J Mol Evol (2007) 65:380–391
DOI 10.1007/s00239-007-9016-x

Complete Mitochondrial DNA Sequence of the Scallop *Placopecten magellanicus*: Evidence of Transposition Leading to an Uncharacteristically Large Mitochondrial Genome

David R. Smith · Marlene Snyder

Received: 24 June 2006 / Accepted: 6 July 2007 / Published online: 6 October 2007

Abstract Complete sequence determination of the mitochondrial (mt) genome of the sea scallop *Placopecten magellanicus* reveals a molecule radically different from that of the standard metazoan. With a minimum length of 30,680 nucleotides (nt; with one copy of a 1.4 kilobase (kb) repeat) and a maximum of 40,725 nt, it is the longest reported metazoan mitochondrial DNA (mtDNA). More than 50% of the genome is noncoding (NC), consisting of dispersed, imperfectly repeated sequences that are associated with tRNAs or tRNA-like structures. Although the genes for *atp8* and two tRNAs were not discovered, the genome still has the potential for encoding 46 genes (the additional genes are all tRNAs), 9 of which encode tRNAs for methionine. The coding portions appear to be evolving at a rate consistent with other members of the pectinid clade. When the NC regions containing "dispersed repeat families" are examined in detail, we reach the conclusion that transposition involving tRNAs or tRNA-like structures is occurring and is responsible for the large size and abundance of noncoding DNA in the molecule. The rarity of enlarged mt genomes in the face of a demonstration that they can exist suggests that a small, compact organization is an actively maintained feature of metazoan mtDNA.

Reviewing Editor: Gail Simmons

D. R. Smith · M. Snyder (✉)
Department of Biology, Acadia University, B4P 2R6 Wolfville, Nova Scotia, Canada
e-mail: marlene.snyder@acadiau.ca

D. R. Smith
Department of Biology, Dalhousie University, Halifax, Nova Scotia, Canada

Keywords *Placopecten magellanicus* · Mitochondrial genome · Transposition · Bivalve · Mollusk · Dispersed repeats

Introduction

Metazoan mitochondrial (mt) genomes were once considered to be an extreme example of genetic economy (Attardi 1985) showing a highly reduced size of 15–17 kilobases (kb) and a gene content of only 13 protein coding-, 2 rRNA coding-, and 22 tRNA coding-genes. Although many of the reported animal mtDNAs fit this model, an increasing number seem to deviate from it substantially. For example, the mt genome of the snail *Biomphalaria glabrata* has a length of only 13.6 kb (Dejong et al. 2004); and those of three species of bark weevil, the brachiopod *Lingula anatina*, and the sea scallop *Placopecten magellanicus* can be as large as 36 kb, 28.8 kb, and 40 kb, respectively (Boyce et al. 1989; Endo et al. 2005; Snyder et al. 1987; Gjetvaj et al. 1992). The gene content of metazoan mtDNA has also been shown to deviate from what was considered the "standard" set. There are examples of genomes with multiple copies of either protein coding-, rRNA coding-, or tRNA coding-genes, and those which lack genes, such as *atp8* and certain tRNAs (Hoffmann et al. 1992; Beagley et al. 1995; Yokobori et al. 2004; Endo et al. 2005; Mizi et al. 2005). There are numerous reports describing duplications of large coding and NC regions (Rand 1993; Yokobori et al. 2004). Dispersed repeated sequences associated with tRNA-like structures have also been observed, raising the possibility of transposition within the mt compartment (Endo et al. 2005). Furthermore, the observations of heteroplasmy, recombination (Ladoukakis and Zouros 2001a,b), introns (Beagley et al. 1995), and

doubly uniparental inheritance mechanisms (Zouros et al. 1994) demonstrate an evolutionarily dynamic system.

Of all the metazoan mt genomes sequenced to date, that of the sea scallop, *Placopecten magellanicus*, departs the most dramatically from what was once considered a standard pattern. The *P. magellanicus* mtDNA is at least twofold larger than most metazoan mt genomes (32–40 kb) (Snyder et al. 1987), and no other bivalves with a similar-sized genome have been recorded. Currently, there are eight complete or nearly complete mt genomes available from bivalves, including the freshwater mussels *Inversidens japanensis* and *Lampsilis ornate*; the marine mussels *Mytilus edulis*, *Mytilus galloprovincialis*, and *Mytilus trossuls*; the oysters *Crassostrea virginica* and *Crassosstrea gigas*; and the Manila clam *Venerupis philippinarum*. Of these sequences, that of *V. philippinarum* is the largest, at 22,676 nt while the rest are all <19,000 nt in length.

It has been proposed that a genome-wide duplication is responsible for the large size of the *P. magellanicus* mt genome (Snyder et al. 1987). This is supported by detection of regions of sequence identity located nearly 180° apart in the genome (LaRoche et al. 1990). However, analysis of the completed sequence reveals that (i) genome-wide duplication is not the underlying cause of the overall large size; (ii) the functional regions have not been involved in any localized duplication events; (iii) the genome contains much more than the normal coding capacity for metazoan mtDNA, with at least 32 genes for tRNAs, but appears to be lacking two tRNAs and the gene for *atp8*; (iv) at least two distinctly different forms of genetic instability are operating in the molecule; (v) transposition events similar in nature to those proposed for the brachiopod *Lingula anatina* (Endo et al. 2005) have occurred; (vi) the transposition events involve expansion of families of tRNA genes; and (vii) the functional regions, in spite of the increased rate of change observed in the NC sequence, seem to be evolving at a normal rate for the pectinid clade.

Materials and Methods

Samples

mtDNA was isolated from live *P. magellanicus* individuals kindly supplied by Dr. Mike Dadswell.

Isolation and Cloning

Isolation of mtDNA was performed according to the method described by Snyder et al. (1987) with the exception that tissue was ground in a Polytron homogenizer at a setting of 30. Intact mitochondria were obtained by brief centrifugation of lysed cells through a sucrose step gradient and then treated with a 0.5% Sarkosyl solution to release DNA. Following phenol-chloroform extraction, mtDNA was separated from contaminating nuclear DNA by CsCl-bisbenzimide isopycnic centrifugation (Turmel et al. 1999). Two bands were present above the nuclear band on the CsCl-bisbenzimide gradient; both were removed and an aliquot of each was digested with restriction enzymes. The lower band gave the pattern diagnostic of *P. magellanicus* mtDNA and was used in all further work. A plasmid library of mtDNA fragments (~1500 nt) was prepared by nebulization followed by ligation into the vector pFBS. Plasmid constructs were cloned in Stratagene XL Gold Ultracompetent cells as described by Lemieux et al. (2000). Recombinant plasmids containing mtDNA inserts were identified by colony hybridization with the original, intact mtDNA as a labeled probe.

DNA Sequencing

Cloned and PCR-amplified mtDNA segments were purified using the QIAprep 8 Miniprep kit (Qiagen) or the QIAquick Gel Extraction kit (Qiagen), respectively. Nucleotide sequences were determined either with the PRISM dye terminator cycle sequencing kit (Applied Biosystems) at the Laboratory for Synthesis and Analysis of DNA at Laval University, Quebec, or at the Florida State University Sequencing Facility using an Applied Biosystems 3100 Genetic Analyzer with Capillary Electrophoresis.

Analysis of Sequence Data

Assembly of the genome was performed using Autoassembler (Version 2.1.1; Applied BioSystems). The 11 initial contigs identified by the assembly program were aligned using the 38 restriction endonuclease recognition sites identified in earlier work (Fuller and Zouros 1993; LaRoche et al. 1990; Snyder et al. 1987) providing a preliminary, incomplete genomic sequence. Gaps between contigs were closed by PCR amplification of intact mtDNA using primers designed from known sequence near the ends of the contigs. This method of closing the gap confirmed the alignment of neighboring contigs. The fragments produced were then sequenced and added to the assembly until all portions of the mtDNA were present in one contig.

The assembly program "collapses" tandemly repeated sequences into a single sequence unless the length of the repeat is significantly less than the approximately 800 nt obtained from a single sequencing reaction (so that multiple copies can be present in a single sequencing reaction). Collapsed sequences cause an artificial increase in redundancy of a repeated region, thus one can scan the assembly by eye and identify regions of potential collapsed sequence.

Two regions were identified as having much higher than the threefold average redundancy. One of these, the previously described 1.4-kb repeat, was anticipated. The other contains the sequences for the *rrnL* gene, parts of which are present in 12 different sequenced fragments. To determine whether or not there are multiple copies of this region present in the genome, primers were constructed from sequence present just outside of the overrepresented portion of the genome. Amplification of intact mtDNA resulted in one band of the length that would result from just one copy of the region; therefore, we have designated it single copy. The dispersed imperfect repeated regions were short enough, with unique sequence on either side, that they were not recognized by the assembly program as related sequence.

The BLAST network services (Altschul et al. 1990) were used for sequence similarity searches. Regions encoding proteins not initially detected in the BLAST search were identified with the Open Reading Frame Finder (ORF Finder; http://www.ncbi.nlm.nih.gov/gorf/gorf.html) using the invertebrate mitochondrial code. Where the ORF finder did not accurately predict the start and/or stop codons, alignment with homologous gene sequences from other mollusks using ClustalW version 3.2 (Thompson et al. 1994) was used. In a few instances, hydropathy profiles, plotted online (http://www.bioinformatics.weizmann.ac.il/hydroph/plot_hydroph.html) using the Kyte-Doolittle method (1982) and a window size of 19, were used to find plausible regions for start and stop codons. The start and end of the rRNA coding genes were determined using sequence comparisons with the rRNAs from the bivalve *Pecten maximus*. Genes encoding tRNAs and pseudo-tRNAs were located either using TRNASCAN-SE (http://www.genetics.wustl.edu/eddy/tRNAscan-SE/ [Lowe and Eddy 1997]) or visually by their potential to form tRNA-like secondary structures fitting the model: 5'-ANNRYNNNNNNYT(anticodon)RN-3'. The program RRTree, version 1.1.11 (Robinson-Rechavi and Huchon 2000), was employed for nucleotide substitution analyses. Repeated regions were identified using JDotter (http://www.athena.bioc.uvic.ca/pbr/jdotter/) with a window size of 50 and a 42-bp stringency, The Tandem Repeat Finder Version 3.01 (Benson 1999), Repeat Finder (http://www.proweb.org/proweb/Tools/selfblast.html), and REPuter (Kurtz et al. 2001).

Results

General Features

The complete *P. magellanicus* mitochondrial genome (GeneBank accession number DQ088274) is 30,680 nt in length when one copy of a 1.4-kb repeat is present. We refer to this as the unit molecule. Individuals have two to eight copies of this 1.4-kb repeat occurring in tandem (LaRoche et al. 1990; Fuller and Zouros 1993). Smaller tandem repeats at two other locations also show a variation in copy number (Fuller and Zouros 1993). Therefore, the true size of the *P. magellanicus* mt genome ranges from ~32,114 to ~40,725 nt. Of the 30,680 nt of sequence in the unit molecule, 15,706 nt are coding DNA, and 16,408 nt are NC DNA, though the latter can be as high as 25,019 nt depending on the number of tandem repeats. The NC DNA is distributed throughout the genome in regions ranging in length from 1 to >10,000 nt.

The molecule has an A + T content of 55.5%. Coding sequences have a value similar to that for the overall molecule (55.7%), indicating that the value is a molecule-wide phenomenon.

Gene Order and Content

The *P. magellanicus* mt genome encodes 12 proteins, 2 rRNAs, and 32 tRNAs, all of which have the same transcriptional polarity. The names and order of these genes are shown in Fig. 1. The genes for *atp8, trnR*, and *trnP* were not identified.

Of the 32 identified tRNA coding genes, 14 are clustered in two separate groups composed of 5 and 9 tRNAs, respectively (Fig. 1). The nine-tRNA cluster is 804 nt long, with the unassigned sequence between tRNA coding genes averaging 25 nt — ranging from a maximum of 80 nt to a minimum of 2 nt. The cluster encoding five tRNAs is 381 nt long, and the average spacing between them is 12 nt, with a maximum of 27 nt and a minimum of 1 nt. The remaining 18 tRNA coding genes are dispersed with at least 100 nt separating them. There are only two instances where a single tRNA coding gene is located between two non-tRNA genes: *rrnS*-(*trnS2*)-*nad4L* and *atp6*-(*trnL2*)-*cytb*. Compared to the tRNA coding genes, the protein coding genes are more dispersed, but most are found within a 12-kb segment of the genome. This region, which begins with *rrnL* and ends with *nad6*, comprises 8 protein coding genes (*nad4L, nad2, nad4, nad1, coxI, nad5, nad3* and *nad6*), both rRNA coding genes (*rrnL* and *rrnS*), and 10 tRNA coding genes (*trnS2*, *trnH*, *trnW*, *trnY*, *trnT*, *trnI*, *trnM3*, *trnC*, *trnL1*, and *trnV*) (Fig. 1). Eighty-seven percent of the bases within this 12-kb area are coding DNA, a much higher percentage than that of the overall genome, which has a coding-to-noncoding ratio of approximately 1:3.

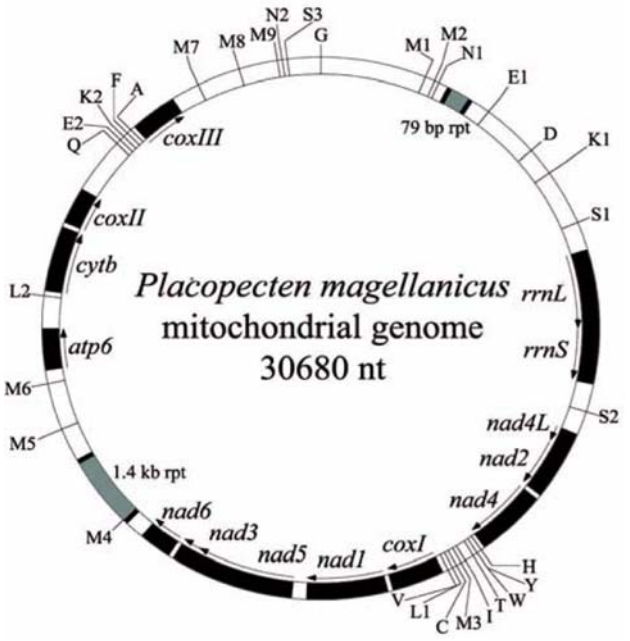

Fig. 1 Gene map of the mitochondrial genome of the bivalve mollusk *Placopecten magellanicus*. Protein and rRNA coding genes are abbreviated as in text and are shown in black on the map. The one-letter amino acid code is used for tRNA designation. Codons recognized by the tRNAs that occur more than once are as follows: N1-UUC, N2-UUC, E1-GAA, E2-GAA, L1-CUA, L2-UUA, M1-AUG, M2-AUG, M3-AUG, M4-AUA, M5-AUA, M6-AUA, M7-AUA, M8-AUA, M9-AUG, S1-UCU, S2-UCU, and S3-AGA. The direction of transcription is depicted by arrows for protein and rRNA coding genes. All tRNA coding genes are transcribed clockwise. All noncoding regions are unlabeled and shown in white on the map, except for those containing the 1.4-kb repeat and the 79-nt repeat, which are shown in gray

The sole occurrence of gene overlap is observed in the four bases shared by the end of *nad5* and the start of *nad3*.

Protein Coding Genes

Within the invertebrate mt code there are three conventional start codons (M-AUG, M-AUA, and I-AUU) and three alternative start codons (I-AUC, L-UUG, and V-GUG) (Wolstenholme 1992). Of the 12 protein coding genes identified in the *P. magellanicus* mt genome, all but 3 use a conventional start codon. The codon GTG was selected as initiator codon for the *nad4L*, *nad2*, and *cytb* genes based on alignments with homologous genes from other mollusks. The abbreviated stop codon T- -, which can become a full stop codon (TAA) after polyadenylation of the mRNA (Ojala et al. 1980), was assigned to the genes *coxI* and *coxII*. A full stop codon was observed for both of these genes, but in both instances it generated a protein with a substantially extended C-terminus (>50 amino acids) relative to its counterparts in other mollusks. No *atp8* gene was identified.

Table 1 Mitochondrial large subunit rDNA substitution rates among *P. magellanicus* (*Placo*), *Pecten maximus* (*Pect*), and *Argopecten irradians* (*Argo*), with *Mytilus edulis* (*Myt*) as a reference

	Evolutionary distance matrix (K)			% GC
	Placo	*Pect*	*Argo*	
Placo	0.000	—	—	42.4
Pect	0.259	0.000	—	40.4
Argo	0.344	0.284	0.000	40.9
Myt	0.658	0.681	0.663	34.6

Note. Rates were calculated using the Kimura (1980) two-parameter model and 1172 sites. *P. magellanicus* (this study); *P. maximus* (GenBank accession number X82501); *A. irradians* (GenBank accession number AF526205); *M. edulis* (Hoffmann et al. 1992; GenBank accession number NC_006161)

Ribosomal RNA Coding Genes

Genes for the large- and small-subunit rRNAs (*rrnL* and *rrnS*) were assigned lengths of 1387 and 970 nt, respectively. Only 51 nt separate these two rRNAs, but unlike the mtDNAs from other organisms, no tRNA was detected in this intergenic space.

The rate of change of functional sequence in *P. magellanicus* was evaluated in two ways: base composition comparison and substitution rate analysis. Since only a few mtDNA sequences are available from other scallops, our study of rates and base composition was limited to 1172 sites from the large subunit rRNA. The data for these analyses is presented in Table 1. Four species were used in the rRNA comparisons: three scallops and the blue mussel *Mytilus edulis* as an outgroup. The base compositions of the scallops are quite similar to each other but differ from that of the mussel. The substitution rates between pairs of scallop species are similar: *P. magellanicus* vs *Argopecten irradians* gives the largest distance (0.344), and *P. magellanicus* vs *Pecten maximus* gives the smallest distance (0.259) (Table 1). The substitution rates of *M. edulis* as compared to each of the three scallops are alike and much higher than the between-scallop values. The highest substitution rate is for *M. edulis* vs *P. maximus* (0.681); the lowest rate is for *M. edulis* vs *P. magellanicus* (0.658).

Transfer RNA Coding Genes

Difficulties arising from two factors were encountered when we attempted to identify tRNA coding sequences in the mtDNA: (i) the abnormally large number of tRNA and tRNA-like structures present in the sequence and (ii) the co-occurrence of most of these structures with six families of dispersed repeats (Fig. 6).

In total, 49 tRNA-like structures were identified (Figs. 2 and 3). Of these, 32 are potentially true tRNA coding genes

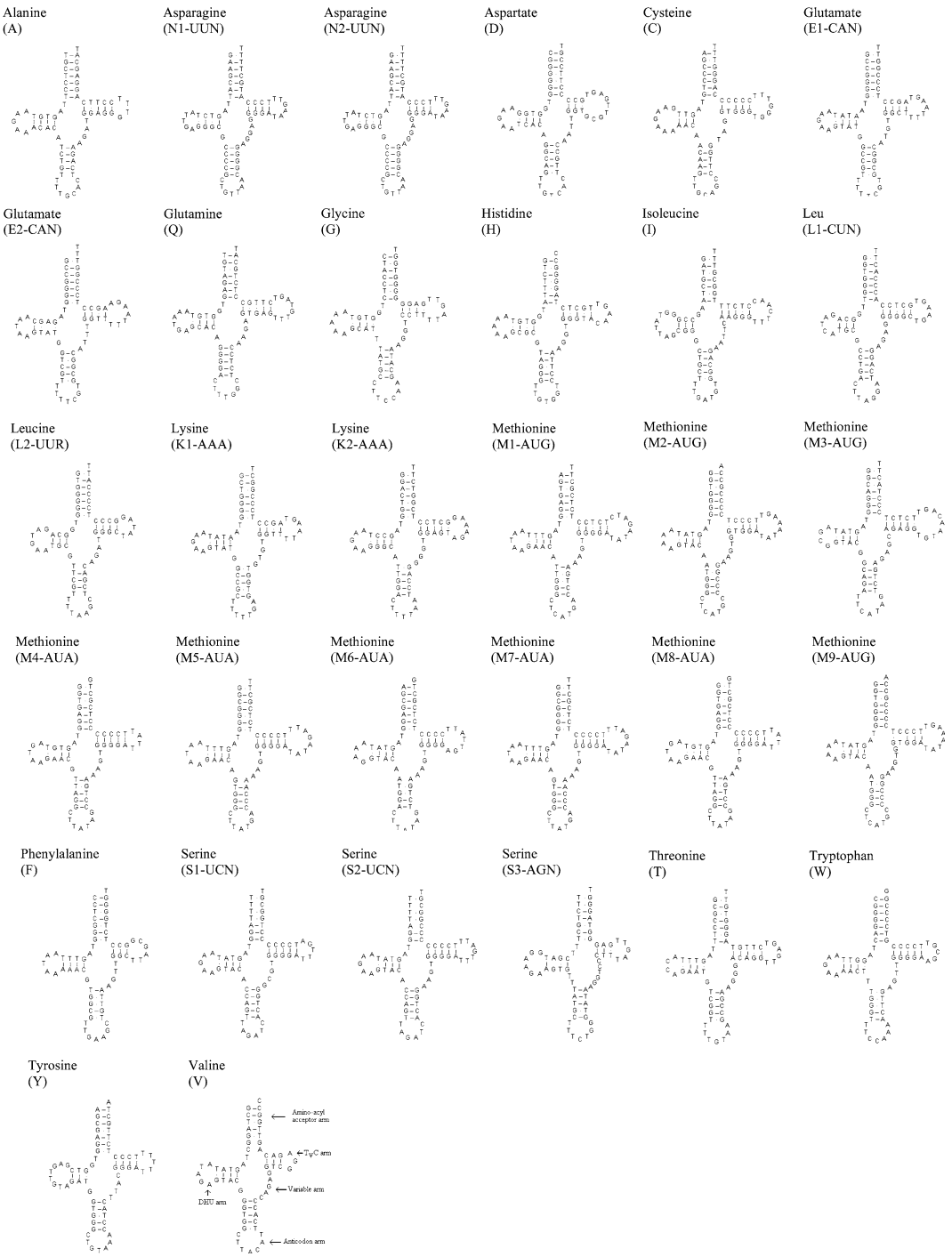

Fig. 2 Putative cloverleaf structures for 32 tRNAs deduced from the *Placopecten magellanicus* mtDNA. Watson-Crick pairing is shown by solid lines and G-T pairing by dots. Nomenclature for the different parts of the tRNA structure is shown for *trnV*

based on secondary structure considerations. Some of the non-canonical conformations commonly observed in mitochondrial encoded tRNAs are present within this group, including reduced loop size (*trnS1* and *trnS2)* and mismatches of base pairing in more than half of the proposed structures in either the aminoacyl acceptor arm or the anticodon stem. In all, 21 different anticodon sequences are represented, 17 of which are present more than once, including 5 copies of *trnM*-AUA, 4 copies of *trnM*-AUG, and 2 each of *trnN*-UUC, *trnE*-GAA, *trnK*-AAA, and *trnS*-UCU. Of the nine tRNA coding genes for methionine, there are three identical pairs: *trnM2/M9, trnM4/M8*, and *trnM5/M7*. When sequences from all nine of the methionine tRNAs are aligned (Fig. 4), their relatedness based on

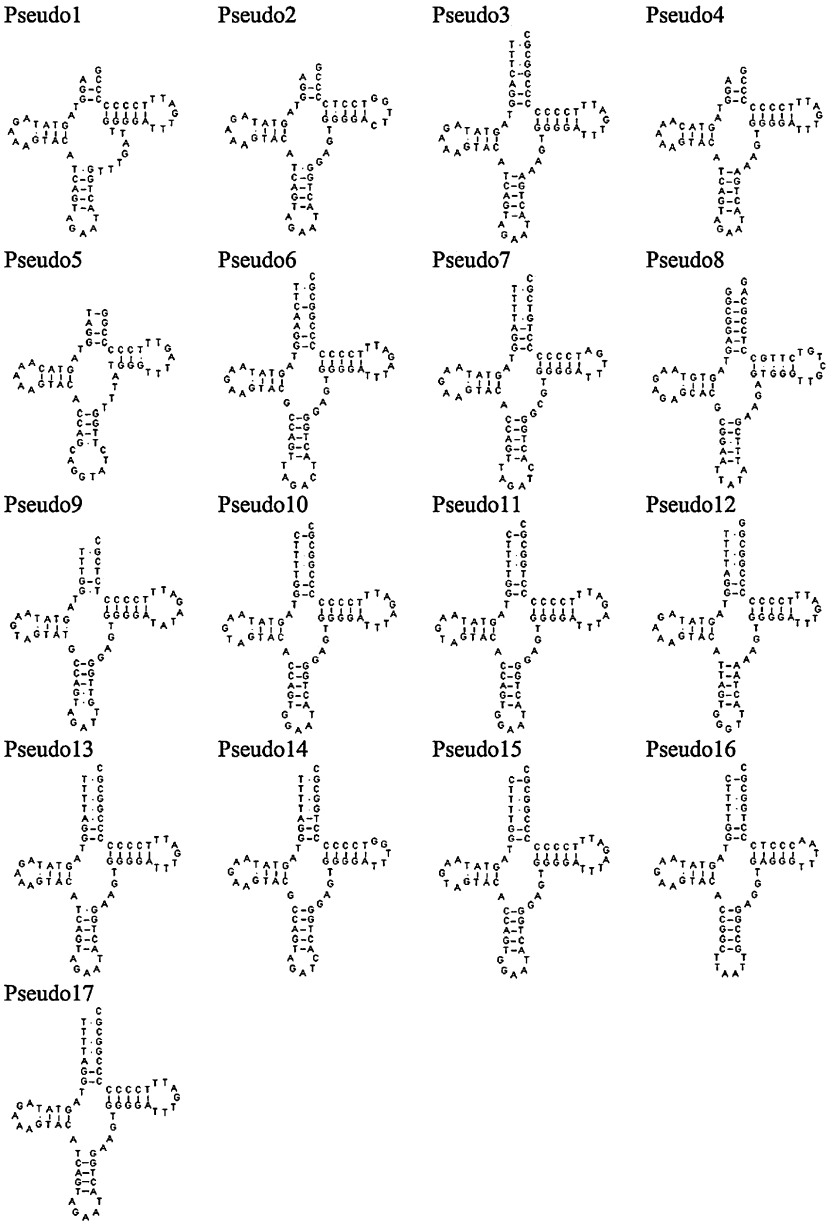

Fig. 3 Cloverleaf structures for 17 pseudo-tRNAs deduced from the *Placopecten magellanicus* mtDNA. Watson-Crick pairing is shown by solid lines; G-T pairing is shown by dots

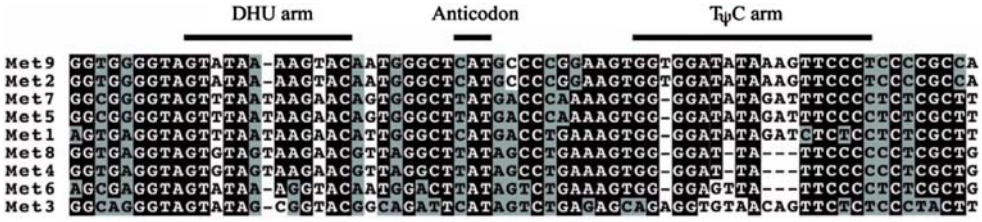

Fig. 4 BOXSHADE representation of the nucleotide sequences of the nine tRNA coding genes denoting methionine. Identical and similar residues are indicated using black and light-gray boxes, respectively. Nomenclature for the different parts of the tRNA structure is shown

unusually long stretches of sequence identity becomes obvious. The two *trnN*-UUC sequences are also identical.

Seventeen of the 49 tRNA-like structures are considered pseudo-tRNAs because they have reduced stems and loops in addition to many mismatches in base pairing. A striking degree of sequence identity exists among all 17 pseudo-tRNAs (Fig. 5). Interestingly, the two *trnS*-UCU sequences show a strong resemblance to the 17 pseudo-tRNAs (Fig. 5).

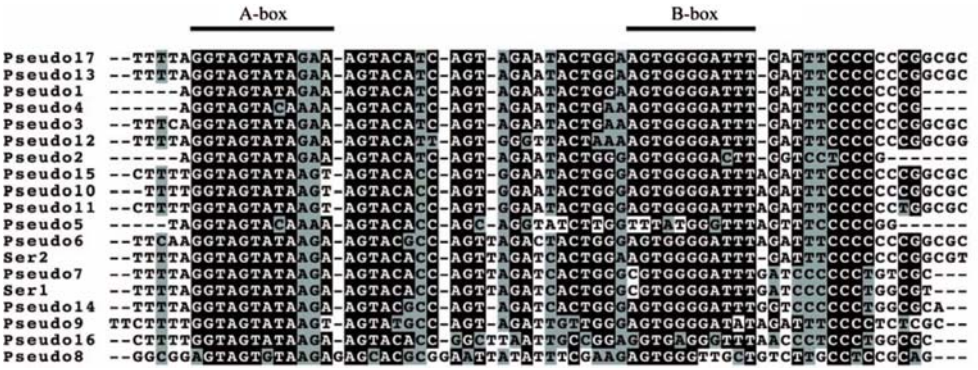

Fig. 5 BOXSHADE representation of the nucleotide sequences from the 17 pseudo-tRNAs and those of *trnS1* and *trnS2*. Identical and similar residues are indicated using black and light-gray boxes, respectively. Regions showing similarity to the A-box and B-box of tRNA-like SINES are labeled

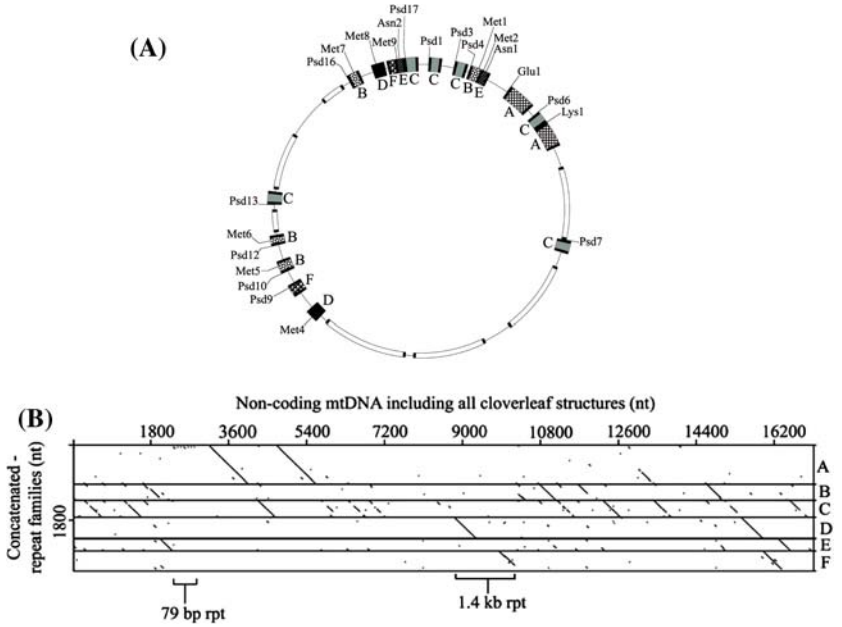

Fig. 6 **A** Map of six repeat families found in the *Placopecten magellancius* mitochondrial genome. The six families of repeat are labeled A, B, C, D, E, and F. They have lengths of 908, 411, 420, 481, 287, and 469 nt, respectively. Repeats belonging to a particular family are indicated as follows: A (checkered pattern), B (dotted pattern), C (light grey shading), D (black shading), E (diagonals), and F (spots). The cloverleaf structures associated with each repeat are shown using the one-letter amino acid code for tRNAs or "Psd" for pseudo-tRNAs. Areas containing coding DNA are shaded in white. **B** Dot matrix of the noncoding DNA including all cloverleaf structures (tRNAs and pseudo-tRNAs) plotted against the concatenated nucleotide sequence of the six repeat families in their respective order. Generated using a 25-nt window and 42-nt stringency. The 1.4-kb and 79-nt repeats are labeled

Analysis of Noncoding DNA

The majority of the NC DNA falls into two distinct categories: those regions containing tandemly organized, perfectly repeated sequences and those containing dispersed, imperfect members of repeat families.

Of the two tandemly organized repeat regions identified, the larger region corresponds to the 1.4-kb repeat that was sequenced and described in previous work (Snyder et al. 1987; LaRoche et al. 1990). With the completion of the sequence reported here, more than six clones containing the 1.4-kb repeat have been sequenced. All copies are identical. The finding of LaRoche et al. (1990) of a seven-base inverted sequence bordering the repeat, beyond which DNA loses resemblance to the repeat, is confirmed by our results. The repeat has an A + T content of 60%, numerous poly(A) tracts, and a perfect 10-base-long, GC-rich stem-and-loop structure. There are no significant open reading frames within the repeat, but the gene *trnM4* is present. A single copy of the repeat is shown in Fig. 1. Two segments, one of 481 nt and another of 469 nt, matching the 1.4-kb repeat were found several thousand base pairs away, near the gene for *coxIII*. These two regions are identified in Fig. 6 as family "D" and family "F," respectively.

The smaller of the two tandemly organized repeats is located nearly 180° from the larger and consists of seven exact copies of a 79-base sequence. It has an A + T content of 58%. Its tandem nature and its variation in copy number

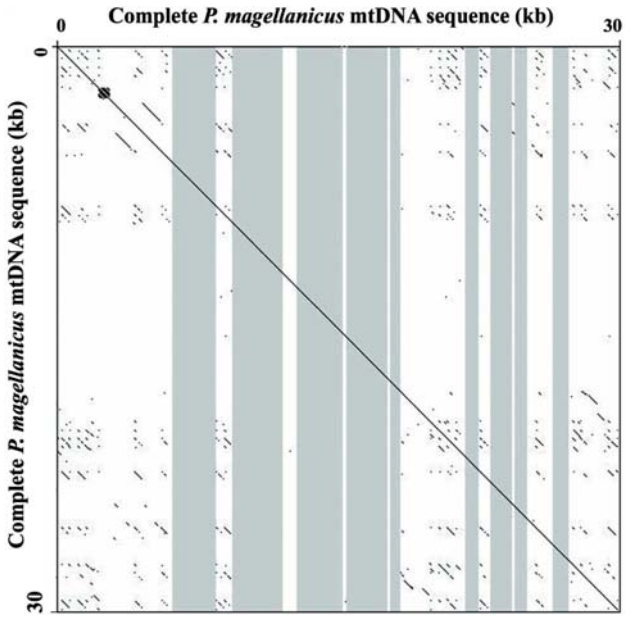

Fig. 7 Dot matrix of the complete *Placopecten magellanicus* mtDNA plotted against itself. Matrix was generated using a 25-bp window and 42-bp stringency. The regions identified as "coding" are shown in their correct positions on the horizontal axis and are extended down through the dot-plot as gray columns. Note the absence of off-diagonal ticks in the gray shaded areas, indicating absence of duplication of coding sequence. Note: The 79-nt repeat results in the presence of a small square superimposed on the diagonal of the dot-plot image

from individual to individual were identified and labeled as "locus III" by Fuller and Zouros (1993). The 79-nt repeat results in the presence of a small square superimposed on the diagonal of the dot-plot image in Fig. 7. The previously identified tandemly repeated sequence of approximately 250 nt, known as "locus II" (Fuller and Zouros 1993), could not be found in our sequence.

Nearly all of the remaining NC DNA (~16 kb of sequence) is made up of six families of dispersed, imperfect repeats, and subsections thereof. The families bear almost no resemblance to each other at the sequence level, but they do share one feature: they are all associated with tRNA or pseudo-tRNA structures. Figure 7 is a dot-plot of the entire mtDNA sequence, with the regions identified as "coding" shown in their correct positions on the horizontal axis and extended down through the dot-plot as gray columns. The diagonal line represents the sequence plotted against itself. Any region of sequence similarity ≥50 nt, which was used as the window size in jdotter, is represented in the plot as a diagonal "tick." No ticks are present in the gray-shaded areas, indicating that no regions containing coding information are repeated elsewhere in the sequence. Multiple individual ticks and short diagonal lines, all of which run from upper left to lower right, are present elsewhere in the dot-plot. The genome was scanned using BLAST, searching for sequence alignment of sections repeated within the molecule. Such a search can identify segments where A shares sequence with B, and B shares sequence with C, but A and C do not share sequence with each other, leading to the identification of families whose members vary in "completeness," similarity, and length. Distributions of the longest members of the six families thus identified are shown in Fig. 6a. The families, called A, B, C, D, E, and F have lengths of 908, 411, 420, 481, 287, and 469 nt, respectively.

Figure 6b presents a second dot-plot analysis of the concatenated sequence of the six families against all the NC DNA. Stretches of sequence corresponding to the labeled segments in Fig. 6a are easy to discern as the major diagonal lines for each of the six families. In addition, it is obvious that smaller segments from the six families are present at scattered locations.

If all DNA that is neither protein coding nor tRNA coding nor rRNA coding in the unit molecule is considered, there are 17,607 nt to account for. Of these, 1854 nt (or 10.5%) are unclassified and fall mainly between genes in the gene-rich region; 1986 nt (or 11.3%) come from the 79-base tandem repeat region plus a portion of the 1.4-kb repeat, and 13,803 nt (or 78.2%) are part of the dispersed repeat families.

Discussion

Analysis of the complete mt genome from *P. magellanicus* provides answers to some questions but raises many others. The abundance, distribution, and composition of NC DNA are highly unusual for a metazoan. Most surprisingly, when sequences of the NC DNA identified as "dispersed sequence families" are examined in detail, we reach the conclusion that transposition of DNA involving tRNA or tRNA-like sequences is occurring and is responsible for the large size and abundance of NC DNA in the molecule. This is the second such finding of imperfect repeats associated with tRNA and tRNA-like sequences in a metazoan mtDNA; Endo et al. (2005) observed a similar phenomenon in the mt genome of the brachiopod *Lingula antina.* These similarities suggest a common mechanism in both systems. However, there are also interesting differences between the genome of *L. anatina* and that of *P. magellanicus*, but these are discussed more fully below.

A large duplication event involving a mtDNA molecule of "normal" size as the underlying cause for the atypical length of the *P. magellanicus* mtDNA can be ruled out because the protein and rRNA coding portions of the sequence show no sequence similarity at any other position in the sequence. Instead, they stand out as islands of unique

sequence embedded in an apparently unstable sequence landscape.

At least two forms of genetic instability, and perhaps a third, can be discerned. One involves variation in copy number of tandemly repeated sequences. This is evident at the population level when adults are surveyed for differences in restriction digest patterns (Snyder et al. 1987; Fuller and Zouros 1993), suggesting a mechanism of expansion and contraction by unequal crossing-over. The repeats involved in these events have lengths of either 79 nt or 1.4 kb. The copy numbers observed range from up to 10 for the 79-nt repeat and from 2 to 8 for the 1.4-kb repeat. Despite a widely held belief that recombination does not occur in metazoan mtDNA (Avise 2000), evidence demonstrating that it does is accumulating (Thyagarajan et al. 1996; Lunt and Hyman 1997; Ladoukakis and Zouros 2001a, b; Hoarau et al. 2002; Burnzynski et al. 2003; Piganeau et al. 2004; Tsaousis et al. 2005). Mechanisms suggested to account for tandemly repeated sequences in the absence of recombination include slipped strand mispairing (Levinson and Gutman 1987), overrunning of replication ahead of the normal termination point (Boore and Brown 1998), and illegitimate priming of DNA synthesis by a tRNA molecule (Jacobs et al. 1989). It is unlikely that any of these recombination-independent mechanisms could account for the sequence space involved in generating up to eight copies of the 1.4-kb repeat observed in *P. magellanicus* mtDNA; a mechanism involving recombination seems much more likely. Concerted evolution appears to be operating on these repeats, again, consistent with a mechanism involving unequal crossing over; the seven sequenced copies of the 79-nt repeat are all identical, as are the multiple copies of the 1.4-kb repeat that have been cloned and sequenced.

The second form of instability involves tRNA and tRNA-like sequences. Of the six imperfectly repeated, dispersed sequence families that were identified, all are associated with cloverleaf structures—either tRNA coding genes or pseudo tRNA genes—but the position and number of these cloverleaf structures vary between families. Although members of a single sequence family share identity throughout their defined regions, the strongest identities occur in the areas surrounding their respective cloverleaf structures. There are 135 bases of sequence identity beyond the tRNA coding sequence for pair *trnM4/M8* (these two tRNAs are identical in sequence; this repeat is "family D"), and 47 bases of sequence identity exist beyond the tRNA coding sequence for pair *trnM5/M7* (this pair is also identical in sequence; this repeat is "family B"). Other repeat families do not display such a high degree of sequence identity, which suggests that some transpositions have occurred more recently than others. Family F, like family D, has one member in the 1.4-kb repeat and another member outside of it. These similarities are responsible for the additional labeled bands observed by LaRoche et al. (1990) when probing with labeled 1.4-kb repeat DNA.

The relationship between *trnS1* and *trnS2* (both have the anticodon UCU) and the 17 pseudo-tRNAs may represent a third form of instability. These 19 sequences show a striking similarity across their entire sequence, but the most remarkable feature is the almost-complete conservation of a nine-base segment that corresponds to the position in the cloverleaf structure representing the junction between the acceptor stem and the DHU arm. This is the position known as Box A in SINES derived from tRNAs, one of two conserved sequences necessary for RNA polIII recognition (Galli et al. 1981) (resemblance between the consensus Box A sequence and the conserved region is not strong). The other position required for RNA Pol III recognition in tRNA-derived SINES is known as Box B and corresponds to part of the loop and stem in the TψC arm. In these 19 sequences, that region is the next most highly conserved. Oddly, neither *trnS1* nor *trnS2* is associated with a repeat family, but all 17 of the pseudo-tRNAs are found in regions of repeated sequence.

The presence of the tRNA and tRNA-like structures in these sequence families suggests the possibility of a SINE-like mechanism of transposition involving reverse transcription. Aside from the tRNA sequence, there are several features commonly associated with tRNA-derived SINES: a region of unrelated DNA, an AT-rich region, and short terminal duplications at the site of insertion (Daniels and Deininger 1985). Of these, we see only the tRNA sequence and region of unrelated DNA adjacent to it. Another interesting peculiarity is that all of the various repeats observed in the *P. magellanicus* mt genome are oriented in the same direction. Insertion mediated by reverse transcription should not be limited to one orientation. Therefore, if reverse transcription is involved, it is operating in a manner different from that documented for tRNA-derived SINES.

The repeated tRNA-like structures observed by Endo et al. (2005) are not as dispersed as the repeat families documented here. Instead, they are localized to two regions of the molecule, show a more clustered nature within those regions, and are interrupted by unrelated sequence. The tRNA-like structures they observe, while similar in number to those detected in *P. magellanicus* (15, versus the 17 reported here), are derived from several different sources, as opposed to the single origin that we infer from sequence similarity. They report one apparent expansion of a single family involving 10 repeats. However, all 10 are clustered in one region, whereas in *P. magellanicus* a very dispersed distribution is observed. They suggest, as do we, that the repeats could be more easily explained using a "cut-and-

paste" mechanism, such as retrotransposition, rather than by tandem duplication and deletion alone.

In spite of the abundance of tRNA sequences in the molecule, two of the "standard set" of tRNAs—*trnR* and *trnP*—were not found. They are either present and unidentified, imported from the cytosol, or generated via the posttranscriptional editing of other mtDNA encoded tRNAs (Rubio et al. 2006; Janke and Pääbo 1993). It is not possible to determine whether or not all 32 identified tRNA coding sequences are functional. The COVE scores for the tRNAs sharing an anticodon are all high (with the exception of *trnS1* [score of 1.85] and *trnE2* [score of 1.43]). The scores for *trnM4* and *trnM8* (which are identical in sequence and both involved in repeated sequences) are higher than the score for *trnM3*, which is in one of the two clusters of tRNA coding genes, and presumably in a more "normal" position for transcription and processing.

The molecule as a whole seems to be composed of two distinct zones. One is "gene-rich" and contains virtually no sequence identity to the six dispersed, imperfect repeat families, and the other is gene-poor. The gene-rich region is marked by the start of the *rrnL* gene and termination of the *nad6* gene, respectively. It contains 8 protein coding genes, the 2 rRNA coding genes, and 10 genes for tRNAs (9 of these in one cluster). In total, it is 87% coding, a value much higher than the overall figure of approximately 30% coding. The average spacing between genes is 87 nt; the longest spacing is 504 nt, and the shortest 2 nt. Compared to other mollusks, it is evident that even in the gene-rich region the average spacing is much higher than that in a more "typical" mtDNA molecule. For example, in the mtDNA of *Mytilus edulis*, the average intergenic distance is 19 nt, and the maximum distance is 284 nt (Boore et al. 2004). In *Mytilus galloprovinicialis*, the average intergenic distance is only 18 nt (Mizi et al. 2005).

The remainder of the *P. magellanicus* mtDNA can be categorized as gene-poor. This region, which includes all the nucleotides found outside the gene-rich region, ranges from 18,185 nt in the unit molecule to nearly 28 kb in the largest observed molecule. Of this, only 3450 nt are protein coding, and 1489 nt potentially encode 22 functional tRNAs. The resulting coding content is 27% (vs. 87% in the gene-rich region).

In reports of other metazoan mtDNA sequences, the longest NC stretch of DNA is generally assumed to be the control region. We obviously cannot make that assumption but can speculate based on sequence features that are typically found in control regions (Wolstenholme and Jeon 1992), such as a high A + T content and hairpin structures. Based on these considerations, the sequence within the 1.4-kb repeat is a good candidate for the control region. It is higher in A and T, at 60%, compared to 55.5% for the molecule as a whole; it has a perfect 10-base pair (bp), GC-rich inverted repeat that has been shown to take on a cruciform structure in solution; and it has A tracts that have been experimentally demonstrated to form bent DNA in solution (LaRoche et al. 1990). There is precedent for more than one copy of a control region existing in a mt genome (Eberhard et al. 2001; Yokobori et al. 2004). Could eight copies be tolerated and provide normal function? The question cannot be answered by sequence analysis alone. Often short tandem repeats are associated with the control region, and from that perspective the 79-nt region and its adjacent 264 nt of NC DNA, one of the few NC DNA tracts not taken up by repeat family sequence, are another potential candidate for the control region. However, no secondary structures could be detected in either the 79-nt repeat or the 264 nt adjacent to it.

The genome appears to be lacking a gene for *atp8*, as do all other published bivalve sequences (Kim et al. 1999; Okazaki and Ueshima, 2001; Boore et al. 2004; Mizi et al. 2005) with the exception of the freshwater mussel, *Lampsilis ornata* (Serb and Lydeard 2003). The gene for *atp8* can often be very difficult to detect because of its lack of sequence conservation between different organisms at both the nucleotide and the amino acid levels. In an attempt to identify the gene, we made use of the two features that help in its recognition: the highly conserved six residues found at the N-terminus of the protein (Met-Pro-Gln-Leu-Ser [or Ala]-Pro) and a distinctive hydropathy profile. All ORFs ranging from 30 to 65 amino acids were translated and hydropathy profiles were plotted. Only one showed a curve characteristic for *atp8*, but the absence of any similarity to the conserved N-terminus sequence indicates that it is not *atp8*.

All 46 genes found in the molecule have the same transcriptional polarity. Though this is rare for metazoan mtDNAs, it is not unusual for bivalves, having been observed in the mt genomes of *C. gigas*, *V. philippinarum, M. edulis*, and *M. galloprovincialis* (Kim et al. 1999; Okazaki and Ueshima 2001; Boore et al. 2004; Mizi et al. 2005). The gene arrangement within the *P. magellanicus* mt genome is distinct from all other available molluscan and metazoan mtDNAs.

The mtDNA of *P. magellanicus* is obviously on an evolutionary trajectory distinct from the majority of reported metazoan mtDNAs. We attempted to determine whether the events leading to the unusual size, which seem to indicate relaxed constraint in some aspect of sequence evolution compared to other metazoans, might also be affecting the rate of evolution of coding portions of the molecule. The lack of available sequences from the pectinid clade limits the scope of the attempt but there are enough to provide some insight. A question that can be

addressed is whether coding portions of the molecule are evolving at an elevated rate compared to that in scallops with mtDNA of a more normal size. The results show that rRNA sequences from *P. magellanicus* are not measurably different on average, in base composition or rate of evolution, compared to those of other scallops. These results stand in interesting contrast to those observed by Endo et al. (2005) for *L. anatina*, where it was observed that protein sequences have diverged much more than anticipated and indicate an elevated rate of sequence evolution. The difference in response of the two molecules to the relaxation of constraint on size implies that there is more than one system of constraint acting on metazoan mtDNAs. The complete sequence of one or all of the three bark weevils (*Pissodes nemorensis, P. strobe*, and *P. terminalis)* (Boyce et al. 1989), whose mt genomes have been observed to attain lengths of 36 kb, will undoubtedly provide further useful comparative information.

As has been observed by others (Boore 1999; Endo et al. 2005), metazoans with unusually large mt genomes are found distributed sporadically across a broad range of taxa; these large genomes have evolved secondarily and independently of each other. The differences and similarities that can be observed between the sequences of *P. magellanicus* and *L. anatina* are informative, providing parameters of tolerance of their deviant states. The rarity of enlarged mt genomes, despite the fact that they can exist, suggests that a small, compact organization is an *active* feature of metazoan mtDNA, not a passively inherited consequence of a streamlined genome clonally inherited in the absence of recombination or transposition. Our observations highlight the need for more sequences of enlarged genomes for comparative studies, as well as additional biochemical approaches to understanding the dynamic maintenance of a streamlined genome.

Acknowledgments We wish to thank Claude Lemieux, Monique Turmel, and Christian Otis for generous assistance during a sabbatical year at Université Laval for M.S.; Mike Dadswell for supplying scallops; and Michael Smith, Kathy Fuller, and Ian Paterson for comments on the manuscript. This work was supported by an AUFA 25.55 grant and by funding from the Howard Gould Trust.

References

Altschul S, Gish W, Miller W, Myers E, Lipman D (1990) Basic local alignment search tool. J Mol Biol 215:403–410

Attardi G (1985) Animal mitochondrial DNA: an extreme example of genetic economy. Int Rev Cytol 93:93–145

Avise JC (2000) Phylogeography: the history and formation of species. Harvard University Press, Cambridge, MA

Beagley CT, Macfarlane JL, Pont-Kingdon GA, Okimoto R, Okada N, Wolstenholme DR (1995) Mitochondrial genomes of Anthozoa (Cnidaria) In: Palmieri F, Papa S, Saccone C, Gadaleta N (eds) Progress in Cell Research—Symposium on Thirty Years of Progress in Mitochondrial Bioenergetics and Molecular Biology (F). Elsevier, Amsterdam, pp 149–153

Benson G (1999) Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res 27:573–580

Boore JL (1999) Animal mitochondrial genomes. Nucleic Acids Res 27:1767–1780

Boore JL, Brown WM (1998) Big trees from little genomes: mitochondrial gene order as a phylogenetic tool. Curr Opin Genet Dev 8:668–674

Boore JL, Medina M, Rosenberg LA (2004) Complete sequences of the highly rearranged molluscan mitochondrial genomes of the scaphopod *Graptacme eborea* and the bivalve *Mytilus edulis.* Mol Biol Evol 21:1492–1503

Boyce TM, Zwich ME, Aquadro CF (1989) Mitochondrial DNA in the bark weevils: size, structure and heteroplasmy. Genetics 123:825–836

Burnzynski A, Zbawicka M, Skibinski DOF, Wenne R (2003) Evidence for recombination of mtDNA in the marine mussel *Mytilus trossulus* from the Baltic. Mol Biol Evol 20:388–392

Daniels GR, Deininger PL (1985) Repeat sequence families derived from mammalian tRNA genes. Nature 317:819–822

DeJong R, Emery AM, Adema CM (2004) The mitochondrial genome of *Biomphalaria glabrata*, intermediate host of *Schistosoma mansoni.* J Parasitol 90:991–997

Eberhard JR, Wright TF, Bermingham E (2001) Duplication and concerted evolution of the mitochondrial control region in the parrot genus amazona. Mol Biol Evol 18:1330–1342

Endo K, Noguchi Y, Ueshima R, Jacobs HT (2005) Novel repetitive structures, deviant protein-encoding sequences and unidentified ORFs in the mitochondrial genome of the brachiopod *Lingula anatina.* J Mol Evol 61:36–53

Fuller K, Zouros E (1993) Dipsersed discrete length polymorphism of mitochondrial DNA in the scallop *Placopecten magellanicus* (Gmelin). Curr Genet 23:365–369

Galli G, Hofstetter H, Birnstiel ML (1981) Two conserved sequence blocks within eukaryotic tRNA genes are major promoter elements. Nature 294:626–631

Gjetvaj B, Cook DI, Zouros E (1992) Repeated sequences and large-scale size variation of mitochondrial DNA: a common feature among scallops (Bivalvia: Pectinidae). Mol Biol Evol 9:106–124

Hoffmann RJ, Boore JL, Brown WM (1992) A novel mitochondrial genome organization for the blue mussel, *Mytilus edulis*. Genetics 131:397–412

Hoarau GS, Holla R, Lescasse W, Stam T (2002) Heteroplasmy and evidence for recombination in the mitochondrial control region of the flatfish *Platichthys flesus*. Mol Biol Evol 19:2261–2264

Jacobs HT, Asakawa W, Araki T, Miura K, Smith MJ, Watanabe K (1989) Conserved tRNA gene cluster in starfish mitochondrial DNA. Curr Genet 15:193–206

Janke A, Pääbo S (1993) Editing of a tRNA anticodon in marsupial mitochondria changes its codon recognition. Nucleic Acids Res 21:1523–1525

Kim SH, Je EY, Park DW (1999) *Crassostrea gigas* mitochondrial DNA. GenBank accession number AF177226

Kimura M (1980) A simple method for estimating evolutionary rate of base substitutions through comparative studies of nucleotide sequences. J Mol Evol 16:111–120

Kurtz S, Choundhuri JV, Ohlebusch E, Schleiermacher C, Stoye J, Giegerich R (2001) The manifold applications of repeat analysis on a genomic scale. Nucleic Acids Res 29:4633–4642

Kyte J, Doolittle RF (1982) A simple method for displaying the hydrophobic character of a protein. Mol Biol 157:105–142

Ladoukakis ED, Zouros E (2001a) Direct evidence for homologous recombination in mussel (*Mytilus galloprovincialis)* mitochondrial DNA. Mol Biol Evol 18:1168–1175

Ladoukakis ED, Zouros E (2001b) Recombination in animal mitochondrial DNA: evidence from published sequences. Mol Biol Evol 18:2127–2131

LaRoche J, Snyder M, Cook DI, Fuller K, Zouros E (1990) Molecular characterization of a repeat element causing large-scale size variation in the mitochondrial DNA of the sea scallop *Placopecten magellanicus.* Mol Biol Evol 7:45–64

Lemieux C, Otis C, Turmel M (2000) Ancestral chloroplast genome in *Mesostigma viride* reveals an early branch of green plant evolution. Nature 403:649–652

Levinson G, Gutman GA (1987) Slipped-strand mispairing: a major mechanism for DNA sequence evolution. Mol Biol Evol 4:203–221

Lowe TM, Eddy SR (1997) tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res 25:955–964

Lunt DH, Hyman BC (1997) Animal mitochondrial DNA recombination. Nature 387:247

Mizi A, Zouros E, Moschonas N, Rodakis GC (2005) The complete maternal and paternal mitochondrial genomes of the Mediterranean mussel *Mytilus galloprovincialis*: implications for the doubly uniparental inheritance mode of mtDNA. Mol Bio Evol 22:952–967

Ojala D, Merkel C, Gelfand R, Attaridi G (1980) The tRNA genes punctuate the reading of genetic information in human mitochondrial DNA. Cell 22:393–403

Okazaki M, Ueshima R (2001) Evolutionary diversity between the gender-associated mitochondrial DNA genomes of freshwater mussels. GenBank accession numbers AB055624 (male haplotype) and AB055625 (female haplotype)

Piganeau G, Gardner M, Eyre-Walker A (2004) A broad survey of recombination in animal mitochondria. Mol Biol Evol 21:2319–2325

Rand D (1993) Endotherms, ectotherms, and mitochondrial genome-size variation. J Mol Evol 37:281–295

Robinson-Rechavi M, Huchon D (2000) RRTree: relative-rate tests between groups of sequences on a phylogenetic tree. Bioinformatics 16:296–297

Rubio MA, Ragone FL, Gaston KW, Ibba M, Alfonzo JD (2006) C to U editing stimulates A to I editing in the anticodon loop of cytoplasmic threonyl tRNA in *Trypanosoma brucei*. J Biol Chem 281:115–120

Serb JM, Lydeard C (2003) Complete mtDNA sequence of the North American freshwater mussel, *Lampsilis ornate* (Unionidae): an examination of the evolution and phylogenetic utility of mitochondrial genome organization in bivalvia (Mullusca). Mol Biol Evol 20:1854–1866

Snyder M, Fraser AR, LaRoche J, Gartner-Kepkay KE, Zouros E (1987) Atypical mitochondrial DNA from the deep-sea scallop *Placopecten magellanicus*. Proc Natl Acad Sci USA 84:7595–7599

Thompson JD, Higgins DG, Gibson TJ (1994) Clustal W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22:4673–4680

Thyagarajan B, Padua RA, Campbell C (1996) Mammalian mitochondria possess homologous DNA recombination activity. J Biol Chem 271:27536–27543

Tsaousis AD, Martin DP, Ladoukakis ED, Posadaand D, Zouros E (2005) Widespread recombination in published animal mtDNA sequences. Mol Biol Evol 22:925–933

Turmel M, Lemieux C, Burger G, Lang BF, Otis C, Plante I, Gray MW (1999) The complete mitochondrial DNA sequences of *Nephroselmis olivacea* and *Pedinomonas minor:* two radically different evolutionary patterns within green algae. Plant Cell 11:1717–1729

Wolstenholme DR, Jeon KW (1992) A survey of cell biology. Int Rev Cytol 141:173–232

Yokobori S, Fukuda N, Nakamura M, Aoyama T, Oshima T (2004) Long-term conservation of six duplicated structural genes in cephalopod mitochondrial genomes. Mol Biol Evol 21:2034–2046

Zouros E, Ball AO, Saavedra C, Freeman KR (1994) Mitochondrial DNA inheritance. Nature 368:818