesize.com). (Note, genome sizes generally refer to the number of base pairs in the haploid genome. Throughout this report the convention of referring to genome size as the haploid size in mega base pairs [Mbp] will be used). For bivalves genome size appears to be more constrained, with available estimates for different species ranging from approximately 600 to 5400 Mbp (www.genomesize.com). No direct measure of the quagga mussel genome size was available, however, an estimate of 1,660 Mbp was available for the zebra mussel (*Dreissena polymorpha*) which is in the same genus (Gregory, 2003). This was considered an appropriate estimate and was used for experimental design. To achieve 100x coverage, sequencing was targeted for an output of 160,000 Mbp from both PacBio Sequel SMRT and Illumina HiSeq PE technologies.

Contracting for sequencing was arranged through Reclamation's Acquisitions and Assistance Management Division (AAMD). Market research was performed during development of the acquisitions package to identify vendors that could provide both PacBio Sequel SMRT and Illumina HiSeq PE sequencing. While Illumina HiSeq PE sequencing is available from numerous commercial laboratories and university core facilities, a much smaller number of laboratories offer PacBio Sequel SMRT. A list of potential vendors was included with the acquisitions package based on this market research. GENEWIZ (South Plainfield, NJ) responded to the posted solicitation from AAMD and was selected for the contract based on statement of capabilities meeting or exceeding the requirements in the solicitation.

Genomic DNA extracted from a male mussel labeled '*Drb016*' was sent to GENEWIZ for PacBio Sequel SMRT and Illumina HiSeq PE sequencing. The *Drb016* extract was chosen because it contained sufficient DNA (50 micrograms) to allow for both sequencing methods to be performed from the same sample. This is critical to accomplish a successful hybrid assembly, as differences in genome sequence between individuals could confound attempts at combining the two datasets.

# Sequence data

## Illumina HiSeq Paired-End (PE) sequencing

GENEWIZ produced Illumin HiSeq PE data totaling 664,901,022 reads. The yield of 199,470 Mbp exceeded the requested yield of 160,000 Mbp requested in the solicitation. The data was provided in FASTQ format, as well as in BCL format which contains the raw data generated by the sequencing instrument. FASTQ files are widely used for sequence assembly because they incorporate quality scores for each base, in addition to the nucleotide sequences themselves. Incorporation of these quality scores allows for automated trimming of low-quality regions of sequences, and subsequent exclusion from analyses and assemblies. FASTQ format quality scores (Q Scores) are calculated as

$$\mathbf{Q = -10\log_{10}(p)}$$

where **p** is the estimated probability of an individual base call being incorrect. The relationship between Q Score and base call accuracy is shown below (Table 1).

Table 1 Illumina quality scores.

| Quality Score | Probability of Incorrect Base Call | Inferred Base Call Accuracy |
|---|---|---|
| 10 (Q10) | 1 in 10 | 90% |
| 20 (Q20) | 1 in 100 | 99% |
| 30 (Q30) | 1 in 1,000 | 99.9% |
| 40 (Q40) | 1 in 10,000 | 99.99% |

*Modified from: https://www.illumina.com/science/technology/next-generation-sequencing/plan-experiments/quality-scores.html*

Q40 is the maximum Q Score. Although base calls could theoretically have an inferred accuracy greater than 99.99%, the Q Score is capped at a value of 40 based on the set of ASCII text characters used to encode the data. The Illumin HiSeq PE data had a mean quality score of 36.82,

which equates to an inferred base call accuracy of 99.98%. The percentage of bases with a Q Score ≥ 30 was 86.99%, which exceeded GENEWIZ's quality guarantee of at least 80% with a Q Score ≥ 30.

### PacBio Sequel sequencing

GENEWIZ produced PacBio Sequel totaling 11,199,225 reads. The yield of 165,845 Mbp exceeded the requested yield of 160,000 Mbp requested in the solicitation. The data was provided in BAM format, which contains raw base-call data generated by the sequencing instrument. Unlike Illumina HiSeq data, raw data from PacBio Sequel runs do not contain quality scores. Because PacBio provides long read lengths from circularized templates, it has the potential to read a single template multiple times. Multiple reads of the template insert (the target DNA, with ligated adapters excluded) can be aligned with one another to form a consensus sequence, referred to as the Circular Consensus Sequence (CCS) read. A quality score can then be calculated for the CCS based on the percent identity between the included subreads. CCS reads from the PacBio Sequel runs had a mean read score of 0.988, with an average of 9.8 read passes per insert. There were 33,350 CCS reads, totaling 308 Mbp of data.

## Genome assembly

All *de novo* genome assembly for this project was performed by Dr. Kevin Kocot at the University of Alabama. Numerous software packages have been developed for *de novo* assembly of large genomes from long-reads, such as those generated by PacBio Sequel SMRT sequencing. Because each genome varies in terms of organization and complexity, and each sequencing project is unique with regard to data quality and quantity, there is no one size fits all solution for genome assembly, particularly with a novel species such as the quagga mussel. Optimization of the assembly is therefore necessarily an iterative process, and a variety of approaches were evaluated in this project to optimize the assembly of the genome. In addition, multiple different software programs are frequently employed in what is referred to as a pipeline, with individual programs used to handle specific tasks, such as assembly, proofreading, and removal of haplotigs (discussed below).

### Metrics of assembly optimization: QUAST and BUSCO

Metrics used to assess the quality of each generated assembly were derived from two programs, QUAST (Gurevich et al., 2013) and BUSCO (Simão et al., 2015). QUAST calculates summary statistics which are internal measures of the assembly. In particular, it calculates the total length of the assembly (total number of bases in all contigs), the N50 value for the assembly, and the total number of contigs in the assembly. N50 is defined as "the largest length *L* such that 50% of all nucleotides are contained in contigs of size at least *L*" (International Human Genome Sequencing Consortium, 2001). N50 is considered a more informative measure of assembly contiguity than either the mean or median of the contig dataset. An optimized assembly should have a total length close to that predicted for the genome, while maximizing the N50 value and minimizing the number of contigs. An assembly that optimizes these values should be closest to the true content of the source genome (minimizing exclusion or duplication of genome content) and minimize fragmentation of the genome sequence into separate contigs.

BUSCO is a tool for measuring assembly completeness by looking for genes in the assembly that are known to exist as conserved single-copy genes in the genomes of related organisms. BUSCO scores are calculated as the total percentage of genes from a reference gene set that are identified in a queried assembly, as well as the proportion that are single copy and complete, duplicated, partial, or fragmented. It is expected that an optimized assembly will maximize the total BUSCO score and the proportion that are single-copy and complete. The BUSCO metazoan gene set (version metazoan_odb9 or metazoan_odb10) was selected as the most appropriate benchmark.

## PacBio Sequel SMRT assembly: Canu and Flye

Canu (Koren et al., 2017) is an assembly program designed for single-molecule long-reads, such as those generated from the PacBio Sequel SMRT runs. Canu has been widely adopted and cited, having been employed for *de novo* assembly for a broad diversity of organisms with different genome sizes. Preliminary assembly of the PacBio Sequel SMRT data using Canu produced an assembly with a total length of 2,843 Mbp in 8,960 contigs, and an N50 of 770,235 bp (Table 2). BUSCO scoring for the Canu assembly identified 93.2% of the orthologs as present and complete, with 28.6% as single-copies and 64.6% duplicated (Table 3).

Flye (Kolmogorov et al., 2019), an alternate long-read assembler, was also tested with the PacBio Sequel SMRT data. The total assembly length from Flye was 2,494 Mb, closer to the predicted quagga mussel genome size than the Canu assembly. However, the assembly was comprised of 13,756 contigs with a an N50 of 29,818 bp. Given that the Flye assembly was more fragmented than the Canu assembly, further testing and analysis with this program was not pursued.

Table 2 Genome assembly statistics

| Assembly | Contigs/Scaffolds | Total Length (bp) | N50 |
|---|---|---|---|
| Canu | 8,960 | 2,843,287,607 | 770,235 |
| Canu + POLCA | 8,960 | 2,843,176,422 | 770,322 |
| Canu+ redundans | 3,520 | 2,146,084,521 | 1,163,125 |
| Canu + purge_haplotigs | 2,096 | 1,674,802,492 | 1,516,778 |
| Canu + purge_dups | 2,714 | 1,612,507,440 | 1,525,161 |
| Canu + purge_dups + POLCA | 2,714 | 1,612,996,050 | 1,525,599 |
| Canu + purge_dups + POLCA + Proximo | 1,125 | 1,613,161,677 | 99,743,722 |

Table 3 Genome assembly BUSCO scores

| Assembly | Complete (%) | Single (%) | Duplicate (%) | Fragmented (%) | Missing (%) |
|---|---|---|---|---|---|
| Canu | 93.2 | 28.6 | 64.6 | 0.6 | 6.2 |
| Canu + Pilon | 93.3 | 25.4 | 67.9 | 0.4 | 6.3 |
| Canu+ redundans | 91.9 | 54.6 | 37.3 | 0.7 | 7.4 |
| Canu + purge_haplotigs | 84.2 | 74.1 | 10.1 | 1.9 | 13.9 |
| Canu + purge_dups | 93.9 | 89.4 | 4.5 | 1.7 | 4.4 |
| Canu + purge_dups + POLCA | 94.4 | 89.8 | 4.6 | 1.5 | 4.1 |
| Canu + purge_dups + POLCA + Proximo | 94.7 | 92.6 | 2.1 | 1.5 | 3.8 |

## Hybrid assembly: MaSuRCA and SPAdes

Hybrid assembly of PacBio Sequel SMART and Illumina HiSeq PE data: MaSuRCA and SPAdes Canu, discussed above, utilized only the PacBio Sequel SMRT data, which has a comparatively high error rate. An alternate approach is to perform a so-called hybrid assembly, which combines from both the PacBio Sequel SMRT (long-reads, low accuracy) and Illumina HiSeq PE (short-reads, high accuracy). Such an approach can ideally improve both the N50 and the accuracy of the assembly. Two widely used hybrid assembly programs, MaSuRCA (Zimin et al., 2013) and hybridSPAdes (Antipov et al., 2016) were tested. Both of these programs were found to be too computationally expensive to run effectively with the available computational resources. Runs with these programs were halted before they reached completion to free resources for other analyses.

## Assembly polishing: Pilon and POLCA

As an alternative to direct hybrid assembly, a two-step approach was tested. Rather than incorporating PacBio Sequel SMRT and Illumina HiSeq PE data with a single program, the two datasets were utilized in a stepwise process wherein the Illumina HiSeq PE data were used to proofread the PacBio Sequel SMRT-based Canu assembly. This process of error correction with high accuracy Illumina short-read data is referred to as polishing. Polishing was initially performed with Pilon (Walker et al., 2014). Polishing the Canu assembly with four rounds of Pilon analysis and correction produced an assembly with a total size of 2,843 Mbp, 8,960 contigs, and an N50 of 770,322. BUSCO scoring for the polished assembly identified 91.9% of the orthologs as present and

complete, with 54.6% as single-copy and 37.3% duplicated (see "Heterozygosity and removal of haplotigs" section below).

Subsequent polishing in later analyses was performed with POLCA (Zimin & Salzberg, 2020), which became available later in the project. POLCA was selected for the final assembly based on data showing that it could outperform Pilon both in error correction, and in minimizing the number of new errors introduced during polishing.

## Heterozygosity and removal of haplotigs

The Pilon-polished Canu assembly size of 2,843 Mbp was approximately 1.7 times larger than the predicted haploid genome size. Although almost all animal genomes are diploid (that is, having two sets of chromosomes, one from each parent) and *de novo* genome assembly projects generally aim to reconstruct a haploid reference genome. Haploid reference genomes are advantageous downstream analyses as they contain only one copy of each locus in the genome. In the case of the Canu assembly, the large size of the assembly was most likely due to the inclusion of haplotigs. Haplotigs are homologous maternal and paternal chromosome regions that differ in sequence composition and are assembled independently. This conclusion was supported by the high percentage (67.9%) of duplicated genes identified in the BUSCO analysis. The presence of haplotigs in the assembly was likely due to high levels of heterozygosity in the sequenced genome. High levels of heterozygosity have been documented in the genomes of other bivalve genomes, with rates generally 10 times or more what is observed in human genomes. To evaluate heterozygosity in the quagga mussel genome we analyzed the Illumina HiSeq PE data with GenomeScope (Vurture et al., 2017), which uses a *k-mer* profile derived from analysis of short reads to estimate genome characteristics, including the heterozygosity rate (Figure 2). GenomeScope estimated the heterozygosity rate of the quagga mussel genome at 2.45%, which is approximately 10 times the rate observed in human genomes.

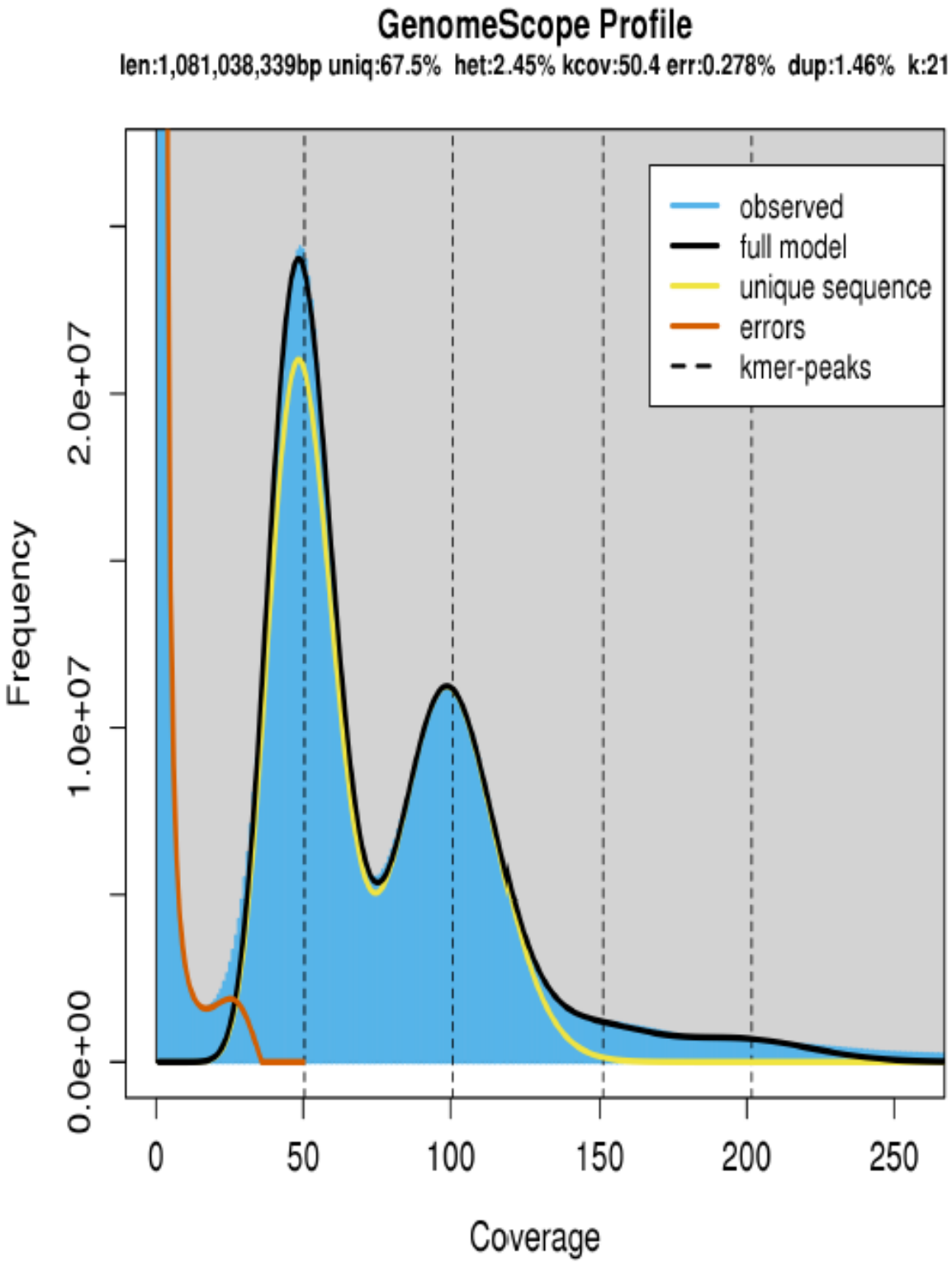


Figure 2 GenomeScope analysis of Illumina HiSeq PE data

To resolve the haplotigs in the Canu assembly, we tested three programs: Redundans (Pryszcz & Gabaldón, 2016), Purge Haplotigs (Roach et al., 2018), and purge_dups (Guan et al., 2020), which are all designed specifically to identify and remove haplotigs from assemblies of highly heterozygous genomes.

### *Redundans*

Redundans filtering of the Pilon-polished Canu assembly reduced a total assembly size to 2,146 Mbp. This is approximately 30% larger than the predicted genome size. In addition, while the number of duplicated BUSCOs were reduced from 67.9%, they remained higher than expected at 37.3%. The number of complete BUSCOs also decreased slightly from 93.3% to 91.9%.

### *Purge Haplotigs*

Purge Haplotigs re-assembly reduced to the total assembly size to 1,675 Mbp, close to the predicted genome size. The number of duplicated BUSCOs was reduced to 10.1%, however the total complete BUSCOs were also reduced to 84.2%. This suggests that the software may have purged the input assembly too aggressively, discarding unique sequences.

### *Purge_dups*

Purge_dups displayed the best performance in removing haplotigs when paired with POLCA polishing of the Canu assembly. The purge-dups re-assembly had a total size of 1,613 Mbp, matching the predicted genome size. N50 of 1,525,599 from this pipeline was also the largest recovered from all methods tested. BUSCOs for the purge_dups re-assembly were 94.4% compete, with only 4.6% duplicated.

## Hi-C Scaffolding: Chromosome-scale assembly

The iterative process described above culminated with the use of Canu for primary assembly from PacBio Sequel SMRT data, POLCA for polishing with Illumina HiSeq PE data, and purge_dups for removal of haplotigs. The resultant genome assembly was deemed to be of high quality based on metrics total assembly size (matching the predicted size for the genome), contiguity (N50 statistic), and completeness (BUSCO scores).

However, in recent years there has been a push to carry data beyond fragmented primary assemblies (even if high-quality) to chromosome-scale reference assemblies. Chromosome-scale assemblies seek to link individual contigs together in an order representative of their positions on the chromosomes that are the largest individual units of genome organization. Development of chromosome-scale assemblies builds upon contig assemblies through the linkage of non-overlapping sequences in a process referred to as scaffolding. The scaffolds resulting from these assemblies are referred to as pseudomolecules.

Classically, chromosome-scale genome assembly was a labor and cost intensive process that was reserved for large-scale projects such as the human genome. In recent years, technical advances coupled with the affordability of NGS have made new approaches to chromosome-scale genome assembly available for less well studied organisms.

For the quagga genome a technology called Hi-C was selected for chromosome-scale assembly. Hi-C relies on the organization of DNA in the chromosome to map the order of DNA sequences. Although DNA is generally thought of as a linear strand of nucleotides coiled in a double helix, in the chromosome DNA is part of a complex 3-dimensional structure called chromatin. Chromatin contain not only DNA, but also proteins and RNA which are bound to it, and which organize the DNA and regulate gene expression. The Hi-C methodology relies on the fact that although chromatin has a complex structure, regions of DNA that are closer to each along the primary strand of the double-helix are most likely to be in close physical proximity to each other in the chromatin complex. Hi-C cross-links the proteins in the chromatin, locking together physically distinct regions of DNA. The DNA can then be fragmented, labeled, and re-ligated. Sequencing of the results labeled DNA fragments and subsequent bioinformatic analysis can then be used to identify fragments of DNA that are at different locations on the primary strand but in close proximity in the chromatin. This information can then be used to scaffold the contigs described above by mapping their relative order and orientation.

Hi-C scaffolding from Phase Genomics was selected for this project based on a track record of successful projects (including for the zebra mussel, *Dreissena polymorpha*), likelihood of success given the specifics of the quagga mussel project, and pricing appropriate to the project budget. To accomplish DNA sequencing, intact frozen tissue was sent to Phase Genomics for processing. Ideally, tissue for Hi-C scaffolding would come from the same individual used from primary sequencing and contig assembly. However, all available tissue from the mussel Drb016 was used for DNA extraction to accomplish PacBio Sequel and Illumina HiSeq. Based on discussions with Phase Genomics it was determined that tissue from another individual could likely be used successfully for the Hi-C assembly, and samples from a second individual.

As the aim of Hi-C scaffolding was to achieve chromosome-scale assembly of the genome, an important piece of information is the actual number of chromosomes present in quagga mussel cells. The complete chromosome count for an organism is termed the karyotype. Karyotyping is generally accomplished through photomicrographic image analysis of the chromosomes, which have been specifically labeled with either a fluorescent dye or a stain to facilitate detection and imaging. An accurate karyotype for the quagga mussel is not currently available. Work on this objective (outside the proposed scope of this project) is underway in collaboration with Biomilabs, LLC, but has not been completed. However, a karyotype is available for the congener zebra mussel, which has been identified as having 16 pairs of chromosomes. For the purpose of Hi-C assembly, Proximo is designed to work with a predicted number of chromosomes. For the purpose of the initial assembly the number of potential scaffolds (corresponding to chromosomes) was constrained to be between 14 and 20. This was based on a literature search to identify the range of chromosome numbers reported in the bivalve clade Heterodonta, of which the quagga mussel is a member. Because Proximo tends to work conservatively and produce a larger number of scaffolds, the initial analysis resulted in an assembly with 20 scaffolds. However, further analysis and refinement by Phase Genomics scientists resolved the assembly into 16 scaffolds. This was determined to represent the optimal scaffold number, and the final assembly was constrained to this value. A list of sizes for scaffolds in the final assembly is shown in Table 4.

Table 4 Hi-C/Proximo scaffolds

| Scaffold Number | Number of Contigs | Length of Scaffold |
|---|---|---|
| 0 | 160 | 141743199 |
| 1 | 133 | 124375728 |
| 2 | 114 | 120602297 |
| 3 | 109 | 115470733 |
| 4 | 115 | 114264553 |
| 5 | 100 | 106848640 |
| 6 | 118 | 99731922 |
| 7 | 104 | 93185342 |
| 8 | 92 | 90724541 |
| 9 | 96 | 88284583 |
| 10 | 73 | 70424728 |
| 11 | 110 | 88467619 |
| 12 | 104 | 84919273 |
| 13 | 93 | 76501195 |
| 14 | 66 | 73228477 |
| 15 | 70 | 55513448 |
| Total | 1657 | 1544286278 |

For the initial Hi-C assembly, Phase Genomics was provided the Canu assembly which had been polished with Pilon and analyzed with Redundans to remove haplotigs. Although this assembly was known to be larger than the expected genome size, and to have a high level of duplicated BUSCOs, discussions with Phase Genomics suggested that Proximo could handle removal of haplotigs to achieve an accurate reference sequence. However, the resultant scaffolded assembly retained 3,267 of the 3,520 input scaffolds and had a total size of 2,129 Mbp. This suggested that haplotigs were not purged during Proximo scaffolding. This was further confirmed by BLAST searches for genes expected to be single copy. In several cases two copies of the gene were identified, matching what was observed in the input assembly. Further investigation demonstrated that in each of these cases the two haplotigs where scaffolded as tandem repeats. That is, matching haplotigs appeared to have been linked in close proximity to one another in the scaffolding, generally within 1 to 2 Mbp. While such tandem repeats can arise as a result of gene duplication during genome evolution, they would be expected to be observed at very low frequency. All the available evidence suggested that Proximo had not effectively detected and removed haplotigs.

Further discussions with Phase Genomics technical staff confirmed that Proximo was not optimized for haplotig identification and removal. It was therefore agreed that the best plan of action was to develop an assembly with more complete removal of haplotigs, which could be resubmitted to Proximo for scaffolding. This prompted further testing of methods for haplotig removal, and ultimately lead to development of the assembly which used purge_dups, as described above. This updated primary assembly was resubmitted to Phase Genomics, which conducted scaffolding on the new dataset with Proximo. The resultant assembly produced 16 pseudomolecule scaffolds with a

total length of 1,544 Mbp. These pseudomolecules included 1,657 of the 2,766 contigs from the input assembly. BUSCO scoring for the complete assembly, including both the 16 scaffolds and the 1,109 unincorporated contigs, identified 94.7% of the orthologs as present and complete, with 92.6% as single-copies and 2.1% duplicated.

## RNASeq: sequencing of gene transcripts

Although the primary goal of this project was the sequencing and assembly of a quagga mussel genome, RNA sequencing (RNASeq) was also pursued to aid in future annotation and characterization of the genome. A main feature of analysis is the identification of genes and the proteins they encode. Protein synthesis involves transcription of genomic gene sequences to form messenger RNA (mRNA) molecules, which are subsequently translated in the polypeptide chain of amino acids that form the protein. Sequencing of mRNA is a powerful tool for gene identification and characterization because it can be accomplished using the same NGS tools that are available for DNA, allowing for the cost-effective generation of vast amounts of sequence data. In addition, mature mRNA molecules lack intronic sequence, non-coding regions that are interspersed in gene loci in most eukaryotes. This facilitates both the annotation of genes in the genome, as well as the characterization of the proteins they encode.

Six samples were prepared for RNASeq, two whole mussels and four individual tissue samples from gill, foot, ovary, and testes and digestive gland (combined). RNA isolation and purification were performed with the RNAEasy Mini kit (QIAGEN Inc., USA), following the manufacturer's protocol. A DNAse treatment was included to degrade DNA in the sample, which can interfere with subsequent sample analysis and sequencing. RNA yield was quantified using the Qubit RNA BR Assay Kit (ThermoFisher, USA) to ensure the quantity of RNA recovered was sufficient for sequencing. RNA quality was initially assessed by spectrophotometric measurement of the 260/280 ratio on the Cary-60, as described above. The expected 260/280 ratio for RNA is 2.0, slightly higher than the value for DNA. Measured 260/280 values for the RNA samples ranged from 1.5 to 2.2 (Table 5).

Table 5 RNA samples

| Sample | Concentration (ng/µl) | 260/280 ratio | RIN value |
|---|---|---|---|
| Whole_mussel_2 | 362 | 2.0 | 9.4 |
| Whole_mussel_3 | 885 | 1.5 | 8.8 |
| Foot | 356 | 2.2 | 9.0 |
| Gill | 317 | 2.2 | 9.7 |
| Ovary | 206 | 1.6 | 10.0 |
| Testes + digestive gland | 550 | 1.9 | 9.1 |

RNA samples were sent to a commercial provider, Macrogen USA, for RNA sequencing. Prior to sequencing, Macrogen USA ran additional quality control checks on the sample. RNA is much less stable than DNA due to its physical and chemical structure and can rapidly be degraded by RNAse enzymes if careful sample storage and handling are not maintained. To evaluate RNA integrity,

samples were run on an Agilent TapeStation system using ScreenTape reagents. The TapeStation is an automated system for fluorescent electrophoresis of samples. The electropherogram produced by the system allows for measurement and quantification of RNA size distributions. In a high-quality sample without degradation, prominent peaks are observed that correspond to the large and small nuclear ribosomal genes, referred to as 28S and 18S rRNA in animals. This class of RNA is in fact excluded during library prep for RNASeq, but it provides an important metric of sample quality. mRNA, which is the target of RNASeq, is much more complex in size and composition, with thousands of different molecules present in a sample. In an undegraded sample these molecules for a broad lower level signal on the electropherogram. In samples that are degraded the height of the 18S and 28S rRNA peaks decrease relative to a broad smear of smaller sized fragments. Agilent has developed a proprietary algorithm for quantification of TapeStation RNA sample results called the RNA Integrity Number (RIN). RIN values are unitless and are measured on a scale of 1 to 10, with a RIN of 10 representing a sample with no measured degradation, and a RIN of 1 representing a sample with complete degradation (Figure 3) (Schroeder et al., 2006). For purposes of sample analysis, RIN numbers higher than 8 are generally considered to represent very high quality, with smaller numbers representing increasing degrees of degradation and concomitant decreases in sample quality. RIN scores for all the samples provided to Macrogen were extremely high, ranging from 8.8 up to 10 (Table 4).

Although the RIN scores for the samples were extremely high, Macrogen placed a hold on further sample processing due to an anomaly in the TapeStation reports. For all the samples there was prominent 18S rRNA peak, but the 28S rRNA peak was entirely absent (Figure 3). As the high RIN scores indicate, there were no other signs to indicate the sample degradation. A literature search suggested an alternate degradation wherein the absence of the 28S rRNA peak in the electropherogram could be due to splitting of the 28S rRNA in separate 'α' and 'β' units (Winnebeck et al., 2010). This phenomenon of "gap deletion" or a "hidden break" has been reported from a variety of taxa (Melen, 1999; S. Sun et al., 2012; Asai et al., 2015; McCarthy et al., 2015; DeLeo et al., 2018; Navarro-Ródenas et al., 2018), including anecdotal reports from marine bivalve mollusks (Barcia et al., 1997; Moreira et al., 2014). Under the heat denaturation conditions used in the TapeStation, it appears that the hydrogen bonds linking the 28Sα and 28Sβ units are broken, and the two fragments run independently on the ScreenTape. The two fragments are smaller than the complete 28S rRNA molecule, and in fact both are close in size to the 18S rRNA molecule. The results is that the 18S, 28Sα, and 28Sβ signals overlap with one another on the ScreenTape electropherogram, making it appear as though there is a very large 18S rRNA band and no 28S rRNA band. RNA extract samples were analyzed with reverse transcription PCR (RT-PCR) to test this hypothesis, but the results were inconclusive. Given the high RIN numbers from the TapeStation analyses, it appeared that 28S rRNA fragmentation, which is a well-known occurrence from diverse invertebrate taxa (DeLeo et al., 2018), was the most likely explanation for the electropherogram results. It was therefore most efficacious to continue with RNASeq processing of the samples that had been sent to Macrogen, and to discontinue investigation of an issue that was ancillary to the main project goals.

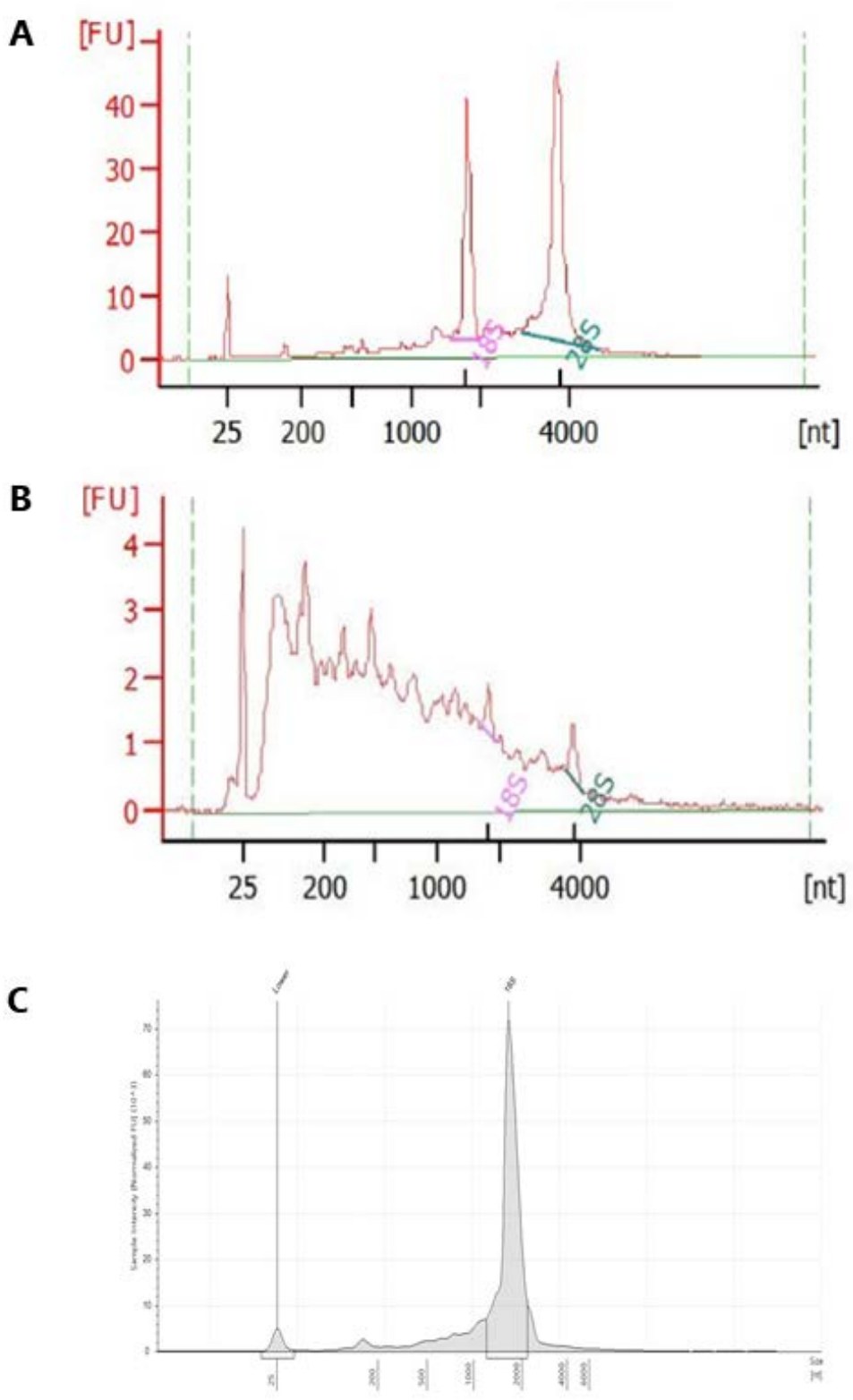


Figure 3 ScreenTape electropherograms. (A) Example of high-quality RNA, RIN = 9.0; (B) example of degraded RNA, RIN = 2.7; (C) quagga mussel RNA from sample 'Whole_mussel_2', RIN = 9.4.

Macrogen performed sample library preparation with the Illumina TruSeq Stranded mRNA kit and ran Illumina 150 bp PE sequencing. The service request was for 100 million PE reads per sample. Data returned by Macrogen met this goal, with read totals (two reads per PE sequence) between 198.6 million and 255.7 million. The proportion of bases with a Q-score ≥ 30 averaged 91.85% across the six samples, with the minimum being 89.95%.

## Gene identification

A key aspect of genome analysis is the identification of protein-coding genes. Although a detailed annotation of the assembled genome was outside the scope for the current project, preliminary gene identification was conducted. Two approaches were pursued in this project to find the broad complement quagga mussel genes. The first was to assemble transcripts from the RNASeq transcriptome project. The second was in *silico* gene prediction from the final Hi-C genome assembly.

### Transcriptome assembly and translation

Assembly of data from the RNASeq project was performed using Trinity (Grabherr et al., 2011). Trinity is a widely used software package which performs well at *de novo* transcriptome assembly from the relatively short reads produced by Illumina PE sequencing. The goal of the Trinity assembly is to reconstruct the complement of RNA transcripts in the source sample used for sequencing. Over 200,000 transcripts were assembled from each library (Table 6). When Trinity was used to assemble all the RNASeq data composited into a single dataset, 842,346 transcripts were recovered. These numbers are not reflective of the number of genes in the genome. Trinity is designed to recover splice-variants, so a single gene may be represented multiple transcripts in the assembly. In addition, although the TruSeq kit used for library construction is designed to select for mRNA containing coding sequence (CDS), some non-coding RNA may have been included in the library and the subsequent assembly. To identify transcripts that are mRNA and to characterize the CDS of these sequences, Transcoder (https://github.com/TransDecoder/TransDecoder) was used to analyze and filter the assemblies (Tables 7 and 8). Transcoder searches for open reading frames and then writes the filtered data to three FASTA format text files. .mRNA files contain the full sequence for all transcripts identified as mRNA, including 5' and/or 3' untranslated regions (UTR), if present. .cds files contain only the CDS portion of the sequences. .pep files contain the predicted peptide sequence from the CDS, translated using the standard genetic code for assignment of amino acids for codons. .bed and .gff3 format files are metadata files containing information on the size of each mRNA transcript, and positional information for the CDS, 5' UTR and 3' UTR.

Table 6 Trinity transcriptome assembly data

| Sample | Total sequences | Total length (nt) | Minimum length (nt) | Average length (nt) | Maximum length (nt) |
|---|---|---|---|---|---|
| Whole_mussel_2 | 273239 | 262308260 | 182 | 960 | 32204 |
| Whole_mussel_3 | 258285 | 242077236 | 180 | 937.2 | 32018 |
| Foot | 215610 | 150394534 | 179 | 697.5 | 22849 |
| Gill | 270155 | 265610218 | 178 | 983.2 | 37470 |
| Ovary | 265100 | 199757350 | 178 | 753.5 | 29988 |
| Testes | 253265 | 239014924 | 187 | 943.7 | 32754 |
| Composite | 842346 | 623066925 | 165 | 739.7 | 35546 |

Table 7 Transcoder filtered transcriptome coding sequence (CDS) data

| Sample | Total sequences | Total length (nt) | Minimum length (nt) | Average length (nt) | Maximum length (nt) |
|---|---|---|---|---|---|
| Whole_mussel_2 | 110217 | 133344471 | 297 | 1209.8 | 30018 |
| Whole_mussel_3 | 101569 | 121633554 | 297 | 1197.5 | 31146 |
| Foot | 59555 | 51454686 | 297 | 864 | 21489 |
| Gill | 94618 | 116021511 | 297 | 1226.2 | 37467 |
| Ovary | 65263 | 79380534 | 297 | 1216.3 | 29055 |
| Testes | 95075 | 112063776 | 297 | 1178.7 | 32625 |
| Composite | 214103 | 222210702 | 297 | 1037.9 | 33216 |

Table 8 Transcoder filtered transcriptome coding mRNA data

| Sample | Total sequences | Complete | 5' Partial | 3' Partial | Internal |
|---|---|---|---|---|---|
| Whole_mussel_2 | 110217 | 57147 | 21712 | 11091 | 20267 |
| Whole_mussel_3 | 101569 | 51896 | 20548 | 10005 | 19120 |
| Foot | 59555 | 22133 | 18691 | 4595 | 14136 |
| Gill | 94618 | 53595 | 18228 | 8427 | 14368 |
| Ovary | 65263 | 35416 | 13757 | 5256 | 10834 |
| Testes | 95075 | 50716 | 19994 | 8166 | 16199 |
| Composite | 214103 | 95569 | 42810 | 24788 | 50936 |

## *In silico* gene prediction

Gene prediction from the Hi-C-scaffolded genome assembly was performed using the BRAKER pipeline (Hoff et al., 2019). BRAKER employs hidden Markov Models (HMMs) for genome analysis and gene prediction. BRAKER was trained on a subset of the RNASeq assembly data to develop

training parameters. BRAKER initially identified 109,614 gene models. BUSCO analysis of this dataset returned a score of 95.6% complete, with 12.6% duplicated. BLASTN comparison of the BRAKER dataset to the composite transcriptome identified 36,773 models that had 99% or greater sequence similarity to sequences derived from RNASeq. This work is still underway as the expected number of genes is on the order of 20,000-30,000 and many of the 109,614 gene models are expected to be false positives.

# Discussion

## Project outcome

The current project successfully sequenced and assembled a reference genome for the quagga mussel (*Dreissena rostriformis bugenesis*). The final assembly size of 1,613 Mbp closely matched the predicted genome size. Scaffolding of the assembly using Hi-C and Proximo analysis resolved the genome into 16 chromosome-scale pseudomolecules and included 95.7% of the bases in the input assembly. BUSCO analyses found high completeness scores for the assembly, with analysis of BRAKER derived gene models having greater than 95% complete BUSCOs. RNASeq data from two complete individuals and four isolated tissues was used to assemble transcriptomes. The transcripts from these datasets provide a resource for gene identification.

## Comparative genomics

During the current project two other studies on dreissenid genomics were published. Another project to sequence a quagga mussel collected in Europe (Calcino et al., 2019) focused primarily on identification of genes involved in osmoregulation. This project relied solely on Illumina HiSeq data, and the resultant assembly had an N50 of 131.4 kb. The total size of the assembly was 1,242 Mbp. The fact that this assembly is smaller than the predicted genome size of 1,600 Mbp may be partially due to collapse of highly repetitive regions that are not distinguished from one another in the relatively short Illumina reads. The BUSCO scores for this assembly were lower than for the scaffolded assembly in the current project, with a completeness of 83.2%. Assembly of a genome for the zebra mussel (*Dreissena polymorpha*) (McCarthy, 2019) also became available through the preprint repository bioRxiv. The zebra mussel project used a similar strategy to the current project, combining data from both PacBio Sequel and Illumina HiSeq, and scaffolding with Hi-C. The resultant assembly was 1,798 Mbp in length, with contigs scaffolded in 16 pseudomolecules. The zebra mussel assembly had a BUSCO completeness score of 92.3%. Both of these genomes will serve as valuable comparators for the quagga mussel genome assembly developed in the current project. Comparison with the zebra mussel genome will serve to identify similarities and differences between these two related, invaders. Comparison of our assembly with the genome from a European quagga mussel will also illuminate what genomic features are common across the species and what features may be unique in the North American population.

Genome assemblies are currently available for at least 24 other bivalves, primarily from economically important marine species. Comparison with these other data shows that the quagga mussel genome assembly is one highest quality bivalve genomes available, both in terms of contiguity and completeness, as based on BUSCO scores for both complete and missing orthologs (Figure 4).

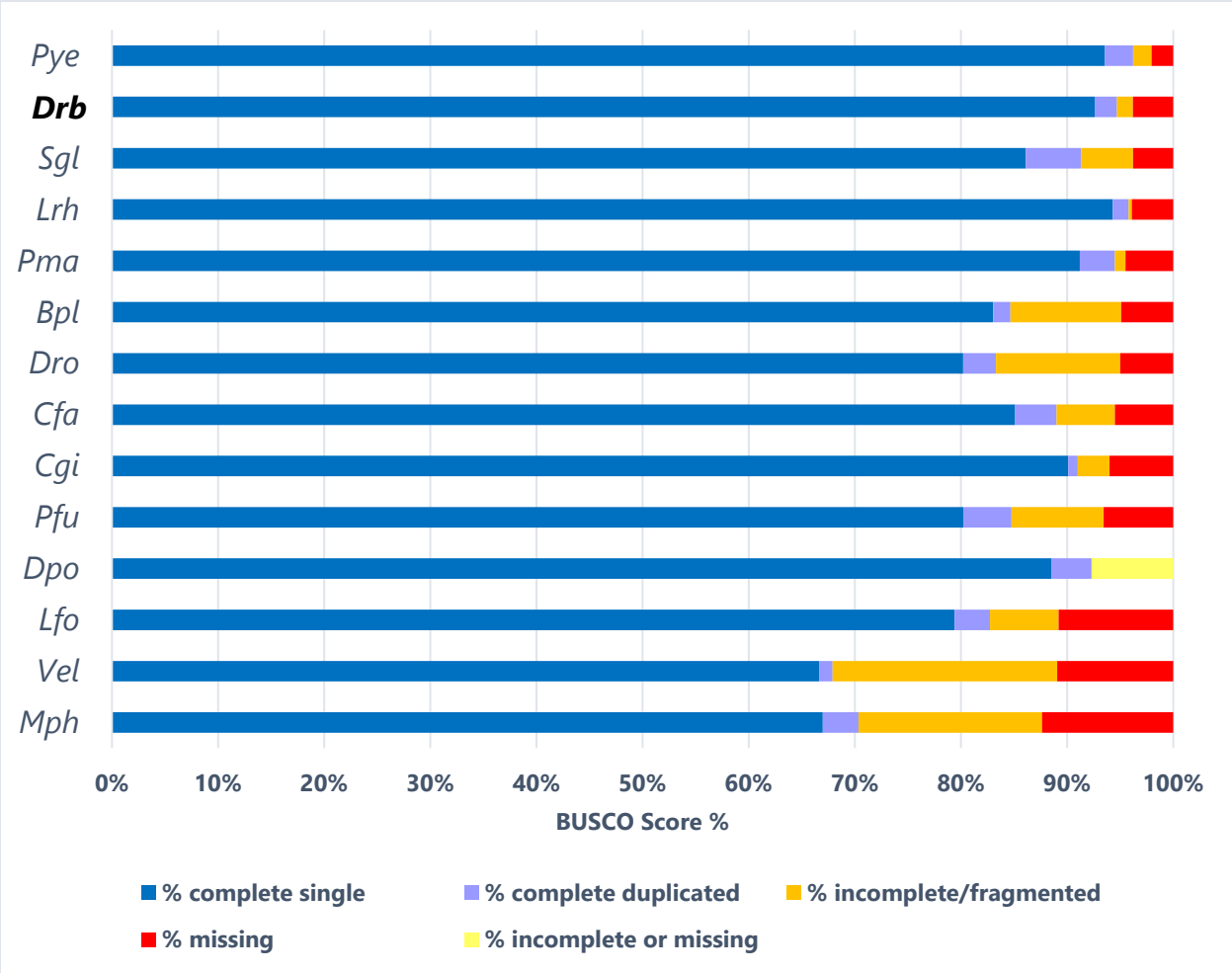


Figure 4 BUSCO scores for *Dreissena rostriformis bugensis* and other representative published bivalve genomes. Species are ordered from lowest to highest percentage of missing BUSCOs. For *Dreissena polymorpha* (*Dpo*), only a combined value for missing and fragmented BUSCOs was available. Species name abbreviations and genome publications are as follows: *Pye -Patinopecten yessoensis* (S. Wang et al., 2017), *Drb – Dreissena rostriformis bugensis* (current study), *Sgl - Saccostrea glomerate* (Powell et al., 2018), *Lrh - Lutraria rhynchaena* (Thai et al., 2019), *Pma – Pecten maximus* (Kenny et al., 2020), *Bpl - Bathymodiolus platifrons* (J. Sun et al., 2017), *Dro - Dreissena rostriformis* (Calcino et al., 2019), *Cfa – Chlamys farreri* (Li et al., 2017), *Cgi – Crassostrea gigas* (X. Wang et al., 2019), *Pfu - Pinctada fucata* (Takeuchi et al., 2016), *Dpo - Dreissena polymporpha* (McCartney et al., 2019), *Lfo - Limnoperna fortunei* (Uliano-Silva et al., 2018), *Vel - Venustaconcha ellipsiformis* (Renaut et al., 2018), *Mph - Modiolus philippinarum* (J. Sun et al., 2017)

Note that BUSCO scores are not all derived from analyses with the same parameters and reference gene set (metazoa_odb9 or metazoan_odb10 were used for different species). Data should be viewed as representative of relative completeness, rather than being directly comparable to one another.

## Future directions

The genome assembly and transcriptomes developed during this project will serve as valuable resources in efforts to develop genetic biocontrols against quagga mussels. These datasets may also be used to better understand quagga mussel ecology, as well and the linkage between source populations and new introductions. To maximize the utility of the genomic data, additional work will be needed. Annotation of the genome is necessary to demarcate where genes, repetitive elements, and other features of interest reside in the genome. Such identification of genes will be a critical next step to understanding the genome, but it will not identify the roles of individual genes, although in some cases this may be posited based on data from other taxa. Functional genomic approaches should be pursued through the development of fine-scale temporal and spatial transcriptome datasets. Such investigations are expected to aid in elucidating the role of genes. This will be an important next step to maximize the utility of the genome data. Public release of annotated genome and transcriptome data through the National Center for Biotechnology Information will open the data to be used in a wide range of applications by researchers outside of Reclamation.

# References

Antipov, D., Korobeynikov, A., McLean, J. S., & Pevzner, P. A. (2016). hybridSPAdes: An algorithm for hybrid assembly of short and long reads. *Bioinformatics*, *32*(7), 1009–1015. https://doi.org/10.1093/bioinformatics/btv688

Asai, S., Ianora, A., Lauritano, C., Lindeque, P. K., & Carotenuto, Y. (2015). High-quality RNA extraction from copepods for Next Generation Sequencing: A comparative study. *Marine Genomics*, *24*, 115–118. https://doi.org/10.1016/j.margen.2014.12.004

Barcia, R., Lopez-García, J. M., & Ramos-Martínez, J. I. (1997). The 28S fraction of rRNA in molluscs displays electrophoretic behaviour different from that of mammal cells. *IUBMB Life*, *42*(6), 1089–1092. https://doi.org/10.1080/15216549700203551

Calcino, A. D., de Oliveira, A. L., Simakov, O., Schwaha, T., Zieger, E., Wollesen, T., & Wanninger, A. (2019). The quagga mussel genome and the evolution of freshwater tolerance. *DNA Research*, *26*(5), 411–422. https://doi.org/10.1093/dnares/dsz019

DeLeo, D. M., Pérez-Moreno, J. L., Vázquez-Miranda, H., & Bracken-Grissom, H. D. (2018). RNA profile diversity across arthropoda: Guidelines, methodological artifacts, and expected outcomes. *Biology Methods and Protocols*, *3*(1). https://doi.org/10.1093/biomethods/bpy012

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., Chen, Z., Mauceli, E., Hacohen, N., Gnirke, A., Rhind, N., di Palma, F., Birren, B. W., Nusbaum, C., Lindblad-Toh, K., … Regev, A. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology*, *29*(7), 644–652. https://doi.org/10.1038/nbt.1883

Gregory, T. R. (2003). Genome size estimates for two important freshwater molluscs, the zebra mussel (*Dreissena polymorpha*) and the schistosomiasis vector snail (*Biomphalaria glabrata*). *Genome*, *46*(5), 841–844.

Guan, D., McCarthy, S. A., Wood, J., Howe, K., Wang, Y., & Durbin, R. (2020). Identifying and removing haplotypic duplication in primary genome assemblies. *Bioinformatics*, *36*(9), 2896–2898. https://doi.org/10.1093/bioinformatics/btaa025

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

Hoff, K. J., Lomsadze, A., Borodovsky, M., & Stanke, M. (2019). Whole-Genome Annotation with BRAKER. In M. Kollmar (Ed.), *Gene Prediction* (Vol. 1962, pp. 65–95). Springer New York. https://doi.org/10.1007/978-1-4939-9173-0_5

International Human Genome Sequencing Consortium. (2001). Initial sequencing and analysis of the human genome. *Nature*, *409*(6822), 860–921. https://doi.org/10.1038/35057062

Kenny, N. J., McCarthy, S. A., Dudchenko, O., James, K., Betteridge, E., Corton, C., Dolucan, J., Mead, D., Oliver, K., Omer, A. D., Pelan, S., Ryan, Y., Sims, Y., Skelton, J., Smith, M., Torrance, J., Weisz, D., Wipat, A., Aiden, E. L., … Williams, S. T. (2020). The gene-rich genome of the scallop Pecten maximus. *GigaScience*, *9*(5). https://doi.org/10.1093/gigascience/giaa037

Kolmogorov, M., Yuan, J., Lin, Y., & Pevzner, P. A. (2019). Assembly of long, error-prone reads using repeat graphs. *Nature Biotechnology*, *37*, 540–546.

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

Li, Y., Sun, X., Hu, X., Xun, X., Zhang, J., Guo, X., Jiao, W., Zhang, L., Liu, W., Wang, J., Li, J., Sun, Y., Miao, Y., Zhang, X., Cheng, T., Xu, G., Fu, X., Wang, Y., Yu, X., … Bao, Z. (2017). Scallop genome reveals molecular adaptations to semi-sessile life and neurotoxins. *Nature Communications*, *8*(1). https://doi.org/10.1038/s41467-017-01927-0

McCarthy, S. D., Dugon, M. M., & Power, A. M. (2015). 'Degraded' RNA profiles in Arthropoda and beyond. *PeerJ*, *3*, e1436. https://doi.org/10.7717/peerj.1436

McCartney, M. A., Auch, B., Kono, T., Mallez, S., Zhang, Y., Obille, A., Becker, A., Abrahante, J. E., Garbe, J., Badalamenti, J. P., Herman, A., Mangelson, H., Liachko, I., Sullivan, S., Sone, E. D., Koren, S., Silverstein, K. A. T., Beckman, K. B., & Gohl, D. M. (2019). *The Genome of the Zebra Mussel,* Dreissena polymorpha*: A Resource for Invasive Species Research* [Preprint]. Genomics. https://doi.org/10.1101/696732

Melen, G. J. (1999). Novel processing in a mammalian nuclear 28S pre-rRNA: Tissue-specific elimination of an `intron' bearing a hidden break site. *The EMBO Journal*, *18*(11), 3107–3118. https://doi.org/10.1093/emboj/18.11.3107

Moreira, R., Pereiro, P., Costa, M. M., Figueras, A., & Novoa, B. (2014). Evaluation of reference genes of *Mytilus galloprovincialis* and *Ruditapes philippinarum* infected with three bacteria strains for gene expression analysis. *Aquatic Living Resources*, *27*(3–4), 147–152. https://doi.org/10.1051/alr/2014015

Navarro-Ródenas, A., Carra, A., & Morte, A. (2018). Identification of an Alternative rRNA Post-transcriptional Maturation of 26S rRNA in the Kingdom Fungi. *Frontiers in Microbiology*, *9*. https://doi.org/10.3389/fmicb.2018.00994

Pedersen, R. A. (1971). DNA content, ribosomal gene multiplicity, and cell size in fish. *Journal of Experimental Zoology*, *177*(1), 65–78. https://doi.org/10.1002/jez.1401770108

Powell, D., Subramanian, S., Suwansa-ard, S., Zhao, M., O'Connor, W., Raftos, D., & Elizur, A. (2018). The genome of the oyster *Saccostrea* offers insight into the environmental resilience of bivalves. *DNA Research*, *25*(6), 655–665. https://doi.org/10.1093/dnares/dsy032

Pryszcz, L. P., & Gabaldón, T. (2016). Redundans: An assembly pipeline for highly heterozygous genomes. *Nucleic Acids Research*, *44*(12), e113–e113. https://doi.org/10.1093/nar/gkw294

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. https://doi.org/10.1093/gbe/evy117

Roach, M. J., Schmidt, S. A., & Borneman, A. R. (2018). Purge Haplotigs: Allelic contig reassignment for third-gen diploid genome assemblies. *BMC Bioinformatics*, *19*(1). https://doi.org/10.1186/s12859-018-2485-7

Schroeder, A., Mueller, O., Stocker, S., Salowsky, R., Leiber, M., Gassmann, M., Lightfoot, S., Menzel, W., Granzow, M., & Ragg, T. (2006). The RIN: an RNA integrity number for assigning integrity values to RNA measurements. *BMC Molecular Biology*, *7*(1), 3. https://doi.org/10.1186/1471-2199-7-3

Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., & Zdobnov, E. M. (2015). BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, *31*(19), 3210–3212. https://doi.org/10.1093/bioinformatics/btv351

Slyusarev, G. S., Starunov, V. V., Bondarenko, A. S., Zorina, N. A., & Bondarenko, N. I. (2020). Extreme Genome and Nervous System Streamlining in the Invertebrate Parasite Intoshia variabili. *Current Biology*, *30*(7), 1292-1298.e3. https://doi.org/10.1016/j.cub.2020.01.061

Sun, J., Zhang, Y., Xu, T., Zhang, Y., Mu, H., Zhang, Y., Lan, Y., Fields, C. J., Hui, J. H. L., Zhang, W., Li, R., Nong, W., Cheung, F. K. M., Qiu, J.-W., & Qian, P.-Y. (2017). Adaptation to deep-sea chemosynthetic environments as revealed by mussel genomes. *Nature Ecology & Evolution*, *1*(5). https://doi.org/10.1038/s41559-017-0121

Sun, S., Xie, H., Sun, Y., Song, J., & Li, Z. (2012). Molecular characterization of gap region in 28S rRNA molecules in brine shrimp Artemia parthenogenetica and planarian Dugesia japonica. *Biochemistry (Moscow)*, *77*(4), 411–417. https://doi.org/10.1134/S000629791204013X

Takeuchi, T., Koyanagi, R., Gyoja, F., Kanda, M., Hisata, K., Fujie, M., Goto, H., Yamasaki, S., Nagai, K., Morino, Y., Miyamoto, H., Endo, K., Endo, H., Nagasawa, H., Kinoshita, S., Asakawa, S., Watabe, S., Satoh, N., & Kawashima, T. (2016). Bivalve-specific gene expansion in the pearl oyster genome: Implications of adaptation to a sessile lifestyle. *Zoological Letters*, *2*(1). https://doi.org/10.1186/s40851-016-0039-2

Thai, B. T., Lee, Y. P., Gan, H. M., Austin, C. M., Croft, L. J., Trieu, T. A., & Tan, M. H. (2019). Whole Genome Assembly of the Snout Otter Clam, Lutraria rhynchaena, Using Nanopore and Illumina Data, Benchmarked Against Bivalve Genome Assemblies. *Frontiers in Genetics*, *10*. https://doi.org/10.3389/fgene.2019.01158

Uliano-Silva, M., Dondero, F., Dan Otto, T., Costa, I., Lima, N. C. B., Americo, J. A., Mazzoni, C. J., Prosdocimi, F., & Rebelo, M. de F. (2018). A hybrid-hierarchical genome assembly strategy to sequence the invasive golden mussel, Limnoperna fortunei. *GigaScience*, *7*(2). https://doi.org/10.1093/gigascience/gix128

Vurture, G. W., Sedlazeck, F. J., Nattestad, M., Underwood, C. J., Fang, H., Gurtowski, J., & Schatz, M. C. (2017). GenomeScope: Fast reference-free genome profiling from short reads. *Bioinformatics*, *33*(14), 2202–2204. https://doi.org/10.1093/bioinformatics/btx153

Walker, B. J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., Cuomo, C. A., Zeng, Q., Wortman, J., Young, S. K., & Earl, A. M. (2014). Pilon: An Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly Improvement. *PLoS ONE*, *9*(11), e112963. https://doi.org/10.1371/journal.pone.0112963

Wang, S., Zhang, J., Jiao, W., Li, J., Xun, X., Sun, Y., Guo, X., Huan, P., Dong, B., Zhang, L., Hu, X., Sun, X., Wang, J., Zhao, C., Wang, Y., Wang, D., Huang, X., Wang, R., Lv, J., … Bao, Z. (2017). Scallop genome provides insights into evolution of bilaterian karyotype and development. *Nature Ecology & Evolution*, *1*(5). https://doi.org/10.1038/s41559-017-0120

Wang, X., Xu, W., Wei, L., Zhu, C., He, C., Song, H., Cai, Z., Yu, W., Jiang, Q., Li, L., Wang, K., & Feng, C. (2019). Nanopore Sequencing and De Novo Assembly of a Black-Shelled Pacific Oyster (Crassostrea gigas) Genome. *Frontiers in Genetics*, *10*. https://doi.org/10.3389/fgene.2019.01211

Winnebeck, E. C., Millar, C. D., & Warman, G. R. (2010). Why Does Insect RNA Look Degraded? *Journal of Insect Science*, *10*(159), 1–7. https://doi.org/10.1673/031.010.14119

Zimin, A. V., Marçais, G., Puiu, D., Roberts, M., Salzberg, S. L., & Yorke, J. A. (2013). The MaSuRCA genome assembler. *Bioinformatics*, *29*(21), 2669–2677. https://doi.org/10.1093/bioinformatics/btt476

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. https://doi.org/10.1371/journal.pcbi.1007981

# Appendix A

During the project several presentations on the work were made to Reclamation employees and to outside audiences through international conferences and webinars. A list of these presentations is provided below.

**Passamaneck, Y.J.** (2020) Genetic biocontrol of quagga and zebra mussels. National Invasive Species Council Task Team on Advanced Biotechnology. Online meeting.

**Passamaneck, Y.J.** Kocot, K. (2019) Sequencing and assembly of the quagga mussel (*Dreissena rostriformis bugensis*) genome: a tool for development of biocontrols. International Conference on Aquatic Invasive Species. Montreal, Canada.

**Passamaneck, Y.J.** (2019) Genomic analysis of the quagga mussel, *Dreissena rostriformis bugensis*: searching for vulnerabilities. Pearls of Wisdom: Synergising Leadership and Expertise in Molluscan Genomics - Royal Society Theo Murphy International Scientific Meeting, Buckinghamshire, UK (invited speaker)

**Passamaneck, Y.J.** (2019) Invasive mussel genomics: sequencing the dreissenid genome. Invasive Mussel Collaborative. Webinar.

**Passamaneck, Y.J.** (2017) Gene sequencing and editing. USBR Invasive Mussel Task Force. Denver, CO.