The Genome of the Zebra Mussel, *Dreissena polymorpha*: A Resource for Invasive Species Research

Michael A. McCartney[1,11], 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], Daryl M. Gohl[2,10,11]

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

[2]University of Minnesota Genomics Center, Minneapolis, MN, USA

[3]Minnesota Supercomputing Institute, University of Minnesota, Minneapolis, MN, USA

[4]Institute of Biomaterials & Biomedical Engineering, University of Toronto, Toronto, ON, Canada

[5]University of Minnesota Informatics Institute, Minneapolis, MN, USA

[6]Phase Genomics, Seattle, WA, USA

[7]Department of Materials Science & Engineering, University of Toronto, Toronto, ON, Canada

[8]Faculty of Dentistry, University of Toronto, Toronto, ON, Canada

[9]Genome Informatics Section, Computational and Statistical Genomics Branch, National Human Genome Research Institute, Bethesda, Maryland, USA

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

[11]To whom correspondence should be addressed: M.A.M: mccartneymichael324@gmail.com; D.M.G: dmgohl@umn.edu

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, with nearly 450 million years of divergence between zebra mussels and its closest sequenced relative. Although the *D. polymorpha* genome is highly repetitive, we have used a combination of long-read sequencing and Hi-C-based scaffolding to generate the highest quality molluscan assembly to date. 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 [1] zebra mussels (*Dreissena polymorpha*, Fig. 1a) have spread throughout European [2,3] and North American [4] fresh waters to become one of the world's most prevalent and damaging aquatic invasive species [5]. Fouling of water intake pipes cost the power generation industry over $3 billion USD from 1993-1999 in the Laurentian Great Lakes region alone [6], where *Dreissena* cause extensive damage to hydropower, recreation and tourism industries, and lakefront property [7,8]. 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 [9-14], and dramatic restructuring of aquatic food webs [15-17]. The congener *D. 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 [18-20].

The ongoing European and North American invasions (Fig. 1c-e) have spurred an explosion in research effort on *Dreissena*, particularly focused on physiology, autecology, and ecosystem impacts [21]. Aside from molecular systematic and population genetic studies [1,22-25], comparatively little genetic work has been accomplished, with transcriptomes from a few tissues [26,27] being the only genomic resources.

Bivalves are a diverse Class of Mollusca with over 10,000 described species in marine and freshwater environments [28,29]. To date, complete genomes have been sequenced and analyzed in only eight species—most of them marine organisms of commercial value (Fig. 1b, Supplemental Table 1). Yet 21 invasive bivalve species cause damage to aquatic and marine ecosystems worldwide [30] and only the golden mussel, *Limnoperna fortunei* [31] has a published genome available (the quagga mussel is also being sequenced at present [32]). Moreover, the divergence time between *Dreissena* and other bivalve species with published genomes is estimated at more than 400 million years ago (Fig. 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 [20,33].

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 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 towards developing control strategies for this economically and ecologically damaging aquatic invader.

Results

To sequence the *D. polymorpha* genome, we used the strategy outlined in Fig. 1f. We generated a size-selected PacBio library with ≥20 kb inserts (Supplemental Fig. 1-2). Using the PacBio Sequel Single Molecule Real Time (SMRT) sequencing platform, we generated 168.97 gigabases (Gb) of sequencing data for an estimated coverage over 100x, assuming a genome size (from densitometry measures of DNA content in stained nuclei) of 1.66 Gb [34]. The subread N50 for the PacBio reads was 16,524 bp, validating the high quality of the input DNA and PacBio sequencing library.

Canu [35] 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 [34] due to the relatively high heterozygosity of the sample (2.13% estimated from GenomeScope and previous Illumina sequencing). Identification of allelic contigs [36] removed redundancy and yielded a 1.8 Gb assembly containing 2,863 contigs with a contig N50 value of 1,111,027 bp (Table 1). Hi-C [37] analysis of the polished assembly generated 16 scaffolds spanning 98.4% of the assembled genome (128 unscaffolded contigs comprised the remaining assembled material, Supplemental Fig. 3). Earlier cytogenetic work found 1N=16 chromosomes for *D. polymorpha* [38,39]. The scaffold N50 value was >107 Mb and the scaffold L50 value was 7, 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 [40] 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, Supplemental Table 2).

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. Functional annotation was carried out by mapping to a number of databases, including PFAM [41], InterPro [42], UniProtKB [43], Merops [44], and CAZymes [45]. Due to the large evolutionary divergence between *D. polymorpha* and other sequenced genomes, the majority of the predicted genes had no annotations assigned. However, 12,772 genes had recognizable orthologs.

Repetitive DNA is abundant in bivalve genomes [46-49], which makes assembly challenging. The *D. polymorpha* genome is also highly repetitive (47.4% repetitive content, Fig. 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 34.3% of the genome, could not be classified (Table 1).

The zebra mussel genome contains several notable gene family expansions (Supplemental Fig. 4, Supplemental Files 1-2). Relative to humans, *D. polymorpha* shows expansions of genes related to cellular stress responses and apoptosis that in several cases even surpass those in Pacific oyster (*Crassostrea gigas*) [49]. Expansions of gene families encoding Hsp70s (heat shock chaperones), caspases (apoptosis), and IAPs (Inhibitor of Apoptosis Proteins) exceeds Pacific oyster; expansion of Cu-Zn superoxide dismutases (antioxidant defense) and C1q domain-containing proteins (innate immunity) show expansions that are, respectively, equal to and smaller than *Crassostrea gigas*, while cytochrome P450s (xeniobiotic detoxification) are contracted relative to humans (Table 2).

Examination of orthology to eastern oyster (*Crassostrea virginica*) identified 10,065 orthologous groups (Supplemental Files 3-4). 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 (5,753; 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 [26], but contained a gap which short-read sequencing and targeted PCR were unable to resolve. PacBio and Oxford Nanopore sequencing (Fig. 2c) reveals that this "gap" is a large highly repetitive segment of nearly 50kb, 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 (Supplemental Fig. 5), with individual repeat elements of approximately 125 bp, 1030 bp, and 86 bp, each copied many times. The 86bp 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 (Supplemental Fig. 6). 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 (Fig. 2b-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 (Supplemental Fig. 5). Eumetazoan mitogenomes, with few exceptions, generally lack length variation and non-coding DNA content [26]. Among these few exceptions are the long enigmatic mitogenomes of scallops[50], 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 as is common in bivalves [50], the order of mitochondrial genes is unique to the species. The reason for this unusual mitochondrial DNA (mtDNA) structure is unknown, but similar repetitive sequences have been observed in the mtDNA of plants and it has been suggested that such repeats may result from increased double-stranded break repair due to the need to cope with desiccation-related DNA damage [51].

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 [52,53]. 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. Further, re-mapping of high-accuracy Illumina reads from the same mussel to the mitochondrial genome revealed no SNPs within the coding region (Supplemental Fig. 6), 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 [53,54].

Steamer-Like Elements

We identified a number of Long Terminal Repeat (LTR) retrotransposons that are similar in structure to *Steamer*, a transposable element (TE) that in the soft-shelled clam *Mya arenaria* causes a leukemia that is transmissible between conspecifics [55,56]. A high incidence of horizontal transmission of TEs (HTT) has spread these Steamer-Like Elements (SLEs) across several bivalves that also contract transmissible cancers, and across phyla to several marine animal species that do not [57]. We identified eight copies of putative SLEs in the *D. polymorpha* genome with intact polycistronic open reading frames (ORFs) that span the conserved Gag-Pol polyprotein and are flanked by LTRs (Fig. 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 *Steamer*-like elements (SLEs: Supplemental Fig. 7). 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[57]. Our phylogenetic analysis identified a minimum of three HTTs leading to their spread to zebra mussels from marine bivalves (Fig. 3b), including an event that is additional to and independent of the two HTTs identified previously [57]. 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 (Supplemental Fig. 8).

Tissue-specific gene expression

We next conducted a number of RNA-Seq experiments in order 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 (Fig. 4a): mantle (the organ that secretes shell), gill (the focal organ for thermal stress response), and the 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) [58] was calculated and 577 genes exceeded the threshold of $\tau = 0.95$ (Fig. 4b, Supplemental Fig. 9, Supplemental Files 5-7). 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 (Supplemental Fig. 9, Supplemental Files 8-10).

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 [59,60]. Positive correlations have been found between shell strength/calcification and ambient Ca^{2+} has been found in some freshwater mollusk species, and selection favoring shell strength for predator defense has been found in others [61,62]. In order 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 other manual annotation methods, we identified orthologs of a set of genes that have been previously implicated in shell formation (Fig. 4c, Supplemental File 11). These include tyrosinases, which are required for DOPA production, and other proteins that likely have structural roles, such as collagen. Six shematrin-like proteins are among the most specific and highly expressed mantle genes. Shematrins are glycine-rich proteins that have been identified in the mantle of other mollusks [63-66]. 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 [67,68]. 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 (Supplemental Fig. 10). Also highly expressed in the mantle were a number of Sushi, von Willebrand factor type A, EGF and pentraxin domain-containing proteins (SVEPs), which 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[69], 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

241 enrichment analysis also showed that the chitin binding molecular function was significantly
242 enriched in the mantle-specific genes, along with a number of peptidase inhibitors (Fig. 4d).

243 Among the most specific and highly expressed mantle genes in *D. polymorpha* were two
244 genes with sequence similarity to temptin, a pheromone that serves as a chemoattractant for
245 mating in *Aplysia* [70]. Zebra mussels attach to one another in clusters known as druses.
246 Settlement of larvae near adults [71] and gregarious post-settlement behaviors[72] create massive
247 aggregations on lake and river bottom. These behaviors increase settlement success, enable
248 "habitat engineering" in mussel beds [72], and may enhance feeding and fertilization success [73-75].
249 Further investigation will be required to determine if the *D. polymorpha* temptin orthologs serve
250 as aggregation cues or play some other unknown sensory role.

251

252 **Insights into byssal thread formation and attachment**

253 The fibers that zebra and quagga mussels use to anchor themselves to hard surfaces are
254 known as byssal threads. These are key innovations (absent from native North American and
255 European freshwater mollusks) used to attach to conspecific mussels, and to native unionid
256 mussels and other benthic animals, which can be smothered and outcompeted. Byssal
257 attachment to boat hulls, docks, boat lifts and other recreational equipment allows rapid rates of
258 spread between water bodies [76-78]. Expression of genes during byssogenesis has been studied
259 in zebra mussels[27] but a majority of mRNAs that are up or down-regulated could not be
260 identified.

261 Previous work identified a full byssal protein cDNA sequence (named Dpfp1) [79,80] and peptide
262 fragments from a second foot byssal protein [81]. More recent proteomic work also identified
263 peptide tags associated with several *D. polymorpha* foot proteins (Dpfps) that are secreted by
264 the foot and together form the stem, threads and attachment plaques (Fig. 5) of the byssus [82,83].
265 The zebra mussel genome provided significant additional information to identify full-length
266 proteins corresponding to the Dpfp peptide tags and to gain further insight into their function
267 (Fig. 5). BLAST was used to align known Dpfp polypeptides against the genome-predicted
268 protein sequences to determine the full Dpfp's (Supplemental File 12). Dpfp2 and Dpfp12 were
269 found to be the C- and N-terminal regions, respectively, of a single protein (Dpfp2). Similarly,
270 Dpfp6 was found to be the C-terminal region of Dpfp1. Surrounding genomic regions on
271 chromosome 8 were searched, and multiple exons were identified that are likely spliced to
272 create mature Dpfp1 and Dpfp 6 (Supplemental File 12). Complete coding sequences were

resolved for Dpfp5 and Dpfp8, and Dpfp8 was found to lack a signal peptide. Dpfp8 shows similarity to *Staphylococcus saprophyticus* cell wall associated fibronectin-binding protein (SCS67603.1). Putative homologs and paralogs to all other previously described byssal proteins were found in the zebra mussel genome, with Dpfp7, Dpfp9, Dpfp10 and Dpfp11 showing similarity to several protein coding DNA regions on multiple chromosomes. This information will inform future characterization of zebra mussel byssus formation and mechanism of adhesion.

We also examined transcripts from the foot following experimental induction of byssogenesis [27] (Supplemental Fig. 11). The foot distal to the byssus was dissected immediately after severing the byssal threads where they exit the shell valves, and four and eight days later. Changes were observed at the day-four time point, after which expression broadly returned to baseline by day eight (Supplemental Fig. 11). Some of the up-regulated genes were consistent with function identified in previous work on byssogenesis in the scallop [48], including tenascin-X (a connective protein) and a gene with phospholipid scramblase activity (Anoctamin-4-like, Supplemental Fig. 11, Supplemental File 13). In addition, there was a clear inhibition of the 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-four time point. While earlier expression studies found otherwise [27,82,83] 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 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 [84,85]. 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 in the river ranges from 5-10°C [86]. 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

306 months [87]. Seasonal scheduling of growth and reproductive effort appears to be responsible for
307 at least some of the adaptation or acclimation to conditions in the lower river, as populations in
308 Louisiana shift their shell and tissue growth to the early spring and stop growing in summer [86]
309 while more northerly populations grow tissue and spawn in summer months [88,89].

310 In order to identify genes involved in the response to thermal stress, we generated
311 transcriptomes from gill tissue in animals exposed to periods of low (24°C), moderate (27°C),
312 and high (30°C) chronic temperature stress (Fig. 6a). Moderate thermal stress led to the
313 induction of a number of genes involved in cellular adhesion or cytoskeletal remodeling,
314 including collagen, gelsolin, MYLIP E3 ubiquitin ligase, and N-cadherin (Fig. 6b, Supplemental
315 File 14). High thermal stress led to strong induction of a large number of chaperones, including
316 HSP70, DNAJ, Calnexin, and HSC70 (several of which were also induced to a lesser extent
317 under moderate thermal stress), as well as the antioxidant protein cytochrome P450 (Fig. 6b-c,
318 Supplemental File 14). The list of down-regulated genes was quite similar for both the moderate
319 and high thermal stress conditions (Fig. 6d-e, Supplemental File 14). In addition to the induction
320 of known stress-response genes, a number of genes with unknown function are also regulated
321 by thermal stress, as is 4-Hydroxyphenylpyruvate Dioxygenase (HPPD), an enzyme which is
322 involved in the catabolism of tyrosine (Fig. 6b-e, Supplemental File 14).

323

324 **Discussion**

325 Here we describe the genome of the zebra mussel. Consistent with the genomes of other
326 bivalves, the *D. polymorpha* genome is highly repetitive and encodes an expanded set of heat-
327 shock and anti-apoptotic proteins, presumably to deal with the challenges of a sessile existence.
328 We examine the genetic underpinnings of several traits that have been linked to population
329 growth and invasiveness, including shell and byssal thread formation, and response to thermal
330 stress. While these analyses uncovered multiple genes and pathways that seem to function in a
331 conserved manner across multiple bivalve species, they also uncovered a large number of
332 genes of unknown function. In the future, it will be of considerable value to compare the zebra
333 mussel genome with that of its congener, the quagga mussel (*D. rostriformis*), in order to gain
334 further insights into ecological displacement of zebra by quagga mussels, and to investigate
335 genetic underpinning of their relative invasiveness, such as comparative work on byssogenesis
336 that may help account for the slower geographic spread of quagga mussels.
337 The existence of genomic resources for *D. polymorpha* and the catalog of genes we have
338 identified will enable multiple new lines of investigation, as well as provide researchers with an

339 improved tool for population genetic experiments, for instance, tracking the spread of mussels
340 using Genotyping-by-Sequencing (GBS) approaches, or designing new targeted assays for the
341 presence or activity of zebra mussels.
342 While it is clear that changes in transportation networks (e.g. canal building, opening of
343 shipping channels, ballast water discharge) were the events that initiated primary invasions of
344 European and North American waters[5,90], several biological characteristics are responsible for
345 the rate of spread of zebra and quagga mussels across both continents, while other traits have
346 limited their suitable habitat range. Genomics offers a path to understanding these traits at the
347 genetic level, which may ultimately guide the development of control methods and management
348 strategies.

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 of 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 20kb (not shown). Further analysis by Pulsed-Field Gel Electrophoresis indicated a broad distribution from 20-120kb, with a modal size of 40kb (not shown).

Thirty µg 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 1X 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 >20kb 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 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 x 125 cycle configuration, yielding 68 Gb of data representing ~37X 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 (Feb 2018) from Phase Genomics using the Proximo™ Hi-C Animal Protocol version 1.0. This method is largely similar to previously published protocols [91].The resulting library was sequenced on a single lane of HiSeq 2500 High Output (SBS V4) in a 2 x 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-Ca^{2+} (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-Ca^{2+}" (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 six hours. 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[92]. 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 X 40 L glass aquaria with flowing well water (4 L/minute) 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 hours for feeding. Tank temperatures were raised to 24-25°C over three days by mixing in heated well water; then temperatures were held constant over seven days for acclimation.

Experimental treatments followed, with each group of 4 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 seven days. For gill transcriptomes, two mussels per each of four treatment tanks were removed, then both ctenidia were dissected and preserved in 750 μL_RNAlater per animal at -20°C. For foot, mussels from Lake Waconia, attached firmly to rocks and maintained for seven 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 four days and eight 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 Supplemental File 15.

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 RNA*later*™ (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 x 50 cycle configuration, generating approximately 15M reads per sample (Mean = 15.8 M, 15% CV).

Genome Assembly

The primary assembly was generated using Canu 1.7[35] from 167.8 Gbp of PacBio subreads over 1kbp 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 [93] and previous Illumina sequencing). BUSCO v3 [40] was run using 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 [36] 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 2,863 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 [94], 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 [95] 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 [96] was used to identify repeat families from the primary haploid genome. The resulted unknown repeat families were combined with the default full RepeatMasker [97]

database. RepeatMasker scanned the primary haploid genome sequences for the combined repeat databases in quick search mode.

Hi-C Scaffolding

Chromatin conformation capture data was 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 [91]. Following the kit protocol, intact cells from two samples were crosslinked using a formaldehyde solution, digested using the *Sau3A*I 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 2x125 cycle configuration, yielding 230,479,044 clusters passing filter.

Reads were aligned to the draft assembly also following the manufacturer's recommendations [98]. Briefly, reads were aligned using BWA-MEM [99] with the -5SP and -t 8 options specified, and all other options default. SAMBLASTER [100] was used to flag PCR duplicates, which were later excluded from analysis. Alignments were then filtered with samtools [101] using the -F 2304 filtering flag to remove non-primary and secondary alignments and further filtered with matlock [102] (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 in Bickhart et al. [37]. As in the LACHESIS method [103], this process computes a contact frequency matrix from the aligned Hi-C read pairs, normalized by the number of *Sau3*AI 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 and construction of to make them as concordant with the observed Hi-C data as possible. This process resulted in a set of 16 chromosome-scale scaffolds containing 1.79 Gb of sequence (98.4% of the contig assembly), with a scaffold N50 of 107.56 Mb and a scaffold N75 of 92.27 Mb.

516

517 **Mitochondrial Genome Assembly, Polishing, Mapping, and Annotation**

518 Mapping of PacBio reads to an initial Canu assembly for the mitochondrial genome indicated a
519 small region of very high coverage (Supplemental Fig. 5). An alternate assembly of the
520 mitochondrial genome was substituted which was generated in parallel in FALCON 0.5
521 (length_cutoff = -1, seed_coverage = 30, genome_size = 2.7G) and which did not collapse this
522 repeat sequence. This assembly was polished for indels via Pilon using Illumina reads as with
523 the nuclear genome, and a single substitution error in the coding region was manually edited
524 (c.14475 C>A, G184W) based on strong support from Illumina reads (data not shown). The
525 mitochondrial genome was annotated based a previously published partial mitochondrial
526 sequence [26] in Geneious using the "Annotate from Database" function with a 98% similarity
527 cutoff. The origin point was set to place the tRNA-Val annotation at base 48, matching the
528 previously published sequence.

529

530 PacBio and Nanopore reads were mapped against a reference file containing two concatenated
531 copies of the mitochondrial genome sequence, in order to allow reads to map across the origin.
532 Alignments were generated with minimap2 -ax using settings map-pb and map-ont,
533 respectively. Visualization of the resulting alignments (Fig. 2c) was performed using a custom
534 tool, ConcatMap (https://github.com/darylgohl/ConcatMap). Illumina reads from the polishing
535 library were mapped (Supplemental Fig. 5) to the final, polished mitochondrial genome using
536 BWA-MEM[94].

537

538 **Hi-C analysis of the mitochondrial contig**

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

547

548 **Transcriptome Assembly**

549 Reads from all zebra mussel RNAseq libraries were pooled for transcriptome assembly. A

550 database of ribosomal RNA was downloaded from SILVA[104-106], restricting the entries to
551 Bivalvia. The combined RNAseq reads were cleaned of putative ribosomal RNA sequences
552 using "BBDuk" from the BBTools suite of scripts[107], treating the Bivalvia ribosomal RNA as
553 potential contaminants, using a k-mer size of 25bp and an edit distance of 1. Reads that passed
554 this filter were then assembled with Trinity 2.8.4 [108] with a "RF" library type, *in silico* read
555 normalization, and a minimum contig length of 500bp. Assembled transcripts from Trinity were
556 then searched against the non-redundant nucleotide sequence database hosted by NCBI,
557 current as of 2018-10-09. A maximum of 20 target sequences were returned for each transcript,
558 restricted by a minimum of 10% identity and a maximum E-value of 1×10^{-5}. Assembled
559 transcripts that matched sequences derived from non-eukaryotes or synthetic constructs were
560 discarded.
561
562 **Differential Expression Analysis**
563 RNAseq reads were checked for quality issues, adapter content, and duplication with FastQC
564 0.11.7. Cleaning for sequencing adapters, trimming of low-quality bases, and filtering for length
565 were performed with Trimmomatic 3.3 [109]. The adapter sequences that were targeted for
566 removal were the standard Illumina sequencing adapters. Quality trimming was performed with
567 a window size of 4bp and a minimum mean quality score of 15. Reads that were shorter than
568 18bp after trimming were discarded.
569
570 Reads were aligned to the HiC-scaffolded genome assembly draft with HISAT2 2.1.0 [110], with
571 putative intron-exon boundaries inferred with genes with functional annotation from the draft
572 annotation and a bundled Python script. Read pairs in which one read failed quality control were
573 not used in alignment and expression analysis. BAM files from HISAT2 were cleaned of reads
574 with a mapping quality score of less than 60 with samtools 1.7. Cleaned alignments were used
575 to generate expression counts with the featureCounts program in the Subread package v. 1.6.2
576 [111]. Both reads in a pair were required to map to a feature and be in the proper orientation for
577 them to be counted. Raw read counts were imported into R 3.5.0 [112] for analysis with edgeR
578 3.24.3 [113]. Genes that were less than 200bp were removed from the counts matrix. Tests for
579 differential expression were performed between experimental conditions within tissue. For each
580 tissue, genes with low expression were filtered in the following way: genes in which at least X
581 samples with fewer than 10 were removed, where X is the size of the condition with the fewest
582 replicates. Tests for differential expression used a negative binomial model for dispersion
583 estimation, and genes showing significant levels of differential expression were identified with a

584 quasi-likelihood F test implemented in edgeR [114]. Genes were identified as differentially
585 expressed if they had a nominal *P*-value of less than 0.01 in the output from the 'glmQLFTest'
586 function.
587
588 **Tissue Specificity Calculation**
589 Filtered, normalized counts were used to calculate τ, a measure of tissue specificity: [58]
590
591
$$\tau = \frac{\sum_{i=1}^{N}(1 - x_i)}{N - 1}$$
595
596 Where *N* is the number of tissues analyzed and x_i are the normalized counts. Normalized and
597 log-transformed counts-per-million (CPM) values for each gene were estimated with edgeR. The
598 mean CPM for samples from each tissue were treated as the expression values for that tissue. τ
599 was then calculated for each gene. Genes with τ of 0.95 or greater were considered to be
600 specific to the tissue with highest expression.
601
602 **Identification of Steamer-Like Elements (SLEs) and phylogenetic analysis**
603 A sequence amplified from *D. polymorpha* using SLE-targeting degenerate primers [57] was used
604 as the basis for an initial BLAST search of the genome assembly. Dotplots of the sequence
605 surrounding hits were analyzed to identify fifty putative LTR sequences, and these were aligned
606 to build a consensus LTR sequence specific to our assembly. A subsequent BLAST search with
607 this consensus sequence was performed, and surrounding sequence context was examined for
608 the presence of long (>3 kb) ORFs between flanking LTRs. Eight intact elements identified with
609 these criteria were aligned based on coding sequence (ClustalW), and annotated based on
610 NCBI Conserved Domain search.
611
612 First, we evaluated phylogenetic evidence that zebra mussel TEs are SLEs. Amino acid
613 sequences for the full-length *Gag-Pol* polyprotein region from these eight elements and from the
614 *Steamer* element from *Mya arenaria* (Accession AIE48224.1) were aligned to a database of the
615 *Gypsy/T3y* family of LTR-retrotransposons [115], using MAFFT[116] and the E-INS-i method. The
616 alignment included 2078 residues and 105 sequences. The model of sequence evolution was

selected based on the AIC option in SMS[117], using the option to estimate amino acid frequencies from the data. A maximum likelihood genealogy was built using PhyML[118], 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 HTT events led to insertions of 20 SLEs that we found in the zebra mussel genome that contained 2 LTRs flanking an intact *Gag-Pol* ORF, including the 8 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[57]. We added 3 sequences of long ORFs from *Gag-Pol* that were cloned from neoplastic tissue[119], 3 that were obtained from *Crassostrea gigas* and *Mizuhopecten yessoensis* genome projects, and the full length *Steamer* clone from *Mya 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 PhylML and the maximum likelihood tree was constructed using the above options (except that in this case, nucleotide frequencies were ML optimized).

Genome Annotation

Functional annotation was carried out with Funannotate 1.0.1[120] 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[96] and soft-masked using RepeatMasker [97]. Second, protein evidence from a UniProtKB/Swiss-Prot-curated database (downloaded on 26 April 2017) was aligned to the genomes using tBLASTn and exonerate[121], and transcript evidence was aligned using GMAP[122]. Analysis *ab initio* used gene predictors AUGUSTUS v3.2.3[123] and GeneMark-ET v4.32[124], trained using BRAKER1[125], and tRNAs were predicted with tRNAscan-SE[126]. Consensus protein coding gene models were predicted using EvidenceModeler[127], 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[128] and Repbase[129] 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[41], InterPro[42], UniProtKB[43], Merops[44], CAZymes[45], and a set of transcription factors based on InterProScan domains[130] to assign functional annotations.

651

652 **Comparison to Eastern Oyster (*Crassostrea virginica*) Proteins**

653 Zebra mussel genes with functional annotation information were used to identify groups of
654 orthologous genes with Eastern Oyster (*Crassostrea virginica*). Annotated protein sequences
655 from *C. virginica* were downloaded from the C_virginica-3.0 assembly and annotation hosted on
656 NCBI. Zebra mussel protein sequences and *C. virginica* protein sequences were grouped into
657 orthologous groups using OrthoFinder version 2.2.7 [131], OrthoFinder was run with BLASTP 2.7.1
658 for similarity searches, MAFFT 7.305 for alignment, MCL 14.137 for clustering, and RAxML
659 8.2.11 for tree inference.

660

661 **Porting annotations to improved Hi-C scaffolds**

662 Annotations were created against a preliminary set of scaffolds and ported to the final set of
663 scaffolds following Juicebox error correction. Several files were used to perform this process.
664 First, from the preliminary set of scaffolds, annotations were generated in GFF format as
665 described above. For both the preliminary and final set of scaffolds, a .assembly file [132,133] and a
666 set of .ordering files [133] were produced by the Proximo pipeline [37]. Finally, the original contig
667 FASTA was obtained. These data were used to identify the location and orientation of each
668 contig in the final scaffolds relative to the preliminary scaffolds, and to generate a new GFF file
669 reflecting the new position of each annotation in the final scaffolds. Annotations were checked
670 with a custom Python script that compares the sequences that correspond to the gene regions
671 on each assembly. Gene sequences were extracted from the assemblies by strand-aware
672 retrieval of the "gene" features in the GFF3 file; that is, if a gene is annotated on the "minus"
673 strand, then the sequence is reverse-complemented. The resulting sequences were then
674 compared between assembly versions to ensure that the re-calculated annotations
675 corresponded to exactly the same sequence. In cases of sequence mismatch, the sequence
676 based on re-calculated annotations was reverse-complemented to diagnose strand conversion
677 issues. Annotations which spanned more than one contig in the preliminary set of scaffolds
678 were discarded during this process; it is likely such annotations, which spanned placeholder
679 unestimated gaps between contigs, were spurious calls [134]. The script used for porting the
680 annotations is available here: https://github.com/phasegenomics/annotation_mover.

681

682 **Accession codes**

683 The *D. polymorpha* genome assembly is available at NCBI (BioProject: PRJNA533175).
684 Sequencing data files are available through the NCBI Sequence Read Archive (BioProject:

PRJNA533175, PRJNA533176). Pending data release by NCBI, the *D. polymorpha* genome and annotations can be downloaded using the following links:
Genome sequence:
https://zebra_mussel.s3.msi.umn.edu/Dpolymorpha_Assembly.V2.Final_wMito.fasta.gz
Annotations:
https://zebra_mussel.s3.msi.umn.edu/Dpolymorpha_Assembly.V2.Final_wMito.gff3.gz

Acknowledgements

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

Author contributions

MAM and DMG conceived and designed experiments, analyzed data, and wrote the paper. BA prepared PacBio and Hi-C libraries, analyzed data, and wrote the paper. TK, YZ, JA, and KATS analyzed data and helped to assemble and annotate the genome. SM designed experiments, collected samples, isolated DNA. JG analyzed data. AO and EDS analyzed byssal thread attachment proteins. AB carried out sequencing of PacBio and Illumina libraries. JPB carried out nanopore sequencing. AH and HM analyzed data. IL, HM, and SS carried out Hi-C-based scaffolding. SK generated the Canu assembly and ran purge haplotigs. KBB conceived and designed experiments.

Competing Financial Interests

IL and SS have a financial interest in and are directors of Phase Genomics, a company commercializing proximity ligation technology. HM is an employee of Phase Genomics.

Figure Legends

Figure 1. Zebra mussel biogeography and genome sequencing strategy.

a) Photo of *D. polymorpha* (by N. 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. Bolded text indicates that a genome sequence for that organism is publicly available. Divergence times and tree construction based on Kumar *et al.* [135]

c-e) Maps depicting the spread of *D. polymorpha* in the United States of America from 1988 through 2018. Data from US Geological Survey, NAS database[136].

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).

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 [26]. 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 panel b.

Figure 3. Steamer-Like Elements in the *D. polymorpha* genome

a) Schematics depicting the eight SLE copies, each with 2 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[137] of DNA sequence evolution was the GTR +G (rates Γ-distributed, α = 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[57]. Arrows label HTT events 1 and 2, identified previously[57] 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[119].

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. Inhalent and exhalant siphon openings are pictured, as is the gill (ctenidium). Modified from [138].
b) Heatmap depicting 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.

Figure 5. Analysis of proteins involved in byssal thread formation.
a) SEM image of byssus, consisting of threads and plaques.
b) Summary of the *D. polymorpha* foot protein (Dpfp) proteome re-analysis using the genome assembly. Proteins on the left were identified previously, from cDNA* sequencing [79,80], and from peptide sequencing using LC-MS/MS of soluble[†, 82] and insoluble[‡, 83] proteins extracted from freshly extruded byssal threads and ESTs from a foot cDNA library [139]. The proteins on the right represent full-length Dpfp proteins identified in the genome, with the number of putative paralogous loci in parentheses.

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 up-regulated during moderate thermal stress by log2 fold-change.
c) Top 20 genes up-regulated during high thermal stress by log2 fold-change.
d) Top 20 genes down-regulated during moderate thermal stress by log2 fold-change.

781 e) Top 20 genes down-regulated during high thermal stress by log2 fold-change.

782

783 **Table 1. Genome assembly statistics.**

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

785

786 **Table 2. *D. polmorpha* gene family expansions.**

787 Selected gene family expansion data comparing *D. polymorpha* to *C. gigas*, *D. melanogaster*,

788 and *H. sapiens*. Data for *C. gigas*, *D. melanogaster*, and *H. sapiens* from Zhang *et al.* [49].

Supplemental Figure Legends

Supplemental Figure 1. PacBio sequencing library and sequencer output.
a) Size-selected PacBio *D. polymorpha* sequencing library.
b) Distribution of subread lengths in the PacBio sequencing data set.

Supplemental Figure 2. Pulsed-field Gel Electrophoresis of input gDNA. Concentrations indicate DNA concentration at time of shearing. 200 ng DNA loaded for gel visualization in each well. Image cropped, inverted, and adjusted for brightness/contrast. Ladder 1: CHEF DNA Size Standard 8-48 kb (1703707) Ladder 2: CHEF DNA Size Standard 5 kb (1703624). White arrowhead indicates 48.5 kb.

Supplemental Figure 3. Hi-C pre- and post-scaffolding heatmaps.
a) Hi-C pre-scaffolding heatmap.
b) Hi-C post-scaffolding heatmap.

Supplemental Figure 4. Divergence trees for select orthogroups.
a) Histogram of paralog group counts for the *D. polymorpha* genome.
b-d) BEAST trees with date estimates for the expansion of several gene families in *D. polymorpha* relative to the Eastern oyster (*C. virginica*).
b) OG0000776.
c) OG0000670.
d) OG0000272.

Supplemental Figure 5. Structure of the *D. polymorpha* mitochondrial genome.
a) Dotplot (self-self) of D. polymorpha mitochondrial genome demonstrating large blocks of repetitive sequence elements. Generated in Geneious based on EMBOSS tool dottup. Word size=15, tile size = 20kb.
b) Comparison of the normalized number of Hi-C links comes from a contig within the same scaffold versus the normalized number of Hi-C links from a contig assembled on a different scaffold. The links observed between the mitochondrial contig and the chromosomal scaffolds is comparable to the number of links between contigs assembled on different scaffolds.

Supplemental Figure 6. Alignment of PacBio and Illumina reads to the mitochondrial genome.

a) Coverage plot of Minimap2-aligned PacBio reads against initial Canu-assembled mitochondrial genome (concatenated) showing area of high coverage which was determined to be a collapsed repeat sequence.

b) Paired-end Illumina reads from mixed somatic/germline male tissue were aligned to the mitochondrial genome, demonstrating a lack of SNPs that might otherwise indicate heteroplasmy. Only the coding region is shown, as unambiguous mapping of short reads to highly repetitive sequences is unreliable. Allele threshold for coverage plot = 0.05.

Supplemental Figure 7. Phylogenetic tree of the *Ty3/Gypsy* family of retrotransposons, including putative SLEs in zebra mussels.

Maximum likelihood phylogenetic tree of amino acid sequences from the entire *Gag-Pol* region. The selected model of amino acid sequence evolution was the LG[140] model +G (rates Gamma-distributed, α = 1.566) +I (estimated proportion of invariant sites =.001)+ F (amino-acid frequencies estimated from the data). The analysis included all sequenced elements from branch 2 of the *Ty3/Gypsy* family (including LTR retrotransposons and non-chromodomain retroviruses[141], but only the Mag clades (A, B, and C in colored boxes) are shown, along with the sister clade. *Steamer* (arrow) from *Mya arenaria* groups with the sea urchin retroelement SURL, in clade C[55]. Here we show that the *D. polymorpha* elements are sister to *Steamer*, confirming that they are SLEs. Bootstrap support values > 70 label the nodes, and the scale bar is expected changes per site from maximum likelihood.

Supplemental Figure 8. Partial SLEs in the *D. polymorpha* genome.

Schematic depicting incomplete SLEs, including LTR-only sequences in the *D. polymorpha* genome. Sequences are centered on an LTR element and additional annotated domains in the SLE ORF are colored as indicated in Fig. 3a.

Supplemental Figure 9. Tissue-specificity scores.

a) Histogram depicting the distribution of tissue-specificity (tau) scores for *D. polymorpha* genes. A cut-off of 0.95 was used to define tissue-specific expression (dashed line).

b) Venn diagram of tissue-specific genes compared to genes that were differentially expressed under different experimental conditions.

Supplemental Figure 10. *D. polymorpha* shematrin-like proteins.

Multiple sequence alignment of the six shematrin-like proteins identified in the *D. polymorpha* genome using CLUSTAL Omega (https://www.ebi.ac.uk/Tools/msa/clustalo/).

Supplemental Figure 11. Foot gene expression during byssal thread formation.

a) Schematic depicting experimental design; byssal threads were severed at day zero, and dissected foot tissue was collected on days zero, four, and eight (n = 4 animals per condition).

b) Gene expression changes (log2 fold-change) relative to the day zero time point.

c) List of up- and down-regulated genes in the foot at the day-four time point.

Supplemental Table 1. Sequenced bivalve genomes.

Supplemental Table 2. RepeatMasker output.

871 **Supplemental Files**

872

873 **Supplemental File 1. Gene IDs for zebra mussel paralogous groups.**

874 ZM_Paralogous_Groups.csv

875

876 **Supplemental File 2. Annotation of zebra mussel paralogous groups.**

877 ZM_GFF_Paralogous_Groups.xlsx

878

879 **Supplemental File 3. Orthogroup gene IDs for zebra mussel and Eastern oyster**

880 **comparison.**

881 ZM_vs_EO_Orthogroups.txt

882

883 **Supplemental File 4. Annotated orthogroups for zebra mussel and Eastern oyster**

884 **comparison.**

885 Orthogroups_EO_ZM_Products.xlsx

886

887 **Supplemental File 5. Mantle-specific genes.**

888 Mantle-specific_genes.csv

889

890 **Supplemental File 6. Gill-specific genes.**

891 Gill-specific_genes.csv

892

893 **Supplemental File 7. Foot-specific genes.**

894 Foot-specific_genes.csv

895

896 **Supplemental File 8. Differentially expressed genes in mantle in mussels collected from**

897 **high and low calcium environments.**

898 Mantle_DEGs.csv

899

900 **Supplemental File 9. Differentially expressed genes in the gill in mussels exposed to**

901 **different thermal stress conditions.**

902 Gill_DEGs_.csv

903

Supplemental File 10. Differentially expressed genes in the foot in response to severing the byssal threads.

Foot_DEGs.csv

Supplemental File 11. Mantle-specific BLAST results.

Mantle-specific_BLAST_results.xlsx

Supplemental File 12. Identification of full length Dpfp proteins.

Dpfp_analysis.docx

Supplemental File 13. Foot DGE BLAST results.

Foot_DEG_BLAST_results.xlsx

Supplemental File 14. Gill DGE BLAST results.

Gill_DEG_BLAST_results.xlsx

Supplemental File 15. Metadata associated with RNA-Seq samples.

Transcriptome_metadata.xlsx

924 **References**

925 1 Stepien, C. A. *et al.* in *Quagga and Zebra Mussels: Biology, Impacts and Control* (eds
926 Thomas F. Nalepa & Don W. Schloesser) 403-444 (CRC Press, 2014).
927 2 Karatayev, A. Y., Burlakova, L. E. & Padilla, D. K. The effects of *Dreissena polymorpha*
928 (Pallas) invasion on aquatic communities in eastern Europe. *Journal of Shellfish*
929 *Research* **16**, 187-203 (1997).
930 3 Karatayev, A. Y., Burlakova, L. E., Padilla, D. K. & Johnson, L. E. Patterns of spread of
931 the zebra mussel (*Dreissena polymorpha* (Pallas)): The continuing invasion of
932 Belarussian lakes. *Biological Invasions* **5**, 213-221, doi:10.1023/A:1026112915163
933 (2003).
934 4 Benson, A. J. in *Quagga and Zebra Mussels : Biology, Impacts, and Control (2nd*
935 *Edition)* (eds Thomas F. Nalepa & Don W. Schloesser) 9-32 (CRC Press, 2014).
936 5 Karatayev, A. Y., Padilla, D. K., Minchin, D., Boltovskoy, D. & Burlakova, L. E. Changes
937 in global economies and trade: The potential spread of exotic freshwater bivalves.
938 *Biological Invasions* **9**, 161-180, doi:10.1007/s10530-006-9013-9 (2007).
939 6 O'Neill, C. R., Jr. in *U.S. House of Representatives Committee on Natural Resources –*
940 *Subcommittee on Water and Power* 1-13 (Washington, D.C., 2008).
941 7 Bossenbroek, J. M., Finnoff, D. C., Shogren, J. F. & Warziniack, T. W. in *Bioeconomics*
942 *of Invasive Species: Integrating Ecology, Economics, Policy, and Management. Oxford*
943 *University Press, Oxford* (eds Rueben P. Keller, David M. Lodge, Mark A. Lewis, &
944 Jason F Shogren) 244-265 (Oxford University Press, 2009).
945 8 Limburg, K. E., Luzadis, V. A., Ramsey, M., Schulz, K. L. & Mayer, C. M. The good, the
946 bad, and the algae: perceiving ecosystem services and disservices generated by zebra
947 and quagga mussels. *Journal of Great Lakes Research* **36**, 86-92,
948 doi:10.1016/j.jglr.2009.11.007 (2010).
949 9 Strayer, D. L., Hattala, K. A. & Kahnle, A. W. Effects of an invasive bivalve (*Dreissena*
950 *polymorpha*) on fish in the Hudson River estuary. *Canadian Journal of Fisheries and*
951 *Aquatic Sciences* **61**, 924-941, doi:10.1139/f04-043 (2004).
952 10 McNickle, G. G., Rennie, M. D. & Sprules, W. G. Changes in benthic invertebrate
953 communities of South Bay, Lake Huron following invasion by zebra mussels (*Dreissena*
954 *polymorpha*), and potential effects on lake whitefish (*Coregonus clupeaformis*) diet and
955 growth. *Journal of Great Lakes Research* **32**, 180-193, doi:10.3394/0380-
956 1330(2006)32[180:CIBICO]2.0.CO;2 (2006).
957 11 Lucy, F., Burlakova, L., Karatayev, A., Mastitsky, S. & Zanatta, D. in *Quagga and zebra*
958 *mussels: biology, impact, and control* (eds T. F. Nalepa & D. W. Schloesser) 623-634
959 (CRC Press, 2014).
960 12 Ward, J. & Ricciardi, A. in *Quagga and zebra mussels: biology, impacts, and control*
961 (eds Thomas F. Nalepa & Don W. Schloesser) 599-610 (CRC Press, 2014).
962 13 Karatayev, A. Y. *et al.* Twenty five years of changes in *Dreissena* spp. populations in
963 Lake Erie. *Journal of Great Lakes Research* **40**, 550-559,
964 doi:https://doi.org/10.1016/j.jglr.2014.04.010 (2014).
965 14 Raikow, D. F. Food web interactions between larval bluegill (*Lepomis macrochirus)* and
966 exotic zebra mussels (*Dreissena polymorpha*). *Canadian Journal of Fisheries and*
967 *Aquatic Sciences* **61**, 497-504, doi:10.1139/f03-171 (2004).
968 15 Bootsma, H. & Liao, Q. in *Quagga and Zebra Mussels : Biology, Impacts, and Control*
969 *(2nd Edition)* (eds Thomas F. Nalepa & Don W. Schloesser) 555-574 (CRC Press,
970 2014).
971 16 Higgins, S. N. & Vander Zanden, M. J. What a difference a species makes: a meta–
972 analysis of dreissenid mussel impacts on freshwater ecosystems. *Ecological*
973 *Monographs* **80**, 179-196, doi:10.1890/09-1249.1 (2010).

17 Mayer, C. *et al.* in *Quagga and Zebra Mussels : Biology, Impacts, and Control* (eds Thomas F. Nalepa & Don W. Schloesser) 575-586 (CRC Press, 2014).
18 Karatayev, A. Y., Mastitsky, S. E., Padilla, D. K., Burlakova, L. E. & Hajduk, M. Differences in growth and survivorship of zebra and quagga mussels: size matters. *Hydrobiologia* **668**, 183-194, doi:10.1007/s10750-010-0533-z (2011).
19 Matthews, J. *et al.* Rapid range expansion of the invasive quagga mussel in relation to zebra mussel presence in The Netherlands and Western Europe. *Biological Invasions* **16**, 23-42, doi:10.1007/s10530-013-0498-8 (2014).
20 Nalepa, T. F. & Schloesser, D. W. 816 (CRC Press, Boca Raton, FL, 2014).
21 Schloesser, D. W. & Schmuckal, C. Bibliography of *Dreissena polymorpha* (zebra mussels) and *Dreissena rostriformis bugensis* (quagga mussels): 1989 to 2011. *Journal of Shellfish Research* **31**, 1205-1263, doi:10.2983/035.031.0432 (2012).
22 Gelembiuk, G. W., May, G. E. & Lee, C. E. Phylogeography and systematics of zebra mussels and related species. *Molecular Ecology* **15**, 1033-1050, doi:10.1111/j.1365-294X.2006.02816.x (2006).
23 May, G. E., Gelembiuk, G. W., Panov, V. E., Orlova, M. I. & Lee, C. E. Molecular ecology of zebra mussel invasions. *Molecular Ecology* **15**, 1021-1031, doi:10.1111/j.1365-294X.2006.02814.x (2006).
24 Brown, J. E. & Stepien, C. A. Population genetic history of the dreissenid mussel invasions: expansion patterns across North America. *Biological Invasions* **12**, 3687-3710 (2010).
25 Mallez, S. & McCartney, M. Dispersal mechanisms for zebra mussels: population genetics supports clustered invasions over spread from hub lakes in Minnesota. *Biological Invasions*, doi:10.1007/s10530-018-1714-3 (2018).
26 Soroka, M. *et al.* Next-generation sequencing of *Dreissena polymorpha* transcriptome sheds light on its mitochondrial DNA. *Hydrobiologia* **810**, 255-263, doi:10.1007/s10750-017-3088-4 (2018).
27 Xu, W. & Faisal, M. Gene expression profiling during the byssogenesis of zebra mussel (*Dreissena polymorpha*). *Molecular Genetics and Genomics* **283**, 327-339, doi:10.1007/s00438-010-0517-8 (2010).
28 Appeltans, W. *et al.* The magnitude of global marine species diversity. *Current Biology* **22**, 2189-2202, doi:https://doi.org/10.1016/j.cub.2012.09.036 (2012).
29 Bogan, A. in *Freshwater Animal Diversity Assessment* Vol. 198 *Developments in Hydrobiology* (eds E. V. Balian, C. Lévêque, H. Segers, & K. Martens) Ch. 16, 139-147 (Springer Netherlands, 2008).
30 Sousa, R., Gutiérrez, J. L. & Aldridge, D. C. Non-indigenous invasive bivalves as ecosystem engineers. *Biological Invasions* **11**, 2367-2385, doi:10.1007/s10530-009-9422-7 (2009).
31 Uliano-Silva, M. *et al.* A hybrid-hierarchical genome assembly strategy to sequence the invasive golden mussel *Limnoperna fortunei*. *GigaScience*, gix128, doi:10.1093/gigascience/gix128 (2017).
32 Calcino, A. D. *et al.* The quagga mussel genome and the evolution of freshwater tolerance. *bioRxiv*, 505305, doi:10.1101/505305 (2018).
33 Lowe, S., Browne, M., Boudjelas, S. & De Poorter, M. 100 of the world's worst invasive alien species: a selection from the global invasive species database. 12 (Aukland, New Zealand, 2000).
34 Gregory, T. R. Genome size estimates for two important freshwater molluscs, the zebra mussel (*Dreissena polymorpha*) and the schistosomiasis vector snail (*Biomphalaria glabrata*). *Genome* **46**, 841-844, doi:10.1139/g03-069 (2003).

35 Koren, S. *et al.* Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res* **27**, 722-736, doi:10.1101/gr.215087.116 (2017).

36 Roach, M. J., Schmidt, S. A. & Borneman, A. R. Purge Haplotigs: allelic contig reassignment for third-gen diploid genome assemblies. *BMC Bioinformatics* **19**, 460, doi:10.1186/s12859-018-2485-7 (2018).

37 Bickhart, D. M. *et al.* Single-molecule sequencing and chromatin conformation capture enable de novo reference assembly of the domestic goat genome. *Nat Genet* **49**, 643-650, doi:10.1038/ng.3802 (2017).

38 Boroñ A., P., W., L., S. & R., Z. Cytogenetic characterization of the zebra mussel *Dreissena polymorpha* (Pallas) from Miedwie Lake, Poland. *Folia Biologica (Kraków)* **52**, 33-38 (2004).

39 Woznicki, P. & Boroń, A. Banding chromosome patterns of zebra mussel *Dreissena polymorpha* (Pallas) from the heated Konin lakes system (Poland). *Caryologia* **56**, 427-430, doi:10.1080/00087114.2003.10589354 (2012).

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

41 Finn, R. D. *et al.* Pfam: the protein families database. *Nucleic Acids Res* **42**, D222-230, doi:10.1093/nar/gkt1223 (2014).

42 Jones, P. *et al.* InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236-1240, doi:10.1093/bioinformatics/btu031 (2014).

43 Apweiler, R. *et al.* UniProt: The Universal Protein knowledgebase. *Nucleic Acids Res* **32**, D115-119, doi:10.1093/nar/gkh131 (2004).

44 Rawlings, N. D., Barrett, A. J. & Finn, R. Twenty years of the MEROPS database of proteolytic enzymes, their substrates and inhibitors. *Nucleic Acids Res* **44**, D343-350, doi:10.1093/nar/gkv1118 (2016).

45 Lombard, V., Golaconda Ramulu, H., Drula, E., Coutinho, P. M. & Henrissat, B. The carbohydrate-active enzymes database (CAZy) in 2013. *Nucleic Acids Res* **42**, D490-495, doi:10.1093/nar/gkt1178 (2014).

46 Sun, J. *et al.* Adaptation to deep-sea chemosynthetic environments as revealed by mussel genomes. *Nature Ecology & Evolution* **1**, 0121, doi:10.1038/s41559-017-0121 (2017).

47 Wang, S. *et al.* Scallop genome provides insights into evolution of bilaterian karyotype and development. *Nature Ecology & Evolution* **1**, 1-12, doi:10.1038/s41559-017-0120 (2017).

48 Li, Y. *et al.* Scallop genome reveals molecular adaptations to semi-sessile life and neurotoxins. *Nature Communications* **8**, 1721, doi:10.1038/s41467-017-01927-0 (2017).

49 Zhang, G. *et al.* The oyster genome reveals stress adaptation and complexity of shell formation. *Nature* **490**, 49-54, doi:https://doi.org/10.1038/nature11413 (2012).

50 Boore, J. L. Survey and summary: Animal mitochondrial genomes. *Nucleic Acids Res* **27**, 1767-1780 (1999).

51 Wynn, E. L. & Christensen, A. C. Repeats of unusual size in plant mitochondrial genomes: Identification, incidence and evolution. *G3 (Bethesda)* **9**, 549-559, doi:10.1534/g3.118.200948 (2019).

52 Doucet-Beaupré, H. *et al.* Mitochondrial phylogenomics of the Bivalvia (Mollusca): searching for the origin and mitogenomic correlates of doubly uniparental inheritance of mtDNA. *BMC Evolutionary Biology* **10**, 50 (2010).

53 Breton, S., Beaupre, H. D., Stewart, D. T., Hoeh, W. R. & Blier, P. U. The unusual system of doubly uniparental inheritance of mtDNA: isn't one enough? *Trends Genet* **23**, 465-474, doi:10.1016/j.tig.2007.05.011 (2007).

54 Breton, S., Stewart, D. T. & Hoeh, W. R. 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, doi:10.1016/j.margen.2010.01.001 (2010).

55 Arriagada, G. *et al.* 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, doi:10.1073/pnas.1409945111 (2014).

56 Metzger, M. J., Reinisch, C., Sherry, J. & Goff, S. P. Horizontal transmission of clonal cancer cells causes leukemia in soft-shell clams. *Cell* **161**, 255-263, doi:10.1016/j.cell.2015.02.042 (2015).

57 Metzger, M. J., Paynter, A. N., Siddall, M. E. & Goff, S. P. Horizontal transfer of retrotransposons between bivalves and other aquatic species of multiple phyla. *Proc Natl Acad Sci U S A* **115**, E4227-E4235, doi:10.1073/pnas.1717227115 (2018).

58 Yanai, I. *et al.* Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. *Bioinformatics* **21**, 650-659, doi:10.1093/bioinformatics/bti042 (2005).

59 Weiner, S. & Traub, W. Macromolecules in mollusc shells and their functions in biomineralization. *Philosophical Transactions of the Royal Society of London. B, Biological Sciences* **304**, 425-434 (1984).

60 Mount, A. S., Wheeler, A. P., Paradkar, R. P. & Snider, D. Hemocyte-mediated shell mineralization in the eastern oyster. *Science* **304**, 297-300 (2004).

61 Russell-Hunter, W., Burky, A. & Hunter, R. Inter-population variation in calcareous and proteinaceous shell components in the stream limpet, *Ferrissia rivularis*. *Malacologia* **20**, 255-266 (1981).

62 Lewis, D. B. & Magnuson, J. J. Intraspecific gastropod shell strength variation among north temperate lakes. *Canadian Journal of Fisheries and Aquatic Sciences* **56**, 1687-1695, doi:10.1139/f99-110 (1999).

63 Lin, Y. *et al.* Cloning and characterization of the shell matrix protein Shematrin in scallop *Chlamys farreri*. *Acta Biochim Biophys Sin (Shanghai)* **46**, 709-719, doi:10.1093/abbs/gmu054 (2014).

64 Yano, M., Nagai, K., Morimoto, K. & Miyamoto, H. 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, doi:10.1016/j.cbpb.2006.03.004 (2006).

65 McDougall, C., Aguilera, F. & Degnan, B. M. Rapid evolution of pearl oyster shell matrix proteins with repetitive, low-complexity domains. *Journal of The Royal Society Interface* **10** (2013).

66 Jackson, D. J. *et al.* Parallel evolution of nacre building gene sets in molluscs. *Molecular Biology and Evolution* **27**, 591-608 (2010).

67 Shoulders, M. D. & Raines, R. T. Collagen structure and stability. *Annu Rev Biochem* **78**, 929-958, doi:10.1146/annurev.biochem.77.032207.120833 (2009).

68 Ringli, C., Keller, B. & Ryser, U. Glycine-rich proteins as structural components of plant cell walls. *Cellular and Molecular Life Sciences CMLS* **58**, 1430-1441 (2001).

69 Takeuchi, T. *et al.* Bivalve-specific gene expansion in the pearl oyster genome: implications of adaptation to a sessile lifestyle. *Zoological Letters* **2**, 3, doi:10.1186/s40851-016-0039-2 (2016).

70 Cummins, S. F. *et al.* 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, doi:10.1074/jbc.M313585200 (2004).

71 Wainman, B. C., Hincks, S. S., Kaushik, N. K. & Mackie, G. L. Biofilm and substrate preference in the dreissenid larvae of Lake Erie. *Canadian Journal of Fisheries and Aquatic Sciences* **53**, 134-140 (1996).

72 Tošenovský, E. & Kobak, J. Impact of abiotic factors on aggregation behaviour of the zebra mussel *Dreissena polymorpha*. *Journal of Molluscan Studies*, eyv033, doi:10.1093/mollus/eyv033 (2015).

73 Nishizaki, M. & Ackerman, J. D. Mussels blow rings: Jet behavior affects local mixing. *Limnology and Oceanography* **62**, 125-136, doi:10.1002/lno.10380 (2017).

74 Quinn, N. P. & Ackerman, J. D. The effect of near-bed turbulence on sperm dilution and fertilization success of broadcast-spawning bivalves. *Limnology and Oceanography: Fluids and Environments* **1**, 176-193, doi:10.1215/21573698-1504517 (2011).

75 Quinn, N. P. & Ackerman, J. D. Biological and ecological mechanisms for overcoming sperm limitation in invasive dreissenid mussels. *Aquatic Sciences* **74**, 415-425, doi:10.1007/s00027-011-0237-0 (2012).

76 Collas, F. P. L., Karatayev, A. Y., Burlakova, L. E. & Leuven, R. S. E. W. Detachment rates of dreissenid mussels after boat hull-mediated overland dispersal. *Hydrobiologia* **810**, 77-84, doi:10.1007/s10750-016-3072-4 (2018).

77 De Ventura, L., Weissert, N., Tobias, R., Kopp, K. & Jokela, J. Overland transport of recreational boats as a spreading vector of zebra mussel *Dreissena polymorpha*. *Biological Invasions* **18**, 1451-1466, doi:10.1007/s10530-016-1094-5 (2016).

78 Johnson, L. E., Ricciardi, A. & Carlton, J. T. Overland dispersal of aquatic invasive species: a risk assessment of transient recreational boating. *Ecological Applications* **11**, 1789-1799, doi:10.1890/1051-0761(2001)011[1789:odoais]2.0.co;2 (2001).

79 Anderson, K. E. & Waite, J. H. A major protein precursor of zebra mussel (*Dreissena polymorpha)* byssus: Deduced sequence and significance. *Biological Bulletin* **194**, 150-160, doi:10.2307/1543045 (1998).

80 Anderson, K. E. & Waite, J. H. Immunolocalization of Dpfp1, a byssal protein of the zebra mussel *Dreissena polymorpha*. *Journal of experimental biology* **203**, 3065-3076 (2000).

81 Rzepecki, L. & Waite, J. The byssus of the zebra mussel, *Dreissena polymorpha*. I: Morphology and *in situ* protein processing during maturation. *Molecular marine biology and biotechnology* **2**, 255-266 (1993).

82 Gantayet, A., Ohana, L. & Sone, E. D. Byssal proteins of the freshwater zebra mussel, *Dreissena polymorpha*. *Biofouling* **29**, 77-85, doi:10.1080/08927014.2012.746672 (2013).

83 Gantayet, A., Rees, D. J. & Sone, E. D. Novel proteins identified in the insoluble byssal matrix of the freshwater zebra mussel. *Marine Biotechnology* **16**, 144-155, doi:10.1007/s10126-013-9537-9 (2014).

84 McMahon, R. F. The physiological ecology of the zebra mussel, *Dreissena polymorpha*, in North America and Europe. *American Zoologist* **36**, 339-363 (1996).

85 Nichols, S. J. Variations in the reproductive cycle of *Dreissena polymorpha* in Europe, Russia, and North America. *American Zoologist* **36**, 311-325 (1996).

86 Allen, Y. C., Thompson, B. A. & Ramcharan, C. W. Growth and mortality rates of the zebra mussel, *Dreissena polymorpha,* in the Lower Mississippi River. *Canadian Journal of Fisheries and Aquatic Sciences* **56**, 748-759, doi:10.1139/f98-212 (1999).

87 Survey, U. G. & System, N. W. I. *USGS daily statistics for the nation*, <https://nwis.waterdata.usgs.gov/nwis/dvstat?search_site_no=05331000&format=sites_selection_links> (2019).
88 Borcherding, J. The annual reproductive cycle of the freshwater mussel *Dreissena polymorpha* Pallas in lakes. *Oecologia* **87**, 208-218, doi:10.1007/BF00325258 (1991).
89 Claxton, W. T. & Mackie, G. L. Seasonal and depth variations in gametogenesis and spawning of *Dreissena polymorpha* and *Dreissena bugensis* in eastern Lake Erie. *Canadian Journal of Zoology* **76**, 2010-2019, doi:10.1139/z98-150 (1998).
90 Pagnucco, K. S. *et al.* The future of species invasions in the Great Lakes-St. Lawrence River basin. *Journal of Great Lakes Research* **41**, 96-107, doi:https://doi.org/10.1016/j.jglr.2014.11.004 (2015).
91 Lieberman-Aiden, E. *et al.* Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289-293, doi:10.1126/science.1181369 (2009).
92 Baldwin, B. S., Carpenter, M., Rury, K. & Woodward, E. Low dissolved ions may limit secondary invasion of inland waters by exotic round gobies and dreissenid mussels in North America. *Biological Invasions* **14**, 1157-1175, doi:10.1007/s10530-011-0146-0 (2012).
93 Vurture, G. W. *et al.* GenomeScope: fast reference-free genome profiling from short reads. *Bioinformatics* **33**, 2202-2204, doi:10.1093/bioinformatics/btx153 (2017).
94 GenomicConsensus v. 2.3.3 (Pacific Biosciences, Inc., Menlo Park, CA, 2019).
95 Walker, B. J. *et al.* Pilon: An integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One* **9**, e112963, doi:10.1371/journal.pone.0112963 (2014).
96 RepeatModeler - 1.0.11 (Institute for Systems Biology, Seattle, WA, 2019).
97 Repeat Masker v. 4.09.2019 (Institute of Systems Biology, Seattle, WA, 2019).
98 Genomics, P. *Aligning and QCing Phase Genomics Hi-C data*, <https://phasegenomics.github.io/2019/09/19/hic-alignment-and-qc.html> (2019).
99 Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**, 589-595, doi:10.1093/bioinformatics/btp698 (2010).
100 Faust, G. G. & Hall, I. M. SAMBLASTER: fast duplicate marking and structural variant read extraction. *Bioinformatics* **30**, 2503-2505, doi:10.1093/bioinformatics/btu314 (2014).
101 Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079, doi:10.1093/bioinformatics/btp352 (2009).
102 Matlock: Simple tools for working with Hi-C data (Phase Genomics, 2018).
103 Burton, J. N. *et al.* Chromosome-scale scaffolding of *de novo* genome assemblies based on chromatin interactions. *Nature Biotechnology* **31**, 1119-1125, doi:10.1038/nbt.2727 (2013).
104 Glockner, F. O. *et al.* 25 years of serving the community with ribosomal RNA gene reference databases and tools. *J Biotechnol* **261**, 169-176, doi:10.1016/j.jbiotec.2017.06.1198 (2017).
105 Quast, C. *et al.* The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res* **41**, D590-596, doi:10.1093/nar/gks1219 (2013).
106 Yilmaz, P. *et al.* The SILVA and "All-species Living Tree Project (LTP)" taxonomic frameworks. *Nucleic Acids Res* **42**, D643-648, doi:10.1093/nar/gkt1209 (2014).
107 BBMap short read aligner and other bioinformatic tools (Joint Genome Institute, Berkeley, CA, 2019).
108 Grabherr, M. G. *et al.* Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology* **29**, 644, doi:10.1038/nbt.1883 (2011).

109 Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114-2120, doi:10.1093/bioinformatics/btu170 (2014).

110 Kim, D., Langmead, B. & Salzberg, S. L. HISAT: a fast spliced aligner with low memory requirements. *Nat Meth* **12**, 357-360, doi:10.1038/nmeth.3317 (2015).

111 Liao, Y., Smyth, G. K. & Shi, W. The Subread aligner: fast, accurate and scalable read mapping by seed-and-vote. *Nucleic Acids Res* **41**, e108, doi:10.1093/nar/gkt214 (2013).

112 R: A language and environment for statistical computing v. 3.5.0 (R Foundation for Statistical Computing, Vienna, Austria, 2018).

113 Package "EdgeR": Empirical analysis of digital gene expression data in R v. 3.24.3 (2013).

114 Lund, S. P., Nettleton, D., McCarthy, D. J. & Smyth, G. K. in *Statistical Applications in Genetics and Molecular Biology* Vol. 11 (2012).

115 Llorens, C. *et al.* The Gypsy Database (GyDB) of mobile genetic elements: release 2.0. *Nucleic Acids Res* **39**, D70-74, doi:10.1093/nar/gkq1061 (2011).

116 Katoh, K., Rozewicki, J. & Yamada, K. D. MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization. *Brief Bioinform*, doi:10.1093/bib/bbx108 (2017).

117 Lefort, V., Longueville, J. E. & Gascuel, O. SMS: Smart Model Selection in PhyML. *Mol Biol Evol* **34**, 2422-2424, doi:10.1093/molbev/msx149 (2017).

118 Guindon, S. & Gascuel, O. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Systematic Biology* **52**, 696-704 (2003).

119 Metzger, M. J. *et al.* Widespread transmission of independent cancer lineages within multiple bivalve species. *Nature* **534**, 705-709, doi:10.1038/nature18599 (2016).

120 Funannotate: a fungal genome annotation and comparative genomics pipeline, Release 1.0.1 (2019).

121 Slater, G. S. & Birney, E. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* **6**, 31, doi:10.1186/1471-2105-6-31 (2005).

122 Wu, T. D. & Watanabe, C. K. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics* **21**, 1859-1875, doi:10.1093/bioinformatics/bti310 (2005).

123 Stanke, M. & Morgenstern, B. AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. *Nucleic Acids Res* **33**, W465-467, doi:10.1093/nar/gki458 (2005).

124 Besemer, J. & Borodovsky, M. GeneMark: web software for gene finding in prokaryotes, eukaryotes and viruses. *Nucleic Acids Res* **33**, W451-454, doi:10.1093/nar/gki487 (2005).

125 Hoff, K. J., Lange, S., Lomsadze, A., Borodovsky, M. & Stanke, M. BRAKER1: Unsupervised RNA-Seq-vased genome annotation with GeneMark-ET and AUGUSTUS. *Bioinformatics* **32**, 767-769 (2016).

126 Lowe, T. M. & Chan, P. P. tRNAscan-SE On-line: integrating search and context for analysis of transfer RNA genes. *Nucleic Acids Res* **44**, W54-57, doi:10.1093/nar/gkw413 (2016).

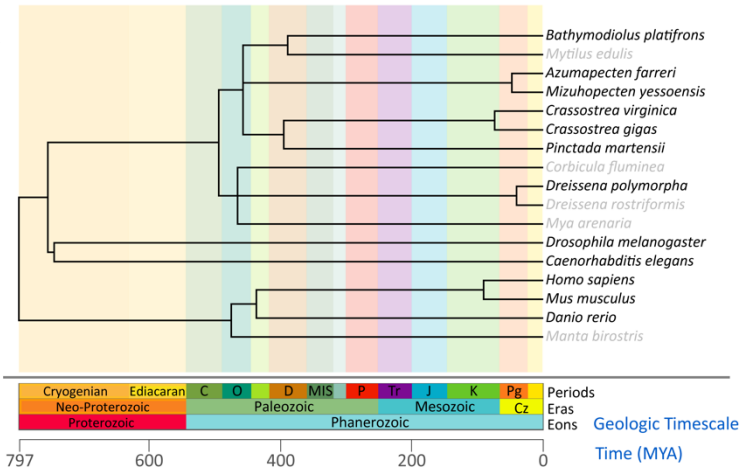
127 Haas, B. J. *et al.* Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biology* **9**, R7, doi:10.1186/gb-2008-9-1-r7 (2008).

128 Haas, B. TransposonPSI: An Application of PSI-Blast to Mine (Retro-)Transposon ORF Homologies. (2010).

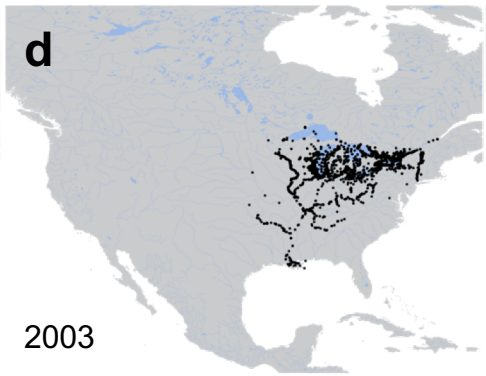
129 Bao, W., Kojima, K. K. & Kohany, O. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mob DNA* **6**, 11, doi:10.1186/s13100-015-0041-9 (2015).

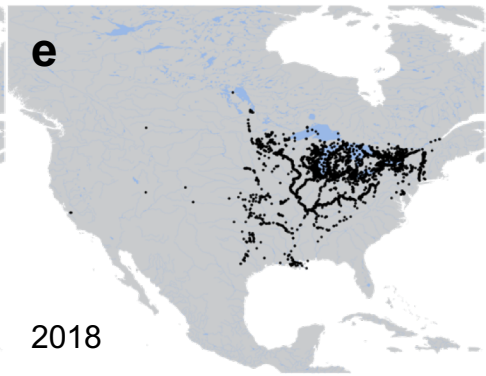
130 Shelest, E. Transcription factors in fungi: TFome Dynamics, three major families, and dual-specificity TFs. *Front Genet* **8**, 53, doi:10.3389/fgene.2017.00053 (2017).
131 Emms, D. M. & Kelly, S. OrthoFinder2: fast and accurate phylogenomic orthology analysis from gene sequences. *bioRxiv*, 466201, doi:10.1101/466201 (2018).
132 Durand, N. C. *et al.* Juicebox Provides a Visualization System for Hi-C Contact Maps with Unlimited Zoom. *Cell Syst* **3**, 99-101, doi:10.1016/j.cels.2015.07.012 (2016).
133 Rao, S. S. *et al.* A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**, 1665-1680, doi:10.1016/j.cell.2014.11.021 (2014).
134 Salzberg, S. L. Next-generation genome annotation: we still struggle to get it right. *Genome Biol* **20**, 92, doi:10.1186/s13059-019-1715-2 (2019).
135 Kumar, S., Stecher, G., Suleski, M. & Hedges, S. B. TimeTree: A resource for timelines, timetrees, and divergence times. *Mol Biol Evol* **34**, 1812-1819, doi:10.1093/molbev/msx116 (2017).
136 Survey, U. S. G. *Specimen observation data for Dreissena polymorpha (Pallas, 1771), Nonindigenous Aquatic Species Database*, < https://nas.er.usgs.gov/viewer/omap.aspx?SpeciesID=5> (2019).
137 Anisimova, M., Gil, M., Dufayard, J. F., Dessimoz, C. & Gascuel, O. Survey of branch support methods demonstrates accuracy, power, and robustness of fast likelihood-based approximation schemes. *Syst Biol* **60**, 685-699, doi:10.1093/sysbio/syr041 (2011).
138 Yonge, C. M. & Campbell, J. I. ii.—on the heteromyarian condition in the bivalvia with special reference to *Dreissena polymorpha* and certain Mytilacea. *Transactions of the Royal Society of Edinburgh* **68**, 21-42, doi:10.1017/S0080456800014502 (2012).
139 Xu, W. & Faisal, M. Putative identification of expressed genes associated with attachment of the zebra mussel (*Dreissena polymorpha*). *Biofouling* **24**, 157-161, doi:10.1080/08927010801975345 (2008).
140 Le, S. Q. & Gascuel, O. An improved general amino acid replacement matrix. *Mol Biol Evol* **25**, 1307-1320, doi:10.1093/molbev/msn067 (2008).
141 Llorens, C., Fares, M. A. & Moya, A. Relationships of gag-pol diversity between Ty3/Gypsy and Retroviridae LTR retroelements and the three kings hypothesis. *BMC Evol Biol* **8**, 276, doi:10.1186/1471-2148-8-276 (2008).

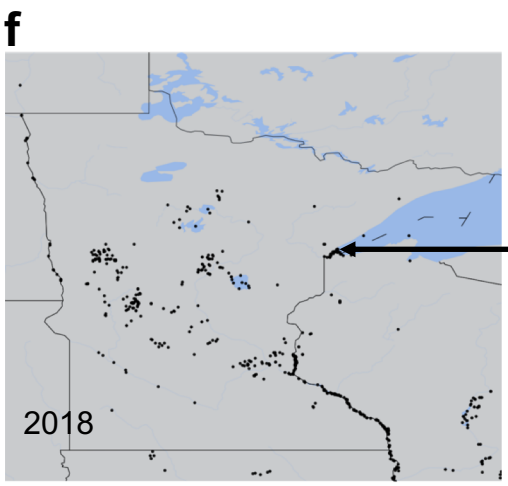

a


b
Bathymodiolus platifrons
Mytilus edulis
Azumapecten farreri
Mizuhopecten yessoensis
Crassostrea virginica
Crassostrea gigas
Pinctada martensii
Corbicula fluminea
Dreissena polymorpha
Dreissena rostriformis
Mya arenaria
Drosophila melanogaster
Caenorhabditis elegans
Homo sapiens
Mus musculus
Danio rerio
Manta birostris
Cryogenian
Ediacaran
C
O
D
MIS
P
Tr
J
K
Pg
Neo-Proterozoic
Paleozoic
Mesozoic
Cz
Proterozoic
Phanerozoic
Periods
Eras
Eons
Geologic Timescale
797
600
400
200
0
Time (MYA)


c
1988


d
2003


e
2018


f
2018

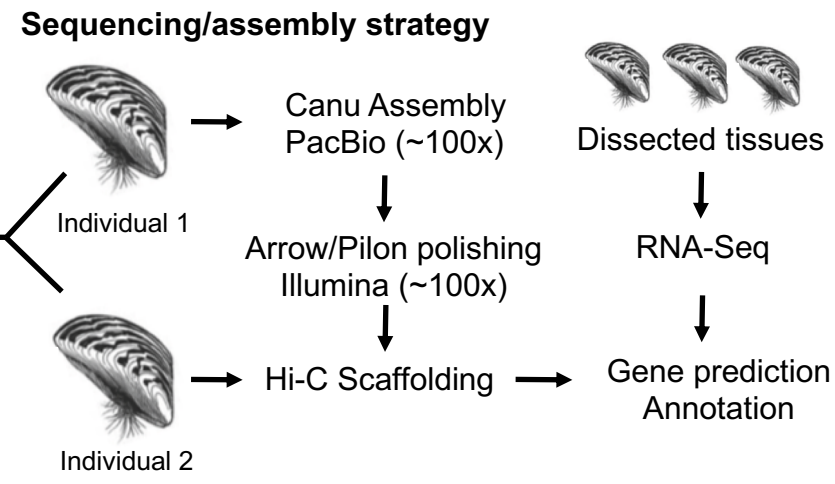

Sequencing/assembly strategy
Individual 1
Individual 2
Canu Assembly
PacBio (~100x)
Arrow/Pilon polishing
Illumina (~100x)
Hi-C Scaffolding
Dissected tissues
RNA-Seq
Gene prediction
Annotation

Figure 1

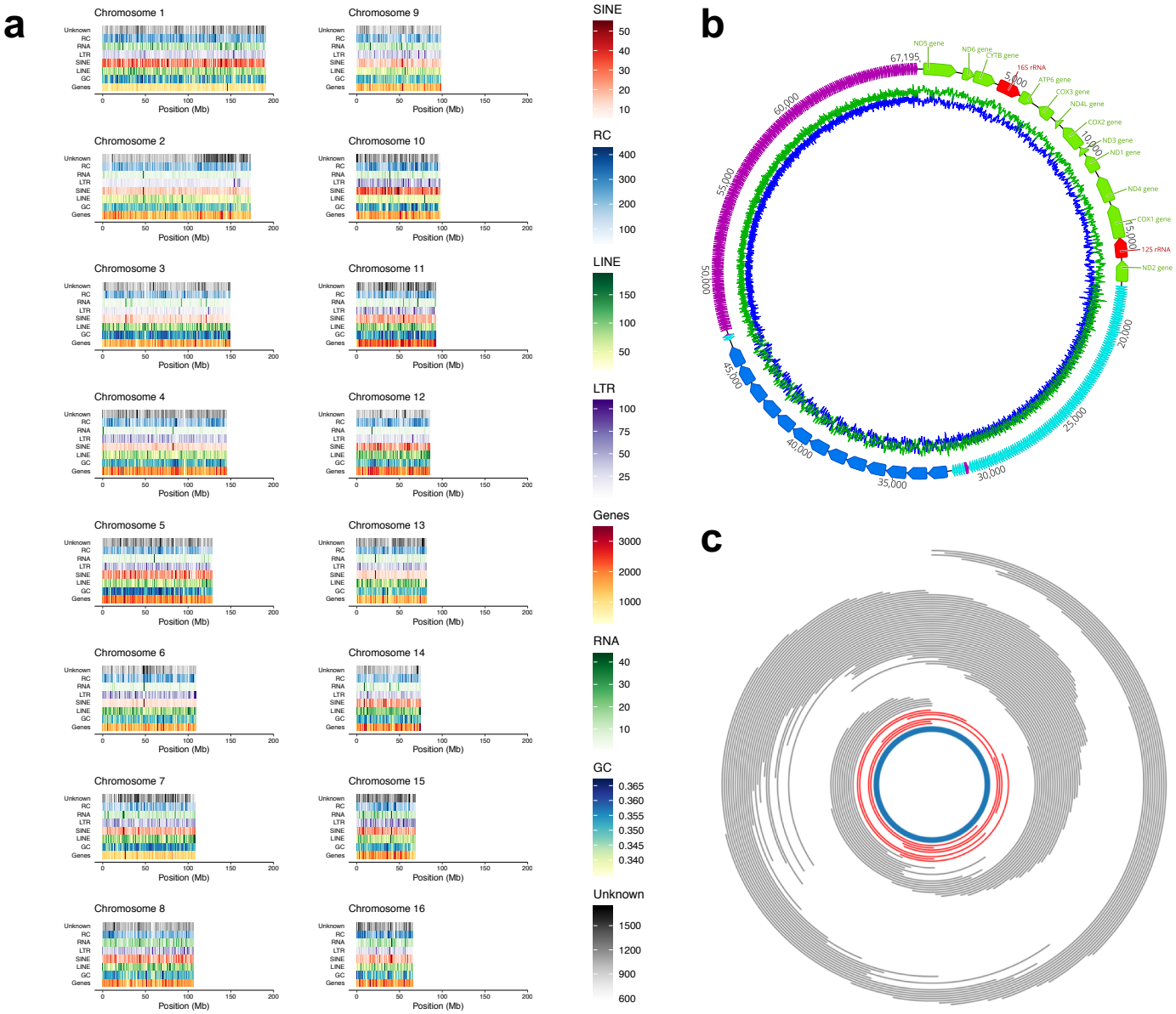

a
Chromosome 1
Chromosome 2
Chromosome 3
Chromosome 4
Chromosome 5
Chromosome 6
Chromosome 7
Chromosome 8
Chromosome 9
Chromosome 10
Chromosome 11
Chromosome 12
Chromosome 13
Chromosome 14
Chromosome 15
Chromosome 16
Unknown
RC
RNA
LTR
SINE
LINE
GC
Genes
Position (Mb)
SINE
RC
LINE
LTR
Genes
RNA
GC
Unknown
b
ND5 gene
ND6 gene
CYTB gene
16S rRNA
ATP6 gene
COX3 gene
ND4L gene
COX2 gene
ND3 gene
ND1 gene
ND4 gene
COX1 gene
12S rRNA
ND2 gene
67,195
5,000
10,000
15,000
20,000
25,000
30,000
35,000
40,000
45,000
50,000
55,000
60,000
c

Figure 2

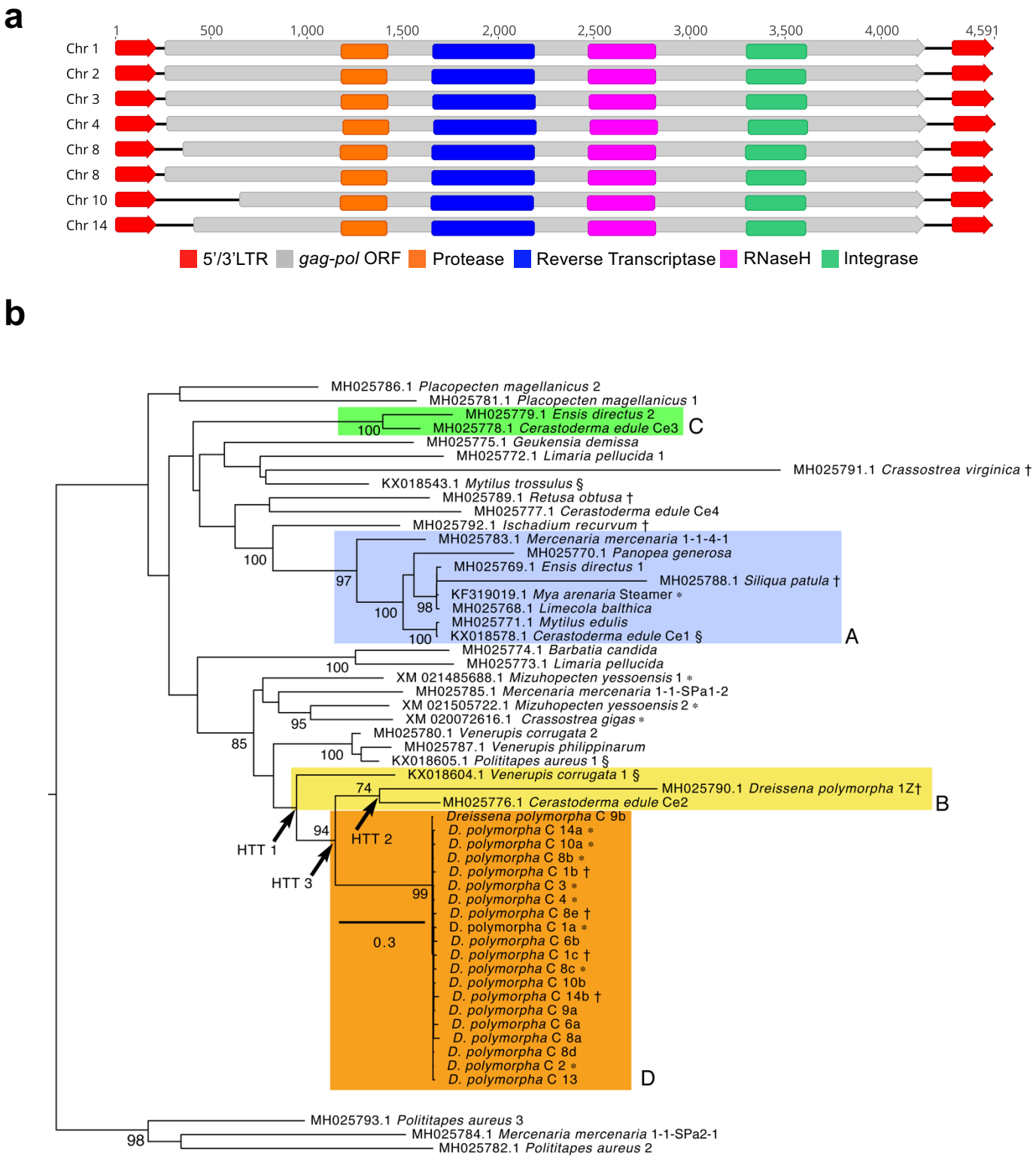
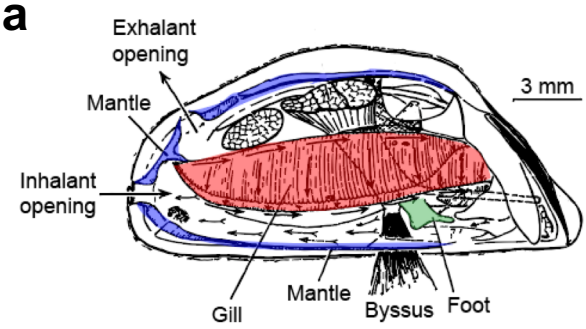
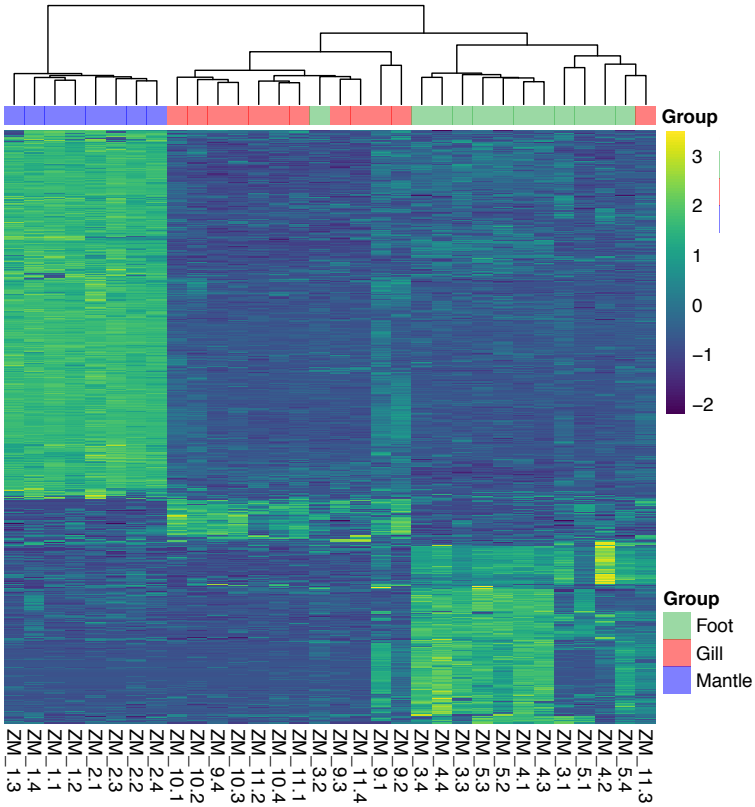

a
1
500
1,000
1,500
2,000
2,500
3,000
3,500
4,000
4,591
Chr 1
Chr 2
Chr 3
Chr 4
Chr 8
Chr 8
Chr 10
Chr 14
5'/3'LTR
gag-pol ORF
Protease
Reverse Transcriptase
RNaseH
Integrase
b
MH025786.1 Placopecten magellanicus 2
MH025781.1 Placopecten magellanicus 1
MH025779.1 Ensis directus 2
MH025778.1 Cerastoderma edule Ce3
C
MH025775.1 Geukensia demissa
MH025772.1 Limaria pellucida 1
MH025791.1 Crassostrea virginica †
KX018543.1 Mytilus trossulus §
MH025789.1 Retusa obtusa †
MH025777.1 Cerastoderma edule Ce4
MH025792.1 Ischadium recurvum †
MH025783.1 Mercenaria mercenaria 1-1-4-1
MH025770.1 Panopea generosa
MH025769.1 Ensis directus 1
MH025788.1 Siliqua patula †
KF319019.1 Mya arenaria Steamer *
MH025768.1 Limecola balthica
MH025771.1 Mytilus edulis
KX018578.1 Cerastoderma edule Ce1 §
A
MH025774.1 Barbatia candida
MH025773.1 Limaria pellucida
XM 021485688.1 Mizuhopecten yessoensis 1 *
MH025785.1 Mercenaria mercenaria 1-1-SPa1-2
XM 021505722.1 Mizuhopecten yessoensis 2 *
XM 020072616.1 Crassostrea gigas *
MH025780.1 Venerupis corrugata 2
MH025787.1 Venerupis philippinarum
KX018605.1 Polititapes aureus 1 §
KX018604.1 Venerupis corrugata 1 §
MH025790.1 Dreissena polymorpha 1Z†
MH025776.1 Cerastoderma edule Ce2
B
Dreissena polymorpha C 9b
D. polymorpha C 14a *
D. polymorpha C 10a *
D. polymorpha C 8b *
D. polymorpha C 1b †
D. polymorpha C 3 *
D. polymorpha C 4 *
D. polymorpha C 8e †
D. polymorpha C 1a *
D. polymorpha C 6b
D. polymorpha C 1c †
D. polymorpha C 8c *
D. polymorpha C 10b
D. polymorpha C 14b †
D. polymorpha C 9a
D. polymorpha C 6a
D. polymorpha C 8a
D. polymorpha C 8d
D. polymorpha C 2 *
D. polymorpha C 13
D
HTT 1
HTT 2
HTT 3
0.3
MH025793.1 Polititapes aureus 3
MH025784.1 Mercenaria mercenaria 1-1-SPa2-1
MH025782.1 Polititapes aureus 2

Figure 3

a



b



c

Mantle-specific genes

Gene ID	BLAST/Manual Annotation
DPMN_173653	Shematrin-like
DPMN_022789	Beta-hexosaminidase
DPMN_090280	Cysteine-rich receptor-like protein kinase
DPMN_173916	Sushi, von Willebrand factor type A, EGF and pentraxin domain-containing protein
DPMN_090297	Hypothetical protein
DPMN_033843	Myosin heavy chain
DPMN_014835	Shematrin-like
DPMN_086009	Cartilage matrix protein
DPMN_126617	Shematrin-like
DPMN_104020	Tyrosinase-like protein
DPMN_025985	Hypothetical protein
DPMN_048164	Shematrin-like
DPMN_074468	O-acetyltransferase
DPMN_123507	Collagen alpha-5(IV) chain
DPMN_173909	Sushi, von Willebrand factor type A, EGF and pentraxin domain-containing protein
DPMN_026871	Zinc finger protein 208-like
DPMN_124127	Kazal type serine protease inhibitor
DPMN_136646	Shematrin-like
DPMN_091682	Hypothetical protein
DPMN_015710	Hypothetical protein
DPMN_105231	CD109 antigen
DPMN_133836	Sushi, von Willebrand factor type A, EGF and pentraxin domain-containing protein
DPMN_012613	No significant similarity found
DPMN_155269	Temptin-like
DPMN_085986	VWA and chitin binding domain-containing protein 1
DPMN_063186	EF-hand calcium-binding domain-containing protein
DPMN_124170	Dentin sialophosphoprotein; Serine protease inhibitor
DPMN_116097	Tenascin-R-like
DPMN_149963	Chitinase-3
DPMN_104042	Tyrosinase-like protein
DPMN_030364	Pacifastin-related precursor
DPMN_128636	Deleted in malignant brain tumors 1
DPMN_038422	Leucine-rich repeat-containing protein

d

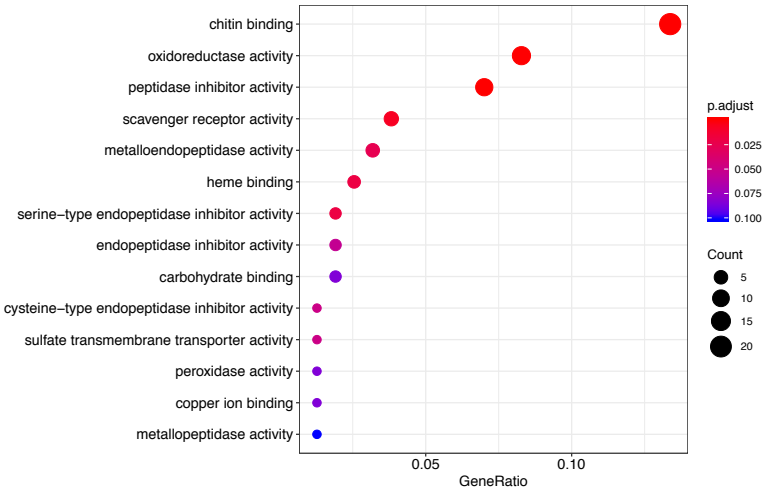


Figure 4

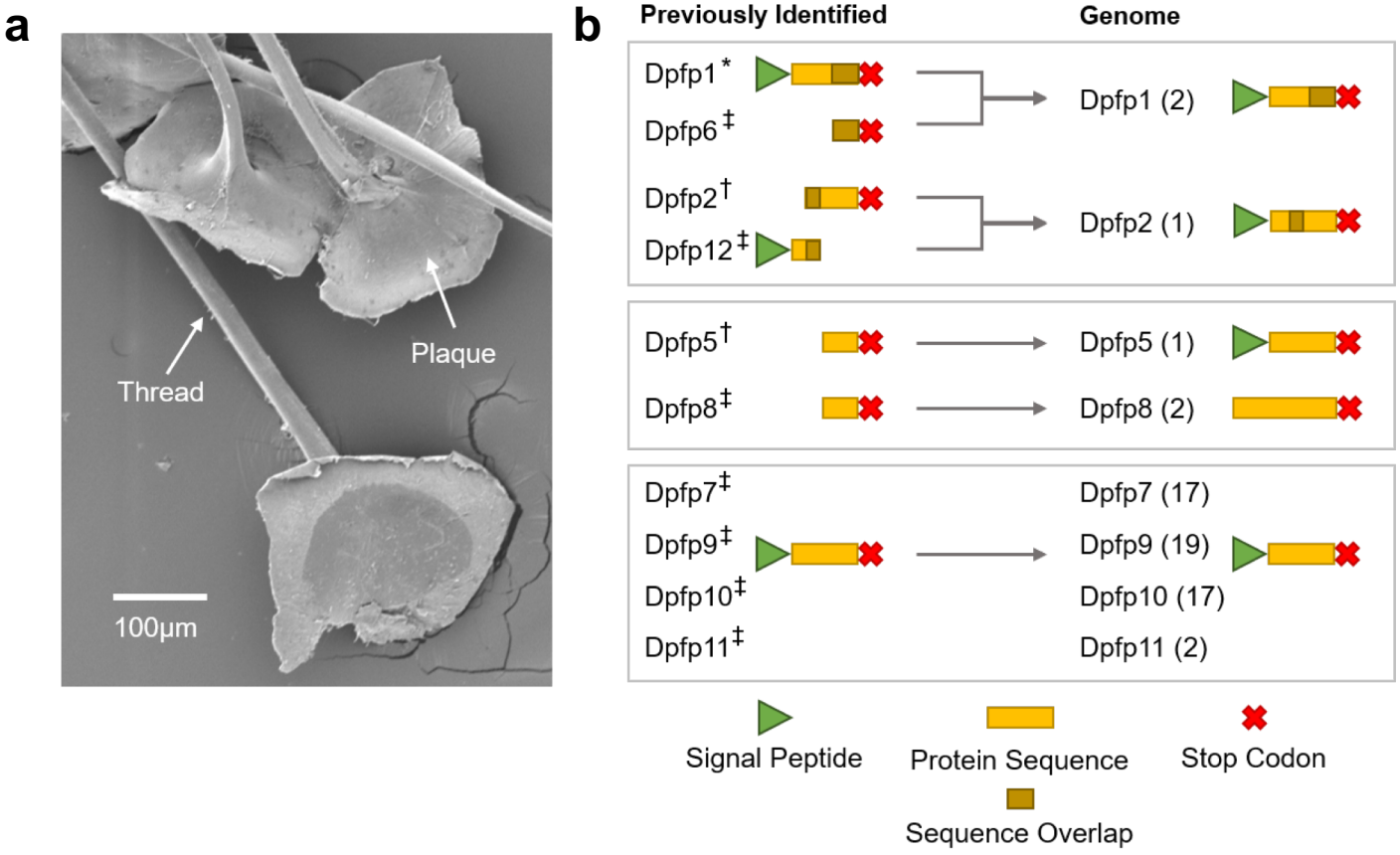

a
Thread
Plaque
100µm
b
Previously Identified
Genome
Dpfp1 *
Dpfp6 ‡
Dpfp1 (2)
Dpfp2 †
Dpfp12 ‡
Dpfp2 (1)
Dpfp5 †
Dpfp8 ‡
Dpfp5 (1)
Dpfp8 (2)
Dpfp7 ‡
Dpfp9 ‡
Dpfp10 ‡
Dpfp11 ‡
Dpfp7 (17)
Dpfp9 (19)
Dpfp10 (17)
Dpfp11 (2)
Signal Peptide
Protein Sequence
Stop Codon
Sequence Overlap

Figure 5

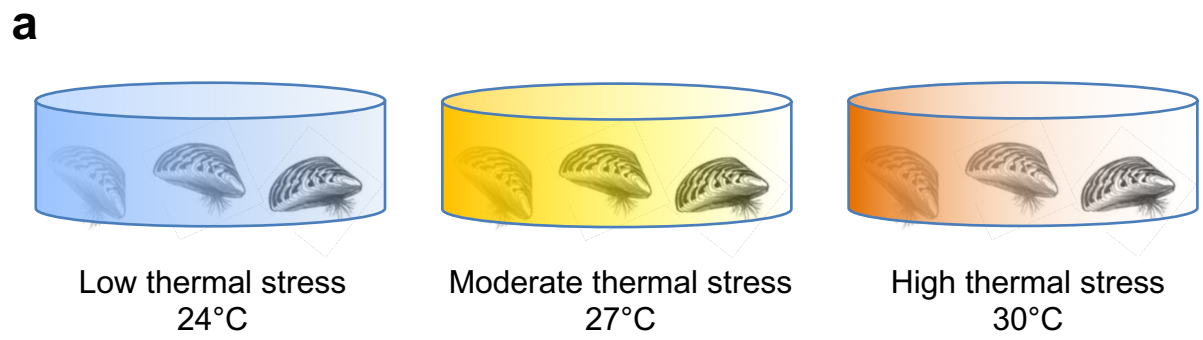

a
Low thermal stress
24°C
Moderate thermal stress
27°C
High thermal stress
30°C
b
DNAJ
Unknown
Unknown
Papalysin-1
HSP70
N-cadherin
Unknown, transmembrane
Exosome component
Unknown, transmembrane
Caspase 3
PEP7
MST1R
ZAP70
TRIM family E3 Ub ligase
Unknown
Unknown
HSP70
MYLIP E3 Ub ligase
Gelsolin
Short-chain collagen C4
0
2
4
6
8
27°C Log2 fold change
c
Fibrillin-1
HSPE1
Unknown
HSC70
Unknown, transmembrane
TTC12
Calnexin
Unknown
CHORDC1
Unknown, DNA binding
ZAP70
Unknown, transmembrane
Unknown, transmembrane
HPPD
Cytochrome P450
DNAJ
Unknown
IAP
HSP70
HSP70
0
2
4
6
8
30°C Log2 fold change
d
Med15 Superfamily domain
Calcium-activated chloride channel
Chitinase-3
Unknown
Unknown
Galectin
Unknown
NACHT domain
Dipeptidase 1
Apolipophorin
Cornifelin
Avidin
PTPR
Metalloproteinase inhibitor
ALDH9A1
MVB12B
RNA exonuclease I
Unknown
Unknown
NLR family CARD domain
-8
-6
-4
-2
0
27°C Log2 fold change
e
Zinc peptidase domain
Alpha protein kinase
Chitinase-3
Calcium-activated chloride channel
Unknown
Unknown
Unknown
Avidin
Unknown
RNA exonuclease I
Hsp70 12A/B
Cathepsin-L
Unknown
Metalloproteinase inhibitor
Apolipophorin
TNF ligand
Unknown
Unknown
Autocrine proliferation repressor A
SMARC
-8
-6
-4
-2
0
30°C Log2 fold change

Figure 6

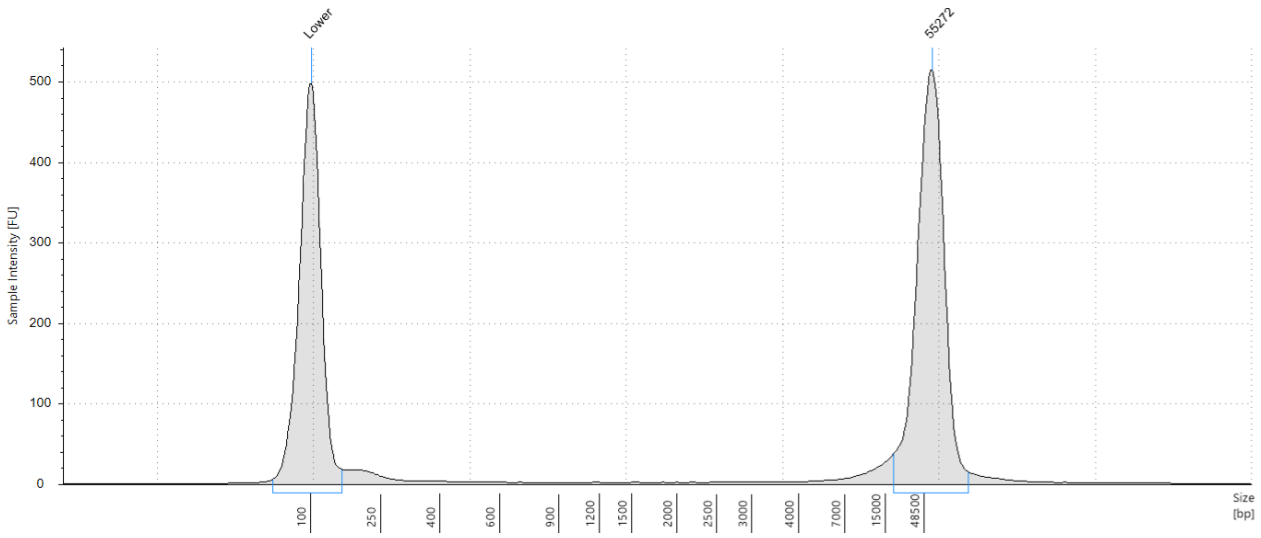
Assembly statistics	
Genome size (bp)	1,797,983,578
GC content	35.1%
Contigs	2,863
Largest contig (bp)	9,337,402
Contig N50 (bp)	1,111,027
Contig L50 (bp)	444
Scaffolds	16
Unscaffolded contigs	128
Largest scaffold (bp)	190,574,438
Scaffold N50 (bp)	107,566,931
Scaffold L50 (bp)	7
BUSCO analysis	
Complete (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	6.0%
Unclassified repeats	34.3%

Table 1

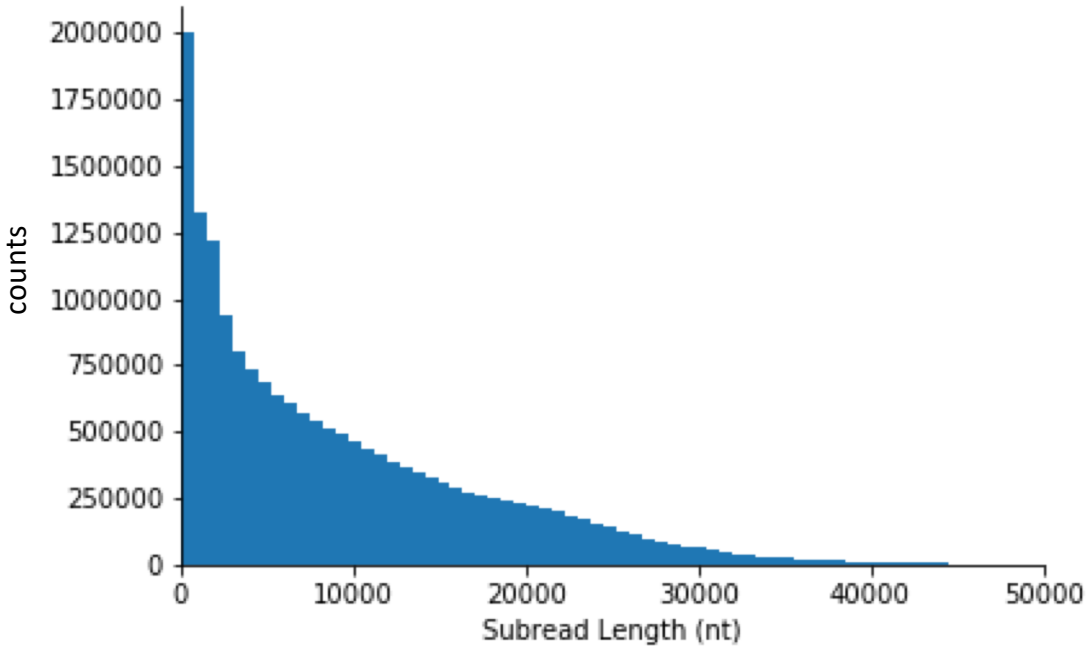
Gene	*H. sapiens*	*D. melanogaster*	*C. gigas*	*D. polymorpha*
IAP	8	4	48	167
Hsp70	17	6	88	97
Caspase	7	7	24	28
Cu-Zn SOD	1	2	6	6
Cyt. P450	57	85	136	56
C1qDC	31	0	321	50

Table 2

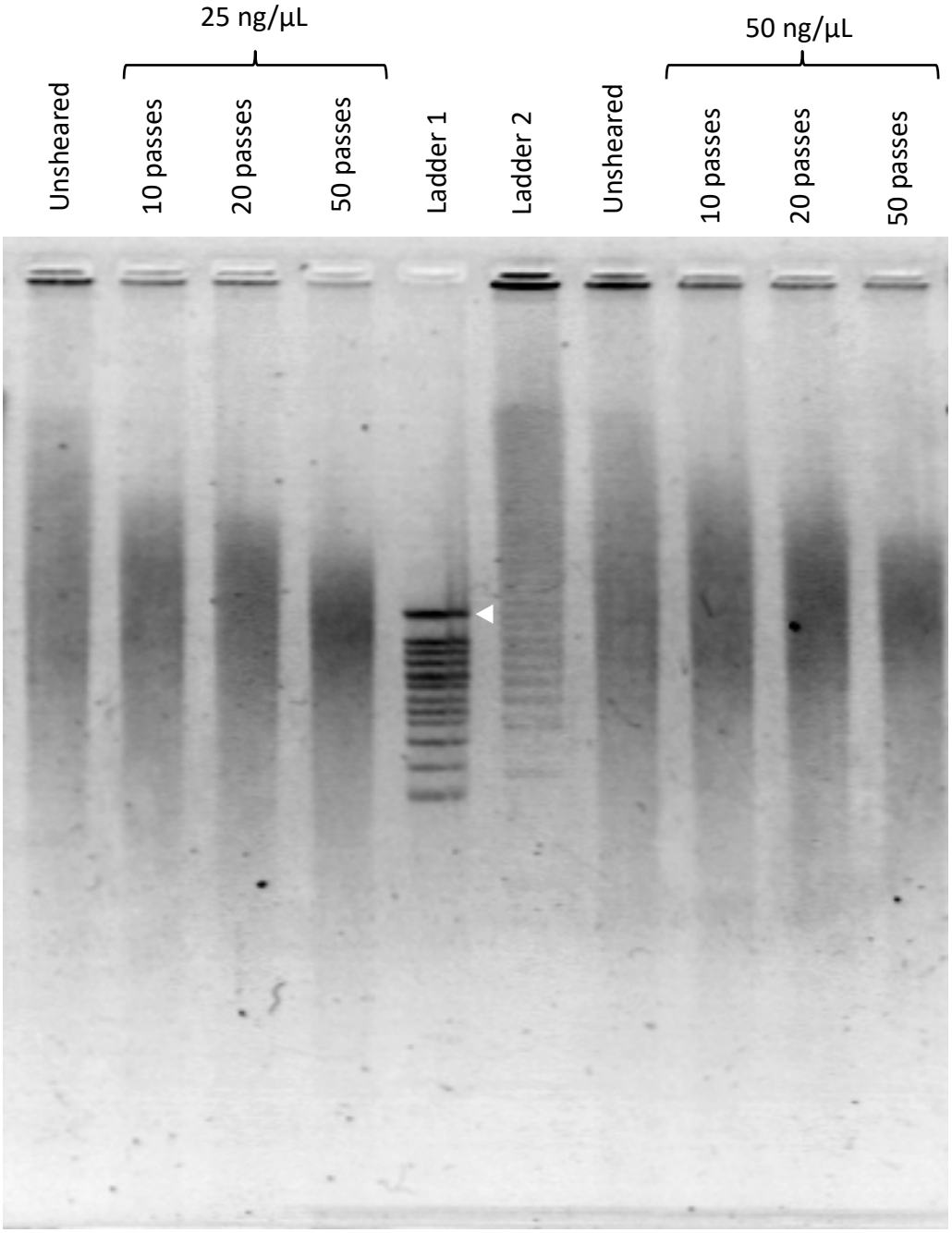
a



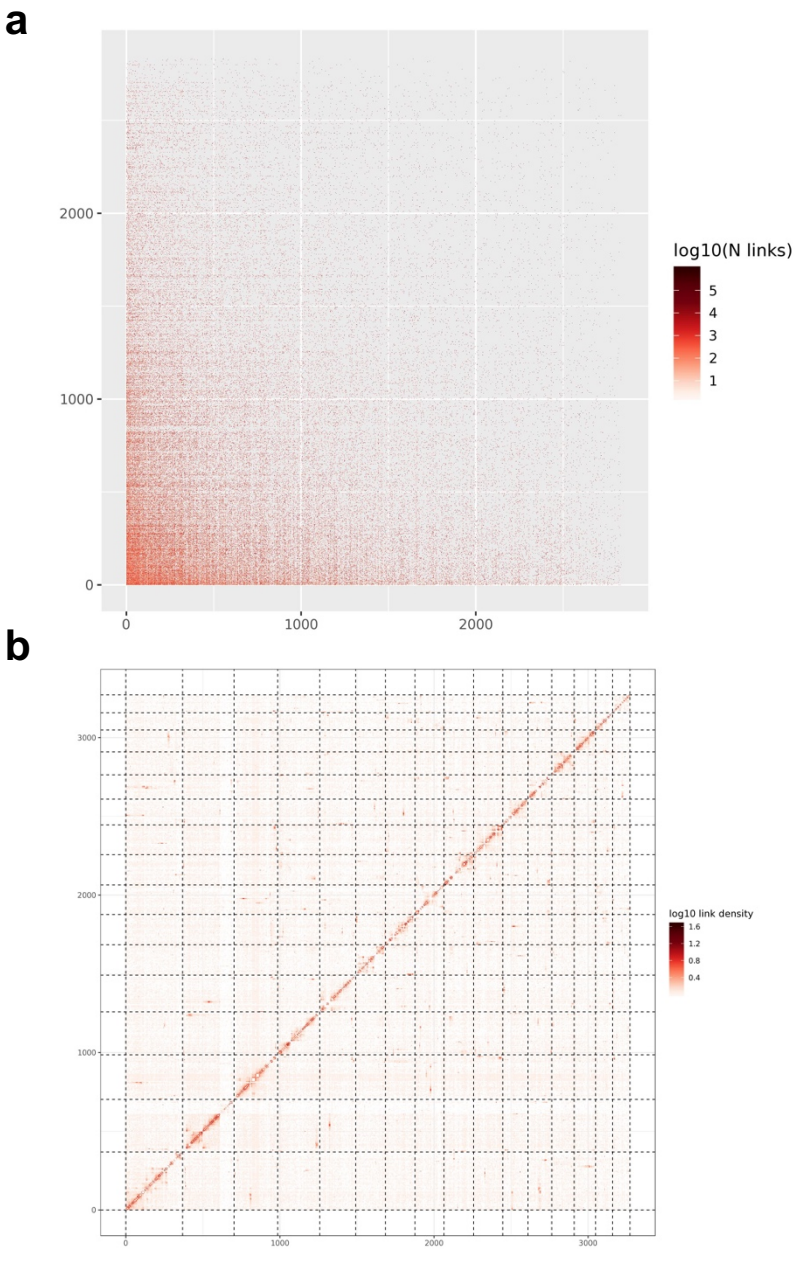
b



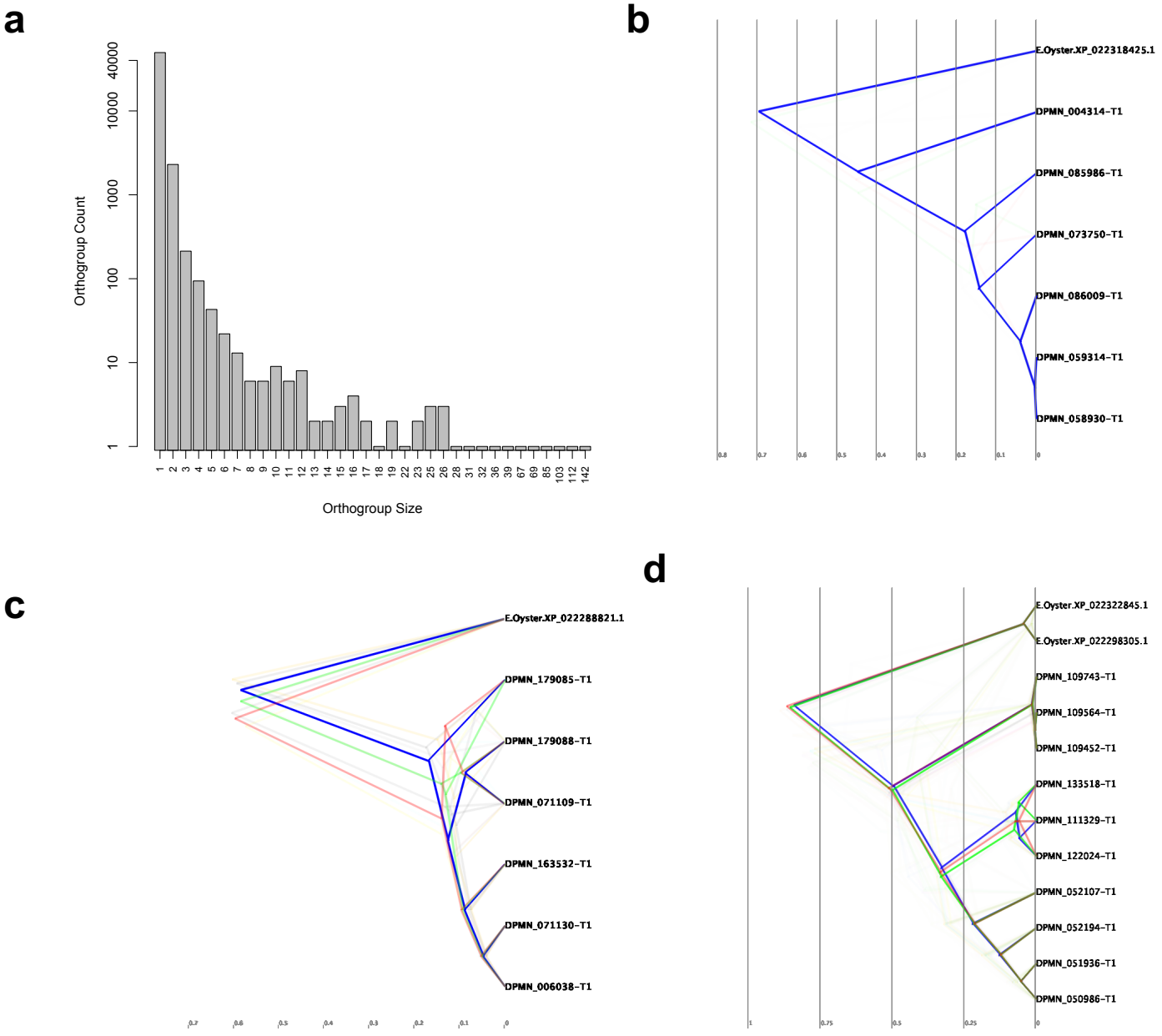
Supplemental Figure 1



Supplemental Figure 2

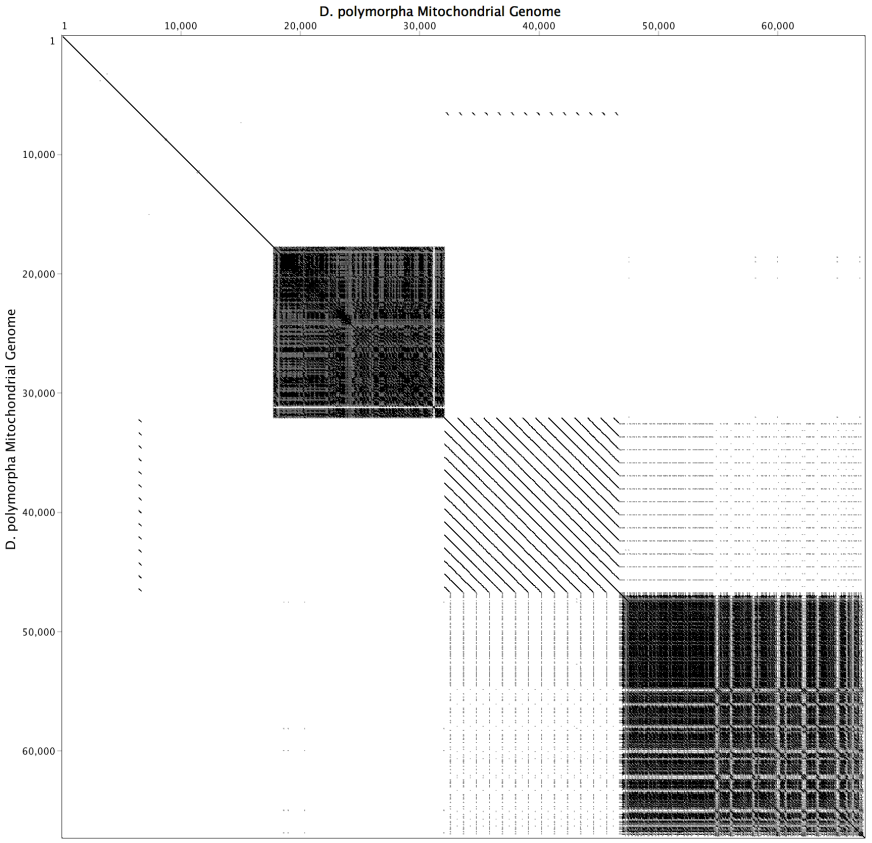


Supplemental Figure 3

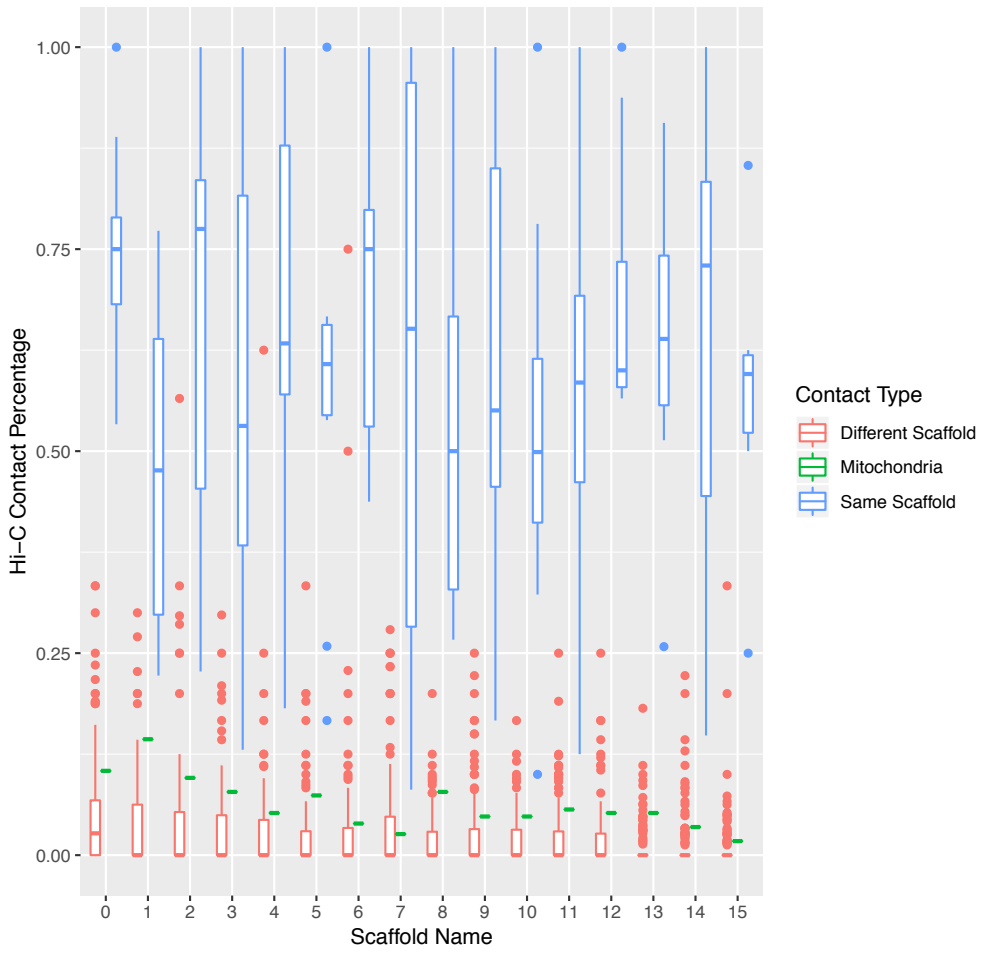


Supplemental Figure 4

a

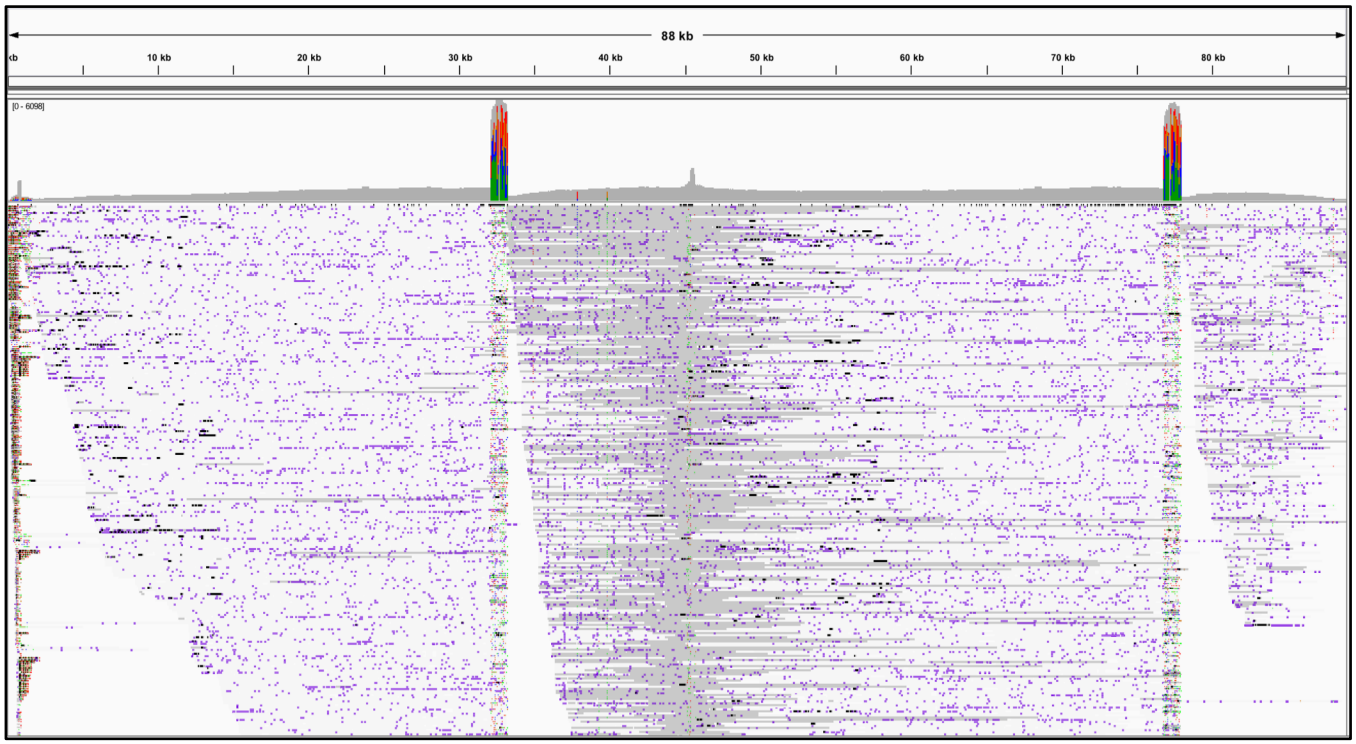


b

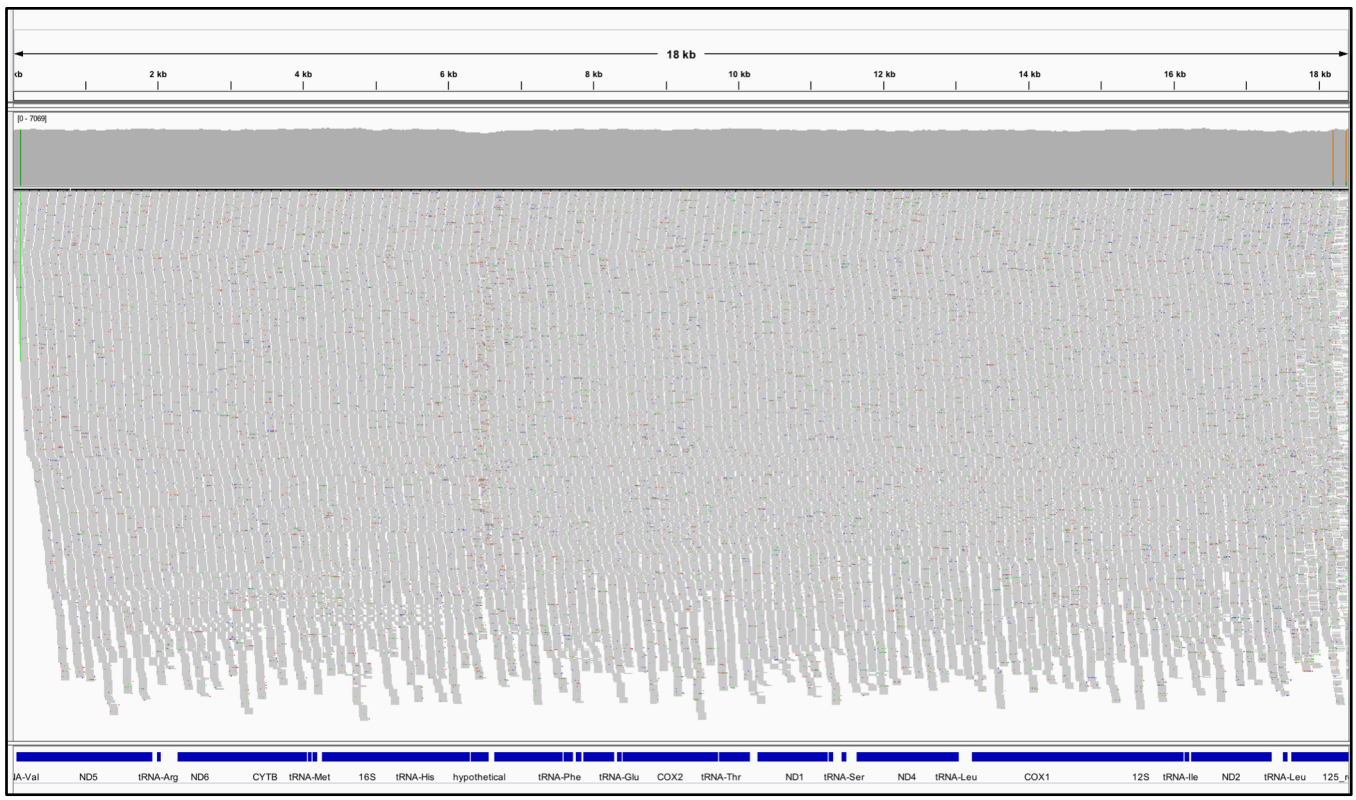


Supplemental Figure 5

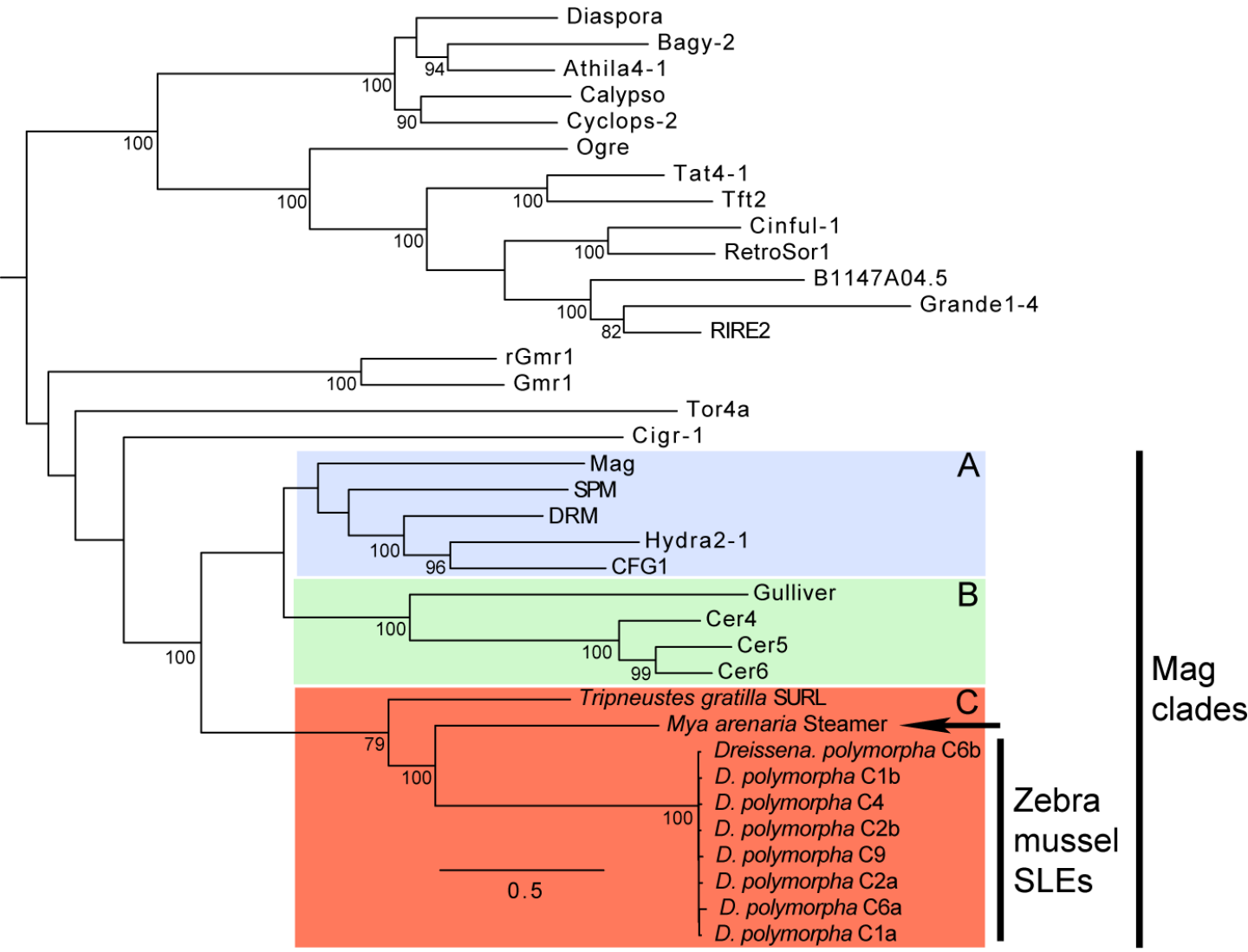
a



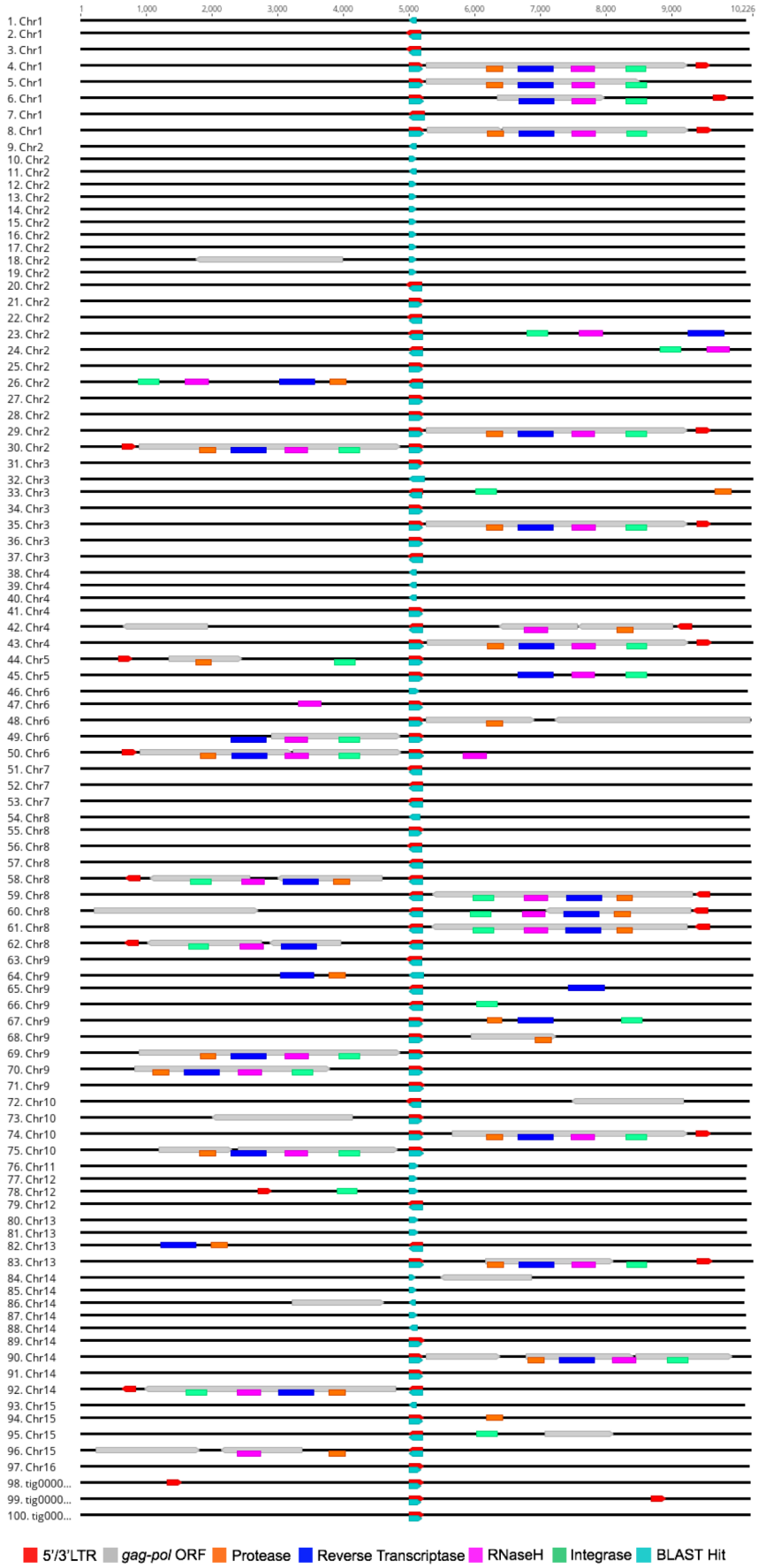
b



Supplemental Figure 6

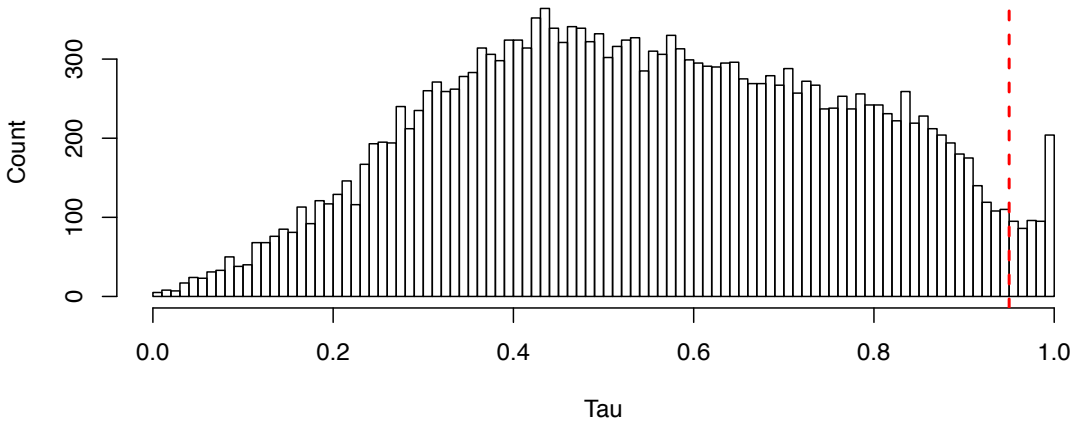


Supplemental Figure 7

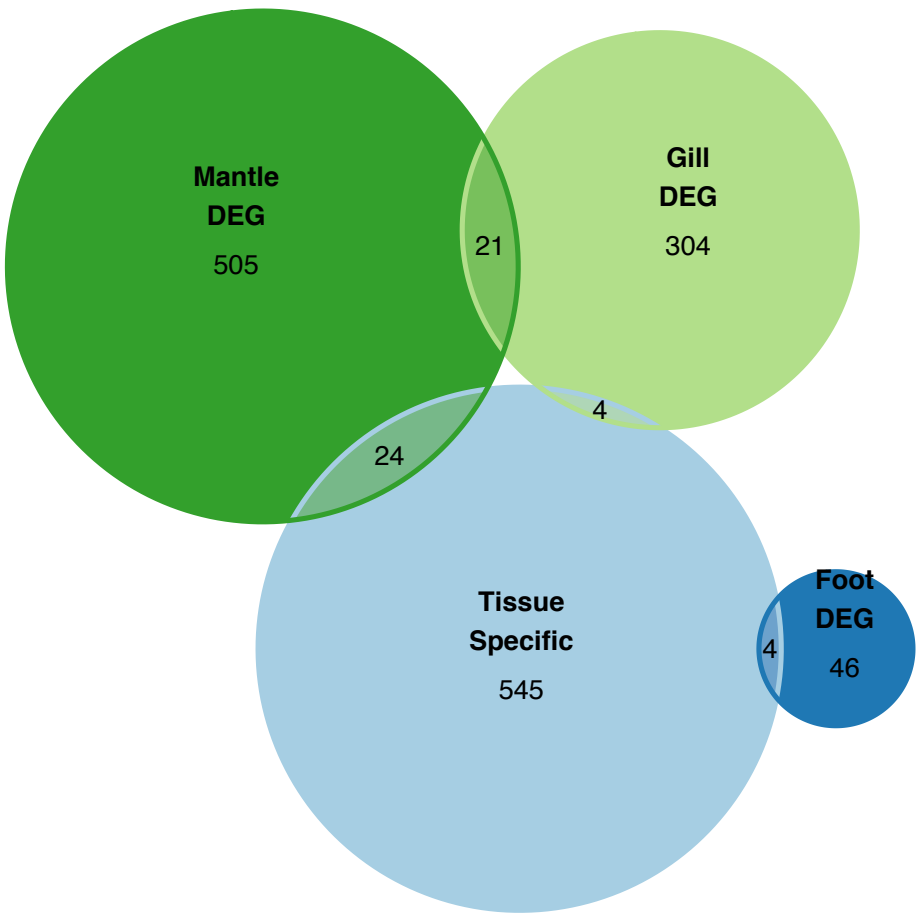


Supplemental Figure 8

a



b

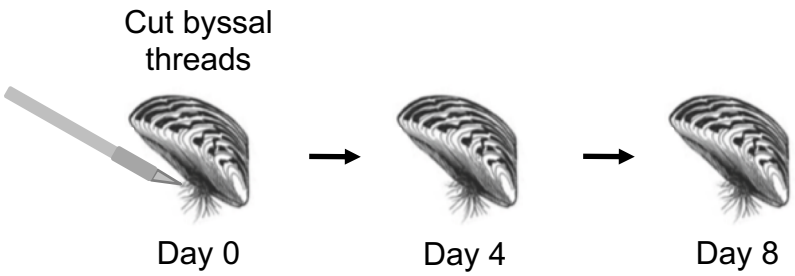


Supplemental Figure 9

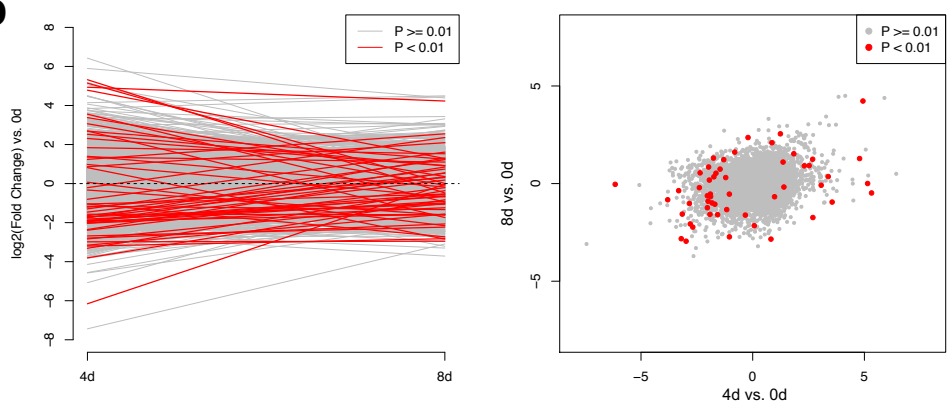
```
DPMN_136646      MANQKVALLLIGAVLFAAIGGLDAQGGGFGGGLVGGLVLGGLLEGGLGGGLGGY------   54
DPMN_014835      -MKTSIALLL-AVFAVSAVSGIAKK--GYGGGYGGGYGGGGGGYGGYGGGGGGYGGYGGG   56
DPMN_048164      ----MKSLAL-IALLVGAVVATP-H--GYGHGYGYGSGL-------YGGGGGGYGGYGG-   44
DPMN_173653      -MNTFVALAL-SCVLLSTAFAVR-K-----KGFWWGG-G-------HGGGYGYGAGYGGG   44
DPMN_173639      -MNACVALAV-SCVLLCTVFAVP-K-----KGHGYVG-G-------YGGGYGGYGGYGGS   44
DPMN_126617      -MNACVALAV-SCVLLCTVFAVP-K-----KGHGYGGGG-------YGGGYGGYGGYGGS   45
                       :* :   . . : .   :      *              *** *

DPMN_136646      ------------GYGGYGFRPPFFGGRFGYPA------------------------YGYG   78
DPMN_014835      YGGGYGGGYGGGDGGGYGGGSFYGGGGLGGGDYGGKKGCHGKHCAFGGGSGGFLGGYGGG  116
DPMN_048164      YGG---------YG--GYG-------------------------------G---LDG----   56
DPMN_173653      YGD---------DG--GYG-------------------------------AVGFAGG----   59
DPMN_173639      YGG---------GD-------------------------------------GYVGG----   54
DPMN_126617      YGG---------GD-------------------------------------GYVGG----   55
                           .

DPMN_136646      FGYPFGGRFF-------------------------------------------------   88
DPMN_014835      YGGGYGGGGYGGG-GYGGGGYGGGGYGGGYGGGEYGGGGLGGYGGGGYGGGYGGGYGGGY  175
DPMN_048164      Y-GGYGGGGYGGYGGFGGYGGYGG-Y---------------------------------   80
DPMN_173653      YGGGYGGG-------Y--GGGYGGGYDDGY-----------------------------   80
DPMN_173639      YGGGYGGGGYGGGGGYGGGGYGGGGYGVGY-----------------------------   84
DPMN_126617      YGGGYGGGGYGG-------------------------------------------------   67
                 :   :**

DPMN_136646      -------------------------------------------------------------   88
DPMN_014835      GGGGYGGKKKCTGKHCGGYGGGGYGGGGYGGGGYGGGGYGGGGYGGGGYGGGYGGGLGGY  235
DPMN_048164      ---------------------------------------------GGL---GGY-------   86
DPMN_173653      -----------------------------------GYGGGSFSGVGFGGGYGGGYGGY   103
DPMN_173639      -----------------------------------GYGGGFDGGVGFGGGDGGYGGGF   107
DPMN_126617      -------------------------------------------------GGYGGGF   74


DPMN_136646      -------------------------------------------------------------   88
DPMN_014835      GGGSYGLGGGGGLGGGGGGFGGVGGFGGSFGHTRQCKCAKRCNKWQKFVGRCPWCQKKCK  295
DPMN_048164      -G-------------GYGGFGHGHGHGWWPHHK---KWRP---------------------  109
DPMN_173653      GG-------------GYGGFGWGN---W--GIRAQCKCAPRCGKFQKYVGRCPWCKKGCK  145
DPMN_173639      GG-------------GYGGYGWGG---KKPTMRAQCKCVKKCGKFQKYVGRCPWCKKGCK  151
DPMN_126617      GG-------------GYGGYGWGG---KKPTMRAQCKCVPKCGKFQKYVGRCPWCKKGCK  118


DPMN_136646      ----------     88
DPMN_014835      LVFCCNRGKY     305
DPMN_048164      ------RKWY     113
DPMN_173653      LVFCCRRKW-     154
DPMN_173639      LVFCCRRKWP     161
DPMN_126617      LVFCCRRKWP     128
```

Supplemental Figure 10

a



b



c

Up-regulated genes

Gene ID	BLAST Annotation	Log2FC (4d)
DPMN_066420	Polyadenylate-binding protein 1	5.32
DPMN_017751	Tax1BP1	5.15
DPMN_005570	Tetraspanin	4.93
DPMN_049419	Vitellogenin	4.78
DPMN_134996	Glutamyl aminopeptidase-like	3.56
DPMN_013285	Monocarboxylate transporter 12	3.38
DPMN_034632	CD109	3.07
DPMN_125763	CD109	2.69
DPMN_011081	Tenascin-X	2.68
DPMN_080226	Metabotropic glutamate receptor	2.53
DPMN_194568	60S ribosomal protein L19	2.31
DPMN_066988	Serine/threonine-protein kinase	1.84
DPMN_013817	Kyphoscoliosis peptidase	1.40
DPMN_027392	Anoctamin-4-like	1.36
DPMN_121885	Arylsulfatase	1.24
DPMN_164933	Cytochrome P450 3A28-like	0.98
DPMN_127641	Cytosolic Fe-S cluster assembly	0.86
DPMN_053135	Ceroid-lipofuscinosis; neuronal 5	0.82
DPMN_122749	Lysocardiolipin acyltransferase 1	0.07

Down-regulated genes

Gene ID	BLAST Annotation	Log2FC (4d)
DPMN_181702	Nucleoside diphosphate kinase	-6.16
DPMN_092058	No similarity found	-3.80
DPMN_114486	Metalloreductase STEAP2-like	-3.31
DPMN_103363	Alpha-tubulin	-3.20
DPMN_028019	von Willebrand factor D and EGF	-3.15
DPMN_084917	AMBP	-2.98
DPMN_186960	Toll-like receptor 4	-2.84
DPMN_030517	von Willebrand factor D and EGF	-2.79
DPMN_118567	Cytokine receptor	-2.68
DPMN_187139	CXorf38 homolog	-2.39
DPMN_114830	Serine/threonine phosphatase	-2.35
DPMN_106288	No similarity found	-2.03
DPMN_097756	No similarity found	-2.03
DPMN_085219	Complement factor B-like protein	-2.00
DPMN_062091	Solute carrier family 23 member	-1.97
DPMN_028623	Matrix metallopeptidase 10	-1.94
DPMN_084246	TNF ligand-like	-1.91
DPMN_004625	DUF4921-domain-containing	-1.90
DPMN_010384	Major facilitator super domain	-1.88
DPMN_102295	Bcl2-like 1	-1.82

Supplemental Figure 11

Species	Family	Common name	Commercial interest	Assembly level	Number of scaffolds	Number of contigs	Contig N50 (bp)	Genome length (Mb)	Reference
Bathymodiolus platifrons	Mytilidae	Hydrothermal vent mussel	None	Scaffold	65,662	272,497	12,602	1,658.2	Sun et al. 2017
Chlamys farreri	Pectinidae	Zhikong (Chinese) scallop	Wild harvest and culture	Scaffold	96,024	148,999	21,500	779.9	Li et al. 2017
Crassostrea gigas	Ostreidae	Pacific oyster	Hatchery culture—leads aquatic animals in global harvest	Scaffold	7,659	30,460	31,239	557.7	Zhang et al. 2012
Crassostrea virginica	Ostreidae	Eastern oyster	Wild harvest and hatchery culture	Chromosome	11	669	1,971,208	684.7	Gómez-Chiarri et al. 2015
Mizuhopecten (Patinopectin) yessoensis	Pectinidae	Yesso scalllop	Culture from wild seed	Scaffold	82,659	120,022	65,014	987.6	Wang et al. 2017
Modiolus philippinarum	Mytilidae	Phillipine horse mussel	None	Scaffold	74,573	301,873	18,389	2,629.6	Sun et al. 2017
Pinctada martensii	Pteriidae	Akoya pearl oyster	Cultured pearls	Chromosome	5,039	85,944	21,518	991.0	Unpublished

Supplemental Table 1

```
==================================================
file name: PGA_Assembly.FINAL.fasta
sequences:           145
total length: 1798019516 bp  (1797746116 bp excl N/X-runs)
GC level:         35.14 %
bases masked:  852870596 bp ( 47.43 %)
==================================================
               number of      length   percentage
               elements*    occupied  of sequence
--------------------------------------------------
SINEs:             50552     11629297 bp    0.65 %
      ALUs             0            0 bp    0.00 %
      MIRs         12195      1557593 bp    0.09 %

LINEs:            115027     77616133 bp    4.32 %
      LINE1         4532      2311960 bp    0.13 %
      LINE2         9224      4723805 bp    0.26 %
      L3/CR1        6363      4054095 bp    0.23 %

LTR elements:      38917     27955936 bp    1.55 %
      ERVL            35         8358 bp    0.00 %
      ERVL-MaLRs       1          120 bp    0.00 %
      ERV_classI     827       507019 bp    0.03 %
      ERV_classII    397        26602 bp    0.00 %

DNA elements:     244050     80624865 bp    4.48 %
     hAT-Charlie     773       239831 bp    0.01 %
     TcMar-Tigger    635       120753 bp    0.01 %

Unclassified:     2068170    617626932 bp   34.35 %

Total interspersed repeats:815453163 bp   45.35 %


Small RNA:          5444       693184 bp    0.04 %

Satellites:         8601      1869040 bp    0.10 %
Simple repeats:   316718     37443122 bp    2.08 %
Low complexity:    29370      1463637 bp    0.08 %
==================================================
```

Supplemental Table 2