Title

Hybrid *de novo* assembly of the draft genome of the freshwater mussel *Venustaconcha ellipsiformis* (Bivalvia: Unionida).

Authors

Sébastien Renaut[1,2,8], Davide Guerra[3], Walter R. Hoeh[4], Donald T. Stewart[5], Arthur E. Bogan[6], Fabrizio Ghiselli[7], Liliana Milani[7], Marco Passamonti[7], Sophie Breton[2,3,8]

Affiliations

[1] Département de Sciences Biologiques, Institut de Recherche en Biologie Végétale, Université de Montréal, Montréal, QC, Canada

[2] Quebec Centre for Biodiversity Science, Montréal, QC, Canada

[3] Département de Sciences Biologiques, Université de Montréal, Montréal, QC, Canada

[4] Department of Biological Sciences, Kent State University, Kent, OH, USA

[5] Department of Biology, Acadia University, Wolfville, NS, Canada

[6] North Carolina Museum of Natural Sciences, Raleigh, NC, USA

[7] Dipartimento di Scienze Biologiche, Geologiche ed Ambientali, University of Bologna, Bologna, Italy

[8] corresponding authors: sebastien.renaut@umontreal.ca, s.breton@umontreal.ca

Abstract

Freshwater mussels (Bivalvia: Unionida) serve an important role as aquatic ecosystem engineers but are one of the most critically imperilled groups of animals. An assembled and annotated genome for freshwater mussels has the potential to be utilized as a valuable resource for many researchers given their ecological value and threatened status. In addition, a sequenced genome will help to answer more fundamental questions of sex-determination and genome evolution in bivalves exhibiting a unique “doubly uniparental inheritance” mode of mitochondrial DNA transmission through comparative genomics approaches. Here, we used a combination of sequencing strategies to assemble and annotate a draft genome of the freshwater mussel *Venustaconcha ellipsiformis*. The genome described here was obtained by combining high coverage short reads (65X genome coverage of Illumina paired-end and 11X genome coverage of mate-pairs sequences) with low coverage Pacific Biosciences long reads (0.3X genome coverage). Briefly, the final scaffold assembly accounted for a total size of 1.54Gb (366,926 scaffolds, N50 = 6.5Kb, with 2.3% of "N" nucleotides), representing 93% of the predicted genome size of 1.66Gb. Over one third of the genome (37.5%) consisted of repeated elements and more than 85% of the core eukaryotic genes were recovered. Finally, we reassembled the full mitochondrial genome and found six polymorphic sites with respect to the previously published reference. This resource opens the way to comparative genomics studies to identify genes related to the unique adaptations of freshwater mussels and their distinctive mitochondrial inheritance mechanism.

Keywords: genome assembly, annotation, High Throughput Sequencing, Freshwater Mussels, Unionida

Introduction

Through their water filtration action, freshwater mussels (Bivalvia: Unionida) serve important roles as aquatic ecosystem engineers (Gutiérrez et al. 2003; Spooner & Vaughn 2006), and can greatly influence species composition (Aldridge et al. 2007). From a biological standpoint, they are also well known for producing obligate parasitic larvae that metamorphose on freshwater fishes (Lopes-Lima et al. 2014), for being slow-growing and long-lived, with several species reaching >30 years old and some species >100 years old (see Haag & Rypel 2011 for a review), and for exhibiting an unusual system of mitochondrial transmission called Doubly Uniparental Inheritance or DUI (see Breton et al. 2007; Passamonti & Ghiselli 2009; Zouros 2013) for reviews). From an economic perspective, freshwater mussels are also exploited to produce cultured pearls (Haag 2012). Regrettably however, habitat loss and degradation, overexploitation, pollution, loss of fish hosts, introduction of non-native species, and climate change have resulted in massive freshwater mussel decline in the last decades (reviewed in Lopes-Lima et al. 2017; 2018). For example, more than 70% of the ~300 North American species are considered endangered at some level (Lopes-Lima et al. 2017).

While efforts are currently underway to sequence and assemble the genome of the marine mussel *Mytilus galloprovincialis* (Murgarella et al. 2016), genomic resources for mussels in general are still extremely scarce. In addition to *M. galloprovincialis*, the genomes of two other marine mytilid mussel species, i.e. the deep-sea vent/seep mussel *Bathymodiolus platifrons* and the shallow-water mussel

Modiolus philippinarum have recently been published (Sun et al. 2017). In all cases, genomes have proven challenging to assemble due to their large size (~1.6 to 2.4Gb) and widespread presence of repeated elements (~30% of the genome, and up to 62% of the genome for the shallow-water mussel *Modiolus philippinarum*, Sun et al. 2017). For example, the *Mytilus* genome remains highly fragmented, with only 15% of the gene content estimated to be complete (Murgarella et al. 2016). With respect to freshwater mussels (order Unionida), no nuclear genome draft currently exists. An assembled and annotated genome for freshwater mussels has the potential to be utilized as a valuable resource for many researchers given the biological value and threatened features of these animals. Such studies are needed to help identifying genes essential for survival (and/or the genetic mechanisms that led to decline) and ultimately for developing monitoring tools for endangered biodiversity and plan sustainable recoveries (Pavey et al. 2016; Savolainen et al. 2013). In addition, a sequenced genome will help answer more fundamental questions of sex-determination (Breton et al. 2011; 2017) and genome evolution through comparative genomics approaches (e.g. Sun et al. 2017).

Given the challenges in assembling a reference genome for saltwater mussels (Sun et al. 2017; Murgarella et al. 2016), we used a combination of different sequencing strategies (Illumina paired-end and mate pair libraries, Pacific Biosciences long reads, and a recently assembled reference transcriptome (Capt et al. 2018) to assemble the first genome draft in the family Unionidae. Hybrid sequencing technologies using long read−low coverage and short read−high coverage offer an affordable strategy with the advantage of assembling repeated regions of the genome (for which short reads are ineffective) and circumventing the relatively higher error rate of long reads (Koren et al. 2012; Miller et al. 2017). Here, we present a *de novo* assembly and annotation of the genome of the freshwater mussel *Venustaconcha ellipsiformis*.

Methods

To determine the expected sequencing effort to assemble the *Venustaconcha ellipsiformis* genome, i.e., the necessary software and computing resources required, we first searched for C-values from other related mussel species. C-values indicate the amount of DNA (in picograms) contained within a haploid nucleus and is roughly equivalent to genome size in megabases. Two closely related freshwater mussel species (*Elliptio* sp., c-value = 3; *Uniomerus* sp., c-value = 3.2), in addition to two other well studied mussel groups (*Mytilus* spp., c-value = 1.3-2.1; *Dreissena polymorpha*, c-value = 1.7) were identified on the Animal Genome Size Database (http://www.genomesize.com). As such, we estimated the *Venustaconcha* genome size to be around ~1.5-3.0Gb, and this originally served as a coarse guide to determine the sequencing effort required, given that when the sequencing for *Venustaconcha* was originally planned, no mussel genome had yet been published.

Mussel specimen sampling, genomic DNA extraction and library preparation

Adult specimens of *Venustaconcha ellipsiformis* were collected from Straight River (Minnesota, USA; Lat 44.006509, Long -93.290899) and sexed by microscopic examination of gonad smears. Gills were dissected from a single female individual and genomic DNA was extracted using a Qiagen DNeasy Blood & Tissue Kit (QIAGEN Inc., Valencia, CA, USA) using the animal tissue protocol. The quality and quantity of DNA, respectively, were assessed by electrophoresis on 1% agarose gel and with a BioDrop mLITE spectrophotometer (a total of 15 µg of DNA was quantified using the spectrophotometer). For whole genome shotgun sequencing and draft genome assembly, we used two sequencing platforms: Illumina (San Diego, CA) Hiseq2000 and Pacific Biosciences (Menlo Park, CA) PacBio RSII. First, three paired-end libraries with insert size of 300b were constructed using Illumina TruSeq DNA Sample Prep Kit. One mate pair library with insert sizes of about 5Kb was constructed for scaffolding process using Illumina Nextera mate-pair library construction protocol. For high-quality genome assembly, Pacific Biosciences system was employed for final scaffolding process using long reads. Pacific Biosciences long reads (>10Kb) were generated using SMRT bell library preparation protocol (ten SMRT cells were sequenced). Construction of sequencing libraries and sequencing analyses were performed at the Genome Quebec Innovation Centre (McGill University, Qc, Canada).

Pre-processing of sequencing reads

We quality trimmed paired-end and mate-pair reads using TRIMMOMATIC 0.32 (Bolger et al. 2014) with the options ILLUMINACLIP:TruSeq3-PE.fa:2:30:10 LEADING:3 TRAILING:3 SLIDINGWINDOW:6:10 MINLEN:36. This allowed removal of base pairs below a threshold Phred score of three at the leading and trailing end, in addition to removing base pairs based on a sliding window calculation of quality (mininum

Phred score of ten over six base pairs). Finally, if trimmed reads fell below a threshold length (36b), both sequencing pairs were removed. We verified visually the quality (including contamination with Illumina paired-end adaptors) before and after trimming using FASTQC (Andrews 2010). This allowed us to only keep high quality reads prior to the assembly steps.

Following quality trimming, we used BFC (Li & Durbin 2009) to perform error correction for the Illumina paired-end sequencing data. BFC suppresses systematic sequencing errors, which helps to improve the base accuracy of the assembly and reduce the complexity of the *de Bruijn* graph based assembly, described below.

Corrected paired-end reads were subsequently used to identify the optimal K value that provides the most distinct genomic k-mers using KMERGENIE v1.7016 (Chikhi & Medvedev 2014). We tested k = 10 to 100, in incremental steps of 10, and we then refined the interval from 20 to 40, in incremental steps of 2 to get a more precise estimate of K. Based on the best K value (k=42), KmerGenie was also used to estimate genome size.

Genome assembly strategy

We used ABYSS 2.0 (Jackman et al. 2017), a modern genome assembler specifically built for large genomes and reads acquired by different sequencing strategies. ABYSS 2.0 works similarly to ABYSS (Simpson et al. 2009), by using a distributed *de Bruijn* graph representation of the genome, therefore allowing parallel computation of the assembly algorithm across a network of computers. In addition, the software makes use of long sequencing reads (Illumina mate-pair libraries and Pacific BioSciences long reads) to bridge gaps and scaffold contigs. Yet, as memory requirements and computing time scale up exponentially with genome size, for large genomes (>1Gb), these rapidly become very large (>100GB of RAM) and unpractical. Consequently, Jackman et al. (2017) introduced ABYSS 2.0, which employs a probabilistic data structure called a Bloom filter (Bloom 1970) to store a *de Bruijn* graph representation of the genome and, consequently, greatly reduces memory requirements and computing time. The Bloom filter allows removing from memory the majority of nearly identical k-mers likely caused by sequencing errors, as k-mers with an occurrence count below a user-specified threshold are discarded. The caveat is that it can generate false positive extension of contigs, but through optimization, this can be kept well below 5%, and in fact, false positives can be corrected later on in the assembly step (Jackman et al. 2017).

In the current study, we combined different types of high throughput sequencing to aid in assembling the genome (**Table 1**). ABYSS 2.0 (Jackman et al. 2017) performs a first genome assembly step without using the paired-end information, by extending unitigs until either they cannot be unambiguously extended or come to an end due to a lack of coverage (*uncorrected unitigs*). This first *de Bruijn* graph representation of the genome is further cleaned of vertices and edges created by sequencing errors (*unitigs*). Paired-end information is then used to resolve ambiguities and merge *contigs*. Following this, mate-pairs are mapped onto the assembly to create *scaffolds*, and finally long reads (Pacific Biosciences long reads) and the *Venustaconcha* reference transcriptome from Capt et al. (2018) were also mapped onto the assembly to create *long-scaffolds*. This reference transcriptome was assembled from a pool of sequences coming from four different male and female individuals and further details are provided in Capt et al. (2018). Although ideally sequencing information would all come from a single individual, the current study design did not allow for this. In addition, given that coding sequences are conserved compared to non-coding regions, it remains highly valuable to use a transcriptome in a *de novo* genome assembly.

We ran the ABYSS 2.0 assembly stage (abyss-bloom-dbg) with a k-mer size of 41 (ABYSS requires an odd number k-mer), a Bloom filter size of 24GB, 4 hash functions and a threshold of k-mer occurrence set at 3. These parameters were chosen after performing several test assemblies, in order to minimize the false positive rate (<5%), maximize the N50 of the assembly and keep the virtual memory (95GB) and CPU (24 CPUs) requirements within a reasonable computational limit for our resources. In addition, we adjusted parameters at the mapping stage to create contigs, scaffolds and long-scaffolds to maximize N50 (overlap required in re-alignments, distance between mate-pairs, nb reads aligned to support assembly, see pipeline available at https://github.com/seb951/venustaconcha_ellipsiformis_genome).

Genome completeness was assessed using BUSCO 3.0.2 (Benchmarking Universal Single-Copy Orthologs, Simao et al. 2015). Briefly, BUSCO uses curated lists of known core single copy orthologs to produce evolutionarily-informed quantitative measures of genome completeness (Simao et al. 2015). Here, we tested both the eukaryotic (303 single copy orthologs) and metazoan (978 single copy orthologs) gene lists to assess the completeness of our genome assembly.

Characterization of repetitive elements

Given that repetitive elements can occupy large proportions of a genome, the characterization of their proportion and composition is an essential step during genome annotation. RepeatModeler open-1.0.10 (Smit & Hubley 2015) was used to create an annotated library of repetitive elements contained in the *Venustaconcha* genome assembly (excluding sequences <1Kb). Then, with RepeatMasker open-4.0.7 (Smit et al. 2015), we extracted libraries of repetitive elements for the taxa “Bivalvia” and “Mollusca” from the RepeatMasker combined database (comprising the databases Dfam_consensus-20170127 and RepBase-20170127) using built-in tools. Sequences classified as “artefact” were removed from the last two libraries before the subsequent steps. The three libraries were used alone and/or in combination (except for the Mollusca+Bivalvia combination) to mask the cut-down assembly again with RepeatMasker, specifying the following options: -nolow (to avoid masking low complexity sequences, which may enhance subsequent exon annotation), -gccalc (to calculate the overall GC percentage of the input assembly), -excln (to exclude runs of ≥20 Ns in the assembly sequences from the masking percentage calculations). Option -species was used to specify the taxon for the runs with Bivalvia and Mollusca libraries, while option -lib used to specify the *Venustaconcha* library and the combined ones.

Results summaries for the latter three runs were refined with the RepeatMasker built-in tools. Linear model fit for genome size and repeats content for all available bivalve genomes were calculated with R version 3.1.0 (R Core Team 2012), using the highest masking value found for *Venustaconcha* .

Genome annotation

We used QUAST (Gurevich et al. 2013) to calculate summary statistics on the genome assembly. In addition, QUAST uses a Hidden Markov Model to identify putative genes in the final assembly (GLIMMERHMM Majoros et al. 2004). Following this, we translated Open Reading Frames identified in the annotation files into protein sequences using BEDTOOLS V2.27.1 (Quinlan & Hall 2010) and EMBOSS TRANSEQ V6.6.0 (Rice et al. 2000) bioinformatics pipelines. These were then compared against the manually curated UniProt database (556,388 reference proteins, downloaded January 11th 2018, e-value cut-off of 10^{-5}) using BLASTp (Altschul et al. 1990). These steps were done on the long-scaffolds assembly, the masked long-scaffolds assembly (with low complexity regions replaced with N), in addition to the broken long-scaffolds assembly (scaffolds broken into smaller contigs by QUAST, based on long stretches of N nucleotides).

Mitochondrial genome

Given the rare mode of mitochondrial inheritance of freshwater mussels and therefore its evolutionary importance, we first aimed to check if the mitochondrial female genome had been properly assembled. Using BLASTn (Altschul et al. 1990) with high stringency (E value <1e-50), we identified a fragmented mitochondrial genome. We then created a mt specific dataset containing 1,396,004 sequence reads by aligning paired-end reads to the reference mt genome of Breton et al. (2009) (GenBank Acc. No. FJ809753) using SAMTOOLS V1.3.1 and BEDTOOLS V2.27.1 (Li et al. 2009; Quinlan & Hall 2010). We then rebuilt the mt genome *de novo* using ABYSS 2.0, testing different k-mers (17-45). In addition, we aligned reads to the reference transcriptome using BWA V0.7.12-R1039 (H Li & Durbin 2009) and identified Single Nucleotide Polymorphisms (SNPs) with respect to the reference mt genome using SAMTOOLS and BCFTOOLS v1.3.1 (Li et al. 2009).

Results and Discussion

We generated 564M paired-end reads (2 X 100b) representing an average 65X coverage of the genome (**Table 1**). This was complemented by 98M mate-pairs (5Kb insert, 11X average genome coverage) and 103,000 Pacific Biosciences long reads (0.3X average genome coverage), and a recently published reference transcriptome comprised of 285,000 contigs (Capt et al. 2018). Filtering and trimming the raw paired-end and mate-pair sequences removed about 5% of the total base pairs from further analyses, indicating that the quality of the raw sequences was high (**Table 1**). K-mer analysis indicated that the number of unique k-mers peaked at 42 and predicted a genome assembly size of 1.66Gb (**Figure 1**), smaller than predicted genome size according to C-value for other Unionida, but in general agreement with the recent draft genome of the marine mussel *Mytilus galloprovincialis* (1.6Gb) and the deep-sea vent/seep mussel (*Bathymodiolus platifrons*, 1.64Gb).

Running the ABySS 2.0 assembly stage (abyss-bloom-dbg) led to a low False Positive Rate (<0.05%). The N50 for the contig assembly was 3.2Kb with 551,875 contigs (discarding contigs <1Kb, given that small contigs likely represent artefacts and provide little information for the overall genome assembly (Pavey et al. 2016; Murgarella et al. 2016, see **Table 2**). Once these were corrected and paired-end, mate-pairs and long read information were added, the scaffolds N50 increased to 5.5Kb, with 2.3% of nucleotides represented as "N" (see **Table 2** for the summary statistics and **Table 3** for overall genome assembly statistics acquired from QUAST analysis). Adding the Pacific Biosciences long reads only slightly improved the scaffolds N50 (from 5.5 to 5.7Kb, **Table 2**) and slightly decreased the number of *long-scaffolds* >1Kb (from 423,853 to 410,237), likely because our long read coverage was quite low (0.3X, **Table 1**). In addition, it is also possible that the more error prone Pacific Biosciences sequences, compared to Illumina paired-end reads, reduced their usability (Miller et al. 2017). Once the reference transcriptome was added, it improved the N50 to 6.5Kb, and substantially decreased the number of long-scaffolds to 366,926. This final long-scaffold assembly accounted for a total size of 1.54Gb (with 2.3% of "N" nucleotides) and represented 93% of the predicted genome size of 1.66Gb. Yet, it remained highly fragmented (366,926 scaffolds, **Table 2**). Genome annotation statistics can also be viewed in html format and downloaded here: https://github.com/seb951/venustaconcha_ellipsiformis_genome/tree/master/annotation_quast_v3

While assembly numbers (N50, number of scaffolds, etc.) are not directly comparable with other recently published genomes given the diversity of sequencing approaches (Illumina, 454, Sanger, PacBio), library types, sequencing depth and unique nature of the genome themselves, they can give a broad perspective of the inherent difficulties of assembling large genomes. The best comparison is probably with the saltwater mussel, *Mytilus galloprovincialis*, giving their similar genome size (1.6Gb for *Mytilus* vs 1.66Gb for *Venustaconcha*) and Illumina paired-end sequencing approaches (32X for *Mytilus* vs 65X for *Venustaconcha*). While the *Mytilu*s genome project (Murgarella et al. 2016) did not utilize mate-pair libraries or Pacific Bioscience long reads, they did make use of sequencing libraries with varying insert sizes (180, 500 and 800b). As such, they obtained a genome assembly quality relatively similar to ours and consisting of 393 thousand scaffolds (>1Kb), with however a substantially lower N50 (2.6Kb compared to 6.5Kb for *Venustaconcha*). The recently reported genome for the deep-sea vent/seep mussel *Bathymodiolus platifrons* (1.64Gb) made use of nine Illumina sequencing libraries with varying insert sizes (180 to 16Kb) and an overall coverage of >300X. With this very thorough sequencing approach, the scaffold N50 obtained was substantially higher (343.4Kb), but again the genome remained highly fragmented, into >65 thousands scaffolds. As exemplified here, high coverage sequencing libraries with varying insert sizes have become a broadly used approach for large and complex genomes. In fact, it is implemented by default in many genome assembly platforms (e.g. SoapdeNovo2, Luo et al. 2012, ALLPATHS-LG, Gnerre et al. 2011). In the future, these libraries will likely be useful to further assemble the *Venustaconcha* genome, at least until these approaches are superseded by affordable, error free, single molecule long read sequencing (Gordon et al. 2016; Badouin 2017) or mapping approaches that allow reaching chromosome level assemblies such as optical mapping (e.g. Bionano Genomics, San Diego, CA).

Results of the BUSCO (Simao et al. 2015) analyses showed that 664 (68%) of the 978 core metazoan genes (CEGs) were considered complete in our assembly. When the BUSCO analysis was extended to include also fragmented matches, 871 (89%) proteins aligned. Results were similar when compared against the 303 core eukaryotic genes (61% complete, 86% complete or fragmented, **Table 4**). When compared to the previously published reference transcriptome for *Venustaconcha* *ellipsiformis* (Capt *et al*. 2018), we found fewer complete genes, but also fewer duplicated genes (97.5% complete, and 24% duplicated in the reference transcriptome, compared to 68.1% complete and 1% duplicated here). This likely reflects the fact that the reference transcriptome is nearly complete, while the current reference genome is still fragmented. However, the reference transcriptome also likely contains multiple isoforms of the same genes, in addition to possible nematode contaminating sequences, despite the authors' best efforts to minimize these problems. Previously analysed molluscan genomes of similar size (Murgarella et al. 2016; Sun et al. 2017) have found that 16% (*Mytilus galloprovincialis*, 1.6Gb), 25% (pearl oyster *Pinctada fucata*, 1.15Gb), 36% (California sea hare *Aplysia californica*, 1.8Gb) of the core eukaryotic genes were complete. For their part Sun and collaborators (2017), identified 96% of the core metazoan genes to be partial or complete in the deep-sea vent/seep mussel *Bathymodiolus platifrons* (1.6Gb), again reflecting that the depth and type of sequencing, in addition to the idiosyncrasies of each genome, can have considerable influence on the end results.

The custom *Venustaconcha* repeat library created *de novo* with RepeatModeler contained 2,068 families, the majority of them (1,498, 72.44% of the total) classified as “unknown”. The genome masking performed with the Bivalvia and Mollusca libraries had scarce performances (masking 2.38% and 2.59%, respectively; details in **Supplementary Table RM1**), possibly because of the phylogenetic distance between *V. ellipsiformis*, which belongs to the early-branching bivalve lineage of Palaeoheterodonta, and the other bivalve and mollusk species represented in the database as well as their relative number of sequences. The custom *Venustaconcha* library masked 37.17% of the genome, while the combined *Venustaconcha*+Bivalvia masked 37.69% of the genome and the *Venustaconcha*+Mollusca reached 37.81%, the highest masking percentage (**Supplementary Table RM2**). After refining, these raw values slightly decreased to respectively 36.29%, 36.80%, and 36.91% (**Supplementary Table RM3**). All these latter values of repeat content fall in the 32-39% range (the median for all species is 37%) where six out of the nine sequenced bivalve species lie, irrespective of their genome size (*M. philippinarum* and *R. philippinarum* are the furthest from this interval) (**Table 5** and **Supplementary Figure 1**). Although the number of species sequenced up to now is still low, this observation indicates that repetitive elements may contribute differently to the total genome size among the different bivalve taxa: indeed, the correlation between genome size and repeats content is weak (**Supplementary Figure 1**). In both the *ab initio* masking with the *Venustaconcha* library and the two combined ones, most of the identified repeats are categorized as “unknown” (22.8% of the assembly), followed by retroelements (LINEs 2.9%, LTR elements 2.3-2.4%, and SINEs 1.7%, for a total of 6.9% of the assembly) and DNA elements (5.4-5.6% of the assembly) (**Supplementary Table RM3**). Direct comparisons of these values with other species should be performed with caution, as the usually large “unclassified” portion of repeats might contain species-specific variants of known elements (Murgarella et al. 2016) that may therefore change the relative weight of each category on the total.

QUAST was used to calculate summary statistics and identify putative genes in the final assembly using a hidden markov model (**Table 3**). Following this, 29,031; 14,195 and 25,544 Open Reading Frames were annotated using BLASTp against UniProt database in the long-scaffolds, broken and masked long-scaffolds assemblies, respectively.

Freshwater mussels, marine mussels, as well as marine clams are the only known exception in the animal kingdom with respect to the maternal inheritance of mitochondrial DNA (see Breton et al. 2007 for a review). Their unique system, characterized by the presence of two gender-associated mitochondrial DNA lineages, has therefore attracted studies to better understand mitochondrial inheritance and the evolution of mtDNA in general. Using BLASTN, we recovered 53 contigs matching to the 15,975b female reference mt genome from Breton *et al.* (2009), indicating that the mt genome was highly fragmented and likely improperly assembled with our current approach, much like what was found in the *Mytilus galloprovincialis* genome draft of Murgarella (Murgarella et al. 2016). As such, we created a dataset of mt specific sequences that could be aligned to the mt genome (1,396,004 reads). This mt specific dataset was then re-assembled *de novo*, using different k-mers (17-45). Using a k-mer similar or larger to the one used in the overall assembly (k≥41) resulted in a failed assembly (no contigs created, data not shown), while using a k-mer <21 generated a highly fragmented mt genome (data not shown). Using a k-mer between 21 and 39 generated one large contig of 16,024b comprising the entire mitogenome, with a 42b insertion in the 16S ribosomal RNA. Given the different rate of evolution of mtDNAs, it is likely that assembly parameters we used for the whole genome were not appropriate for the *V. ellipsiformis* female mt genome. Finally, we also re-aligned the mt specific dataset to the original mt genome of Breton et al. (2009) and found high coverage (mean = 7,256X, SD = 682) for most positions, while for three regions coverage dropped below 300X (**Figure 2**). Six SNPs with respect to the reference were also identified, indicating possible polymorphism, or sequencing error in the original mt reference genome (**Figure 2**).

Conclusion

High throughput sequencing has the power to produce draft genomes that were only reserved to model systems ten years ago. Here we report the first *de novo* draft assembly of the *Venustaconcha ellipsiformis* genome, a freshwater mussel from the bivalve order Unionida. Our assembly covers over 93% of the genome and contains nearly 90% of the core eukaryotic orthologs, indicating that it is nearly complete. However, as for other mussel genomes recently published, our genome remains fragmented, showing the limits of high throughput sequencing and the necessity to combine different sequencing approaches to augment the scaffolding and overall genome quality, especially when a large fraction of the genome is comprised of repetitive elements. In the future, the *Venustaconcha* genome will benefit from a larger number of long read sequences, varying library size for paired-end sequencing, and the use of genetic, physical or optimal maps to subsequently order scaffolded contigs into pseudomolecules or chromosomes.

Abbreviations

BLAST: Basic Local Alignment Search Tool

b: base pairs

Kb: Kilobases

M: Million

Gb: Gigabases

GB: gigabytes

CPU: Central Processing Unit

DNA: Deoxyribonucleic acid

LINEs: Long interspersed elements

LTR: Long terminal repeats

ORF: Open Reading Frames

N80/50/20: weighted median statistic such that 80/50/20% of the entire assembly is contained in contigs/scaffolds equal to or larger than this value.

L50 = minimum number of sequences required to represent 50% of the entire assembly

RAM: Random Access Memory

SINEs: Short interspersed elements

Data availability

Supporting data for this Genome Report will be made available on datadryad.org

Raw sequences are available in the SRA database with number SRP132483 (submission SUB3624229 to be release upon publication) and Bioproject accession PRJNA433387. All scripts used in the analyses are available on github (https://github.com/seb951/venustaconcha_ellipsiformis_genome).

Acknowledgments

Computations were made on the supercomputer briaree from Université de Montréal, managed by Calcul Québec and Compute Canada. The operation of this supercomputer is funded by the Canada Foundation for Innovation (CFI), the ministère de l'Économie, de la science et de l'innovation du Québec (MESI) and the Fonds de recherche du Québec - Nature et technologies (FRQ-NT).

References

Aldridge DC, Fayle TM, Jackson N. 2007. Freshwater mussel abundance predicts biodiversity in UK lowland rivers. Aquatic Conserv: Mar. Freshw. Ecosyst. 17:554–564. doi: 10.1002/aqc.815.

Altschul SF, Gish W, Miller W, Myers EW, Lipman, DJ, 1990. Basic local alignment search tool. Journal of molecular biology, 215(3), pp.403-410.

Andrews S. 2010. FastQC: a quality control tool for high throughput sequence data. httpswww.bioinformatics.babraham.ac.ukprojectsfastqc.

Badouin H. 2017. The sunflower genome provides insights into oil metabolism, flowering and Asterid evolution. Nature. 1–20.

Bloom BH. 1970. Space/time trade-offs in hash coding with allowable errors. Communications of the ACM. 13:422–426. doi: 10.1145/362686.362692.

Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 30:2114–2120. doi: 10.1093/bioinformatics/btu170.

Breton S et al. 2009. Comparative Mitochondrial Genomics of Freshwater Mussels (Bivalvia: Unionoida) With Doubly Uniparental Inheritance of mtDNA: Gender-Specific Open Reading Frames and Putative Origins of Replication. Genetics. 183:1575–1589. doi: 10.1534/genetics.109.110700.

Breton S et al. 2011. Novel Protein Genes in Animal mtDNA: A New Sex Determination System in Freshwater Mussels (Bivalvia: Unionoida)? Mol. Biol. Evol. 28:1645–1659. doi: 10.1093/molbev/msq345.

Breton S, Beaupre HD, Stewart DT, Hoeh WR, Blier PU. 2007. The unusual system of doubly uniparental inheritance of mtDNA: isn't one enough? Trends Genet. 23:465–474. doi: 10.1016/j.tig.2007.05.011.

Breton S, Capt C, Guerra D, Stewart D. 2017. Sex Determining Mechanisms in Bivalves. Preprints. 1–23. doi: 10.20944/preprints201706.0127.v1.

Capt C et al. 2018. Deciphering the Link between Doubly Uniparental Inheritance of mtDNA and Sex Determination in Bivalves: Clues from Comparative Transcriptomics. Genome Biology and Evolution. 10:577–590. doi: 10.1093/gbe/evy019.

Chikhi R, Medvedev P. 2014. Informed and automated k-mer size selection for genome assembly. Bioinformatics. 30:31–37. doi: 10.1093/bioinformatics/btt310.

Gnerre S et al. 2011. High-quality draft assemblies of mammalian genomes from massively parallel sequence data. PNAS. 108:1513–1518. doi: 10.1073/pnas.1017351108.

Gordon D et al. 2016. Long-read sequence assembly of the gorilla genome. Science. 352:aae0344–aae0344. doi: 10.1126/science.aae0344.

Gurevich A, Saveliev V, Vyahhi N, Tesler G. 2013. QUAST: quality assessment tool for genome assemblies. Bioinformatics. 29:1072–1075. doi: 10.1093/bioinformatics/btt086.

Gutiérrez JL, Jones CG, Strayer DL, Iribarne OO. 2003. Mollusks as ecosystem engineers: the role of shell production in aquatic habitats. Oikos. 101:79–90. doi: 10.1034/j.1600-0706.2003.12322.x.

Haag WR. 2012. *North American freshwater mussels: natural history, ecology, and conservation*.

Haag WR, Rypel AL. 2011. Growth and longevity in freshwater mussels: evolutionary and conservation implications. Biol Rev. 86:225–247. doi: 10.1111/j.1469-185X.2010.00146.x.

Jackman SD et al. 2017. ABySS 2.0: resource-efficient assembly of large genomes using a Bloom filter. Genome Res. 27:768–777. doi: 10.1101/gr.214346.116.

Koren S et al. 2012. Hybrid error correction and de novo assembly of single-molecule sequencing reads. Nat Biotechnol. 30:693–700. doi: 10.1038/nbt.2280.

Li H et al. 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 25:2078–2079. doi: 10.1093/bioinformatics/btp352.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 25:1754–1760. doi: 10.1093/bioinformatics/btp324.

Li Y et al. 2017. Scallop genome reveals molecular adaptations to semi-sessile life and neurotoxins. Nature Communications. 1–11. doi: 10.1038/s41467-017-01927-0.

Lopes-Lima M et al. 2014. Biology and conservation of freshwater bivalves: past, present and future perspectives. Hydrobiologia. 735:1–13. doi: 10.1007/s10750-014-1902-9.

Lopes-Lima M et al. 2018. Conservation of freshwater bivalves at the global scale: diversity, threats and research needs. Hydrobiologia. 1–14. doi: 10.1007/s10750-017-3486-7.

Lopes-Lima M et al. 2017. Conservation status of freshwater mussels in Europe: state of the art and future challenges. Biol Rev. 92:572–607. doi: 10.1111/brv.12244.

Luo R et al. 2012. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. Gigascience. 1. doi: 10.1186/2047-217X-1-18.

Majoros WH, Pertea M, Salzberg SL. 2004. TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders. Bioinformatics. 20:2878–2879. doi: 10.1093/bioinformatics/bth315.

Miller JR et al. 2017. Hybrid assembly with long and short reads improves discovery of gene family expansions. 1–12. doi: 10.1186/s12864-017-3927-8.

Mun S et al. 2017. The Whole-Genome and Transcriptome of the Manila Clam (Ruditapes philippinarum). Genome Biology and Evolution. 9:1487–1498. doi: 10.1093/gbe/evx096.

Murgarella M et al. 2016. A First Insight into the Genome of the Filter-Feeder Mussel Mytilus galloprovincialis Craft, JA, editor. PLoS ONE. 11:e0151561. doi: 10.1371/journal.pone.0151561.

Passamonti M, Ghiselli F. 2009. Doubly Uniparental Inheritance: Two Mitochondrial Genomes, One Precious Model for Organelle DNA Inheritance and Evolution. Dna and Cell Biology. 28:79–89. doi: 10.1089/dna.2008.0807.

Pavey SA et al. 2016. Draft genome of the American Eel (Anguilla rostrata). Molecular Ecology Resources. 17:806–811. doi: 10.1111/1755-0998.12608.

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 26:841–842. doi: 10.1093/bioinformatics/btq033.

R Core Team. 2017. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing: Vienna, Austria.

Rice P, Longden I, Bleasby A. 2000. EMBOSS: the European molecular biology open software suite. Trends Genet. 16:276–277.

Savolainen O, Lascoux M, Merilä J. 2013. Ecological genomics of local adaptation. Nat Rev Genet. 14:807–820. doi: 10.1038/nrg3522.

Simao FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. 31:3210–3212. doi: 10.1093/bioinformatics/btv351.

Simpson JT et al. 2009. ABySS: a parallel assembler for short read sequence data. Genome Res. 19:1117–1123. doi: 10.1101/gr.089532.108.

Smit A, Hubley R. RepeatModeler Open-1.0. (2008-2015). httpwww.repeatmasker.org.

Smit A, Hubley R, Green P. RepeatMasker Open-4.0.(2013-2015).

Spooner DE, Vaughn CC. 2006. Context□dependent effects of freshwater mussels on stream benthic communities. Freshwater Biology. 51:1016–1024. doi: 10.1111/j.1365-2427.2006.01547.x.

Sun J et al. 2017. Adaptation to deep-sea chemosynthetic environments as revealed by mussel genomes. Nat. ecol. evol. 1:0121–7. doi: 10.1038/s41559-017-0121.

Takeuchi T et al. 2012. Draft Genome of the Pearl Oyster Pinctada fucata: A Platform for Understanding Bivalve Biology. Dna Research. 19:117–130. doi: 10.1093/dnares/dss005.

Wang S et al. 2017. Scallop genome provides insights into evolution of bilaterian karyotype and development. Nat. ecol. evol. 1:0120–12. doi: 10.1038/s41559-017-0120.

Zhang G et al. 2012. The oyster genome reveals stress adaptation and complexity of shell formation. Nature. 490:49–54. doi: 10.1038/nature11413.

Zouros E. 2013. Biparental Inheritance Through Uniparental Transmission: The Doubly Uniparental Inheritance (DUI) of Mitochondrial DNA. Evolutionary Biology. 40:1–31. doi: 10.1007/s11692-012-9195-2.

Table 1: DNA sequencing strategy.

			Raw	**Raw**	**Trimmed reads (%)**					**SRA**
Type	**Insert size (bp)**	**Read Length (bp)**	**No. Reads (paired)**	**Total length (mb)**	**No. Reads (paired)**	**Total length (mb)**	**Total length (% raw)**	**read length**	**coverage**	**accession**
Paired-end	300	2X100	189,876,842	37,975	185,721,156	36,274	95.5	97.6		
Paired-end	300	2X100	195,394,768	39,079	191,002,987	37,319,	95.5	97.7		
Paired-end	300	2X100	178,820,287	35,764	174,954,230	34,224	95.6	98.9		
Total			564,091,897	112,818	551,678,373	107,818	95.6	98.1	65X	

Mate pair	5000	2X100	97,801,148	19,560	94,350,168	18,717	95.7	99.3	11X
Pacific Bioscience Long reads		4,406.4 (average)	103,096	454					0.27X
assembled transcriptome	1,170.9 (average) 301-50,048 (min-max)		285,260	334					

Table 2: Assembly statistics (ABySS2.0).

assembly	**n (x10e6)**	**n:1000**	**L50**	**min**	**N80**	**N50**	**N20**	**max**	**sum (x10e6)**
raw unitigs	39.8	347,879	101,624	1,000	1,361	2,181	3,891	25,883	707
unitigs	18.5	444,734	127,617	1,000	1,485	2,452	4,273	25,944	984
contigs	14.0	551,875	141,012	1,000	1,704	3,117	5,817	39,408	1,449
scaffolds	13.7	423,853	92,607	1,000	2,303	5,477	9,099	45,260	1,539
long scaffolds (PacBio)	13.7	410,237	86,661	1,000	2,391	5,708	9,893	47,610	1,548
long scaffolds (PacBio + transcriptome)	13.6	366,926	58,906	1,000	2,534	6,523	16,660	298,135	1,549

n = number of contigs, n:1,000 = number of contigs of mininum length of 1,000, L50 = minimum number of sequences required to represent 50% of the entire assembly, min = mininum length of sequences analysed, N80, N50, N20 = weighted median statistic such that 80/50/20% of the entire assembly is contained in contigs equal to or larger than this value, max = maximum size of contig, sum = sum of all contigs of size > min, assembly stage (*raw unitigs* = raw assembly, not taking into account paired-end information, *unitigs* = filtering, merging and popping bubbles in *De Bruijn* graph, *contigs* = unitigs with paired-end information mapped, *scaffolds* = contigs with mate-pairs information mapped, *long scaffolds* = scaffolds with PacBio / transcriptome information integrated).

Table 3: Assembly and annotation statistics for the long scaffold assembly.

QUAST **Assembly statistics**	**long_scaffolds**	**long_scaffolds (> 1kb scaffolds broken based on N streches)**	**long_scaffolds (> 1kb scaffolds, masked assembly)**
Number of contigs (>= 0 b)	13,635,758	821,266	374,245
Number of contigs (>= 1 kb)	371,706	549,364	374,245
Number of contigs (>= 5 kb)	94,238	50,209	95,019
Number of contigs (>= 10 kb)	26,952	5,151	27,030
Number of contigs (>= 25 kb)	5,073	23	4,976
Number of contigs (>= 50 kb)	1,456	0	1,427
Total length (>= 0 b)	2,638,723,663	1,554,026,338	1,596,234,060
Total length (>= 1kb)	1,590,292,198	1,425,294,273	1,596,234,060
Total length (>= 5 kb)	1,000,983,904	360,423,103	1,003,000,325
Total length (>= 10 kb)	541,545,133	64,766,821	538,648,016
Total length (>= 25 kb)	231,252,884	687,249	226,147,564
Total length (>= 50 kb)	107,178,666	0	104,739,660
Number of contigs	371,706	821,266	37,4245
Largest contig	313,274	44,597	31,3274
Total length	1,590,292,198	1,554,026,338	1,596,234,060

Estimated reference length	1,660,000,000	1,660,000,000	1,660,000,000
GC (%)	34.19	34.19	33.49
N50	6,656	2,812	6,627
number of N's per 100 kb	2,293.33	13.17	39,200.22
number of predicted genes (unique)	201,068	277,765	123,457
number of predicted genes (>= 300 b)	74,820	82,359	41,697
number of predicted genes (>= 1.500 kb)	18,539	14,338	11,897
number of predicted genes (>= 3 kb)	6,511	3,289	4,375
number of annotated ORF (uniprot)	29,031	14,198	25,544

All statistics are based on contigs of size >= 1 kb, unless otherwise noted (e.g., "# contigs (>= 0 b)" and "Total length (>= 0 b)" include all contigs.).

Table 4: Analysis of genome completeness using BUSCO 3.0.2 (Benchmarking Universal Single-Copy Orthologs, (Simao et al. 2015)).

	metazoa	**eukaryota**
Complete orthologs (C)	664 (68%)	185 (61%)
Complete and single-copy orthologs (S)	652 (67%)	181 (60%)
Complete and duplicated orthologs (D)	12 (1%)	4 (1%)
Fragmented orthologs (F)	207 (21%)	76 (25%)
Missing orthologs (M)	107 (11%)	42 (14%)
Total ortholog groups searched	978	303

Table 5: Gene size and repeat elements

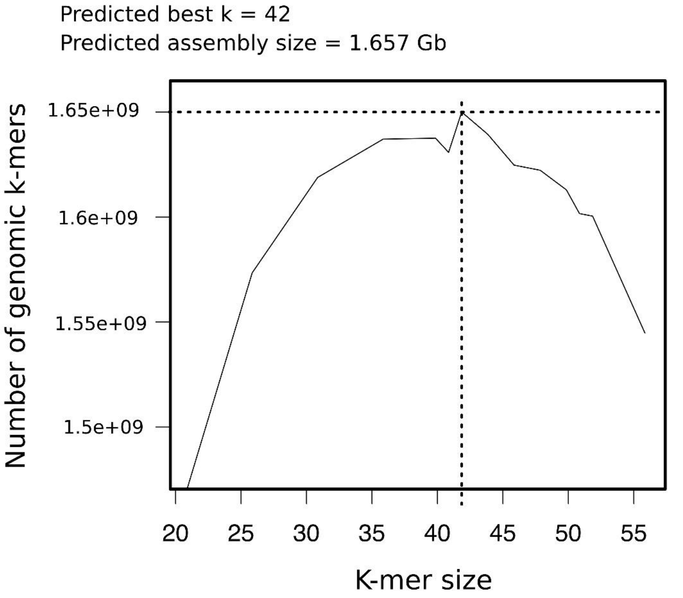
Subclass	Order	Family	Species	Estimated genome size (Gb)	% of repeated elements
Palaeoheterodonta	Unionida	Unionidae	*Venustaconcha ellipsiformis*	1.66	37.81
Heterodonta	Veneroida	Veneridae	*Ruditapes philippinarum*	1.37	26.38
Pteriomorphia	Mytiloida	Mytilidae	*Bathymodiolus platifrons*	1.64	47.90
			Modiolus philippinarum	2.38	62.00
			Mytilus galloprovincialis	1.60	36.13
	Ostreoida	Ostreidae	*Crassostrea gigas*	0.55	36.00
		Pectinidae	*Chlamys farreri*	0.95	32.10
			Patinopecten yessoensis	1.43	38.87
	Pterioida	Pteriidae	*Pinctada fucata*	1.15	37.00
Pteriomorphia mean (s.d.)				1.39 (0.58)	41.43 (10.29)
	Mytiloida mean (s.d.)			1.87 (0.44)	48.68 (12.95)
	Ostreoida mean (s.d.)			0.98 (0.44)	35.66 (3.40)
		Pectinidae mean (s.d.)		1.19 (0.34)	35.49 (4.79)
all subclasses mean (s.d.)				1.41 (0.51)	39.35 (10.23)

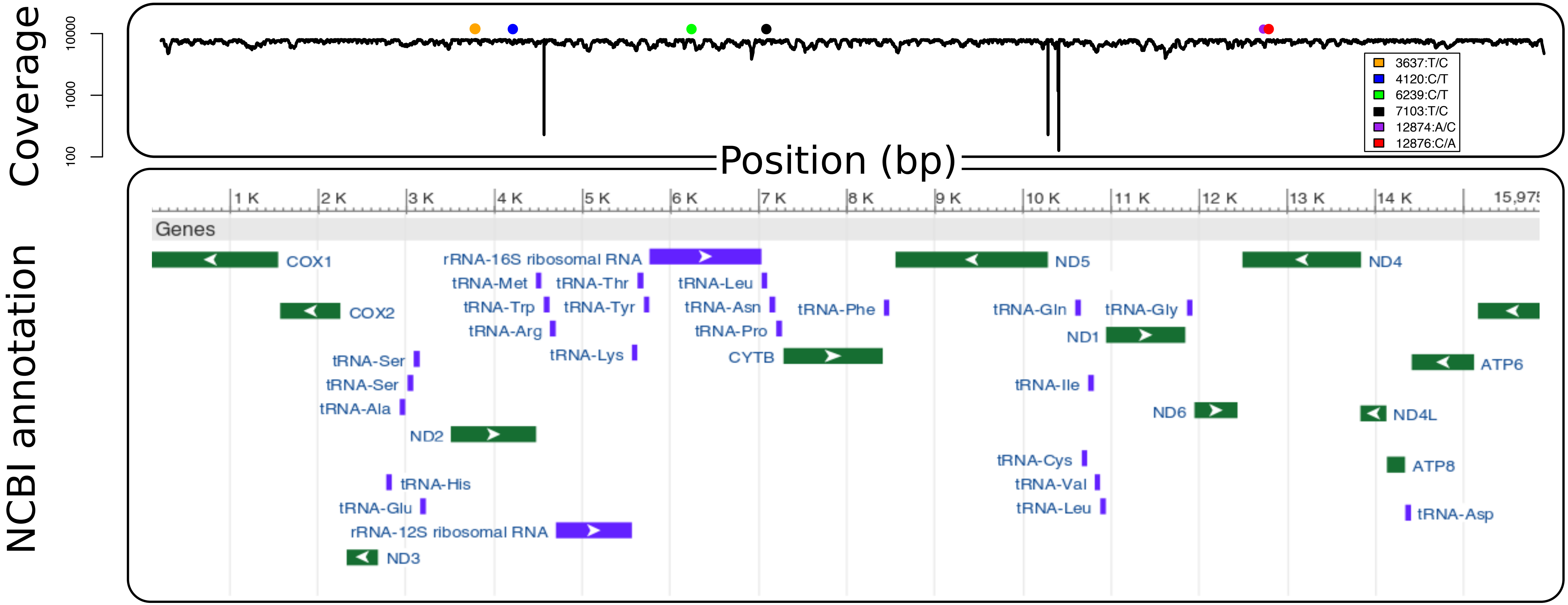
Estimates of genome size and percentage of repeated elements in the currently available bivalve nuclear genomes. Data for each single species was retrieved from the literature (Takeuchi et al. 2012; Zhang et al. 2012; Murgarella et al. 2016; Mun et al. 2017; Yuli Li et al. 2017; Wang et al. 2017). The genome size for *V. ellipsiformis* was based on k-mer analysis (see methods and Fig. 1). Mean and standard deviation (s.d.) values are also shown for the taxa comprising more than one species and for all subclasses, i.e. the class Bivalvia.

Figure Legends

Figure 1: KmerGenie report for best k + predicted genome size.

Figure 2: Mitochondrial coverage based on sequence alignment and annotation (from NCBI). Six nucleotide positions were identified in the legend as fixed for an alternative allele compared to the reference of Breton et al. (2009).


Predicted best k = 42
Predicted assembly size = 1.657 Gb
Number of genomic k-mers
1.65e+09
1.6e+09
1.55e+09
1.5e+09
20
25
30
35
40
45
50
55
K-mer size


Coverage
10000
1000
100
🟧 3637:T/C
🟦 4120:C/T
🟩 6239:C/T
⬛ 7103:T/C
🟪 12874:A/C
🟥 12876:C/A
Position (bp)
NCBI annotation
1 K
2 K
3 K
4 K
5 K
6 K
7 K
8 K
9 K
10 K
11 K
12 K
13 K
14 K
15,975
Genes
COX1
rRNA-16S ribosomal RNA
ND5
ND4
tRNA-Met
tRNA-Thr
tRNA-Leu
tRNA-Trp
tRNA-Tyr
tRNA-Asn
tRNA-Phe
tRNA-Gln
tRNA-Gly
COX2
tRNA-Arg
tRNA-Pro
ND1
tRNA-Lys
CYTB
tRNA-Ser
ATP6
tRNA-Ser
tRNA-Ile
tRNA-Ala
ND6
ND4L
ND2
tRNA-Cys
ATP8
tRNA-His
tRNA-Val
tRNA-Glu
tRNA-Leu
tRNA-Asp
rRNA-12S ribosomal RNA
ND3