

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28

**Title**

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

**Authors**

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

**Affiliations**

<sup>1</sup> Département de Sciences Biologiques, Institut de Recherche en Biologie Végétale, Université de Montréal, Montréal, QC, Canada

<sup>2</sup> Quebec Centre for Biodiversity Science, Montréal, QC, Canada

<sup>3</sup> Département de Sciences Biologiques, Université de Montréal, Montréal, QC, Canada

<sup>4</sup> Department of Biological Sciences, Kent State University, Kent, OH, USA

<sup>5</sup> Department of Biology, Acadia University, Wolfville, NS, Canada

<sup>6</sup> North Carolina Museum of Natural Sciences, Raleigh, NC, USA

<sup>7</sup> Dipartimento di Scienze Biologiche, Geologiche ed Ambientali, University of Bologna, Bologna, Italy

<sup>8</sup> corresponding authors: [sebastien.renaut@umontreal.ca](mailto:sebastien.renaut@umontreal.ca), [s.breton@umontreal.ca](mailto:s.breton@umontreal.ca)

29 **Abstract**

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

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

## 53 **Introduction**

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

70

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

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

92

93         Given the challenges in assembling a reference genome for saltwater mussels  
94 (Sun et al. 2017; Murgarella et al. 2016), we used a combination of different  
95 sequencing strategies (Illumina paired-end and mate pair libraries, Pacific Biosciences  
96 long reads, and a recently assembled reference transcriptome (Capt et al. 2018) to  
97 assemble the first genome draft in the family Unionidae. Hybrid sequencing  
98 technologies using long read–low coverage and short read–high coverage offer an

99 affordable strategy with the advantage of assembling repeated regions of the genome  
100 (for which short reads are ineffective) and circumventing the relatively higher error  
101 rate of long reads (Koren et al. 2012; Miller et al. 2017). Here, we present a *de novo*  
102 assembly and annotation of the genome of the freshwater mussel *Venustaconcha*  
103 *ellipsiformis*.

104

## 105 **Methods**

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

118

### 119 ***Mussel specimen sampling, genomic DNA extraction and library preparation***

120 Adult specimens of *Venustaconcha ellipsiformis* were collected from Straight River  
121 (Minnesota, USA; Lat 44.006509, Long -93.290899) and sexed by microscopic

122 examination of gonad smears. Gills were dissected from a single female individual and  
123 genomic DNA was extracted using a Qiagen DNeasy Blood & Tissue Kit (QIAGEN  
124 Inc., Valencia, CA, USA) using the animal tissue protocol. The quality and quantity of  
125 DNA, respectively, were assessed by electrophoresis on 1% agarose gel and with a  
126 BioDrop mLITE spectrophotometer (a total of 15 µg of DNA was quantified using the  
127 spectrophotometer). For whole genome shotgun sequencing and draft genome  
128 assembly, we used two sequencing platforms: Illumina (San Diego, CA) Hiseq2000  
129 and Pacific Biosciences (Menlo Park, CA) PacBio RSII. First, three paired-end  
130 libraries with insert size of 300b were constructed using Illumina TruSeq DNA Sample  
131 Prep Kit. One mate pair library with insert sizes of about 5Kb was constructed for  
132 scaffolding process using Illumina Nextera mate-pair library construction protocol. For  
133 high-quality genome assembly, Pacific Biosciences system was employed for final  
134 scaffolding process using long reads. Pacific Biosciences long reads (>10Kb) were  
135 generated using SMRT bell library preparation protocol (ten SMRT cells were  
136 sequenced). Construction of sequencing libraries and sequencing analyses were  
137 performed at the Genome Quebec Innovation Centre (McGill University, Qc, Canada).  
138

### 139 *Pre-processing of sequencing reads*

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

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

150

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

155

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

162

### 163 ***Genome assembly strategy***

164 We used ABYSS 2.0 (Jackman et al. 2017), a modern genome assembler specifically  
165 built for large genomes and reads acquired by different sequencing strategies. ABYSS  
166 2.0 works similarly to ABYSS (Simpson et al. 2009), by using a distributed *de Bruijn*  
167 graph representation of the genome, therefore allowing parallel computation of the

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

182

183 In the current study, we combined different types of high throughput  
184 sequencing to aid in assembling the genome (**Table 1**). ABYSS 2.0 (Jackman et al.  
185 2017) performs a first genome assembly step without using the paired-end information,  
186 by extending unitigs until either they cannot be unambiguously extended or come to an  
187 end due to a lack of coverage (*uncorrected unitigs*). This first *de Bruijn* graph  
188 representation of the genome is further cleaned of vertices and edges created by  
189 sequencing errors (*unitigs*). Paired-end information is then used to resolve ambiguities  
190 and merge *contigs*. Following this, mate-pairs are mapped onto the assembly to create



191 *scaffolds*, and finally long reads (Pacific Biosciences long reads) and the  
192 *Venustaconcha* reference transcriptome from Capt et al. (2018) were also mapped onto  
193 the assembly to create *long-scaffolds*. This reference transcriptome was assembled  
194 from a pool of sequences coming from four different male and female individuals and  
195 further details are provided in Capt et al. (2018). Although ideally sequencing  
196 information would all come from a single individual, the current study design did not  
197 allow for this. In addition, given that coding sequences are conserved compared to  
198 non-coding regions, it remains highly valuable to use a transcriptome in a *de novo*  
199 genome assembly.

200

201 We ran the ABYSS 2.0 assembly stage (abyss-bloom-dbg) with a k-mer size of  
202 41 (ABYSS requires an odd number k-mer), a Bloom filter size of 24GB, 4 hash  
203 functions and a threshold of k-mer occurrence set at 3. These parameters were chosen  
204 after performing several test assemblies, in order to minimize the false positive rate  
205 (<5%), maximize the N50 of the assembly and keep the virtual memory (95GB) and  
206 CPU (24 CPUs) requirements within a reasonable computational limit for our  
207 resources. In addition, we adjusted parameters at the mapping stage to create contigs,  
208 scaffolds and long-scaffolds to maximize N50 (overlap required in re-alignments,  
209 distance between mate-pairs, nb reads aligned to support assembly, see pipeline  
210 available at [https://github.com/seb951/venustaconcha\\_ellipsiformis\\_genome](https://github.com/seb951/venustaconcha_ellipsiformis_genome)).

211

212 Genome completeness was assessed using BUSCO 3.0.2 (Benchmarking  
213 Universal Single-Copy Orthologs, Simao et al. 2015). Briefly, BUSCO uses curated lists

214 of known core single copy orthologs to produce evolutionarily-informed quantitative  
215 measures of genome completeness (Simao et al. 2015). Here, we tested both the  
216 eukaryotic (303 single copy orthologs) and metazoan (978 single copy orthologs) gene  
217 lists to assess the completeness of our genome assembly.

218

### 219 *Characterization of repetitive elements*

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

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

241

### 242 ***Genome annotation***

243 We used QUAST (Gurevich et al. 2013) to calculate summary statistics on the genome  
244 assembly. In addition, QUAST uses a Hidden Markov Model to identify putative genes  
245 in the final assembly (GLIMMERHMM Majoros et al. 2004). Following this, we  
246 translated Open Reading Frames identified in the annotation files into protein  
247 sequences using BEDTOOLS v2.27.1 (Quinlan & Hall 2010) and EMBOSS TRANSEQ  
248 v6.6.0 (Rice et al. 2000) bioinformatics pipelines. These were then compared against  
249 the manually curated UniProt database (556,388 reference proteins, downloaded  
250 January 11<sup>th</sup> 2018, e-value cut-off of  $10^{-5}$ ) using BLASTp (Altschul et al. 1990). These  
251 steps were done on the long-scaffolds assembly, the masked long-scaffolds assembly  
252 (with low complexity regions replaced with N), in addition to the broken long-  
253 scaffolds assembly (scaffolds broken into smaller contigs by QUAST, based on long  
254 stretches of N nucleotides).

255

### 256 ***Mitochondrial genome***

257 Given the rare mode of mitochondrial inheritance of freshwater mussels and therefore  
258 its evolutionary importance, we first aimed to check if the mitochondrial female  
259 genome had been properly assembled. Using BLASTn (Altschul et al. 1990) with high

260 stringency (E value <1e-50), we identified a fragmented mitochondrial genome. We  
261 then created a mt specific dataset containing 1,396,004 sequence reads by aligning  
262 paired-end reads to the reference mt genome of Breton et al. (2009) (GenBank Acc.  
263 No. FJ809753) using SAMTOOLS v1.3.1 and BEDTOOLS v2.27.1 (Li et al. 2009; Quinlan  
264 & Hall 2010). We then rebuilt the mt genome *de novo* using ABYSS 2.0, testing  
265 different k-mers (17-45). In addition, we aligned reads to the reference transcriptome  
266 using BWA v0.7.12-R1039 (H Li & Durbin 2009) and identified Single Nucleotide  
267 Polymorphisms (SNPs) with respect to the reference mt genome using SAMTOOLS and  
268 BCFTOOLS v1.3.1 (Li et al. 2009).  
269

270

## 271 **Results and Discussion**

272 We generated 564M paired-end reads (2 X 100b) representing an average 65X  
273 coverage of the genome (**Table 1**). This was complemented by 98M mate-pairs (5Kb  
274 insert, 11X average genome coverage) and 103,000 Pacific Biosciences long reads  
275 (0.3X average genome coverage), and a recently published reference transcriptome  
276 comprised of 285,000 contigs (Capt et al. 2018). Filtering and trimming the raw  
277 paired-end and mate-pair sequences removed about 5% of the total base pairs from  
278 further analyses, indicating that the quality of the raw sequences was high (**Table 1**).  
279 K-mer analysis indicated that the number of unique k-mers peaked at 42 and predicted  
280 a genome assembly size of 1.66Gb (**Figure 1**), smaller than predicted genome size  
281 according to C-value for other Unionida, but in general agreement with the recent draft  
282 genome of the marine mussel *Mytilus galloprovincialis* (1.6Gb) and the deep-sea  
283 vent/seep mussel (*Bathymodiolus platifrons*, 1.64Gb).

284

285       Running the ABySS 2.0 assembly stage (abyss-bloom-dbg) led to a low False  
286 Positive Rate (<0.05%). The N50 for the contig assembly was 3.2Kb with 551,875  
287 contigs (discarding contigs <1Kb, given that small contigs likely represent artefacts  
288 and provide little information for the overall genome assembly (Pavey et al. 2016;  
289 Murgarella et al. 2016, see **Table 2**). Once these were corrected and paired-end, mate-  
290 pairs and long read information were added, the scaffolds N50 increased to 5.5Kb,  
291 with 2.3% of nucleotides represented as “N” (see **Table 2** for the summary statistics  
292 and **Table 3** for overall genome assembly statistics acquired from QUAST analysis).  
293 Adding the Pacific Biosciences long reads only slightly improved the scaffolds N50

294 (from 5.5 to 5.7Kb, **Table 2**) and slightly decreased the number of *long-scaffolds*  
295 >1Kb (from 423,853 to 410,237), likely because our long read coverage was quite low  
296 (0.3X, **Table 1**). In addition, it is also possible that the more error prone Pacific  
297 Biosciences sequences, compared to Illumina paired-end reads, reduced their usability  
298 (Miller et al. 2017). Once the reference transcriptome was added, it improved the N50  
299 to 6.5Kb, and substantially decreased the number of long-scaffolds to 366,926. This  
300 final long-scaffold assembly accounted for a total size of 1.54Gb (with 2.3% of "N"  
301 nucleotides) and represented 93% of the predicted genome size of 1.66Gb. Yet, it  
302 remained highly fragmented (366,926 scaffolds, **Table 2**). Genome annotation  
303 statistics can also be viewed in html format and downloaded here:  
304 [https://github.com/seb951/venustaconcha\\_ellipsiformis\\_genome/tree/master/annotatio](https://github.com/seb951/venustaconcha_ellipsiformis_genome/tree/master/annotation_quast_v3)  
305 [n\\_quast\\_v3](https://github.com/seb951/venustaconcha_ellipsiformis_genome/tree/master/annotation_quast_v3)

306  
307 While assembly numbers (N50, number of scaffolds, etc.) are not directly  
308 comparable with other recently published genomes given the diversity of sequencing  
309 approaches (Illumina, 454, Sanger, PacBio), library types, sequencing depth and  
310 unique nature of the genome themselves, they can give a broad perspective of the  
311 inherent difficulties of assembling large genomes. The best comparison is probably  
312 with the saltwater mussel, *Mytilus galloprovincialis*, giving their similar genome size  
313 (1.6Gb for *Mytilus* vs 1.66Gb for *Venustaconcha*) and Illumina paired-end sequencing  
314 approaches (32X for *Mytilus* vs 65X for *Venustaconcha*). While the *Mytilus* genome  
315 project (Murgarella et al. 2016) did not utilize mate-pair libraries or Pacific Bioscience  
316 long reads, they did make use of sequencing libraries with varying insert sizes (180,

317 500 and 800b). As such, they obtained a genome assembly quality relatively similar to  
318 ours and consisting of 393 thousand scaffolds (>1Kb), with however a substantially  
319 lower N50 (2.6Kb compared to 6.5Kb for *Venustaconcha*). The recently reported  
320 genome for the deep-sea vent/seep mussel *Bathymodiolus platifrons* (1.64Gb) made  
321 use of nine Illumina sequencing libraries with varying insert sizes (180 to 16Kb) and  
322 an overall coverage of >300X. With this very thorough sequencing approach, the  
323 scaffold N50 obtained was substantially higher (343.4Kb), but again the genome  
324 remained highly fragmented, into >65 thousands scaffolds. As exemplified here, high  
325 coverage sequencing libraries with varying insert sizes have become a broadly used  
326 approach for large and complex genomes. In fact, it is implemented by default in many  
327 genome assembly platforms (e.g. Soapdenovo2, Luo et al. 2012, ALLPATHS-LG, Gnerre  
328 et al. 2011). In the future, these libraries will likely be useful to further assemble the  
329 *Venustaconcha* genome, at least until these approaches are superseded by affordable,  
330 error free, single molecule long read sequencing (Gordon et al. 2016; Badouin 2017)  
331 or mapping approaches that allow reaching chromosome level assemblies such as  
332 optical mapping (e.g. Bionano Genomics, San Diego, CA).

333

334 Results of the BUSCO (Simao et al. 2015) analyses showed that 664 (68%) of  
335 the 978 core metazoan genes (CEGs) were considered complete in our assembly.  
336 When the BUSCO analysis was extended to include also fragmented matches, 871  
337 (89%) proteins aligned. Results were similar when compared against the 303 core  
338 eukaryotic genes (61% complete, 86% complete or fragmented, **Table 4**). When  
339 compared to the previously published reference transcriptome for *Venustaconcha*

340 *ellipsiformis* (Capt *et al.* 2018), we found fewer complete genes, but also fewer  
341 duplicated genes (97.5% complete, and 24% duplicated in the reference transcriptome,  
342 compared to 68.1% complete and 1% duplicated here). This likely reflects the fact that  
343 the reference transcriptome is nearly complete, while the current reference genome is  
344 still fragmented. However, the reference transcriptome also likely contains multiple  
345 isoforms of the same genes, in addition to possible nematode contaminating sequences,  
346 despite the authors' best efforts to minimize these problems. Previously analysed  
347 molluscan genomes of similar size (Murgarella *et al.* 2016; Sun *et al.* 2017) have found  
348 that 16% (*Mytilus galloprovincialis*, 1.6Gb), 25% (pearl oyster *Pinctada fucata*,  
349 1.15Gb), 36% (California sea hare *Aplysia californica*, 1.8Gb) of the core eukaryotic  
350 genes were complete. For their part Sun and collaborators (2017), identified 96% of  
351 the core metazoan genes to be partial or complete in the deep-sea vent/seep mussel  
352 *Bathymodiolus platifrons* (1.6Gb), again reflecting that the depth and type of  
353 sequencing, in addition to the idiosyncrasies of each genome, can have considerable  
354 influence on the end results.

355

356 The custom *Venustaconcha* repeat library created *de novo* with RepeatModeler  
357 contained 2,068 families, the majority of them (1,498, 72.44% of the total) classified  
358 as “unknown”. The genome masking performed with the Bivalvia and Mollusca  
359 libraries had scarce performances (masking 2.38% and 2.59%, respectively; details in  
360 **Supplementary Table RM1**), possibly because of the phylogenetic distance between  
361 *V. ellipsiformis*, which belongs to the early-branching bivalve lineage of  
362 Palaeoheterodonta, and the other bivalve and mollusk species represented in the



363 database as well as their relative number of sequences. The custom *Venustaconcha*  
364 library masked 37.17% of the genome, while the combined *Venustaconcha*+Bivalvia  
365 masked 37.69% of the genome and the *Venustaconcha*+Mollusca reached 37.81%, the  
366 highest masking percentage (**Supplementary Table RM2**). After refining, these raw  
367 values slightly decreased to respectively 36.29%, 36.80%, and 36.91%  
368 (**Supplementary Table RM3**). All these latter values of repeat content fall in the 32-  
369 39% range (the median for all species is 37%) where six out of the nine sequenced  
370 bivalve species lie, irrespective of their genome size (*M. philippinarum* and *R.*  
371 *philippinarum* are the furthest from this interval) (**Table 5** and **Supplementary Figure**  
372 **1**). Although the number of species sequenced up to now is still low, this observation  
373 indicates that repetitive elements may contribute differently to the total genome size  
374 among the different bivalve taxa: indeed, the correlation between genome size and  
375 repeats content is weak (**Supplementary Figure 1**). In both the *ab initio* masking with  
376 the *Venustaconcha* library and the two combined ones, most of the identified repeats  
377 are categorized as “unknown” (22.8% of the assembly), followed by retroelements  
378 (LINEs 2.9%, LTR elements 2.3-2.4%, and SINEs 1.7%, for a total of 6.9% of the  
379 assembly) and DNA elements (5.4-5.6% of the assembly) (**Supplementary Table**  
380 **RM3**). Direct comparisons of these values with other species should be performed  
381 with caution, as the usually large “unclassified” portion of repeats might contain  
382 species-specific variants of known elements (Murgarella et al. 2016) that may  
383 therefore change the relative weight of each category on the total.  
384

385           QUAST was used to calculate summary statistics and identify putative genes in  
386 the final assembly using a hidden markov model (**Table 3**). Following this, 29,031;  
387 14,195 and 25,544 Open Reading Frames were annotated using BLASTp against  
388 UniProt database in the long-scaffolds, broken and masked long-scaffolds assemblies,  
389 respectively.

390

391           Freshwater mussels, marine mussels, as well as marine clams are the only  
392 known exception in the animal kingdom with respect to the maternal inheritance of  
393 mitochondrial DNA (see Breton et al. 2007 for a review). Their unique system,  
394 characterized by the presence of two gender-associated mitochondrial DNA lineages,  
395 has therefore attracted studies to better understand mitochondrial inheritance and the  
396 evolution of mtDNA in general. Using BLASTN, we recovered 53 contigs matching to  
397 the 15,975b female reference mt genome from Breton *et al.* (2009), indicating that the  
398 mt genome was highly fragmented and likely improperly assembled with our current  
399 approach, much like what was found in the *Mytilus galloprovincialis* genome draft of  
400 Murgarella (Murgarella et al. 2016). As such, we created a dataset of mt specific  
401 sequences that could be aligned to the mt genome (1,396,004 reads). This mt specific  
402 dataset was then re-assembled *de novo*, using different k-mers (17-45). Using a k-mer  
403 similar or larger to the one used in the overall assembly ( $k \geq 41$ ) resulted in a failed  
404 assembly (no contigs created, data not shown), while using a k-mer  $< 21$  generated a  
405 highly fragmented mt genome (data not shown). Using a k-mer between 21 and 39  
406 generated one large contig of 16,024b comprising the entire mitogenome, with a 42b  
407 insertion in the 16S ribosomal RNA. Given the different rate of evolution of mtDNAs,

408 it is likely that assembly parameters we used for the whole genome were not  
409 appropriate for the *V. ellipsiformis* female mt genome. Finally, we also re-aligned the  
410 mt specific dataset to the original mt genome of Breton et al. (2009) and found high  
411 coverage (mean = 7,256X, SD = 682) for most positions, while for three regions  
412 coverage dropped below 300X (**Figure 2**). Six SNPs with respect to the reference were  
413 also identified, indicating possible polymorphism, or sequencing error in the original  
414 mt reference genome (**Figure 2**).

415

## 416 **Conclusion**

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

430

## 431 **Abbreviations**

432 BLAST: Basic Local Alignment Search Tool

433 b: base pairs

434 Kb: Kilobases

435 M: Million

436 Gb: Gigabases

437 GB: gigabytes

438 CPU: Central Processing Unit

439 DNA: Deoxyribonucleic acid

440 LINEs: Long interspersed elements

441 LTR: Long terminal repeats

442 ORF: Open Reading Frames

443 N80/50/20: weighted median statistic such that 80/50/20% of the entire assembly is

444 contained in contigs/scaffolds equal to or larger than this value.

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

446 RAM: Random Access Memory

447 SINEs: Short interspersed elements

448

449 **Data availability**

450 Supporting data for this Genome Report will be made available on [datadryad.org](https://datadryad.org)

451 Raw sequences are available in the SRA database with number SRP132483

452 (submission SUB3624229 to be release upon publication) and Bioproject accession

453 PRJNA433387. All scripts used in the analyses are available on github

454 ([https://github.com/seb951/venustaconcha\\_ellipsiformis\\_genome](https://github.com/seb951/venustaconcha_ellipsiformis_genome)).

455

456 **Acknowledgments**

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

462

## 463 **References**

- 464 Aldridge DC, Fayle TM, Jackson N. 2007. Freshwater mussel abundance predicts  
465 biodiversity in UK lowland rivers. *Aquatic Conserv: Mar. Freshw. Ecosyst.* 17:554–  
466 564. doi: 10.1002/aqc.815.
- 467 Altschul SF, Gish W, Miller W, Myers EW, Lipman, DJ, 1990. Basic local alignment  
468 search tool. *Journal of molecular biology*, 215(3), pp.403-410.
- 469 Andrews S. 2010. FastQC: a quality control tool for high throughput sequence data.  
470 <https://www.bioinformatics.babraham.ac.uk/projects/fastqc>.
- 471 Badouin H. 2017. The sunflower genome provides insights into oil metabolism,  
472 flowering and Asterid evolution. *Nature*. 1–20.
- 473 Bloom BH. 1970. Space/time trade-offs in hash coding with allowable errors.  
474 *Communications of the ACM*. 13:422–426. doi: 10.1145/362686.362692.
- 475 Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina  
476 sequence data. *Bioinformatics*. 30:2114–2120. doi: 10.1093/bioinformatics/btu170.
- 477 Breton S et al. 2009. Comparative Mitochondrial Genomics of Freshwater Mussels  
478 (*Bivalvia: Unionoida*) With Doubly Uniparental Inheritance of mtDNA: Gender-  
479 Specific Open Reading Frames and Putative Origins of Replication. *Genetics*.  
480 183:1575–1589. doi: 10.1534/genetics.109.110700.
- 481 Breton S et al. 2011. Novel Protein Genes in Animal mtDNA: A New Sex  
482 Determination System in Freshwater Mussels (*Bivalvia: Unionoida*)? *Mol. Biol. Evol.*  
483 28:1645–1659. doi: 10.1093/molbev/msq345.
- 484 Breton S, Beaupre HD, Stewart DT, Hoeh WR, Blier PU. 2007. The unusual system of  
485 doubly uniparental inheritance of mtDNA: isn't one enough? *Trends Genet.* 23:465–  
486 474. doi: 10.1016/j.tig.2007.05.011.
- 487 Breton S, Capt C, Guerra D, Stewart D. 2017. Sex Determining Mechanisms in  
488 Bivalves. *Preprints*. 1–23. doi: 10.20944/preprints201706.0127.v1.
- 489 Capt C et al. 2018. Deciphering the Link between Doubly Uniparental Inheritance of  
490 mtDNA and Sex Determination in Bivalves: Clues from Comparative Transcriptomics.  
491 *Genome Biology and Evolution*. 10:577–590. doi: 10.1093/gbe/evy019.
- 492 Chikhi R, Medvedev P. 2014. Informed and automated k-mer size selection for  
493 genome assembly. *Bioinformatics*. 30:31–37. doi: 10.1093/bioinformatics/btt310.
- 494 Gnerre S et al. 2011. High-quality draft assemblies of mammalian genomes from  
495 massively parallel sequence data. *PNAS*. 108:1513–1518. doi:  
496 10.1073/pnas.1017351108.

- 497 Gordon D et al. 2016. Long-read sequence assembly of the gorilla genome. *Science*.  
498 352:aae0344–aae0344. doi: 10.1126/science.aae0344.
- 499 Gurevich A, Saveliev V, Vyahhi N, Tesler G. 2013. QUASt: quality assessment tool  
500 for genome assemblies. *Bioinformatics*. 29:1072–1075. doi:  
501 10.1093/bioinformatics/btt086.
- 502 Gutiérrez JL, Jones CG, Strayer DL, Iribarne OO. 2003. Mollusks as ecosystem  
503 engineers: the role of shell production in aquatic habitats. *Oikos*. 101:79–90. doi:  
504 10.1034/j.1600-0706.2003.12322.x.
- 505 Haag WR. 2012. *North American freshwater mussels: natural history, ecology, and*  
506 *conservation*.
- 507 Haag WR, Rypel AL. 2011. Growth and longevity in freshwater mussels: evolutionary  
508 and conservation implications. *Biol Rev*. 86:225–247. doi: 10.1111/j.1469-  
509 185X.2010.00146.x.
- 510 Jackman SD et al. 2017. ABySS 2.0: resource-efficient assembly of large genomes  
511 using a Bloom filter. *Genome Res*. 27:768–777. doi: 10.1101/gr.214346.116.
- 512 Koren S et al. 2012. Hybrid error correction and de novo assembly of single-molecule  
513 sequencing reads. *Nat Biotechnol*. 30:693–700. doi: 10.1038/nbt.2280.
- 514 Li H et al. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*.  
515 25:2078–2079. doi: 10.1093/bioinformatics/btp352.
- 516 Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler  
517 transform. *Bioinformatics*. 25:1754–1760. doi: 10.1093/bioinformatics/btp324.
- 518 Li Y et al. 2017. Scallop genome reveals molecular adaptations to semi-sessile life and  
519 neurotoxins. *Nature Communications*. 1–11. doi: 10.1038/s41467-017-01927-0.
- 520 Lopes-Lima M et al. 2014. Biology and conservation of freshwater bivalves: past,  
521 present and future perspectives. *Hydrobiologia*. 735:1–13. doi: 10.1007/s10750-014-  
522 1902-9.
- 523 Lopes-Lima M et al. 2018. Conservation of freshwater bivalves at the global scale:  
524 diversity, threats and research needs. *Hydrobiologia*. 1–14. doi: 10.1007/s10750-017-  
525 3486-7.
- 526 Lopes-Lima M et al. 2017. Conservation status of freshwater mussels in Europe: state  
527 of the art and future challenges. *Biol Rev*. 92:572–607. doi: 10.1111/brv.12244.
- 528 Luo R et al. 2012. SOAPdenovo2: an empirically improved memory-efficient short-  
529 read de novo assembler. *Gigascience*. 1. doi: 10.1186/2047-217X-1-18.
- 530 Majoros WH, Pertea M, Salzberg SL. 2004. TigrScan and GlimmerHMM: two open

- 531 source ab initio eukaryotic gene-finders. *Bioinformatics*. 20:2878–2879. doi:  
532 10.1093/bioinformatics/bth315.
- 533 Miller JR et al. 2017. Hybrid assembly with long and short reads improves discovery  
534 of gene family expansions. 1–12. doi: 10.1186/s12864-017-3927-8.
- 535 Mun S et al. 2017. The Whole-Genome and Transcriptome of the Manila Clam  
536 (*Ruditapes philippinarum*). *Genome Biology and Evolution*. 9:1487–1498. doi:  
537 10.1093/gbe/evx096.
- 538 Murgarella M et al. 2016. A First Insight into the Genome of the Filter-Feeder Mussel  
539 *Mytilus galloprovincialis* Craft, JA, editor. *PLoS ONE*. 11:e0151561. doi:  
540 10.1371/journal.pone.0151561.
- 541 Passamonti M, Ghiselli F. 2009. Doubly Uniparental Inheritance: Two Mitochondrial  
542 Genomes, One Precious Model for Organelle DNA Inheritance and Evolution. *Dna*  
543 and *Cell Biology*. 28:79–89. doi: 10.1089/dna.2008.0807.
- 544 Pavey SA et al. 2016. Draft genome of the American Eel (*Anguilla rostrata*).  
545 *Molecular Ecology Resources*. 17:806–811. doi: 10.1111/1755-0998.12608.
- 546 Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing  
547 genomic features. *Bioinformatics*. 26:841–842. doi: 10.1093/bioinformatics/btq033.
- 548 R Core Team. 2017. *R: A Language and Environment for Statistical Computing*. R  
549 Foundation for Statistical Computing: Vienna, Austria.
- 550 Rice P, Longden I, Bleasby A. 2000. EMBOSS: the European molecular biology open  
551 software suite. *Trends Genet*. 16:276–277.
- 552 Savolainen O, Lascoux M, Merilä J. 2013. Ecological genomics of local adaptation.  
553 *Nat Rev Genet*. 14:807–820. doi: 10.1038/nrg3522.
- 554 Simao FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. 2015.  
555 BUSCO: assessing genome assembly and annotation completeness with single-copy  
556 orthologs. *Bioinformatics*. 31:3210–3212. doi: 10.1093/bioinformatics/btv351.
- 557 Simpson JT et al. 2009. ABySS: a parallel assembler for short read sequence data.  
558 *Genome Res*. 19:1117–1123. doi: 10.1101/gr.089532.108.
- 559 Smit A, Hubley R. RepeatModeler Open-1.0. (2008-2015).  
560 <http://www.repeatmasker.org>.
- 561 Smit A, Hubley R, Green P. RepeatMasker Open-4.0.(2013-2015).
- 562 Spooner DE, Vaughn CC. 2006. Context-dependent effects of freshwater mussels on  
563 stream benthic communities. *Freshwater Biology*. 51:1016–1024. doi: 10.1111/j.1365-  
564 2427.2006.01547.x.



- 565 Sun J et al. 2017. Adaptation to deep-sea chemosynthetic environments as revealed by  
566 mussel genomes. *Nat. ecol. evol.* 1:0121–7. doi: 10.1038/s41559-017-0121.
- 567 Takeuchi T et al. 2012. Draft Genome of the Pearl Oyster *Pinctada fucata*: A Platform  
568 for Understanding Bivalve Biology. *Dna Research.* 19:117–130. doi:  
569 10.1093/dnares/dss005.
- 570 Wang S et al. 2017. Scallop genome provides insights into evolution of bilaterian  
571 karyotype and development. *Nat. ecol. evol.* 1:0120–12. doi: 10.1038/s41559-017-  
572 0120.
- 573 Zhang G et al. 2012. The oyster genome reveals stress adaptation and complexity of  
574 shell formation. *Nature.* 490:49–54. doi: 10.1038/nature11413.
- 575 Zouros E. 2013. Biparental Inheritance Through Uniparental Transmission: The  
576 Doubly Uniparental Inheritance (DUI) of Mitochondrial DNA. *Evolutionary Biology.*  
577 40:1–31. doi: 10.1007/s11692-012-9195-2.
- 578

579  
580  
581  
582  
583

**Table 1:** DNA sequencing strategy.

Type	Raw		Raw		Trimmed reads (%)			SRA		
	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		
<b>Total</b>			564,091,897	112,818	551,678,373	107,818	95.6	98.1	65X	

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

584

585

586

587 **Table 2:** Assembly statistics (ABYSS2.0).

588

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

589

590 n = number of contigs, n:1,000 = number of contigs of minimum length of 1,000, L50 = minimum number of sequences required to  
 591 represent 50% of the entire assembly, min = minimum length of sequences analysed, N80, N50, N20 = weighted median statistic such  
 592 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 =  
 593 sum of all contigs of size > min, assembly stage (*raw unitigs* = raw assembly, not taking into account paired-end information, *unitigs*  
 594 = filtering, merging and popping bubbles in *De Bruijn* graph, *contigs* = unitigs with paired-end information mapped, *scaffolds* =  
 595 contigs with mate-pairs information mapped, *long scaffolds* = scaffolds with PacBio / transcriptome information integrated).

596

597

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

599

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

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

600

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

603

604

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

606

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

607

608

609 **Table 5:** Gene size and repeat elements

Subclass	Order	Family	Species	Estimated genome size (Gb)	% of repeated elements
Palaeoheterodonta	Unionida	Unionidae	<i>Venustaconcha ellipsiformis</i>	1.66	37.81
Heterodonta	Veneroida	Veneridae	<i>Ruditapes philippinarum</i>	1.37	26.38
Pteriomorpha	Mytiloida	Mytilidae	<i>Bathymodiolus platifrons</i>	1.64	47.90
			<i>Modiolus philippinarum</i>	2.38	62.00
			<i>Mytilus galloprovincialis</i>	1.60	36.13
	Ostreoida	Ostreidae	<i>Crassostrea gigas</i>	0.55	36.00
		Pectinidae	<i>Chlamys farreri</i>	0.95	32.10
			<i>Patinopecten yessoensis</i>	1.43	38.87
Pterioidea	Pteriidae	<i>Pinctada fucata</i>	1.15	37.00	
Pteriomorpha 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)

610 Estimates of genome size and percentage of repeated elements in the currently available bivalve nuclear genomes. Data for each single  
 611 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  
 612 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



613 standard deviation (s.d.) values are also shown for the taxa comprising more than one species and for all subclasses, i.e. the class

614 Bivalvia.

615

616

617

618 **Figure Legends**

619

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

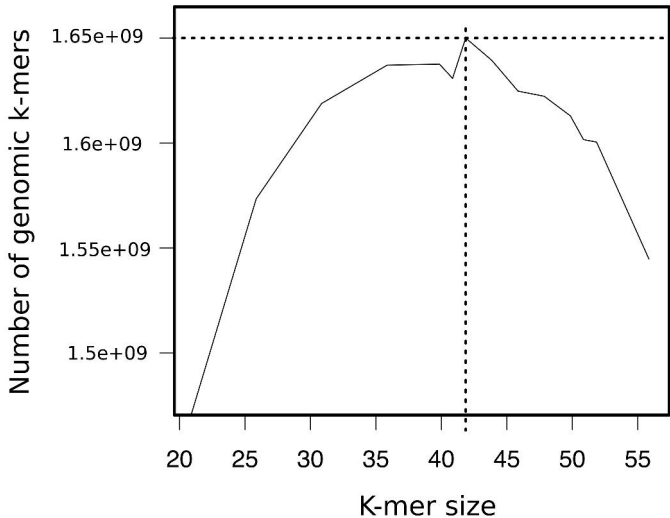
621

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

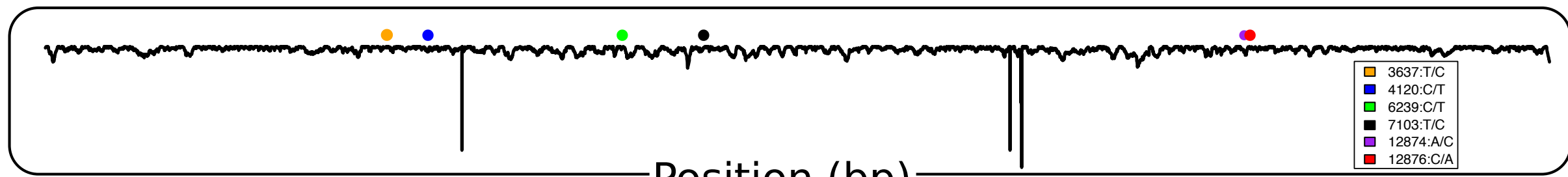
624

Predicted best k = 42

Predicted assembly size = 1.657 Gb



Coverage

10000  
1000  
100

NCBI annotation

