Comparative mitochondrial genomics of freshwater mussels (Bivalvia: Unionoida) with Doubly Uniparental Inheritance of mtDNA: gender-specific Open Reading Frames (ORFs) and putative origins of replication

Sophie Breton,* Hélène Doucet Beaupré,† Donald T. Stewart,‡ Helen Piontkivska,* Moumita Karmakar,* Arthur E. Bogan,§ Pierre U. Blier,† and Walter R. Hoeh* 

*Department of Biological Sciences, Kent State University, Kent, OH 44242 USA
†Département de Biologie, Université du Québec à Rimouski, Rimouski, Qc G5L 3A1 Canada
‡Department of Biology, Acadia University, Wolfville, NS B4P 2R6 Canada
§North Carolina State Museum of Natural Sciences, Raleigh, NC 27699-1626 USA
Running Head: Comparative mitochondrial genomics of species with DUI

Key Words: mtDNA, Doubly uniparental inheritance (DUI), comparative genomics, control region, Open Reading Frame (ORF)

Corresponding author:
Sophie Breton
Evolutionary, Population, and Systematic Biology Group
Department of Biological Sciences
Kent State University
Kent, OH 44242
USA
418-732-4837 (cell)
330-672-3713 (fax)
Email: breton.sophie@gmail.com
ABSTRACT

Doubly uniparental inheritance (DUI) of mitochondrial DNA in marine mussels (Mytiloida), freshwater mussels (Unionoida), and marine clams (Veneroida) is the only known exception to the general rule of strict maternal transmission of mtDNA in animals. DUI is characterized by the presence of gender-associated mitochondrial DNA lineages that are inherited through males (male-transmitted or M types) or females (female-transmitted or F types), respectively. This unusual system constitutes an excellent model for studying basic aspects of mitochondrial DNA inheritance and the evolution of mtDNA genomes in general.

Here we compare published mitochondrial genomes of unionoid bivalve species with DUI, with an emphasis on characterizing unassigned regions, to identify regions of the F and M mtDNA genomes that could (i) play a role in replication or transcription of the mtDNA molecule, and/or (ii) determine whether a genome will be transmitted via the female or the male gamete. Our results reveal the presence of one F-specific and one M-specific open reading frames (ORFs), and we hypothesize that they play a role in the transmission and/or gender-specific adaptive functions of the M and F mtDNA genomes in unionoid bivalves.

Three major unassigned regions shared among all F and M unionoid genomes have also been identified, and our results indicate that (i) two of them are potential heavy-strand control regions (OH) for regulating replication and/or transcription, and that (ii) multiple and potentially bidirectional light-strand origins of replication (OL) are present in unionoid F and M mitochondrial genomes. We propose that unassigned regions are the most promising candidate sequences in which to find regulatory and/or gender-specific sequences that could determine whether a mitochondrial genome will be maternally or paternally transmitted.
INTRODUCTION

Marine mussels (Mytiloida), freshwater mussels (Unionoida), and marine clams (Veneroida) are the only known animals that do not transmit their mitochondrial DNA exclusively maternally (see White et al. 2008 for a review of exceptional cases of paternal leakage in animals). The system of mitochondrial DNA transmission in these bivalves is referred to as “Doubly Uniparental Inheritance” or DUI and is characterized by the presence of two gender-associated mitochondrial DNA lineages that are inherited through males (male-transmitted or M types) or females (female-transmitted or F types), respectively (see Breton et al. 2007 and Passamonti and Ghielli 2009 for reviews of DUI). DUI constitutes an excellent model system for studying basic aspects of mitochondrial DNA inheritance and the evolution of mtDNA genomes in general. Because DUI is the exception to the rule, understanding how bivalves evolved distinct male and female mtDNA lineages can provide important insights into the evolutionary forces that maintain strictly maternal inheritance in most animals.

To date, complete F and M mtDNA sequences have been determined for the mytiloid mussels Mytilus edulis, M. galloprovincialis and M. trossulus (AY484747, Boore et al. 2004; AY497292 and AY363687, Mizi et al. 2005; AY823623 and AY323624, Breton et al. 2006; DQ198231 and DQ198225, Zbawicka et al. 2007), the veneroid clam Venerupis philippinarum (AB065374 and AB065375, M Okazaki and R Ueshima personal communication), and 7 unionoid bivalve species (i.e., the F genome of Lampsilis ornata (Unionoida: Ambleminae: Lampilini) [AY365193, Serb and Lydeard 2003], the F genome of Hyriopsis cumingii (Unionoida: Ambleminae: Gonideini) [FJ529186, RL Zheng and JL Li personal communication], the F genome of Cristaria plicata (Unionoida: Anodontinae: Anodontini) [FJ986302, WP Jiang, RL Zheng and JL Li personal communication], the F and M genomes of Inversidens japonensis (Unionoida: Ambleminae: Gonideini) [AB055624 and
AB055625, M OKAZAKI and R UESHIMA personal communication], and recently, we have sequenced the F and M genomes of *Venustaconcha ellipsiformis* (Unionoida: Ambelminae: Lampilini), *Quadrula quadrula* (Unionoida: Ambelminae: Quadrulini) and *Pyganodon grandis* (Unionoida: Anodontinae: Anodontini) [FJ809750 to FJ809755, H DOUCET BEAUPRÉ et al. unpublished data)] (Table 1). Overall, these studies have shown: (i) a high level of nucleotide sequence divergence but nearly identical gene content between F and M genomes within each species, (ii) an accelerated rate of molecular evolution of both the M and F genomes compared to other animal mitochondrial genomes, (iii) an accelerated rate of molecular evolution of M genomes compared to F genomes, (iv) an absence of *atp8* in mytiloid mussels, (v) the presence of a second *trnM* (i.e., transfer RNA gene for methionine) in mytiloid and veneroid bivalves, (vi) recombination between M and F genomes in mytiloid mussels, (vii) periodic “role reversals” (= masculinization) of female-transmitted mtDNA in mytiloid mussels that are subsequently transmitted through sperm, (viii) different gene order between F and M genomes in unionoids, and (ix) the presence of a unique extension of the cytochrome *c* oxidase subunit II gene in the M (but not the F) genome in unionoids (reviewed in BRETON et al. 2007, see also CHAPMAN et al. 2008). An important hypothesis that has emerged from sequencing studies of species with DUI is that gender-specific sequences and/or sequences that exhibit the highest level of nucleotide divergence between the F and M genomes (i.e., regions that are under different, potentially gender-specific selective constraints and could, therefore, have different roles in either genome) are the most likely candidates for determining whether a mitochondrial genome will be transmitted maternally or paternally (ZOUROS 2000; BURZYŃSKI et al. 2003; CAO et al. 2004a; BRETON et al. 2006; BRETON et al. 2007; VENETIS et al. 2007; THEOLOGIDIS et al. 2007; CAO et al. 2009). For example, it has been demonstrated in marine mussels that masculinized type genomes (i.e. an F genome that "masculinizes" and takes on the role of the previous M genome) are essentially
recombinants composed of an F genome’s coding and control regions, with an additional standard M-type control region (BURZYŃSKI et al. 2003; BRETON et al. 2006; VENETIS et al. 2007). This has led to the proposition that an M-type control region, particularly its most variable domain called VD1 by CAO et al. (2004a), might be necessary to confer the paternal role on genomes that are otherwise F-like (BURZYŃSKI et al. 2003; BRETON et al. 2006; VENETIS et al. 2007). Alternatively, the absence of masculinization or role-reversal events in freshwater bivalves coincides with the presence of a unique M genome-specific 3’ extension of the cytochrome c oxidase subunit 2 gene (Mcox2e; CUROLE and KOCHER 2002) that could facilitate the transmission of the M genomes in freshwater bivalves (BRETON et al. 2007; CHAKRABARTI et al. 2007; CHAPMAN et al. 2008). To date, the control region has not been confirmed in unionoid bivalves (see SERB and LYDEARD 2003). Identifying the unionoid F and M control regions is essential for a more complete understanding of DUI. Studies of mytiloid genomes that have switched from maternal to paternal transmission have provided evidence that specific sequences of the mtDNA genome that control the mode of inheritance (i.e., male or female transmission) are located in the control region (BURZYŃSKI et al. 2003; BRETON et al. 2006; VENETIS et al. 2007; CAO et al. 2009). It is therefore critical that the control region in unionoids be confirmed to facilitate comparative studies of the developmental control of the F and M genomes across all bivalve species that possess DUI.

Here we present a comparative analysis of complete F and M mitochondrial genomes of unionoid bivalve species with DUI. Our objectives are to highlight both unique features and characteristics shared among different species, with an emphasis on characterizing unassigned regions (i.e., non-coding regions that are functionally unassigned) to identify F and M sequences that could play crucial roles in replication or transcription of the mtDNA molecule and to point out particular regions of the genomes that could determine whether a genome will be transmitted by eggs or sperm. One F-specific and one M-specific ORF have
been identified and, given their expression and antiquity in unionoid bivalves, we hypothesize that they are involved in the different modes of transmission and/or gender-specific adaptive functions of the M and F mtDNA genomes in unionoid bivalves. Additionally, our results reveal that unionoid mitochondrial control regions are not well defined and their locations could be variable among unionoid mitochondrial genomes. We propose that the currently unassigned regions are the most favorable candidates in which to find regulatory or gender-specific sequences that could determine whether a genome is transmitted maternally or paternally.

**MATERIALS AND METHODS**

Complete F and M mitochondrial sequences of unionoid bivalve species with DUI were obtained from the National Center for Biotechnology Information (NCBI) GenBank entries for the 11 genomes listed in Table 1 (it has to be noted that the M genomes of *C. plicata*, *H. cumingii* and *L. ornata* have not been sequenced). ClustalW (THOMPSON *et al.* 1994) was used to align sequences and MEGA 3.0 (KUMAR *et al.* 2004) was used to calculate the proportion of nucleotide and amino acid differences (p-distances) between F and M genes and AT-skew = (A – T)/(A + T) (PERNA and KOCHER 1995) at fourfold redundant sites for each mitochondrial protein-coding gene. Repeated elements were identified using REPFIND (BETLEY *et al.* 2002), and conserved motifs were identified using Dialign version 2.2.1 (SUBRAMANIAN *et al.* 2008) and Jalview version 2 (WATERHOUSE *et al.* 2009). DNA secondary structures in unassigned regions were predicted using Mfold version 3.2 (ZUKER 2003). Examination of open reading frames (ORFs) was performed using the NCBI ORF Finder program (http://www.ncbi.nlm.nih.gov/projects/gorf/) with the invertebrate mitochondrial genetic code. Sequence similarity searches were performed in GenBank using BLASTN (BENSON *et al.* 2004), BLASTX and PSI-BLAST (ALTSCHUL *et al.* 1997). Gender-
specific ORFs were examined using Fickett’s (1982) test code algorithm. Transmembrane helices and other potential protein features of gender-specific ORFs were identified using the ConPred II (Arai et al. 2004) and PredictProtein (Rost et al. 2004). Western blot analyses of eggs and testes extracts of the species *V. ellipsiformis* were performed as described by Chakrabarti et al. (2006).

In addition to the customary characteristics used to identify the mitochondrial control region of replication in animals (i.e., the largest non-coding region, increased AT content, presence of repetitive elements and secondary structures) (Boore 1999; Saccone et al. 2002; Cao et al. 2004a; Saito et al. 2005; Kuhn et al. 2006; Oliveira et al. 2007; Brugler and France 2008), we also used AT-skew values of protein-coding genes at fourfold redundant sites to locate the origins of heavy (O_H) and light (O_L) strand replication in freshwater bivalves. In most metazoans, the mitochondrial DNA genome replicates with a strand-asynchronous, asymmetric mechanism during which the parental heavy (H) strand becomes temporarily single-stranded DNA (ssDNA) while the nascent H-strand is synthesized, and when the heavy strand synthesis reaches 2/3 of the genome, it exposes the O_L and initiates the synthesis of a new light (L) strand in the opposite direction (Clayton 1982; Reyes et al. 1998). Strong biases toward A+C for the L strand and G+T for the H strand are common in mitochondrial genomes and are associated with this asymmetrical replication and the extended time that the parental heavy strand spends in the mutagenically susceptible single-stranded state during the process (Reyes et al. 1998; Saccone et al. 2002; Faith and Pollock 2003). This mutational bias appears to be due to: (i) spontaneous deamination of C on the single-stranded H strand that produces U, which DNA polymerase basepairs with A rather than G (consequently the percentage of C decreases and that of T increases on the H strand according to single strand exposure), (ii) deamination of A that produces hypoxanthin, which basepairs with C rather than T (A decreases and G increases on the H strand), and (iii)
oxidation of guanine on the H strand that produces 8-hydroxyguanine, which basepairs with A rather than C (G decreases and T increases on the H strand) (REYES et al. 1998; SACCONI et al. 2002; FAITH and POLLOCK 2003; RODAKIS et al. 2007). The net effect on nucleotide frequency is that cytosine and adenine may only decrease on the H strand, whereas thymine may only increase and guanine could either increase or decrease as a result of single-strand exposure (RODAKIS et al. 2007). Because the increase or decrease of G on the H strand will influence the amount of C on the L strand, we will only consider AT-skew values in the following section.

According to the formula \( \text{AT-skew} = (A - T)/(A + T) \), skew values are distributed in the range of -1 to +1 and, thus, the compositional asymmetry increases when the absolute skew values approach one, and decreases when the skew values approach zero (SACCONI et al. 2002). The genes located in the vicinity of the O_H remain in a single-stranded state for a relatively long time during the replication process and these regions therefore experience more mutations (FAITH and POLLOCK 2003; FONSECA et al. 2008). Consequently, the compositional asymmetry should be greater for the genes closer to the O_H (the genes closer to the O_H should be less A-rich and more T-rich when encoded on the heavy strand and more A-rich and less T-rich when encoded on the light strand), and thus AT-skew values should be greater for these genes. Alternatively, the genes closer to the O_L, in the direction of the L-strand synthesis, remain exposed to mutation for less time so the genes closer to the O_L should present a lower compositional asymmetry with lower skew values (these genes should be more A-rich and less T-rich when encoded on the heavy strand and less A-rich and more T-rich when encoded on the light strand) (FAITH and POLLOCK 2003; FONSECA et al. 2008). If the asymmetrical model of mtDNA replication also applies to unionoid bivalves, we can speculate that O_H would be located in the vicinity of the protein-coding genes showing the
greatest AT-skew value at fourfold redundant sites, while O$_L$ would be located in the vicinity of the protein-coding genes showing the lowest AT-skew values.

**RESULTS AND DISCUSSION**

**Unassigned regions in F and M unionoid mitochondrial genomes**

All analyzed unionoid mitochondrial genomes contain the 37 genes typically found in metazoans (Fig. 1). Eleven genes (i.e., *cox1*, *cox3*, *atp6*, *atp8*, *trnD*, *nd4l*, *nd4*, *trnH*, *nd5*, *nd3*, *cox2/cox2e*) are encoded on the heavy strand (H-strand; G+T = ~ 65 % and 67 % for the F and M genomes, respectively), while the remaining 26 genes are encoded on the light strand (L-strand; G+T = ~ 35 % and 33 % for the F and M genomes, respectively) (Fig. 1). The only exception observed within the M lineage is the absence of a complete *atp8* gene in *P. grandis* (H DOUCET BEAUPRÉ et al. unpublished data). In the unionoid F lineage, the mtDNAs of *I. japonensis* and *Hyriopsis cumingii* share a different gene order between *cox2* and *rrnS* that could represent a derived characteristic of the F lineage in the Gonideini (Unionidae: Ambleminae). The transposition of *trnH*, the gene order inversion of *trnD* and *atp8* and one F- and one M-specific open reading frame (gender-specific ORFs; see below) are responsible for the organizational differences between F and M genomes in freshwater bivalves (Fig. 1). The M-specific *cox2* extension also contributes to differences between F and M genomes (Fig. 1).

In total, twenty-two to thirty-three unassigned regions are found in F and M unionoid genomes, encompassing a total of 1148 – 2241 bp (F genomes) and 1526 – 2096 bp (M genomes), i.e., about ~ 10 % of the total length of the mitochondrial genomes (Fig. 1 and Supplementary Table S1). Similar results have been observed in mytiloid F and M genomes (i.e., ~ 10 % unassigned sequences) (MIZI et al. 2005; BRETON et al. 2006; ZBAWICKA et al. 2007). In comparison, a higher proportion of unassigned sequences (i.e., > 15.8 % for F and >
21.3 % for M) is observed in the F and M genomes of the veneroid clam *V. philippinarum* (M OKAZAKI and R UESHIMA personal communication). Multiple intergenic regions are not unusual in mitochondrial genomes that have undergone significant rearrangements (BOORE 1999). These regions often represent vestiges of pseudogenes generated by gene duplication followed by random deletions, a process that appears to be particularly important in molluscan taxa (SERB and LYDEARD 2003; AKASAKI *et al.* 2006; BOORE 2006).

An examination of all complete F and M genomes shows three unassigned regions in the same relative genomic positions that are greater than 20 nucleotides in length in all species (Fig. 1A, B, C and Supplementary Table S1). We refer to these as “shared unassigned regions”. They include the regions between *trnF* and *nd5*, *nd3* and *trnA*, and *nd5* and *trnQ* (referred here as *trnF-nd5*, *nd3-trnA* and *nd5-trnQ*). Because the *nd3-trnA* region contains *trnH* in F genomes of *L. ornata*, *C. plicata*, *P. grandis*, *Q. quadrula*, and *V. ellipsiformis*, we refer to it as *nd3-trnA+H* in F genomes. Because the *nd5-trnQ* region contains *trnH* in M genomes, we refer to it as *nd5-trnQ+H* in M genomes (see Fig. 1B, C, A and Supplementary Table S1). Interestingly, these three regions correspond to segments of the genome with a change in the direction of transcription (Fig. 1). As will be explained below, this arrangement marks these three regions as potential candidates for control regions (i.e., regions that contains elements involved in the regulation of replication and/or transcription of mtDNA; e.g., BOORE 2006). In contrast, some unassigned regions are unique to either the F or to the M unionoid genome. One relatively large unassigned region specific to the F genomes is found between *trnE-nd2* (the F ORF, see below), except for the F mtDNAs of *H. cumingii* and *I. japonensis*, which possess a different gene order between *cox2* and *12SrRNA* and instead contain their large gender-specific unassigned regions between *trnE-trnW*. All unionoid M genomes possess two relatively large unassigned regions located between *nd4l-trnD* and *atp8-
Further examination of the three shared and gender-specific “F trnE-nd2” (trnE-trnW for F Inversidens and Hyriopsis) and “M nd4l-trnD” unassigned regions reveals two categories of sequences. The first category contains sequences that possess an open reading frame (ORF) of considerable length (i.e., the ORF makes up most of the length of the unassigned sequence) and the second category contains sequences exhibiting many characteristics typically associated with the animal mitochondrial control regions, such as presence of repeat units and sequences that can form stem-loop and hairpin structures. Because the gender-specific regions fall into the former category and the shared unassigned regions fall into the second category, we interpret the latter sequences as potential control regions. The results of the analyses of F- and M-specific ORFs and shared unassigned regions are presented in two different subsections below.

**Gender-specific ORFs in F and M unionoid genomes**

The identification of gender-specific ORFs of significant length in unassigned regions (“F trnE-nd2” [trnE-trnW for F Inversidens and Hyriopsis] and “M trnD-nd4l”; Fig. 1 and 2) is particularly interesting because they could represent regulatory sequences or proteins that could be responsible for the different mode of transmission of the mtDNAs and/or gender-specific adaptive functions of the M and F mtDNA genomes in unionoid bivalves. A role in the differential segregation patterns of sperm mitochondria in mussel embryos has already been proposed for the M-specific cox2 extension that is only present in unionoid bivalve M genomes (BRETON et al. 2007; CHAKRABARTI et al. 2007; CHAPMAN et al. 2008). The unionoid MCOX2 protein, which is extremely variable in both its amino acid sequence and number of near C-terminus transmembrane helices among different species (CUROLE and
KOCHER 2002, 2005; CHAPMAN et al. 2008), has been localized to both inner and outer mitochondrial membranes in sperm (CHAKRABARTI et al. 2007) and appears to function in reproduction (CHAKRABARTI et al. 2006, 2007; CHAPMAN et al. 2008). It has been proposed that the localization of MCOX2 to the outer mitochondrial membranes likely “tags” the outer surface of unionoid M genome-bearing mitochondria and facilitates the distinct movements of the M genome-containing mitochondria, derived from the fertilizing sperm, in male and female embryos (as observed in *Mytilus* [CAO et al. 2004b; COGSWELL et al. 2006]). As is the case for MCOX2 (CHAPMAN et al. 2008), no significant amino acid sequence similarity is detected with known proteins for the two new gender-specific ORFs using BLASTX and PSI-BLAST, and currently their identities/functions remain unclear. The only exception is the M ORF sequence of *V. ellipsiformis*, in which a putative conserved seryl-tRNA synthetase domain has been detected. Specifically, the M ORF of *V. ellipsiformis* exhibits moderate degree of sequence similarity (E-value of 0.003) with the N-terminal nucleotide-binding domain of the seryl-tRNA synthetase, indicating that it could potentially be a DNA- or RNA-binding protein. Interestingly, many of the “extra” ORFs discovered in invertebrate mitochondrial genomes contain amino acid patterns characteristic of interaction with DNA (e.g., PONT-KINGDON et al. 1995, 1998; SHAO et al. 2006; GISSI et al. 2008). Sequence comparisons among M ORFs reveal high variability in length (Fig. 2) and low extent of amino acid sequence similarity (~ 20% similarity between each species pair comparison). Notably, a single transmembrane helix (TMH) is predicted in the 5’ half of each M ORF, and several positively charged amino acids are observed in the region following the predicted TMH (Fig. 2). Amino acid sequence comparisons among F ORFs reveal that a greater degree of similarity exists, compared with M ORFs. The highest identity (~ 60% amino acid identity) is observed between F ORFs from two the most closely-related species (*L. ornata* and *V. ellipsiformis*; for comparison, 72% amino acid identity is found in the highly variable
Between these two species, the divergence at the nucleotide level follows the pattern expected for a protein-coding gene evolving under purifying selection, i.e., where 3rd and 2nd codon position exhibit the largest and smallest numbers of substitutions, respectively. Specifically, 25 (0.082 %), 18 (0.059 %) and 31 (0.102 %) substitutions are identified in the first, second and third codon positions, respectively.

Analyses of F ORF protein sequences indicate variability in length (Fig. 2), yet, a single predicted transmembrane helix is present in a conserved position (i.e., in the 5’ half of the F ORFs; Fig. 2) that is typically followed by a casein kinase II phosphorylation motif (data not shown).

There are multiple lines of evidence indicating functionality of both novel gender-specific ORFs, likely as expressed proteins. First evidence comes from the testcode analysis (FICKETT 1982), which is used to identify coding regions on the basis of non-random nucleotide distributions at the third codon positions within a reading frame. Except for the F ORFs of *I. japanensis* and *P. grandis* (both ORFs having 40 % probability of being protein-coding sequences), all of the unionoid F and M ORFs have been classified as protein-coding or as having > 77 % probability of being protein-coding sequences. More importantly, Western blot analyses, using antibodies generated against peptides synthesized from the predicted F and M ORF amino acid sequences of *V. ellipsiformis*, indicated that both the F- and M-specific proteins are effectively expressed in female and male gonads, respectively (Fig. 3).

To date, all analyzed F and M unionoid genomes were found to harbor an F ORF and M ORF, respectively. The hypothesized universal presence of the two ORFs in unionoid F and M mitochondrial genomes, when viewed in the light of the unionoid fossil record (WATTERS 2001), indicates an origin for the ORFs >100 million years (MY). Again, the retention of these ORFs in the conserved position in multiple distantly related genomes for
over 100 MY, as well as evidence for evolutionary relatedness of two F ORFs from two most closely related species (i.e., 60% identity at the protein level) (CHUNG and SUBBIAH 1996; PEARSON 1996) argue in favor of their functionality. However, as is the case for the M-specific \textit{cox2} extension (CHAPMAN et al. 2008), the original sources of the DNA comprising the F and M ORF regions may remain unidentified given the extremely high substitution rate in these regions and the relatively large amount of time over which these region has accumulated substitutions.

Further, while comparisons of ORFs from distantly related taxa yields percent identity values in the so-called “twilight zone” of 20-25%, it is possible that they retain similar 3D folds (CHUNG and SUBBIAH 1996). Indeed, the prediction of one transmembrane helix in similar positions in all F and M ORFs (i.e., in the 5’ half of the ORFs), which is also a significant support for identifying these ORFs as protein-coding genes, suggests a higher conservation of the secondary structure compared to the primary sequence. Interestingly, the mitochondrially-encoded \textit{ATP8} protein is also characterized by extremely variable length and by a higher conservation of the secondary structure compared to the primary sequence (GISSI et al. 2008). Specifically, typical animal \textit{ATP8} proteins are characterized by a hydrophobic N-terminus domain and a positively charged C-terminus domain (GRAY 1999; GISSI et al. 2008). Because both F and M ORFs possess one TMH followed by positively charged amino acids (Fig. 2), it is tempting to speculate that they originated from within the mitochondrial genome and that the \textit{atp8} gene could have been the original copy, but no significant sequence similarity has been found between the ORFs and metazoan \textit{atp8} sequences. Presently, it is only possible to suggest that selection might maintain the general characteristics (i.e., one TMH followed by positively charged amino acids) of two functional F and M ORFs in the face of an extremely high overall amino acid substitution rate. This was also the suggestion for the M-specific \textit{cox2} extension, which, as mentioned above, appears to have a reproductive
function (Chakrabarti et al. 2006, 2007; Chapman et al. 2008). Given that relatively rapid rates of evolution are frequently observed for proteins involved in reproduction (e.g., Metz et al. 1998; Swanson and Vacquier 2002; Clark et al. 2006), it is possible that the F and M ORFs might also function in reproduction. Further analyses of the “F trnE-nd2” and “M trnD-nd4l” gender-specific unassigned regions in additional species of freshwater bivalves and further protein-based analyses will be necessary to characterize the biological significance of these ORFs.

Identifying the F and M control regions and origins of replication in unionoid bivalves

An attempt has been made to identify the F and M control regions and/or potential signaling elements that are conserved (i) among F, (ii) among M, and/or (iii) among F and M unassigned regions in unionoid bivalves. As mentioned above, there are three shared unassigned regions longer than 20 nucleotides in length in all species that could play crucial roles in replication or transcription of the mtDNA molecule. These include the regions trnF-nd5, nd3-trnA [this region contains trnH in L. ornata, C. plicata, P. grandis, Q. quadrula, and V. ellipsiformis F genomes], and nd5-trnQ [Note: because this region contains trnH in M genomes, we refer to it as nd5-trnQ+H in M genomes]) (see Fig. 1B, C, A).

Typically, the main control region of animal mitochondrial genomes corresponds to (i) the longest non-coding region, (ii) a region in which repetitive elements and secondary structures frequently occur, (iii) a region containing relatively high A+T content and/or (iv) a region associated with abrupt changes in base composition bias (Lewis et al. 1994; Boore 1999; Sacccone et al. 2002; Serb and Lydeard 2003; Cao et al. 2004a; Saito et al. 2005; Kuhn et al. 2006; Oliveira et al. 2007; Brugler and France 2008). In the mytiloid mussel Mytilus, for example, the putative control region has been identified as such because it is the largest non-coding region, and it is capable of producing characteristic secondary structures
(CAO et al. 2004a). In addition, it contains conserved motifs thought to play a crucial role in replication and transcription, and it is broadly similar (with specific nucleotidic regions exhibiting 60 to 90% percent identity) to the mammalian control region (CAO et al. 2004a). To our knowledge, the main control region of the veneroid clam *V. philippinarum* has not yet been localized or described.

*The length, repeats/secondary structure and A+T content criteria*

Based on the length criterion alone, “F nd5-trnQ” is the most likely candidate for the control region in all F genomes while “M nd5-trnQ+H” is the most likely candidate for the control region in all M genomes (except possibly for *P. grandis* where the largest non-coding regions is found in *nd3-trnA* region) (Fig. 1A and Table 2). The lengths observed for “F nd5-trnQ” and “M nd5-trnQ+H” (i.e., from 81 bp for M *P. grandis* to 1196 bp for F *I. japanensis*) are comparable to the size of known metazoan control regions [e.g., 109 bp in deep-sea Bamboo corals (BRUGLER and FRANCE 2008); ~1160 bp for the F mtDNA of *M. edulis* (CAO et al. 2004a)]. The repeats/secondary structures criterion also suggests that the “F nd5-trnQ” and “M nd5-trnQ+H” sequences constitute the main control regions in F and M unionoid mtDNAs. The predicted location of the control region in the large unassigned “F nd5-trnQ” and “M nd5-trnQ+H” regions is particularly well supported by the non-coding features of *I. japanensis* (Table 3). In the F genome of this species, “F nd5-trnQ” is 1196 bp long (i.e., 910 bp longer than the second largest unassigned region in *trnE-trnW*) and it has a higher A+T content (67.5%) compared to the other parts of the genome (Tables 1 and 3; A+T for *trnE-trnW* = 51%). High A+T content is also a characteristic typically used to identify origins of replication (e.g., LEWIS et al. 1994; SERB and LYDEARD 2003; KUHN et al. 2006). There are eight consecutive repeats of a 106-bp element with the potential to form stem-loop structures, with an additional incomplete repeat on each side of the eight-repeat cluster. Similarly, the *I.
“M nd5-trnQ+H” is the largest region (698 bp; i.e., 334 bp longer than the second largest unassigned region between nd4l-trnD) and also is A+T-rich (63 %) compared to the other parts of the genome (Tables 1 and 3; A+T for nd4l-trnD = 58 %). It also harbors five consecutive repeats of a 107-bp element with the potential to form stem-loop structures, with an additional incomplete repeat at the 3’ end of the five-repeat cluster. Likewise, “F nd5-trnQ” and “M nd5-trnQ+H” are the longest unassigned regions of the F and M genomes of Q. quadrula (346 bp and 555 bp long, respectively) and both regions contain stem-loop structures and are A+T-rich compared to the other part of the genomes (65 % and 70 % for “F nd5-trnQ” and “M nd5-trnQ+H” respectively). Two consecutive repeats of a 98 nucleotide element are found in Q. quadrula “M nd5-trnQ+H”, while short repetitive sequences (< 8 bp) are observed in the F genome of the same species (Table 3). Short repetitive identical fragments are also found in all other “F nd5-trnQ” and “M nd5-trnQ+H” regions of unionoid genomes (Table 3). Overall, the extent of nucleotide divergence observed between both regions within each species (43-50 %) is slightly higher than that observed for the other parts of the genomes (Tables 1 and 3). Interspecifically, the lowest amount of nucleotide sequence divergence (31-35 %) has been detected between the two most closely-related species (i.e., between the F unassigned regions “nd3-trnH”, “trnF-nd5” and “nd5-trnQ” of L. ornata and V. ellipsiformis). For the more distantly-related species, pairwise comparisons within each gender revealed, as expected, higher levels of nucleotide sequence divergence (> 45 %).

Identification of conserved motifs or regions with high similarity to the other DUI-species control regions

Previous analyses of the mytiloid mussel Mytilus have shown that the F and M main control regions can be divided into three domains on the basis of indels and patterns of nucleotide variation (CAO et al. 2004a; CAO et al. 2009). The middle domain of the control
region encodes a hairpin structure and is the most slowly evolving part of the mitochondrial genome. In contrast, the first and last domains are among the most divergent parts of the M and F genomes. It has been suggested that this tripartite structure, which is also a characteristic of the mammalian control region, demonstrates that different parts of the control region evolve under different selective constraints (CAO et al. 2004a; CAO et al. 2009). In freshwater mussels (Unionoida), the high degree of sequence divergence observed among species and also between intraspecific F and M genomes presents a challenge for characterizing the structure of any of the shared unassigned regions. Alignments indicate that the portion of the shared “F nd5-trnQ” and “M nd5-trnQ+H” unassigned region closest to the trnQ gene is the most conserved part of the sequence. Interestingly, this portion consistently contains hairpin or stem-loop structures in all F and M genomes examined here, but neither the flanking regions nor the morphology of the structures appear to be conserved (Supplementary Fig. S1). The identification of motifs of significant sequence similarity (> 60 %) with elements known to have specific functions in the sea urchin and the mammalian control region was possible in the mytiloid mussel Mytilus (CAO et al. 2004a). However, this approach has not been successful in defining potential regulatory elements in unionoid genomes. We performed a search for possible signaling elements by comparing “F nd5-trnQ” and “M nd5-trnQ+H” sequences for any block of 10 nucleotides or longer with nucleotide identity of at least 65 % (e.g., BOORE 2006). Apart from the A+T-rich region capable of forming stem-loop structures and several (A)_n, (G)_n and (T)_n homopolymer tracts, we found no conserved motif within each gender or between “F nd5-trnQ” and “M nd5-trnQ+H”. Similarly, the search for possible signaling elements in other long, shared regions (i.e., between nd3-trnA [nd3-trnH and trnH-trnA in F genomes] and trnF-nd5) did not reveal any conserved motif. These results indicate that F and M unionoid genomes might share more similar organization in this region at the level of secondary structures despite low extent of
similarity identifiable at the nucleotide level. In other words, a high level of DNA sequence variability might be compensated for by specific secondary structure configurations, with homopolymer tracts and stem-loop structures, which might be responsible for the regulation of mtDNA replication.

*Using mitochondrial AT-skews to identify the F and M origins of replication in unionoid bivalves*

As previously mentioned, a consequence of the asymmetrical model of mtDNA replication, which is thought to occur in most metazoans, is that during replication of the molecule different genes will be exposed to the single-strand state (more susceptible to mutation) for different lengths of time, depending on their position on the genome (CLAYTON 1982). Recently, RODAKIS *et al.* (2007) observed a positive correlation between single-strand state duration and nucleotide composition bias (AT-skews) at the less constrained protein-coding gene positions (fourfold degenerate sites) in the F and M mitochondrial genomes of the mytiloid mussel *Mytilus*, indicating that the replication proceeds via the asymmetric strand-displacement model, with the origins of heavy (**O**<sub>H</sub>) and light (**O**<sub>L</sub>) strand replication separated by ~ 2/3 of the genome.

Herein, we use mitochondrial AT-skews to help identify the F and M origins heavy and light strand replication in unionoid bivalves. If the asymmetrical model of mtDNA replication also applies to unionoid bivalves and if their mtDNAs are also exposed to the same type of mutations as are vertebrate mtDNAs, then we hypothesize that **O**<sub>H</sub> would be located in the vicinity of the protein-coding genes showing the greatest AT-skew values at fourfold redundant sites, while **O**<sub>L</sub> would be located in the vicinity of the protein-coding genes showing the lowest AT-skew values. Figure 3 displays AT-skew values at fourfold redundant sites for 12 F and M mitochondrial genes (we excluded atp8) within each species
whose F and M genomes have been completely sequenced, and the deduced locations of the putative \( O_H \) and \( O_L \) in each genome. According to our combined skew analyses, mitochondrial control regions are not well defined and their locations could be variable among unionoid bivalve mitochondrial genomes. In \( I. japonensis \), \( O_H \) would be located between “F nd5-trnQ” in the F genome and “M nd5-trnQ+H” in the M genome, a result that corresponds to our predictions based on the length, A+T content and repeat/secondary structures criteria. Indeed, the heavy strand encoded \( nd5 \) and the light strand encoded \( nd1 \) and \( nd6 \) show the greatest AT-skew values (negative or positive) in both F and M \( I. japonensis \) genomes, indicating that this region would be exposed to mutagenic pressures for the longest time and thus we can infer that it is located near the \( O_H \) (Fig. 4A). In the F genome of this species, a marked difference is observed between the skew values of the heavy strand encoded \( nd3 \) and all other heavy strand encoded genes. The low AT-skew value in \( nd3 \), together with the low AT-skew value of the light strand encoded \( nd2 \) gene suggest that \( O_L \) would be located between \( nd3 \) and \( cox2 \) with the direction of the L strand synthesis toward \( nd2 \) in \( I. japonensis \) F mtDNA. It should be noted that a similar pattern is observed for the F genome of \( H. cumingii \) (data not shown). Interestingly, this region (i.e., between \( trnH \) and \( trnS2 \)) contains a non-coding region on the heavy strand capable of forming a stem-loop structure similar to that of other metazoan \( O_L S \) (e.g., \textit{RODAKIS et al. 2007; SELIGMANN 2008}) (Supplementary Figure S2A and B). No clear skew pattern is observed for the M genome of \( I. japonensis \), therefore, the location of its \( O_L \) remains unclear.

The position of \( O_L \) between \( nd3 \) and \( cox2 \) in the F genomes of \( I. japonensis \) and \( H. cumingii \) marks the location of the sole observed gene rearrangement among F unionoid genomes (\( I. japonensis \) and \( H. cumingii \) have a different gene order in this segment compared to other F mtDNAs; see Fig. 1 and 4). It has previously been demonstrated that novel gene orders occur more frequently with movement of \( O_L \) than would be expected by chance.
(Macey et al. 1997). Apparently, the region encompassing \textit{cox2-nd3-nd2} is not associated with O\textsubscript{L} in the F genomes of \textit{Q. quadrula}, \textit{V. ellipsiformis} and \textit{P. grandis} (Fig. 4B and C; a similar pattern is observed for the F genome of \textit{L. ornata} and \textit{C. plicata} [data not shown]). The rather high skew values observed for \textit{nd3} and \textit{cox2} suggest that this region would remain exposed to mutagenic pressures for a long time and could be associated with the O\textsubscript{H} in these genomes. The same pattern is also observed for the M genomes of \textit{Q. quadrula} and \textit{V. ellipsiformis}, but not for \textit{P. grandis} (Fig. 4B and C). Although shorter in length as compared with “F \textit{nd5-trnQ}” and “M \textit{nd5-trnQ+H}”, the unassigned regions located between \textit{nd3-trnA} in M genomes and \textit{trnH-trnA} in F genomes are also favourable candidates that could be involved in replication initiation. Indeed, they possess characteristics typically used to identify origins of replication, i.e., they are A+T rich, they are capable of forming stem-loop and hairpin structures, and they contain several (A)\textsubscript{n}, (G)\textsubscript{n} and (T)\textsubscript{n} homopolymer tracts (Table 4).

Interestingly, two putative O\textsubscript{L}s can be detected in the F genomes of \textit{Q. quadrula} and \textit{V. ellipsiformis} (Fig. 4B). According to its very low AT-skew value, one O\textsubscript{L} would be located in the vicinity of the \textit{cytb} gene. If we assume that the H strand synthesis is unidirectional and similar to that of the \textit{I. japonensis} F genome, a “unidirectional” O\textsubscript{L} could only be located within the mitochondrial tRNA gene cluster \textit{trnL1-trnN-trnP} or in the 16SrRNA and the direction of the L strand synthesis is toward \textit{cytb} (see Fig. 1 and 4B). Seligmann et al. (2006) reported that heavy strand encoded tRNA genes can sometimes function as an O\textsubscript{L} by forming alternative secondary structures other than their classical cloverleaf structures. The possibility therefore exists that the heavy strand “mirror” sequences of \textit{trnL1-trnN-trnP} could serve as origins of light strand replication in the F genomes of \textit{Q. quadrula} and \textit{V. ellipsiformis} (Seligmann et al. 2006; Seligmann 2008). It is of interest to note that the prominent difference in AT-skew value for the \textit{cytb} gene has not been observed in the M genomes of \textit{Q. quadrula}. 
Quadrula and *V. ellipsiformis*, in the F genome of *C. plicata* (data not shown) or in the F or M genome of *P. grandis* (Fig. 4). This is consistent with the previous suggestion that an alternative O_L can frequently evolve and disappear in mtDNA sequences (Seligmann *et al.* 2006). Alternatively, our results can be explained by (i) a putative bidirectional O_L in the unassigned region *trnF-nd5* (Fig. 1 and 4B; bidirectional replication has previously been demonstrated in mitochondrial genomes of vertebrates [Reyes *et al.* 2005; Seligmann *et al.* 2006]), and/or (ii) differences in transcription processes and/or selective constraints experienced by the F genomes of *V. ellipsiformis* and *Q. quadrula*.

According to our results, a second putative O_L in the F genomes of *Q. quadrula* and *V. ellipsiformis* would be located in the vicinity of the *atp6* gene, a region that corresponds to the gene order inversion of *trnD* and *atp8* and which represents one of the two organizational differences between F and M genomes in unionoid bivalves (the other difference is the transposition of the *trnH* gene; see Fig. 1 and 4B). The M genomes of the same species do not share a similar pattern (i.e., the O_L is not clearly indicated in these M genomes), but the results obtained for the F genome of *C. plicata* (data not shown) and the F and M genomes of *P. grandis* also suggest the presence of a putative O_L in the vicinity of *atp6* (Fig. 4B and C). Again, our results are in line with and confirm previous studies showing that mitochondrial regions near origins of replication are hot spots for rearrangement and that gene rearrangements occur more frequently with displacement of O_L than expected by chance (e.g., Macey *et al.* 1997; Mueller and Boore 2005; Brugler and France 2008). We also note that the AT-skew values suggest that replication could proceed in both directions from this O_L (Fig. 4B and C). Because no unassigned region is found between *atp6* and *nd4l* in the F genome of *V. ellipsiformis*, one likely candidate for an O_L in this species would be the heavy strand encoded *trnD* gene. Consistent with the observations of Seligmann *et al.* (2006), the F-type *trnD* (but not the M-type) of *V. ellipsiformis* is able to form an O_L-like structure,
thereby suggesting it could have an alternative $O_L$ function (Supplementary Figure S2C and D, respectively). A similar result is also observed for the $trnD$ of the F genomes of $Q. quadrula$ and $P. grandis$, but not for the M genome of $P. grandis$ (data not shown). In the $P. grandis$ M genome, an $O_L$-like structure is found in the unassigned region between $trnD$ and the truncated $atp8$. Moreover, this genome could possess a second putative $O_L$ in the vicinity of the $nd2$ gene (Fig. 4). Once more, this suggests that an alternative $O_L$ can frequently evolve and disappear in mtDNA sequences (SELGANN et al. 2006).

Overall, our AT-skew values suggest that the classical asymmetrical model of mtDNA replication, with one $O_H$ and one $O_L$, might not apply to all unionoid bivalves. Our data suggest that the locations of the origins of replication in these species are variable, there are potentially multiple locations, and locations could be uni- or bidirectional. Other studies of distantly-related invertebrate taxa also reported that control region locations can be highly variable (e.g., insect species [SAITO et al. 2005] and corals [BRUGLER and FRANCE 2008; CHEN et al. 2008]). Furthermore, multiple and bidirectional mitochondrial origins of replication have been previously suggested for vertebrates (REYES et al. 2005). However, our results could also be associated with differences in transcription processes and/or selective constraints experienced by the F and M genomes, leading to potential misinterpretation of the localizations of the $O_H$ and $O_L$. According to the length, A+T content and repeat/secondary structures criteria, “F $nd5-trnQ$” and “M $nd5-trnQ+H$” remain the most likely candidate $O_H$ control regions for regulating replication and/or transcription (i.e., origin of replication, initiation or termination sites for transcription) in unionoid bivalves. Additional studies, both functional and comparative, will be required to determine the precise position of the mitochondrial origins of replication in unionoid bivalves.

Conclusion
The description and comprehensive analysis of complete F and M mitochondrial
genomes of bivalve species with DUI, more specifically the unassigned regions of unionoid
species, have led to new insights on the mitogenomics of species with DUI. Our results reveal
that at least three shared regions in F and M unionoid genomes could contain regulatory
signals involved in the replication and/or transcription of mtDNA. According to the length,
A+T content and repeat/secondary structures criteria, likely candidate regions for the O_H
origin of replication would be “F nd5-trnQ” and “M nd5-trnQ+H”. AT-skew values of
protein-coding genes at fourfold degenerate sites have also been used to identify the location
of O_H and O_L control regions. Our results reveal that (i) two regions (i.e., “F nd5-trnQ” and
“M nd5-trnQ+H” and “M nd3-trnA” and “F trnH-trnA”) are potential O_H control regions for
regulating replication and/or transcription, and (ii) multiple and potentially bidirectional O_L
origins of replication are present in unionoid F and M mitochondrial genomes. In other words,
unionoid mitochondrial control regions are seldom well defined using skew values and their
locations could be variable among genomes and species.

Finally, although uncharacterized mitochondrial ORFs are essentially absent in
vertebrates, their presence in other animal groups is not unprecedented (BURGER et al. 2003a;
GISSI et al. 2008). For example, ORFs that exhibit no significant sequence similarity to
known proteins have been recently discovered in the Cnidaria and Porifera (e.g., SHAO et al.
2006; FLOT and TILLIER 2007; WANG and LAVROV 2008). Such metazoans ORFs could be (i)
homologous to ancestral bacterial protein-coding genes, (ii) the product of mitochondrial gene
duplication events or (iii) the result of DNA transferred from nuclear genomes. However,
testing the above hypotheses will likely be difficult due to the ORFs' highly divergent
sequences (BURGER et al. 2003b). Analyses of complete mitochondrial genomes from
additional bivalves species, particularly basal taxa, and further protein-based studies are
needed to elucidate the number, taxonomic distribution, evolution and function of
uncharacterized ORFs in this group as well as the molecular features that are associated with the developmental regulation and transmission genetics of DUI.

SUPPLEMENTARY MATERIAL

Table S1 of number of nucleotides at gene boundaries in F and M mitochondrial genomes of freshwater bivalves (Bivalvia: Unionidae). Figure S1 of Hairpin and stem-loop structures located in the portion of “F nd5-trnQ” and “M nd5-trnQ+H” closest to trnQ. Figure S2 of Secondary structures for potential origin of light strand replication (OL) in unionoid mussels.

ACKNOWLEDGMENTS

This study was supported by research grants from the National Science Foundation (NSF, to W.R. Hoeh), the National Sciences and Engineering Research Council (NSERC, to P. U. Blier and D. T. Stewart) and by A.E. Bogan’s research funds. S. Breton and H. Doucet Beaupré were respectively financially supported by NSERC fellowship and FQRNT scholarship. R. Shandilya and S. Vijayaraghavan are thanked for their assistance in the laboratory.

REFERENCES

AKASAKI, T., M. NIKAIDO, K. TSUCHIYA, S. SEGAWA, M. HASEGAWA and N. OKADA, 2006


BOORE, J. L., 2006 The complete sequence of the mitochondrial genome of Nautilus macromphalus (Mollusca: Cephalopoda). BMC Genomics 7: 182.


CAO, L., E. KENCHINGTON, E. ZOUROS and G. C. RODAKIS, 2004a Evidence that the large noncoding sequence is the main control region of maternally and paternally transmitted mitochondrial genomes of the marine mussel (Mytilus spp.). Genetics 167: 835-850.


THEOLOGIDIS, I., C. SAAVEDRA and E. ZOUROS, 2007 No evidence for absence of paternal mtDNA in male progeny from pair matings of the mussel *Mytilus galloprovincialis*. Genetics **176**: 1367-1369.


Jalview Version 2 - a multiple sequence alignment editor and analysis workbench.
Bioinformatics 25: 1189-1191.

WATTERS, G. T., 2001 The evolution of the Unionacea in North America, and its implications
for the worldwide fauna, pp. 281-307 in Ecology and Evolution of the Freshwater

WHITE, D. H., J. N. WOLFF, M. PIERSON and N. J. GEMMELL, 2008 Revealing the hidden

ZBAWICKA, M., A. BURZYŃSKI and R. WENNE, 2007 Complete sequences of mitochondrial

ZOUROS, E., 2000 The exceptional mitochondrial DNA system of the mussel family

ZUKER, M., 2003 Mfold web server for nucleic acid folding and hybridization prediction.
Nucl. Acids Res. 31: 3406-341.
FIGURE LEGENDS

**Figure 1.** Gene maps of the gender-associated mitochondrial genomes of unionoid mussels. Gene identities: *nd1-6* and *nd4l*, NADH dehydrogenase subunits 1-6 and 4L (Complex I in green); *cytb*, cytochrome b (Complex III in light blue); *cox1-3*, cytochrome c oxidase subunits I-III (Complex IV in blue); *atp6* and *atp8*, ATP synthase subunits 6 and 8 (Complex V in light purple); *12SrRNA* and *16SrRNA*, small and large subunits of ribosomal RNA (in purple). Transfer RNA genes (in grey) are depicted by one-letter amino acid codes; L1, L2, S1, and S2 are differentiated by their anticodon sequences: CUA, UAA, AGA and UCA, respectively. F ORF, F-specific open reading frame (orange); M ORF, M-specific open reading frame (red). Genes positioned inside the plain circle are transcribed counter-clockwise and genes outside the circle are transcribed clockwise. Arrows A, B and C indicate shared unassigned regions > 20 bp between F and M unionoid genomes (i.e., A = *nd5-trnQ* [this region contains *trnH* in M genomes]); B = *trnF-nd5*; C = between *nd3-trnA* [this region contains *trnH* in *P. grandis*, *Q. quadrula*, and *V. ellipsiformis* F genomes]).

**Figure 2.** Amino acid sequences of the potential peptides encoded by gender-specific open reading frames (F and M ORFs). The amino acids that constitute the putative transmembrane helix are indicated in bold and bigger characters. Positively charged amino acids are in red. The synthesized antigenic peptides used to generate antibodies for the Western blots are underlined.

**Figure 3.** Expression of the F and M ORFs in female and male gonads, respectively, from *V. ellipsiformis*. Western Blots revealed with anti-F ORF and anti-M ORF antibodies. The origins of the tissue samples for each lane are indicated in the figure. The positions of marker proteins (kDa) are shown.

**Figure 4.** AT-skew values in 12 mitochondrial genes (excluding *atp8*) of unionoid mussels and the deduced locations of the putative O_H and O_L in each genome. Because they share a similar pattern, *V. ellipsiformis* and *Q. quadrula* have been analysed together. The big black arrow indicates the unassigned region located between “F *nd5-trnQ*” and “M *nd5-trnQ+H*” and the potential direction of DNA synthesis, whereas the small black arrow indicates the unassigned region located between “M *nd3-trnA*” and “F *trnH-trnA*” and the potential direction of DNA synthesis. A = *Inversidens japanensis*; B = *Quadrula quadrula* and *V. ellipsiformis*; C = *Pyganodon grandis*.

**Supplementary Figure S1.** Hairpin and stem-loop structures located in the portion of “F *nd5-trnQ*” and “M *nd5-trnQ+H*” closest to *trnQ*.

**Supplementary Figure S2.** Secondary structures for (A) a classical origin of light strand replication (O_L) in metazoans, (B) the unassigned region between *trnH* and *trnS2* associated to O_L in the F genome of *I. japanensis*, (C) the alternative O_L-like structure for *trnD* in the F genomes of *V. ellipsiformis*, *Q. quadrula* and *P. grandis*, (D) the typical cloverleaf structure for *trnD* in the M genomes of *V. ellipsiformis*, *Q. quadrula* and *P. grandis*. 
Table 1. Gender-Associated Mitochondrial Genomes in Unionoid Bivalve Species with DUI

<table>
<thead>
<tr>
<th>Species</th>
<th>Gender</th>
<th>Genome Size</th>
<th>A+T Content</th>
<th>% Coding + RNA</th>
<th>GenBank Accession Number</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Freshwater mussels</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(Unionoida)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>C. plicata</em></td>
<td>Female</td>
<td>15,712 bp</td>
<td>63.8 %</td>
<td>93.0 %</td>
<td>FJ986302</td>
<td>WP JIANG, JL LI and RL ZHENG personal communication</td>
</tr>
<tr>
<td><em>H. cumingii</em></td>
<td>Female</td>
<td>15,954 bp</td>
<td>59.5 %</td>
<td>91.6 %</td>
<td>FJ529186</td>
<td>RL ZHENG and JL LI pers. comm.</td>
</tr>
<tr>
<td><em>I. japonensis</em></td>
<td>Female</td>
<td>16,826 bp</td>
<td>57.2 %</td>
<td>86.7 %</td>
<td>AB055625</td>
<td>M OKAZAKI and R UESHIMA pers. comm.</td>
</tr>
<tr>
<td><em>I. japonensis</em></td>
<td>Male</td>
<td>16,966 bp</td>
<td>57.2 %</td>
<td>91.0 %</td>
<td>AB055624</td>
<td>M OKAZAKI and R UESHIMA pers. comm.</td>
</tr>
<tr>
<td><em>L. ornate</em></td>
<td>Female</td>
<td>16,060 bp</td>
<td>62.4 %</td>
<td>92.0 %</td>
<td>AY365193</td>
<td>SERB and LYDEARD 2003</td>
</tr>
<tr>
<td><em>P. grandis</em></td>
<td>Female</td>
<td>15,848 bp</td>
<td>64.2 %</td>
<td>90.8 %</td>
<td>FJ809754</td>
<td>H DOUCET BEAUPRE <em>et al.</em> unpublished data</td>
</tr>
<tr>
<td><em>P. grandis</em></td>
<td>Male</td>
<td>17,071 bp</td>
<td>64.8 %</td>
<td>87.7 %</td>
<td>FJ809755</td>
<td>H DOUCET BEAUPRE <em>et al.</em> unpublished data</td>
</tr>
<tr>
<td><em>Q. quadrula</em></td>
<td>Female</td>
<td>16,033 bp</td>
<td>62.6 %</td>
<td>91.5 %</td>
<td>FJ809750</td>
<td>H DOUCET BEAUPRE <em>et al.</em> unpublished data</td>
</tr>
<tr>
<td><em>Q. quadrula</em></td>
<td>Male</td>
<td>16,970 bp</td>
<td>62.1 %</td>
<td>90.7 %</td>
<td>FJ809751</td>
<td>H DOUCET BEAUPRE <em>et al.</em> unpublished data</td>
</tr>
<tr>
<td><em>V. ellipsiformis</em></td>
<td>Female</td>
<td>15,975 bp</td>
<td>62.6 %</td>
<td>92.2 %</td>
<td>FJ809753</td>
<td>H DOUCET BEAUPRE <em>et al.</em> unpublished data</td>
</tr>
<tr>
<td><em>V. ellipsiformis</em></td>
<td>Male</td>
<td>17,174 bp</td>
<td>63.5 %</td>
<td>88.2 %</td>
<td>FJ809752</td>
<td>H DOUCET BEAUPRE <em>et al.</em> unpublished data</td>
</tr>
</tbody>
</table>
### Table 2. Major Unassigned Regions in F and M Mitochondrial Genomes of Freshwater Mussels (Bivalvia: Unionoida)

<table>
<thead>
<tr>
<th>Species</th>
<th>Gender</th>
<th>F $nd5$-$trnQ$</th>
<th>F $nd3$-$trnA$</th>
<th>F $trnF$-$nd5$</th>
<th>F $trnE$-$nd2$ (F ORF)</th>
<th>F $trnE$-$trnW$ (F ORF)</th>
<th>M $nd4$-$trnD$ (M ORF)</th>
<th>M $atp8$-$atp6$</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>C. plicata</em></td>
<td>F</td>
<td>289 bp</td>
<td>77 bp</td>
<td>42 bp</td>
<td>329 bp</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td><em>H. cumingii</em></td>
<td>F</td>
<td>202 bp</td>
<td>-</td>
<td>44 bp</td>
<td>-</td>
<td>423 bp</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td><em>I. japanensis</em></td>
<td>M</td>
<td>698 bp</td>
<td>91 bp</td>
<td>34 bp</td>
<td>-</td>
<td>-</td>
<td>364 bp</td>
<td>112 bp</td>
</tr>
<tr>
<td><em>L. ornata</em></td>
<td>F</td>
<td>247 bp</td>
<td>130 bp</td>
<td>135 bp</td>
<td>282 bp</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td><em>P. grandis</em></td>
<td>F</td>
<td>450 bp</td>
<td>71 bp</td>
<td>56 bp</td>
<td>305 bp</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td><em>P. grandis</em></td>
<td>M</td>
<td>81 bp</td>
<td>248 bp</td>
<td>36 bp</td>
<td>-</td>
<td>1103 bp</td>
<td>58 bp</td>
<td>-</td>
</tr>
<tr>
<td><em>Q. quadrula</em></td>
<td>F</td>
<td>346 bp</td>
<td>79 bp</td>
<td>51 bp</td>
<td>287 bp</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td><em>Q. quadrula</em></td>
<td>M</td>
<td>555 bp</td>
<td>62 bp</td>
<td>58 bp</td>
<td>-</td>
<td>325 bp</td>
<td>65 bp</td>
<td>-</td>
</tr>
<tr>
<td><em>V. ellipsiformis</em></td>
<td>F</td>
<td>308 bp</td>
<td>88 bp</td>
<td>68 bp</td>
<td>283 bp</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td><em>V. ellipsiformis</em></td>
<td>M</td>
<td>120 bp</td>
<td>93 bp</td>
<td>33 bp</td>
<td>-</td>
<td>833 bp</td>
<td>54 bp</td>
<td>-</td>
</tr>
</tbody>
</table>

**NOTE.**—Shared unassigned regions are in bold. a The F genomes of *I. japanensis* and *H. cumingii* possess a different gene order in this region and are not included in this table. Moreover, this region contains *trnH* in *L. ornata, C. plicata, P. grandis, Q. quadrula,* and *V. ellipsiformis* F genomes, the length has been calculated for the segment *trnH-trnA* in these genomes. b *I. japanensis* and *H. cumingii* F genomes have a different gene order between *cox2* and *trnW* and are thus the only genomes to possess a major unassigned region between *trnE-trnW*. 
Table 3. General Characteristics of “F \(nd5-trnQ\)” and “M \(nd5-trQ+H\)”

<table>
<thead>
<tr>
<th>Species/Taxon</th>
<th>Length</th>
<th>A+T Content</th>
<th>Number of Repetitive elements (&gt; 5) bp</th>
<th>Longest Repetitive element</th>
<th>Copy Number of Repetitive Elements</th>
<th>Other Characteristics</th>
<th>F/M Divergence</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>C. plicata</strong> F</td>
<td>289 bp</td>
<td>71.1 %</td>
<td>8</td>
<td>8 bp</td>
<td>2 - 8</td>
<td>Stem-loop &amp; hairpin structures, ((A)n, (G)n, (T)n)</td>
<td>-</td>
</tr>
<tr>
<td><strong>H. cumingii</strong> F</td>
<td>202 bp</td>
<td>69.0 %</td>
<td>5</td>
<td>10 bp</td>
<td>2-3</td>
<td>Stem-loop structures</td>
<td>-</td>
</tr>
<tr>
<td><strong>I. japonensis</strong> F</td>
<td>1196 bp</td>
<td>67.5 %</td>
<td>1</td>
<td>106 bp</td>
<td>9</td>
<td>Stem-loop structures</td>
<td>43 %</td>
</tr>
<tr>
<td><strong>I. japonensis</strong> M</td>
<td>698 bp</td>
<td>63.0 %</td>
<td>1</td>
<td>107 bp</td>
<td>5</td>
<td>Stem-loop structures, ((G)n)</td>
<td></td>
</tr>
<tr>
<td><strong>L. ornata</strong> F</td>
<td>247 bp</td>
<td>64.4 %</td>
<td>6</td>
<td>7 bp</td>
<td>2 - 3</td>
<td>Hairpin structure, ((A)n)</td>
<td>-</td>
</tr>
<tr>
<td><strong>P. grandis</strong> F</td>
<td>450 bp</td>
<td>70.9 %</td>
<td>10</td>
<td>23 bp</td>
<td>2 - 3</td>
<td>Stem-loop &amp; hairpin structures, ((A)n, (G)n)</td>
<td>43 %</td>
</tr>
<tr>
<td><strong>P. grandis</strong> M</td>
<td>81 bp</td>
<td>76.5 %</td>
<td>2</td>
<td>7 bp</td>
<td>2</td>
<td>Stem-loop structures, ((A)n)</td>
<td>43 %</td>
</tr>
<tr>
<td><strong>Q. quadrula</strong> F</td>
<td>346 bp</td>
<td>65.0 %</td>
<td>17</td>
<td>8 bp</td>
<td>2 - 5</td>
<td>Stem-loop &amp; hairpin structures</td>
<td>49 %</td>
</tr>
<tr>
<td><strong>Q. quadrula</strong> M</td>
<td>555 bp</td>
<td>70.3 %</td>
<td>33</td>
<td>98 bp</td>
<td>2 - 7</td>
<td>Stem-loop &amp; hairpin structures, ((A)n, (T)n)</td>
<td>50 %</td>
</tr>
<tr>
<td><strong>V. ellipsoidformis</strong> F</td>
<td>308 bp</td>
<td>67.9 %</td>
<td>5</td>
<td>7 bp</td>
<td>2 - 3</td>
<td>Hairpin structure, ((A)n)</td>
<td></td>
</tr>
<tr>
<td><strong>V. ellipsoidformis</strong> M</td>
<td>120 bp</td>
<td>70.0 %</td>
<td>4</td>
<td>10 bp</td>
<td>2 - 4</td>
<td>Hairpin structure, ((G)n, (T)n)</td>
<td></td>
</tr>
</tbody>
</table>

**NOTE.**—Sequence divergences are given in percentages for the total number of aligned nucleotides. Stretches of nucleotides are considered when \(> 6\) bp.
Table 4. General characteristics of “F trnH-trnA” and “M nd3-trnA”

<table>
<thead>
<tr>
<th>Species/Taxon</th>
<th>Length</th>
<th>A+T Content</th>
<th>Number of Repetitive elements &gt; 5 bp</th>
<th>Longest Repetitive element</th>
<th>Copy Number of Repetitive Elements</th>
<th>Other Characteristics</th>
<th>F/M Divergence</th>
</tr>
</thead>
<tbody>
<tr>
<td>C. plicata F</td>
<td>77 bp</td>
<td>80.8 %</td>
<td>1</td>
<td>5 bp</td>
<td>3 - 6</td>
<td>Hairpin &amp; Stem-loop structures, (A)_n, G)_n</td>
<td>-</td>
</tr>
<tr>
<td>I. japonensis M</td>
<td>91 bp</td>
<td>62.6 %</td>
<td>6</td>
<td>6 bp</td>
<td>2 - 3</td>
<td>Hairpin &amp; Stem-loop structures</td>
<td>-</td>
</tr>
<tr>
<td>L. ornata F</td>
<td>130 bp</td>
<td>66.7 %</td>
<td>3</td>
<td>7 bp</td>
<td>2 - 4</td>
<td>Hairpin &amp; Stem-loop structures, (G)_n</td>
<td>-</td>
</tr>
<tr>
<td>P. grandis F</td>
<td>71 bp</td>
<td>93.0 %</td>
<td>1</td>
<td>8 bp</td>
<td>2</td>
<td>Hairpin structure, (A)_n, (T)_n</td>
<td>39.6 %</td>
</tr>
<tr>
<td>P. grandis M</td>
<td>248 bp</td>
<td>75.4 %</td>
<td>4</td>
<td>10 bp</td>
<td>2 - 3</td>
<td>Hairpin &amp; Stem-loop structures, (A)_n, (G)_n, (T)_n</td>
<td>53.7 %</td>
</tr>
<tr>
<td>Q. quadrula F</td>
<td>79 bp</td>
<td>68.4 %</td>
<td>3</td>
<td>8 bp</td>
<td>2</td>
<td>Hairpin &amp; Stem-loop structures</td>
<td>52.8 %</td>
</tr>
<tr>
<td>Q. quadrula M</td>
<td>62 bp</td>
<td>85.5 %</td>
<td>1</td>
<td>5 bp</td>
<td>2</td>
<td>Hairpin &amp; Stem-loop structures</td>
<td></td>
</tr>
<tr>
<td>V. ellipsiformis F</td>
<td>88 bp</td>
<td>69.3 %</td>
<td>4</td>
<td>7 bp</td>
<td>2 - 4</td>
<td>Hairpin &amp; Stem-loop structures</td>
<td></td>
</tr>
<tr>
<td>V. ellipsiformis M</td>
<td>93 bp</td>
<td>67.7 %</td>
<td>3</td>
<td>6 bp</td>
<td>2</td>
<td>Hairpin &amp; Stem-loop structures</td>
<td></td>
</tr>
</tbody>
</table>

NOTE.— Sequence divergences are given in percentages for the total number of aligned nucleotides. Stretches of nucleotides are considered when > 6 bp. The F genomes of I. japonensis and H. cumingii possess a different gene order in this region and are not included in this table.
Inversidens japonensis & Hyriopsis cumingii
F genomes

Typical unionoid
F genome

F nd5-trnQ
F trnF-nd5

Typical unionoid
M genome

M nd5-trnQ+H
M trnF-nd5

C F nd3-trnH+A
C M nd3-trnA
FORF

*C. plicata* (83 a.a.)

**C. plicata**

<table>
<thead>
<tr>
<th>MSKKA</th>
<th>LNL</th>
<th>FLI</th>
<th>IIU</th>
<th>TSL</th>
<th>TTQ</th>
</tr>
</thead>
<tbody>
<tr>
<td>TTNP</td>
<td>L</td>
<td>DEF</td>
<td>W</td>
<td>MT</td>
<td>NQ</td>
</tr>
<tr>
<td>STELN</td>
<td>N</td>
<td>K</td>
<td>T</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**H. cumingii** (89 a.a.)

**H. cumingii**

| MLTT | LTT | LLI | LLI | T | TF | L | M | FF | T | LS | AY | H | D | F | L | S | L | E | T | D | P | E | M | A | G | E | S | Q | K |
|-------|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| TTT | N | S | G | D | T | P | S | T | S | N | H | N | L | V | A | S | P | G | H | T |
| NIS | Q | TAP | PNT | P | Q | K |

**I. japonensis** (66 a.a.)

**I. japonensis**

<table>
<thead>
<tr>
<th>I</th>
<th>L</th>
<th>GL</th>
<th>C</th>
<th>L</th>
<th>LL</th>
<th>CG</th>
<th>I</th>
<th>Y</th>
<th>HG</th>
<th>P</th>
<th>A</th>
<th>N</th>
<th>S</th>
</tr>
</thead>
</table>

**L. ornata** (92 a.a.)

**L. ornata**

<table>
<thead>
<tr>
<th>IT</th>
<th>P</th>
<th>M</th>
<th>V</th>
<th>K</th>
<th>T</th>
<th>K</th>
<th>T</th>
<th>R</th>
<th>IM</th>
<th>S</th>
<th>M</th>
<th>S</th>
<th>K</th>
<th>H</th>
<th>T</th>
<th>M</th>
<th>Y</th>
</tr>
</thead>
<tbody>
<tr>
<td>K</td>
<td>T</td>
<td>A</td>
<td>A</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>M</td>
<td>L</td>
<td>L</td>
<td>M</td>
<td>L</td>
<td>I</td>
<td>L</td>
<td>S</td>
<td>N</td>
<td>P</td>
<td>F</td>
</tr>
<tr>
<td>V</td>
<td>M</td>
<td>M</td>
<td>P</td>
<td>K</td>
<td>V</td>
<td>P</td>
<td>Y</td>
<td>T</td>
<td>E</td>
<td>L</td>
<td>S</td>
<td>L</td>
<td>D</td>
<td>N</td>
<td>P</td>
<td>L</td>
<td>D</td>
</tr>
<tr>
<td>K</td>
<td>N</td>
<td>Q</td>
<td>P</td>
<td>V</td>
<td>N</td>
<td>T</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>G</td>
<td>Y</td>
<td>Y</td>
<td>P</td>
<td>I</td>
<td>K</td>
<td>N</td>
<td>S</td>
<td>P</td>
<td>A</td>
<td>T</td>
<td>N</td>
<td>I</td>
<td>S</td>
</tr>
</tbody>
</table>

**P. grandis** (85 a.a.)

**P. grandis**

| MS | L | E | M | S | K | V | I | L | K | P | S | S | K | L | F | L | M | S | I | F | T | V | S | F | F | T | I | K |
| V | F | S | D | H | F | W | L | M | D | Q | I | L | C | S | M | E | D | D | D | V | S | Q | A | D | D | H | P | V | L | P | S | K |
| A | S | T | D | L | T | K | P | N | T | S | L |

**Q. quadrula** (80 a.a.)

**Q. quadrula**

| MN | K | F | R | N | K | T | T | W | D | L | I | I | V | A | I | S | L | M | L | V | L | F | P | N | L | L | T |

**V. ellipsiformis** (89 a.a.)

**V. ellipsiformis**

| L | V | M | K | T | Q | J | M | N | L | N | K | M | V | Q | K | L | I | I | F | T | T | G | L | M | I | L | L | P | S | P | F | F | L |

M ORF

**I. japonensis** (118 a.a.)

**I. japonensis**


**P. grandis** (234 a.a.)

**P. grandis**


**Q. quadrula** (101 a.a.)

**Q. quadrula**


**V. ellipsiformis** (226 a.a.)

**V. ellipsiformis**


E
Female gonad & FORF antibody
Male gonad & MORF antibody
A. *Inversidens japonensis*

B. *Quadrula quadrula* and *Venustaconcha ellipsiformis*

C. *Pyganodon grandis*