The Complete Mitogenome of Two Australian Lampreys: *Mordacia mordax* and *Mordacia praecox*

James Garbutt
B.Sc. Forensic Biology & Toxicology
School of Veterinary and Life Sciences
Murdoch University
2015

This thesis is submitted for the degree of Honours in Molecular Biology at Murdoch University.

I declare this thesis is my own account of my research and contains, as its main content, work which has not been previously submitted for a degree at any tertiary education institution.

James Garbutt.

Abstract

As one of only two surviving outgroups to all jawed vertebrates, lampreys (Petromyzontiformes) can provide important information in our understanding of early vertebrate evolution. However, with many phylogenetic aspects of lamprey evolution still uncertain, the ability to use contemporary lampreys in this role depends on robust phylogenetic hypotheses regarding the interrelationship of the three lamprey families, as well as the relationship between lampreys and hagfishes. To achieve this, complete mitogenome data of Southern Hemisphere lampreys is required. Another contentious issue in lamprey taxonomy is the status of paired species. Whilst many studies have focused on Northern Hemisphere species pairs, this study is the first to compare *Mordacia mordax* and *Mordacia praecox,* two lamprey species endemic to Australia and the only species pair in the Southern Hemisphere.

The complete mitochondrial genome of *Mordacia mordax* and *Mordacia praecox* was determined twice independently, in a single shotgun sequencing run on an Ion Torrent PGM, and using a combination of Sanger sequencing of short range PCR products and Roche 454 GS Junior pyrosequencing of long range PCR products. Both of the mitogenomes contain the 37 typical vertebrate genes. Their gene order and contents are identical to those of previously described lamprey mitogenomes, with the exception of a novel tandem repeat array located between *Cyt b* and tRNA proline. The tandem repeat array, referred to as NCIII, contains pseudogenes of tRNA proline and phenylalanine, indicating that it has arisen by tandem duplication of the tRNA proline – phenylalanine region. Characterisation of NCIII revealed that the number of repeat copies was polymorphic between individuals of both species, and was a source of both intra-individual and inter-individual variation. Consistent with other studies of lamprey species pairs, the mitogenome of *M. mordax* and *M. praecox* are nearly identical.

Phylogenetic analyses were carried out using the newly determined mitogenomes, together with five additional lamprey species and two hagfishes. Most tree topologies obtained strongly support the hypothesis that Petromyzontidae plus Geotriidae are a clade whose sister group is Mordaciidae. Additionally, lamprey divergence times were estimated by a temporally-calibrated phylogenetic analysis that included 20 vertebrate mitogenomes and was done using nine well-established fossil calibration points. The recovered topology strongly supported the hypothesis that lampreys separated from hagfishes about 409 MYA, and that lamprey divergence involved the early radiation of Mordaciidae (about 132 MYA), followed by the monophyletic divergence of Geotriidae plus Petromyzontidae about 85 MYA. Taken together, the results in this study provide robust hypotheses regarding the interrelationship of lamprey families and the relationship between hagfish and lampreys, whilst providing an estimation of their divergence times.

Acknowledgements

I would like to thank my supervisors, Dr Howard Gill, Dr David Berryman, Frances Brigg and Prof. Ian Potter, for their guidance and support throughout the tumultuous years. I would also like to thank the staff of the SABC, particularly Frances Brigg, for their expertise and supplying chemicals and reagents for next-generation sequencing. A special thank you to Karen Olkowski (Honours Coordinator) and Dr Howard Gill for their patience and understanding with my predicaments during my honours thesis.

This thesis was the bane of my life. It broke me mentally, financially and emotionally. Together with the passing of my pop, this ‘rollercoaster of an experience’ took me to a dark place. I’m glad to finally close this chapter of my life. It took a while, but I finally have closure in knowing that I didn’t give up, that I survived. I could not have done this without the love and support of my mum and nan. Most importantly, I could not have done this without my partner Justine, whom during those dark times, felt like the only source of light.

In Loving Memory of Richard Garbutt

Table of Contents

Section 3 Results

List of Figures

encompass the complete mitogenome. Also shown for comparison is the region determined by short range PCR (light blue as annotated). Note: Light-strand (5'– 3') is shown for simplicity. Arrowheads indicate the direction of transcription for coding genes (green arrows as annotated), rRNA (red arrows as annotated) and tRNAs (pink arrows).

Figure 11. Comparison of total nucleic acids before and after RNAse digestion. **Lane 1)** 100bp plus ladder, **Lane 2)** before RNAse digestion, and **Lane 3)** after RNAse digestion. Agarose gel (0.8%) electrophoresis for 1 hr and stained with SYBR safe. Comparison shows the two small bands (approximately 1750 bp and 1100 bp) and associated smearing (approximately between 2000 – 250 bp) in Lane 2 is RNA and that the large band (> 10 kb), present in Lanes 2 and 3, is intact DNA.

Figure 12. The *Cyt b – 16S* mitochondrial region and the short range PCR– Sanger sequencing strategy used in this study. The sequenced region (dotted box) was determined by Sanger sequencing of 12 amplicons (blue boxes). Note: the gene organisation of *M. mordax* as determined by Haouchar (2009) is shown. Arrowheads indicate the direction of transcription for each gene. The amplicon name consists of the primer pairs used to generate it.

Figure 13. Example of the electrophoresis profile for X50F/R amplicons of both *M. mordax* and *M. praecox* individuals in this study. Depending on the individual, the band is either approximately 950 or 800 bp in size (Table 9). **Lane 1)** 100 plus ladder, **Lane 2)**, *M. praecox* #3, **Lane 3)** *M. mordax* #4, and **Lane 4)** negative template control. Agarose gel (1.5%) after electrophoresis for 1hr at 90v and stained with SYBRsafe.

Note: the expected band size of 478 bp based on the complete *M. mordax* mitogenome of Haouchar (2009) was never seen in this study.

Figure 14. Example of the electrophoresis profile for X51F/R amplicons of both *M. mordax* and *M. praecox* individuals in this study. Both had two bands approximately 500 bp and 650 bp in size. The presence of the secondary band (650 bp) was unexpected based on the complete *M. mordax* mitogenome of Haouchar (2009). **Lane 1)** 100 plus ladder, **Lanes 2, 4 and 6)** empty wells, **Lane 3)** *M. praecox* individual #3, **Lane 5)** *M. mordax* individual #4, and **Lane 7)** negative template control. Agarose gel (0.8%) after electrophoresis for 1hr at 90v and stained with SYBRsafe.

Figure 15. Nucleotide alignment of the *Cyt b – 12S* mitochondrial region of *M. mordax* (top- Haouchar 2009) determined by Haouchar (2009) to that determined in *M. mordax* #1D in this study. The *Cyt b – 12S* mitochondrial region of *M. mordax* #1D was amplified using primer pair X50F/R and primer pair X51F/R. Gene regions are indicated by the coloured bar above the top sequence and are as annotated. The location and direction of the individual primers for primer pair X50F/R (blue text as annotated) and primer pair X51F/R (green text as annotated) are shown. Alternative (imperfect) primer binding sites for primer X51F and primer X50R (red text as annotated) are also shown. Highlighted in yellow is the additional 48 bp in *Cyt b* found in all six individuals sequenced in this study. The two repeats comprising noncoding region III of *M. mordax* #1D are underlined and annotated in bold. Note: light-strand (5’ – 3’) is shown for simplicity. Arrows indicate direction of transcription for each gene. Dots indicate identical nucleotides whilst dashes indicate deletions. // indicates continuation of the sequence/region.

Figure 16. The gene arrangement and location of NCIII (orange boxes) as determined in this study for **A)** *M. praecox* #3 and *M. mordax* #23, and **B)** *M. mordax* individuals #1D, #4, #24 and #25. The repeat copies comprising NCIII are annotated 5', i (= internal) and 3' as outlined in the text. A hatched orange box indicates partial repeat, a solid orange box indicates complete repeat. Note: arrowheads indicate the direction of transcription for each gene. Proline and phenylalanine tRNA genes are abbreviated as P and F, respectively.

Figure 17. L-strand nucleotide alignment of tRNA Proline – tRNA Phenylalanine mitochondrial region to the tandem repeat copy units of NCIII found in *M. mordax and M. praecox* individuals sequenced in this study. The tRNA Proline – tRNA Phenylalanine mitochondrial region (Pro-Phe Region) is identical in all individuals sequenced in this study. The repeats occur in a tandem orientation but are aligned here for comparison. Dots indicate identical bases while dashes indicate deletions. 5', i (= internal) and 3' refer to the order in which copies occur on the L-strand. PR = partial repeat and CR =complete repeat. Genes and other features are marked above the sequence and are abbreviated as used in the text. Transfer RNA genes are marked above the sequence and their features marked with: arrows (< or >) indicating direction of transcription; | indicating nucleotides part of the tRNA arm and are as annotated; and the anticodon is bold and underlined. The bind site for primer X51F is indicated by the red text and the bind site for primer X50R is indicated by the blue text.

Figure 18. Comparison of the predicted secondary structure of **A)** tRNA phenylalanine, and **B)** pseudoPhe. Blue and red dots signify Watson-Crick base pairing. Note: only the most stable secondary structure (i.e. largest -ΔG value) predicted by mfold is shown - these vary slightly to the secondary structures predicted by tRNAscan-SE due to

inherent differences in programming employed by the two webservers (i.e allowance of non-Watson-Crick base pairings such as G-U (pink dots).

Figure 19. Observed size of the amplicons (blue boxes) generated by primer pair X50F/R and primer pair X51F/R due to NCIII (orange boxes). **A)** Observed amplicon sizes in individuals with NCIII comprising three repeats, and **B)** Observed amplicon sizes in individuals with NCIII comprising two repeats. Note: primer pair X51F/R yields two amplicons due to an alternative X51F primer bind site (#2) located in the 3' repeat of NCIII. The repeat copies comprising NCIII are annotated *5′*, i (= internal) and 3' as outlined in the text. A hatched orange box indicates partial repeat, a solid orange box indicates complete repeat. Arrowheads indicate the direction of transcription for each gene. Proline and phenylalanine tRNA genes are abbreviated as P and F, respectively.

Figure 20. Expected size of the amplicon (blue box) generated by primer pair X50F/X51R based on the presence or absence of NCIII (orange boxes). **A)** Expected size if NCIII was a PCR artefact, **B)** Expected size if NCIII contains two repeats, and **C)** Expected size if NCIII contains three repeats. Note: the repeat copies comprising NCIII are annotated 5′, i (= internal) and 3' as outlined in the text. A hatched orange box indicates partial repeat, a solid orange box indicates complete repeat. Arrowheads indicate the direction of transcription for each gene. Proline and phenylalanine tRNA genes are abbreviated as P and F, respectively.

Figure 21. Comparison of X50F/X51R amplicon size between *M. praecox* and *M. mordax* individuals. Depending on the individual lamprey, the band is either approximately 1350 or 1200 bp in size. **Lane 1)** 100 plus ladder, **Lane 2)** *M. praecox*

#3, **Lanes 3 – 6)** *M. mordax* individuals #4, #23, #1D and #2, and **Lane 7)** negative template control. Agarose gel (0.8%) after electrophoresis for 1hr at 90v and stained with SYBRsafe.

Figure 22. Sequencing coverage of NCIII in *M. praecox* #3 (line 1). Due to the priming sites of primers X50R (yellow arrow as annotated) and X51F (blue arrow as annotated) being located within the tRNA genes phenylalanine (pink F arrow) and proline (pink P arrow), respectively, it was possible that the tandem repeats of NCIII (three orange boxes) may have been due to alternative (imperfect) priming sites (blue and yellow arrows annotated alt). Amplification was carried out using primers X50F (dark blue arrow as annotated) and X51R (gold arrow as annotated) as their binding sites flanked NCIII. Sanger sequencing and inclusion of X50F/X50R sequence (line 4) shows that the region is not due to alternative priming as it independently supports the sequences of amplicon X50F/R (line 3) and amplicon X51F/R (line 5). Note: L-strand consensus sequences are shown (5' – 3') for simplicity. Each amplicon consensus sequence has two times coverage. Direction of primer arrows indicates amplification direction. Only a portion of the *Cyt b* (green) and *12S* rRNA (red) gene are shown. The repeat copies comprising NCIII are annotated 5′, i (= internal) and 3' as outlined in the text.

Figure 23. The overlapping long range PCR method of generating the expected 17kb mitogenome of *M. mordax and M. Praecox.* The 5 long range PCR primer pairs would generate amplicons of 5 – 6 Kb in size (dark blue as annotated) that would encompass the complete mitogenome. Also shown for comparison is the region determined by short range PCR (light blue as annotated). Note: L-strand (5' – 3') is shown for simplicity. Arrowheads indicate the direction of transcription for each coding gene (green arrows as annotated), rRNA (red arrows as annotated) and tRNAs (pink arrows).

Figure 24. Example of the 454 sequencing reads spanning tRNA isoleucine of *M. praecox* #16. The consensus sequence (top - line 1) indicates that the cytosine homopolymer (C stretch enclosed in the black box) is six bases in length as the number of 454 sequencing reads with 6C's (lines 2 – 6) is greater than the number of 454 sequencing reads with 5C's (lines 7 – 10). Note: only nine sequencing reads are shown as an example. Only the tRNA isoleucine portion of each read is shown for simplicity. As indicated, the sequences are shown in the 5' – 3' direction.

Figure 25. The predicted secondary structure of tRNA isoleucine and the predicted effect of the homopolymer (bold C). Shown is the secondary structure for *M. mordax* #5, *M. mordax* #25 , and *M. praecox* #1 in which the 5C homopolymer forms the T loop and T arm. The orange box shows the 4C homopolymer increasing the size of the T loop for *M. praecox* #17. The blue box shows the 6C homopolymer found in *M. praecox* #16 causing a secondary loop in the T arm. Blue and red dots denote Watson-Crick base pairing. Note: in *M. mordax* #25 the purple C is a U.

Figure 26. Example of the 454 sequencing reads spanning first 138 bp of *ND4* of *M. mordax* #4. The consensus sequence (top - line 1) indicates that the adenine homopolymer (A stretch enclosed in the black box) is six bases in length as the number of 454 sequencing reads with 6A's (lines 3 – 8) is greater than the number of 454 sequencing reads with 7A's (lines 9 – 11). However, 6A's is incorrect as it results in premature stop codons (black boxes; line 2), whilst 7A's is correct as it changes the reading frame to eliminate the stop codons (line 11). Note: only nine sequencing reads are shown as an example. Only 138 bases of each read are shown for simplicity. As indicated, the sequences are shown in the 5' – 3' direction.

Figure 27. Homopolymer length and its effect on *ND6* amino acid sequence. Nucleotide alignment of the first 99 bases of *ND6* gene of *M. mordax* #25 (line 1) to the first 98 bases of *M. praecox* #1 (line 2). Both sequences were determined by 454 sequencing and differ only by a single nucleotide due to variation in the thymine homopolymer (T stretch enclosed in the orange box). Underestimation of the homopolymer (6T's) in *M. praecox* #1 causes a frame shift, resulting in premature stop codons (black boxes: line 4) and a gene length of only 52 bases (13 amino acids). In comparison, the homopolymer in *M. mordax* #25 is 7 bases (line 1) and results in the expected gene length of 522 bases (174 amino acids). Note: as indicated, the sequences are shown in the 3' – 5' direction due to *ND6* being located on the L-strand. Dash in pink box indicates nucleotide is not present.

Figure 28: The complete annotated circular mitogenome of **A)** *M .praecox* individual #3, and **B)** *M. mordax* individual #25. Note: L-strand shown in 5' – 3' direction. Arrowheads indicate the direction of transcription. Genes located outside the circle are encoded by the heavy-strand. Serine and leucine tRNA genes are also identified by codon family (in parentheses). See Table 18 for exact gene locations and abbreviations.

Figure 29. Schematic comparison of the absence and presence of NCIII in the mitochondrial genome of lampreys sequenced to date. Noncoding region III (represented by the orange boxes) is located between genes Cytochrome b (green arrow box as annotated) and tRNA Proline (first pink arrow box as annotated). **A)** The gene arrangement and absence of NCIII in all lampreys published to date. **B)** The gene arrangement and location of NCIII as determined in this study for *M. mordax* #25. The 289 base pair NCIII of this individual contains two tandem repeats: a 177 bp repeat (annotated 3’) and a 112 bp near identical partial repeat (annotated 5’). **C)** The gene arrangement and location of NCIII as determined in this study for *M. praecox* #3. The 437 base pair NCIII contains three tandem repeats: a 164 bp repeat (annotated 3’), a 164 bp identical repeat (annotated i) and a 109 bp near identical partial repeat (annotated 5’). Note: arrowheads indicate the direction of transcription for each gene. Proline and phenylalanine tRNA genes are abbreviated as P and F, respectively.

Figure 30. Nucleotide comparison of the two tandem repeat copies comprising the 289 bp NCIII of *M. mordax* individual #25. Bold and underlined is the 3 base pair insertion that differentiates the partial copy from the copied sequence. Note: L-strand (5’ – 3’) is shown for simplicity. 5' and 3’ refer to the order in which the repeat copies occur on the L-strand, and PR= partial repeat and CR= complete repeat.

Figure 31. Nucleotide comparison of the three tandem repeats comprising the 437 bp NCIII of *M. praecox* #3. Bold and underlined is the single purine transition change. Note: L-strand (5’ – 3’) is shown for simplicity. 5' , i , and 3' refer to the order in which the repeat copies occur on the L-strand, and PR= partial repeat and CR= complete repeat.

Figure 32. Nucleotide alignment of NCIII of *M. mordax* (#25) and *M. praecox* (#3). Whilst the overall region is only 59.2% identical between the two individuals, the 5’-partial repeat (annotated 5’PR) is 97.3% identical and the 3’-complete repeat (3’CR) is 89.8% identical. Bold and underlined are the differences between the nucleotide sequences. Dots indicate identical bases while dashes indicate deletions. The repeat copies are annotated (below the sequence) 5', i (= internal) and 3' , referring to the

Figure 37. Ancestral arrangement: H-strand encoded *Cyt b* overlaps L-strand encoded tRNA proline; the template sequence of tRNA proline contains stop codon for *Cyt b*. Intermediate arrangement: Duplication of Pro-Phe region followed by Pseudogenization of the original tRNA genes. Pseudogenization of the original Pro (by insertions and deletions) shifts the stop codon for *Cyt b*, thus enabling *Cyt b* to increase from 1191 bp (seen in other lampreys) to 1218 bp (seen in this study). Further tandem duplications are enabled by strand slippage during DNA replication to contract and expand the number of tandem repeats, giving rise to the current arrangement in individuals containing a tandem array of 2 or 3 repeats. Note: pseudogenes represented by orange boxes, tRNA genes indicated by pink boxes and are annotated as P (proline) and F (phenylalanine).

List of Tables

Section 1: Introduction

1.1 Importance of Lampreys

Lampreys are ancient 'fishes' belonging to the order Petromyzontiformes and, together with hagfishes (Myxiniformes), are the only living jawless (agnathan) chordates (Hubbs & Potter 1971; Hardisty 1982; Potter & Gill 2003). As the closest living outgroup to all jawed vertebrates (gnathastomes), these two groups provide important information in our understanding of the evolution, biology and genome of early vertberates (Lee & Kocher 1995; Kuratani *et al.* 2002; Richardson & Wright 2003; Osório & Rétaux 2008; Smith *et al.* 2010). This is particularly the case for lampreys, in which all life stages are far more accessible than hagfishes (Gess *et al.* 2006). Lampreys are now playing a major role in our understanding of early vertebrate vision (Lamb *et al.* 2007; Collin *et al.* 2009) and the development of treatments for spinal cord injury (McClellan 2013), whilst the recent discovery of genes linked to human neurodegenerative disorders (such as Alzheimer's and Autism) in the sea lamprey (Smith *et al.* 2013) will ensure that lampreys remain of great significance in the near future.

1.2 Cyclostome Debate

Despite their scientific importance, much of the evolutionary history of hagfishes and lampreys is uncertain (Lee & Kocher 1995) with many phylogenetic aspects of "considerable controversy" (Hardisty 1982; Linzey 2012) and heavily debated. Such is the case regarding the phylogenetic relationship between living lampreys, hagfishes and gnathostomes. The debate between the "cyclostome theory" (i.e. hagfish and lamprey are a sister group to gnathostomes (Figure 1A)) and the "vertebrate theory" (i.e. hagfish is the sister to lampreys and gnathostomes (Figure 1B)) (reviewed by Hardisty 1982; Forey & Janvier 1993; Delarbre *et al.* 2000; Meyer & Zardoya 2003; Near 2009; Janvier

2010; Smith *et al.* 2010) dates back to the late 1970s, and whilst the debate is still not fully resolved, the work by Heimberg *et al.* (2008) has changed the opinions of some sceptics (Janvier 2010) in favour of cyclostome monophyly.

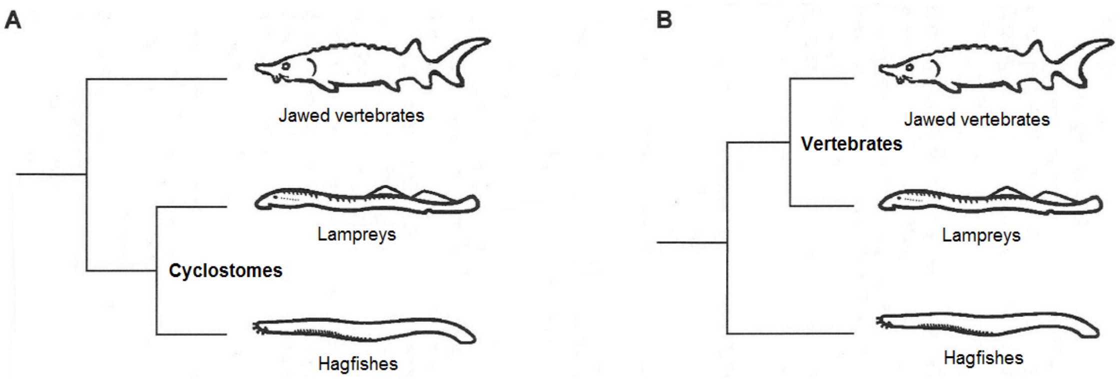


Figure 1. Two competing phylogenetic hypotheses regarding the relationship between extant hagfishes, lampreys and jawed vertebrates. **A)** The cyclostome hypothesis: hagfishes and lampreys are more closely related to one another than either is to jawed vertebrates. **B)** The vertebrate hypothesis: lampreys are more closely related to jawed vertebrates than either is to hagfishes. Adapted from Meyer and Zardoya (2003).

1.3 Lamprey Fossil History

Due to lampreys having little mineralized tissue, such as bone or calcified cartilage, lamprey fossils are seldom well preserved and consequently the evolutionary history of lampreys is poorly known (Chang *et al.* 2006). However, four unambiguous and well preserved lamprey species are known from fossils (Renaud 2011): *Mayomyzon pieckoensis* Bardack and Zangeri (1968), *Hardistiella montanensis* Janvier & Lund (1983) , *Mesomyzon mengae* Chang *et al.* (2006) and *Priscomyzon riniensis* Gess *et al.* (2006) (Figure 2). Arguably the two most important of these fossil finds are those of *P. riniensis and M. mengae*. The fossil of *P. riniensis* dates back 360 million years ago (MYA) to the Famennian period of the Late Devonian Epoch in South Africa (Gess *et al.* 2006). This marine/estuarine lamprey had key specializations found in modern day parasitic lampreys such as annular cartilage, a large oral disc and, unlike in the slightly younger *M. pieckoensis* and *H. montanensis,* obvious circumoral teeth. This finding was particularly important as the fossil indicated that lamprey morphology and parasitic feeding habit had been stable for at least 360 million years, and that jawless fish morphologically 'close' to extant lampreys had evolved before the end of the Devonian period, some 35 million years earlier than previously thought (Gess *et al.* 2006; Janvier 2006).

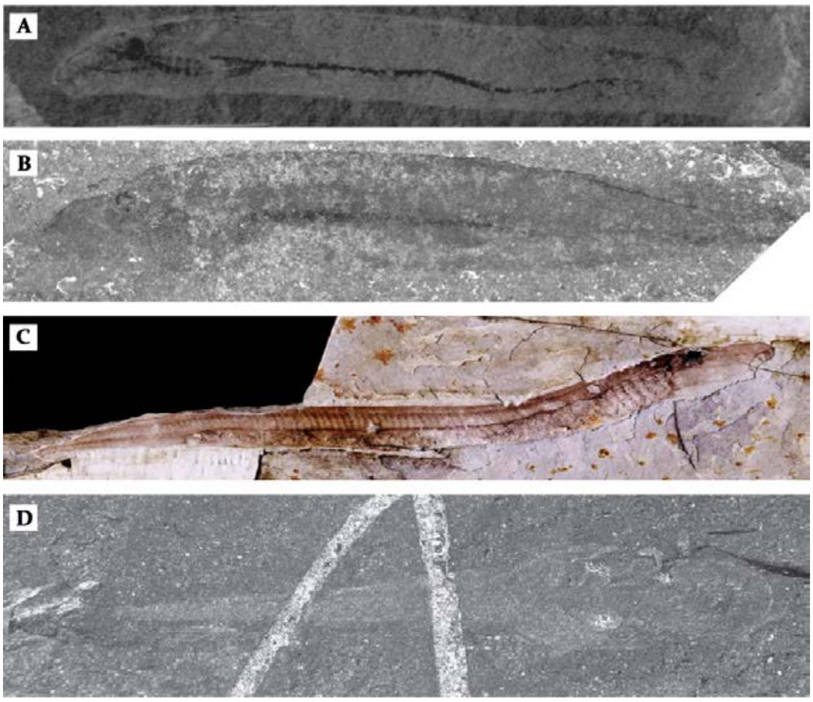


Figure 2. Four unambiguous fossil lampreys. Composite image of **A)** *Mayomyzon pieckoensis*, **B)** *Hardistiella montanensis*, **C)** *Mesomyzon mengae* and **D)** *Priscomyzon riniensis*. Images are from Bardack and Zangeri (1968), Janvier and Lund (1983), Chang *et al*. (2006) and Gess *et al.* (2006), respectively.

The fossil of *M. mengae*, a landlocked freshwater lamprey dating back 120 MYA to the Early Cretaceous Epoch in China (Chang et al. 2006), bridged the gap between extant lampreys and the two previous Carboniferous (c.a 300-330 MYA) lampreys *H. montanensis* and *M. pieckoensis* (Osório & Rétaux 2008). Morphologically nearly identical and considered as the closest fossil relative to living lampreys (Janvier 2006), *M. mengae* indicates that lampreys had approximated their current morphology by the Early Cretaceous and have undergone few major morphological changes over the past 100 million years (Chang *et al.* 2006). Furthermore, this fossil was found in freshwater deposits, was under 100mm in total length and possessed well developed gonads (Chang *et al.* 2006), suggesting that nonparatism in lampreys has been established for at least 120 million years (see later for description of life cycles).

1.4 Lamprey Distribution

Living lampreys have an anti-tropical distribution that is probably related to their relatively low thermal tolerance (max 31.4°C) (Potter & Beamish 1975; Hardisty 1982; Renaud 2011). For example, Meeuwig *et al.* (2005) showed that larval survival in *Entosphenus tridentatus* (Pacific lamprey, as *Lampetra tridentata*) and *L. richardsoni* (western brook lamprey) was significantly reduced and developmental abnormalities increased significantly at 22°C. Furthermore, as there are no records of adult anadromous species being caught within tropical marine habitats, it is also likely that adults have low thermal tolerances (Figure 3). In the Northern Hemisphere, the 37 species of Petromyzontidae are found in the cooler waters of North America, Europe, and Asia, with the distributions of the 8 genera varying in location (Figure 3) and range (Hubbs & Potter 1971). The Southern Hemisphere contains the monogeneric Mordaciidae (3 species) and monospecific Geotriidae, with Mordaciidae restricted to temperate waters of south-eastern Australia (*Mordacia mordax* and *Mordacia praecox*) and southern Chile (*Mordacia lapicida*), while *Geotria australis* is widely distributed, occurring in the waters of New Zealand and the southern regions of both South America and Australia (Potter & Strahan 1968) (Figure 3). Given that there is far more land mass situated above the tropics in the Northern Hemisphere, and thus more habitats suitable for larvae development, it is not surprising that the majority of lampreys are petromyzontids.

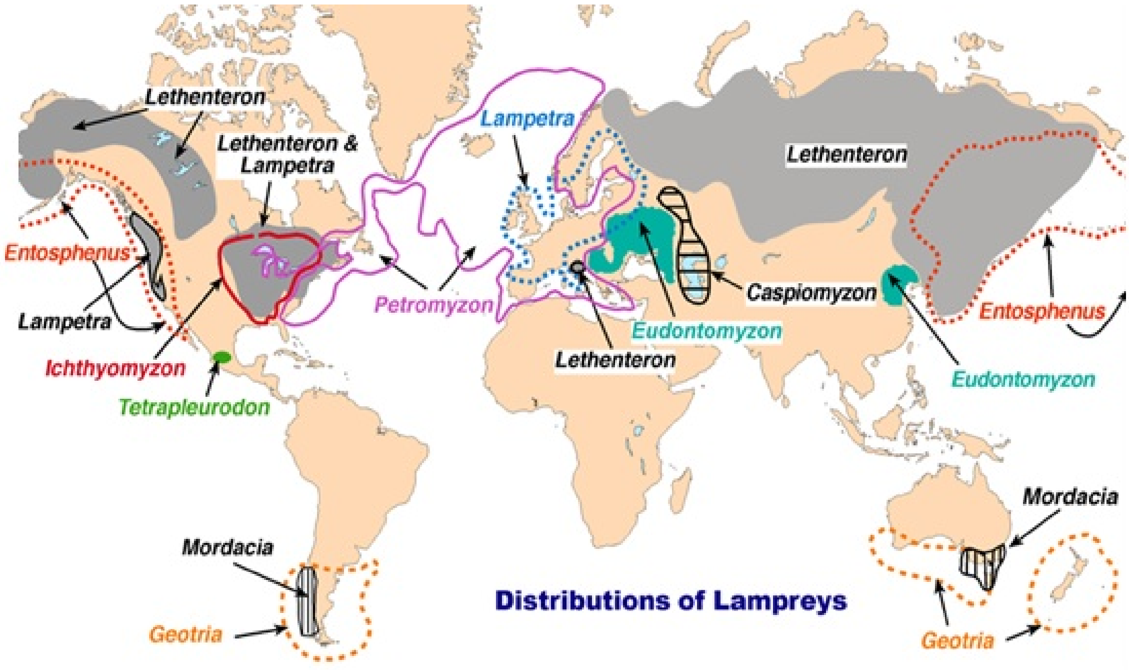


Figure 3. World distribution of living lamprey genera (with permission of Rick Mayden).

1.5 Lamprey Life Cycles

The larval phase of all 41 species of lampreys is spent in freshwater as a blind, filter-feeding, microphagous larva known as an ammocoete (Hardisty & Potter 1971a; Docker 2009; Potter *et al.* 2014). The morphology of an ammocoete is remarkably different from that of an adult lamprey and lacks the important characteristics used in species identification, e.g. oral disc dentition (Hardisty & Potter 1971b; Potter 1980; Youson 1980; Potter *et al.* 1982; Potter *et al.* 2014). After a number of years (depending on both interspecific and intraspecific variation (Renaud 2011)) the ammocoete undergoes metamorphosis into an adult lamprey (post-metamorphosed individual) which can last several months (Figure 4). During this time, the eyes become uncovered and fully functional, the dorsal fins, branchial openings and pigmentation become fully developed and the oral hood of the ammocoete is transformed into a tooth-bearing, circular suctorial disc (Hardisty & Potter 1971b; Potter *et al.* 1982).

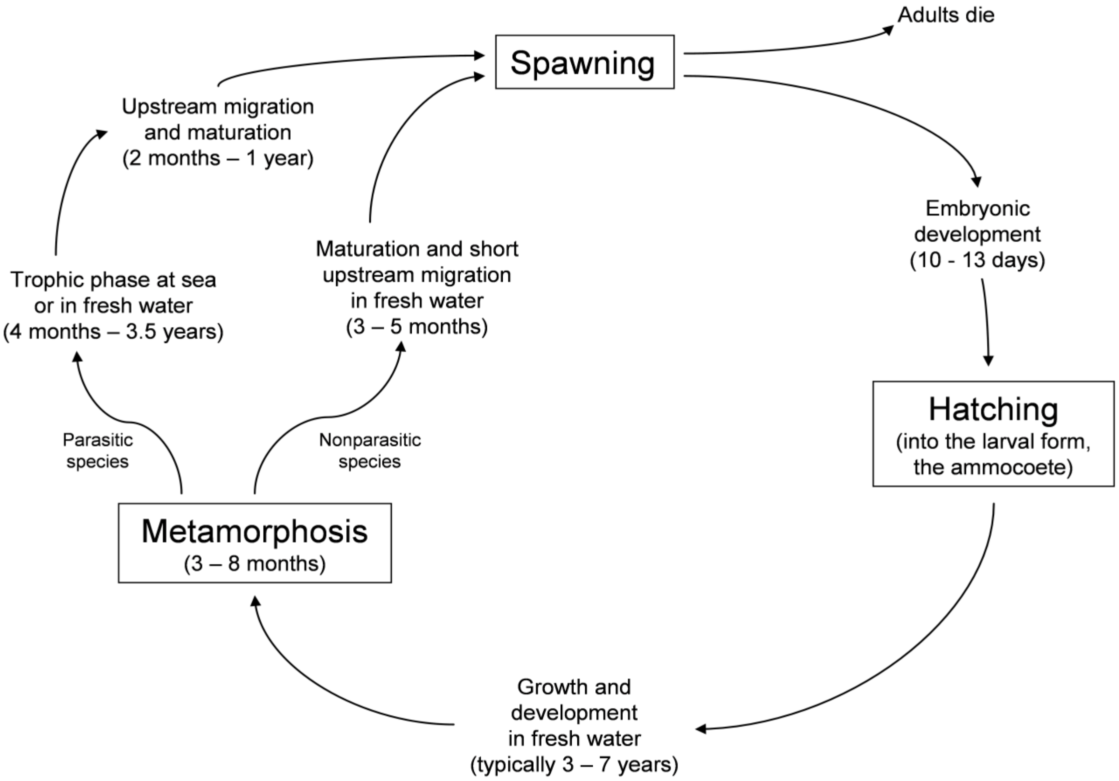


Figure 4. The general life cycle of lampreys (with permission of Ian Potter, Claude Renaud and Howard Gill).

Depending on the species, lampreys can be classified into three types based on mode of life: freshwater nonparasitic or brook lampreys (23 species); strictly freshwater parasitic (9 species); and anadromous parasitic (9 species) (Table 1). In 10 cases, a single parasitic form is believed to have given rise to one or multiple nonparasitic species (Table 1). This hypothesis is based on the fact that these species, which are referred to as paired species, have ammocoetes which are essentially indistinguishable from one another and whose adults have very similar patterns of dentition (a major taxonomic character in lampreys) (Potter *et al.* 2014). Indeed, the main distinguishing feature is the far larger size of the parasite in comparison to the presumed nonparasitic derivative at sexual maturation (Hardisty & Potter 1971c; Hardisty 2006; Docker 2009).

Table 1. Mode of life for all 41 living lamprey species that comprise three families. Modified from Potter *et al.* (2014).

Family	Anadromous Parasitic	Freshwater Parasitic	Freshwater Nonparasitic
Mordaciidae	*Mordacia mordax* [1]		*Mordacia praecox* [1]
	Mordacia lapicida		
Geotriidae	*Geotria australis*		
Petromyzontidae	*Caspiomyzon wagneri*		
	Petromyzon marinus		
		Ichthyomyzon unicuspis [2]	*Ichthyomyzon fossor* [2]
		Ichthyomyzon castaneus [3]	*Ichthyomyzon gagei* [3]
		Ichthyomyzon bdellium [4]	*Ichthyomyzon greeleyi* [4]
		Tetrapleurodon spadiceus [5]	*Tetrapleurodon geminis* [5]
	Entosphenus tridentatus [6]		*Entosphenus lethophagus* [6]
		Entosphenus minimus	
		Entosphenus similis	
		Entosphenus macrostomus	
			Entosphenus folletti
			Entosphenus hubbsi
	Lethenteron camtschaticum [7]		*Lethenteron alaskense* [7]
			Lethenteron appendix [7]
			Lethenteron reissneri [7]
			Lethenteron kessleri [7]
			Lethenteron zanandreai
			Lethenteron ninae
		Eudontomyzon danfordi [8]	*Eudontomyzon mariae* [8]
			Eudontomyzon stankokaramani [8]
		Eudontomyzon morii	
			Eudontomyzon hellenicus
			Eudontomyzon graecus
	Lampetra ayresii [9]		*Lampetra pacifica* [9]
			Lampetra richardsoni [9]
			Lampetra aepyptera
	Lampetra fluviatilis [10]		*Lampetra planeri* [10]
			Lampetra lanceolata [10]

Numbered superscripts indicate paired species.

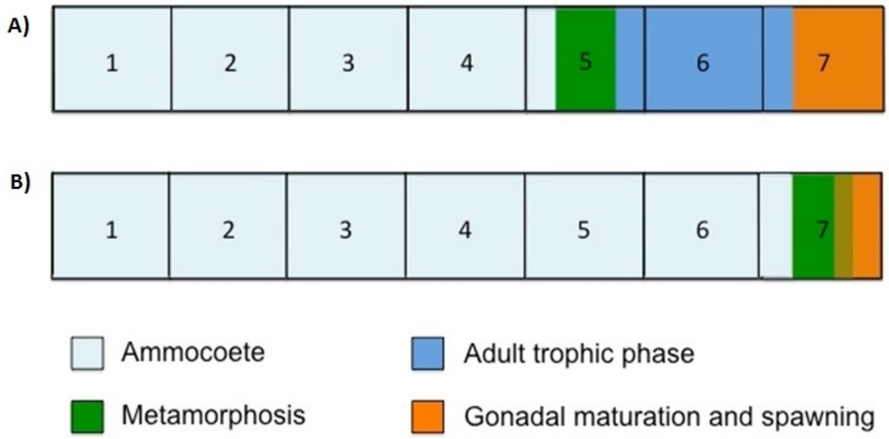


Figure 5. A comparison of the durations of the different phases in the life cycles of **A)** the anadromous parasitic *Lampetra fluviatilis,* and **B)** its nonparasitic derivative *Lampetra planeri*. From Potter *et al.* (2014).

After completion of metamorphosis, parasitic adult lampreys undergo a trophic phase at sea (anadromous parasitic) or in fresh water (freshwater parasitic). During this adult feeding phase, parasitic lampreys can either feed on blood, flesh, or both (Renaud *et al.* 2009). Lasting up to several years (Figures 4 and 5), the lamprey grows considerably in size (two to three times greater than at metamorphosis (Docker 2009)) before becoming sexually mature and migrating upstream for spawning (Figure 4). In contrast, the larval stage of nonparasitic lampreys is prolonged by several years in lieu of an adult trophic phase, sexual maturation begins during metamorphosis and the upstream migration is shorter than that of the parasitic ancestor. Notwithstanding these differences, members of a species pair from similar environments have life cycles of the same duration (Figure 5) (Hardisty & Potter 1971c). It is worth noting that, due to the eggs in a species pair being of similar size, the shorter nonparasite has considerably reduced fecundity (Hardisty & Potter 1971c; Docker 2009). Prior to the act of spawning, nest building in the stream bed is initiated, in most species, by the male lamprey via a combination of ‘vibratory activity’ and stone-moving behavior, with the nest further enlarged by copulatory pairing acts (Hardisty & Potter 1971b; Hardisty 2006). With the female attached to a stone near the leading edge of the nest, the male, using its oral disc, attaches himself to the back of the female’s head. The act of spawning involves the male squeezing the female’s body with his tail, thereby forcing the oocytes through the cloacal opening. To ensure effective fertilization of the eggs, it is generally believed that the male must be of similar body length due to the dependency on precise and intricate positioning of the tail of the male to cloacal opening of the female (Hardisty & Potter 1971b, c; Docker 2009). For this reason, lamprey speciation is believed to be due to size-assortative mating resulting in immediate reproductive isolation, therefore, paired species are generally considered to be “good biological species”, especially for those living in sympatry (but see Docker 2009). Death of the adult lamprey after spawning is predominately due to exhaustion of its energy reserves from 6 – 9 months of starvation (Hardisty 2006), which cannot be replaced due to severe atrophy of the liver and intestines.

1.6 Lamprey Phylogeny

There has been much debate regarding the relationships of the various taxa of lampreys. For example, Hubbs and Potter (1971) considered that Petromyzontidae comprised 7 genera with *Lampetra* comprising 3 subgenera, whereas Vladykov and Kott (1979) considered *Entosphenus*, *Lethenteron* and *Lampetra* at the generic level but failed to recognize *Okkelbergia* as a valid genus (Figure 6A, B). In contrast to both of these authors, Bailey (1980) considered that *Tetrapleurodon*, *Entosphenus*, *Okkelbergia*, *Eudontomyzon* and *Lethenteron* were all subgenera of *Lampetra* (Figure 6C). In addition to differences in what constituted genera and subgenera, these authors also differed in their approach to families, subfamilies and tribes, and neither Hubbs and

Potter (1971) or Bailey (1980) could determine the relationships of the two Southern Hemisphere families (as defined in this study) to one another or to the Northern Hemisphere lampreys. Vladykov and Kott (1979) failed to include the Southern Hemisphere lampreys in their considerations (Figure 6B). The first cladistic analyses based on morphological characteristics of all parasitic species of lampreys and that utilized fossils as an outgroup (Gill *et al.* 2003), also failed to determine the relationships of the three families but produced a hypothesis of the relationships of the Northern Hemisphere that was similar to that of Hubbs and Potter (1971) (Figure 7). The studies above were all based on predominately morphological characteristics, but as noted, none could resolve the relationships of the three families. This is not, however, surprising as 1) hagfishes and gnathostomes possess virtually none of the characters that are phylogenetically informative within lampreys, which results in the inability to select appropriate outgroups (Gill *et al.* 2003) and thus an inability for models to polarize characters using outgroup analysis (Stock & Whitt 1992). 2) Very few phylogenetically informative characters are preserved in lamprey fossils, thus reducing their usefulness as outgroups, which again leads to the inability to polarize many characters (Gill *et al.* 2003). 3) Within the extant lampreys, there is a paucity of phylogenetically informative characters, for example, Gill *et al.* (2003) found that only 20 characters from the 32 utilized were phylogenetically informative. Furthermore, during that study, well over 100 morphological characters were explored (Gill, H., pers. comm., 2014). In addition to these problems, the incredible similarity of most parasitic lampreys and their presumed nonparasitic derivative(s) means that morphological characteristics provide virtually no additional phylogenetic information (Gill, H., pers. comm., 2014).

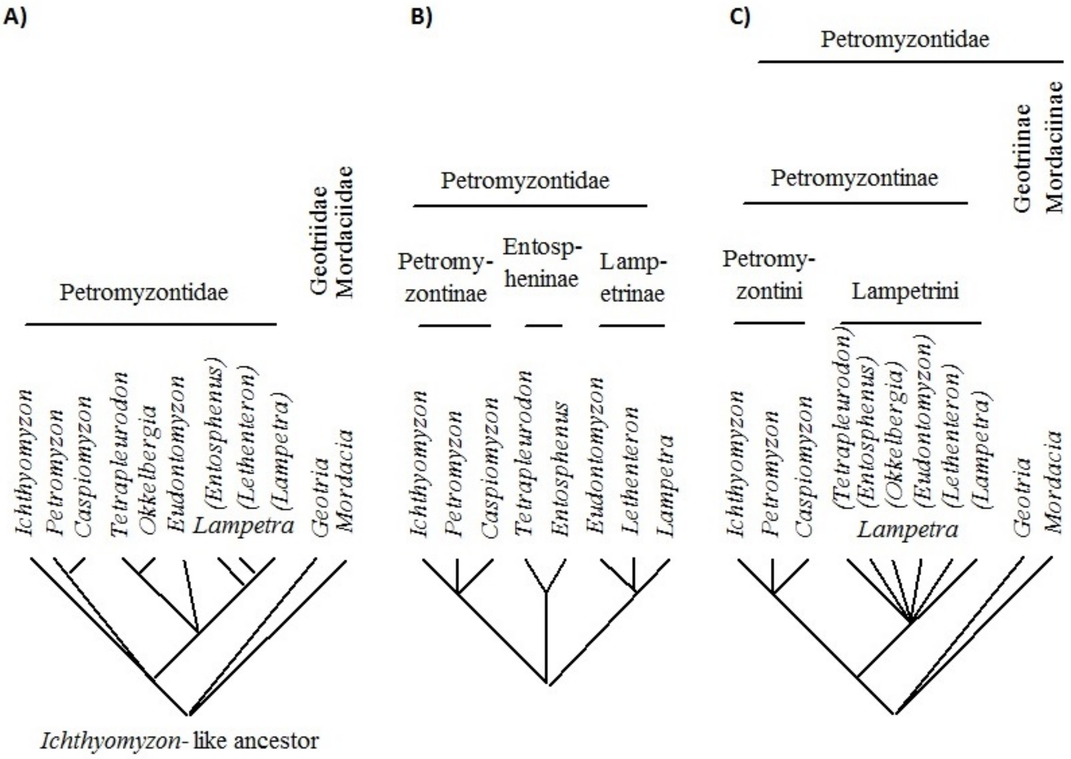


Figure 6. Taxonomic schemes and either the proposed or implicit phylogenies of the extant lampreys as proposed by **A)** Hubbs and Potter (1971), **B)** Vladykov and Kott (1979) and **C)** Bailey (1980).

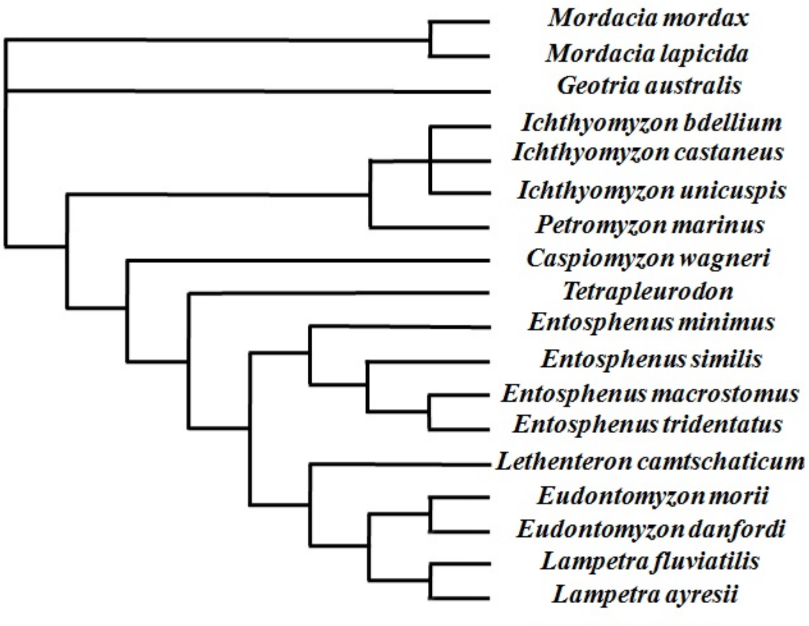


Figure 7. Phylogeny of the extant parasitic lampreys based on predominately morphological characteristics. From Gill *et al*. (2003).

In an attempt to resolve the relationships between the three extant lamprey families, and the relationships between paired species, Lang *et al*. (2009) generated sequence data for Cytochrome b (*Cyt b*) of 40 species of lampreys. Of the 1133 nucleotides that could be aligned, 610 were phylogenetically informative. Both parsimony and Bayesian analyses failed to provide robust and well resolved phylogenies. For example, although Bayesian analysis suggested that *Geotria* was sister to the Petromyzontidae, this relationship was poorly supported in the parsimony analysis, whilst both analyses suggested that Mordaciidae, *Lampetra*, *Lethenteron*, and *Eudontomyzon* are all polyphyletic taxa. Furthermore, the trees produced by the two analyses differed in their topologies. For example, in the parsimony analysis, *Caspiomyzon wagneri* + *Eudontomyzon hellenicus* are sister to all other Northern Hemisphere lampreys, whereas in the Bayesian analysis, they are sister to *Petromyzon* + *Ichthyomyzon.*

Whilst other phylogenetic studies using molecular markers have investigated certain groups of the Northern Hemisphere lampreys (e.g. Docker 1999; Espanhol *et al.* 2007; Mateus *et al.* 2011; Boguski *et al.* 2012) , none has included representatives of either of the two families of Southern Hemisphere lampreys, and only two have used more than one complete mitochondrial gene (Espanhol *et al.* 2007; Boguski *et al.* 2012). The inability for morphology alone, or the use of a small number of genetic markers to resolve the interrelationships of lampreys, highlights the importance of using a range of molecular markers in any future attempts to develop robust hypotheses regarding lamprey relationships, a view strongly supported by Potter *et al*. (2014, 2015). In the past, development of a large number of molecular markers was both cost and time prohibitive, however, with the development of next-generation sequencing, the ability to undertake studies that require a large number of markers is now possible.

1.7 Next-Generation Sequencing

Whilst an in-depth discussion of next-generation sequencing (NGS) is far beyond the scope of this thesis, this section attempts to provide a brief overview of the technologies used in this study. For further reading, the reader is referred to: Morozova and Marra (2008); Metzker (2010); Glenn (2011) and Ku *et al.* (2013), which are just a few of the dozens of excellent review papers on the topic.

Dideoxy Sanger sequencing (often referred to as Sanger or first-generation sequencing) by dideoxy chain termination (Sanger *et al.* 1977) was the dominant method of DNA sequencing until the introduction of the first NGS platform in 2005 (Morozova & Marra 2008; Metzker 2010; Niedringhaus *et al.* 2011; Ku *et al.* 2013). Since then, developments in high throughput sequencing technologies have been rapid and, at time of writing, there are at least 12 NGS platforms on the market which can be further classified as second-generation and third-generation sequencing technologies. Indeed, the advances of the platforms themselves have been so fast that comparative reviews about performance, accuracy, run costs, read lengths and sequencing output data are quickly outdated (Hert *et al.* 2008; Shendure & Ji 2008). A decade later, Sanger sequencing instruments (like the ABI 3730xl from Applied Biosystems used in this study) are now largely supplanted by NGS platforms for large sequencing projects, e.g. whole genomes or large gene regions from one or several individuals. This is due to the implications of throughput (Box 1, page 16), that is, the amount of DNA that can be sequenced in a single sequencing reaction, and its impact on time and cost efficiency. For Sanger sequencing machines, the limitations imposed by capillary array electrophoresis results in a maximum throughput of approximately 96 thousand bases (kilobases or kb) using a 96 capillary array, due to only 96 reads (DNA sequences) of up to 1000 bases being sequenced individually in parallel in a single run. In contrast, the throughput of NGS platforms range from 10 million bases (megabases or Mb) to >1 billion bases (gigabases or Gb) due to their characteristic ability to perform “massively parallel sequencing” of hundreds of millions of DNA reads (Ku *et al.* 2013). Consequently, NGS is significantly more time/labor efficient and cost effective (cost per Mb) for large scale sequencing than Sanger sequencing due to the considerable reduction in both number of sequencing runs and consumable costs (Niedringhaus *et al.* 2011). For example, the sequencing of an entire human genome by Sanger sequencing is currently estimated to cost at least $5 million USD and, if determined using a single machine, would take approximately 60 years to complete (Rizzo & Buck 2012). In contrast, the latest NGS platform (HiSeqX^{TM} by Illumina) is capable of sequencing 32 human genomes a week for close to $1000 each, with each nucleotide read by the machine an average of 30 times due to a maximum throughput of 1.8 terabases (Illumina 2014).

However, NGS platforms are not without their limitations. Currently, lack of accuracy in comparison to Sanger (Table 2) is the major limitation of NGS and, for this reason, Sanger sequencing is still considered the gold standard method (due to raw base accuracy of up to 99.999%) (Niedringhaus *et al.* 2011; Ku *et al.* 2013) and often used to validate NGS findings (Vogl *et al.* 2012; Ku *et al.* 2013). Another limitation of NGS is the shorter read lengths (usually <400 bases) in comparison to Sanger (Table 2). Consequently, NGS can have issues with read assembly, especially in *de novo* assembly, due to the difficulties associated with the assembly of billions of short reads into large contiguous DNA sequences (contigs). Repetitive regions, particularly those that are larger than the read length, are especially problematic as they can cause ambiguities in read alignment, which can result in misassemblies that can ultimately result in the erroneous reconstruction of the sequenced region (Pop & Salzberg 2002; Schatz *et al.* 2010; Treangen & Salzberg 2012). For example, reads from two copies of a repeat can be incorrectly combined (collapsed) such that the misassembled contig only contains a single repeat, and, if there is a unique region flanked by two repeats, it can be erroneously discarded as an unmappable contig (Figure 8).

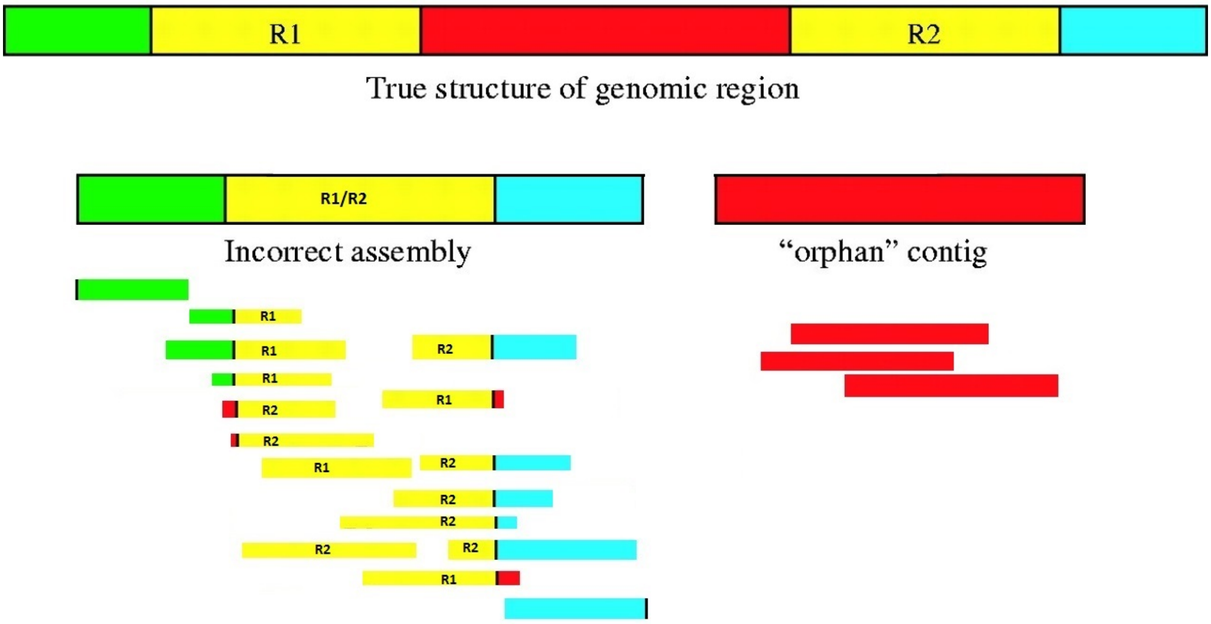


Figure 8. Schematic representation of a misassembled genomic region due to small sequencing reads. The genomic region being sequenced (top) contains two identical repeats (yellow; annotated R1 and R2) which are flanked by unique regions (green, red and blue). Due to the sequencing reads being smaller than the repeats, the individual reads do not extend sufficiently into the unique sequences. Consequently, the assembly program has incorrectly combined the repeat (yellow) reads into a single repeat (yellow; annotated R1/R2), resulting in an incorrect assembly of the genomic region, and the red flanking region being erroneously discarded as an unmappable "orphan" contig. Adapted from Pop and Salzberg (2002)

To overcome the limitations of each sequencing technology, multiple technologies are often used in tandem to ensure accurate *de novo* determination of genomes, such as the use of Sanger sequencing and NGS to determine complete mitogenomes (e.g. (Elbrecht *et al.* 2013; Gaitán-Espitia *et al.* 2013; Kalchhauser *et al.* 2014).

Box 1. Commonly used sequencing terminology

Assembly: Computational reconstruction of the target DNA sequence from sequencing reads using bioinformatic algorithms. Relies on the basic assumption that two reads whose nucleotide sequence are of high similarity originated from the same locus. Two methods:

<u>Reference-based assembly</u>: use of a reference sequence from the same organism or a closely related species as a map to guide the assembly process by aligning the fragments being assembled.

<u>De novo assembly</u>: assembly without using a reference sequence, usually due to being novel sequence data.

Consensus sequence: Produced by aligning multiple overlapping reads. Accuracy of the consensus sequence depends largely upon sequence coverage.

Contig: A contiguous stretch of DNA sequence that is the result of assembling multiple overlapping sequencing reads into a single consensus sequence. A contig requires a complete tiling set of overlapping sequencing reads spanning a DNA region without gaps.

De novo sequencing: The sequencing of a novel DNA region or genome of an organism. This term is also used whenever a DNA region, genome or sequence data set is assembled by De Novo assembly (see assembly).

Read: see sequencing read.

Read length: see sequencing read length.

Read Mapping or Alignment: Assembly process of aligning (mapping) each sequence read to its corresponding loci/position (see assembly).

Reference sequence: The formally recognized, official sequence of a known genome or gene. Usually stored in a public database and may be referred to by an accession number. Often used for qualitative assessment and comparison of the experimentally determined sequence.

Re-sequencing: The sequencing of a DNA region or genome for which a reference sequence is available. This term is also used whenever a DNA region, genome or sequence data set is assembled by referenced-based assembly (see assembly).

Sequence coverage or depth: The average number of sequencing reads that cover a given nucleotide position (i.e. average number of times a base is sequenced in a given experiment). E.g. 30x coverage means that each nucleotide of a target DNA sequence has been sequenced 30 times on average (Rizzo & Buck 2012).

Sequencing read: A contiguous length of nucleotide bases (nucleotide sequence) that is generated by Sanger or NGS sequencing platforms. Usually less than 1000 bases but see sequencing read length.

Sequencing read length: The number of nucleotide bases that comprises an individual sequencing read. Read length is dependent on the sequencing technology used.

Throughput: The amount of DNA sequence (number of nucleotide bases) that can be sequenced in a single sequencing run. Can also be referred to as sequencing output.

Next-Generation Sequencing (NGS): A Blanket term used to refer collectively to the high-throughput DNA sequencing technologies available which are capable of sequencing large numbers of different DNA sequences in a single reaction (i.e. in parallel). All NGS technologies monitor the sequential addition of nucleotides to immobilized and spatially arrayed DNA templates but differ substantially in how these templates are generated and how the sequences are determined.

Unmappable Reads- read sequences that cannot be assembled and matched to the reference sequence

re significant differences in the cost of the platform, cost per run, throughput, sequencing approach and detection of nucleotide incorporation (see Table 2 for examples).

Table 2. Comparison of five commercially available sequencing platforms. Platforms used in this study are highlighted in blue.

Sequencer	Sanger 3730xl [a]	454 GS Junior	Ion Torrent [b]	SOLiD v4 [c]	HiSeqX [d]
Platform cost [e]	$95000 [c]	$108000 [f]	$80490 [f]	$495000	$1000000
Max. reads / run	96	~ 100000 [g]	~ 3 million [h]	~ 1.4 billion	~ 6 billion
Sequencing Mechanism	Dideoxy chain termination [c]	Pyro-sequencing [c]	Ion Semiconductor [f]	Ligation	Sequencing by synthesis*
Supported Read length (bases/read)	400[c] - 1000[i]	~ 550 [j]	Average ~ 200 [h]	100	300
Max. throughput (bases/ run)	~ 96, 000	~ 55 million [j]	~ 500 million [j]	~ 120 billion	~ 1800 billion
Reagent cost/ Megabase	~ $10, 500 [j]	~ $15.45 [j]	~ $1.72 [j]	~ $0.13	~ $0.007
Reagent cost/ Run	~ $1008 [j]	~ $850 [J]	~ $858 [j]	~ $15600	~ $12750 [k]
Time/ Run	2 hrs [j]	8 - 10 hrs [g, j]	5 hrs [j]	7-14 days	3 days
Accuracy	99.9 - 99.999% [c,g,i,]	99% [g]	~ 99% [g]	99.94%	99.95 - 99.98% [l]

[a] Based on a 96 capillary ABI3730xl Sanger sequencer

[b] Based on a Ion 316 v2 Chip using 400bp chemistry

[c] Data from Liu *et al.* (2012) , including all SOLiD v4 data.

[d] Data is for a single machine but are sold as a set of ten. Data based on (Illumina 2014) unless stated otherwise.

[e] All platform costs are in $USD

[f] Data from Loman *et al.* (2012)

[g] Data from Glenn (2011)

[h] Data based on (Life Technologies 2014). Read lengths of up to 400bp are supported but not typical-average is 200bp.

[i] Data from Shendure and Ji (2008)

[j] Data from the SABC, Murdoch University. Costs are in $AUD. Data is based on 'typical' performance runs.

[k] Data from Sheridan (2014).

[l] Data from Novogene (2014)

* Also referred to as cyclic reversible termination

1.8 Aims

Within the Southern Hemisphere lampreys, the entire mitochondrial genome of *G. australis* was sequenced several years ago (Milton 2003). However, sequencing of the entire mitogenome for *M. mordax* has proved problematic (Milton 2003; Riddington 2007; Haouchar 2009), whilst only 1133 bp of the *Cyt b* gene of *M. praecox* have been previously sequenced (Lang *et al.* 2009). In order to determine a robust hypothesis, Potter *et al*. (2014) recommend that multiple markers should be employed in any future study determining the relationships of lampreys.

The aims of this thesis are to:

A) Determine the complete mitogenome of *M. mordax* and *M. praecox* using a combination of Sanger and next-generation sequencing (454 and Ion Torrent) in order to provide, for the first time, the complete mitogenome of *M. praecox* and to check the sequence of *M. mordax* which had previously proved difficult to sequence.

B) Compare the similarities in the nucleotide data generated for *M. mordax* and *M. praecox* to determine how closely related the parasitic *M. mordax* is to its presumed nonparasitic derivative *M. praecox*.

C) And, with the aid of Dr Matthew Phillips, use the nucleotide data determined in this study in conjunction with that previously determined for *G. australis* and those published representatives of the Northern Hemisphere lampreys, hagfishes and 20 other deutrostomes in phylogenetic analyses to produce robust hypotheses regarding the relationships between the single Northern Hemisphere family (Petromyzontidae) and two Southern Hemisphere families (Mordaciidae and Geotriidae) of lampreys and provide an estimation of their divergence times.

Section 2: Materials and Methods

2.1 Materials

The chemicals and reagents used in this study are detailed in Appendix 1. The required buffers and solutions are detailed in Appendix 2.

2.2 Sample Collection and Identification

Lampreys were collected by Dean Gilligan from the Wallagaraugh River, south of Eden in New South Wales. A total of 11 adult lampreys were obtained euthanized (by immersion in tricaine methane sulfonate solution [MS-222]) from the University of Western Australia. To minimise DNA degradation, all lampreys were stored on ice immediately after being euthanized and, prior to total nucleic acid extraction (Section 2.3), were either kept on ice for immediate use or stored at -40°C until required. All lampreys were given a number for later identification (Table 3). Adult lampreys were identified as *M. mordax* (5) and *M. praecox* (6) by Dr Howard Gill on the basis of their dentition and presence (in *M. praecox*) / absence (in *M. mordax*) of readily identifiable gonadal material.

Transverse body tissue sections from an individual *M. mordax* specimen, referred to in this study as *M. mordax* #1D, was obtained from Dalal Haouchar (Veterinary and Life Sciences, Murdoch University) towards the end of this project. In a previous study, total nucleic acid had been extracted from this specimen to obtain the complete mitochondrial sequence of *M. mordax* (Haouchar 2009), especially the 454 sequence data (Haouchar, D., pers. comm., 2013). This lamprey had been provided by Shaun Collin from the University of Queensland (2009).

Table 3. Identification of individual lampreys used in this study and date obtained.

Date obtained	Identification	
	#	Species
25/6/13	1	*M. praecox*
	2	*M. praecox*
	3	*M. praecox*
	4	*M. mordax*
	5	*M. mordax*
26/7/13	16	*M. praecox*
	17	*M. praecox*
7/8/13	22	*M. praecox*
	23	*M. mordax*
	24	*M. mordax*
	25	*M. mordax*
16/8/13*	1D	*M. mordax*

* Date obtained from D. Haouchar

2.3 Extraction of Total Nucleic Acids

Total nucleic acid was extracted from lamprey muscle tissue or liver using the method of Milton (2003) with modification as follows:

Muscle tissue (20 mg) or whole livers (15 – 25 mg) were finely diced (approximately 0.1 cm^2 pieces) with a sterile scalpel blade and placed in a 1.5 mL microcentrifuge tube containing 200 µL of tissue digestion buffer (0.02M EDTA, 0.05M Tris, 0.12M NaCl, pH 8.0, 0.05M SDS) and 20 µL of Proteinase K (20 mg/mL). Each sample was incubated for 1 hr at 55°C with constant shaking at 300 rpm (Bioer ThermoCell MB-102) and vortexed for 15 sec after 30 min. Equal volume (220 µL) of 4 M ammonium acetate was added, the mixture vortexed for 10 sec, followed by incubation at 37°C for 15 min under constant shaking at 300rpm. The precipitated proteins were pelleted by centrifugation at 20800 x g for 10 min. Equal volume aliquots of the supernatant were removed and transferred to a new 1.5 mL microcentrifuge tube containing 2.5 x volume of ice cold 100% ethanol, then stored at -70^{o}C for one hour and pelleted by centrifugation at 20800 x g for 30 min at 4°C. The supernatant was then discarded and the pellet washed with ice cold 70% ethanol before centrifugation at 20800 x g for 5 min at 4°C. The supernatant was removed and the pellet air dried for 30 min to evaporate any excess ethanol before resuspension in 100 µL of 1 x TE buffer overnight at 4°C. Extracts were used immediately or stored at -20°C until analysis.

2.4 Quantification of Nucleic Acid

Nucleic acid was quantified by either UV spectrophotometry using a NanoDrop ND-1000 (Thermo Scientific) or by double stranded DNA-specific Broad Range Assay using a Qubit 2.0 Fluorometer (Life Technologies) following the manufacturer's instructions.

2.5 Agarose Gel Electrophoresis

Agarose gels (0.8% or 1.5% weight per volume) were prepared by dissolving DNA grade agarose powder (Progen Biosciences) in 1 x Tris-Acetate-Ethylenediaminetetraacetic acid (1 x TAE). Gels were stained with 0.8 µL or 1.8 µL of 10000 x SYBR Safe DNA Gel Stain (Invitrogen) prior to setting for 0.8% and 1.5% gels, respectively. Electrophoresis was carried out in 1 x TAE electrophoresis buffer at 90 V for 1 hr in either a Bio-Rad Mini Sub™ DNA cell or a Bio-Rad Wide Mini Sub DNA cell electrophoresis tank.

For analyses, 5 µL of the sample to be electrophoresed was loaded using 2 µL of 6 x Blue/Orange loading dye (Promega) or 1 µL of 6 x loading dye (Fermentas). DNA Molecular weight markers (DNA ladders); either 100 bp DNA ladder (Promega), GeneRuler 100 bp Plus DNA ladder (Fermentas) or 1 Kb DNA ladder (Promega), were electrophoresed on all gels except band extraction gels (Section 2.5.1). Electrophoresed samples were visualised using a Vilber Lourmat EXC-F20 Skylight transilluminator with Bio-Vision 1000 gel imaging system (Fisher Biotec) and images were captured and stored using VisionCapt v15.06 for Windows (Vilber Lourmat).

2.5.1 Agarose Gel Band Extraction and DNA Recovery

For samples requiring DNA sequencing (Section 2.7.1 and 2.7.2), the PCR sample was electrophoresed (Section 2.5) to confirm the presence of the band size of interest (target band). A portion (15 – 30 μL) of the sample was then electrophoresed on a 0.8% 1 x TAE agarose gel (Section 2.5). To minimise cross contamination, PCR products from the same individual lamprey were run on the same gel with one blank lane between each sample. Electrophoresed samples were visualised and the images captured (Section 2.5). A Dark Reader™ DR-45M transilluminator (Clare Chemical) was used to visualise target bands during agarose gel band extraction, which were excised using Axygen gel cutting tips (Fisher Biotec), trimmed to remove excess agarose, and the amplified DNA recovered by the following method adapted from http://www.protocol-online.org/biology-forums/posts/16813.html.

A 200 μL filter pipette tip (Axygen, Fisher Biotec) was cut off at the first line below the filter such that the bottom 32 mm of the tip was removed. The top of the filter tip was then placed in a 1.5 mL Eppendorf tube with the trimmed excised gel band placed inside the filter tip (on top of the filter). The tube was then centrifuged at 20800 x g at room temperature for 1 min. The filter tip was then discarded and the resulting tube, containing the DNA in 1 x TAE buffer, was either temporarily stored at 4^{0}C or at -20^{0}C until analysis.

2.6 PCR Primers

2.6.1 Primer Design

Where possible, overlapping primers were designed to be 18-30 nucleotides in length, have an annealing temperature between 50 – 65^{o}C, a GC content of 40 – 60% (where possible), and a melting temperature (Tm) difference of less than 5^{o}C between primers in a pair. Priming sites were selected to yield amplicons of 400 – 1300 bp for short range PCR and 5000 – 6000 bp in size for long range PCR. Potential primer pairs from the aligned unpublished complete mitogenomes of *G. australis* (Milton 2003; Riddington 2007) and *M. mordax* (Haouchar 2009) were identified using MitoPrimerV1 software (Yang *et al.* 2011).

MitoPrimer determined primer pairs were checked using the partial *Cyt b* sequence of *M. praecox* (Accession #GQ206186) and the unpublished *M. Mordax* mitogenome of Haouchar (2009). Geneious v7.1.5 for Windows (Biomatters) was used to identify potential replacement primers for MitoPrimer determined primers that had a low GC content (below 40% where possible) and Tm difference of greater than 5^{o}C. The nucleotide sequence of each primer was checked for specificity using Basic Local Alignment Search Tool (BLAST)(www.ncbi.nlm.nih.gov). Primers were ordered online from Integrated DNA Technologies PTY LTD (IDT) (www.idtdna.com) (Iowa, USA).

2.6.2 Resuspending and Diluting Lyophilized Primers

The lyophilized primers were resuspended in injection water (medical grade, AstraZeneca) to a stock primer concentration of 100 µM following the manufacturer's instructions. For example, primer X1F was 29.1 nmoles and suspended in 291 µl of injection water. Working stock solutions were made by diluting the concentrated primer stock 1:10 to obtain a working concentration of 10 µM. For example, 2 µl of 100 µM primer X1F stock was diluted in 18 µl of water for a 20 µl 10 µM working stock for use in subsequent PCR reactions. All primer stocks were stored at -20^{o}C until use.

2.7 Molecular Approaches Undertaken

To obtain the complete mitochondrial genome sequence of *M. mordax* and *M .praecox*, three different approaches were undertaken in this experiment:

1. Short range PCR and subsequent Sanger sequencing (Section 2.7.1)
2. Long range PCR and subsequent 454 sequencing (Section 2.7.2)
3. Ion Torrent (Life Technologies) sequencing of total nucleic acid extract (Section 2.7.3)

The reasons for this are briefly outlined. Past studies (unpublished) involving short range PCR of *M. mordax* and *Geotria australis* (Milton 2003; Riddington 2007) had failed to determine the complete mitogenome of the two lamprey species. Whilst long range PCR had been used successfully to determine the complete mitogenome of lampreys *Lethenteron camtschaticum* (Hwang *et al.* (2013a) as *Lampetra japonica*) and *Lethenteron reissneri* (Hwang *et al.* (2013b) as *Lampetra reissneri*), the difficulties (e.g. low amplification success rates and difficulties in amplification consistency) associated with the method are well known (Briscoe *et al.* 2013). Lastly, Ion Torrent sequencing was new to the Western Australian State Agricultural Biotechnology Centre (SABC) at Murdoch University. All three approaches were initially undertaken as a fail-safe approach.

2.7.1 Short Range PCR and Sanger Sequencing Strategy

Overlapping short range PCR primers (Table 4) were designed (Section 2.6.1) to yield amplicons of 400 – 1300 bp in size that, upon subsequent Sanger sequencing (Section

2.7.1.2), would result in a 3.5 Kb contiguous (contig) mitochondrial DNA sequence extending from within Cytochrome b (*Cyt b*) to within *16S* RNA (Figure 9).

Table 4. Short range PCR primer pairs used to amplify a 3.5 kb contiguous mitochondrial region extending from within *Cyt b* to within *16S* RNA.

Primer Pair	Primer	Sequence (5' - 3')	Annealing Temp (°C)	Amplicon Size (bp)
X48F/R	X48F X48R	TGCCGAGACGTGAATAGTGG GATATGAGTGGGGGTGCTTAGG	55	597
X49F/R	X49F X49R	CGTTCTACCATGAGGCCAAATATC AAGGGTCGAAATTGAGCACCAC	55	567
X50F/R	X50F X50R	CATACGCTATCCTACGATCAATCC ACCATCTAAGCATCTTCAGTGC	55	478
X51F/R	X51F X51R	AAAGAGAATTAGAATCTCTATTACTAGCC GAAATATTTAAGCGTGTCTATTCAACT	52	500
X50F/51R	X50F X51R	CATACGCTATCCTACGATCAATCC GAAATATTTAAGCGTGTCTATTCAACT	55	859
X1F/R	X1F X1R	AGAAAAAGAGCTGGCATTAGGC AGGCTCCTCTAGGTGGGTTTAG	60	448
X2F/R	X2F X2R	TTACACGAGGGGCTCAAGTT CCATGTTACGACTTTTCTCCAG	55	628
X3F/R	X3F X3R	GTCACTCTCCTTCCTCATTCC ACCAGCTATCACTAGGCTCG	60	524
X4F/R	X4F X4R	AAGCAAACCCGAATTTACC GTGTGTTGTGTAAGTGGGAGG	55	567
X4F/6R	X4F X6R	AAGCAAACCCGAATTTACC TAATCGTTGAACAAACGAACC	60	1216
X6F/R	X6F X6R	CCTGTTTACCAAAAACATCGC TAATCGTTGAACAAACGAACC	52	554
FB1F/3R	FB1F FB3R	AAACCTCGTGCCAGCCA GTAATCCCAGGGTAGCTCGTC	60	1037

Note: amplicon size is based on the unpublished mitogenome of *M. mordax* (Haouchar, 2010). Primer pair FB1F/3R was designed by Frances Brigg, SABC, Murdoch University.

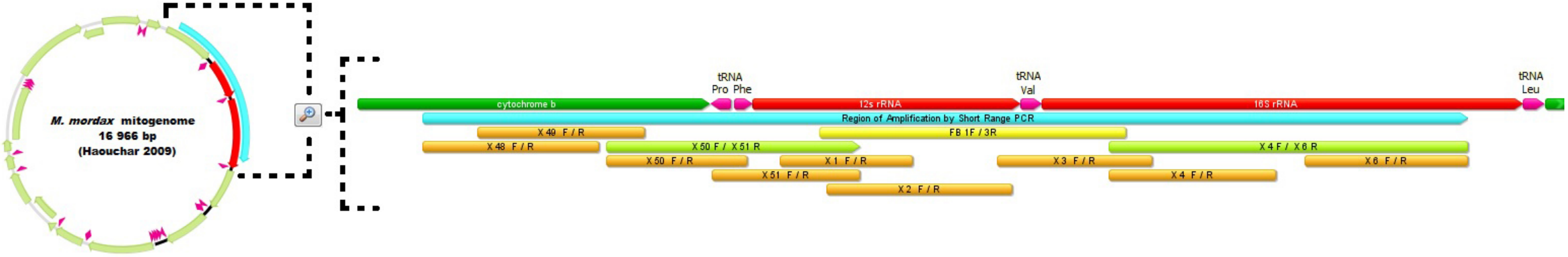


Figure 9. The overlapping short range PCR method of generating the expected 3.5 Kb mitogenome region for *M. mordax and M. praecox.* Left) Rough location in respect to the complete *M. mordax* mitogenome as determined by Haouchar (2009). Right) Exact location and degree of amplicon coverage. The short range PCR mitochondrial contig (light blue) starts from within *Cyt b* (dark green as annotated) and finishes within *16S* rRNA (red as annotated). Overlapping primer pairs (orange and yellow as annotated) were designed based on the complete *M. mordax* sequence of Haouchar (2009) (far left). Combined primer pairs (lime as annotated) and primer pair FB1F/3R (yellow) were used to provide extra coverage to ensure successful *de novo* assembly.

2.7.1.1 Short Range PCR Amplification.

Short range PCR amplifications were performed in a 50 µL reaction volume containing 10 µL of MyTaq 5 x Reaction Buffer (Bioline), 1 µL of MyTaq DNA polymerase (Bioline), 2 µL each of one forward and one reverse primer (10 µM), 3 µL of DNA (1 – 3 ng) and 32 µL of injection water. Thermal cycling conditions were: initial denaturation at 96°C for 2 min, 32 cycles of [denaturation at 96°C for 10 sec, annealing temperature of the primer pair for 10 sec, and extension at 72°C for 1.5 min], followed by a final extension at 72°C for 10 min with a 12°C hold. Products were electrophoresed and visualised on a 1.5% 1 x TAE agarose gel (Section 2.5).

2.7.1.2 Dideoxy Sanger Sequencing of Short Range PCR Products

Short range PCR products were sequenced in both directions using a protocol modified for GC rich template on an ABI3730xl capillary sequencer (Applied Biosystems). Sequencing was carried out in a 10 µL reaction volume containing 1 µL of the PCR primer (3.2 µM), 4 µL of a 1:3 ratio of dGTP BigDye Terminator v3.0 to BigDye Terminator v3.1, and either 5 µL of neat or 2 µL of diluted (1 in 5) agarose band extracted DNA and injection water to make up the volume. Thermal cycling was carried out in either a Veriti 96-well or model 2720 thermal cycler (Applied Biosystems) with the following thermal cycling conditions: initial denaturation at 96°C for 2 min, 30 cycles of denaturation at 96°C for 10 sec and combined annealing and extension at 60°C for 4 min, followed by an 8°C hold. To remove salts and unincorporated dyes, post-reaction purification consisted of the ethanol/EDTA/sodium acetate precipitation protocol for 10 µL reactions in 96-well plates (Applied Biosystems, 2002). Sequences were analysed using Geneious v7.1.5 for Windows (Biomatters).

2.7.2 Long Range PCR and 454 Sequencing Strategy

Long range PCR primers (Table 5) were designed (Section 2.6.1) to yield amplicons of 5000 – 6000 bp in size that, when sequenced (Section 2.7.2.2), would yield a contiguous fragment of the complete mitogenome (Figure 10).

2.7.2.1 Long Range PCR Amplification

Long Range PCR amplification was carried out using three oligonucleotide primer pairs Long 1F/R, Long 2F/R, and Long 3F/R (Table 5) for subsequent NGS on a 454 platform (Section 2.7.2.2).

Table 5. Long Range PCR Primer pairs used to amplify 5 – 6 Kb contiguous mitochondrial fragments extending the entire mitogenome.

Primer Pair	Primer	Sequence (5' – 3')	Annealing Temp (°C)	Amplicon Size* (bp)
Long 1F/R	Long 1F	AAGCAATGGACAAAAGTCCGAC	59	5167
	Long 1R	AGTAGAGAAAAATCAGCGAGTGAGG		
Long 2F/R	Long 2F	GAGAGGAATTAAACCCCTGTAAG	57	5724
	Long 2R	ATATTTTTTGTTAGGGGGATAAATAG		
Long 3F/R	Long 3F	TCACGTAGAAGCTCCAATTGCAG	55	5567
	Long 3R	GCTTAGGGGGTTAGCGTAAATAAAA		
Long 4F/R	Long 4F	ACCAAACAGGCTCAAATAATCCGTTAG	N/D	5917
	Long 4R	GTCTGTTAGGGCTTGTGTAGATCAAAGC		
Long 5F/r	Long 5F	CAACCCCAAACCTCCCGTTA	N/D	5084
	Long 5R	GCCTTCGTTTCTCTGGTCCT		

*Amplicon size is based on the unpublished mitogenome of *M. mordax* (Haouchar 2010).

N/D indicates that an extension temperature at which the primer pair worked could not be determined.

These were performed in a 50 µL reaction volume containing 10 µL of Velocity 5x HiFi Reaction Buffer (Bioline), 1 µL of Velocity DNA polymerase (Bioline), 1.5 µL of Velocity DMSO (Bioline), 1.25 µL of either Roche Expand Long Range dNTP mix (Applied Sciences) or Promega dNTP mix (10 mM), 2 µL each of one forward and one reverse primer (10 μM), 3 μL of DNA (100 ng) and 31.25 μL of injection water. Thermal cycling conditions were: initial denaturation at 98°C for 2 min, 35 cycles of [denaturation at 98°C for 30 sec, annealing temperature of the primer pair for 30 sec, and extension at 72°C for 6 min] followed by a final extension at 72°C for 10 min with a 12°C hold. Products were electrophoresed and visualised on a 1.5% 1 x TAE agarose gel (Section 2.5). Each PCR product was then extracted and recovered (Section 2.5.1) prior to sequencing (Section 2.7.2.2).

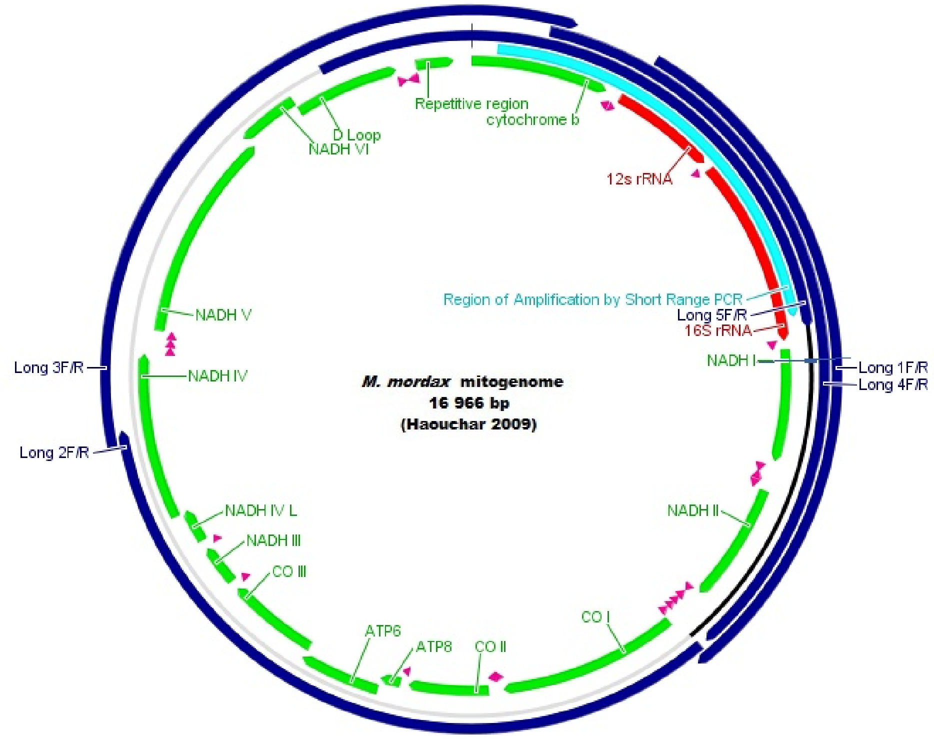


Figure 10. The overlapping long range PCR method of generating the approximately 17kb mitogenome of *M. mordax and M. praecox.* The 5 long range PCR primer pairs (dark blue as annotated) would generate amplicons 5 – 6 Kb in size that would encompass the complete mitogenome. Also shown for comparison is the region determined by short range PCR (light blue as annotated). Note: Light-strand (5'– 3') is shown for simplicity. Arrowheads indicate the direction of transcription for coding genes (green arrows as annotated), rRNA (red arrows as annotated) and tRNAs (pink arrows).

2.7.2.2 454 Sequencing of Long Range PCR Products

Long range PCR fragments were submitted to Frances Brigg at the SABC at Murdoch University for library preparation and pyrosequencing on a 454 GS Junior (Roche) next-generation sequencing platform. For method reproducibility, the key steps are briefly outlined.

Long Range PCR products (500 ng) were subjected to DNA fragmentation by ultrasonication using a Covaris M220 Focused-ultrasonicator (Trend Bio) for 260 seconds duration in 50uL screw cap vials at 50 W Peak Incident Power, 200 Cycles Per Burst with a Duty Facto of 20 to yield fragments with an average size of 600 – 800 bp. Fragment end-repair, AMPure XP bead purification (Amersham International), adaptor ligation, quality assessment using the Agilent 2100 Bioanalyzer DNA High Sensitivity kit, and library quantification by quantitative PCR (qPCR) were carried out according to manufacturer’s instructions. During the adapter ligation stage, each long range PCR product from each individual was given a unique barcode using the Rapid Library MID adapter kit (Roche) prior to the libraries being pooled in equimolar ratios. Emulsions PCR, enrichment and 454 GS Junior pyrosequencing were carried out as per manufacturer's protocols using 1.2 million copies per bead and Lib L chemistry.

2.7.3 Ion Torrent Sequencing Strategy.

The mitogenome of *M. praecox* (individual #3) and *M. mordax* (individual #25) were sequenced by Frances Brigg at the SABC at Murdoch University from total nucleic acid extracted in this study (Section 2.3). Ion Torrent sequencing comprised of 4 steps: library construction (Section 2.7.3.1); template preparation (Section 2.7.3.2); sequencing (Section 2.7.3.3); and data analysis (Section 2.7.3.4). Brief details are provided for method reproducibility.

2.7.3.1 Library Construction for Ion Torrent

DNA in the total nucleic acid extracts (Section 2.3) were quantified using Qubit 2.0 Fluorometer (Section 2.4). DNA (500 ng) was fragmented using a Covaris M220 Focused-ultrasonicator in conjunction with the Ion Xpress Plus Fragment Library Kit (Life Technologies) following the manufacturer's recommendations. Size and quality of the fragmented DNA were determined with the Agilent High Sensitivity DNA Kit on the Agilent 2100 Bioanalyzer following the manufacturer's recommendations. Fragments were end-repaired before being purified and size selected using Agencourt AMPure XP beads (Beckman Coulter Genomics). Ion Torrent adapters were ligated onto the fragment ends and nick-repaired to complete the linkage between adaptors and DNA. The adapter-ligated fragments were size selected using E-Gel SizeSelect 2% Agarose Gel and 50 bp ladder on a E-Gel EX (Invitrogen) to give an average 480 bp (400 bp chemistry, *M. mordax*) and 330 bp (200 bp chemistry, *M. praecox*) including adapters and assessed using a High Sensitivity DNA Chip.

2.7.3.2 Template Preparation for Ion Torrent

Libraries were quantified by qPCR and 250 million copies (200 bp chemistry) and 125 million copies (400 bp chemistry) of the library were used in emulsion PCR using the Ion OneTouch 2 system (Life Technologies) following the manufacturer's recommendations. For clonal amplification, DNA was localized to Ion Sphere particles (Life Technologies), which were automatically enriched with the Ion OneTouch ES system. Quality was assessed using the Qubit 2.0 Fluorometer (Section 2.4) following the manufacturer's recommendations.

2.7.3.3 Ion Torrent Sequencing

Sequencing was performed using the Ion Torrent Personal Genome Machine (PGM, Life Technologies) using Ion Express Template 200 bp or 400 bp chemistry (Life Technologies) on a 316 v2 chip following the manufacturer’s instructions. Three 316 v2 chips were run in total, with two chips used for the 200 bp chemistry and one used for the 400 bp chemistry.

2.7.3.4 Ion Torrent Data Analysis

PGM sequences were analysed with the Ion Torrent Software Suite v.3.2. Alignment could only be visualized with separate tools for alignment and assembly viewing, such as CLC Genomics Workbench v7.3 (CLC Bio) for Linux and Tablet v1.1.4 (Milne *et al.* 2010) for Windows, which accepts BAM (binary alignment map), BAI (binary alignment index files) and other file formats. *De novo* assemblies were carried out using CLC Genomics Workbench.

2.8 Mitogenome Annotation

Mitogenome sequences were submitted to the metazoan mitochondrial genome annotation web server MITOS (Bernt *et al.* 2013) for preliminary annotation of functional genes. The boundaries of the annotations were checked manually by comparison to published lamprey mitochondrial genomes (Table 6). The 22 transfer RNAs (tRNAs) were confirmed using both tRNAscan-SE v1.21 (Schattner *et al.* 2005) and Rfam v11.0 (Burge *et al.* 2013) web servers.

2.9 Phylogenetic Analyses

Phylogenetic analyses were carried out using the complete mitogenomes of five lamprey species together with the two newly sequenced mitogenomes (this study) of

M. mordax and *M. praecox* (Table 6). Two hagfishes, *Myxine glutinosa* and *Eptatretus burgeri*, served as the outgroup and are the only two available complete hagfish mitogenomes on GenBank.

Table 6. List of the 9 taxa used in the phylogenetic analyses.

Species	Common name	GenBank ID	Mitogenome	Reference
			size (bp)	
Mordacia mordax	Short-headed lamprey	Unpublished	17092	This study
Mordacia praecox	Australian brook lamprey	Unpublished	17251	This study
Geotria australis	Pouched lamprey	Unpublished	17059	Unpublished*
Lampetra fluviatilis	European river lamprey	Y18683	16159	Delarbre *et al*. 2009
Lethenteron camtschaticum	River lamprey	KC353468	16277	Hwang *et al*. 2013a
Lethenteron reissneri	Sand lamprey	KC353466	16207	Hwang *et al.* 2013b
Petromyzon marinus	Sea lamprey	PMU11880	16201	Lee *et al*. 1995
Myxine glutinosa	Atlantic hagfish	NC002639	18909	Delarbre *et al*. 2001
Eptatretus burger	Inshore hagfish	NC002807	17168	Delarbre *et al*. 2002

* Honours thesis of Milton (2003) and Riddington (2007)

2.9.1 Sequence Alignments

Preliminary sequence alignments identified premature stop codons in the *COIII* gene of *G. australis*. This error, which caused by two single base insertions (Appendix 3), was fixed by alignment to the original Sanger sequencing reads from Milton (2003).

The nucleotide and amino acid sequences of the 13 protein coding genes (PCGs) from each taxa were used to reconstruct the phylogenetic relationships in lampreys. To ensure that any insertion of gaps during the alignment process did not change the reading frame, all 13 PCGs (*ATP6, ATP8, Cyt b, CO1, CO2, CO3, ND1, ND2, ND3, ND4, ND4L, ND5, ND6)* were individually aligned (i.e. 9 x *ATP6* gene sequences) at the amino acid level using Geneious. This was done using the translational alignment algorithm with the genetic code set to vertebrate mitochondrial and protein alignment options set to ClustalW with default settings (Cost matrix: BLOSUM, gap open cost 10, gap extend cost 0.1). Ambiguously aligned regions were identified by eye and excluded from the alignment. The stop codon of each protein coding gene was also excluded. The final size of each alignment for each dataset are summarised in Table 7. Alignment at the nucleotide level was inferred from the amino acid alignments by back-translation into their corresponding nucleotide sequences. Nucleotide alignments of the PCGs were analysed as three datasets: 1) individually (all codon positions); 2) individually with 3rd codon removed; and 3) as a concatenated alignment (without 3rd codon). Amino acid alignments of the PCGs were analysed as two datasets: 1) individually; and 2) as a concatenated alignment. Geneious was used to create the concatenated alignment, which comprised of the 13 individually aligned PCG's as a single alignment.

Table 7. Final size of each alignment used in the phylogenetic analyses for each dataset.

			Final Size After Elimination of Gaps		
Alignment	**Nucleotide Size**	**No. of Gaps**	**Nucleotides** (all positions)	**Nucleotides** (excluding 3rd codon)	**Amino Acids**
ATP6	681	3	678	452	226
ATP8	162	0	162	108	54
CO1	1551	3	1548	1032	516
CO2	675	3	672	448	224
CO3	783	0	783	522	261
Cyt b	1134	9	1125	750	375
ND1	957	9	948	632	316
ND2	1008	12	996	664	332
ND3	348	3	345	230	115
ND4	1389	36	1353	902	451
ND4L	288	0	288	192	96
ND5	1809	27	1782	1188	594
ND6	483	24	459	306	153
CONCAT	11268	129	11139	7426	3713

Note: the size of each alignment is after the removal of unambiguously aligned regions and the stop codon for each gene. Abbreviations: Concat = concatenated dataset containing all 13 PCGs; ATP6 and 8 = ATP synthase subunits 6 and 8; COI, II, and III = Cytochrome C oxidase subunits I, II, and III; Cyt b = Cytochrome b; ND1–6 = NADH dehydrogenase subunits 1–6.

2.9.2 Phylogenetic Inference Method

Following sequence alignment (Section 2.9.1), each alignment file was converted (using Geneious) to a Nexus file format for phylogenetic analyses. Phylogenetic trees were estimated for each data set using Neighbor-Joining (Saitou & Nei 1987) and Bayesian inference (BI) methods. The treatment of alignment gaps prior to phylogenetic analysis is a contentious issue for which there is no general consensus. Whilst MEGA6 allows sequence gaps to contribute phylogenetic information, MrBayes v3.2.1 does not (it eliminates all positions containing gaps) unless complex *ad hoc* modelling is undertaken and binary character coding is carried out (e.g. Dwivedi and Gadagkar (2009). For better comparison of the phylogenetic trees generated by NJ and BI methods, all positions containing gaps were eliminated (complete deletion) in both NJ and BI analyses. Where possible, third codon positions were excluded from analyses to reduce problems caused by substitution saturation which can cause biased results in deep phylogenies (Curole & Kocher 1999).

2.9.2.1 Neighbor-Joining Analyses

Phylogenetic relationships inferred using the Neighbor-Joining (NJ) method were performed using MEGA6 (Tamura *et al.* 2013) for Windows. Bootstrap support was calculated from 5000 bootstrap replicates. Each analysis involved 9 amino acid sequences (one for each taxa) that had previously been aligned in this study using Geneious (Section 2.9.1).

For nucleotide datasets, evolutionary distances were computed using the Jukes-Cantor nucleotide substitution model (Jukes & Cantor 1969) as suggested by Nei and Kumar (guideline point 2, page 112, 2000). The rate variation among sites was modelled with a gamma distribution (default shape parameter mean = 1) to accommodate nucleotide substitution rate variation. A mean of 1 was used as it is a standard approach to assume that rates vary accordingly to a gamma distribution with mean = 1 (Ronquist *et al.* 2009). The codon positions included in the NJ analyses were either 1st +2nd+3rd (all positions analyses) or 1st+2nd (3rd codon removed analyses).

For amino acid datasets, the amino acid sequence was first translated in MEGA6 using the vertebrate mitochondrial genetic code table and the evolutionary distances then computed using the Poisson correction method (Zuckerkandl & Pauling 1965). The rate variation among sites was modelled with a gamma distribution (shape parameter =1).

2.9.2.2 Bayesian Inference Analyses

Bayesian analyses were conducted using MrBayes v3.2.1 (Ronquist *et al.* 2012) for Windows. Prior to the analyses the appropriate nucleotide and amino acid substitution model for each dataset was tested using MEGA6 (Table 8). The substitution model with the lowest Bayesian Information Criterion (Schwarz 1978) score was used in phylogenetic analyses as it is considered to describe the substitution model the best (Tamura *et al.* 2013). The molecular clock hypothesis was also tested for each data set using MEGA6 (Table 8). Briefly, the program does this by comparing the maximum likelihood value for the given topology of an automatically generated NJ tree with and without the molecular clock constraints under the previously determined appropriate substitution model (Table 8). The statistical significance is tested by comparing twice the difference in log-likelihood values to a chi-squared 5% significance threshold value with s-2 degrees of freedom, where s is the number of sequences in the alignment (Tamura *et al.* 2013).

Table 8. Substitution models used in the Bayesian Inference phylogenetic analyses.

Gene	Nucleotide Model	Nucleotide Model	Amino Acid Model
	(1st+2nd+3rd codon)	(3rd codon removed)	
ATP6	HKY+G+Clock	HKY+G I	MtREV+G
ATP8	HKY+G+Clock	HKY+G	MtREV+Clock
CO1	GTR+G+I+Clock	HKY+G+Clock	MtREV+I+Clock
CO2	GTR+G	HKY+G	MtREV+G
CO3	GTR+G+I	HKY+G	MtREV+G
Cyt b	GTR+G	HKY+G	MtREV+G+Clock
ND1	GTR+G+Clock	HKY+G	MtREV+G+Clock
ND2	HKY+G+I+Clock	HKY+I+Clock	MtREV+G+Clock
ND3	HKY+G	HKY+G+Clock	MtREV+G
ND4	HKY+G+I	HKY+G+Clock	MtREV+G+Clock
ND4L	HKY+G	HKY+G	MtREV+Clock
ND5	GTR+G+I	HKY+G	MtREV+G+Clock
ND6	HKY+G	HKY+G	MtREV+G
Concat	GTR+G+I	GTR+G+Clock	MtREV+G

Abbreviations: All genes abbreviated as stated in footnote of Table 7; Concat = concatenated dataset containing all 13 PCGs; GTR = General Time Reversible model; HKY = Hasegawa-Kishino-Yano model; MtREV = General Reversible Mitochondrial model; +G = plus gamma; +I = plus invariable sites; +Clock = plus 'strict' molecular clock. For example GTR+G+I is a GTR model with the addition of invariant sites and a gamma distribution of rates across sites.

Each Bayesian analysis consisted of two simultaneously independent Markov Chain Monte Carlo (MCMC) runs of 1 million generations, started from a random tree, with four independent (3 hot and 1 cold) chains (default temperature of 0.1) and tree sampling once every 500 generations. The first 25% of trees were discarded as burn-in and the remaining trees were used to calculate Bayesian posterior probabilities. All remaining parameters were left as default values. The substitution models used for each dataset are summarised in Table 8, but the substitution model priors, such as gamma shape (*shapepr*) and transition/transverion ratio (*tratiopr*), were flat (i.e. left as default values) to enable MrBayes to estimate the parameters from the data during the run, as recommended (Ronquist *et al* 2009; MrBayes v3.2.1 user manual). Markov chain stationarity was considered to be reached when the average standard deviation of split frequencies fell below 0.01 and potential scale reduction factor values approached 1.0 (MrBayes v3.2.1 user manual).

Section 3: Results

3.1 Extraction of Total Nucleic Acid

Total nucleic acid (DNA + RNA) was extracted from both *M. mordax* and *M. praecox* by the modified method of Milton (2003) as described in Section 2.3. When electrophoresed, extracts showed a band greater than 10 kb in size and faint smearing from approximately 2000 bp to 250 bp in which two weak bands approximately 1750 bp and 1100 bp in size were observed (Figure 11; lane 2). Following RNAse digestion, the two weak bands and associated smearing were no longer present and only the single DNA band greater than 10 kb was observed (Figure 11; lane 3). This indicated that the two smaller bands and associated smearing were RNA, most likely ribosomal, and that extracts contained DNA of sufficient size for subsequent analysis.

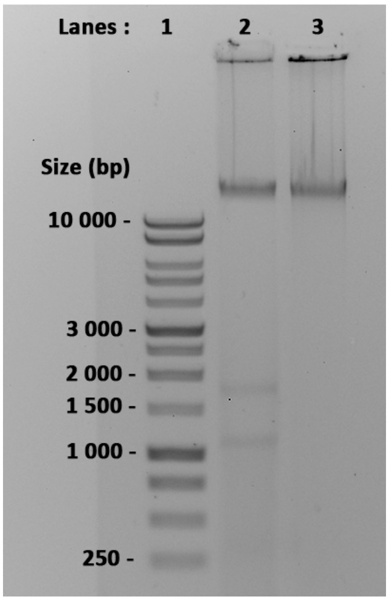


Figure 11. Comparison of total nucleic acids before and after RNAse digestion. **Lane 1)** 100bp plus ladder, **Lane 2)** before RNAse digestion, and **Lane 3)** after RNAse digestion. Agarose gel (0.8%) electrophoresis for 1 hr and stained with SYBR safe. Comparison shows the two small bands (approximately 1750 bp and 1100 bp) and associated smearing (approximately between 2000 – 250 bp) in Lane 2 is RNA and that the large band (> 10 kb), present in Lanes 2 and 3, is intact DNA.

Extracts were quantified by Nanodrop (Section 2.4) prior to RNAse digestion and contained between 115 – 475 ng of nucleic acid per micro litre of extract. Total nucleic acid extracted from the liver approximately 3 – 5 hrs after euthanasia tended to yield more nucleic acid than those that were frozen after euthanasia. However, this depended on factors such as the size of the liver and the ease of grinding liver tissue in contrast to muscle tissue during the extraction method.

3.2 *Cyt b – 16S* region of the Mitochondrion

During the initial stages of this study, the *Cyt b – 16S* mitochondrial region of one *M. mordax* individual (#4) and one *M. praecox* individual (#3) were amplified using 12 primer pairs (Section 2.7.1; Table 4) and sequenced by Sanger sequencing (Figure 12). The *Cyt b – 16S* mitochondrial region was selected as *Cyt b* is often used in phylogenetic analyses and the region would serve as a starting point for sequencing the entire mitogenome of these individuals using the primers listed in Figure 12.

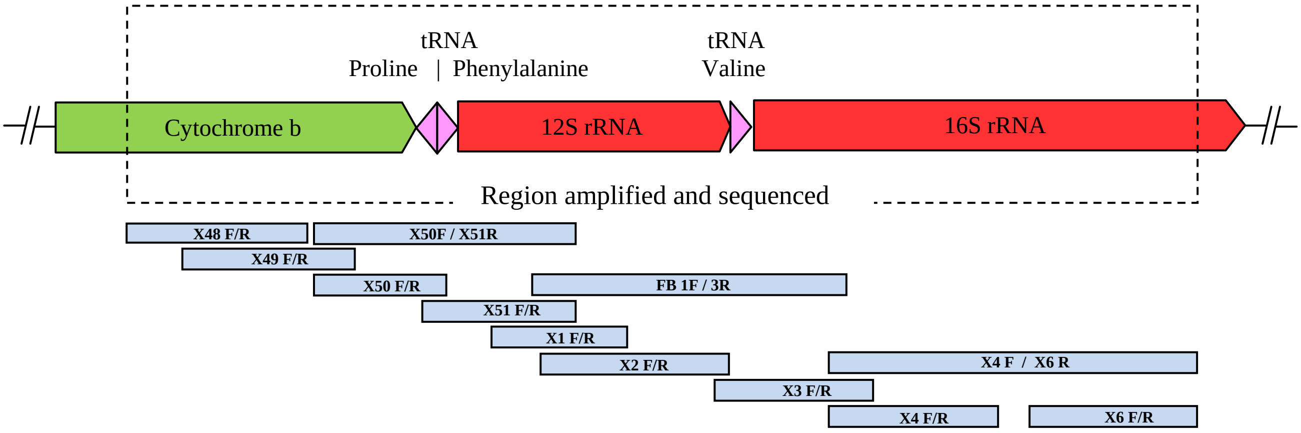


Figure 12. The *Cyt b – 16S* mitochondrial region and the short range PCR– Sanger sequencing strategy used in this study. The sequenced region (dotted box) was determined by Sanger sequencing of 12 amplicons (blue boxes). Note: the gene organisation of *M. mordax* as determined by Haouchar (2009) is shown. Arrowheads indicate the direction of transcription for each gene. The amplicon name consists of the primer pairs used to generate it.

3.2.1 Electrophoresis of Amplified Products

Except for primer pair X50F/R and primer pair X51F/R, all primer pairs generated amplicons of the expected size based on the complete *M. mordax* mitogenome sequence of Haouchar (2009). However, X50F/R amplicons were either 800 or 950 bp depending on the individual (Table 9) and X51F/R amplicons were 500 and 650 bp in size for all individuals (Table 9).

Although X50F/R amplicons were expected to be 478 bp in size, no amplicons of this size were produced in this study. Instead, electrophoresis of X50F/R amplicons resulted in a single band approximately 800 bp in size for *M. mordax* #4 (Figure 13, lane 3) and 950 bp in size for *M. praecox* #3 (Figure 13, lane 2).

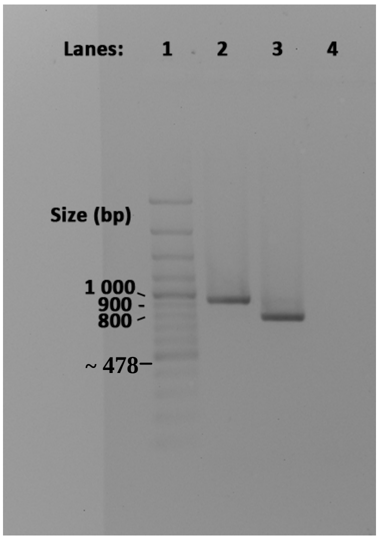


Figure 13. Example of the electrophoresis profile for X50F/R amplicons of both *M. mordax* and *M. praecox* individuals in this study. Depending on the individual, the band is either approximately 950 or 800 bp in size (Table 9). **Lane 1**) 100 plus ladder, **Lane 2)** *M. praecox* #3, **Lane 3)** *M. mordax* #4, and **Lane 4)** negative template control. Agarose gel (1.5%) after electrophoresis for 1hr at 90v and stained with SYBRsafe. Note: the expected band size of 478 bp based on the complete *M. mordax* mitogenome of Haouchar (2009) was never seen in this study.

In order to determine why X50F/R amplicons of 478 bp in size were not observed in *M. mordax* #4 (800 bp) and *M. praecox* #3 (950 bp), as well as determine whether the observed 150 bp difference between the two individuals was species specific, X50F/R amplification was carried out on all available lampreys (Table 9). Two *M. praecox* and one *M. mordax* individual were repeatedly found to have the 950 bp band, whilst the other nine lamprey (four *M. praecox* and five *M. mordax*) individuals were repeatedly found to have the 800 bp band (Table 9). This finding suggested that the observed difference was not species specific, instead, it suggested that there was individual variation in both species and that it was confined to *Cyt b* – tRNA proline region amplified by primer pair X50F/R (Figure 12).

Table 9. Size comparison of X50F/R bands and X51F/R bands for all lampreys in this study.

	PCR Amplicon size (base pairs)	
Individual	X50F/R	X51F/R
M. praecox #1	800	500 and 650
M. praecox #2	800	500 and 650
M. praecox #3	950	500 and 650
M. praecox #16	800	500 and 650
M. praecox # 17	800	500 and 650
M. praecox #22	950	500 and 650
M. mordax #4	800	500 and 650
M. mordax #5	800	500 and 650
M. mordax #23	950	500 and 650
M. mordax #24	800	500 and 650
M. mordax #25	800	500 and 650
M. mordax #1D	800	500 and 650

Individuals with 950 bp bands for X50F/R are highlighted in blue.

PCR using primer pair X51F/R was expected to yield a single amplicon of 500 bp in size, however, electrophoresis resulted in two bands: the expected 500 bp band and an additional band approximately 650 bp in size (Figure 14). PCR optimisation was carried out by increasing the annealing temperature and decreasing the primer concentration. All optimised X51F/R amplifications resulted in amplicons of 500 bp and 650 bp in size. The *Cyt b – 12S* region of the *M. mordax* mitogenome sequence (Haouchar 2009) was manually inspected using Geneious software for any additional regions complimentary to the X50 and X51 primers. No alternative priming sites could be identified that would produce amplicons of 800 / 950 bp in size for primer pair X50F/R and 650 bp in size for primer pair X51F/R.

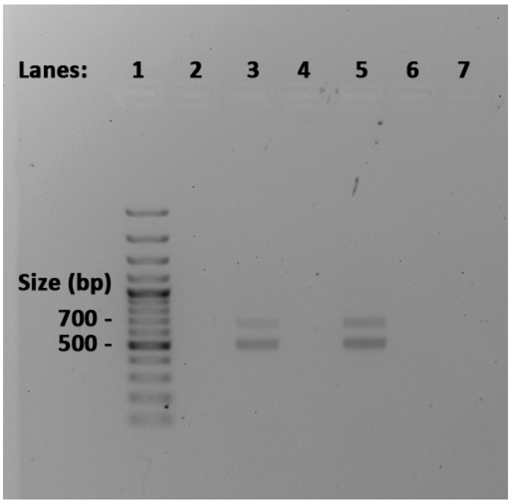


Figure 14. Example of the electrophoresis profile for X51F/R amplicons of both *M. mordax* and *M. praecox* individuals in this study. Both had two bands approximately 500 bp and 650 bp in size. The presence of the secondary band (650 bp) was unexpected based on the complete *M. mordax* mitogenome of Haouchar (2009). **Lane 1)** 100 plus ladder, **Lanes 2, 4 and 6)** empty wells, Lane 3) *M. praecox* individual #3, **Lane 5)** *M. mordax* individual #4, **Lane 7)** negative template control. Agarose gel (0.8%) after electrophoresis for 1hr at 90v and stained with SYBRsafe.

To resolve the difference between expected and observed results for primer pair X50F/R and primer pair X51F/R, muscle tissue from the same lamprey individual used by Haouchar (2009) (referred to as *M. mordax* #1D in this study) was obtained from D. Haouchar (School of Veterinary and Life Sciences, Murdoch University, 2013). Total nucleic acid was freshly extracted and amplification was carried out using primer pair X50F/R and primer pair X51F/R. Because *M. mordax* #1D was used to generate the complete mitogenome sequence of *M. mordax* (Haouchar 2009), PCR of this individual with these two primer pairs was expected to yield single bands of 478 bp for X50F/R and 500 bp for X51F/R. However, electrophoresis of X50F/R amplicons resulted in a single band of 800 bp in size (Table 9) and X51F/R amplicons resulted in two bands of 500 bp and 650 bp in size (Table 9). This finding was consistent with that observed for the other 11 lampreys reported in Table 9.

3.2.2 Sequence Comparison of *M. mordax* #1D and *M. mordax* (Haouchar 2009)

Since amplification of *M. mordax* #1D using primer pair X50F/R and primer pair X51F/R failed to resolve the difference between expected and observed results, the *Cyt b – 16S* mitochondrial region of *M. mordax* #1D individual was re-sequenced in this study by Sanger sequencing.

Based on the complete *M. mordax* mitogenome sequence of Haouchar (2009), the total length of the *Cyt b – 16S* mitochondrial region amplified by primers X48F – X6R was expected to be 3541 bp. However, the *Cyt b – 16S* mitochondrial region of *M. mordax* #1D re-sequenced in this study was 3818 bp, 277 bp larger than reported (Haouchar 2009). Furthermore, the region was only 91.2% identical (3503 identical sites) to that reported by Haouchar (2009). Pairwise nucleotide alignment identified the 277 bp difference in expected and observed length of the *Cyt b – 16S* mitochondrial region to be due to three things: *M. mordax* #1D contained an additional 48 bp within *Cyt b* (Figure 15); *M. mordax* #1D contained an additional 253 bp located between *Cyt b* and tRNA proline (Figure 15) due to a repetitive third non coding region (NCIII); and the *Cyt b – 16S* mitochondrial region of *M. mordax* (Haouchar 2009) contained an additional 24 bp due to 12 insertions within the *12S – 16S* region relative to *M. mordax* #1D (not shown).

```
                   // → Cytochrome b
Haouchar 2009   // CATACGCTATCCTACGATCAATCCCAAATAAACTTGGAGGAGTTATAGCTCTAGCCATATCTATTCTTATTCTCTTTTTTATCCCATTTCTACACACTTCACACCAACGTGGTGCTCAATTTCGACCCTT  130
M. mordax #1D   // ..................................................................................................................................  130
                       Primer X50F →

Haouchar 2009   TACCCAGTTAATTTTTTGAACTATAATTGCCAATTTGACTATACTAACATGATTAGGAGGTGAACCGGCAGAATACCCATTCATCTTAATAACACAAATCGCCTCCACAGTATACTTTGCAATTTTTATT  260
M. mordax #1D   ..................................................................................................................................  260

                                                                                                       end of Cytochrome b →||- Noncoding
Haouchar 2009   GTTATTTTTCCATTACTAGG------------------------------------------------AGAAAAAGAAAAACAGAAAATAAAAATATCATCTAAGATCATTAATTAAATTATTACCATAA  342
M. mordax #1D   ....................TCTCCTAGAAAACCCTATTCTAGCTCACAAACCCGAGGAAGAGGAAGA..............................................................  390
                                                                                                                                   |  Repeat 1

                  region III
Haouchar 2009   GTTAAAG-----------------------------------------------------------------------------------------------------------------------------  349
M. mordax #1D   .......TAGCTTAAGTTTAAAGCAAAGCTCGAAGATGCTAGTATGATGGTCCAAACAAATGACCGCCCTTTAACATCAACCCCCTCAATCAAACATCAAAGAAAGAGAATTAGAATCTCTATTACTAGG  520
                (partial)                 ← alt. X50R bind site                    end of Repeat 1 (partial)  || Repeat 2  alt. X51F bind site →

                                                                                                                        end of Noncoding region III
Haouchar 2009   ----------------------------------------------------------------------------------------------------------------------------------  349
M. mordax #1D   CCCCCCCCAACAACAGTATTTTTTAATTATTACCATAAGTTAAAGTAGCTTAAGTTTAAAGCAAAGCTCGAAGATGCTAGTATGATGGTCCAAACAAATGACCGCCCTTTAACATCAACCCCCTCAATCA  650
                                                                            ← alt. X50R bind site                                     Repeat 2

                   ||←      tRNA Proline                                    ←|      |→       tRNA Phenylalanine
Haouchar 2009   AACATCAAAGAAAGAGAATTAGAATCTCTATTACTAGCCCCCAAAGCTAGTATTTTTAAATTAAATTATTCTTTGAATAAGTTAAAATAGCTTAAATTTAAAGCAAAGCACTGAAGATGCTTAGATGGTC  479
M. mordax #1D   ..................................................................................................................................  780
                end|         Primer X51F →                                                                                         ← Primer X50R

                             →||→   12S rRNA
Haouchar 2009   TAATGGGCCCTTTAACATAAAGGATTAGTTCTAGCCTTAATATCATCTATTTACAAAATTACACATGCAAGTATCCGCACCCCCGTGAGGACCCCCTTTAACTATCGCAATAGAAAAAGAGCTGGCATTA  609
M. mordax #1D   ..................................................................................................................................  910

Haouchar 2009   GGCTCACATTACTAGCCCACAACGCCTAGCCACCCACACCCTCAAGGGTATCCAGCAGTGATAAACTTTAAGCAATGGACAAAAGTCCGACTAAGTTATGTATCTCAGAGCCGGTAAACCTCGTGCCAGC  739
M. mordax #1D   ..................................................................................................................................  1040

                                                                                                                         //
Haouchar 2009   CACCGCGGTTACACGAGGGGCTCAAGTTGATATTTACGGCGCAAAGCGTGATTAAAAATATGTTCCATCACTATAGAAGCCATTATGCCTACTAGTTGAATAGACACGCTTAAATATTTC //   859
M. mordax #1D   ........................................................................................................................ //  1160
                                                                                                    ← Primer X51R
```

Figure 15 (page 45). Nucleotide alignment of the *Cyt b – 12S* mitochondrial region of *M. mordax* (top-Haouchar 2009) determined by Haouchar (2009) to that determined in *M. mordax* #1D in this study. The *Cyt b – 12S* mitochondrial region of *M. mordax* #1D was amplified using primer pair X50F/R and primer pair X51F/R. Gene regions are indicated by the coloured bar above the top sequence and are as annotated. The location and direction of the individual primers for primer pair X50F/R (blue text as annotated) and primer pair X51F/R (green text as annotated) are shown. Alternative (imperfect) primer binding sites for primer X51F and primer X50R (red text as annotated) are also shown. Highlighted in yellow is the additional 48 bp in *Cyt b* found in all six individuals sequenced in this study. The two repeats comprising noncoding region III of *M. mordax* #1D are underlined and annotated in bold. Note: light-strand (5' – 3') is shown for simplicity. Arrows indicate direction of transcription for each gene. Dots indicate identical nucleotides whilst dashes indicate deletions. // indicates continuation of the sequence/region.

This finding suggested two possibilities: *M. mordax* #1D was not the individual sequenced by Haouchar (2009) or the mitogenome sequence of *M. mordax* (Haouchar 2009) contained errors. Given that the expected band size of 478 bp for X50F/R amplicons was never observed in this study and that none of the six lampreys (one *M. praecox* and five *M. mordax*) sequenced in this study (Section 3.2.3) were greater than 91.2% identical to that reported by Haouchar (2009), it is most likely that the observed differences were due to errors in the mitogenome of *M. mordax* (Haouchar 2009). This view is supported by the fact that most nucleotide differences were in regions that had been edited. For example, the 12 insertions within the *12S – 16S* region of the *M. mordax* mitogenome (Haouchar 2009), none of which were present in the six lamprey individuals sequenced in this study, had all been manually edited (inserted) by Haouchar. Additionally, examination of the Haouchar (2009) digital appendix resulted in further findings that strongly suggest that the *M. mordax* mitogenome (Haouchar 2009) contains errors. Firstly, the *M. mordax* mitogenome (Haouchar 2009) is 'missing'

48 bp in *Cyt b* (as seen in Figure 15) despite being present in four *M. mordax* individuals sequenced in the same study (Appendix 4). Secondly, assembly of the 454 sequencing reads from the Haouchar (2009) study to M. mordax #1D identified seven reads that spanned 143 bp of NCIII, four of which also contained the 'missing' 48 bp of *Cyt b* (Appendix 5).

3.2.3 Comparison of *Cyt b – 16S* Mitochondrial Region of Individuals Sequenced in this Study

To characterise the *Cyt b – 16S* mitochondrial region, three additional *M. mordax* individuals (#23, #24 and #25) were sequenced by Sanger sequencing. In total, the *Cyt b – 16S* mitochondrial region of six individuals (*M. praecox* #3 and *M. mordax* individuals #1D, #4, #23, #24 and #25) were determined.

Depending on the sequenced individual, the length of the *Cyt b – 16S* mitochondrial region varied between 3815 – 3978 bp in length (Table 10). Individual length variation of the *Cyt b – 16S* mitochondrial region was due to NCIII. Noncoding region III is characterised in Section 3.2.4. Excluding NCIII, the *Cyt b – 16S* mitochondrial region of these five lampreys were 99.8% – 99.9 % identical to *M. mordax* #1D (Table 10). In contrast, nucleotide alignment of the *Cyt b – 16S* mitochondrial region of these five lampreys to that reported by Haouchar (2009) resulted in only 87.5 – 91.2% similarity (Table 10).

Table 10. Expected versus observed size of the *Cyt b – 16S* mitochondrial region.

	Size of *Cyt b – 16S* mitochondrial region		**Nucleotide similarity**		
	Expected* (bp)	Observed (bp)	% identical to Haouchar (2009)	% identical to *M. mordax* #1D	% identical to *M. mordax #1D* excluding NCIII
Individual					
M. praecox #3	3541	3978	87.5	95.7	99.9
M. mordax #4	3541	3815	91.2	99.8	99.9
M. mordax #23	3541	3976	87.6	95.7	99.9
M. mordax #24	3541	3830	90.9	99.2	99.8
M. mordax #25	3541	3830	90.9	99.2	99.8
M. mordax #1D	3541	3818	91.2	-	

* Expected size is based on the complete *M. mordax* mitogenome sequence of Haouchar (2009). Individuals highlighted in blue have a NCIII comprising 3 repeats. Non highlighted individuals have a NCIII comprising 2 repeats

3.2.4 NCIII

Noncoding region III (NCIII) is a tandem repeat array located between *Cyt b* and tRNA proline (Figure 16). Depending on the sequenced individual, the length of NCIII varied between 274 – 437 bp in length (Table 11). Length variation was primarily due to repeat copy number variation, which was polymorphic between individuals of both species. The NCIII of *M. praecox* #3 and *M. mordax* #23 consisted of three repeats comprising two complete copies and one partial copy (Figure 16A). In contrast, the NCIII of *M. mordax* individuals #1D, #4, #24 and #25 contained two repeats corresponding to one complete copy and one partial copy (Figure 16B). Repeat copies were designated as 5’, internal, or 3’ based on their position on the light-strand (L-strand) relative to the *Cyt b* gene. The size of the repeats for each sequenced individual is summarised in Table 11.

Table 11. Number of repeats, overall size of NCIII and the size of each component comprising of each repeat unit.

	NCIII Repeat Unit and Component Size (bp)															NCIII Region	
	5' - Partial Repeat					Internal Repeat					3' - Complete Repeat					Total	No. of
Individuals	Pro*	IGN	Phe*	RES	Total	Pro*	IGN	Phe*	RES	Total	Pro*	IGN	Phe*	RES	Total	Size (bp)	Repeats
M. praecox #3	9	4	76	20	109	64	4	76	20	164	64	4	76	20	164	437	3
M. mordax #23	9	4	76	20	109	63	4	76	20	163	63	4	76	20	163	435	3
M. mordax #1D	9	4	76	20	109	-	-	-	-	-	68	4	76	20	168	277	2
M. mordax #4	9	4	76	20	109	-	-	-	-	-	65	4	76	20	165	274	2
M. mordax #24	9	4	76	23	112	-	-	-	-	-	71	4	76	26	177	289	2
M. mordax #25	9	4	76	23	112	-	-	-	-	-	71	4	76	26	177	289	2
M. mordax [&]	9	4	7^	-	20^	-	-	-	-	-	-	-	-	4^	4^	24^	-
M. mordax!	9	4	76	22	111	-	-	-	-	-	32^	ND	ND	ND	ND	143^	1^

*Psuedogene

[&] Haouchar (2009)

^ Partial

! based on the re-examination in this study (Appendix 5) of 454 sequence data from Haouchar (2009) study

IGN = intergenic nucleotides

RES = Repeat End Sequence - as referred to in Section 3.2.4 and detailed in Section 3.2.4.3

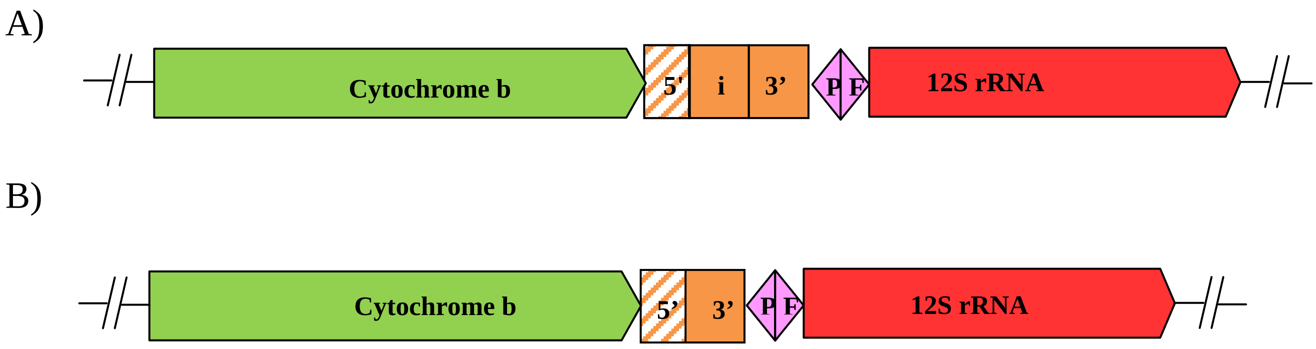


Figure 16. The gene arrangement and location of NCIII (orange boxes) as determined in this study for **A)** *M. praecox* #3 and *M. mordax* #23, and **B)** *M. mordax* individuals #1D, #4, #24 and #25. The repeat copies comprising NCIII are annotated 5', i (= internal) and 3' as outlined in the text. A hatched orange box indicates partial repeat, a solid orange box indicates complete repeat. Note: arrowheads indicate the direction of transcription for each gene. Proline and phenylalanine tRNA genes are abbreviated as P and F, respectively.

Each repeat copy comprised a tandem duplication of the tRNA proline – tRNA phenylalanine mitochondrial gene region and was followed by a C rich sequence. This C rich sequence is hereafter referred to as the Repeat End Sequence (RES) as it was located at the 3'-end of each repeat unit. Intra-individual repeat variation (if present) was confined to the RES and due to either an A⇆G base change or variation in cytosine homopolymer length (Figure 17). Inter-individual repeat variation was mostly associated with the putative tRNA proline pseudogene (pseudoPro) by minor insertions/deletions (Figure 17). Each component of the repeat units is now briefly described in context of the L-strand (major sense strand 5'–3') sequence.

```
                      3’ Arm                           tRNA proline                                5’ Arm     IGN sequence
                      ||||||||<<<<< TC Loop     <<<<<    <<<< <    AC Loop<<<<< <<<<D Loop <<<<  || | ||||
    Pro-Phe Region    TCAAAGAAAGAGAATTAGAA------TCTCTATTACTAG-CC---CCCAA-AGCTAGTATTTTTAAATTAAATTATT-C-TTTG   AATAA
M.Praecox #3 5’PR                                                                            ......A.C----   -....
M.Praecox #3 iCR      ....................------.............-..---.....C.A.-.........----T-.......A.C----   -....
M.Praecox #3 3’CR     ....................------.............-..---.....C.A.-.........----T-.......A.C----   -....
M.mordax #1D 5’PR                                                                             ......A.C----   -....
M.mordax #1D 3’CR     ....................------.............G..CCC.....C.A.-.........----T-.......A.C----   -....
M.mordax #4  5’PR                                                                             ......A.C----   -....
M.mordax #4  3’CR     ....................------.............-..—-......C.A.-.........----T-.......A.C----   -....
M.mordax #23 5’PR                                                                             ......A.C----   -....
M.mordax #23 iCR      .................A..------........-....-..---.....C.A.-.........----T-.......A.C----   -....
M.mordax #23 3’CR     .................A..------........-....-..---.....C.A.-.........----T-.......A.C----   -....
M.mordax #24 5’PR                                                                             ......A.C----   -....
M.mordax #24 3’CR     ....................TTAGAA.............G..---.....-.............----T-.......A.C----   -....
M.mordax #25 5’PR                                                                             ......A.C----   -....
M.mordax #25 3’CR     ....................TTAGAA.............G..---.....-.............----T-.......A.C----   -....

                     5’ Arm                  tRNA phenylalanine                            3’ Arm
                     |||||||  >>>> D loop>>>> >>>>> ACloop>>>>  >     >>>> TC loop    >>>> |||||||         RES sequence
     Pro-Phe Region  GTTAAAATAGCTTAAATTTAAAGCAAAGCACTGAAGATGCT--TA-GATGGTC-----TAATG—-GGCCCTTTAACA
M.Praecox #3 5’PR    ......G........G.............T.-.........AG..T.......CAAACA....ACC...........   TCAA------CCCCCTCAATCAAACA
M.Praecox #3 iCR     ......G........G.............T.-.........AG..T.......CAAACA....ACC...........   TCAA------CCCCCTCAATCAAACA
M.Praecox #3 3’CR    ......G........G.............T.-.........AG..T.......CAAACA....ACC...........   TCAG------CCCCCTCAATCAAACA
M.mordax #1D 5’PR    ......G........G.............T.-.........AG..T.......CAAACA....ACC...........   TCAA------CCCCCTCAATCAAACA
M.mordax #1D 3’CR    ......G........G.............T.-.........AG..T.......CAAACA....ACC...........   TCAA------CCCCCTCAATCAAACA
M.mordax #4 5’PR     ......G........G.............T.-.........AG..T.......CAAACA....ACC...........   TCAA------CCCCCTCAATCAAACA
M.mordax #4 3’CR     ......G........G.............T.-.........AG..T.......CAAACA....ACC...........   TCAG------CCCCCTCAATCAAACA
M.mordax #23 5’PR    ......G........G.............T.-.........AG..T.......CAAACA....ACC...........   TCAA------CCCCCTCAATCAAACA
M.mordax #23 iCR     ......G........G.............T.-.........AG..T.......CAAACA....ACC...........   TCAA------CCCCCTCAATCAAACA
M.mordax #23 3’CR    ......G........G.............T.-.........AG..T.......CAAACA....ACC...........   TCAA------CCCCCTCAATCAAACA
M.mordax #24 5’PR    ......G........G.............T.-.........AG..T.......CAAACA....ACC...........   TCAA---CCCCCCCCTCAATCAAACA
M.mordax #24 3’CR    ......G........G.............T.-.........AG..T.......CAAACA....ACC...........   TCAACCCCCCCCCCCTCAATCAAACA
M.mordax #25 5’PR    ......G........G.............T.-.........AG..T.......CAAACA....ACC...........   TCAA---CCCCCCCCTCAATCAAACA
M.mordax #25 3’CR    ......G........G.............T.-.........AG..T.......CAAACA....ACC...........   TCAACCCCCCCCCCCTCAATCAAACA
```

Figure 17. L-strand nucleotide alignment of tRNA Proline – tRNA Phenylalanine mitochondrial region to the tandem repeat copy units of NCIII found in *M. mordax* and *M. praecox* individuals sequenced in this study. The tRNA Proline – tRNA Phenylalanine mitochondrial region (Pro-Phe Region) is identical in all individuals sequenced in this study. The repeats occur in a tandem orientation but are aligned here for comparison. Dots indicate identical bases while dashes indicate deletions. 5', i (= internal) and 3' refer to the order in which copies occur on the L-strand. PR = partial repeat and CR =complete repeat. Genes and other features are marked above the sequence and are abbreviated as used in the text. Transfer RNA genes are marked above the sequence and their features marked with: arrows (< or >) indicating direction of transcription; | indicating nucleotides part of the tRNA arm and are as annotated; and the anticodon is bold and underlined. The bind site for primer X51F is indicated by the red text and the bind site for primer X50R is indicated by the blue text.

3.2.4.1 PseudoPro

Because the 5’-partial repeat copy was located adjacent to the 3’-end of *Cyt b* (Figure 16), pseudoPro in the partial repeat copy was only 9 nucleotides in length and corresponded to the complementary 5’-end sequence of the pseudogene (i.e the 3’-end of the L-strand pseudogene sequence). In complete repeat copies, pseudoPro was 63 – 71 nucleotides in length and had 81.1 – 85.7 % pairwise similarity to the 78 nucleotide tRNA proline sequence (Figure 17). As seen in Figure 17, inter-individual size variation of pseudoPro was due minor insertions/deletions and/or variation in the cytosine homo-nucleotide length of the AC loop.

PseudoPro is likely non-functional as it is unable to form the canonical tRNA clover-leaf secondary structure due to mutations/deletions - especially the 4 base deletion in the complementary sequence of the 5’-terminus arm of the pseudogene (Figure 17) which prevents formation of the acceptor stem. However, nucleotide alignment to tRNA proline (Figure 17) shows that pseudoPro is capable of forming the TC loop and, in the case of *M. mordax* #24 and #25, capable of forming the anti-codon loop. Whether these loops can play a role or function in transcription/translation of the precursor RNA transcripts remains unclear.

3.2.4.2 PseudoPhe

The putative tRNA phenylalanine pseudogene (pseudoPhe) was identical in all six sequenced individuals (Figure 17). PseudoPhe (76 nucleotides) is also 79.2% identical (61 identical sites) to the 67 nucleotide tRNA phenylalanine (Figure 17). Based on the predictions of webservers mfold (Zuker 2003) and tRNAscan-SE, pseudoPhe is capable of forming a stable clover-leaf secondary structure (Figure 18). In comparison to tRNA phenylalanine, the secondary structure of pseudoPhe has a variant TC loop, shorter AC stem and larger variable loop. How these changes affect the functionality of pseudoPhe is unclear, thus, pseudoPhe is assumed to be non-functional until transcription studies can be carried out.

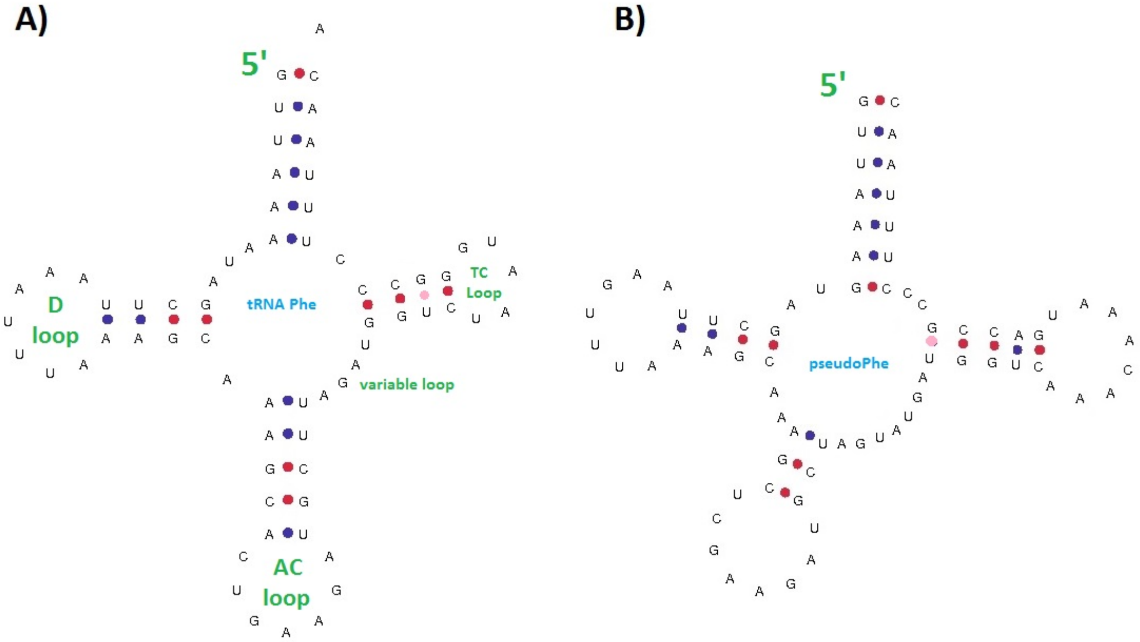


Figure 18. Comparison of the predicted secondary structure of **A)** tRNA phenylalanine, and **B)** pseudoPhe. Blue and red dots signify Watson-Crick base pairing. Note: only the most stable secondary structure (i.e. largest -ΔG value) predicted by mfold is shown - these vary slightly to the secondary structures predicted by tRNAscan-SE due to inherent differences in programming employed by the two webservers (i.e allowance of non-Watson-Crick base pairings such as G-U (pink dots).

3.2.4.3 RES

Within the 3’-end of each repeat unit was a C rich sequence of 20 – 26 nucleotides in length (Figure 17). The 20 nucleotide consensus sequence found in most individuals was TCAACCCCCTCAATCAAACA. This sequence contained the key motif CAACCCCCT found in conserved sequence blocks (CSB) II and III (CBS detailed in Section 3.6.3), which are located in NCI of lampreys and are thought to play a role in heavy-strand replication (Lee & Kocher 1995).

3.2.4.4 Sequencing of X50F/R and X51F/R Amplicons

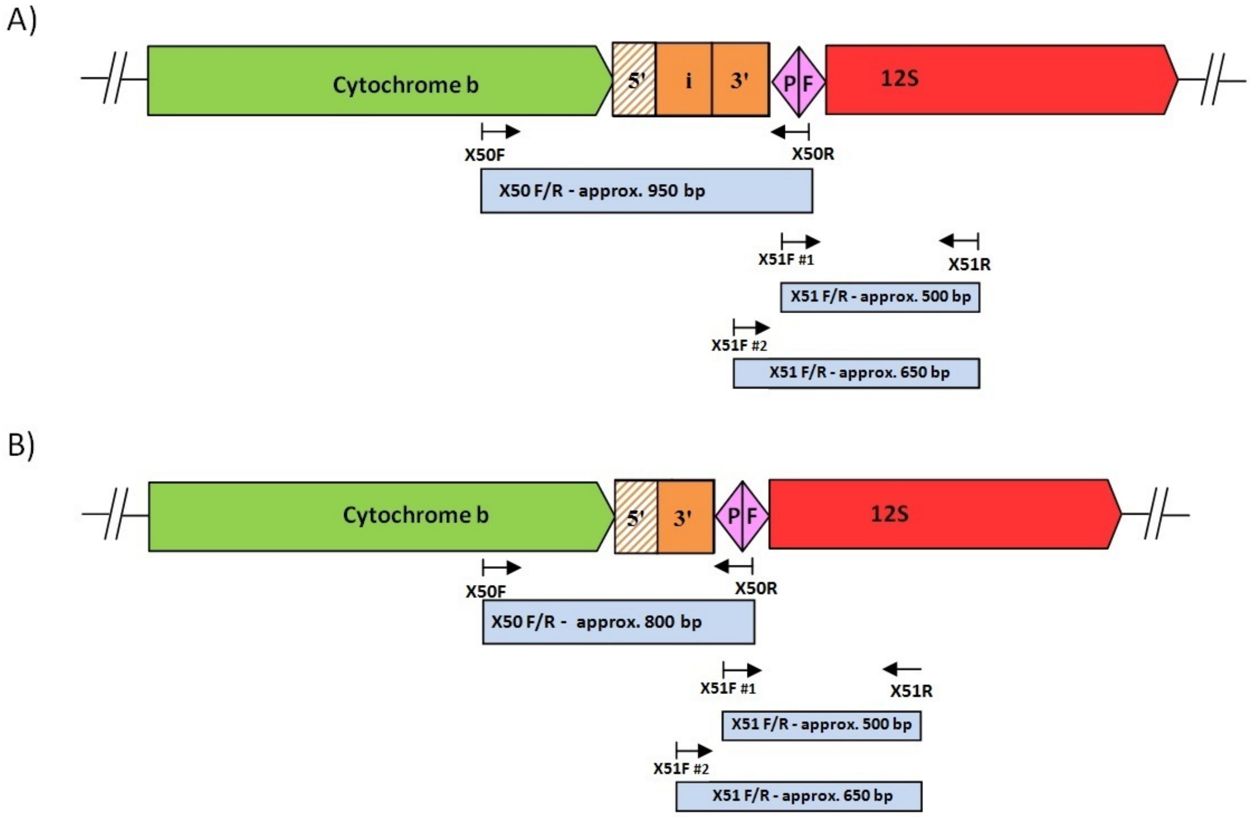


Figure 19. Observed size of the amplicons (blue boxes) generated by primer pair X50F/R and primer pair X51F/R due to NCIII (orange boxes). **A)** Observed amplicon sizes in individuals with NCIII comprising three repeats, and **B)** Observed amplicon sizes in individuals with NCIII comprising two repeats. Note: primer pair X51F/R yields two amplicons due to an alternative X51F primer bind site (#2) located in the 3' repeat of NCIII. The repeat copies comprising NCIII are annotated *5',* i (= internal) and 3' as outlined in the text. A hatched orange box indicates partial repeat, a solid orange box indicates complete repeat. Arrowheads indicate the direction of transcription for each gene. Proline and phenylalanine tRNA genes are abbreviated as P and F, respectively.

Sequencing of X50F/R amplicons revealed a direct correlation between the observed band sizes seen for X50F/R amplicons (reported in Table 9) and the number of repeats comprising NCIII of an individual. As shown in Figure 19, sequenced individuals (*M. praecox* #3 and *M. mordax* #23) with observed X50F/R amplicon size of approximately 950 bp had a NCIII comprised of three repeats, whilst sequenced individuals (*M. mordax* #1D, #4, #24 and #25) with observed X50F/R amplicon size of approximately 800 bp had only 2 repeats. The approximate 150 bp difference in X50F/R band size observed between individuals in this study (reported in Table 9) was therefore due to the additional complete repeat copy in 3 tandem repeat individuals (Figure 19), the size of which was 164 bp in *M. praecox* #3 and 163 bp in *M. mordax* #23.

Sequencing revealed the cause of the secondary (650 bp) X51F/R band to be due to an alternative bind site for the forward primer X51F (Figure 19) located in pseudoPro of the 3’-complete repeat adjacent to tRNA proline. Due to each tandem repeat containing sequence similar to that of tRNA proline and phenylalanine, each repeat contained possible alternative (imperfect) bind sites for primer X50R and primer X51F.

3.2.4.5 Confirmation of NCIII by PCR and Sanger Sequencing

Because sequencing had identified alternative (imperfect) priming sites for primer X50R primer X51F within the tandem repeats of NCIII, it was possible that NCIII was a PCR artefact resulting from alternative priming. To determine if this was the case, PCR was carried out using primer X50F and primer X51R (as their binding sites flanked NCIII) and the resulting amplicon was sequenced by Sanger sequencing (Figure 20). If NCIII was indeed a PCR artefact, the amplicon generated by primer pair X50F/X51R was expected to be approximately 900 bp (Figure 20A). If NCIII was not an artefact of PCR, individuals with NCIII comprising two repeats would yield amplicons approximately 1200 bp in size (Figure 20B) and individuals with NCIII comprising three repeats would yield amplicons approximately 1350 bp in size (Figure 20C).

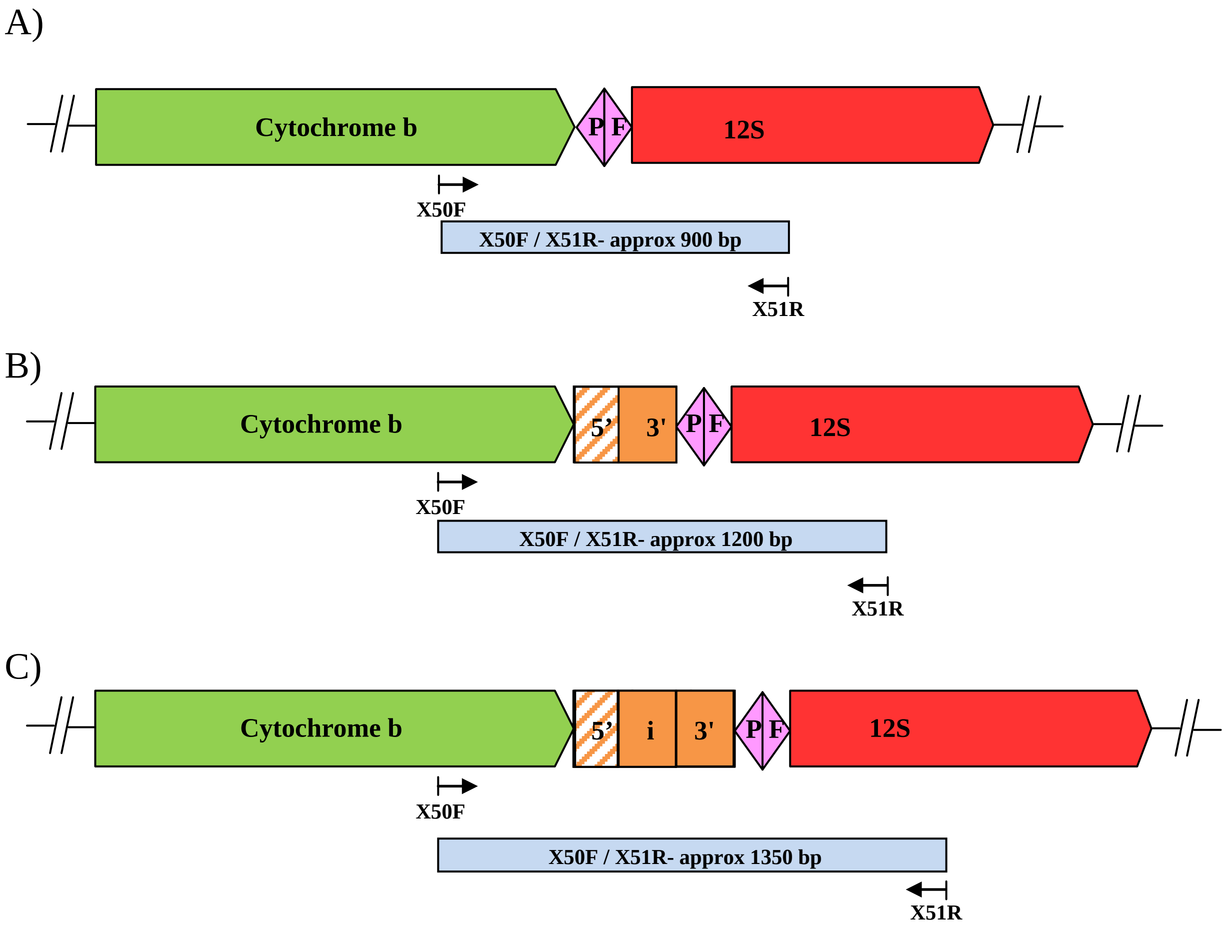


Figure 20. Expected size of the amplicon (blue box) generated by primer pair X50F/X51R based on the presence or absence of NCIII (orange boxes). **A)** Expected size if NCIII was a PCR artefact, **B)** Expected size if NCIII contains two repeats, and **C)** Expected size if NCIII contains three repeats. Note: the repeat copies comprising NCIII are annotated 5', i (= internal) and 3' as outlined in the text. A hatched orange box indicates partial repeat, a solid orange box indicates complete repeat. Arrowheads indicate the direction of transcription for each gene. Proline and phenylalanine tRNA genes are abbreviated as P and F, respectively.

Agarose gel electrophoresis of X50F/X51R amplicons repeatedly resulted in single bands of approximately 1350 bp in size for those individuals with 3 repeats (Figure 21; lane 2 and 4) and approximately 1200 bp for individuals with 2 repeats (Figure 21: lanes 3, 5, and 6). This finding indicated that NCIII was not an artefact of PCR.

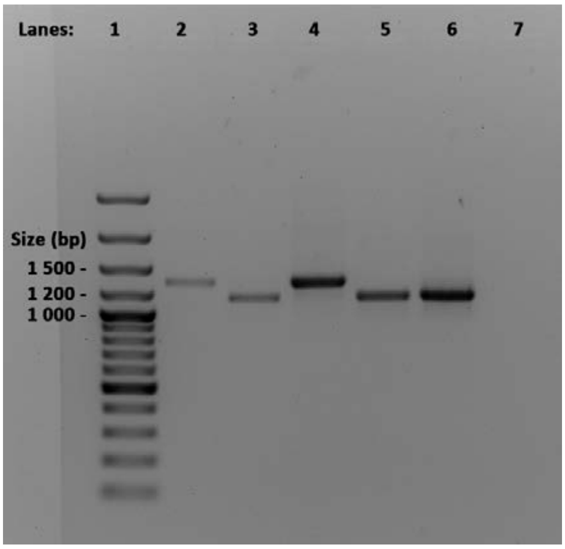


Figure 21. Comparison of X50F/X51R amplicon size between *M. praecox* and *M. mordax* individuals. Depending on the individual lamprey, the band is either approximately 1350 or 1200 bp in size. **Lane 1)** 100 plus ladder, **Lane 2)** *M. praecox* #3, **Lanes 3 – 6)** *M. mordax* individuals #4, #23, #1D and #2, and **Lane 7)** negative template control. Agarose gel (0.8%) after electrophoresis for 1hr at 90v and stained with SYBRsafe.

Following PCR, Sanger sequencing of X50F/X51R amplicons was carried out in both directions (Section 2.7.1.2). Whilst X50F/X51R amplicons could not be read in their entirety due to size limitations of the ABI sequencer, each sequenced strand provided sufficient coverage and overlap to result in a trimmed consensus sequence between 1044 bp and 1208 bp in length. Subsequent inclusion into the *de novo* assembly of the *Cyt b – 12S* mitochondrial region, whether as individual or consensus reads, resolved any issue with alignment posed by the repeats of NCIII and confirmed that this region was not due to alternative priming (Figure 22).

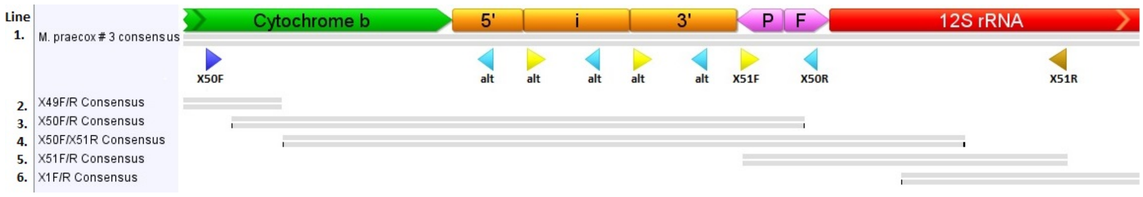


Figure 22. Sequencing coverage of NCIII in *M. praecox* #3 (line 1). Due to the priming sites of primers X50R (yellow arrow as annotated) and X51F (blue arrow as annotated) being located within the tRNA genes phenylalanine (pink F arrow) and proline (pink P arrow), respectively, it was possible that the tandem repeats of NCIII (three orange boxes) may have been due to alternative (imperfect) priming sites (blue and yellow arrows annotated alt). Amplification was carried out using primers X50F (dark blue arrow as annotated) and X51R (gold arrow as annotated) as their binding sites flanked NCIII. Sanger sequencing and inclusion of X50F/X50R sequence (line 4) shows that the region is not due to alternative priming as it independently supports the sequences of amplicon X50F/R (line 3) and amplicon X51F/R (line 5). Note: L-strand consensus sequences are shown (5' – 3') for simplicity. Each amplicon consensus sequence has two times coverage. Direction of primer arrows indicates amplification direction. Only a portion of the *Cyt b* (green) and *12S* rRNA (red) gene are shown. The repeat copies comprising NCIII are annotated 5*'*, i (= internal) and 3' as outlined in the text.

3.3 Long Range PCR

To obtain the complete mitochondrial genome of multiple lamprey individuals, long range PCR (Section 2.7.2) was attempted on all twelve lamprey individuals using five primer pairs (Long 1F/R; Long 2F/R; Long 3F/R; Long 4F/R; and Long 5F/R) that would generate overlapping amplicons of 5000-6000 bp in size and encompass the complete mitogenome (Figure 23). Subsequent sequencing on a 454 sequencing platform would enable the mitogenome sequences of multiple individuals to be determined simultaneously in a single sequencing run, thereby bypassing the laborious and expensive method of primer walking and Sanger sequencing.

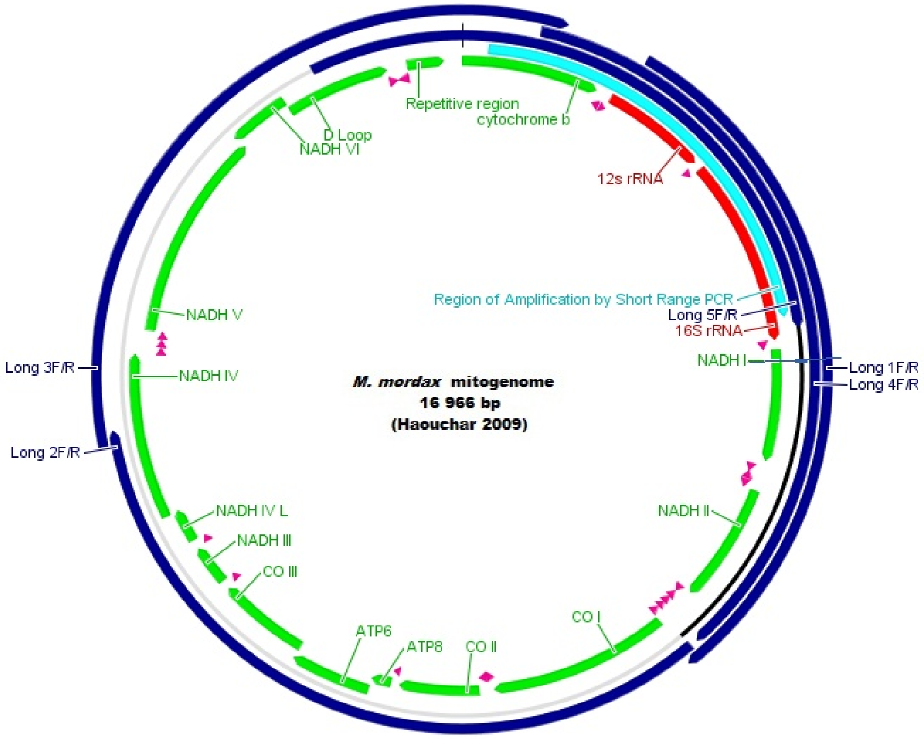


Figure 23. The overlapping long range PCR method of generating the expected 17kb mitogenome of *M. mordax and M. Praecox.* The 5 long range PCR primer pairs would generate amplicons of 5 – 6 Kb in size (dark blue as annotated) that would encompass the complete mitogenome. Also shown for comparison is the region determined by short range PCR (light blue as annotated). Note: L-strand (5' – 3') is shown for simplicity. Arrowheads indicate the direction of transcription for each coding gene (green arrows as annotated), rRNA (red arrows as annotated) and tRNAs (pink arrows).

3.3.1 Amplification

Amplification with long range primers was inconsistent and successful amplification was intermittent. For example, DNA extracts which had been successfully amplified often failed when used as a positive control in later amplification rounds - even when the variables were controlled (optimised conditions, same thermocycler, same primer and DNA stocks). Over one hundred long range PCRs were carried out using a total of three Velocity long range PCR kits. Long range PCR was intermittently successful using primer pair Long1F/R, primer pair Long2F/R and primer pair Long3F/R (Table 12). Amplification using primer pair Long4F/R and primer pair Long5F/R consistently failed despite the four primers working in different primer pair combination when amplifying short range products (results not shown). Primer pairs Long4F/R and Long5F/R were abandoned due to time constraints. In total, the mitogenome of three *M. mordax* individuals and three *M. praecox* individuals were partially amplified (Table 12). Only the mitogenome of *M. mordax* #25 and *M. praecox* #1 could be amplified using the three working primer pairs (Table 12).

Table 12. Overview of long range PCR success.

		Primer pair	
Individual	Long 1F/R	Long 2F/R	Long 3F/R
M. mordax #4	x	√	X
M. mordax #5	√	x	X
M. mordax #25	√	√	√
M. praecox #1	√	√	√
M. praecox #16	√	x	X
M. praecox #17	√	x	X

3.3.2 Electrophoresis of Long Range PCR Products.

Electrophoresis of amplicons generated by successful long range PCR (as reported in Table 12) resulted in single bands that were approximately of the expected size based on the *M. mordax* mitogenome of Haouchar (2009) : approximately 5000 bp for Long 1F/R amplicons; approximately 6000 bp for Long 2F/R amplicons; and approximately 5500 bp for Long 3F/R amplicons. Prior to sequencing (Section 2.7.2.2), single amplicon bands were extracted and the DNA recovered by the method described in Section 2.5.1.

3.4 454 Sequencing of Long Range PCR Products

The recovered amplicon DNA was sent to Frances Brigg at the Western Australian State Agricultural Biotechnology Centre (SABC) at Murdoch University for sequencing on a 454 sequencing platform. Sequencing resulted in 1347 – 6526 mappable reads per amplicon per individual. Minimum depth of coverage for each amplicon ranged from 28 – 362 reads per nucleotide (Table 13). Long 1F/R and Long 2FR amplicons had a minimum depth of coverage of 103 and 145 reads per nucleotide, respectively, whilst Long 3F/R amplicons had the lowest minimum depth of coverage with 28 reads per nucleotide for one individual (Table 13). The low coverage for Long 3F/R amplicons may have been caused by unequal DNA estimation. However, it is more likely to be due to the low GC content (found to be <16% in this study) of the second control region (NCII) and the difficulties associated with sequencing GC poor regions on a NGS platform (Chen *et al.* 2013)**.**

Table 13. Summary statistics of 454 sequencing

	Long 1F/R			**Long 2FR**			**Long 3FR**		
	No. of mappable reads	Depth of coverage*		No. of mappable reads	Depth of coverage*		No. of mappable reads	Depth of coverage*	
		Min.	Max.		Min.	Max.		Min.	Max.
Lamprey									
M4	-	-	-	3369	145	479	-	-	-
M5	4989	270	778	-	-	-	-	-	-
M25	3459	191	490	5763	244	876	2433	89	359
P1	2071	103	318	5831	236	892	1347	28	155
P16	6526	362	933	-	-	-	-	-	-
P17	3218	168	493	-	-	-	-	-	-

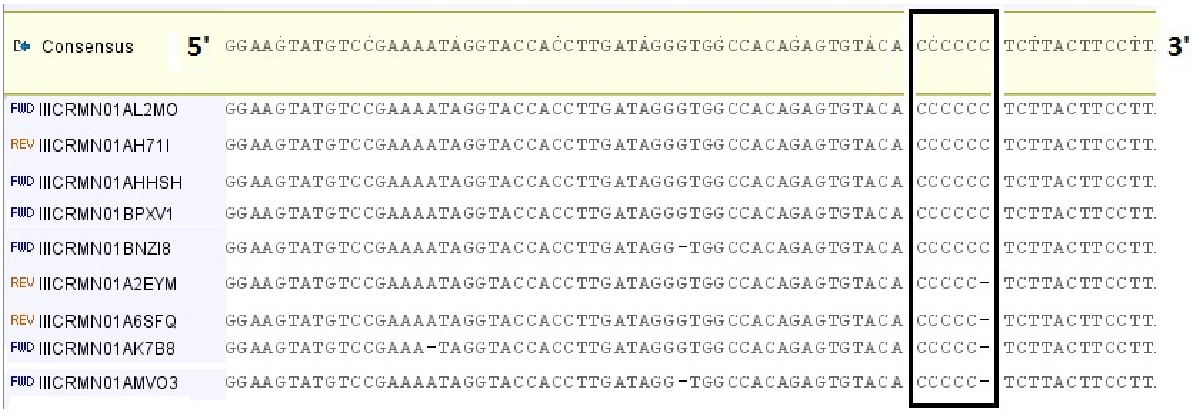
* Number of reads per nucleotide. - indicates not applicable

M4, M5 and M25 = *M. mordax* individuals #4, #5, and #25, respectively.

P1, P16 and P17 = *M. praecox* individuals #1, #16 and #17, respectively.

3.4.1 *12S* – tRNA Cysteine Mitochondrial Region

The *12S* – tRNA cysteine mitochondrial region was determined by sequencing Long 1F/R amplicons of two *M. mordax* individuals (#5 and #25) and three *M. praecox* individuals (#1, #16 and #17). The size of the *12S* – tRNA cysteine mitochondrial region was 5090 bp for both *M. mordax* individuals and *M. praecox* #1. The region was one base pair smaller for *M. praecox* #16 at 5089 bp in size, and one base pair larger for *M. praecox* #17 at 5091 bp in size. Nucleotide alignment of the *12S* – tRNA cysteine mitochondrial region of these five lampreys showed that the variance in length observed in *M. praecox* #16 and #17 was due to a cytosine (C) homopolymer (stretch of repeating bases) located in tRNA isoleucine. In *M. mordax* #5, #25 and *M. praecox* #1 the homopolymer was 5 bases, but in *M. praecox* #16 it was 4 bases and in *M. praecox* #17 it was 6 bases. The individual sequencing reads were checked to ensure correct base calling (Figure 24).



```
Consensus      5' GGAAGTATGTCCGAAAATAGGTACCACCTTGATAGGGTGGCCACAGAGTGTACA CCCCCC TCTTACTTCCTT. 3'
FWD IIICRMN01AL2MO  GGAAGTATGTCCGAAAATAGGTACCACCTTGATAGGGTGGCCACAGAGTGTACA CCCCCC TCTTACTTCCTT.
REV IIICRMN01AH71I  GGAAGTATGTCCGAAAATAGGTACCACCTTGATAGGGTGGCCACAGAGTGTACA CCCCCC TCTTACTTCCTT.
FWD IIICRMN01AHHSH  GGAAGTATGTCCGAAAATAGGTACCACCTTGATAGGGTGGCCACAGAGTGTACA CCCCCC TCTTACTTCCTT.
FWD IIICRMN01BPXV1  GGAAGTATGTCCGAAAATAGGTACCACCTTGATAGGGTGGCCACAGAGTGTACA CCCCCC TCTTACTTCCTT.
FWD IIICRMN01BNZI8  GGAAGTATGTCCGAAAATAGGTACCACCTTGATAGG-TGGCCACAGAGTGTACA CCCCCC TCTTACTTCCTT.
REV IIICRMN01A2EYM  GGAAGTATGTCCGAAAATAGGTACCACCTTGATAGGGTGGCCACAGAGTGTACA CCCCC- TCTTACTTCCTT.
REV IIICRMN01A6SFQ  GGAAGTATGTCCGAAAATAGGTACCACCTTGATAGGGTGGCCACAGAGTGTACA CCCCC- TCTTACTTCCTT.
FWD IIICRMN01AK7B8  GGAAGTATGTCCGAAA-TAGGTACCACCTTGATAGGGTGGCCACAGAGTGTACA CCCCC- TCTTACTTCCTT.
FWD IIICRMN01AMVO3  GGAAGTATGTCCGAAAATAGGTACCACCTTGATAGG-TGGCCACAGAGTGTACA CCCCC- TCTTACTTCCTT.
```

Figure 24. Example of the 454 sequencing reads spanning tRNA isoleucine of *M. praecox* #16. The consensus sequence (top - line 1) indicates that the cytosine homopolymer (C stretch enclosed in the black box) is six bases in length as the number of 454 sequencing reads with 6C's (lines 2 – 6) is greater than the number of 454 sequencing reads with 5C's (lines 7 – 10). Note: only nine sequencing reads are shown as an example. Only the tRNA isoleucine portion of each read is shown for simplicity. As indicated, the sequences are shown in the 5' – 3' direction.

In *M. praecox* #16, the homopolymer had a minimum depth of coverage of 458 reads of which over 400 had the homopolymer as being 6 bases long (Figure 24). In *M. praecox*

#17, the homopolymer had a minimum depth of 188 reads of which over 170 had the homopolymer as being 4 bases long. This suggested that the observed differences are unlikely to be due to sequencing errors and that *M. praecox* #16 and #17 are polymorphic. The secondary structure of tRNA isoleucine was predicted using the web server tRNAscan-SE to determine if the polymorphism would affect the tRNA. The predicted secondary structures show that the polymorphism affects the T arm and the size of the T loop (Figure 25). Given that the T arm is the recognition site for the ribosome (Redko *et al.* 2007), it is unclear as to what effect it would have on the formation of the tRNA isoleucine – ribosome complex and thus tRNA function.

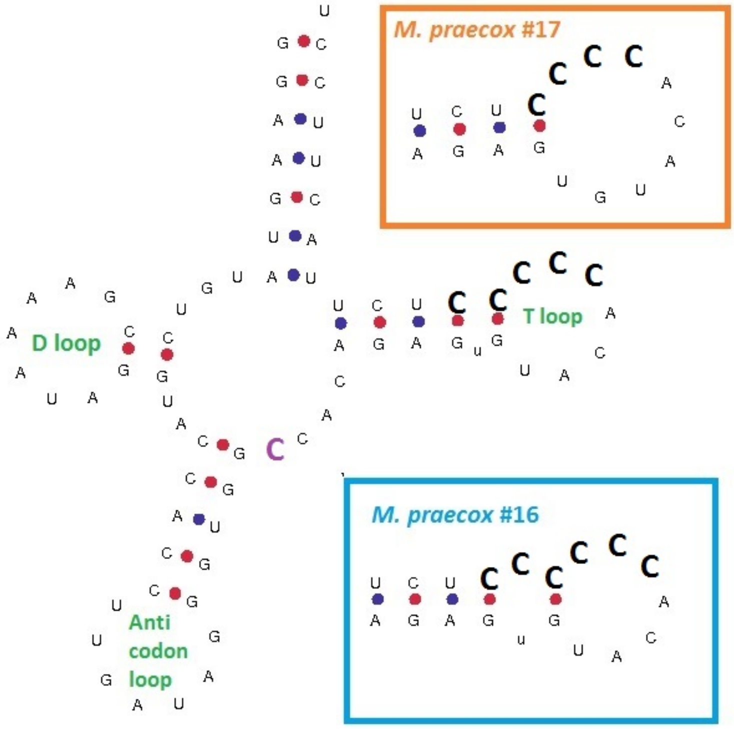


Figure 25. The predicted secondary structure of tRNA isoleucine and the predicted effect of the homopolymer (bold C). Shown is the secondary structure for *M. mordax* #5, *M. mordax* #25 , and *M. praecox* #1 in which the 5C homopolymer forms the T loop and T arm. The orange box shows the 4C homopolymer increasing the size of the T loop for *M. praecox* #17. The blue box shows the 6C homopolymer found in *M. praecox* #16 causing a secondary loop in the T arm. Blue and red dots denote Watson-Crick base pairing. Note: in *M. mordax* #25 the purple C is a U.

Given that lamprey tRNAs are fairly conserved between species (Delarbre *et al.* 2000), the tRNA isoleucine sequences of *P. marinus* (Lee & Kocher 1995), *P. fluviatilis* (Delarbre *et al.* 2000), *L. camtschaticum* (Hwang *et al.* 2013a) and *L. reissneri* (Hwang *et al.* 2013b) were aligned to determine the importance of the homopolymer. The homopolymer was found to be conserved in all 4 species and was five bases in length. This finding suggests that the five base homopolymer may be important in tRNA isoleucine - ribosome interactions.

Despite the observed polymorphic difference in tRNA isoleucine, the *12S* – tRNA cysteine mitochondrial region of the five individuals were nearly identical (at least 99.9%) with no more than 8 individual differences observed (Table 14).

Table 14. Number of nucleotide differences (below diagonal) and percentage similarity (above diagonal) between *12S* – tRNA cysteine mitochondrial sequences.

Individual	1.	2.	3.	4.	5.	6.
1. *M. mordax**	-	98.0%	98.0%	98.0%	98.0%	98.0%
2. *M. mordax* #5	103	-	99.9%	99.9%	99.9%	99.9%
3. *M. mordax* #25	105	7	-	99.9%	99.9%	99.8%
4. *M. praecox* #1	102	1	6	-	99.9%	99.9%
5. *M. praecox* #16	102	2	7	1	-	99.9%
6. *M. praecox* #17	103	3	8	2	3	-

* *12S* – tRNA cysteine sequence reported by Haouchar (2009)

3.4.2 tRNA Tyrosine – *ND4* Mitochondrial Region

The tRNA tyrosine – *ND4* mitochondrial region of *M. praecox* #1 and *M. mordax* #4 and #25 was determined by sequencing Long 2F/R amplicons. The size of this region was 5776, 5775 and 5777 bp in size, respectively. Nucleotide alignment of the tRNA tyrosine – *ND4* mitochondrial region of these three lampreys showed that the variance in length was due to two homopolymer stretches (one adenine and the other thymine) located in *ND4*. The incorrect base calling in both *M. mordax* #4 (both sites) and *M. praecox* #1 (T site only) caused shifting of the reading frame, resulting in premature stop codons throughout the gene, giving an unexpected gene size of 69 and 454 bp in size, respectively. In contrast, the correct base call observed in *M. mordax* #25 at both sites resulted in no premature stop codons and the expected gene size of 799 bp. The individual sequencing reads were checked to ensure correct base calling (Figure 26).

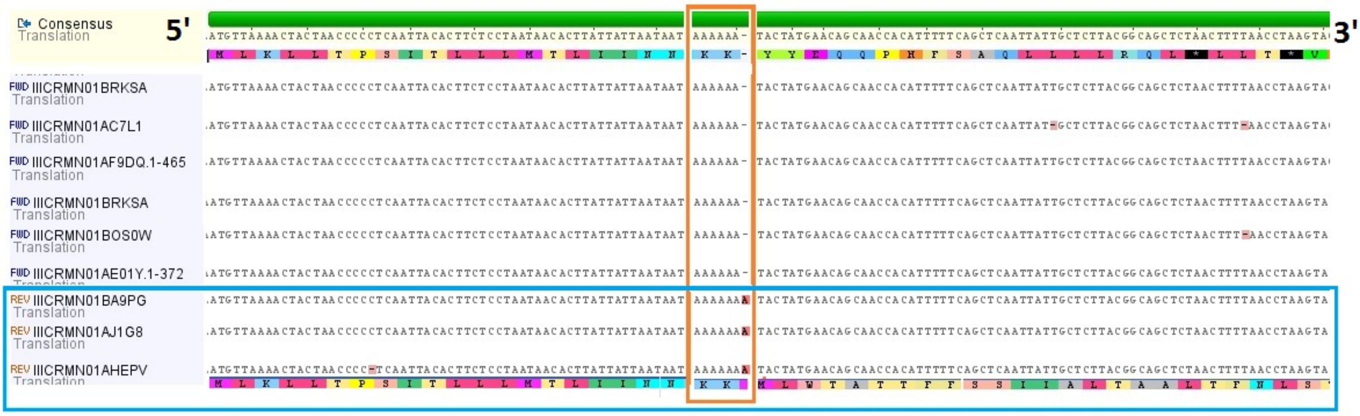


Figure 26. Example of the 454 sequencing reads spanning first 138 bp of *ND4* of *M. mordax* #4. The consensus sequence (top - line 1) indicates that the adenine homopolymer (A stretch enclosed in the black box) is six bases in length as the number of 454 sequencing reads with 6A's (lines 3 – 8) is greater than the number of 454 sequencing reads with 7A's (lines 9 – 11). However, 6A's is incorrect as it results in premature stop codons (black boxes; line 2), whilst 7A's is correct as it changes the reading frame to eliminate the stop codons (line 11). Note: only nine sequencing reads are shown as an example. Only 138 bases of each read are shown for simplicity. As indicated, the sequences are shown in the 5' – 3' direction.

In *M. mordax* #4, the adenine homopolymer had a minimum read depth of 161 reads of which the ratio of reads with 6A's : 7A's was approximately 60 : 40. The thymine homopolymer had a minimum read depth of 434 reads in which the ratio of reads with 5T's : 6T's was again approximately 60:40. This same approximate ratio was also seen in T homopolymer reads of *M. praecox* #1 - despite having twice the minimum read depth at 813 reads. Given that mitochondrial protein coding genes are involved in cellular respiration and thus vital for life, such interruption to the reading frame of the gene would be deleterious to the organism and most likely represent sequencing errors associated with homopolymer stretches (Jex *et al.* 2010). When corrected, the tRNA tyrosine – *ND4* mitochondrial region of all three lampreys was 5777 bp in length. Furthermore, the nucleotide sequences were nearly identical (at least 99.8%) with less than 13 individual differences observed (Table 15).

Table 15. Number of nucleotide differences (below diagonal) and percentage similarity (above diagonal) between tRNA tyrosine – *ND4* mitochondrial sequences.

Individual	**1.**	**2.**	**3.**	**4.**
1. *M. mordax**	-	98.2%	98.0%	98.2%
2. *M. mordax* #4	107	-	99.8%	99.9%
3. *M. mordax* #25	116	13	-	99.8%
4. *M. praecox* #1	106	1	12	-

* tRNA tyrosine - *ND4* sequence reported by Haouchar (2009)

3.4.3 *ND4L – Cyt b* Mitochondrial Region

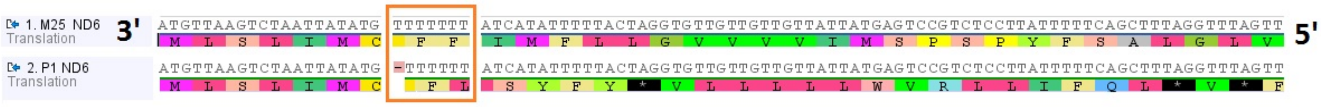
The *ND4L – Cyt b* mitochondrial region of *M. mordax* #25 and *M. praecox* #1 amplified by primer pair Long 3F/R were 5403 bp and 5413 bp in size, respectively. However, this was not unexpected as the two noncoding regions (NCI and NCII) found in lampreys are known to vary in size and composition between lamprey species (Delarbre *et al.* 2000; Hwang *et al.* 2013b). The *ND4L – Cyt b* mitochondrial region of *M. mordax* #25 and *M. praecox* #1 was 99.2% identical (Table 16). Whilst 43 nucleotide differences were observed (Table 16), this was mostly associated with the repetitive nature of NCII. Two of the nucleotide differences were due to homopolymers: one located in *ND6* and one in NCI. The thymine homopolymer in *ND6* was 6 bases long in *M. praecox* #1 and 7 bases long in *M. mordax* #25.

Table 16. Number of nucleotide differences (below diagonal) and percentage similarity (above diagonal) between *ND4L – Cyt b* mitochondrial sequences.

Individual	***M. mordax****	***M. mordax* #25**	***M. praecox* #1**
M. mordax*	-	96.1%	96.1%
***M. mordax* #25**	217	-	99.2%
***M. praecox* #1**	215	42	-

* *ND4L – Cyt b* sequence reported by Haouchar (2009)

However, underestimation of the homopolymer length in *M. praecox* #1 resulted in a frame shift that caused premature stop codons throughout the gene (Figure 27), such that *ND6* would have only been 42 bp in length relative to 522 bp in length in *M. mordax* #25. Surprisingly, when the individual sequencing reads of *M. praecox* #1 were checked, 50 reads supported the incorrect 6 base length and only 6 reads supported the correct 7 base length.



```
1. M25 ND6   3'  ATGTTAAGTCTAATTATATG TTTTTTT ATCATATTTTTACTAGGTGTTGTTGTTGTTATTATGAGTCCGTCTCCTTATTTTTCAGCTTTAGGTTTAGTT  5'
Translation       M  L  S  L  I  M  C    F  F   I  M  F  L  L  G  V  V  V  V  I  M  S  P  S  P  Y  F  S  A  L  G  L  V
2. P1 ND6        ATGTTAAGTCTAATTATATG -TTTTTT ATCATATTTTTACTAGGTGTTGTTGTTGTTATTATGAGTCCGTCTCCTTATTTTTCAGCTTTAGGTTTAGTT
Translation       M  L  S  L  I  M  C     F  L   S  Y  F  Y  *  V  L  L  L  L  L  W  V  R  L  L  I  F  Q  L  *  V  *  F
```

Figure 27. Homopolymer length and its effect on *ND6* amino acid sequence. Nucleotide alignment of the first 99 bases of *ND6* gene of *M. mordax* #25 (line 1) to the first 98 bases of *M. praecox* #1 (line 2). Both sequences were determined by 454 sequencing and differ only by a single nucleotide due to variation in the thymine homopolymer (T stretch enclosed in the orange box). Underestimation of the homopolymer (6T's) in *M. praecox* #1 causes a frame shift, resulting in premature stop codons (black boxes: line 4) and a gene length of only 52 bases (13 amino acids). In comparison, the homopolymer in *M. mordax* #25 is 7 bases (line 1) and results in the expected gene length of 522 bases (174 amino acids). Note: as indicated, the sequences are shown in the 3' – 5' direction due to *ND6* being located on the L-strand. Dash in pink box indicates nucleotide is not present.

The cytosine homopolymer in NCI was 10 bases long in *M. praecox* #1 and 11 bases long in *M. mordax* #25. Because NCI is a noncoding region, it is harder to determine whether the one base difference is due to homopolymer underestimation or individual variation as inserts/deletions in noncoding regions do not result in frameshift mutations. When the individual sequencing reads were checked for *M. praecox* #1, the homopolymer had 75 x read depth with 57 reads supporting 10C and only 18 reads supporting 11C. In comparison, the homopolymer in *M. mordax* #25 had 156 x read depth with 114 reads supporting 11C. This finding tends to suggest that the single base difference may indeed be due to individual variation (but see Section 3.5).

3.4.4 Comparison to *M. mordax* (Haouchar 2009)

Based on the complete *M. mordax* mitogenome of Haouchar (2009), none of the mitochondrial regions determined by 454 sequencing in this study were of expected size

(Table 17). The *12S* – tRNA cysteine mitochondrial region was expected to be 5167 bp in length, but was found in this study to be 5090 bp, 77 bp shorter than expected. The tRNA tyrosine – *ND4* mitochondrial region was expected to be 5724 bp in length but was found in this study to be 5777bp, 53 bp longer than expected. The *ND4L* – *Cyt b* mitochondrial region was shorter than expected by 154 bp in *M. mordax* #25 and 164 bp in *M. praecox* #1.

Alignment of the three mitochondrial regions sequenced by 454 sequencing in this study to their respective homologous regions in the *M. mordax* mitogenome of Haouchar (2009) resulted in only 96.1 – 98.0 % similarity. Between 102 (Table 14) – 217 (Table 16) nucleotide differences were observed in *M. mordax* (Haouchar 2009).

Table 17. Expected versus observed size of the three mitochondrial regions determined in this study by 454 sequencing.

	Mitochondrial region size (bp)		
	12S – tRNA cysteine	tRNA tyrosine – *ND4*	*ND4L- Cyt b*
Expected *	5167	5724	5567
Observed	5090	5777	5403 and 5413
Difference	- 77	53	-164 and -154

* based on the *M. mordax* mitogenome sequence of Haouchar (2009)

Negative difference value indicates shorter than expected; positive value indicates longer than expected

Consistent with the findings in Section 3.2.2, most nucleotide differences were in regions that had been manually edited by Haouchar, suggesting that the observed differences were due to errors in the mitogenome of *M. mordax* (Haouchar 2009). Furthermore, the mitogenome of *M. mordax* (Haouchar 2009) contained two sizable features that were never observed in this study: a 150 bp noncoding region (referred to as 'NCIII' in the Haouchar (2009) study) located between NCII and *Cyt b* of which the first 104 bp corresponds to a direct copy of the 5'-end of *Cyt b*; and a 41 bp insertion between *ND1* and tRNA isoleucine.

3.5 Ion Torrent Sequencing

Ion Torrent sequencing became available midway through this study. When compared to 454 sequencing, Ion Torrent has approximately 30 fold greater read depth and 10 fold greater throughput (Section 1.7; Table 2.). This enables the complete mitogenome of a single individual to be determined with sufficient depth of coverage directly from DNA extracts – bypassing the need for long range PCR.

The complete mitogenomes of *M. praecox* and *M. mordax* were successfully determined by Ion Torrent sequencing of the total nucleic extract of *M. praecox* #3 and *M. mordax* #25 (Section 2.7.3). Using CLC Genomics Workbench v7.3, the *M. praecox* mitogenome assembly first involved a *de novo* assembly of the approximate 4.7 million individual Ion Torrent sequencing reads that were generated by two 200 bp chemistry sequencing libraries. The largest contiguous sequence (contig) was 17249 bp and was used to reassemble the approximately 5900 mappable sequencing reads. This resulted in a 17249 bp contiguous sequence with just over 76 times average coverage depth per nucleotide (minimum of 13 reads per nucleotide). The same iterative *de novo* assembly method was used to obtain the *M. mordax* mitogenome but involved approximately 1.8 million individual reads, generated by a single 400 bp chemistry sequencing library, of which approximately 6600 reads were mappable. This resulted in a 17094 bp contiguous sequence with just over 122 times average coverage depth (minimum 12 reads per nucleotide).

3.5.1 Accuracy of the *de novo* Mitogenome of *M. praecox* #3

Ion Torrent sequencing of total nucleic acid results in both the mitogenome and the nuclear genome being sequenced. Therefore, the assembled mitogenome must be verified to ensure that it does not comprise of nuclear genomic sequences such as nuclear mitochondrial pseudogenes (numts). The accuracy and verification of the *de novo* mitogenome of *M. praecox* #3 was checked via alignment of the Sanger determined *Cyt b – 16S* of *M. praecox* #3 (Section 3.2) and the near-complete mitogenome of *M. praecox* #1 determined by 454 (Section 3.4).

Comparison of the *de novo* determined *Cyt b – 16S* mitochondrial sequence and Sanger determined *Cyt b – 16S* mitochondrial sequence of the same individual revealed only 2 nucleotide differences. In the *de novo* sequence, these two differences were due to single nucleotide under-calls for two homopolymers: 4C instead of 5C in NCIII; and 2G instead of 3G in *16S* gene. By checking the individual Ion Torrent reads, the NCIII homopolymer error was identified to be due to low read depth (25x) in which the ratio of 4C : 5C reads was approximately 2 : 1. The *16S* homopolymer error had higher read depth (98x), however the ratio of 2G : 3G reads was approximately 60 : 40 - the same approximate ratio observed for other homopolymer errors reported in this study (Section 3.4.2) and is supported by the fact that the *16S* gene of all five lampreys sequenced by 454 sequencing contained 3G’s (Section 3.4.1). Once these two errors were fixed, the final *M. praecox* #3 mitogenome was 17251 bp.

When the error-corrected mitogenome of *M. praecox* #3 (17251 bp) was compared to the near-complete mitogenome of *M. praecox* #1 (16071 bp), the two mitogenome sequences were nearly identical (>99.9%). Only three differences were observed in the *de novo* sequence, two of which were base changes (1 transversion and 1 transition), and were located in NCI. The third difference was due to the C homopolymer (reported in Section 3.4.3) of *M. praecox* #3 being 1 base longer (11C) than *M. praecox* #1 (10C). When the individual Ion Torrent sequencing reads of *M. praecox* #3 were checked for correct base calling, the homopolymer had 68 x depth coverage of which the ratio of 10C : 11C reads were approximately 20 : 50. Given that the homopolymer in *M. praecox* #3 was the same length (11C) as that reported in *M. mordax* #25 (Section 3.4.3), this finding suggests that the single base difference in *M. praecox* #1 (Section 3.4.3) is due to individual variation.

3.5.2 Accuracy of the *de novo* Mitogenome of *M. mordax* #25

The accuracy and verification of the *de novo M. mordax* #25 mitogenome (17094 bp) was checked by alignment to the complete Sanger/454 consensus mitogenome of *M. mordax* #25 (17091 bp). The two mitogenome sequences were nearly identical (>99.9%) with only four differences observed. In the *de novo* sequence, three of the four differences were due to single base differences in homopolymer lengths and were located in the gene *ND5* and noncoding regions NCI and NCIII. The fourth difference was a two base insert in the *de novo* sequence relative to the Sanger/454 consensus. The single base over-call of the adenosine homopolymer in *ND5* of the *de novo* sequence (8A) relative to the Sanger/454 consensus (7A) was identified to be an error as it caused a shift in the reading frame that resulted in premature stop codons throughout the gene (like that seen in Figure 26). Furthermore, when the individual Ion Torrent sequencing reads of the *de novo* sequence were checked for correct base calling, the homopolymer had a minimum read depth of 80 reads in which the ratio of 8A : 7A reads was approximately 60 : 40 – the common ratio observed for homopolymer errors in this study. The NCI cytosine homopolymer was one base longer in the *de novo* sequence (12C) than determined by long range PCR for the same individual (Section 3.4.3).

However, this was identified to be an error associated with the 2.5 fold lower Ion Torrent coverage depth (60 x) relative to the 454 sequencing reads (150 x). The two observed differences in NCIII were due to a 2 base insert (underlined) in a small cytosine homopolymer tract (GCTCCCT) in the *de novo* sequence relative to the Sanger/454 consensus (GCCCT), and a 1 base under-call of a large cytosine homopolymer length (10C) relative to that determined by Sanger (11C). When the individual Ion Torrent sequencing reads were checked for incorrect base calling, the errors were identified to be due to low read depth (20 x). Once these four errors were fixed, the final *M. mordax* #25 mitogenome was 17091 bp.

3.6 Bioinformatics - the Mitogenomes of *M. mordax* and *M. praecox*

A total of 37 structural genes were identified in both *M. praecox* and *M. mordax* (Figure 28 and Table 18), comprising 13 protein-coding genes, 2 ribosomal RNAs (rRNAs) and 22 transfer RNAs (tRNAs). Equivalent genes in *M. mordax* and *M. praecox* were identical in size. The gene arrangement of the mitochondrion is compact with approximately 90% of the mitogenome encoding for structural genes in both species. Only 160 intergenic nucleotides (IGN) are interspersed between the genes; the largest (74 bases) being located between tRNAs serine (TCN) and aspartic acid (Table 18; IGN). Despite being of a similar size to lamprey tRNAs, this 74 base sequence was not identified as a tRNA (including when BLASTn searched). Eight cases of gene overlap were observed (Table 18; negative IGN values), seven of which have been previously reported in other lampreys (Lee & Kocher 1995; Delarbre *et al.* 2000; Hwang *et al.* 2013a, b). The single base pair overlap between tRNAs isoleucine and glutamine is currently unique to *M. mordax* and *M. praecox*.

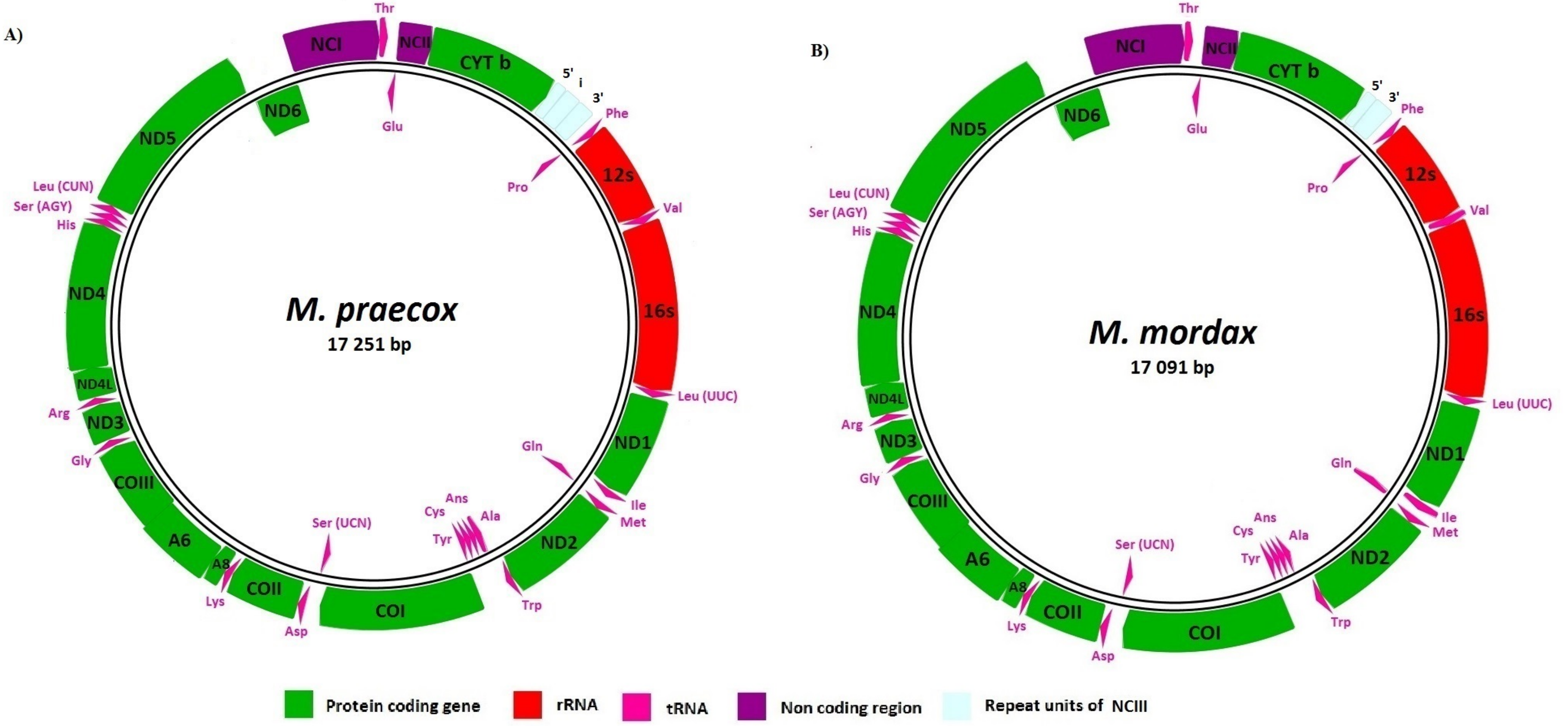


Figure 28. The complete annotated circular mitogenome of **A)** *M .praecox* individual #3, and **B)** *M. mordax* individual #25. Note: L-strand shown in 5' – 3' direction. Arrowheads indicate the direction of transcription. Genes located outside the circle are encoded by the heavy-strand. Serine and leucine tRNA genes are also identified by codon family (in parentheses). See Table 18 for exact gene locations and abbreviations.

Table 18. Location and coding strand of each gene in the mitochondrial genome of *M. praecox* (#3) and *M. mordax* (#25).

Strand	**Gene**	***M. praecox***		***M. mordax***		***M. praecox* and *M. mordax***			
		Location	Size*	Location	Size	Size	Codons		IGN
			Base Pairs		Base pairs	Amino Acids	Start	Stop	
Heavy	Cytochrome b	1 – 1218	1218	1 – 1218	1218	406	ATG	TAA	0
-	Noncoding region III	1219 – 1655	437	1219 – 1507	289				-
Light	tRNA – Proline	1656 – 1726	71	1508 – 1578	71				5
Heavy	tRNA – Phenylalanine	1732 – 1798	67	1584 – 1650	67				0
Heavy	12s rRNA	1799 – 2699	901	1651 – 2551	901				0
Heavy	tRNA – Valine	2700 – 2770	71	2552 – 2622	71				0
Heavy	16s rRNA	2771 – 4380	1610	2623 – 4232	1610				0
Heavy	tRNA – Leucine (UUR)	4381 – 4454	74	4233 – 4306	74				3
Heavy	ND1	4458 – 5417	960	4310 – 5269	960	320	ATG	TAA	12
Heavy	tRNA – Isoleucine	5430 – 5499	70	5282 – 5351	70				-1
Light	tRNA – Glutamine	5449 – 5569	71	5351 – 5421	71				1
Heavy	tRNA – Methionine	5571 – 5640	70	5423 – 5492	70				2
Heavy	ND2	5643 – 6692	1050	5495 – 6544	1050	350	ATG	TAA	-1
Heavy	tRNA – Tryptophan	6692 – 6758	67	6544 – 6610	67				2
Light	tRNA – Alanine	6761 – 6830	70	6613 – 6682	70				1
Light	tRNA– Asparagine	6832 – 6899	68	6684 – 6751	68				0
Light	tRNA – Cysteine	6900 – 6964	65	6752 – 6816	65				3
Light	tRNA – Tyrosine	6968 – 7039	72	6820 – 6891	72				1
Heavy	COI	7041 – 8591	1551	6893 – 8443	1551	517	GTG	AGA	-5
Light	tRNA – Serine (UCN)	8587 – 8657	71	8439 – 8509	71				74
Heavy	tRNA – Aspartic Acid	8732 – 8880	69	8584 – 8652	69				0
Heavy	COII	8801 – 9490	690	8653 – 9342	690	230	ATG	TAA	6
Heavy	tRNA – Lysine	9497 – 9562	66	9349 – 9414	66				1
Heavy	ATP8	9564 – 9731	168	9416 – 9583	168	56	ATG	TAG	-10
Heavy	ATP6	9722 – 10435	714	9574 – 10287	714	238	ATG	AGA	-35
Heavy	CO III	10401 – 11186	786	10253 – 11038	786	262	ATG	TAA	7
Heavy	tRNA – Glycine	11194 – 11262	69	11046 – 11114	69				1
Heavy	ND3	11264 – 11614	351	11116 – 11466	351	117	ATA	TAA	7

Table 18 continued.

Heavy	tRNA – Arginine	11622 – 11687	66	11474 – 11539	66				1
Heavy	ND4L	11689 – 11979	291	11541 – 11831	291	97	ATG	TAA	-7
Heavy	ND4	11973 – 13352	1380	11825 – 13204	1380	460	ATG	TAA	0
Heavy	tRNA – Histamine	13353 – 13421	69	13205 – 13273	69				-1
Heavy	tRNA – Serine (AGY)	13421 – 13490	70	13273 – 13342	70				0
Heavy	tRNA – Leucine (CUN)	13491 – 13561	71	13343 – 13413	71				0
Heavy	ND5	13562 – 15352	1791	13414 – 15204	1791	597	ATG	AGA	-16
Light	ND6	15337 – 15858	522	15189 – 15710	522	174	ATG	AGA	0
-	Noncoding region I	15859 – 16753	895	15711 – 16604	894				-
Heavy	tRNA – Threonine	16754 – 16826	73	16605 – 16677	73				33
Light	tRNA – Glutamic acid	16860 – 16930	71	16711 – 16781	71				0
-	Noncoding region II	16931 – 17251	321	16782 – 17091	310				-

Note: protein coding genes are highlighted in blue. Serine and leucine tRNA genes are also identified by codon family (in parentheses).

Abbreviations: ATP 6 and 8 = ATP synthase subunits 6 and 8; COI, II, and III = Cytochrome C oxidase subunits I, II, and III; Cyt b = Cytochrome b; ND1–6 = NADH dehydrogenase subunits 1–6; * indicates the size includes stop codons; dash (-) indicates not applicable; IGN is the number of intergenic nucleotides separating two adjacent genes with negative values indicating the number of nucleotides that overlap.

3.6.1 Protein Coding Genes

The protein coding genes of *M. mordax* and *M. praecox* are highly conserved and nearly identical at both the nucleotide and amino acid level (Table 19). At the nucleotide level, the two species differ by 24 nucleotide substitutions (22 transitions and 2 transversions), of which pyrimidine transitions (T ⇆ C substitutions) were found to be more frequent (54.5%) than purine transitions (A ⇆ G substitutions). At the amino acid level, only five out of 24 nucleotide substitutions resulted in an amino acid change (nonsynonomous or non silent substitution) and comprised of three pyrimidine (60%) and two purine transitions (40%). In *M. praecox*, the three pyrimidine transitions resulted in nonsynonymous substitutions of Threonine → Methionine that were found to occur twice in *ND5* and once in *COIII*, whilst the two purine transitions resulted in amino acid changes of Glycine→ Glutamic acid in *COII* and Alanine→Valine in *ND6*. All transversions (two in total) were synonymous substitutions occurring in *COI* (T ⇆ A) and *ND4* (C ⇆ G).

Table 19. Comparison between mitochondrial protein coding gene sequences of *M. mordax* and *M. praecox*

			No. of substitutions			
			Transitions (non silent)		Transversions (non silent)	
Protein Coding Gene	**% Identity** Nucleotides	Amino acids	Purine	Pyrimidine		**Total**
Cyt b	99.7	100	1	3	0	**4**
ND1	100	100	0	0	0	**0**
ND2	99.7	100	2	1	0	**3**
COI	99.7	100	2	2	1	**5**
COII	99.7	99.6	1 (1)	1	0	**2**
ATP8	100	100	0	0	0	**0**
ATP6	100	100	0	0	0	**0**
COIII	99.7	99.6	1	1 (1)	0	**2**
ND3	100	100	0	0	0	**0**
ND4L	99.7	100	1	0	0	**1**
ND4	99.9	100	0	1	1	**2**
ND5	99.8	99.7	1	3 (2)	0	**4**
ND6	99.8	99.4	1 (1)	0	0	**1**
			Combined Total No. of Substitutions			
			10 (2)	12 (3)	2 (0)	**24**

The two species also use the same initiation and termination codons (Table 18). *Mordacia mordax* and *M. praecox* are the only lampreys sequenced to date to use start codon ATA for *ND3* instead of ATG; and stop codons: TAA for *Cyt b* and *ND2* instead of AGA and TAG, respectively. Together with *G. australis*, these three Southern Hemisphere lampreys are currently the only lampreys known to date to use TAA instead of AGA as the termination codon for *ND4*.

When comparing complete lamprey mitochondrial genomes, there is no consensus as to the position of the stop codon for genes *ATP6* and *COI*. In *P. marinus,* Lee and Kocher (1995) found that *ATP6* overlapped *COIII* gene by 35 nucleotides (11 amino acids). Through cloning experiments, Delarbre *et al.* (2000) found that the mRNA coded by the *ATP6* gene in *L. fluviatilis* contained an incomplete stop codon (T--) that was polyadenylated immediately before the initiation codon of the *COIII* gene and thus the two genes did not overlap. The group suggested that the same process was used in *P. marinus*. On this basis, Milton (2003) and Riddington (2007) reported no overlap in *G. australis*, but Hwang *et al.*(2013a,b) reported a 35 nucleotide overlap in both *L. camtschaticum* and *L. reissneri*. Given that no cloning experiments have been done in this study nor any previous study involving *M. mordax* (Milton 2003; Riddington 2007, Haouchar 2010) , the identified *ATP6* stop codon AGA in *M. mordax* and *M. praecox* results in a 35 nucleotide overlap between the two genes as previously reported (Lee & Kocher 1995; Hwang *et al.* 2013a, b)

For *COI*, which overlaps tRNA Serine (TCN) by 10 nucleotides in both *P. marinus* and *L. fluviatilis*, Hwang *et al*. (2013a,b) report that the *COI* gene in *L. camtschaticum* and *L. reissneri* has an incomplete stop codon (T--) immediately prior to the tRNA and therefore no overlap occurs. Given that no cloning experiment was carried out in this study, the AGA stop codon in the *COI* gene of *M. mordax* and *M. praecox* results in a 5 nucleotide overlap between the two genes.

Interestingly, if the finding of Delarbre *et al*. (2000) or Hwang *et al.* (2013) is proven to be utilised by all lampreys, then, with the exception of *COI* for *M. mordax* and *M. praecox*, the gene size of *ATP6* and *COI* is highly conserved among all lampreys sequenced to date.

3.6.2 RNAs

The rRNA genes *12S* and *16S* are also highly conserved between *M. mordax* and *M. praecox* with *12S* being identical and *16S* only differing by a single pyrimidine transition. All tRNAs were identical except isoleucine and arginine. These two tRNAs contained single transition substitutions which are predicted to be in the variable loop of isoleucine and TC loop of arginine. The location of the 22 tRNAs are the same as in the other lamprey species sequenced to date and are similar in size. A noteworthy finding yet to be reported in lampreys is that tRNA leucine$^{(\mathrm{UUR})}$ contains a putative terminator sequence for mitochondrial transcription termination factor (mTERF). The tridecamer sequence in *M. mordax* and *M. praecox* (TGGCAGAGCCCAG) contains only a two nucleotide difference (underlined) relative to that determined in human mitochondria (TGGCAGACCCCGG) (Christianson & Clayton 1988; Hyvärinen *et al.* 2007). The putative consensus mTERF terminator sequence for lampreys is TGGCAGAGYTCAG, where Y is IUPAC code for C or T.

3.6.3 Noncoding Regions

As reported in all lamprey species sequenced to date, the putative control region (or displacement loop) for *M. mordax* and *M. praecox* is located between *ND6* and *Cyt b* and is split into two parts (noncoding regions I and II) by tRNAs threonine and glutamine (Figure 28). The lengths of noncoding region I (NCI) and noncoding region II (NCII) are quite different between the five sequenced lamprey species (Table 20). The differences in length primarily result from the presence of tandem repeats within the 5’-end of NC and 3’-end of NCII, in which the tandem repeat length and copy number are variable (Table 20). For example, the NCI of Northern Hemisphere species is comprised of a 39 base repeat with the consensus sequence TATGCCTMTATGGCAT AGGTATATMTAATGRCATAGGT (where M is IUPAC code for A or C and R is for G or A) and is repeated at least three times, whereas *G. australis* has a 289 base sequence that is repeated twice (Table 20). However, despite *M. mordax* and *M. praecox* containing no tandem repeats in the 5’-end of NCI, *M. praecox* has the largest NCI (895 bp) currently sequenced and is one base larger than that of *M. mordax* (Table 20). Furthermore, the NCII of *M. mordax* and *M. praecox* is larger than most lampreys except *G. australis*, which has the largest NCII sequenced (Table 20).

Table 20. Size comparison of mitogenome, NCI and NCII of 7 complete lamprey mitogenomes.

Species	Mitogenome size (bp)	NCI (bp)	Repeat (total bp)	NCII (bp)	Reference
Mordacia mordax	17091	894	0	310	This study
Mordacia praecox	17251	895	0	321	This study
Geotria australis	17059	881	2 x 289 (578)	605	Unpublished*
Lampetra fluviatilis	16159	491	3 x 39 (117)	151	Delarbre *et al*. 2009
Lethenteron camtshaticum	16277	518	3 x 39 (117)	239	Hwang *et al*. 2013b
Lethenteron reissneri	16207	479	4 x 39 (156)	209	Hwang *et al.* 2013a
Petromyzon marinus	16201	491	3 x 39 (117)	199	Lee *et al*. 1995

* Honours theses- Milton (2003) and Riddington (2007)

Conserved sequence blocks (CSB) II and III, sequences normally associated with the replication of the heavy-strand (Pham *et al.* 2006; Scarpulla 2008), were identified in

NCI as previously reported (Lee & Kocher 1995). The 21 base CSBII was identical to that reported in *P. marinus* and located at position 16510 – 16530 in *M. mordax* and 16659 – 16679 in *M. praecox*. In both *M. mordax* and *M. praecox,* the 14 base CSBIII was identical and located 21 bases upstream of CSBII. The CSBIII sequence of TCGACAACCCCCTT in *M. mordax* and *M. praecox* contains two base changes (underlined) relative to the CSBIII sequence of CCGTCAACCCCCTT reported in *P. marinus* (Lee & Kocher 1995).

A unique finding in this study is the presence of a repetitive third noncoding region (NCIII) located between *Cyt b* and tRNA proline in both *M. mordax* and *M. praecox* (Figure 29). This region was first identified by the short range PCR and Sanger sequencing approach (Section 3.2) and later confirmed by Ion Torrent sequencing and by reanalysis in this study of the 454 sequencing data of Haouchar (2009). NCIII is now discussed in context of the two complete mitogenomes, but see Section 3.2.4 for inter and intraspecies variation. In *M. mordax*, NCIII is 289 bp in length and contains two tandem repeats (Figure 29B): a 177 bp repeat and a 112 bp near identical partial repeat that differed by a three base pair insertion (Figure 30). In contrast, the NCIII of *M. praecox* is 437 bp in length and contains three tandem repeats (Figure 29C): two identical 164 bp repeats, and a 109 bp near identical partial repeat that differed by a single base pair change (Figure 31). Noncoding region III of *M. mordax* is 59.2% identical to that of *M. praecox* (Figure 32). However, the tandem repeats that comprise NCIII are highly conserved. Firstly, the 5'-partial repeat (112 bp) is 97.3% identical (109 identical sites) to the 109 bp partial repeat in *M. praecox – differing* only by the insertion of three cytosine homopolymers. Lastly, the 3'- repeat in *M. mordax* (177bp) is 89.8% identical (159 identical sites) to the 164 bp sequence in *M. praecox* – differing by 13 insertions and five base changes (Figure 32).

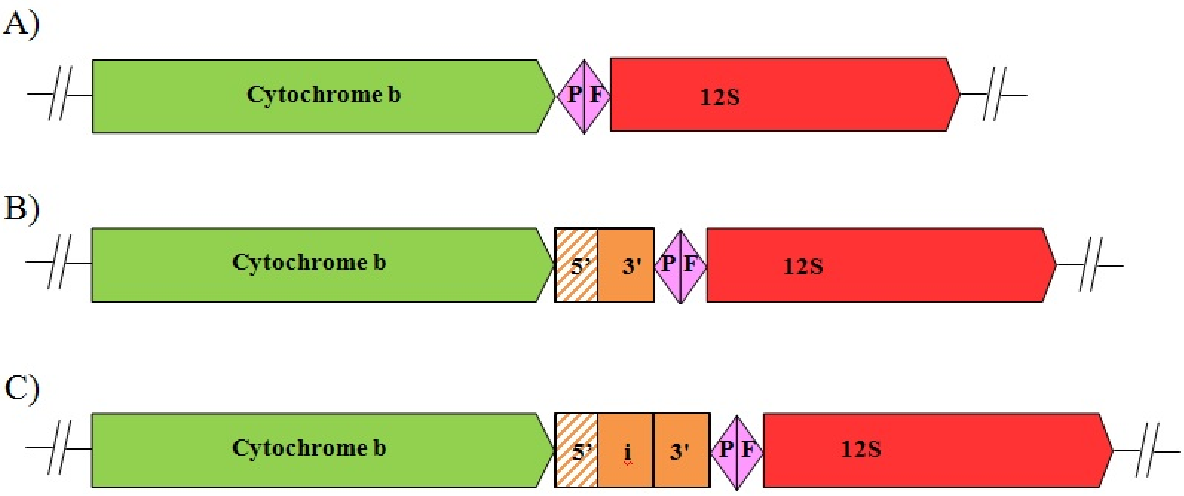


Figure 29. Schematic comparison of the absence and presence of NCIII in the mitochondrial genome of lampreys sequenced to date. Noncoding region III (represented by the orange boxes) is located between genes Cytochrome b (green arrow box as annotated) and tRNA Proline (first pink arrow box as annotated). **A)** The gene arrangement and absence of NCIII in all lampreys published to date. **B)** The gene arrangement and location of NCIII as determined in this study for *M. mordax* #25. The 289 base pair NCIII of this individual contains two tandem repeats: a 177 bp repeat (annotated 3') and a 112 bp near identical partial repeat (annotated 5'). **C)** The gene arrangement and location of NCIII as determined in this study for *M. praecox* #3. The 437 base pair NCIII contains three tandem repeats: a 164 bp repeat (annotated 3'), a 164 bp identical repeat (annotated i) and a 109 bp near identical partial repeat (annotated 5'). Note: arrowheads indicate the direction of transcription for each gene. Proline and phenylalanine tRNA genes are abbreviated as P and F, respectively.

```
5’PR   --------------------------------------------------    0
3’CR   TCAAAGAAAGAGAATTAGAATTAGAATCTCTATTACTAGGCCCCCAAAGC   50

5’PR   ------------ATTATTACCATAAGTTAAAGTAGCTTAAGTTTAAAGCA   38
3’CR   TAGTATTTTTTAATTATTACCATAAGTTAAAGTAGCTTAAGTTTAAAGCA  100

5’PR   AAGCTCGAAGATGCTAGTATGATGGTCCAAACAAATGACCGCCCTTTAAC   88
3’CR   AAGCTCGAAGATGCTAGTATGATGGTCCAAACAAATGACCGCCCTTTAAC  150

5’PR   ATCAA---CCCCCCCCTCAATCAAACA  112
3’CR   ATCAACCCCCCCCCCCTCAATCAAACA  174
```

Figure 30. Nucleotide comparison of the two tandem repeat copies comprising the 289 bp NCIII of *M. mordax* individual #25. Bold and underlined is the 3 base pair insertion that differentiates the partial copy from the copied sequence. Note: L-strand (5’ – 3’) is shown for simplicity. 5' and 3’ refer to the order in which the repeat copies occur on the L-strand, and PR= partial repeat and CR= complete repeat.

```
5’PR  --------------------------------------------------    0
iCR   TCAAAGAAAGAGAATTAGAATCTCTATTACTAGCCCCCAACAACAGTATT   50
3’CR  TCAAAGAAAGAGAATTAGAATCTCTATTACTAGCCCCCAACAACAGTATT   50


5’PR  -----ATTATTACCATAAGTTAAAGTAGCTTAAGTTTAAAGCAAAGCTCG   45
iCR   TTTTAATTATTACCATAAGTTAAAGTAGCTTAAGTTTAAAGCAAAGCTCG  100
3’CR  TTTTAATTATTACCATAAGTTAAAGTAGCTTAAGTTTAAAGCAAAGCTCG  100


5’PR  AAGATGCTAGTATGATGGTCCAAACAAATGACCGCCCTTTAACATCAACC   95
iCR   AAGATGCTAGTATGATGGTCCAAACAAATGACCGCCCTTTAACATCAACC  150
3’CR  AAGATGCTAGTATGATGGTCCAAACAAATGACCGCCCTTTAACATCAGCC  150


5’PR  CCCTCAATCAAACA   109
iCR   CCCTCAATCAAACA   164
3’PR  CCCTCAATCAAACA   164
```

Figure 31. Nucleotide comparison of the three tandem repeats comprising the 437 bp NCIII of *M. praecox* #3. Bold and underlined is the single purine transition change. Note: L-strand (5’ – 3’) is shown for simplicity. 5' , i , and 3' refer to the order in which the repeat copies occur on the L-strand, and PR= partial repeat and CR= complete repeat.

```
M. mordax (#25) ATTATTACCATAAGTTAAAGTAGCTTAAGTTTAAAGCAAAGCTCGAAGATGCTAGTATGATGGTCCAAACAAATGACC     78
M. praecox (#3) ATTATTACCATAAGTTAAAGTAGCTTAAGTTTAAAGCAAAGCTCGAAGATGCTAGTATGATGGTCCAAACAAATGACC     78
                |- 5’PR

M. mordax (#25) GCCCTTTAACATCAACCCCCCCCTCAATCAAACA-----------------------------------------        112
M. praecox (#3) GCCCTTTAACATCAACCCCC---TCAATCAAACATCAAAGAAAGAGAATTAGAATCTCTATTACTAGCCCCCAAC        150
                                           5’PR -||- iCR

M. mordax (#25) ---------------------------------------------------------------------------        112
M. praecox (#3) AACAGTATTTTTTAATTATTACCATAAGTTAAAGTAGCTTAAGTTTAAAGCAAAGCTCGAAGATGCTAGTATGAT        225

M. mordax (#25) ------------------------------------------------TCAAAGAAAGAGAATTAGAATTAGAAT        139
M. praecox (#3) GGTCCAAACAAATGACCGCCCTTTAACATCAACCCCCTCAATCAAACATCAAAGAAAGAGAATTAGAA------T        294
                                                          iCR -||- 3’CR

M. mordax (#25) CTCTATTACTAGGCCCCCAAAGCTAGTATTTTTTAATTATTACCATAAGTTAAAGTAGCTTAAGTTTAAAGCAAA        214
M. praecox (#3) CTCTATTACTA-GCCCCCAACAACAGTATTTTTTAATTATTACCATAAGTTAAAGTAGCTTAAGTTTAAAGCAAA        368

M. mordax (#25) GCTCGAAGATGCTAGTATGATGGTCCAAACAAATGACCGCCCTTTAACATCAACCCCCCCCCCCTCAATCAAACA        289
M. praecox (#3) GCTCGAAGATGCTAGTATGATGGTCCAAACAAATGACCGCCCTTTAACATCAGCCCCC------TCAATCAAACA        437
                                                                                     3’CR-|
```

Figure 32. Nucleotide alignment of NCIII of *M. mordax* (#25) and *M. praecox* (#3). Whilst the overall region is only 59.2% identical between the two individuals, the 5’-partial repeat (annotated 5’PR) is 97.3% identical and the 3’-complete repeat (3’CR) is 89.8% identical. Bold and underlined are the differences between the nucleotide sequences. Dots indicate identical bases while dashes indicate deletions. The repeat copies are annotated (below the sequence) 5', i (= internal) and 3' , referring to the order in which copies occur on the L-strand, and PR = partial repeat and CR =complete repeat. Note: L-strand sequence (5’ – 3’) is shown for simplicity.

3.7 Lamprey Phylogenetics

To examine the relationship among the three lamprey families, phylogenetic analyses were carried out using five complete lamprey mitogenomes together with the two newly sequenced mitogenomes (this study) of *M. mordax* and *M. praecox* (Section 2.9; Table 6). Two hagfishes, *Myxine glutinosa* and *Eptatretus burgeri*, served as the outgroup and are the only two complete hagfish mitogenomes available on GenBank.

3.7.1 Analyses of Concatenated Protein Coding Genes

The Bayesian Inference (BI) and Neighbor-Joining (NJ) phylogenetic analyses of the concatenated mitochondrial protein coding genes, analysed as nucleotides (without third codon) and amino acids, resulted in four trees with identical topologies (Figure 33). All nodes had maximum Bayesian posterior probability and NJ bootstrap support (100%) except for a single node (99%) in Figure 33. In all four trees, the seven lamprey species were unequivocally recovered as a monophyletic clade. *Mordacia mordax* and *M. praecox* formed a monophyletic group (Mordaciidae) that was basal to the monophyletic clade comprised of the Southern Hemisphere *G. australis* and the remaining four Northern Hemisphere species (Petromyzontidae). The placement of *G. australis* basal to the Northern Hemisphere species indicates that Geotriidae and Petromyzontidae have a sister-group relationship and that they are in turn sister to Mordaciidae. In other words, Geotriidae shared a common ancestor with the Petromyzontidae more recently than Mordaciidae.

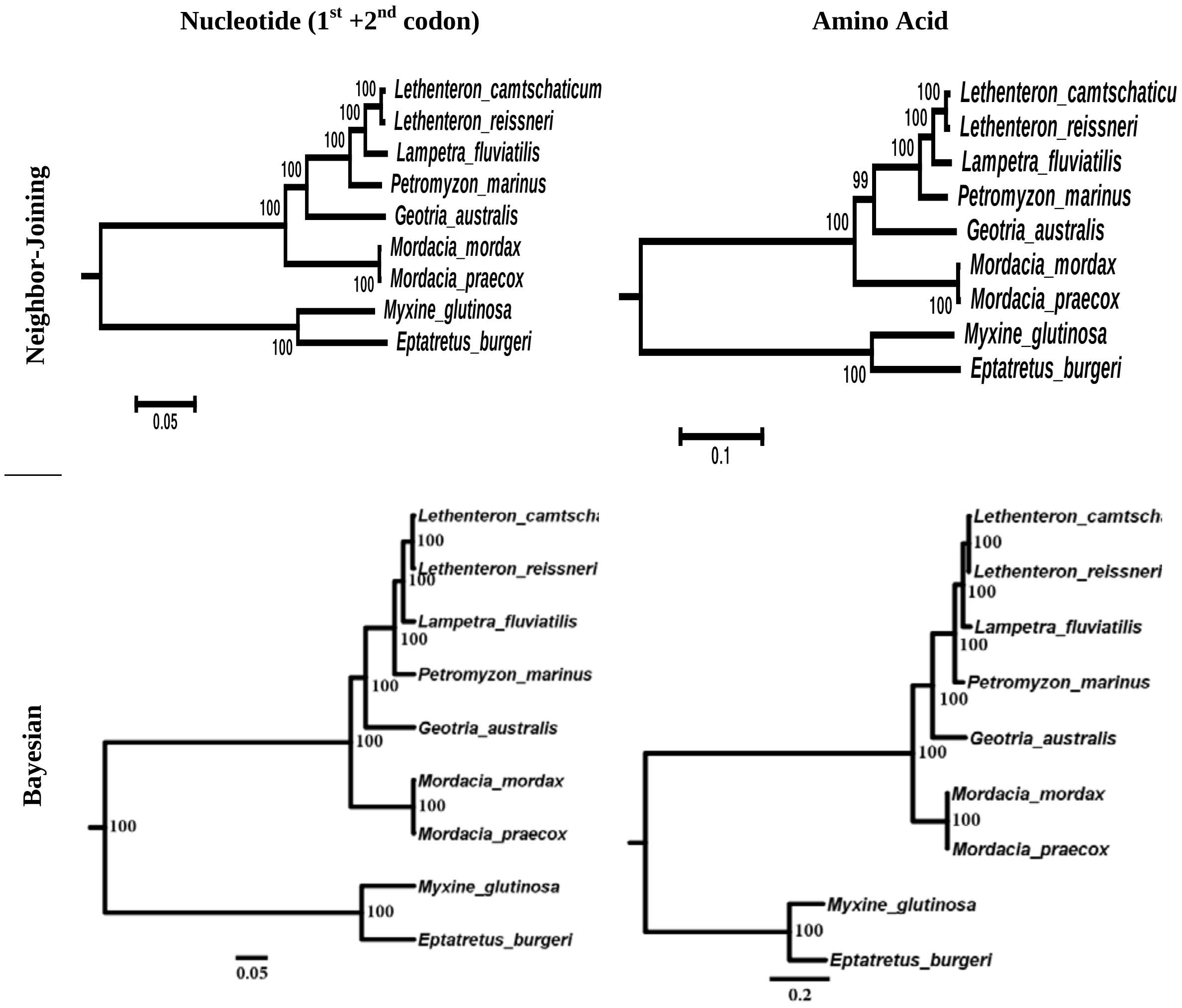


Figure 33. Phylogenetic relationships among the lamprey families based on concatenated molecular data set comprising 13 mitochondrial protein coding genes. The data set was analysed as nucleotides without third codon (left column) and as amino acid sequence (right column), and inferred using NJ (top row) and BI (bottom row).

3.7.2 Analyses of Individual Protein Coding Genes

Bayesian Inference and NJ phylogenetic analyses of the 13 individual mitochondrial protein coding genes (with and without third codon and as amino acids) resulted in 78 trees (Appendix 6). Three different resolved and several unresolved topologies regarding the interrelationships of Geotriidae, Mordaciidae and Petromyzontidae were recovered. The three resolved tree topologies (generalised) are shown in Figure 34. The number of trees supporting each of the three topologies is summarised in Table 21.

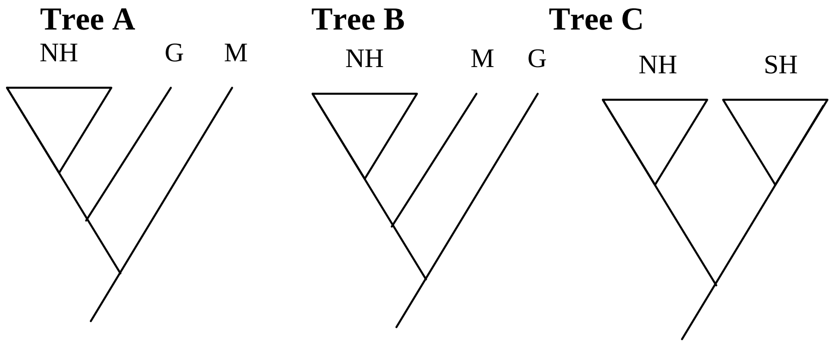


Figure 34. The three recovered hypotheses regarding the interrelationships of Geotriidae, Mordaciidae and Petromyzontidae. **Tree A**: hypothesis 1 suggests that Petromyzontidae (i.e. Northern Hemisphere species [NH]) + Geotriidae (G) are a clade whose sister group is Mordaciidae (M). **Tree B:** hypothesis 2 suggests that (Northern Hemisphere species + Mordaciidae) are a clade whose sister group is Geotriidae. **Tree C:** hypothesis 3 suggests that that Northern Hemisphere species are a clade whose sister group is the Southern Hemisphere species (SH).

Table 21. Number of phylogenetic trees supporting each tree topology in Figure 34.

	Topology			
Dataset	**Tree A**	**Tree B**	**Tree C**	**Unresolved**
Nuc $^{\text{all pos}}$	13	4	6	3
Nuc$^{\text{1st+2nd}}$	16	2	3	5
AA	18	1	5	2
Total	**47**	**7**	**14**	**10**

Like the concatenated gene dataset trees in Figure 33, most individual gene trees (47; 60%) supported the placement of *G. australis* basal to the Northern Hemisphere species (i.e Tree A). Of the remaining trees, 7 trees (9%) supported the placement of *G.*

australis as the most distal lamprey (i.e. Tree B), 14 trees (18%) supported the monophyly of Southern Hemisphere lampreys (i.e. Tree C) and 10 trees (13%) were unresolved. With the notable exception of *ND1*, trees supporting Tree B often had weak (>70% Bayesian posterior probability) nodal support for the monophyly of Geotriidae and Mordaciidae. Using the two phylogenetic methods (Bayesian Inference and NJ), only genes *COI*, *Cyt b* and *ND2* recovered the same topology (Tree A) in all 3 datasets (Table 22). The 'worst' performing genes were *ND3*, *ND4L* and *ND6*. Most of the trees recovered by these four genes were unresolved due to polytomies at the basal lamprey node irrespective of the two methods used (Table 22 and Appendix 6).

Table 22. Summary of the phylogenetic results.

	Data set					
	Nuc - All		**Nuc- 1st + 2nd**		**Amino Acid**	
Gene	Mr Bayes	NJ Jukes	Mr Bayes	NJ Jukes	Mr Bayes	NJ Jukes
ATP6	Bx	A	**Polytomy***	A	A	A
ATP8	A	A	A	A	C	A
COI	A	A	A	A	A	A
CO2	C x	C x	A	A	A	A
CO3	C x	A	C x	A	A **+**	A
Cyt b	A	A	A **!**	A	A	A **&**
ND1	C	C x	**polytomy#**	C x	C	C
ND2	A	A	A	A	A	A
ND3	**polytomy^**	B	A **!**	A **&**	**polytomy^**	C x
ND4	B	A	A	A	A	A **@**
ND4L	**polytomy^**	A	**polytomy^**	**Unresolved**	A	B
ND5	B	B	B	B	A	A
ND6	**polytomy***	A	**Polytomy***	C x	**Polytomy***	C x
CONCAT	-	-	A	A	A	A

A, B, C = Tree A, Tree B, Tree C as in Figure 34.

^ = polytomy at basal node between *P. marinus*, *L. fluviatilis* (both Lethenteron) and (SH)

* = polytomy at basal node between (NH), (Mordaciidae) and *G. australis*

# = polytomy at basal node between (*L. fluviatilis* + (both Lethenteron)), *P. marinus* and (SH)

! = polytomy at node between *L. fluviatilis* and *P. marinus*

+ = polytomy at node between *L. reissneri*, *L. camtschaticum* and *L. fluviatilis*

& = *L. fluviatilis* basal relative to *P. marinus*

@ = (*L. fluviatilis* + *P .marinus*) form a clade

x = low support (<90 Bayesian probability)

NH= Northern Hemisphere; SH= Southern Hemisphere

3.7.3 Estimation of Lamprey Divergence Times

To examine the relationship among living agnathans and gnathostomes, as well as provide and estimation of their divergence times, a temporally-calibrated BI analysis was carried out by Dr Matthew Phillips (Australian National University) using a concatenated dataset comprising the 13 mitochondrial protein coding genes of 29 taxa (7 lampreys, 2 hagfishes and 20 deuterostomes; see Appendix 7) and 10 fossil priors (see Appendix 8). Lampreys and hagfishes were unequivocally recovered as a monophyletic clade ("Cyclostomata") with maximum nodal support (100%), sister-group to gnathostomes (Figure 35).

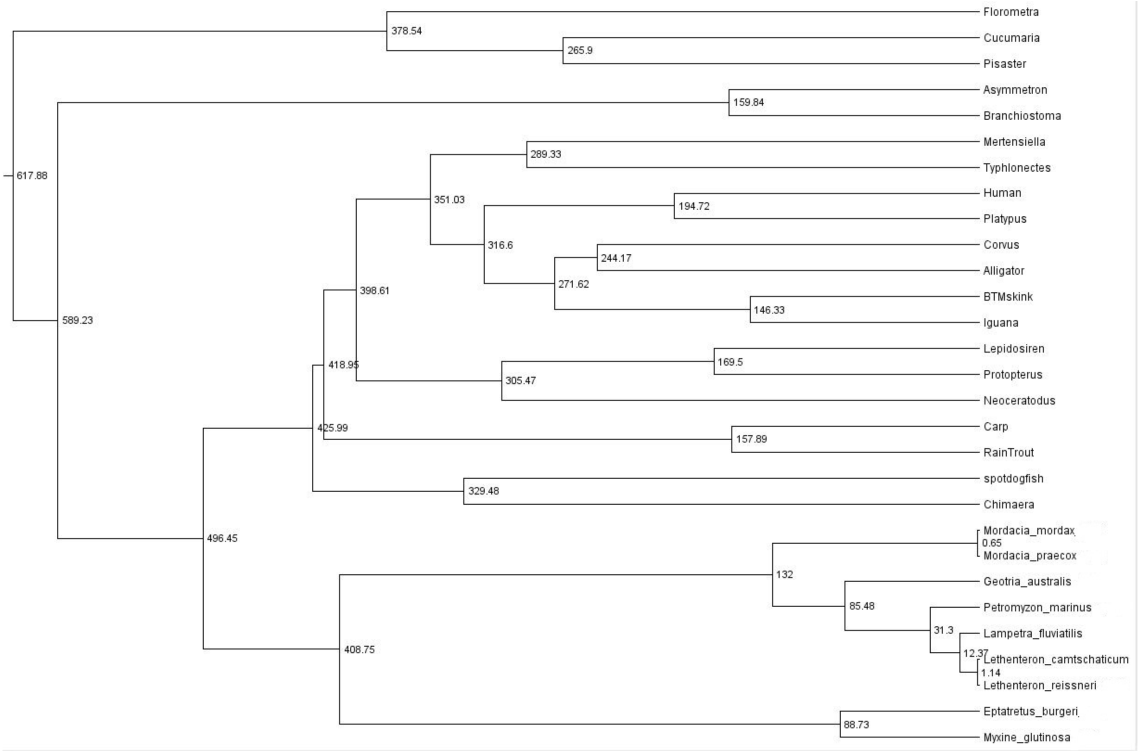


Figure 35. Bayesian estimation of lamprey divergence times based on concatenated dataset comprising the 13 mitochondrial protein coding genes of 29 taxa. Node values are MYA and nodal support is 100% unless otherwise stated in the text. Note: see Appendix 6 for list of taxa used and see Appendix 7 for list of fossil priors.

Consistent with the findings in Section 3.7.1 and 3.7.2, Geotriidae was recovered as a sister-group to Petromyzontidae (99% nodal support) and they are in turn sister to Mordaciidae (100% nodal support). It also hypothesises that: 1) lampreys diverged from hagfishes at 409 MYA (95% Highest Posterior Density [HPD] 360-466); 2) Mordaciidae from Geotriidae plus Petromyzontidae at 132 MYA (95% HPD 73-227); 3) Geotriidae from Petromyzontidae at 85 MYA (95% HPD 37-145); 4) *Lampetra* plus *Lethenteron* from *Petromyzon* at 31 MYA (95% HPD 14-56); 5) *Lampetra* from *Lethenteron* at 12 MYA (95% HPD 4-26); and 6) *M. mordax* from *M. praecox* at 0.65 MYA (95% HPD 0.02 – 2.3).

Discussion and Conclusion

4.1 Lamprey Phylogenetics and Divergence

The main finding of this thesis (Section 3.7) clearly indicates that Geotriidae and Petromyzontidae share a more recent common ancestry than either does with Mordaciidae. This inferred phylogenetic hypothesis is consistent with previous studies utilising amino acid composition of lactate dehydrogenase (Baldwin *et al.* 1988), Bayesian Inference of *Cyt b* (Lang *et al.* 2009), and previous honours studies involving Maximum Parsimony (Milton 2003) and Bayesian Inference (Haouchar 2009) of multiple individual mitochondrial protein coding genes.

Additionally, the phylogenetic results in this thesis support the hypothesis that cyclostomes are monophyletic. This inferred phylogenetic hypothesis is supported by recent morphological data (Oisi *et al.* 2013) and most molecular studies utilising nuclear rRNA gene sequences (Mallat & Sullivan 1998; Mallat *et al.* 2001; Mallat & Winchell 2007), nuclear protein coding gene sequences (Takezaki *et al.* 2003; Blair & Hedges 2005; Kuraku & Kuratani 2006), mitochondrial DNA sequences (Delarbre *et al.* 2002; Haouchar 2009), a combination of mitochondrial and nuclear gene sequences (Furlong & Holland 2002), and microRNA sequences (Heimberg *et al.* 2008).

Furthermore, the phylogenetic finding in this thesis are in concordance with the preliminary phylogenetic estimation of lamprey divergence times carried out by Matthew Phillips (Phillips *et al*, unpublished; Section 3.7.3), which infers that lampreys separated from hagfishes about 409 MYA, and that lamprey divergence involved the early radiation of Mordaciidae (about 132 MYA), followed by the monophyletic divergence of Geotriidae plus Petromyzontidae about 85 MYA.

4.2 Determining the Complete Mitogenomes

The use of Sanger sequencing has long been associated with mitochondrial gene characterisation. However, sequencing the complete mitogenome using this process is a laborious method often requiring dozens of short overlapping primer pairs (short range PCR). This study used a combination of Sanger sequencing and two NGS sequencing methods to re-sequence the complete mitogenome of *M. mordax* and, for the first time, determine the complete mitogenome of *M. praecox.* There was good consensus between sequencing results from the three sequencing platforms and very few differences were observed. These differences were mostly attributable to variant/miscalled homopolymer lengths of greater than 5 bases.

Errors associated with homopolymer length are inherent in both 454 and Ion Torrent sequencing due to the nonlinear increase in signal strength with increasing homopolymer length (Berglund *et al.* 2011; De Beuf *et al.* 2012; Vogl *et al.* 2012). This makes it difficult for the base-calling algorithm to interpret the correct number of nucleotides from the signal (‘flow’) values, which, is further complicated when signal values are rounded to the nearest integer (De Beuf *et al.* 2012; Golan & Medvedev 2013; Martin & Rahmann 2013). For example, if a 6C homopolymer resulted in a flow value of 5.45, 5.47, and 5.55 in three separate wells containing the same sequence, the base-caller would incorrectly under-call 5C in well 1 and 2 and correctly call 6C in well 3. This may explain why a read ratio of approximately 60:40 was commonly observed for homopolymer errors with greater than 60x read depth. A search of the literature about such a ratio could not be found. However, this highlights the importance of read depth in NGS to overcome such errors. For example, whilst accuracy of a read can be low (64%) for a 9 base homopolymer, the consensus basecall accuracy can be as high as 99% with over sampling (Leamon & Rothberg 2009).

In this study, the biggest advantage of the Ion Torrent sequencing method was the ability to obtain the complete mitogenome directly from total nucleic acid extracts with sufficient depth of coverage, such that ‘target enrichment’ by long range PCR was not required. Despite the 454 sequencing method providing greater sequencing depth relative to Ion Torrent, this benefit was negated by the unreliability of long range PCR in this study and the associated time, labour and cost that went into obtaining the few partial mitogenome sequences by this method. Indeed, if long range PCR had not been problematic, the 454 sequencing method would have provided the complete mitogenome of multiple individuals and this would have enabled a more in-depth mitochondrial analysis at both the intraspecific and interspecific level – a shortcoming of this study.

4.2.1 Numts

In any analysis involving mitochondrial DNA, numts, that is, mitochondrial DNA that has been incorporated into the nuclear genome (Triant & DeWoody 2007), must be considered as a possible source of contamination as they can often be coamplified with, or amplified instead of, the mitochondrial targets in PCR and can lead to erroneous findings (Zhang & Hewitt 1996; Bensasson *et al.* 2001; Triant & DeWoody 2007).

In this study, whilst numt contamination is a possibility given that mitochondrial DNA was sequenced from total nucleic acid and not from purified mitochondrial DNA, it is considered unlikely due to the following preventative measures undertaken: 1) DNA was extracted from liver and muscle tissue, both of which are rich in mitochondrial DNA relative to nuclear DNA (Triant & DeWoody 2007). 2) The PCR primers used in this study were target-specific and designed based on lamprey sequence, therefore, less likely to preferentially amplify numts relative to primers that are universal or degenerate

(Zhang & Hewitt 1996). 3) Long range PCR was carried out and is less likely to amplify numt DNA (Triant & DeWoody 2007). 4) Regarding NCIII, the *Cyt b* sequence flanking NCIII did not contain any premature stop codons or shifts in the reading frame - features that are often associated with non-functional sequences of nuclear origin due to lack of selection pressure (Zhang & Hewitt 1996; Bensasson *et al.* 2001; Triant & DeWoody 2007). 5) NCIII was confirmed twice independently, by Ion Torrent sequencing and Sanger sequencing, it is therefore highly unlikely that NCIII is a numt.

4.2.2 *Mordacia mordax* versus *M. praecox*

There has been much controversy over the taxonomic status of lamprey paired species regarding whether or not mode of life (parasitic vs. nonparasitic) in the adult constitutes a criterion for species differentiation (Renaud 2011; Mateus *et al.* 2012). This study is the first to enable complete mitogenome comparison between the Southern Hemisphere species pairing of *M. mordax* and *M. praecox*.

The mitogenome of the parasitic *M. mordax* is nearly identical to that of the nonparasitic *M. praecox.* The thirteen protein coding genes are at least 99.7% identical at the nucleotide level and at least 99.4% identical at the amino acid level. This finding is not unexpected as significant genetic differences in mitochondrial genes have not been found between paired lamprey species to date (Docker *et al.* 1999; Yamazaki *et al.* 2006; Espanhol *et al.* 2007; Blank *et al.* 2008; Boguski *et al.* 2012) . For example, even though allozyme analysis showed that the parasitic *L. japonicum* was genetically divergent from the nonparasitic *L. reissneri*, analysis of *Cyt b* and *ND3* genes failed to find fixed differences in mitochondrial sequence (Yamazaki *et al.* 2006) and the complete mitogenomes of this species pair (Hwang *et al.* 2013a, b) are nearly identical.

The lack of genetic difference between *M. mordax* and *M. praecox* in this study is also not unexpected given that individuals in this study were from a single location in which the two species occur sympatrically. Furthermore, breeding is thought to occur during the same time of the year (Aug - Oct) between the two species (Hughes & Potter 1968) and, not only are three sets of species pairs in the Northern Hemisphere known to produce viable offspring by artificial hybridization (Piavis *et al.* 1970; Beamish & Neville 1992), research has also shown that pre-zygotic barriers to gene flow in the form of strong assortative mating does not occur between sympatric populations of *L. fluviatilis* and *L. planeri* (Hume *et al.* 2013). In context, the lack of genetic difference between *M. mordax* and *M. praecox* suggests ongoing gene flow and supports the hypothesis that the two species are alternate life-history forms (ecotypes) of a single species. However, the hypothesis that the two species are closely related, reproductively isolated species that have undergone recent speciation cannot be ruled out. It should be noted that, whilst the preliminary phylogenetic estimation of lamprey divergence times infers that *M. mordax* and *M. praecox* separated about 0.65 MYA, the error bars are too wide (95% HPD 0.02 – 2.3 MYA) to be reliable.

Population studies of *M. praecox* and *M. mordax* need conducting to test the third hypothesis of multiple independent origins of the nonparasite from parasite followed by reproductive isolation. However, lamprey population studies to date have been unable to resolve the species pair debate due to one or more mitochondrial haplotypes being shared between two taxa of a species pair, and intraspecies and geographical mitochondrial variation being higher than interspecies variation, (Docker *et al.* 1999; Yamazaki *et al.* 2006; Espanhol *et al.* 2007; Blank *et al.* 2008; Boguski *et al.* 2012).

4.2.3 NCIII

Noncoding region III (NCIII) is a unique finding in this study and is not reported in any lamprey sequenced to date. Despite not being reported in the complete *M. mordax* mitogenome of Haouchar (2009), examination (this study) of the 454 sequencing data from the Haouchar (2009) study found seven overlapping sequencing reads that confirmed its presence (Appendix 5). These reads were probably missed due to the low 454 sequencing coverage in the Haouchar (2009) study and lack of a closely related reference mitogenome (use of *G. australis*) at the time.

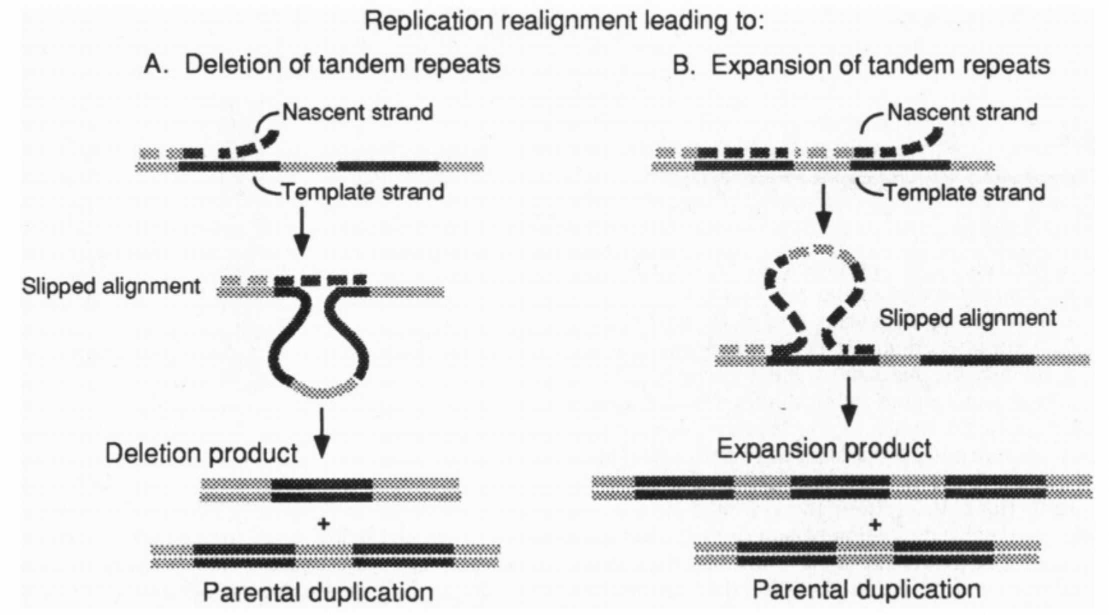


Figure 36. Replication slippage model for **A)** contraction, and **B)** expansion of tandem repeats. A slipped alignment of the nascent-strand with respect to template during mitochondrial replication can generate daughter mitochondria containing a deletion or expansion of a tandem repeat and any intervening DNA. From Lovett and Feschenko (1996).

Mordacia mordax and *M. praecox* individuals in this study share two mitochondrial polymorphisms in which NCIII contains either 2 tandem repeats (one full and one partial) or 3 tandem repeats (two full and one partial). This expansion (insertion) and contraction (deletion) of a single repeat can readily be explained by slipped-strand mispairing of either the heavy-strand (H-strand) or the light-strand (L-strand) during DNA replication, with slippage of the nascent-strand resulting in expansion whilst slippage of the template-strand results in contraction (Figure 36). In *M. mordax* and *M. praecox*, strand slippage is likely facilitated by the ability of tRNA proline and phenylalanine to form stable secondary structures.

The finding of tRNA proline and phenylalanine pseudogenes in each repeat unit of NCIII suggests that it originates from a tandem duplication of the tRNA proline – phenylalanine region. Transfer RNA pseudogenes have been reported from a number of animal mitogenomes such as frogs (Kurabayashi *et al.* 2008), salamanders (Mueller & Boore 2005), lizards (Moritz & Brown 1987), caecilians (San Mauro *et al.* 2006), snakes (Kumazawa *et al.* 1998) and fish (Mabuchi *et al.* 2004). In fact, NCIII in this study shares similarities to the tandem duplication reported in the reptile *Bipes biporus* in that: 1) it involves two adjacent tRNA genes (being tRNA threonine and proline in *Bipes biporus*); 2) the additional copies have undergone pseudogene formation; and 3) the tandemly duplicated pair of tRNA genes are separated by sequence that is also repeated (Macey *et al.* 1998). It is interesting to note that the pattern of pseudogene formation varies among populations of *Bipes biporus* (Macey *et al.* 2004). This is also possible for *M. mordax* and *M. praecox* populations given the current arrangement of NCIII. Furthermore, subsequent loss of repeated genes under the Tandem Duplication Random Loss model could enable a genomic rearrangement (San Mauro *et al.* 2006) in *M. mordax* and *M. praecox.* Depending upon which gene copies are eliminated, a new gene arrangement may arise (*Cyt b* – F – P – *12S*) or the original gene order may be restored (*Cyt b* – P – F – *12S*).

4.2.3.1 Hypothesis Regarding the Origin of NCIII

A likely hypothesis regarding the origin of NCIII is now explained based on preliminary mitogenome results of *M. lapicida* (Brigg and Berryman, unpublished), which is the earliest *Mordaciidae* lineage to evolve based on phylogenetic analysis of *Cyt b* (Haouchar 2009; Lang *et al.* 2009). In *M. lapicida,* the H-strand encoded *Cyt b* overlaps the L-strand encoded tRNA Pro by 18 nucleotides. Consequently, the coding sequence of *Cyt b* contains the template sequence corresponding to the 3'-terminus arm, the 3'-TC loop stem, and part of the TC loop of tRNA proline. Such an overlap is likely to have major functional implications on both genes, with any insertions/deletions within the 18nt overlap most likely being at the expense of the other gene. To overcome this burden, the tRNA proline – tRNA phenylalanine region was duplicated (Figure 37, page 99). Mutations in the original tRNA proline gene (that overlapped *Cyt b*) resulted in the increased size of *Cyt b* (1218 bp) seen in this study relative to other lampreys (1191 bp) and helps to explain the difficulty in aligning the 3'-end of *Cyt b* between the three lamprey families. Lastly, replication slippage enables the current arrangements found in this study (Figure 37).

4.3 Future Studies

Recently, *M. praecox* has reportedly been found in Queensland (Gill, H., pers. comm., 2014). Given that the distribution of *M. mordax* is confined to south-east Australia and Tasmania, the Queensland *M. praecox* population are likely allopatric to *M. mordax* and thus more likely to have fixed mitochondrial differences as a result of lineage sorting, genetic drift and local adaptation. Therefore, the complete mitogenome of the Queensland *M. praecox* should be determined and incorporated in any future *Mordacia* phylogenetic study.

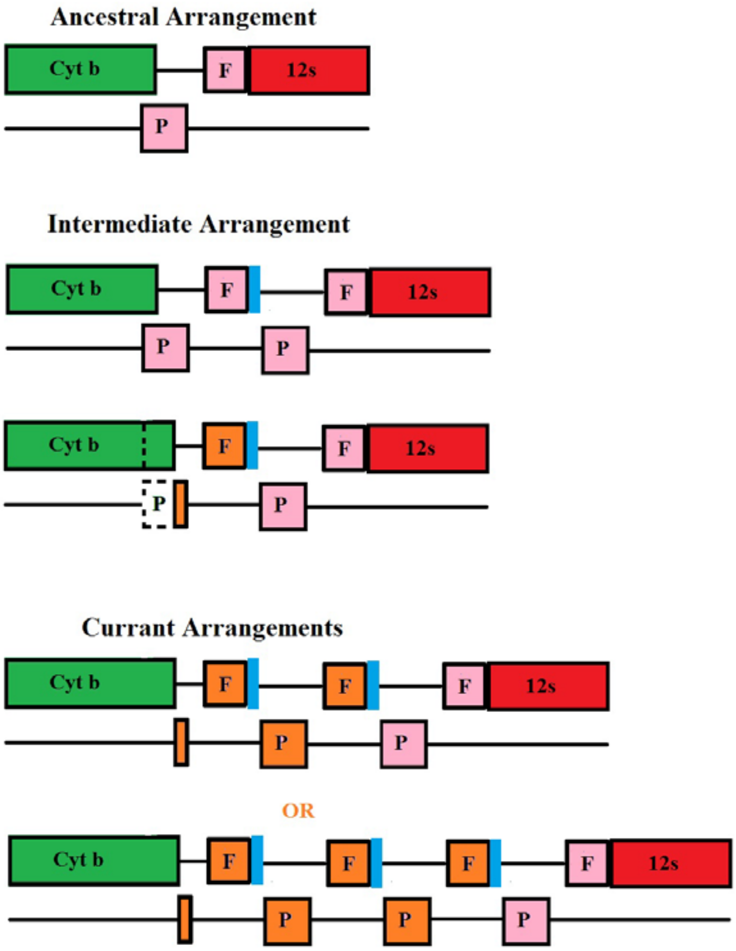


Figure 37. Ancestral arrangement: H-strand encoded *Cyt b* overlaps L-strand encoded tRNA proline; the template sequence of tRNA proline contains stop codon for *Cyt b*. Intermediate arrangement: Duplication of Pro-Phe region followed by Pseudogenization of the original tRNA genes. Pseudogenization of the original Pro (by insertions and deletions) shifts the stop codon for *Cyt b*, thus enabling *Cyt b* to increase from 1191 bp (seen in other lampreys) to 1218 bp (seen in this study). Further tandem duplications are enabled by strand slippage during DNA replication to contract and expand the number of tandem repeats, giving rise to the current arrangement in individuals containing a tandem array of 2 or 3 repeats. Note: pseudogenes represented by orange boxes, tRNA genes indicated by pink boxes and are annotated as P (proline) and F (phenylalanine).

Future lamprey phylogenetic studies should also involve both mitochondrial and nuclear markers. Nuclear restriction site-associated DNA markers could particularly help in elucidating the taxonomic status of lamprey species pairs. This has recently been demonstrated in a study by Mateus *et al*. (2013), in which a population study of sympatric *L. fluviatilis* and *L. planeri* in Portugal found genetic differentiation and diagnostic single nucleotide polymorphisms in several genes (such as saltwater-freshwater osmoregulation) that have been implicated in the adaptation from an anadromous to freshwater resident lifestyle in lampreys and bony fishes (Mateus *et al.* 2013). Furthermore, the findings in Section 3.7.3 highlight the risk of relying on a single mitochondrial gene, such as *ND1*. Therefore, the use of multiple genes in any future phylogenetic study is recommended.

Noncoding region III has raised many questions that can only be elucidated by future studies. These questions are now presented in the following sections and briefly discussed.

4.3.1 Does NCIII Play a Functional Role in Transcription?

With tandemly duplicated genes, there should be little selection against random mutations that alter the copies. In this study, the homogeneity of pseudoPhe at both the intra-individual and inter-individual level suggests that pseudoPhe is under positive selection – which is likely only if it plays some sort of functional role. In mammals, tRNA phenylalanine plays an important role in mitochondrial transcription, containing HSP2, which is the transcriptional promoter responsible for the mRNA transcript containing all H-strand genes (Scarpulla 2008; Lodeiro *et al.* 2012; Zollo *et al.* 2012). Given the finding of a putative mTERF termination sequence in tRNA leucine^{UUR} of *M. mordax* and *M. praecox* that closely resembles that in human mitochondria (Section

3.6.2), it is plausible that mitochondrial transcription in lampreys is similar and thus tRNA phenylalanine and/or pseudoPhe could contain HSP2 and/or contain recognition sites for transcriptional machinery (e.g. mtRNA polymerase). Transcriptional studies will need to be carried out to: 1) elucidate mitochondrial transcription in lampreys and whether it indeed resembles that in mammals; 2) determine if pseudoPhe and pseudoPro are indeed non-functional; and 3) determine if NCIII plays any role in transcription.

4.3.2 Are NCIII-containing Mitogenomes Dominant?

Theoretically, mitogenomes containing tandem duplications should be negatively selected against due to their larger size resulting in slower replication. However, this is overcome if the extra DNA confers a replicative advantage such as an alternative replication site - even if it has a negative effect on mitochondrial function (i.e. serious impairment of respiration or disease (Holt & Reyes 2012). Thus, if population studies were to find NCIII-containing mitogenomes to be the dominant genotype, this would suggest that NCIII may confer a replicative advantage. Population sampling for NCIII could be easily be carried out using primer pair X50F and X51R as a molecular marker to determine both the existence and size of NCIII in an individual with/without the need for sequencing.

4.3.3 Does NCIII Elucidate the Vicinity of O_L?

Given that nearly all gene duplications in Chordata occur near one of the two origins of replication (Boore & Brown 1998), it is strongly thought that genes flanking the origins of strand replication are more likely to be duplicated (Mauro *et al* 2005 and references therein). In most vertebrates, such as the dogfish (Delarbre *et al.* 1997), the replication origin of the L-strand (O_L) is found between tRNAs asparagine and cysteine (Wanrooij *et al.* 2012). The O_L forms a loop that is recognised by the enzyme responsible for L-strand synthesis (Delarbre *et al.* 1997; Pereira 2000; Wanrooij *et al.* 2012). Lamprey studies have been unable to find any obvious O_L in either *L. fluviatilis* or *P. marinus* (Lee & Kocher 1995; Delarbre *et al.* 1997; Delarbre *et al.* 2000), nor has the extensive metazoan study by Wanrooij *et al*. (2012). In the absence of O_L, tRNA stem-and-loop structures could be recognised (Pereira 2000). It is therefore possible that NCIII is located near O_L or that stem-and-loop structures of either tRNA proline (anticodon loop) or phenylalanine (TC loop) is acting as the $O_{L.}$ Replication studies need to be carried out to elucidate mitochondrial replication in lampreys. In doing so, many unanswered questions relating to lamprey mitochondria can be answered. For example, does lamprey mitochondria replication occur via the strand displacement model or the RITOLS model? Where are O_L and O_H located? What role does NCII play in replication if NCI acts as the d-loop?

References

Anderson J.S., Reisz R.R., Scott D., Frobisch N.B., Sumida S.S. (2008) A stem batrachian from the Early Permian of Texas and the origin of frogs and salamanders. *Nature*, **453**, 515-518.

Bailey R.M. (1980) Comments on the classification and nomenclature of lampreys- an alternative view. *Can. J. Fish. Aquat. Sci.*, **37**, (11), 1626-1629.

Baldwin J., Mortimer K., Patak A. (1988) Evolutionary relationships among lamprey families: Amino acid composition analysis of lactate dehydrogenase. *Biochem. Syst. Ecol.*, **16**, (3), 351-353.

Bardack D., Zangeri R. (1968) First fossil lamprey: a record from the Pennsylvanian of Illinois. *Science*, **162**, 1265-1267.

Beamish R.J., Neville C.M. (1992) The importance of size as an isolating mechanism in lampreys. *Copeia*, **1992**, 191-196.

Bensasson D., Zhang D., Hartl D.L., Hewitt G.M. (2001) Mitochondrial pseudogenes: evolution's misplaced witnesses. *Trends. Ecol. Evol.*, **16**, (6), 314-321.

Benton M.J., Donoghue P., Asher R.J. (2009) Calibrating and constraining molecular clocks. In: *The Timetree of Life* (eds. Hedges SB, Kumar S), pp. 35-86. Oxford University Press, Oxford.

Berglund E.C., Kiialainen A., Syvänen A.C. (2011) Next-generation sequencing technologies and applications for human genetic history and forensics. *Investig. Genet.*, **2**, 23.

Bernt M., Donath A., Jühling F., *et al.* (2013) MITOS: Improved *de novo* metazoan mitochondrial genome annotation. *Mol. Phylogenet. Evol.*, **69**, (2), 313-319.

Blair J.E., Hedges S.B. (2005) Molecular phylogeney and divergence times of deuterostome animals. *Mol. Biol. Evol.*, **22**, (11), 2275-2284.

Blank M., Jürss K., Bastrop R. (2008) A mitochondrial multigene approach contributing to the systematics of the brook and river lampreys and the phylogenetic position of *Eudontomyzon mariae*. *Can. J. Fish. Aquat. Sci.*, **65**, 2780-2790.

Boguski D.A., Reid S.B., Goodman D.H., Docker M.F. (2012) Genetic diversity, endemism and phylogeny of lampreys within the genus *Lampetra sensu stricto* (Petromyzontiformes: Petromyzontidae) in western North America. *J. Fish. Biol.*, **81**, (6), 1891-1914.

Boore J., Brown W.M. (1998) Big trees from little genomes: mitochondrial gene order as a phylogenetic tool. *Curr. Opin. Genetics. Dev.*, **8**, 668-674.

Briscoe A., Goodacre S., Masta S.E., *et al.* (2013) Can long-range PCR be used to amplify genetically divergent mitochondrial genomes for comparative phylogenetics? A case study within spiders (Arthropoda: Araneae). *PLoS One*, **8**, (5), e62404.

Burge S.W., Daub J., Eberhardt R., *et al.* (2013) Rfam 11.0: 10 years of RNA families. *Nucleic Acids Res.*, **41**, (Database Issue), D226-D232.

Chang M., Zhang J., Miao D. (2006) A lamprey from the Cretaceous Jehol biota of China. *Nature*, **441**, 972-974.

Chen Y., Liu T., Yu C., Chiang T., Hwang C. (2013) Effects of GC bias in next-generation-sequencing data on *de novo* genome assembly. *PLoS One*, **8**, (4), e62856.

Christianson T.W., Clayton D.A. (1988) A tridecamer DNA sequence supports human mitochondrial RNA 3'-end formation in vitro. *Mol. Cell. Biol.*, **8**, (10), 4502-4509.

Collin S.P., Davies W.L., Hart N.S., Hunt D.M. (2009) The evolution of early vertebrate photoreceptors. *Philos. Trans. R. Soc. Lond. B. Sci.*, **364**, (1531), 2925-2940.

Curole P.J., Kocher T.D. (1999) Mitogenomics: digging deeper with complete mitochondrial genomes. *TREE*, **14**, (10), 394-398.

De Beuf K., De Schrijver J., Thas O., *et al.* (2012) Improved base-calling and quality scores for 454 sequencing based on a Hurdle Poisson model. *Bioinformatics*, **13**, 303.

Delarbre C., Barriel V., Tillier S., Janvier P., Gachelin G. (1997) The main features of the craniate mitochondial DNA between the *ND1* and the *COI* genes were established in the common ancestor with the lancelet. *Mol. Biol. Evol.*, **14**, (8), 807-813.

Delarbre C., Escriva H., Gallut C., *et al.* (2000) The complete nucleotide sequence of the mitochondrial DNA of Agnathan *Lampetra fluviatilis*: bearings on the phylogeny of cyclostomes. *Mol. Biol. Evol.*, **17**, (4), 519-529.

Delarbre C., Gallut C., Barriel V., Janvier P., Gachelin G. (2002) Complete mitochondrial DNA of the hagfish, *Eptatretus burgeri*: the comparative analysis of mitochondrial DNA sequences strongly supports the cyclostome monophyly. *Mol. Phylogenet. Evol.*, **22**, (2), 184-192.

Docker M.F. (2009) A review of the evolution of nonparasitism in lampreys and an update of the paired species concept. In: *Biology, Management, and Conservation of Lampreys in North America* (eds. Brown LR, Chase SD, Mesa MG, Beamish RJ, Moyle PB), pp. 71-114. American Fisheries Society Symposium 72, Bethesda, Maryland.

Docker M.F., Youson J.H., Beamish R.J., Devlin R.H. (1999) Phylogeny of the lamprey genus *Lampetra* inferred from mitochondrial cytochrome *b* and ND3 gene sequences. *Can. J. Fish. Aquat. Sci.*, **56**, 2340-2349.

Dwivedi B., Gadagkar S.R. (2009) Phylogenetic inference under varying proportions of indel-induced alignment gaps. *BMC Evol. Biol.*, **9**, 211.

Elbrecht V., Poettker L., John U., Leese F. (2013) The complete mitochondrial genome of the stonefly *Dinocras cephalotes* (Plecoptera, Perlidae). *Mitochondrial DNA*, **26**, (3), 469-470.

Espanhol R., Almeida P.R., Alves M.J. (2007) Evolutionary history of lamprey paired species *Lampetra fluviatilis* (L.) and *Lampetra planeri* (Bloch) as inferred from mitochondrial DNA variation. *Mol. Ecol.*, **16**, (9), 1909-1924.

Forey P., Janvier P. (1993) Agnathans and the origin of jawed vertebrates. *Nature*, **361**, 129-134.

Furlong R.F., Holland P.W. (2002) Bayesian phylogenetic analysis supports monophyly of ambulacraria and of cyclostomes. *Zoolog. Sci.*, **19**, (5), 593-599.

Gaitán-Espitia J.D., Nespolo R.F., Opazo J.C. (2013) The complete mitochondrial genome of the land snail *Cornu aspersum* (Helicidae:Mollusca): intra-specific divergence of protein-coding genes and phylogenetic considerations within Euthyneura. *PLoS One*, **8**, (6), e67299.

Gess R.W., Coates M.I., Rubidge B.S. (2006) A lamprey from the Devonian period of South Africa. *Nature*, **443**, 981-984.

Gill H.S., Renaud C.B., Chapleau F., Mayden R.L., Potter I.C. (2003) Phylogeny of living parasitic lampreys (Petromyzontiformes) based on morphological data. *Copeia*, **4**, 687-703.

Glenn T.C. (2011) Field guide to next-generation DNA sequencers. *Mol. Ecol. Res.*, **11**, 759-769.

Golan D., Medvedev P. (2013) Using state machines to model the Ion Torrent sequencing process and to improve read error rates. *Bioinformatics*, **29**, (13), i344-i351.

Haouchar D. (2009) *The use of mitochondrial genomes to study the evolutionary history of southern hemisphere lampreys and the position of Petromyzontiformes in a vertebrate phylogeny* Honours Thesis, Murdoch University.

Hardisty M.W. (1982) Lampreys and hagfishes: analysis of cyclostome relationships. In: *The Biology of Lampreys* (eds. Hardisty MW, Potter IC), pp. 165-259. Academic Press, London.

Hardisty M.W. (2006) *Lampreys: Life without Jaws* Forrest Text, London.

Hardisty M.W., Potter I.C. (1971a) The behaviour, ecology and growth of larval lampreys. In: *The Biology of Lampreys* (eds. Hardisty MW, Potter IC), pp. 85-125. Academic Press, London.

Hardisty M.W., Potter I.C. (1971b) The general biology of adult lampreys. In: *The Biology of Lampreys* (eds. Hardisty MW, Potter IC), pp. 127-206. Academic Press, London.

Hardisty M.W., Potter I.C. (1971c) Paired species. In: *The Biology of Lampreys* (eds. Hardisty MW, Potter IC), pp. 249-277. Academic Press, London.

Heimberg A.M., Sempere L.F., Moy V.N., Donoghue P.C., Peterson K.J. (2008) MicroRNAs and the advent of vertebrate morphological complexity. *PNAS*, **105**, (8), 2946-2950.

Hert D.G., Fredlake C.P., Barron A.C. (2008) Advantages and limitations of next-generation sequencing technologies: a comparison of electrophoresis and non-electrophoresis methods. *Electrophoresis*, **29**, (23), 4618-4626.

Holt I.J., Reyes A. (2012) Human mitochondrial DNA replication. *Cold Spring Harb. Perspect. Biol*, **4**, (12), a012971.

Hubbs C.L., Potter I.C. (1971) Distribution, phylogeny and taxonomy. In: *The Biology of Lampreys* (eds. Hardisty MW, Potter IC), pp. 1-65. Academic Press, London.

Hughes R.L., Potter I.C. (1968) Studies of gametogenesis and fecundity in the lampreys *Mordacia praecox* and *M. mordax* (Petromyzonidae). *Aust. J. Zool.*, **1969**, (17), 447-464.

Hume J.B., Adams C.E., Mable B., Bean C.W. (2013) Sneak male mating tactics between lampreys (Petromyzontiformes) exhibiting alternative life-history strategies. *J. Fish. Biol.*, **82**, (3), 1093-1100.

Hwang D., Byeon H., Lee J. (2013a) Complete mitochondrial genome of the river lamprey, Lampetra japonica (Petromyzontiformes, Petromyzonidae). *Mitochondrial DNA*, **Early Online**, 1-2.

Hwang D., Byeon H., Lee J. (2013b) Complete mitochondrial genome of the sand lamprey, Lampetra reissneri (Petromyzontiformes, Petromyzontidae). *Mitochondrial DNA*, **Early Online**, 1-3.

Hyvärinen A.K., Pohjoismäki J.L., Reyes A., *et al.* (2007) The mitochondrial transcription termination factor mTERF modulates replication pausing in human mitochondrial DNA. *Nucleic Acids Res.*, **35**, (19), 6458-6474.

Illumina (2014) *HiSeqX Series of Sequencing Systems Specification Sheet.* Illumina, Inc, Online. http://www.illumina.com/documents/products/datasheets/datasheet-hiseq-x-ten.pdf

Janvier P. (2006) Palaeontology: modern look for ancient lamprey. *Nature*, **443**, 921-924.

Janvier P. (2010) microRNAs revive old views about jawless vertebrate divergence and evolution. *PNAS*, **107**, (45), 19137-19138.

Janvier P., Lund R. (1983) *Hardistiella montanensis* N. gen. et. sp, (Petromyzontida) from the Lower Carboniferous of Montana, with remarks on the affinities of the lampreys. *J. Vert. Paleontol.*, **2**, (4), 407-413.

Jex A., Hall R., Littlewood D., Gasser R. (2010) An integrated pipeline for next-generation sequencing and annotation of mitochondrial genomes. *Nucleic Acids Res.*, **38**, (2), 522-533.

Jukes T.H., Cantor C.R. (1969) Evolution of protein molecules. In: *Mammalian Protein Metabolism* (ed. Munro H), pp. 21-132. Academic Press, New York.

Kalchhauser I., Kutschera V., Burkhardt-Holm P. (2014) The complete mitochondrial genome of the invasive Ponto-Caspian goby *Ponticola kessleri* obtained from high-throughput sequencing using the Ion Torrent Personal Genome Machine. *Mitochondrial DNA*, **Early Online**, 1-3.

Ku C., Pawitan Y.W., M., Roukos D.H., Cooper D.N. (2013) The evolution of high-throughput sequencing technologies: from Sanger to single-molecule sequencing. In: *Next Generation Sequencing in Cancer Research* (eds. Wu W, Choudhry H), pp. 1-30. Springer-Verlag, New York.

Kumazawa Y., Ota H., Nishida M., Ozawa T. (1998) The complete nucleotide sequence of a snake (*Dinodon semicarinatus*) mitochondrial genome with two identical control regions. *Genetics*, **150**, 313-329.

Kurabayashi A., Sumida M., Yonekawa H., *et al.* (2008) Phylogeny, recombination, and mechanisms of stepwise mitochondrial genome reorganization in mantellid frogs from Madagascar. *Mol. Biol. Evol.*, **25**, (5), 874-891.

Kuraku S., Kuratani S. (2006) Time scale for cyclostome evolution inferred with a phylogenetic diagnosis of hagfish and lamprey cDNA sequences. *Zoolog. Sci.*, **23**, (12), 1053-1064.

Kuratani S., Kuraku S., Murakami Y. (2002) Lamprey as an evo-devo model: lessons from the comparative embryology and molecular phylogenetics. *Genesis*, **34**, (3), 175-183.

Lamb T.D., Collin S.P., Pugh E.N. (2007) The evolution of the vertebrate eye: opsins, photoreceptors, retina and eye cup. *Nat. Rev. Neurosci.*, **8**, (12), 960-976.

Lang N.J., Roe K.J., Renaud C.B., *et al.* (2009) Novel relationships among lampreys (Petromyzontiformes) revealed by taxonomically comprehensive molecular data set. *Am. Fish. Soc. Symp.*, **72**.

Leamon R.H., Rothberg J.M. (2009) DNA Sequencing and Genomics. In: *Desk Encyclopedia of Microbiology* (ed. Schaechter M), pp. 369-382. Elsevier, Oxford, U.K.

Lee W., Kocher T. (1995) Complete sequence of a Sea Lamprey (*Petromyzon Marinus*) mitochondrial genome: early establishment of the vertebrate genome organization. *Genetics*, **139**, 873-887.

Life Technologies (2014) *Ion PGM System Specifications.* Life Technologies, Online. http://www.lifetechnologies.com/au/en/home/life-science/sequencing/next-generation-sequencing/ion-torrent-next-generation-sequencing-workflow/ion-torrent-next-generation-sequencing-run-sequence/ion-pgm-system-for-next-generation-sequencing/ion-pgm-system-specifications.html.html

Linzey D.W. (2012) Early Chordates and Jawless Fishes. In: *Vertebrate Biology*, pp. 79-94. Johns Hopkins University Press, Baltimore.

Liu L., Li Y., Li S., *et al.* (2012) Comparison of next-generation sequencing systems. *J. Biomed. Biotechnol.*, **2012**, 251364.

Lodeiro M.F., Uchida A., Bestwick M., *et al.* (2012) Transcription from the second heavy-strand promoter of human mtDNA is repressed by transcription factor A in vitro. *Proc. Natl. Acad. Sci. USA*, **109**, (17), 6513-6518.

Loman N.J., Misra R.V., Dallman T.J., *et al.* (2012) Performance comparison of benchtop high-throuhput sequencing platforms. *Nat. Biotechnol.*, **30**, (5), 434-439.

Lovett S.T., Feschenko V.V. (1996) Stabilization of diverged tandem repeats by mismatch repair: evidence for deletion formation via a misaligned replication intermediate. *Proc. Natl. Acad. Sci. USA*, **93**, 7120 - 7124.

Mabuchi K., Miya M., Satoh T.P., Westneat M.W., Nishida M. (2004) Gene rearrangements and evolution of tRNA pseudogenes in the mitochondrial genome of the parrotfish (Toleostei: Perciformes: Scaridae). *J. Mol. Evol.*, **59**, (3), 287-297.

Macey J.R., Papenfuss T.J., Kuehl J.V., Fourcade H.M., Boore J. (2004) Phylogenetic relationships among amphisbaenian reptiles based on complete mitochondrial genomic sequences. *J. Mol. Phylogenet. Evol.*, **33**, (1), 22-31.

Macey J.R., Schulte J.A., Larson A., Papenfuss T.J. (1998) Tandem duplication via light-strand synthesis may provide a precursor for mitochondrial genomic rearrangement. *Mol. Biol. Evol.*, **15**, (1), 71 - 75.

Mallat J., Sullivan J. (1998) 28S and 18S rDNA sequences support the monophyly of lampreys and hagfishes. *Mol. Biol. Evol.*, **15**, (12), 1706-1718.

Mallat J., Sullivan J., Winchell C.J. (2001) The relationship of lampreys to hagfishes: a spectral analysis of ribosomal DNA sequences. In: *Major Events in Early Vertebrate Evolution: Palaeontology, Phylogeny, and Development* (ed. Ahlberg PE), pp. 106-118. Taylor and Francis, London.

Mallat J., Winchell C.J. (2007) Ribosomal RNA genes and deuterostome phylogeny revisited: more cyclostomes, elasmobranchs, reptiles, and a brittle star. *Mol. Phylogenet. Evol.*, **43**, (3), 1005-1022.

Martin M., Rahmann S. (2013) Aligning flowgrams to DNA sequences. In: *German Conference on Bioinformatics 2013* eds. Beissbarth T, Kollmar M, Leha A, *et al.*), pp. 125-135. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Wadern.

Mateus C.S., Almeida P.R., Quintella B.R., Alves M.J. (2011) MtDNA markers reveal the existence of allopatric evolutionary lineages in the threatened lampreys *Lampetra fluviatilis* (L.) and *Lampetra planeri* (Bloch) in the Iberian glacial refugium. *Conserv. Genet.*, **12**, 1061-1074.

Mateus C.S., Rodriguez-Munoz R., Quintella B.R., Alves M.J., Almeida P.R. (2012) Lampreys of the Iberian Peninsula: distribution, population status and conservation. *Endang. Species Res.*, **16**, 183-189.

Mateus C.S., Stange M., Berner D., *et al.* (2013) Strong genome-wide divergence between sympatric European river and brook lampreys. *Curr. Biol.*, **23**, (15), R649-R650.

McClellan A.D. (2013) Spinal Cord Injury: The Lamprey Model. In: *Animal Models of Spinal Cord Repair* (ed. Aldskogius H), pp. 63-108. Humana Press, London.

Meeuwig M.H., Bayer J.M., Seelye J.G. (2005) Effects of temperature on survival and development of early life stage Pacific and Western Brook lampreys. *Trans. Am. Fish. Soc.*, **134**, (1), 19-27.

Metzker M.L. (2010) Sequencing technologies- the next generation. *Nat. Rev. Genet.*, **11**, (1), 31-46.

Meyer A., Zardoya R. (2003) Recent advances in the (molecular) phylogeny of vertebrates. *Annu. Rev. Ecol. Evol. Syst.*, **34**, 311-338.

Milne I., Bayer M., Cardle L., *et al.* (2010) Tablet- next generation sequence assembly visualization. *Bioinformatics*, **26**, (3), 401-402.

Milton P.L. (2003) *The phylogeny of lampreys : an assessment based on the gene order and nucleotide sequence for the mitochondrial genomes of Geotria australis and Mordacia mordax.* Honours Thesis, Murdoch University.

Moritz C., Brown W.M. (1987) Tandem duplications in animal mitochondrial DNAs: variation in incidence and gene content among lizards. *Proc. Natl. Acad. Sci. USA*, **84**, 7183-7187.

Morozova O., Marra M.A. (2008) Applications of next-generation sequencing technologies in functional genomics. *Genomics*, **92**, 255-264.

Mueller R.L., Boore J.L. (2005) Molecular mechanisms of extensive mitochondrial gene rearrangement in Plethodontid salamanders. *Mol. Biol. Evol.*, **22**, (10), 2104-2112.

Near T.J. (2009) Conflict and resolution between phylogenies inferred from molecular and phenotypic data sets for hagfish, lampreys, and gnathostomes. *J. Exp. Zool. B Mol. Dev. Evol.*, **312**, (7), 749-761.

Nei M., Kumar S. (2000) *Molecular Evolution and Phylogenetics* Oxford University Press, New York.

Niedringhaus T.P., Milanova D., Kerby M.B., Snyder M.P., Barron A.E. (2011) Landscape of next-generation sequencing technologies. *Anal. Chem.*, **83**, 4327-4341.

Novogene (2014) *HiSeq X Ten Human Sequencing Data Disclosed*. Online. http://www.novogene.com/en/index.php?m=content&c=index&a=show&catid=97&id=188

Oisi Y., K.G. O., Kuraku S., S. F., Kuratani S. (2013) Craniofacial development of hagfishes and the evolution of vertebrates. *Nature*, **493**, 175-180.

Osório J., Rétaux S. (2008) The lamprey in evolutionary studies. *Dev. Genes Evol.*, **218**, 221-235.

Pereira S.L. (2000) Mitochondrial genome organization and vertebrate phylogenetics. *Genet. Mol. Biol.*, **23**, (4), 745-752.

Pham X.H., Farge G., Shi Y., *et al.* (2006) Conserved sequence box II directs transcription termination and primer formation in mitochondria. *J. Biol. Chem.*, **281**, (34), 24647-24652.

Phillips M.J., Bennet T.H., Lee M.S. (2009) Molecules, morphology, and ecology indicate a recent, amphibious ancestry for echidnas. *Proc. Natl. Acad. Sci. USA*, **106**, (40), 17089-17094.

Piavis G.W., Howell J.H., Smith A.J. (1970) Experimental hybridization among five species of lampreys from the Great Lakes. *Copeia*, **1970**, (1), 29-37.

Pop M., Salzberg S.S., M. (2002) Genome sequence assembly: algorithms and issues. *IEEE Comput.*, **35**, 47-54.

Potter I.C. (1980) Ecology of larval and metamorphosing lampreys. *Can. J. Fish. Aquat. Sci.*, **37**, 1641-1657.

Potter I.C., Beamish F.W.H. (1975) Lethal temperatures in ammocoetes of four species of lampreys. *Acta Zool.*, **56**, (1), 85-91.

Potter I.C., Gill H.S. (2003) Adaptive radiation of Lampreys. *J. Great Lakes Res.*, **29** (Supplement 1), 95-112.

Potter I.C., Gill H.S., Renaud C.B. (2014) Petromyzontidae: Lampreys. In: *Freshwater Fishes of North America* (eds. Burr BM, Warren ML), pp. 105-139. Johns Hopkins University Press, Baltimore.

Potter I.C., Gill H.S., Renaud C.B., Haouchar D. (2015) The taxonomy, phylogeny and distribution of lampreys. In: *Lampreys: Biology, Conservation and Control* (ed. Docker MF), pp. 35-73. Springer, Berlin.

Potter I.C., Hilliard R.W., Bird D.J. (1982) Stages in metamorphosis. In: *The Biology of Lampreys* (eds. Hardisty MW, Potter IC), pp. 137-164. Academic Press, London.

Potter I.C., Strahan R. (1968) The taxonomy of the lampreys *Geotria* and *Mordacia* and their distribution in Australia. *Proc. Linn. Soc. London*, **179**, 229-240.

Redko Y., Sierra-Gallay I., Condon C. (2007) When all's zed and done: the structure and function of RNase Z in prokaryotes. *Nat. Rev. Microbiol.*, **5**, (4), 278- 286.

Renaud C.B. (2011) Lampreys of the world. An annotated and illustrated catalogue of lamprey species known to date. In: *FAO Species Catalogue for Fishery Purposes No. 5*. FAO, Rome.

Renaud C.B., Gill H.S., Potter I.C. (2009) Relationships between the diets and characteristics of the dentition, buccal glands and velar tentacles of the adults of the parasitic species of lamprey. *J. Zool.*, **278**, (3), 231-242.

Richardson M.K., Wright G.M. (2003) Developmental transformations in a normal series of embryos of the sea lamprey *Petromyzon marinus* (Linnaeus). *J. Morphol.*, **257**, (3), 348-363.

Riddington J. (2007) *The Mitochondrial DNA Sequence of Geotria australis and Mordacia mordax: Phylogeny and the Establishment of a Lamprey Consensus Gene Order* Honours Thesis, Murdoch University.

Rizzo J.M., Buck M.J. (2012) Key principles and clinical applications of "next-generation" DNA sequencing. *Cancer Prev. Res.*, **5**, (7), 887-900.

Ronquist F., Teslenko M., van der Mark P., *et al.* (2012) MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Syst. Biol.*, **61**, (3), 539-542.

Ronquist F., van der Mark P., Huelsenbeck J.P. (2009) Bayesian Phylogenetic Analysis using MrBayes. In: *The Phylogenetic Handbook: A Practical Approach to Phylogenetic Analysis and Hypothesis Testing* (eds. Lemey P, Salemi M, Vandamme A), p. 232. Cambridge University Press, New York.

Saitou N., Nei M. (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.*, **4**, (4), 406-425.

San Mauro D., Gower D.J., Zardoya R., Wilkinson M. (2006) A hotspot of gene order rearrangement by tandem duplication and random loss in the vertebrate mitochondrial genome. *Mol. Biol. Evol.*, **23**, (1), 227-234.

Sanger F., Nicklen S., Coulson A.R. (1977) DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. USA*, **74**, (12), 5463-5467.

Scarpulla R.C. (2008) Transcriptional paradigms in mammalian mitochondrial biogenesis and function. *Physiol. Rev.*, **88**, (2), 611-638.

Schattner P., Brooks A.N., Lowe T.M. (2005) The tRNAscan-SE, snoscan and snoGPS web servers for the detection of tRNAs and snoRNAs. *Nucleic Acids Res.*, **33**, (Web Server Issue), w686-w689.

Schatz M.C., Delcher A.L., Salzberg S.L. (2010) Assembly of large genomes using second-generation sequencing. *Genome. Res.*, **20**, (9), 1165-1173.

Schwarz G. (1978) Estimating the dimension of a model. *Ann. Stat.*, **6**, (2), 461-464.

Shendure J., Ji H. (2008) Next-generation DNA sequencing. *Nat. Biotechnol.*, **26**, (10), 1135-1145.

Sheridan C. (2014) Illumina claims $1,000 genome win. *Nat. Biotechnol.*, **32**, (2), 115.

Smith J.J., Kuraku S., Holt C., *et al.* (2013) Sequencing of the sea lamprey (*Petromyzon marinus*) genome provides insights into vertebrate evolution. *Nat. Genet.*, **45**, (4), 415-421.

Smith J.J., Saha N.R., Amemiya C.T. (2010) Genome biology of the cyclostomes and insights into the evolutionary biology of the vertebrate genomes. *Integr. Comp. Biol.*, **50**, (1), 130-137.

Stock D.W., Whitt G.S. (1992) Evidence from 18S ribosomal RNA sequences that lampreys and hagfishes form a natural group. *Science*, **257**, (5071), 787-789.

Takezaki N., Figueroa F., Zaleska-Rutczynska Z., Klein J. (2003) Molecular phylogeny of early vertebrates: monophyly of the agnathans as revealed by sequences of 35 genes. *Mol. Biol. Evol.*, **20**, (2), 287-292.

Tamura K., Stecher G., Peterson D., Filipski A., Kumar S. (2013) MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. *Mol. Biol. Evol.*, **30**, (12), 2725.

Treangen T.J., Salzberg S.L. (2012) Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nat. Rev. Genet.*, **13**, (1), 36-46.

Triant D.A., DeWoody J.A. (2007) The occurrence, detection, and avoidance of mitochondrial translocations in mammalian systematics and phylogeography. *J. Mammol.*, **88**, (4), 908-920.

Vladykov V.D., Kott E. (1979) Satellite species among the Holartic lampreys (Petromyzontidae). *Can. J. Zool.*, **57**, 860-867.

Vogl I., Eck S.H., Benet-Pagès A., *et al.* (2012) Diagnostic applications of next generation sequencing: working towards quality standards. *J. Lab. Med.*, **36**, (4), 227-239.

Wanrooij S., Fuste J.M., Stewart J.B., *et al.* (2012) *In vivo* mutagenesis reveals that OriL is essential for mitochondrial DNA replication. *EMBO Rep.*, **13**, (12), 1130-1137.

Yamazaki Y., Yokoyama R., Nishida M., Goto A. (2006) Taxonomy and molecular phylogeny of *Lethenteron* lampreys in eastern Eurasia. *J. Fish Biol.*, **68**, (Supplement B), 251-269.

Yang C.H., Chang H.W., Ho C.H., Chou Y.C., Chuang L.Y. (2011) Conserved PCR primer set designing for closely-related species to complete mitochondrial genome sequencing using a sliding window-based PSO algorithm. *PLoS One*, **6**, (3), e17729.

Youson J.H. (1980) Morphology and physiology of lamprey metamorphosis. *Can. J. Fish. Aquat. Sci.*, **37**, (11), 1687-1710.

Zhang D., Hewitt G.M. (1996) Nuclear integrations: challenges for mitochondrial DNA markers. *Trends. Ecol. Evol.*, **11**, (6), 247-251.

Zollo O., Tiranti V., Sondheimer N. (2012) Transcriptional requirements of the distal heavy-strand promoter of mtDNA. *Proc. Natl. Acad. Sci. USA*, **109**, (17), 6508-6512.

Zuckerkandl E., Pauling L. (1965) Evolutionary divergence and convergence in proteins. In: *Evolving Genes and Proteins* (eds. Bryson V, Vogel HJ), pp. 97-166. Academic Press, New York.

Zuker M. (2003) Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res.*, **31**, (13), 3406-3415.

Appendices

Appendix 1.

Products and their manufacturer, used in this study.

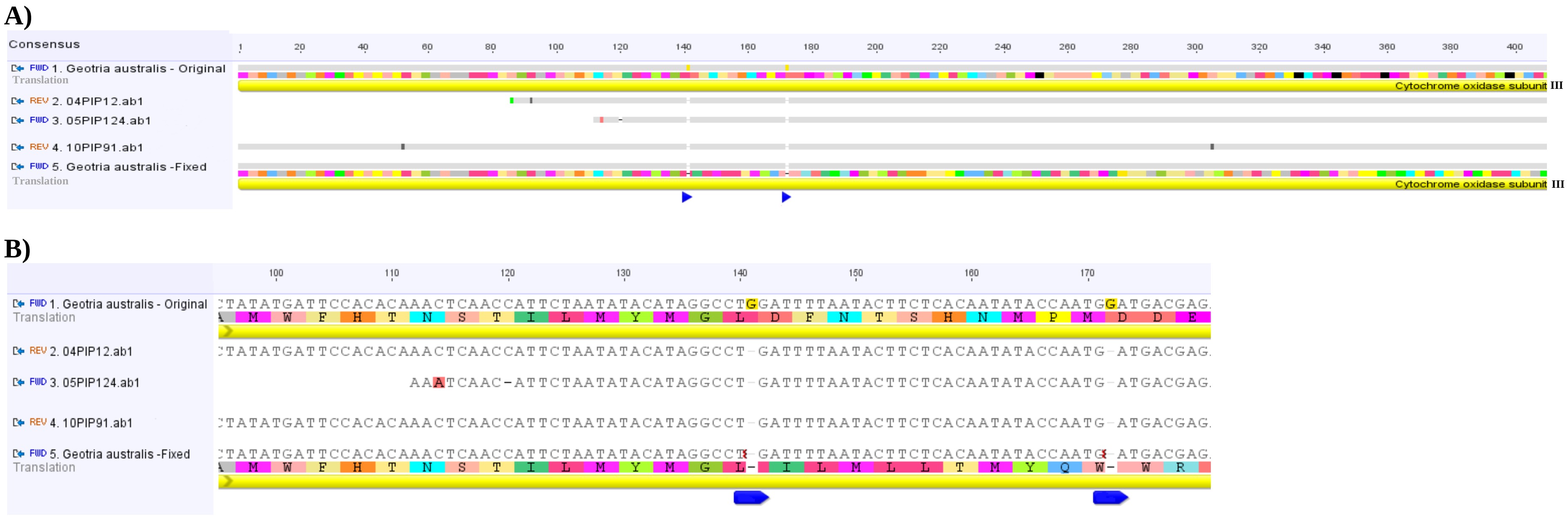
Product	Manufacturer
1 Kb DNA ladder	Promega (USA)
6 x Blue/Orange loading dye	Promega (USA)
6 x Loading Dye	Fermentas (Australia)
100 bp DNA ladder	Promega (USA)
100 bp Plus GeneRuler DNA Ladder	Fermentas
10 000 x SYBR Safe DNA gel stain	Invitrogen (USA)
Agarose powder (DNA grade)	Progen Biosciences (Australia)
Ammonium acetate	Univar (Australia)
DNA polymerase 10X reaction buffer	Fisher Biotech (Australia)
dNTP mix 10mM	Fisher Biotech (Australia) Promega?
EliminatoR	Fisher Biotech (Australia)
Ethanol absolute	Merck Chemicals (Germany)
Ethylenediaminetetraacetic acid (EDTA)	Sigma Alderich (USA)
Expand Long Range dNTPack PCR nucleotide mix	Roche (Germany) Applied Sciences?
Expand Long Range dNTPack polymerase blend	Roche (Germany)
Expand Long Range dNTPack 15X PCR buffer	Roche (Germany)
Hydrochloric acid (HCl)	Merck Chemicals (Australia)
Magnesium Chloride (25mM)	Fisher Biotech (Australia)
MyTaq 5 x Reaction Buffer	Bioline (Australia)
MyTaq DNA Polymerase	Bioline (Australia)
PCR polymerisation 10X buffer	Fisher Biotech (Australia)
PCR primers	Integrated DNA Technologies (USA)
Proteinase-K	Promega (USA)
Sucrose	Merck Chemicals (Australia)
Sodium acetate	Chem-Supply (Australia)
Sodium Chloride ($NaCl_2$)	Sigma Alderich (USA)
Sodium dodecyl sulphate (SDS)	BDH Laboratory Supplies
Sodium hydroxide (NaOH)	AnalaR (Australia)
Tris base	Invitrogen (USA)
UltraPure water	Fisher Biotech (Australia)
Velocity 5 x HiFi Reaction Buffer	Bioline
Velocity DNA Polymerase	Bioline
Velocity DMSO	Bioline
Water For Injections BP (medical grade water)	AstraZeneca (Australia)

Appendix 2.

Commonly used buffers and solutions, used in this study.

Ammonium acetate (4M)	30.832g ammonium acetate dissolved in 100mL dH_2O.
Proteinase-K (20mg/mL)	20mg proteinase-K dissolved in 1mL dH_2O.
Resuspension Buffer	42.76g sucrose, 0.1486g EDTA and 0.6064g Tris base dissolved in 400mL dH_2O. Adjust to pH 7.5 with conc. HCl and made up to 500mL with dH_2O.
SDS (10% w/v)	10g SDS dissolved in 100mL of dH_2O.
Sodium Acetate (4M)	32.812g Sodium Acetate dissolved in 100mL dH_2O.
TAE buffer (50X)	242g Tris base, 57.1mL glacial acetic acid and 37.2g EDTA dissolved in 1L dH_2O.
Tissue Digestion Solution	1.46g EDTA, 1.51g Tris base and 1.75g $NaCl_2$ added to200mL dH_2O. Adjust to pH 8.0 with sodium hydroxide pellets. Fill to 214mL with dH_2O and Autoclave. Add 36mL of 10% SDS.
TE Buffer	Tris-HCl (10 mM) mixed with 1 mM EDTA, pH adjusted to 8.0

Appendix 3.



```
1. Geotria australis - Original  CTATATGATTCCACACAAACTCAACCATTCTAATATACATAGGCCTGGATTTTAATACTTCTCACAATATACCAATGGATGACGAG.
Translation                      A M W F H T N S T I L M Y M G L D F N T S H N M P M D D E
2. 04PIP12.ab1                   CTATATGATTCCACACAAACTCAACCATTCTAATATACATAGGCCT-GATTTTAATACTTCTCACAATATACCAATG-ATGACGAG.
3. 05PIP124.ab1                                            AAATCAAC-ATTCTAATATACATAGGCCT-GATTTTAATACTTCTCACAATATACCAATG-ATGACGAG.
4. 10PIP91.ab1                   CTATATGATTCCACACAAACTCAACCATTCTAATATACATAGGCCT-GATTTTAATACTTCTCACAATATACCAATG-ATGACGAG.
5. Geotria australis -Fixed      CTATATGATTCCACACAAACTCAACCATTCTAATATACATAGGCCT-GATTTTAATACTTCTCACAATATACCAATG-ATGACGAG.
Translation                      A M W F H T N S T I L M Y M G L - I L M L L T M Y Q W - W R
```

Appendix 3. Identification and correction (this study) of errors causing premature stop codons in the *COIII* gene of *G. australis* (Milton 2003). The *COIII* nucleotide sequence from the original G. australis mitogenome (*G. australis* -Original) is aligned to three Sanger sequencing reads (rows ending in .ab1) – determined in the Milton (2003) study – and the error corrected *COIII* sequence (*G. australis* -Fixed). **A)** Zoomed out view of the alignment showing 4 premature stop codons (represented by black boxes) interspersed throughout the first 400 bp of sequence. **B)** Zoomed in version showing the Sanger sequencing reads confirming that the two G insertions (at base 141 and 172) are errors in the original sequence. The amino acid translation for the fixed sequence changes the amino acid sequence such that the premature stop codons are removed. Note: the original *COIII* sequence contained 13 premature stop codons. The original *COIII* size was 788 bp, the fixed *COIII* size is 786 bp. L-strand (5' – 3') is shown for simplicity. Only selected parts of the *COIII* gene are shown for simplicity. Dashes indicate deletions; coloured bases indicate insertions or base changes. The blue annotation arrows are a reference point in the alignment between **A)** and **B)**. IUPAC Single letter codes are used for the amino acid sequence.

Appendix 4.

```
M. mordax (Haouchar) // GTTATTTTTCCATTACTAGG------------------------------------------------AGAAAAAGAAAAACAGAAAATAAAAATATCATCTAAGATCATTAATTAA //
Wallagorough (F195)  // GTTATTTTTCCATTACTAGGTCTCCTAGAAAACCCTATTCTAGCTCACAAACCCGAGGAAGAGGAAGAAGAAAAAGAAAAACAGAAA
Bega (F122)          // GTTATTTTTCCATTACTAGGTCTCCTAGAAAACCCTATTCTAGCTCACAAACCCGAGGAAGAGGAAGAAGAAAAAGAAAAACAGAAAATAAAAATATCATCTAAGATCATTAATTAA //
Bega (F121)          // GTTATTTTTCCATTACTAGGTCTCCTAGAAAACCCTATTCTAGCTCACAAACCCGAGGAAGAGGAAGAAGAAAAAGAAAAACAGAAAATAAAAA
Bega (F120)          // GTTATTTTTCCATTACTAGGTCTCCTAGAAAACCCTATTCTAGCTCACAAACCCGAGGAAGAGGAAGAAGAAAAAGAAAAACAGAAAATAAAAATATCATCTAAGATCATTAATTAA //
M. mordax #1D        // GTTATTTTTCCATTACTAGGTCTCCTAGAAAACCCTATTCTAGCTCACAAACCCGAGGAAGAGGAAGAAGAAAAAGAAAAACAGAAAATAAAAATATCATCTAAGATCATTAATTAA //
Amino acid sequence  //  V  I  F  P  L  L  G  L  L  E  N  P  I  L  A  H  K  P  E  E  E  E  E  E  K  E  K  Q  K  M  K  M  S  S  K  I  I  N  *
```

Appendix 4. *Cyt b* nucleotide alignment of *M. mordax* (Haouchar 2009; line 1) and four *M.* mordax individuals from different localities (Same study, unpublished; lines 2-5). The sequences are aligned to *M. mordax* #1D (this study; line 6) for comparison and the amino acid sequence is also shown. As can be seen in the figure, the *Cyt b* sequence of *M. mordax* (Haouchar 2009) is ‘missing’ 48 bp despite being present (highlighted yellow) in other individuals sequenced in the same study, and being present in *M. mordax* #1D. Note: only 117 bp corresponding to the L-strand 3’-end of *Cyt b* is shown. Dashes indicate deletions. // indicates continuation of the sequence. IUPAC Single letter codes are used for the amino acid sequence

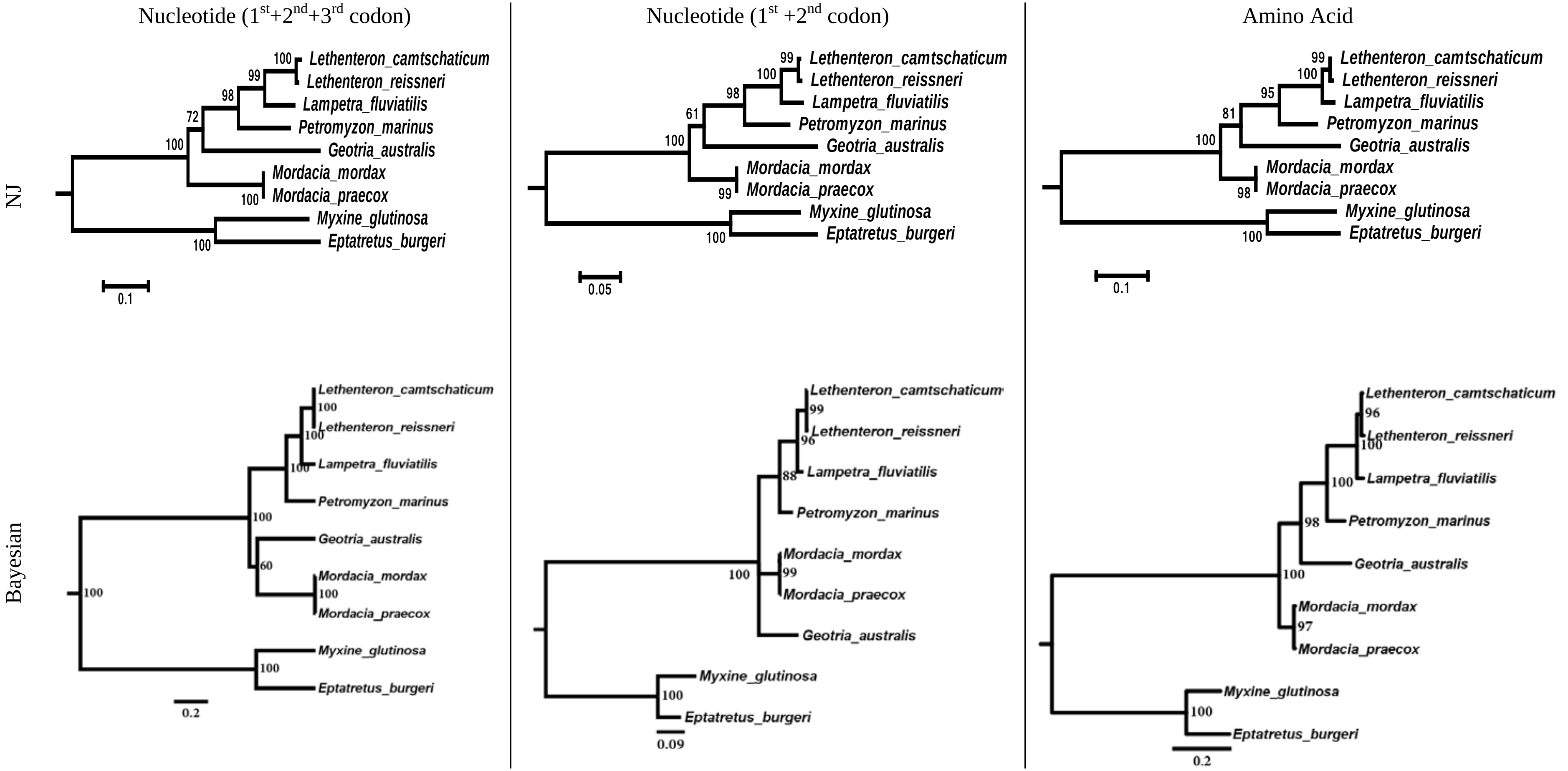
Appendix 5.

```
                //                                                                                         Cytochrome b
Haouchar 2009   // TTACTAGG------------------------------------------------AG-AAAAAGAAAAACAGAAAATAAAAATATCATCTAAGATCATT
M. mordax #1D   // TTACTAGGTCTCCTAGAAAACCCTATTCTAGCTCACAAACCCGAGGAAGAGGAAGAAG-AAAAAGAAAAACAGAAAATAAAAATATCATCTAAGATCATT
FSIDF9J02B0GQA  // TTACTAGGTCTCCTAGAAAACCCTATTCTAGCTCACAAACCCGAGGAAGAGGAAGAAGAAAAAAGAAAAACAGAAAATAAAAATATCATCTAAGATCATT
FSIDF9J02CCOH7  // TTACTAGGTCTCCTAGAAAACCCTATTCTAGCTCACAAACCCGAGGAAGAGGAAGAAGAAAAAAGAAAAACAGAAAATAAAAATATCATCTAAGATCATT
FSIDF9J02CJD4N  // TTACTAGGTCTCCTAGAAAACCCTATTCTAGCTCACAAACCCGAGGAAGAGGAAGAAG-AAAAAGAAAAACAGAAAATAAAAATATCATCTAAGATCATT
FSIDF9J02B0VSA                                                                                                         TT

                -----> || 5’-Repeat
Haouchar  2009  AA--TTAAATTATTACCATAAGTTAAAG------------------------------------------------------------------------
M. mordax #1D   AA--TTAAATTATTACCATAAGTTAAAGTAGCTTAAG-TTT–AAAGCAAAGCTCGAAGATGCTAGTATGATGGTCCAAACAAATGACCGCCCTTTAACAT
FSIDF9J02B0GQA  AA--TTAAATTATTACCATAAGTTAAAGTAGCTTAAG-TTT-AAAGCAAAGCTCGAAGATGCTAG-----------------------------------
FSIDF9J02CCOH7  AA--TTAAATTATTACCATAAGTTAAAGTAGCTTAAG-TTT-AAAGCAAAG-------------------------------------------------
FSIDF9J02CJD4N  AA--TTAAATTATTACCATAAGTTAAAGTAGCTTAAG-TTT-AAAGCAAAGCTCGAAGATGCTAGTATGATGGTCCAAACAAATGACC------------
FSIDF9J02B0VSA  AATTTTAAATTATTAC-ATAAGTTAAAGTAGCTTAAG-TTT-AAAGCAAAGCTCGAAGATGCTAGTATGATGGTCCAAACAAATGACCGCCCTTTAACAT
FSIDF9J02CCBCF       TAAATTATTACCATAAGTTAAAGTAGCTTAAGTTTTAAAAGCAAAGCTCGAAGATGCTAGTATGATGGTCCAAACAAATGACCGCCCTTTAACAT
FSIDF9J02BXHHK       TAAATTATTACCATAAGTTAAAGTAGCTTAAG-TTT-AAAGCAAAGCTCGAAGATGCTAGTATGATGGTCCAAACAAATGACCGCCCTTTAACAT
FSIDF9J02CAPFF       TAAATTATTACCATAAGTTAAAGTAGCTTAAG-TTT-AAAGCAAAGCTCGAAGATGCTAGTATGATGGTCCAAACAAATGACCGCCCTTTAACAT

                  end of 5’-Repeat   || 3’-Repeat                                                                   //
Haouchar  2009  ---------------------------------------------------------------------------------------------------- //
M. mordax #1D   CAAC---CCCCTCAATCAAACATCAAAGAAAGAGAATTAGAATCTCTATTACTAGGCCCCCCCCAACAACAGTATTTTTTAATTATTACCATAAGTTAAA //
FSIDF9J02B0VSA  CAACCCCCCCC-CAATCAAACATCAAAGAAAGAGAATTAGAATCTCTATTACTA
FSIDF9J02BXHHK  CAACCCCC----A
FSIDF9J02CAPFF  CAACCCCC
```

Appendix 5. Identification of 454 sequencing reads from the Haouchar (2009) study that confirm the presence of NCIII as found in *M. mordax* #1D by Sanger sequencing in this study. The seven 454 sequencing reads (rows that begin with FSIDF9J02) of Haouchar (2009) are aligned to the homologous region of *M. mordax* (Same study, row that begins with Haouchar 2009) and *M. mordax* #1D (this study) . Gene regions are indicated by the coloured bar above the top sequence and are as annotated. The orange bar indicates NCIII and its comprising repeat units are annotated as 5’ and 3’ as described in the text (Section 3.2.4). The four sequence components that comprise each repeat unit of NCIII are highlighted in the *M. mordax* #1D sequence as follows: Pseudo Pro; IGN; PseudoPhe; and RES (abbreviated as in Section 3.2.4). Note: L-strand (5’ – 3’) is shown for simplicity. Dashes indicate deletions. // indicates continuation of the sequence/region. Sequences are shown in blocks of 100 bp.

Appendix 6.

ATP6



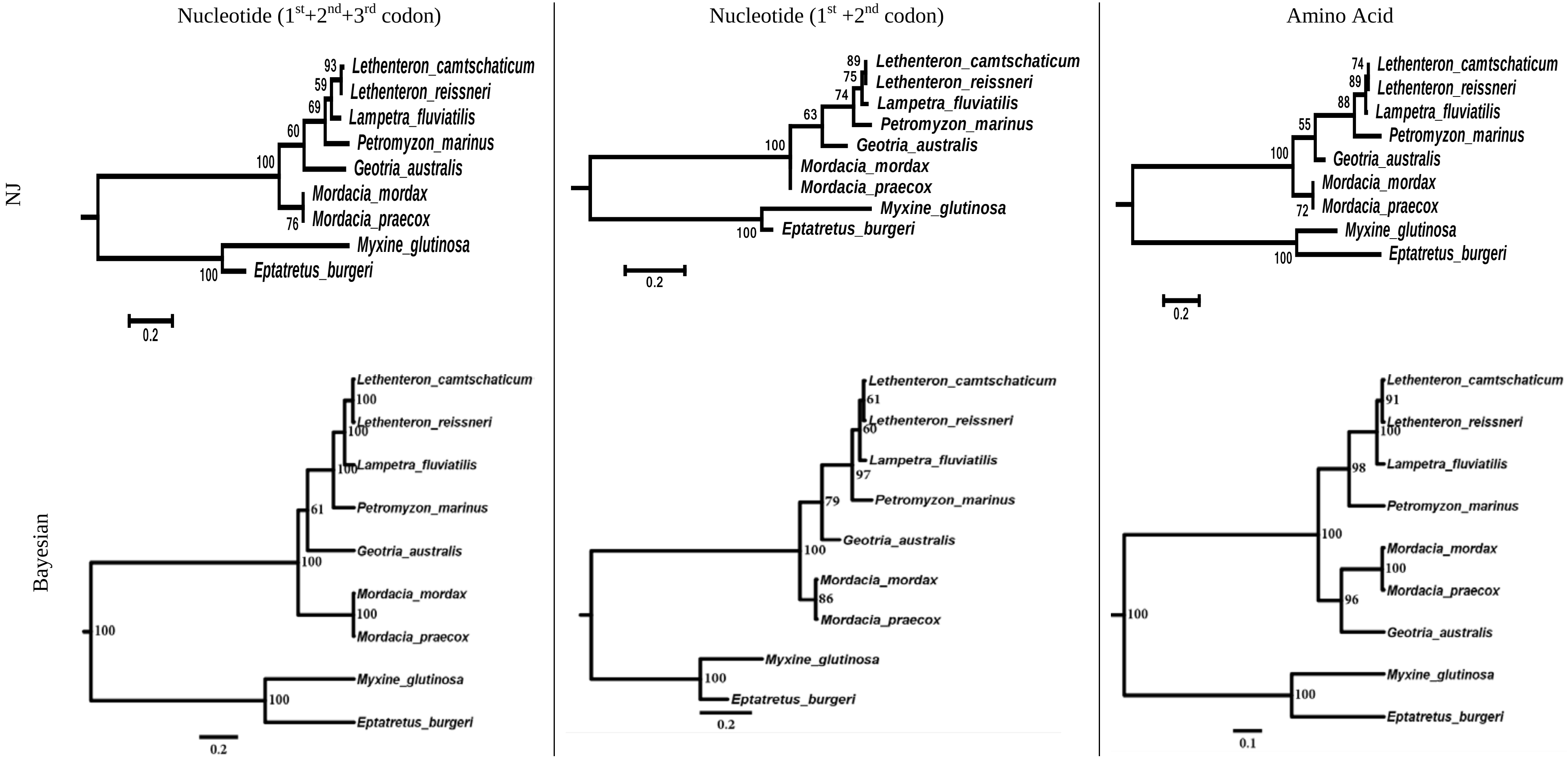
Nucleotide (1st+2nd+3rd codon)

Nucleotide (1st +2nd codon)

Amino Acid

Appendix 6 continued.

ATP8



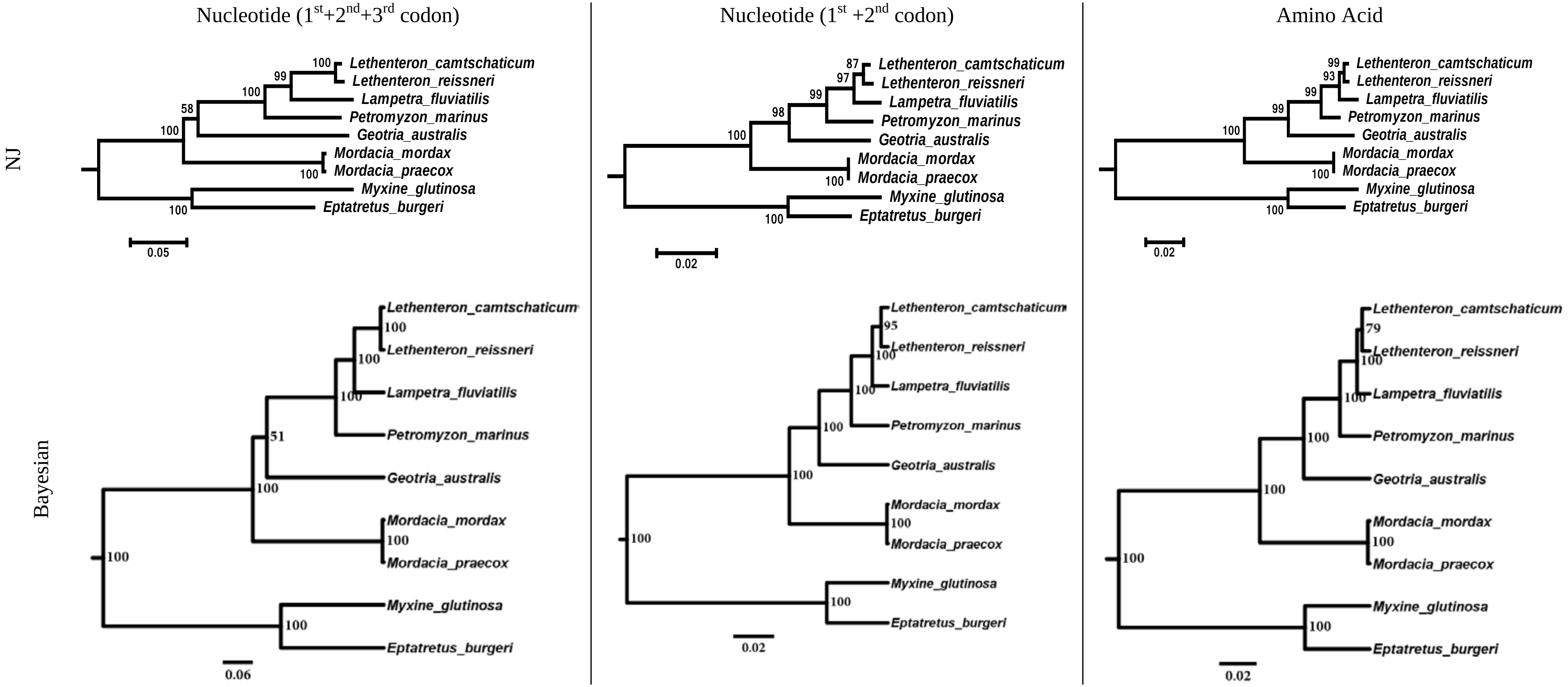
Nucleotide (1st+2nd+3rd codon)

Nucleotide (1st +2nd codon)

Amino Acid

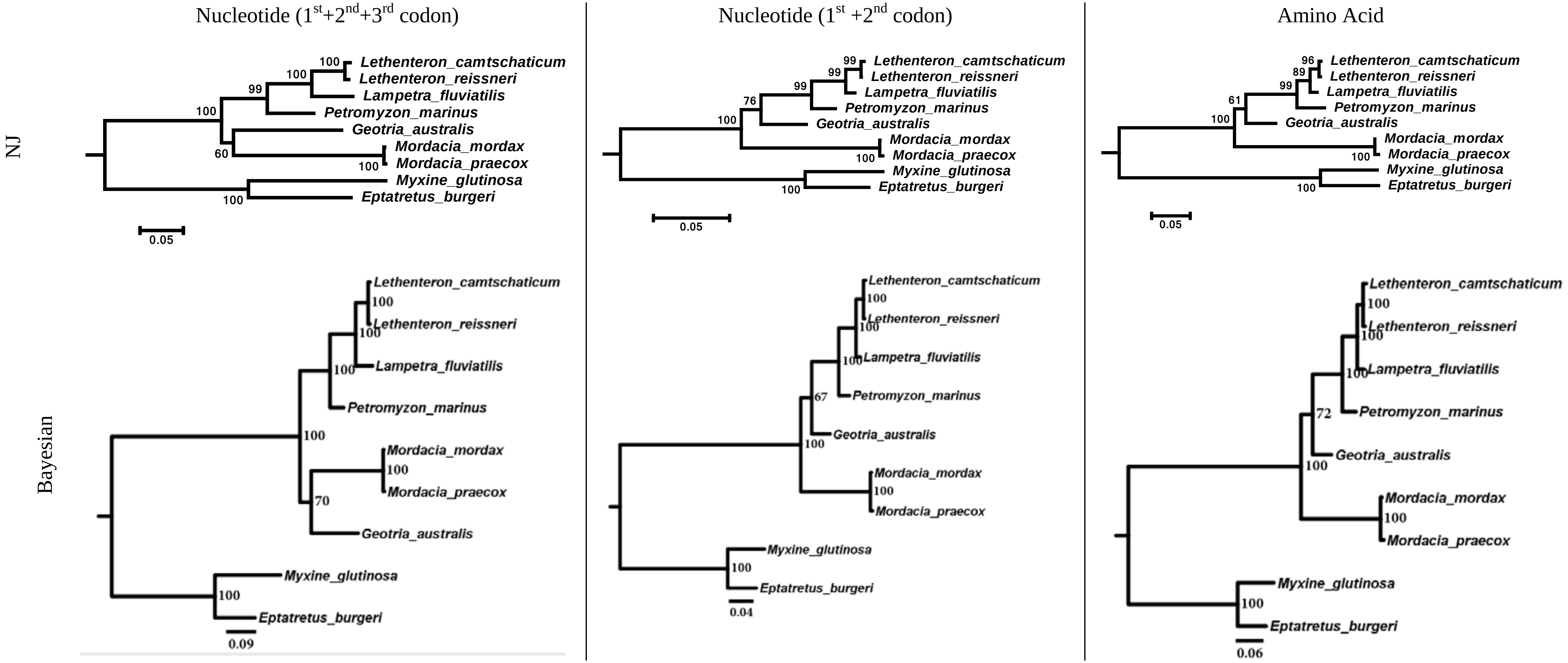
Appendix 6 continued.

COI



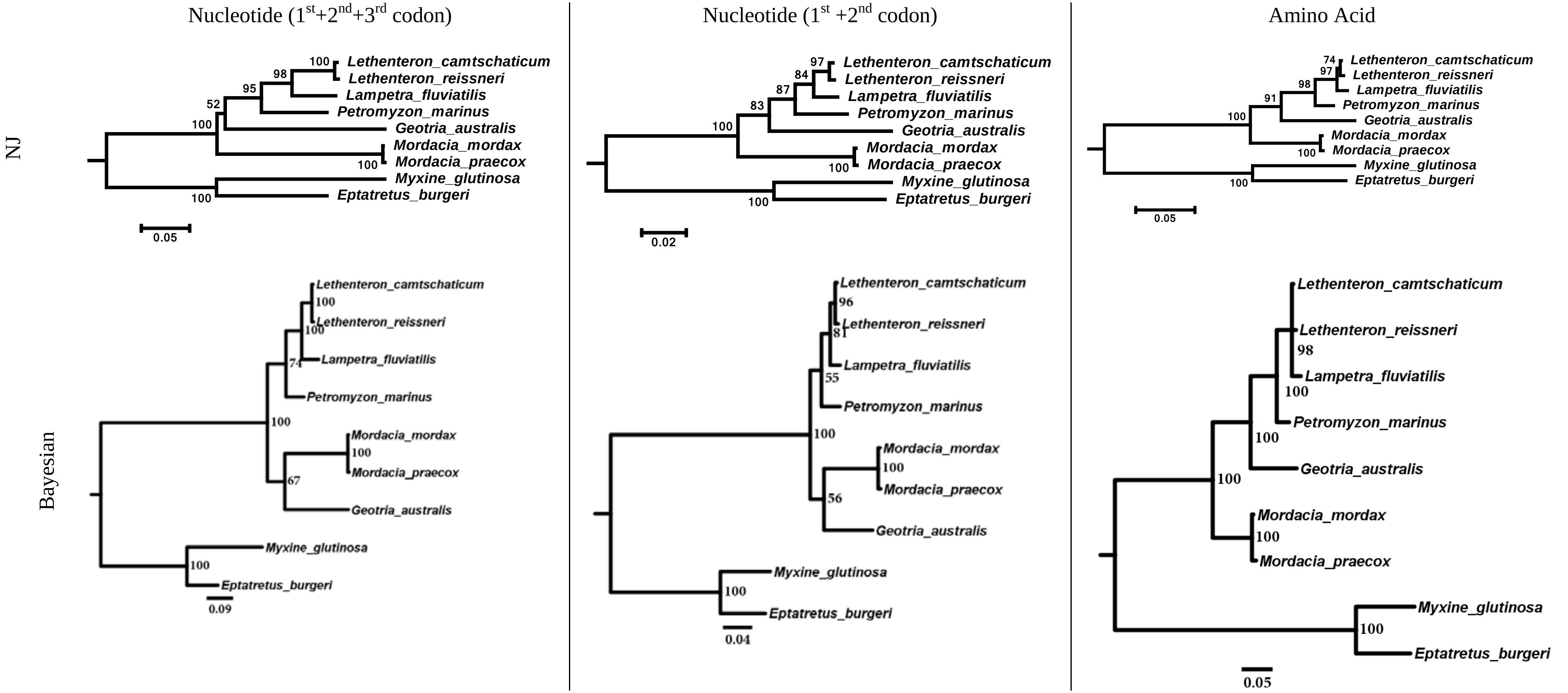
Appendix 6 continued.

COII



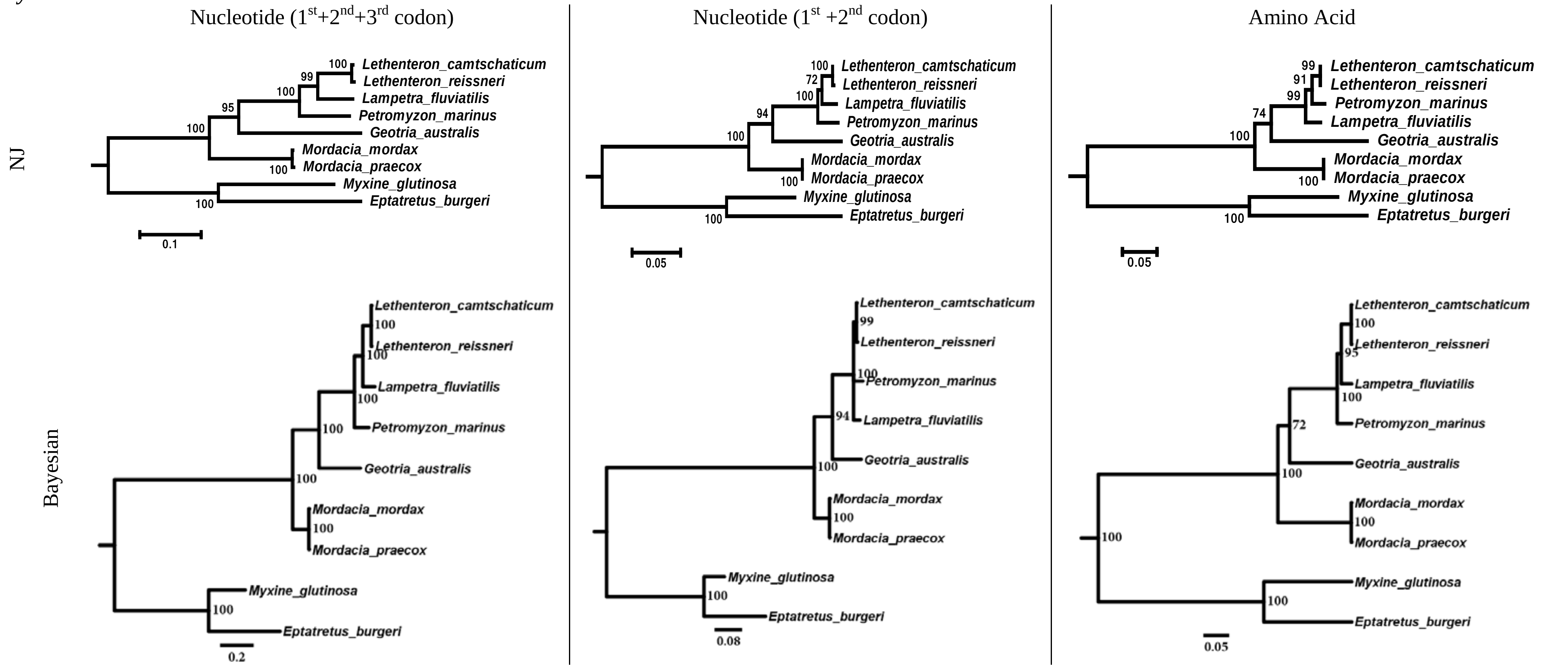
Appendix 6 continued.

COIII



Appendix 6 continued.

Cyt b



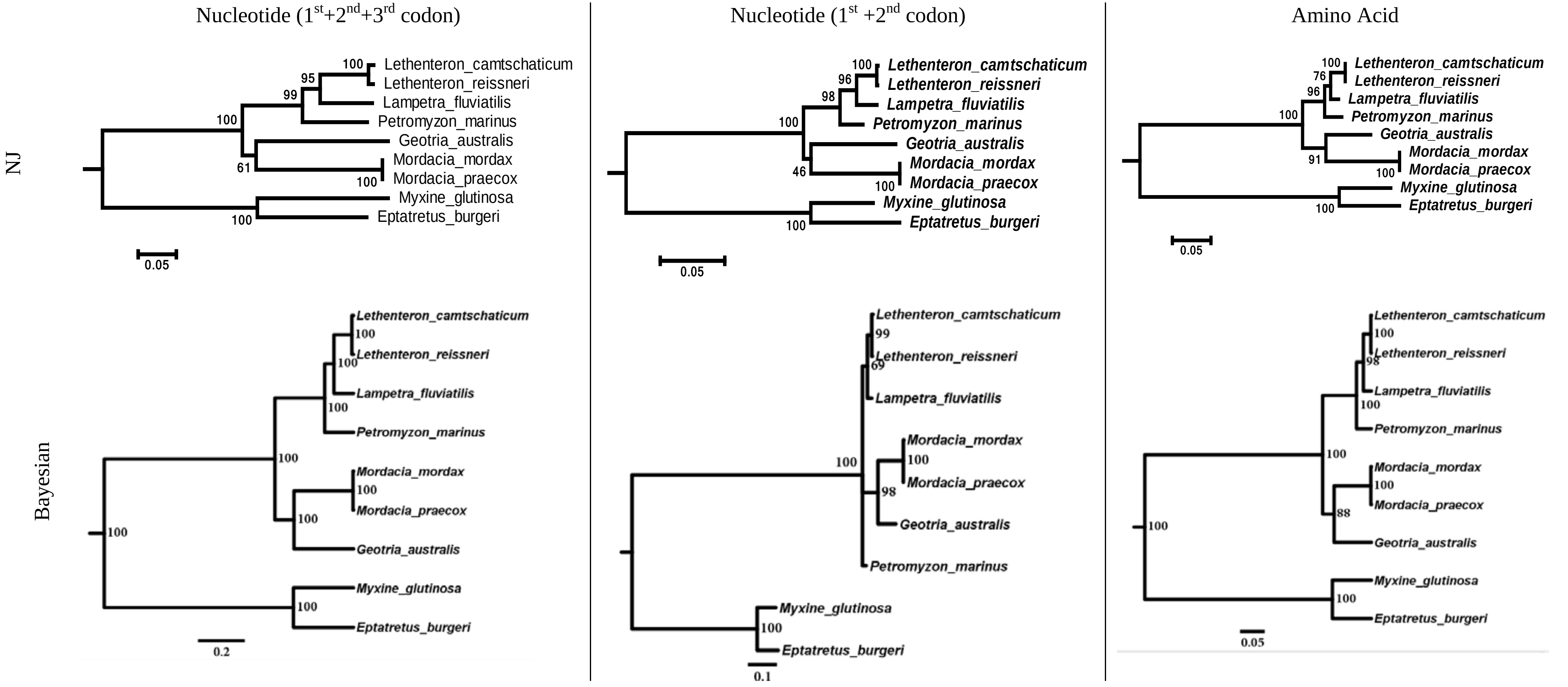
Nucleotide ($1^{st}+2^{nd}+3^{rd}$ codon)

Nucleotide ($1^{st}+2^{nd}$ codon)

Amino Acid

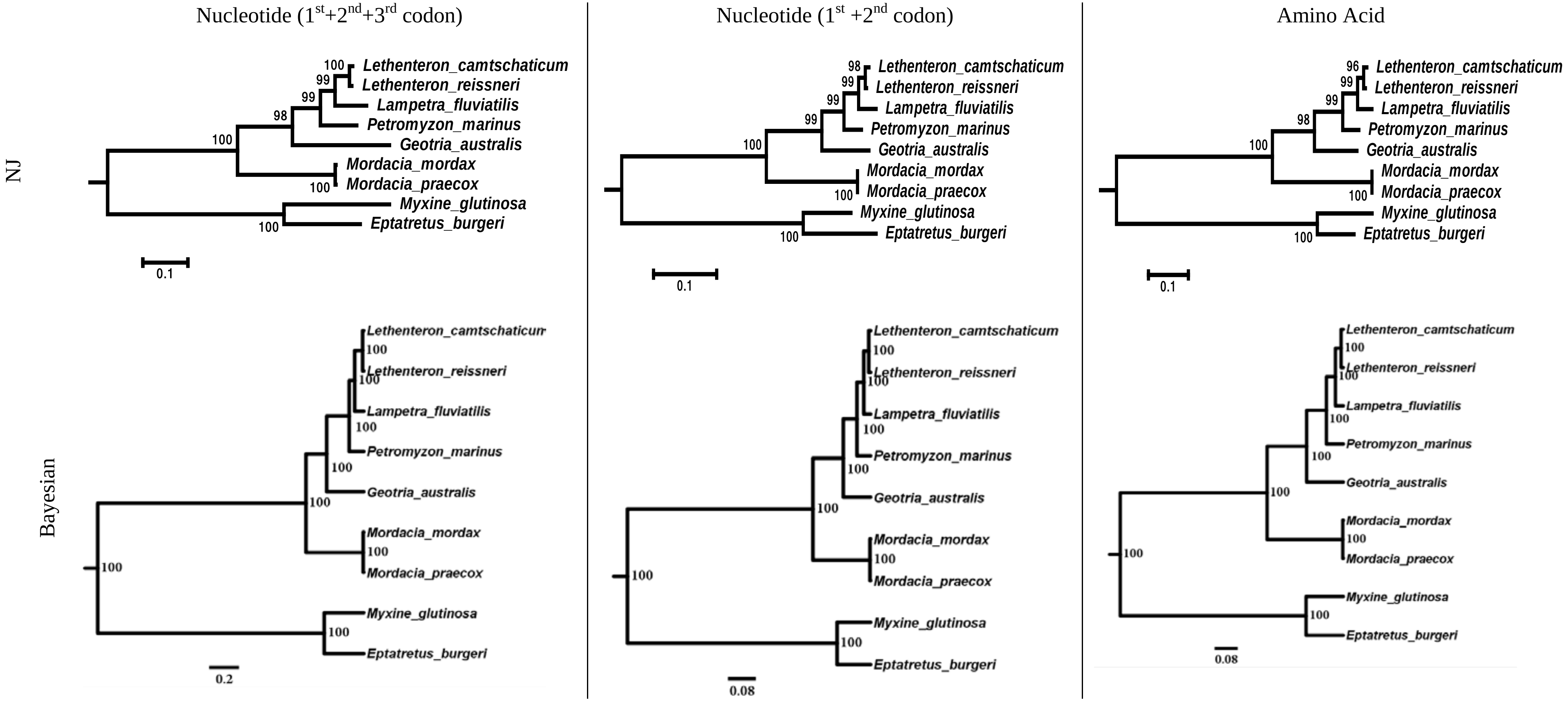
Appendix 6 continued.

ND1



Appendix 6 continued.

ND2



Nucleotide (1st+2nd+3rd codon)

Nucleotide (1st +2nd codon)

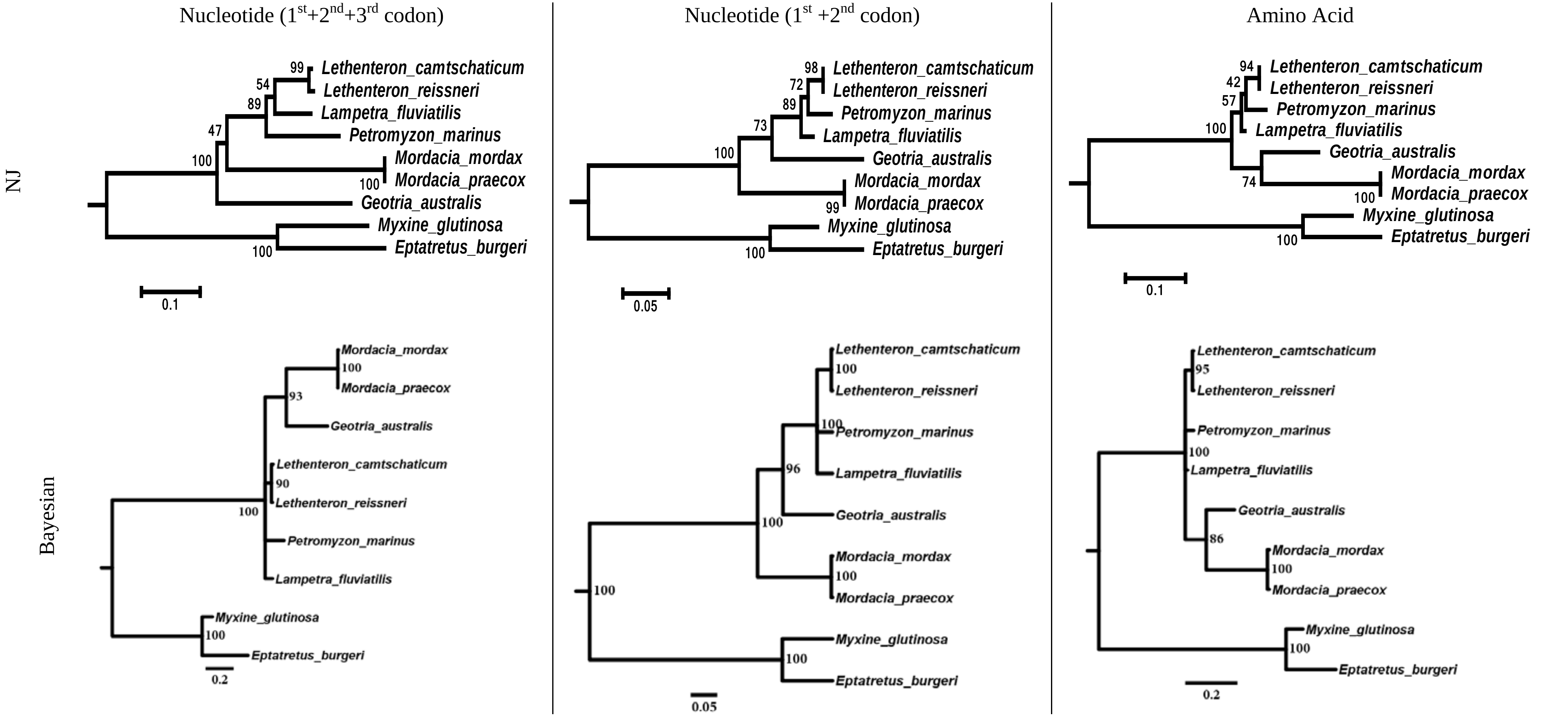
Amino Acid

NJ

Bayesian

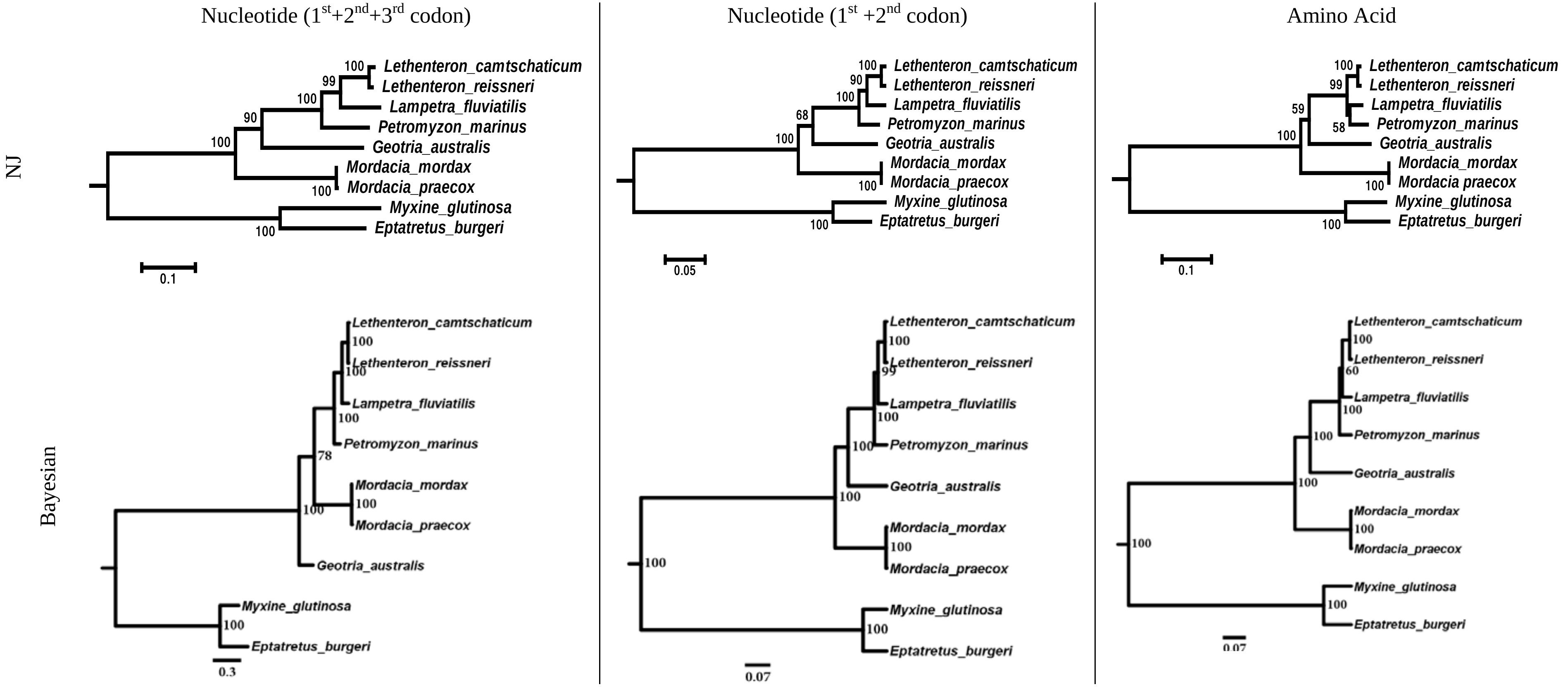
Appendix 6 continued.

ND3



Appendix 6 continued.

ND4



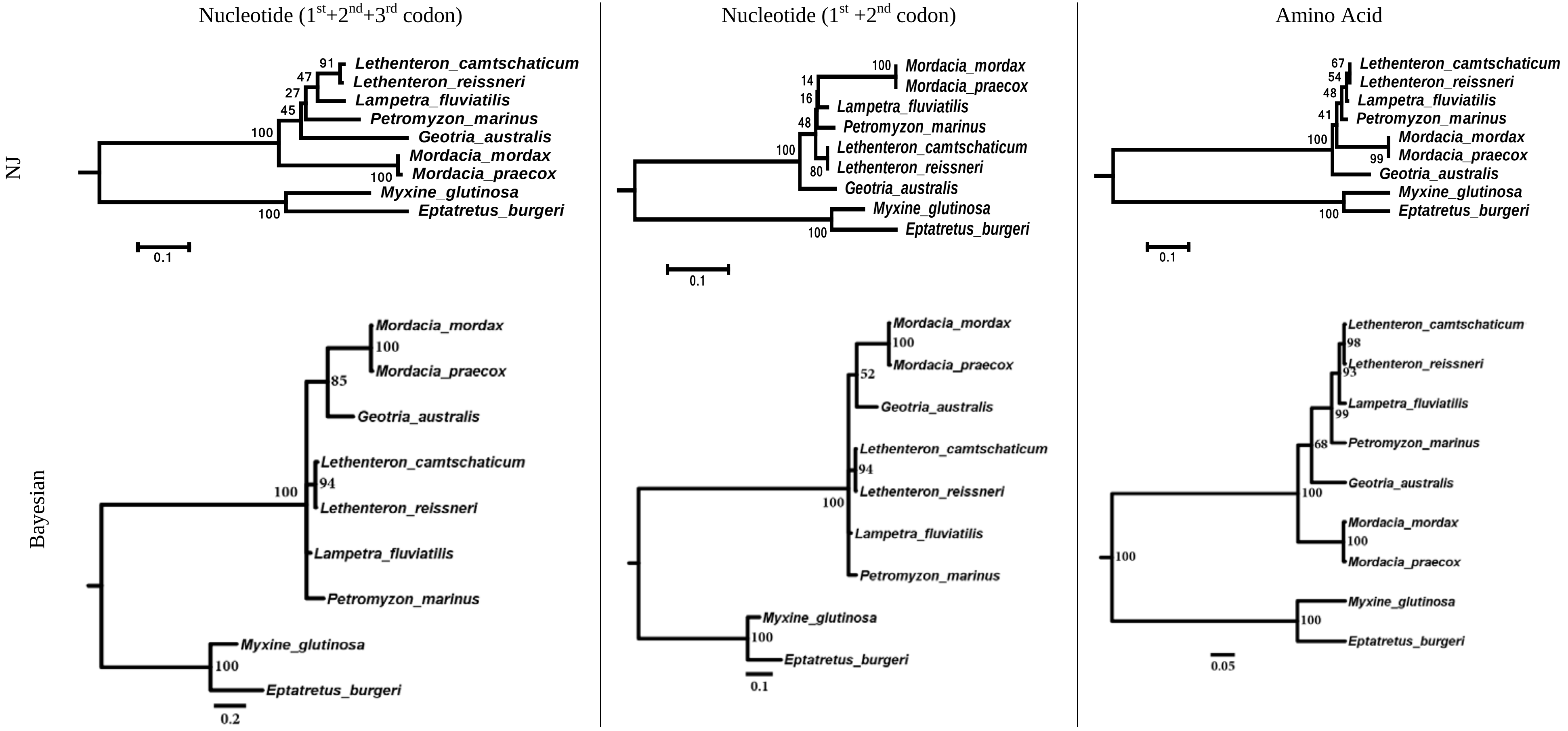
Nucleotide (1st+2nd+3rd codon)

Nucleotide (1st +2nd codon)

Amino Acid

Appendix 6 continued.

ND4L



Nucleotide (1st+2nd+3rd codon)

Nucleotide (1st +2nd codon)

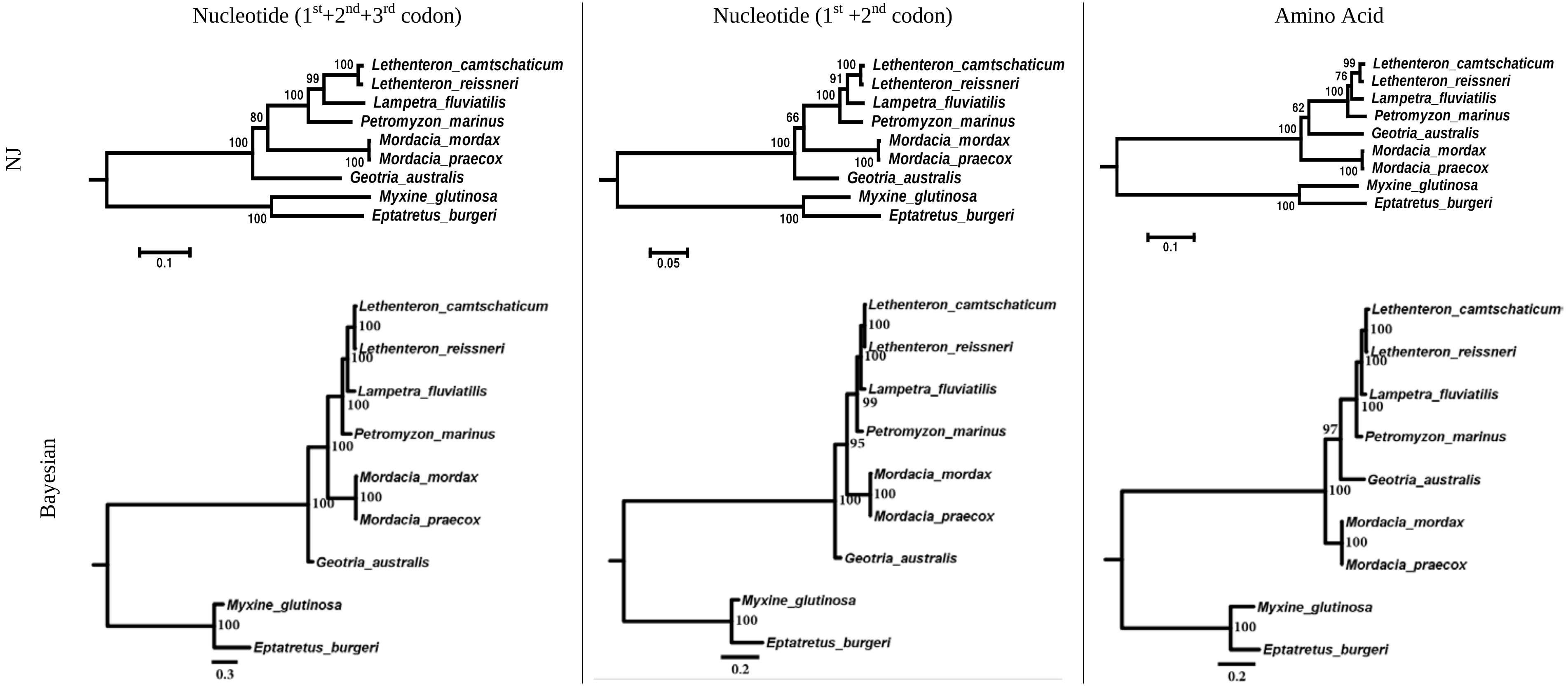
Amino Acid

NJ

Bayesian

Appendix 6 continued.

ND5



Nucleotide (1st+2nd+3rd codon)

Nucleotide (1st +2nd codon)

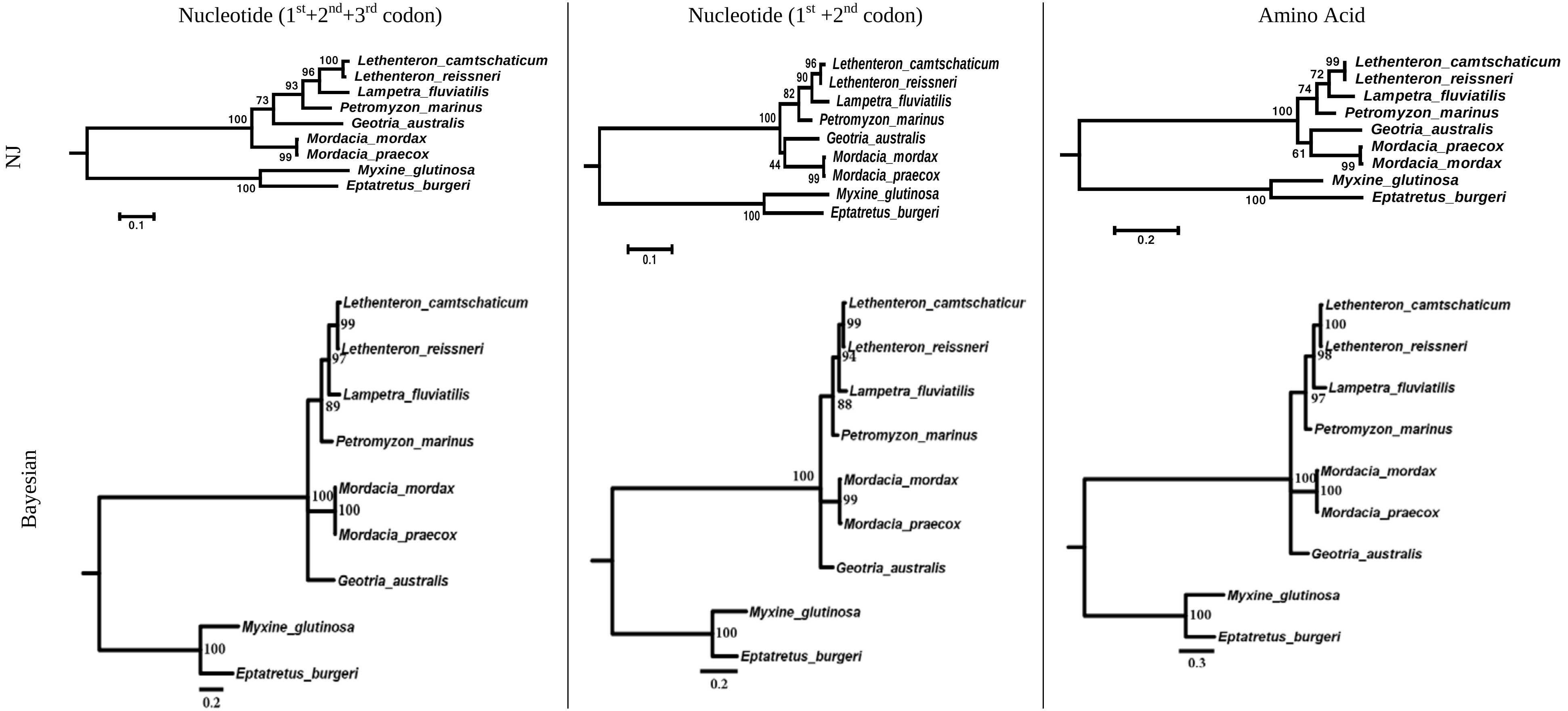
Amino Acid

NJ

Bayesian

Appendix 6 continued.

ND6



Nucleotide (1st+2nd+3rd codon)

Nucleotide (1st +2nd codon)

Amino Acid

NJ

Bayesian

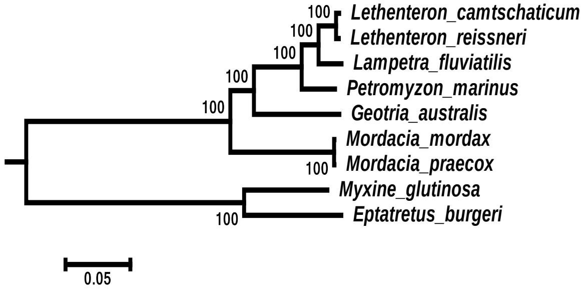
Appendix 6 continued.

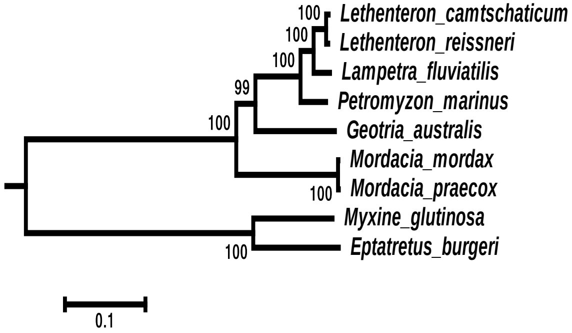
CONCAT

Nucleotide (1^{st} +2^{nd} codon)

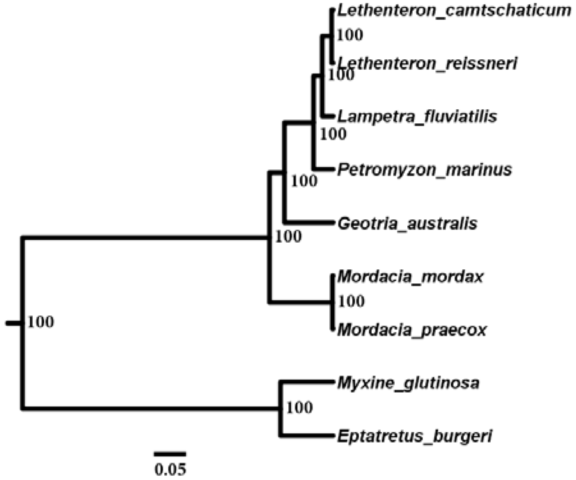
Amino Acid

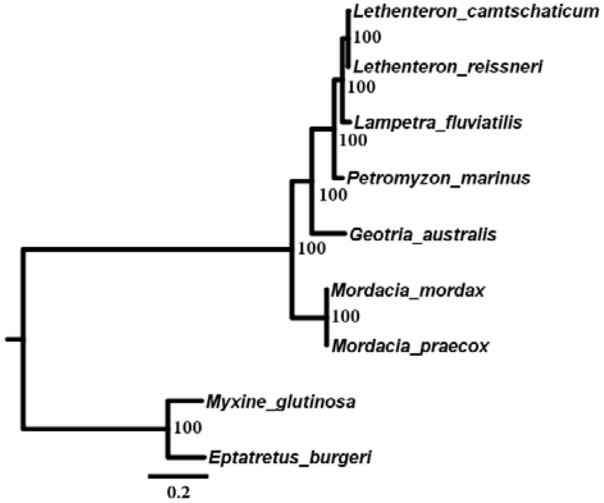
NJ





Bayesian





Appendix 7.

Calibrations used for the temporally-calibrated phylogenetic reconstruction of 29 taxa (Figure 35) with justifications for priors used.

Node	Prior bounds	Reference / Justification
Deuterostomia	Uniform prior (518.0-635.0 Ma)	(Benton *et al.* 2009)
Vertebrata (inc. hagfish)	Uniform prior (460.6-581.0 Ma)	(Benton *et al.* 2009)
Amniota	Normally distributed prior (95% prior bounds 305.0-330.4 Ma)	(Phillips *et al.* 2009)
Clupeocephala	Normally distributed prior (95% prior bounds 149.85-165.2 Ma)	(Benton *et al.* 2009)
Mammalia	Normally distributed prior (95% prior bounds 162.9-229 Ma)	(Benton *et al.* 2009) Maximum bound modified to cover Late Triassic putative mammals and haramyids.
Osteichthyes	Normally distributed prior (95% prior bounds 416-421.5 Ma)	(Benton *et al.* 2009)
Sauropsida	Lognormal prior (minimum 255.9, mean 265 and 97.5% soft maximum 299.8Ma)	(Phillips *et al.* 2009)
Cyclostomata	Uniform prior (360-581.0 Ma)	Minimum based on the oldest fossil Lamprey *Priscomyzon riniensis* (Gess *et al.* 2006) and the maximum being the upper bound for Vertebrata (see above).
Archosauria	Normally distributed prior (95% prior bounds 239-250.4 Ma)	(Benton *et al.* 2009)
Amphibia	Uniform prior (270.0-313.0 Ma)	Minimum age being the earliest well accepted amphibian crown fossil, *Gerobatrachus* (Anderson *et al.* 2008) and the maximum being the oldest age of Moscovian stage, which is well sampled but contains no putative crown amphibians.

Appendix 8.

List of the 29 taxa used for the temporally-calibrated phylogenetic reconstruction (Figure 35).

Outgroup	GenBank ID
Cucumaria miniata	AY182376
Florometra serratissima	AF049132
Pisaster ochraceus	X55514
Asymmetron lucayanum	NC_006464
Branchiostoma floridae	NC_000834
Homo sapiens	D38112
Ornithorhynchus anatinus	NC_000891
Corvus frugilegus	NC_002069
Alligator mississippiensis	NC_001922
Plestiodon egregius	AB016606
Mertensiella caucasica	EU880319
Typhlonectes natans	AF154051
Iguana iguana	NC_002793
Mustelus manazo	AB015962
Chimaera monstrosa	NC_003136
Cyprinus carpio	X61010
Oncorhynchus mykiss	L29771
Neoceratodus forsteri	AF302933
Lepidosiren paradoxa	AF302934
Protopterus annectens	L42813
Myxine glutinosa	NC_002639
Eptatretus burgeri	NC_002807
Lampetra fluviatilis	Y18683
Lethenteron camtschaticum	KC353468
Lethenteron reissneri	KC353466
Petromyzon marinus	PMU11880
Mordacia mordax	This study – *M. mordax* #25
Mordacia praecox	This study – *M. praecox* #3
Geotria australis	Honours thesis of Milton (2003) and Riddington (2007)*

* *COIII* gene error (see Appendix 3) was fixed prior to being sent to Matthew Phillips for analysis.