







The genome of the zebra mussel, *Dreissena polymorpha*: a resource for comparative genomics, invasion genetics, and biocontrol

Michael A. McCartney,^{1,*} Benjamin Auch ,² Thomas Kono ,³ Sophie Mallez,¹ Ying Zhang,³ Angelico Obille ,⁴ Aaron Becker,² Juan E. Abrahante,⁵ John Garbe,² Jonathan P. Badalamenti ,² Adam Herman ,³ Hayley Mangelson,⁶ Ivan Liachko,⁶ Shawn Sullivan,⁶ Eli D. Sone,^{4,7,8} Sergey Koren,⁹ Kevin A. T. Silverstein,³ Kenneth B. Beckman,² and Daryl M. Gohl ^{2,10,*}

¹Department of Fisheries, Wildlife and Conservation Biology, Minnesota Aquatic Invasive Species Research Center, University of Minnesota, St. Paul, MN 55108, USA,

²University of Minnesota Genomics Center, Minneapolis, MN 55455, USA,

³Minnesota Supercomputing Institute, University of Minnesota, Minneapolis, MN 55455, USA,

⁴Institute of Biomaterials & Biomedical Engineering, University of Toronto, Toronto, ON M5S 3G9, Canada,

⁵University of Minnesota Informatics Institute, Minneapolis, MN 55455, USA,

⁶Phase Genomics, Seattle, WA 98109, USA,

⁷Department of Materials Science & Engineering, University of Toronto, Toronto, ON M5S 3E4 Canada,

⁸Faculty of Dentistry, University of Toronto, Toronto, ON M5G 1G6, Canada,

⁹Genome Informatics Section, Computational and Statistical Genomics Branch, National Human Genome Research Institute, Bethesda, MD 20892, USA, and

¹⁰Department of Genetics, Cell Biology, and Developmental Biology, University of Minnesota, Minneapolis, MN 55455, USA

*Corresponding author: Email: mcartneymichael324@gmail.com (M.A.M.); dmgoehl@umn.edu (D.M.G.)

Abstract

The zebra mussel, *Dreissena polymorpha*, continues to spread from its native range in Eurasia to Europe and North America, causing billions of dollars in damage and dramatically altering invaded aquatic ecosystems. Despite these impacts, there are few genomic resources for *Dreissena* or related bivalves. Although the *D. polymorpha* genome is highly repetitive, we have used a combination of long-read sequencing and Hi-C-based scaffolding to generate a high-quality chromosome-scale genome assembly. Through comparative analysis and transcriptomics experiments, we have gained insights into processes that likely control the invasive success of zebra mussels, including shell formation, synthesis of byssal threads, and thermal tolerance. We identified multiple intact steamer-like elements, a retrotransposon that has been linked to transmissible cancer in marine clams. We also found that *D. polymorpha* have an unusual 67 kb mitochondrial genome containing numerous tandem repeats, making it the largest observed in Eumetazoa. Together these findings create a rich resource for invasive species research and control efforts.

Keywords: *Dreissena polymorpha*; zebra mussel; genome; RNA-Seq; thermal tolerance; stress response; shell formation

Introduction

Native to a small region of southern Russia and Ukraine (Stepien et al. 2014), zebra mussels (*Dreissena polymorpha*, Figure 1A) have spread throughout European (Karatajev et al. 1997, 2003) and North American (Benson 2014) fresh waters to become one of the world's most prevalent and damaging aquatic invasive species (Karatajev et al. 2007). Fouling of water intake pipes cost the power generation industry over \$3 billion USD from 1993 to 1999 in the Laurentian Great Lakes region alone (O'Neill Jr. 2008), where *Dreissena* cause extensive damage to hydropower, recreation and tourism industries, and lakefront property (Bossenbroek et al. 2009; Limburg et al. 2010). Dense infestations smother and outcompete native benthic species and remove large amounts of phytoplankton from lakes and rivers, causing population declines and extinctions of native freshwater mussels and other

invertebrates, damage to fish populations (Raikow 2004; Strayer et al. 2004; McNickle et al. 2006; Karatajev et al. 2014; Lucy et al. 2014; Ward and Ricciardi 2014), and dramatic restructuring of aquatic food webs (Higgins and Vander Zanden 2010; Bootsma and Liao 2014; Mayer et al. 2014). The congener *Dreissena rostriformis* (the quagga mussel), while far less widespread than zebra mussels in inland waters, has ecologically replaced zebra mussels in much of the Laurentian Great Lakes proper and in deep European lakes, and may cause even greater ecological damage in those systems (Karatajev et al. 2011b; Matthews et al. 2014; Nalepa and Schloesser 2014).

The ongoing European and North American invasions (Figure 1, C–E) have spurred an explosion in research effort on *Dreissena*, particularly focused on physiology, autecology, and

Received: September 15, 2021. Accepted: December 02, 2021

© The Author(s) 2021. Published by Oxford University Press on behalf of Genetics Society of America.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

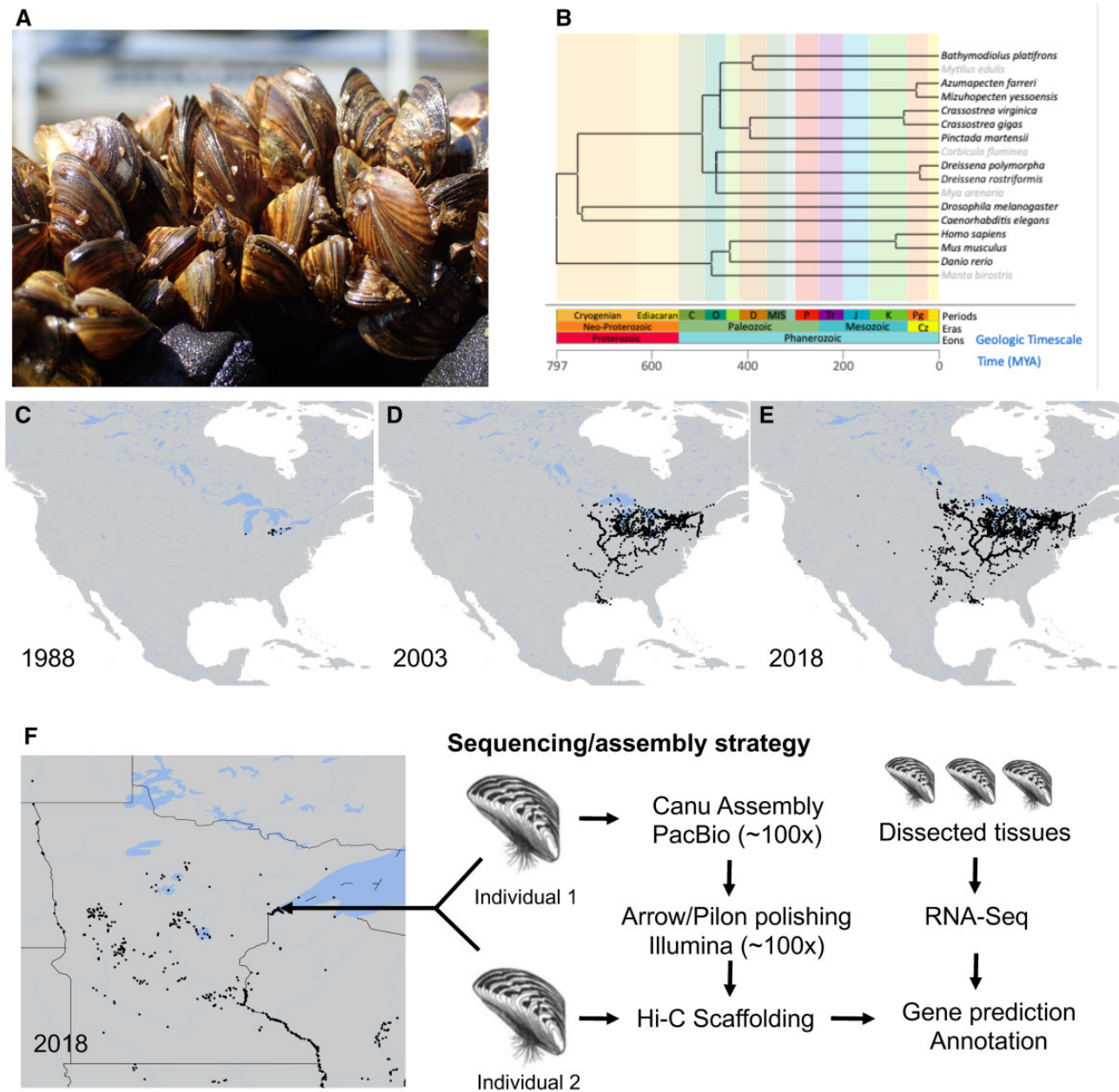


Figure 1 Zebra mussel biogeography and genome sequencing strategy. (A) Photo of *D. polymorpha* (by Naomi Blinick). (B) Phylogenetic tree showing the evolutionary divergence between *D. polymorpha* and other sequenced bivalve genomes. For context, the evolutionary divergence of humans, mice, zebrafish, manta rays, nematodes, and fruit flies are shown. Grey text indicates that a genome sequence for that organism is not publicly available. Divergence times and tree construction based on Kumar et al. (2017). (C–E) Maps depicting the spread of *D. polymorpha* in the United States of America from 1988 through 2018. Data from US Geological Survey, Non-indigenous Aquatic Species database (USGS 2019a). (F) Map showing the extent of zebra mussel infestation in Minnesota lakes as of 2018 and depicting the location where the specimens for genome sequencing and scaffolding were collected (left). Summary of the sequencing and annotation strategy (right).

ecosystem impacts (Schloesser and Schmuckal 2012). Aside from molecular systematic and population genetic studies (Gelembiuk et al. 2006; May et al. 2006; Brown and Stepien 2010; Stepien et al. 2014; Mallez and McCartney 2018), comparatively little genetic work has been accomplished, with transcriptomes from a few tissues (Xu and Faisal 2010; Soroka et al. 2018) being the only genomic resources available for zebra mussels.

Bivalves are a diverse Class of Mollusca with over 10,000 described species in marine and freshwater environments (Bogan 2008; Appeltans et al. 2012). To date, complete genomes have been sequenced and analyzed in 30 species—many of them marine organisms of commercial value (Figure 1B, Supplementary

Table S1) (Zhang et al. 2012; Gómez-Chiarri et al. 2015; Du et al. 2017; Li et al. 2017; Sun et al. 2017; Wang et al. 2017; Powell et al. 2018; Renaut et al. 2018; Uliano-Silva et al. 2018; Calcino et al. 2019; Ran et al. 2019; Yan et al. 2019; Gerdol et al. 2020; Kenny et al. 2020; Li et al. 2020; Liu et al. 2020; Wei et al. 2020; Bao et al. 2021; Gomes-Dos-Santos et al. 2021; Inoue et al. 2021; Ip et al. 2021; Rogers et al. 2021; Smith 2021; Song et al. 2021; Yang et al. 2021). Yet 21 invasive bivalve species cause damage to aquatic and marine ecosystems worldwide (Sousa et al. 2009) and of these only the golden mussel, *Limnoperna fortunei* (Uliano-Silva et al. 2018) and recently, the quagga mussel (Calcino et al. 2019) *D. rostriformis* have so far been sequenced. Moreover, the divergence time

between *Dreissena* and other bivalve species with published genomes is estimated at more than 400 million years ago (Figure 1B). Sequencing of the zebra mussel genome will provide a resource for comparative genomic and other studies of an underexplored lineage of bivalves that includes two of the world's most notorious and damaging invasive species (Lowe et al. 2000; Nalepa and Schloesser 2014).

Here, we present the genome sequence of *D. polymorpha*. Using short and long-read sequencing technologies as well as Hi-C-based scaffolding, we generated a chromosome-scale genome assembly with high contiguity and completeness. Through comparative analysis and RNA-sequencing (RNA-seq) experiments, we provide insights into the process of shell formation, the formation of byssal thread attachment fibers, and mechanisms of thermal tolerance—three processes of critical importance to continued spread. The genomic resources we describe lay the groundwork for further investigation of the traits that allow zebra mussels to thrive as an invasive species and are a step toward developing control strategies for this economically and ecologically damaging aquatic invader.

Methods

Genomic DNA extraction and PacBio library creation

Zebra mussel individuals were collected by SCUBA from off the Duluth waterfront beach (46.78671°N, -92.09114°W), in Lake Superior in June 2017. Mature adults were dissected. To sex the animals, gonad squashes were prepared and examined under a compound microscope for gametes, and a set of large males (25–30 mm shell length) were selected for genome sequencing and analysis. Genomic DNA was extracted using the Qiagen Genomic Tip 100/G kit, with all tissues (except gut) from each selected individual split across six total extractions to prevent clogging of Genomic Tips. Pooled extractions from one chosen individual yielded >100 ug genomic DNA as assessed by PicoGreen DNA quantification (ThermoFisher). The Agilent TapeStation Genomic DNA assay indicated that the majority of gDNA extracted was well over 20 kb (not shown). Further analysis by Pulsed-Field Gel Electrophoresis indicated a broad distribution from 20 to 120 kb, with a modal size of 40 kb (not shown).

Thirty micrograms of gDNA was sheared by passing a solution of 50 ng/uL DNA through a 26G blunt-tipped needle for a total of 20 passes. This sheared DNA was cleaned and concentrated using AMPurePB beads with a 1 × bead ratio, and further library preparation was performed following the PacBio protocol for >30kb libraries using the SMRTbell® Template Prep Kit 1.0. Size-selection of the final library was carried out using the >20 kb high-pass protocol on the PippinHT (Sage Science), and an additional PacBio DNA Damage Repair treatment was performed following size-selection.

PacBio sequencing

Sequencing was carried-out on a PacBio Sequel between November 2017 and February 2018 using 1M v2 Single Molecule Real-Time (SMRT) Cells with 2.1 chemistry and diffusion loading.

Nanopore library creation and sequencing

Genomic DNA from the individual used for PacBio sequence was prepared for Nanopore sequencing using the Oxford Nanopore Ligation Sequencing Kit (SQK-LSK109). The resulting library was sequenced on a single Oxford Nanopore R9.4.1 flowcell on a GridION X5. Reads were collected in MinKNOW for GridION

release 18.07.9 (minknow-core-gridion v. 1.15.4) and basecalled live with guppy v. 1.8.5-1.

Illumina polishing library creation and sequencing

High-molecular-weight DNA from the individual used for the PacBio sequencing was also used as input for Illumina TruSeq DNA PCR-Free library creation, targeting a 350 bp insert size. The resulting library was sequenced on a single lane of HiSeq 2500 High Output (SBS V4) in a 2 × 125 cycle configuration, yielding 68 gigabases (Gb) of data representing ~37 × coverage of the genome.

Hi-C library creation and sequencing

A previously frozen male individual from the same collection date and site in Lake Superior was thawed and mantle, gonad, and gill tissues were dissected using a razor blade. This was a different mussel, because insufficient tissue remained after earlier DNA extractions of the other mussel for genome assembly and polishing. Hi-C library creation was carried out with a Proximo™ Hi-C kit (February 2018) from Phase Genomics using the Proximo™ Hi-C Animal Protocol version 1.0. This method is largely similar to previously published protocols (Lieberman-Aiden et al. 2009). The resulting library was sequenced on a single lane of HiSeq 2500 High Output (SBS V4) in a 2 × 125 cycle configuration, yielding 234M clusters passing filter.

Sample collection for transcriptome studies

Mantle

Adult zebra mussels (20–25 mm shell length) were collected from a high-Ca²⁺ (35–38 mg/L) site: the Lake Ore-Be-Gone mine pit in Gilbert, MN (47.4836°N, -92.4605°W) and from a “low-Ca²⁺” (14.4 mg/L) site: Lake Superior near the Duluth Lift Bridge (46.7867°N, 92.0911°W). Mussels and water were collected underwater by SCUBA, and mussels were stored on ice and returned to the laboratory for dissection within 6 h. This approach was used in lieu of experimental manipulations, because chronic exposure to low calcium concentrations are difficult to achieve in the laboratory—slow shell growth and poor survival have been observed in these marginal (< 15 mg/L) concentrations (Baldwin et al. 2012). Calcium concentration in unfiltered, undigested lake water was determined by 15-element ICP-OES on the iCAP 7600 (Thermo-Fisher, Waltham, MA).

Gill and foot

For these transcriptomes, experiments were used to study differential gene expression in adult mussels that were housed in aquaria for several weeks where they were acclimated, fed laboratory diets, then exposed to experimental treatments. Zebra mussels (15–22 mm shell length) were collected from sites in Lake Minnetonka (44.9533° N, -93.4870° W and 44.8980° N, -93.6688° W) and Lake Waconia (44.8711° N, -93.7596° W) then transported in coolers to the University of Minnesota, where they were acclimated, 100 mussels per each of 12 × 40L glass aquaria with flowing well water (4L/min) at 20°C (unheated). Temperature was checked twice daily with digital probes. Mussels were fed 1.8 ml per tank of liquid shellfish diet (Reed Mariculture, Campbell, CA) once daily, with water flow shut off for 1.5 h for feeding. Tank temperatures were raised to 24–25°C over 3 days by mixing in heated well water; then temperatures were held constant over 7 days for acclimation.

Experimental treatments followed, with each group of four tanks raised 1°C per day (using a 200 W aquarium heater in each tank) to target temperatures of 25, 27, and 30°C then maintained

at target for 7 days. For gill transcriptomes, two mussels per each of four treatment tanks were removed, then both ctenidia were dissected and preserved in 750 μ L RNAlater per animal at -20°C . For foot, mussels from Lake Waconia, attached firmly to rocks and maintained for 7 days in each of two of the 25°C tanks above were selected. Byssal threads were severed where they enter the shell valves to induce byssus growth and reattachment. Immediately thereafter, foot tissue (distal tip region) was dissected from each of eight animals (for a time-zero control) and preserved in RNAlater. Byssus-cut animals were painted with nail polish and placed onto rocks in each of two tanks at 25°C . Mussels that firmly attached overnight were observed for 4 days and 8 days after reattachment, and four firmly attached mussels per time point were selected and foot tissue was dissected and preserved as above. Metadata for transcriptome samples is in [Supplementary File S16](#).

RNA-Seq sample preparation, library creation, and sequencing

Zebra mussel tissue RNA was extracted using the Qiagen RNeasy Plus Universal kit from tissues stored at -20°C in RNAlaterTM (Ambion, Carlsbad, CA). RNA concentration was assessed using Nanodrop, and quantified fluorometrically with the RiboGreen RNA assay kit (ThermoFisher). Further evaluation was based on RNA Integrity Number (RIN) scores generated by the Agilent TapeStation 2200 Eukaryotic RNA assay. Samples with RIN >9.0 and RNA mass >500 ng were used as input for library preparation. Libraries were prepared using the TruSeq[®] Stranded mRNA kit (Illumina) and sequenced on a HiSeq 2500 High Output (SBS V4) run in a 2×50 cycle configuration, generating approximately 15 M reads per sample (Mean = 15.8 M, 15% CV).

Genome assembly

The primary assembly was generated using Canu 1.7 (Koren et al. 2017) from 167.8 Gbp of PacBio subreads over 1 kbp in length with the command:

```
canu -p asm -d asm 'genomeSize=2g' 'correctedErrorRate=0.105'
'corMinCoverage=4' 'corOutCoverage=100' 'batOptions=-dg
3 -db 3 -dr 1 -ca 500 -cp 50' 'corMhapSensitivity=normal'.
```

The assembly used heterozygous parameters due to the relatively high heterozygosity of the sample [2.13% estimated from Genoscope (Vurture et al. 2017) and previous Illumina sequencing]. Benchmarking Universal Single-Copy Orthologs (BUSCO) analysis was run using BUSCO v3 (Simao et al. 2015) and the metazoa_odb9 gene set with the command:

```
python run_BUSCO.py -c 16 -blast_single_core -f -in asm.
contigs.fasta -o SAMPLE -l -m metazoa_odb9 genome.
```

The assembly had 93.9% core metazoan complete genes with 35.2% single copy complete and 58.7% duplicated complete genes. Purge haplotigs (Roach et al. 2018) was run to remove redundancy in the assembly with the commands:

```
minimap2 -ax map-pb -secondary=no -t 16 asm.contigs.fasta
reads.fasta.gz > reads.sam
samtools view -b -T asm.contigs.fasta -S reads.sam > reads.bam
samtools sort -O bam -o reads.sorted.bam -T tmp reads.bam
samtools index reads.sorted.bam
purge_haplotigs readhist reads.sorted.bam
purge_haplotigs contigcov -i reads.sorted.bam.genecov -l 15 -m
80 -h 120 -j 200
```

```
purge_haplotigs purge -t 32 -g asm.contigs.fasta -c coverage_
stats.csv -b reads.sorted.bam -windowmasker
```

Unassigned contigs were removed from the primary set leaving 1.80 Gbp in 2863 contigs with an N50 of 1,111,027 bp.

Genome polishing

The resulting contigs were re-analyzed using the PacBio standard polishing pipeline—GenomicConsensus v2.3.3 (Seifert and Alexander 2019), which derives a better genomic consensus through long read mapping and variant calling using an improved Hidden Markov Model implemented in the algorithm Arrow. The polished draft assembly was further corrected for Indels using Pilon (Walker et al. 2014) with setting: `-fix indels -threads 32 -verbose -changes -tracks`. A single contig corresponding to the PacBio sequencing control was removed from the final assembly.

Repeat analysis

RepeatModeler (Smit and Hubley 2008-2015) was used to identify repeat families from the primary haploid genome. The resulted unknown repeat families were combined with the default full RepeatMasker (Smit et al. 2019) database. RepeatMasker scanned the primary haploid genome sequences for the combined repeat databases in quick search mode.

Hi-C scaffolding

Chromatin conformation capture data were generated using a Phase Genomics (Seattle, WA) Proximo Hi-C Animal Kit v1.0, which is a commercially available version of the Hi-C protocol (Lieberman-Aiden et al. 2009). Following the kit protocol, intact cells from two samples were crosslinked using a formaldehyde solution, digested using the *Sau3AI* restriction enzyme, and proximity-ligated with biotinylated nucleotides to create chimeric molecules composed of fragments from different regions of the genome that were physically proximal in vivo, but not necessarily proximal in the genome. Continuing with the manufacturer's protocol, molecules were pulled down with streptavidin beads and processed into an Illumina-compatible sequencing library. Sequencing was performed in a single lane of Illumina HiSeq 2500 High Output (SBS V5) in a 2×125 cycle configuration, yielding 230,479,044 clusters passing filter.

Reads were aligned to the draft assembly also following the manufacturer's recommendations (Phase Genomics 2019). Briefly, reads were aligned using BWA-MEM (Li and Durbin 2010) with the `-5SP` and `-t 8` options specified, and all other options default. SAMBLASTER (Faust and Hall 2014) was used to flag PCR duplicates, which were later excluded from analysis. Alignments were then filtered with samtools (Li et al. 2009) using the `-F 2304` filtering flag to remove non-primary and secondary alignments and further filtered with matlock (Sullivan 2018) (default options) to remove alignment errors, low-quality alignments, and other alignment noise due to repetitiveness, heterozygosity, and other ambiguous assembled sequences.

Phase Genomics' Proximo Hi-C genome-scaffolding platform was used to create chromosome-scale scaffolds from the corrected assembly as described (Bickhart et al. 2017). As in the LACHESIS method (Burton et al. 2013), this process computes a contact frequency matrix from the aligned Hi-C read pairs, normalized by the number of *Sau3AI* restriction sites (GATC) on each contig, and constructs scaffolds in such a way as to optimize expected contact frequency and other statistical patterns in Hi-C data. Approximately 140,000 separate Proximo runs were

performed to optimize the number of scaffolds to make them as concordant as possible with the observed Hi-C data. This process resulted in a set of 16 chromosome-scale scaffolds containing 1.76 Gbp of sequence (97.9% of the contig assembly), with a scaffold N50 of 117.5 Mbp and a scaffold N90 of 75.4 Mbp.

Mitochondrial genome assembly, polishing, mapping, and annotation

Mapping of PacBio reads to an initial Canu assembly for the mitochondrial genome indicated a small region of very high coverage (Supplementary Figure S5). An alternate assembly of the mitochondrial genome was substituted which was generated in parallel in FALCON 0.5 (length_cutoff = -1, seed_coverage = 30, genome_size = 2.7G) and which did not collapse this repeat sequence. This assembly was polished for indels via Pilon using Illumina reads as with the nuclear genome, and a single substitution error in the coding region was manually edited (c.14475 C > A, G184W) based on strong support from Illumina reads (data not shown). The mitochondrial genome was annotated based a previously published partial mitochondrial sequence (Soroka et al. 2018) in Geneious using the “Annotate from Database” function with a 98% similarity cutoff. The origin point was set to place the tRNA-Val annotation at base 48, matching the previously published sequence.

PacBio and Nanopore reads were mapped against a reference file containing two concatenated copies of the mitochondrial genome sequence to allow reads to map across the origin. Alignments were generated with minimap2 -ax using settings map-pb and map-ont, respectively. Visualization of the resulting alignments (Figure 2C) was performed using a custom tool, ConcatMap (<https://github.com/darylgohl/ConcatMap>). Illumina reads from the polishing library were mapped (Supplementary Figure S5) to the final, polished mitochondrial genome using BWA-MEM (Li and Durbin 2010).

Hi-C analysis of the mitochondrial contig

Ten contigs ranging in size from 50 kb to 100 kb were selected from each of the pseudo-chromosome scaffolds. The number of Hi-C contacts between each selected contig and each pseudo-chromosome was determined. The same analysis was performed using the mitochondrial contig, then all Hi-C link counts were normalized by dividing the number of contacts between a contig and pseudo-chromosome by the total number of Hi-C contacts associated with the contig. The resulting normalized data were visualized using ggplot2 to develop boxplots that compare the number of links for contigs based on their association with each pseudo-chromosome.

Transcriptome assembly

Reads from all zebra mussel RNA-seq libraries were pooled for transcriptome assembly. A database of ribosomal RNA was downloaded from SILVA (Quast et al. 2013; Yilmaz et al. 2014; Glockner et al. 2017), restricting the entries to Bivalvia. The combined RNA-seq reads were cleaned of putative ribosomal RNA sequences using “BBDuk” from the BBTools suite of scripts (Bushnell 2019), treating the Bivalvia ribosomal RNA as potential contaminants, using a k-mer size of 25 bp and an edit distance of 1. Reads that passed this filter were then assembled with Trinity 2.8.4 (Grabherr et al. 2011) with a “RF” library type, in silico read normalization, and a minimum contig length of 500 bp. Assembled transcripts from Trinity were then searched against the non-redundant nucleotide sequence database hosted by NCBI, current as of October 9, 2018. A maximum of 20 target

sequences were returned for each transcript, restricted by a minimum of 10% identity and a maximum E-value of 1×10^{-5} . Assembled transcripts that matched sequences derived from non-eukaryotes or synthetic constructs were discarded.

Differential expression analysis

RNA-seq reads were checked for quality issues, adapter content, and duplication with FastQC 0.11.7. Cleaning for sequencing adapters, trimming of low-quality bases, and filtering for length were performed with Trimmomatic 3.3 (Bolger et al. 2014). The adapter sequences that were targeted for removal were the standard Illumina sequencing adapters. Quality trimming was performed with a window size of 4 bp and a minimum mean quality score of 15. Reads that were shorter than 18 bp after trimming were discarded.

Reads were aligned to the HiC-scaffolded genome assembly draft with HISAT2 2.1.0 (Kim et al. 2015), with putative intron-exon boundaries inferred with genes with functional annotation from the draft annotation and a bundled Python script. Read pairs in which one read failed quality control were not used in alignment and expression analysis. BAM files from HISAT2 were cleaned of reads with a mapping quality score of less than 60 with samtools 1.7. Cleaned alignments were used to generate expression counts with the featureCounts program in the Subread package v. 1.6.2 (Liao et al. 2013). Both reads in a pair were required to map to a feature and be in the proper orientation for them to be counted. Raw read counts were imported into R 3.5.0 (R Core Team 2018) for analysis with edgeR 3.24.3 (Robinson et al. 2013). Genes that were less than 200 bp were removed from the counts matrix. Tests for differential expression were performed between experimental conditions within tissue. For each tissue, genes with low expression were filtered in the following way: genes in which at least X samples with fewer than 10 were removed, where X is the size of the condition with the fewest replicates. Tests for differential expression used a negative binomial model for dispersion estimation, and genes showing significant levels of differential expression were identified with a quasi-likelihood F test implemented in edgeR (Lund et al. 2012). Genes were identified as differentially expressed if they had a nominal P-value of less than 0.01 in the output from the “glmQLFTest” function.

Tissue specificity calculation

Filtered, normalized counts were used to calculate τ , a measure of tissue specificity (Yanai et al. 2005):

$$\tau = \frac{\sum_{i=1}^N (1 - x_i)}{N - 1},$$

where N is the number of tissues analyzed and x_i are the normalized counts. Normalized and log-transformed counts-per-million (CPM) values for each gene were estimated with edgeR. The mean CPM for samples from each tissue were treated as the expression values for that tissue. τ was then calculated for each gene. Genes with τ of 0.95 or greater were considered to be specific to the tissue with highest expression.

Identification of steamer-like elements and phylogenetic analysis

A sequence amplified from *D. polymorpha* using Steamer-like element (SLE)-targeting degenerate primers (Metzger et al. 2018) was used as the basis for an initial BLAST search of the genome assembly. Dotplots of the sequence surrounding hits were analyzed

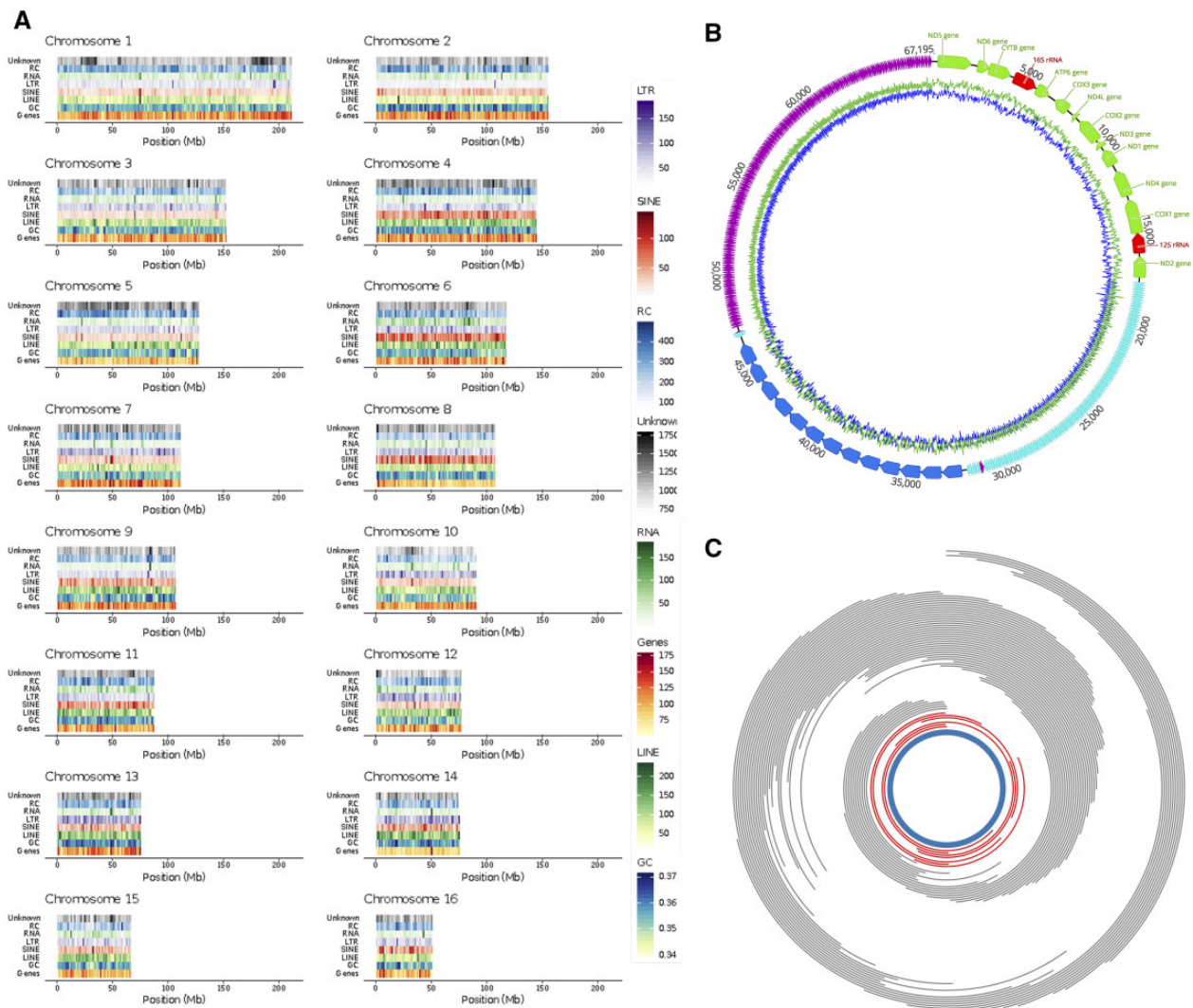


Figure 2 *D. polymorpha* genome and mitogenome structure and content. (A) Plots depicting the gene content, repeat and transposon density, and GC content of the 16 *D. polymorpha* chromosomal scaffolds. (B) Proposed circular mitochondrial genome structure. GC content plots (blue) based on 40 bp sliding window. Annotations based on sequence similarity to previously published partial mitochondrial genome (Soroka et al. 2018). Coding regions are in green and red, and the three large repeat blocks are colored turquoise, blue, and purple. (C) Plot of long (>25 kb) Oxford Nanopore (red) and PacBio (grey) reads supporting the proposed 67 kb circular mitogenome structure. Orientation of mitochondrial genome (blue) is the same as in (B).

to identify 50 putative Long Terminal Repeat (LTR) sequences, and these were aligned to build a consensus LTR sequence specific to our assembly. A subsequent BLAST search with this consensus sequence was performed, and surrounding sequence context was examined for the presence of long (>3 kb) open reading frames (ORFs) between flanking LTRs. Eight intact elements identified with these criteria were aligned based on coding sequence (ClustalW) and annotated based on NCBI Conserved Domain search.

First, we evaluated phylogenetic evidence that zebra mussel TEs are SLEs. Amino acid sequences for the full-length *Gag-Pol* polyprotein region from these eight elements and from the *Steamer* element from *Mya arenaria* (Accession AIE48224.1) were aligned to a database of the *Gypsy/T3y* family of LTR-retrotransposons (Llorens et al. 2011), using MAFFT (Katoh et al. 2017) and the E-INS-i method. The alignment included 2078 residues and 105 sequences. The model of sequence evolution was selected based on the AIC option in SMS (Lefort et al. 2017), using the option to estimate amino acid frequencies from the data. A maximum likelihood genealogy was built using PhyML (Guindon

and Gascuel 2003), using the NNI tree topology search and the BIONJ starting tree options, and support for nodes was evaluated based on 100 bootstrap replications.

Next, we used DNA sequence genealogies to further investigate whether horizontal transmission of TE (HTT) events led to insertions of 20 SLEs that we found in the zebra mussel genome that contained two LTRs flanking an intact *Gag-Pol* ORF, including the eight elements above. From GenBank, we downloaded sequences from multiple bivalve species, from the region located between the RNase H and integrase domains of *Gag-Pol* that was amplified using degenerate primers (Metzger et al. 2018). We added three sequences of long ORFs from *Gag-Pol* that were cloned from neoplastic tissue (Metzger et al. 2016), three that were obtained from *Crassostrea gigas* and *Mizuhopecten yessoensis* genome projects, and the full length *Steamer* clone from *M. arenaria*. We used MAFFT and the G-INS-1 progressive method to align nucleotide sequences based on the translated amino acid sequences and trimmed the ends. The alignment of 54 sequences and 1074 nucleotide positions was loaded into PhyML and the maximum likelihood tree was constructed using the above

options (except that in this case, nucleotide frequencies were optimized using maximum likelihood).

Genome annotation

Functional annotation was carried out with Funannotate 1.0.1 (Palmer 2019) in haploid mode using transcript evidence from RNA-seq alignments, de novo Trinity assemblies, and genome-guided Trinity assemblies. First, repeats were identified using RepeatModeler (Smit and Hubley 2008-2015) and soft-masked using RepeatMasker (Smit and Hubley 2019). Second, protein evidence from a UniProtKB/Swiss-Prot-curated database (downloaded on April 26, 2017) was aligned to the genomes using tBLASTn and exonerate (Slater and Birney 2005), and transcript evidence was aligned using GMAP (Wu and Watanabe 2005). Analysis *ab initio* used gene predictors AUGUSTUS v3.2.3 (Stanke and Morgenstern 2005) and GeneMark-ET v4.32 (Besemer and Borodovsky 2005), trained using BRAKER1 (Hoff et al. 2016), and tRNAs were predicted with tRNAscan-SE (Lowe and Chan 2016). Consensus protein coding gene models were predicted using EvidenceModeler (Haas et al. 2008), and finally gene models were discarded if they were more than 90% contained within a repeat masked region and/or identified from a BLASTp search of known transposons against the TransposonPSI (Haas 2010) and Repbase (Bao et al. 2015) repeat databases. Any fatal errors detected by tbl2asn (<https://www.ncbi.nlm.nih.gov/genbank/asndisc/>) were fixed. Functional annotation used the following databases and tools: PFAM (Finn et al. 2014), InterPro (Jones et al. 2014), UniProtKB (Apweiler et al. 2004), Merops (Rawlings et al. 2016), CAZymes (Lombard et al. 2014), and a set of transcription factors based on InterProScan domains (Shelest 2017) to assign functional annotations.

Comparison to eastern oyster (*Crassostrea virginica*) proteins

Zebra mussel genes with functional annotation information were used to identify groups of genes orthologous to eastern oyster (*Crassostrea virginica*). Annotated protein sequences from *C. virginica* were downloaded from the *C. virginica*-3.0 assembly and annotation hosted on NCBI. Zebra mussel protein sequences and *C. virginica* protein sequences were grouped into orthologous groups using OrthoFinder version 2.2.7 (Emms and Kelly 2018), OrthoFinder was run with BLASTP 2.7.1 for similarity searches, MAFFT 7.305 for alignment, MCL 14.137 for clustering, and RAXML 8.2.11 for tree inference.

Results

To sequence the *D. polymorpha* genome, we used the strategy outlined in Figure 1F. We generated a size-selected PacBio library with ≥ 20 kb inserts (Supplementary Figures S1 and S2). Using the PacBio Sequel SMRT sequencing platform, we generated 168.97 Gb of sequencing data for an estimated coverage over 100 \times , assuming a genome size (from densitometry measures of DNA content in stained nuclei) of 1.66 Gb (Gregory 2003). The subread N50 for the PacBio reads was 16,524 bp, validating the high quality of the input DNA and PacBio sequencing library.

Canu (Koren et al. 2017) yielded a 2.92 Gb assembly, with 15,311 contigs and a contig N50 of 549,263 bp. The assembly was 1.3 Gb larger than previously estimated (Gregory 2003) due to the relatively high heterozygosity of the sample (2.13% estimated from GenomeScope and previous Illumina sequencing). Identification of allelic contigs (Roach et al. 2018) removed redundancy and yielded a 1.8 Gb assembly containing 2863 contigs

Table 1 Genome assembly statistics

| | |
|--------------------------|---------------|
| Assembly statistics | |
| Genome size (bp) | 1,798,019,516 |
| GC content | 35.1% |
| Contigs | 2,863 |
| Largest contig (bp) | 9,337,402 |
| Contig N50 (bp) | 1,111,027 |
| Contig L50 (bp) | 444 |
| Scaffolds | 16 |
| Un scaffolded contigs | 179 |
| Largest scaffold (bp) | 211,287,978 |
| Scaffold N50 (bp) | 117,515,128 |
| Scaffold L50 (bp) | 6 |
| BUSCO analysis | |
| Complete (Eukaryotic) | 92.7% |
| Duplicated (Eukaryotic) | 4.6% |
| Complete (Metazoan) | 92.3% |
| Duplicated (Metazoan) | 3.8% |
| Remapping rates | |
| Illumina DNA-Seq | 98.5% |
| Illumina RNA-Seq | 88.3% |
| Predicted genome content | |
| Predicted genes | 68,018 |
| Repetitive content | 47.4% |
| LINES | 4.3% |
| SINES | 0.7% |
| Known transposons | 4.5% |
| Unclassified repeats | 34.4% |

Statistics summarizing the contiguity, completeness, and content of the *D. polymorpha* genome.

with a contig N50 value of 1,111,027 bp (Table 1). Hi-C (Bickhart et al. 2017) analysis of the polished assembly generated 16 scaffolds spanning 97.9% of the assembled genome (179 un scaffolded contigs comprised the remaining assembled material, Table 1, Supplementary Figure S3). Earlier cytogenetic work found 1N=16 chromosomes for *D. polymorpha* (Boroń et al. 2004; Woznicki and Boroń 2012). The scaffold N50 value was >117 Mb and the scaffold L50 value was 6, consistent with a chromosome-scale assembly. The resulting scaffolds and contigs were checked for contamination from bacterial genomic DNA and sequencing adapters, and a single contig was removed because it mapped to the PacBio sequencing control.

BUSCO analysis (Simao et al. 2015) demonstrated that in addition to having high contiguity, the *D. polymorpha* genome assembly is highly complete, with >92% of eukaryotic and metazoan BUSCOs identified and <5% duplication (Table 1). Also consistent with high completeness, 98.5% of the Illumina DNA sequencing reads mapped to the *D. polymorpha* assembly (Table 1, Supplementary Table S2).

Features of the *D. polymorpha* genome

The genome assembly was annotated using de novo as well as protein and transcript-guided methods. This analysis resulted in a list of 68,018 genes. Based on the number of genes typically present in other eukaryotic genomes, we believe this list is an overestimate of the number of bona fide zebra mussel genes. *Ab initio* gene prediction can introduce errors such as splitting genes based on allelic variation, fragmentation within the assembly, or failure to join exons (Denton et al. 2014). The number of genes in the human genome was initially overestimated and this estimate has been refined over time using both experimental and computational methods (Pertea and Salzberg 2010). Gene number estimates from other sequenced bivalves range from 24,045 (Bai et al. 2019) to over 200,000 (Renaut et al. 2018), with an average of

around 41,000 estimated genes (Smith 2021). Functional annotation was carried out by mapping to a number of databases, including PFAM (Finn et al. 2014), InterPro (Jones et al. 2014), UniProtKB (Apweiler et al. 2004), Merops (Rawlings et al. 2016), and CAZymes (Lombard et al. 2014). Due to the large evolutionary divergence between *D. polymorpha* and other sequenced genomes, most of the predicted genes had no annotations assigned. However, 12,772 genes had recognizable orthologs.

Repetitive DNA is abundant in bivalve genomes (Zhang et al. 2012; Li et al. 2017; Sun et al. 2017; Wang et al. 2017), which makes assembly challenging. The *D. polymorpha* genome is also highly repetitive (47.4% repetitive content, Figure 2A, Table 1) and AT-rich (35.1% GC). While a portion of this repetitive content could be assigned to long or short interspersed elements (LINEs or SINEs), or to known transposons. The majority of the repeats, or 34.4% of the genome, could not be classified (Table 1).

The zebra mussel genome contains several notable gene family expansions (Supplementary Figure S4, Files S1 and S2). *D. polymorpha* shows expansions of genes related to cellular stress responses and apoptosis that surpass humans and in several cases Pacific oyster (*C. gigas*; Zhang et al. 2012), including genes that encode the Hsp70s (heat shock chaperones), caspases (apoptosis), and Inhibitor of Apoptosis Proteins. Families of genes encoding the Cu-Zn superoxide dismutases (antioxidant defense) and C1q domain-containing proteins (innate immunity) show expansions that are, respectively, equal to and smaller than *C. gigas*, while cytochrome P450s (xenobiotic detoxification) are contracted relative to humans (Table 2). Given the large number of annotated genes in *D. polymorpha*, it should be noted that gene family sizes may have been overestimated.

Examination of orthology to eastern oyster (*C. virginica*) identified 10,065 orthologous groups (Supplementary Files S3 and S4). A total of 26.3% of zebra mussel genes that were used for orthologous group identification were assigned to a group within *C. virginica*. This is consistent with the low sequence similarity between zebra mussel and *C. virginica*, even at the amino acid level. A majority (5753; 57.16%) of the orthologous groups involved equal numbers of genes from zebra mussel and *C. virginica*. Of orthologous groups of unequal size, there were far more groups with contracted than expanded gene families in zebra mussel, relative to this distantly related bivalve (76.86% contracted and 23.14% expanded).

In the initial assembly, we recovered a single contig containing the *D. polymorpha* mitochondrial genome (Figure 2B). A partial *D. polymorpha* mitogenome sequence was previously published (Appeltans et al. 2012), but contained a gap which short-read sequencing and targeted PCR were unable to resolve. PacBio and Oxford Nanopore sequencing (Figure 2C) reveals that this “gap” is a large highly repetitive segment of nearly 50 kb, making the *D. polymorpha* mitogenome the largest reported so far from

Eumetazoa at 67,195 bp. The repetitive segment consists of three distinct blocks of direct tandem repeats (Supplementary Figure S5), with individual repeat elements of approximately 125 bp, 1030 bp, and 86 bp, each copied many times. The 86 bp repeat element was discovered only after re-mapping of long reads to the initial assembly, which indicated an area of especially high coverage and read-clipping (Supplementary Figure S6). An alternate mitochondrial assembly generated using FALCON revealed this anomaly to be an additional repeat sequence, to which the PacBio and Oxford Nanopore reads mapped seamlessly. Thus, the FALCON mitogenome assembly has been used in datasets associated with this paper (Figure 2, B and C). We further validated that the mitochondrial contig was not associated with chromosomal sequences by examining Hi-C data, where the association between the mitochondrial contig and the *D. polymorpha* chromosomes was much lower than the association between contigs on the same scaffold and was comparable to background levels of crosslinking seen between contigs on different scaffolds (Supplementary Figure S5). Eumetazoan mitogenomes, with few exceptions, generally lack length variation and non-coding DNA content (Soroka et al. 2018). Among these few exceptions are the long enigmatic mitogenomes of scallops (Boore 1999), but unlike scallops, the coding genes of *D. polymorpha* remain contiguous, instead of being interrupted by interspersed repeats. Typical of animals, the coding region in *D. polymorpha* is compact (~17.5 kb), but the order of mitochondrial genes is unique to the species, a finding that is common in bivalves (Boore 1999). The reason for this unusual mitochondrial DNA (mtDNA) structure is unknown, but similar repetitive sequences have been observed in the mtDNA of plants where it has been suggested that such repeats may result from increased double-stranded break repair in response to desiccation-related DNA damage (Wynn and Christensen 2019).

Some mussels exhibit doubly uniparental inheritance (DUI) of mtDNA, or transmission of two gender-associated mitogenomes: an F-type through eggs and M-type through sperm (Breton et al. 2007; Doucet-Beaupré et al. 2010). DUI is present in *Venerupis*; i.e. in Superorder Imparidentia, containing Dreissenidae. We found no evidence for a second divergent mitogenome. We located no other contigs (via tblastx) that contain mitochondrial genes. Furthermore, re-mapping of high-accuracy Illumina reads from the same mussel to the mitochondrial genome revealed no SNPs within the coding region (Supplementary Figure S6), indicative of homoplasmy. The tissues used for DNA extraction included ripe male gonad with abundant motile sperm. With DUI, extracts would be expected to contain both mtDNAs, as the M-type is transmitted exclusively through male germline, while in somatic tissues, the F-type is predominant (Breton et al. 2007, 2010).

Steamer-like elements

We identified a number of LTR retrotransposons that are similar in structure to *Steamer*, a transposable element (TE) that in the soft-shelled clam *M. arenaria* causes a leukemia that is transmissible between conspecifics (Arriagada et al. 2014; Metzger et al. 2015). A high incidence of HTT has spread these SLEs across several bivalves that also contract transmissible cancers, and across phyla to several marine animal species that do not (Metzger et al. 2018). We identified eight copies of putative SLEs in the *D. polymorpha* genome with intact polycistronic ORFs that span the conserved Gag-Pol polyprotein and are flanked by LTRs (Figure 3A). The *D. polymorpha* elements were aligned to the full length ORFs of 99 Ty3/Gypsy LTR-retrotransposons. Phylogenetic analysis confirmed that the TEs in *D. polymorpha* are SLEs (Supplementary

Table 2 *Dreissena polymorpha* gene family expansions

| Gene family | <i>H. sapiens</i> | <i>D. melanogaster</i> | <i>C. gigas</i> | <i>D. polymorpha</i> |
|-------------|-------------------|------------------------|-----------------|----------------------|
| IAP | 8 | 4 | 48 | 167 |
| Hsp70 | 17 | 6 | 88 | 97 |
| Caspase | 7 | 7 | 24 | 28 |
| Cu-Zn SOD | 1 | 2 | 6 | 6 |
| Cyt. P450 | 57 | 85 | 136 | 56 |
| C1qDC | 31 | 0 | 321 | 50 |

Selected gene family expansion data comparing *D. polymorpha* to *Crassostrea gigas*, *Drosophila melanogaster*, and *Homo sapiens*. Data for *C. gigas*, *D. melanogaster*, and *H. sapiens* from Zhang et al. (2012).

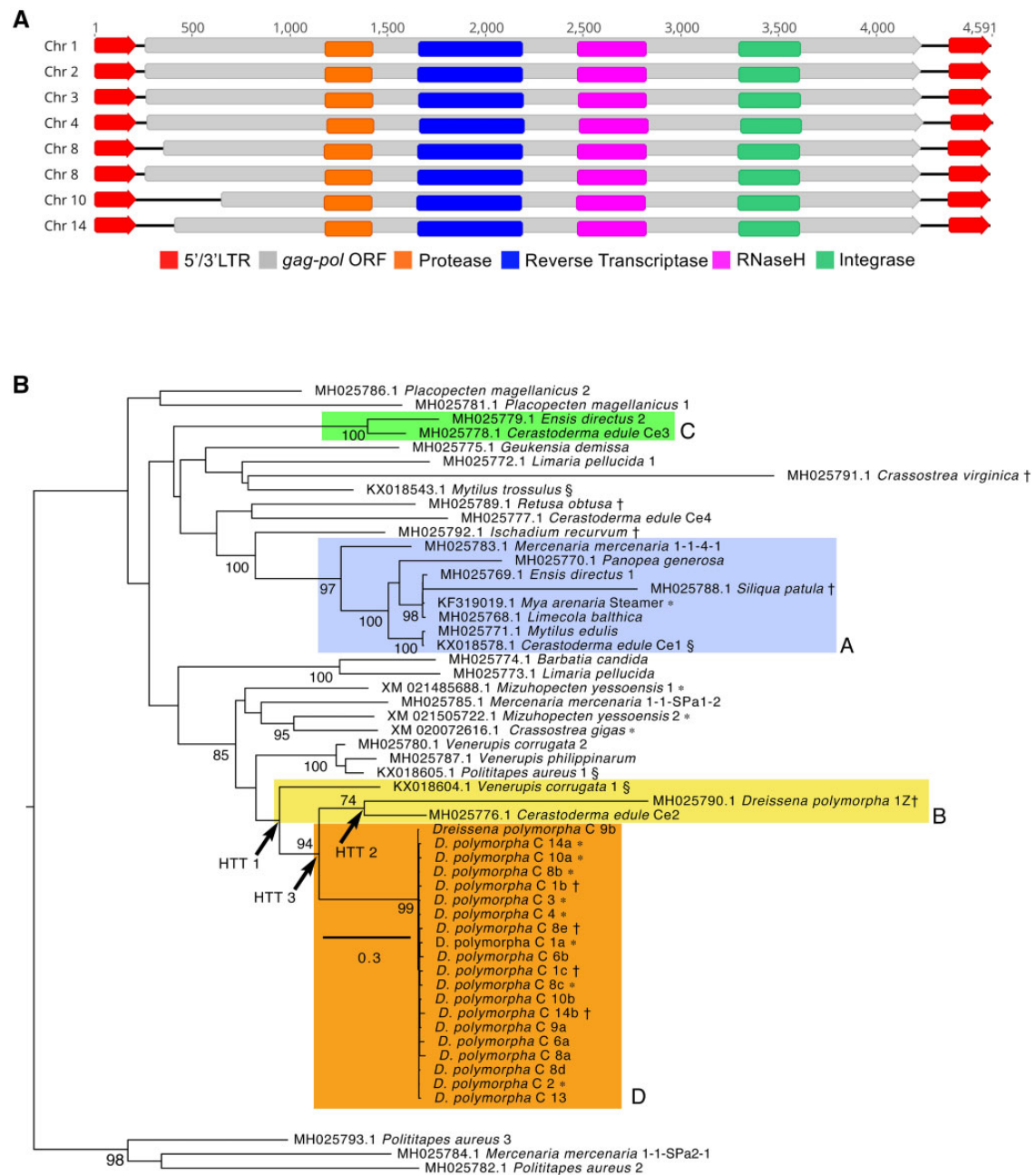


Figure 3 SLEs in the *D. polymorpha* genome. (A) Schematics depicting the eight SLE copies, each with two LTRs flanking the longest ORFs among all similar elements in the *D. polymorpha* genome. (B) Maximum likelihood phylogenetic tree of nucleotide sequences from the RNaseH-integrase domain of Gag-Pol in *D. polymorpha* and other bivalve SLEs. The selected model (Anisimova et al. 2011) of DNA sequence evolution was the GTR + G (rates Γ -distributed, $\alpha = 1.190$) + I (estimated proportion of invariant sites = 0.011). The tree was rooted on the *Polititapes aureus* 2/3/*Mercenaria mercenaria* branch (bottom) and bootstrap support values > 70 are shown. Colored boxes A, B, and C contain taxa involved in all HTT events within bivalves that were identified previously (Metzger et al. 2018). Arrows label HTT events 1 and 2, identified previously (Metzger et al. 2018) and HTT 3, which we identified based on the same criteria. Together these account for two independent insertions of SLEs into zebra mussels. Clade D contains SLE sequences from the zebra mussel genome; “*D. polymorpha* C” = chromosomal location of the SLE, with letters to order multiple insertion sites. Taxon labels include NCBI Accession number, taxon, followed by isolate number or code. * = Sequence is from full length ORF encoding Gag-Pol, † = pseudogene sequence (one or more stop codons), § = sequence derived from neoplastic hemocytes (Metzger et al. 2016).

Figure S7). The *D. polymorpha* elements grouped within the Mag C clade with 100% bootstrap support, and sister to *Steamer*. Next, we performed phylogenetic analysis of the *D. polymorpha* elements and amplicons from within the RNaseH-integrase domain of Gag-Pol from 47 other bivalve species, characterized in an earlier study of HTT events (Metzger et al. 2018). Our phylogenetic analysis identified a minimum of three HTTs leading to their spread to zebra mussels from marine bivalves (Figure 3B),

including an independent event in addition to the two HTTs identified previously (Metzger et al. 2018). It is unknown whether SLEs are currently undergoing active transposition within zebra mussels. However, the high levels of sequence similarity between Gag-pol regions of different SLE loci, and between the two LTRs of each SLE, indicates that the latest wave of transposition in this genome was recent. We also identified numerous degenerate copies that are missing portions of Gag-Pol or LTR sequences, as

well as isolated LTR scars on most chromosomes (Supplementary Figure S8).

Tissue-specific gene expression

We next conducted several RNA-Seq experiments to identify genes that are expressed in a tissue-specific manner, or genes that are regulated in response to different experimental conditions. We examined gene expression in the following tissues (Figure 4A): mantle (the organ that secretes shell), gill (the focal organ for thermal stress response), and foot (the organ that forms and attaches the byssal threads). RNA-Seq data from these three tissues was mapped to the reference containing the 68,018 annotated genes. A tissue-specificity index (τ) (Yanai et al. 2005) was calculated and 577 genes exceeded the threshold of $\tau = 0.95$

(Figure 4B, Supplementary Figure S9 and Files S5–S7). Mantle contained the most tissue-specific genes—359 or 62.2% of the total unique transcripts. Tissue-specific genes had relatively little overlap with genes that were differentially expressed under the experimental conditions tested, suggesting that most tissue-specific genes are carrying out core as opposed to regulated functions (Supplementary Figure S9 and Files S8–S10).

Mantle gene expression and shell formation

In dreissenids and other bivalves, the shell is constructed of calcium carbonate of different crystal forms (typically calcite in adult and aragonite in larval shells) that are deposited in an organic matrix, either through an extracellular or cell-mediated mechanism (Weiner and Traub 1984; Mount et al. 2004). Positive

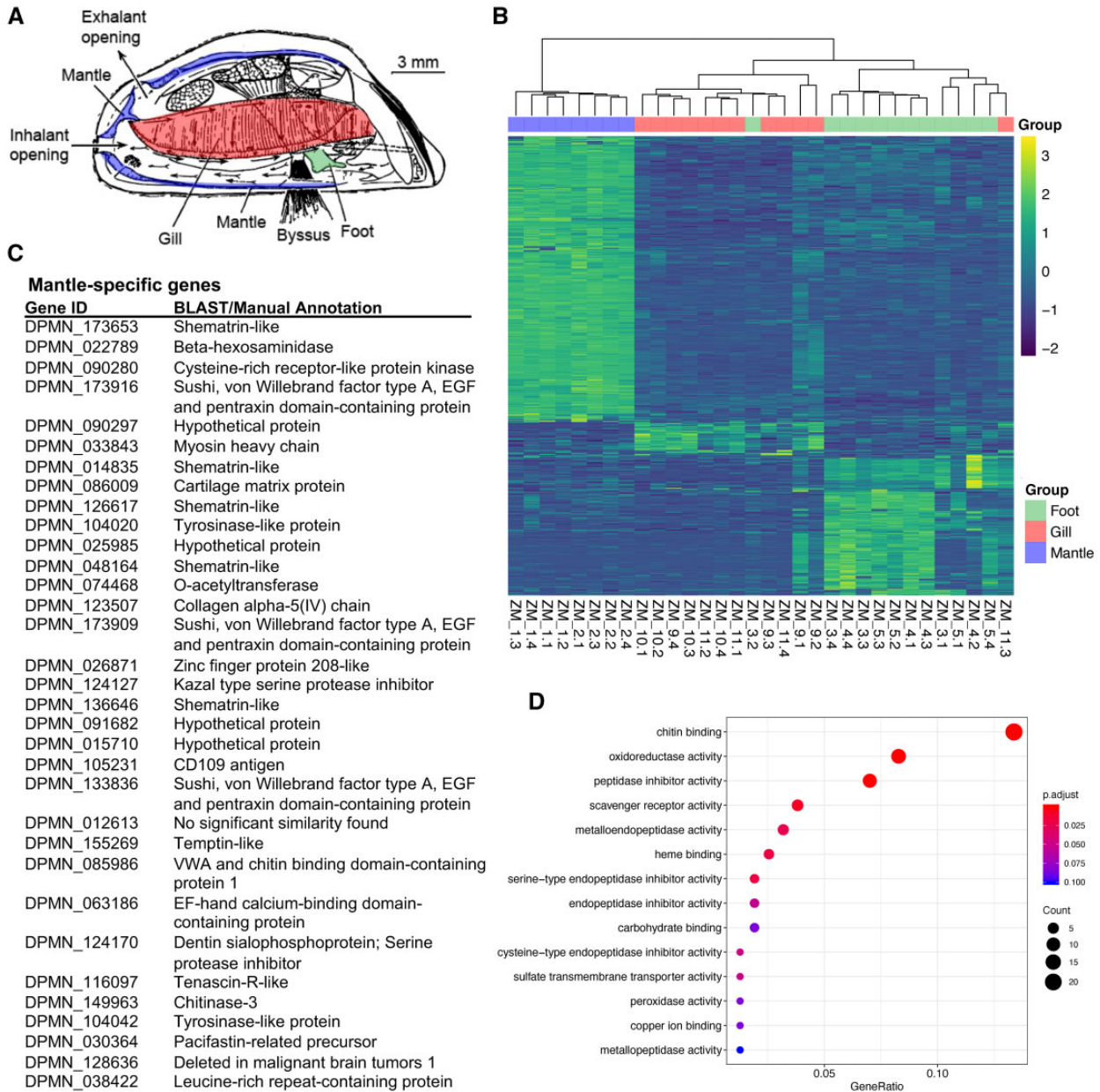


Figure 4 Tissue-specific gene expression patterns: mantle gene expression analysis. (A) *D. polymorpha*: lateral view of the left valve with the right valve and the covering mantle fold removed to reveal the organs dissected for transcriptomes. In purple is the margin of the mantle tissue within the left valve. In *D. polymorpha*, the mantle tissue is fused to form the siphons. Inhalant and exhalant siphon openings are pictured, as is the gill (ctenidium). Modified from Yonge and Campbell (2012). (B) Heatmap depicting Z-scores for tissue-specific gene expression in the foot, gill, and mantle. (C) List of the most highly expressed mantle-specific genes ($\tau > 0.95$). (D) Gene ontology term enrichment analysis for the mantle-specific genes.

correlations between ambient Ca^{2+} and shell strength and calcification have been found in some freshwater mollusk species, and selection favoring shell strength to aid in predator defense has been detected in others (Russell-Hunter et al. 1981; Lewis and Magnuson 1999). To identify biomineralization-related genes, we dissected mantle from adult zebra mussels. We collected mussels from both a calcium-rich (Lake Ore-be-gone: 35.4 mg/L) and a calcium-poor (Lake Superior: 14.4 mg/L) water body.

By inspecting highly expressed mantle-specific genes using the automated annotations as well as BLASTp and comparison with published gene lists (Zhang et al. 2012), we identified orthologs of a set of genes that have been previously implicated in shell formation (Figure 4C, Supplementary File S11). These include tyrosinases, which are required for DOPA production, and other proteins that likely have structural roles, such as collagen. Transcripts for six shematrin-like proteins were among the most specific and highly expressed in mantle. Shematrins are glycine-rich shell matrix proteins that are expressed in the mantle of other mollusks (Yano et al. 2006; Jackson et al. 2010; McDougall et al. 2013; Lin et al. 2014). Glycine-rich peptides in other organisms include structural proteins in rigid plant cell walls (60–70% glycine residues) as well as the major connective tissue in animals, collagen (Shoulders and Raines 2009; Ringli et al. 2001). The exact function of shematrins in shell formation is not clear, but their high expression levels and unusual structure is intriguing; *D. polymorpha* shematrins are characterized by arrays of G(n)Y repeats (Supplementary Figure S10). The zebra mussel shematrin proteins cannot be aligned to shematrins of pearl oyster *Pinctada fucata*, from which they were first characterized. However, the proteins in both genera share features. All are basic, with long runs of compositional bias including glycine-rich tandem repeats (Supplementary Figures S11 and S12, File S12). Functional studies of bivalve shematrin-like proteins are greatly needed.

Also highly expressed in the mantle were transcripts that encode a number of Sushi, von Willebrand factor Type A, EGF, and pentraxin domain-containing proteins that have been implicated in osteogenesis in mammals and have been identified in the mantle of other bivalves. In contrast to shell formation in pearl oysters (Takeuchi et al. 2016), no *nacrein* genes were identified in the zebra mussel genome and a tBLASTn search of the zebra mussel genome with *P. fucata nacrein* yielded no hits. Gene ontology term enrichment analysis also showed that the chitin-binding molecular function was significantly enriched in the mantle-specific genes, along with a number of peptidase inhibitors (Figure 4D).

Among the most specific and highly expressed mantle genes in *D. polymorpha* were two genes with sequence similarity to *temptin*, which encodes a pheromone that serves as a chemoattractant for mating in the sea hare *Aplysia* (Cummins et al. 2004). Zebra mussels attach to one another in clusters known as druses. Settlement of larvae near adults (Wainman et al. 1996) and gregarious post-settlement behaviors (Tošenovský and Kobak 2015) create massive aggregations on lake and river bottom. These behaviors increase settlement success, enable “habitat engineering” in mussel beds (Tošenovský and Kobak 2015), and may enhance feeding and fertilization success (Quinn and Ackerman 2011, 2012; Nishizaki and Ackerman 2017). BLAST searches of the genomes of *D. polymorpha*, other bivalves, and *Aplysia* (Supplementary Figures S13 and S14) found several additional proteins that share the temptin calcium-binding epidermal growth factor-like domain. Further studies are needed to determine if *D. polymorpha* temptin-like proteins serve chemosensory roles, for instance in synchronizing spawning, in sperm

attraction, in settlement of larvae near adults (Wainman et al. 1996), or in gregarious post-settlement behaviors (Tošenovský and Kobak 2015).

Insights into byssal thread formation and attachment

The fibers that zebra and quagga mussels use to anchor themselves to hard surfaces are known as byssal threads. These are key innovations (absent from native North American and European freshwater mollusks) used to attach to conspecific mussels, and to native unionid mussels and other benthic animals that can be smothered and outcompeted. Byssal attachment to boat hulls, docks, boat lifts, and other recreational equipment allows rapid rates of spread between water bodies (Johnson et al. 2001; De Ventura et al. 2016; Collas et al. 2018). Expression of genes during byssogenesis has been studied in zebra mussels (Xu and Faisal 2010) but a majority of mRNAs that are up or down-regulated could not be identified.

Previous work identified a full byssal protein cDNA sequence (named Dpfp1) (Anderson and Waite 1998, 2000) and peptide fragments from a second byssal protein in the foot, the structure that secretes and anchors the threads (Rzepecki and Waite 1993). More recent proteomic work also identified peptide tags associated with several *D. polymorpha* foot proteins that are secreted by the foot and together form the stem, threads, and attachment plaques (Figure 5) of the byssus (Gantayet et al. 2013). Sequences and chromosomal locations of all the genes encoding these byssal proteins are resolved in the zebra mussel genome (Supplementary File 13). The byssalome includes 37 loci on 10 of the 16 zebra mussel chromosomes (Figure 5B). Duplications have generated multiple copies of the byssal genes; some in clusters on single chromosomes, others dispersed onto different chromosomes on both strands (Figure 5B). Duplications are especially abundant in the Dpfp7 and Dpfp9 families, generating substantial amino acid coding variation between the paralogs (not shown). A recent publication provides further detail on the characterization of the zebra mussel byssal thread genes (McCartney 2021).

We also examined transcripts from the foot following experimental induction of byssogenesis (Xu and Faisal 2010) (Supplementary Figure S15). The foot distal to the byssus was dissected immediately after severing the byssal threads, and 4 and 8 days later. Changes were observed at the day-4 time point, after which expression broadly returned to baseline by day 8 (Supplementary Figure S15). Some of the up-regulated genes were consistent with function identified in previous work on byssogenesis in the scallop (Li et al. 2017), including tenascin-X (a connective protein) and a gene with phospholipid scramblase activity (Anoctamin-4-like, Supplementary Figure S15 and File S14). In addition, there was a clear inhibition of the tumor necrosis factor (TNF) pathway, with down-regulation of a TNF-ligand-like protein and up-regulation of Tax1BP1 (a negative regulator of TNF-signaling). The TNF pathway regulates inflammation and apoptosis, suggesting that production of the byssal thread may induce stress in the surrounding tissues and that this stress response may be actively suppressed. Consistent with this, both a cytokine receptor and the pro-apoptotic Bcl2-like gene are down-regulated at the day-4 time point. While earlier expression studies found otherwise (Xu and Faisal 2010; Gantayet et al. 2013, 2014), some byssal proteins were absent from our differentially expressed gene set. And while some of these proteins are differentially distributed across the byssus, localized expression in the foot has not been studied. Nevertheless, one explanation is that

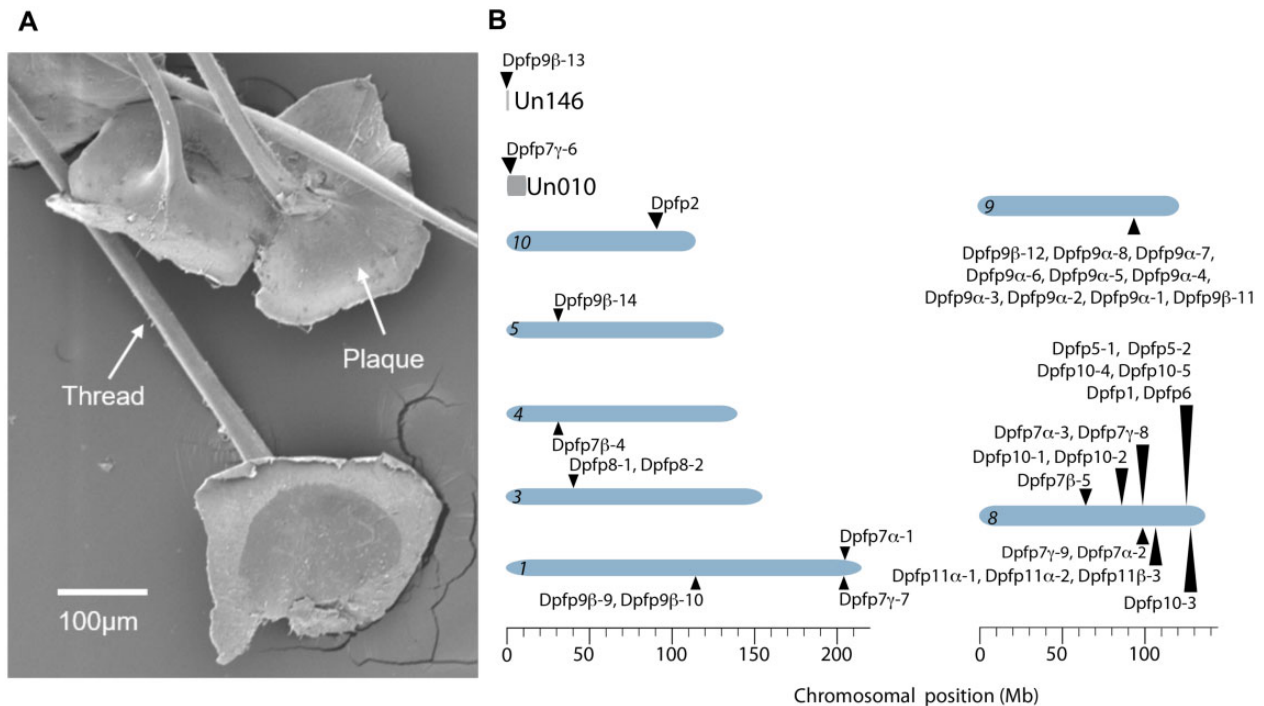


Figure 5 Byssal genes. (A) SEM image of byssus, consisting of threads and plaques. (B) Chromosomal location of the 37 loci predicted to encode 38 byssal protein variants. Chromosomal contigs (blue shaded ovals) are numbered (italics) in order of decreasing size. Byssal genes labeled above the chromosomes are on (+) strands; below are on (-) strands. Byssal protein Dpfp7 has three (α , β , γ) and Dpfp 11 has two (α , β) classes of divergent variants. Chromosome lengths and gene coordinates are in megabases (Mb). To the right of panel B are chromosomes 8 and 9, on which byssal genes are abundant. Modified from McCartney (2021).

our dissections missed the secretory cells more proximal to the threads, a possibility that awaits testing.

Thermal tolerance and chronic heat stress

In *Dreissena*, broad thermal tolerance and ability to adjust to local conditions have clearly played a role in invasion success. Zebra mussels have higher lethal temperature limits and spawn at higher water temperatures in North America than in Europe (McMahon 1996; Nichols 1996). In the Lower Mississippi River, zebra mussels are found south to Louisiana. There they lack cooler water refuges, and persist near their lethal limit of 29–30°C for 3 months during the summer, while for 3 months, temperatures in the river range from 5 to 10°C (Allen et al. 1999). In contrast, zebra mussels in the Upper Mississippi River encounter water temperatures > 25°C for just 1 month of the year, and <2°C for about 3 months (USGS 2019b). Seasonal scheduling of growth and reproductive effort appears to be responsible for at least some of the adaptation or acclimation to conditions in the lower river, as populations in Louisiana shift their shell and tissue growth to the early spring and stop growing in summer (Allen et al. 1999) while more northerly populations grow tissue and spawn in summer months (Borcherding 1991; Claxton and Mackie 1998).

To identify genes involved in the response to thermal stress, we generated transcriptomes from gill tissue in animals exposed to periods of low (24°C), moderate (27°C), and high (30°C) chronic temperature stress (Figure 6A). Moderate thermal stress led to the induction of several genes involved in cellular adhesion or cytoskeletal remodeling, including collagen, gelsolin, MYLIP E3 ubiquitin ligase, and N-cadherin (Figure 6B, Supplementary File S15). High thermal stress led to strong induction of a large number of chaperones, including HSP70, DNAJ, Calnexin, and HSC70

(several of which were also induced to a lesser extent under moderate thermal stress), as well as the antioxidant protein cytochrome P450 (Figure 6, B and C, Supplementary File S15). The list of down-regulated genes was quite similar for both the moderate and high thermal stress conditions (Figure 6, D–E, Supplementary File S15). In addition to the induction of known stress-response genes, a number of genes with unknown function are also regulated by thermal stress, as is 4-Hydroxyphenylpyruvate Dioxygenase, an enzyme which is involved in the catabolism of tyrosine (Figure 6, B–E, Supplementary File S15).

Discussion

Here, we describe the genome of the zebra mussel. Consistent with the genomes of other bivalves, the *D. polymorpha* genome is highly repetitive and encodes an expanded set of heat-shock and anti-apoptotic proteins, presumably to deal with the challenges of a sessile existence. We examine the genetic underpinnings of several traits that have been linked to population growth and invasiveness, including shell and byssal thread formation, and response to thermal stress. While these analyses uncovered multiple genes and pathways that seem to function in a conserved manner across multiple bivalve species, they also uncovered many genes of unknown function. In the future, it will be of considerable value to compare the zebra mussel genome with that of its congener, the quagga mussel (*D. rostriformis*), in order to gain further insights into ecological displacement of zebra mussels by quagga mussels, and to investigate genetic underpinning of their relative invasiveness, such as comparative work on byssogenesis that may help account for the slower geographic spread of quagga mussels (Karatayev et al. 2011a).

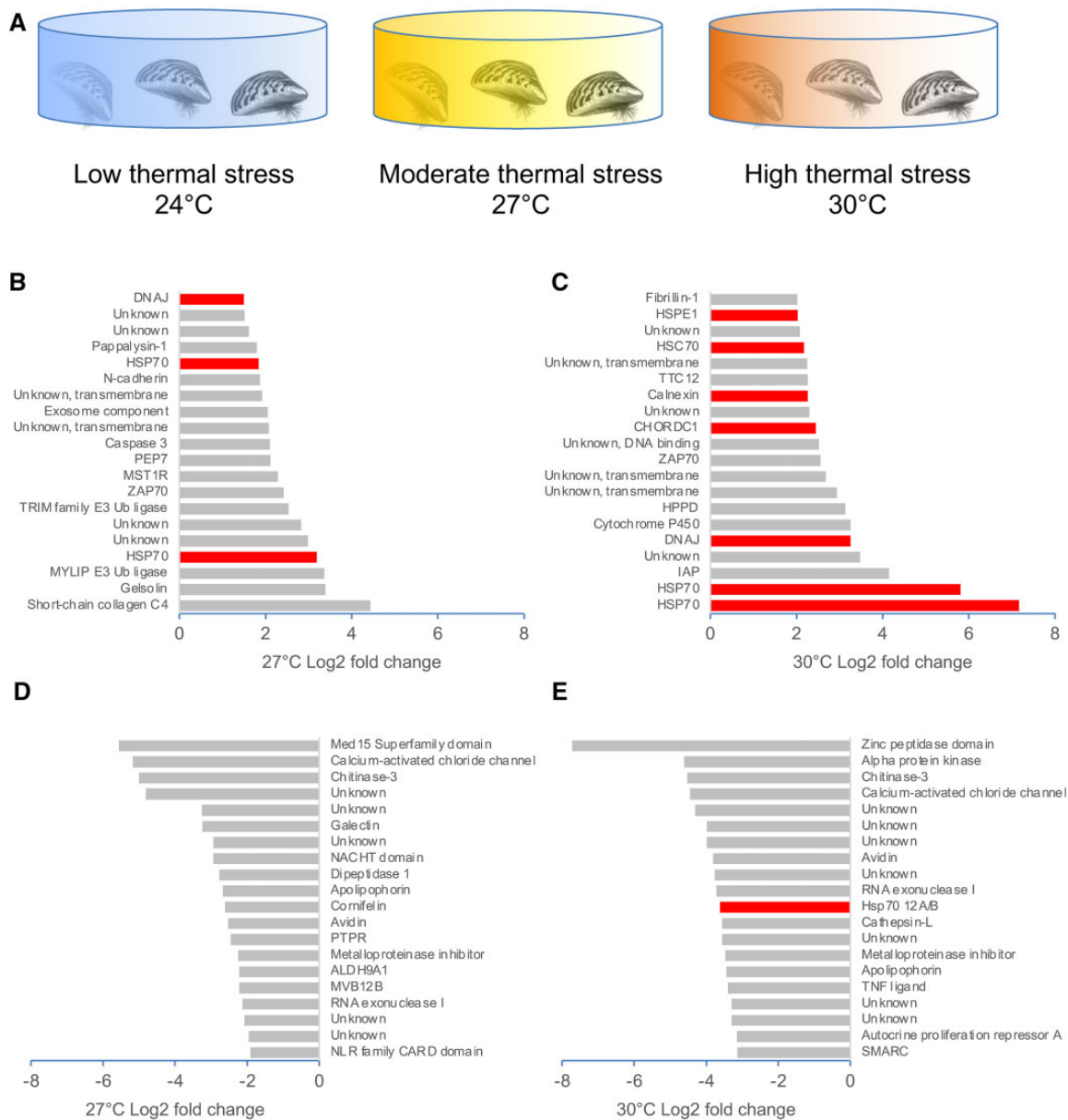


Figure 6 Response of *D. polymorpha* to thermal stress. (A) Overview of experimental set-up. Animals were subjected to low (24°C), moderate (27°C), and high (30°C) thermal stress ($n = 4$ animals per condition). (B) Top 20 genes upregulated during moderate thermal stress by log₂ fold-change. (C) Top 20 genes upregulated during high thermal stress by log₂ fold-change. (D) Top 20 genes downregulated during moderate thermal stress by log₂ fold-change. (E) Top 20 genes downregulated during high thermal stress by log₂ fold-change. Genes highlighted in red encode chaperone proteins.

The existence of genomic resources for *D. polymorpha* and the catalog of genes we have identified will enable multiple new lines of investigation, as well as provide researchers with an improved tool for population genetic experiments, for instance, tracking the spread of mussels using Genotyping-by-Sequencing approaches, or designing new targeted assays for the presence or activity of zebra mussels.

While it is clear that changes in transportation networks (*e.g.* canal building, opening of shipping channels, ballast water discharge) were the events that initiated primary invasions of European and North American waters (Karatayev et al. 2007; Pagnucco et al. 2015), several biological characteristics are responsible for the rate of spread of zebra and quagga mussels across both continents, while other traits have limited their

suitable habitat range. Genomics offers a path to understanding these traits at the genetic level, which may ultimately guide the development of control methods and management strategies.

Data availability

The *D. polymorpha* genome assembly is available at NCBI (BioProject: PRJNA533175). Sequencing data files are available through the NCBI Sequence Read Archive (BioProject: PRJNA533175, PRJNA533176). This Whole Genome Shotgun project has been deposited at DDBJ/ENA/GenBank under the accession JAIWYP000000000. The version described in this paper is version JAIWYP010000000.

Supplementary material is available at G3 online.

Acknowledgments

This work utilized the computational resources of the NIH HPC Biowulf cluster (<https://hpc.nih.gov>) and the Minnesota Supercomputing Institute (<https://www.msi.umn.edu>).

Author Contributions

M.A.M. and D.M.G. conceived and designed experiments, analyzed data, and wrote the paper. B.A. prepared PacBio and Hi-C libraries, analyzed data, and wrote the paper. T.K., Y.Z., J.E.A., and K.A.T.S. analyzed data and helped to assemble and annotate the genome. S.M. designed experiments, collected samples, isolated DNA. J.G. analyzed data. A.O. and E.D.S. analyzed byssal thread attachment proteins. A.B. carried out sequencing of PacBio and Illumina libraries. J.P.B. carried out nanopore sequencing. A.H. and H.M. analyzed data. I.L., H.M., and S.S. carried out Hi-C-based scaffolding. S.K. generated the Canu assembly and ran purge haplotigs. K.B.B. conceived and designed experiments.

Funding

S.K. is supported by the Intramural Research Program of the National Human Genome Research Institute, National Institutes of Health. Funding was from the Minnesota Environment and Natural Resources Trust Fund and the Minnesota Aquatic Invasive Species Research Center.

Conflicts of interest

I.L. and S.S. have a financial interest in and are directors of Phase Genomics, a company commercializing proximity ligation technology. H.M. is an employee of Phase Genomics.

Literature cited

- Allen YC, Thompson BA, Ramcharan CW. 1999. Growth and mortality rates of the zebra mussel, *Dreissena polymorpha*, in the Lower Mississippi River. *Can J Fish Aquat Sci.* 56:748–759.
- Anderson KE, Waite JH. 1998. A major protein precursor of zebra mussel (*Dreissena polymorpha*) byssus: deduced sequence and significance. *Biol Bull.* 194:150–160.
- Anderson KE, Waite JH. 2000. Immunolocalization of Dpfp1, a byssal protein of the zebra mussel *Dreissena polymorpha*. *J Exp Biol.* 203:3065–3076.
- Anisimova M, Gil M, Dufayard JF, Dessimoz C, Gascuel O. 2011. Survey of branch support methods demonstrates accuracy, power, and robustness of fast likelihood-based approximation schemes. *Syst Biol.* 60:685–699.
- Appeltans W, Ahyong ST, Anderson G, Angel MV, Artois T, et al. 2012. The magnitude of global marine species diversity. *Curr Biol.* 22:2189–2202.
- Apweiler R, Bairoch A, Wu CH, Barker WC, Boeckmann B, et al. 2004. UniProt: the Universal Protein knowledgebase. *Nucleic Acids Res.* 32:D115–D119.
- Arriagada G, Metzger MJ, Muttray AF, Sherry J, Reinisch C, et al. 2014. Activation of transcription and retrotransposition of a novel retroelement, *Steamer*, in neoplastic hemocytes of the mollusk *Mya arenaria*. *Proc Natl Acad Sci U S A.* 111:14175–14180.
- Bai CM, Xin LS, Rosani U, Wu B, Wang QC, et al. 2019. Chromosomal-level assembly of the blood clam, *Scapharca (Anadara) broughtonii*, using long sequence reads and Hi-C. *Gigascience.* 8:giz067.
- Baldwin BS, Carpenter M, Rury K, Woodward E. 2012. Low dissolved ions may limit secondary invasion of inland waters by exotic round gobies and dreissenid mussels in North America. *Biol Invasions.* 14:1157–1175.
- Bao W, Kojima KK, Kohany O. 2015. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mob DNA.* 6:11.
- Bao Y, Zeng Q, Wang J, Zhang Z, Zhang Y, et al. 2021. Genomic insights into the origin and evolution of molluscan red-bloodedness in the blood clam *Tegillarca granosa*. *Mol Biol Evol.* 38:2351–2365.
- Benson AJ. 2014. Chronological history of zebra and quagga mussels (*Dreissenidae*) in North America, 1988–2010. In: TF Nalepa and DW Schloesser, editors. *Quagga and Zebra Mussels: Biology, Impacts, and Control*. 2nd ed. Boca Raton, FL: CRC Press. p. 9–32.
- Besemer J, Borodovsky M. 2005. GeneMark: web software for gene finding in prokaryotes, eukaryotes and viruses. *Nucleic Acids Res.* 33:W451–W454.
- Bickhart DM, Rosen BD, Koren S, Sayre BL, Hastie AR, et al. 2017. Single-molecule sequencing and chromatin conformation capture enable de novo reference assembly of the domestic goat genome. *Nat Genet.* 49:643–650.
- Bogan A. 2008. Global diversity of freshwater mussels (Mollusca, Bivalvia) in freshwater. In: EV Balian, C Lévêque, H Segers, K Martens, editors. *Freshwater Animal Diversity Assessment*. Netherlands: Springer. p. 139–147.
- Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics.* 30:2114–2120.
- Boore JL. 1999. Survey and summary: animal mitochondrial genomes. *Nucleic Acids Res.* 27:1767–1780.
- Bootsma HA, Liao Q. 2014. Nutrient cycling by dreissenid mussels: controlling factors and ecosystem response. In: TF Nalepa and DW Schloesser, editors. *Quagga and Zebra Mussels: Biology, Impacts, and Control*. 2nd ed. Boca Raton, FL: CRC Press. p. 555–574.
- Borcherding J. 1991. The annual reproductive cycle of the freshwater mussel *Dreissena polymorpha* Pallas in lakes. *Oecologia.* 87:208–218.
- Boroń A, Woźnicki P, Skuza L, Zieliński R. 2004. Cytogenetic characterization of the zebra mussel *Dreissena polymorpha* (Pallas) from Miedwie Lake, Poland. *Folia Biol (Krakow).* 52:33–38.
- Bossenbroek JM, Finnoff DC, Shogren JF, Warziniack TW. 2009. Advances in ecological and economical analysis of invasive species: dreissenid mussels as a case study. In: RP Keller, DM Lodge, MA Lewis, JF Shogren, editors. *Bioeconomics of Invasive Species: Integrating Ecology, Economics, Policy, and Management*. New York: Oxford University Press. p. 244–265.
- Breton S, Beaupre HD, Stewart DT, Hoeh WR, Blier PU. 2007. The unusual system of doubly uniparental inheritance of mtDNA: isn't one enough? *Trends Genet.* 23:465–474.
- Breton S, Stewart DT, Hoeh WR. 2010. Characterization of a mitochondrial ORF from the gender-associated mtDNAs of *Mytilus* spp. (Bivalvia: Mytilidae): identification of the “missing” ATPase 8 gene. *Mar Genomics.* 3:11–18.
- Brown JE, Stepien CA. 2010. Population genetic history of the dreissenid mussel invasions: expansion patterns across North America. *Biol Invasions.* 12:3687–3710.
- Burton JN, Adey A, Patwardhan RP, Qiu R, Kitzman JO, et al. 2013. Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions. *Nat Biotechnol.* 31:1119–1125.
- Bushnell B. 2019. BBMap short read aligner and other bioinformatic tools. Berkeley, CA: Joint Genome Institute <https://sourceforge.net/projects/bbmap/files/>.
- Calcino AD, de Oliveira AL, Simakov O, Schwaha T, Zieger E, et al. 2019. The quagga mussel genome and the evolution of freshwater tolerance. *DNA Res.* 26:411–422.

- Claixton WT, Mackie GL. 1998. Seasonal and depth variations in gametogenesis and spawning of *Dreissena polymorpha* and *Dreissena bugensis* in eastern Lake Erie. *Can J Zool.* 76:2010–2019.
- Collas FPL, Karatayev AY, Burlakova LE, Leuven RSEW. 2018. Detachment rates of dreissenid mussels after boat hull-mediated overland dispersal. *Hydrobiologia.* 810:77–84.
- Cummins SF, Nichols AE, Amare A, Hummon AB, Sweedler JV, et al. 2004. Characterization of *Aplysia* enticin and temptin, two novel water-borne protein pheromones that act in concert with attractin to stimulate mate attraction. *J Biol Chem.* 279:25614–25622.
- De Ventura L, Weissert N, Tobias R, Kopp K, Jokela J. 2016. Overland transport of recreational boats as a spreading vector of zebra mussel *Dreissena polymorpha*. *Biol Invasions.* 18:1451–1466.
- Denton JF, Lugo-Martinez J, Tucker AE, Schrider DR, Warren WC, et al. 2014. Extensive error in the number of genes inferred from draft genome assemblies. *PLoS Comput Biol.* 10:e1003998.
- Doucet-Beaupré H, Breton S, Chapman EG, Blier PU, Bogan AE, et al. 2010. Mitochondrial phylogenomics of the Bivalvia (Mollusca): searching for the origin and mitogenomic correlates of doubly uniparental inheritance of mtDNA. *BMC Evol Biol.* 10:50.
- Du X, Fan G, Jiao Y, Zhang H, Guo X, et al. 2017. The pearl oyster *Pinctada fucata martensii* genome and multi-omic analyses provide insights into biomineralization. *Gigascience.* 6:1–12.
- Emms DM, Kelly S. 2018. OrthoFinder2: fast and accurate phylogenomic orthology analysis from gene sequences. *bioRxiv* doi: <http://dx.doi.org/10.1101/466201>.
- Faust GG, Hall IM. 2014. SAMBLASTER: fast duplicate marking and structural variant read extraction. *Bioinformatics.* 30:2503–2505.
- Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, et al. 2014. Pfam: the protein families database. *Nucleic Acids Res.* 42:D222–D230.
- Gantayet A, Ohana L, Sone ED. 2013. Byssal proteins of the freshwater zebra mussel, *Dreissena polymorpha*. *Biofouling.* 29:77–85.
- Gantayet A, Rees DJ, Sone ED. 2014. Novel proteins identified in the insoluble byssal matrix of the freshwater zebra mussel. *Mar Biotechnol (NY).* 16:144–155.
- Gelembiuk GW, May GE, Lee CE. 2006. Phylogeography and systematics of zebra mussels and related species. *Mol Ecol.* 15:1033–1050.
- Gerdol M, Moreira R, Cruz F, Gómez-Garrido J, Vlasova A, et al. 2020. Massive gene presence-absence variation shapes an open pan-genome in the Mediterranean mussel. *Genome Biol.* 21:275.
- Glockner FO, Yilmaz P, Quast C, Gerken J, Beccati A, et al. 2017. 25 years of serving the community with ribosomal RNA gene reference databases and tools. *J Biotechnol.* 261:169–176.
- Gomes-Dos-Santos A, Lopes-Lima M, Machado AM, Marcos Ramos A, Usié A, et al. 2021. The Crown Pearl: a draft genome assembly of the European freshwater pearl mussel *Margaritifera margaritifera* (Linnaeus, 1758). *DNA Res.* 28:1–10.
- Gómez-Chiarri M, Warren WC, Guo X, Proestou D. 2015. Developing tools for the study of molluscan immunity: the sequencing of the genome of the eastern oyster, *Crassostrea virginica*. *Fish Shellfish Immunol.* 46:2–4.
- Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, et al. 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol.* 29:644–652.
- Gregory TR. 2003. Genome size estimates for two important freshwater molluscs, the zebra mussel (*Dreissena polymorpha*) and the schistosomiasis vector snail (*Biomphalaria glabrata*). *Genome.* 46:841–844.
- Guindon S, Gascuel O. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol.* 52:696–704.
- Haas B. 2010. TransposonPSI: an application of PSI-blast to mine (retro-)transposon ORF homologies.
- Haas BJ, Salzberg SL, Zhu W, Pertea M, Allen JE, et al. 2008. Automated eukaryotic gene structure annotation using EvidenceModeler and the program to assemble spliced alignments. *Genome Biol.* 9:R7.
- Higgins SN, Vander Zanden MJ. 2010. What a difference a species makes: a meta-analysis of dreissenid mussel impacts on freshwater ecosystems. *Ecol Monogr.* 80:179–196.
- Hoff KJ, Lange S, Lomsadze A, Borodovsky M, Stanke M. 2016. BRAKER1: unsupervised RNA-Seq-based genome annotation with GeneMark-ET and AUGUSTUS. *Bioinformatics.* 32:767–769.
- Inoue K, Yoshioka Y, Tanaka H, Kinjo A, Sassa M, et al. 2021. Genomics and transcriptomics of the green mussel explain the durability of its byssus. *Sci Rep.* 11:5992.
- Ip JCH, Xu T, Sun J, Li R, Chen C, et al. 2021. Host-endosymbiont genome integration in a deep-sea chemosymbiotic clam. *Mol Biol Evol.* 38:502–518.
- Jackson DJ, McDougall C, Woodcroft B, Moase P, Rose RA, et al. 2010. Parallel evolution of nacre building gene sets in molluscs. *Mol Biol Evol.* 27:591–608.
- Johnson LE, Ricciardi A, Carlton JT. 2001. Overland dispersal of aquatic invasive species: a risk assessment of transient recreational boating. *Ecol Appl.* 11:1789–1799.
- Jones P, Binns D, Chang HY, Fraser M, Li W, et al. 2014. InterProScan 5: genome-scale protein function classification. *Bioinformatics.* 30:1236–1240.
- Karatayev AY, Burlakova LE, Mastitsky SE, Padilla DK, Mills EL. 2011a. Contrasting rates of spread of two congeners, *Dreissena polymorpha* and *Dreissena rostriformis bugensis*, at different spatial scales. *J Shellfish Res.* 30:923–931.
- Karatayev AY, Burlakova LE, Padilla DK. 1997. The effects of *Dreissena polymorpha* (Pallas) invasion on aquatic communities in eastern Europe. *J Shellfish Res.* 16:187–203.
- Karatayev AY, Burlakova LE, Padilla DK, Johnson LE. 2003. Patterns of spread of the zebra mussel (*Dreissena polymorpha* (Pallas)): the continuing invasion of Belarussian lakes. *Biol Invasions.* 5:213–221.
- Karatayev AY, Burlakova LE, Pennuto C, Ciborowski J, Karatayev VA, et al. 2014. Twenty five years of changes in *Dreissena* spp. populations in Lake Erie. *J Great Lakes Res.* 40:550–559.
- Karatayev AY, Mastitsky SE, Padilla DK, Burlakova LE, Hajduk M. 2011b. Differences in growth and survivorship of zebra and quagga mussels: size matters. *Hydrobiologia.* 668:183–194.
- Karatayev AY, Padilla DK, Minchin D, Boltovskoy D, Burlakova LE. 2007. Changes in global economies and trade: the potential spread of exotic freshwater bivalves. *Biol Invasions.* 9:161–180.
- Katoh K, Rozewicki J, Yamada KD. 2017. MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization. *Brief Bioinform.* 20:1160–1166.
- Kenny NJ, McCarthy SA, Dudchenko O, James K, Betteridge E, et al. 2020. The gene-rich genome of the scallop *Pecten maximus*. *Gigascience.* 9:1–13.
- Kim D, Langmead B, Salzberg SL. 2015. HISAT: a fast spliced aligner with low memory requirements. *Nat Methods.* 12:357–360.
- Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, et al. 2017. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res.* 27:722–736.
- Kumar S, Stecher G, Suleski M, Hedges SB. 2017. TimeTree: a resource for timelines, timetrees, and divergence times. *Mol Biol Evol.* 34:1812–1819.
- Lefort V, Longueville JE, Gascuel O. 2017. SMS: smart model selection in PhyML. *Mol Biol Evol.* 34:2422–2424.

- Lewis DB, Magnuson JJ. 1999. Intraspecific gastropod shell strength variation among north temperate lakes. *Can J Fish Aquat Sci.* 56:1687–1695.
- Li H, Durbin R. 2010. Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics.* 26:589–595.
- Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, et al.; 1000 Genome Project Data Processing Subgroup. 2009. The sequence alignment/map format and SAMtools. *Bioinformatics.* 25:2078–2079.
- Li Y, Nong W, Baril T, Yip HY, Swale T, et al. 2020. Reconstruction of ancient homeobox gene linkages inferred from a new high-quality assembly of the Hong Kong oyster (*Magallana hongkongensis*) genome. *BMC Genomics.* 21:17.
- Li Y, Sun X, Hu X, Xun X, Zhang J, et al. 2017. Scallop genome reveals molecular adaptations to semi-sessile life and neurotoxins. *Nat Commun.* 8:1721.
- Liao Y, Smyth GK, Shi W. 2013. The Subread aligner: fast, accurate and scalable read mapping by seed-and-vote. *Nucleic Acids Res.* 41:e108.
- Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragozcy T, et al. 2009. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science.* 326:289–293.
- Limburg KE, Luzadis VA, Ramsey M, Schulz KL, Mayer CM. 2010. The good, the bad, and the algae: perceiving ecosystem services and disservices generated by zebra and quagga mussels. *J Great Lakes Res.* 36:86–92.
- Lin Y, Jia G, Xu G, Su J, Xie L, et al. 2014. Cloning and characterization of the shell matrix protein Shematrin in scallop *Chlamys farreri*. *Acta Biochim Biophys Sin (Shanghai).* 46:709–719.
- Liu X, Li C, Chen M, Liu B, Yan X, et al. 2020. Draft genomes of two Atlantic bay scallop subspecies *Argopecten irradians irradians* and *A. i. concentricus*. *Sci Data.* 7:99.
- Llorens C, Futami R, Covelli L, Dominguez-Escriba L, Viu JM, et al. 2011. The Gypsy Database (GyDB) of mobile genetic elements: release 2.0. *Nucleic Acids Res.* 39:D70–D74.
- Lombard V, Golaconda Ramulu H, Drula E, Coutinho PM, Henrissat B. 2014. The carbohydrate-active enzymes database (CAZy) in 2013. *Nucleic Acids Res.* 42:D490–D495.
- Lowe S, Browne M, Boudjelas S, De Poorter M. 2000. 100 of the World's Worst Invasive Alien Species: A Selection from the Global Invasive Species Database. New Zealand: The Invasive Species Specialist Group (ISSG) a specialist group of the Species Survival Commission (SSC) of the World Conservation Union (IUCN).
- Lowe TM, Chan PP. 2016. tRNAscan-SE On-line: integrating search and context for analysis of transfer RNA genes. *Nucleic Acids Res.* 44:W54–W57.
- Lucy FE, Burlakova LE, Karatayev AY, Mastitsky SE, Zanatta DT. 2014. Zebra mussel impacts on unionids: a synthesis of trends in North America and Europe. In: TF Nalepa and DW Schloesser, editors. *Quagga and Zebra Mussels: Biology, Impact, and Control*. Boca Raton, FL: CRC Press. p. 623–634.
- Lund SP, Nettleton D, McCarthy DJ, Smyth GK. 2012. Detecting differential expression in RNA-sequence data using quasi-likelihood with shrunken dispersion estimates. *Stat Appl Genet Mol Biol.* 11:1–42.
- Mallez S, McCartney M. 2018. Dispersal mechanisms for zebra mussels: population genetics supports clustered invasions over spread from hub lakes in Minnesota. *Biol Invasions.* 20:2461–2484.
- Matthews J, Van der Velde G, Bij de Vaate A, Collas FPL, Koopman KR, et al. 2014. Rapid range expansion of the invasive quagga mussel in relation to zebra mussel presence in The Netherlands and Western Europe. *Biol Invasions.* 16:23–42.
- May GE, Gelembiuk GW, Panov VE, Orlova MI, Lee CE. 2006. Molecular ecology of zebra mussel invasions. *Mol Ecol.* 15:1021–1031.
- Mayer CM, Burlakova LE, Eklöv P, Fitzgerald D, Karatayev AY, et al. 2014. Benthification of freshwater lakes: exotic mussels turning ecosystems upside down. In: TF Nalepa and DW Schloesser, editors. *Quagga and Zebra Mussels: Biology, Impacts, and Control*. Boca Raton, FL: CRC Press. pp. 575–586.
- McCartney MA. 2021. Structure, function and parallel evolution of the bivalve byssus, with insights from proteomes and the zebra mussel genome. *Philos Trans R Soc Lond B Biol Sci.* 376:20200155.
- McDougall C, Aguilera F, Degnan BM. 2013. Rapid evolution of pearl oyster shell matrix proteins with repetitive, low-complexity domains. *J R Soc Interface.* 10:20130041.
- McMahon RF. 1996. The physiological ecology of the zebra mussel, *Dreissena polymorpha*, in North America and Europe. *Am Zool.* 36:339–363.
- McNickle GG, Rennie MD, Sprules WG. 2006. Changes in benthic invertebrate communities of South Bay, Lake Huron following invasion by zebra mussels (*Dreissena polymorpha*), and potential effects on lake whitefish (*Coregonus clupeaformis*) diet and growth. *J Great Lakes Res.* 32:180–193.
- Metzger MJ, Paynter AN, Siddall ME, Goff SP. 2018. Horizontal transfer of retrotransposons between bivalves and other aquatic species of multiple phyla. *Proc Natl Acad Sci U S A.* 115:E4227–E4235.
- Metzger MJ, Reinisch C, Sherry J, Goff SP. 2015. Horizontal transmission of clonal cancer cells causes leukemia in soft-shell clams. *Cell.* 161:255–263.
- Metzger MJ, Villalba A, Carballal MJ, Iglesias D, Sherry J, et al. 2016. Widespread transmission of independent cancer lineages within multiple bivalve species. *Nature.* 534:705–709.
- Mount AS, Wheeler AP, Paradkar RP, Snider D. 2004. Hemocyte-mediated shell mineralization in the eastern oyster. *Science.* 304:297–300.
- Nalepa TF, Schloesser DW. 2014. *Quagga and Zebra Mussels: Biology, Impacts, and Control*. 2nd ed. Boca Raton, FL: CRC Press.
- Nichols SJ. 1996. Variations in the reproductive cycle of *Dreissena polymorpha* in Europe, Russia, and North America. *Am Zool.* 36:311–325.
- Nishizaki M, Ackerman JD. 2017. Mussels blow rings: jet behavior affects local mixing. *Limnol Oceanogr.* 62:125–136.
- O'Neill CR, Jr. 2008. The silent invasion: finding solutions to minimize the impacts of invasive quagga mussels on water rates, water infrastructure and the environment. Washington, DC: Hearing of the US House of Representatives Committee on Natural Resources—Subcommittee on Water and Power. <https://www.seagrant.sunysb.edu/ais/pdfs/zebraquaggamusseltestimony062408.pdf>.
- Pagnucco KS, Maynard GA, Fera SA, Yan ND, Nalepa TF, et al. 2015. The future of species invasions in the Great Lakes-St. Lawrence River basin. *J Great Lakes Res.* 41:96–107.
- Palmer JM. 2019. Funannotate: a fungal genome annotation and comparative genomics pipeline, Release 1.0.1. <https://github.com/nextgenusfs/funannotate.git>.
- Perlea M, Salzberg SL. 2010. Between a chicken and a grape: estimating the number of human genes. *Genome Biol.* 11:206.
- Phase Genomics. 2019. *Aligning and QCing Phase Genomics Hi-C data*. Seattle, WA: Phase Genomics Documentation. <https://phasegenomics.github.io>.
- Powell D, Subramanian S, Suwansa-Ard S, Zhao M, O'Connor W, et al. 2018. The genome of the oyster *Saccostrea* offers insight into the environmental resilience of bivalves. *DNA Res.* 25:655–665.

- Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, et al. 2013. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res.* 41:D590–D596.
- Quinn NP, Ackerman JD. 2011. The effect of near-bed turbulence on sperm dilution and fertilization success of broadcast-spawning bivalves. *Limnol Oceanogr.* 1:176–193.
- Quinn NP, Ackerman JD. 2012. Biological and ecological mechanisms for overcoming sperm limitation in invasive dreissenid mussels. *Aquat Sci.* 74:415–425.
- R Core Team. 2018. R: a language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. <https://www.R-project.org/>.
- Raikow DF. 2004. Food web interactions between larval bluegill (*Lepomis macrochirus*) and exotic zebra mussels (*Dreissena polymorpha*). *Can J Fish Aquat Sci.* 61:497–504.
- Ran Z, Li Z, Yan X, Liao K, Kong F, et al. 2019. Chromosome-level genome assembly of the razor clam *Sinonovacula constricta* (Lamarck, 1818). *Mol Ecol Resour.* 19:1647–1658.
- Rawlings ND, Barrett AJ, Finn R. 2016. Twenty years of the MEROPS database of proteolytic enzymes, their substrates and inhibitors. *Nucleic Acids Res.* 44:D343–D350.
- Renaut S, Guerra D, Hoeh WR, Stewart DT, Bogan AE, et al. 2018. Genome survey of the freshwater mussel *Venustaconcha ellipsiformis* (Bivalvia: Unionida) using a hybrid de novo assembly approach. *Genome Biol Evol.* 10:1637–1646.
- Ringli C, Keller B, Ryser U. 2001. Glycine-rich proteins as structural components of plant cell walls. *Cell Mol Life Sci C.* 58:1430–1441.
- Roach MJ, Schmidt SA, Borneman AR. 2018. Purge Haplotigs: allelic contig reassignment for third-gen diploid genome assemblies. *BMC Bioinformatics.* 19:460.
- Robinson MD, McCarthy DJ, and Smyth GK. 2010. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 26:139–140.
- Rogers RL, Grizzard SL, Titus-McQuillan JE, Bockrath K, Patel S, et al. 2021. Gene family amplification facilitates adaptation in freshwater unionid bivalve *Megaloniaias nervosa*. *Mol Ecol.* 30:1155–1173.
- Russell-Hunter WD, Burky AJ, Hunter RD. 1981. Inter-population variation in calcareous and proteinaceous shell components in the stream limpet, *Ferrissia rivularis*. *Malacologia.* 20:255–266.
- Rzepecki LM, Waite JH. 1993. The byssus of the zebra mussel, *Dreissena polymorpha*. I: morphology and in situ protein processing during maturation. *Mol Mar Biol Biotechnol.* 2: 255–266.
- Schloesser DW, Schmuckal C. 2012. Bibliography of *Dreissena polymorpha* (zebra mussels) and *Dreissena rostriformis bugensis* (quagga mussels): 1989 to 2011. *J Shellfish Res.* 31:1205–1263.
- Seifert D, Alexander DH. 2019. GenomicConsensus variant and consensus caller Menlo Park, CA: Pacific Biosciences, Inc. <https://github.com/PacificBiosciences/GenomicConsensus>.
- Shelest E. 2017. Transcription factors in fungi: TFome Dynamics, three major families, and dual-specificity TFs. *Front Genet.* 8:53.
- Shoulders MD, Raines RT. 2009. Collagen structure and stability. *Annu Rev Biochem.* 78:929–958.
- Simao FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics.* 31:3210–3212.
- Slater GS, Birney E. 2005. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics.* 6:31.
- Smit AFA, Hubley R Green, P. 2019. Repeat Masker 4.0.9. Seattle, WA: Institute for Systems Biology. <http://repeatmasker.org/>.
- Smit AFA, Hubley R. 2008-2015. RepeatModeler Open-1.0. Seattle, WA: Institute for Systems Biology. <http://repeatmasker.org/>.
- Smith CH. 2021. A high-quality reference genome for a parasitic bivalve with doubly uniparental inheritance (Bivalvia: Unionida). *Genome Biol Evol.* 13:evab029.
- Song H, Guo X, Sun L, Wang Q, Han F, et al. 2021. The hard clam genome reveals massive expansion and diversification of inhibitors of apoptosis in Bivalvia. *BMC Biol.* 19:20.
- Soroka M, Rymaszewska A, Sańko T, Przyłucka A, Lubośny M, et al. 2018. Next-generation sequencing of *Dreissena polymorpha* transcriptome sheds light on its mitochondrial DNA. *Hydrobiologia.* 810:255–263.
- Sousa R, Gutiérrez JL, Aldridge DC. 2009. Non-indigenous invasive bivalves as ecosystem engineers. *Biol Invasions.* 11:2367–2385.
- Stanke M, Morgenstern B. 2005. AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. *Nucleic Acids Res.* 33:W465–W467.
- Stepien CA, Grigorovich IA, Gray MA, Sullivan TJ, Yerga-Woolwine S, et al. 2014. Evolutionary, biogeographic, and population genetic relationships of dreissenid mussels, with revision of component taxa. In: TF Nalepa and DW Schloesser, editors. *Quagga and Zebra Mussels: Biology, Impacts and Control*. London, CRC Press. p. 403–444.
- Strayer DL, Hattala KA, Kahnle AW. 2004. Effects of an invasive bivalve (*Dreissena polymorpha*) on fish in the Hudson River estuary. *Can J Fish Aquat Sci.* 61:924–941.
- Sullivan S. 2018. Matlock: simple tools for working with Hi-C data. Seattle, WA: Phase Genomics, Inc. <https://github.com/phasegenomics/matlock>.
- Sun J, Zhang Y, Xu T, Zhang Y, Mu H, et al. 2017. Adaptation to deep-sea chemosynthetic environments as revealed by mussel genomes. *Nat Ecol Evol.* 1:121.
- Takeuchi T, Koyanagi R, Gyoja F, Kanda M, Hisata K, et al. 2016. Bivalve-specific gene expansion in the pearl oyster genome: implications of adaptation to a sessile lifestyle. *Zool Lett.* 2:3.
- Tošenovský E, Kobak J. 2015. Impact of abiotic factors on aggregation behaviour of the zebra mussel *Dreissena polymorpha*. *J Mollus Stud.* 82:55–65.
- Uliano-Silva M, Dondero F, Dan Otto T, Costa I, Lima NCB, et al. 2018. A hybrid-hierarchical genome assembly strategy to sequence the invasive golden mussel, *Limnoperna fortunei*. *Gigascience.* 7:1–10.
- USGS. 2019a. Specimen observation data for *Dreissena polymorpha* (Pallas, 1771), Nonindigenous Aquatic Species Database. Gainesville, FL: United States Geological Survey, Wetland and Aquatic Research Center. <https://nas.er.usgs.gov/queries/default.aspx>.
- USGS. 2019b. USGS Surface-Water Daily Statistics for the Nation, Site 05331000, Mississippi River at St. Paul, MN, 10/01/1956 - 10/01/1990. retrieved from https://waterdata.usgs.gov/nwis/dv?referred_module=sw&site_no=05331000.
- Vurtture GW, Sedlazeck FJ, Nattestad M, Underwood CJ, Fang H, et al. 2017. GenomeScope: fast reference-free genome profiling from short reads. *Bioinformatics.* 33:2202–2204.
- Wainman BC, Hincks SS, Kaushik NK, Mackie GL. 1996. Biofilm and substrate preference in the dreissenid larvae of Lake Erie. *Can J Fish Aquat Sci.* 53:134–140.
- Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, et al. 2014. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One.* 9:e112963.
- Wang S, Zhang J, Jiao W, Li J, Xun X, et al. 2017. Scallop genome provides insights into evolution of bilaterian karyotype and development. *Nat Ecol Evol* 1:1–12.
- Ward JM, Ricciardi A. 2014. Impacts of Dreissena on benthic macroinvertebrate communities—Predictable patterns revealed by invasion history. In: TF Nalepa and DW Schloesser, editors. *Quagga*

- and Zebra Mussels: Biology, Impacts, and Control. Boca Raton, FL: CRC Press. p. 599–610.
- Wei M, Ge H, Shao C, Yan X, Nie H, et al. 2020. Chromosome-level clam genome helps elucidate the molecular basis of adaptation to a buried lifestyle. *iScience*. 23:101148.
- Weiner S, Traub W. 1984. Macromolecules in mollusc shells and their functions in biomineralization. *Philos Trans R Soc London B Biol Sci*. 304:425–434.
- Woznicki P, Boroń A. 2012. Banding chromosome patterns of zebra mussel *Dreissena polymorpha* (Pallas) from the heated Konin lakes system (Poland). *Caryologia*. 56:427–430.
- Wu TD, Watanabe CK. 2005. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics*. 21: 1859–1875.
- Wynn EL, Christensen AC. 2019. Repeats of unusual size in plant mitochondrial genomes: identification, incidence and evolution. *G3 (Bethesda)*. 9:549–559.
- Xu W, Faisal M. 2010. Gene expression profiling during the byssogenesis of zebra mussel (*Dreissena polymorpha*). *Mol Genet Genomics*. 283:327–339.
- Yan X, Nie H, Huo Z, Ding J, Li Z, et al. 2019. Clam genome sequence clarifies the molecular basis of its benthic adaptation and extraordinary shell color diversity. *iScience*. 19:1225–1237.
- Yanai I, Benjamin H, Shmoish M, Chalifa-Caspi V, Shklar M, et al. 2005. Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. *Bioinformatics*. 21:650–659.
- Yang JL, Feng DD, Liu J, Xu JK, Chen K, et al. 2021. Chromosome-level genome assembly of the hard-shelled mussel *Mytilus coruscus*, a widely distributed species from the temperate areas of east Asia. *Gigascience*. 10:1–13.
- Yano M, Nagai K, Morimoto K, Miyamoto H. 2006. Shematrin: a family of glycine-rich structural proteins in the shell of the pearl oyster *Pinctada fucata*. *Comp Biochem Physiol B Biochem Mol Biol*. 144:254–262.
- Yilmaz P, Parfrey LW, Yarza P, Gerken J, Pruesse E, et al. 2014. The SILVA and “All-species Living Tree Project (LTP)” taxonomic frameworks. *Nucleic Acids Res*. 42:D643–D648.
- Yonge CM, Campbell JI. 2012. II.—On the heteromyarian condition in the bivalvia with special reference to *Dreissena polymorpha* and certain Mytilacea. *Trans R Soc Edinb*. 68:21–42.
- Zhang G, Fang X, Guo X, Li L, Luo R, et al. 2012. The oyster genome reveals stress adaptation and complexity of shell formation. *Nature*. 490:49–54.

Communicating editor: D. Macqueen