PhD in Aquatic Biosciences // No. 25 - 2018

Molecular and evolutionary characterization of the Atlantic cod mitochondrial genome

Tor Erik Jørgensen

FACULTY OF BIOSCIENCES AND AQUACULTURE

NORD University

www.nord.no

Molecular and evolutionary characterization of the Atlantic cod mitochondrial genome

Tor Erik Jørgensen

A thesis for the degree of
Philosophiae Doctor (PhD)

PhD in Aquatic Biosciences no. 25 (2018)
Faculty of Biosciences and Aquaculture

PhD in Aquatic Biosciences no. 25 (2018)

Tor Erik Jørgensen
Molecular and evolutionary characterization of the Atlantic cod mitochondrial genome

ISBN: 978-82-93165-24-8

Print: Trykkeriet NORD

Nord University
N-8049 Bodø
Tel: +47 75 51 72 00
www.nord.no

Preface

This dissertation is submitted in the fulfillment of the requirements for the degree of Philosophiae Doctor (PhD) at the Faculty of Biosciences and Aquaculture (FBA), Nord University, Bodø, Norway. The PhD project was carried out from September 2009 to May 2018, mostly as a part time project. This work was funded by Nord University, UiT: The Arctic University of Norway, and the Research Council of Norway (GenoFisk), and supervised by Professor Steinar D. Johansen.

The project team included the following members:

Tor Erik Jørgensen, Siv.ing, FBA, Nord University: PhD student

Steinar D. Johansen, Professor, FBA, Nord University, and Faculty of Health Sciences, UiT: primary supervisor

Truls Moum, Professor, FBA, Nord University: co-supervisor

Jarle T. Nordeide, Professor, FBA, Nord University: co-supervisor

Tor Erik Jørgensen

Bodø, June 2018

Acknowledgements

Looking back, the first step towards my PhD started probably as early as 2004 when I was interviewed by Professor Steinar D. Johansen for a position as a technician in molecular biology. Since then, many people have been involved, shared their knowledge and given valuable scientific advices. They have all earned my gratitude, but mentioning all here would take too much time and space.

First and foremost I want to thank my primary supervisor Professor Steinar D. Johansen for the continuous support of my PhD study, for his patience, motivation, enthusiasm and immense knowledge. His guidance has been invaluable and has helped me with the research and writing of this thesis. I could not imagine having a better mentor and supervisor for my PhD study.

I want to thank my co-supervisors Professor Truls Moum and Professor Jarle T. Nordeide for their support, patience and always being available for discussions and questions. Their guidance, advices and comments have been invaluable and greatly appreciated.

Special thanks to Bård Ove Karlsen, for his friendship and collaboration through all these years. We have worked together on several projects, including the major undertaking of getting the DeepSeq facility up and running. Your enthusiasm for research and technology is highly appreciated and rewarding.

I am thankful to all lab colleagues in the Genomics group in Bodø and the RNA group in Tromsø. I want to thank Åse Emblem and Teshome T. Bizuaeyehu in particular for many years of friendship and help.

I want to thank all the co-authors that have taken part in the research of this thesis, especially Lars Martin Jakt for interesting discussions and valuable comments.

I am grateful to my mother and father for their encouragement and support throughout my study. Most importantly, I thank my daughter Runa, and my wife May-Wenche, for all their love, support and endless patience. We have been through some

very challenging, but rewarding years. Adoption, moving, renovations, and my long-lasting PhD project. I somehow feel life will be a bit simpler and less hectic after the completion of my PhD.

Tor Erik Jørgensen Bodø, June 2018

Table of contents

List of figures and tables

List of papers

List of papers included in this thesis:

I. Jørgensen TE, Karlsen BO, Emblem Å, Breines R, Andreassen M, Rounge TB, Nederbragt AJ, Jakobsen KS, Nymark M, Ursvik A, Coucheron DH, Jakt LM, Nordeide JT, Moum TB, Johansen SD (2018). **Mitochondrial genome variation of Atlantic cod.** Manuscript accepted for publication in BMC Research Notes.

II. Karlsen BO, Emblem Å, Jørgensen TE, Klingan KA, Nordeide JT, Moum T, Johansen SD (2014). **Mitogenome sequence variation in migratory and stationary ecotypes of North-east Atlantic cod**. Mar Genomics, 15: 103-108.

III. Jørgensen TE, Bakke I, Ursvik A, Andreassen M, Moum T, Johansen SD (2014). **An evolutionary preserved intergenic spacer in gadiform mitogenomes generates a long noncoding RNA.** BMC Evol Biol, 14: 182.

IV. Jørgensen TE, Karlsen BO, Emblem Å, Jakt LM, Nordeide JT, Moum TB, Johansen SD (2018). **The Atlantic cod mitochondrial long noncoding RNA lncCR-H harbours complex heteroplasmic tandem repeat motifs.** Manuscript. Submitted to Mitochondrial DNA Part A.

Abbreviations

ATP	adenosine triphosphate
CoQ	coenzyme Q
CR	control region
CSB	conserved sequence block
D-loop	displacement loop
HN	humanin
HTR	heteroplasmic tandem repeat
kb	kilo bases
lncRNA	long non-coding RNA
MDPs	mitochondrial derived peptides
mt-rRNA	mitochondrial ribosomal RNA
mt-mRNA	mitochondrial messenger RNA
mt-tRNA	mitochondrial transfer RNA
mtDNA	mitochondrial genome
MTERF1	mitochondrial termination factor 1
mtLSU	mitochondrial large subunit
mtPAP	mitochondrial polyA polymerase
mtSSU	mitochondrial small subunit
nDNA	nuclear DNA
numt	nuclear copy of mtDNA

ORF	open reading frame
OxPhos	oxidative phosphorylation
POLRMT	mitochondrial (specific) RNA polymerase
ROS	reactive oxygen species
RNase Z	endoribonuclease
TAS	termination associated sequence
TFAM	mitochondrial transcription factor A
TFB(1/2)M	mitochondrial transcription factor B, isoform 1 or 2
tRFs	tRNA derived fragments
UQ	ubiquinone
VNTR	variable numbers of tandem repeat

Abstract

The mitochondrion is a cytoplasmic organelle in the eukaryotic cell, and is the main supplier of energy carrying molecule adenosine triphosphate (ATP) via the metabolic pathway machinery called oxidative phosphorylation (OxPhos). The mitochondrion contains its own small circular genome (mtDNA), which in vertebrates is highly conserved in structure, gene content and function. It encodes 37 conventional gene products essential for OxPhos and translation, and is typically 16 – 17 kb in size. Organization of the Atlantic cod (*Gadus morhua*) mtDNA is similar to most other vertebrate species, but there are some distinguishing molecular hallmarks. These are the noncoding intergenic T-P spacer, located between mt-tRNA^{Thr} and mt-tRNA^{Phe}, and the heteroplasmic tandem repeat (HTR) array located in the 5' domain of the noncoding control region (CR). In this study, we investigated molecular features of the mitochondrial genome, more precisely mtDNA sequence variation, mtDNA heteroplasmy and mitochondrial noncoding RNAs, in the Atlantic cod and related codfishes, based on large-scale DNA and RNA sequencing.

We analysed complete mtDNA sequences of > 200 specimens of Atlantic cod from the eastern and western parts of the species distribution area in the North Atlantic Ocean. In one study (Paper I) we analysed complete Atlantic cod mtDNAs at the individual level, and all specimens harboured a unique mitochondrial haplotype. Polymorphic sites were unevenly distributed across cod mtDNAs, but the vast majority were located within the protein coding genes. The sequence variation distribution was found to be similar to that of other fish species. Furthermore, mt-tRNA and mt-rRNA genes were overall much more conserved compared to protein coding genes. The noncoding regions showed a composite mode of sequence variation. While OriL and the central domain of CR were almost invariant, the 5' domain of CR and T-P spacer contained high sequence variation. In a second study (Paper II) we investigated mtDNA sequence variation in pooled samples of specimens representing two defined ecotypes (the Norwegian coastal cod and the North East Atlantic cod). The result was coherent

with the individual samples with regards to qualitative and quantitative variations. Phylogeographic analyses revealed no clear structuring, but population genetic parameters showed a lower nucleotide diversity for cod in the northwest compared to the stationary and migratory cod in the northeast. Furthermore, there was a lack of mitochondrial genetic differentiation between the northeast ecotypes, whereas eastern and western cod showed genetic differentiation, supporting deviating population histories.

In the second part of this thesis, we investigated the structural and evolutionary features of the intergenic noncoding T-P spacer among gadiform species, and the HTR within the Atlantic cod mtDNA. Sequence analysis of T-P spacer in gadiform species (Paper III) identified a complex and size variable element ranging from 16 bp to 532 bp, and the T-P spacer was found to be a preserved feature across all taxa, which indicates a biological function. Intraspecific variation of the T-P spacer was further assessed in 225 Atlantic cod specimens, and as much as 26 haplotypes were identified. A next-generation pyrosequencing run was performed on a polyA-enriched Atlantic cod transcriptome, and the data revealed an abundant H-strand specific long noncoding RNA (*lncCR-H*) of approximately 375 nt. The T-P spacer corresponded to the 5' part of the transcript, and terminated tail-to-tail with the L-strand specific transcript (*lncCR-L* or 7S RNA). In Atlantic cod, an HTR array of approximately 100 to 300 bp occupies the CR, and the *lncCR-H* transcript harbours this array. We performed a more detailed analysis of the HTR array in 134 Atlantic cod specimens (Paper IV). All specimens investigated harboured the HTR array and 26 different HTR array motifs of 39-42 bp were identified. The majority contained 2-5 copies of the same HTR array motif, but some contained two or more array motifs. Downstream of the HTR array we identified a heteroplasmic C-run motif consisting of 8-12 C nucleotides. This C-run appeared as a conserved CR feature among Gadidae mtDNAs, and is also an integral part of the *lncCR-H* long noncoding RNA in Atlantic cod. Mitochondrial heteroplasmy in the Atlantic cod CR is a common feature, and resulted in length heterogeneity in lncCR-H. The functional roles of *lncCR-H* and *lncCR-L* are not known, but we speculate that they

could be involved in antisense regulation of L-strand and H-strand transcription or function as precursor RNA of small regulatory RNAs.

1 Introduction

1.1 Vertebrate mitochondrial genomes

Mitochondria originate from α-proteobacteria, and produce most of the eukaryotic cell's energy in the form of ATP (Gray et al. 1999; Taanman 1999; Butow and Chen 2005; Martin et al. 2015). Essential protein components of the ATP-producing metabolic machinery of mitochondria, are encoded by the mitochondrial genome (mtDNA). Most somatic mammalian cells contain about ~1,000–10,000 copies of mtDNA, distributed among ~500–1000 mitochondria (Wiesner et al. 1992; Scarpulla 2008). A typical vertebrate mtDNA molecule (Figure 1) is circular, 16 - 17 kb in size, and maternally inherited. It typically codes for 37 conventional gene products, which all are involved in oxidative phosphorylation or translation. MtDNAs are organized and packed in protein-DNA complexes called mitochondrial nucleoids (Butow and Chen 2005). Super-resolution microscopy revealed that each mammalian nucleoid on average contain approximately 1.4 mtDNA molecules (Kukat et al. 2011). However, there are inconsistencies among estimates of mtDNA numbers depending on the methods used (Bogenhagen 2012; Kukat and Larsson 2013).

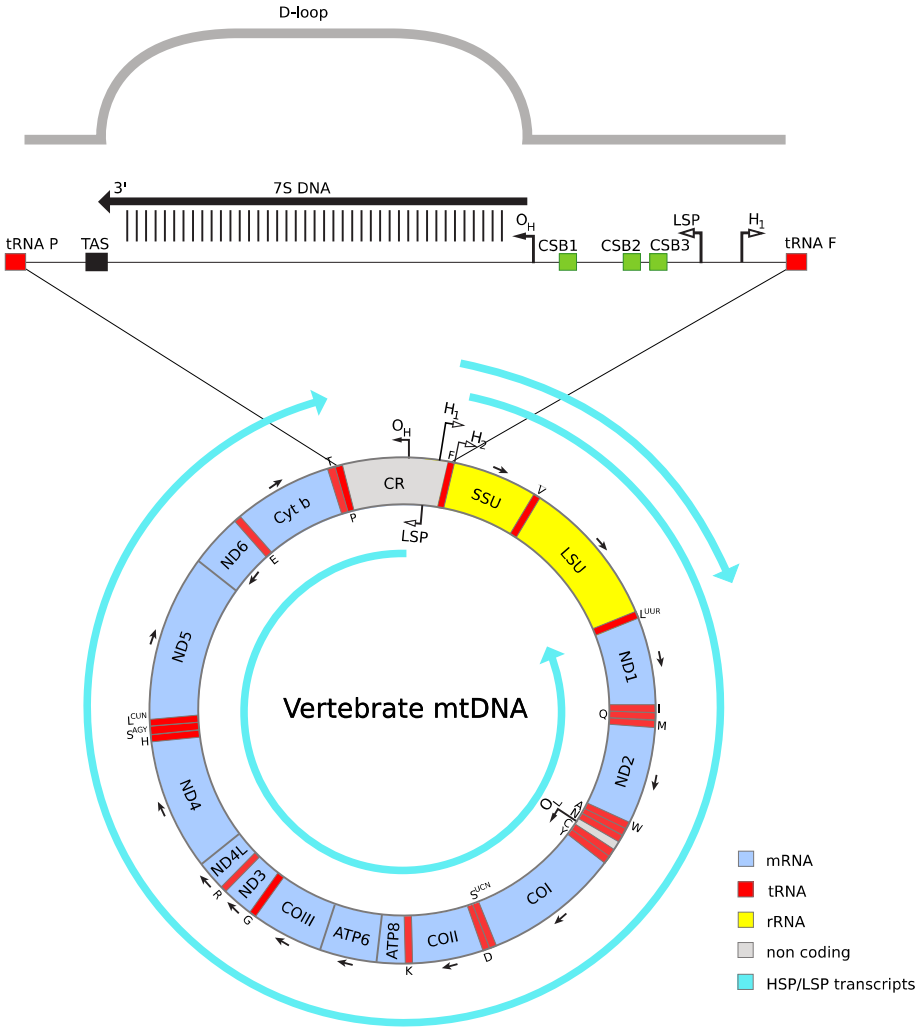


Figure 1: Typical gene content and organization of the vertebrate mitochondrial genome. Detailed version of the control region (CR) shown on top. Termination of H-strand replication at termination-associated sequences (TAS) produces the 7S DNA, which forms the stable displacement loop (D-loop) in the CR. Light-strand promoter (LSP) transcription produces ND6 mt-mRNA and eight mt-tRNAs. Heavy-strand promoter (HSP) transcription is initiated from two sites; H_1 and H_2. H_1 transcription produces a short transcript and terminates at the 3' end of LSU. H_2 transcripton initiates near the 5' end of SSU and produces a polycicstronic molecule consisting of almost the entire H-strand. The classical origins of heavy- and light-strand replication, O_H and O_L respectively. CSB, conserved sequence block/box; SSU and LSU, mitochondrial small- and large-subunit rRNA genes; COI, cytochrome c oxidase subunit I; COII,

cytochrome c oxidase subunit II; COIII, cytochrome c oxidase subunit III; Cyt b, cytochrome b; ND1, NADH dehydrogenase subunit 1; ND2, NADH dehydrogenase subunit 2; ND3, NADH dehydrogenase subunit 3; ND4, NADH dehydrogenase subunit 4; ND4L, NADH dehydrogenase subunit 4L; ND5, NADH dehydrogenase subunit 5; ND6, NADH dehydrogenase subunit 6; ATP6, ATPase subunit 6; ATP8, ATPase subunit 8.

1.1.1 Organization and gene-order conservation

Among the 37 conventional gene products encoded by the vertebrate mtDNA, 13 are hydrophobic protein subunits essential for the oxidative phosphorylation (OxPhos) system, two are ribosomal RNAs (rRNAs) of the mitochondrial ribosome (mito-ribosome), and 22 are transfer RNAs (tRNAs) necessary for the mitochondrial translation. The OxPhos system consists of five protein complexes embedded in the inner mitochondrial membrane. Four of these complexes (Complex I – IV) are part of the electron transport chain or respiratory chain, and a fifth complex, the ATP synthase, drives the synthesis of ATP (Figure 2) (DiMauro et al. 2006; Kühlbrandt 2015). The majority of the OxPhos proteins are encoded by the nuclear genome and imported into the mitochondria by protein translocases (Kühlbrandt 2015). Of the 13 mtDNA encoded proteins seven are Complex I (ND1-ND4, ND4L, ND5, ND6) subunits, one is a Complex III (Cyt b) subunit, three are Complex IV (COI-COIII) subunits and two are Complex V (ATP6, ATP8) subunits (DiMauro et al. 2006; Wu et al. 2016). For a schematic view of the OxPhos system see Figure 2.

The first complex of the electron transport chain, Complex I (NADH:UQ oxidoreductase), uses NADH from the glycolysis to harvest electrons via the membrane located ubiquinone (UQ), also known as coenzyme Q (CoQ). This electron transfer is coupled with pumping of protons from the matrix to the intermembrane space of the mitochondria. The mammalian Complex I is the largest of the five complexes, consisting of 45 (7 mtDNA and 38 nDNA encoded) different subunits, and has a mass of almost one MDa (Rich and Maréchal 2010; Fiedorczuk et al. 2016).

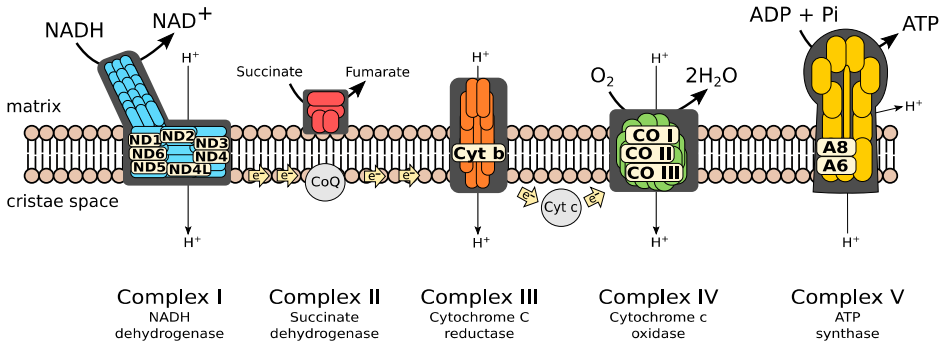


Figure 2: Mitochondrial respiratory chain complexes embedded in the inner mitochondrial membrane. ATP is generated by oxidative phosphorylation. Electrons are transferred from electron donors (NADH and succinate) to O_2 via Complex I (blue), Complex II (red), coenzyme Q (grey), Complex III (orange), cytochrome c (grey), and Complex IV (green). Complexes I, III and IV pump protons across the inner mitochondrial membrane during the electron transfer. This proton transfer generates the electrochemical potential used by Complex V (yellow) to synthetize ATP. Complex II does not pump protons. (Adapted from Davies and Daum 2013.)

Complex II (succinate:UQ oxireductase) differs from the other major complexes in some important ways. First, it contains no mtDNA encoded subunits, only nuclear encoded protein components. In addition, it does not contribute to the proton gradient. Complex II harvests electrons by catalysing the reduction of succinate to fumarate and transfer them to UQ. The mammalian Complex II is the smallest (124 kDa) enzymatic OxPhos complex, composed of four protein subunits (Sun et al. 2005; Rich and Maréchal 2010).

The mitochondrial cytochrome bc_1 complex, or Complex III, (UQ:cytochrome c oxidoreductase) accepts electrons from the Complex I and II reduced UQ, and transfer them to cytochrome c, a small and soluble electron carrier. During this process,

protons are translocated across the membrane. The 11-subunit mammalian Complex III contain only one mtDNA encoded protein, cytochrome b (Iwata et al. 1998; Rich and Maréchal 2010).

Finally, cytochrome c delivers the electrons to Complex IV, the cytochrome c oxidase, which catalyses the transfer to the final electron acceptor oxygen (O_2), forming water (H_2O). During this process, more protons are translocated from the matrix to the intermembrane space. Mammalian Complex IV contains 13 subunits where three core subunits are encoded by the mtDNA (Rich and Maréchal 2010; Wu et al. 2016).

The last step of the OxPhos chain is where ATP synthase, also called F_1F_O-ATP synthase (Complex V), uses the electrochemical potential generated by the electron transport chain to synthesize ATP. In mammalian mitochondria, the enzyme is an approximately 600 kDa membrane protein complex containing two mtDNA encoded subunits (Wittig and Schägger 2008). Complex V is composed of a catalytic and soluble F_1 portion, and a membrane bound F_O portion. The two regions have distinct functions; F_O is a proton channel and F_1 performs catalysis of ATP hydrolysis. The energy from the proton diffusion through F_O drives the F_1 synthesis of ATP (Baker et al. 2012; Junge and Nelson 2015).

The vertebrate mitochondria use their own translation machinery to synthesize the 13 OxPhos proteins. This machinery consists of mito-ribosomes (which includes mt-SSU and mt-LSU), mt-tRNAs, and several translational factors. So, mitochondrial protein synthesis relies on import of many protein components, but the essential RNA components necessary for translation appear supplied by the mitochondrion (Suzuki et al. 2011; Hällberg and Larsson 2014; Ott et al. 2016). For a more detailed description of mt-tRNA and mt-rRNA, see Section 1.2.2.

A common feature of vertebrate mtDNA is a compact gene organization with few intergenic regions, and lack of introns. The control region (CR), usually located between the $tRNA^{Phe}$ and $tRNA^{Pro}$ genes, is the major non-coding region. CR contains

transcriptional promoters, the displacement loop (D-loop), and at least one of the origins of replication (Figure 1).

The individual mtDNA strands differ in nucleotide composition. One strand is guanine poor and one is guanine rich. They are denoted as the light (L) and heavy (H) strands, respectively, due to their difference in buoyant density in a cesium chloride gradient (Berk and Clayton 1974). Most mitochondrial genes are encoded by the H-strand, including two rRNAs, 14 tRNAs and 12 protein coding genes. The L-strand encodes eight tRNAs and one protein. The mtDNA substitution rate is several times higher than that of nuclear DNA (Wallace and Chalkia 2013); in Atlantic cod about 14 times higher (**Paper II**/ Karlsen et al. 2014). This is believed to be due to the lack of protective histones, exposure to ROS (reactive oxidative/oxygen species), frequent replication, and a less efficient DNA repair machinery compared to nDNA (Ingman and Gyllensten 2006; Alexeyev et al. 2013; Payne et al. 2013). In addition, the mitochondrial genetic code differs slightly from that of nDNA (Ott et al. 2016).

To date (May 2018), mtDNAs from approximately 4,800 vertebrate species have been completely sequenced and published in the Organelle Genome Resource database of NCBI (National Center for Biotechnology Information; Table 1). The majority of these mitochondrial genomes show the typical vertebrate gene organization, encoding 37 gene products, and are 16 – 17 kb in size. Even though almost 50% of the sequenced mtDNAs taxonomically belong to fishes, the amphibians show the biggest genome size variance. One of the largest vertebrate mtDNAs known, belongs to the common rain frog (*Breviceps adspersus*), and is 28.8 kb in size (Kurabayashi and Sumida 2013). Interestingly, *Breviceps* contains the typical vertebrate 37 genes only, despite the relatively large mtDNA. The extended size is mainly due to duplications and triplications of DNA segments (Kurabayashi and Sumida 2013).

Table 1: An overview of the number of sequenced vertebrate mitochondrial genomes and their size ranges reported in NCBI's Organelle Genome Resources database (https://www.ncbi.nlm.nih.gov/genome/organelle/). Accessed May 2018.

Group	Sub group	#mt-genomes[1]	Max mtDNA size in bp	Min mtDNA size in bp[2]
Animals		7905		
Vertebrates		4889		
	Amphibians	248	28757	15897
	Birds	625	22737	15568
	Fishes	2626	24889	15564
	Mammals	1080	17755	15011
	Reptiles	310	25972	15181

1. Some species are reported more than once in the database (e.g. *Homo sapiens*, *Mus musculus* etc.) so the number of distinct species is apparently slightly less.
2. Not all mtDNAs in the NCBI Organelle Genome database are complete; the CR is quite often missing.

The mitochondria have several important cellular roles besides energy production, including molecular signaling to and from the mitochondria (Whelan and Zuckerbraun 2013). The mitochondria to nucleus signals are collectively called retrograde signals, but little is known about how information is transmitted. Recent reports have shown that vertebrate mtDNA have the potential of encoding short mitochondrial-derived peptides (MPDs) (Kim et al. 2017). The best knowns are the 21/24-amino-acid peptide Humanin (HN) and the 16-amino-acid peptide MOTS-c, both characterized in mammalian tissues and cells. The genes coding for HN and MOTS-c are located as small ORFs (open reading frames) within the mtLSU rRNA and mtSSU rRNA genes, respectively (Lee et al. 2013; Lee et al. 2015). HN was discovered over a decade ago and suggested to play biological roles in cellular stress protection, apoptosis, and metabolism (Hashimoto et al. 2001; Lee et al. 2013; Capt et al. 2016). The MOTS-c

MDP was proposed to regulate metabolic homeostasis and insulin sensitivity (Lee et al. 2015).

Segments of mtDNA will sometimes integrate in the nuclear genome, termed numts (also referred to as mitochondrial pseudogenes). Such events are well documented, and continue to play a part in the molecular evolution of eukaryotes (Richly and Leister 2004; Leister 2005). Numts have been identified in almost all of the species studied so far (Hazkani-Covo et al. 2010). In humans the rate of transposition of mitochondrial DNA sequences was estimated to $5x10^{-6}$ per germ cell per generation (Dayama et al. 2014). This process is not only contributing to adding more neutral polymorphism in the nuclear genome, but can potentially be associated with malfunctions and diseases due to the insertion of mtDNA into critical genomic positions (Leister 2005; Dayama et al. 2014). The numts range in size from short fragments to the entire length of the mtDNA, and the frequency of insertion events varies greatly among species. In vertebrates numts contribute less than 0.1% of the nuclear genome (Richly and Leister 2004; Hazkani-Covo et al. 2010). Following the release of new whole genome sequences, numts have also been identified in various fish species (Antunes and Ramos 2005). However, so far no numt sequences have been detected and reported in the Atlantic cod genome (Star et al. 2011; Karlsen et al. 2013).

1.1.2 Control region and replication

The control region (CR), located between $tRNA^{Pro}$ and $tRNA^{Phe}$, is the major non-coding region of vertebrate mtDNA, with a size of approximately 1kb. A part of the CR represents the most variable mtDNA segment (Sbisà et al. 1997). While mitochondrial intergenic regions are short, scarce and erratic, the approximately 30-bp spacer between $tRNA^{Asp}$ and $tRNA^{Cys}$, which contains the origin of L-strand synthesis (OriL), is functionally conserved in most vertebrates (Wanrooij et al. 2012; Bailey and Doherty 2017). The CR harbours the genetic control elements for the origin of H-strand replication (OriH), the transcription initiation sites for H-strand and L-strand, as well as the D-loop (Figure 1). A large part of the CR is able to generate a D-loop structure. The D-loop, which is located between OriH and the termination associated sequence (TAS),

is formed by premature termination of DNA synthesis at OriH, resulting in a triple stranded structure (Nicholls and Minczuk 2014). Here, the pre-terminated replicon (7S DNA) remains annealed to the template strand. However, the function of the 7S DNA is still largely unknown. CR also harbours the TAS domain, and conserved sequence block (CSB) domain. TAS has been proposed to be in involved in termination of newly synthesized H-strand DNA during replication (Jemt et al. 2015). The CSB domain contains protein binding sites involved in OriH, terminates transcription prematurely, and is believed to be involved in priming H-strand replication (Jemt et al. 2015; Uhler and Falkenberg 2015).

The mtDNA replisome represents a unique enzymatic machinery responsible for replicating mtDNA, distinctly different from the apparatus copying the nuclear genome (Wanrooij and Falkenberg 2010). MtDNA replication is not dependent on the cell cycle, is continuously recycled, and takes place in the mitochondrial matrix. An essential part of the replisome is DNA polymerase-γ (POLγ) that is the only replicative polymerase in mitochondria (Uhler and Falkenberg 2015; Gustafsson et al. 2016). In addition to POLγ the replisome consists of the replicative DNA helicase TWINKLE, the mitochondrial single stranded DNA-binding protein mtSSB, and the mitochondrial RNA polymerase (POLRMT), which also are essential parts of the mtDNA transcription machinery (see below) (Wanrooij and Falkenberg 2010). For a more detailed view of mtDNA replication and the replisome, see review by Gustafson et al. (2016). Three different models of how mtDNA is primed and replicated have been proposed. The most favored model is the strand-displacement model (SDM) (Wanrooij and Falkenberg 2010; Fusté et al. 2014; Gustafsson et al. 2016), according to which DNA synthesis replication is initiated at OriH. Initiation of H-strand replication is primed by short RNAs that originate from pre-terminated LSP transcripts. After initiation, the synthesis proceeds unidirectionally to produce a new H-strand. Once the replication machinery reaches the OriL, a single stranded stem-loop structure forms, from where POLRMT initiates primer synthesis. After a few nucleotides of primer synthesis POLRMT is replaced by POLγ and L-strand synthesis from OriL begins. The L-strand synthesis

proceeds in the opposite direction of H-strand synthesis. H- and L-strand synthesis will continue until each strand is replicated, and two daughter mtDNA molecules are made. Two other models for mtDNA replication have been proposed: the model of ribonucleotide incorporation throughout the lagging strand (RITOLS), and the model of strand-coupled mtDNA replication (Bailey and Doherty 2017). Recent genetic and biochemical evidence favour the strand displacement mechanism, but these issues are under debate.

1.1.3 Mitochondrial heteroplasmy

Mitochondrial heteroplasmy is defined as the existence of more than one mtDNA sequence variant or haplotype in a cell or an individual (Taylor and Turnbull 2005). The mtDNA molecules could differ by point mutations or by length. Homoplasmy is the situation when all mtDNA molecules within a cell or an individual are identical.

In a vertebrate cell there is several hundreds to thousands of copies of mtDNA, depending on the cell type and cell state (Wai et al. 2010; Kukat et al. 2011). Recent research has shown that single cells quite commonly contain different haplotypes of mtDNA (Stewart and Chinnery 2015; Rensch et al. 2016, Hedberg et al. 2018). Mitochondrial heteroplasmy may be a result of elevated mutation rate in the mitochondria, due to limited protection and repair mechanisms present in the mitochondria. High mutation rate contributes to increased diversity of mtDNA, and the generation of somatic heteroplasmy during lifetime.

The 5’ part of the CR may be highly variable both in length and nucleotide composition. The variable length of the CR is usually due to the presence of variable numbers of tandem repeats (VNTRs). These VNTRs are in many cases heteroplasmic due to point mutations or indels, and are often referred to as heteroplasmic tandem repeats (HTR). Tandem repeats in vertebrate mitochondrial CR are widespread, and the repeat number can vary extensively, even within an individual. The size of a repeat motif may range from about four bp to 200 bp, and with a copy number from two to more than one hundred (Wang et al. 2015). They are mainly located in the hyper

variable part of the CR. HTRs are probably generated by slipped-strand mispairing and/or DNA recombination. In fish, HTRs have been reported in several groups (Johansen et al. 1990; Lunt et al. 1998; Mjelle et al. 2008). Especially among flatfishes and percid species, these tandem repeats seem to be more common (Nesbø et al. 1998; Hoarau et al. 2002; Mjelle et al. 2008).

1.2 Mitochondrial transcriptomes in vertebrates

The mitochondrial transcription process in vertebrates has mainly been studied in human cells and tissues (Mercer et al. 2011, Gustafsson et al. 2016). Typically, mtDNA is transcribed as long polycistronic RNA precursors from the L-strand promoter (LSP) and the H-strand promoters (HSP1 and HSP2) located within or close to the CR (Figure 1). The transcripts encompass all coding genes of each strand. While the transcript from LSP is processed into the ND6 mt-mRNA and eight mt-tRNAs, the H-strand transcription process is more complex. The HSP1 specific transcript is highly abundant and covers both mt-rRNAs genes (encoding mtSSU and mtLSU rRNAs) as well as two mt-tRNA genes ($tRNA^{Val}$ and $tRNA^{Phe}$) (Shutt et al. 2010; Suzuki et al. 2011). The transcription products from the HSP2 transcript, on the other hand, cover both mt-rRNA, eight monocistronic and two bicistronic mt-mRNAs, and 13 mt-tRNAs (Asin-Cayuela and Gustafsson 2007).

A significant progress has been made during the past decade in defining the key components of vertebrate mitochondrial transcription, especially in mammals (Mercer et al. 2011; Rackham et al. 2012; Hällberg and Larsson 2014; Gustafsson et al. 2016; Pearce et al. 2017). The enzymatic machinery performing transcription in the mitochondrion appears completely distinct from that in the nucleus (Falkenberg et al. 2007). It requires organelle specific RNA polymerase (POLRMT), and transcription factors. POLRMT binds to specific promoter sequence elements, but cannot initiate transcription without the assistance of mitochondrial transcription factor A (TFAM), and one of the mitochondrial transcription factor B paralogues (TFB2M or TFB1M) (Hällberg and Larsson 2014). TFAM binds to the specific promoters (HSP/LSP) and initiates transcription (Falkenberg et al. 2007; Shutt et al. 2010; Gustafsson et al. 2016).

TFAM then interacts with and recruits POLRMT. Finally TFBM2 binds to POLRMT and the transcription complex is fully assembled (Gustafsson et al. 2016). Among the three transcription units only the one starting at HSP1 has an established termination site, which is located at the end of the mtLSU rRNA gene. Termination of transcription at this site is believed to be carried out by mitochondrial termination factor 1 (MTERF1). This factor binds to its 28-bp recognition site at the 3' end of $tRNA^{Leu(UUR)}$ (Asin-Cayuela and Gustafsson 2007; Hällberg and Larsson 2014). Recent research has challenged this H-strand two-transcription unit model (Terzioglu et al. 2013, Hällberg and Larsson 2014, Agaronyan et al. 2015). A knockout study of the mouse MTERF1 gene by Terzioglu et al. (2013) indicates that MTERF1 is not a key regulator of mammalian mitochondrial transcription, and the authors point out that there is no *in vivo* evidence to challenge this conclusion. Instead, the results suggested that there was only one transcription unit of H-strand genes, regulated by a single HSP. MTERF1 seems instead to terminate and prevent the L-strand transcription from producing antisense mt-rRNA transcripts.

The polycistronic transcripts need to be post-transcriptionally processed to generate mt-rRNAs, mt-tRNAs and mt-mRNAs. Central to this process are mitochondrial foci called RNA granules, where early transcript processing is believed to take place (Jourdain et al. 2013). Mt-RNA processing occurs according to the "tRNA punctuation" model by Ojala et al. (1981b). Mt-mRNAs and mt-rRNAs are usually punctuated with mt-tRNAs, and the mt-tRNAs form cloverleaf structures that act as signals for the processing machinery. The 5' cleavage of mt-tRNAs are performed by mitochondrial RNase P, consisting of three protein subunits (MRPP1-MRPP3), while the 3' ends of mt-tRNAs are cleaved by RNase Z, also known as ELAC2 (Temperley et al. 2010; Sanchez et al. 2011; Jourdain et al. 2013; Antonicka and Shoubridge 2015). As shown in Figure 1, not all mt-mRNAs are flanked by mt-tRNAs, like ATP6-COIII and CytB-ND5, but it is still unclear how these non-flanked mt-mRNAs are processed. Upon the release of the different mt-mRNAs, mt-rRNAs and mt-tRNAs, they will subsequently undergo further maturation like polyadenylation, base modification and

CAA addition (Hällberg and Larsson 2014; Rorbach and Minczuk 2012). The tRNA punctuation model appears highly conserved among mammal and fish mitochondria, exemplified by the codfishes Atlantic cod and coalfish (Coucheron et al. 2011).

1.2.1 Protein coding mRNA

Mt-mRNAs in vertebrates are excised from their respective precursor transcripts, and undergo further maturation steps (Coucheron et al. 2011; Mercer et al. 2011; Rackham et el. 2012). Compared to the maturation steps of nuclear encoded mRNAs, these events are less complex since mt-mRNAs neither contain a 5' cap modification, nor introns. In humans polyadenylation by mitochondrial polyA polymerase (mtPAP) adds about 40-50 adenosines to the 3' terminus of mt-mRNAs (Temperley et al. 2010; Rorbach and Minczuk 2012). A polyA tail has been identified in all mt-mRNAs, except for ND6 mt-mRNA (Temperley et al. 2010). An essential role of the polyadenylation is to create mt-mRNA stability, and in some cases to complete the UAA stop codons (Ojala et al. 1981b; Temperley et al. 2010). Seven UAA stop codons are created in the human mitochondria by polyadenylation RNA editing (Hällberg and Larsson 2014).

Of the 11 mature mt-mRNAs generated from vertebrate mtDNA, all are monocistronic except two bicistronic mt-mRNAs, corresponding to ND4/ND4L and ATP8/ATP6 subunits (Hällberg and Larsson 2014). Some mt-mRNAs lack 5' and 3' untranslated regions (UTR) completely, and include mt-COIII, mt-CYB, mt-ND2 and mt-ND3 (Temperley et al. 2010; Coucheron et al. 2011). Most mt-mRNAs, however, have 5' or 3' UTRs of a few nucleotides. The mt-COI has a 3' UTR of 72 nt corresponding to the complete anti-sense sequence of mt-tRNA$^{Ser(UCN)}$, mt-ND5 has an anti-sense mt-ND6 3' UTR, and mt-COII has a 3' UTR of variable length (15-24 nt) depending on cell type (Temperley et al. 2010; Coucheron et al. 2011). Eight of the 13 mt-mRNAs start codons appear at the exact 5' terminal nucleotide, while in other mt-mRNAs the translation initiates a few bases downstream the mature mt-mRNA. The most common start codon is AUG, but AUA, AUU, and GUG are also used, while UAA terminates most human mt-mRNAs (Temperley et al. 2010; Hällberg and Larsson 2014).

1.2.2 Structural rRNAs and tRNAs

All essential mt-rRNAs and mt-tRNAs involved in mitochondrial translation are encoded by the mtDNA. During and after processing from the polycistronic precursor RNAs, both mt-rRNAs and mt-tRNAs are subject to substantial post-transcriptional modifications, before they become active elements in the translation machinery in mitochondria (Suzuki et al. 2011; Rorbach and Minczuk 2012). For mt-tRNAs these modifications can be divided into two types: modifications important for RNA structure and stability, and modifications important for function (i.e. codon-anticodon recognition). Of the variety of post-transcriptional chemical modifications, base methylation is the most abundant mt-tRNA base modification (Powell et al. 2015). Nucleotide bases also undergo isomeriations, thiolations, formylations and ribosylations among others. These modifications occur at many important positions along the "clover leaf" mt-tRNA structure to ensure proper function (i.e structure, translation accuracy, correct aminoacylation) (Haute et al. 2015; Powell et al. 2015; Salinas-Giegé et al. 2015; Suzuki et al. 2011). Finally, a 3' addition of the universally conserved CAA sequence to all tRNAs occurs. This is the major post-transcriptional polymerization event, and is catalysed by the essential mitochondrial CAA adding enzyme TRNT1 (tRNA-nucleotidyltransferase 1). The addition of CAA ensures that the mt-tRNA is no longer substrate for RNase Z since it works as an antideterminant (Mohan et al. 1999).

In all organisms a mature mitochondrial ribosome (mito-ribosomes) consists of a small subunit (mtSSU) and a large subunit (mtLSU). The mtLSU is expected to be involved in catalyzing the peptidyl-transferase reaction, while mtSSU is involved in mt-mRNA binding and decoding. The mammalian mito-ribosome (55S) is composed of a small subunit made of mtSSU rRNA (ca 950 nt) and approximately 30 mitochondrial ribosomal proteins (MRP), and a large subunit made of mtLSU rRNA (ca 1600 nt) and about 50 MRPs (De Silva et al. 2015; Sharma et al. 2003). As is the case for mt-tRNAs, mt-rRNAs are subject to post-transcriptional modifications to ensure correct folding, stability and ribosome assembly (Rorbach and Minczuk 2012; Haute et al. 2015). Nine

modifications have been reported in mammalian mt-rRNAs at conserved regions. This number is very low compared to the more than 200 modifications reported in eukaryotic cytosolic rRNAs (Haute et al. 2015; Pearce et el. 2017). The mt-rRNA modifications are of 3 major types; 1) nucleotide base methylation, 2) 2'-O-ribose methylation, and 3) pseudouridylation (De Silva et al. 2015; Pearce et el. 2017). These modifications can take place co-transcriptionally, immediately after transcription, or as the mt-rRNAs are assembled into pre-ribosomal particles (De Silva et al. 2015, Haute et al. 2015; Hällberg and Larsson 2014).

The mito-ribosome has been reported to co-evolved alongside the mitochondrial genome and the OxPhos molecules (Levin et al. 2014; van der Sluis et al. 2015), and loss of the 5S rRNA species during evolution had a dramatic effect on the mtLSU (Ott et al. 2016). Recent developments within high-resolution cryo electron microscopy (EM) have provided structural evidence that a mitochondrially encoded tRNA replaces the 5S rRNA component of cytosolic ribosomes (De Silva et al. 2015; Rorbach et al. 2016). While mt-tRNA^{Val} is predominant in human and rat, mt-tRNA^{Phe} is the most common in porcine and bovine (Rorbach et al. 2016). Other mt-tRNA species have not been identified as structural rRNAs in mammalian mito-ribosomes (Rorbach et al. 2016).

1.2.3 Non-coding RNAs

Until recently very little was known about mitochondrial non-coding RNAs. Long non-coding RNAs (lncRNA), about 200 nucleotides or longer, are key regulators of vertebrate gene expression and have recently been identified and characterized in vertebrate mitochondria (Rackham et al. 2011; Fatica and Bozzoni 2014; Kumarswamy et al. 2014; Dietrich et al. 2015; Gao et al. 2017). The primary transcript encoding the H-strand genes apparently generates very few lncRNAs. However, the L-strand transcript only contains genes for seven mt-tRNAs and one mt-mRNA (ND6), separated by several thousand nucleotides of non-coding sequences released upon processing (Figure 1). L-strand specific non-coding RNA might be functional significant. The first mitochondrial lncRNA to be described was the human L-strand 7S RNA originating from within the CR (Ojala et al. 1981a; Chang and Clayton 1984). Additional human

mitochondrial-derived lncRNAs were identified more recently. These are lncND5, lncND6, lncCytB and the cardiac disease related LIPCAR lncRNAs antisense to CytB and COII mt-mRNAs (Rackham et al. 2011; Kumarswamy et al. 2014; Dietrich et al. 2015). There are apparently a large number of small non-coding RNAs generated from mitochondrial transcripts (Mercer et al. 2011; Ro et al. 2013; Ma et al. 2016). A recent study by Riggs and Podrabsky (2017) linked a set of mitochondria-derived small RNAs to hypoxia stress response in killifish (*Austrofundulus limnaeus*) embryos.

1.3 Atlantic cod and codfishes

1.3.1 Classification, taxonomy, and genomic contents

The Atlantic cod (*Gadus morhua*) is a benthopelagic fish in the Gadidae family, belonging to the order of Gadiformes (see Figure 3) (Bakke and Johansen 2005; Froese and Pauly 2017). The geographical distribution ranges from the North American west coast, throughout the North Atlantic Ocean, and all the way to the Barents Sea in the east (Nordeide et al. 2011; FAO 2017).

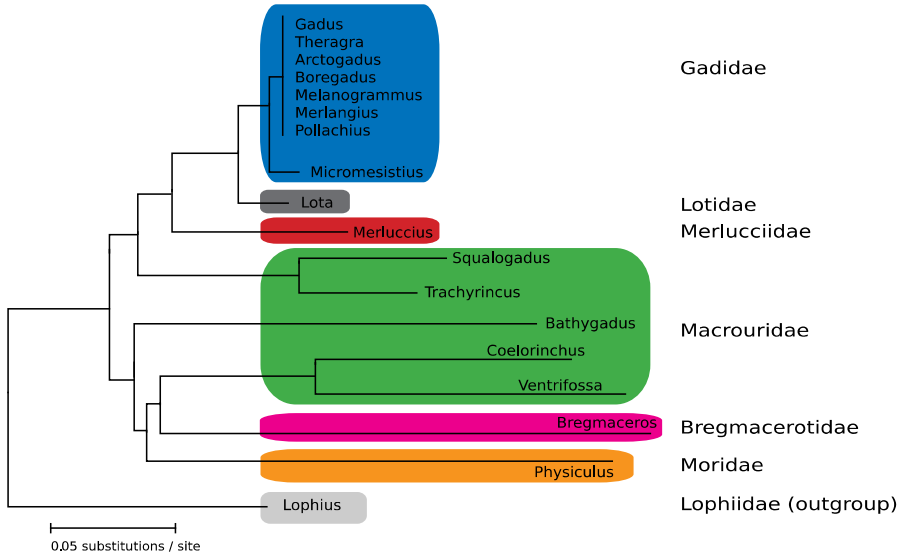


Figure 3: Gadiform phylogeny based on mitogenome-derived amino acid sequences. Phylogenetic tree based on the alignment of 13 concatenated proteins corresponding to 3814 amino acids. Different families are color-coded. Note that Macrouridae appears paraphyletic.

The haploid genome of the Atlantic cod consists of 23 nuclear chromosomes, and a mitochondrial DNA component (Johansen et al. 2009). The genome size was initially estimated to 850 Mbp, and annotation resulted in an estimate of approximately

20.000 protein-coding genes (Star et al. 2011). An improved genome assembly was recently released, in which the genome size and protein-coding gene number were estimated to 650 Mbp and 23.000, respectively (Tørresen et al. 2017). The Atlantic cod mtDNA is a 16.7 kb multicopy and circular double stranded DNA containing 37 canonical genes encoding products involved in OxPhos and mitochondrial translation (Johansen et al. 1990; Johansen and Bakke 1996). Among the 37 genes are 13 protein-coding genes, 22 tRNA genes and 2 rRNA genes (Figure 4). Structure and gene content of the Atlantic cod mtDNA, as well as the mitochondrial transcriptome is similar to that of humans and most vertebrates, with a few exceptions (Johansen et al. 1990; Johansen and Bakke 1996; Coucheron et al. 2011).

1.3.2 Importance for society

Atlantic cod is one of the most important species in commercial fisheries, and has historically been a major contributor to the economic growth and cultural identity of people living on and off the Norwegian coast. The global capture of Atlantic cod has declined from about 3.5 million tons in the 1970s to approximately 1.3 million tons the last few years (FAO FishStat; FAO 2016). This decline in capture is most likely due to overexploitation of the wild cod populations, environmental climate change, and a more strict management policy (Johansen et al. 2009; ICES 2017). Atlantic cod has also become a potential aquaculture species especially in Norway, being the largest producer of farmed Atlantic cod. This interest in farmed cod is also shared by other countries like Iceland, Faroe Islands (Denmark), UK and Canada. From early 2000 until today, the global aquaculture production of cod has grown significantly, peaking in 2009 and 2010 in excess of 22,000 tons. Norway is the main contributor with a production of approximately 21,000 tons (Johansen et al. 2009; FAO FishStat).

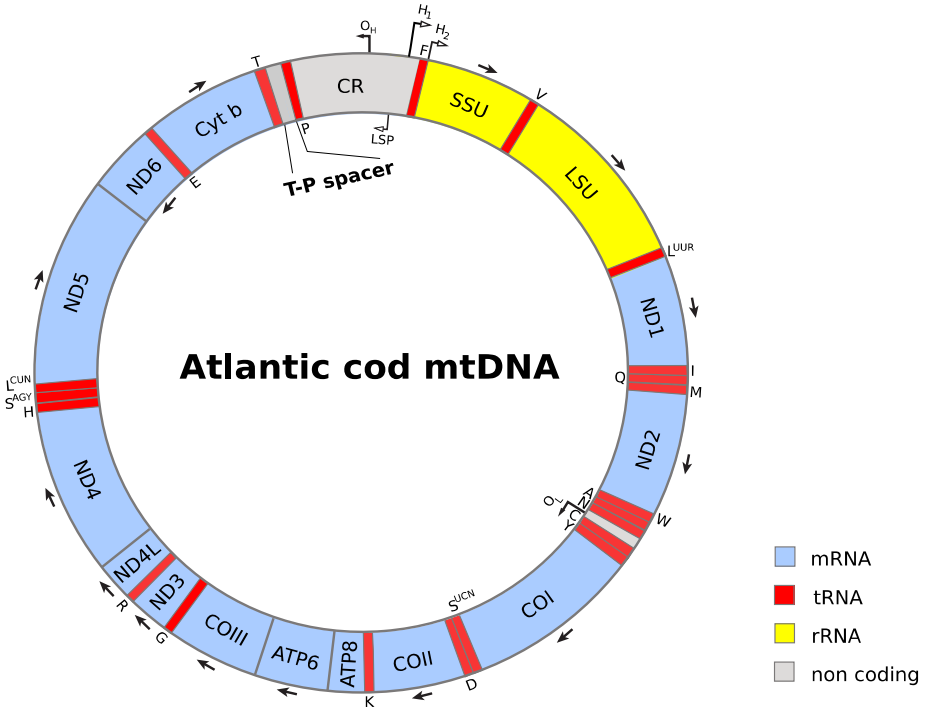


Figure 4: Organization and gene content of the Atlantic cod mitochondrial genome. Ribosomal RNA (SSU and LSU) and protein genes, except ND6, are encoded by the H-strand (clockwise arrows), and initiated from two sites (H1 and H2). The H-strand encoded mt-tRNAs are indicated with the standard one-letter symbols for amino acids and marked on the outside. Light-strand promoter (LSP) transcription produces ND6 mt-mRNA and eight mt-tRNAs. SSU and LSU, mitochondrial small- and large-subunit rRNA genes; COI, cytochrome c oxidase subunit I; COII, cytochrome c oxidase subunit II; COIII, cytochrome c oxidase subunit III; Cyt b, cytochrome b; ND1, NADH dehydrogenase subunit 1; ND2, NADH dehydrogenase subunit 2; ND3, NADH dehydrogenase subunit 3; ND4, NADH dehydrogenase subunit 4; ND4L, NADH dehydrogenase subunit 4L; ND5, NADH dehydrogenase subunit 5; ND6, NADH dehydrogenase subunit 6; ATP6, ATPase subunit 6; ATP8, ATPase subunit 8. CR, control region containing the D-loop, transcription promoters and H-strand origin of replication (OH). The mitochondrial genome size in cod is about 16.7 kb but is not fixed due to the occurrence of heteroplasmic tandem repeats within the CR and the length heteroplasmic intergenic T-P spacer.

A decent management plan for wild cod populations and a healthy cod brood stock development will help ensuring a future growth of the aquaculture industry. In part this depends on a thorough and in-depth analysis of the Atlantic cod genome and transcriptome. The Atlantic cod genome was released in 2011 (Star et al. 2011) by the Norwegian consortium GenoFisk (THE COD GENOME PROJECT 2014, see Section 1.3.1). The genomic information provides researchers with valuable genomic resources for developing genetic tools, like SNPs, QTL and ESTs.

2 Aim of study

The aim of this study was to strengthen the mitochondrial genomics resources for Atlantic cod and to investigate sequence variation and molecular features of cod mtDNA, based on large-scale DNA and RNA sequencing.

The specific objectives of the four manuscripts were as follows:

1. Determine complete mitochondrial genome sequences and analyse sequence variation from 124 individual specimens (Paper I) and two pooled populations (Paper II) of Atlantic cod.

2. Investigate structural and evolutionary features of an intergenic noncoding spacer located between the $tRNA^{Thr}$ and $tRNA^{Pro}$ genes in gadiform species (Paper III).

3. Investigate heteroplasmic tandem repeats within the Atlantic cod mitochondrial genome as an integral feature of a long noncoding RNA (Paper IV).

3 Summary of papers

Paper I: Mitochondrial genome variation of Atlantic cod.

By a combination of Sanger, Illumina and Roche 454 sequencing technologies we determined the complete mitochondrial genome sequences from 124 Atlantic cod specimens, representing major parts of the Atlantic cod's distribution area in the North Atlantic Ocean. All specimens harboured a unique mitochondrial haplotype. We identified 952 polymorphic sites across the mitochondrial genome, and 756 were confined within the 13 protein coding genes. Of these, 109 were non-synonymous substitutions. We also found 18 variable sites as indels solely distributed in structural RNA genes and non-coding regions. The protein-coding genes had a higher density of SNPs per basepair compared to the RNA genes. We performed phylogenetic analyses and calculated population genetic parameters on 156 available Atlantic cod mitogenomes. We found no clear phylogeographic structuring. The nucleotide diversity was apparently lower for cod in the northwest Atlantic compared to the migratory and stationary cod in the northeast Atlantic. There were also a lack of mitochondrial genetic differentiation between the two northeast ecotypes, but eastern and western cod were differentiated, suggesting deviating population histories.

Paper II: Mitogenome sequence variation in migratory and stationary ecotypes of northeast Atlantic cod.

We sequenced pools of individuals from two Atlantic cod ecotypes, the migratory northeast Arctic (NA) and the stationary Norwegian coastal (NC), using the SOLiD™ (Sequencing by Oligonucleotide Ligation and Detection) technology. Genomic DNA from 44 individuals of each of the NA and NC cod was used to construct one library for each of the two ecotypes. The sequencing represented 1100 and 1400 times mitochondrial genome coverage of NC and NA, respectively, and more than 25 times coverage of each individual. By using stringent settings for scoring SNPs, we found a total of 365 SNP loci. Here, 121 were shared between the ecotypes, and 151 and 93 were private within NA and NC cod, respectively. Three hundred and twenty three SNP

loci were located in the protein coding genes and 29 of these were non-synonymous. Although the migratory population had more substitutions than the stationary population, about 9% of the SNPs in both populations were non-synonymous substitutions. We determined the mitochondrial substitution rates to be about 14 times higher than that of the nuclear genome. Estimates of F_{ST} showed little differentiation of mitochondrial SNPs between the two ecotypes. The ND1 and ND2 protein coding genes showed elevated estimates compared to the remaining mitochondrial genes. The neutrality statistic Tajima's D and the nucleotide diversity parameter π were estimated. While Tajima's D was negative and slightly more so in NA than in NC, nucleotide diversities π were similar in the two ecotypes. Based on our analyses, we have shown that there is little differentiation between the mtDNA of NA and NC cod.

Paper III: An evolutionary preserved intergenic spacer in gadiform mitogenomes generates a long noncoding RNA.

Vertebrate mitochondrial genomes are small and compact. They lack introns and usually also spacing between genes. The main exceptions are the approximately 1-kb control region (CR) and the much smaller origin of light (L) strand replication (OriL). A mtDNA intergenic spacer, located between $tRNA^{Thr}$ and $tRNA^{Phe}$ (termed the T-P spacer), has been reported in some taxa, including gadiform species. We sequenced the complete 17,078 bp mitochondrial genome of European hake (*Merluccius merluccius*), as a first representative of the Merlucciidae family (order Gadiformes). The sequence revealed a large and complex intergenic T-P spacer, ranging in size from 223-532 bp due to heteroplasmic length variations. We then studied the T-P spacers from a total of 32 gadiform species, representing 8 families and 28 genera. We found T-P spacers to be a consistent feature among taxa, but significant size variations were observed. Intraspecific variation of the T-P spacer was further assessed by analyzing 225 Atlantic cod specimens. We identified 26 T-P spacer haplotypes, and the dominant haplotype (74 bp) was represented by 54%. Next-generation pyrosequencing data, from a polyA-enriched Atlantic cod transcriptome, was mapped to the Atlantic cod

mitochondrial genome. We identified two abundant transcripts covering most of the CR; a light strand specific lncRNA (*lncCR-L*) of approximately 500 nt, and a heavy strand specific lncRNA (*lncCR-H*) of approximately 375 nt. The T-P spacer was located within the 5' end of the *lncCR-H* transcript, while termination site was at the TAS. The L-strand and H-strand specific lncRNAs were found to terminate in a tail-to-tail configuration, and covered most of the CR. We conclude that the T-P spacer is evolutionary preserved within gadiform mitochondrial genomes, and part of a putative regulatory RNA.

Paper IV: The Atlantic cod mitochondrial long noncoding RNA *lncCR-H* harbours complex heteroplasmic tandem repeat motifs.

Variable numbers of tandem repeats (VNTRs) in mtDNA are in many cases heteroplasmic due to point mutations or indels, and are often referred to as heteroplasmic tandem repeats (HTR). Tandem repeats in vertebrate mitochondrial CR are widespread, and the repeat number usually vary among as well as within individuals. In Atlantic cod an approximately 80-400 bp HTR array occupies the CR, and we recently reported a long noncoding RNA (lncRNA), named *lncCR-H*, that harbours this HTR array. In the current paper we investigated and analysed in detail the HTR feature within mtDNA of 134 Atlantic cod specimens. All specimens investigated contained the HTR array, and among the 402 repeat motifs assessed in the 134 specimens, we identified 26 different HTR array motifs of 39-42 bp in size. The majority of the specimens contained 2-5 repeats of the same 40 bp HTR array motif, while 22 specimens contained two or more HTR array motifs. We further assessed HTR profiles between different tissue types of a single individual and found it to be invariant. In addition, through an Atlantic cod mating experiment, we found HTR motifs and profiles to be strictly maternally inherited. A heteroplasmic C-run motif, located downstream of the HTR, and also part of the *lncCR-H* transcript, consisted of 8-12 C nucleotides. This C-run appeared as a conserved CR feature among Gadidae mtDNAs. We conclude that mitochondrial heteroplasmy in CR, which involves a heteroplasmic HTR array and a C-run, is very common in Atlantic cod. In addition, the mtDNA of Atlantic cod is more

complex than previously recognized. These heteroplasmic features are integral parts of the *lncCR-H* long noncoding RNA, and results in significant length heterogeneities.

4 General Discussion

The main focus of this thesis has been to study molecular features of the mitochondrial genomes in codfishes, and more specifically the Atlantic cod (*Gadus morhua*). We have applied DNA sequencing as a major tool in analyzing sequence variability. In addition to traditional Sanger sequencing, we have applied various next-generation sequencing platforms. High-throughput SOLiD® sequencing represented the main technology for mitochondrial genome and small RNA sequencing. Furthermore, we also applied Ion Torrent™ (PGM™) and Illumina® as an additional support in mitochondrial genome sequencing.

Specific results and findings are commented and discussed in the individual **Papers I-IV**. Below is a summary of major findings and a general discussion that emphasis mitochondrial DNA sequence variation, mitochondrial DNA heteroplasmy, and mitochondrial noncoding RNAs.

4.1 Sequence variation

Sequence variation in the Atlantic cod mitochondrial genome was assessed in **Paper I** and **Paper II**. In **Paper I** we analysed 124 Atlantic cod specimens from the eastern and western parts of the species' distribution area in the North Atlantic Ocean. We identified 952 variable sites within the 16.7 kb mitochondrial genome, corresponding to 5.7% of the mtDNA nucleotide positions. Only 18 of those were identified as indels. Parsimony informative sites constituted 349 of the 952 sites, which corresponded to 36.7% of all variable sites.

Paper II investigated mitochondrial sequence variation in pooled samples of specimens, representing two defined ecotypes (the Norwegian Coastal cod and the North East Atlantic cod). SOLiD® deep sequencing was applied to obtain high coverage of the complete mitochondrial genome, and to identify variable sites. The individual specimens analysed in **Paper I** and the pooled specimens analysed in **Paper II** gave coherent results regarding qualitative and quantitative variation.

4.1.1 Protein-coding genes

A large part of the Atlantic cod mitochondrial genome carry protein-coding genes, encoding hydrophobic OxPhos proteins located at the mitochondrial inner membrane. The protein-coding genes were shown to contain 756 of the 952 variable sites (79.4%) in the study presented in **Paper I**, and all sites were nucleotide substitutions (no indels were observed). Variable positions were identified in all 13 protein-coding genes. Interestingly, 109 nucleotide substitutions resulted in amino acid changes (found in all 13 proteins), of which most were conservative replacements.

We found Complex VI protein-coding genes (belonging to the cytochrome oxidase complex) to be more conserved than the other mitochondrial protein-coding genes. Genes coding for subunits of the NADH dehydrogenase complex (Complex I) were less conserved. Four genes (ND2, ND4, ND5 and CytB) showed the highest density of parsimony sites between species, and thus represented interesting and informative genetic marker candidates in population- or stock assessments. The ND2 gene appeared the most informative genetic marker due to high intraspecific variability. The ND2 was found to contain > 7 times the parsimony sites per position than e.g. the Complex IV COII gene.

The sequence variation distribution pattern for protein coding genes observed in Atlantic cod was found to be similar to that of other fish species like the closely related *Theragra* spp. (Yanagimoto et al. 2004; Ursvik et al. 2007), halibuts (Mjelle et al. 2008), and zebrafish (Flynn et al. 2016), strongly indicating a conserved feature among the teleost fishes.

4.1.2 Structural RNA genes

The mitochondrial structural genes included 22 tRNA genes and two rRNA genes. The tRNA genes in Atlantic cod were highly conserved, and 4 tRNA genes ($tRNA^{Ile}$, $tRNA^{Ser(UCN)}$, $tRNA^{Ser(AGY)}$ and $tRNA^{Cys}$) were 100% identical among all specimens analysed in **Paper I** and **Paper II**. The few substitutions noted in mitochondrial tRNA genes were

all located to variable loop regions, or as compensatory base-pair changes in tRNA helical regions.

The mitochondrial rRNA genes contained less than half the number of the SNPs per nucleotide position compared to protein genes, and approximately 5 times less parsimony/variable sites per position. Ribosomal RNA genes possess a mosaic structure of conserved and variable regions, where the variable parts usually encode RNA located at the outer surface of the mito-ribosome (Kaushal et al. 2014; Greber et al. 2014). Substitutions that we identified in the mt-SSU rRNA and mt-LSU rRNA genes were mostly singletons located in variable surface regions (**Paper I**). Only seven polymorphic/parsimony sites were noted in each of the mt-SSU rRNA and mt-LSU rRNA genes. While the latter (mt-LSU rRNA) contained polymorphic/parsimony sites in three different domains (*Domains II*, *III*, *V*), six of the seven polymorphic/parsimony sites were found in a single domain (*Domain 3' major*) of the mt-SSU rRNA. Why *Domain 3' major* appeared more variable and informative as a genetic marker compared to the remaining mt-SSU rRNA, and if this feature is general or unique to Atlantic cod, is currently not known.

4.1.3 Noncoding regions

The Atlantic cod mitochondrial genome contained four noncoding intergenic regions longer than 10 bp (Johansen and Bakke 1996; Coucheron et al. 2011). The two shortest regions, the 14-bp spacer between $tRNA^{Asp}$ and COII, and the 30-bp spacer between $tRNA^{Asn}$ and $tRNA^{Cys}$, were found to be highly conserved and invariant between cod specimens (**Paper I** and **Paper III**). Both spacers form stable hairpin structures at the nucleotide level, constituting a 5' UTR of the COI mRNA and the light-strand replication origin, respectively.

More sequence variations, however, were noted within the two longest mitochondrial noncoding regions; the 74 bp T-P spacer (**Paper III**) and the approximately 900 bp CR (**Paper I**). The T-P spacer contained 16 variable sites and a 29 bp sequence duplication (in three specimens), constituting 26 haplotypes among 225

specimens assessed. The Atlantic cod sequence contained two conserved 17 bp sequence motifs (Box I and Box II), each forming a potential hairpin structure at the RNA level (**Paper III**). Interestingly, all 32 gadiform species analysed (eight families, 28 genera) contained a T-P spacer. The gadiform T-P spacers were found to be highly divergent in sequence, including heteroplasmic features in two species (*Merluccius merluccius* and *Gadiculus argenteus*; **Paper III**). However, all species contained the conserved Box motif, which strongly suggests a biological role. The T-P spacer, at least in Atlantic cod, appeared to be part of a long noncoding RNA (*lncCR-H*), and the precursor for a 25 nt RNA (tRF-1) and a 19 nt C-run specific small RNA. These features are discussed in more detail in Section 4.3.2.

The CR has a composed sequence variation feature. The 3' and central domains were found to be highly conserved between Atlantic cod specimens, suggesting functional or structural roles in mitochondrial genome function. Here, the 3' domain contains protein binding sites for various initiation and elongation factors important in mtDNA replication and transcription (see Section 1.1.2). In **Paper I**, however, we reported the 5' domain to be highly variable in sequence among specimens, and in fact the most variable domain within the mitochondrial genome (> 3 times that of the average substitution rate). This sequence variation was due to hot-spot substitution sites, homo-polymeric heterogeneity, and complex length heteroplasmy. The later feature is discussed in more detail below in Section 4.2.

4.2 Heteroplasmy

Presence of more than one mitochondrial genome haplotype in a single specimen (mitochondrial heteroplasmy) is a consistent feature in Atlantic cod. We observed different modes of mitochondrial heteroplasmy. (1) Heteroplasmic pyrimidine-rich motif (C-run) within the Atlantic cod CR reported in **Paper IV**. Here, investigated specimens contained 8-12 C-nucleotides. (2) Extensive heteroplasmic tandem repeats (HTR) within the Atlantic cod CR reported in **Paper IV.** The latter feature is discussed below in Section 4.2.1.

4.2.1 Heteroplasmic tandem repeats (HTR)

The Atlantic cod CR contained extensive HTR with an approximately 40-bp motif (**Paper IV**). Interestingly, closely related codfishes belonging to the *Gadus* and the *Theragra* genera lacked HTR arrays (Yanagimoto et al. 2004; Ursvik et al. 2007). We identified 26 sequence variants of the HTR motif among 134 Atlantic cod individuals, but one motif variant was dominating (90%). Furthermore, we detected 2-6 copies of the HTR motif in most arrays, but some harboured as many as 8 to 9 copies (**Paper IV**). These observations corroborate the conclusions by Árnason and Rand (1992) and by Kijewska et al. (2009). Finally, we found that HTR motifs and copy number profiles were highly conserved between different tissues of the same specimen, and strictly maternally inherited (**Paper IV**). Why somatic and germ line inheritance strictly preserves the HTR copy number profiles in Atlantic cod is not known.

Somewhat similar HTR features, but at a different location within the CR, were reported in European flounder (Hoarau et al. 2002) and in three species of halibut (Mjelle et al. 2008). Atlantic halibut (*Hippoglossus hippoglossus*), Pacific halibut (*H. stenolepis*) and Greenland halibut (*Reinhardtius hippoglossoides*) contained two conserved sequence motifs of 11 bp and 61 bp, which were highly heteroplasmic in copy number. Similar to what we observed in Atlantic cod (**Paper IV**), some individuals of all three halibut species, as well as the European flounder, contained HTR motifs in a compound organization that included two or more HTR motifs in the same array (Hoarau et al. 2002; Mjelle et al. 2008). Slipped strand mispairing during mtDNA replication has previously been suggested as a molecular mechanism in explaining HTR features (Levinson and Gutman 1987). However, this mechanism cannot fully explain the HTR observations in Atlantic cod and flatfishes. MtDNA recombination has been proposed as a plausible explanation in halibut and flounder, a mechanism that we favour also in Atlantic cod.

4.3 Noncoding RNA

In addition to the 37 canonical genes, and the newly proposed MDPs (Kim et al. 2017), vertebrate mitochondrial genomes encode a series of noncoding RNAs

(ncRNAs). At least eight different long ncRNAs (lncRNAs) and a larger number of small ncRNAs have been proposed (Mercer et al. 2011; Dietrich et al. 2015; Ma et al. 2016; Gao et al. 2017). Currently, none of these ncRNAs have been assigned to a specific function based on experimental evidence, but at least two lncRNAs (LIPCAR; lncCR-L) have been associated with human disease conditions (Kumarswamy et al. 2014; Hedberg et al. 2018).

4.3.1 Mitochondrial long noncoding RNAs in Atlantic cod

So far, four lncRNAs (*lncND5*, *lncCytB*, *lncCR-H* and *lncCR-L*) have been identified in the Atlantic cod mitochondrial genome (Coucheron et al. 2011; **Paper III**; **Paper IV**). Among these, the CR-specific lncRNAs, *lncCR-H* and *lncCR-L*, were investigated in the current thesis. Both lncRNAs were found to be polyadenylated, but transcribed from opposite strands within the CR (**Paper III**). In **Paper I** and **Paper III** we showed that the Atlantic cod *lncCR-L* has a substitution rate and expression level that corresponded to Complex I mRNAs (see also Coucheron et al. 2011). The *lncCR-L* corresponds to the 7S RNA in human mitochondria (Ojala et al. 1981a), and recently we showed *lncCR-L* to be differential expressed in human cancer cell lines (Hedberg et al. 2018).

The *lncCR-H*, on the other hand, was found to be variable in sequence and structure between the Atlantic cod specimens (**Paper I**, **Paper III**, **Paper IV**). A schematic overview of the *lncCR-H* RNA is presented in Figure 5 (and Figure 3 in **Paper III**). Here, the noncoding T-P spacer sequence was present at the 5' end of the RNA. In **Paper III** we reported 16 variable sites and a 29 bp sequence feature duplication in the 74 bp T-P spacer of Atlantic cod. In the *lncCR-H*, the T-P spacer is followed by an antisense $tRNA^{Pro}$, and then the HTR repeat motifs. The HTR copy number vary between 2 (80 bp) and > 8 (>320 bp), rendering the *lncCR-H* RNA highly variable in size. Downstream of the HTRs we noted a second heteroplasmic region (the C-run; Figure 5), as well as several variable single nucleotide positions, which further contributed to the size variability (**Paper IV**). In **Paper I** we reported this part of the lncCR-H as the most variable region of the mitochondrial genome. Finally, the *lncCR-H* RNA terminated in a short polyA tail at TAS. Thus, *lncCR-H* has no fixed length in Atlantic cod, varying in size

between approximately 300 and 500 nt. Interestingly, the TAS motif consisted of a perfect palindromic sequence (**TTTATACAT**ATGTATAAA). We found *lncCR-L* to terminate and poly-adenylate at the same site, but on the opposite strand (**Paper III**).

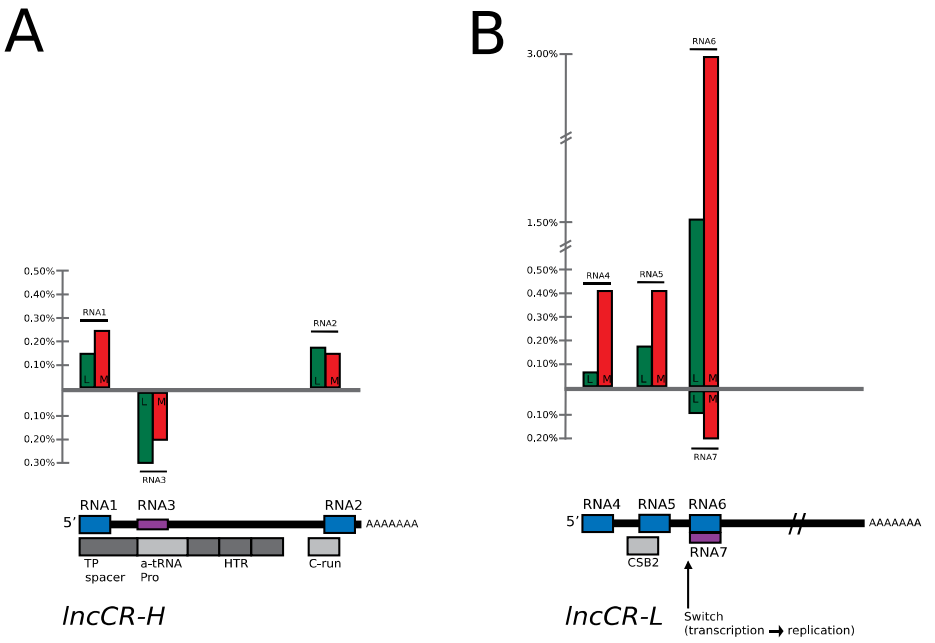


Figure 5: Abundant small RNA mapping on CR-specific lncRNA sequences. A. Mapping small RNAs on lncCR-H identified three RNA candidates (RNA1-3) present in both liver (green bars) and muscle (red bars) tissues. Note that RNA1 and 2 (blue boxes) and RNA3 (purple box) are sense and antisense, respectively, to the lncRNA. A total of 48,191 and 123,954 small RNA reads from liver and muscle, respectively, mapped to the Atlantic cod mitogenome. Fraction (%) of each specific small RNA (of total mitochondrial small RNA) is indicated. **B.** Mapping small RNAs on lncCR-L identified four RNA candidates (RNA4-7) present in both liver and muscle tissues. RNA7 is antisense to the lncRNA. HTR, heteroplasmic tandem repeat; CSB2, conserved sequence block 2. Switch, transcription to replication switch in CR (see main text for additional information).

4.3.2 Do the mitochondrial long noncoding RNAs *lncCR-L* and *lncCR-H* function as precursors of small noncoding RNAs?

In a larger survey of the human mitochondrial transcriptome a number of small ncRNAs were proposed (Mercer et al. 2011). Many of the same RNA species were also detected in rainbow trout (Ma et al. 2016) and in Atlantic cod in SOLiD™ small RNA sequencing experiments (Genomic group, Nord University – our unpublished data). In the latter experiment a sub-set of the mitochondrial tRNAs generated tRNA derived fragments (tRFs), including tRF-1, tRF-3, and tRF-5 variants. Interestingly, the great majority of the tRFs detected in Atlantic cod corresponded to H-strand tRNAs (our unpublished data).

Paper III and **Paper IV** reported mitochondrial CR-specific lncRNAs. The functional roles of *lncCR-L* and *lncCR-H* in Atlantic cod are not known. However, one possibility could be involvement in antisense regulation by controlling the overall level of L-strand and H-strand transcripts. Another, and probably more plausible explanation is that *lncCR-L* and *lncCR-H* may function as precursor RNA of small RNAs with regulatory roles.

The *lncCR-L* RNA was identified in Atlantic cod (**Paper III**; Coucheron et al. 2011), and in four independent studies of human mitochondrial RNA (Ojala et al. 1981; Mercer et al. 2011; Gao et al. 2017; Hedberg et al. 2018). SOLiD™ small RNA sequencing experiments in cod detected several abundant small RNA mapping to the mitochondrial CR. We found four small RNA candidates both in Atlantic cod liver and muscle tissue, mapping to the *lncCR-L* gene region (Figure 5). Three of the small RNAs map to the sense strand, suggesting that *lncCR-L* could be the precursor RNA. Interestingly, two of the small RNAs (RNA 5 and RNA6, Figure 5) corresponded to conserved regulatory motifs in vertebrate mtDNA. CSB2 and Switch site are involved in transcription-replication regulation in part by binding the transcription elongation factor TFEM (Agaronyan et al. 2015).

The *lncCR-H* was identified and studied in **Paper III** and **Paper IV**. The SOLiD™ small RNA sequencing experiment identified three small RNAs mapping to the *lncCR-H* gene region (Figure 5). Similar to *lncCR-L*, two small RNAs corresponded to the sense strand. The most interesting small RNA is probably RNA2, which corresponds to the heteroplasmic pyrimidine-rich motif (C-run). The other one (RNA1) appeared to be a tRF-1 RNA, derived from $tRNA^{Thr}$ (Figure 5).

We propose that the mitochondrial *lncCR-L* and *lncCR-H* lncRNAs may serve as precursors for small RNAs. The specific functions these small RNAs in the mitochondria are not known, but we speculate that they involve regulatory roles in transcription elongation, mtDNA replication, or ribosome function. However, more research needs to be done in order to elucidate the potential role of the small RNAs in mitochondrial function.

5 Further perspectives

Additional sequencing of complete mitochondrial genomes from more Atlantic cod specimens may not contribute significantly to a better resolution in sequence variation analysis. Further work and investigations, however, should focus more on the cellular aspects of the mitochondrial genome function. The canonical 37 mitochondrial genes in the Atlantic cod mitochondrial genome most likely function similarly to those in all other vertebrates, including humans. In 2011, we reported that the Atlantic cod mitochondrial transcripts followed the same main rules of expression and processing as those in human cells (Coucheron et al. 2011).

Mitochondrial-derived small RNAs need to be profiled in various adult tissue types during normal and stress conditions, and during developmental stages. Preliminary and unpublished data from our laboratory indicate differentially expression patterns of the mitochondrial tRFs during Atlantic cod early stage development. A first step could be to investigate the cellular location of mitochondrial tRFs and other small RNAs by in situ RNA hybridization. Next, it would be relevant to investigate whether mitochondrial small RNAs are confined to the mitochondrial compartment, or exported to the cytosol, or the nucleus.

It would also be very interesting to determine the cellular localization of the mitochondrial lncRNAs, especially those that are generated by the CR. Is the biological function as suggested above to serve as precursors for small RNAs, or do they have other additional roles in the mitochondrial function? These questions are not easily experimentally approached since a genetic transformation system of mitochondria is currently lacking. Perhaps studies in Atlantic cod need to be performed alongside a more established model system, such as the zebrafish.

The newly discovered MDPs are intriguing, in particular so because at least some of the mitochondrial peptides appear highly conserved between fish and human (Jørgensen and Johansen 2018). It would be very interesting to investigate Atlantic cod MDPs by western blot analysis, similar to that performed in human cells (Cobb et al.

2016). Perhaps there are undetected networks of mitochondrial transcripts and gene products, involved in the fine-tuning of mitochondrial functions.

6 References

Agaronyan K, Morozov YI, Anikin M, Temiakov D (2015) Mitochondrial biology. Replication-transcription switch in human mitochondria. Science 347:548–551. doi: 10.1126/science.aaa0986

Alexeyev M, Shokolenko I, Wilson G, LeDoux S (2013) The maintenance of mitochondrial DNA integrity—critical analysis and update. Cold Spring Harb Perspect Biol 5:a012641. doi: 10.1101/cshperspect.a012641

Antonicka H, Shoubridge EA (2015) Mitochondrial RNA granules are centers for posttranscriptional RNA processing and ribosome biogenesis. Cell Rep 10:920–932. doi: 10.1016/j.celrep.2015.01.030

Antunes A, Ramos MJ (2005) Discovery of a large number of previously unrecognized mitochondrial pseudogenes in fish genomes. Genomics 86:708–717. doi: 10.1016/j.ygeno.2005.08.002

Árnason E, Rand DM (1992) Heteroplasmy of short tandem repeats in mitochondrial DNA of Atlantic cod, *Gadus morhua*. Genetics 132:211–220.

Asin-Cayuela J, Gustafsson CM (2007) Mitochondrial transcription and its regulation in mammalian cells. Trends Biochem Sci 32:111–117. doi: 10.1016/j.tibs.2007.01.003

Bailey LJ, Doherty AJ (2017) Mitochondrial DNA replication: a PrimPol perspective. Biochem Soc Trans 45:513–529. doi: 10.1042/BST20160162

Baker LA, Watt IN, Runswick MJ, Walker JE, Rubinstein JL (2012) Arrangement of subunits in intact mammalian mitochondrial ATP synthase determined by cryo-EM. Proc Natl Acad Sci USA 109:11675–11680. doi: 10.1073/pnas.1204935109

Bakke I, Johansen SD (2005) Molecular phylogenetics of Gadidae and related Gadiformes based on mitochondrial DNA sequences. Mar Biotechnol 7:61–69. doi: 10.1007/s10126-004-3131-0

Berk AJ, Clayton DA (1974) Mechanism of mitochondrial DNA replication in mouse L-cells: Asynchronous replication of strands, segregation of circular daughter molecules, aspects of topology and turnover of an initiation sequence. J Mol Biol 86:801–824 . doi: 10.1016/0022-2836(74)90355-6

Bogenhagen DF (2012) Mitochondrial DNA nucleoid structure. Biochim Biophys Acta - Gene Regul Mech 1819:914–920. doi: 10.1016/j.bbagrm.2011.11.005

Butow RA, Chen XJ (2005) The organization and inheritance of the mitochondrial genome. Nat Rev Genet 6:815. doi: 10.1038/nrg1708

Capt C, Passamonti M, Breton S (2016) The human mitochondrial genome may code for more than 13 proteins. Mitochondrial DNA Part A 27:3098–3101. doi: 10.3109/19401736.2014.1003924

Chang DD, Clayton DA (1984) Precise identification of individual promoters for transcription of each strand of human mitochondrial DNA. Cell 36:635–643. doi: 10.1016/0092-8674(84)90343-X

Cobb LJ, Lee C, Xiao J, Yen K, Wong RG, Nakamura HK, Mehta HH, Gao Q, Ashur C, Huffman DM, Wan J, Muzumdar R, Barzilai N, Cohen P (2016) Naturally occurring mitochondrial-derived peptides are age-dependent regulators of apoptosis, insulin sensitivity, and inflammatory markers. Aging 8:796–809. doi: 10.18632/aging.100943

Coucheron DH, Nymark M, Breines R, Karlsen BO, Andreassen M, Jørgensen TE, Moum T, Johansen SD (2011) Characterization of mitochondrial mRNAs in codfish reveals unique features compared to mammals. Curr Genet 57:213–222. doi: 10.1007/s00294-011-0338-2

Davies KM, Daum B (2013) Role of cryo-ET in membrane bioenergetics research. Biochem Soc Trans 41:1227–1234. doi: 10.1042/BST20130029

Dayama G, Emery SB, Kidd JM, Mills RE (2014) The genomic landscape of polymorphic human nuclear mitochondrial insertions. Nucleic Acids Res 42:12640–12649. doi: 10.1093/nar/gku1038

De Silva D, Tu Y-T, Amunts A, Fontanesi F, Barrientos A (2015) Mitochondrial ribosome assembly in health and disease. Cell Cycle Georget Tex 14:2226–2250. doi: 10.1080/15384101.2015.1053672

Dietrich A, Wallet C, Iqbal RK, Gualberto JM, Lotfi F (2015) Organellar non-coding RNAs: Emerging regulation mechanisms. Biochimie 117:48–62. doi: 10.1016/j.biochi.2015.06.027

DiMauro S, Hirano M, Schon EA (2006) Approaches to the treatment of mitochondrial diseases. Muscle Nerve 34:265–283. doi: 10.1002/mus.20598

Falkenberg M, Larsson N-G, Gustafsson CM (2007) DNA replication and transcription in mammalian mitochondria. Annu Rev Biochem 76:679–699. doi: 10.1146/annurev.biochem.76.060305.152028

FAO, 2016. The State of World Fisheries and Aquaculture 2016, 1–204.

FAO (2017). FAO Fisheries & Aquaculture - species fact sheets - *Gadus morhua* (Linnaeus, 1758). http://www.fao.org/fishery/species/2218/en

FAO FishStat. Fisheries and aquaculture software. FishStatJ - software for fishery statistical time series. 21 July 2016. http://www.fao.org/fishery/ Contributing to food security and nutrition for all. Rome. 200 pp

Fatica A, Bozzoni I (2014) Long non-coding RNAs: new players in cell differentiation and development. Nat Rev Genet 15:7–21. doi: 10.1038/nrg3606

Fiedorczuk K, Letts JA, Degliesposti G, Kaszuba K, Skehel M, Sazanov LA (2016) Atomic structure of the entire mammalian mitochondrial complex I. Nature 538:406–410. doi: 10.1038/nature19794

Froese, R. and D. Pauly. Editors. 2017. FishBase. World Wide Web electronic publication. www.fishbase.org, version (10/2017).

Flynn T, Signal B, Johnson SL, Gemmell NJ (2016) Mitochondrial genome diversity among six laboratory zebrafish (*Danio rerio*) strains. Mitochondrial DNA Part A 27:4364–4371. doi: 10.3109/19401736.2015.1089536

Fusté JM, Shi Y, Wanrooij S, Zhu X, Jemt E, Persson Ö, Sabouri N, Gustafsson CM, Falkenberg M (2014) *In vivo* occupancy of mitochondrial single-stranded DNA binding protein supports the strand displacement mode of DNA replication. PLOS Genet 10:e1004832. doi: 10.1371/journal.pgen.1004832

Gao S, Tian X, Chang H, Sun Y, Wu Z, Cheng Z, Dong P, Zhao Q, Ruan J, Bu W (2017) Two novel lncRNAs discovered in human mitochondrial DNA using PacBio full-length transcriptome data. Mitochondrion. doi: 10.1016/j.mito.2017.08.002

Gray MW, Burger G, Lang BF (1999) Mitochondrial evolution. Science 283:1476–1481. doi: 10.1126/science.283.5407.1476

Greber BJ, Boehringer D, Leibundgut M, Bieri P, Leitner A, Schmitz N, Aebersold R, Ban N (2014) The complete structure of the large subunit of the mammalian mitochondrial ribosome. Nature 515:283 . doi: 10.1038/nature13895

Gustafsson CM, Falkenberg M, Larsson N-G (2016) Maintenance and expression of mammalian mitochondrial DNA. Annu Rev Biochem 85:133–160. doi: 10.1146/annurev-biochem-060815-014402

Hällberg BM, Larsson N-G (2014) Making proteins in the powerhouse. Cell Metab 20:226–240. doi: 10.1016/j.cmet.2014.07.001

Hashimoto Y, Niikura T, Tajima H, Yasukawa T, Sudo H, Ito Y, Kita Y, Kawasumi M, Kouyama K, Doyu M, Sobue G, Koide T, Tsuji S, Lang J, Kurokawa K, Nishimoto I (2001) A rescue factor abolishing neuronal cell death by a wide spectrum of familial Alzheimer's disease genes and Aβ. Proc Natl Acad Sci USA 98:6336–6341. doi: 10.1073/pnas.101133498

Haute LV, Pearce SF, Powell CA, D'Souza AR, Nicholls TJ, Minczuk M (2015) Mitochondrial transcript maturation and its disorders. J Inherit Metab Dis 38:655–680. doi: 10.1007/s10545-015-9859-z

Hazkani-Covo E, Zeller RM, Martin W (2010) Molecular poltergeists: mitochondrial DNA copies (numts) in sequenced nuclear genomes. PLoS Genet 6:e1000834. doi: 10.1371/journal.pgen.1000834

Hedberg A, Knutsen E, Løvhaugen AS, Jørgensen TE, Perander M, Johansen SD (2018). Cancer-specific SNPs originate from low-level heteroplasmic variants in a human mitochondrial genome. Mitochondrial DNA Part A 29. doi: 10.1080/24701394.2018.1461852

Hoarau G, Holla S, Lescasse R, Stam WT, Olsen JL (2002) Heteroplasmy and evidence for recombination in the mitochondrial control region of the flatfish *Platichthys flesus*. Mol Biol Evol 19:2261–2264. doi: 10.1093/oxfordjournals.molbev.a004049

ICES 2017. International Council for the Exploration of the Sea, Copenhagen, Denmark. (2017) Stock assessment graphs: http://standardgraphs.ices.dk/stockList.aspx

Ingman M, Gyllensten U (2006) Vertebrate mitochondrial DNA. In: Reviews in Cell Biology and Molecular Medicine. Wiley-VCH Verlag GmbH & Co. KGaA

Iwata S, Lee JW, Okada K, Lee JK, Iwata M, Rasmussen B, Link TA, Ramaswamy S, Jap BK (1998) Complete structure of the 11-subunit bovine mitochondrial Cytochrome bc1 Complex. Science 281:64–71. doi: 10.1126/science.281.5373.64

Jemt E, Persson Ö, Shi Y, Mehmedovic M, Uhler JP, Dávila López M, Freyer C, Gustafsson CM, Samuelsson T, Falkenberg M (2015) Regulation of DNA replication at the end of the mitochondrial D-loop involves the helicase TWINKLE and a conserved sequence element. Nucleic Acids Res 43:9262–9275. doi: 10.1093/nar/gkv804

Johansen S, Bakke I (1996) The complete mitochondrial DNA sequence of Atlantic cod (*Gadus morhua*): relevance to taxonomic studies among codfishes. Mol Mar Biol Biotechnol 5:203–214

Johansen SD, Coucheron DH, Andreassen M, Karlsen BO, Furmanek T, Jørgensen TE, Emblem Å, Breines R, Nordeide JT, Moum T, Nederbragt AJ, Stenseth NC, Jakobsen KS (2009) Large-scale sequence analyses of Atlantic cod. New Biotechnol 25:263–271. doi: 10.1016/j.nbt.2009.03.014

Johansen S, Guddal PH, Johansen T (1990) Organization of the mitochondrial genome of Atlantic cod, *Gadus morhua*. Nucleic Acids Res 18:411–419

Jourdain AA, Koppen M, Wydro M, Rodley CD, Lightowlers RN, Chrzanowska-Lightowlers ZM, Martinou J-C (2013) GRSF1 regulates RNA processing in mitochondrial RNA granules. Cell Metab 17:399–410. doi: 10.1016/j.cmet.2013.02.005

Junge W, Nelson N (2015) ATP synthase. Annu Rev Biochem 84:631–657. doi: 10.1146/annurev-biochem-060614-034124

Jørgensen TE, Johansen SD (2018) Expanding the coding potential of vertebrate mitochondrial genomes: lesson learned from the Atlantic cod. In: Mitochondrial DNA – New Insights (M. Mattila, ed). InTech Open. ISBN 978-953-51-6167-7

Karlsen BO, Emblem Å, Jørgensen TE, Klingan KA, Nordeide JT, Moum T, Johansen SD (2014) Mitogenome sequence variation in migratory and stationary ecotypes of North-east Atlantic cod. Mar Genomics 15:103–108. doi: 10.1016/j.margen.2014.01.001

Karlsen BO, Klingan K, Emblem Å, Jørgensen TE, Jueterbock A, Furmanek T, Hoarau G, Johansen SD, Nordeide JT, Moum T (2013) Genomic divergence between the migratory and stationary ecotypes of Atlantic cod. Mol Ecol 22:5098–5111. doi: 10.1111/mec.12454

Kaushal PS, Sharma MR, Booth TM, Haque EM, Tung C-S, Sanbonmatsu KY, Spremulli LL, Agrawal RK (2014) Cryo-EM structure of the small subunit of the mammalian mitochondrial ribosome. Proc Natl Acad Sci USA 111:7284–7289 . doi: 10.1073/pnas.1401657111

Kijewska A, Burzynski A, Wenne R (2009) Variation in the copy number of tandem repeats of mitochondrial DNA in the North-East Atlantic cod populations. Mar Biol Res 5:186–192. doi: 10.1080/17451000802277408

Kim S-J, Xiao J, Wan J, Cohen P, Yen K (2017) Mitochondrially derived peptides as novel regulators of metabolism. J Physiol 595:6613–6621. doi: 10.1113/JP274472

Kühlbrandt W (2015) Structure and function of mitochondrial membrane protein complexes. BMC Biol 13:89. doi: 10.1186/s12915-015-0201-x

Kukat C, Larsson N-G (2013) mtDNA makes a U-turn for the mitochondrial nucleoid. Trends Cell Biol 23:457–463. doi: 10.1016/j.tcb.2013.04.009

Kukat C, Wurm CA, Spåhr H, Falkenberg M, Larsson N-G, Jakobs S (2011) Super-resolution microscopy reveals that mammalian mitochondrial nucleoids have a uniform size and frequently contain a single copy of mtDNA. Proc Natl Acad Sci USA 108:13534–13539. doi: 10.1073/pnas.1109263108

Kumarswamy R, Bauters C, Volkmann I, Maury F, Fetisch J, Holzmann A, Lemesle G, de Groote P, Pinet F, Thum T (2014) Circulating long noncoding RNA, LIPCAR, predicts survival in patients with heart failure. Circ Res 114:1569–1575. doi: 10.1161/CIRCRESAHA.114.303915

Kurabayashi A, Sumida M (2013) Afrobatrachian mitochondrial genomes: genome reorganization, gene rearrangement mechanisms, and evolutionary trends of duplicated and rearranged genes. BMC Genomics 14:633 . doi: 10.1186/1471-2164-14-633

Lee C, Yen K, Cohen P (2013) Humanin: a harbinger of mitochondrial-derived peptides? Trends Endocrinol Metab 24:222–228. doi: 10.1016/j.tem.2013.01.005

Lee C, Zeng J, Drew BG, Sallam T, Martin-Montalvo A, Wan J, Kim S-J, Mehta H, Hevener AL, de Cabo R, Cohen P (2015) The mitochondrial-derived peptide MOTS-c promotes metabolic homeostasis and reduces obesity and insulin resistance. Cell Metab 21:443–454. doi: 10.1016/j.cmet.2015.02.009

Leister D (2005) Origin, evolution and genetic effects of nuclear insertions of organelle DNA. Trends Genet 21:655–663. doi: 10.1016/j.tig.2005.09.004

Levin L, Blumberg A, Barshad G, Mishmar D (2014) Mito-nuclear co-evolution: the positive and negative sides of functional ancient mutations. Front Genet 5:448. doi: 10.3389/fgene.2014.00448

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

Lunt DH, Whipple LE, Hyman BC (1998) Mitochondrial DNA variable number tandem repeats (VNTRs): utility and problems in molecular ecology. Mol Ecol 7:1441–1455. doi: 10.1046/j.1365-294x.1998.00495.x

Ma H, Weber GM, Wei H, Yao J (2016) Identification of mitochondrial genome-encoded small RNAs related to egg deterioration caused by postovulatory aging in rainbow trout. Mar Biotechnol 18:584–597. doi: 10.1007/s10126-016-9719-3

Martin WF, Garg S, Zimorski V (2015) Endosymbiotic theories for eukaryote origin. Phil Trans R Soc B 370:20140330. doi: 10.1098/rstb.2014.0330

Mercer TR, Neph S, Dinger ME, Crawford J, Smith MA, Shearwood A-MJ, Haugen E, Bracken CP, Rackham O, Stamatoyannopoulos JA, Filipovska A, Mattick JS (2011) The human mitochondrial transcriptome. Cell 146:645–658. doi: 10.1016/j.cell.2011.06.051

Mjelle KA, Karlsen BO, Jørgensen TE, Moum T, Johansen SD (2008) Halibut mitochondrial genomes contain extensive heteroplasmic tandem repeat arrays involved in DNA recombination. BMC Genomics 9:10. doi: 10.1186/1471-2164-9-10

Mohan A, Whyte S, Wang X, Nashimoto M, Levinger L (1999) The 3' end CCA of mature tRNA is an antideterminant for eukaryotic 3'-tRNase. RNA 5:245–256

Nesbø CL, Arab MO, Jakobsen KS (1998) Heteroplasmy, length and sequence variation in the mtDNA control regions of three percid fish species (*Perca fluviatilis*, *Acerina cernua*, *Stizostedion lucioperca*). Genetics 148:1907–1919

Nicholls TJ, Minczuk M (2014) In D-loop: 40 years of mitochondrial 7S DNA. Exp Gerontol 56:175–181. doi: 10.1016/j.exger.2014.03.027

Nordeide JT, Johansen SD, Jørgensen TE, Karlsen BO, Moum T (2011) Population connectivity among migratory and stationary cod *Gadus morhua* in the Northeast Atlantic—A review of 80 years of study. Mar Ecol Prog Ser 435:269–283. doi: 10.3354/meps09232

Ojala D, Crews S, Montoya J, Gelfand R, Attardi G (1981a) A small polyadenylated RNA (7 S RNA), containing a putative ribosome attachment site, maps near the origin of human mitochondrial DNA replication. J Mol Biol 150:303–314. doi: 10.1016/0022-2836(81)90454-X

Ojala D, Montoya J, Attardi G (1981b) tRNA punctuation model of RNA processing in human mitochondria. Nature 290:470–474.

Ott M, Amunts A, Brown A (2016) Organization and regulation of mitochondrial protein synthesis. Annu Rev Biochem 85:77–101. doi: 10.1146/annurev-biochem-060815-014334

Payne BAI, Wilson IJ, Yu-Wai-Man P, Coxhead J, Deehan D, Horvath R, Taylor RW, Samuels DC, Santibanez-Koref M, Chinnery PF (2013) Universal heteroplasmy of human mitochondrial DNA. Hum Mol Genet 22:384–390. doi: 10.1093/hmg/dds435

Pearce SF, Rebelo-Guiomar P, D'Souza AR, Powell CA, Haute LV, Minczuk M (2017) Regulation of mammalian mitochondrial gene expression: recent advances. Trends Biochem Sci 42:625–639. doi: 10.1016/j.tibs.2017.02.003

Powell CA, Nicholls TJ, Minczuk M (2015) Nuclear-encoded factors involved in post-transcriptional processing and modification of mitochondrial tRNAs in human disease. Front Genet. doi: 10.3389/fgene.2015.00079

Rackham O, Mercer TR, Filipovska A (2012) The human mitochondrial transcriptome and the RNA-binding proteins that regulate its expression. Wiley Interdiscip Rev RNA 3:675–695. doi: 10.1002/wrna.1128

Rackham O, Shearwood A-MJ, Mercer TR, Davies SMK, Mattick JS, Filipovska A (2011) Long noncoding RNAs are generated from the mitochondrial genome and regulated by nuclear-encoded proteins. RNA 17:2085–2093. doi: 10.1261/rna.029405.111

Rensch T, Villar D, Horvath J, Odom DT, Flicek P (2016) Mitochondrial heteroplasmy in vertebrates using ChIP-sequencing data. Genome Biol 17:139. doi: 10.1186/s13059-016-0996-y

Rich PR, Maréchal A (2010) The mitochondrial respiratory chain. Essays Biochem 47:1–23. doi: 10.1042/bse0470001

Richly E, Leister D (2004) NUMTs in sequenced eukaryotic genomes. Mol Biol Evol 21:1081–1084. doi: 10.1093/molbev/msh110

Riggs CL, Podrabsky JE (2017) Small noncoding RNA expression during extreme anoxia tolerance of annual killifish (*Austrofundulus limnaeus*) embryos. Physiol Genomics 49:505–518. doi: 10.1152/physiolgenomics.00016.2017

Ro S, Ma H-Y, Park C, Ortogero N, Song R, Hennig GW, Zheng H, Lin Y-M, Moro L, Hsieh J-T, Yan W (2013) The mitochondrial genome encodes abundant small noncoding RNAs. Cell Res 23:759–774. doi: 10.1038/cr.2013.37

Rorbach J, Gao F, Powell CA, D'Souza A, Lightowlers RN, Minczuk M, Chrzanowska-Lightowlers ZM (2016) Human mitochondrial ribosomes can switch their structural RNA composition. Proc Natl Acad Sci U S A 113:12198–12201 . doi: 10.1073/pnas.1609338113

Rorbach J, Minczuk M (2012) The post-transcriptional life of mammalian mitochondrial RNA. Biochem J 444:357–373. doi: 10.1042/BJ20112208

Salinas-Giegé T, Giegé R, Giegé P (2015) tRNA biology in mitochondria. Int J Mol Sci 16:4518–4559. doi: 10.3390/ijms16034518

Sanchez MIGL, Mercer TR, Davies SMK, Shearwood A-MJ, Nygård KKA, Richman TR, Mattick JS, Rackham O, Filipovska A (2011) RNA processing in human mitochondria. Cell Cycle Georget Tex 10:2904–2916. doi: 10.4161/cc.10.17.17060

Sbisà E, Tanzariello F, Reyes A, Pesole G, Saccone C (1997) Mammalian mitochondrial D-loop region structural analysis: identification of new conserved sequences and their functional and evolutionary implications. Gene 205:125–140. doi: 10.1016/S0378-1119(97)00404-6

Scarpulla RC (2008) Transcriptional Paradigms in Mammalian Mitochondrial Biogenesis and Function. Physiol Rev 88:611–638. doi: 10.1152/physrev.00025.2007

Sharma MR, Koc EC, Datta PP, Booth TM, Spremulli LL, Agrawal RK (2003) Structure of the mammalian mitochondrial ribosome reveals an expanded functional role for its component proteins. Cell 115:97–108. doi: 10.1016/S0092-8674(03)00762-1

Shutt TE, Lodeiro MF, Cotney J, Cameron CE, Shadel GS (2010) Core human mitochondrial transcription apparatus is a regulated two-component system in vitro. Proc Natl Acad Sci USA 107:12133–12138. doi: 10.1073/pnas.0910581107

Star B, Nederbragt AJ, Jentoft S, Grimholt U, Malmstrøm M, Gregers TF, Rounge TB, Paulsen J, Solbakken MH, Sharma A, Wetten OF, Lanzén A, Winer R, Knight J, Vogel J-H, Aken B, Andersen Ø, Lagesen K, Tooming-Klunderud A, Edvardsen RB, Tina KG, Espelund M, Nepal C, Previti C, Karlsen BO, Moum T, Skage M, Berg PR, Gjøen T, Kuhl H, Thorsen J, Malde K, Reinhardt R, Du L, Johansen SD, Searle S, Lien S, Nilsen F, Jonassen I, Omholt SW, Stenseth NC, Jakobsen KS (2011) The genome sequence of Atlantic cod reveals a unique immune system. Nature 477:207–210. doi: 10.1038/nature10342

Stewart JB, Chinnery PF (2015) The dynamics of mitochondrial DNA heteroplasmy: implications for human health and disease. Nat Rev Genet 16:530–542. doi: 10.1038/nrg3966

Sun F, Huo X, Zhai Y, Wang A, Xu J, Su D, Bartlam M, Rao Z (2005) Crystal structure of mitochondrial respiratory membrane protein complex II. Cell 121:1043–1057. doi: 10.1016/j.cell.2005.05.025

Suzuki T, Nagao A, Suzuki T (2011) Human mitochondrial tRNAs: biogenesis, function, structural aspects, and diseases. Annu Rev Genet 45:299–329. doi: 10.1146/annurev-genet-110410- 132531

Taanman J-W (1999) The mitochondrial genome: structure, transcription, translation and replication. Biochim Biophys Acta - Bioenerg 1410:103–123. doi: 10.1016/S0005-2728(98)00161-3

Taylor RW, Turnbull DM (2005) Mitochondrial DNA mutations in human disease. Nat Rev Genet 6:389. doi: 10.1038/nrg1606

Temperley RJ, Wydro M, Lightowlers RN, Chrzanowska-Lightowlers ZM (2010) Human mitochondrial mRNAs--like members of all families, similar but different. Biochim Biophys Acta 1797:1081–1085. doi: 10.1016/j.bbabio.2010.02.036

Terzioglu M, Ruzzenente B, Harmel J, Mourier A, Jemt E, López MD, Kukat C, Stewart JB, Wibom R, Meharg C, Habermann B, Falkenberg M, Gustafsson CM, Park CB, Larsson N-G (2013) MTERF1 binds mtDNA to prevent transcriptional interference at the light-strand promoter but is dispensable for rRNA gene transcription regulation. Cell Metab 17:618–626. doi: 10.1016/j.cmet.2013.03.006

Tørresen OK, Star B, Jentoft S, Reinar WB, Grove H, Miller JR, Walenz BP, Knight J, Ekholm JM, Peluso P, Edvardsen RB, Tooming-Klunderud A, Skage M, Lien S, Jakobsen KS, Nederbragt AJ (2017) An improved genome assembly uncovers prolific tandem repeats in Atlantic cod. BMC Genomics 18:95. doi: 10.1186/s12864-016-3448-x

Uhler JP, Falkenberg M (2015) Primer removal during mammalian mitochondrial DNA replication. DNA Repair 34:28–38. doi: 10.1016/j.dnarep.2015.07.003

Ursvik A, Breines R, Christiansen JS, Fevolden S-E, Coucheron DH, Johansen SD (2007) A mitogenomic approach to the taxonomy of pollocks: *Theragra chalcogramma* and *T. finnmarchica* represent one single species. BMC Evol Biol 7:86. doi: 10.1186/1471-2148-7-86

van der Sluis EO, Bauerschmitt H, Becker T, Mielke T, Frauenfeld J, Berninghausen O, Neupert W, Herrmann JM, Beckmann R (2015) Parallel structural evolution of mitochondrial ribosomes and OXPHOS complexes. Genome Biol Evol 7:1235–1251. doi: 10.1093/gbe/evv061

Wai T, Ao A, Zhang X, Cyr D, Dufort D, Shoubridge EA (2010) The role of mitochondrial DNA copy number in mammalian fertility. Biol Reprod 83:52–62. doi: 10.1095/biolreprod.109.080887

Wallace DC, Chalkia D (2013) Mitochondrial DNA genetics and the heteroplasmy conundrum in evolution and disease. Cold Spring Harb Perspect Biol 5:a021220. doi: 10.1101/cshperspect.a021220

Wang X, Liu N, Zhang H, Yang X-J, Huang Y, Lei F (2015) Extreme variation in patterns of tandem repeats in mitochondrial control region of yellow-browed tits (Sylviparus modestus, Paridae). Sci Rep 5:13227. doi: 10.1038/srep13227

Wanrooij S, Falkenberg M (2010) The human mitochondrial replication fork in health and disease. Biochim Biophys Acta - Bioenerg 1797:1378–1388. doi: 10.1016/j.bbabio.2010.04.015

Wanrooij S, Miralles Fusté J, Stewart JB, Wanrooij PH, Samuelsson T, Larsson N-G, Gustafsson CM, Falkenberg M (2012) In vivo mutagenesis reveals that OriL is essential for mitochondrial DNA replication. EMBO Rep 13:1130–1137. doi: 10.1038/embor.2012.161

Whelan SP, Zuckerbraun BS (2013) Mitochondrial signaling: forwards, backwards, and in between. Oxid Med Cell Longev 2013:351613. doi: 10.1155/2013/351613

Wiesner RJ, Rüegg JC, Morano I (1992) Counting target molecules by exponential polymerase chain reaction: Copy number of mitochondrial DNA in rat tissues. Biochem Biophys Res Commun 183:553–559. doi: 10.1016/0006-291X(92)90517-O

Wittig I, Schägger H (2008) Structural organization of mitochondrial ATP synthase. Biochim Biophys Acta - Bioenerg 1777:592–598. doi: 10.1016/j.bbabio.2008.04.027

Wu M, Gu J, Guo R, Huang Y, Yang M (2016) Structure of mammalian respiratory supercomplex I1III2IV1. Cell 167:1598–1609.e10. doi: 10.1016/j.cell.2016.11.012

Yanagimoto T, Kitamura T, Kobayashi T (2004) Complete nucleotide sequence and variation of mitochondrial DNA from 10 individuals of walleye pollock, *Theragra chalcogramma*. Fish Sci 70:885–895. doi: 10.1111/j.1444-2906.2004.00883.x

Paper I

Mitochondrial genome variation of Atlantic cod

Tor Erik Jørgensen[1], **Bård Ove Karlsen**[2], **Åse Emblem**[3], **Ragna Breines**[3], **Morten Andreassen**[3], **Trine B. Rounge**[4], **Alexander J. Nederbragt**[4], **Kjetill S. Jakobsen**[4], **Marianne Nymark**[3], **Anita Ursvik**[3], **Dag H. Coucheron**[3], **Lars Martin Jakt**[1], **Jarle T. Nordeide**[1], **Truls B. Moum**[1], **Steinar D. Johansen**[1,3*]

***For correspondence:**
steinar.d.johansen@nord.no (SDJ)

Email adresses:
tor.e.jorgensen@nord.no (TEJ)
bard.ove.karlsen@nordlandssykehuset.no (BOK)
ase.emblem@gmail.com (ÅE)
ragna.breines@uib.bo (RB)
morten.andreassen@gmail.com (MA)
trine.rounge@kreftregisteret.no (TBR)
lex.nederbragt@ibv.uio.no (AJN)
k.s.jakobsen@ibs.uio.no (KSJ)
marianne.nymark@ntnu.no (MN)
anita.ursvik@uit.no (AU)
dag.coucheron@uit.no (DHC)
lars.m.jakt@nord.no (LMJ)
jarle.t.nordeide@nord.no (JTN)
truls.b.moum@nord.no (TBM)

[1]Genomics Research Group, Faculty of Biosciences and Aquaculture, Nord University, Bodø, Norway; [2]Research Laboratory and Department of Laboratory Medicine, Nordland Hospital, Bodø, Norway; [3]Department of Medical Biology, Faculty of Health Sciences, UiT – Arctic University of Norway, Tromsø, Norway; [4]Centre for Ecological and Evolutionary Syntheses (CEES), Department of Biosciences, University of Oslo, Oslo, Norway

Abstract

Objective

The objective of this study was to analyse intraspecific sequence variation of Atlantic cod mitochondrial DNA, based on a comprehensive collection of completely sequenced mitochondrial genomes.

Results

We determined the complete mitochondrial DNA sequence of 124 cod specimens from the eastern and western part of the species' distribution range in the North Atlantic Ocean. All specimens harboured a unique mitochondrial DNA haplotype. Nine hundred and fifty-two polymorphic sites were identified, including 109 non-synonymous sites within protein coding regions. Eighteen variable sites were identified as indels, exclusively distributed in structural RNA genes and non-coding regions. Phylogeographic analyses based on 156 available cod mitochondrial genomes did not reveal a clear structure. There was a lack of mitochondrial genetic differentiation between two ecotypes of cod in the eastern North Atlantic, but eastern and western cod were differentiated and mitochondrial genome diversity was higher in the eastern than the western Atlantic, suggesting deviating population histories. The geographic distribution of mitochondrial genome variation seems to be governed by demographic processes and gene flow among ecotypes that are otherwise characterized by localized genomic divergence associated with chromosomal inversions.

Introduction

The Atlantic cod (*Gadus morhua*) is one of the most important species for fisheries in the North Atlantic Ocean (Johansen et al. 2009), and recently the nuclear genome was reported (Star et al. 2011; Tørresen et el. 2017). The mitochondrial genome (mitogenome) is considered the second genome of the cell, and its gene content is conserved among most vertebrates (Boore 1999). The 16.7 kb Atlantic cod mitogenome encodes the standard set of 13 hydrophobic membrane proteins, 2 ribosomal RNAs (rRNAs), and 22 transfer RNAs (tRNAs), as well as peptides and long non-coding RNAs, and is organized similar to that of humans (Johansen & Bakke 1996; Jørgensen et al. 2014; Jørgensen & Johansen 2018).

On avarage, the Atlantic cod mitogenome evolves about 14 times more rapidly at the nucleotide level than the nuclear genome (Karlsen et al. 2014), and mitochondrial sequence variation in cod was previously used to trace population structures and patterns of mitogenome

evolution (Árnason 2004; Carr & Marshall 2008; Marshall et al. 2009; Nordeide et al. 2011; Karlsen et al. 2014). Árnason (2004) investigated sequence variants of a 250-bp cytochrome b (CytB) gene fragment in 1278 Atlantic cod specimens throughout the distribution range, and identified trans-Atlantic haplotype clines with more diversity in northeastern and mid-Atlantic cod as compared to northwestern cod. Carr and co-workers (Carr & Marshall 2008; Marshall et al. 2009) reported on mitogenome variation based on 32 cod specimens and identified 298 single nucleotide polymorphic (SNP) sites. They found similar diversities in northwestern and northeast Atlantic cod, but their sample from the northeast Atlantic consisted of six specimens only (Carr & Marshall 2008). In the present study, we sequenced the complete mitogenome of 124 individuals, generating a mitochondrial sequence resource for future studies of Atlantic cod. We analysed relationships among the 156 cod mitogenomes currently available and compared mitogenome variation of the offshore migratory and stationary coastal cod ecotypes (Nordeide et al. 2011), both from the northeast Atlantic, and cod from the northwest Atlantic.

Main Text

Methods

Tissue samples, nucleic acid extraction, PCR amplification, and plasmid cloning

Atlantic cod tissue samples were collected from the western (off Nova Scotia and Newfoundland) and eastern parts (off the British Isles, in the Baltic Sea, Irish Sea, North Sea, along the Norwegian coast and fjords, and in the Barents Sea) of the North Atlantic Ocean (Additional file 1: Table S1). DNA was extracted from fresh muscle tissue or ethanol preserved tissue (stored at -20°C) using the High Pure PCR Template Preparation kit (Roche) or the MasterPure™ Complete DNA and RNA Purification Kit (Epicentre®) according to manufacturer's protocols. Complete mitogenomes were PCR amplified in five overlapping fragments of approximately 4 to 4.5 kb in size using LaTaq polymerase (TAKARA BIO INC). The PCR products were purified using USB® ExoSAP-IT® reagent (Affymetrix). Agarose gel electrophoresis and gel extraction using Invitrogen™ PureLink® Quick Gel Extraction Kit or Invitrogen™ PureLink® PCR Purification Kit were performed according to manufacturer's protocols. PCR and sequencing primers used in this study have been described previously (Breines et al. 2008). Plasmid cloning of the control region (CR) was performed by using Invitrogen™ TOPO® TA Cloning® Kit with One Shot® TOP10 E. coli competent cells. Positive clones were cultivated and plasmid DNA was purified using Invitrogen™ PureLink® Quick Plasmid Miniprep Kit.

Mitogenome sequencing and data analysis

The complete mitogenome sequences of 124 Atlantic cod specimens were determined, using Sanger, Illumina, and Roche 454 technologies (117, six, and one specimen, respectively). The latter, based on pyrosequencing was reported previously (Star et al. 2011). The Illumina GAII sequencing was performed according to protocols in (Kirubakaran et al. 2016) using 2 x 108 bp paired end reads, library inserts of 550 to 575 bp, and 3.1 to 6.6 times (average 4.8 times) whole genome coverage (Norwegian Sequencing Centre - Oslo, Norway). Ninety-five mitogenomes were determined by Sanger sequencing provided by Eurofins MWG Operon (Germany). Two Sanger sequenced mitogenomes (NF1 and NC3) have been reported previously (Ursvik et al. 2007; Karlsen et al. 2014). The 20 remaining mitogenomes were sequenced in-house by Sanger technology directly on purified PCR products or plasmid DNA using the BigDye kit (Applied Biosystems). The complete 16,696-bp NC3 Atlantic cod mitogenome (HG514359) was used as a reference for assembly and mapping of mitogenome sequences and reads. Computer analyses of Sanger-generated mitogenomes were performed using software from DNASTAR® Lasergene software. For mitogenome sequences generated by Roche 454 and Illumina platforms, reference mappings were performed in the CLC Genomics Workbench (QIAGEN®).

A total of 156 available mitogenomes were used to calculate population genetic parameters and reconstruct molecular relationships among Atlantic cod specimens. The CR, tRNA-Phe, and half of the tRNA-Pro sequence were excluded from these comparisons, as these sequences were not available for the 32 specimens previously reported (Carr & Marshall 2008). Population

genetic statistics and measures of genetic differentiation were estimated for the following three subsets of specimens, defined by their geographic origin and ecotype: cod from the northwest Atlantic (NW; N=32), cod from the north east Atlantic of the coastal stationary ecotype (NC; N=25), and Arctic cod from the Barents Sea of the migratory ecotype (NA; N=97) (Additional file 1: Table S1). Two specimens from the Baltic Sea were excluded from these analyses, since differentiation from NC due to vicariance is likely. Nucleotide sequence alignments were generated using T-coffee v/9 software (Notredame et al. 2001) with manual refinements. The tree-building method of maximum likelihood (ML) in MEGA version 6 (Tamura et al. 2013) was used to reconstruct molecular relationships. The ML trees were built from best-fit models of nucleotide evolution generated MEGA 6 (Bayesian information criterion calculations resulted in Tamura and Nei 1993 with a proportion of invariable sites (I) and gamma distributed rate heterogenity among sites (G) as best-fit model). The topologies of the ML trees were evaluated by bootstrap analyses (2000 replications). We analysed nucleotide diversity indices, Tajima's D statistic, and genetic differentiation indices F_{ST} and Da (the average number of net nucleotide substitutions), as implemented in the DnaSP version 6 software (Rozas et al. 2017)

Results

Sequence variation among Atlantic cod mitogenomes

Complete mitogenome sequences of approximately 16.7 kb were obtained for 124 cod specimens sampled in the western and eastern parts of the North Atlantic Ocean. Mitochondrial sequence variation was initially assessed by considering nucleotide variants of the CytB gene fragment (250 bp) previously reported for 1278 Atlantic cod specimens throughout the species' range (Árnason 2004). Eleven haplotypes from that study, including all main haplotypes (A, C, D, E and G) were identified among the 124 specimens, as well as 12 other singleton haplotypes (Additional file 2: Table S2).

The 124 mitogenome sequences were unambiguously aligned from all specimens using the Norwegian costal NC3 (HG514359) (Karlsen et al. 2014) as an Atlantic cod reference, resulting in an alignment of 16,551 positions. The total number of polymorphic sites was 952 (5.7% of mitogenome positions), and these were distributes across the two rRNA genes, all 13 protein coding genes, 18 of the 22 tRNA genes, and major non-coding regions (TP-spacer and CR) (Figure 1). Only 18 variable sites (1.9%) were identified as indels (6 in structural RNA genes and 12 in non-coding regions). Protein coding genes contained 756 (79.4%) substitutions, of which 109 were non-synonymous (14.4%) resulting in amino acid changes in all 13 proteins (Additional file 3: Table S3)

Key features of mitogenome sequence variation are summarized in Table 1, and several features are noted. Protein genes have 2.3 times more SNPs per nucleotide position than RNA genes. This reflect high sequence conservation in RNA genes at species level. An interesting observation is that variable sites in the mitochondrial small subunit rRNA gene are mainly clustered within the 3'M structural domain (Additional file 4: Figure S1). A similar feature was not noted in the mitochondrial large subunit rRNA (Additional file 5: Figure S2). The cytochrome oxidase (CO) genes and proteins (Complex IV) are generally more conserved than the NADH dehydrogense (ND) genes and proteins (Complex I). Here, the ND2 gene contains 7.2 times more polymorphic sites per position than e.g. COII or the RNA genes. The ND2, ND4, ND5, and CytB genes contain the highest density of polymorphic sites.

Mitogenome diversity, intraspecific relationship and genetic differentation

The alignment of 156 available cod mitogenomes resulted in 15,592 common sites following the exclusion of alignment gaps. There were 1002 polymorphic sites, and 1034 substitutions in total among the mitogenome sequences. The nucleotide diversity index was low for cod in the northwest Atlantic (0.203%) compared to stationary and migratory cod in the northeast Atlantic (0.291% and 0.285%, respectively; Table 2). Tajima's D statistic was significantly negative for all population subsets, rejecting the null hypothesis of stable populations with no selection. A representative maximum likelihood (ML) tree is shown in Figure 2. Twelve clades were supported in >80% of bootstrap replications, but there was not a clear geographic structuring of clades. However, while some clades were dominated by NA and NC cod individuals, others

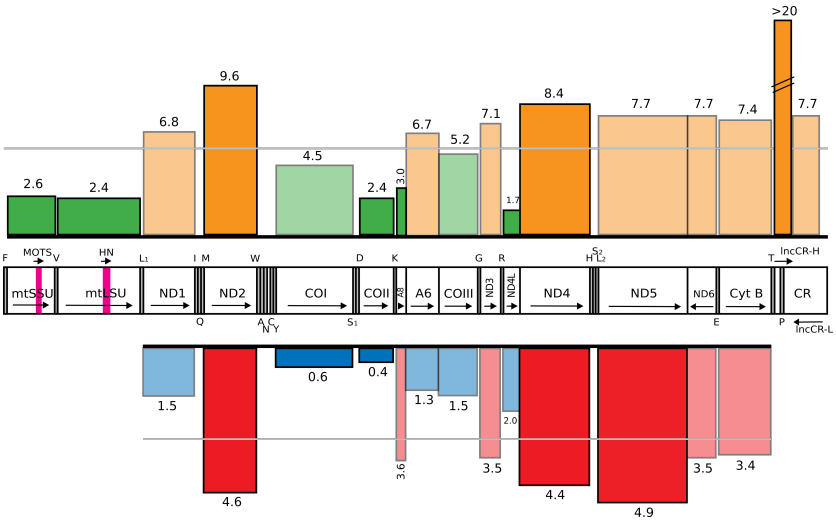


Figure 1. Gene content and variability of Atlantic cod mitogenomes. Mitogenome presented as a linear map of the circular mtDNA. Single nucleotide polymorphisms (SNPs per 100 bp) in gene regions detected among the 124 completely sequenced mitogenomes are indicated above the gene map. Grey horizontal line denotes the average number of SNP (5.5) per 100 bp that include all genes. Genes above and below this average are shown as orange and green bars, respectively. Amino acid substitutions in mitochondrial proteins are presented below the gene map. Grey horizontal line denotes the average number of substitutions (2.9) per 100 amino acids that include all proteins. Proteins above and below this average are shown as red and blue bars, respectively. Abbreviations: mtSSU and mtLSU, mitochondrial small- and large-subunit ribosomal RNA genes; ND1-6, NADH dehydrogenase subunit 1 to 6; COI-III, cytochrome c oxidase subunit I to III; A6 and A8, ATPase subunit 6 and 8; Cyt B, cytochrome b; MOTS, putative MOTS-c peptide; HN, putative humanin peptide; lncCR-H and lncCR-L, long non-coding RNAs coded by the control region (CR). tRNA genes are indicated by the standard one-letter symbols for amino acids. All genes are H-strand encoded, except Q, A, N, C, Y, S_1, E, P, ND6, and lncCR-L (L-strand encoded).

harboured mainly NW cod (Figure 2). Measures of pairwise genetic differentiation were negative (interpreted as nil) between NA and NC cod in the northeast Atlantic, while NW cod were differentiated from both NC and NA cod (F_{ST} of approximately 0.06 and 0.09, respectively; Table 3).

Discussion

Here we provide a comprehensive SNP map of the Atlantic cod mitochondrial genome based on 124 completely sequenced mtDNAs. The 952 variable sites identified among the 124 specimens were not equally distributed along the mitogenome sequence. Structural RNA genes have a significant lower density of overall SNPs per site and variable sites per position compared to protein coding genes. Furthermore, the ND2 gene and the COII gene were the least and most conserved, respectively, among the protein coding genes. This feature was also observed at the protein level. Thus, the Atlantic cod mitogenome follow a similar pattern of conservation as seen for e.g zebrafish (Flynn et al. 2016) or humans (Ingman & Gyllenstein 2001; Diroma et al. 2014).

One hundred and twenty-four cod individuals harboured substantial sequence variation in their mitogenomes, including 349 phylogenetically informative parsimony sites. Phylogenetic analysis of 156 available mitogenomes identified ten haplotype clusters supported by high bootstrap values, but little phylogeographic structuring. Mitogenome evolution in cod seems to be nearly neutral (Marshall et al. 2009; Karlsen et al. 2014), suggesting that the significantly negative Tajima's D statistic mainly signifies recent demographic change, rather than selection. The differentiation of certain cod populations into so-called ecotypes defined by migratory and stationary behaviour, most notably NC and NA cod in the northeastern Atlantic, has long been

Table 1. Key features of SNPs in 124 complete Atlantic cod mitogenomes.

Gene	SNPs	SNPs/100bp	Singleton sites[1]
ND1	66	6,8	38
ND2	100	9,6	55
ND3	25	7,1	15
ND4L	5	1,7	4
ND4	117	8,4	71
ND5	141	7,7	79
ND6	40	7,7	25
COI	69	4,5	49
COII	17	2,4	13
COIII	41	5,2	26
ATPase6	46	6,7	36
ATPase8	5	3,0	5
CytB	84	7,4	48
SSU	25	2,6	18
LSU	52	3,1	45
tRNAs	41	2,7	32
Protein genes	756	6,6	464
RNA genes	118	2,8	95

[1] Non-informative site. Nucleotide variant that appears only once in the data set.

Table 2. Population genetic parameters of Atlantic cod based on the alignment of nearly complete mitochondrial DNA sequences.

	π %	S	η	k	TD
NW	0.203	306	308	31.60	-2.26**
NC	0.291	339	344	45.41	-2.01*
NA	0.285	724	743	44.47	-2.36**
Total (N=156)	0.276	1002	1034	42.99	-2.51***

Sites with alignment gaps were excluded from the alignment of 16,551 positions in all subsets resulting in 15,592 common sites. NW, cod from the north west Atlantic (N=32); NC, cod from the north east Atlantic of the coastal stationary ecotype (N=25); NA, Arctic cod from the Barents Sea of the migratory ecotype (N=97); π %, percent nucleotide diversity; S, number of segregating sites; η, total number of substitutions; k, average number of pairwise nucleotide differences; TD, Tajima's D statistic. * P < 0.05; ** P < 0.01; *** P < 0.001.

Table 3. Genetic differentiation among population subsets defined by geography and ecotype based on nearly complete mitochondrial DNA sequences.

Comparison	F_{ST}	Dxy	Da
NW vs NC	0.05851	0.00262	0.00015
NW vs NA	0.08832	0.00268	0.00024
NC vs NA	-0.00499	0.00287	-0.00001

Sites with alignment gaps were excluded from the alignment of 16,551 positions in all subsets resulting in 15,592 common sites. Population subsets NW, NC, and NA: see Table 2 and main text. F_{ST}, fixation index; Dxy, average number of nucleotide substitutions between subsets; Da, average number of net nucleotide substitutions between subsets.

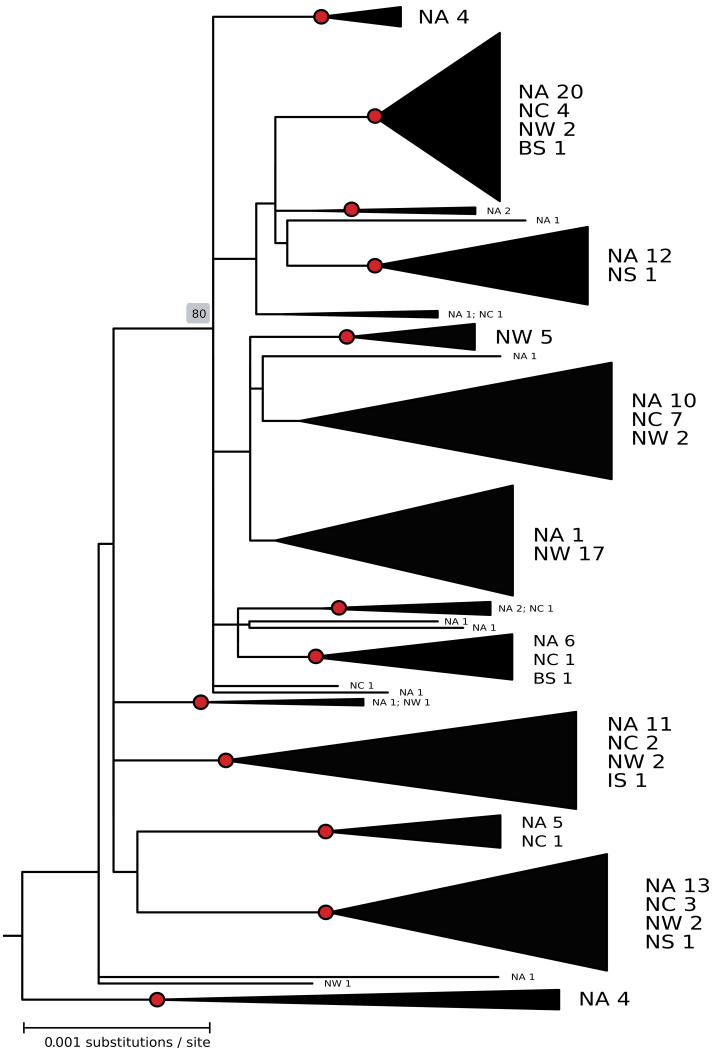


Figure 2. Mitogenome relationships in Atlantic cod. Maximum likelighood (ML) phylogenetic tree based on complete mitogenome haplotype sequences (15,592 common nucleotide positions) of 156 Atlantic cod specimens. *Theragra finnmarchica* (Norwegian Pollock; AM489718) mitogenome was used as an outgroup in tree construction. Bootstrap values (%) from 2000 replicates, all over 70%, are shown at branches. Red filled circles indicate highly significant branch points of bootstrap values above 80% in ML analysis. Closely related haplotype clades are collapsed (bootstrap values above 60%). NA, Northeast Arctic cod; NC, Norwegian costal cod, NW, Northwest cod; BS, Baltic Sea cod; IS, Irish Sea cod; NS, North Sea cod.

a conundrum (Nordeide et al. 2011). Recently, it was shown that these ecotypes are associated with genomic islands of differentiation, inferred to reside within chromosomal inversions in at least four linkage groups (Karlsen et al. 2013; Berg et al. 2016). It is conceivable that such genomic regions could preclude recombination and break-up of co-adapted genes within them, and thus make it possible for locally adapted ecotypes to persist in the face of continued gene flow. Similar chromosomal inversions, suggesting a common ancestry, were subsequently found to contribute to ecotype divergence in the western Atlantic as well (Berg et al. 2017). The mitogenome data indicate some differentiation between western and eastern cod, but a lack of differentiation between NC and NA cod. This would be consistent with isolation by distance

and some gene flow between ecotypes in their mitochondrial genes and neutrally evolving parts of the nuclear genome. Thus, the geographic structuring of mitogenome variation in cod seems to be governed mainly by demographic and stochastic processes in a species with high fecundity and variance in offspring number, much in line with Árnason's conclusions based on CytB sequences (Árnason 2004).

Conclusion

Our study provides a mitochondrial genome resource from Atlantic cod specimens captured in the North Atlantic Ocean. Phylogeographic analyses based on 156 mitochondrial genomes did not reveal a clear structure, but eastern and western cod were differentiated. Mitochondrial genome diversity was higher in the eastern than the western Atlantic, suggesting deviating population histories.

Limitations

The SNP map of the Atlantic cod mitochondrial genome consisted of 952 polymorphic sites among the 124 specimens studied here, and 1002 polymorphic sites among 156 available mitogenomes from the western and eastern parts of the Atlantic Ocean. A more exhaustive SNP map of the cod mitogenome would most certainly require a substantial increase in the number of mitogenomes collected from the vast distribution range of the species.

Abbreviations

CO: cytochrome oxidase
CR: control region
CytB: cytochrome b
LSU: large subunit
ML: maximum likelighood
mtDNA: mitochondrial DNA
ND: NADH dehydrogenase
SNP: single nucleotide polymorphism
SSU: small subunit

Declarations

Author's contributions

TEJ, BOK, ÅE, RB, MA, MN, AU, DHC, TM and SDJ organized the sequencing of the mitochondrial genomes. TBR, AJN and KSJ provided mitogenomes sequenced by Illumina and 454 sequencing. TEJ, BOK, LMJ, JTN, TM, SDJ contributed to mtDNA sequence analyses, and all authors to the discussion of results. SDJ directed the research in collaboration with JTN and TM. SDJ wrote the paper in collaboration with all authors. All authors read and approved the final version of the manuscript.

Acknowledgements

We thank Esben Olsen, Morten Krogstad, Espen Søreng, Ove Nicolaisen, Agnieszka Kijewska, Bill Hutchinson, Sherrylynn Rowe, Asgeir Aglen, and Francois Botha for the contribution of Atlantic cod specimen samples.

Competing interests

The authors declare that they have no competing interests.

Availability of data

Accession numbers of mitogenomes are available at ENA under the study accession number PRJEB23234/ERP104973.

Consent for publication

Not applicable.

Ethics approval and consent to participate

Fish tissue samples were obtained at site of fisheries, and do not involve research on animals. In general, this study was carried out in accordance with ethical guidelines stated by the Norwegian Ministry of Agriculture and Food through the Animal Welfare Act.

Funding

This work was supported by grants from the Norwegian Research Council (GenoFisk), Nord University, and the Arctic University of Norway - UiT.

Additional files

Additional file 1: Table S1. Geographical distribution of completely sequenced Atlantic cod mitogenomes.

Additional file 2: Table S2. Mitochondrial CytB haplotypes generated from 124 complete Atlantic cod mitogenomes.

Additional file 3: Table S3. Non-synonymous substitutions in 124 Atlantic cod mitogenomes.

Additional file 4: Figure S1. Complete secondary structure diagram of Atlantic cod mitochondrial small subunit rRNA. Variable positions among the 124 complete Atlantic cod mitogenomes are indicated, as well as frequency (%) and variable sites (red boxes).

Additional file 5: Figure S2. Complete secondary structure diagram of Atlantic cod mitochondrial large subunit rRNA. (A) Domains O and II. (B) Domains III and IV. (C) Domains V and VI. Variable positions among the 124 complete Atlantic cod mitogenomes are indicated, as well as frequency (%) and variable sites (red boxes).

References

Berg PR, Star B, Pampoulie C, Bradbury IR, Bentzen P, Hutchings JA, Jentoft S, Jakobsen KS. Trans-oceanic genomic divergence of Atlantic cod ecotypes is associated with large inversions. Heredity. 2017; 119(6):418–428.

Berg PR, Star B, Pampoulie C, Sodeland M, Barth JMI, Knutsen H, Jakobsen KS, Jentoft S. Three chromosomal rearrangements promote genomic divergence between migratory and stationary ecotypes of Atlantic cod. Scientific Reports. 2016; 6:23246.

Boore JL. Animal mitochondrial genomes. Nucleic Acids Res. 1999; 27(8):1767–1780.

Breines R, Ursvik A, Nymark M, Johansen SD, Coucheron DH. Complete mitochondrial genome sequences of the Arctic Ocean codfishes Arctogadus glacialis and Boreogadus saida reveal oriL and tRNA gene duplications. Polar Biol. 2008; 31(10):1245–1252.

Carr SM, Marshall HD. Intraspecific Phylogeographic Genomics From Multiple Complete mtDNA Genomes in Atlantic Cod (Gadus morhua): Origins of the "Codmother," Transatlantic Vicariance and Midglacial Population Expansion. Genetics. 2008; 180(1):381–389.

Diroma MA, Calabrese C, Simone D, Santorsola M, Calabrese FM, Gasparre G, Attimonelli M. Extraction and annotation of human mitochondrial genomes from 1000 Genomes Whole Exome Sequencing data. BMC Genomics. 2014; 15(3):S2.

Flynn T, Signal B, Johnson SL, Gemmell NJ. Mitochondrial genome diversity among six laboratory zebrafish (Danio rerio) strains. Mitochondrial DNA A DNA Mapp Seq Anal. 2016; 27(6):4364–4371.

Ingman M, Gyllensten U. Analysis of the complete human mtDNA genome: methodology and inferences for human evolution. J Hered. 2001; 92(6):454–461.

Johansen S, Bakke I. The complete mitochondrial DNA sequence of Atlantic cod (Gadus morhua): relevance to taxonomic studies among codfishes. Mol Marine Biol Biotechnol. 1996; 5(3):203–214.

Johansen SD, Coucheron DH, Andreassen M, Karlsen BO, Furmanek T, Jørgensen TE, Emblem Å, Breines R, Nordeide JT, Moum T, Nederbragt AJ, Stenseth NC, Jakobsen KS. Large-scale sequence analyses of Atlantic cod. New Biotechnology. 2009; 25(5):263–271.

Jørgensen TE, Bakke I, Ursvik A, Andreassen M, Moum T, Johansen SD. An evolutionary preserved intergenic spacer in gadiform mitogenomes generates a long noncoding RNA. BMC Evol Biol. 2014; 14:182.

Jørgensen TE, Johansen SD. Expanding the coding potential of vertebrate mitochondrial genomes: Lesson learned from the Atlantic cod. In: Matilla M, editor. *Mitochondrial DNA - New Insights* InTech Open; 2018.

Karlsen BO, Emblem Å, Jørgensen TE, Klingan KA, Nordeide JT, Moum T, Johansen SD. Mitogenome sequence variation in migratory and stationary ecotypes of North-east Atlantic cod. Mar Genomics. 2014; 15:103–108.

Karlsen BO, Klingan K, Emblem Å, Jørgensen TE, Jueterbock A, Furmanek T, Hoarau G, Johansen SD, Nordeide JT, Moum T. Genomic divergence between the migratory and stationary ecotypes of Atlantic cod. Mol Ecol. 2013; 22(20):5098–5111.

Kirubakaran TG, Grove H, Kent MP, Sandve SR, Baranski M, Nome T, De Rosa MC, Righino B, Johansen T, Otterå H, Sonesson A, Lien S, Andersen Å. Two adjacent inversions maintain genomic differentiation between migratory and stationary ecotypes of Atlantic cod. Mol Ecol. 2016; 25(10):2130–2143.

Marshall HD, Coulson MW, Carr SM. Near Neutrality, Rate Heterogeneity, and Linkage Govern Mitochondrial Genome Evolution in Atlantic Cod (Gadus morhua) and Other Gadine Fish. Mol Biol Evol. 2009; 26(3):579–589.

Nordeide JT, Johansen SD, Jørgensen TE, Karlsen BO, Moum T. Population connectivity among migratory and stationary cod Gadus morhua in the Northeast Atlantic—A review of 80 years of study. Marine Ecology Progress Series. 2011; 435:269–283.

Notredame C, Higgins DG, Heringa J. T-coffee: a novel method for fast and accurate multiple sequence alignment11Edited by J. Thornton. Journal of Molecular Biology. 2000; 302(1):205–217.

Rozas J, Ferrer-Mata A, Sánchez-DelBarrio JC, Guirao-Rico S, Librado P, Ramos-Onsins SE, Sánchez-Gracia A. DnaSP 6: DNA Sequence Polymorphism Analysis of Large Data Sets. Mol Biol Evol. 2017; 34(12):3299–3302.

Star B, Nederbragt AJ, Jentoft S, Grimholt U, Malmstrøm M, Gregers TF, Rounge TB, Paulsen J, Solbakken MH, Sharma A, Wetten OF, Lanzén A, Winer R, Knight J, Vogel JH, Aken B, Andersen Å, Lagesen K, Tooming-Klunderud A, Edvardsen RB, et al. The genome sequence of Atlantic cod reveals a unique immune system. Nature. 2011; 477(7363):207–210.

Tamura K, Stecher G, Peterson D, Filipski A, Kumar S. MEGA6: Molecular Evolutionary Genetics Analysis Version 6.0. Mol Biol Evol. 2013; 30(12):2725–2729.

Tørresen OK, Star B, Jentoft S, Reinar WB, Grove H, Miller JR, Walenz BP, Knight J, Ekholm JM, Peluso P, Edvardsen RB, Tooming-Klunderud A, Skage M, Lien S, Jakobsen KS, Nederbragt AJ. An improved genome assembly uncovers prolific tandem repeats in Atlantic cod. BMC Genomics. 2017; 18:95.

Ursvik A, Breines R, Christiansen JS, Fevolden SE, Coucheron DH, Johansen SD. A mitogenomic approach to the taxonomy of pollocks: Theragra chalcogramma and T. finnmarchica represent one single species. BMC Evolutionary Biology. 2007; 7:86.

Árnason E. Mitochondrial Cytochrome b DNA Variation in the High-Fecundity Atlantic Cod: Trans-Atlantic Clines and Shallow Gene Genealogy. Genetics. 2004; 166(4):1871–1885.

Additional file 1: Table S1.
Geographical distribution of completely sequenced Atlantic cod mitogenomes

Management unit/ Stock/ Location	Number of specimens	Reference
North East Arctic cod NA	97	PRJEB23234/ERP104973
Norwegian coastal cod NC	21	HG514359; PRJEB23234/ERP104973; EU877736–EU877741
North West Atlantic cod NW	32	AM489716; PRJEB23234/ERP104973; EU877710–EU877735
Other locations:		
Baltic Sea cod BS	2	PRJEB23234/ERP104973
North Sea cod NS	3	PRJEB23234/ERP104973
Irish Sea cod IS	1	PRJEB23234/ERP104973

Additional file 2: Table S2. Mitochondrial CytB haplotypes generated from 124 complete Atlantic cod mitogenomes.

Haplo-type[1]	4 4 4 4 4 4 4 4 4 4 5 5 5 5 5 5 5 5 5 5 5 5 5 5 6 6 6 6 6 6 6 6 6 6 6 6 6 7 6 6 6 6 7 8 8 8 9 9 0 0 1 1 2 2 3 4 4 4 4 5 6 6 8 8 9 0 3 3 4 4 5 5 6 7 7 8 9 9 9 0 3 6 7 8 5 1 7 8 0 6 2 8 4 7 2 3 5 1 2 4 7 6 2 5 6 9 5 1 1 7 3 9 5 8 4 3 6 5 1 5 7 6	N
A	C A A T A A A C A A A T A C A G A A G C C T A C T A T T G T A C T T T A G C G G C T	39
E	. . . . . C	35
D	. . . . . A . . . . . T	16
G	. . . . . T . . .	10
C	. . . . . T	3
NI	. . . . . A . . . . . T . . .	3
XI	. . . . . G . . C	2
	. . . . . T	1
	. . . . . G . T	1
	. . . . . T . A . .	1
	. . . . . C	1
P	. . . . . C	1
S	. . . . . T	1
DI	. . . . . A . . . . . T A . . .	1
	. . . . . G . . . . . T	1
	. . . . . G . . . . . T . . .	1
	. . . . . G . . . . . T . . .	1
TI	. . . . . G . . . . . T . . .	1
	. . . . . G . . C . . . A	1
	. . . . . G C	1
	. . . . . C . . . . G	1
	. . . . . C . . . . . C	1
	. . . . . C . . . . . G	1

Note: 1 Position 14463 – 14706, positions are according to the reference sequence NC3 (HG514359)

Additional file 3: Table S3 Non-synonymous substitution in 124 Atlantic cod mitogenomes.

[1]Position	[2]Gene	Codon	Amino acid	Position	Gene	Codon	Amino acid
3154	ND1	ATT → ACT	I → T	5583	COI	CTT → CCT	L → P
3177	ND1	AGT → GGT	S → G	6662	COI	GAT → AAT	D → N
3607	ND1	GAA → GGA	E → G	7278	COII	GTC → ATC	V → I
3646	ND1	TTA → TCA	L → S	7936	ATPase8	GCC → ACC	A → T
3651	ND1	ATG → GTG	M → V	7955	ATPase8	TTC → TCC	F → S
4039	ND2	ATC → ACC	I → T	8316	ATPase6	CTC → TTC	L → F
4138	ND2	ATC → ACC	I → T	8637	ATPase6	GCA → ACA	A → T
4236	ND2	AGC → GGC	S → G	8638	ATPase6	GCA → GTA	A → V
4293	ND2	TTC → CTC	F → L	9050	COIII	CTC → TTC	L → F
4462	ND2	AAC → AGC	N → S	9113	COIII	ACA → GCA	T → A
4491	ND2	ACC → GCC	T → A	9254	COIII	ACA → GCA	T → A
4545	ND2	ATC → GTC	I → V	9442	COIII	TCC → CCC	S → P
4681	ND2	TCT → TGT	S → C	9638	ND3	CTT → TTT	L → F
4754	ND2	ATT → ATG	I → M	9650	ND3	GCT → ACT	A → T
4834	ND2	GAT → GGT	D → G	9662	ND3	ATT → GTT	I → V
4893	ND2	CTC → ATC	L → I	9869	ND3	ACC → GCC	T → A
4974	ND2	GCA → ACA	A → T	10088	ND4L	TTA → TTC	L → F
4975	ND2	GCA → GTA	A → V	10312	ND4L	AAT → AGT	N → S
4996	ND2	ATT → ACT	I → T	[3]10322	ND4	ATG → GTG	M → V
5014	ND2	GCT → GTT	A → V	10370	ND4	TTA → ATA	L → M
5041	ND2	ACC → ATC	T → I	10440	ND4	ACC → ATC	T → I
5453	COI	ACC → TCC	T → S	10484	ND4	TCA → CCA	S → P

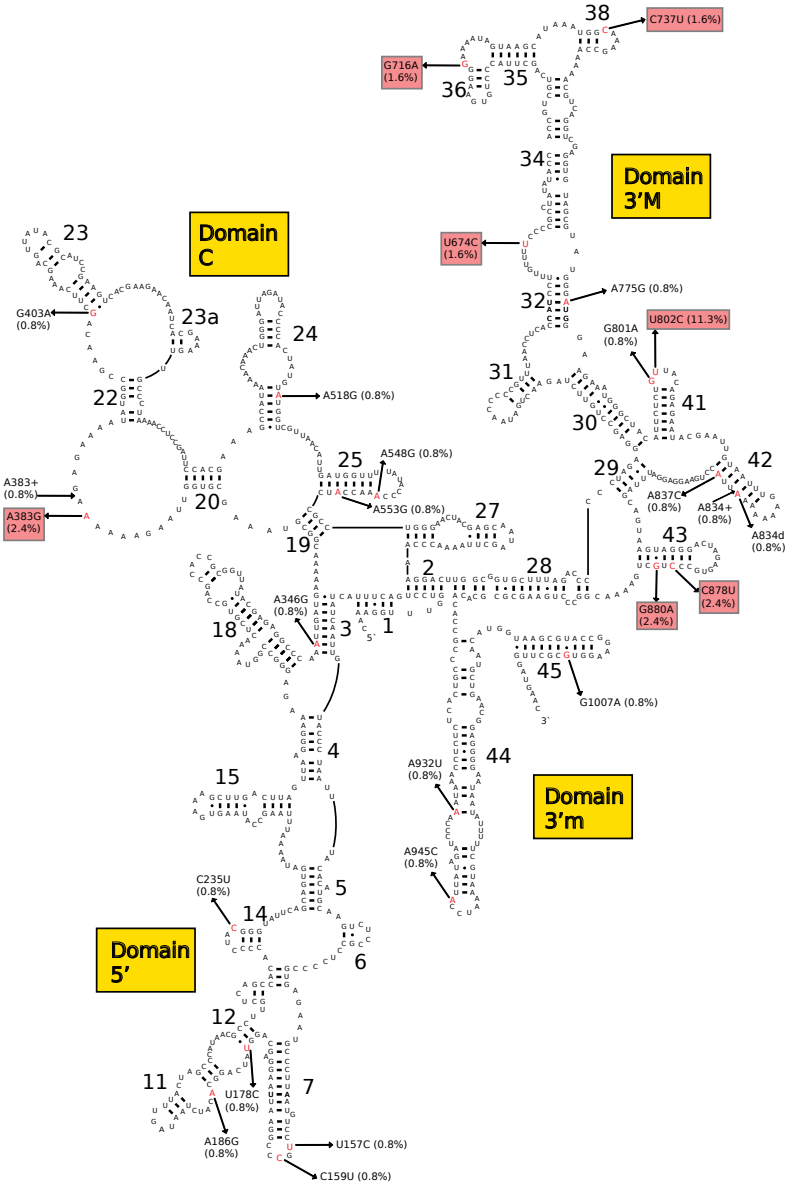
Position	Gene	Codon	Amino acid	Position	Gene	Codon	Amino acid
10591	ND4	ATA → ATT	M → I	12721	ND5	ATG → ATC	M → N
10593	ND4	AAC → AGC	N → S	12737	ND5	AAT → GAT	N → V
10632	ND4	CAG → CGG	Q → R	12906	ND5	GCC → GTC	A → V
10833	ND4	AAC → AGC	N → S	12908	ND5	TTT→ TCT	F → S
10859	ND4	ATC → CTC	I → L	13028	ND5	GCC → ACC	A → T
10889	ND4	CCC → TCC	P → S	13113	ND5	GCA → GTA	A → V
10898	ND4	AGT → GGT	S → G	13168	ND5	ACC → GCC	T → A
11247	ND4	GGA → GAA	G → E	13208	ND5	GTT → ATT	V → I
11264	ND4	ATT → GTT	I → V	13247	ND5	CCC → TCC	P → S
11405	ND4	ATT → GTT	I → V	13274	ND5	CCA → TCA	P → S
11466	ND4	ATT → ACT	I → T	13382	ND5	GCA → ACA	A → T
11484	ND4	AAC → AGC	N → S	13382	ND5	GCA → CCA	A → P
11504	ND4	ATT → GTT	I → V	13383	ND5	GCA → GTA	A → V
11597	ND4	GCC → ACC	A → T	13415	ND5	ATT → GTT	I → V
11678	ND4	GCC → ACC	A → T	13436	ND5	CTT → TTT	L → F
11691	ND4	GGC → GCC	G → A	13458	ND5	AGC → AAC	S → N
12024	ND5	GCC → GTC	A → V	13472	ND5	GTT → ATT	V → I
12074	ND5	GCT → ACT	A → T	13476	ND5	ACA→ ATA	T → M
12096	ND5	AGC → AAC	S → N	13520	ND5	GTC → ATC	V → I
12110	ND5	GCT → ACT	A → T	13601	ND5	ACA → GCA	T → A
12266	ND5	CTC → TTC	L → F	13640	ND5	GCT → ACT	A → T
12489	ND5	GCA → GTA	A → V	13673	ND5	ATA → GTA	M → V
12548	ND5	AAT → GAT	N → D	13739	ND5	ATT → GTT	I → V

Position	Gene	Codon	Amino acid	Position	Gene	Codon	Amino acid
13969	ND6	GCA → GTA	A → V	14828	Cyt b	CAA → CGA	Q → R
13975	ND6	GTG → GCG	V → A	14843	Cyt b	GGT → GAT	G → D
13976	ND6	GTG → ATG	V → M	14980	Cyt b	TCA → CCA	S → P
14144	ND6	TTA → GTA	L → V	15004	Cyt b	CAC → AAC	H → N
14177	ND6	CTG → GTG	L → V	15140	Cyt b	CCT → CAT	P → H
14251	ND6	ACT → AGT	T → S	15322	Cyt b	GTC → ATC	V → I
14542	Cyt b	GTC → ATC	V → I	15382	Cyt b	TTC → CTC	F → L
14589	Cyt b	ATA → ATT	M → I	15409	Cyt b	GTA → ATA	V → M
14695	Cyt b	GTC → ATC	V → I	15457	Cyt b	ACT → GCT	T → A
14719	Cyt b	GCC → ACC	A → T				

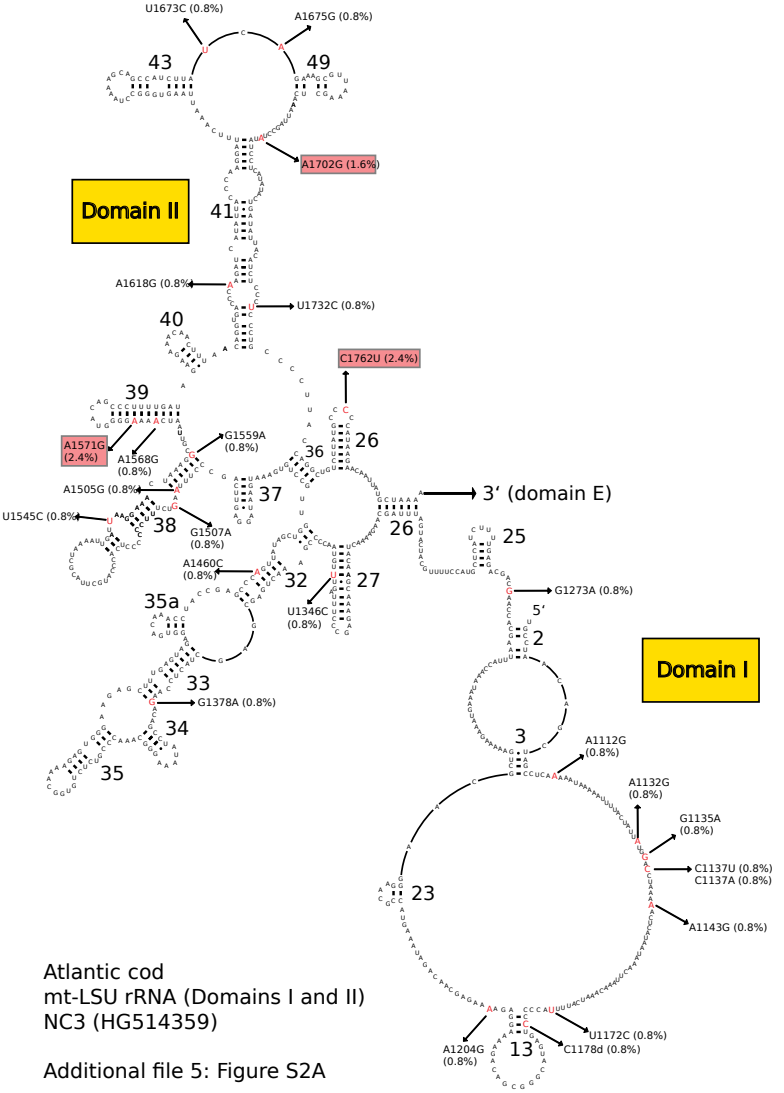
Note: 1 Positions are according to the reference sequence NC3 (HG514359)

2 ND1 - NADH dehydrogenase subunit 1; ND2 - NADH dehydrogenase subunit 2; ND3 - NADH dehydrogenase subunit 3; ND4 – NADH dehydrogenase subunit 4; ND4L - NADH dehydrogenase subunit 4L; ND5 - NADH dehydrogenase subunit 5; ND6 – NADH dehydrogenase subunit 6; COI - cytochrome c oxidase subunit I; COII - cytochrome c oxidase subunit II; COIII - cytochrome c oxidase subunit III; ATPase8 - ATP synthase 8; ATPase6 -ATP synthase 6 ; Cyt b - Cytochrome B.

3 Start codon of ND4

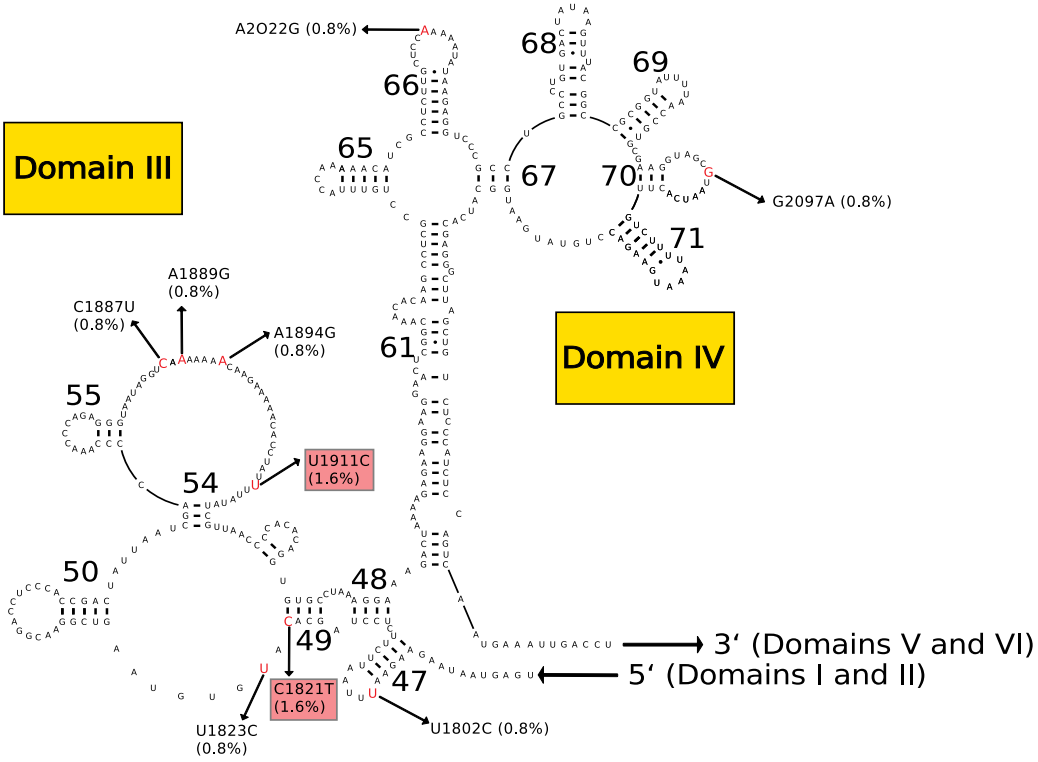


Additional file 4: Figure S1



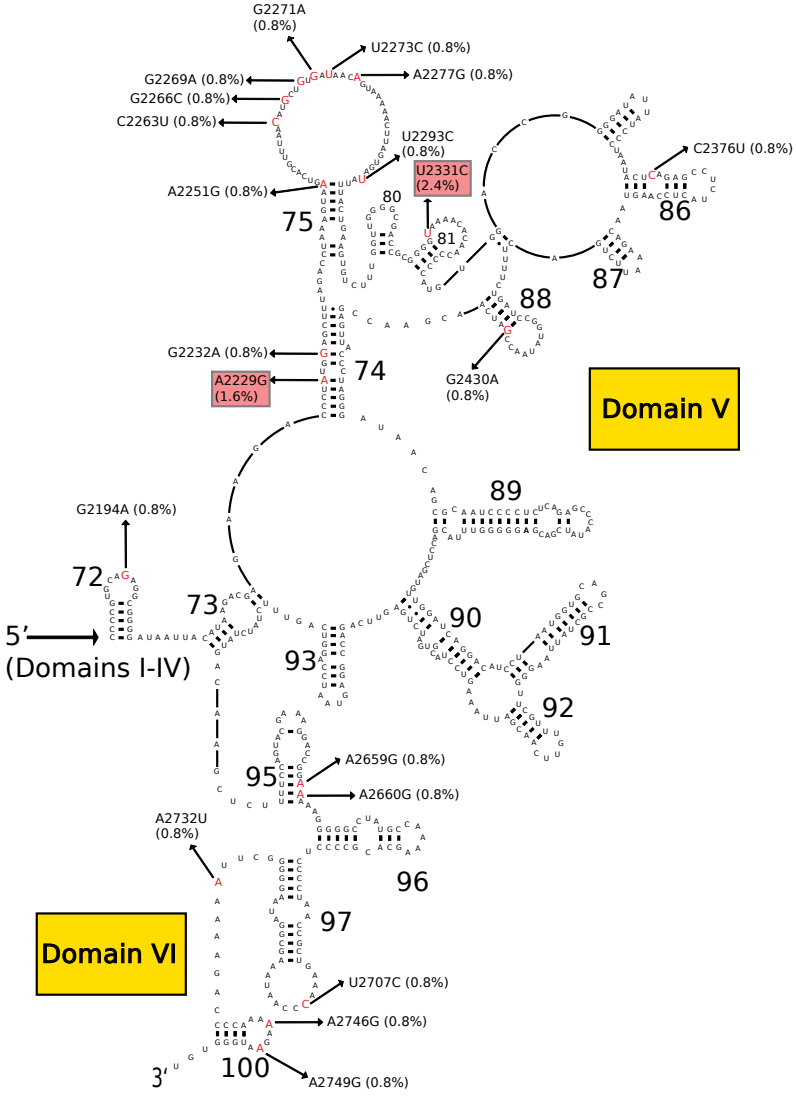
Atlantic cod
mt-LSU rRNA (Domains I and II)
NC3 (HG514359)

Additional file 5: Figure S2A



Atlantic cod
mt-LSU rRNA (Domains III and IV)
NC3 (HG514359)

Additional file 5: Figure S2B



Atlantic cod
mt-LSU rRNA (Domains V and VI)
NC3 (HG514359)

Additional file 5: Figure S2C

Paper II

Marine Genomics 15 (2014) 103–108

Contents lists available at ScienceDirect

Marine Genomics

journal homepage: www.elsevier.com/locate/margen

Mitogenome sequence variation in migratory and stationary ecotypes of North-east Atlantic cod

Bård O. Karlsen [a,b,1,2], Åse Emblem [a,2], Tor E. Jørgensen [b], Kevin A. Klingan [b], Jarle T. Nordeide [b], Truls Moum [b], Steinar D. Johansen [a,b,*]

[a] Department of Medical Biology — RNA Lab, Faculty of Health Sciences, University of Tromsø, N-9037 Tromsø, Norway
[b] Marine Genomics Group, Faculty of Biosciences and Aquaculture, University of Nordland, N-8059 Bodø, Norway

ARTICLE INFO

Article history:
Received 4 October 2013
Received in revised form 11 December 2013
Accepted 6 January 2014
Available online 21 January 2014

Keywords:
Atlantic cod
Marine genomics
mtDNA
Population mitogenomics
SOLiD sequencing

ABSTRACT

Sequencing of mitochondrial gene fragments from specimens representing a wide range of geographical locations has indicated limited population structuring in Atlantic cod (*Gadus morhua*). We recently performed whole genome analysis based on next-generation sequencing of two pooled ecotype samples representing offshore migratory and inshore stationary cod from the North-east Atlantic Ocean. Here we report molecular features and variability of the 16.7 kb mitogenome component that was collected from the datasets. These sequences represented more than 25 times coverage of each individual and more than 1100 times coverage of each ecotype sample. We estimated the mitogenome to have evolved 14 times more rapidly than the nuclear genome. Among the 365 single nucleotide polymorphism (SNP) sites identified, 121 were shared between ecotypes, and 151 and 93 were private within the migratory and stationary cod, respectively. We found 323 SNPs to be located in protein coding genes, of which 29 were non-synonymous. One synonymous site in *ND2* was likely to be under positive selection. F_{ST} measurements indicated weak differentiation in *ND1* and *ND2* between ecotypes. We conclude that the Atlantic cod mitogenome and the nuclear genome apparently evolved by distinct evolutionary constraints, and that the reproductive isolation observed from whole genome analysis was not visible in the mtDNA sequences.

1. Introduction

Knowledge on population structures is important in order to preserve fishery resources, local populations, and to assess the genetic pool of marine fish species. The Atlantic cod is an important species for fisheries and has in recent years become attractive for aquaculture (Johansen et al., 2009). Atlantic cod is widely distributed across the North Atlantic Ocean and complex patterns of genotypic and phenotypic variation have been revealed in recent years by means of single nucleotide polymorphisms (SNPs), microsatellites, mitochondrial genomes (mitogenomes; mtDNA), allozymes, restriction fragment length polymorphism loci, and phenotypic characters (e.g. Pogson et al., 1995; Árnason, 2004; Pogson and Mesa, 2004; Coulson et al., 2006; Carr and Marshall, 2008; Moen et al., 2008; Johansen et al., 2009; Hubert et al., 2010). Apparently, the population genetic structuring in Atlantic cod is both due to isolation by distance and vicariance, as well as adaptive variation along environmental clines and adaptive divergence among sympatric populations (e.g. Nielsen et al., 2009; Bradbury et al., 2010, 2013; Karlsen et al., 2013). The status of two phenotypically divergent ecotypes of cod that occur sympatrically in the North-east Atlantic, the offshore migratory North-east Arctic (NA) cod and inshore stationary Norwegian coastal (NC) cod, has been addressed by a number of studies and evidence concerning their connectedness is still incomplete (reviewed in Nordeide et al., 2011). Differential markers of NA and NC cod have been reported at blood type E, hemoglobin (*Hb* I) and pantophysin (*Pan* I) loci, while few significant differences have been found in microsatellite and SNP data, although temperature associations have been coupled with some SNPs supporting an adaptive diversity (Bradbury et al., 2010; Andersen, 2012).

The mitogenome of Atlantic cod is small (about 16.7 kb) and circular, and holds a gene content identical to that of most vertebrates (Johansen et al., 1990, 2009; Johansen and Bakke, 1996). The use of mitochondrial gene sequences as biomarkers for intra- and interspecies studies has become popular. Presumably near-neutral evolution, maternal inheritance, and a small compact size have made mitogenomes an easy-to-use population tool. However, in the last years it has been argued that mitogenomes may be less suited for vertebrate population history assessments due to increasing evidence of recombination events, positive selection, erratic evolutionary rate, and heteroplasmy (e.g. Brown et al., 1979; Bazin et al., 2006; Mjelle et al., 2008; Galtier et al., 2009; He et al., 2010; Ameur et al., 2011). Árnason (2004) examined a

* Corresponding author at: Department of Medical Biology — RNA Lab, Faculty of Health Sciences, University of Tromsø, N-9037 Tromsø, Norway.
E-mail address: steinar.johansen@uit.no (S.D. Johansen).
[1] Current address: Department of Laboratory Medicine, Division of Molecular Diagnostics, Nordland Hospital Trust, N-8092 Bodø, Norway.
[2] Contributed equally.

1874-7787/$ – see front matter © 2014 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.margen.2014.01.001

250 bp segment of the mitochondrial *CytB* from almost 1300 Atlantic cod specimens representing the species' range in the North Atlantic Ocean. A shallow genealogy and lack of clear population structuring were detected, and he found that sequence variation was mainly due to synonymous substitutions and that limited selection was operating on these substitutions. Marshall et al. (2009) studied the complete mtDNA of 32 cod specimens and found that variation is consistent with mutation rate heterogeneity across the mitogenome and complies with the nearly neutral theory of molecular evolution. Elevated substitution rates were found in ND genes, but the rate heterogeneity among genes was mainly due to changes at synonymous sites (Marshall et al., 2009).

Sequencing of pooled samples of specimens offers an attractive approach to investigate mutation rate and heterogeneity of mitogenomes within and between populations, while the genealogical information content of individual haplotypes is sacrificed (Futschik and Schlotterer, 2010; Kofler et al., 2011). Next-generation sequencing (NGS) is well suited for such analyses and reveals sequence information of a number of mitochondrial genomes at the same time, keeping labor and running cost at a minimum. The pooled sample analyses of mitogenomes presented here is part of a population genomic project that apply NGS in assessing genetic diversity and divergence between the migratory NA and stationary NC cod ecotypes. Recently we reported the divergence in nuclear DNA of the NA and NC cod ecotypes, where mitochondrial genome sequences were not included in the genome assembly (Karlsen et al., 2013). Here we aim to address the utility and limitation of mitogenome sequences as a biomarker in population assessments. Each ecotype was represented by 44 specimens, which were pooled and deep sequenced by SOLiD ligation sequencing (Karlsen et al., 2013). Reads originating from the mitogenomes were mapped to a Sanger sequenced reference Atlantic cod mtDNA. Genetic variation and differentiation between ecotypes were then calculated from variation at SNP loci applying BayeScan and the PoPoolation toolbox.

2. Materials and methods

2.1. Sample collection, DNA isolation, and DNA sequencing

These procedures are essentially as described by Karlsen et al. (2013). In summary, muscle tissue was collected from 44 Atlantic cod specimens captured in the Barents Sea north of 72°N (migratory NA cod) and 44 specimens from Holandsfjord, 66° 71′N 13° 68′E (stationary NC cod). DNA was isolated applying Roche High Pure PCR Template Preparation Kit and Nexttec™ Genomic DNA Isolation Kit for Tissue and Cells according to manufacturers' specifications. Purified total DNA from each individual was quality verified, and equal amounts (100 ng) from each individual in one ecotype were titrated, pooled, and sheared to an average size of 125 bp with a Covaris S2 system. SOLiD Fragment libraries were prepared according to the standard protocol given by the manufacturer and sequencing performed on a SOLiD 3 Plus system (Applied Biosystems). The two libraries were sequenced on separate SOLiD runs. The NC run produced 298,850,963 usable reads and the NA run produced 328,107,403 usable reads (Table 1). Sanger sequencing of the complete mitogenome isolated from muscle tissue from the Norwegian coastal cod specimen NC3 (69° 44′N 19° 37′E) was prepared and performed according to the previously described protocol (Ursvik et al., 2007).

2.2. Data analysis

The NC3 Atlantic cod mitogenome (16,696 bp) was used as a reference for mapping of the NA and NC pooled mitochondrial genome reads. Reference mapping was performed in the SOLiD Bioscope Software 1.3 with standard settings for fragment libraries. The CLC Genomics Workbench 4.7.2 was used to create SNP statistics for mapping of pooled samples with the following stringency, and for each position there had to be at least 400 reads mapping to the reference nucleotide. For the setting of quality window length 11 (default), max gaps 2 (default), minimum quality of central base 25 (default: 20), minimum average quality of surround bases 25 (default: 15), and minimum variant frequency 2%. Gff3 files were created with GMAP version 2012-01-11 (Wu and Watanabe, 2005) for the mitochondrial genes mapping to NC3. SNPs were then extracted from Bam files dictated by the gff3 files.

Table 1
Summary statistics for SOLiD sequencing and mitogenome mapping.

	NA ecotype[a]	NC ecotype[b]
Total raw reads for mapping[c]	328,107,403	298,850,963
NC3 reference length (bp)[d]	16,696	16,696
Total mapped reads	497,307	387,233
Average mapped read length (nt)	46.81	47.52
Average coverage	1,394	1,102

[a] Migratory NA cod (n = 44).
[b] Stationary NC cod (n = 44).
[c] From Karlsen et al. (2013).
[d] Sanger sequenced NC3 mitogenome (HG514359) applied for mapping in CLC Genomic Workbench.

2.3. F_{ST} statistics and BayeScan

F_{ST} measures the genetic differentiation between populations, with values ranging from 0 to 1. If the F_{ST} is high, the allele frequencies are highly differentiated while low F_{ST} signifies little differentiation. Large populations with migration generally show low F_{ST} values while small populations with little migration tend to show higher F_{ST} values (Holsinger and Weir, 2009). We created an R script (v 2.13.1, R Development Core Team, 2011), provided in Karlsen et al. (2013), to calculate F_{ST} value for each of the detected SNPs. To identify outlier loci putatively under selection, the Bayesian approach implemented in the BayeScan Version 2.01 (Foll and Gaggiotti, 2008) was applied (default settings). This method is based on the finite-island model of Beaumont and Nichols (1996), and infers significance from the posterior probability of a locus.

2.4. Tajima's D and π analysis

Tajima's D is a statistical test that aims to distinguish between neutral evolution of a locus (genetic drift) versus a non-random process like natural selection. It compares the average number of nucleotide differences between pairs of sequences to the total number of segregating sites. A neutrally evolving locus in a population in equilibrium, where the allele frequencies remain constant over time, will have a D value of approximately zero. A locus with increased genetic variation will generate a positive Tajima's D value and indicates balancing selection, while a locus with decreased genetic variation gives a negative D value and indicates purifying selection. Tajima's D is also influenced by demographic processes such that a population reduction or bottleneck can give the same locus a transient positive value, while a population expansion (i.e. after a bottleneck) can create a low or negative D value (Tajima, 1989; Oleksyk et al., 2010). The Tajima's π (nucleotide diversity) is the mean number of pairwise differences. Tajima's D and π analyses were performed applying the PoPoolation toolbox (Kofler et al., 2011), with the pileup files created with SAMtools (Li et al., 2009) as instructed in the PoPoolation manual. Window size was set to 30 nt and step size to 15 nt. We used Gnuplot Version 4.4 patchlevel 3 to plot the genome-wide scan of Tajima's D and π. The Gnuplot scripts for statistic plotting were taken from the webpage: http://www.phyast.pitt.edu/~zov1/gnuplot/html/statistics.html.

3. Results and discussion

3.1. Mitogenome sequencing, mapping and SNP detection

Pan I genotyping of the 44 NA cod and the 44 NC cod specimens supported that these ecotype samples were representative of the migratory and stationary North-east Atlantic cod, respectively (Karlsen et al., 2013). Approximately 300 million single reads from each pooled ecotype sample were generated from SOLiD ligation sequencing (Table 1). This corresponded to approximately 1400 and 1100 times coverage of the NA and NC cod mitogenomes, respectively. When applying stringent SNP statistics criteria in mapping to the NC3 reference sequence we found 323 SNP loci in protein coding genes (365 SNPs in total), and 9% of these substitutions (29 SNPs) were non-synonymous (Supplementary Table S1). About 50% of the non-synonymous substitutions (14 out of 29 SNPs) were located in *ND5* (Supplementary Table S2). Most SNPs were transitions (Supplementary Table S3). Among the latter, one position (A14144C) resulted in a conservative (L to V) amino acid shift in ND6 of NA cod (Supplementary Table S2), and four positions (T3113, G4499, A4748, and A9832) harbored tri-allelic SNPs. A total of 17 SNPs were detected in the control region (containing the D-loop), and the rRNA genes and tRNA genes were as expected highly conserved (Supplementary Table S4). These structural RNA gene SNPs were present in low frequencies among the NA and NC ecotypes, corresponding to one or very few individuals. Twelve of the 22 annotated tRNA genes, however, contained no SNP sites.

We detected more mitogenome SNPs within the NA cod sample (272 SNPs) than the NC cod sample (214 SNPs) (Table 2). A major fraction (91%) of the SNP sites translates into synonymous amino acids. Furthermore, about 33% of the mitogenome SNPs were shared in the NA and NC cod ecotypes, which is similar proportion as observed for the nuclear genome SNPs. The distribution of private SNPs between NA and NC cod, however, is significantly different compared to that of the nuclear genome (Table 2). An SNP was considered private when detected in only one of the ecotypes. However, it cannot be excluded that some private SNPs might have originated from a single individual. In the nuclear genome the NA and NC cod harbored 27% and 40% private SNPs, respectively. The correspondent fractions in the mitogenome were 41% and 26%. This difference suggests that mitogenomes and the nuclear genomes have been influenced by different evolutionary mechanisms. Similar observations were made in cichlid and marine goby where phylogenetic reconstructions of mitochondrial and nuclear data revealed that the mitogenome retained patterns of historical events, while nuclear data indicated genetic cohesion among different groups and present patterns congruent with phenotypic population indicators (Egger et al., 2007; Larmuseau et al., 2010).

Nuclear copies of mitochondrial DNA sequences (NUMTs) have been reported in several animal systems (Bensasson et al., 2001; Tsuji et al., 2012), and could potentially lead to an overestimation of the assessed SNP diversity. We expect some NUMTs to be present in the Atlantic cod genome, but so far not reported (Star et al., 2011). Inspections of mapped reads in our study revealed no suspicious stop-codons, breakpoints or unusual SNPs consistent with the presence of nuclear mitochondrial sequences. Whereas the possibility of NUMT in our data set may not be theoretically excluded, we conclude that at least the great majority of assessed SNPs represent authentic mitogenome sequences.

Table 2
Comparison of SNP distribution in the mitogenome and nuclear genome of NA and NC cod ecotypes.

	Mitogenome	Nuclear genome[a]
Total SNPs in NA and NC	365	961,619
Shared SNPs in NA and NC	121 (33.2%)[b]	321,342 (33.4%)
Total SNPs in NA	272 (74.5%)	581,962 (60.5%)
Private SNPs in NA	151 (41.4%)	260,620 (27.1%)
Total SNPs in NC	214 (58.6%)	700,999 (72.9%)
Private SNP in NC	93 (25.5%)	379,657 (39.5%)
Size of reference sequence	16.7 Kb	611,000 Kb
SNPs per Kb reference sequence	21.9	1.6

[a] From Karlsen et al. (2013).
[b] All % values are fractions of total NA + NC SNPs.

3.2. Genetic variability in the mitogenome of the NA and NC cod ecotypes

F_{ST} values were calculated for the NA and NC cod datasets (Fig. 1). Estimation of F_{ST} mean values for all protein coding mitochondrial genes showed that *ND1* and *ND2* have an elevated differentiation between ecotypes than other mitochondrial genes investigated (Fig. 1a). BayeScan was used to identify F_{ST} outlier loci postulated to be under natural selection with a posterior probability threshold of 0.99. This scan indicated that the A4748 site was under positive selection (Fig 1b). The A4748 locus represented a tri-allelic synonymous SNP in the NC cod ecotype (Supplementary Table S3) and a synonymous substitution in the NA cod ecotype. The F_{ST} value (0.054) was significant and could be an indication of balancing selection at this locus. Increased heterogeneity in *ND1* and *ND2* has previously been reported in Atlantic cod mitogenomes (Marshall et al., 2009), and a departure from neutrality has been suggested in human *ND2* (Wise et al., 1998). Additional Atlantic cod samples from populations at various geographical locations within the species' range will probably illuminate this potential selection further. The neutral theory predicts that the ratio of synonymous and non-synonymous SNPs should be the same within and between species (McDonald and Kreitman, 1991), which agrees with our findings (8% and 9% in NA and NC cod ecotypes, respectively) (Supplementary Table S1).

3.3. Tajima's D and π sweep analyses

The statistical tests of Tajima's D and π are presented as a genomic scan plot (Fig. 2). Tajima's D is negative in all cases while Tajima's π is always positive. We found *CytB*, control region, and ND genes (the latter boxed in Fig. 2) to have higher D and π values than the ribosomal RNA, CO and ATP genes, exceeding the standard deviation at a more frequent basis. The Tajima's D and pi values were largely correlating between the two ecotypes, with a few exceptions and notably in *ND2*. Tajima's D and π values were low and suggested little variation between the ecotypes. The NA cod and NC cod D and π values were found to be highly similar, but with few notable differences (Fig. 2). We observed that the ND genes exceed the standard deviation on a more frequent basis, but the negative Tajima's D values (Fig. 3a) contradict the hypothesis of balancing selection in *ND2*. The *ND2* does however obtain the highest Tajima's D values in both populations (Figs. 2, 3, Table 2) indicating a relaxed purifying selection relative to other mitochondrial protein coding genes.

4. Concluding remarks

The genomic divergence analyses of the offshore migratory NA and inshore stationary NC cod ecotypes based on SOLiD high-throughput sequencing represented a unique opportunity to study and directly compare the mitochondrial and the nuclear genomes in the same experimental dataset. By performing statistical analysis on the mitogenome we calculated the differentiation (F_{ST}) between the NA and NC cod ecotypes, the nucleotide diversity (Tajima's π), and the frequency spectrum of nucleotide polymorphisms (Tajima's D). We found that pooled population sequencing provides an economical means to generate valuable information of the cod mitogenome variability. A 14 time higher substitution rate in the mitogenome compared to that of the nuclear genome was estimated, which is similar to the order of magnitude in mutation rate difference reported for animals in general (e.g. Bar-Yaacov et al., 2012). The nuclear genome uses other mechanisms

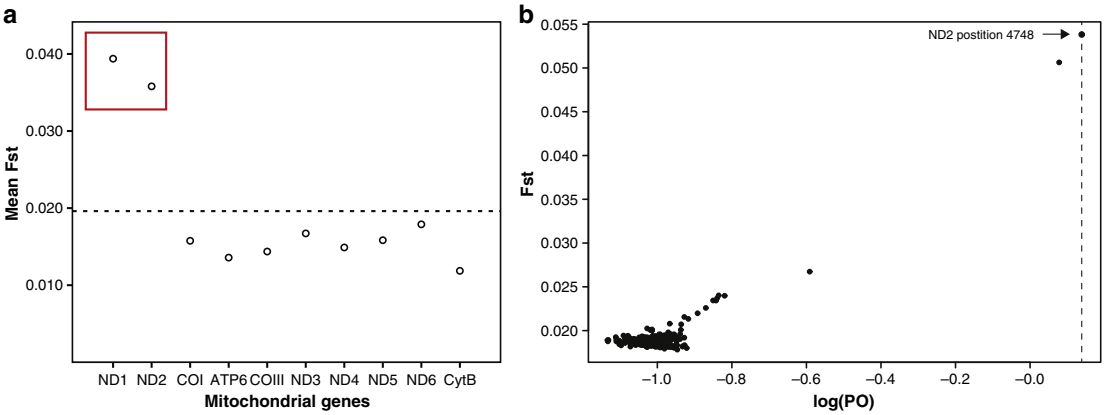


Fig. 1. Graphical presentations of statistical assessments. (a) F_{ST} values from the mitogenome. Mean F_{ST} value for each of the mitochondrial protein coding genes. *COII*, *ATP8* and *ND4L* values were not calculated due to few SNPs. A dashed horizontal line represents the overall mean F_{ST} value. *ND1* and *ND2* values are indicated (red box). (b) Scan for identification of F_{ST} outlier loci postulated to be under natural selection. BayeScan was used to scan for potential outliers in the mitochondrial protein coding genes of the NA and NC cod ecotypes. Dots indicate loci with SNPs found in both ecotypes (matching) and only one of the ecotypes (private). The dashed vertical line corresponds to a posterior probability of 0.99 for a locus to be under selection. One SNP locus, corresponding to mitogenome position A4748 in *ND2*, reached this threshold. PO; posterior odds.

providing rapid genomic changes and adaptations, such as homologous recombination and transposable element mobility, which can cause further rapid genomic adaptation by large genomic shifts (Fedoroff, 2012). A high mitochondrial substitution rate will therefore not necessarily reflect adaptations apparent in the nuclear genome. We found that the Atlantic cod mitogenome hardly reflected any strong adaptations or ongoing differentiation events between the migratory and stationary ecotypes despite the fact that a significant number of private

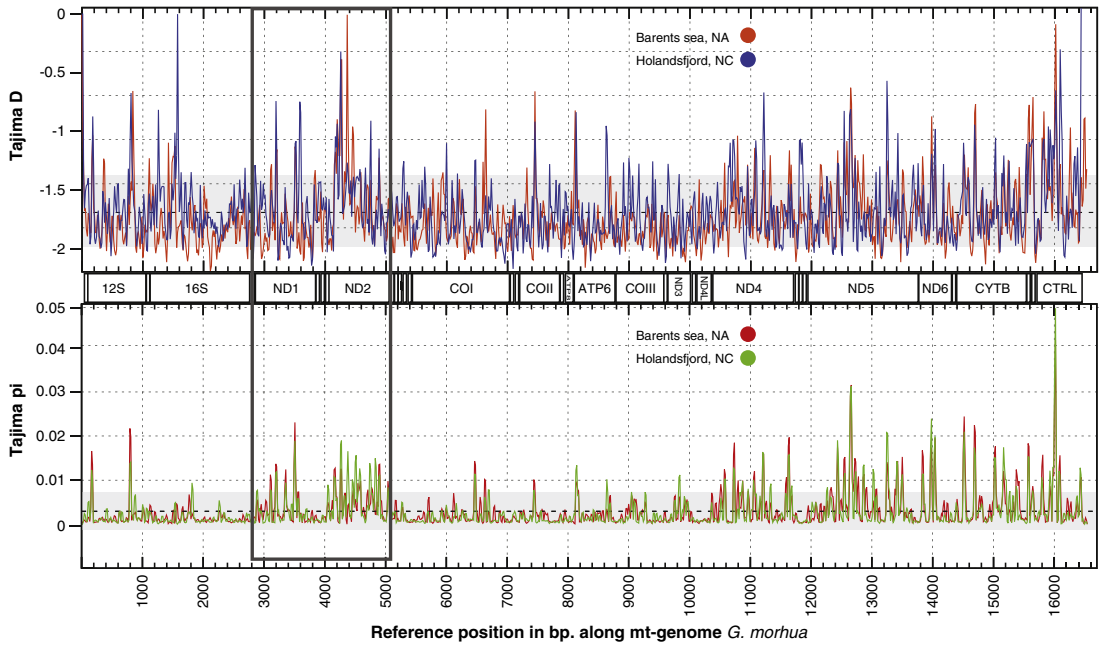


Fig. 2. Plot of Tajima's D and π along the 16,696 bp NC3 mitogenome reference with gene markers and boundaries. The NA and NC cod ecotype mitogenomes were scanned in sliding windows of 30 nt with 15 nt overlap, Tajima' D (departures from neutrality) and Tajima' π (mean pairwise differences) were calculated for each sliding window. NA cod ecotype, red plots; NC cod ecotype, blue/green plots. Light blue shading indicates one standard deviation. The dashed box illustrates *ND1* and *ND2*. *ND1–6* (NADH dehydrogenase subunit 1–6 genes), *COI–III* (cytochrome *c* oxidase subunit I–III genes), *ATP6* and *ATP8* (ATPase subunit 6 and 8 genes) and CytB (cytochrome B gene), and *12S* and *16S* (small and large subunit ribosomal RNA genes, respectively). CTRL (control region containing the D-loop). The 22 transfer RNA genes are boxed but not annotated in the figure.

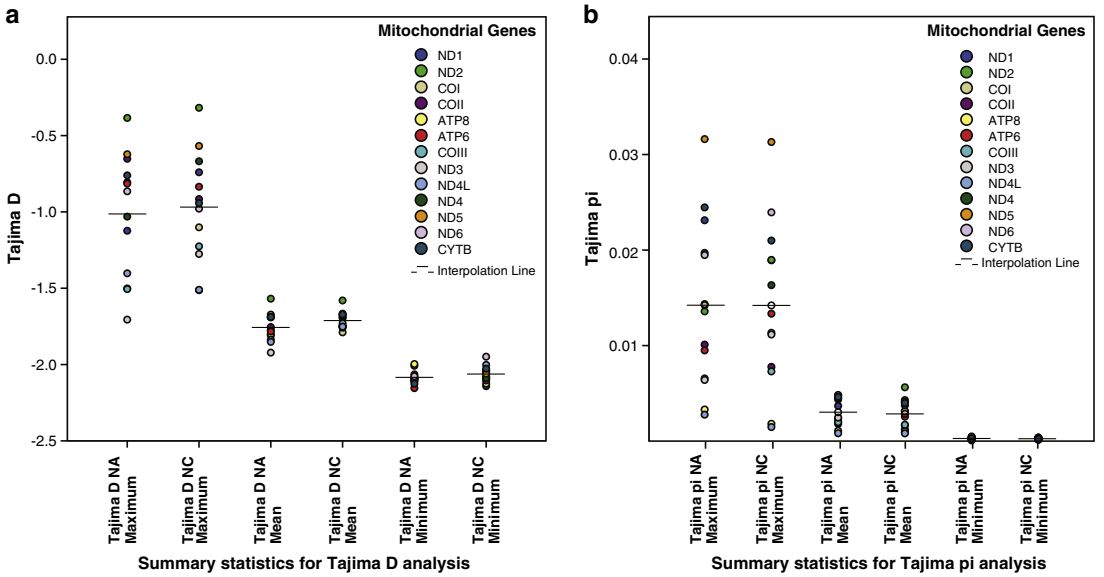


Fig. 3. Summary statistics for Tajima's D and π values from the mitogenome sliding window analysis. (a) Maximum, mean and minimum values of Tajima's D. (b) Maximum, mean and minimum values of Tajima's π of all mitochondrial protein coding genes in the NA and NC cod ecotypes.

mitochondrial SNPs were identified in each ecotype. These results diverge from those reported from the whole nuclear genome of NA and NC cod where significant SNP divergences were seen in 25 scaffolds (Karlsen et al., 2013).

Acknowledgments

We thank the members from the Marine Genomics Lab at University of Nordland (UiN) and the RNA Lab at University of Tromsø (UiT) for encouragement and discussions, and Andrew Morse at Life Technologies for his excellent contribution to the installation of Bioscope on the Stallo supercluster. Ove Nicolaisen and Espen Søreng are acknowledged for sampling Atlantic cod specimens in Holandsfjord. The Institute of Marine Research and Asgeir Aglen kindly allowed Francois Botha to collect muscle samples of Atlantic cod at their demersal fish survey by "RV G.O. Sars" in the Barents Sea (WD #8, AFWG 2007). This work was supported by a grant from UiN and UiT.

Sequencing data accessibility

This mitogenome pooled sample study in Atlantic cod is a part of a genomic project applying NGS of two Atlantic cod ecotypes. The genome sequence data is available at NCBI — Bioproject, under the accession number PRJNA78311. The complete mitogenome sequence from the Atlantic cod NC3 reference specimen is available from the European Nucleotide Archive (ENA) under the accession number HG514359.

Appendix A. Supplementary data

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.margen.2014.01.001.

References

Ameur, A., Stewart, J.B., Freyer, C., Hagstrom, E., Ingman, M., Larsson, N.G., Gyllensten, U., 2011. Ultra-deep sequencing of mouse mitochondrial DNA: mutational patterns and their origins. PLoS Genet. 7, e1002028.

Andersen, Ø., 2012. Hemoglobin polymorphisms in Atlantic cod — a review of 50 years of study. Mar. Genomics 8, 59–65.

Árnason, E., 2004. Mitochondrial cytochrome B DNA variation in the high-fecundity Atlantic cod: trans-Atlantic clines and shallow gene genealogy. Genetics 166, 1871–1885.

Bar-Yaacov, D., Blumberg, A., Mishmar, D., 2012. Mitochondrial-nuclear co-evolution and its effects on OXPHOS activity and regulation. Biochim. Biophys. Acta 1819, 1107–1111.

Bazin, E., Glemin, S., Galtier, N., 2006. Population size does not influence mitochondrial genetic diversity in animals. Science 312, 570–572.

Beaumont, M.A., Nichols, R.A., 1996. Evaluating loci for use in the genetic analysis of population structure. Proc. R. Soc. Lond. B Biol. 263, 1619–1626.

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

Bradbury, I.R., Hubert, S., Higgins, B., Borza, T., Bowman, S., Paterson, I.G., Snelgrove, P.V., Morris, C.J., Gregory, R.S., Hardie, D.C., Hutchings, J.A., Ruzzante, D.E., Taggart, C.T., Bentzen, P., 2010. Parallel adaptive evolution of Atlantic cod on both sides of the Atlantic Ocean in response to temperature. Proc. Biol. Sci. 277, 3725–3734.

Bradbury, I.R., Hubert, S., Higgins, B., et al., 2013. Genomic islands of divergence and their consequences for the resolution of spatial structure in an exploited marine fish. Evol. Appl. 6, 450–461.

Brown, W.M., George Jr., M., Wilson, A.C., 1979. Rapid evolution of animal mitochondrial DNA. Proc. Natl. Acad. Sci. U. S. A. 76, 1967–1971.

Carr, S.M., Marshall, H.D., 2008. Intraspecific phylogeographic genomics from multiple complete mtDNA genomes in Atlantic cod (*Gadus morhua*): origins of the 'codmother', transatlantic vicariance and midglacial population expansion. Genetics 180, 381–389.

Coulson, M.W., Marshall, H.D., Pepin, P., Carr, S.M., 2006. Mitochondrial genomics of gadine fishes: implications for taxonomy and biogeographic origins from whole-genome data sets. Genome 49, 1115–1130.

Egger, B., Koblmüller, S., Sturmbauer, C., Sefc, K.M., 2007. Nuclear and mitochondrial data reveal different evolutionary processes in the Lake Tanganyika cichlid genus *Tropheus*. BMC Evol. Biol. 7, 137.

Fedoroff, N.V., 2012. Transposable elements, epigenetics, and genome evolution. Science 338, 758–767.

Foll, M., Gaggiotti, O., 2008. A genome-scan method to identify selected loci appropriate for both dominant and codominant markers: a Bayesian perspective. Genetics 180, 977–993.

Futschik, A., Schlotterer, C., 2010. The next generation of molecular markers from massively parallel sequencing of pooled DNA samples. Genetics 186, 207–218.

Galtier, N., Nabholz, B., Glémin, S., Hurst, G.D., 2009. Mitochondrial DNA as a marker of molecular diversity: a reappraisal. Mol. Ecol. 18, 4541–4550.

He, Y., Wu, J., Dressman, D.C., Iacobuzio-Donahue, C., Markowitz, S.D., Velculescu, V.E., Diaz Jr., L.A., Kinzler, K.W., Vogelstein, B., Papadopoulos, N., 2010. Heteroplasmic mitochondrial DNA mutations in normal and tumour cells. Nature 464, 610–614.

Holsinger, K.E., Weir, B.S., 2009. Genetics in geographically structured populations: defining, estimating and interpreting F_{ST}. Nat. Rev. Genet. 10, 639–650.

Hubert, S., Higgins, B., Borza, T., Bowman, S., 2010. Development of a SNP resource and a genetic linkage map for Atlantic cod (*Gadus morhua*). BMC Genomics 11, 191.

Johansen, S., Bakke, I., 1996. The complete mitochondrial DNA sequence of Atlantic cod (*Gadus morhua*): relevance to taxonomic studies among codfishes. Mol. Mar. Biol. Biotechnol. 5, 203–214.

Johansen, S., Guddal, P.H., Johansen, T., 1990. Organization of the mitochondrial genome of Atlantic cod, *Gadus morhua*. Nucleic Acids Res. 18, 411–419.

Johansen, S.D., Coucheron, D.H., Andreassen, M., Karlsen, B.O., Furmanek, T., Jørgensen, T.E., Emblem, Å., Breines, R., Nordeide, J.T., Moum, T., Nederbragt, A.J., Stenseth, N.C., Jakobsen, K.S., 2009. Large-scale sequence analyses of Atlantic cod. Nat. Biotechnol. 25, 263–271.

Karlsen, B.O., Klingan, K., Emblem, Å., Jørgensen, T.E., Jüterbock, A., Furmanek, T., Hoarau, G., Johansen, S.D., Nordeide, J.T., Moum, T., 2013. Genomic divergence between migratory and stationary ecotypes of Atlantic cod. Mol. Ecol. 22, 5098–5111.

Kofler, R., Orozco-terWengel, P., De Maio, N., Pandey, R.V., Nolte, V., Futschik, A., Kosiol, C., Schlötterer, C., 2011. PoPoolation: a toolbox for population genetic analysis of next generation sequencing data from pooled individuals. PLoS One 6, e15925.

Larmuseau, M.H., Raeymaekers, J.A., Hellemans, B., Van Houdt, J.K., Volckaert, F.A., 2010. Mito-nuclear discordance in the degree of population differentiation in a marine goby. Heredity 105, 532–542.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., et al., 2009. The sequence alignment/map (SAM) format and SAMtools. Bioinformatics 25, 2078–2079.

Marshall, H.D., Coulson, M.W., Carr, S.M., 2009. Near neutrality, rate heterogeneity, and linkage govern mitochondrial genome evolution in Atlantic cod (*Gadus morhua*) and other gadine fish. Mol. Biol. Evol. 26, 579–589.

McDonald, J.H., Kreitman, M., 1991. Adaptive protein evolution at the Adh locus in *Drosophila*. Nature 351, 652–654.

Mjelle, K.A., Karlsen, B.O., Jørgensen, T.E., Moum, T., Johansen, S.D., 2008. Halibut mitochondrial genomes contain extensive heteroplasmic tandem repeat arrays involved in DNA recombination. BMC Genomics 9, 10.

Moen, T., Hayes, B., Nilsen, F., Delghandi, M., Fjalestad, K.T., Fevolden, S.E., Berg, P.R., Lien, S., 2008. Identification and characterization of novel SNP markers in Atlantic cod: evidence for directional selection. BMC Genet. 9, 18.

Nielsen, E.E., Hemmer-Hansen, J., Poulsen, N.A., Loeschcke, V., Moen, T., Johansen, T., Mittelholzer, C., Taranger, G.L., Ogden, R., Carvalho, G.R., 2009. Genomic signatures of local directional selection in a high gene flow marine organism; the Atlantic cod (*Gadus morhua*). BMC Evol. Biol. 9, 276.

Nordeide, J.T., Johansen, S.D., Jørgensen, T.E., Karlsen, B.O., Moum, T., 2011. Population connectivity among migratory and stationary cod *Gadus morhua* in the Northeast Atlantic — a review of 80 years of study. Mar. Ecol. Prog. Ser. 435, 269–283.

Oleksyk, T.K., Smith, M.W., O'Brien, S.J., 2010. Genome-wide scans for footprints of natural selection. Philos. Trans. R. Soc. Lond. B Biol. Sci. 365, 185–205.

Pogson, G.H., Mesa, K.A., 2004. Positive Darwinian selection at the pantophysin (Pan I) locus in marine gadid fishes. Mol. Biol. Evol. 21, 65–75.

Pogson, G.H., Mesa, K.A., Boutilier, R.G., 1995. Genetic population structure and gene flow in the Atlantic cod *Gadus morhua*: a comparison of allozyme and nuclear RFLP loci. Genetics 139, 375–385.

Star, B., Nederbragt, A.J., Jentoft, S., Grimholt, U., Malmstrøm, M., Gregers, T.F., Rounge, T.B., Paulsen, J., Solbakken, M.H., Sharma, A., Wetten, O.F., Lanzén, A., Winer, R., Knight, J., Vogel, J.H., Aken, B., Andersen, O., Lagesen, K., Tooming-Klunderud, A., Edvardsen, R.B., Tina, K.G., Espelund, M., Nepal, C., Previti, C., Karlsen, B.O., Moum, T., Skage, M., Berg, P.R., Gjøen, T., Kuhl, H., Thorsen, J., Malde, K., Reinhardt, R., Du, L., Johansen, S.D., Searle, S., Lien, S., Nilsen, F., Jonassen, I., Omholt, S.W., Stenseth, N.C., Jakobsen, K.S., 2011. The genome sequence of Atlantic cod reveals a unique immune system. Nature 477, 207–210.

Tajima, F., 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123, 585–595.

Tsuji, J., Frith, M.C., Tomii, K., Horton, P., 2012. Mammalian NUMT insertion is non-random. Nucleic Acids Res. 40, 9073–9088.

Ursvik, A., Breines, R., Schou Christiansen, J., Fevolden, S.E., Coucheron, D.H., Johansen, S.D., 2007. A mitogenomic approach to the taxonomy of pollocks: *Theragra chalcogramma* and *T. finnmarchia* represent one single species. BMC Evol. Biol. 7, 86.

Wise, C.A., Sraml, M., Easteal, S., 1998. Departure from neutrality at the mitochondrial NADH dehydrogenase subunit 2 gene in humans, but not in chimpanzees. Genetics 148, 409–421.

Wu, T.D., Watanabe, C.K., 2005. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. Bioinformatics 21, 1859–1875.

Supporting information

Supplementary Table S1
Mitogenome divergence in protein coding genes between NA and NC ecotypes.

Gene	Population	SNP [a]	ns-SNP [b]	Average D_T [c]	Average π [d]	Average θ [e]
ND1	NA	17	2	-1.807	0.00369	0.04629
ND1	NC	17	1	-1.696	0.00315	0.03890
ND2	NA	30	2	-1.540	0.00477	0.03585
ND2	NC	27	1	-1.573	0.00577	0.03722
COI	NA	18	0	-1.773	0.00202	0.03951
COI	NC	13	0	-1.789	0.00167	0.03880
COII	NA	9	1	-1.754	0.00180	0.03752
COII	NC	3	0	-1.748	0.00124	0.03506
ATP8	NA	0	0	-1.800	0.00107	0.03763
ATP8	NC	0	0	-1.748	0.00100	0.03388
ATP6	NA	9	0	-1.783	0.00234	0.04222
ATP6	NC	9	0	-1.758	0.00253	0.03937
COIII	NA	11	1	-1.840	0.00196	0.04305
COIII	NC	7	1	-1.732	0.00169	0.03598
ND3	NA	5	0	-1.922	0.00247	0.05076
ND3	NC	7	0	-1.758	0.00311	0.04162
ND4L	NA	0	0	-1.852	0.00082	0.04159
ND4L	NC	0	0	-1.752	0.00081	0.03368
ND4	NA	37	3	-1.693	0.00432	0.04074
ND4	NC	29	2	-1.666	0.00371	0.03777
ND5	NA	55	8	-1.672	0.00479	0.04073
ND5	NC	38	10	-1.682	0.00426	0.03915
ND6	NA	13	1	-1.688	0.00452	0.04093
ND6	NC	7	0	-1.677	0.00413	0.03788
CytB	NA	32	1	-1.689	0.00462	0.04104
CytB	NC	23	2	-1.672	0.00398	0.03712

[a] Total number of SNPs. [b] Non-synonymous SNPs. [c] D_T; Tajima's D (departures from neutrality) using genome wide scans in sliding windows. [d] π; average number of pairwise nucleotide differences per site. [e] θ; average number of nucleotides segregating per site.

Supplementary Table S2
Non-synonymous mitochondrial SNPs, their frequencies, and corresponding amino acid shift detected in the pooled NA and NC cod samples.

Ref SNP position [a]	SNP variant	Freq. % [b]	AA shift [c]	Gene	Ecotype
G2886	A	2.3	V → I	ND1	NC
A3177	G	4.1	S → G	ND1	NA
A3607	G	2.4	E → G	ND1	NA
T4293	C	2.3	F → L	ND2	NA
A4491	G	2.5	T → A	ND2	NA
T4837	C	2.1	I → T	ND2	NC
G7278	A	2.9	V → I	COII	NA
C9050	T	95.1/97.0	L → F	COIII	NA/NC
T10370	A	3.3/3.1	L → M	ND4	NA/NC
G10808	A	3.3	V → I	ND4	NA
A10833	G	5.7	N → S	ND4	NC
G11597	A	2.7	A → T	ND4	NA
C12024	T	5.5	A → V	ND5	NC
G12074	A	4.4	A → T	ND5	NC
A12108	G	2.5	E → G	ND5	NC
C12266	T	4.4	L → F	ND5	NC
A12548	G	23.6/18.0	N → D	ND5	NA/NC
C12906	T	4.3	A → V	ND5	NA
G13028	A	3.7/19.1	A → T	ND5	NA/NC
C13113	T	2.5	A → V	ND5	NA
G13230	A	2.4	G → D	ND5	NC
C13247	T	2.5/8.0	P → S	ND5	NA/NC
C13274	T	3.3	P → S	ND5	NA
G13382	A	9.5	A → T	ND5	NA
A13415	G	4.1	I → V	ND5	NC
G13640	A	4.3/2.6	A → T	ND5	NA/NC
A14144	C	23.0	L → V	ND6	NA

G14390	A	3.0	S → N	CytB	NC
G14719	A	100/100	A → T	CytB	NA/NC

[a] Position in the NC3 reference mitogenome (HG514359).

[b] Frequency (%) of SNP in NA and/or NC cod ecotypes.

[c] Amino acid shift using one-letter abbreviations. on-synonymous SNPs.

Supplementary Table S3
SNP transversions and their frequencies in whole mitogenome (except control region) detected in the pooled NA and NC cod samples.

Ref SNP position [a]	SNP variant	Freq. % [b]	Gene	Ecotype
T3113	G/C	83.5/87.9	ND1	NA/NC
C4112	G	97.3/96.6	ND2	NA/NC
G4499	A/C	82.9	ND2	NC
G4712	C	95.7	ND2	NA
A4748	G/C	78.9	ND2	NC
G4871	T	95.9	ND2	NA
T8570	A	95.9	ATP6	NA
G8953	C	97.8	COIII	NA
T9097	G	96.4	COIII	NC
G9268	C	95.1/89.6	COIII	NA/NC
A9832	C/G	83.9	ND3	NC
T10370	A	96.7/96.9	ND4	NA/NC
T10735	A	89.7	ND4	NA
G11062	C	86.6/90.6	ND4	NA/NC
T12472	G	97.2	ND5	NC
T12865	G	86.8/79.7	ND5	NA/NC
T13072	A	88.2/90.0	ND5	NA/NC
A13363	C	98.0	ND5	NA
C13827	G	74.9/80.4	ND6	NA/NC
C13947	A	96.5	ND6	NA
A14144	C	97.7	ND6	NA
T14691	G	79.5/85.7	CytB	NA/NC
G14692	T	96.6/96.8	CytB	NA/NC
C15036	A	96.5	CytB	NA
A15222	C	96.8	CytB	NA
C15426	A	97.7	CytB	NA

[a] Position in the NC3 reference mitogenome (HG514359).

[b] Frequency (%) of SNP in NA and/or NC cod ecotypes.

Supplementary Table S4
SNP substitutions and their frequencies in structural RNA genes and TP-spacer detected in the pooled NA and NC cod samples.

Ref SNP position [a]	SNP variant	Freq. % [b]	Gene	Ecotype
A47	G	2.8	Phe	NC
T106	C	6.2	SSU	NC
G411	A	5.9	SSU	NC
T674	C	3.4	SSU	NC
C878	T	9.2	SSU	NC
A1027	G	2.3	Val	NA
G1072	A	2.2	Val	NC
A1132	G	2.2	LSU	NA
A1204	G	3.0	LSU	NC
A1332	G	2.1	LSU	NC
A1505	G	4.8	LSU	NA
G1507	A	2.2	LSU	NA
C1762	T	2.1	LSU	NA
C1821	T	11.2	LSU	NC
A2277	G	2.5/3.5	LSU	NA/NC
T2331	C	2.5	LSU	NA
A2346	G	2.5	LSU	NC
A2660	G	2.7	LSU	NA
A2746	G	2.5	LSU	NC
G4000	A	98.0/99.7	Met	NA/NC
A5115	G	2.4	Trp	NC
A5156	G	3.3	Ala	NC
A5167	G	7.6	Ala	NA
A5272	G	4.5/6.7	Asn	NA/NC
G9597	A	3.3	Gly	NA
G9967	A	5.2/4.5	Arg	NA/NC
G9978	A	2.1	Arg	NC
G10027	A	4.0	Arg	NA

C11719	G	2.3	His	NC
A11754	G	2.3	His	NA
A14300	G	2.8	Glu	NC
A15574	G	17.3/16.6	TP-spacer	NA/NC
G15575	A	9.3/5.5	TP-spacer	NA/NC
A15577	C	2.4	TP-spacer	NC
A15621	G	3.8	TP-spacer	NA
A15622	G	6.2/6.3	TP-spacer	NA/NC

[a] Position in the NC3 reference mitogenome (HG514359).

[b] Frequency (%) of SNP in NA and/or NC cod ecotypes.

Paper III

Jørgensen et al. BMC Evolutionary Biology 2014, 14:182
http://www.biomedcentral.com/1471-2148/14/182

BMC Evolutionary Biology

RESEARCH ARTICLE Open Access

An evolutionary preserved intergenic spacer in gadiform mitogenomes generates a long noncoding RNA

Tor Erik Jørgensen[1], Ingrid Bakke[2], Anita Ursvik[3], Morten Andreassen[3], Truls Moum[1] and Steinar D Johansen[1,3*]

Abstract

Background: Vertebrate mitogenomes are economically organized and usually lack intergenic sequences other than the control region. Intergenic spacers located between the $tRNA^{Thr}$ and $tRNA^{Pro}$ genes ("T-P spacers") have been observed in several taxa, including gadiform species, but information about their biological roles and putative functions is still lacking.

Results: Sequence characterization of the complete European hake *Merluccius merluccius* mitogenome identified a complex T-P spacer ranging in size from 223–532 bp. Further analyses of 32 gadiform species, representing 8 families and 28 genera, revealed the evolutionary preserved presence of T-P spacers across all taxa. Molecular complexity of the T-P spacers was found to be coherent with the phylogenetic relationships, supporting a common ancestral origin and gain of function during codfish evolution. Intraspecific variation of T-P spacer sequences was assessed in 225 Atlantic cod specimens and revealed 26 haplotypes. Pyrosequencing data representing the mito-transcriptome poly (A) fraction in Atlantic cod identified an abundant H-strand specific long noncoding RNA of about 375 nt. The T-P spacer corresponded to the 5′ part of this transcript, which terminated within the control region in a tail-to-tail configuration with the L-strand specific transcript (the 7S RNA).

Conclusions: The T-P spacer is inferred to be evolutionary preserved in gadiform mitogenomes due to gain of function through a long noncoding RNA. We suggest that the T-P spacer adds stability to the H-strand specific long noncoding RNA by forming stable hairpin structures and additional protein binding sites.

Keywords: Atlantic cod, European hake, Heteroplasmy, lncRNA, Mitogenome

Background

One characteristic feature of vertebrate mitochondrial genomes (mitogenomes) is the economical gene organization. Vertebrates usually lack mitochondrial intergenic sequences of any appreciable size other than the approximately 1-kb control region (CR) and the 30-bp origin of light (L) strand replication (OriL). The CR harbors the origin of heavy (H) strand replication (OriH) and initiation sites for H-strand and L-strand transcription, as well as the displacement loop (D-loop), but does not contain any canonical protein coding or RNA coding genes [1].

Among the 2600 completely sequenced vertebrate mitogenomes, a great majority shares an identical arrangement of 37 genes. However, alternative gene orders due to minor rearrangements and sequence duplications have been noted [2,3]. A common feature of most rearrangements is DNA duplications that involve tRNA genes and intergenic sequences. Mitogenomes of several snake species contain two CR copies [4,5], and the four gene orders known in birds differ in CR copy number and CR location [6-8]. Related rearrangements have also been noted in bony fishes, and include both CR, OriL and tRNA gene duplications [9-12].

A mitogenome intergenic spacer between the $tRNA^{Thr}$ and $tRNA^{Pro}$ genes (the T-P spacer), located close to the CR, has been reported in a few species representing distantly related vertebrate taxa. Short T-P spacers have

* Correspondence: Steinar.Johansen@uit.no
[1]Marine Genomics group, Faculty of Biosciences and Aquaculture, University of Nordland, Bodø, Norway
[3]Department of Medical Biology, Faculty of Health Sciences, UiT – Norwegian Arctic University, MH-building Breivika, N-9037 Tromsø, Norway
Full list of author information is available at the end of the article

been found in ostrich birds [13], in some toad linages including *Xenopus* [14,15], and in the mole lizard [16]. Larger T-P spacers have been reported in various species of the salamander family Ambystomatidae [17,18]. These complex spacers vary in size from 240 bp to 680 bp and tandem repeat sequence motifs appear common, but no functional role has yet been assigned. A T-P spacer in a fish species was first discovered in the Atlantic cod (*Gadus morhua*) mitogenome [19-21], and later noted in the related Walleye pollock (*Theragra chalcogramma*) [22] and six additional gadiform species representing three families (Gadidae, Lotidae and Phycidae) [23]. These T-P spacers vary in size from 25 bp in fourbeard rockling (*Enchelyopus cimbrius*) to 99 bp in haddock (*Melanogrammus aeglefinus*), and contain one or two copies of a conserved 17-bp motif (Box-motif) [23]. T-P spacers demonstrate interspecific sequence variation and have been applied as genetic markers in population studies of both Walleye pollock [22] and Atlantic cod [24]. More recently T-P spacers were reported in two additional gadid species [25]. While Greenland cod (*Gadus ogac*) contains two size variants of the spacer (73 bp and 102 bp), highly similar in sequence to that of Atlantic cod, the Arctic cod (*Arctogadus glacialis*) T-P spacer appears more complex and variable in size due to short heteroplasmic tandem repeat motifs [25].

The vertebrate mitochondrial transcriptome (mito-transcriptome) has been investigated in human cells and tissues [26,27]. The H and L strand polycistronic precursor transcripts are processed into the 22 tRNAs, 2 rRNAs, and 11 mRNAs [28]. Here, the L-strand specific promoter (LSP) generates a transcript that gives rise to the ND6 mRNA and 8 tRNAs. The H-strand specific transcription, however, is more complex since two promoters (HSP_1 and HSP_2) are involved. Whereas the HSP_1-specific transcript is short and highly abundant, and is the main source of mitochondrial rRNAs, the HSP_2-specific transcript is processed into 8 monocistronic and two bicistronic mRNAs [26,28]. This organization appears highly conserved among vertebrates since Atlantic cod and saithe (*Pollachius virens*) mito-transcriptomes are very similar to that of humans [29].

Long noncoding RNAs (lncRNAs), typically longer than 200 nt, are highly abundant in vertebrates where they possess key roles in gene regulation linked to tissue specificity, development and disease [30,31]. The human mitogenome also code for lncRNAs in addition to their 37 canonical genes. The first mitochondrial lncRNA to be discovered was the 7S RNA, transcribed from LSP within the CR [32,33]. Additional mitochondrial lncRNAs have since then been identified in human cells, and include lncND5, lncND6, lncCytB [27], as well as the prognostic LIPCAR lncRNAs antisense to CytB and COII mRNAs [34]. The lncND5 RNA is of particular interest due to its antisense organization to ND5 mRNA, and the fact that ND5 is the only tightly regulated protein gene in vertebrate mitochondria [35,36]. Mitochondrial lncRNAs appear highly conserved among vertebrates, and we recently reported noncoding RNA corresponding to lncND5 in the mito-transcriptomes of Atlantic cod and saithe [29].

Here we report two non-overlapping mitochondrial lncRNAs (lncCR-H and lncCR-L) transcribed from opposite strands within the Atlantic cod CR, both terminating at the termination-associated sequence (TAS). While the 500 nt lncCR-L RNA apparently corresponds to the human 7S RNA [29], the 375 nt lncCR-H RNA has not been reported previously. The 5' end of lncCR-H corresponds to the T-P spacer that apparently adds stability to the noncoding RNA. The T-P spacer was found to be present in all 32 gadiform species investigated (representing 8 families and 28 genera), and feature analysis indicates directional evolution of spacers from a simple organization in early branching gadiform families (e.g. Moridae or Macrouridae) to complex structures including heteroplasmy in Merluciidae and Gadidae.

Results

Complete mitogenome sequence of the European hake

The 17.078 bp mitogenome of European hake (*Merluccius merluccius*) was determined and represents the first complete mitogenome sequence in the family Merlucciidae (Order Gadiformes). The circular mitogenome contains the same set of two ribosomal RNA genes, 13 protein coding genes, and 22 transfer RNA genes as reported in all codfishes to date (Figure 1A), and is arranged according to the general vertebrate organization [1]. The 871 bp CR was found to be similar in size and related in sequence to that of the southern hake *M. australis* (Additional file 1: Figure S1). We noted an unusual 42 bp insertion in the ND6 gene, corresponding to 14 amino acids, in European hake compared to that of other gadiforms (Additional file 2: Figure S2). The insertion is located in a non-transmembrane region known to contain insertions in mammals [37]. Finally, a large and complex intergenic T-P spacer (532 bp) was found between the $tRNA^{Thr}$ and $tRNA^{Pro}$ genes (Figure 1A).

The phylogenetic position of European hake among Gadiformes was assessed from concatenated mitochondrial protein sequences (3814 amino acid positions) derived from complete mitogenome sequences (Additional file 3: Table S1), and a representative neighbor-joining tree is presented in Additional file 4: Figure S3. Eight gadiform families were included in the analysis and European hake (family Merlucciidae) was found to occupy a more basal position compared to Gadidae and Lotidae, but apparenty more recent than Macrouridae, Bregmacerotidae and Moridae. These observations are in general agreement with previous works on gadiform phylogeny [38-40].

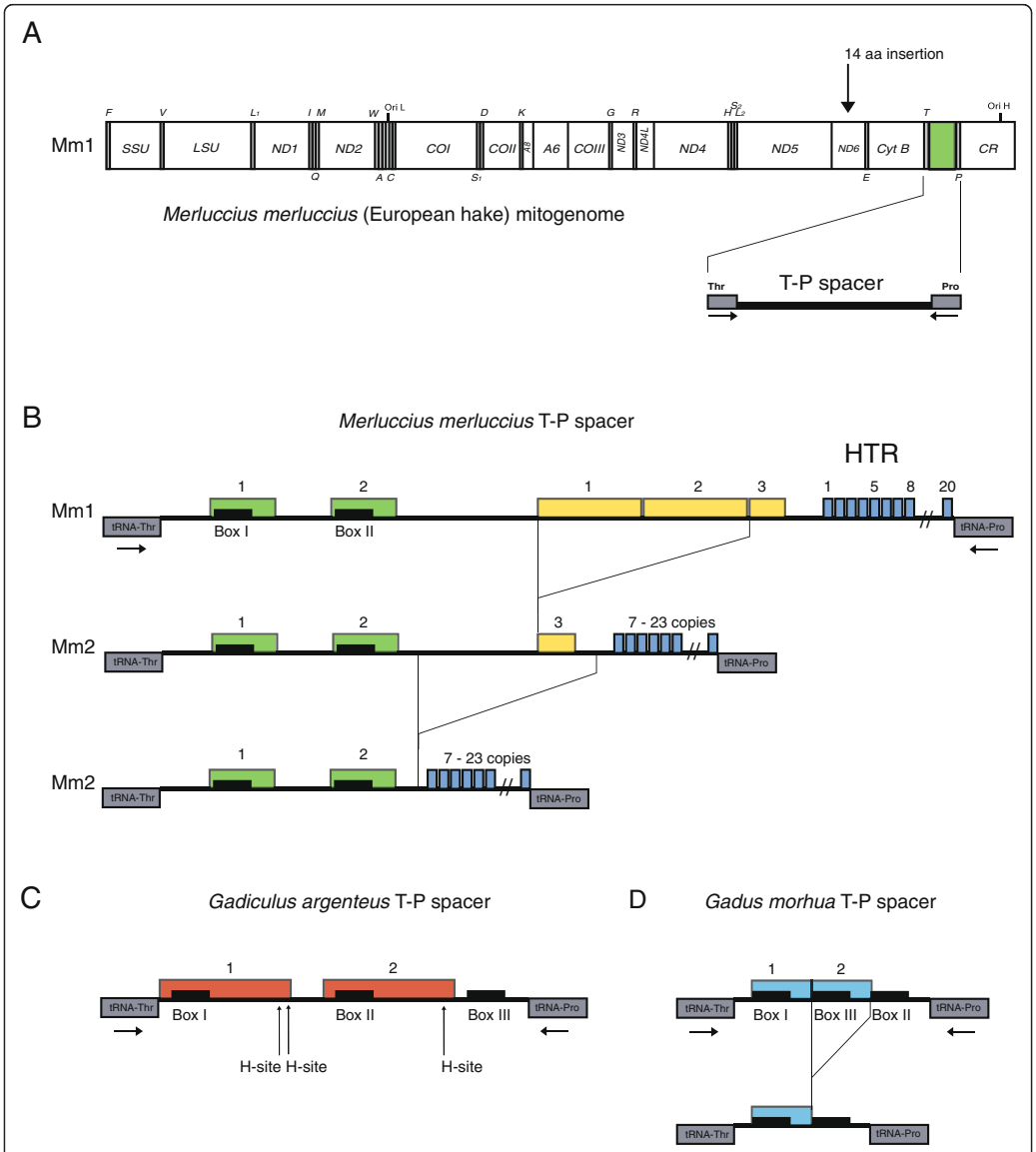


Figure 1 Organization of T-P spacer. A. Gene content and organization of European hake mitochondrial genome (Mm1 specimen) presented as a linear map of the circular mtDNA. All protein genes, except ND6, are encoded by the H-strand. Abbreviations: SSU and LSU, mitochondrial small- and large-subunit ribosomal RNA genes; ND1-6, NADH dehydrogenase subunit 1 to 6; COI-III, cytochrome c oxidase subunit I to III; A6 and A8, ATPase subunit 6 and 8; CytB, cytochrome B; oriH and oriL, origin of H-strand and L-strand replication; CR, control region containing the D-loop; tRNA genes are indicated by the standard one-letter symbols for amino acids. H-strand and L-strand encoded tRNA genes are indicated above and below the diagram, respectively. T-P spacer is indicated below the diagram, and position of the 42-bp insertion in the ND6 gene corresponding to 14 amino acids is indicated by arrow above. **B**. Organization of the European hake T-P spacer in specimens Mm1 and Mm2. Different direct repeat motifs are indicated by green (conserved repeats), yellow (optional repeats), and blue (heteroplasmic tandem repeats, HTR) boxes. Two copies of the Box motif are present. **C**. Organization of the Silvery pout (*G. argenteus*) T-P spacer containing three Box-motifs. The third copy of the direct repeat (red boxes) is truncated. Heteroplasmic sites (H) are indicated. **D**. Organization of the Atlantic cod (*G. morhua*) T-P spacer. About 1% (3/225) of analyzed specimens harbor a direct repeat duplication (blue box).

T-P spacer in European hake contains heteroplasmic tandem repeats

The T-P spacer included in the complete European hake mitogenome sequence (Mm1) was determined to 532 bp. This sequence consists of a 34 bp conserved direct repeat containing the 17-bp Box-motif [23], an optional direct repeat of about 60 bp, and 20 copies of an 8-bp heteroplasmic tandem repeat (HTR) (Figure 1B; Additional file 5: Figure S4A). To investigate individual sequence variations, a T-P spacer region from another specimen (Mm2) was analyzed by PCR amplification, plasmid cloning, and subsequent DNA sequencing. The Mm2 T-P spacer showed extensive length heteroplasmy, varying in size from 258 bp to 430 bp (Figure 1B; Additional file 5: Figure S4B). Size variation was partly caused by the 8-bp HTR motif, which varied from 7 to 23 copies among six sequenced clones. Furthermore, four clones contained an additional 76 bp sequence (Figure 1B). All plasmid clones investigated gave identical sequences, except the HTR copy number and two single nucleotide polymorphic sites (Additional file 5: Figure S4B).

Distribution and variability of T-P spacers among gadiform species

To address the phylogenetic distribution of T-P spacers among gadiform species, we determined the DNA sequences for the corresponding region in 9 additional species (Table 1). We also retrieved available sequences from the NCBI database (Table 1, Additional file 6: Figure S5), adding up to a total of 32 species representing eight families (Gadidae, Lotidae, Ranicipitidae, Merlucciidae, Phycidae, Macrouridae, Moridae and Bregmacerotidae). Several interesting features were noted: (1) all of the gadiform species investigated harbor mitochondrial T-P spacers. (2) Extensive size variation was observed among species, from 7 bp in Anadara (*Bathygadus antrodes*) to 532 bp in European hake. (3) A conserved 17 bp sequence (Box-motif) in 1 to 3 copies was present in 6 of 8 gadiform families. (4) Heteroplasmic features were observed in three species, namely Arctic cod [25], European hake, and silvery pout (*Gadiculus argenteus*), a codfish species belonging to the family Gadidae. In a similar approach to that of the European hake, the silvery pout spacer was PCR amplified, cloned and sequenced. Four independent clones revealed small length variations of homopolymeric stretches and one heteroplasmic SNP site (Additional file 5: Figure S4C). The T-P spacer was found to be 261–263 bp in size, carrying three copies of the Box-motif (Figure 1C).

Intraspecific variation of the Atlantic cod T-P spacer

Intraspecific variation of the T-P spacer was investigated in Atlantic cod. The T-P spacer of Atlantic cod is normally 74 bp in size and contains two copies of the Box-motif (Figure 1D). A total of 225 specimens were included in the analysis, representing all main localities in the North Atlantic Ocean [41]. Among these, 115 sequences were published previously [19,24,42,43] and 110 new sequences were obtained by PCR amplifications and sequencing. We found variable sites at 16 of the 74 positions, which defined 26 T-P spacer haplotypes (Figure 2). The Box-motif II appeared more conserved than motif I, but no heteroplasmic sites were detected. The dominant haplotype (Gm-I) was represented by 53.8% (121 of 225 individuals), including the whole genome sequenced NEAC_001 specimen [44]. These findings are in general agreement with intraspecific variation previously reported in the T-P spacer of 110 individuals of Walleye pollock [23]. Three individuals representing two haplotypes (Gm-XXV and Gm-XXVI) contained a 29 bp insert that includes the 17-bp Box-motif copy (Figure 1D). A similar duplication of Box-motif I was observed in Greenland cod [25].

Analysis of mito-transcriptome data from Atlantic cod liver tissue

We reported previously the mapping of 4698 pyrosequencing reads of mitochondrial transcripts from Atlantic cod liver tissue [29]. Here, all mitochondrial mRNAs as well as the ribosomal RNA transcript were represented by multiple reads each. We also observed a significant and stable L-strand specific lncRNA (lncCR-L) of about 500 nt covering parts of the CR, and apparently similar to that of the 7S RNA in human mitochondria [29]. A closer inspection of the mito-transcriptome data identified a number of reads mapping to the opposite part of the CR including the T-P spacer (Figure 3A), and corresponding to a H-strand specific lncRNA. Interestingly, this approximately 375 nt transcript (lncCR-H) starts exactly 3 nt upstream of the $tRNA^{Thr}$ gene and includes the T-P spacer, proceeds into the mirror $tRNA^{Pro}$ and further into the HTR domain of the CR, and terminates by a non-template polyA tail immediately downstream of the TAS. Based on read map coverage, the relative abundance of lncCR-H was comparable to that of mitochondrial mRNAs (e.g. ND4L mRNA [29]). LncCR-H varies in size due to different numbers of HTRs included in the RNA. We observed 1–4 HTR copies in lncCR-H, with a majority of 3 copies (Figure 3A). A secondary structure diagram of lncCR-H is presented in Figure 3B, indicating structured domains at both the 5' and 3' ends. Interestingly, the Box and TAS motifs are proposed to be part of tetraloop hairpin structures.

Discussion

We report a complex intergenic spacer between the $tRNA^{Thr}$ and $tRNA^{Pro}$ genes in the European hake mitochondrial genome. The T-P spacer contains HTRs, and significant size variation was observed among specimens. Furthermore, we found T-P spacers to be present in

Table 1 Key features of gadiform T-P spacers

Species	Common name	[1]Size (bp)	[2]Box	[3]Reference
Order: Gadiformes				
Family: Gadidae				
Gadus morhua	(Atlantic cod)	74/103	2/3	AM489716; [22,23]
Gadus ogac	(Greenland cod)	73/102	2/3	FJ396453; [24]
Theragra chalcogramma	(Alaska Pollock)	70-72	2	Y17984; [22]
Theragra finnmarchica	(Norwegian Pollock)	72	2	AM489718
Boreogadus saida	(Polar cod)	70	2	Y17985; [22]
Arctogadus glacialis	(Arctic cod)	*291	2	[24]
Melanogrammus aeglefinus	(Haddock)	99	2	Y17986; [22]
Merlangius merlangius	(Whiting)	69	2	DQ020496
Pollachius virens	(Saithe)	59	1	FR751399
Pollachius pollachius	(Pollack)	50	1	FR751400
Trisopterus esmarkii	(Norwegian pout)	77	1	This work
Trisopterus minutes	(Poor cod)	75	1	This work
Micromesistius poutassou	(Blue whiting)	48	1	FR751401; [22]
Gadiculus argenteus	(Silvery pout)	*261-263	3	This work
Family: Lotidae				
Lota lota	(Burbot)	103	1	AP004412
Brosme brosme	(Tusk)	48	1	Y17988; [22]
Molva molva	(Ling)	83	1	This work
Molva dipterygia	(Blue ling)	59	1	This work
Family: Phycidae				
Enchelyopus cimbrius	(Fourbeaed rockling)	25	1	Y17989; [22]
Gaidropsarus argentatus	(Tree-bearded rockling)	48	1	Y17990; [22]
Phycis blennoides	(Greater forkbeard)	28	1	This work
Family: Ranicipitidae				
Raniceps raninus	(Tadpole fish)	39	1	This work
Family: Merlucciidae				
Macruronus novaezelandiae	(Hoki)	59	1	This work
Merluccius merluccius	(European hake)	*223-532	2	FR751402; This work
Family: Macrouridae				
Squalogadus modificatus	(Tadpole whiptail)	47	1	AP008989
Trachyrincus murragi	(Roughnose grenadier)	64	1	AP008990
Bathygadus antrodes	(Anadara)	7	-	AP008988
Caelorinchus kishinouyei	(Mugara grenadier)	12	-	AP002929
Coryphaenoides rupestris	(Roundnose grenadier)	21	-	This work
Ventrifossa garmani	(Sagami grenadier)	72	-	AP008991
Family: Bregmacerotidae				
Bregmaceros nectabanus	(Smallscale codlet)	(+)	-	AP004411
Family: Moridae				
Physiculus japonicus	(Japanese codling)	16	-	AP004409
Laemonema longipes	(Longfin codling)	33	-	AB108839

Table 1 Key features of gadiform T-P spacers *(Continued)*

Order: Lophiiformes				
Family: Lophiidae				
Lophius piscatorius	(Frogfish)	0	-	This work
Lophius americanus	(American angler)	0	-	AP00414

Notes: [1]Sequence and size variants of T-P spacers are given in Additional file 6: Figure S5. T-P spacers of *M. merluccius*, *G. argenteus*, and *A. glacialis* are heteroplasmic (*). The T-P spacer of *B. nectabanus* (+) is involved in gene order rearrangements. [2]Box-motif copy number. [3]Data base accession number and key references.

mitogenomes of all the 32 gadiform species investigated, representing 8 families and 28 genera, and the majority of spacers harbor a conserved 17-bp Box-motif. Analysis of the Atlantic cod liver mito-transcriptome identified a long noncoding RNA (lncCR-H) that contains the T-P spacer within its 5' end.

Mitogenomes of most bony fishes conform to the general vertebrate gene organization, and lack intergenic nucleotides between the $tRNA^{Thr}$ and $tRNA^{Pro}$ genes. In contrast, we found the presence of intergenic spacers at this location to be a consistent feature among gadiform species. Based on key features and the distribution pattern of gadiform T-P spacers (Table 1), as well as the current understanding of gadiform phylogeny [38-40], we propose the following evolutionary scenario for the gadiform T-P spacer (Figure 4). This scenario implies gain of function for this sequence element during gadiform evolution. (1) A spacer sequence was introduced at a basal point of gadiform phylogenesis. (2) Once established, the Box-motif became a preserved sequence feature of descendant taxa after the split from Bathygadus, Caelorinchus, Coryphaenoides, Ventrifossa, Physiculus, Laemonema, and Bregmaceros. The majority of gadiform genera, representing 6 families, contain a single copy of the Box-motif. (3a) Duplication of the Box-motif then occurred within the family Gadidae, leaving six genera (*Gadus*, *Theragra*, *Arctogadus*, *Boregadus*, *Melanogrammus* and *Merlangius*) with two consecutive Box-motifs. (3b) The Box-motif was also independently duplicated in silvery pout (*Gadiculus*), and (3c) within the family Merluciidae, resulting in the complex European hake T-P spacer. (4) A second duplication occurred recently in *Gadus*. Here, a subset of Atlantic cod and Greenland cod specimens contains three copies of the Box-motif due to a duplication of Box-motif I. T-P spacer heteroplasmy was only noted in genera with duplicated Box-motifs (*Merluccius*, *Gadiculus* and *Arctogadus*).

Higher order phylogenetic relationships within Gadiformes are still weakly supported by molecular data [38-40].

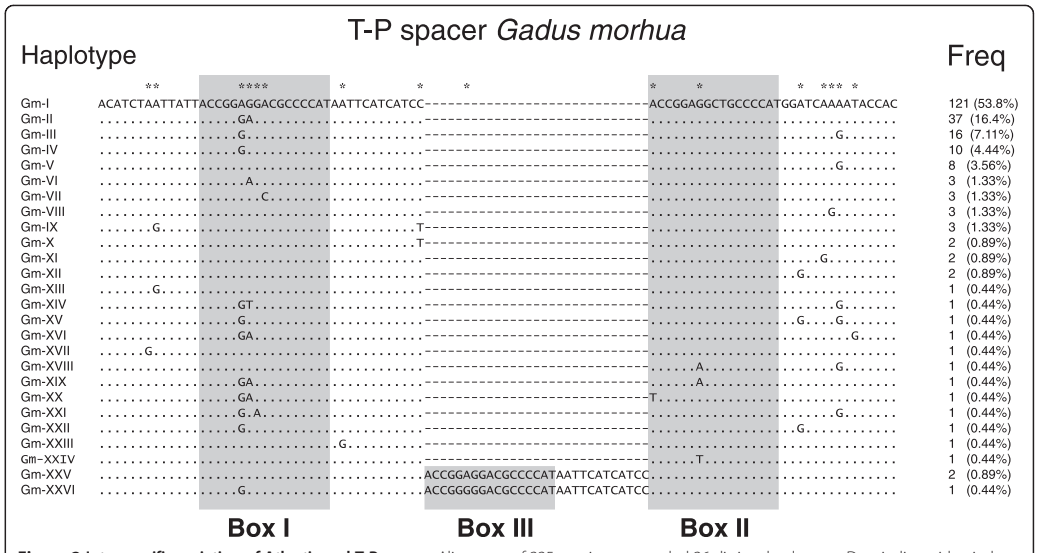


Figure 2 Intraspecific variation of Atlantic cod T-P spacer. Alignment of 225 specimens revealed 26 distinct haplotypes. Dots indicate identical positions to the reference Gm-I haplotype and dashes indicate deletions. Two haplotypes (Gm-XXV and Gm-XXVI) harbor a 29 bp insertion. The conserved 17-bp motif is boxed, and the 16 variable sites are indicated by stars above the Gm-I haplotype sequence. Right; numbers of specimens and frequency (%) belonging to each haplotype.

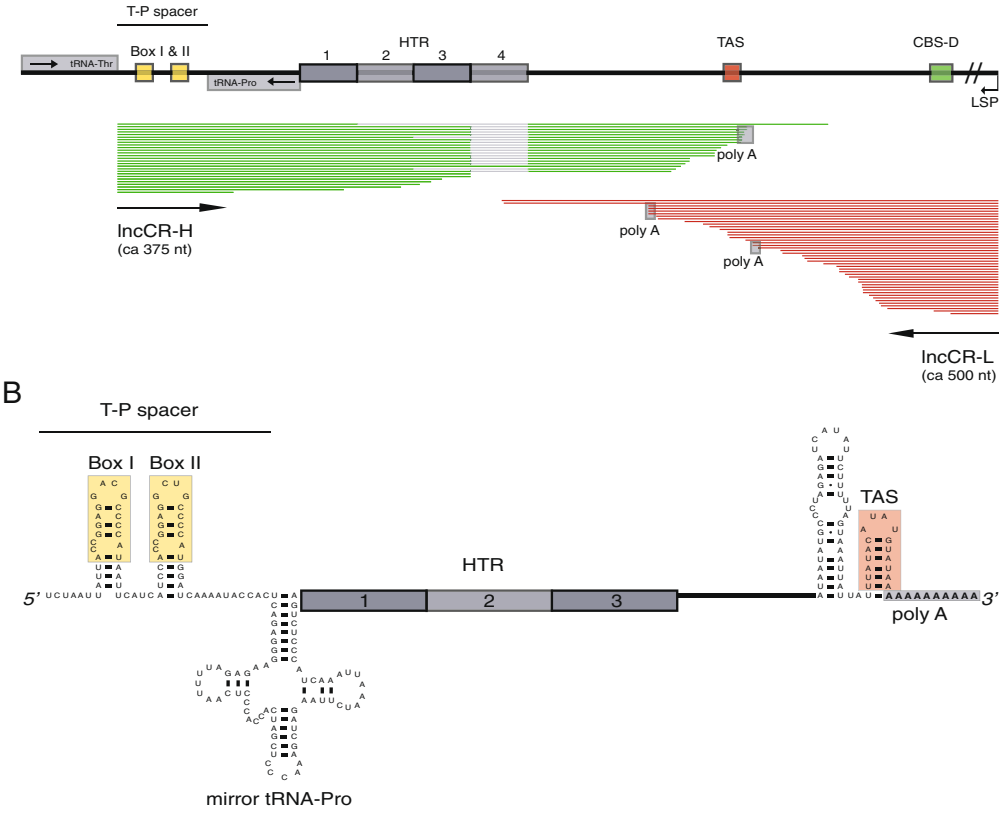


Figure 3 Long noncoding RNAs generated from CR and T-P spacer. A. 454 read coverage of the Atlantic cod liver mitochondrial transcriptome at the T-P spacer and control region. H-strand specific reads (green) start exactly 3 nt downstream the first position of the T-P spacer, and represent lncCR-H. L-strand specific reads (red) represent lncCR-L. Grey line in lncCR-H indicates missing HTR motif due to HTR copy number variation. **B**. Secondary structure diagram of lncCR-H in Atlantic cod. The T-P spacer is located at the 5' end of the RNA, and Box-motifs are indicated (yellow boxes). The mirror $tRNA^{Pro}$ structure and TAS hairpin structure (red box) are located upstream and downstream, respectively, of the heteroplasmic tandem repeat (HTR) motifs.

Other branching patterns than those depicted here would imply instances of secondary loss of the Box-motif, which can of course not be precluded. However, the Box-motif apparently represents a selective feature, of which the exact functional role is still unknown. It is interesting to note that complex T-P spacers in salamanders [17,18] contains strikingly similar Box-motifs to those reported in codfishes (Additional file 7: Table S2), suggesting that an evolutionary conserved mitochondrial factor may interact with the motif.

Analysis of the Atlantic cod liver mito-transcriptome identified a relatively abundant lncRNA (lncCR-H) of about 375 nt, that contains the T-P spacer sequence within its 5' end. The first nucleotide of lncCR-H corresponds to 3 nt downstream the $tRNA^{Thr}$ gene, and the 5' end of the lncRNA was probably generated after RNase Z processing of the tRNA precursor [45]. The 3' end of lncCR-H appears heterogenous and polyadenylated, and corresponds to sequences in proximity to the TAS box element within the CR. LncCR-H harbours secondary structure elements, including a mirror $tRNA^{Pro}$ and tetraloop helices, probably adding stability to the RNA. Interestingly, recent RNA-Seq experiments in breast cell lines identified an expressed human homolog to lncCR-H (our unpublished results), and HeLa cell RNA profiles supported the existence of this mitochondrial lncRNA (see Figure 1B in [27]). Thus,

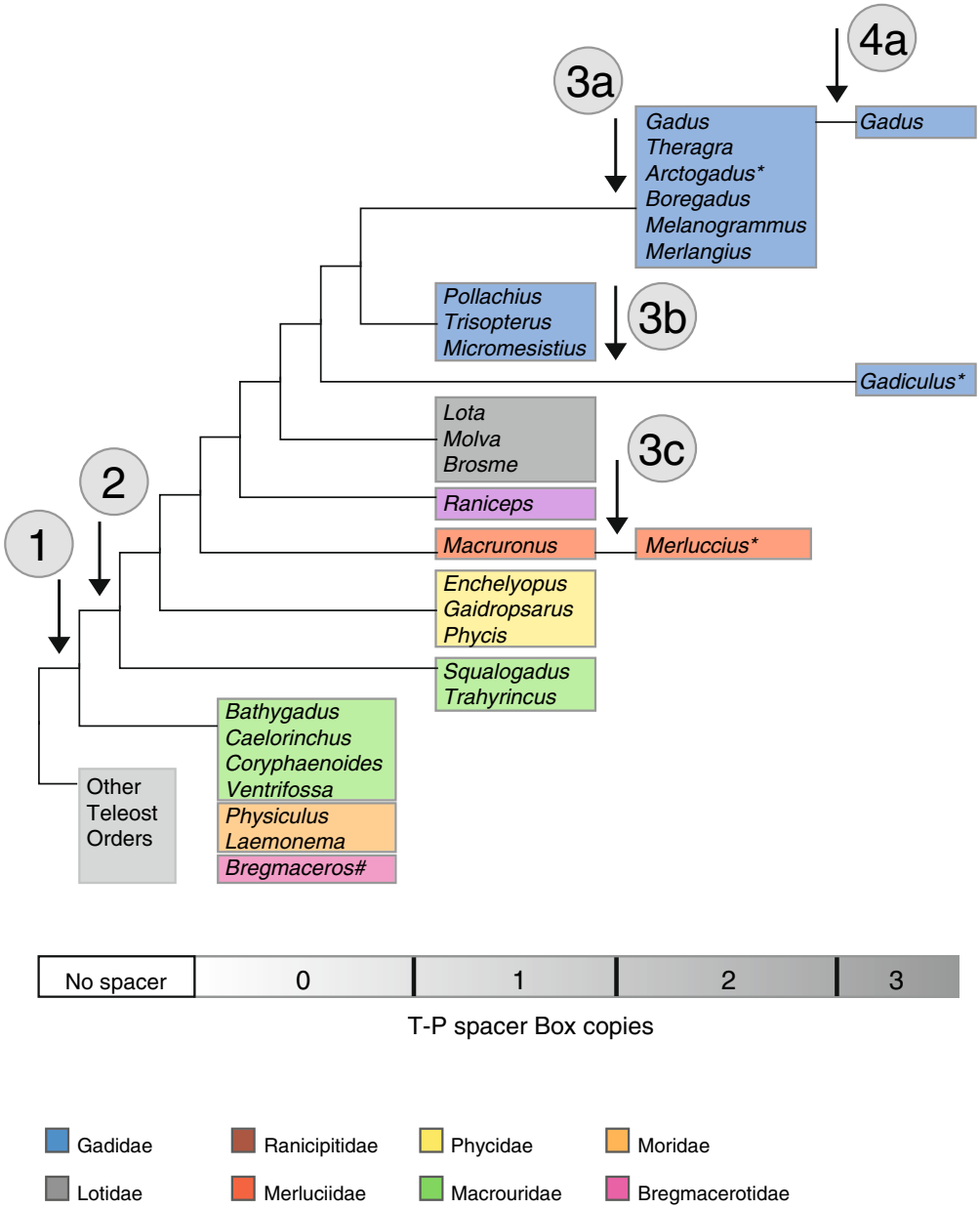


Figure 4 Evolutionary history of gadiform T-P spacer. Key events in the evolutionary history of the T-P spacer are indicated by numbers 1–4 and explained in the main text. In short: 1, first occurrence of spacer; 2, first occurrence of Box-motif; 3abc, duplication of Box-motif; 4, second duplication of Box motif. Genus containing T-P spacer with heteroplasmic feature is indicated by star (*). The *Bregmaceros* spacer is involved in gene order rearrangements (#). The 32 gadiform species represent 20 genera and 8 families (color-coded). Pattern of relationship between gadiform families is extrapolated from [38–40].

lncCR-H appears conserved at a wide taxonomic range, which implies a functional role in vertebrate mitochondria.

Mitochondrial lncRNAs corresponding to antisense transcripts of mRNAs have been characterized in human cells [26,27,34] as well as in Atlantic cod [29]. In that study we identified a second mitochondrial lncRNA within CR, named lncCR-L. The L-strand specific lncCR-L, which corresponds to the human 7S RNA [32,33], is apparently initiated at LSP and terminated at TAS. Our finding implicates that TAS is a transcription termination site both for H-strand and L-strand specific lncRNAs. TAS in Atlantic cod consists of a perfect inverted repeat motif (*UUAUACAU*AUGUAUAA) and thus forms identical tetraloop hairpin structures at the H- and L-strand transcripts. TAS-binding proteins have been reported in human, rat and bovine mitochondria [46,47]. How lncCR-H and lncCR-L are terminated and regulated in Atlantic cod, as well as their biological functions, are currently unknown. However, we speculate that the biological role of the T-P spacer is to add stability to an evolutionary conserved lncRNA where the Box-motif forms an RNA structure signal to be recognized by mitochondrial proteins. lncCR-H may have a role in antisense regulation by premature termination of L-strand transcription at TAS, resulting in a coordination of H and L transcript levels. Anyway, these possibilities need further investigations in different cell and tissue types, and in different species.

Conclusions

Two long noncoding RNA candidates from the mitochondrial control region were described in gadiform fishes. These RNAs were transcribed from different DNA strands, and found to terminate at the TAS region in a tail-to-tail configuration. Whereas the L-strand specific lncRNA (lncCR-L) corresponded to the previously described mitochondrial 7S RNA, lncCR-H was found to be unique by containing a 5' domain of an intergenic spacer. Analyses of 32 gadiform species revealed that a T-P spacer was present in all investigated species, representing 8 families and 28 genera. Some species contained complex spacer sequences due to conserved motifs, duplications, and heteroplasmy. We propose that the T-P spacer adds stability to lncCR-H by forming stable hairpin structures and additional protein binding sites, but a biological role of lncCR-H is currently unknown.

Methods

DNA extraction and PCR amplification

DNA was extracted from muscle tissue samples as previously described [11,38,42,48]. DNA for complete mitogenome sequencing of European hake was extracted from frozen muscle tissue by using the mtDNA Extractor CT Kit from Wako [42]. The PCR and sequencing primers (Additional file 8: Table S2) applied in the European hake sequencing were designed from the published Atlantic cod mitogenome sequence [21] and other available gadiform sequences (Additional file 3: Table S1). These primers were used to amplify the mitogenomes in 1 to 4 kb fragments essentially as described by [11]. The primer pair L15498/H15666 was used for amplification of the T-P spacer from most species. For European hake, a specific primer set (L15498/H15642) was designed for T-P spacer amplifications. The PCR products were treated with exonuclease and shrimp alkaline phosphatase (USB) prior to plasmid cloning and sequencing. The amplified fragments were subcloned into the *Sma*I-site of pUC18 vector using the Sure-Clone Ligation kit (Pharmacia Biotech).

DNA sequencing

Sanger sequencing of the European hake mitogenome and Atlantic cod T-P spacers was performed directly on PCR products as previously described [11] using the BigDye kit (Applied Biosystems). Sequencing primers are found in Additional file 8: Table S2. Manual sequencing reactions were performed on most T-P spacers using the Thermo Sequencing Terminator Cycle Sequencing kit (Amersham) and [α-^{33}P] ddNTP (Amersham) as the label. Sequencing reactions were run on 7 M urea/6% polyacrylamide gels after denaturation at 85°C for 2 minutes. Roche 454 pyrosequencing of the Atlantic cod transcriptome was performed as a service given by Eurofins MWG Operon (Germany). A poly(A) enriched normalized liver-specific cDNA library (based on total cellular RNA) was generated by random hexamer first strand synthesis. About 1.2 million reads were obtained from pooling equal amounts of total RNA from 10 Atlantic cod specimens [29].

Data analyses

A sequence alignment representing complete mitogenomes of 20 gadiform species and one out-group was generated using T-Coffee v/9 [49] with manual refinements. The alignment was build from the complete set of protein codons (except stop codons) creating a concatenated sequence of 11442 nt positions (3814 codons) corresponding to the 13 protein genes. The tree-building methods of neighbor-joining (NJ), maximum parsimony (MP) and maximum likelihood (ML) in MEGA version 5 [50] were used to reconstruct molecular phylogenies. NJ trees were build using Jones-Taylor-Thornton model. MP trees were reconstructed using the Close-Neighbor-Interchange search option. ML trees were built from best-fit models of protein evolution generated by MEGA 5 and ProtTest 3 [51]. The topologies of the ME, MP and ML trees were evaluated by bootstrap analyses (2000 replications). Analyses of mito-transcriptome 454 pyrosequencing reads were as previously described [29].

Additional files

Additional file 1: Figure S1. DNA sequence alignment of hake mitochondrial control region. Dots indicate identical positions to the European hake reference sequence and dashes indicate deletions. Conserved sequence elements (red letters) recognized among gadiform mitochondrial CRs [21] are indicated. TAS, termination-associated sequence; Py-RUN, pyrimidine-rich segment; CSB 2, 3, and D, conserved sequence blocks. European hake – *Merluccius merluccius* (Mmer), FR751402; Southern hake – *M. australis* (Maus), FJ423612.

Additional file 2: Figure S2. Amino acid sequence alignment of gadiform ND6 protein. The 14 amino acid insertion between trans-membrane domains (TMD) IV and V is indicated by red letters. The TMD annotations are according to [37].

Additional file 3: Table S1. Complete gadiform mitogenomes.

Additional file 4: Figure S3. Gadiform phylogeny based on mitogenome-derived amino acid sequences. Neighbor joining (NJ) phylogenetic tree based on the alignment of 13 concatenated proteins corresponding to 3814 amino acids. Bootstrap values (%) from 2000 replicates, all over 40%, are shown at branches. The values are from NJ, maximum parsimony (MP) and maximum likelihood (ML) analyses. Red filled circles indicate highly significant branch points of bootstrap values of 100% in the NJ, MP and ML tree construction methods. Different families are color-coded. References to the complete mitogenome sequences are found in Additional file 3: Table S1. Note that Macrouridae appears paraphyletic.

Additional file 5: Figure S4. European hake and Silvery pout T-P spacer sequence variants. A) DNA sequence of the European hake T-P spacer in specimen Mm1 (FR751402). Different direct repreat motifs are indicated. DR-a (green), conserved direct repeat containing the 17-bp Box-motif sequence; DR-b (yellow), optional direct repeat, including a truncated (*) copy; HTR (blue), heteroplasmic tandem repeat. B) Long heteroplasmic variant in the Mm2 specimen. Two single nucleotide positions (red) found to be heteroplasmic are indicated. C) Short heteroplasmic variant in the Mm2 specimen. D) Silvery pout T-P spacer including heteroplasmic direct repeats (DR) boxed in red.

Additional file 6: Figure S5. Gadiform T-P spacer sequences. Box-motif is underlined. The different gadiform families are shown in A to G.

Additional file 7: Table S2. Box-motif sequence compilation.

Additional file 8: Table S3. PCR and sequencing primer specifications.

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
TEJ and IB participated in DNA sequencing and design of the study, performed analysis, and contributed in the discussions of the results. AU and MA participated in mitogenome sequencing and transcriptome analysis, respectively. TM participated in the design of this study and in discussions of results. SDJ directed the research, participating in the analysis, and wrote the manuscript in collaboration with all the authors. All authors read and approved the final manuscript.

Acknowledgments
We thank Willy Richardsen, Kees O. Ekeli, Odd M. Smedstad, Wim Vader, Stefan Hugel, Jon Einar Hansen, Bjørn Steinar Sætre and Åse Emblem for kindly providing fish samples. We also thank Ragna Breines, Dag H. Coucheron, Marianne Nymark, Bård O. Karlsen, Åse Emblem and Jarle T. Nordeide for discussions. This work was supported by grants to SDJ from the Norwegian Research Council and the University of Tromsø.

Author details
[1]Marine Genomics group, Faculty of Biosciences and Aquaculture, University of Nordland, Bodø, Norway. [2]Department of Biotechnology, Norwegian University of Science and Technology, Trondheim, Norway. [3]Department of Medical Biology, Faculty of Health Sciences, UiT – Norwegian Arctic University, MH-building Breivika, N-9037 Tromsø, Norway.

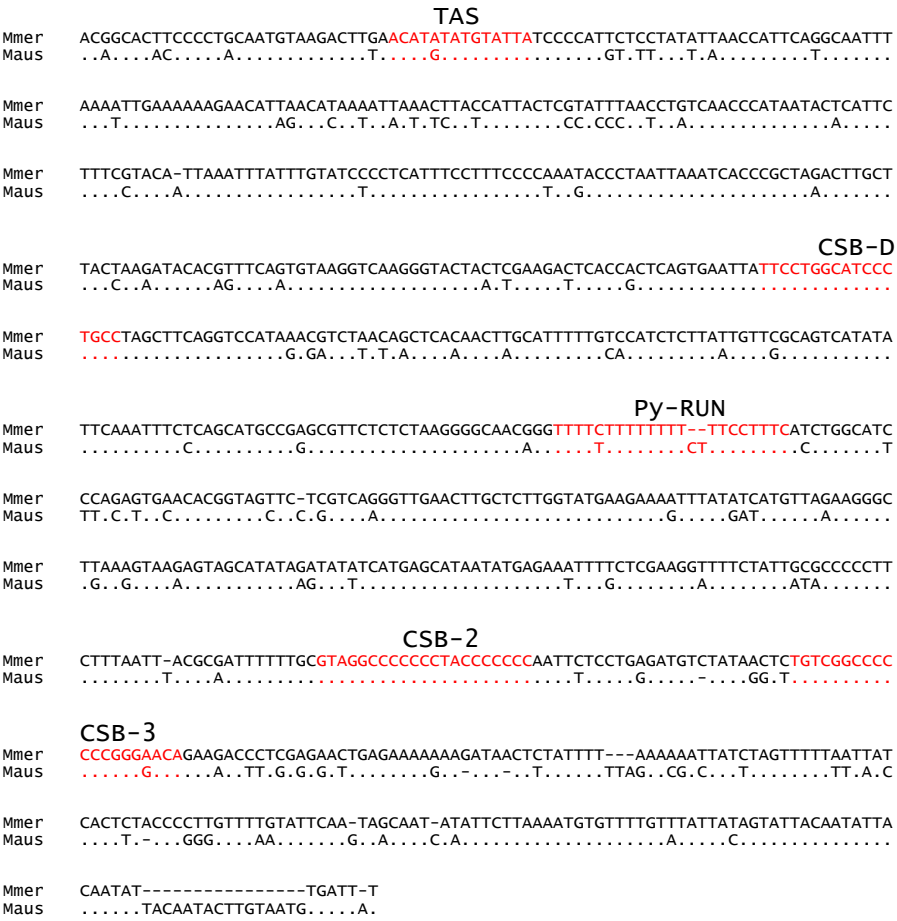
Received: 3 May 2014 Accepted: 5 August 2014
Published: 22 August 2014

References
1. Wallace DC: **Why do we still have a maternally inherited mitochondrial DNA? Insights from evolutionary medicine.** *Annu Rev Biochem* 2007, **76:**781–821.
2. Brown C, Dowling TE, Brown WM: **Evolution of animal mitochondrial DNA: relevance for population biology and systematics.** *Annu Rev Ecol Syst* 1987, **18:**269–292.
3. San Mauro D, Gower DJ, Zardoya R, Wilkinson M: **A hotspot of gene order rearrangement by tandem duplication and random loss in the vertebrate mitochondrial genome.** *Mol Biol Evol* 2006, **23:**227–234.
4. Kumazawa Y, Ota H, Nishida M, Ozawa T: **Gene rearrangements in snake mitochondrial genomes: highly concerted evolution of control-region-like sequences duplicated and inserted into a tRNA gene cluster.** *Mol Biol Evol* 1996, **13:**1242–1254.
5. Kumazawa Y, Ota H, Nishida M, Ozawa T: **The complete nucleotide sequence of a snake (*Dinodon semicarinatus*) mitochondrial genome with two identical control regions.** *Genetics* 1998, **150:**313–329.
6. Mindell DP, Sorenson MD, Dimcheff DE: **Multiple independent origins of mitochondrial gene order in birds.** *Proc Natl Acad Sci U S A* 1998, **95:**10693–10697.
7. Eberhard JR, Wright TF, Bermingham E: **Duplication and concerted evolution of the mitochondrial control region in the parrot genus *Amazona*.** *Mol Biol Evol* 2001, **18:**1330–1342.
8. Abbott CL, Double MC, Trueman JW, Robinson A, Cockburn A: **An unusual source of apparent mitochondrial heteroplasmy: duplicate mitochondrial control regions in *Thalassarche albatrosses*.** *Mol Ecol* 2005, **14:**3605–3613.
9. Inoue JG, Miya M, Tsukamoto K, Nishida M: **Complete mitochondrial DNA sequence of *Conger myriaster* (Teleostei: Anguilliformes): novel gene order for vertebrate mitochondrial genomes and the phylogenetic implications for Auguilliform families.** *J Mol Evol* 2001, **52:**311–320.
10. Satoh TP, Miya M, Endo H, Nishida M: **Round and pointed-head grenadier fishes (Actinopterygii: Gadiformes) represent a single sister group: evidence from complete mitochondrial genome sequences.** *Mol Phylogen Evol* 2006, **40:**129–138.
11. Breines R, Ursvik A, Nymark M, Johansen SD, Coucheron DH: **Complete mitochondrial genome sequences of the Arctic Ocean codfishes *Arctogadus glacialis* and *Boreogadus saida* reveal oriL and tRNA gene duplications.** *Polar Biol* 2008, **31:**1245–1252.
12. Poulsen JY, Byrkjedal I, Willassen E, Rees D, Takeshima H, Satoh TP, Shinohara G, Nishida M, Miya M: **Mitogenomic sequences and evidence from unique gene rearrangements corroborate evolutionary relationships of myctophiformes (Neoteleostei).** *BMC Evol Biol* 2013, **13:**111.
13. Härlid A, Janke A, Arnason U: **The mtDNA sequence of the ostrich and the divergence between Paleognathous and Neognathous birds.** *Mol Biol Evol* 1997, **14:**754–761.
14. Roe BA, Ma DP, Wilson RK, Wong JF: **The complete nucleotide sequence of the *Xenopus laevis* mitochondrial genome.** *J Biol Chem* 1985, **260:**9759–9774.
15. Lloyd RE, Foster PG, Guille M, Littlewood DTJ: **Next generation sequencing and comparative analyses of *Xenopus* mitogenomes.** *BMC Genomics* 2012, **13:**496.
16. Macay JR, Schulte JA, Larson A, Papenfuss TJ: **Tandem duplications via light-strand synthesis may provide a precursor for mitochondrial genomic rearrangement.** *Mol Biol Evol* 1998, **15:**71–75.
17. McKnight ML, Shaffer HB: **Large, rapidly evolving intergenic spacers in the mitochondrial DNA of the salamander family Ambystomatidae (Amphibia: Caudata).** *Mol Biol Evol* 1997, **14:**1167–1176.
18. Samuels AK, Weisrock DW, Smith JJ, France KJ, Walker JA, Putta S, Voss SR: **Transcriptional and phylogenetic analysis of five complete ambystomatid salamander mitochondrial genomes.** *Gene* 2005, **349:**43–53.
19. Johansen S, Guddal PH, Johansen T: **Organization of the mitochondrial genome of Atlantic cod, *Gadus morhua*.** *Nucleic Acids Res* 1990, **18:**411–419.
20. Johansen S, Johansen T: **Sequence analysis of 12 structural genes and a novel non-coding region from mitochondrial DNA of Atlantic cod, *Gadus morhua*.** *Biochem Biophys Acta* 1994, **1218:**213–217.
21. Johansen S, Bakke I: **The complete mitochondrial DNA sequence of Atlantic cod (*Gadus morhua*): relevance to taxonomic studies among codfishes.** *Mol Mar Biol Biotechnol* 1996, **5:**203–214.

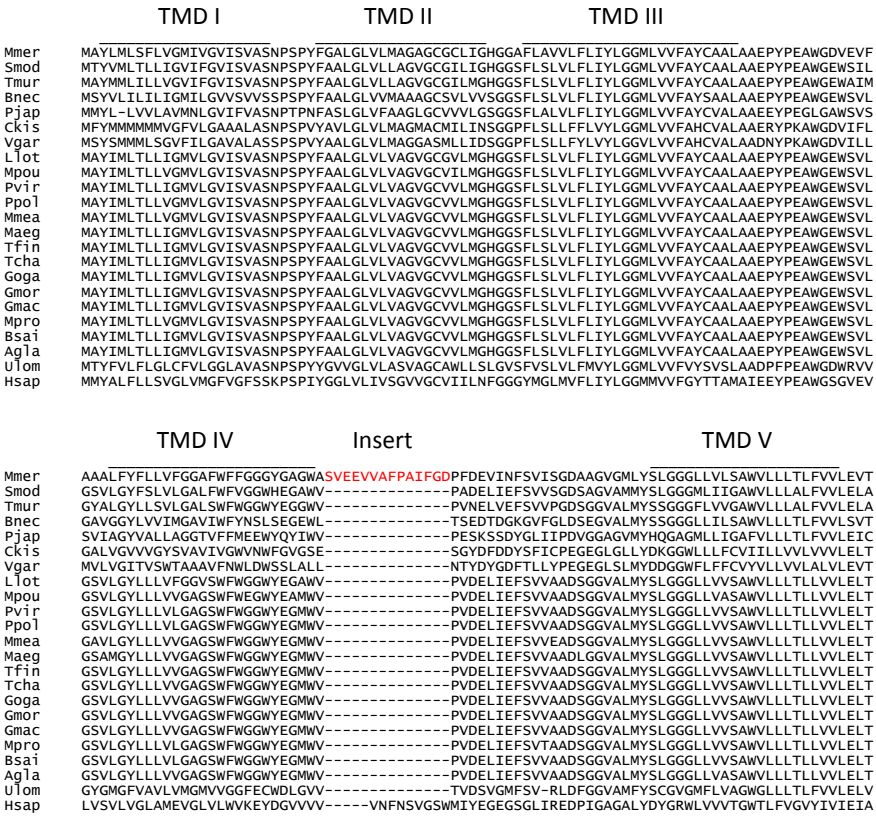
22. Shields GF, Gust JR: **Lack of geographic structure in mitochondrial DNA sequences of Bering Sea walleye Pollock, *Theragra chalcogramma*.** *Mol Mar Biol Biotechnol* 1995, **4:**69–82.
23. Bakke I, Shields GF, Johansen S: **Sequence characterization of a unique intergenic spacer in gadiformes mitochondrial DNA.** *Mar Biotechnol* 1999, **1:**411–415.
24. Sigurgíslason H, Árnason E: **Extent of mitochondrial DNA sequence variation in Atlantic cod from the Faroe Islands: a resolution of gene genealogy.** *Heredity* 2003, **91:**557–564.
25. Pálsson S, Paulsen J, Árnason E: **Rapid evolution of the intergenic T-P spacer in the mtDNA of Arctic cod *Arctogadus glacialis*.** *Mar Biotechnol* 2008, **10:**270–277.
26. Mercer TR, Neph S, Dinger ME, Crawford J, Smith MA, Shearwood AMJ, Haugen E, Bracken CP, Rackham O, Stamatoyannopoulos JA, Filipovska A, Mattick JS: **The human mitochondrial transcriptome.** *Cell* 2011, **146:**645–658.
27. Rackham O, Shearwood AMJ, Mercer TR, Davies SMK, Mattick JS, Filipovska A: **Long noncoding RNAs are generated from the mitochondrial genome and regulated by nuclear-encoded proteins.** *RNA* 2011, **17:**2085–2093.
28. Temperley RJ, Wydro M, Lightowlers RN, Chrzanowska-Lightowlers ZM: **Human mitochondrial mRNAs – like members of all families, similar but different.** *Biochim Biophys Acta* 2010, **1797:**1081–1085.
29. Coucheron DH, Nymark M, Breines R, Karlsen BO, Andreassen M, Jørgensen TE, Moum T, Johansen SD: **Characterization of mitochondrial mRNAs in codfish reveals unique features compared to mammals.** *Curr Genet* 2011, **57:**213–222.
30. Fatica A, Bozzoni I: **Long non-coding RNAs: new players in cell differentiation and development.** *Nature Rev Genet* 2014, **15:**7–21.
31. Necsulea A, Soumillon M, Warnefors M, Liechti A, Daish T, Zeller U, Baker JC, Grutzner F, Kaessmann H: **The evolution of lncRNA repertoires and expression patterns in tetrapods.** *Nature* 2014, **505:**635–640.
32. Ojala D, Crews S, Montoya J, Gelfand R, Attardi G: **A small polyadenylated RNA (7S RNA), containing a putative ribosome attachment site, maps near the origin of human mitochondrial DNA replication.** *J Mol Biol* 1981, **150:**303–314.
33. Chang DD, Clayton DA: **Precise identification of individual promoter for transcription of each strand of human mitochondrial DNA.** *Cell* 1984, **36:**635–643.
34. Kumarswamy R, Bauters C, Volkmann I, Maury F, Fetisch J, Holzmann A, Lemesle G, de Groote P, Pinet F, Thum T: **Circulating long noncoding RNA, LIPCAR, predicts survival in patients with heart failure.** *Circ Res* 2014, **114:**1569–1575.
35. Bai Y, Shakeley RM, Attardi G: **Tight control of respiration by NADH dehydrogenase ND5 subunit gene expression in mouse mitochondria.** *Mol Cell Biol* 2000, **20:**805–815.
36. Chomyn A: **Mitochondrial genetic control of assembly and function of complex I in mammalian cells.** *J Bioenerg Biomembr* 2001, **33:**251–257.
37. Moum T, Willassen NP, Johansen S: **Intragenic rearrangements in the mitochondrial NADH dehydrogenase subunit 6 gene of vertebrates.** *Curr Genet* 1994, **25:**554–557.
38. Bakke I, Johansen SD: **Molecular phylogenetics of gadidae and related gadiformes based on mitochondrial DNA sequences.** *Mar Biotechnol* 2005, **7:**61–69.
39. Teletchea F, Laudet V, Hanni C: **Phylogeny of the Gadidae (sensu Svetovidov, 1948) based on their morphology and two mitochondrial genes.** *Mol Phylogen Evol* 2006, **38:**189–199.
40. Roa-Varón A, Ortí G: **Phylogenetic relationships among families of Gadiformes (Teleostei, Paracanthopterygii) based on nuclear and mitochondrial data.** *Mol Phylogenet Evol* 2009, **52:**688–704.
41. Johansen SD, Coucheron DH, Andreassen M, Karlsen BO, Furmanek T, Jørgensen TE, Emblem Å, Breines R, Nordeide JT, Moum T, Nederbragt AJ, Stenseth NC, Jakobsen KS: **Large-scale sequence analyses of Atlantic cod.** *N Biotechnol* 2009, **25:**263–271.
42. Ursvik A, Breines R, Christiansen JS, Fevolden S-E, Coucheron DH, Johansen SD: **A mitogenomic approach to the taxonomy of pollocks: *Theragra chalcogramma* and *T. finnmarchica* represent one single species.** *BMC Evol Biol* 2007, **7:**87.
43. Karlsen BO, Emblem A, Jørgensen TE, Klingan KA, Nordeide JT, Moum T, Johansen SD: **Mitogenome sequence variation in migratory and stationary ecotypes of North-east Atlantic cod.** *Mar Genomics* 2014, **15:**103–108.
44. Star B, Nederbragt AJ, Jentoft S, Grimholt U, Malmstrøm M, Gregers TF, Rounge TB, Paulsen J, Solbakken MH, Sharma A, Wetten OF, Lanzén A, Winer R, Knight J, Vogel JH, Aken B, Andersen O, Lagesen K, Tooming-Klunderud A, Edvardsen RB, Tina KG, Espelund M, Nepal C, Previti C, Karlsen BO, Moum T, Skage M, Berg PR, Gjøen T, Kuhl H, *et al.*: **The genome sequence of Atlantic cod reveals a unique immune system.** *Nature* 2011, **477:**207–210.
45. Rossmanith W: **Of P and Z: Mitochondrial tRNA processing enzymes.** *Biochim Biophys Acta* 2012, **1810:**1017–1026.
46. Madsen CS, Ghivizzani SC, Hauswirth WW: **Protein binding to a single termination-associated sequence in the mitochondrial DNA D-loop region.** *Mol Cell Biol* 1993, **13:**2162–2171.
47. Roberti M, Musicco C, Polosa PL, Milella F, Gadaleta MN, Cantatore P: **Multiple protein-binding sites in the TAS-region of human and rat mitochondrial DNA.** *Biochem Biophys Res Com* 1998, **243:**36–40.
48. Bakke I, Johansen S: **Characterization of mitochondrial ribosomal RNA genes in gadiformes: sequence variations, secondary structural features, and phylogenetic implications.** *Mol Phylogen Evol* 2002, **25:**87–100.
49. Notredame C, Higgins DG, Heringa J: **T-Coffee: A novel method for fast and accurate multiple sequence alignment.** *J Mol Biol* 2001, **302:**205–217.
50. Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S: **MEGA5: Molecular Evolutionary Genetics Analysis Using Maximum Likelihood, Evolutionary Distance, and Maximum Parsimony Methods.** *Mol Biol Evol* 2011, **28:**2731–2739.
51. Abascal F, Zardoya R, Posada D: **ProtTest: selection of best-fit models of protein evolution.** *Bioinform* 2005, **21:**2104–2105.

doi:10.1186/s12862-014-0182-3
Cite this article as: Jørgensen *et al.*: An evolutionary preserved intergenic spacer in gadiform mitogenomes generates a long noncoding RNA. *BMC Evolutionary Biology* 2014 **15**:182.

Additional file 1: Figure S1



Additional file 2: Figure S2

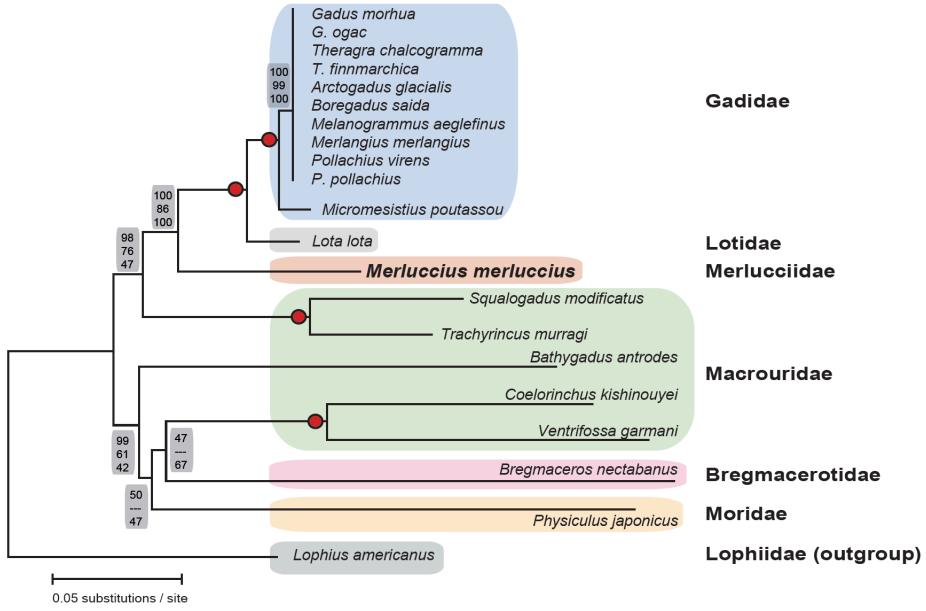


Additional file 3: Table S1

Complete gadiform mitogenomes

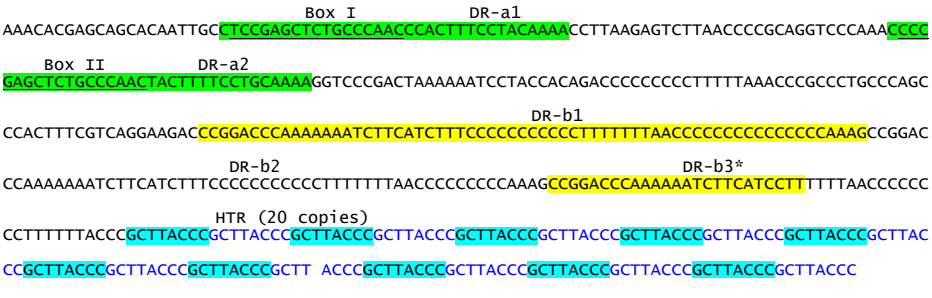
Species	Common name	Acc #
Order: Gadiformes		
Family: Gadidae		
Gadus morhua	(Atlantic cod)	AM489716
Gadus ogac	(Greenland cod)	DQ489716
Theragra chalcogramma	(Alaska Pollock)	AB094061
Theragra finnmarchica	(Norwegian Pollock)	AM489718
Boreogadus saida	(Polar cod)	AM919428
Arctogadus glacialis	(Arctic cod)	AM919429
Melanogrammus aeglefinus	(Haddock)	AM489717
Merlangius merlangius	(Whiting)	DQ020496
Pollachius virens	(Saithe)	FR751399
Pollachius pollachius	(Pollack)	FR751400
Micromesistius poutassou	(Blue whiting)	FR751401
Family: Lotidae		
Lota lota	(Burbot)	AP004412
Family: Merlucciidae		
Merluccius merluccius	(European hake)	FR751402
Family: Macrouridae		
Bathygadus antrodes	(Anadara)	AP008988
Coelorinchus kishinouyei	(Mugara grenadier)	AP002929
Squalogadus modificatus	(Tadpole whiptail)	AP008989
Trachyrincus murragi	(Roughnose grenadier)	AP008990
Ventrifossa garmani	(Sagami grenadier)	AP008991
Family: Bregmacerotidae		
Bregmaceros nectabanus	(Smallscale codlet)	AP004411
Family: Moridae		
Physiculus japonicus	(Japanese codling)	AP004409
Order: Lophiiformes		
Family: Lophiidae		
Lophius americanus	(American angler)	AP004414

Additional file 4: Figure S3

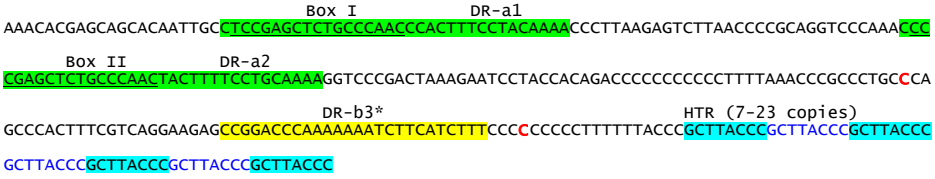


Additional file 5: Figure S4

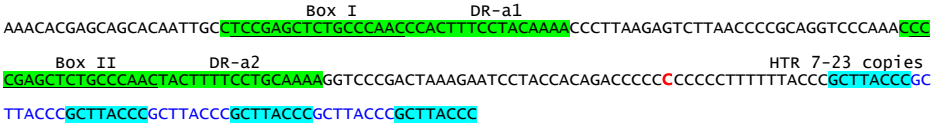
A) European hake (*M. merluccius*); Mm1



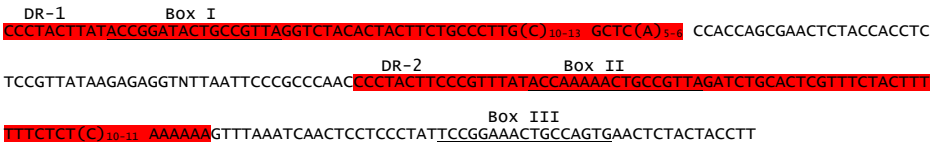
B) European hake (*M. merluccius*); Mm2 - long



C) European hake (*M. merluccius*); Mm2 - short



D) Silvery pout (*G. argenteus*)



Additional file 7: Table S2

Box-motif sequence compilation

Species	Box-motif [1]	Copy [2]
Gadiform codfishes		
Gadus morhua	ACCGGAGGACGCCCCAT	I
Gadus ogac	ACCGGGTGACGCCCCAT	I
Theragra chalcogramma	ACCGGATGACGCCCCAT	I
Theragra finnmarchica	ACCGGATGACGCCCCAT	I
Boreogadus saida	ACCGGATGTAGCCCCAT	I
Arctogadus glacialis	ACCGGGTGTAGCCCCAC	I
Melanogrammus aeglefinus	ACCGGTAACCGCCCCCC	I
Merlangius merlangius	ACCGGTTGCCGCCACAT	I
Gadus morhua	ACCGGAGGCTGCCCCAT	II
Gadus ogac	ACCGGAGGCTGCCCCAG	II
Theragra chalcogramma	ACCGGA-GCTGCCCCAT	II
Theragra finnmarchica	ACCGGAGGCTGCCCCAT	II
Boreogadus saida	ACCGGTAGCTGCCCCAT	II
Arctogadus glacialis	ACCGGTAGCCGCCCCAT	II
Melanogrammus aeglefinus	ACCGGTAGCCGCCCCAC	II
Merlangius merlangius	GCCGGAAGCTGCCCTAT	II
Pollachius virens	ACCGGATGCTGCCCCAT	
Pollachius pollachius	ACCGGATGCTGCCCCAT	
Trisopterus esmarkii	ACCGGATACTGCCCCAG	
Trisopterus minutes	ACCGGAAACTGCCCCAT	
Micromesistius poutassou	TCCGGGCACTGCCCCAT	
Gadiculus argenteus	ACCGGATACTGCCGTTA	I
Gadiculus argenteus	ACCAAAAACTGCCGTTA	II
Gadiculus argenteus	TCCGGAAACTGCCAGTG	III
Lota lota	ACCGGACACTGCCACCC	
Brosme brosme	TCCGGGCACTGCCCCAT	
Molva molva	ACCGGGCTACGCCCCAA	
Molva dipterygia	GCCGGGCTACGCCCCAA	
Enchelyopus cimbrius	ACCGGTTTAGGCCCCAC	
Gaidropsarus argentatus	CCCGTGAGGCGCCTCAC	
Phycis blennoides	CCCGAGCTCTGCCTCAC	
Macruronus novaezelandiae	ACCGGATGCTGCCCCAT	
Merluccius merluccius	TCCGAGCTCTGCCCAAC	I
Merluccius merluccius	CCCGAGCTCTGCCCAAC	II
Raniceps raninus	ACCGGACACCGCCCCTT	
Squalogadus modificatus	ACCGGGCTCTGCCCCAA	
Trachyrincus murragi	TCCAAGCTCTGCCCCAA	
CONSENSUS	aCCggangctGCCccat	
Ambystomatid salamanders		
Ambystoma amblycephalum	ACCGGGCTATGC-TACG	
Ambystoma andersoni	ACCGGGCTATGCCTACG	
Ambystoma annulatum	ACCGGGCTATGC-TACG	
Ambystoma californiense	ACCAGGCTATGCCTACA	
Ambystoma cingulatum	ACCAGGCTCTGCTTACG	
Ambystoma dumerilii	ACCGGGCTATGCCTACG	
Ambystoma mexicanum	ACCGGGCTATGCCTACG	
Ambystoma opacum	ACCGGGCTATGC-TGCA	
Ambystoma talpoideum	ACCTAACTATGC-TGCA	
Ambystoma texanum	ACCAGGCAATGCCTACG	
Ambystoma tigrinum	ACCGGGCTATGCCTACG	
Ambystoma velasci	ACCGGGCTATGC-TACG	
CONSENSUS	ACCgggCtaTGCcTaCg	

Notes: [1] Box-motif sequences obtained from T-P spacers given in Additional file 6: Figure S5 (gadiform species) and references in the main text [17,18] (ambystomatid salamanders). Consensus sequences are presented for gadiform and salamander Box-motifs. Uppercase letters indicate 100% conserved nucleotide positions. Lowercase letters indicate majority nucleotide positions. 'n' indicates non-conserved nucleotide positions. Positions conserved between codfish and salamander are indicated by red letters. [2] Box-motif copy as given in Additional file 6: Figure S5.

Additional file 6: Figure S5

A) Family: Gadidae

Gadus morhua (Atlantic cod); 74 bp; AM489716:
ACATCTAATTATTACCGGAGGACGCCCCATAATTCATCATCCACCGGAGGCTGCCCCATGGATCAAAATACCAC

Gadus ogac (Greenland cod); 73 bp; FJ396453:
ACATCTAATATTACCGGGTGACGCCCCATATATCATCATCTACCGGAGGCTGCCCCAGAGATCAAAATACCAC

Theragra chalcogramma (Alaska Pollock); 71 bp; Y17984:
ACATCTAACTATTACCGGATGACGCCCCATAATTATCAGCTACCGGAGCTGCCCCATGGTAAAAATACCAC

Theragra finnmarchica (Norwegian Pollock); 72 bp; AM489718:
ACATCTAACTATTACCGGATGACGCCCCATAATTATCAGCTACCGGAGGCTGCCCCATGGTAAAAATACCAC

Boreogadus saida (Polar cod); 70 bp; Y17985:
GCATCCAATTATTACCGGATGTAGCCCCATTTATATTAACCGGTAGCTGCCCCATGGATCAAAATACCGC

Arctogadus glacialis (Arctic cod); 291 bp; Ref [25]:
AGCATCTAATTATTACCGGGTGTAGCCCCACAATAAATAATAAATAATAAATAAATAATAAATAATAAATAATAAATAA
ATAATAAATAAATAATAAATAAATAATAAATAAATAATAAATAAATAATAAATAAATAATAAATAAATAATAAATAAATA
ATAAATAAATAATAAATAAATAATAAATAAATAATAAATAAATAATAAATAAATAATAAATAAATAATAAATAAATAAAA
TTTATTTATTTATTTTATTTACCGGTAGCCGCCCCATGGTCAAAATACCAC

Melanogrammus aeglefinus (Haddock); 99 bp; Y17986:
GCAGCACACTACCGAAGCCTCATTATTTTCAACCGGTAACCGCCCCCCCATTAACATCAGCAACCATATTCTCAACCGGT
AGCCGCCCCACCAATAACA

Merlangius merlangius (Whiting); 69 bp; DQ020496:
ACAGCTTCTTATACCGGTTGCCGCCACATAATATTCTATGCCGGAAGCTGCCCTATATAACCTTCATAC

Pollachius virens (Saithe); 59 bp; FR751399:
TTTCGCGCGCGAATATTCTACTATCACCGGATGCTGCCCCATAATGTCGAAACATGTCT

Pollachius pollachius (Pollack); 50 bp; FR751400:
CAGCACGTCTTACTTTACCGGATGCTGCCCCATAATGTCGAAACACGTTC

Trisopterus esmarkii (Norwegian pout); 77 bp; This work:
TCCACTTTATCCACCCAACCTGCAGATGCAAAGACTACCGGATACTGCCCCAGGATACCCCGATCGACCAAACTTTT

Trisopterus minutes (Poor cod); 75 bp; This work:
CCGCCTTAGTAAATAGCTCTGCATAAGAAAGACTACCGGAAACTGCCCCATAACACCCAAATCGACCAAACTTTT

Micromesistius poutassou (Blue whiting); 48 bp; FR751401:
ACGCGCATATGCTCCGGGCACTGCCCCATAATGCTAAAACATTCCTGT

Gadiculus argenteus (Silvery pout); 258 nt; This work:
CCCTACTTATACCGGATACTGCCGTTAGGTCTACACTACTTCTGCCCTTGCCCCCCCCCCGCTCAAAAACCACCAGCGAA
CTCTACCACCTCTCCGTTATAAGAGAGGTNTTAATTCCCGCCCAACCCCTACTTCCCGTTTATACCAAAAACTGCCGTTA
GATCTGCACTCGTTTCTACTTTTTTCTCTCCCCCCCCCCAAAAAAGTTTAAATCAACTCCTCCCTATTCCGGAAACTGCC
AGTGAACTCTACTACCTT

B) Family: Lotidae

Lota lota (Burbot); 103 bp; AP004412:
TATAGACACATAAATATTTAGTAAAGAGATTCAAAAATCTCTGATGCCACGAGCCGCGGCTCGTGGCTTTTCAACCGGAC
ACTGCCACCCCCCCCCCCCCAAC

Brosme brosme (Tusk); 48 bp; Y17988:
ACGCGCATATGCTCCGGGCACTGCCCCATAATGCTAAAACATTCCCTG

Molva molva (Ling); 83 bp; This work:
TCCAACATAAAGTGACTACCTCAGAGAGCTACTACACTTTCTCCAAACCGGGCTACGCCCCAAATAACCAATAAACGCCC
GCC

Molva dipterygia (Blue ling); 59 bp; This work:
CTAACAATAATTTTTCCCTTCGGGCCGGGCTACGCCCCAAATAGCCGAAACCGCCCGCC

C) Family: Phycidae

Enchelyopus cimbrius (Fourbeaed rockling); 25 bp; Y17989:
CATTTACCGGTTTAGGCCCCACAAA

Gaidropsarus argentatus (Tree-bearded rockling); 48 bp; Y17990:
CTTTCCCGTGAGGCGCCTCACTAACCTCGTGGAGATGAGCCCGCACCC

Phycis blennoides (Greater forkbeard); 28 bp; This work:
TTTTACCCCGAGCTCTGCCTCACAATAC

D) Family: Merlucciidae

Macruronus novaezelandiae (Hoki); 59 bp; This work:
TTTCGCGCGCGAATATTCTACTATCACCGGATGCTGCCCCATAATGTCGAAACATGTCT

Merluccius merluccius (European hake); 532 bp; FR751402:
AAACACGAGCAGCACAATTGCCTCCGAGCTCTGCCCAACCCACTTTCCTACAAAACCTTAAGAGTCTTAACCCCGCAGGT
CCCAAACCCCGAGCTCTGCCCAACTACTTTTCCTGCAAAAGGTCCCGACTAAAAAATCCTACCACAGACCCCCCCCCTTT
TTAAACCCGCCCTGCCCAGCCCACTTTCGTCAGGAAGACCCGGACCCAAAAAAATCTTCATCTTTCCCCCCCCCCTTTT
TTTAACCCCCCCCCCCCCCAAAGCCGGACCCAAAAAAATCTTCATCTTTCCCCCCCCCCTTTTTTTAACCCCCCCCCA
AAGCCGGACCCAAAAAATCTTCATCCTTTTTTAACCCCCCCCTTTTTTACCCGCTTACCCGCTTACCCGCTTACCCGCTT
ACCCGCTTACCCGCTTACCCGCTTACCCGCTTACCCGCTTACCCGCTTACCCGCTTACCCGCTTACCCGCTTACCCGCTT
ACCCGCTTACCCGCTTACCCGCTTACCCGCTTACCCGCTTACCCGCTTACCC

E) Family: Ranicipitidae

Raniceps raninus (Tadpole fish); 39 bp; This work:
ATCTCAACTTTTTATACCGGACACCGCCCCTTAACACAA

F) Family: Macrouridae

Squalogadus modificatus (Tadpole whiptail); 47 bp; AP008989:
ACCTAGCACATTTTTTTACCGGGCTCTGCCCCAAATATTAAAATCAA

Trachyrincus murragi (Roughnose grenadier); 64 bp; AP008990:
TCACTTTTTTTCCAAGCTCTGCCCCAAATCAATGGAAATAAAGCTCAAAGGGGGAAGATTCGAA

Bathygadus antrodes (Anadara); 7 bp; AP008988:
ATATTTT

Caelorinchus kishinouyei (Mugara grenadier); 12 bp; AP002929:
CTAAACTACCCA

Coryphaenoides rupestris (Roundnose grenadier); 21 bp; This work:
CACCGACGTTCACCACCCACA

Ventrifossa garmani (Sagami grenadier); 72 bp; AP008991:
CTTCCCCACTCAGTTTTTTATTCCGAGTTCCGCCACACATATTTTAACGCAAAACGGCCCCCCGCCGGATGA

G) Family: Moridae

Physiculus japonicas (Japanese codling); 16 bp; AP004409:
CACCGCCCTCTGCCAC

Laemonema longipes (Longfin codling); 33 bp; AB108839:
TCCCTTACCCCCGGACACCGCCACACTCTTCAA

Additional file 8: Table S3

PCR and sequencing primer specifications

Primer [1]	Sequence (5′→3′)	Gene region
(T-P spacer PCR and sequencing primers):		
L15498	GAAACTGCCCTAGTAGCCA	tRNA-Thr
H15666	GTTTAATTTAGAATTCTAGCTTTGG	tRNA-Pro
(European hake PCR and sequencing primers):		
H112	CGGAGACTTGCATGTGT	SSU
L347	ACGAGCCCAAGTTGATA	SSU
H617	TAGAACAGGCTCCTCTAG	SSU
L1223	CGCAAGGGAACGCTGAAA	LSU
L2538	AATGGTGCAGCCGCTAT	LSU
L2776	TGTGTTAGAGTGGCAGAGC	tRNA-Leu1
H2812	CTCTGTAGAAAGGGCTTAGG	tRNA-Leu1
H3454	CTCTCCTTCAGTCAGGTCGA	ND1
L3760	TGGCACTAGTGATTTGACAT	ND1
L5072	GATAAACTAGACCAAGGGCCT	tRNA-Trp
H5481	GTGCCAATGTCTTTGAGATT	COI
L5862	CTTGCAAGTAATCTTGCCC	COI
L6842	CTAATCTCACTTCTAGCAAGT	COI
H7079	GGCAGTGTGATATTTCACCAG	tRNA-Asp
L7208	ATCACCCGTAATAGAAGAGT	COII
L7849	ACTTGAAGACGCCTCACTAAG	tRNA-Lys
L8016	GATAAACTAGACCAAGGGCCT	ATP8
H8073	CTTAGTGTCATGGTCAGTTTC	ATP8
L8864	CCTGCCGTATGGTTCCA	COIII
L9547	CAATCTATTGATGAGGCTC	tRNA-Gly
H10980	CCAGAGGTGTATTCCGTAGA	ND4
L11759	CTCTAATCCACCGAGAGAGGC	tRNA-His
H11887	ACTTGGAGTTGCACCAAGAG	tRNA-Leu2
L12809	CTGTGCCTCTGCCTCGGCG	ND5
H13950	TTGGGCCTCAGTTGAGGAGG	ND6
L14626	ACGGATGACTAATCCGCAAC	Cyt-B
H14746	AATTACGGTAGCTCCTCAGAATGATATTTGTCCTCA	Cyt-B
L15491	CTTGAATGAAACTGCCCTAG	tRNA-Thr
H15642	AGTTAGTGATGGGAGTTTAAGTC	tRNA-Pro
H16000	GGGTTGACAGGTTAAATACG	CR

Note: [1] The primers are numbered according to the 3' nucleotide of the complete Atlantic cod mitogeneome sequence [21].

List of previously published theses for PhD in Aquaculture / PhD in Aquatic Biosciences,

Nord University

No. 1 (2011)
PhD in Aquaculture
Chris André Johnsen
Flesh quality and growth of farmed Atlantic salmon (*Salmo salar* L.) in relation to feed, feeding, smolt type and season
ISBN: 978-82-93165-00-2

No. 2 (2012)
PhD in Aquaculture
Jareeporn Ruangsri
Characterization of antimicrobial peptides in Atlantic cod
ISBN: 978-82-93165-01-9

No. 3 (2012)
PhD in Aquaculture
Muhammad Naveed Yousaf
Characterization of the cardiac pacemaker and pathological responses to cardiac diseases in Atlantic salmon (*Salmo salar* L.)
ISBN: 978-82-93165-02-6

No. 4 (2012)
PhD in Aquaculture
Carlos Frederico Ceccon Lanes
Comparative Studies on the quality of eggs and larvae from broodstocks of farmed and wild Atlantic cod
ISBN: 978-82-93165-03-3

No. 5 (2012)
PhD in Aquaculture
Arvind Sundaram
Understanding the specificity of the innate immune response in teleosts: Characterisation and differential expression of teleost-specific Toll-like receptors and microRNAs
ISBN: 978-82-93165-04-0

No. 6 (2012)
PhD in Aquaculture
Teshome Tilahun Bizuayehu
Characterization of microRNA during early ontogeny and sexual development of Atlantic halibut (*Hippoglossus hippoglossus* L.)
ISBN: 978-82-93165-05-7

No. 7 (2013)
PhD in Aquaculture
Binoy Rajan
Proteomic characterization of Atlantic cod skin mucosa – Emphasis on innate immunity and lectins
ISBN: 978-82-93165-06-04

No. 8 (2013)
PhD in Aquaculture
Anusha Krishanthi Shyamali Dhanasiri
Transport related stress in zebrafish: physiological responses and bioremediation
ISBN: 978-82-93165-07-1

No. 9 (2013)
PhD in Aquaculture
Martin Haugmo Iversen
Stress and its impact on animal welfare during commercial production of Atlantic salmon (*Salmo salar* L.)
ISBN: 978-82-93165-08-8

No. 10 (2013)
PhD in Aquatic Biosciences
Alexander Jüterbock
Climate change impact on the seaweed *Fucus serratus*, a key foundational species on North Atlantic rocky shores
ISBN: 978-82-93165-09-5

No. 11 (2014)
PhD in Aquatic Biosciences
Amod Kulkarni
Responses in the gut of black tiger shrimp *Penaeus monodon* to oral vaccine candidates against white spot disease
ISBN: 978-82-93165-10-1

No. 12 (2014)
PhD in Aquatic Biosciences
Carlo C. Lazado
Molecular basis of daily rhythmicity in fast skeletal muscle of Atlantic cod *(Gadus morhua)*
ISBN: 978-82-93165-11-8

No. 13 (2014)
PhD in Aquaculture
Joanna Babiak
Induced masculinization of Atlantic halibut (*Hippoglossus hippoglossus* L.): towards the goal of all-female production
ISBN: 978-82-93165-12-5

No. 14 (2015)
PhD in Aquaculture
Cecilia Campos Vargas
Production of triploid Atlantic cod: A comparative study of muscle growth dynamics and gut morphology
ISBN: 978-82-93165-13-2

No. 15 (2015)
PhD in Aquatic Biosciences
Irina Smolina
Calanus in the North Atlantic: species identification, stress response, and population genetic structure
ISBN: 978-82-93165-14-9

No. 16 (2016)
PhD in Aquatic Biosciences
Lokesh Jeppinamogeru
Microbiota of Atlantic salmon (*Salmo salar L.*), during their early and adult life
ISBN: 978-82-93165-15-6

No. 17 (2017)
PhD in Aquatic Biosciences
Christopher Edward Presslauer
Comparative and functional analysis of microRNAs during zebrafish gonadal development
ISBN: 978-82-93165-16-3

No. 18 (2017)
PhD in Aquatic Biosciences
Marc Jürgen Silberberger
Spatial scales of benthic ecosystems in the sub-Arctic Lofoten-Vesterålen region
ISBN: 978-82-93165-17-0

No. 19 (2017)
PhD in Aquatic Biosciences
Marvin Choquet
Combining ecological and molecular approaches to redefine the baseline knowledge of the genus Calanus in the North Atlantic and the Arctic Oceans
ISBN: 978-82-93165-18-7

No. 20 (2017)
PhD in Aquatic Biosciences
Torvald B. Egeland
Reproduction in Arctic charr – timing and the need for speed
ISBN: 978-82-93165-19-4

No. 21 (2017)
PhD in Aquatic Biosciences
Marina Espinasse
Interannual variability in key zooplankton species in the North-East Atlantic: an analysis based on abundance and phenology
ISBN: 978-82-93165-20-0

No. 22 (2018)
PhD in Aquatic Biosciences
Kanchana Bandara
Diel and seasonal vertical migrations of high-latitude zooplankton: knowledge gaps and a high-resolution bridge
ISBN: 978-82-93165-21-7

No. 23 (2018)
PhD in Aquatic Biosciences
Deepti Manjari Patel
Characterization of skin immune and stress factors of lumpfish, *Cyclopterus lumpus*
ISBN: 978-82-93165-22-4

No. 24 (2018)
PhD in Aquatic Biosciences
Prabhugouda Siriyappagouder
The intestinal mycobiota of zebrafish – community profiling and exploration of the impact of yeast exposure early in life
ISBN: 978-82-93165-23-1

Mitochondria are cytoplasmic organelles present in almost all eukaryotic cells and organisms. Mitochondria are the main supplier of the cellular energy (ATP), and have an important role in cell function regulation including innate immunity, stress responses, and apoptosis. The mitochondria contain residual small circular genomes that apparently are highly conserved among vertebrates. In this thesis we have investigated molecular features of the mitochondrial genome in the Atlantic cod *(Gadus morhua)* and related codfishes using large-scale DNA and RNA sequencing. We have studied mitochondrial genome sequence variation among > 200 specimens of the Atlantic cod representing different ecotypes and geographic locations. The analyses were focused on long noncoding RNAs, and molecular hallmarks such as heteroplasmic tandem repeats and noncoding spacers. The Atlantic cod mitochondrial genome was found to be surprisingly variable with a gene content extending the 37 conventional genes known to be highly conserved among vertebrates. This work provides new knowledge on vertebrate mitochondrial genomes in general, and Atlantic cod and codfishes more specifically.

NORD University

ISBN: 978-82-93165-24-8
Trykk: Trykkeriet, Nord universitet
www.nord.no