Natural selection and speciation in Atlantic cod and related cod-fish

Katrín Halldórsdóttir
Natural selection and speciation in Atlantic cod and related cod-fish

Katrín Halldórsdóttir

Dissertation submitted in partial fulfillment of a Philosophiae Doctor degree in Biology

Advisor
Arnar Pálsson

PhD Committee
Arnar Pálsson
Pétur Henry Petersen
Helga Ögmundsdóttir

Opponents
Matthew W. Hahn
Michael Matschiner

Faculty of Life and Environmental Sciences
School of Engineering and Natural Sciences
University of Iceland
Reykjavik, April 2016
Natural selection and speciation in Atlantic cod and related cod-fish
Natural selection and speciation in cod
Dissertation submitted in partial fulfillment of a Philosophiae Doctor degree in Biology

Copyright © Katrín Halldórsdóttir 2016
All rights reserved

Faculty of Life and Environmental Sciences
School of Engineering and Natural Sciences
University of Iceland
Sæmundargötu 2
101 Reykjavik, Reykjavik
Iceland

Telephone: 525-4000

Bibliographic information:
Katrín Halldórsdóttir, 2016, Natural selection and speciation in Atlantic cod and related cod-fish, PhD dissertation, Faculty of Life and Environmental Sciences, University of Iceland, 270 pp.

ISBN 978-9935-9307-4-3

Printing: Samskipti ehf.
Reykjavík, Iceland, April 2016
Abstract

Natural selection is the main force in evolution. Population genetics, the theoretical explanation of evolution in modern Darwinism, is a study of the frequencies and interaction of alleles and genes in populations and how they change under the influence of evolutionary forces. Population genomics extends the field to the whole genomes of the organisms. The fingerprints of natural selection can be detected in molecular data. The organismic models used for studying selection are of importance. The Atlantic cod, *Gadus morhua*, is well known for its extreme fecundity. Each female spawns millions of eggs each time. An organism with such a high fecundity should be able to withstand considerable selection and respond more quickly to environmental pressures than less fecund organisms. Therefore, Atlantic cod is an excellent model for studying natural selection at the molecular level in wild populations. There have been large improvements in molecular techniques and in methods of data generation in evolutionary biology in recent years which significantly enhance population genetics and genomics. Here in this thesis several of these methods are applied to study selection in Atlantic cod and related cod-fish species from the Atlantic and Pacific ocean. Candidate genes under selection were studied and the work was then extended to sequencing of whole genomes. Comparison between organisms of related taxa can be useful in estimating divergence and admixture and in understanding which selective factors are important in Darwinian fitness of the organisms. The cod-fish analysed are in addition to Atlantic cod, the Pacific cod, *Gadus macrocephalus*; Greenland cod, *Gadus ogac*; walleye pollock, *Gadus chalcogrammus*; Arctic cod, *Boreogadus saida* and Polar cod, *Arctogadus glacialis*. A study was done on structure and arrangement of candidate selected globin genes in the Atlantic cod genome. Balancing selection is one of the main forces in maintaining genetic variation in populations. Evidence of trans-species polymorphism, which is a most important evidence of balancing selection, was found in Cathelicidin innate immunity genes. New inference methods based on $Λ$ coalescents are used as neutral null models to study selection at a $Ckma$ (Creatine Kinase Muscle A) gene. The results show that multiple merger coalescents better suit as null models for organisms with high fecundity and type III survivorship than bifurcating Kingman coalescent in describing their gene genealogy. Finally, for the first time a population study using whole genome sequencing data of individual cod-fish was performed. In this study new statistical methods of genotype likelihoods, appropriate for low coverage sequence data, were applied in the analysis of the genomic data. The results revealed new knowledge about speciation in Atlantic cod and the origin and admixture between Atlantic and Arctic cod. The findings are important for understanding the evolutionary status of the Pacific species. Walleye pollock is shown to be a hybrid between Arctic cod and Atlantic cod. Various subgroups of Atlantic cod were found to be genetically distinct. A new hypothesis based on a model of divergence-after-speciation is proposed for the evolutionary status of the known behavioural ecotypes of cod around Iceland, the frontal and coastal behavioural types. The two types are proposed to be separate species, adapted to deep and shallow waters respectively, which hybridized and formed a new homoploid hybrid species. The homoploid hybrid species transgresses the ecology of the parental types.
Útdráttur


Dedication

To my children; Oddur, Nanna Bríet and Tjörvi.
# Table of Contents

Abstract iii  
Útdráttur v  
Dedication vii  
Table of Contents ix  
List of Figures xi  
List of Original Papers xiii  
Acknowledgments xv  

1 Introduction 1  
1.1 Population genetics in the age of genomics .......................... 1  
1.2 Detecting fingerprints of natural selection ............................... 5  
1.3 Cod-fish .............................................................................. 7  
1.3.1 Atlantic cod — *Gadus morhua* ........................................ 7  
1.3.2 Pacific and Arctic species .................................................. 9  
1.3.3 Summary ......................................................................... 10  

2 Objectives and methodology 13  

3 Results and discussion 15  
3.1 Candidate genes approach — (Papers I-IV) ............................. 15  
3.1.1 The hemoglobin loci of Atlantic cod .................................... 15  
3.1.1.1 Synopsis .................................................................. 18  
3.1.2 Innate immunity genes — Cathelicidin genes ......................... 18  
3.1.2.1 Synopsis .................................................................. 20  
3.1.3 Screening for candidate genes in a genomic BAC-Library ........... 20  
3.2 Coalescence approach - (Paper IV) ........................................ 24  
3.2.1 Site frequency spectrum ................................................... 25  
3.2.2 Appropriate coalescent null models ..................................... 26  
3.2.3 Synopsis ......................................................................... 28  
3.3 Genome sequencing approach - (Paper V) .............................. 28  
3.3.1 Partial/reduced genome sequencing approach ......................... 29  
3.3.1.1 RAD-sequencing ....................................................... 30  
3.3.2 Whole genome sequencing ............................................... 34  
3.3.3 Speciation ........................................................................ 34  
3.3.3.1 Speciation in the sea ................................................... 35  
3.3.3.2 Homoploid hybrid speciation ....................................... 35  
3.3.3.3 A hypothesis of walleye pollock origin ............................ 38  
3.3.3.4 Species and models of speciation in Atlantic cod .................. 38  
3.3.4 Synopsis ......................................................................... 42  

4 Future Perspectives 45  
References 47  

Paper I 61
Table of Contents

Paper II .......................... 77
Paper III .......................... 93
Paper IV .......................... 125
Paper V .......................... 159
Appendix I. Supplementary data for Paper I .......................... 171
Appendix II. Supplementary data for Paper II .......................... 179
Appendix III. Supplementary data for Paper III .......................... 207
Appendix IV. Supplementary data for Paper IV .......................... 233
Appendix V. Supplementary data for Paper V .......................... 251
List of Figures

3.1 $HbI^1$ allele frequency cline along the Norwegian coast. .......................... 16
3.2 The $HbI$ locus with neighboring globin genes. ................................. 17
3.3 Sample locations around Iceland. .................................................. 22
3.4 Map of sampling sites of Atlantic cod and closely related species. ....... 23
3.5 $F_{ST}$ outlier detection. ............................................................... 24
3.6 Type III survivorship curve. ......................................................... 26
3.7 Schematic genealogies for three types of coalescent models. ............... 27
3.8 Inferred cod-fish tree with mixture events. ..................................... 31
3.9 Inferred cod-fish tree with different number of mixture events. .......... 32
3.10 Within- and between-species variation along a single chromosome. ...... 33
3.11 Schematic of recombinational chromosomes in hybrid speciation. ....... 36
3.12 Schematic of unbalanced chromosomes in hybrid speciation. ............ 37
3.13 A model of hybrid speciation in Atlantic cod. .................................. 41
List of Original Papers


**Paper III:** Halldórsdóttir, K. and Árnason E. 2015. Trans-species polymorphism at antimicrobial innate immunity cathelicidin genes of Atlantic cod and related species. *PeerJ* 3 e976

**Paper IV:** Árnason E. and Halldórsdóttir K. 2015. Nucleotide variation and balancing selection at the *Ckma* gene in Atlantic cod: analysis with multiple merger coalescent. *PeerJ* 3 e786

**Paper V:** Halldórsdóttir, K. and Árnason E. 2015. Whole genome sequencing uncovers cryptic and hybrid species among Atlantic and Pacific cod-fish. *In review*
**Acknowledgments**

I would like to thank my thesis committee for their helpful advice and discussions. I especially thank Arnar Pálsson for his help with writing, critical comments on the manuscripts and on my thesis, discussions, and his general positive attitude.

I thank the people of lab 387 in Askja throughout the years for their company. Especially my friends in the pop-gen group Guðni Magnúss Eiríksson, Svava Ingimarsdóttir and Ubaldo Hernandez for the science discussions on the project but particularly all the foolery, laughing, and the good times.

I thank the people who have provided us with samples: Jarle Mork, Grant Pogson, Michael Canino and Kristján Kristjánsson.

Special thanks go to Dick Lewontin for giving me the opportunity to work at Harvard in an atmosphere of hospitality where I finally had time to finish most of my writing. His charm and presence, enthusiasm in science and critical discussions were encouraging. It was invaluable to have access to Hopi Hoekstra’s lab and the Bauer Core Facility to do the genomic work. I thank Hopi Hoekstra and Kyle Turner for their assistance.

I also thank Svala Árnadóttir and Dick Lewontin for their financial support for the work at Harvard.

I thank my family and friends for their support throughout the years; special thanks to Gutta and Guðlaug for moral support in difficult times, my parents and my most precious people in the world, Oddur, Nanna, and Tjörvi, for always being there for me. I also would like to thank my sisters in Kvenfélagið Ung og Aðlaðandi for reminding me of all the other things in life outside science.

Last but not least. My dear, dear Einar. As Lennon put it: “Life is What Happens To You While You’re Busy Making Other Plans”. I guess “life” just happened to us! To live and work with someone and there is never a dull moment, truly is a privilege. Your inspiring and genius mind keeps me busy all the time. Thank you for your endless support and love, I wouldn’t have finished this without your encouragement.

The work was supported by the Icelandic Science Foundation grant of excellence nr. 4030311 and The University of Iceland Research Fund.
1 Introduction

It is interesting to contemplate an entangled bank, clothed with many plants of many kinds, with birds singing on the bushes, with various insects flitting about, and with worms crawling through the damp earth, and to reflect that these elaborately constructed forms, so different from each other, and dependent on each other in so complex a manner, have all been produced by laws acting around us. These laws, taken in the largest sense, being Growth with Reproduction; Inheritance which is almost implied by reproduction; Variability from the indirect and direct action of the external conditions of life, and from use and disuse; a Ratio of Increase so high as to lead to a Struggle for Life, and as a consequence to Natural Selection, entailing Divergence of Character and the Extinction of less-improved forms. Thus, from the war of nature, from famine and death, the most exalted object which we are capable of conceiving, namely, the production of the higher animals, directly follows. There is grandeur in this view of life, with its several powers, having been originally breathed into a few forms or into one; and that, whilst this planet has gone cycling on according to the fixed law of gravity, from so simple a beginning endless forms most beautiful and most wonderful have been, and are being, evolved. (Charles Darwin 1859; On the Origin of Species by Means of Natural Selection).

In the last paragraph of his book, Darwin condenses the theory of evolution by natural selection into this beautiful poetic metaphor of an entangled bank. One can almost smell the flowers and hear the bugs buzzing. Underlying this fascinating apparent chaotic scene is the certainty that it is explicable by the simple natural laws. This is where modern evolutionary biology as a discipline begins, a new philosophical thinking that revolutionized our worldview. Darwin brought the tools that we are still using to build an understanding of life on earth and of the relatedness between individuals and organisms around us.

1.1 Population genetics in the age of genomics

Population genetics is the theoretical explanation of evolution in modern Darwinism. The field is the study of the frequencies and interaction of alleles and genes in populations and how they change under the influence of the four main evolutionary processes: natural selection, random genetic drift, mutation, and gene flow. It attempts to explain such phenomena as adaptation and speciation. The fundamental works of Ronald A. Fisher (Fisher, 1930), J.B.S. Haldane (Haldane, 1932) and Sewall Wright (Wright, 1931) provided the theoretical underpinnings to population genetics in the 1920s and 1930s and formed much of the paradigm still used today (Wakeley,
The field has evolved from being a field of biology with its theoretical problems and issues well defined, yet with little empirical data, into a data-driven field which tests the tolerance limits of the models available (Pool et al., 2010; Wakeley, 2004). The primary goal of population and evolutionary genetics is to understand processes and patterns of genetic variation within populations and patterns of divergence between species. Population genetics and its evolutionary interpretations provide context to interpret the wealth of molecular data produced by modern revolutionary molecular tools.

Population genomics—a new field emerging with the genomic revolution—combines techniques, concepts, and thinking of genomics with the principles of population genetics to further our understanding of evolution (Black et al., 2001; Luikart et al., 2003; Nielsen et al., 2009b). The field emerges with the sequencing of the genomes of several model organisms and can be defined broadly as the simultaneous study of large numbers of genetic loci or genomic regions with the aim of elucidating the role which the various forces of evolution play in molding variation across both genomes and populations in time and space.

A historic debate has been going on between neutralist and selectionists from the early days of molecular data. According to the neutral theory (Kimura, 1968; King and Jukes, 1969) the great bulk of molecular variation found between individuals is selectively neutral, caused by mutation and influenced mainly by genetic drift. On the other hand the selectionists consider the majority of variation to be actively maintained by balancing Darwinian selection, affecting the fitness of the organism. Even though the functional consequences of the variation are in many or most cases not known, it must affect the fitness of the genotypes (Gillespie, 1991) under the selectionist view.

Measurements of allozyme variability in proteins within and among populations, data gained with electrophoresis that became available and accumulated after the pioneering work of Hubby and Lewontin (1966), revealed much more variability within populations than previously thought to exist. Every morphological or physiological character was thought by selectionists to have evolved by natural selection and maintained by balancing selection. Hubby and Lewontin (1966) showed that 12% of loci in the genome of an average individual *Drosophila* were variable. If each of them was maintained by balancing selection independent of other loci, the genetic load would be too high for any population to bear (genetic load is the cost for a population of having individuals of less favoured genotype with low fitness compared to the most favoured genotype). This was the contradiction in the selectionist view of polymorphism (Nei, 2005). Sved et al. (1967), Milkman (1967) and King (1967) criticized the idea of selection acting independently on each locus. Instead, they argued that cumulative effects of genes at many loci in an individual as a whole will make individual fitness. And in continuation Kimura (1968) and King and Jukes (1969) proposed the neutral theory as an explanation of the great amount of variation found in the genomes within a species and of molecular divergence between species.

Gould and Lewontin (1979) criticized what they called the adaptationist program, the tendency to explain every phenotype of organism as an adaptation. Their fundamental critique is that form does not prove purpose and function and observations on functional advantage cannot always be explained by adaptation. Their inventive idea to use the spandrels, spaces that inevitably develop with the construction of an arch, which are often artfully decorated in churches, as a metaphor for the adaptationist storytelling indeed changed the debate in evolutionary biology (Gould and Lewontin,
1.1 Population genetics in the age of genomics

1979; Nielsen, 2009). As Nielsen (2009) describes in his commentary, even though we find selection acting on genes behind functional effects we cannot claim that the selection is not caused by some pleiotropic effect of the mutation or hitch-hiking due to linkage (Maynard Smith and Haigh, 1974). From the point of view of evolutionary biology selective differences are not equal to adaptive changes. A link between genetic variation and fitness differences among individuals is required (Kawecki and Ebert, 2004).

Sober (1993) urged us to differentiate between ‘selection of’ and ‘selection for’. The difference reflects if the selection is acting on the particular phenotype or if the phenotype is a consequence of selection acting on other targets. As an example of the problems involved in discussions about selection of versus selection for we can examine the well known hemoglobin polymorphism in Atlantic cod Gadus morhua. Karpov and Novikov (1980) showed different norms-of-reaction of oxygen-binding affinity among hemoglobin genotypes. Andersen et al. (2009) described the polymorphic Hb-β1 locus in Atlantic cod where two non-synonymous substitutions were found unambiguously associated to the formerly discovered, by protein electrophoresis, hemoglobin phenotypes HbI-1/1 and HbI-2/2 (Sick, 1961). Andersen et al. (2009) state that the amino acid change at position 55 Met/Val and at position 62 Lys/Ala at the α1/β1 subunit interface and heme pocket affect the oxygen-binding properties and electrostatic features resulting in different oxygen affinity between the phenotypes. They genotyped several individuals using Single Nucleotide Polymorphism (SNP) variation in the β1 gene from cDNA. Their conclusion is that fish possessing Val55-Ala62 are better adapted to cold and low-oxygen waters. Brix et al. (2004) showed that heterozygous individuals, HbI-1/2, increased synthesis of the product of the appropriate allele depending on the temperature of the water they were exposed to. However, it has not been shown whether this polymorphism in the β1 gene is the selective target or whether the effects seen are due to linkage to other variation in the genome. Borza et al. (2009) also published a thorough genomic study of the Hb genes expressed in Atlantic cod. They base their results on genomic sequencing data and describe extensive allelic variation at nine hemoglobin loci. They described eight alleles of the β1 gene that are highly differentiated showing 18 SNPs and multiple indel differences and two or three amino acid differences. The two main allele types (Met55/Lys62 - phenotype HbI-1/1 and Val55/Ala62 - phenotype HbI-2/2) are represented by two subtypes which differ at synonymous sites and intronic sequences (Borza et al., 2009). The additional variation at this locus (and elsewhere in the genome) gives us reasons to be cautious in interpreting selective effects as selection for because they could be selection of.

SNP data sets give comparative studies the opportunity to answer and test many predictions of the theory of population genetics about positive and negative selection (Nielsen, 2005a). SNP data should in principle be favourable to analyse regions of the genome detected by selection. A problem of ascertainment bias (Kreitman and Di Rienzo, 2004) immediately arises in SNP data due to the two stage process: first there is discovery of SNP markers and second there is their characterization in populations. Informative SNPs are searched for in a few individuals followed by genotyping large samples (Nielsen, 2005a; Bradbury et al., 2011). This creates a bias because only high frequency variants are picked up in the discovery stage. How the SNPs were chosen affects the estimates of association between loci and linkage disequilibrium (LD) in the data, as well as estimated demographic parameters such
as population subdivision (Nielsen, 2004, 2005a). Increasing use of SNPs to study demographic and adaptive processes in non-model organisms, such as Atlantic cod, demand a critical evaluation of the ascertainment bias (Hubert et al., 2010; Helyar et al., 2011; Bradbury et al., 2011). Bradbury et al. (2011) report that ascertainment bias affects the assignment success in a panel of 1641 expressed sequence tag-derived SNPs developed from northwest Atlantic cod. In their study they excluded SNPs possibly influenced by selection. In contrast Nielsen et al. (2009a) used a genome scan to produce a SNP data set of candidate genes under selection. Their SNP discovery for 12 new SNPs from candidate genes for adaptive evolution came from four different areas of Atlantic cod distribution. They also used previously described SNPs from Moen et al. (2008). Nielsen et al. (2009a) revealed directional selection for local adaptations along multiple environmental dimensions. On the other hand they found no evidence of balancing selection in cod, which is difficult to detect in data with low levels of population structure (Nielsen et al., 2009a). They conclude that the effect of ascertainment bias is a minor problem. The fish Moen et al. (2008) used for ascertainment in their study were 15 Norwegian coastal cod and 500 embryos. It is not indicated to what extent the EST sequences from the embryo libraries could possibly have differed from adult libraries. The question arises if embryo EST cDNA library is representative for the samples that were genotyped using these markers (Nielsen et al., 2004; Wang and Nielsen, 2012). That is, whether the SNPs found are from some important developmental sequences essential for embryos but with limited or no relevance to adult fish.

Using SNPs for studying population genetics of non-model organisms is challenging. Helyar et al. (2011) discuss the issues of increased number of markers available, ascertainment bias and effects of non-neutral loci. Because of the issues outlined above SNPs with low minor allele frequencies are less likely to be discovered. This results in an overestimate of average diversity of polymorphic sites and underestimate of average diversity across all sites, introducing a bias towards common alleles (Helyar et al., 2011). Correction of the allele frequency spectrum using specific models have been proposed (Nielsen et al., 2004). However, the correction has consequences for further analysis that depend on the allele frequency spectrum (Helyar et al., 2011). Wang and Nielsen (2012) propose a solution to this problem in large genome-wide SNP data set with more than 10,000 SNPs (such as in humans) by using an outgroup population for ascertainment of the SNPs. However, this would rarely be possible for non-model organisms such as cod.

To make general inferences about genome-wide neutral evolutionary processes, e.g. genetic drift and gene flow, neutral markers are needed. SNPs in gene regions are more likely to show signs of both positive and negative selection (Helyar et al., 2011) than are SNPs located outside gene regions. SNPs extracted from EST sequences and cDNA libraries are therefore biased towards genic regions and might be difficult to use for inference about neutral evolutionary forces. However, if long noncoding RNAs (lncRNA) turn out to be commonly transcribed as indicated lately (Ponting et al., 2009) they will be included in ESTs which will then represent a larger part of the genome, thereby minimizing this problem (Ponting et al., 2009; Helyar et al., 2011).

As pointed out by Nielsen et al. (2009b) sequence-based analyses in marine fishes are very few but they provide detailed information on action of evolutionary forces such as the study by Pogson (2001). This approach should also lower the effect of ascertainment bias. If the aim of a study is to identify candidate genes under selection
1.2 Detecting fingerprints of natural selection

It is of great importance to study positive selection because it is associated with adaptation and the evolution of new form and function. In analysis of selection the functional differences among genotypes have to be related to the variation found in DNA sequences to understand the direct cause and effect relationship between DNA sequence variation and fitness differences among genotypes. According to Árnason and Barker (2000) the analysis of selection involves three stages: i) Detection of selection, ii) determining what is the relative fitness of different genotypes and iii) assessing how the fitness differences relate to the biology and ecology of the species.

In his seminal study Kreitman (1983) described nucleotide diversity at the Alcohol dehydrogenase Adh gene in Drosophila melanogaster. The results of mainly silent polymorphisms in exons and introns of the two electromorphs Adh-f and Adh-s, which differ by one amino acid, were interpreted as a strong indication that mutations causing amino acid replacement have been selectively deleterious. Kreitman’s explanation was that different strengths of natural selection are working on synonymous and non-synonymous mutations in coding regions and that these regions of the genome are evolutionarily constrained by natural selection (Schaeffer and Aguadé, 2000; Hedrick, 2005; Kreitman, 1983). Kreitman’s thesis is the basis of many current methods used to detect selection. The paper considered the extent of silent polymorphism compared to amino acid replacement polymorphism (the ratio of synonymous to non-synonymous polymorphism within and between species). Since then the study of molecular evolution has been mainly at the DNA level and the main improvement of population genetics last decade is the amount of easily gained DNA sequencing data. Population genetics is currently a data driven field. All the new data have nevertheless not solved the controversy between selectionists and neutralists. Instead it has taken on a new form which is partially caused by new techniques of within population research that focuses on the detection of positive selection using neutrality as a null model.

Interestingly, the issues of selection, neutrality and function are hotly debated in connection with results from the Encyclopedia Of DNA Elements (ENCODE) project (Dunham et al., 2012). The ENCODE proponents argue, using a new definition of
a function for non-coding DNA (previously referred to as junk DNA), that they are able to assign biochemical functions to more than 80% of variable sites in the human genome (Dunham et al., 2012). This view has been heavily criticized and the main objections are based on the concept of the C-paradox. If it is correct that almost every nucleotide site in the genome has a function all these sites as such are targets of natural selection to act upon. Then the expectation is that more complex organisms should have larger genomes (Doolittle, 2013; Graur et al., 2013). But on contrary single-celled amoebae have some of the largest genomes, up to 100-fold larger than the human genome (Eddy, 2012) and the genome of the onion is four or five times the size of the human genome (Graur et al., 2013). This debate is interesting because this reflects how deep a division is between opposite views in the field. In effect the old debate of the neutralist/selectionist controversy is unresolved.

There has thus been an increasing interest in finding genes and genomic region exhibiting locus specific effects, that have been or are being targeted by natural selection (Nielsen, 2005a,b; Hemmer-Hansen et al., 2007a,b; Akey et al., 2002; Nielsen et al., 2005, 2007; Bustamante et al., 2005; Sabeti et al., 2007; Akey et al., 2004; Williamson et al., 2007; Andolfatto, 2005; Nielsen et al., 2009b). This is based on the realization that inference on natural selection can provide important functional information. Genes and genomic regions influenced by selection are important for understanding different ecological functions and selective forces important in the ecology of the organism. Genes and variation may be targeted under different modes of selection: positive, negative, balancing, frequency and density dependent. Selection may in many cases increase the degree of population differentiation (Hedrick, 2005). In such cases there will be a local or global geographic component to the differentiation, perhaps detectable by \( F_{ST} \) outliers (Lewontin and Krakauer, 1973; Beaumont and Nichols, 1996; Beaumont, 2005; Foll and Gaggiotti, 2008; Gompert and Buerkle, 2011). In other cases ecological/environmental selective pressures responsible may be part of the niche structure of the organisms unrelated to geography. Of course geography may be a proxy for contrasting natural environments (e.g. salt-, brackish, and freshwater habitats, Vasemägi et al., 2005). Furthermore, the niche-variation hypothesis (Van Valen, 1965) states that genetic variation may be a strategy for increasing population fitness in temporally and spatially variable habitats (Somero and Soulé, 1974). The concept is that species of ecological generalists are in fact a collection of specialized individuals giving increased intra-population variation that is key in frequency-dependent interactions that influence population dynamics and the ecology of the organism (Bolnick et al., 2007). Thus, although there may be environmental correlates of genetic variation there is not necessarily a geographic or a spatial population differentiation.

Several problems complicate the detection of selection. First, the effects of demography (e.g. population growth or bottlenecks) may mimic selection (Simonsen et al., 1995; Nielsen et al., 2007). Tests of selection also are sensitive to variation in recombination rates of different genomic regions. These confounding effects can be disentangled using genomic information to some extent at least (Jensen et al., 2005; Williamson et al., 2007; Nielsen et al., 2009c). There is also the problem of ascertainment bias described above (Kreitman and Di Rienzo, 2004; Nielsen, 2005b,a). This may be hard to correct for (Nielsen, 2005a; Bradbury et al., 2011).
1.3 Cod-fish

Cod-fish belong to the subfamily Gadinae of the Gadidae family of the class Teleostei (see for example Teletchea et al., 2006). They are a group of benthopelagic fishes inhabiting cold water, distributed throughout the northern hemisphere. They occupy coastal zones, continental shelves and slopes of the sea (Teletchea et al., 2006). Further morphological classifications proposed by Howes (1991) divided the Gadidae family into subgroups; Gadinae, Eleginae, Gadiculinae and Microgadinae. Gadinae constitute four genera; Gadus, Theragra, Boreogadus and Micromesistius. Later Teletchea et al. (2006) established a phylogeny using genetic data on mtDNA and Arctogadus was classified belonging to this group.

In my study the main focus, of which I have the largest data set, was Atlantic cod. Other species studied belong to a genus in the subfamily of Gadinae i.e. the closest relatives of Atlantic cod. In addition to Atlantic cod they are the following species: Pacific cod, Gadus macrocephalus; Greenland cod, Gadus ogac; walleye pollock, Gadus chalcogrammus; Arctic cod, Boreogadus saida, and Polar cod, Arctogadus glacialis. For clarification it is worth mentioning that there are different conventions in common names of Boreogadus saida and Arctogadus glacialis. In some instances (often used in Europe) the common names are reversed, i.e. Boreogadus saida is called Polar cod and Arctogadus glacialis is called Arctic cod. Here the former naming convention will be followed.

I shall now describe salient features of the biology of these taxa that are relevant for the questions addressed in this thesis.

1.3.1 Atlantic cod — *Gadus morhua*

Atlantic cod is the most harvested fish stock in the North Atlantic. It has been exploited by fishermen from Europe and America since the Viking age but extensively since around 1400 (Kurlansky, 1997). During the last century the fishing vessels and fishing gear have become more advanced which has led to heavy exploitation and severe declines in population sizes from the 1970 onwards (Caddy and Cochrane, 2001) and even to collapses of major fisheries (Hutchings, 2000). The fisheries have been shown to have implications for life history traits in Atlantic cod (e.g. Árnason et al., 2009; Pardoe et al., 2009) and fisheries can act as selective agents i.e. fisheries induced selection (Árnason et al., 2009; Jakobsdóttir et al., 2011). Or, as stated by Árnason et al. (2009), man became a “techno beast” in the interplay between prey and predators, fish and the fishing industry.

Atlantic cod is widely distributed throughout the North Atlantic. It lives from the shoreline down to about 600 m depth, occupying the continental shelf and banks. It shows both stationary and migratory behaviour. It is extremely fecund, each female is capable of spawning millions of eggs per season. It can reach up to two meters in length and live up to 25 years (Salvanes et al., 2004). Cod are continuously growing and increase the proportion of their body mass devoted to reproduction with age. In short, their age-specific reproduction (\(m_x\) schedule), increases with age (Marteinsdottir and Begg, 2002).

In Iceland the main spawning areas of cod are on the southwest and west coasts. Some minor spawning areas are in the fjords on the west, north and the east coasts. After fertilization the larvae drift with water currents to settling areas west, north and
east of the island (Jonsson and Valdimarsson, 2005). In some years they may drift all the way to Greenland (Begg and Marteinsdottir, 2000). The spawning population at the main spawning ground consists of cod from the north and the east coast of Iceland and in some years cod from Greenland (Jamieson and Jónsson, 1971). A question arises whether cod from Greenland are Icelandic cod “returning” to Iceland or Greenland cod “exploring” new terrain.

Population structure in Atlantic cod has been extensively studied and population substructure was described early on. In their paper on the Greenland component of spawning cod at Iceland, Jamieson and Jónsson (1971) found extensive polymorphism in hemoglobin. They found large differences between simultaneous samples of neighboring localities as well as between temporal samples from the same locality. They explained the spawning stock as a moving mosaic of genetic isolates. Møller (1968) described genetic diversity among Atlantic cod along the Norwegian coast. The results similarly were thought to demonstrate subpopulations among Atlantic cod in Norway. Jónsson (1996) carried out tagging experiment of cod around Iceland and showed that fish from the North moved to the spawning ground in the South. Most of the mature spawners migrated post spawning to the feeding ground north-west and north-east of Iceland. The data also are thought to show evidence of both sedentary and migratory behaviour. For example Pálsson and Thorsteinsson (2003) tagging experiments using data storage tags, gave much more accurate data on the migration and behaviour of the fish. Their research suggested two foraging strategies, two ecotypes, among cod at the spawning ground south of Iceland. The two distinct ecotypes have since been shown to differ in behaviour and to some extent in spawning sites (Grabowski et al., 2011). The data storage tags show different patterns in feeding behaviour and migration, a conclusion drawn from temperature and depth records from the data (Pálsson and Thorsteinsson, 2003; Pampoulie et al., 2007; Thorsteinsson et al., 2012). The behavioural ecotypes are coastal, stationary fish, occupying shallow water and frontal, offshore migratory fish, occupying colder deep water. The temperature and depth data show that the typical frontal type migrates down to 600 meters and forages at temperature fronts. The typical shallow water coastal type stays above 100 meters during feeding period (Thorsteinsson et al., 2012). The genetic differences between the two ecotypes has been studied and some loci under selection have been shown to differentiate between them. In particular the Pan I locus, initially characterized by Pogson (2001), has been extensively used (Pampoulie et al., 2006, 2007; Hemmer-Hansen et al., 2013) and results of that locus have been interpreted as evidence for subpopulation structure. Árnason et al. (2009) showed how the polymorphism found at this locus clearly is related to depth and argued the pattern found is because of habitat selection rather than historical population genetic structure. Likewise Karlsson and Mork (2003) showed the variation found at Pan I in Trondheimsfjord in Norway to be maintained through complex forms of natural selection. The continuous clines connected with depth in the polymorphism is also seen as a cline in behaviour. The clear coastal and frontal types shown for example in Thorsteinsson et al. (2012) are the extreme behavioural types, however, intermediates are found. Pampoulie et al. (2006) showed the typical behavioural types to be closely tied to the two alternative homozygotes of Pan I implying genetically differentiated coastal and frontal subpopulations. The heterozygotes and the Pan I homozygotes showing opposite or intermediate behaviour have not been explained under this hypothesis. In Norway, the migratory North-east Arctic population and the stationary coastal Norwegian population are a comparable
system to what is found in Icelandic waters. Hemmer-Hansen et al. (2013) found similar differences between the two ecotypes in the two environments using polymorphisms from expressed sequenced SNPs in genes under selection. The main objection to regard the two forms as being genetically isolated is the lack of evidence in neutral loci (Eiríksson and Árnason, 2013; Karlsen et al., 2013).

Other proposed subpopulations of Atlantic cod are for example found in the western Atlantic and Canadian waters, in the North Sea, Skagerak, the western and eastern Baltic Sea and in the Faroe Islands Bank and the Faroe Island plateau (Therkildsen et al., 2010; Stroganov, 2015; Hüsey et al., 2016). The various supposed subpopulations are adapted to different environments, such as gradients of salinity, depth and temperature. These different responses make Atlantic cod a very attractive organism to study evolution, natural selection and speciation.

The Atlantic cod is a top predator in its environment. Its enormous fecundity and plasticity enables the species to sustain the biggest fisheries in the North Atlantic. The stock collapse in the western Atlantic has been regarded as a consequence of over fishing. There is almost no recovery of the population in spite of a moratorium on fishing (Hutchings, 2000) which presents an enigma. Changes in the ocean are happening with unpredictable consequences for both the ecosystem and the fisheries (Gewin, 2015). Recent studies suggest the rapid warming of the sea to be the main cause of collapse in the Gulf of Maine in western Atlantic (Pershing et al., 2015). Possibly cod did not have time to adapt to the rapid changes in temperature. Another possible explanation is that ecotypes or subspecies that sustained the fisheries have been overexploited to extinction.

1.3.2 Pacific and Arctic species

With the climate warming and melting of ice in the northern hemisphere, species interchanges between the North Pacific and North Atlantic are predicted (Wisz et al., 2015). It is therefore of interest to study related species in the Arctic and Pacific oceans to understand and predict what secondary contact of these species might possibly bring about. Figure 3.4 shows sampling sites of the species in this study.

**Pacific cod,** *Gadus macrocephalus* is a transoceanic fish in the Pacific. It ranges from California to the Bering Strait and the Yellow Sea and migrates between Bering Sea and the Gulf of Alaska where it is common. It is highly fecund. It is demersal and its habitat ranges from shoreline down to about 500 m. The fishery management is divided between the Bering Sea/Aleutian Islands region and the Gulf of Alaska region. However, there is scant data showing the genetic differences between those entities. Pacific cod is considered a speciation from an Atlantic cod invasion into the Pacific at approximately 4 million years ago (mya) based on genomic mtDNA data (Cohen et al., 1990; Coulson et al., 2006; Carr et al., 1999; Pogson and Mesa, 2004). **Greenland cod,** *Gadus ogac* is regarded as a recent re-invasion of Pacific cod into the Arctic and Atlantic oceans. (Coulson et al., 2006; Carr et al., 1999; Pogson and Mesa, 2004). It ranges from Alaska to Hudson Bay to west Greenland and southwards along the Canadian coast to the Gulf of St. Lawrence. Greenland cod is bottom dwelling and stays close to the shoreline. It is smaller than both Pacific and Atlantic cod (Cohen et al., 1990).

**Walleye pollock,** formerly *Theragra chalcogramma*, is now classified within the Gadidae family as *Gadus chalcogrammus*. The commercial fishery started around 1950.
with great increases in 1970 when it was discovered suitable for industrialization of surimi production. Today it is one of the most valuable fisheries in the US. Bailey wrote its story and named the book “The billion dollar fish” (Bailey, 2013).

Walleye pollack is a semipelagic, semidemersal schooling fish widely distributed in the North Pacific Ocean, playing the largest ecological role in the Bering Sea ecosystem. The north Pacific groundfish stock assessments divide the Bering Sea stock into three substocks mainly for management purposes. Largest concentration of walleye pollock is found in the eastern Bering Sea stocks. Other concentration are the Aleutian Islands and Bogoslof Island stocks. The pollock stock found in the Gulf of Alaska is thought to be independent of the Bering sea stock and is managed as a single stock. The largest spawning concentrations are in the southeast, north of Unimak Pass in the eastern Bering Sea, in Shelikof Strait, and in the Shumagin Islands in the Gulf of Alaska (Cohen et al., 1990).

Based on mitochondrial data, it was concluded that walleye pollock is a speciation from an Atlantic cod invasion into the Pacific at 3.8 mya (Coulson et al., 2006; Carr et al., 1999; Pogson and Mesa, 2004).

**Arctic cod, Boreogadus saida,** is a small circumpolar fish in the Arctic. Its range in the North Atlantic is in the White Sea, Iceland, and southern Greenland into the Miramichi River, New Brunswick in Canada. In the North Pacific it is found in Bering Sea to Cape Olyutorski, the Pribilof Islands, and Bristol Bay. It is cryopelagic or epontic (living near sea ice), brackish, demersal and lives from the surface down to 400 m. It reaches a maximum of 40 cm in length and is commonly around 25 cm. It has a lifetime of about 7 years. The spawning seasons differ between habitats, from November in the Beaufort Sea to February in the White Sea. It only spawns once in its lifetime and female fecundity is 11,900 eggs on average. This small fish has limited commercial value because it is small and the flesh is of low quality (Cohen et al., 1990).

**Polar cod, Arctogadus glacialis.** This is the least studied species in the study. It is an Arctic species found in the western half of Canadian Arctic coast, Arctic coasts of Siberia and off northern and southern coast of Greenland but it is hardly found south of the Polar circle (Cohen et al., 1990; Jordan et al., 2003). It is a pelagic fish found around packed ice at sea and at the edge of the continental shelf. It’s commercial use is mainly for production of fish meal and oil.

### 1.3.3 Summary

The different fish taxa in this study differ in their geographic range, size, and spawning habits. The small circumpolar Arctic cod overlaps with the northern region of the other species range. Adapted to cold area it is an r-selected species (Pianka, 1970) with life history traits of early maturation, small size, short life span and high offspring number in only one spawning event in its lifetime. The Atlantic cod is a long lived species with type-III survivorship and extremely high fecundity. In our paper V (Halldórsdóttir and Árnason, 2015b) we discuss that these two fish species have hybridized and produced a new homoploid species, namely the walleye pollock. Interestingly, our results may explain why trained fisheries survey people have difficulties distinguishing between Arctic cod and walleye pollock at certain ages (Short et al., 2006). The Pacific cod is somewhat smaller than the Atlantic cod and not as abundant. The Greenland cod, is about half the size of the Pacific cod and has a shorter lifespan and is only found in
coastal regions. The extent to which characteristics of these taxa reflect phenotypic plasticity and to what extent they will be explicable by admixture between the various taxa remains to be studied. Expanding on these questions will be one aspect of future work.
2 Objectives and methodology

The main objectives of this thesis are:

• To provide some answers to critical questions on the role of natural selection and local adaptation in forming population structure in Atlantic cod.

• To apply thinking, and techniques of population genetics and genomics including new extensions of coalescence theory to detect fingerprints of natural selection in molecular data in non-model high-fecundity organisms.

• To compare genes and genomes of closely related species to understand evolution and speciation among populations of closely related fish taxa and to disentangle the units which selection is acting upon.

• To study admixture and potential homoploid hybrid speciation in these taxa.

To achieve my goal I have approached these objectives using different methods of the rapidly evolving molecular techniques used in evolutionary biology. The study is in a way a story of the large improvements in methods of data generation and the increased possibilities which the field of population genetics and genomics has thereby gained over the last few years. In what follows in this introduction I shall discuss the results according to the different methodologies used in approaching the subject. I will add details to the theoretical underpinnings and the questions posed in the various papers. I expand on explanations that do not appear in the published papers. Detailed technical information of the specific methods can be found in the method chapter of each paper. Finally, I provide a summary of the main results from each chapter.
3 Results and discussion

3.1 Candidate genes approach — (Papers I-IV)

A basic tenet of population genetics is that breeding structure on average affects all genes and all alleles in the genome in the same way (Wright, 1931; Lewontin and Krakauer, 1973). Therefore, effects of random genetic drift, historical bottlenecks of population numbers, gene flow, and inbreeding or outbreeding will be genome-wide. In contrast the effects of natural selection, mutation, assortative mating, and recombination will be locus-specific or limited to specific genomic or chromosomal regions (Lewontin and Krakauer, 1973; Black et al., 2001). The distinction between locus-specific and genome-wide effects is crucial in population genomics and genetics. When considering population structure we can look at polymorphism at neutral loci for inference about gene flow between different populations. Locus-specific effects, however, are important for inferences about fitness and adaptation, the physiological and ecological functioning of the organism in its environment. Therefore, to study natural selection and adaptation, it is important to find genes showing evidence of selection.

3.1.1 The hemoglobin loci of Atlantic cod

Sick (1965) described a system consisting of two major zones of the hemoglobin proteins of Atlantic cod, HbI and HbII, by agar gel electrophoresis of whole blood. The HbI zone showed variation interpreted genetically as a polymorphism with a pair of co-dominant alleles giving rise to $HbI-1/1$ ($SS$, slow moving), $HbI-2/2$ ($FF$, fast moving) homozygotes and $HbI-1/2$ ($FS$) heterozygotes. Allozyme research, including the $HbI$ locus of hemoglobin, indicated heterogeneity of cod populations in the North Atlantic (Mork et al., 1985).

Frydenberg et al. (1965) showed an apparent cline in $HbI$ allele frequency along the Norwegian coast (Figure 3.1). Clines are often indicative of natural selection and gene flow. If the locus is under selection it would limit the use of the locus for studies on population structure because the observed patterns presumably are a response to specific environmental factors. To study historical population structure neutral variation would be more appropriate. However, a locus showing clear selective effects is of interest for studies of adaptation to various environmental conditions.
The HbI locus was the first locus used for inference about population structure and selection in Atlantic cod. At the beginning of my thesis work, the locus had not been described at the DNA level. It was therefore a first step to find and to characterize the genes responsible for these patterns. In paper I one set of linked $\alpha/\beta$ genes was described (Halldórsdóttir and Árnason, 2009a). Andersen et al. (2009) and Borza et al. (2009) then described the molecular structure of the HbI system as a two amino-acid replacements, the Met55Val and Lys62Ala in the $\beta_1$ gene. The haplotypes Met55/Lys62 and Val55/Ala62 were found to be associated with the HbI-1 and HbI-2 variants determined by agarose gel electrophoresis. A third non-synonymous substitution was only found polymorphic in the HbI-2/2 genotype (Val55/Ala62) at position 123, Leu123Met. Borza et al. (2009) also described nine distinct Hb transcripts or genes, four of which corresponded to the $\alpha$ Hb gene family and five to the $\beta$ Hb gene family. Borza et al. (2010) mapped the nine globin genes onto two linkage groups: $\alpha_1$, $\alpha_4$, $\beta_1$ and $\beta_5$ on linkage group 2 (LG02) and $\alpha_2$, $\alpha_3$, $\beta_2$, $\beta_3$ and $\beta_4$ on linkage group 18 (LG18) respectively. Wetten et al. (2010) further determined the genomic organization of the hemoglobin loci which turned out to be congruent with my own unpublished results. That is, the organization of the four globin genes in the 7kb region of the HbI containing linkage group was $\alpha_4$, $\beta_1$, $\alpha_1$, and $\beta_5$ with $\alpha_1$ oriented in the opposite transcriptional direction to the other genes (see Figure 3.2 based on a poster that I presented at a RVON symposium in 2010).
3.1 Candidate genes approach — (Papers I-IV)

FGENESH 2.6 Prediction of potential genes in Fish genomic DNA
Seq name: eanyttogeascaffold00026oge6-c26-c6-scaffold00146.fasta
Length of sequence: 16971
Number of predicted genes 5: in +chain 2, in -chain 3.
Number of predicted exons 22: in +chain 12, in -chain 10.

Positions of predicted genes and exons: Variant 1 from 1, Score:258.725586

<table>
<thead>
<tr>
<th></th>
<th>CDSf</th>
<th>CDSi</th>
<th>CDSl</th>
<th>CDSo</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>PolA</th>
<th>TSS</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 3.2. The sequence of the linked β1 and α1 globin genes of the HbI locus together with neighboring globin genes. 1) α4, 2) β1, 3) α1 and 4) β5 (gene naming as in Borza et al., 2009). This is Figure 2 from my poster at RVON symposium 2010. Unpublished data congruent with the findings in Wetten et al. (2010).

Hubert et al. (2010) made a 1641 gene-associated SNP resource from EST sequences based on western Atlantic cod and made a genetic linkage map of Atlantic cod. Signal of selection in linkage group LG02, where the HbI locus resides, was not associated with the migratory ecotype as had been suggested by Star et al. (2011). Using a subset of these loci Bradbury et al. (2010, 2014) found parallel evolution of a temperature-associated cline in both western and eastern Atlantic on linkage groups LG02, LG07 and LG12. The loci showing clinal variation were neither linked to the HbI polymorphism nor to the Pan I polymorphism. Thus the clinal differentiation that the allozyme research on this locus had showed (i.e. Frydenberg et al., 1965) was not explained by the DNA work.

The number and orientation of the α and β globin gene sets differ between fish species (Gillemans et al., 2003; Wagner et al., 1994). Given my results detailed in paper I it was of interest to estimate the number of globin gene sets in the genome. The method that I used in paper II was in a way a genomic approach in a non-model organism before the publication of the sequence of the genome (Halldórsdóttir and Árnason, 2009b). The strategy was to cast the net widely and search for different globin genes within individuals of each of the HbI genotypes. I fished for globin genes using a PCR based strategy. Several clones of PCR products of individuals previously genotyped for the known allozyme variation were cloned and sequenced.
3 Results and discussion

3.1.1.1 Synopsis

The main finding was a characterization of globin gene structure of Atlantic cod which at that time was not known. A hemoglobin locus was the first and main locus from the allozyme era previously used for population inferences and studies of selection in cod. By using PCR, cloning and screening of a λ library of Atlantic cod I described how an α globin and a β globin gene were linked on the same chromosome with a putative intergene in the intergenic region between them. This is described in Paper I.

I then used the data from paper I to estimate how many sets of paired α/β globin loci there might be in the genome of Atlantic cod. In Paper II, I used a PCR based strategy, of genomic exploration, using a strict criterion of phylogenetically informative sites to exclude PCR errors. The approach is a mixture of a PCR based strategy and population genetic and molecular evolution analysis for genomic exploration to inquire into the number of α and β linked globin gene sets and about potential allelic variation of these genes in Atlantic cod. Ten clusters were described, different paralogous genes and potentially some of them allelic variation of the same orthologous locus.

3.1.2 Innate immunity genes — Cathelicidin genes

Star et al. (2011) published the genome sequence of Atlantic cod and fundamentally advanced the field of cod genetics and genomics. They described a lack of genes encoding for MHC-II and CD4, important parts of the adaptive immunity in vertebrates. Vertebrates have both adaptive and innate immunity to fight against microbial infections. The mechanisms of the two system are different. The adaptive immunity produces antibodies which specifically recognize and bind to particular antigens. In contrast, the innate immunity consist of molecules with less specialized functions i.e. pathogen recognition receptors which bind to patterns found in pathogens and not in the host. Innate immunity also is based on antimicrobial peptides which attack the cells of the pathogens and disrupt their cell membrane.

The MHC-I and MHC-II molecules are cell surface glycoproteins that deliver peptides from different cellular compartments to the surface of an infected cell. MHC-I bind cytosolic peptides from viruses and some bacteria that replicate in the cytosol. The MHC-I molecules are recognized by CD8 T cells which kill the infected cell. MHC-II bind peptides from vesicles, e.g. some bacteria and parasites engulfed by phagocytic cells which are able to proliferate within the endocytic vesicles. The MHC-II bind their peptides and are recognized by CD4 T_H1 cells which activate macrophages to kill the intravesicular pathogens. Proteins from extracellular pathogens can enter the vesicular system by endocytosis. MHC-II bind these proteins and are recognized by CD4 T_H2 cells which stimulate B cells to produce antibody (Murphy et al., 2007). Given the loss of these genes the poor antibody response of Atlantic cod (Pilström et al., 2005; Magnadottir, 2010) is not surprising and research is ongoing to understand how the cod compensates for this loss (Magnadottir, 2014).

There are several gene families that function in innate immunity (Kimbrell and Beutler, 2001; Alberts et al., 2002). As was shown to be the case with the MHC I genes, some of them have highly expanded gene numbers in Atlantic cod, e.g. the toll-like receptors (TLR) gene family (Star et al., 2011; Sundaram et al., 2012). TLR are non-catalytic transmembrane molecules with broad pathogen recognition receptors which
bind to pathogen-associated molecular patterns or PAMPs. On binding they initiate intra-cellular signaling cascades to activate immune response. In some instances they induce transcription of antimicrobial peptides (Hansson, 2005; Blasius and Beutler, 2010). Interestingly, composition of the Atlantic cod TLR is unique, with expanded teleost specific TLR but lacking most of the bacterial recognizing mammalian TLR (Sundaram et al., 2012). This may be indicative of duplicated genes gaining new function in pathogen recognition compensating for the specialized cell surface molecules. Antimicrobial peptides are an important mechanism of the innate immunity (Ganz, 2003; Tomasinsig and Zanetti, 2005; Ruangsri et al., 2012). These small molecules have different structures which in combination with the amphipathicity and cationic charge produce diversity for distinct functions and antimicrobial activity. The Cathelicidin family is one of them. The Cathelicidins are synthesized as inactive precursor forms. The active part activates after cleavage from the precursor with appropriate stimuli. The inactive part, known as the N-terminus of the protein, has evolutionarily conserved parts while the C-terminus which cleaves off and becomes the active form is highly divergent (Tomasinsig and Zanetti, 2005).

High-frequency polymorphisms in immunity genes are frequently detected when genomes are scanned for selection (Nielsen et al., 2007; Leffler et al., 2013). Cathelicidins in cod had been characterized by Maier et al. (2008) and they were an interesting system to study evolutionarily on a population level. This was the subject of paper III (Halldórsdóttir and Árnason, 2015a)

A critique of our paper III by an anonymous reviewer of the manuscript (Reviewer-2, 2015), raised the important question whether the Cathelicidin genes identified were orthologous or paralogous genes. The reviewer also raised question about the interpretation of trans-species polymorphism indicative of balancing selection. Here is some discussion on these topics.

Orthologous genes are genes that diverge after a speciation event. They have the same origin but segregate differently from the origin after the split of the species. They are truly homologous genes. Paralogous genes, on the other hand, originated with duplication of genes in a genome independent of speciation events. Different paralogous genes will be passed on to descendants but some may be lost in different lineages complicating analysis of genes in gene families. If duplication events happens before a lineage splits into separate species, such as the α and β globin genes in mammals, then each of the paralogous genes are orthologous between the daughter lineages. When looking at those genes between species, the β genes will cluster together and the α genes as well. This does not represent a trans-species polymorphisms. However, if alleles among for example β genes are shared between species, such that alleles from an orthologous β gene in two species are more similar to each other than are alleles of the same β gene within either of the species, it would represent a case of trans-species polymorphism.

Balancing selection is a potent force actively maintaining variation in a population (Hedrick, 2005). It occurs when natural selection prevents variants from going either extinct or to fixation. Heterozygote advantage and negative frequency dependent selection are important mechanisms of balancing selection. A classical example of the former phenomenon is the molecular background of Sickle-cell anemia in humans. A non-synonymous transition of A to T at the sixth amino acid codon in the β globin gene, results in substitution of glutamic-acid to valin, causing abnormal shape of the red blood cells that lead to less oxygen uptake and lowering the viability of the affected
Results and discussion

It is a recessive deleterious syndrome that is nevertheless frequent in populations from areas where malaria is common. The heterozygotes are more tolerant against malaria than normal homozygotes so there is some antagonistic pleiotropy between the two components of fitness of the individual. The result is heterozygous advantage maintaining the recessive allele in the population in malaria areas. Even if heterozygous advantage, the selectionists warhorse of balancing selection (Lewontin, 1974), may be rare, negative frequency dependent selection remains a potent force of balancing selection.

The molecular signatures of balancing selection can be hard to detect and distinguish from other types of selection (Quintana-Murci and Clark, 2013). The size of the genomic region around balanced polymorphism with its associated neutral variation can be relatively short. The rate of recombination affects our ability to detect balancing selection. For example if recombination is suppressed and if multiple sites are under balancing selection it will be easier to detect (Wiuf et al., 2004). A most important evidence for balancing selection is a trans-species polymorphism (Charlesworth, 2006). Trans-species polymorphisms of any type are relatively unusual. The best known cases have been previously reported in the MHC genes (Fan et al., 1989; Nei and Hughes, 1991) and in the disease resistance R genes in Arabidopsis (Bakker et al., 2006).

The innate immune system is at the forefront of the interactions between hosts and infective agents. It serves as a first defense, acting immediately and non-specifically, while the adaptive immune system has some lag time between exposure and reaction. The innate immune system is evolutionarily older, found in all metazoa but the adaptive immune system is only found in vertebrates (Murphy et al., 2007). The interactions of host and infective agents makes innate immunity genes good candidates to study selection. It is of particular interest from an evolutionary perspective in an organism lacking the adaptive immunity where the innate immunity genes seem to have expanded drastically.

3.1.2.1 Synopsis

In Paper III we describe extensive polymorphism in the Cathelicidin genes of Atlantic cod and related species (Halldórsson and Árnason, 2015a). This study addressed what evolutionary forces are shaping the trans-species polymorphisms and the problem of discriminating between orthologous and paralogous genes. The orthologous variation was trans-species among the different paralogs. This fact was the main reason for interpreting the results as evidence for balancing selection. With comparison to closely related taxa from the Pacific and the Arctic we show extensive nucleotide and amino acid polymorphism in the Cathelicidin gene family in cod-fish. There were three major clusters of variants which we called Cath1, Cath2, and Cath3. The variation clustered by alleles and not by species in phylogenetic trees and in discriminant analysis of principal components, DAPC (Jombart et al., 2010). Variation within the three groups showed trans-species polymorphism that was older than speciation and these patterns are suggestive of balancing selection maintaining the variation.

3.1.3 Screening for candidate genes in a genomic BAC-Library

A dense genomic map of genetic variation in humans (and in model organisms) allows us to scan the genome for signatures of natural selection (Voight et al., 2006; Sabeti
et al., 2007; Storz, 2005). Asking what percentage of the human genome shows footprints of selection depends on the density of the maps and sensitivity of the various methods used (Voight et al., 2006; Sabeti et al., 2007; Storz, 2005). It is safe to say that only a small percentage of single nucleotide polymorphisms (SNPs) show footprints of selection in humans. For microsatellite loci, 2% (13/624) were detected as $F_{ST}$ outliers when African and non-African human populations were compared (Storz et al., 2004). In contrast in Atlantic cod 11% (26 out of 235, Moen et al., 2008) and 4% (70 out of 1641 Bradbury et al., 2010) of independent SNP loci are flagged as outliers undergoing selection and one fourth of microsatellite loci (Nielsen et al., 2006), supporting the thesis that a considerable fraction of the genome of Atlantic cod is simultaneously under selection for different adaptations (Árnason, 2004; Árnason and Halldórsdóttir, 2015; Halldórsdóttir and Árnason, 2015a).

A next step in my study was to find other candidate genes under selection. To facilitate the analysis I wanted to sequence longer regions around these genes and to get phased sequencing reads. To meet these objectives a genomic BAC library of the Atlantic cod was prepared. To identify BAC clones we probed the library using short sequences around SNPs and genes of interest. The candidate genes mostly were outlier loci from Moen et al. (2008). After screening the positive clones were sequenced using 454 sequencing at a sequencing company. The sequences of scaffolds from the 454 assembly were used to design primers for PCR of large fragments around the selected sites.

We screened for several outlier genes found by Moen et al. (2008). The outlier detection approach used by Moen et al. (2008) was based on Beaumont (2005). Briefly, Moen et al. (2008) sampled several individuals from four locations in Norway. They genotyped several hundred SNPs and calculated $F_{ST}$ among the populations. The mean $F_{ST}$ was used in computer simulations (Beaumont, 2005) calculating the distribution of $F_{ST}$ under the assumption of neutrality and using the island model of evolution (Wright, 1943). The observed $F_{ST}$ for each locus was compared to this neutral distribution of $F_{ST}$ to evaluate if all variation is neutral and the main evolutionary force acting was genetic drift. If an $F_{ST}$ for a locus is an outlier in this distribution we can infer that some forces are acting differentially on that locus compared to the rest. Either an outlier locus is under selection and all the other loci are neutral or vice versa, that the outlier locus is neutral and the rest is under selection (Lewontin and Krakauer, 1973). For the latter scenario we would deduce that the populations were small because the effects of genetic drift is higher in small populations. However, usually the outlier loci are taken to be under selection.
Because of indel variation there is in general a difficulty in directly sequencing PCR products from a diploid organism. The PCR products were therefore TOPO-TA cloned and several clones from each individual were then sequenced. However, this creates a problem of how to deal with PCR errors in cloned products. The solution was to sequence three clones from each individual. A consensus of the three gives the sequence of one of the alleles from an individual. This is because two of the three clones will be the same allele; the third may or may not be the same allele. The third allele will provide a consensus to screen out errors in the two other clones. The strategy is thoroughly explained in paper IV.

An extensive collection of Atlantic cod samples from throughout the distribution are available in the laboratory and the approach was to take a blind sample from that collection stratified so as to sample from throughout the geographical distribution. We sampled about 100 individuals from around Iceland using METACOD divisions (Pampoulie et al., 2006, Figure 3.3) and 300 individuals from the distribution range of Atlantic cod and of several closely related species from the Arctic and the Pacific Oceans (Figure 3.4).
Three SNPs which had the highest $F_{ST}$ in the study of Moen et al. (2008), with extreme differentiation of 0.83, 0.82 and 0.83 (Figure 3.5) turned out to be a part of the $Ckma$ (Creatine Kinase Muscle A) gene. Creatine kinases are important in bioenergetic processes in cells. Creatine kinase generates Phosphocreatine (PCr) at the sites of ATP production in glycolysis and oxidative phosphorilation in mitochondria and regenerates ATP from PCr at subcellular sites of ATP use by ATPases (Wallimann et al., 1992, 2011). The physiological advantage is to provide a spatial and temporal energy buffer storing and releasing energy in and from PCr. The rate of intracellular diffusion of both Creatine (Cr) and PCr is one and three orders of magnitude faster than diffusion of ATP and ADP respectively (Wallimann et al., 1992, 2011). Thus, it is an important feature of energy use for Atlantic cod, a predator which has been shown to travel several hundred meters up and down the water column as well as long distances geographically.

The $Ckma$ gene is an example of a candidate gene with a well known function. It is a gene showing extreme spatial differentiation. The results of a population genetic study on $Ckma$ are described in paper IV. The results summary of paper IV is presented after discussion of results about the appropriate coalescent theory which also was a part of paper IV.
3 Results and discussion

Figure 3.5. Outlier detection. Figure 4 from Moen et al. (2008). The most extreme outliers were found in the Ckma gene (Creatine Kinase Muscle A) which we studied in paper IV. (Moen et al. paper published under CC BY 4.0 license).

3.2 Coalescence approach - (Paper IV)

Negative, positive and balancing selection will affect levels of variation within and among species. A wealth of statistical methods exist for detecting signatures of selection from molecular data. Many population genetics methods, with neutrality as a null model, use a comparative approach within and among species (Hudson et al., 1987). An example is the MK test proposed by McDonald and Kreitman (1991) of neutral protein evolution that asks if the ratio of non-synonymous to synonymous mutation is equivalent for fixed differences between species and polymorphisms within species. The Hudson, Kreitman, and Aguadé (HKA) (Hudson et al., 1987) test, uses intraspecific and interspecific nucleotide data to test the null hypothesis that the ratio of polymorphism to divergence is equivalent among loci. The HKA test rejects a neutral model if the ratio of polymorphism to divergence differs significantly among independent loci. Hughes and Nei (Hughes and Nei, 1988), Tajima (Tajima, 1989), Fu and Li (Fu and Li, 1993) and others also have proposed various tests based on allele frequency spectra. A good review of these statistical tests can be found in Nielsen (2005a). The ratio of non-synonymous to synonymous mutations ($d_N/d_S$) is a proxy for both negative and positive selection (Nielsen, 2005a). Intraspecific studies are required for detecting recent and ongoing selection (Nielsen et al., 2007). A positively selected mutation sweeping to fixation (Maynard Smith and Haigh, 1974) carries other variants along and leaves a mark on a population allele frequency spectrum. These can for example be detected as a decrease in Tajima’s $D$ (Tajima, 1989) and the number of segregating sites and an increase in linkage disequilibrium (Nielsen, 2005a; Nielsen et al.,
Selection can also be detected through its effects on population differentiation: local adaptation leads to permanent differentiation of populations while global adaptation may temporarily produce differentiation as a mutation sweeps to fixation throughout a species range (Nielsen et al., 2007). Traditionally the null model used was the Fisher-Wright model (Fisher, 1930; Wright, 1931) which is based on prediction forward in time, i.e. how allele frequencies change with time.

Population genetics, beginning as a theoretical discipline, has been transformed with the abundance of modern data and has become more data oriented. As such it has laid the foundation for modern genomic analyses (Tellier and Lemaire, 2014). Coalescent theory (Kingman, 1982b,a) represents a fundamental shift in testing hypothesis away from progressive models, looking forward in time, to coalescent models which generate expected times back to a common ancestor of a set of sequences, that is the gene genealogy of a sample. Rapid nucleotide sequencing in experimental population genetics together with coalescent models in theoretical population genetics are powerful tools to detect fingerprints of natural selection on a genetic locus (Hein et al., 2004). The variation in frequency on synonymous and non-synonymous sites can be compared with expectations of a neutral model as the null hypothesis to measure the effect of positive Darwinian selection (Schaeffer and Aguadé, 2000). However, recent developments in coalescent theory (Wakeley, 2013) raise questions about what are the appropriate coalescent models in each case.

3.2.1 Site frequency spectrum

Common estimators of variability found in DNA sequence data are $S$, the number of segregating sites, and $\pi$, the average number of pairwise differences or nucleotide diversity. Both of them are useful summary statistics and can be used for estimating $\theta$ ($4N_e\mu$), the scaled effective population size. The site frequency spectrum (SFS) is another very useful summary statistic giving information on allele frequencies of all mutations in the sample. With information of ancestral state, i.e. which allele is ancestral and which are derived based on an outgroup, the unfolded SFS can be estimated. Often the ancestral state is not known from the sequence data and the folded SFS can then be applied. In folded SFS the derived and the ancestral alleles of opposite frequencies are added i.e. $f^*_j = f_j + f_{n-j}$. The expected SFS is calculated under the coalescence model as a null model. For the Kingman coalescent (Kingman, 1982b) the expected SFS frequency is $\theta \sum_{k=1}^{n-1} 1/k$ for all mutations in the tree (Nielsen and Slatkin, 2013) where $k$ is the number of lineages and $n$ is the sample size. The gene genealogies described by the Kingman coalescent are said to be bifurcating, i.e. only two alleles can find a common ancestor at any time, i.e. only two lineages can coalesce at any time (Tellier and Lemaire, 2014; Wakeley, 2013).

Fecundity and survivorship

Many marine animals have Type III survivorship with high fecundity and high early mortality (Figure: 3.6). This leads to variable recruitment to the adult populations (Hedgecock and Pudovkin, 2011). High fecundity translates into large excess reproductive capacity that would allow organisms to withstand substantial natural selection
and to bear the entailing high genetic load (Kimura, 1995). High-fecundity organisms relative to low-fecundity organisms should at any time be able to adapt a larger proportion of their genome to meet various environmental challenges. The genetic diversity found in such large marine stocks is not in concordance with the expected diversity in very large populations. This has been explained by comparison to sweepstakes-like chances matching reproductive capacity with environmental factors and contribution to next spawning population (Hedgecock and Pudovkin, 2011). However, high fecundity and type-III survivorship, coupled with sweepstakes-like reproduction leads to high variance in individual reproductive success (Williams, 1975; Hedgecock, 1994; Hedgecock and Pudovkin, 2011). Their offspring distribution is heavy-tailed (Birkner et al., 2013a).

![Figure 3.6. Type III survivorship curve. The ordinate is on a logarithmic scale.](image)

### 3.2.2 Appropriate coalescent null models

The Kingman coalescent theory is based on models of low offspring number populations and that individual genotypes will not affect the distribution of offspring numbers. Therefore, for high fecundity marine organism like the Atlantic cod, with high variance in offspring number, the Kingman coalescent may not be appropriate null model. However, recent development of gene trees and coalescent theories (Wakeley, 2009, 2013; Zhu et al., 2015) brings a wealth of methods and programs to bear on analysing molecular data of this nature both within and between populations and species (Felsenstein, 2004; Hein et al., 2004). These are models of multiple-merger coalescents (Wakeley, 2013). These models are appropriate null models of highly fecund organisms that have heavy-tailed offspring distributions (Birkner et al., 2013b; Árnason and Halldórsson, 2015).

One such model is the $\Lambda$ coalescent which differs from the Kingman coalescent in allowing multiple mergers at a time (Eldon and Wakeley, 2006; Tellier and Lemaire, 2014) instead of binary mergers at a generation. Furthermore, the $\Xi$ coalescent allows simultaneous multiple mergers at any one generation (see Figure 3.7) (Schweinsberg, 2000; Zhu et al., 2015). Generation in this context means coalescent generation, i.e. the time between two coalescent events and is not the same as the generation time of the organism studied.
3.2 Coalescence approach - (Paper IV)

In paper IV we estimated the unfolded site frequency spectra for the *Ckma* gene in Atlantic cod by using Pacific cod as an outgroup to determine the ancestral state. We compared our results to expectations of the Kingman coalescent and two different $\Lambda$ coalescent models (Figure 3.7 A and B), the $\text{Beta}(2 - \alpha, \alpha)$ (Schweinsberg, 2003) and the point mass model (Eldon and Wakeley, 2006). One reviewer of the paper (Matschiner, 2015) raised the question of why the multiple merger coalescents models were more appropriate than Kingman coalescent for empirical data sets from large geographic areas. His concern was that none of the fish from distant regions were siblings, offspring from the same high-fecundity individual. This is an interesting point which I would like to elaborate on.

The two $\Lambda$ models measure different parameters. The $\text{Beta}(2 - \alpha, \alpha)$ estimates $\alpha$ whereas the pointmass estimates $\psi$. The $\text{Beta}(2 - \alpha, \alpha)$ model yields the Kingman coalescent in the limit with $\alpha \geq 2$. The parameter $\alpha$ in $\text{Beta}(2 - \alpha, \alpha)$ is related to the probability of observing large families and the context is as follows. Under this model the probability of getting a family size with $k$ or more offspring decays like $k^{-\alpha}$. For the Kingman coalescent $\alpha \geq 2$ and for $\text{Beta}(2 - \alpha, \alpha)$, $1 \leq \alpha < 2$. The lower is the parameter $\alpha$, the greater is the chance of seeing large families. Thus
there is little chance of observing large families under the Kingman coalescent. For instance, for $k = 1000$ and $\alpha = 1$ the probability under the Beta$(2 - \alpha, \alpha)$ model of getting a family with $1000$ mature offspring is $0.001$ (one in $1000$) but for the Kingman coalescence with $\alpha = 2$ it is $0.000001$ (one in a million). Thus under a Kingman coalescent multiple merger coalescence is highly unlikely for marine species with highly skewed variance in offspring number (Hedgecock and Pudovkin, 2011). Coalescent time scales are functions of the effective population size $N_e$. Even though coalescence times are shorter under multiple merger than under the Kingman coalescent, they are still longer than a biological generation time. A theoretically extreme case in a sweepstakes-like reproduction is when all individuals are siblings and all samples coalesce in the previous generation, eliminating variation from the population except for new mutation (Árnason and Halldórsdóttir, 2015). For Atlantic cod, with highly skewed offspring number and high fecundity, the real coalescent, however, lies somewhere between the two extreme of $\alpha = 1$ and $\alpha \geq 2$. Thus we do not necessarily expect to find siblings distributed throughout the species range. In coalescent trees coalescent events are traced not distribution of individuals in a family tree.

### 3.2.3 Synopsis

In **Paper IV** we apply new methods of multiple merger $\Lambda$-coalescents to study the nature of selection. We use this method to analyse in detail nucleotide variation at the $Chma$ gene in Atlantic cod. In their study of variation among four Atlantic cod populations along the Norwegian coast, Moen et al. (2008) showed this locus to be an outlier showing an extreme differentiation of $F_{ST} = 0.83$ the highest $F_{ST}$ of any locus in their study. By screening a BAC library of Atlantic cod and having positive clones sequenced using the 454 sequencing technique, we obtained sequences around the SNPs and we used those to make primers for studying the entire gene. The $\alpha$ globin gene ($HbA2$) from the first two papers and a myoglobin gene ($Myg$) were used as neutral markers for comparison. In this paper we used different coalescents as null models. We argue that the new $\Lambda$-coalescents are appropriate null models of neutrality for high fecundity-organisms with high variance in offspring number. We detected what appears to be balancing selection on a global scale in the distribution of Atlantic cod in the eastern Atlantic.

### 3.3 Genome sequencing approach - (Paper V)

The revolution brought about by next-generation sequencing (NGS) has made easy the sequencing of whole genomes or of large regions of the genome either random or targeted. Briefly, in the most popular Illumina platform used today, the genomic DNA is fragmented into small overlapping pieces that are each sequenced multiple times (representing the coverage of an individual). The sequences are then either aligned to a reference genome or assembled *de novo*. Variable sites, e.g. SNPs, are detected among individuals in a sample and the genotype of each individual is determined. The data have to be filtered in various ways, e.g. bases with low quality scores are filtered out.

Genome wide studies in non-model organisms like Atlantic cod are limited in
comparison with studies in model organisms like humans. The results and the problems that have arisen in extensive studies of model organisms can be used to guide our exploration of variation in non-model organisms. It is important to study non-model organisms, such as the Atlantic cod, because they may show different adaptation to various environmental factors, different from what can be seen in model organisms. They may also, as already stated, experience natural selection simultaneously at a number of loci and thus be good models to study selection (Árnason and Halldórsdóttir, 2015). Furthermore, studying a marine organism such as Atlantic cod may increase our understanding of the complicated environment of the sea, an environment which we, as land crabs, may have limited insight into.

Although NGS is readily available it is still relatively expensive. This is particularly true for population level studies requiring sequences from a large number of individuals. For acceptable power of statistical inference population genetic studies require data on a number of individuals. Therefore, a compromise must be struck between the number of individuals and the budget. There are currently two ways to address this problem. The first is to sequence a random part of the genome (i.e. the approach of reduced representation, Peterson et al., 2012). The second is to sequence each individual at a lower coverage and use statistical methods to handle the increased uncertainty of the genotypes that results from lower coverage (the approach of genotype likelihood, Nielsen et al., 2011). In paper V we used the latter method for whole genome sequencing of cod-fish.

3.3.1 Partial/reduced genome sequencing approach

Hubert et al. (2010) developed a SNP resource for Atlantic cod from Canadian waters. They prepared a chip microarray-based platform from 1641 SNPs generated from ESTs (expressed sequence tags) i.e. from genes. They used the SNPs to build a linkage map of 23 linkage groups. Some of the SNPs were only polymorphic in the western Atlantic but not in the European or eastern Atlantic. The problem of ascertainment bias, already discussed, arises with these techniques. Bradbury et al. (2011) addressed this problem and evaluated the ascertainment bias in range-wide studies of Atlantic cod using the 1640 SNP panel. They observed a clear reduction in diversity in the east Atlantic in an assignment test, showing that ascertainment bias can be a real problem using this panel.

This platform and subsets of the SNPs generated by Hubert et al. (2010) have since been extensively used for studying selection and admixture among populations of Atlantic cod. Bradbury et al. (2010) described parallel clines of adaptation on both sides of the Atlantic in response to temperature among loci spanning three linkage groups. Therkildsen et al. (2013) used a subset of these loci for a spatio-temporal analysis in Greenland. They described four genetically distinct groups of cod in the waters of Greenland and Iceland. Most individuals in their study could be assigned with a high probability to a single cluster but some individuals showed equal membership probabilities to two clusters. The sample from western Atlantic was too distinct for their analysis and therefore they excluded western cod from the analysis. Bradbury et al. (2013) characterized genomic islands of elevated divergence across multiple linkage groups, indicating non random adaptive variation in the cod genome at outlier loci. In a follow up, Bradbury et al. (2014) described inter-chromosomal linkage disequilibrium between the genomic islands and also hybridization between North
and South populations primarily in the western Atlantic. They found $F_1$ individuals but very few $F_2$ or backcrossed individuals. Their interpretation was that there is an ongoing cryptic speciation with gene flow i.e. there is considerable gene flow between groups but the loci showing linkage disequilibrium across linkage groups are speciation genes which can respond quickly to selective factors and show differentiation faster than neutral loci (Bradbury et al., 2014). In paper V we refute this hypothesis of speciation with gene flow for cod-fish (see also discussion on speciation below).

Among Atlantic cod, the shallow-water coastal and deep-water migratory frontal behavioural ecotypes have been characterized and studied using storage-tags data (DST) (Pálsson and Thorsteinsson, 2003; Thorsteinsson et al., 2012). The same ecotypes have been described in Norway as Norwegian coastal cod, occupying the shallow water in the fjords along Norwegian coast, and Northeast Arctic cod found in the colder and deeper waters in the Barents sea (Nordeide et al., 2011). Using the SNPs and the platform from Hubert et al. (2010), Hemmer-Hansen et al. (2013) investigated the difference between these two ecotypes in two different environments. They found a genomic island in LG01 of strong differentiation between the two populations or ecotypes in both environments. Their interpretation is that the genomic island may represent genomic signatures underlying ecological divergence in the speciation process from panmixia to reproductive isolation in a species with high gene flow. They appear to be referring to a model of speciation-with-gene-flow (Turner et al., 2005; Nosil, 2008).

In another study of the two behavioural ecotypes Karlsen et al. (2013) examined pooled DNA samples from each of two subgroups (coastal cod and North East Arctic cod) in Norway. This method is more affordable than whole genome sequencing of individuals but variation among individuals is lost from the analysis and only mean summary statistics can be estimated. In their study Karlsen et al. (2013) suggest a high degree of reproductive isolation between the ecotypes. However, individual genotypes cannot be determined from pooled data limiting the usefulness of such data.

The well known Pantophysin I or Pan I locus characterized by Pogson (2001) has been linked to the different behavioural ecotypes (Pampoulie et al., 2006, 2007). Furthermore the relationship of Pan I to depth and habitat selection and the consequences of intensive fisheries induced selection in specific habitats has been reported (Árnason et al., 2009). Interestingly Pan I resides in the genomic island in LG01 (Borza et al., 2010; Hernandez and Árnason, 2015). Therefore we use Pan I as a proxy for the behavioural ecotypes in our analysis in paper V.

### 3.3.1.1 RAD-sequencing

When taking the approach of reducing the genomic fraction sequenced from each individual in high throughput sequencing it is crucial for population studies to ensure that the same fragments are investigated among the various individuals. One solution to this is to use restriction endonuclease enzymes to cut the DNA at specific sites and build libraries of regions around the restriction sites. Originally these methods relied on access to a reference genome for otherwise much of the sequenced data would have to be discarded. Peterson et al. (2012) further developed this technique by using two restriction enzymes which then allowed more precise choice of the size of the fragments to use in the library. The precise size selection maximizes the number of the same genomic fragments that were sequenced in a very large number of individuals.
The method does not rely on genomic knowledge or reference genome and thus is appropriate for non-model organisms.

We used the ddRAD technique (Peterson et al., 2012) to build a library based on 32 individuals of Atlantic cod, Pacific cod, walleye pollock, Greenland cod and Arctic cod. From this we obtained information on roughly 400,000 single base variations between species and about 100,000 SNPs among Atlantic cod individuals. The results were somewhat contradictory from these data. For example using TreeMix (Pickrell and Pritchard, 2012) which models mixtures on maximum likelihood trees, we estimated splits and admixture between the species (Figure 3.8 and Figure 3.9, Halldórsdóttir and Árnason unpublished data). According to this analysis, migration between the Atlantic and Pacific oceans appeared to be common. This pattern was hard to explain under the currently accepted biogeographic hypothesis of two independent invasions of Atlantic cod into the Pacific (Coulson et al., 2006).

![Figure 3.8](image)

**Figure 3.8.** Inferred cod-fish tree with mixture events. Plotted is the maximum likelihood tree found using TreeMix analysis (Pickrell and Pritchard, 2012) of the ddRAD data under the parameter of migration \( m = 3 \). Horizontal branch length are proportional to the amount of genetic drift that has occurred on the branch. Migration events \( m \) are added to reduce the residual variance of the model and improve the fit. The number of migration events is decided a priori. The arrow indicates the direction of migration. The amount of migration is shown with the heatmap colorbar. Localities and species codes are as in Figure 3.4.
Figure 3.9. Inferred cod-fish tree with different number of mixture events. Plotted are the maximum likelihood trees found using TreeMix analysis (Pickrell and Pritchard, 2012) of the ddRAD data under various values of the parameter of migration $m$. A) $m = 1$, B) $m = 2$, C) $m = 4$, and D) $m = 5$. See Figure 3.8 for explanation of axis and identifiers.

Considering the origin of variation in closely related species it is worth examining how different loci can show different patterns. This is clearly illustrated and explained in Figure 3.10 adapted from Pääbo (2003). The possibility is, therefore, when using
reduced samples of this kind, that the haplotypes consist of different loci between individuals complicating results. In light of this it is possible that our ddRAD analysis has by chance detected genomic regions showing migration patterns between the Pacific and Atlantic that are not characteristic of the whole genome. This raises the question about how appropriate the reduced representation of sequences is in answering questions about biogeography and relationship between taxa. We therefore abandoned further ddRAD sequencing in favour of low coverage whole genome sequencing.

Figure 3.10. Within- and between-species variation along a single chromosome. a, The interspecies relationships of five chromosome regions to corresponding DNA sequences in a chimpanzee and a gorilla. Most regions show humans to be most closely related to chimpanzees (red) whereas a few regions show other relationships (green and blue). b, The among-human relationships of the same regions are illustrated schematically for five individual chromosomes. Most DNA variants are found in people from all three continents, namely Africa (Af), Asia (As) and Europe (Eu). But a few variants are found on only one continent, most of which are in Africa. Note that each human chromosome is a mosaic of different relationships. For example, a chromosome carried by a person of European descent may be most closely related to a chromosome from Asia in one of its regions, to a chromosome from Africa in another region, and to a chromosome from Europe in a third region. For one region (red), the extent of sequence variation within humans is low relative to what is observed between species. The relationship of this sequence among humans is illustrated as star-shaped owing to a high frequency of nucleotide variations that are unique to single chromosomes. Such regions may contain genes that contribute to traits that set humans apart from the apes. (Figure 2 and Figure legend from Pääbo (2003). © Copyright Nature Publishing Group. Printed with permission).
3.3.2 Whole genome sequencing

The second approach to address the high costs of sequencing a large number of individuals is to reduce the coverage of each individual in favour of more individuals. This approach, however, has a higher error rate. This problem can be alleviated by using new statistical methods of genotype likelihoods (Nielsen et al., 2011) permitting the use of low coverage data for a large number of individuals.

We apply this approach in paper V. In this study we prepared ten libraries of high coverage sequencing data, six Atlantic cod individuals and one individual from each of the four closely related species. For a population level study we generated libraries for 191 additional individuals, 153 Atlantic cod from throughout the species range and eight Arctic cod, nine Pacific cod, twelve walleye pollock, eight Greenland cod and one Polar cod with low sequencing coverage of approximately $2 \times$ coverage.

The high error rates of NGS data, especially the low coverage data, affect all downstream analysis which rely on calling genotypes by counting alleles. To reduce these effects algorithms within a probabilistic framework, called genotype likelihoods (Nielsen et al., 2011), have achieved the same accuracy without SNP and genotype calling as commonly done on high coverage data. The quality scores are used as a posterior probability for each genotype. The essence of the method is how to calculate the genotype likelihood. The probability $p(X_i|G)$ for genotype $G$ and read $i$ of individual $X$ is achieved by rescaling of the quality score of $X_i$. The product of $p(X_i|G)$ for all $i$ is then used to calculate the genotype likelihood $p(X|G)$ for each site i.e. the likelihood of observing the data given the unobserved genotype (Nielsen et al., 2011). In our analysis we used the ANGSD (Korneliussen et al., 2014) and related software (e.g., Skotte et al., 2013).

To study population structure we applied NGSadmix (Skotte et al., 2013) for admixture analysis and ngsCovar (Vieira et al., 2015) to estimate a covariance matrix that we used for principal component analysis (PCA) and discriminant analysis of principal components (DAPC) (Jombart and Ahmed, 2011). These programs are all adapted to genotype likelihoods from low coverage data.

The ddRAD and whole genome sequencing data presented here (ddRAD data) and in paper V (whole genome sequencing) are new and with these techniques we detected patterns of variation and divergence that have not been described before in the species under study. The main results of paper V is the finding of cryptic and hybrid species which I discuss in the last chapter.

3.3.3 Speciation

The concept of species, how to define a species, has long been argued. The biological species concept by Ernst Mayr (Mayr, 1963) is still the most commonly used: “Species are groups of interbreeding natural populations that are reproductively isolated from other such groups”. However, many alternatives to this basic concept have been proposed such as using genetic or phenotypic cohesion, evolutionary cohesion, or evolutionary history instead of interbreeding (Coyne and Orr, 2004).

Recently a hypothesis of speciation islands in face of gene flow has been introduced (Nosil, 2008; Turner and Hahn, 2010). Large-scale sequencing has revealed regions of divergence that are represented as loci involved in reproductive isolation between hybridizing populations. Under this hypothesis other parts of the genome show little
or no signs of divergence because of gene flow (Turner and Hahn, 2010; Cruickshank and Hahn, 2014). Strong selection is thought to be working against introgression of the divergent genes between species (Wu, 2001). This kind of speciation has been described in hybrid zones of divergent species e.g. between the M and S form of mosquitoes in the species complex *Anopheles gambiae* (Turner et al., 2005) and in the *Heliconius* butterflies (Nadeau et al., 2011). As mentioned before, this mechanism of divergence has been proposed to be at work in the Atlantic cod (Hemmer-Hansen et al., 2013; Bradbury et al., 2014).

### 3.3.3.1 Speciation in the sea

In allopatric speciation, populations split into two different environments with physical barriers between the populations reducing or blocking migration of breeding individuals between them (migration, \(m\), is zero or much reduced). In sympatric speciation the barriers are biological features of the organisms not geography or distance (Coyne and Orr, 2004).

In the sea, fish populations often have high dispersal potential and the marine environment appears to have few barriers to gene flow. Thus allopatric divergence may be slow. Speciation in the marine environment frequently involves behavioural differences as known in sympatric speciation, e.g. in spawning time and mate recognition, gametic incompatibility, and habitat specialization such as salt tolerance (Palumbi, 1994; Matschiner et al., 2010).

Cryptic and sibling species, forms that are very similar morphologically, are common in the sea. They may reflect adaptive divergence of habitat use, life-history, and chemical recognition without morphological divergence (Knowlton, 1993; Bickford et al., 2007). Possibly we are not detecting or using the correct phenotypic or morphological traits to distinguish between forms or species and to define genera for marine organisms. Molecular genetic studies have revealed cryptic species in what we see as morphologically identical species (Vrijenhoek, 2009). This may be an indication of our lack of understanding of key selective factors in an environment such as the sea. The situation is even more complicated because of developmental plasticity in response to different environment under which the same genotype can produce several phenotypes as seen among some deep-sea organisms (Vrijenhoek, 2009).

### 3.3.3.2 Homoploid hybrid speciation

Homoploid hybrid speciation is when two (sympatric) species or lineages hybridize and produce a fertile breeding new species. The hybrid species is reproductively isolated from the parental species but has the same ploidy (Rieseberg et al., 2003). This form of speciation differs from polyploid speciation in which the new hybrid species has three or more sets of chromosomes (Coyne and Orr, 2004). A few examples of homoploid hybrid species are known or suspected among animals e.g. in *Heliconius* butterflies (Mavárez et al., 2006) and swordfish (Cui et al., 2013). Homoploid hybrid speciation can be hard to detect and that might explain why it has been considered extremely rare. With increasing genomic data this view has been changing in the last decade and now homoploid hybrid speciation is thought to be an important mechanism in evolutionary biology (Coyne and Orr, 2004; Schumer et al., 2014).

The mechanism of structural chromosomal rearrangement can trigger reproductive
isolation in hybrid species. This is best explained by examples and illustrations. First consider hybridization between species with genotypes $A A B B$ and $Ab Ab Ba Ba$ with a single reciprocal translocation between $Ab$ and $Ba$, the translocated chromosomes. If they hybridize the $F_1$ will be $A Ab B Ba$. It can form the two parental gametes, $A B$ and $Ab Ba$, and two new gametes, $A Ba$ and $Ab B$, which will be aneuploid suffering from duplications and deficiencies. In this case the $F_1$ is only half as fertile as the parents and the viable $F_2$ that are produced if selfing occurs, are of the parental genotypes. Under this scenario hybridization does not lead to a new hybrid lineage. Also there would be continuous selection against hybridization because fertility is reduced by one half (Coyne and Orr, 2004).

![Diagram](image)

**Figure 3.11.** A model of recombinational or hybrid speciation. Two parental species with the same diploid chromosome number differ by two reciprocal translocations (line one). The $F_1$ hybrid is heterozygous for these rearrangements (line two). The $F_1$ hybrid can produce 16 different gametes, four of which are genetically balanced (line three), see Figure 3.12 for examples of other possibilities. Two of these gametic forms are identical to the parental types and two are novel. By selfing, the $F_1$ hybrids will give rise in the $F_2$ to the two homozygous parental genotypes (the same as are shown in first line) and to two new fertile genotypes (fourth line) that are infertile with the parental types. Redrawn and modified from Rieseberg (1997).
However, in a scenario in which two species differ by two reciprocal translocations the possibility of hybrid species arises. Figure 3.11 illustrates this speciation model. The two species have genotypes AA BB CC DD on one hand and with reciprocal translocations between A and B and between C and D: AbAb BaBa CdCd DcDc. They can hybridize and form an $F_1$ which can produce 16 different forms of gametes. However, only four of them will be viable and balanced, two parental and two recombinant types. Thus under inbreeding the $F_1$ can produce four fit homokaryotypic $F_2$ lineages (homozygous for the translocated genomic region). Two of them will represent the original parental genotypes, AA BB CC DD and AbAb BaBa CdCd DcDc and two will represent novel true-breeding hybrid lines AA BB CdCd DcDc and AbAb BaBa CC DD (Figure 3.11, and see Rieseberg et al., 2003; Coyne and Orr, 2004). The $F_1$ barrier is a large threshold (see Figure 3.12 for examples of unbalanced forms) but high fecundity organism, such as cod-fish, may be able to pass through the $F_1$ barrier possibly facilitated by a hybrid swarm (Nolte and Tautz, 2010).

![Diagram of genome sequencing approach](image)

**Figure 3.12. Examples of unbalanced gametes in hybrid speciation.** Four examples of the 12 genetically unbalanced aneuploid gametes that the $F_1$ in Figure 3.11 will produce (line three). Two examples of the unbalanced aneuploid $F_2$ karyotypes resulting from selfing are shown (line four). Redrawn and modified from Rieseberg (1997).
The extent to which hybrid speciation (Mallet, 2007) contributes to both enigmatic and cryptic biodiversity in the sea (Knowlton, 1993, 2000) is unknown. Identifying and understanding cryptic species is important for evaluation of biodiversity. Identifying cryptic species complexes in commercially exploited organisms also is important for conservation and the protection and management of natural resources such as the fish stocks studied here.

3.3.3.3 A hypothesis of walleye pollock origin

How can homoploid hybrid speciation be explained? One explanation is that the hybrid species have a recombinant phenotype of the parental forms which allows them to transgress via ecological selection the parental niches which are unavailable to the parents (Mallet, 2007; Schumer et al., 2014). Under this situation the hybrid species could expand dramatically. This might be a scenario applicable to the walleye pollock described in paper V. Whole genome sequencing revealed that the walleye pollock genome arose by admixture of Atlantic cod and Arctic cod (Figure 2 in paper V). It is morphologically different from both parental types and formerly it was classified as a pollock. The mitochondrial comparison of Coulson et al. (2006) showed its relatedness to Atlantic cod but interestingly it is both morphologically and genetically different from Pacific and Atlantic cod. Both Pacific cod and walleye pollock were considered to be the results of Atlantic cod invading the Pacific about 200,000 years apart (Coulson et al., 2006). In the Pacific the two species occupy different habitats and niches. The walleye pollock is one of the most abundant fish stock in the Pacific. It is the second most harvested ocean fish stock in the world and the most important fishery in United States (Bailey, 2013). This species fulfills two of three criteria which Schumer et al. (2014) proposed as required to demonstrate the occurrence of homoploid hybrid speciation: 1) reproductive isolation from the parental types, and 2) evidence of hybridization in the genome. We do not have clear evidence for the third criterion: 3) evidence of hybridization being the cause of reproductive isolation, which will always be hard to obtain (Schumer et al., 2014). In this case the hybridization seems to have contributed to adaptive evolution and hybrid vigour, possibly associated with changes in gene expression (Hegarty et al., 2008).

3.3.3.4 Species and models of speciation in Atlantic cod

As mentioned above storage-tag data reveal two behavioural ecotypes that occupy different thermal niches (Thorsteinsson et al., 2012). Grabowski et al. (2011) also showed fine-scale differences between the two behavioural types in spawning habitat selection at the main spawning area suggestive of reproductive isolation. Homozygotes of the Pan I gene on linkage group LG01 have been shown to distinguish between the two ecotypes, however, heterozygotes complicate the picture and contradict the conclusions of reproductive isolation between the two behavioural ecotypes. The Pan I heterozygotes and some homozygotes show a range of behavioural patterns from shallow-water coastal to the deep-water migratory frontal behavioural type (see for example a supplemental figure in Pampoulie et al., 2007). Despite the obvious differences between the typical (homozygous) types the intermediate behaviour of heterozygotes (and some homozygotes individuals) has more or less been ignored in analysis (Pampoulie et al., 2007; Grabowski et al., 2011; Hemmer-Hansen et al., 2013).
In several studies the Pan I locus, which we can look at as a proxy for the two different ecotypes, has been shown to deviate from Hardy-Weinberg equilibrium. Although heterozygote deficiency (an apparent Wahlund effect) is seen in some instances (Árnason et al., 2009) there is a general excess of heterozygotes in many instances (Árnason et al., 2009; Jakobsdóttir et al., 2011; Karlsen et al., 2013), the opposite of Wahlund effects. One explanation for heterozygote excess is overdominance in fitness and another explanation is differential allele frequencies among the sexes (Robertson, 1965; Árnason et al., 2009). Árnason et al. (2009), however, suggested a third explanation namely a behavioural response in that heterozygotes might be more mobile and moving between habitats. In paper V we argue that the well-known homozygous behavioural ecotypes of Atlantic cod around Iceland (and by extension elsewhere) are two distinct species which hybridized (line one in Figure 3.11 and Figure 3.13) and produced a F1 hybrid (line two in Figure 3.11 and Figure 3.13) which will be heterozygous for all chromosomes including the Pan I locus (Figure 3.13). We suggest that the general heterozygote excess may be a consequence of hybrid vigour of the hybrid species (Mallet, 2007). It has trangressed the parental niches and constructed a new niche. The hybrid species may have a special Darwinian fitness in this new niche which both extends and encompasses the two parental niches. Variation of genes other than the speciation genes (the translocations in Figures 3.11 and 3.13) will segregate already in the F2 generation. Subsequently allele frequencies will change through time with evolution of the new hybrid species.

Other explanations are possible. The most obvious one and previously favoured view, which we are challenging, is that there is one species of Atlantic cod in Icelandic waters and elsewhere in the North Atlantic. It has two behavioural ecotypes adapted to different environments which reproduce and form heterozygotes of the costal and frontal ecotypes. Under this scenario there could be overdominance in fitness or other form of balancing selection which would maintain variation within this single species. Furthermore, the fish will reproduce independent of ecotype, i.e. all Pan I and all LG01 genotypes will reproduce panmictically at the breading grounds. However, our results on whole genome analysis shown in Figure 4 in paper V show that the frontal ecotype separates from the other groups immediately under the admixture model of $k = 3$ populations. Also there is little evidence of genetic material from the coastal type in the frontal type which is hard to understand if the three groups represent one panmictic species. Also under this hypothesis it is hard to explain the homozygotes which group with the heterozygotes in our discriminant analysis of principal components (DAPC analysis in Figure 5 in paper V).

Another possible hypothesis is that the ecotypes are two species which continuously reproduce and form sterile hybrids. Bradbury et al. (2014) describe the finding of $F_1$ individuals between N and S types of Atlantic cod in western Atlantic, with very few instances of $F_2$ and of backcrossed individuals. They show high interchromosomal linkage disequilibrium. This implies that the $F_1$ are infertile (mules) and one would expect strong selection against such costly waste leading to a rapid response to avoid interbreeding, especially in a high fecundity organism with large excess reproductive capacity. The model of one reciprocal translocation discussed above, might explain the situation as well. Under this model the fertile hybrids only produce the parental types but with reduction of fertility by one half. This scenarios would reflect recent admixture. Alternatively, by using the model of two or more reciprocal translocations, the hybrids might be a new homoploid species as we propose for the behavioural
ecotypes in Icelandic waters. This model better explains the interchromosomal linkage disequilibrium (Cruickshank and Hahn, 2014). Nielsen et al. (2003) showed evidence of a similar hybrid zone in the Baltic and our whole genome data also support admixture in the Baltic (Figure 4 and Supplemental Figure S3 in Halldórsdóttir and Árnason, 2015b). This might imply that hybrid speciation is prevalent in cod-fish. Future work will address these issues.

Using the model of homoploid hybrid speciation, involving two or more reciprocal translocations, the intermediates (including Pan I heterozygotes) can be explained as a new species (novel $F_2$ karyotypes in Figure 3.11 and Figure 3.13). If the Pan I locus and the entire LG01 is another chromosome outside of the translocated regions of the hybrid species, the individuals in the $F_2$ population will segregate out the parental types as well as the heterozygotes with normal Mendelian segregation (Figure 3.13). This could explain the enigmatic intermediate behavioural types among all genotypes of the Pan I locus. In other words, the Pan I locus is not a speciation locus in this new hybrid species (Figure 3.13). Instead this locus and other genes on LG01 reflect ecological adaptations. These genes allow the new hybrid species to construct a new niche while also utilizing the parental niches. There are two inversions on LG01 (Kirubakaran et al., 2016) which suppress recombination and are responsible for the high observed LD (Hernandez and Árnason, 2015; Kirubakaran et al., 2016). In this model the inversions segregate normally within the hybrid species.

A model of speciation-with-gene-flow (Turner et al., 2005; Nosil, 2008) has been proposed as an explanation for population differentiation found within Atlantic cod (Hemmer-Hansen et al., 2013; Bradbury et al., 2014). This model assumes low levels of differentiation in large parts of the genome because of ongoing gene flow which homogenizes those genomic regions. Under this model 'islands of genomic divergence', genomic regions showing high differentiation, are loci connected to isolating traits between the incipient species, speciation genes in effect. Introgression of islands is prevented by selection but recombination homogenizes the rest of the genome. This model assumes the establishment of interchromosomal linkage disequilibrium although chromosomes assort independently. Maintaining such a strong association at unlinked loci at different chromosomes implies that most recombinant genotypes are less viable. Only genotypes with the correct allelic combination survive. Even with a model of only few islands participating the number of gametes that can form viable zygotes will be very small (Cruickshank and Hahn, 2014). Although the high-fecundity Atlantic cod has high reproductive excess and should be able to withstand large selection pressures (Árnason and Halldórsdóttir, 2015) such strong selective forces should lead to the evolution of effective reproductive barriers between the forms.

As an alternative to the model of speciation-with-gene-flow Cruickshank and Hahn (2014) presented a model of divergence-after-speciation. Under this model there is no differential gene flow among loci. Instead reproductive isolation is rapid, even instantaneous and complete (such as could occur in homoploid hybrid speciation). The divergence-after-speciation model explains the low level of divergence throughout most part of the genome as shared ancestral polymorphisms. The more recent the split the more homogenized are the genomes. The elevated differentiation in the 'islands of divergence' is explained as being due to heterogeneous natural selection over the genome but not due to heterogenous migration rates. The genomic islands are loci of ecological adaptations. This model also explains linkage disequilibrium between islands on different chromosomes because ecological adaptation within species involving genes
Figure 3.13. A model of homoploid hybrid speciation in Atlantic cod. The model in Figure 3.11 is extended to include a chromosome for linkage group LG01 containing the Pan I locus which is a proxy for the behavioural ecotypes (red: coastal and blue: frontal). The model assumes that the LG01 chromosome does not carry a translocation that is involved in the speciation. Therefore, the $F_2$ hybrid species will have all three genotypes of LG01. Two balanced gametes are shown ($G_1$ and $G_2$). Either of them could produce a $F_2$ homoploid hybrid species that produces all three genotypes of Pan I or LG01 inversion blocks (Kirubakaran et al., 2016) by Mendelian segregation. Redrawn and modified from Rieseberg (1997).

As explained above when considering two reciprocal translocations, the $F_1$ by selfing will possibly produce two balanced gametes (line three in Figure 3.13). In this model the LG01 (the chromosome with the Pan I gene) is not one of the chromosomes carrying a translocation. Fusions of these gametes may produce via the reciprocal translocations a new homoploid $F_2$ hybrid (Figure 3.11 and 3.13, line four), which is reproductively isolated from the parental types. As explained in Figure 3.12, part(s) of the genome will be missing in other gametic types and the offspring consequently...
will be inviable or with reduced fitness depending on the size of the translocations. The new balanced gametes produced by the $F_1$ (line three in Figure 3.13) can have different composition of other parts of the genome. The model presented here is that the gametes with balanced reciprocal translocations have either \textit{Pan I A} or \textit{Pan I B} allele. Therefore, on selfing (interbreeding among the $F_1$) the three genotypes of \textit{Pan I}, \textit{AA}, \textit{AB}, \textit{BB}, will be produced. After this speciation event the LG01 (\textit{Pan I}) homozygotes may further diverge from the homozygotes from the parental species. Evidence for this can be seen in Figure 5 in paper V, in which a group of \textit{Pan I} homozygotes clusters with the heterozygotes.

In paper V we used a modified model of divergence-after-speciation (Cruickshank and Hahn, 2014), assuming divergence before and further divergence after speciation, to interpret the results of whole genome sequencing of the different ecotypes in Atlantic cod. In light of our results we therefore favour the idea of at least three species of Atlantic cod in Icelandic waters and elsewhere in the North Atlantic. These are the coastal and frontal ecotypes and the homoploid hybrid species which transgresses the parental niches. This hypothesis also has gained support with new findings (Kirubakaran et al., 2016) which show two inversions in LG01 involved in the divergence of two ecotypes.

\subsection*{3.3.4 Synopsis}

To date genome wide studies in Atlantic cod comprise analysis of about 1600 EST (expressed sequence tags) SNPs (e.g. Hubert et al., 2010; Borza et al., 2010; Hemmer-Hansen et al., 2013; Bradbury et al., 2014). Subsets of the outlier loci show clines of adaptation to various environments, spatial patterns or complex spatio-temporal patterns while no differentiation is found among neutral loci. I decided to extend this by sequencing 32 individuals using the ddRAD method (double digest restrictions associated sequencing, Peterson et al., 2012). This was partial genome sequencing of individuals of Atlantic cod, Pacific cod, walleye pollock, Greenland cod and Arctic cod. From this data we gained information on about 400.000 single base variations between species and on about 100.000 SNPs among individual Atlantic cod. The ddRAD method overcomes ascertainment bias and randomly selects regions from the genome. The results showed that it would be of great interest to extend this to whole genome sequencing and cover the entire genome. We, therefore, designed a 2×coverage whole genome sequencing study of 191 individuals. In addition six individual Atlantic cod were sequenced at a 30×coverage and one of each of the closely related species was sequenced at a 20 to 30×coverage. For statistical analysis we applied the methods of genotype likelihoods (Nielsen et al., 2011) implemented in ANGSD (Korneliussen et al., 2014) and related software (e.g. Skotte et al., 2013). The main idea is to use genotype likelihood of the individuals instead of called genotypes, thus circumventing problems of errors associated with low coverage next generation sequencing.

The results revealed new knowledge about speciation in Atlantic cod and the origin and admixture between Atlantic and Arctic cod. The results are important for understanding the evolutionary status of the Pacific species. We show that the evolutionary status of walleye pollock is a hybrid between Arctic cod and Atlantic cod which has transgressed the ecology of its parents. The evolutionary status of various subgroups of Atlantic cod is found to be genetically distinct e.g. the western cod, eastern or coastal cod and the frontal ecotype of cod from Iceland are distinct entities.
in an admixture analysis with $k = 3$ where $k$ is the number of ancestral populations in the model. We propose a new hypothesis about the evolutionary status of the known ecotypes of cod around Iceland, and by extension in the North Atlantic, the frontal and coastal behavioural types. We propose that the two types are separate species, adapted to deep and shallow water respectively, which hybridized and formed a new homoploid hybrid species. Thereby we reject the scenario of Atlantic cod being one species with subpopulations. Regarding the scenario of there being two reproducing species which produce infertile $F_1$ we are sceptical of that explanation because it entails high genetic load and selective pressures that should quickly lead to avoidance of interbreeding. We favour the model proposed by Cruickshank and Hahn (2014) of divergence-after-speciation and explain our results accordingly. These issues are addressed in Paper V which is a submitted manuscript under review.
4 Future Perspectives

The results in this thesis, especially paper V, raise many interesting hypotheses about hybrid speciation in general and about cod-fish biology in particular, which remain to be tested. For example, we need to examine the criteria of hybrid speciation (Schumer et al., 2014) and hybrid sterility, as well as structural chromosomal variation (c.f. Fan and Fox, 1991; Kirubakaran et al., 2016), formal tests of the age and extent age of hybridization (c.f. ABBABABA, Durand et al., 2011; Patterson et al., 2012), biogeography and natural selection. Our results call for a re-evaluation of previous work on cod-fish with implications for resource management. The hybrid nature of both the valleym pollock and Atlantic cod raises the question concerning the extent to which very profitable fisheries (Kurlansky, 1997; Bailey, 2013) depend on hybrid vigour. For example, did the decimation of the hybrid and frontal fish influence the collapse and non-recovery of the Newfoundland fishery (Hutchings, 2000)? Changes in the oceans and their impact of ecosystems and fisheries have unpredictable consequences (Gewin, 2015). The ongoing climate warming and northern hemisphere ice melting predict species interchange between the north Pacific and the Atlantic (Wisz et al., 2015). The potential exists for further hybridization of Pacific, Arctic and Atlantic taxa, with unknown consequences. Hybrid speciation may be more common among animals in general and marine fishes in particular than previously thought (Schumer et al., 2014). It will be exciting to extend the work in these directions.
References


Begg, G. and Marteinsdottir, G., 2000. Spawning origins of pelagic juvenile cod *Gadus morhua* inferred from spatially explicit age distributions: potential influences on


References


References


Matschiner, M., 2015. Peer review #1 of Nucleotide variation and balancing selection at the Ckma gene in Atlantic cod: analysis with multiple merger coalescent models (v0.1). *PeerJ* doi:10.7287/peerj.786v0.1/reviews/1.


References


Reviewer-2, 2015. Peer review #2 of Trans-species polymorphism at antimicrobial innate immunity cathelicidin genes of Atlantic cod and related species (v0.1). *PeerJ* doi:10.7287/peerj.976v0.1/reviews/2.


Wallimann, T., Wyss, M., Brdiczka, D., Nicolay, K., and Eppenberger, H., 1992. Intracellular compartmentation, structure and function of creatine kinase isoenzymes in tissues with high and fluctuating energy demands: the ‘phosphocreatine circuit’
References


Organization of a $\beta$ and $\alpha$ Globin Gene Set in the Teleost Atlantic Cod, *Gadus morhua*.

Katrín Halldórsdóttir and Einar Árnason, 2009

*Biochemical Genetics*, 47, 817–830

Author contributions: Katrín Halldórsdóttir and Einar Árnason conceived and designed the experiments, analysed the data, contributed analysis tools, prepared figures and/or tables and wrote the paper. KH performed the experiments. EÁ contributed reagents and materials.

© Copyright Springer. Printed with permission.
Organization of a β and α Globin Gene Set in the Teleost Atlantic Cod, *Gadus morhua*

Katrín Halldórsdóttir · Einar Árnason

Received: 19 October 2008 / Accepted: 8 May 2009 / Published online: 26 July 2009 © Springer Science+Business Media, LLC 2009

**Abstract** Developmental globin gene expression and gene switching in vertebrates have been extensively studied. Globin gene regions have been characterized in some fish species and show linked α and β loci. Understanding coordinated expression between α and β globin genes in fish is of importance for further insights into globin gene regulation in teleosts and higher vertebrates. We characterize linked β and α globin genes in Atlantic cod, pulled from the Atlantic cod genome with a PCR research strategy, by screening a genomic λ library and primer walking. The genes are oriented tail-to-head (5′–3′), differing from the head-to-head orientation in transcriptional polarity characteristic of teleostean globin genes. Four tandem repeats are found in an intergenic region of 1500 base pairs. One microsatellite, which consists primarily of atg tandem repeats, has an open reading frame. The globin genes and open reading frame have a CCAAT promoter element and TATA boxes. The promoters of the open reading frame and the β gene share an 89-bp block (with 100% identity) that probably regulates transcription.

**Keywords** Hemoglobin · β/α globin genes · Regulatory elements · Atlantic cod · *Gadus morhua*

**Introduction**

Hemoglobins are found in all groups of organisms and are encoded for by orthologous genes descended from an ancient ancestral gene (Hardison 1998).

Electronic supplementary material The online version of this article (doi:10.1007/s10528-009-9280-0) contains supplementary material, which is available to authorized users.

K. Halldórsdóttir (✉) · E. Árnason
Institute of Biology, University of Iceland, Sturlugata 7, Reykjavik, Iceland
E-mail: katrinhalldorsdottir@gmail.com

Springer
Globin genes in mammals and other higher vertebrates are arranged in distinct clusters, the \(\alpha\)-like and \(\beta\)-like clusters, located on separate chromosomes (Karlsson and Nienhuis 1985). The arrangement of the genes in the clusters reflects the order of expression during development. Proximal \textit{cis}-acting regulatory elements, as well as a distal locus control region (LCR), play a vital role in this regulation (Ji et al. 2000; Shen et al. 2002).

Globin gene regions have been characterized in some model fish species, such as pufferfish (\textit{Fugu rubripes}, Gillemans et al. 2003), embryonic and adult zebrafish (\textit{Danio rerio}, Brownlie et al. 2003; Chan et al. 1997), and medaka (\textit{Oryzias latipes}, Maruyama et al. 2004), as well as in the semimodel Atlantic salmon (\textit{Salmo salar}, Wagner et al. 1994). In these fish, the \(\alpha\) and \(\beta\) loci are a linked set and are located on the same chromosome. An understanding of coordinated expression of \(\alpha\) and \(\beta\) globin genes in fish, at the base of the vertebrate tree, is important to further our insights into the regulation of globin genes in higher vertebrates in which developmental globin gene expression and gene switching have been studied extensively.

In most species, transcription proceeds in the same direction—in the 5'–3' orientation—for the various globin genes at a given locus. In the African frog (\textit{Xenopus laevis}), the genes are linked on the same chromosome with three \(\alpha\) genes followed by three \(\beta\) genes. The direction of transcription is 5'–3', and embryonic genes are located 5' to adult genes (Hosbach et al. 1983). The arrangement of these genes in fish is different, however. In the zebrafish, pairs of linked \(\alpha\) and \(\beta\) genes are found for embryonic genes and adult genes, both on the same chromosome. The direction of transcription is tail-to-tail (5'–3', 3'–5') in the embryonic pair but head-to-head (3'–5', 5'–3') in the adult globin pair (Brownlie et al. 2003). Similarly, the globin genes in pufferfish are closely linked and directed in opposite transcriptional orientations (Gillemans et al. 2003). In salmon (Wagner et al. 1994), the genes are oriented 5'–3', 3'–5' (tail-to-tail orientation), similar to the zebrafish embryonic pair. In medaka, the globin genes are grouped in two clusters, one containing embryonic genes and the other adult globin genes. Both clusters have pairs of \(\alpha\) and \(\beta\) genes in a head-to-head orientation. One pair of embryonic genes, however, had the genes oriented 5'–3', similar to the results described here. It is interesting to note that these genes show an expression pattern different from other embryonic genes (Maruyama et al. 2004).

In humans, gene order and distance relative to the LCR affect regulation of transcription (Harju et al. 2005; Johnson et al. 2005). The LCR, a cluster of five DNase I hypersensitive sites (HS) located upstream of the 5' \(\epsilon\)-globin gene, is a major \textit{cis} regulator of all five \(\beta\)-like globin genes (Ji et al. 2000; Feng et al. 2005). The hypersensitive sites have different roles. HS2 and HS3, for example, are involved in silencing activity of the LCR during development, but HS4 is not (Feng et al. 2005). Also, each individual gene has its promoter, and proximal regulatory elements that play pivotal roles in globin gene switching by interacting with the LCR (Shen et al. 2002). In addition, sequence-specific transcription factors like EKLF (erythroid Kruppel-like factor) are thought to regulate activation and repression via proximal \textit{cis} elements of individual genes. Moreover, direct repeat elements in the promoters of \(\epsilon\) and \(\gamma\) genes influence silencing of the genes in the
adult stage (Omori et al. 2005). The mechanisms developmental regulation is, in the context of chromatin structure, where transcriptional regulators must modify and open or close the chromatin (Shen et al. 2002). Nuclear transcripts that extend across the LCR and intergenic regions are found in erythroid cell lines but not in the cytoplasm; thus, the LCR and intergenic transcripts appear to be nuclear specific. These transcripts are either a consequence of open chromatin or they act to open the chromatin for transcription of the genes (Ashe et al. 1997).

Gillemans et al. (2003) mapped the α-β locus in pufferfish for DNase I hypersensitive sites to look for LCR and search for transcription factor binding sites. They found an EKLF consensus site and a hypersensitive site in the promoter of the β-globin gene and concluded that remote regulatory elements do exist in fish, although they are still not characterized (Gillemans et al. 2003). Maruyama et al. (2007), however, found a homolog of DNase-I Hypersensitive site-401 kb downstream of the embryonic cluster in medaka. This hypersensitive site is found in intron 5 of the C16orf35 gene in humans and in medaka as well. They found an NF-E2 binding site located between GATA boxes, which characterizes this locus in humans and mice (Maruyama et al. 2007). Recently, whole genome comparison between humans and pufferfish has shown highly conserved noncoding sequences covering large areas in and around genes that are involved in the regulation of development. These sequences are presumed to have cis-regulatory function (Sandelin et al. 2004; Woolfe et al. 2005), and their conservation between human and teleost genomes implies an evolutionarily stable function in regulatory processes. Thus, it is important to look for conserved sequence blocks and repeat elements in and around genes of interest (Maruyama et al. 2007).

Duplicated copies of a gene can be fixed and maintained in a population if their functions diverge or are subdivided (Lynch and Force 2000). Often, a gene may gain a new function, even without changing the protein (for cell type-specific genes), with changes in promoter and enhancer sequences (Shimeld 1999). Mutations in promoter sequences can lead to severe disease, such as β-thalassemia in humans (Agarwall et al. 2006), as well as to novel functions. Therefore, to characterize genes in multigene families, it is important to determine the control regions and promoters of the genes and to compare them.

The biology of the Atlantic cod, Gadus morhua, a nonmodel organism, has several features that make it suitable for comparison with model organisms such as zebrafish and pufferfish. It is one of the most fecund vertebrates and may thus be subject to natural selection of greater intensity than low fecundity organisms (Arnason 2004). Second, it lives from the surface of the sea to a depth of 600 m, and thus its hemoglobin system must deal with highly variable physiological conditions of oxygen demand and supply. The Atlantic cod has, for example, very high root-effect hemoglobin, the highest among teleosts (Berenbrink et al. 2005), which indicates the potential to deal successfully with variable pressure in the water column.

An important step in understanding and explaining the structure and function of hemoglobin proteins and their genes in the physiology of the Atlantic cod is to characterize the molecular components. The first step is to characterize the location
and orientation of the \( \alpha \) and \( \beta \) genes, as well as potential control regions regulating transcription. We report here, as a first step toward such a goal, a study revealing a set of linked \( \beta \) and \( \alpha \) genes and an intergenic region containing tandem repeats, as well as proximal promoter sequences and conserved noncoding sequence blocks associated with the genes.

**Materials and Methods**

**DNA Isolation**

Both blood and muscle tissue samples of individuals already genotyped by protein isoelectric focusing as HbI-1/1 (or SS), HbI-2/2 (or FF), and HbI-1/2 (or FS) (Sick 1965) were obtained from Jarle Mork, NTNU, Norway. We extracted DNA with a Chelex/proteinase K extraction method (Walsh et al. 1991).

**PCR**

Primer pairs were designed for the \( \alpha \) gene and for the \( \beta \) gene (Supplemental Table S1). Based on knowledge about fish globin genes [e.g., that Atlantic salmon \( \alpha \) and \( \beta \) globin genes are linked tail-to-tail (Wagner et al. 1994), and that in several fish, globin genes are located in tandem on the same chromosome, frequently in opposite directions (Brownlie et al. 2003; Chan et al. 1997)], we designed a PCR strategy for finding similar features in Atlantic cod. We tested amplification on genomic DNA with sets of forward and reverse \( \beta \) and \( \alpha \) globin primers. We successfully amplified a large fragment using a forward \( \beta \) and a reverse \( \alpha \) primer. In order to improve specificity of amplification, we made the primers longer [about 40 base pairs (bp)]. GmHBBL29_long was extended to amino acid 7 in exon 1 of the \( \beta \) gene, and GmHBAR553_long was extended in the 3' untranslated region of the \( \alpha \) gene. The PCR was carried out with an initial denaturation step at 94°C for 2 min; followed by 10 cycles of denaturation at 94°C for 20 s, annealing at 72°C for 30 s, and extension at 68°C for 10 min; followed by 25 cycles of denaturation at 94°C for 20 s, annealing at 68°C for 30 s, and extension at 68°C for 10 min, which was increased by 10 s every cycle. A final extension step was at 68°C for 10 min. PCR amplifications were performed in 25 µl reactions using Long PCR Enzyme Mix (Fermentas K0181). This unusual PCR protocol is based on the principle of touchdown PCR, with an initial high annealing temperature to enhance specificity. The long primers had a high melting temperature, and thus initial annealing was done at 72°C. However, extension was carried out at 68°C, as recommended by the manufacturer of the Long PCR Enzyme Mix. This PCR gave clear bands on a gel, whereas other protocols gave a smear. Chelex-extracted DNA (3 µl, 1:19 dilution) was used as template without quantification.

In order to look for the 3' region of the \( \alpha \) genes, we designed three primers (GmHBAL371, GmHBAL389, and GmHBAL398; Supplemental Table S1) in exon 3 of the \( \alpha \) gene to use with a DNA Walking Speedup Premix Kit (Seegene K1501). We were able to amplify a 500-bp fragment with a standard PCR protocol.
recommended in the kit. The fragment contained exon 3 of the a gene, starting at amino acid 13, and continued 3' to the coding sequence.

Cloning

The PCR products from individual SS104.1 and Speedup a 3' UTR-Clone 1 were purified from 0.7% TAE agarose gel with an Ultra Agarose Spin Kit (ABgene). We cloned gel fragments into the PCR 4-TOPO vector using the TOPO TA Cloning Kit for Sequencing (Invitrogen K4530-20). Plasmid DNA purification, from overnight cultures, was done with a Qiaprep Spin Miniprep Kit (Qiagen). Plasmids from SS104.1 containing inserts of about 3000 bp were sequenced with a set of 10 sequencing primers (Supplemental Table S2). Plasmids from Speedup a 3' UTR-Clone 1 were sequenced with primers M13 Forward long and M13 Reverse long (Supplemental Table S2).

Screening of λ Genomic Library and Subcloning

In order to isolate other globin genes from Atlantic cod, and in particular to search for a 5' region of the β genes, we screened a genomic library constructed in the phage λ vector GEM-11 by Scan Biotec using standard procedures (Sambrook et al. 1989) (for details see Supplemental materials) and obtained a 13 kb λ clone giving positive hybridization signals with the β/β probe. We subcloned it into pUC19 and sequenced a BamHI digested fragment. The clone, λ1.6BamHIClone21, contained a 5' region and part of the coding sequence of a β gene (GenBank accession no. EF644912).

Sequencing and Data Analysis

The sequencing reaction products were run on an ABI 3100 automated capillary sequencer (Applied Biosystems). For analysis of sequence data, we used the Phred/Phrap/Consed software (Ewing et al. 1998; Ewing and Green 1998; Gordon et al. 1998). Each base in the contigs of the assembly of sequenced data from SS104.1 and Speedup a 3' UTR-Clone 1 had quality values higher than 40 or 99.99% accuracy of base call, and most were higher than 60 (or 99.9999% accuracy). Quality values of each base pair of subclone λ1.6BamHIClone21 were greater than 60. DNA sequence data have GenBank accession nos. EF644912 (1386 bp), EF644886 (3025 bp), and FJ858154 (428 bp).

Gene Prediction

For prediction of coding sequences of the contigs, we used the New Genscan Web Server at MIT software (http://genes.mit.edu/Genscan.html; Burge and Karlin 1997). Information from Genscan prediction was used for various data analyses, using R (R Development Core Team 2006) and the Ape package (Paradis et al. 2005) in particular. We also used the FGenes and FGenesH programs from
Softberry (www.softberry.com) for gene prediction to compare with and augment Genscan predictions.

For further evaluation and authentication of the predicted genes, we performed Blast searches, using Blastn, of the contigs. We aligned the clones isolated from the genomic λ library and the clones derived from PCR amplification of genomic DNA using Bl2seq (Altschul et al. 1990; Gish and States 1993; Zhang et al. 2000). For analyzing the structure of sequence data, we used the Emboss program Etandem (Rice et al. 2000), which looks for tandem repeats in a nucleotide sequence.

**Results**

**A Globin Gene Set**

The 3 kb clone SS104.1 had a β globin locus with three exons and two introns at the 5' end, followed by about 1500 bp of intergenic region, and an α globin locus with three exons and two introns at the 3' end (Fig. 1). Four different tandem repeats or microsatellite repeat loci predicted by Etandem, one tetrameric (caaa) and three trimeric (atg, aat, taa), were located in the intergenic region (m1–m4; Fig. 1). The number and identity of tandem repeats varied, with the highest identity of 93% for atg (Table 1). The β gene (gene 1; Supplemental Table S3) was oriented 5'–3' at the 5' end of the 3000-bp fragment and 5' to the α gene (Fig. 1). Genscan predicted a probability of 1 for all β exons, but the probability was 56% for exon 1 and 89% for exon 2 and exon 3 of the predicted α gene (Supplemental Table S3). The α gene likewise was oriented 5'–3' at the 3' end of the 3000-bp fragment and 3' to the β locus (Fig. 1). Genscan predicted a single exon open reading frame that largely coincided with the atg microsatellite locus in the intergenic region, with a poly D amino acid sequence by codon gat. The microsatellite locus can therefore be called a gat repeat.

**Promoters and Putative Control Elements**

The λ clone contained exon 2 (where a BamHI restriction site exists), intron 1, and exon 1 of a β gene and continued for 1700 bp upstream from the coding sequence of the β gene (λ1.6BamHIClone21; Fig. 1). A TTTAAA box was 5' to the β gene (TSS 1 in Supplemental Table S4). An aataaa poly-A signal was 32 bp downstream from the coding sequence of the β gene (pA and black stripe 3' to β exon 3; Fig. 1). An alignment of λ1.6BamHIClone21 with SS104.1 showed three highly conserved noncoding regions: the λ clone nucleotide position (np) 99–120 with the SS104.1 np 980–1001 had 22/22 (100%) identity; the λ clone np 275–363 with the SS104.1 np 1024–1112 had 89/89 (100%) identity; the λ clone np 581–622 with the SS104.1 np 1767–1808 had 36/42 (85%) identity (c1, c2, and c3, respectively; Fig. 1). Furthermore, caaa and ata tandem repeats (m1 and m2; Fig. 1) also were found in the 5' β region, and they as well as the conserved sequence blocks occurred in the same relative positions in the 5' β region and in the β/α intergenic region 3' to the β gene.
Fig. 1 Schematic of a 5'–3' tail-to-head β–α gene set in Atlantic cod. Two clones are represented: The PCR clone SS104.1S is a TOPO-TA cloned 3000-bp PCR fragment of genomic DNA from individual SS104. λ1.6 BamHI Clone 21 is a subclone from a larger clone pulled from an Atlantic cod genomic λ library. Speedup α 3' UTR Clone 1 is a sequence gained by primer walking, using a DNA Walking Speedup Premix Kit, showing the 3' UTR of the z gene. Zero (0) represents the SS104 β gene initiation codon. e (blue box) indicates three exons in β- and α-like genes. i (thin line between boxes) indicates introns. e1 (red box) is a single-exon gene in the intergenic region. Striped blue boxes show parts of exons that are identical in different clones. t (yellow box) represents TATA boxes in the 5' UTR promoter regions of the three genes. pA (thin black bar) is polyadenylation signals in the 3' UTR region of all genes. m (orange box), tandem repeats (microsatellites) of the first two clones predicted by Etandem. c (green box), regions of conserved or nearly conserved sequences in the β/α intergenic region of Clone SS104.1 and in the 5′ region of the β gene in λ1.6 BamHI Clone 21. Similarities were found to be part of an unrelated gene, the pleurocidin-like gene, designated o (magenta box), q, r, and s (three top levels in SS104.1) indicate locations of short fragments (about 10–12 bp) at α exon 1, intron 1, and exon 2 junctions, showing identity to fragments in the intergenic region (color figure online).

Table 1 Tandem repeats in β and α globin genes of Atlantic cod, predicted by the Etandem program

<table>
<thead>
<tr>
<th>Position</th>
<th>Start</th>
<th>End</th>
<th>Score</th>
<th>Size</th>
<th>Count</th>
<th>Identity</th>
<th>Consensus repeat</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>908</td>
<td>979</td>
<td>26</td>
<td>4</td>
<td>18</td>
<td>71</td>
<td>caaa</td>
</tr>
<tr>
<td>2</td>
<td>1140</td>
<td>1292</td>
<td>30</td>
<td>3</td>
<td>51</td>
<td>61</td>
<td>taa</td>
</tr>
<tr>
<td>3</td>
<td>1872</td>
<td>1913</td>
<td>31</td>
<td>3</td>
<td>14</td>
<td>91</td>
<td>aat</td>
</tr>
<tr>
<td>4</td>
<td>1446</td>
<td>1568</td>
<td>102</td>
<td>3</td>
<td>41</td>
<td>93</td>
<td>atg</td>
</tr>
</tbody>
</table>

a Number of matches and mismatches to consensus

b Percentage identity of bases to a perfect consensus repeat
The largest conserved box, c2, partly coincided with the promoter predicted for the putative single-exon open reading frame in the intergenic region. The first 17 bp of the promoter sequence were identical at the end of the c2 box. The overlap region included the CCAAT box, a known type of promoter in vertebrate globin genes (Hardison 1998; Fang et al. 2004). The promoter for the \( \beta \) locus was, therefore, classified as a CCAAT promoter with a TTTAAA box. A Blast of the c2 sequence onto the zebrafish and pufferfish genomes found no similarities. The conserved sequence blocks, however, found up to 20 bp of similarities to sequences on several chromosomes in the human genome.

A 40-bp promoter with a TATA box was located 5' of the \( \alpha \) gene (t3; Fig. 1). At 29 bp 5' of this promoter, we also found a CCAAT box. The cloned fragment ended right after the 3' end of the \( \alpha \) coding sequence, and there was no evidence for a polyA signal 3' to the \( \alpha \) gene. Using a DNA Walking Speedup Premix Kit (Seegene K1501), we walked 500 bp 3' of the \( \alpha \) locus and found a polyA signal 237 bp downstream from the coding sequence (Speedup \( \alpha \) 3' UTR-Clone 1; Fig. 1).

Similarities to the intergenic region were found in three consecutive 11-bp coding regions from the \( \alpha \) gene (q, r, and s in SS104.1; Fig. 1). The q fragment with amino acid sequence ALSR was found at the end of \( \alpha \) exon 1 (q2) and at np 1766–1776 in the intergenic region (q1). Differing by one substitution, this same fragment was found at the end of \( \alpha \) intron 1 (q3). The r fragment with amino acid sequence VAV was found in direct continuation of the q3 fragment at the beginning of exon 2 (r2 amino acid numbers 2–4 of that exon). It was also found 3' to the \( \beta \) gene at np 789–794 (r1). The s2 fragment partly overlapping and in continuation of fragment r2 with amino acid sequence VYPQ was found in exon 2 (s2 amino acid numbers 4–7) and also 5' to the \( \alpha \) gene at np 2207–2217 (s1). Thus, different parts of the contiguous 30-bp region at the boundary of \( \alpha \) intron 1 and exon 2 show similarities with various parts of the intergenic region.

The single-exon open reading frame at np numbers 1442–1579 (gene number 2, Supplemental Table S3, e1; Fig. 1) had all the elements of a functional gene. A strange-looking polyD predicted protein based on the atg microsatellite revealed no similarities to any known protein by Blasting. Genscan probability for this exon was 55% (Supplemental Table S3); however, it had a methionine start codon, a termination codon, and both a promoter (t2, 3' to \( \beta \) locus; Fig. 1) and a putative polyA signal (pA, 3' to single-exon gene; Fig. 1, Supplemental Table S3). The 40-bp promoter region had a CCAAT box, and we found a TTTAAA box 85 bp 5' to the exon.

Discussion

Linked \( \beta \) and \( \alpha \) Globin-like Genes in Atlantic Cod

A distinct arrangement of globin genes characterizes vertebrates in general. In humans (and other mammals), \( \beta \) and \( \alpha \) globin clusters are located on different chromosomes. The \( \beta \) cluster includes five functional genes and one pseudogene. Their arrangement on the chromosome frequently corresponds to their order of
expression in development, all of them in a 5′–3′ direction of transcription. The β
cluster in mice has the structural genes in tandem including three embryonic genes
and two genes expressed in fetus and adult animals. In chicken, however, the adult β
genes are flanked by the embryonic genes (Sjakste and Sjakste 2002; Hardison
1998). This demonstrates variations on the theme of arrangement and order of
expression in development. Teleosts have sets of linked α and β genes oriented
head-to-head or tail-to-tail. The tail-to-head orientation of globin genes in the
Atlantic cod described here differs from the main pattern found in the model and
semimodel fish species described. The tail-to-head orientation is similar to that of
higher vertebrates, but it has also been described for one gene set out of five in the
embryonic gene cluster in medaka (Maruyama et al. 2004). The genes of a set are
 coordinately expressed. Thus, α3/β3 and α4/β4 in medaka are expressed simulta-
neously in early stage embryos (Maruyama et al. 2004). This is known in zebrafish
as well (Brownlie et al. 2003). The adult α1 globin gene in medaka, however, was
not coexpressed with adjacent adult β1 globin gene (Maruyama et al. 2004),
demonstrating variation in patterns of expression.

We predicted the amino acid sequence of our gene set to figure out whether our
genes belonged to embryonic or adult globin genes. Molecular phylogenetic
analysis of Atlantic cod and model teleosts did not reveal a close relationship to
either adult or embryonic genes. Also, the genes were not more related to the
embryonic α2/β2 in medaka, which has the same transcriptional polarity, than to
other genes (Supplemental Figs. S1 and S2).

In the zebrafish, the embryonic cluster is located on the same chromosome as the
adult cluster separated by an 8.5-kb region (Brownlie et al. 2003). On the other
hand, the embryonic and adult clusters are located on distinct chromosomes in the
medaka (Maruyama et al. 2004). This demonstrates variation on the arrangement
among teleosts as well. It raises a question about how teleost globin genes with such
an arrangement are controlled and expressed.

Open Reading Frame and Microsatellites in Intergenic Region

A single-exon open reading frame or gene is predicted in the region between the β
and α genes. The protein exon of that would consist mainly of aspartic acid (D). The
gene has its own promoter, initiation and termination codons, and a polyA signal and
is thus a putative functional gene. No similarities have been found to any known
protein in GenBank. The question remains whether this gene produces a functional
protein or whether it has something to do with controlling expression or other
functions of the globin genes or proteins. Li et al. (2004) state that simple sequence
repeats (SSR) in protein coding regions or variations of SSR in 5′-UTRs can regulate
gene expression. Triplet SSR expansion in 3′-UTRs can cause heterochromatin
silencing by decreasing promoter accessibility (Li et al. 2004). Possibly the atg
microsatellite tandem repeats in the intergenic region could act in that fashion. We
have been further analyzing expansion and contraction of the repeats, which could be
involved in regulating transcription of one or both of the linked globin genes. We will
publish the results elsewhere (Halldórsdóttir and Árnason, unpublished manuscript).
Indications of SSR within coding genes and their untranslated regions participating
in regulation of gene expression are known (Li et al. 2004). Variable copy numbers of the tandem repeats linked to specific genes are also thought to be in relation to an immediate response to environmental challenges (Li et al. 2004). The open reading frame gene could possibly also function to open the chromatin for transcription of the globin genes (Ashe et al. 1997).

The highly variable and even fluctuating environment of fish such as Atlantic cod may create a high and varied demand for oxygen. The organization of the globin genes in a gene set may accelerate the expression of the chains grounding the hemoglobin protein. It is possible that the control elements and tandem repeats flanking the genes are somehow control elements for expression of both genes. That would prevent delay in forming the protein to ensure equal amounts of both control elements. It can be speculated that this organization of genes in a set facilitates the immediate response to high demand for oxygen.

The presence of duplicate genes is sometimes beneficial because additional amounts of protein or RNA products are provided that are important for a specific function, mainly to genes with products that are in high demand. Strong purifying selection against mutations that modify gene function can prevent such duplicated genes from diverging. Similarly, gene conversion can prevent divergence of such genes. Paralogous genes will have similar sequences after gene conversion. Thus, genes serving a single function in high demand are preserved (Zhang 2003). This mechanism may explain the high conservation of globin genes between fish and higher vertebrates.

Conserved Sequence Blocks in 5′ Region of β and α Genes

Conserved sequences are found both 5′ and 3′ to the β gene, and they are likely control elements of some kind. Their links to promoters and their same relative order with respect to microsatellites may indicate that. What role the conserved sequences and the conservation of the relative order of these various elements play, however, is not known. Nonetheless, the 5′–3′ order of expression of globin genes by arrangement in development in humans (Dickerson and Geis 1983) might give a clue. The conserved structure may be related to a joint expression of linked adjoining β and α genes in Atlantic cod. The 89-bp sequence (c2; Fig. 1) contains the CCAAT box, a conserved motif in globin promoters (Filipe et al. 1999). Therefore, these conserved sequences (c1, c2, c3; Fig. 1), located 5′ to both the β and α genes, might serve as binding sites for some transcriptional activators or complexes. Maruyama et al. (2004) looked for binding sites for some transcription factors and found them 3′ to many α genes but not in the corresponding region of the linked β genes. They found the conserved binding site NF-E2 and another putative binding site 3′ to the embryonic locus. In humans, HS-40 contains two such binding sites. This site is located in the intron of the same gene in different vertebrates, the C16orf35 gene. The length of the intron and the number of the C16orf35 homologs, however, varied between different teleosts. Thus, the arrangement in the Atlantic cod might differ from other teleosts analyzed. It can thus be hypothesized that the conserved sequences found in Atlantic cod could be a binding site for an unknown transcription factor. This hypothesis remains to be tested.
A 5′ Control Region of a β Gene

A part of the 5′ UTR sequence, about 1000 bp upstream of the β gene in the λ clone, showed 97% similarity to a pleurocidin-like gene in the winter flounder (Pseudopleuronectes americanus, GenBank AY282498.1). This indicates that our clone contains the most 5′ region of a globin pair, perhaps similar to the one found in pufferfish (Gillems et al. 2003). Alternatively, our β/α gene set may be one of a number of such gene sets that have large genetic regions between them, and such regions may in turn contain unrelated genes, as is known in zebrafish, mice, and humans (Brownlie et al. 2003; Bulger et al. 1999). The teleosts are members of evolutionary lineages that are considered to have undergone a genomic duplication via tetraploidization after the divergence of ray-finned and lobe-finned fishes (Hoegg et al. 2004). Independent gene or chromosome duplications, however, are significantly more frequent in each lineage of euteleosts than in mammals, or are lost less frequently (Robinson-Rechavi et al. 2001). Thus, information about gene duplication obtained in one fish lineage cannot be extended systematically to another (Robinson-Rechavi et al. 2001). The Atlantic cod lineage may thus be showing a pattern of globin gene arrangement different from that previously described among fish.

Variation in orientation of globin genes among different fish, as well as different orientation of various gene sets within a species, could be a coincidental result of gene and/or chromosome duplications. The role that different orientations play, however, will not be fully understood until the involvement of the control elements has been explained.

Proximal Regulatory Regions of the Genes

Different functions of hemoglobins in all kingdoms of organisms illustrate the acquisition of new roles by a pre-existing structural gene. Temporal and environmental regulation of expression is usually controlled by promoters and enhancers (Hardison 1998; Dickerson and Geis 1983).

In humans, a distal LCR is located 16 kb upstream of the β cluster (Sjakste and Sjakste 2002). An interaction between DNase I hypersensitive sites of the LCR and a promoter of a globin gene is believed to switch on expression (Sjakste and Sjakste 2002). The role of proximal regulatory regions in regulation of transcription in a gene system such as the globin clusters is of great importance.

In our study, a CCAAT regulatory motif is located 5′ to the first exon of all of the three genes reported. This motif is found in all the vertebrate globin gene promoters and can be bound by the CP1 complex (Hardison 1998). The CP1 binds more strongly to the CCAAT box in the α globin gene promoter than in the β globin gene promoter (Cohen et al. 1986; Hardison 1998).

The β gene has a TTTAAX box, the open reading frame in the intergenic region has a TTTAAX box, and the α gene has a TATA box in the promoter region. All of them are known as conserved sequences for the TATA binding protein in initiation of transcription.
To sum up, the proximal regulatory elements for linked β and α globin genes in Atlantic cod are CCAAT with TATA box promoters. This corresponds to the pattern found in other fish; however, the orientation of the genes in Atlantic cod is rare in the model species.

Hemoglobin loci have been thoroughly investigated in many vertebrates. The organization and arrangement of the genes differ between species. Nevertheless, many elements are conserved among them. The evolution of control elements is of importance to understand the coordinated transcription of the two chains that constitute the hemoglobin molecule. Teleosts are a diverse group, and a variety of control elements, promoters, and binding sites for transcription factors may have evolved among them. Thus, the conserved sequences found 5' to both the α and β gene in the Atlantic cod, as well as the microsatellites in the intergenic region, are important findings. They may represent variations of control elements found among teleosts. An understanding of that diversity can be gained with the comparative method, using Atlantic cod, a nonmodel teleost of great biological interest, and model systems.

Acknowledgments We thank Prof. Lars Pilstrom, Uppsala University, for giving us copies of a λ genomic library, and Prof. Jarle Mork at NTNU in Trondheim for tissue samples of genotyped individuals. The study was supported by a grant from the Icelandic Research Fund. We thank Dr. Petur Henry Petersen, and an anonymous reviewer for critical and helpful comments on the manuscript.

References

R Development Core Team (2006) R: a language and environment for statistical computing Vienna, Austria ISBN 3-900051-07-0


Multiple linked \( \beta \) and \( \alpha \) globin genes in Atlantic cod: a PCR based strategy of genomic exploration.
Katrín Halldórsdóttir and Einar Árnason, 2009
*Marine Genomics, 2*, 169–181

Author contributions: Katrín Halldórsdóttir and Einar Árnason conceived and designed the experiments, analysed the data, contributed analysis tools, prepared figures and/or tables and wrote the paper. KH performed the experiments. EÁ contributed reagents and materials.

© Copyright Elsevier. Printed with permission.
Multiple linked β and α globin genes in Atlantic cod: A PCR based strategy of genomic exploration☆

Katrín Halldórsdóttir *, Einar Árnason

Institute of Biology, University of Iceland, Sturlugata 7, Reykjavík, Iceland

A R T I C L E  I N F O

Article history:
Received 7 January 2009
Received in revised form 18 September 2009
Accepted 5 October 2009

Keywords:
Hemoglobin
Hemoglobin variability
PCR errors
β/α globin genes
Genomic exploration
Linkage disequilibrium

A B S T R A C T

Allozyme variation in Atlantic cod hemoglobins shows various signs of natural selection. We report a genomic exploration of globin genes in this non-model organism. Applying a PCR based strategy with a strictly phylogenetically informative sites we estimate the number of linked β and α globin genes. We estimate PCR error rate by PCR of cloned DNA and recloning and by analysis of singleton variable sites among clones. Based on the error rate we exclude variable sites so that the remaining variation meets successively stricter criteria of doubleton and triplet variable site. Applying these criteria we find ten clusters of linked β/α globin genes in the genome of Atlantic cod. Six variable amino acid changes in both genes were found in linkage disequilibrium with silent nucleotide substitutions. A phylogenetic tree, based on our strictly phylogenetically informative sites among 57 clones from 19 individuals, is split into two major branches by an amino acid change in a β gene. This change is supported by extensive linkage disequilibrium between the amino acid change and numerous other phylogenetically informative silent nucleotide sites. The different gene sets in the genome may represent different loci encoding different globins and/or allelic variation at some loci.

1. Introduction

Various environmental factors and niches encountered by fish make individual fitness depend on oxygen capacity. Thus, muscular activity during active swimming puts demands on oxygen supply. Also the capacity of maintaining neutral buoyancy at different depths is solved with a swim bladder and special types of hemoglobins. Root effect (Root, 1931) hemoglobins can deliver oxygen into the swim bladder, an organ containing a high partial pressure of oxygen (Berenbrink et al., 2005). Additionally, for fish having an avascular retina, vision demands oxygen particularly vision under low light conditions such as at great depths, and the retina of fish eye has high demand for oxygen (Pelster and Decker, 2004), Sick (1965) described a system consisting of two major zones of hemoglobin in Atlantic cod, Hbl and Hbl. The Hbl zone shows variation interpreted genetically as a polymorphism with a pair of co-dominant alleles giving rise to Hbl-1/1, Hbl-2/2 homozygotes and Hbl-1/2 heterozygotes. Andersen et al. (2009) explained the system by two amino acid replacements, the Met55/Val and Lys62/Ala. On the allozyme level this appeared to be a simple single locus two-allele polymorphism. By applying the more sensitive technique of iso-electric focusing (Brix et al., 2004; Fyhn et al., 1994), patterns of at least five major zones of hemoglobins are found with several minor zones as well. In Atlantic salmon, 17 electrophoretically distinct hemoglobin proteins have been characterized, grouped according to their migration towards anode and cathode as anodal or cathodal proteins. Expression of cathodal proteins increases with growth. Non-Bohr hemoglobin is one of the cathodal proteins, a hemoglobin for which pH does not affect O₂ affinity. Such multiple hemoglobins are quite common in fish (McMorrow et al., 1997). However, the developmental regulation of their expression is not known as in higher vertebrates (McMorrow et al., 1997).

The allozyme research on hemoglobin in Atlantic cod gives a strong indication of natural selection influencing the Hbl hemoglobin locus. Thus Frydenberg et al. (1965) described clinal variation in the frequencies of these phenotypes along the Norwegian coast. Norms-of-reaction differ between the FF and SS homozygotes (Karpov and Novikov, 1980; Petersen and Steffensen, 2003). Thus saturation of hemoglobin depends on temperature and oxygen pressure, which differ at different depth of the sea, and the genotypes show different responses under these conditions (Karpov and Novikov, 1980; Petersen and Steffensen, 2003). Andersen et al. (2009) also found strong relation between haplotypes and different habitats in the sea. Given the multiple hemoglobin isoforms observed and environmental correlation it is important to apply a genomic perspective to the globin system of Atlantic cod.

It is not known which of the hemoglobin isoforms are the Root or Bohr effect hemoglobins, or which are important at various developmental stages. In some species the two effects are shown by a single
hemoglobin e.g. in hemoglobin IV in trout (Binotti et al., 1971; Decker and Nadja, 2006). Verde et al. (2006) also showed that HB 3 displays a Bohr effect, which is enhanced by organophosphates, and a root effect, which is enhanced by ATP. The various different environmental conditions which Atlantic cod experience likely are met with a variety of hemoglobins. It is thought that all Root effect hemoglobins also show Bohr effect. But the reverse is not true. Therefore, the possibility exist that among some species there could be specific genes involved in the specific Root effects. The Atlantic cod hemoglobins apparently have e.g. extremely low histidine content, low buffer value and the highest pH-sensitive Root effect hemoglobins among teleosts (Berenbrink et al., 2005). Perhaps this is due to specific globins instead of effectors.

The genetic polymorphism found in the hemoglobin also is clearly related to selective factors in relation to various environmental conditions. These environmental challenges are likely met at the genetic level by various structural elements, control elements or both. Clearly, in order to understand how expression of different isoforms is regulated, a first step is to understand the structure of the gene involved at genomic level.

Model organisms are an extensively studied set of species of great importance for understanding of various topics in biology. Many important methods derive from studies of model organisms and for many model organisms complete genomic sequence is known. However, many non-model organisms have biological features or inhabit niches which are not found in any model organism. Therefore, they are of great comparative interest. For example the heterogeneous environment which Atlantic cod live in and its high variance in offspring number, likely due to selective forces (Arnason, 2004), makes it an interesting organism for studying fitness and natural selection. Understanding of the hemoglobin system in organism living from the surface of the sea to a depth of 600 m and thus has to deal with range in depth and pressure in its life history, has great value for our characterization of hemoglobin function in other teleost and in higher vertebrates. Also the model fish species, e.g. Zebrafish and Pufferfish, live in narrower depth range and thus might have less need for complex oxygen carrier molecules which make the comparison between their systems and the cod system informative.

In our previous study (Halldórsdóttir and Arnason, 2009) we characterized a set of linked \( \beta \) and \( \alpha \) globin genes in cod. These genes are oriented tail to head (5’ to 3’) with a putative single exon gene in the 1500 bp intergenic region. In this study, we address a question of genomic basis of the multiplicity of hemoglobin genes in Atlantic cod.

The approach is a mixture of PCR based strategy, population genetic and molecular evolution analysis for genomic exploration inquiring into the number of \( \beta \) and \( \alpha \) linked globin gene sets and potential allelic variation in Atlantic cod.

2. Materials and methods

2.1. Molecular work

All molecular work, PCR, sequencing and cloning conditions and methods have been described in Halldórsdóttir and Arnason (2009). Briefly we PCR amplified approximately 3000bp fragments using genomic DNA as template, TOPO-TA cloned the fragments, sequenced the clones with a set of primers giving overlapping sequencing and base called and assembled sequences into contigs using the Phred/Phrap/Consed software (Ewing and Green, 1998; Ewing et al., 1998; Gordon et al., 1998). (DNA sequence data presented in this work have been submitted to GenBank with accession numbers EF644859-\textendash}EF644911).

2.2. Source of genomic DNA

Samples from individuals, already genotyped for the Hbl locus (Sick, 1965) by iso-electric focusing technique, were obtained from Professor Jarle Mork at the University of Trondheim. The Hbl hemoglobin shows variation considered to be a polymorphism with a pair of co-dominant alleles (Sick, 1965). The homozygous genotypes Hbl-1/1 and Hbl-2/2 and a heterozygote genotype Hbl-1/2 were named after their relative movements in agar gel electrophoresis: SS (slow slow), FF (fast fast) and FS (fast slow) respectively.

2.3. cDNA

Total RNA was isolated from fresh blood from cod caught in Iceland using TRIZOL® Reagent (Invitrogen). The RNA was treated with DNase, using TURBO DNA Kit according to manufacturer’s instructions (Applied Biosystems) to remove traces of genomic DNA. First-strand cDNA synthesis was subsequently performed using ReverTaid™ First Strand cDNA Synthesis Kit (Fermentas K1621).

2.4. A PCR based strategy of genomic exploration

PCR products were amplified using conserved primers of a linked \( \beta \) and \( \alpha \) gene set (Halldórsdóttir and Arnason, 2009) from a number of individuals already genotyped for the Hbl locus and cloned into a PCR®4-TOPO vector (Invitrogen). Several clones were derived from three individuals (twelve from an FF, eleven from an SS, and twelve from an FS individual, Supplemental Table S1 in the Appendix). Two to four clones were derived from a few additional individuals and a single clone was derived from each of several other individuals. We fully sequenced clones from seven FF individuals, six SS individuals and six FS individuals, a total 57 clones from nineteen individuals. Thus the strategy was to cast the net widely and search for different globin genes within an individual of each genotype. By taking other individuals we looked for independent confirmation of variant gene sets from the three individuals. If the \( j/j \) gene sets (Halldórsdóttir and Arnason, 2009) occur as linked sets the expectation was that products from the FS heterozygotes would group with either FF or SS homozygotes. Singletons refer to variable sites found in a single individual thus representing a single PCR reaction. Doubletons refer to variable sites found in two or more individuals, thus representing two separate PCR reactions. Triplets refer to variable sites found among three or more individuals thus representing three PCR reactions.

2.5. Potential PCR and cloning errors

There was a potential for two kinds of errors in the PCR and cloning procedure. First, there exists a possibility for the polymerase to insert incorrect nucleotides in the molecules which were later cloned and sequenced. Second, the PCR elongation step may be terminated prematurely generating a molecule which might prime another PCR products and sequenced a single clone (repeat) (Fig. 1). PCR conditions were the same between original and repeat except for the source of DNA. We did this for seven clones and thus had seven pairs of sequences from original and repeat clones. The differences between the original and repeat sequences of each pair are due to errors in PCR and cloning. Furthermore, on the assumption that singleton variants observed among the clones represented PCR errors we also calculated the PCR error rate from singletons found among clones. To study the second kind of error we inspected the sequences generated to look for signs of chimeric molecules.
2.6. A strict criterion for phylogenetically informative sites

To further evaluate the polymorphism seen among clones of each individual as well as among individuals we applied stricter criteria. To be considered a phylogenetically (parsimony) informative site each nucleotide or an amino acid that varied from the rest had to appear at least in two individuals thus being derived from at least two separate PCR and cloning events (doubletons). A variable site with three or more different base pairs was also taken as phylogenetically informative site (nucleotide position 2585 in Table 2 for example). A variant observed in two or more clones of the same individual was not regarded as phylogenetically informative. Phylogenetically informative sites were acquired using MEGA (Kumar et al., 2004) and the results edited to conform to our strict criteria before being submitted to further analysis. In addition, we applied even stricter criteria for variant having to appear in three or more individuals thus being derived from three or more PCR reactions (triplets).

2.7. Data analysis

We aligned nucleotide and protein sequences with CLUSTALW (Thompson et al., 1994), blasted (http://www.ncbi.nlm.nih.gov/BLAST/) and used GENSCAN (Burge and Karlin, 1997) for gene prediction. The DNAPARS, DNADIST, NEIGHBOR and PROTDIST programs from the PHYLIP package were used to make parsimony, maximum likelihood and distance trees (Felsenstein, 2002) and the program MEGA (Kumar et al., 2004) to make Neighbor-joining phylogeny trees. The online ProtTest server (http://darwin.uvigo.es/software/prottest_server.html) was used to predict the best model for building trees based on Bayesian method and the program MrBayes was used to calculate the Bayesian trees. Variation in tandem repeat copy number among genotypes evaluated with etandem program (Rice et al., 2000) was tested with ANOVA. Nucleotide diversity (Π) and number of segregating sites (S) were estimated along the sequence in a sliding window of 100 nucleotides and step-size of 50 nucleotides using DnaSP (Rozas et al., 2003). This was done for all 57 original clones and the seven pairs of original and repeat clones and averages compared to evaluate PCR error rate. Linkage disequilibrium (Lewontin and Kojima, 1960) among sites was estimated with Fisher's exact test using DnaSP (Rozas et al., 2003) with Bonferroni adjustment for multiple tests (Rice, 1995).

3. Results

3.1. Estimates of PCR errors

Nucleotide diversity, Π, is contingent on frequency of different nucleotides at a position whereas the parameter segregating sites, S, is not. Nucleotide diversity Π estimated in a sliding window was an order of magnitude higher among the original clones than between original and repeat clones of the pairs of the PCR error study (Fig. 1). Peaks of nucleotide diversity seen among the original clones (Fig. 1) were not seen between the original and repeat pairs. In particular nucleotide diversity peaked at nucleotide positions 400–800 in exons 2 and 3 and 3′ untranslated region of the j1 gene (Fig. 1) because of relatively high frequency of clones carrying variation in this region.

The diversity seen between the original and repeat pairs was roughly equal over the 3071bp. This was an estimate of PCR errors (Fig. 1). Segregating sites did not show peaks but were more even over the fragment than Π with an order of magnitude difference between segregating sites of original clone and mean difference between original and repeats (Supplemental Figs. 51 in the Appendix). To further evaluate that our data represented a real signal and not PCR errors the exact error rate was calculated. There were altogether 25 variable sites among all seven original and repeat pairs, on average 3.57 variable sites in the 3071bp, 1.2×10\(^{-4}\) error rate per base pair. The probability of two such events to happen in two independent PCR from two individuals was 1.44×10\(^{-4}\) predicting 0.004 variable sites in two individuals. Similarly, the probability of the same error occurring in three individuals was 1.728×10\(^{-6}\) or miniscule. Even if we assumed a ten times higher error the probability of two independent events is 1.4×10\(^{-5}\) or 0.4bp in 3071bp.

Another way to estimate the PCR error rate was to look at the observed variation among all the clones, on the assumption that this variation represents the error in the experiment. The alignment of all 57 clones was 3071bp. A total of 399 segregating sites were found. Of these 133 resulted from length variation of microsatellites among individuals. Of the 266 remaining sites, 210 were singletons found in a single clone or in clones from a single individual. 56 were doubletons found in two or more individuals (Table 2) and of these 36 were triplets found among three or more individuals (Supplemental Table S2, Supplemental Fig. S2 in the Appendix). Assuming that singletons represent errors the PCR error rate was calculated. There were altogether 25 variable sites among seven original and repeat pairs, on average 3.57 variable sites in the 3071 bp, 1.2×10\(^{-4}\) error rate per base pair. The probability of two such events to happen in two independent PCR from two individuals was 1.44×10\(^{-4}\) predicting 0.004 variable sites in two individuals. Similarly, the probability of the same error occurring in three individuals was 1.728×10\(^{-6}\) or miniscule. Even if we assumed a ten times higher error the probability of two independent events is 1.4×10\(^{-5}\) or 0.4bp in 3071bp.

Transitions/transversions ratio in singletons were 185/25 or 12% transversions. Transitions/transversions ratio in singletons/doubletons were 185/25, 37/19 respectively (Table 1). The difference was highly significant (Χ\(^2\)=13.98, df=1, P=1.8×10\(^{-4}\)) Transitions/transversions

<table>
<thead>
<tr>
<th>Variable site</th>
<th>Transitions</th>
<th>Transversions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Singletons</td>
<td>15 75 24 71 4 6 3 7 5</td>
<td>6 3 2 4 2 2 1</td>
</tr>
<tr>
<td>Doubletons</td>
<td>6 11 11 9 5 2 3 4 2 2 1</td>
<td>4 2 2 1 2 1</td>
</tr>
<tr>
<td>Triples</td>
<td>4 8 7 1 4 2 2 4 1 2 1</td>
<td>2 2 1 2 1</td>
</tr>
</tbody>
</table>
ratio in singletons/triplets were 185/25, 20/16 respectively (Table 1). Again, the difference was highly significant ($X^2 = 21.14$, $df = 1$, $P = 4.26 \times 10^{-6}$).

Chimeric sequences (mimicking in vitro recombination) could also have occurred during DNA amplification (Gonzalez et al., 2005). By inspection of the data two instances had the possibility of indicating this. In Fig. 2 the FF127.12 in Cluster 2 and the SS103.11 in Cluster 5 shared the amino acid R instead of K (labelled x in Fig. 2). Similarly the FF127.4 in Cluster 2 and FS43.2 in Cluster 5 had the amino acid P in common instead of L (labelled xx in Fig. 2). An inspection of nucleotide sites in Table 2, we see that in order to account for the data in terms of chimeric sequences, a premature termination must have happened after the nucleotide number 2297 for FF127.4/FS43.2 (xx in Fig. 2) and at position 1620 or higher for FF127.12/SS103.11 (x in Fig. 2). We found no other obvious signs of possible chimeric sequences in our data.

3.2. Open reading frame in intergenic region

GENSCAN predicts a single exon open reading frame that largely coincided with the atg microsatellite locus in the intergenic region,
Table 2: Doubleton phylogenetically informative sites of an approximately 3 kb β/α globin gene region among 57 cloned contigs from genomic DNA of Atlantic cod.

<table>
<thead>
<tr>
<th>Clone</th>
<th>Segregating site</th>
<th>Intron region</th>
<th>Intergenic region</th>
<th>Segregating site</th>
<th>Intron region</th>
<th>Intergenic region</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>β gene</td>
<td></td>
<td></td>
<td>a gene</td>
<td></td>
<td></td>
</tr>
<tr>
<td>FF27.1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FF48.1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Cluster 2</td>
<td></td>
<td></td>
<td>a gene</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Cluster 3</td>
<td></td>
<td></td>
<td>a gene</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Cluster 4</td>
<td></td>
<td></td>
<td>a gene</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Cluster 5</td>
<td></td>
<td></td>
<td>a gene</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Cluster 6</td>
<td></td>
<td></td>
<td>a gene</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Cluster 7</td>
<td></td>
<td></td>
<td>a gene</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Cluster 8</td>
<td></td>
<td></td>
<td>a gene</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Cluster 9</td>
<td></td>
<td></td>
<td>a gene</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(continued on next page)
with a poly D (aspartic acid) amino acid sequence by codon gat. The microsatellite locus can therefore be called a gat repeat. The predicted gene has its own promoter, initiation and termination codons and a polyA signal and is thus a putative functional gene. No similarities are found to any known protein in GenBank. In some of the clones GENSCAN predicts the single exon of this putative gene to be the first exon of the α gene as an addition joined to the normal exon 1. In order to find out if this gene was really transcribed, we made primers which started 5′ in the interglobin gene and ended at specific amino acids predicted (individual FF127.9 and primer N and primer V in Table 4. We used cDNA made from DNAse treated RNA from whole blood as a template. As a control we made a forward primer in the interglobin gene based on sequence from an individual for which GENSCAN predict the interglobin exon as a single exon gene (individual SS104.1 and primer L in Table 4). We used these primers with a reverse α primer in a PCR reaction. Primer sequences are described in Halldórsdóttir and Árnason (2009). There was no amplification using primer L, but both primer N and primer V amplified a fragment. The amplified fragments were cloned and sequenced with universal primers M13F and M13R. The sequencing results show the amino acid sequence predicted (Table 4), and confirm that an mRNA combining interglobin and α gene is formed in vivo.

3.3. Phylogenetically informative sites

Our PCR based strategy was to cast the net widely within three individuals in the hope of finding different gene sets within an individual. By then casting the net in different individuals we hoped to find gene sets matching the different gene sets found within the three individuals. There were 56 phylogenetically informative doubleton sites according to our strict criteria, 13 at the β locus and 10 at the α locus (Table 2). The remaining phylogenetically informative sites were in the intergenic region.

3.4. Phylogenetic trees and clusters

We made phylogenetic trees using our strictly phylogenetically informative sites of all 57 clones (Table 2, Fig. 2). They were a maximum likelihood tree (DNAML Supplemental Fig. S3), a parsimony tree, (DNAPARS Supplemental Fig. S4) and a neighbor-joining tree.
of genetic distances (DNADIST and NEIGHBOUR Supplemental Fig. S5). The topologies of the trees were almost identical. Inspection of the trees revealed distinct clusters. The clones could be clustered ad hoc into ten clusters when maximum split up was done (Fig. 2 and Table 2). Clusters one to six all had an Aspartic acid, D, as the second variable amino acid variable in the β gene. The remaining cluster had an E. 

In Fig. 2 the sequence of amino acid variation (TETVYL a common pattern) were given for each clone in a cluster. The β chain change from D to E, which split the tree into two major branches, was supported with many silent variable sites in linkage disequilibrium with the amino acid change and several other sites (Fig. 3 and Table 2). Similarly, linkage disequilibrium between various sites supported most clusters (Fig. 3 and Table 2).

It is important to study the relationship of the genes we found to other genes in Atlantic cod and their relationship to adult and embryonic globin in other teleosts. A consensus sequence was made from clones in each cluster and used to build a phylogeny trees with known embryonic and adult α and β globin genes from several fish species (Neighbor-joining trees in Figs. 4 and 5). Consensus sequences of our ten clusters group with genes 2, 3, and 4 of Andersen and Birley and 2 of Verde. Hbt Andersen and Verde.1 are clearly different. There was, however, no obvious relationship of clusters to either embryonic of adult globins of other teleosts. The topology of the trees builds with the Maximum likelihood method (Supplemental Figs. S6 and S7) which show the same relationship among the globins as the distance and maximum likelihood trees.

The polymorphism found in coding region among amino acid residues were considerably higher at the β locus than at the α locus (Table 2). This was consistent with the sliding window pattern of
The nucleotide diversity (Fig. 1). The peaks of nucleotide diversity were higher at the β locus than the α locus. The nucleotide diversity peaked at the end of β exon 3 close to the location of the D to E change (Figs. 1 and 2). There was also considerable linkage disequilibrium among sites in this region of the fragment (Fig. 3). The non-synonymous polymorphism found in the α gene was less frequent than in the β gene (Fig. 1). The amino acid changes in the α gene were dispersed over the tree (Fig. 2).

The mean number of $\text{atg}$ tandem repeats (which forms the basis of the interglobin gene) were significantly different among the clusters (Table 3). Seven of the 56 strictly phylogenetically informative sites (Table 2) were at this microsatellite locus. These sites are part of the data used to make the trees and clusters. Therefore the degrees of freedom in the ANOVA we made were overestimated to some extent because of partial correlation of parts with whole. However, correcting for this is not likely to reverse this determinant and highly significant results.

4. Discussion

In this study we show multiple putative globin gene sets in Atlantic cod, a non-model organism. By our PCR based strategy of genomic exploration we found ten clusters of linked β and α globin like genes. The diversity of clusters observed is considered to represent multiple β and α gene sets as well as allelic variation of some of the loci of the gene sets.

Table 3

<table>
<thead>
<tr>
<th>$\text{atg}$ tandem repeat</th>
<th>df</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cluster</td>
<td>9</td>
<td>1187.1</td>
<td>131.90</td>
<td>44.6</td>
<td>0</td>
</tr>
<tr>
<td>Clones within cluster</td>
<td>47</td>
<td>139.1</td>
<td>3.0</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
4.1. PCR errors

To evaluate the authenticity of the variation defining the ten clusters we estimated potential PCR errors in order to exclude results due to technical artifacts. We used two methods to estimate PCR errors. First, the particular conditions in the PCR reactions with elongation step of ten minutes and using long primers (Hallósdóttir and Árnason, 2009), might facilitate polymerase errors. Pusch and Bachmann (2004) showed that ancient DNA may induce mutations in a PCR on a human DNA. Based on this we hypothesize that the extract of genomic DNA isolation may include some materials from the chelex isolation and precipitation procedures and various extraneous DNA sequences which can make the polymerase more error prone. Our first error estimate was based on cloned DNA as a template in a new PCR and subsequent cloning. All conditions are the same the only different is the DNA template. The cloned DNA is presumably cleaner than the genomic DNA and contains little or no extraneous DNA sequence which could induce PCR errors. This method estimates the minimum PCR error rate and is a measure of errors occurring in the PCR reactions and cloning procedures. The error rate according to this method is $1.2 \times 10^{-5}$ per base pair. On the assumption that all singleton variable sites among the original clones represent PCR errors, the error rate is $7.1 \times 10^{-6}$ or 60 times higher than from our estimate of original and repeat pairs. However, this probably is an overestimate because this high an error rate would alter the sequence beyond recognition. Using this rate, thus erring on the conservative side, the predicted number of errors of the 3000 bp fragment occurring in two separate PCR reactions from two individuals is 15. By this argument 15 of the 56 doubleton variable sites found in two or more individuals (Supplemental Table S2) would be considered PCR errors. Taking an even stricter stand by considering triplet variable sites found among three or more individuals thus having occurred in three or more independent PCR reactions the variable sites are reduced to 36. According to this maximal error rate there should be one triplet error found in three independent PCR reactions (Supplemental Table S2). Furthermore, according to Bracho et al. (1998), studying Taq polymerase induced errors in RNA virus diversity, the ratio of transition/transversion nucleotide substitutions because of Taq polymerase PCR errors is $83/19$ or 18.6%. This ratio among singleton sites is 185.25 or 12% (Table 1) in our data, a non-significant difference ($X^2 = 1.19$, df = 1, $P = 0.28$). However, the transition/transversion ratio is clearly different among our doubleton and triplet sites (Table 1). It is likely that they are different because of purifying natural selection. PCR errors, on the other hand, have not been subjected to selection. This difference justifies our assumption of considering singleton errors as therefore, we consider it likely that our doubleton and triplet variable sites represent authentic naturally occurring variation to a large extent. A network based on our strict criteria of independence of variable sites being found in two (Fig. 2) and three (Supplemental Fig. S2) individuals have the same topology of ten clusters. Therefore, we regard the ten clusters being defined by authentic variable sites. Although the possibility of PCR errors always remains, we consider it minimal among doubletons and triplets.

4.2. Potential allelic variation or different gene sets

Allelic variation at the Hbl locus is of interest because of the apparent balancing selection of the protein variants (Frydenberg et al., 1965). Therefore, by using genotyped individuals we could study nucleotide and amino acid variation in relation to the Hbl genotypes. Most of the clones of each of the three multiclone individuals cluster together in one cluster (Fig. 2). However, clones from each individual also are found in two other clusters. Thus, they all show more gene sets (clusters) than there are alleles at a single locus. The three multiclone individuals have their representative clones in cluster 8. The tree, based on our strictly phylogenetically informative sites, is split into two major branches with the amino acid change D--E in the $\beta$ gene. Although the amino acid replacement does not have large consequences for the protein the difference is in linkage disequilibrium with numerous other phylogenetically informative nucleotide sites. Our method of using repeat clones from an individual for genomic exploration is a pseudo-replication statistically. As two of the multiclone individuals contain D as the second variable amino acid of the $\beta$ gene (Fig. 2 and Table 2) the D--E change may be overrepresented. If this major split in the tree represents allelic variation at a single locus it would imply that we have found an ancient balanced polymorphism because the extensive linkage disequilibrium observed shows that many sites are involved (Fig. 3). The various clusters observed may possibly be a sign of allelic variation at some loci. However, neither this split nor other clusters are related to the Hbl polymorphism (Sick, 1965; Frydenberg et al., 1965), as is also clearly shown by Andersen et al. (2009). Also most of the FS individuals cluster together and have amino acid E instead of D as the second variable amino acid of the $\beta$ gene. The FS heterozygote individuals should cluster either with the FF or the SS homozygotes if the D--E split represented the two allelic model of Hbl (Sick, 1965). Furthermore, individual clones are found in clusters in both sites of the tree. The association of microsatellite variation and clusters (Table 3) suggests that the various clusters represent different gene sets or allelic variation at a locus other than the Hbl. Clones from the three multiclone individuals each appear in three clusters (Table 5). Therefore, because each individual only has two alleles of each gene the variation is not solely allelic variation at a single locus. Consider two scenarios. First, on the assumption that cluster 8, in which the three individuals have their representative clone, is a separate gene set the remaining six clusters might represent alleles of another $\alpha$ gene set. If it is the same
homologous gene set among the three individuals they would represent three different heterozygotes for one gene set: FF127: cluster1/cluster2; SS103: cluster5/cluster4; FS113: cluster9/cluster3 (Table 5). Under this assumption this particular gene set has at least six alleles. Other combinations of allelic variation are possible as well. Second, on the assumption that cluster 8 represents allelic variation of a gene set common to the three individuals in heterozygous condition with one of the other clusters of each individual there would be four alleles and up to four separate gene sets. Whichever way one counts, the results likely represent both allelic variation of at least one locus as well as at least two \( \alpha \)/\( \beta \) gene sets. These are all separate from the Hb locus (Andersen et al., 2009).

The more individuals we analyzed the more new clusters appeared. We analyzed multiple clones from three individuals and later added single clones from many individuals. Clusters 6, 7 and 10 do not contain any of the original clones. They contain clones from both homozygous and heterozygous HbI individuals. Clusters 3 and 8 have all genotypes, representing loci clearly independent of HbI genotype (Fig. 2 and Table 2). This raises the question of how many linked \( \beta \) and \( \alpha \) globin gene sets there are in the genome of Atlantic cod?

In the Zebrafish an \( \alpha \)/\( \beta \) embryonic globin locus is linked to an \( \alpha \)/\( \beta \) adult globin locus on the same chromosome with a 24k intergenic region (Brownlie et al., 2003). Chan et al. (1997) show that the linked \( \alpha \) and \( \beta \) genes are coordinately expressed. From this we can deduce that the \( \alpha \)/\( \beta \) gen sets in Atlantic cod possibly encode different embryonic or adult or both \( \alpha \)/\( \beta \) gene set loci. Known globin gene families (Karlsson and Nienhuis, 1985; Sjäkte and Sjäkte, 2002) consist of genes expressed at different developmental stages. From these facts we argue that some of our loci may contain embryonic or larval globin genes. However, blast analysis shows partial homology of our sequences to both embryonic and adult forms in different taxa. Consensus sequences from our clusters show no strong relation to either embryonic or adult genes in phylogeny trees we made using \( \alpha \) and \( \beta \) globin sequences in model and semi-model fish species (Figs. 4 and 5 and Supplemental Figs. S6, S7, S8 and S9). Also our clusters are very similar to genes 2, 3, and 4 of Andersen et al. (2009) which are expressed in adult fish. There is thus no clear indication for our genes that they represent embryonic or adult forms and they are not in concert with the gene encoding the HbI locus which also is expressed in the adult. Our results that the interglobin gene is expressed as part of the \( \alpha \) gene (Table 4) is clear indication of multiple and complex system of gene sets in the genome. The length of this single exon varies among the clones and the GENSCAN gene prediction changes according to the length of this single exon in the intergenic region. The potential function of such an elongated protein is not known. However, several hemoglobin pathologies involving elongated \( \alpha \) globin are known in humans, showing at least partial functioning of such products (Brennan et al., 2008). In primates, the \( \beta \) globin locus contains five genes which are arranged in the same order in which they are expressed during development. It has been suggested that distance from the locus control region (LCR) controls the order of expression of these genes (Johnson et al., 2005). Based on this suggestion, some regulation of expression of the globin genes could be related to the \( \alpha \to \beta \) tandem repeats in the intergenic region which form the main part of the single exon gene. Other research has also indicated that simple sequence repeat expansions and/or contractions can regulate gene expression and thus should be subjected to strong selective pressures (Li et al., 2004). However, in our case the repeated sequence has a promoter and a polyA signal and is thus a putative functional gene.

4.3. Known Bohr/Root effect sites

The question arises if some of our gene set encode for Root effect hemoglobin (Root, 1931). What causes the Root effect is still not clear. Perutz and Brunori (1982) suggested that the serine to cysteine P\( \gamma \)i replacement caused the Root effect. This has been questioned (Brittain, 2005). All our clones have serine at this site. The Asp99\( \beta \) and Asp101\( \beta \) have been considered to be the minimum structural requirement for establishing the Root effect in the Antarctic fish Trematodus newnesi (Mazzarella et al., 2006). All the clones have these three aspartates. In contrast to the Antarctic fish our clones have His146\( \beta \) and His147\( \beta \) which are important in Bohr and Root effects but are lacking in Root effect hemoglobin in Trematodus newnesi (Myl stock et al., 1996) and in trout hemoglobin I (Brittain, 2005).

All 57 clones had Ile67 E11 at the distal side of the heme which reduces the accessibility of oxygen to the iron (Mazzarella et al., 2006). Interestingly it is replaced by Ala in Andersen.1, Verde.1, A. glacialis.1b and B. saida.1b (Fig. 7). The Ile residue at this site is thought to block access of oxygen because of its bulky side chain (Giordano et al., 2007).

The phyloinformative amino acids in this study are not at sites of known Root effect (Figs. 6 and 7). Nevertheless some of the identified key amino acids residues specific to Root effect hemoglobins are found in the sequences. Based on the sequence variant we cannot unambiguously state if these are Root or Bohr effect hemoglobin or both. Andersen et al. (2009) described the genetic basis of the polymorphism at the HbI locus as two amino acid replacements, the Met55Val and Lys62I at the a\( \gamma \)i subunit interface. The haplotype of our clones is LecD5 and Alad2 like Andersen.3 and Andersen.4. In Fig. 7 Andersen.1 represents HbI-1/1 and Verde.1 represents HbI-2/2. The \( \alpha \) sequences fall into two distinct groups (6). They are our clusters along with Andersen.2, B. saida.1 and A. glacialis.1. The second group is Verde.1, Andersen.1, B. saida.2 and A. glacialis.1. Curiously, the B. saida and A. glacialis of group one differ from their group by two amino acids in exon 1 which they share with group 2. This likely represents convergent evolution.

The \( \beta \) sequences fall into three major groups (Fig. 7). The first group contains our clusters, Andersen.3 and 4. The second group contains Andersen.2, Verde.2, B. saida.2 and A. glacialis.2. The third group contains the HbI of Andersen.1, Verde.1, B. saida.1 and A. glacialis.1. However, the sequences within each group also differ at many sites. For example Andersen.2 and Verde.2 (in group two) have the haplotype Met55 and Ala62 and are therefore like a chimera of the two HbI haplotypes defined by Andersen et al. (2009). The sequences of Andersen.1 and Verde.1, two different haplotypes of the HbI locus, also differ at site 44 besides the difference in site 55 and 62. The two individuals from the two studies fulfill the criteria we made for strictly phylogenetically informative sites. Therefore, the sequences found by Verde and by Andersen either represents different gene sets or allelic variation. These facts and our results reflect the great amount of variation found among these genes.

A genomic perspective clearly has added a new dimension to globin variation in cod. Only with a further genomic study can we hope to fully resolve the complexity of the multiple isoforms.
Fig. 7. Structural alignment of Atlantic cod amino acid sequences of the β chains, including representative different clones of protein variants from this study, two from Andersen et al. (2009), and two from Verde et al. (2006) including also related species. Red boxes are phylogenetically informative sites. Helical and non-helical regions are indicated as in mammalian hemoglobins.
Acknowledgements

We thank Professor Jarle Mork at NTNU in Trondheim for tissue samples of genotyped individuals. The work was supported by a grant from the Icelandic Research Fund. We thank Dr. Petur Henry Petersen and anonymous reviewers for critical comments on the manuscript.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.margen.2009.10.001.

References


91
Trans-species polymorphism at antimicrobial innate immunity cathelicidin genes of Atlantic cod and related species.
Katrín Halldórsdóttir and Einar Árnason, 2015
PeerJ 3, e786

Author contributions: Katrín Halldórsdóttir and Einar Árnason conceived and designed the experiments, analysed the data, contributed analysis tools, prepared figures and/or tables and wrote the paper. KH performed the experiments. EÁ contributed reagents and materials.

© Copyright 2015 Halldórsdóttir and Árnason. Distributed under Creative Commons CC-BY 4.0
Trans-species polymorphism at antimicrobial innate immunity cathelicidin genes of Atlantic cod and related species

Katrín Halldórsdóttir and Einar Árnason

Institute of Life and Environmental Sciences, University of Iceland, Reykjavík, Iceland

ABSTRACT

Natural selection, the most important force in evolution, comes in three forms. Negative purifying selection removes deleterious variation and maintains adaptations. Positive directional selection fixes beneficial variants, producing new adaptations. Balancing selection maintains variation in a population. Important mechanisms of balancing selection include heterozygote advantage, frequency-dependent advantage of rarity, and local and fluctuating episodic selection. A rare pathogen gains an advantage because host defenses are predominantly effective against prevalent types. Similarly, a rare immune variant gives its host an advantage because the prevalent pathogens cannot escape the host’s apostatic defense. Due to the stochastic nature of evolution, neutral variation may accumulate on genealogical branches, but trans-species polymorphisms are rare under neutrality and are strong evidence for balancing selection. Balanced polymorphism maintains diversity at the major histocompatibility complex (MHC) in vertebrates. The Atlantic cod is missing genes for both MHC-II and CD4, vital parts of the adaptive immune system. Nevertheless, cod are healthy in their ecological niche, maintaining large populations that support major commercial fisheries. Innate immunity is of interest from an evolutionary perspective, particularly in taxa lacking adaptive immunity. Here, we analyze extensive amino acid and nucleotide polymorphisms of the cathelicidin gene family in Atlantic cod and closely related taxa. There are three major clusters, Cath1, Cath2, and Cath3, that we consider to be paralogous genes. There is extensive nucleotide and amino acid allelic variation between and within clusters. The major feature of the results is that the variation clusters by alleles and not by species in phylogenetic trees and discriminant analysis of principal components. Variation within the three groups shows trans-species polymorphism that is older than speciation and that is suggestive of balancing selection maintaining the variation. Using Bayesian and likelihood methods positive and negative selection is evident at sites in the conserved part of the genes and, to a larger extent, in the active part which also shows episodic diversifying selection, further supporting the argument for balancing selection.
INTRODUCTION

Vertebrates fight microbial infections using both innate immunity and adaptive responses. MHC molecules, cell surface molecules with broad (MHC-I) and specialized (MHC-II) pathogen recognition features (Murphy, Travers & Walport, 2007), show trans-species polymorphisms, variation indicative of adaptive balancing selection. For example, certain MHC-II alleles of humans are more closely related to certain alleles of chimpanzee than to other human alleles (Fan et al., 1989; Nei & Hughes, 1991). An ancient balanced polymorphism will generate long genealogical branches. Neutral variation will accumulate at sites close to the balanced polymorphic sites (Charlesworth, 2006). However, depending on recombination, the size of the genomic region can be quite short, making trans-species polymorphism hard to detect. Obvious and pervasive trans-species polymorphism, in contrast, is most likely due either to multiple sites under balancing selection or to suppression of recombination or to both (Winf et al., 2004). The models that have been proposed for detecting balancing selection in molecular data frequently assume that there is a single site under balancing selection. The silent and non-coding polymorphisms surrounding that site are taken as a signature of selection (Gao, Przeworski & Sella, 2013; Leffler et al., 2013). With the wealth of genomic data currently being generated, it is evident that many selective effects are related to immune defenses (Nielsen et al., 2007; Quintana-Murci & Clark, 2013; Teixeira et al., 2014; Osborne et al., 2013). Our understanding of balancing selection will be much improved by these new data, and important insights will be gained from genetic data without embarking on functional studies (Charlesworth, 2006).

Unique among vertebrates, the Atlantic cod (Gadus morhua) genome reveals the evolutionary loss of MHC-II and CD4, major parts of the adaptive immune system (probably they also are lost in other gadids, Star et al., 2011). Yet cod are healthy, playing a major ecological role in the North Atlantic, and are capable of sustaining large commercial fisheries. However, the way in which cod compensate for the lack of an adaptive immune response is unknown (Pilström, Warr & Strömberg, 2005; Magnadottir, 2010; Star & Jentoft, 2012). Host and parasite/pathogen interactions are very interesting in evolutionary terms. Pathogens set selective pressures on hosts and the response of the host is crucial for its own survival as well as the survival of the parasite. The innate immune system is at the forefront of this battle. It is of special interest to investigate evolution and variation of the innate immunity genes responsible for host defense.

Various families of antimicrobial peptides are an essential part of innate immunity. The cathelicidin family, first described in various mammals (Zanetti, Gennaro & Romeo, 1995), has been extensively studied in many organisms, e.g., primates (Zelezetsky et al., 2006) and fish (Maier et al., 2008; Kapralova et al., 2013). Important tools, such as Clnp-/- knockout mice, are available for functional studies (see e.g., Zhang et al., 2012). The number of genes coding for this protein varies among species. For example, there is a single gene in human (Gudmundsson et al., 1996) whereas there are ten in pig (Dawson et al., 2013). The protein is characterized by an N-terminus, a signal sequence, a conserved cathelin-like domain (exons 1, 2 and 3) and a C-terminal domain with antimicrobial activity (exon...
4). The N-terminus of the protein has certain conserved features that characterize all cathelicidins, i.e., four cysteine residues forming two disulfide bridges (Tomasinsig & Zanetti, 2005) (Fig. S1). This evolutionarily conserved part is, nevertheless, targeted by positive selection (Zhu, 2008) (Fig. S1). The C-terminus is highly variable within multigene families and among species, most likely due to diversifying balancing selection (Tomasinsig & Zanetti, 2005). Many innate immune molecules have been described in Atlantic cod, e.g., piscidin (Fernandes, Ruangsri & Kiron, 2010), beta-defensin (Ruangsrri et al., 2013) and the expanded toll-like receptor family (Sundaram et al., 2012), showing novel forms and patterns indicating importance of antimicrobial peptides and their genes for the immunity of these fish.

Several hypotheses have been proposed for the selective maintenance of high diversity at the MHC-II loci in vertebrates. These hypotheses include the heterozygote advantage hypothesis, the frequency-dependent rare-allele advantage hypothesis, and the fluctuating selection hypothesis under which the intensity of selective pressure can vary in accordance with the stimulus from pathogens. Thus, pathogen-driven episodic selection may vary in different environments and at different time periods (Clarke, 1962; Spurgin & Richardson, 2010; Sommer, 2005). However, the molecular signatures behind such balancing selection can be hard to detect and distinguish from other types of selection (Quintana-Murci & Clark, 2013).

Another example of unusually high polymorphism are the disease resistance R genes in Arabidopsis (Bakker et al., 2006). The mechanism behind extremely high gene copy number has been explained by the advantage of fixed heterozygosity based on duplicated genes each carrying different variants. This would give the advantage of overdominance without incurring any segregation load. In another study on R genes in the Arabidopsis, Shen et al. (2006) showed the effect of balancing selection in evolution of presence/absence polymorphism. In their study the R genes show different allele frequencies reflecting frequency-dependent selection at different stages of the evolutionary process.

Most genome-wide studies, scanning for variation, show high-frequency polymorphisms in genes related to immunity (Nielsen et al., 2007; Jeffler et al., 2013). In this study, we examine the Cathelicidin family of innate immunity genes in Atlantic cod in individuals from throughout the distributional range (Fig. 1), and in closely related species. We report large variation within and among species. We report a distinctive data set discovered when we attempted to amplify a particular Cathelicidin gene with a pair of primers designed from Atlantic cod sequences. Our initial aim was to study population variation at the single codCath1 locus previously described (Maier et al., 2008) and also found in the Atlantic cod genome sequence (Star et al., 2011). With only a single pair of primers we found extreme variation in 97 clones from 27 individuals. The amount and patterns of variation both within and among species cannot be explained as single locus variation. We discuss paralogous variation and the orthologous variation within paralogs in terms of trans-species polymorphism.
MATERIALS AND METHODS

Sampling

We used 97 clones from 27 individuals in the study. We isolated DNA from gill filament tissue for samples from Iceland and from fin clips tissue for all other specimens. There were 19 individuals of Atlantic cod (mnemonic: Gmo) from throughout the distributional range of the species: two each from Greenland (Gre), Barents Sea (Bar), Celtic Sea (Cel), Baltic Sea (Bal), Norway (Nor), Faroe Islands (Far), and Canada (Can) and five from around Iceland (Ice). We randomly sampled the individuals from our large sample collection (Árnason & Halldórsson, 2015) containing thousands of samples so as to cover a wide geographic area. We also included two individuals of each of the closely related species (Fig. 1) the Pacific cod Gadus macrocephalus (Gma), Greenland cod Gadus ogac (Gog), Walleye pollock Gadus chalcogrammus (Gch), and Polar cod Boreogadus saida (Bsa), which is more distantly related. Pacific cod is considered a speciation from an Atlantic cod invasion into the Pacific (Pac) at approximately 4 mya based on genomic mtDNA data, Greenland cod is a recent re-invasion of Pacific cod into the Arctic and Atlantic oceans, and Walleye pollock is a speciation from an Atlantic cod invasion into the Pacific at 3.8 mya (Coulson et al., 2006) (and see Carr et al., 1999; Pogson & Mesa, 2004). Labeling is as follows: Individuals are labeled with a six digit barcode, clones with a dash and a one or two digit clone number, species is labeled with the species mnemonic, and locality with the locality mnemonic.

The Icelandic Committee for Welfare of Experimental Animals, Chief Veterinary Office at the Ministry of Agriculture, Reykjavik, Iceland has determined that the research...
conducted here is not subject to the laws concerning the Welfare of Experimental Animals. (The Icelandic Law on Animal Protection, Law 15/1994, last updated with Law 157/2012.) DNA was isolated from tissue taken from dead fish on board research vessels. Fish were collected during the yearly surveys of the Icelandic Marine Research Institute. All research plans and sampling of fish, including the ones for the current project, have been evaluated and approved by the Marine Research Institute Board of Directors. The Board comprises the Director General, Deputy Directors for Science and Finance and heads of the Marine Environment Section, the Marine Resources Section, and the Fisheries Advisory Section. Samples were also obtained from dead fish from marine research institutes in Norway, the Netherlands, Canada and the US that were similarly approved by the respective ethics boards. The samples from the US used in this study have been described in Cunningham et al. (2009) and the samples from Norway in Árnason & Pálsson (1996). The samples from Canada consisted of DNA isolated from the samples described in Pogson (2001). The samples from the Netherlands were obtained from the Beam-Trawl-Survey (http://www.wageningenur.nl/en/Expertise-Services/Research-Institutes/imares/Weblogs/Beam-Trawl-Survey.htm) of the Institute for Marine Resources & Ecosystem Studies (IMARES), Wageningen University, the Netherlands, which is approved by the IMARES Animal Care Committee and IMARES Board of Directors.

Molecular analysis

We extracted genomic DNA using a Chelex/proteinase K extraction method (Walsh, Metzger & Higuchi, 1991). PCR was performed using Long PCR Enzyme Mix (Thermo Scientific/Fermentas #K0181) according to the manufacturer’s two-step cycling protocol. The PCR program was as follow: initial denaturation step of 3 min at 94 °C; 10 cycles of 20 s denaturation at 94 °C, 30 s annealing at 60 °C and 1.5 min extension at 68 °C. The annealing temperature was reduced by 2 °C in next two cycles and 1 °C in the following cycle, reaching annealing temperature of 55 °C. Additional 22 cycles were run under this condition with a 7 min final extension. Total 35 cycles.

The primers used for PCR were CodCathF1: 5′-TGTTCAGCACAAAGCCAAACT-3′ from Maier et al. (2008) and CodCathR4: 5′-GAGACAGGCTCAAGCCAATG-3′ a new reverse primer made for this study. The CodCathR4 primer was designed using Primer3 (Untergasser et al., 2012) and 3′ UTR part of GenBank sequence with accession number EU707291.1 as template.

Universal M13F and M13R primers were used for sequencing, using the BigDye® Terminator v3.1 Cycle Sequencing Kit (Applied Biosystems) according to the manufacturer’s protocols for plasmid sequencing except that we used 1/16 of the manufacturer’s recommended amount.

The PCR amplification fragments were gel extracted and cloned with PCR® 4-TOPO vector (Invitrogen™) and Sanger sequenced using an AB-3500XL Genetic Analyser (Applied Biosystems) (Halldórsdóttir & Árnason, 2009). All sequences were analyzed using the Phred/Phrap/Consed software suite (Ewing et al., 1998; Ewing & Green, 1998; Gordon, Abajian & Green, 1998) and had top-quality Phred score values (>30). Our initial goal
Table 1  Number of clones and number of forms or alleles in clones from different individuals. Individuals are labeled by species and sampling locality. Individuals showing three different forms or alleles are marked with **.

<table>
<thead>
<tr>
<th>nr</th>
<th>Barcode</th>
<th>Origin</th>
<th>Number of clones sequenced</th>
<th>Number of forms or alleles</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>105746</td>
<td>Gmo.Gre</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>104931</td>
<td>Gmo.Gre</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>140254</td>
<td>Gmo.Bar</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>140272</td>
<td>Gmo.Bar</td>
<td>8</td>
<td>3 **</td>
</tr>
<tr>
<td>5</td>
<td>118507</td>
<td>Gmo.Ice</td>
<td>12</td>
<td>3</td>
</tr>
<tr>
<td>6</td>
<td>125968</td>
<td>Gmo.Ice</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>7</td>
<td>118214</td>
<td>Gmo.Ice</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>8</td>
<td>117795</td>
<td>Gmo.Ice</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>9</td>
<td>117757</td>
<td>Gmo.Ice</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>10</td>
<td>140179</td>
<td>Gmo.Cel</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>11</td>
<td>140176</td>
<td>Gmo.Cel</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>12</td>
<td>140219</td>
<td>Gmo.Bal</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>13</td>
<td>140233</td>
<td>Gmo.Bal</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>14</td>
<td>152921</td>
<td>Gmo.Nor</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>15</td>
<td>152924</td>
<td>Gmo.Nor</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>16</td>
<td>115574</td>
<td>Gmo.Far</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>17</td>
<td>114718</td>
<td>Gmo.Far</td>
<td>6</td>
<td>2</td>
</tr>
<tr>
<td>18</td>
<td>200093</td>
<td>Gmo.Can</td>
<td>6</td>
<td>2</td>
</tr>
<tr>
<td>19</td>
<td>200079</td>
<td>Gmo.Can</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>20</td>
<td>103659</td>
<td>Bsa.Gre</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>21</td>
<td>104725</td>
<td>Bsa.Gre</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>22</td>
<td>103852</td>
<td>Gog.Gre</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>23</td>
<td>104947</td>
<td>Gog.Gre</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>24</td>
<td>152074</td>
<td>Gma.Pac</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>25</td>
<td>152050</td>
<td>Gma.Pac</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>26</td>
<td>152018</td>
<td>Gch.Pac</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>27</td>
<td>152027</td>
<td>Gch.Pac</td>
<td>3</td>
<td>3 **</td>
</tr>
</tbody>
</table>

Closely related species

| nr | Barcode | Origin | Number of clones sequenced | Number of forms or alleles |

was to sequence three clones from each individual to eliminate PCR errors according to a strategy that we discuss below and in Arnason & Halldórsdóttir (2015). The amplified fragment contained the whole gene, four exons and three introns with part of the 5’ and 3’ UTR (Fig. S2). We sequenced the gene and the 3’ UTR. EcoR1 digest of the clones run on agarose gels showed different sizes of the fragments in clones from some individuals. The size differences were confirmed upon sequencing. Therefore, we added and sequenced more clones from chosen individuals to further study the different sized fragments (see Table 1).
Data analysis

Errors occur during PCR amplification and inevitably will be found, mostly as singletons, in the sequences of the cloned DNA. To remove this source of variation from the data we initially had planned to use the strategy of Árnason & Halldórsson (2015) to get a consensus sequence for each individual from its three clones. However, the results showed that sequences of clones from some individuals were very different from each other, too divergent to be variation due to PCR errors. In some instances they belonged on the amino acid level to already described paralogous genes (Maier et al., 2008). Therefore, we revised the strategy for eliminating PCR errors by screening out singleton sites as follows. The three clones from each of the 27 individuals yielded 81 clones and, as already stated, we added extra clones for some individuals for a total of 97 clones. Singleton sites among the various clones from each individual that belonged to a certain cluster were considered PCR errors and not counted if that site was not found variable in clones from another individual (or other individuals). However, a singleton variant among the clones of an individual was considered a real SNP and was retained if that site was similarly variable in clones from other individuals (see Halldórsson & Árnason, 2009, for estimation of errors in replicate PCR reactions). If a single clone from a particular individual represented a different cluster (paralogous gene) from the rest of the clones from that individual, then that clone was included in the analysis. If the same form was present in all clones from an individual only one sequence was included in the analysis (a consensus sequence for that individual). Using this strategy we had 43 clones. Each singleton site in the data of the 43 clones analyzed here was considered a real variant because it was found in more than one clone in the original data of 97 clones. The 43 clones analyzed here contain a single representative clone from each individual for either each allele or each gene. We also present an analysis of the 97 clones for comparison. New sequences generated in this study have GenBank accession numbers KJ831349–KJ831391.

EST sequences from the Canadian Atlantic Cod Genomics and Broodstock Development project (Bowman et al., 2011) were used in the analysis for comparison on the protein level. These were GenBank Accession numbers EY975127.1 (based on mRNA from a Gadus morhua spleen SSH library enriched for genes up-regulated by formalin-killed atypical Aeromonas salmonicida), FG312333.1 (based on Gadus morhua blood library injected with polyriboinosinic-polyribocytidylic acid and formalin-killed Aeromonas salmonicida), and ES786338.1 (Gadus morhua spleen SSH library enriched for genes up-regulated by polyriboinosinic-polyribocytidylic acid). We also used GW862872.1 (based on mRNA from thymus from a Norwegian coastal cod, already charactized as cod Cathelicidin 2 in Maier et al., 2008) and EU707291.1 (complete cds from mRNA isolated from a wild caught cod from Iceland, previously characterized as cod Cathelicidin 1, codcath1 in Maier et al., 2008). Finally, we also included the complete gene sequence from GeneScaffold 2759 from the North East Arctic Atlantic cod genome sequence (Star et al., 2011) available on the Ensembl browser (Flicek et al., 2014).

Phylogenetic maximum likelihood trees were built using Muscle (Edgar, 2004) aligned sequences with a branch support of 100 bootstrap replicates using phyML (Guindon &
Table 2. Codon-based maximum likelihood and Bayesian analysis for positively selected sites in exon 4 and exons 1, 2, and 3 combined. Statistics with significance level $p < 0.05$, posterior probability $>0.9$ and Bayes Factor $>50$ are boldfaced. Consensus column summarizes methods which found the codon positively selected with significance level $p < 0.2$. Analysis was made using the Datamonkey server www.datamonkey.org (Delport et al., 2010; Pond, Frost & Muse, 2005).

<table>
<thead>
<tr>
<th>Codon</th>
<th>SLAC $d_N - d_S$</th>
<th>SLAC $p$-value</th>
<th>FEL $d_N - d_S$</th>
<th>FEL $p$-value</th>
<th>REL $d_N - d_S$</th>
<th>REL $p$-value</th>
<th>MEME Bayes F</th>
<th>MEME $p$-value</th>
<th>MEME $d_N - d_S$</th>
<th>FUBAR $p$-value</th>
<th>FUBAR $d_N - d_S$</th>
<th>Consensus</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$d_N - d_S$</td>
<td>$p$-value</td>
<td>$d_N - d_S$</td>
<td>$p$-value</td>
<td>$d_N - d_S$</td>
<td>$p$-value</td>
<td></td>
<td></td>
<td>$d_N - d_S$</td>
<td>$p$-value</td>
<td>$d_N - d_S$</td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>3.97</td>
<td>0.63</td>
<td>9.15</td>
<td>0.20</td>
<td>1.60</td>
<td>2.00</td>
<td>&gt;100</td>
<td>0.13</td>
<td>0.40</td>
<td>0.83</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>28</td>
<td>4.79</td>
<td>0.49</td>
<td>10.80</td>
<td>0.15</td>
<td>1.63</td>
<td>2.14</td>
<td>&gt;100</td>
<td>0.19</td>
<td>0.40</td>
<td>0.83</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>30</td>
<td>6.63</td>
<td>0.46</td>
<td>9.98</td>
<td>0.31</td>
<td>1.46</td>
<td>1.41</td>
<td>&gt;100</td>
<td>0.02</td>
<td>0.41</td>
<td>0.87</td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>45</td>
<td>−3.49</td>
<td>0.92</td>
<td>−12.31</td>
<td>0.35</td>
<td>−0.45</td>
<td>0.01</td>
<td>&gt;100</td>
<td>0.04</td>
<td>0.51</td>
<td>0.32</td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>47</td>
<td>4.92</td>
<td>0.46</td>
<td>7.18</td>
<td>0.21</td>
<td>1.61</td>
<td>2.21</td>
<td>&gt;100</td>
<td>0.046</td>
<td>0.37</td>
<td>0.83</td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>48</td>
<td>4.61</td>
<td>0.52</td>
<td>8.18</td>
<td>0.30</td>
<td>1.46</td>
<td>1.58</td>
<td>&gt;100</td>
<td>0.13</td>
<td>0.30</td>
<td>0.81</td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>49</td>
<td>5.50</td>
<td>0.41</td>
<td>13.24</td>
<td>0.12</td>
<td>1.65</td>
<td>2.23</td>
<td>&gt;100</td>
<td>0.16</td>
<td>0.55</td>
<td>0.86</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>51</td>
<td>6.80</td>
<td>0.37</td>
<td>12.22</td>
<td>0.10</td>
<td>1.71</td>
<td>2.36</td>
<td>&gt;100</td>
<td>0.11</td>
<td>0.84</td>
<td>0.94</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>55</td>
<td>4.79</td>
<td>0.48</td>
<td>10.78</td>
<td>0.13</td>
<td>1.68</td>
<td>2.45</td>
<td>&gt;100</td>
<td>0.03</td>
<td>0.42</td>
<td>0.84</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>57</td>
<td>7.27</td>
<td>0.32</td>
<td>12.63</td>
<td>0.11</td>
<td>1.68</td>
<td>2.21</td>
<td>&gt;100</td>
<td>0.14</td>
<td>0.74</td>
<td>0.92</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>59</td>
<td>4.94</td>
<td>0.45</td>
<td>8.24</td>
<td>0.19</td>
<td>1.63</td>
<td>2.23</td>
<td>&gt;100</td>
<td>0.22</td>
<td>0.29</td>
<td>0.81</td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>62</td>
<td>0.43</td>
<td>0.72</td>
<td>4.43</td>
<td>0.72</td>
<td>−0.01</td>
<td>0.02</td>
<td>&gt;100</td>
<td>0.01</td>
<td>0.40</td>
<td>0.66</td>
<td>+</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$d_N - d_S$</td>
<td>$p$-value</td>
<td>$d_N - d_S$</td>
<td>$p$-value</td>
<td>$d_N - d_S$</td>
<td>$p$-value</td>
<td></td>
<td></td>
<td>$d_N - d_S$</td>
<td>$p$-value</td>
<td>$d_N - d_S$</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>$d_N - d_S$</td>
<td></td>
<td>$d_N - d_S$</td>
<td>+</td>
</tr>
<tr>
<td>42</td>
<td>10.51</td>
<td>0.38</td>
<td>240.59</td>
<td>0.11</td>
<td>7.67</td>
<td>17.23</td>
<td>&gt;100</td>
<td>0.14</td>
<td>4.37</td>
<td>0.97</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>89</td>
<td>8.90</td>
<td>0.61</td>
<td>189.86</td>
<td>0.29</td>
<td>7.25</td>
<td>10.16</td>
<td>&gt;100</td>
<td>0.30</td>
<td>3.43</td>
<td>0.93</td>
<td>+</td>
<td>+</td>
</tr>
</tbody>
</table>

Gascuel, 2003) through Seaview (Gouy, Guindon & Gascuel, 2010). Translations of our original nucleotide data were performed with EMBOSs Transseq (http://www.ebi.ac.uk/Tools/st/emboss_transseq/). We used DIAP (Librado & Rozas, 2009) and R (R Core Team, 2014) and the ape, pegas, seqinr, ade4, adegenet and LDheatmap packages (Paradis, Claude & Strimmer, 2004; Paradis, 2010; Charif & Lobry, 2007; Dray & Dufour, 2007; Jombart & Ahmed, 2011; Shin et al., 2006) for population genetic and statistical analysis. We performed Discriminant Analysis of Principal Components (DAPC) with functions from the adegenet package. We used TexShade (Beitz, 2000) for visual presentation of alignments. We used ggtree (Yu, 2015, see https://github.com/GuangchuangYu/ggtree) based on ggplot2 (Wickham, 2009) to present secondary structure data.

For codon-based likelihood and Bayesian analysis of selected sites, we used the website www.datamonkey.org (Delport et al., 2010; Pond, Frost & Muse, 2005). The following methods were used to search for positively and negatively selected sites: MEME (Murrell et al., 2012), SLAC, FEL and REL (Kosakovsky Pond & Frost, 2005) and FUBAR (Murrell et al., 2013). Indels were excluded from the analysis of exon 4, and, therefore, only sites found in Cath2 that were common to the three genes were analyzed. The $p$-values were set at 0.2 for the SLAC, FEL and MEME programs to generate the results in Tables 2 and 3. The REL Bayes Factor was 50, and the FUBAR Posterior Probability was 0.9.
Table 3 Codon-based maximum likelihood and Bayesian analysis for negatively selected sites in exon 4 and in exons 1, 2, and 3 combined. Statistics with significance level \( p < 0.05 \), posterior probability \( > 0.9 \) and Bayes Factor \( > 50 \) are boldfaced. Consensus column summarizes methods which found the codon positively selected with significance level \( p < 0.2 \). Analysis was made using the Datamonkey server www.datamonkey.org (Delport et al., 2010; Pond, Frost & Muse, 2005).

<table>
<thead>
<tr>
<th>Codon</th>
<th>SLAC ( d_N - d_S )</th>
<th>SLAC ( p )-value</th>
<th>FEL ( d_N - d_S )</th>
<th>FEL ( p )-value</th>
<th>REL ( d_N - d_S )</th>
<th>REL Bayes F</th>
<th>FUBAR ( d_N - d_S )</th>
<th>FUBAR Post.Pr.</th>
<th>Consensus</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>−5.63</td>
<td>0.31</td>
<td>−15.55</td>
<td>0.08</td>
<td>−1.41</td>
<td>513.1</td>
<td>−0.63</td>
<td>0.79</td>
<td>−</td>
</tr>
<tr>
<td>7</td>
<td>−3.09</td>
<td>0.51</td>
<td>−5.92</td>
<td>0.58</td>
<td>−0.38</td>
<td>147.9</td>
<td>−0.133</td>
<td>0.51</td>
<td>−</td>
</tr>
<tr>
<td>13</td>
<td>−2.35</td>
<td>0.57</td>
<td>−2.43</td>
<td>0.81</td>
<td>−0.36</td>
<td>10777.7</td>
<td>−0.208</td>
<td>0.53</td>
<td>−</td>
</tr>
<tr>
<td>14</td>
<td>−12.35</td>
<td>0.14</td>
<td>−55.82</td>
<td>0.02</td>
<td>−0.67</td>
<td>16734.7</td>
<td>−3.53</td>
<td>0.96</td>
<td>−</td>
</tr>
<tr>
<td>31</td>
<td>−2.26</td>
<td>0.58</td>
<td>−2.17</td>
<td>0.83</td>
<td>−0.35</td>
<td>13704.8</td>
<td>−0.146</td>
<td>0.51</td>
<td>−</td>
</tr>
<tr>
<td>45</td>
<td>−3.49</td>
<td>0.50</td>
<td>−12.31</td>
<td>0.35</td>
<td>−0.45</td>
<td>102.0</td>
<td>−0.509</td>
<td>0.62</td>
<td>−</td>
</tr>
<tr>
<td>46</td>
<td>−9.93</td>
<td>0.12</td>
<td>−17.62</td>
<td>0.06</td>
<td>−1.28</td>
<td>2384580.0</td>
<td>−1.53</td>
<td>0.94</td>
<td>−</td>
</tr>
<tr>
<td>58</td>
<td>−5.88</td>
<td>0.37</td>
<td>−8.66</td>
<td>0.26</td>
<td>−0.65</td>
<td>197.32</td>
<td>−0.62</td>
<td>0.66</td>
<td>−</td>
</tr>
<tr>
<td>60</td>
<td>−2.18</td>
<td>0.56</td>
<td>−0.63</td>
<td>0.97</td>
<td>−0.02</td>
<td>112811000.0</td>
<td>−0.03</td>
<td>0.34</td>
<td>−</td>
</tr>
<tr>
<td>62</td>
<td>0.43</td>
<td>0.73</td>
<td>4.43</td>
<td>0.72</td>
<td>−0.01</td>
<td>52.6</td>
<td>0.40</td>
<td>0.25</td>
<td>−</td>
</tr>
<tr>
<td>68</td>
<td>0.04</td>
<td>0.71</td>
<td>1.10</td>
<td>0.92</td>
<td>−0.08</td>
<td>13381.4</td>
<td>0.09</td>
<td>0.32</td>
<td>−</td>
</tr>
</tbody>
</table>

Negatively selected sites in exons 1, 2, and 3 combined

<table>
<thead>
<tr>
<th>Codon</th>
<th>SLAC ( d_N - d_S )</th>
<th>SLAC ( p )-value</th>
<th>FEL ( d_N - d_S )</th>
<th>FEL ( p )-value</th>
<th>REL ( d_N - d_S )</th>
<th>REL Bayes F</th>
<th>FUBAR ( d_N - d_S )</th>
<th>FUBAR Post.Pr.</th>
<th>Consensus</th>
</tr>
</thead>
<tbody>
<tr>
<td>15</td>
<td>−11.59</td>
<td>0.24</td>
<td>−210.0</td>
<td>0.10</td>
<td>−7.08</td>
<td>338.4</td>
<td>−4.50</td>
<td>0.97</td>
<td>−</td>
</tr>
<tr>
<td>17</td>
<td>−15.07</td>
<td>0.11</td>
<td>−345.8</td>
<td>0.02</td>
<td>−7.15</td>
<td>579.5</td>
<td>−5.65</td>
<td>0.98</td>
<td>−</td>
</tr>
<tr>
<td>26</td>
<td>−8.34</td>
<td>0.39</td>
<td>−128.4</td>
<td>0.45</td>
<td>−6.38</td>
<td>68.0</td>
<td>−2.93</td>
<td>0.73</td>
<td>−</td>
</tr>
<tr>
<td>47</td>
<td>−7.54</td>
<td>0.33</td>
<td>−131.9</td>
<td>0.15</td>
<td>−7.1</td>
<td>501.4</td>
<td>−2.06</td>
<td>0.83</td>
<td>−</td>
</tr>
</tbody>
</table>

Due to the trans-species nature of variation, some analysis that are developed for intraspecific variation were made on the trans-species variation. The assumption here is that trans-species variation is representative of intraspecific variation that could be found with larger sample sizes within each species.

RESULTS AND DISCUSSION

Clusters of Cathelicidin variation

The variation clusters by tree building into three major groups (Fig. 2) that we call Cath1, Cath2, and Cath3. Cath1 has already been described as a single gene and characterized by Maier et al. (2008); Cath2, was originally described by Maier et al. (2008) based on a partial sequence from Canadian cDNA databank, and which we fully sequenced here. Cath2 was described as a paralogue of Cath1 (Maier et al., 2008). The third major group, Cath3, was novel and has not been described before. Only one of these genes, Cath1, was found in the cod genome assembly (www.ensemble.org, Star et al., 2011). However, the Atlantic cod genome sequence is incomplete with 611 Mb of 830 Mb assembled into scaffolds (Star et al., 2011) and probably is missing genes (Zhuang et al., 2012). Maier et al. (2008) had named a variant, for which they had found a cDNA sequence in GenBank, and that was characterized relative to Cath1 by a 10 amino acid indel, as Cath3. We found the same variant (117757_1_Gmo_Ice, Figs. 2 and 3) in our data as a variant of Cath1. Therefore,
Figure 2 Maximum likelihood phylogenetic tree of exon 4 with bootstrap values. Phylogenetic tree built on amino acid sequences of exon 4, the active peptide in cathelicidin, from 43 clones of various individuals of Atlantic cod and four sister taxa. Bsa.Gre (Boreogadus saida), Gch.Pac (Gadus chalcogrammus), Gma.Pac (Gadus macrocephalus), Gog.Gre (Gadus ogac) and Gmo (Gadus morhua) from various locations: Iceland (Gmo.Ice), Greenland (Gmo.Gre), Barents Sea (Gmo.Bar), Celtic Sea (Gmo.Cel), Baltic Sea (Gmo.Bal), Norway (Gmo.Nor), Faeroe Islands (Gmo.Far), Canada (Gmo.Can).

we drop the Cath3 label for this variant of Cath1 and henceforth use Cath3 for one of the major clusters of Figs. 2 and 3.

Orthologs and paralogs
An obvious question is whether these clusters represent orthologous or paralogous genes and alleles. Cath1 and Cath2 have already been established as paralogs (Maier et al., 2008). In our data clones from individual 118507.Gmo.Ice belonged to all three major clusters, Cath1, Cath2, and Cath3 (Fig. 2). Allelic variation at a single locus would only yield two forms in a diploid organism. Therefore, the three clusters must represent at least two paralogous genes. Similarly clones from Walleye pollock individual 152027.Gch.Pac also belonged to the three clusters (Fig. 2). Cath2 was most divergent. The Cath2 sequences,
Figure 3 Maximum likelihood phylogenetic tree of all clones with bootstrap values. Phylogenetic tree built on nucleotide sequences found in 97 clones from various individuals of Atlantic cod and four closely related taxa: Bsa.Gre (Boreogadus saida), Gch.Pac (Gadus chalcogrammus), Gma.Pac (Gadus macrocephalus), Gog.Gre (Gadus ogac) and Gmo (Gadus morhua) from various locations: Iceland (Gmo.Ice), Greenland (Gmo.Gre), Barents Sea (Gmo.Bar), Celtic Sea (Gmo.Cel), Baltic Sea (Gmo.Bal), Norway (Gmo.Nor), Faeroe Islands (Gmo.Far), Canada (Gmo.Can).

individuals in row 9–16 in Fig. 4 and Fig. S2, were considerably shorter than both Cath1 and Cath3 sequences or about 1210 bp long compared to about 1310–1368 bp (and see discussion on length variation below). Individual variation was found in repeats at the beginning of intron 3 and an indel in exon 4 in Atlantic cod from Celtic sea (individual 140179.Gmo.Cel). Compared to the other two groups Cath2 had deletions in intron 3 and exon 4 (Fig. S2). The amino acids sequence in exon 4, the active peptide, also were different.
Figure 4 Alignment of exon 4, the major peptide in cathelicidin, from various individuals of Atlantic cod and four closely related taxa. The sequences are grouped in accordance with the clades shown in Fig. 2. The first two groups are Cath3, the third group is Cath2, and the last group represents Cath1. Bsa.Gre (Boreogadus saida), Och.Pac (Gadus chalcogrammus), Gma.Pac (Gadus macrocephalus), Gog.Gre (Gadus ogac) and Gmo (Gadus morhua) from various locations; Iceland (Gmo.Ice), Greenland (Gmo.Gre), Barents Sea (Gmo.Bar), Celtic Sea (Gmo.Cel), Baltic Sea (Gmo.Bal), Norway (Gmo.Nor), Faeroe Islands (Gmo.Far), Canada (Gmo.Can). Up arrows represent positively selected sites and down arrows negatively selected sites in Tables 2 and 3. (Fig. S1 shows the same for the conserved part in exons 1–3).

Halldórsdóttir and Árnason (2015), PeerJ, DOI 10.7717/peerj.976

<table>
<thead>
<tr>
<th>Sequence</th>
<th>Codon</th>
<th>Residue</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bsa.Gre</td>
<td>1</td>
<td>Asp</td>
</tr>
<tr>
<td>Och.Pac</td>
<td>1</td>
<td>Arg</td>
</tr>
<tr>
<td>Gma.Pac</td>
<td>1</td>
<td>Lys</td>
</tr>
<tr>
<td>Gog.Gre</td>
<td>1</td>
<td>His</td>
</tr>
<tr>
<td>Gmo.Ice</td>
<td>1</td>
<td>Pro</td>
</tr>
<tr>
<td>Gmo.Gre</td>
<td>1</td>
<td>Tyr</td>
</tr>
<tr>
<td>Gmo.Bar</td>
<td>1</td>
<td>Tyr</td>
</tr>
<tr>
<td>Gmo.Cel</td>
<td>1</td>
<td>Tyr</td>
</tr>
<tr>
<td>Gmo.Bal</td>
<td>1</td>
<td>Tyr</td>
</tr>
<tr>
<td>Gmo.Nor</td>
<td>1</td>
<td>Tyr</td>
</tr>
<tr>
<td>Gmo.Far</td>
<td>1</td>
<td>Tyr</td>
</tr>
<tr>
<td>Gmo.Can</td>
<td>1</td>
<td>Tyr</td>
</tr>
<tr>
<td>Cath3</td>
<td>1</td>
<td>Ser</td>
</tr>
<tr>
<td>Cath2</td>
<td>1</td>
<td>Ser</td>
</tr>
<tr>
<td>Cath1</td>
<td>1</td>
<td>Ser</td>
</tr>
<tr>
<td>Cath3</td>
<td>2</td>
<td>Tyr</td>
</tr>
<tr>
<td>Cath2</td>
<td>2</td>
<td>Tyr</td>
</tr>
<tr>
<td>Cath1</td>
<td>2</td>
<td>Tyr</td>
</tr>
<tr>
<td>Cath3</td>
<td>3</td>
<td>Tyr</td>
</tr>
<tr>
<td>Cath2</td>
<td>3</td>
<td>Tyr</td>
</tr>
<tr>
<td>Cath1</td>
<td>3</td>
<td>Tyr</td>
</tr>
<tr>
<td>Cath3</td>
<td>4</td>
<td>Tyr</td>
</tr>
<tr>
<td>Cath2</td>
<td>4</td>
<td>Tyr</td>
</tr>
<tr>
<td>Cath1</td>
<td>4</td>
<td>Tyr</td>
</tr>
<tr>
<td>Cath3</td>
<td>5</td>
<td>Tyr</td>
</tr>
<tr>
<td>Cath2</td>
<td>5</td>
<td>Tyr</td>
</tr>
<tr>
<td>Cath1</td>
<td>5</td>
<td>Tyr</td>
</tr>
<tr>
<td>Cath3</td>
<td>6</td>
<td>Tyr</td>
</tr>
<tr>
<td>Cath2</td>
<td>6</td>
<td>Tyr</td>
</tr>
<tr>
<td>Cath1</td>
<td>6</td>
<td>Tyr</td>
</tr>
<tr>
<td>Cath3</td>
<td>7</td>
<td>Tyr</td>
</tr>
<tr>
<td>Cath2</td>
<td>7</td>
<td>Tyr</td>
</tr>
<tr>
<td>Cath1</td>
<td>7</td>
<td>Tyr</td>
</tr>
<tr>
<td>Cath3</td>
<td>8</td>
<td>Tyr</td>
</tr>
<tr>
<td>Cath2</td>
<td>8</td>
<td>Tyr</td>
</tr>
<tr>
<td>Cath1</td>
<td>8</td>
<td>Tyr</td>
</tr>
<tr>
<td>Cath3</td>
<td>9</td>
<td>Tyr</td>
</tr>
<tr>
<td>Cath2</td>
<td>9</td>
<td>Tyr</td>
</tr>
<tr>
<td>Cath1</td>
<td>9</td>
<td>Tyr</td>
</tr>
<tr>
<td>Cath3</td>
<td>10</td>
<td>Tyr</td>
</tr>
<tr>
<td>Cath2</td>
<td>10</td>
<td>Tyr</td>
</tr>
<tr>
<td>Cath1</td>
<td>10</td>
<td>Tyr</td>
</tr>
<tr>
<td>Cath3</td>
<td>11</td>
<td>Tyr</td>
</tr>
<tr>
<td>Cath2</td>
<td>11</td>
<td>Tyr</td>
</tr>
<tr>
<td>Cath1</td>
<td>11</td>
<td>Tyr</td>
</tr>
</tbody>
</table>

Figure 4: Alignment of exon 4, the major peptide in cathelicidin, from various individuals of Atlantic cod and four closely related taxa. The sequences are grouped in accordance with the clades shown in Fig. 2. The first two groups are Cath3, the third group is Cath2, and the last group represents Cath1. Bsa.Gre (Boreogadus saida), Och.Pac (Gadus chalcogrammus), Gma.Pac (Gadus macrocephalus), Gog.Gre (Gadus ogac) and Gmo (Gadus morhua) from various locations: Iceland (Gmo.Ice), Greenland (Gmo.Gre), Barents Sea (Gmo.Bar), Celtic Sea (Gmo.Cel), Baltic Sea (Gmo.Bal), Norway (Gmo.Nor), Faeroe Islands (Gmo.Far), Canada (Gmo.Can). Up arrows represent positively selected sites and down arrows negatively selected sites in Tables 2 and 3. (Fig. S1 shows the same for the conserved part in exons 1–3).
from the two other groups (Fig. 4). Thus, we consider Cath2 to be paralogous to the other clusters in accordance with Maier et al. (2008).

Furthermore, clones from individual 140272.Gmo.Bar belonged to both Cath2 and Cath3 (Fig. 2). Two Cath2 clones from this individual that differed by several sites, probably representing allelic variation at Cath2. This is further support that the Cath2 and Cath3 clusters represent paralogous genes. Clones from individual 140179.Gmo.Cel belonged to Cath2 and Cath3 (Fig. 2). The two Cath2 clones were identical and differed from the Cath2 of individual 140272 by several sites and an indel that is indicative of the variation among clones within the Cath2 cluster.

Clones from individual 104931.Gmo.Gre belonged to Cath1 and Cath2. There was only singleton variation, probably PCR error, between the two Cath2 clones. The Cath1 clone had very similar amino acid sequence to Cath1 clones from other individuals (Fig. 2) yet it differed somewhat at the nucleotide level (Fig. 3).

Clones from Pacific cod individual 152074.Gma.Pac belonged to both Cath1 and Cath3. If Cath1 and Cath3 are orthologous it would imply deeply divergent alleles at that locus. Similarly, clones from Pacific cod individual 152050.Gma.Pac belonged to both Cath1 and Cath3. The Cath3 clones (clones 1 and 2; Fig. 3) had identical amino acid sequence to clones from three other individuals: Greenland cod 103852.Gog.Gre, Atlantic cod 105746.Gmo.Gre and the other Pacific cod already mentioned 152074.Gma.Pac. At the nucleotide level the two Cath3 clones of 152050.Gma.Pac differed from each other by a few singleton sites that were probably due to PCR errors. It clustered with the other Pacific cod clones showing similar singleton variation at the nucleotide level (Fig. 3 and Fig. S2).

The sequences for different groups/alleles were of different sizes. The Cath3 cluster showed two subgroups or clades (A and B) that had some length differences. The first four clones in the alignment (Fig. 4 and alignment of the whole sequence in Fig. S2) are 1322 bp long except the clones of individual 152074.Gma.Pac which were 1237 bp long because of an indel in intron 3 and exon 4. The second subgroup or clade of Cath3 (the next four sequences in Fig. 4) were 1321, 1281, 1281 and 1276 bp long respectively due to length variation in intron 3 (Fig. S2). The Cath1 sequences, which constitute the rest of the sequences in Fig. 4, were from 1318 to 1368 bp long. Some variation was found in intron 3 (Fig. S2). For example, individual 152027.Gch.Pac had a long insertion but individuals 104947.Gog.Gre and 152050.Gma.Pac had deletions. Some minor variations were found in other individuals in intron 3, e.g., a repeats at the beginning of the intron. Individuals 104947.Gog.Gre and 152050.Gma.Pac had deletions in exon 4 but individuals 114718.Gmo.Far, 117757.Gmo.Ice, 105746.Gmo.Gre and 152074.Gma.Pac had insertions.

The three clusters probably represent functional genes. The cDNA sequences that we included are based on expressed sequences and they belonged to the Cath1 and Cath2 clusters. There were no signs of lack of function for Cath3.

From these considerations, we consider Cath2 to be a paralog of the Cath1 and Cath3 clusters. Based on the tree, the overall divergence between Cath1 and Cath3 was similar to the divergence of Cath2 from both Cath1 and Cath3 (Fig. 2). However, the sequence similarity is much higher between Cath1 and Cath3 than between Cath1 or Cath3 on one
hand and Cath2 on the other, both at the nucleotide and amino acid levels (Fig. 4, Figs. S1 and S2). Cath1 and Cath3 probably are paralogs although we do not have conclusive evidence for that deduction. However, if they are orthologs it will strengthen our main hypothesis of trans-species level of variation. Furthermore, one could argue that the two Cath3 clades represented paralogous genes. If that were the case, it would also strengthen our hypothesis of trans-species polymorphism because variation within both (A and B) forms of Cath3 clusters by alleles and not by species. The discriminant analysis of principal components (DAPC) lends further support that the variation clusters by alleles (Fig. 5) and not by species (Fig. 6). The DAPC cleanly separated groups defined by alleles but groups based on species were largely overlapping. We thus conclude that there are three paralogous genes, Cath1, Cath2, and Cath3, and that the variation within each cluster represents allelic variation of each gene. The most important result is the trans-species nature of the variation in that each allele group contains representatives of various species.

In some individuals we found representatives of only one gene or even of only a single allele. In some instances, we looked more closely at several clones of such individuals without detecting more alleles. This may be a chance event or it may be due to variation in primer binding sites. In that case, our data would have ascertainment bias from using only a single primer pair for PCR amplification. If that were the case, we are missing even more alleles. Similarly, a single Cathelicidin, Cath1, is found in the cod genome assembly (www.ensemble.org, Star et al., 2011) which may indicate a single gene in that individual. However, the incompleteness of the genome assembly also may explain that. A further exploration of the possibility of copy number variation is one avenue for further studies. For example, whole genome or targeted sequencing of individuals showing different forms of Cathelicidins could reveal if there is copy number variation. If so, selection might be on the level of gene number as is the case in presence/absence polymorphism in R genes in Arabidopsis (Shen et al., 2006). If a duplicated gene is being selected for or against, copy number variation may confound the detection of selection by the various methods we have used.

Trans-species polymorphic variation
The major feature of the results is that within each parologue the clones cluster by alleles and not by species. This is the hallmark of a trans-species polymorphism (Leffler et al., 2012; Leffler et al., 2013; Eimes et al., 2015). We have found trans-species polymorphisms of the cathelicidin genes and their alleles of Atlantic cod and closely related taxa that are akin to the human vs. chimpanzee MHC-II (Fan et al., 1989). The same topology was found for trees based on amino acid sequences of exon 4, the active part (Fig. 2), the amino acid sequences of exons 1–3, the conserved part, and, based on the nucleotide sequences for the whole genes (Figs. S2 and S3) for the 43 clones used. The tree based on nucleotide sequences of the complete genes for all 97 clones (Fig. 3) also showed the three distinctive groups clustering by alleles and not by species as also seen in the DAPC results as already stated. Thus, the profuse nucleotide and amino acid variation within each of the three paralogous genes fell into distinct clades with forms or alleles of the closely related species intertwined (Figs. 2–6 and Figs. S1–S4).
Figure 5 Discriminant Analysis of Principle Components (DAPC) scatterplot of the five allele clusters. Ten principle components and three discriminant functions were retained in the analysis. Scatterplot of the first two discriminant functions with eigenvalues used in black. The alleles are represented as dots of different shapes and colors representing the a priori groups Bsa (Boreogadus saida), and the Cath1, Cath2, Cath3-A and Cath3-B clusters of Fig. 2.

Signatures of gene conversion

Although no recombination was found by GARD, and visual inspection did not show four gametes, the sequences showed signatures of gene conversion (Lamb, 1984; Chen et al., 2007) (Fig. S2).

For instance, the individual clone 152027-1.Gch.Pac (individual eight in the Cath1 group in Fig. 4) clusters within Cath1. However, the first two highlighted amino acids (aa) are the same as in Cath3. The third aa highlighted in this individual, aa 42 (S), resembled that found in Boreogadus saida (the most distantly related taxon) and aa 48 (K) is identical to that of Cath2 for 152018-3.Gch.Pac. That aa is therefore unique for the Gadus chalcogrammus (Gch) species.

Halldórsdóttir and Árnason (2015), PeerJ, DOI 10.7717/peerj.976
Figure 6: Discriminant Analysis of Principle Components (DAPC) scatterplot of the five species clusters. Ten principle components and three discriminant functions were retained in the analysis. Scatterplot of the first two discriminant functions with eigenvalues used in black. The species are represented as dots of different shapes and colors representing the a priori groups of species: Bsa (Boreogadus saida), Gch Gadus chalcogrammus, Gma Gadus macrocephalus, Gmo Gadus morhua, and Gog Gadus ogac.

The peptides of clones of individuals 105746-3.Gmo.Gre and 152074-3.Gma.Pac in the Cath1 group (first two individuals in the Cath1 group in Fig. 4) have an insertion of five aa after site 24; they have L in site 51, as found in Cath2, a unique I in position 61 and K in position 66. There was thus unique allele of Cath1 found in two different species a clear case of trans-species variation.

The peptides of clones of individuals 152050-3.Gma.Pac (Gadus macrocephalus) and 104947-2.Gog.Gre (Gadus ogac) (individuals three and four in Fig. 4) show the same gap (or deletion) as in Cath2 (between sites 32 and 45) and R in position 24, also found in Cath2 and Cath3, they share unique aa in sites 54 and 66 (S and K) but after that position they resemble Cath1. These patterns are indicative of gene conversion. In this case, we have Halldórsdóttir and Árnason (2015), PeerJ, DOI 10.7717/peerj.976.
two alleles in Cath1 that are found in different species. These alleles are more closely related to each other than to other alleles from the same species, i.e., again a trans-species level of variation.

The aa sequence AYSIN at the C-terminus of the peptide is characteristic of the second of the two alleles of Cath3 (B) in our data (individual four to eight in the alignment in Fig. 4; the other allele (A) was characterized by the similar sequence AYIIN). However, this aa sequence also is found in the EST sequence FG312333.1 from Canada (individual six in Cath1 group in Fig. 4), which is clearly a Cath1 sequence elsewhere. This again is indicative of gene conversion and an indication of trans-species level of variation.

The peptide of individual 117757-1.Gmo.Ice (individual eleven in Cath1 group in Fig. 4) has the nine aa insertion that previously had been described as a paralogous gene Cath3 (Maier et al., 2008). According to our data it is an allelic variant of Cath1. Therefore, we drop the Cath3 label for this variant and reserve that for the major cluster (Fig. 2). Interestingly a shorter insertion of six aa (similar but not identical) was also found in individual 114718-4.Gmo.Far, an Atlantic cod from the Faeroe Islands.

**Population genetic statistics**

We estimated the nucleotide diversity \( \pi \), the scaled mutation rate \( \theta \) and Tajima’s \( D \) in a sliding window of 100 bp over the genes coding for Cath1 and Cath3, noncoding regions and both synonymous and non-synonymous sites in coding regions. For Cath1, \( \theta \) was higher than \( \pi \), giving a negative \( D \) over the whole gene (Figs. S5 and S6) with a high peak in exon 4 implying either purifying selection or demographic population expansion. Negative Tajima’s \( D \) can also indicate a selective sweep of positive selection and at several sites \( D < -2 \) was statistically significant. In contrast, for Cath3, \( \pi \) was generally larger than \( \theta \), giving a positive \( D \) for almost all sites, with high and significant peaks (\( D > 2 \)) in exon 4 (Figs. S7 and S8). This implies balancing selection or demographic population subdivision and bottlenecks. There also was much variation in non-coding regions, predominantly in introns. Intronic variation in the distinct clusters were in linkage disequilibrium with the non-synonymous variation found in exon 4 (Figs. 3 and Fig. S2).

We estimated linkage disequilibrium \( D' \) among highly polymorphic sites (with a minor allele frequency at least three sequences out of 36; Fig. 8 and two out of 22 in Fig. 7). We excluded low polymorphic sites for clarity. Cath1 alone showed linkage disequilibrium between sites in the active part (exon 4) and the conserved part (exon 1–3) and sites in intron 3 (Fig. 7). If we consider Cath1 and Cath3 as one orthologous gene and consider the variants from the various species simply as representative of allelic variation within any single species, we can estimate linkage disequilibrium among that group of clones (all alleles from Cath1 and the two Cath3 clusters independent of species Fig. 8). With these assumptions, we found even stronger linkage disequilibrium between sites in exon 4 and intron 3. Overall, this may indicate the presence of control sequences in intron 3. However, these overall summary statistics may miss important details of selection. Therefore, we decided to examine what a codon-based analysis, skipping intronic variation, might reveal about selection.
**Codon based analysis**

In order to screen for purifying or positive selection acting on the protein we used several routines in Datamonkey server: [www.datamonkey.org](http://www.datamonkey.org) (Delport et al., 2010; Pond, Frost & Muse, 2005). This server provides several methods for detecting various forms of selection (Tables 2 and 3). We screened alignments for recombination with GARD (Kosakovsky Pond et al., 2006) and found no sign of recombination.

We analyzed exons 1–3, the conserved part, separately from exon 4, which constitutes the active peptide. Sites containing gaps were excluded from this analysis. Therefore, the analysis was done only on sites found in all three groups. The analysis estimated synonymous \( (S) \) and non-synonymous \( (N) \) changes within each codon and calculated either the ratio \( d_S/d_N \) or the difference \( d_S - d_N \). For the codons with significant results, described below, both \( d_S \) and \( d_N \) were greater than zero. We compared several methods,
SLAC, REL, FEL, MEME and FUBAR (Kosakovsky Pond & Frost, 2005) to detect amino acid sites under selection (Tables 2 and 3).

The SLAC (Single Likelihood Ancestor Counting) program, the most conservative compared with the empirical Bayesian and likelihood approaches, found no evidence of selection. Similarly, FEL (Fixed Effects Likelihood), which is less conservative, found no evidence of selection. However, REL (Random Effects Likelihood) found no positively selected sites but found 11 and four negatively selected sites in exon 4 and exons 1–3, respectively. A REL Bayes factor higher than 10 is strong evidence of selection, giving support to positively selected sites in exons 1–3, as also found by FUBAR. REL is highly sensitive but has a tendency to produce false positives because of an a priori defined distribution of rates to be fitted; therefore, it may misinterpret a new distribution of rates (Kosakovsky Pond & Frost, 2005). FUBAR (Fast Unconstrained Bayesian AppRoximation, Murrell et al., 2013) uses MCMC to avoid constraints on the distribution of the selection
parameter. For FUBAR we consider a posterior probability of 0.95 as a stringent cutoff, 0.90 as a strong cutoff, and 0.80 as a suggestive cutoff. FUBAR found two positively and two negatively selected sites both in exon 4 and in exons 1–3 using the strong cutoff. Ten of twelve sites (Table 2) have posterior probabilities (for $\omega = \beta/\alpha > 1$ at a site) higher than the suggestive cutoff 0.8 (more than six-fold higher than the expected number of false positives of 1.6 with CI [0–4]). MEME (Mixed Effects Model of Evolution Marrelli et al., 2012) might be the most appropriate method for our data because this method detects selection varying across lineages and identifies episodic and pervasive positive selection. MEME detected five sites with $p \leq 0.05$ indicative of selection (Table 2). It is noteworthy that sites that are significant by one method (MEME) are not significant by another method (FUBAR) (sites 51 and 57 and the other way around for site 45). MEME can identify diversifying evolution in a subset of branches, where more restricted methods identify only purifying selection. Examples of this situation are sites 45 and 62 (Fig. 4 and Table 2), positively selected with $p < 0.05$ by MEME but negatively selected by REL.

Overall, the results of the exploratory codon-based analysis are in line with the results of the summary statistics ($\pi$ and Tajima’s $D$) indicating positive and balancing selection mainly in exon 4, the active part. Both results add support for the inference of balancing selection based on the trans-species nature of the within paralogs variation.

**Secondary structure predictions**

Given the support for diversifying selection it is worthwhile to ask if predictions of protein structure of the active peptide would add support for the role of selection. We used the RaptorX protein structure server (http://raptorx.uchicago.edu/, Källberg et al., 2012) to predict secondary structure of exon 4, the active peptide. This program can give some predictions of structure without the use of close homologs in the protein structure databases. Because of how diverse the peptides are, it is difficult to use more accurate programs like pymol which rely on close homology of the predicted and template proteins from protein structure databases.

The results of the analysis showed that most sequences were predicted as rod-like linear Glycine rich structures. In all three groups there were sequences which predicted $\alpha$ helical structures and among Cath3 sequences there also were predictions of beta-hairpin structures (Fig. 9). The sequence variation of the Glycine, Serine, and Arginine rich part of the peptide (Fig. 4) may be responsible for these differences in predicted structure.

The exact impact on the protein structure, of mutations between the highly different alleles, will not be described here. However, robust prediction of the secondary structures for the mature antimicrobial peptide part of the gene, show variation that may indicate different biological function of the proteins of these alleles to a variety of microbes (Fig. 9) (Tomasinsig & Zanetti, 2005; Zhu & Gao, 2009). The predicted peptides described here are highly cationic. Their size ranges from 50 to 81 amino acid residues. The more positively charged the peptides, the stronger they bind to bacterial membranes (Bals & Wilson, 2003). Most of the peptides have linear secondary structure which presumably prevent $\alpha$-helical conformation as is known for Proline rich peptides (Tomasinsig & Zanetti, 2005).
Figure 9: Predicted secondary structures of peptides in each group on a maximum likelihood phylogenetic tree of amino acid sequence of exon 4. Secondary structure predictions were made using the RaptorX protein structure server (http://raptorx.uchicago.edu, Källberg et al., 2012).
In mammals there is at least one cathelicidin peptide with α-helical conformation. This peptide folds into an amphipathic helical structure in connection with biological membranes (Tomasinsig & Zanetti, 2005). The first Cathelicidin identified in fish was from the Atlantic hagfish, Mysine glutinosa, with the mature peptide showing α-helical conformation (Uzzell et al., 2003). Few or any other Cathelicidins in fish have so far been shown to adopt α-helical conformation. In our data, we have prediction of peptides in all three groups i.e., Cath1, Cath2 and Cath3, which adopt this α-helical structure. Broekman et al. (2011a) made developmental expression studies with antibody from the mature peptide of Cathelicidin 1 in Atlantic cod. They showed that the peptide has broad activity against different stimuli (Broekman et al., 2011b). Interestingly, the antibody they use was raised against the 14 amino acids which do not differentiate the three groups that we describe here (Broekman et al., 2011b). Therefore, it will be of interest to test whether the different forms described here have different activities and whether that could explain the broad activity they found. These future studies of the activity of the different peptides, will also be very interesting in the context of the rising interest in fish antimicrobial peptides in clinical dermatology (Rakers et al., 2013) and therapeutic antimicrobials (Masso-Silva & Diamond, 2014).

Spatial population differentiation

There has been a long-standing debate about the possible population differentiation of Atlantic cod (Jónsdóttir et al., 1999; Árnason, 2004; Eiríksson & Árnason, 2013). The genes behind primary defense against pathogens, like cathelicidin, are presumably under strong selection. It is expected that such loci will show pattern of geographic subdivision in contrast to loci with genome wide effect which relay demographic effects. However, there is no particular geographic structure evident among localities by visual inspection. For example, three individuals of Atlantic cod from Faroes, Norway and Canada show one of the alleles found in Cath1 (three aa highlighted in individuals 115574-2.Gmo.Far 152924-2.Gmo.Nor 200093-5.Gmo.Can in Fig. 4). In general the different specific variants were widely dispersed as expected of allelic variation of an ancient polymorphism.

Balancing selection

The shared polymorphism within paralogs found in our data, e.g., between Atlantic cod and Walleye pollock, suggests long-lasting maintenance by balancing selection. A trans-species polymorphism is in general a most important indication of balancing selection (Charlesworth, 2006). With an approximately five-year generation time and an effective population size (Ne) of approximately 10,000 in Atlantic cod (Árnason, 2004), the approximately 4 mya divergence time between the species (Coulson et al., 2006) is 20Ne, or five times higher than the average 4Ne fixation time for neutral variation (Clark, 1997). Such long-lasting trans-species polymorphism is often thought to be indicative of balancing selection (Hughes, 2002; Sommer, 2005). These considerations are based on the time scale of the Kingman coalescent (Kingman, 1982). The faster time scales of the multiple-merger coalescent, which are more appropriate for the high fecundity Atlantic cod.
Cod (Birkner, Blath & Eldon, 2013; Árnason & Halldórsson, 2015), would make this even more significant.

We show that the polymorphism is older than speciation given that divergent alleles of different paralogous genes can be found in different species. The balancing selection hypothesis is a plausible explanation because a scenario of concerted evolution between paralogous genes would otherwise be expected (Liao, 1999).

CONCLUSION

Trans-species polymorphism is in general strong evidence for balancing selection. We found a highly variable polymorphism at antimicrobial Cathelicidin loci with trans-species level of variation that suggests maintenance by some form of balancing selection. Given the functional role of the cathelicidin peptides and the diverse structures predicted the system may play an important role in a host/pathogen arms race. This may imply that negative frequency dependent and possibly episodic selection may be responsible for the balancing selection.

Further experiments are needed to test the activity of various cathelicidin peptides against a variety of microbes to both elucidate the mechanisms of selection (Nielsen et al., 2007; Quintana-Murci & Clark, 2013) and to better understand the expression of the various genes in relation to microbial infection. Further intra- and interspecific experiments are also needed to find out if there are more paralogous genes in the genome (c.f. pigs Dawson et al., 2013) and to establish their paralogous and orthologous relationships. This should include population genetics studies using stringent experimental protocols to avoid PCR and cloning artifacts (c.f. Lenz & Becker, 2008).

Using a phylogenetic analysis Star & Jentoft (2012) show an expansion of MHC-I and various Toll like receptor genes coinciding with the loss of MHC-II (Star et al., 2011). Atlantic cod may thus compensate evolutionarily for the loss of MHC-II. Our results imply evolutionary forces shaping variable innate immunity under selection pressure from contacts between hosts and microbes (Barreiro & Quintana-Murci, 2010; Quintana-Murci & Clark, 2013) in a manner similar to what is known for the MHC-II genes conferring adaptive immunity in other vertebrates. Such an extensive polymorphism of antimicrobial peptides has not been previously described in fish. Here, dynamic natural selection at hotspots of individual primary defenses may indicate the added importance of innate immunity when adaptive immunity is lacking.

ACKNOWLEDGEMENTS

We thank Jarle Mork (Norwegian University of Science and Technology), Kristján Kristjánsson (Marine Research Institute in Reykjavik), Grant Pogson (University of California at Santa Cruz), Remment ter Hofste (Institute for Marine Resources and Ecosystem Studies in the Netherlands), and Michael Canino (National Oceanic and Atmospheric Administration) for help in securing some of the samples. We thank Valerie H. Maier and Arnar Pálsson for their valuable comments on the manuscript. We thank RC Lewontin for office space and encouragement and stimulating environment and discussion while writing the manuscript.
ADDITIONAL INFORMATION AND DECLARATIONS

Funding
Funding was provided by Icelandic Science Foundation grant of excellence (nr. 40303011), a University of Iceland Research Fund grant, a Svala Árnadóttir private foundation grant to Einar Árnason and a doctoral grant from the University of Iceland Research Fund to Katrín Halldórsdóttir. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Grant Disclosures
The following grant information was disclosed by the authors:
Icelandic Science Foundation grant of excellence: 40303011.
University of Iceland Research Fund.
Svala Árnadóttir private foundation.

Competing Interests
The authors declare there are no competing interests.

Author Contributions
• Katrín Halldórsdóttir and Einar Árnason conceived and designed the experiments, performed the experiments, analyzed the data, contributed reagents/materials/analysis tools, wrote the paper, prepared figures and/or tables, reviewed drafts of the paper.

Animal Ethics
The following information was supplied relating to ethical approvals (i.e., approving body and any reference numbers):
The Icelandic Committee for Welfare of Experimental Animals, Chief Veterinary Office at the Ministry of Agriculture, Reykjavík, Iceland has determined that the research conducted here is not subject to the laws concerning the Welfare of Experimental Animals. (The Icelandic Law on Animal Protection, Law 15/1994, last updated with Law 157/2012.) DNA was isolated from tissue taken from dead fish on board research vessels. Fish were collected during the yearly surveys of the Icelandic Marine Research Institute. All research plans and sampling of fish, including the ones for the current project, have been evaluated and approved by the Marine Research Institute Board of Directors. The Board comprises the Director General, Deputy Directors for Science and Finance and heads of the Marine Environment Section, the Marine Resources Section, and the Fisheries Advisory Section. Samples were also obtained from dead fish from marine research institutes in Norway, the Netherlands, Canada and the US that were similarly approved by the respective ethics boards. The samples from the US used in this study have been described in Cunningham et al. (2009) and the samples from Norway in Árnason & Pálsson (1996). The samples from Canada consisted of DNA isolated from the samples described in Pogson (2001). The samples from the Netherlands were obtained from the Beam-Trawl-Survey (http://www.wageningenur.nl/en/Expertise-Services/Research-Institutes/imares/Weblogs/)
DNA Deposition
The following information was supplied regarding the deposition of DNA sequences:

New sequences generated in this study have GenBank accession numbers KJ831349–KJ831391.

Supplemental Information
Supplemental information for this article can be found online at http://dx.doi.org/10.7717/peerj.976#supplemental-information.

REFERENCES


---


Nucleotide variation and balancing selection at the Ckma gene in Atlantic cod: analysis with multiple merger coalescent models.
Einar Árnason and Katrín Halldórsdóttir, 2015
PeerJ 3, e786

Author contributions: Katrín Halldórsdóttir and Einar Árnason conceived and designed the experiments, analysed the data and wrote the paper. KH conceived and directed the molecular work and BAC library screening. KH performed part of that work and prepared and submitted sequences to GenBank. EÁ prepared figures and/or tables and contributed reagents/materials/analysis tools.

© Copyright 2015 Árnason and Halldórsdóttir. Distributed under Creative Commons CC-BY 4.0
Nucleotide variation and balancing selection at the Ckma gene in Atlantic cod: analysis with multiple merger coalescent models

Einar Árnason and Katrín Halldórsdóttir
Institute of Life and Environmental Sciences, University of Iceland, Reykjavík, Iceland

ABSTRACT

High-fecundity organisms, such as Atlantic cod, can withstand substantial natural selection and the entailing genetic load of replacing alleles at a number of loci due to their excess reproductive capacity. High-fecundity organisms may reproduce by sweepstakes leading to highly skewed heavy-tailed offspring distribution. Under such reproduction the Kingman coalescent of binary mergers breaks down and models of multiple merger coalescent are more appropriate. Here we study nucleotide variation at the Ckma (Creatine Kinase Muscle type A) gene in Atlantic cod. The gene shows extreme differentiation between the North (Canada, Greenland, Iceland, Norway, Barents Sea) and the South (Faroe Islands, North-, Baltic-, Celtic-, and Irish Seas) with $F_{ST} > 0.8$ between regions whereas neutral loci show no differentiation. This is evidence of natural selection. The protein sequence is conserved by purifying selection whereas silent and non-coding sites show extreme differentiation. The unfolded site-frequency spectrum has three modes, a mode at singleton sites and two high frequency modes at opposite frequencies representing divergent branches of the gene genealogy that is evidence for balancing selection. Analysis with multiple-merger coalescent models can account for the high frequency of singleton sites and indicate reproductive sweepstakes. Coalescent time scales vary with population size and with the inverse of variance in offspring number. Parameter estimates using multiple-merger coalescent models show that times scales are faster than under the Kingman coalescent.

INTRODUCTION

High fecundity translates into large excess reproductive capacity that would allow organisms to withstand substantial natural selection enabling them to bear the entailing high genetic load. Based on the concept of the cost of natural selection (Haldane, 1957) high-fecundity organisms relative to low-fecundity organisms should at any time be able to adapt a larger proportion of their genome to meet various environmental challenges. Trying to explain the paradox of sexual reproduction Williams (1975) in his Sex and Evolution book argues that high-fecundity coupled with heavy mortality of young...
(type III survivorship) may be able to pay the 50% fitness cost of meiosis. He developed several models, such as the Elm/Oyster and the Cod/Starfish models, which emphasize the importance of high-fecundity for selection. Williams also discussed the concept of reproductive sweepstakes. There is no heritability of fitness and sexual reproduction continuously assembles Sisyphean genotypes (from Sisyphus who was punished to roll a boulder up a hill only to see it roll back down, and having to repeat his actions forever). The distribution of offspring numbers is highly skewed, heavy-tailed and with high variance (lognormal). That is Williams’s fitness distribution. The environment factors are envisioned as acting in a sequence of selective filters. With only a few factors (e.g., temperature, salinity, etc.) there nevertheless can be an enormous number of different sequences of selective filters (environments) that do not recur. Hence a winning genotype is not permanent and must be continuously reassembled. Natural selection increases the variance in offspring number and thereby reduces effective population size genome-wide. Neutral variation will therefore drift faster under pervasive natural selection.

Coalescent theory (Kingman, 1982a; Kingman, 1982b) traces the genealogy of a sample and is very useful for making statistical inferences from molecular population genetic data. However, in an extreme case under a winner-take-all sweepstakes reproduction, all samples would coalesce immediately in the previous generation (Arnason, 2004) and there would be no variation. However, this extreme case is not realistic. The Kingman coalescent, which is derived from Wright/Fisher models of low fecundity non-skewed offspring distributions, assumes a bifurcating genealogy and is not appropriate for reproduction of this kind (Eldon & Wakeley, 2006; Schweinsberg, 2003; Wakeley, 2013; Tellier & Lemasse, 2014). Some organisms may exhibit both high fecundity and highly skewed offspring distributions. The coalescent for such organisms will lie somewhere between the extreme of a winner-take-all sweepstakes coalescent and the Kingman coalescent. For these organisms the \( \Lambda \) coalescent allowing multiple mergers of ancestral lineages at any one generation (Pitman, 1999; Sagitov, 1999; Donnelly & Kurtz, 1999; Eldon & Wakeley, 2006; Schweinsberg, 2003; Sargsyan & Wakeley, 2008) or \( \Xi \) coalescent allowing simultaneous multiple mergers of ancestral lineages (Schweinsberg, 2000; Möhle & Sagitov, 2001) may be more appropriate. Wakeley (2013) gives an overview of the development of coalescent theory in new directions. There is also active development of statistical inference methods associated with multiple merger coalescents (e.g., Birkner, Blath & Eldon, 2013b; Eldon et al., 2015). Studies on the high fecundity organisms Pacific oyster Crassostrea gigas (Hedgecock & Pudovkin, 2011) and Atlantic cod Gadus morhua (Linnaeus, 1758) (Arnason & Pálsson, 1996; Arnason, Pálsson & Petersen, 1998; Arnason et al., 2000; Carr & Marshall, 1991a; Carr et al., 1995; Pepin & Carr, 1993; Arnason, 2004) have provided data for a number of tests of some of the new coalescent models (Eldon & Wakeley, 2006; Eldon, 2011; Eldon & Degnan, 2012; Steinrücken, Birkner & Blath, 2013; Birkner, Blath & Eldon, 2013b). A high number of singletons is a feature of sequence studies of high fecundity organisms such as the Atlantic cod. This is expected under models of multiple merger coalescents and, therefore, they perform better than the Kingman coalescent by capturing the high frequency of singletons. Atlantic cod thus provides a model for studies applying the multiple merger coalescent.
In this paper we apply some of these new methods for $\Lambda$ coalescents as appropriate neutral null models for high fecundity organisms in a study of balancing selection at a gene showing extreme spatial differentiation in Atlantic cod.

In general the time scale of multiple merger coalescent models can be much shorter than for a Kingman coalescent. For example, under the Beta$(2 - \alpha, \alpha)$ coalescent model time scales depend on $N^{\alpha - 1}$ (Schweinsberg, 2003; Eldon et al., 2015). Under the extreme winner-take-all sweepstakes coalescent mentioned above, all individuals would be sibs differing only by new mutations. However, this is an extreme case. Real world multiple merger coalescents lie somewhere between this extreme and the Kingman coalescent. Applying multiple merger coalescents does not imply that we are sampling siblings or that samples from say Greenland and Norway share the same parents. Möhle (1998) and elsewhere shows that for large $N$ the Kingman coalescent is robust because the influence of structure of various types (selfing, age structure, geographic structure) occurs on a shorter time scale than the time scale of the coalescent. Under the Kingman coalescent, the expected time to the most recent coalescence of a sample of $n$ individuals is $2/(\alpha(n - 1))$ ($\times 2N_e$ generations). Although the general robustness of the Kingman coalescent breaks down under multiple merger coalescents, coalescence times will nevertheless be longer than a single generation. Although time scales under multiple merger coalescents are shorter than under Kingman coalescent, and for example our estimates are square root or cube root of $N_e$, they are longer than a single generation of the winner-take-all sweepstakes coalescent.

A dense genomic map of genetic variation in humans (and in model organisms) allows scanning the genome for signatures of natural selection (Voight et al., 2006; Sabeti et al., 2007; Storz, 2005). The density of the genetic maps and sensitivity of the various methods used influences what percentage of the human genome we observe to show footprints of selection (Voight et al., 2006; Sabeti et al., 2007; Storz, 2005). It is safe to say that only a small percentage of single nucleotide polymorphisms (SNPs) show footprints of selection in the low fecundity humans (Akey, 2009; Pickrell et al., 2009). For microsatellite loci, 2% (13/624) were detected as outliers when African and non-African human populations were compared (Storz, Payseur & Nachman, 2004). In contrast, comparable genome level studies in Atlantic cod find that 11% (26 out of 235) of independent SNPs (Moen et al., 2008) are $F_{ST}$ outliers (by method of Beaumont & Nichols, 1996) and 4% SNPs (70 out of 1641 Bradbury et al., 2010) are Bayscan outliers (by method of Foll & Gaggiotti, 2008) likely undergoing selection. Similarly one fourth of microsatellite loci in Atlantic cod (Nielsen, Hansen & Meldrup, 2006) are $F_{ST}$ outliers. This supports our thesis that a considerable fraction of the Atlantic cod genome may be simultaneously under selection for different adaptations.

More than half of the 70 outliers in Bradbury et al. (2010) study of Atlantic cod show adaptive parallel clines related to temperature on both the western and eastern side of the Atlantic Ocean. They show that multiple genes, located in three independent linkage groups, are involved. There are single genes as well as blocks of genes in "genomic islands" (Bradbury et al., 2013; Hemmer-Hansen et al., 2013). Some of the genes or blocks of genes show clear spatial patterns while other genes show complex spatio-temporal patterns in
contrast to no differentiation of non-outlier (neutral) loci (Poulsen et al., 2011; Therkildsen et al., 2013). For example, a locality in West Greenland shows great similarity to coastal areas in Iceland, implying either parallel adaptation on a fine scale or patterns of gene flow that are hard to reconcile with geographic distance. Another study (Hemmer-Hansen et al., 2014) adds even more complexity of population structure at outlier loci with little or no difference at non-outlier neutral loci.

A study of differentiation among four Atlantic cod populations along the coast of Norway (Moen et al., 2008) showed no differentiation among presumably neutral non-outlier loci with an average $\bar{F}_{ST} = 0.0012$. In contrast, among the outlier loci, presumably under selection, the $F_{ST}$ ranged from 0.08 to extreme differentiation of 0.83 with an average $\bar{F}_{ST} = 0.27$. Here we analyze in detail nucleotide variation at a large fragment of the Ckma gene (encoding a muscle isoform A of creatine kinase) showing extreme spatial differentiation (Moen et al., 2008) to understand the nature of selection.

Creatine kinases (CK) are crucially important in bioenergetic processes in cells and tissues (Wallimann et al., 1992; Wallimann, Tokarska-Schlattner & Schlattner, 2011). The creatine kinase/phosphocreatine system (CK/PCr) is an intracellular energy shuttle. CK generates Phosphocreatine (PCr) at the sites of ATP production in glycolysis and oxidative phosphorylation in mitochondria and regenerates ATP from PCr at subcellular sites of ATP use by ATPases. The physiological advantage is to provide a spatial and temporal energy buffer storing and releasing energy in and from PCr. Importantly, the rate of intracellular diffusion of both Creatine (Cr) and PCr is one and three orders of magnitude faster than diffusion of ATP and ADP respectively (see Wallimann et al., 1992; Wallimann, Tokarska-Schlattner & Schlattner, 2011, for a detailed account of the CK/PCr system).

We thus have here a gene with a well defined and well understood function. The gene shows extreme spatial differentiation most likely due to selection considering the contrasting behavior of neutral non-outliers. We apply methods of multiple merger $\Lambda$ coalescents, as a new and appropriate null model of neutrality for organisms with highly skewed heavy-tailed offspring distributions, to nucleotide variation of the gene to better understand the nature of selection.

**MATERIALS AND METHODS**

**Population sampling**

We randomly sampled 180 individual cod from various localities from the distributional range of Atlantic cod (Fig. S1). The samples come from our large sample database of greater than 20,000 individuals. All localities are represented with at least 100 individuals (except the White Sea with 24 individuals). The localities are the waters around Newfoundland (New), Greenland (Gre), Iceland (Ice), Faroe Islands (Far), Norway (Nor), and the Barents Sea (Bar), North Sea (Nse), Celtic Sea (Cel), Irish Sea (Iri), Baltic Sea (Bal), and the White Sea (Whi). We took a large sample from Iceland and stratified the sampling to get about 8–10 individuals from the other localities to cover the widest geographic range possible with our database. After cloning, sequencing and quality
checking as detailed below we had 122 individuals covering a wide geographic area from the Southwest/Northwest to the Northeast and South.

We included samples of the closely related taxa Arctic cod *Boreogadus saida* (Lepechin, 1774) (Bsa) and Greenland cod *G. ogac* (Richardson, 1836) (Gog) both sampled in Greenland waters as well as Pacific cod *G. macrocephalus* (Tilesius, 1810) (Gma) and Walleye pollock *Theragra chalcogramma* (Pallas, 1811) (Gch) sampled from the Pacific ocean as outgroups. Carr et al. (1999) and Pogson & Mesa (2004) discuss the relationship and biogeography of these taxa. Coulson et al. (2006) provide the most comprehensive account based on mitochondrial genomics. They consider Arctic cod to be an outgroup for all these taxa. Atlantic cod and Walleye pollock are closely related taxa and Pacific cod slightly more distant. Pacific cod and Walleye pollock represent two separate but nearly simultaneous invasions of the Pacific. The Atlantic cod vs. Pacific cod split is dated at 4 mya and the Atlantic cod vs. Walleye pollock split is dated at 3.8 mya using conventional rates of mtDNA evolution (see time scales below). Coulson et al. (2006) suggested a nomenclature revision from *Theragra chalcogramma* to *Gadus chalcogrammus* (Pallas, 1814) for Walleye pollock that has been accepted by the American Fisheries Society (Page et al., 2013). We follow the new nomenclature hereafter. Greenland cod is a recent reinvasion of Pacific cod into the Arctic and Coulson et al. (2006) consider it to be a subspecies of Pacific cod.

The Icelandic Committee for Welfare of Experimental Animals, Chief Veterinary Office at the Ministry of Agriculture, Reykjavik, Iceland has determined that the research conducted here is not subject to the laws concerning the Welfare of Experimental Animals (The Icelandic Law on Animal Protection, Law 15/1994, last updated with Law 157/2012). DNA was isolated from tissue taken from dead fish on board research vessels. Fish were collected during the yearly surveys of the Icelandic Marine Research Institute. All research plans and sampling of fish, including the ones for the current project, have been evaluated and approved by the Marine Research Institute Board of Directors. Samples were also obtained from dead fish from marine research institutes in Norway, the Netherlands, Canada and the US that were similarly approved by the respective ethics boards. The samples from the US used in this study have been described in Cunning et al. (2009) and the samples from Norway in Árnason & Pálsson (1996). The samples from Canada consisted of DNA isolated from the samples described in Pogson (2001). The samples from the Netherlands were obtained from the Beam-Trawl-Survey (http://www.wageningenur.nl/en/Expertise-Services/Research-Institutes/imares/Weblogs/Beam-Trawl-Survey.htm) of the Institute for Marine Resources & Ecosystem Studies (IMARES), Wageningen University, the Netherlands, which is approved by the IMARES Animal Care Committee and IMARES Board of Directors.

Molecular analysis
We used sequences associated with the Moen et al. (2008) high *F*<sub>ST</sub> SNP's (*Gm366-0514* with an *F*<sub>ST</sub> = 0.83, *Gm366-1022* with an *F*<sub>ST</sub> = 0.82, and *Gm366-1073* with an *F*<sub>ST</sub> = 0.82) to make probes to search our Atlantic cod BAC library (GAMH, made for us by Amplicon Express, www.amplicon-express.com). We had positive clones 454...
sequenced (Microsynth, Framingham, Massachusetts, USA) and obtained a 34,223 bp scaffold containing the gene of interest. From this sequence we generated primers (Table S1) for PCR amplifying a 4,000 bp fragment for population studies. Our scaffold largely but not entirely aligned to GeneScaffold 4232 of the Atlantic cod genome sequence (Star et al., 2011) (www.ensemble.org). We confirmed our primers using the Atlantic cod genome sequence. Our BAC library was made from a single individual from Bay of Faxa (Reykjavík) Iceland and the genomic sequence is based on a specimen from the North East Arctic cod in Norway (Star et al., 2011). The conformity of primer sequences between the two specimens from widely separated geographic localities means that the primers should amplify the fragment of interest in individuals taken from widely separate geographic areas. However, it does not preclude the possibility of ascertainment bias for example due to variation in primer binding sites. The amplification primers were long (Table S1), which may facilitate annealing in spite of some variation in primer binding site. Samples from all localities were PCR amplified without issue and there were no signs of ascertainment bias in the molecular results.

We Topo-TA cloned fragments into pCR XL-TOPO vector (Invitrogen, Waltham, Massachusetts, USA). We sequenced clones with M13 primers and sequencing primers (Table S1) using BigDye Terminator kit (Applied Biosystems, Waltham, Massachusetts, USA) and performed sequencing on ABI 3100 and ABI3500XL (Applied Biosystems) automated sequencers.

We applied the same methods and sequenced 711 bp of the Hemoglobin α 2 (HbA2) locus (Hallárdóttir & Arnason, 2009a; Hallárdóttir & Arnason, 2009b; Borza et al., 2009) and 1021 bp of the myoglobin (M yg) locus (Lurman et al., 2007). The previous studies on these genes (Hallárdóttir & Arnason, 2009a; Hallárdóttir & Arnason, 2009b; Borza et al., 2009; Lurman et al., 2007) had not found any signs of selection, and therefore we used them for neutral locus comparisons. The HbA2 data were of 114 Atlantic cod individuals and 13 individuals of various closely related taxa. The Myg data were from 45 Atlantic cod individuals and two individuals of Pacific cod. Other closely related taxa did not amplify for Myg. The HbA2 and Myg individuals covered much the same geographic localities as Ckma.

All sequences have been deposited in Genbank with Ckma accession numbers KM624178–KM624309, HbA2 accession numbers KM624310–KM624436, and Myg accession numbers KM624437–KM624483.

Population genetic analysis

We base called, assembled and edited sequence reads using phred, phrap and consed (Ewing et al., 1998; Ewing & Green, 1998; Gordon, Abajian & Green, 1998). We aligned sequences using muscle (Edgar, 2004), inspected alignments using seaview (version 4) (Gouy, Guindon & Gascuel, 2009) and generated maximum likelihood trees with phyml (Guindon & Gascuel, 2003) under seaview. We used R (R Core Team, 2013) and the ape, pegas, seqinr, ade4, adegenet, and LDheatmap packages (Paradis, Claude & Strimmer, 2004; Paradis, 2010; Charif & Lobry, 2007; Dray & Dufour, 2007; Jombart & Ahmed, 2011;
Shin et al., 2006) and various functions written by us for managing, analyzing, and plotting the data. We used the MLHKA program (Wright & Charlesworth, 2004) for a maximum likelihood HKA test (Hudson, Kreitman & Aguade, 1987) based on the Kingman coalescent.

By PCR amplifying and cloning of fragments, polymerase copy errors in the PCR reaction inevitably will be found in clones. The coalescent methods are especially sensitive to singleton variants and errors that would enter into the data as singleton variants should be removed. To remove PCR errors and ensure authenticity of natural variation among individuals, we sequenced three clones from each individual. We claim that taking three clones is sufficient to eliminate PCR errors among clones of an individual and yield a consensus sequence of one allele from that individual. We are taking three copies (clones) of two items (chromosomes or alleles A and a). Any two of these will always be the same allele (A and A or a and a). A third clone (order is not important) will be of that same allele with probability 1/2 and of the alternative allele from the other chromosome with probability 1/2. One of the three has probability 1/2 of being different from the two that are the same. In the first case, a consensus sequence will be a true consensus of that allele. In the second case, a consensus sequence will be a true consensus except at sites where the third clone (alternative allele) matches one of the other clones. That is when a naturally occurring site variant or a PCR error in the third clone matches a PCR error in one of the other two clones. This scenario is expected to be a rare event. The effect of such a rare event would be to generate variation that would look like recombination thus, if anything, reducing measures of linkage disequilibrium.

We thus got consensus sequences for a number of individuals. In some cases parts of a clone had low quality sequence. We visually inspected all variant sites using the above-mentioned tools. To maximize the number of individuals and the size of the sequenced fragment, we struck a balance between number of individuals and quality of sequence. We removed individuals with short sequences and removed individuals that were not covered by three clones. Also, we eliminated regions with a phred quality less than 30. We thus obtained consensus sequences of three clones from each of 122 Atlantic cod and 10 individuals of closely related taxa covering three fragments of the gene (Fig. S2) concatenated to give a total sequence of 2,500 bp.

We analyzed sequence variation for statistics of neutrality and selection using DNAsp (Rozas et al., 2003) and R functions. Site frequency spectra are a most important summary statistics for coalescent analysis of nucleotide data (Wakeley, 2009). We analyzed site frequency spectra using the Kingman coalescent (Kingman, 1982a) and statistical methods developed for multiple merger Λ coalescents (Birkner, Blath & Eldon, 2013b). We used software from Bjarki Eldon (Birkner, Blath & Eldon, 2013b) (http://page.math.tu-berlin.de/~eldon/programs.html) to estimate various parameters of the multiple merger Λ coalescents. In particular, we used the minimum ℓ2 distance (Birkner, Blath & Eldon, 2013b) (sum of squares) to estimate the parameter α of the Beta(2 − α,α) coalescent (Schweinsberg, 2003) and the ψ parameter of the point-mass coalescent (Eldon & Wakeley, 2006). Using these estimates, we generated expected site-frequency spectra for the models and compared them to our observed spectra using a likelihood ratio G test with the
multiple-merger coalescent models nested within the Kingman coalescent. We also used the overall $\ell_2$ distance (square root of $\ell^2$) to compare the observed and expected site frequency spectrum of the three genes, Ckma, HbA2, and Myg. We used software from Bjarki Eldon to estimate parameters of algebraic ($A, \gamma$) and exponential ($E, \beta$) growth models (Eldon et al., 2015) and compared the observed site frequency spectra for the three genes to expectations based on these growth models using the $\ell_2$ distance.

RESULTS

Gene and protein
The Ckma gene encodes creatin kinase muscle isoform A (CKMA). The locus is 3604 base pairs (bp) in GeneScaff 4232 (coordinates 332764 to 336367, gene name ENSMOG00000008778 in the cod genome, www.ensemble.org, Star et al. (2011)). The gene has seven exons (Fig. S2). Ensemble reports 382 amino acids (aa). However, both genescan (http://genes.mit.edu/GENSCAN.html) and fgenesh (www.softberry.com) predicted 381 aa. The www.ensemble.org sequence adds a Glycine (G) residue in position 323 apparently due to incorrect splicing at the junction of the last two exons.

For mapping the gene, the SNP locus cgpGmo-S497 at position 19.5 cM (see Appendix S3 in Supplemental data of Borza et al., 2010) in linkage group CGP16 is found in a partial cDNA mRNA sequence (Genbank accession number EX184243) (Hubert et al., 2010; Borza et al., 2010) matching the Ckma gene.

There are seven paralogous genes found in the Atlantic cod genome (www.ensemble.org) encoding mitochondrial, brain and muscle isoforms of Creatine Kinase. The protein sequence of the two alleles A and B in Atlantic cod and of all the closely related taxa studied were of the CKMA isoform (Fig. S3). The variation reported is thus from orthologous genes.

Nucleotide variation and divergence
The variants of Ckma in Atlantic cod fell into two distinct and divergent groups, which we refer to as A and B alleles or haplogroups (Fig. 1 and Fig. S4). They were fixed for a C vs T at site 1,732 in the concatenated sequence (Table S2). The alleles also differed at 19 additional sites (Fig. 2 and Table S2). However, there was variation at these 19 sites that was segregating at a low frequency within one or both alleles which was evidently from recombination.

The divergence of the A and B alleles has arisen after the speciation between Gadus morhua and its Pacific closely related species G. macrocephalus or G. chalcogrammus. The gross, $D_{XY}$, and net, $D_a$, nucleotide divergence (see also Cruickshank & Hahn, 2014) between the A and B alleles was about one half that of the divergence between the closely related taxa (Table S3). The Ckma and HbA2 divergences between the closely related taxa are very similar, but the Myg divergence is about twice that (Fig. S5 and Table S3). The variance of times to coalescence is large so it is not unexpected to find differences in divergence among genes. There is nothing in the behavior of Myg and HbA2 to indicate deviation from the multiple merger null hypothesis of neutrality. In contrast with the
Figure 1 Maximum likelihood tree of Ckma variation (A and B alleles) among 122 individual Atlantic cod and 10 individuals of four closely related outgroup O taxa, Boreogadus saida Bsa, Gadus chalcogramma Gch, Gadus macrocephalus Gma, and Gadus ogac Gog. Localities and color codes for Atlantic cod are the waters of Canada (Nova Scotia and Newfoundland) Can, Greenland Gre, Iceland Ice, Norway Nor, Faroe Islands Far, and from the Barents Sea Bar, White Sea Whi, North Sea Nse, Baltic Sea Bal, Celtic Sea Cel, and Irish Sea Iri.

Árnason and Halldórsdóttir (2015), PeerJ, DOI 10.7717/peerj.786
results of Coulson et al. (2006), the maximum likelihood tree for Ckma (Fig. 1) and divergence estimates (Table S3) imply that separation of G. chalcogrammus predates the separation of G. macrocephalus and G. morhua. Similarly, the Hba2 locus showed the same pattern that G. chalcogrammus is outside of G. macrocephalus and G. morhua (Fig. S6). Unfortunately the Myg locus did not yield sequences for G. chalcogrammus.

All summary statistics showed high variation for Ckma (Table 1). In particular, nucleotide diversity $\hat{\pi}$ was high relative to the scaled population size $\hat{\Theta}$, resulting in a non-significant Tajima's $\hat{D}$. This was due to the great number of high heterozygosity sites
Table 1 Summary statistics of polymorphism of 2,500 bp fragment of the Ckma gene, 711 bp fragment HbA2 gene and 1,021 bp fragment of the Myg gene in Atlantic cod.

<table>
<thead>
<tr>
<th>Group</th>
<th>n</th>
<th>S</th>
<th>H</th>
<th>h</th>
<th>k</th>
<th>ˆh</th>
<th>ˆθS</th>
<th>ˆπ</th>
<th>ˆD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ckma all</td>
<td>122</td>
<td>87</td>
<td>72</td>
<td>0.959</td>
<td>10.62</td>
<td>0.0067</td>
<td>0.0043</td>
<td>-1.13 ns</td>
<td></td>
</tr>
<tr>
<td>Ckma North</td>
<td>86</td>
<td>65</td>
<td>51</td>
<td>0.941</td>
<td>5.12</td>
<td>0.0054</td>
<td>0.0015</td>
<td>-1.97 ns</td>
<td></td>
</tr>
<tr>
<td>Ckma South</td>
<td>36</td>
<td>45</td>
<td>23</td>
<td>0.891</td>
<td>3.61</td>
<td>0.0045</td>
<td>0.0015</td>
<td>-2.43 ns</td>
<td></td>
</tr>
<tr>
<td>Ckma A allele</td>
<td>43</td>
<td>49</td>
<td>28</td>
<td>0.930</td>
<td>3.10</td>
<td>0.0044</td>
<td>0.0013</td>
<td>-2.33 **</td>
<td></td>
</tr>
<tr>
<td>Ckma B allele</td>
<td>79</td>
<td>53</td>
<td>44</td>
<td>0.908</td>
<td>3.10</td>
<td>0.0044</td>
<td>0.0013</td>
<td>-2.33 **</td>
<td></td>
</tr>
<tr>
<td>HbA2 all</td>
<td>114</td>
<td>11</td>
<td>11</td>
<td>0.347</td>
<td>0.39</td>
<td>0.0025</td>
<td>0.0005</td>
<td>-1.95 *</td>
<td></td>
</tr>
<tr>
<td>HbA2 North</td>
<td>95</td>
<td>9</td>
<td>9</td>
<td>0.347</td>
<td>0.39</td>
<td>0.0025</td>
<td>0.0005</td>
<td>-1.95 *</td>
<td></td>
</tr>
<tr>
<td>HbA2 South</td>
<td>19</td>
<td>3</td>
<td>4</td>
<td>0.298</td>
<td>0.32</td>
<td>0.0016</td>
<td>0.0005</td>
<td>-0.95 ns</td>
<td></td>
</tr>
<tr>
<td>Myg all</td>
<td>45</td>
<td>30</td>
<td>24</td>
<td>0.901</td>
<td>2.74</td>
<td>0.0071</td>
<td>0.0028</td>
<td>-2.03 *</td>
<td></td>
</tr>
<tr>
<td>Myg North</td>
<td>36</td>
<td>28</td>
<td>20</td>
<td>0.894</td>
<td>2.65</td>
<td>0.0069</td>
<td>0.0027</td>
<td>-2.12 *</td>
<td></td>
</tr>
<tr>
<td>Myg South</td>
<td>9</td>
<td>10</td>
<td>7</td>
<td>0.944</td>
<td>3.22</td>
<td>0.0037</td>
<td>0.0033</td>
<td>-0.58 ns</td>
<td></td>
</tr>
</tbody>
</table>

Notes.

n, Sample size; S, number of segregating sites; H, number of haplotypes; h, haplotype diversity; k, average number of pairwise differences; ˆθS, scaled population size from S; ˆπ, nucleotide diversity; ˆD, Tajima’s. ns is not significant.

* represents P < 0.05.
** represents P < 0.01.

differing between the two alleles (Fig. 2 and Table S2). Considering the North and South population and the A and B alleles separately, there was much less variation. Although there were several polymorphic sites within both A and B alleles (Fig. 2 and Table S2), nucleotide diversity was lower than for the entire sample, and the relative difference of ˆh and ˆθS for each allele was greater resulting in negative and significant Tajima’s ˆD. The HbA2 gene had a very low haplotype and nucleotide diversity but disparity with ˆθS gave overall a negative and significant Tajima’s ˆD. In congruence with divergence measures, the Myg locus had high haplotype and nucleotide diversity, albeit lower than Ckma, but overall a negative and significant Tajima’s ˆD.

There were five non-synonymous changes segregating as singleton sites within Atlantic cod (Tables S2 and S4). Two of these were also segregating as singletons within B. saida and G. macrocephalus, and one other singleton was also found in G. macrocephalus. B. saida was fixed for a Glycine (GGT codon) for which the other taxa have a Glutamine (CAG codon) with changes in all three sites of the respective codon (aa number 242). Assuming independent mutations and depending on the path of evolution of that particular codon, all three changes may have been non-synonymous.

There was considerable linkage disequilibrium (measured as D’) throughout the gene (Figs. S7 and S8). Linkage disequilibrium measures are sensitive to allele frequency (Hedrick, 1987) and in general there is no measure that is independent of allele frequencies (Lewontin, 1988). In Fig. S8 we have therefore excluded singleton sites because they will always show maximum linkage disequilibrium. However, low frequency sites generate noise so the signal of linkage disequilibrium is hard to see. We therefore used sites with minor allele frequency greater than an arbitrary frequency of 0.1 (Fig. S7), which includes
all the intermediate allele frequency (high heterozygosity) polymorphisms and gets rid of low frequency variants that generate noise in the linkage disequilibrium plots. The high frequency sites gave the clearest sign of two blocks of sites with almost full linkage disequilibrium both among sites within and between the blocks. The two blocks are separated by a site of recombination (site 691 in Table S2). On both the A and B allele backgrounds, both the ancient c major allele and the derived t minor allele at site 691 were geographically widespread (the t on an A allele background was found in individuals from the Baltic and from Iceland and the c on the B allele background was found in individuals from throughout the North ranging from the White Sea, Barents Sea, Norway, Iceland, Greenland, and Canada).

Two other sites (site three and 12 in Fig. S7 that are sites 578 and 1,444 in Table S2) showed slight reduction in linkage disequilibrium (Fig. S7) and therefore some signs of recombination. Other sites (such as sites 509, 660, 1,075 in Table S2) also showed some evidence of recombination. In all these cases, the recombinant gametic types with respect to the A and B allelic backgrounds were geographically widespread in general agreement with the result above for site 691.

The results of a maximum likelihood HKA test of selection that is based on the Kingman coalescent (Wright & Charlesworth, 2004) gave a selection parameter $k = 2.12$ in the direction of balancing selection (Table S5). However, the results were not statistically significant possibly because of too high variation among the presumed neutral loci ($HbA2$ and $Myg$) used for comparison in the test.

**Spatial differentiation**

The variation of the $Ckma$ gene was spatially patterned. The A allele was overall at a high frequency of 97% in an area that we call South (Faroe Islands, North Sea, Baltic Sea, Celtic Sea and Irish Sea) (Table S6). Conversely, the B allele was at a high frequency of 92% in an area that we call North ranging from the Northwest (Nova Scotia and Newfoundland in Canada) through Greenland, Iceland, Norway, Barents Sea and the White Sea. There was variation among localities within each region, with some localities having zero frequency, presumably due to low sample sizes. We do not have genotypic data and cannot test for Hardy-Weinberg equilibrium. The differentiation of North and South was evident in interlocality $F_{ST}$ values (Table S7) and an overall $F_{ST} = 0.763$ between North and South. There was no significant differentiation among localities within either the North or the South, but very high and significant differentiation between North and South localities. Similarly, there was great differentiation between the A and B alleles with an $F_{ST} = 0.804$. This was in stark contrast to the lack of differentiation between North and South at the $HbA2$ ($F_{ST} = 0.004$) and $Myg$ ($F_{ST} = −0.029$) loci.

The high differentiation was mostly due to the great number of high heterozygosity sites differing between the two alleles (Fig. 2 and Table S2). Three of the sites were the SNPs already found by Moen et al. (2008) with an $F_{ST} = 0.82$ for north and south localities along the coast of Norway. The high frequency sites showed indications of recombination.
between the A and B alleles (see for example patterns of segregating sites for individuals 105698, 124401, 105657, 200500, 118129, 119535, 118147, and 106620 in Table S2).

There were also several high heterozygosity polymorphic sites within both the A and B alleles (Fig. 2). This variation, however, did not show geographical patterns (Table S2). For example sites 1,050 and 1,428 mutated relative to outgroup within the A alleles were found among individuals from Iceland, White Sea, Celtic Sea, Faroe Islands and the Baltic. Similarly within the B alleles high heterozygosity sites 656, 691, 1,340, and 1,444, which were mutated relative to the outgroup, were all widespread among North localities ranging from the Northwest to the Northeast Atlantic (Fig. S1).

Site frequency spectra
The unfolded site frequency spectrum for the Ckma gene was trimodal (Fig. 3), with a mode at singleton sites, a mode at 43, and a mode at 79. The latter modes were at opposite frequencies out of a total of 122 and represented the A and B lineages of the genealogy. The Kingman coalescent did not fit the data well. Both the Beta\((2 - \alpha, \alpha)\) and point-mass coalescent models gave a much better fit (Table S8), in particular by capturing the singleton class. None of the coalescent models captured the modes at 43 and 79.

In contrast the site frequency spectra for the HbA2 and Myg genes were L shaped with a high peak at singleton sites (Figs. S9 and S10). Again, the Kingman coalescent did not fit
Table 2 Parameter values minimizing the $\ell^2$ distance (sum of squares) between observed and expected unfolded site frequency spectra for nuclear genes and for mtDNA variation of various localities.

<table>
<thead>
<tr>
<th>Source</th>
<th>$\hat{\alpha}$</th>
<th>$\hat{\psi}$</th>
<th>$\ell^2(\hat{\alpha})$</th>
<th>$\ell^2(\hat{\psi})$</th>
<th>$\ell^2(0)$</th>
<th>n</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nuclear locus</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hba2</td>
<td>1.000</td>
<td>0.230</td>
<td>0.035</td>
<td>0.016</td>
<td>0.431</td>
<td>113</td>
<td>This study</td>
</tr>
<tr>
<td>Myg</td>
<td>1.000</td>
<td>0.225</td>
<td>0.010</td>
<td>0.018</td>
<td>0.230</td>
<td>45</td>
<td>This study</td>
</tr>
<tr>
<td>Ckma</td>
<td>1.280</td>
<td>0.070</td>
<td>0.006</td>
<td>0.007</td>
<td>0.141</td>
<td>122</td>
<td>This study</td>
</tr>
<tr>
<td>CkmaA</td>
<td>1.100</td>
<td>0.170</td>
<td>0.017</td>
<td>0.012</td>
<td>0.161</td>
<td>43</td>
<td>This study</td>
</tr>
<tr>
<td>CkmaB</td>
<td>1.140</td>
<td>0.120</td>
<td>0.006</td>
<td>0.015</td>
<td>0.189</td>
<td>79</td>
<td>This study</td>
</tr>
<tr>
<td>Locality for mtDNA</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Newfoundland</td>
<td>1.550</td>
<td>0.015</td>
<td>0.014</td>
<td>0.028</td>
<td>0.084</td>
<td>378</td>
<td>Carret al.</td>
</tr>
<tr>
<td>Greenland</td>
<td>1.945</td>
<td>0.005</td>
<td>0.072</td>
<td>0.071</td>
<td>0.072</td>
<td>78</td>
<td>Arnason et al. (2000)</td>
</tr>
<tr>
<td>Iceland</td>
<td>1.550</td>
<td>0.010</td>
<td>0.006</td>
<td>0.050</td>
<td>0.078</td>
<td>519</td>
<td>Arnason et al. (2000)</td>
</tr>
<tr>
<td>Norway</td>
<td>1.895</td>
<td>0.015</td>
<td>0.093</td>
<td>0.089</td>
<td>0.095</td>
<td>100</td>
<td>AP 1996</td>
</tr>
<tr>
<td>White Sea</td>
<td>2.000</td>
<td>0.005</td>
<td>0.551</td>
<td>0.554</td>
<td>0.551</td>
<td>109</td>
<td>Arnason, Pálsson &amp; Petersen (1998)</td>
</tr>
<tr>
<td>Faroe Islands</td>
<td>1.555</td>
<td>0.050</td>
<td>0.059</td>
<td>0.055</td>
<td>0.093</td>
<td>74</td>
<td>SA 2003</td>
</tr>
<tr>
<td>Baltic Sea</td>
<td>2.000</td>
<td>0.005</td>
<td>0.105</td>
<td>0.109</td>
<td>0.105</td>
<td>109</td>
<td>Arnason, Pálsson &amp; Petersen (1998)</td>
</tr>
<tr>
<td>Atlantic</td>
<td>1.530</td>
<td>0.010</td>
<td>0.006</td>
<td>0.055</td>
<td>0.249</td>
<td>1278</td>
<td>Arnason (2004)</td>
</tr>
</tbody>
</table>

Coalescent parameter estimates

Following Birkner, Blath & Eldon (2013b), we used the $\ell^2$ distance, the sum of the squared differences between the observed and expected site frequency spectrum (scaled with the number of segregating sites), for estimating parameters of two $\Lambda$ coalescent models, $\hat{\alpha}$ for the Beta($2 - \alpha, \alpha$) and $\hat{\psi}$ for the point-mass coalescent (Table 2, Figs. S12 and S13). The Kingman coalescent, a null model for which $\alpha = 2.0$, had the highest $\ell^2$ indicating worst fit among the models. The Hba2 and Myg loci had an $\hat{\alpha} = 1.00$ and a $\hat{\psi} = 0.23$. The Ckma locus had overall a considerably higher $\alpha$ and lower $\psi$. The parameter estimates for the Ckma alleles separately were similar to those of the presumed neutral loci Hba2 and Myg.

For comparison we also estimated the parameters for the entire dataset of mtDNA variation in the North Atlantic (Arnason, 2004) and the various subsamples making up

well but both multiple merger coalescent models captured the high frequency of singleton sites.

The site frequency spectra of the A and B alleles alone were bimodal, with a high singleton class and peaks around 40 and 78 respectively (Fig. S11). The high frequency modes for the two alleles, at 40 and 78 respectively, resulted because most of the high frequency and high heterozygosity sites that separate the two alleles were not fixed within each allele presumably due to recombination between the alleles (Table S2, and see examples presented above).
that total sample using the unfolded site frequency spectrum with *G. macrocephalus* as the outgroup (Table 2 and Fig. S13). These have been previously analysed using the folded site frequency spectrum (see for example Birkner, Blath & Eldon, 2013b; Steinrücken, Birkner & Blath, 2013). For the total sample, spanning a similar geographic range as the nuclear genes, the parameter estimates differed from the nuclear loci with $\hat{\alpha} = 1.53$ and $\hat{\psi} = 0.01$. The large samples from Newfoundland and Iceland and the sample from the Faroe Islands gave similar values. The values for Greenland, Norway, White Sea, and Baltic Sea were much closer to the results for the Kingman coalescent ($\alpha = 2.0$). For these localities homoplasies were more frequent in the data than for the total and the large samples. Homoplasies will reduce the number of singletons and move such sites towards the right tail of the site frequency distribution. This explains the higher values for these localities.

**Models of multiple merger coalescents and population growth**

It is important to see how a locus deviates from a null model of neutrality to understand selection. Here the null model is multiple merger $\Lambda$ coalescents instead of the Kingman coalescent. Following Birkner, Blath & Eldon (2013b) we used the $\ell_2$ distance, the square root of the sum of the squared differences between the observed and expected site frequency spectrum. The overall $\ell_2$ distance for the three loci between the observed site frequency spectrum and expectations based on the two $\Lambda$ coalescent models are in Table S9. The $Ckma$ had the highest overall distance (the worst fit). There is clearly something special about the $Ckma$ locus that was not seen among the other loci. In particular, the trimodal site frequency spectrum is a sign of natural selection. We did not see these for the other genes. Admittedly, this is not a formal test of selection; however, $Ckma$ behaved differently. This is a locus specific behavior that is most likely a sign of selection.

The high frequency of singletons is predicted both by population growth and by $\Lambda$ and $\Xi$ multiple merger coalescents. Eldon et al. (2015) found that the weight of the right tail of the site frequency spectrum may have features allowing one to distinguish between population growth and $\Lambda$ coalescents. Eldon et al. (2015) have developed methods for such analysis which we apply here. Using both the $\ell_2$ distance and approximate log likelihood, we find (Table S10) that the algebraic ($A, \gamma$) and exponential ($E, \beta$) growth models gave very similar fits for each of the three genes. Again, as with the multiple merger coalescent models (Fig. S11), the $Ckma$ gene stood out and had the worst fit. The $Myg$ gene showed equally good fit to the the two growth models and the Beta($2 - \alpha, \alpha$) coalescent model. For both the $Ckma$ and $HbA2$ genes, the growth models showed worse fit than the coalescent models. However, this comparison of $\ell_2$ distances does not constitute a formal test as stated above.

**DISCUSSION**

**Genes and proteins**

The CKMA protein is highly conserved among the investigated taxa. The single aa difference between *B. saida* and the other species presumably is adaptive, with all sites
of the codon having changed. The few aa variants were all singletons in the sample. In fact, most of the variation is in non-coding regions and all the high heterozygosity sites in coding regions are synonymous changes. Given the high conservation of the protein and the high variation among silent and non-coding sites that are indicative of the mutational pressure, the singleton non-synonymous changes are likely slightly deleterious and will be removed by purifying selection. Some or even all of the silent and non-coding differences between the A and B alleles may be functional control elements important in expression in different tissues or under different environments. The potential functional differences remain to be studied.

The HbA2 and Myg genes have well-defined functions. They are probably under purifying selection. They were taken as independent genes in separate linkage groups for comparison. A caveat is that genetic variation at unlinked sites may be correlated and not independent in high fecundity populations with skewed distribution of offspring (Eldon & Wakeley, 2008; Birkner, Blath & Eldon, 2013a). The question remains, however, whether and to what extent such dependence impacts inference.

Three hypotheses
We discuss three possible explanations for the observed patterns of great divergence of the A alleles and B alleles, their spatial differentiation, and the trimodal site-frequency spectrum. The first explanation is the isolation/admixture hypothesis, the second is the coalescent of simultaneous multiple mergers in any one generation, and the third is the balancing selection hypothesis. Our interpretation is that the evidence favors balancing selection.

Ancient isolation and recent admixture
First, there is the possibility of recent admixture of anciently separated and divergent gene pools that have come together in a hybrid zone of secondary contact (Bowcock et al., 1991; Bernardi, Sordino & Powers, 1993; Gainand, Lemaire & Bonhomme, 2004). The spatial patterns of genetic separation between the South (Faroe Islands, North Sea, Baltic Sea, Celtic Sea and Irish Sea) and the North (Nova Scotia and Newfoundland, Greenland, Iceland, Norway, Barents Sea, and White Sea) could be taken as evidence for this. The South is a shallow water environment whereas the North has more diversity of depth ranging from shallow to deep waters. Differences in temperature, salinity and other environmental factors are correlated with the North South difference. The great nucleotide divergence between the North and the South would imply either that this is an ancient divergence (not a Pleistocene event) or even a not-so-ancient divergence driven by strong selection over a shorter time. If the time of separation of G. morhua and G. macrocephalus and G. chalcogrammus is taken at 3.8–4.0 Mya (Coulson et al., 2006) the time of separation of the A and B clades would then be 2 Mya based on the nucleotide divergence of the A and B clades which we show is one half that of the closely related taxa. An even lower divergence time of 2.1 Mya has been suggested (Pogson & Mesa, 2004) that would still leave the divergence of the A and B clade at 1 Mya. These divergence times, however, are all based
Paper IV

on the Kingman coalescent and the faster time scales of the multiple merger coalescent are discussed below.

A counter argument is that isolation and admixture are part of the breeding structure of a population leaving genome-wide impacts (Wright, 1931). Therefore, different genes should be concordant in their behavior (Bernardi, Sardino & Powers, 1993). This should be true for neutral genes that randomly drift apart in the different isolated areas. Genes under selection adapting to the different environments of the isolated areas should show even greater divergence. The HbA2 and the Myg show no differentiation between the North and the South. Also, the non-outlier SNPs in Moen et al. (2008) show no differentiation whereas three SNPs of the Ckma gene show high and extreme $F_{ST}$. The correspondence between our results and those of Moen et al. (2008), with very similar $F_{ST}$ between our North vs. South and the north vs. south along the coast of Norway in Moen et al. (2008), is strong independent verification of our main result. The Ckma was, after all, the most extreme outlier in Moen et al. (2008). Similarly, Bradbury et al. (2010) found that non-outlier SNPs show no differentiation although other SNPs show differentiation from parallel adaptation to temperature on the eastern and western side of the Atlantic Ocean. Nielsen et al. (2003) described a pattern of microsatellite variation in a transition area between the Baltic and Danish Belt Sea which they interpret as a hybrid zone. There is no evidence for a hybrid zone at that location in the Ckma data. In fact, specific variants within the A allele are widely distributed among localities in the South including the Baltic Sea. This implies gene flow among localities in the South. Similar patterns within B alleles imply gene flow among localities in the North. If indeed there is a hybrid zone for the Ckma gene, it would lie between the Faroe Islands on one hand and Iceland and north and middle Norway on the other hand. Considering the North East Arctic and Coastal cod in Norway as an admixture of isolated populations (Pogson & Fevolden, 2003; Arnason & Palsson, 1996) would add a third hybrid zone within the distribution of the species. It is not a parsimonious explanation to consider there to be multiple hybrid zones of secondary contact within distribution of the species.

For comparison one can consider the Pan I locus (Fevolden & Pogson, 1995; Fevolden & Pogson, 1997) that clearly is under selection (Pogson, 2001; Pogson & Mesa, 2004) related to depth and fisheries (Sarvas & Fevolden, 2005; Case et al., 2005; Arnason, Hernandez & Kristinsson, 2009). At face value, the locus shows similar differentiation between North and South (Sarvas & Fevolden, 2005) as the Ckma locus. However, the details differ and the parallels between the Pan I and Ckma genes are more apparent than real. Pogson & Fevolden (2003) argue that specific neutral alleles found within a functional class (the Pan I A allele) should show differences between historically isolated regions. Under the historical (isolation/admixture) hypothesis different neutral alleles will drift to high frequencies or fixation in geographic regions isolated from each other. Under the selection hypothesis, they should move seamlessly among localities within the putative isolated regions. Pogson & Fevolden (2003) tested the “historical” and “selection” hypotheses (c.f. Arnason & Palsson, 1996) of Atlantic cod in northern Norway by studying presumed neutral variation among the Pan I A alleles in coastal and Arctic localities. In short, they found no evidence
supporting the historical hypothesis. In fact there were greater differences among coastal localities and between the two Arctic localities than overall between the Arctic and coastal areas. Because of the heterogeneity among coastal localities, Pogson & Fevolden (2003) also rejected the selection hypothesis because neutral mutations would move freely among localities within a region and should not show any structure. However, under a skewed offspring distribution and sweepstakes reproduction, there can be substantial population structure as measured by \( F_{ST} \) in the face of considerable gene flow (Eldon & Wakeley, 2009). Thus their results do not seem at odds with a multiple merger coalescent model.

For \( Pan \ I \) the \( B \) allele is largely absent in the South. But the absence of an allele from a certain region cannot be used as evidence for the isolation of populations from that region from populations in other regions. Instead under the isolation/admixture (historical) hypothesis one would expect (Pogson & Fevolden, 2003) specific \( Pan \ I A \) alleles to be present characterizing the South and another set of \( A \) alleles characterizing the North. But that is not the case; among the various \( A \) alleles there is no specific clade of \( Pan \ I A \) alleles in the South (U Hernandez & E Árnason, 2014, unpublished data). However, for the \( Ckma \) gene there is a specific allele, namely the \( A \) allele, that is at a high frequency and characterizes the South.

The \( Pan \ I B \) allele which is adapted to the deep (Pampoulie et al., 2007; Árnason, Hernandez & Kristinsson, 2009) is largely absent from the South. The \( Pan \ I B \) allele, which is found in the North and in deep water, is much less variable than the \( Pan \ I A \) alleles (Pogson, 2001). This is opposite to what we find for the \( Ckma \) \( A \) alleles (the South allele), which has less variation than the \( Ckma \) \( A \) allele (Fig. 1) although this is not seen in the summary statistics (Table 1) because of greater recombinational variation at the base of the \( A \) clade (Table S2). Also, the \( Pan \ I \) locus variation is more related to depth than to geography (Árnason, Hernandez & Kristinsson, 2009). Under the admixture hypothesis, these two loci and all loci showing genome wide effects are expected to show the same pattern.

Under the isolation/admixture hypothesis, one would expect recombinant types to be restricted geographically to the zone of secondary contact. This was not the case. Therefore, we think it is more likely that the two blocks of nucleotide sites are held together in linkage disequilibrium by epistatic fitness interactions and that there has been a build up of linkage disequilibrium over time.

Overall, therefore, we find that the \( Ckma \) gene does not fit the hypothesis of ancient divergence of gene pools and admixture in secondary contact.

\[ \Xi \] colascent and site frequency spectra

The trimodal site frequency spectrum is not predicted by any of the coalescent models considered here: the Kingman coalescent and the two \( \Lambda \) coalescent models, the Beta(2 − \( \alpha \), \( \alpha \)) (Schweinsberg, 2003), and the point-mass coalescent (Eldon & Wakeley, 2006). Under the \( \Lambda \) coalescent, at most a single multiple merger event occurs at any one time. The distribution of family size is of interest, and the parameter \( \alpha \) influences the probability of getting large families. Under the Beta(2 − \( \alpha \), \( \alpha \)) coalescent model, the probability of a family size of \( k \) or more viable offspring decays like \( k^{-\alpha} \) (Schweinsberg, 2003) in the limit of a large \( k \). The pool of viable offspring is then resampled to form the next generation.
under the same conditions. For the Kingman coalescent, $\alpha \geq 2$ and there is little chance of seeing large families. For the Beta($2 - \alpha$, $\alpha$) coalescent, $1 \leq \alpha < 2$ and the lower $\alpha$ the greater is the chance of seeing a large family (Schweinsberg, 2003). The $\psi$ parameter of the point-mass coalescent (Eldon & Wakeley, 2006) similarly measures the proportion of the population that is the offspring of a single individual and is thus an indicator of reproductive sweepstakes. Our estimates of $\psi$ indicate reproductive sweepstakes at the neutral loci and within the A and B alleles of Ckma. Balancing selection at Ckma lessens the effects of sweepstakes reproduction. Sweepstakes reproduction has been detected in other high-fecundity organisms (Hedgecock & Pudovkin, 2011; Harrang et al., 2013).

Under the more general $\Xi$ coalescent $0 < \alpha < 1$ (Schweinsberg, 2000) there can be many large families independently in each generation. It would seem that this process could generate multimodal site frequency spectra. Indeed in simulations of $\Xi$ coalescence site frequency spectra can display multiple modes (B Eldon, pers. commun., 2014). This question needs further theoretical work. In terms of the concept of sweepstakes reproduction, multiple local sweepstakes could have this effect on the site frequency spectrum. Under local sweepstakes, genetic structure may be ephemeral (Johnson & Wernham, 1999). Whether this affects the location of the modes and the exact shape of the site frequency spectrum under $\Xi$ coalescent is not known. However, one would not expect build-up of sites around a specific mode of the site frequency spectrum or of two modes at opposite frequencies as at Ckma. Also, there should be no particular or regular geographical pattern. Therefore, we think that bumps in the site frequency spectrum under $\Xi$ coalescent is not a good explanation for the Ckma spectrum.

Models of population growth can account for the high frequency of singletons. However, these models also do not predict the trimodal site frequency spectrum observed at Ckma. This is a locus specific behavior that is most likely due to balancing selection.

It is of course possible that population growth and sweepstakes could be occurring at the same time. We do not at this time have methods that estimate simultaneous multiple merger coalescents and population growth and evaluate the relative contribution of each. It is likely that disentangling the effects changes in population size and sweepstakes reproduction may be hard. For example, Birkner et al. (2009) discussed how recurrent bottlenecks may construct simultaneous multiple merger $\Xi$ coalescent.

**Balancing selection**

Balancing selection generates long branches in the genealogy and neutral variation accumulates on the branches. The balanced functional types (the Ckma A and B alleles in this case) act as they were separate and isolated populations accumulating neutral variation. Recombination can bring variation from one branch to another acting like migration that brings alleles from one population to another (Charlesworth, Nordborg & Charlesworth, 1997; Charlesworth, Charlesworth & Barton, 2003; Charlesworth, 2006). However, the molecular signatures of balancing selection depend on many factors. Is it a long standing, even trans-species, polymorphism such as MHC in human and chimpanzee (Fan et al., 1989; Nei & Hughes, 1991) or is it very recent? Examples of the latter are human glucose 6 phosphate dehydrogenase (G6PD) (Verrelli et al., 2002), and hemoglobin $\beta$
S (Currat et al., 2002) and hemoglobin β (Ohashi et al., 2004) and spatially divergent selection of lactase persistence (Tishkoff et al., 2007; Ranciaro et al., 2014) in which a particular allele sweeps a chromosomal segment to an intermediate equilibrium frequency. In these instances, recombination has not had time to break up linkage disequilibrium, which can extend over large regions. There is very little variation among the new alleles while the alternative chromosomes show much more variation in this region representing the standing variation in the population at the start of the partial sweep.

The effects of a long standing single locus balancing selection will extend only short distances with free recombination and will be difficult to detect (Wiuf & Hein, 1999). If, however, there are obvious signs of a long standing balanced polymorphism it is likely due to a build-up of co-adapted complexes of epistatic interactions among multiple sites and/or suppression of recombination (Wiuf & Hein, 1999). The concept of a supergene of multiple co-adapted sites possibly locked together by structural variation (Thompson & Jiggins, 2014) such as found in butterfly mimicry (Joron et al., 2011) is relevant. There also can be both partial and complete selective sweeps of new types within each allele of a supergene. Such intra-allelic selective sweeps would reduce variation within and increase variation between alleles. Such reduction of variation could look similar to that for a recent balanced polymorphism, except that it would not be limited to one functional type. Thus Pogson (2001) argues that he has detected on-going partial sweeps within each of the two Pan I alleles of Atlantic cod.

Pogson & Mesa (2004) further argue that the Pan I polymorphism is older than speciation of Atlantic cod and Walleye pollock, the closest relatives. The Pan I locus is in a “genomic island” (Bradbury et al., 2013; Hemmer-Hansen et al., 2013) a potential supergene of co-adapted complexes possibly locked together by structural variation.

U Hernandez & Árnason (2014, unpublished data) find large number of differences between the two functional Pan I types in a 12.5 kb region around the PanI gene that are too extensive to be a partial sweep of a new allele. Such variation is likely to have built up over some time by selection (see time scales below). This is in face of considerable gene flow implied by lack of differentiation of neutral loci (Moen et al., 2009; Bradbury et al., 2010; Eiriksson & Árnason, 2013; Hemmer-Hansen et al., 2014). Similarly, the wide distribution of variants within both the A and B alleles of Ckma implies gene flow among localities within South and within North areas. The recombinant haplotypes between the A and B alleles of Ckma imply gene flow between the South and the North localities.

The observation that the amino acid sequences are conserved might be taken as evidence that there is only purifying selection at the locus. However, claiming balancing selection does not necessarily imply amino acid differences. There is evidence for positive selection in non-coding DNA in other systems (e.g., Drosophila, Andolfatto, 2005) and methods have been developed to detect positive and balancing selection in non-coding regions (e.g., Zhen & Andolfatto, 2012). Balancing selection has also been detected in regulatory regions in other systems. For example, the 5′ cis regulatory region of CCR5 shows evidence for balancing selection (Bamshad et al., 2002), as does the promoter region of the human Interleukin 10 gene (Wilson et al., 2006), a regulatory region upstream from the human
UGT2B4 gene (Sun et al., 2011), the NE1 locus in modern Humans and Neanderthals (Gokcumen et al., 2013), and the 5’ UTR’s of upregulated genes and genes for effector proteins of a plant-pathogenic fungus (Rech et al., 2014). We have not identified a specific target of selection and we speculate that there is selection on regulatory regions (5’, 3’, intronic, and even silent sites that may influence regulation) of the Ckma gene.

Ckma had the highest FST among all loci studied by Moen et al. (2008) and, therefore, the focus of selection is likely either the gene itself or a very tightly linked locus. We have looked in www.ensemble.org what genes are in the close neighborhood. There are no obvious candidates among them for a gene under strong selection. However, we think that an answer to this question must await a more detailed analysis of a larger region around the Ckma gene.

The Kingman and multiple-merger Λ coalescent models that we apply here are models of neutrality. One could argue that it is not appropriate to apply such neutral models to the Ckma locus that is already suspected to be under selection. However, understanding how the locus deviates from neutrality is important for understanding the pattern of selection. Under the neutral theory (Kimura, 1983), polymorphism within species is the transient phase of molecular evolution that leads to divergence between species. This is the rational for the HKA test of selection or neutrality (Hudson, Kreitman & Aguadé, 1987) that neutrally evolving genomic regions should have the same proportion of polymorphism to divergence. Balancing selection would tend to increase the level of polymorphism within species relative to divergence between them. The results of HKA test are in the direction of balancing selection. The HKA test shows a relative slowing down of divergence to rate of polymorphism at the Ckma locus.

Similarly, we consider the peaks in the site frequency spectrum of the Ckma gene to be evidence for balancing selection. The trimodal site frequency spectrum with two high frequency peaks at opposite frequencies that fold into one peak in a folded site frequency spectrum points to the build-up of variation over time. Under a recent balanced polymorphism scenario, such as G6PD and β globins in humans, there would be one peak at a particular frequency in the site frequency spectrum representing all sites at which the new allele differs from the ancient alleles. There could be multiple peaks representing high frequency polymorphisms among the ancient alleles. However, they are not expected to be at opposite frequencies to the frequency of the new allele. Therefore, we argue that the pattern at Ckma represents a balanced polymorphism that has been built up over time.

**Coalescent parameter estimates and time scales**

The question of coalescent time scale, however, must be considered. Under the Kingman coalescent, time is measured in terms of $N/\sigma^2$, with population size scaled by the variance of family size (Sagitov, 1999; Arnason, 2004; Tavaré, 2004). With a Poisson distribution of family size, $\sigma^2 = 1$ for a constant size haploid population and, therefore, time scales with $N$ under the Kingman coalescent. In an extreme winner-take-all sweepstakes, $\sigma^2 = N$ and a sample would coalesce in the previous generation and there would be no variation (Arnason, 2004). In more realistic multiple merger coalescent models, the time scale is
the quantity $c_N = \frac{E(\nu_1 - 1)^2}{N-1}$ where $c_N$ is the probability of two lineages coalescing in the previous generation in a haploid population of fixed size $N$ and $\nu_1$ is the random number of offspring of individual 1 (Sagitov, 1999). In general, the time scale of multiple merger coalescent models can be much shorter than for Kingman coalescent. Under the Beta$(2 - \alpha, \alpha)$ coalescent model time scales with $N^{\alpha-1}$ (Schweinsberg, 2003; Eldon et al., 2015). For this model, our estimates of $\alpha$ for the nuclear genes are quite low which implies very short time scales. The neutral genes would seem to coalesce in the very recent past. The $A$ and $B$ alleles of Ckma run on very similar time scales to the neutral genes and the locus itself at a slower rate due to the balancing selection with a time scale approximately the cube root of the effective population size $N_e$. The mitochondrial DNA runs at yet another and slower time scale. For mtDNA time scales with approximately the square root of $N$. Predicted turnover of alleles is faster and ages of alleles shorter under multiple merger coalescent (Eldon, 2012). Different populations and species may run on different time scales (Eldon & Degnan, 2012) complicating divergence time estimates. Estimates based on Kingman coalescent of divergence times of Atlantic cod populations (Bigg et al., 2008) or divergence of gadid taxa (Coulson et al., 2006) may therefore be too high and may need revision.

Conclusion

The Ckma protein coding sequence is conserved between all but the most distantly related Arctic cod. The amino acid variants are all singletons in the sample. Based on these facts, we conclude that the protein coding sequence is under purifying selection. At the same time, silent and non-coding variation at the locus shows extreme spatial differentiation with an $F_{ST}$ greater than 0.8 between the North and the South regions. The regulatory function of this variation is unclear. We argue that the high and locus-specific $F_{ST}$, the highest seen so far for any locus and any spatial comparison in Atlantic cod, indicates that selection and not admixture of anciently divergent gene pools is responsible. Selection is likely to be very strong. It follows that Ckma (or an extremely tightly linked locus) is the focus of selection because the highest $F_{ST}$ indicates the site of action of selection (Nielsen, 2005). Some of the variation may be neutral having risen in frequency within the balanced functional allele where it arose (Charlesworth, 2006). Alternatively, some of the variation may be due to selection building co-adapted complexes (Thompson & Jiggins, 2014). In addition to a high peak at singleton sites, higher than that predicted by the Kingman coalescent and characteristic of the multiple-merger coalescent, the site frequency spectrum has two high-frequency modes at opposite but matching frequencies representing the two branches of the genealogy. This pattern is further support for balancing selection. Our estimates of parameters of multiple-merger $\Lambda$ coalescent show that time-scales are fast in accordance with theoretical expectations.

ACKNOWLEDGEMENTS

We would like to thank Jarle Mork (Norwegian University of Science and Technology), Kristján Kristjánsson (Marine Research Institute in Reykjavik), Grant Pogson (University of California at Santa Cruz), Remment ter Hofstede (Institute for Marine Resources
and Ecosystem Studies in the Netherlands), and Michael Canino (National Oceanic and Atmospheric Administration) for help in securing some of the samples. We would like to thank Brenda Ciervo Adarna, Guðni Magnúss Eiriksson, Lilja Stefánsdóttir, Ragnheiður Fossdal, Svava Ingimarsdóttir, Ubaldo Benitez Hernandez for help with some of the laboratory work. We would like to thank Bjarki Eldon for programs and help with coalescent parameter estimation and for critical comments on the manuscript. We would like to thank R.C. Lewontin for discussions and critical comments on the manuscript. We would also like to thank reviewer Michael Matschiner and an anonymous reviewer for critical comments that helped us improve the paper.

ADDITIONAL INFORMATION AND DECLARATIONS

Funding
Funding was provided by Icelandic Science Foundation grant of excellence (nr. 40303011), a University of Iceland Research Fund grant, a Svala Árnadóttir private foundation grant to Einar Árnason and a doctoral grant from the University of Iceland Research Fund to Katrín Halldórsdóttir. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Grant Disclosures
The following grant information was disclosed by the authors:
Icelandic Science Foundation: 40303011.
University of Iceland Research Fund.
Svala Árnadóttir private foundation.

Competing Interests
The authors declare there are no competing interests.

Author Contributions
• Einar Árnason conceived and designed the experiments, performed the experiments, analyzed the data, contributed reagents/materials/analysis tools, wrote the paper, prepared figures and/or tables, reviewed drafts of the paper.
• Katrín Halldórsdóttir conceived and designed the experiments, performed the experiments, analyzed the data, wrote the paper, reviewed drafts of the paper, submitted sequences to GenBank.

Animal Ethics
The following information was supplied relating to ethical approvals (i.e., approving body and any reference numbers):
The Icelandic Committee for Welfare of Experimental Animals, Chief Veterinary Office at the Ministry of Agriculture, Reykjavik, Iceland has determined that the research conducted here is not subject to the laws concerning the Welfare of Experimental Animals (The Icelandic Law on Animal Protection,
Law 15/1994, last updated with Law 157/2012). DNA was isolated from tissue taken from dead fish on board research vessels. Fish were collected during the yearly surveys of the Icelandic Marine Research Institute. All research plans and sampling of fish, including the ones for the current project, have been evaluated and approved by the Marine Research Institute Board of Directors. The Board comprises the Director General, Deputy Directors for Science and Finance and heads of the Marine Environment Section, the Marine Resources Section, and the Fisheries Advisory Section. Samples were also obtained from dead fish from marine research institutes in Norway, the Netherlands, Canada and the US that were similarly approved by the respective ethics boards. The samples from the US used in this study have been described in Cunningham et al. (2009) and the samples from Norway in Árnason & Pálsson (1996). The samples from Canada consisted of DNA isolated from the samples described in Pogson (2001). The samples from the Netherlands were obtained from the Beam-Trawl-Survey (http://www.wageningenur.nl/en/Expertise-Services/Research-Institutes/imares/Weblogs/Beam-Trawl-Survey.htm) of the Institute for Marine Resources & Ecosystem Studies (IMARES), Wageningen University, the Netherlands, which is approved by the IMARES Animal Care Committee and IMARES Board of Directors.

DNA Deposits
The following information was supplied regarding the deposition of DNA sequences:
GenBank
KM624178–KM624309
KM624310–KM624436
KM624437–KM624483.

Data Deposit
The following information was supplied regarding the deposition of related data:
DataDryad
http://datadryad.org/.

Supplemental Information
Supplemental information for this article can be found online at http://dx.doi.org/10.7717/peerj.786#supplemental-information.

REFERENCES

Árnason and Halldórsdóttir (2015), PeerJ, DOI 10.7717/peerj.786


Cunningham KM, Canino MF, Spies IB, Hauser L. 2009. Genetic isolation by distance and localized fjord population structure in Pacific cod (Gadus macrocephalus): limited effective


Whole genome sequencing uncovers cryptic and hybrid species among Atlantic and Pacific cod-fish
Katrín Halldórsdóttir and Einar Árnason, 2015
In review

Author contributions: Einar Árnason and Katrín Halldórsdóttir conceived the experiments, conducted the experiments and analysed the results. Both authors wrote and reviewed the manuscript.

© Copyright 2015 Halldórsdóttir and Árnason. Made available under a CC-BY-NC-ND 4.0 International license.
Whole-genome sequencing uncovers cryptic and hybrid species among Atlantic and Pacific cod-fish

Katrín Halldórsdóttir¹ and Einar Árnason¹,²

¹Institute of Life and Environmental Sciences, University of Iceland, Sturlugata 7, 101, Reykjavík, Iceland
²Department of Organismal and Evolutionary Biology, Harvard University, Cambridge, MA 02138

ABSTRACT

Speciation often involves the splitting of a lineage and the adaptation of daughter lineages to different environments. It may also involve the merging of divergent lineages, thus creating a stable homoploid hybrid species¹ that constructs a new ecological niche by transgressing² the ecology of the parental types. Hybrid speciation may also contribute to enigmatic and cryptic biodiversity in the sea.³,⁴ The enigmatic walleye pollock, which is not a pollock at all but an Atlantic cod that invaded the Pacific 3.8 Mya,⁵ differs considerably from its presumed closest relatives, the Pacific and Atlantic cod. Among the Atlantic cod, shallow-water coastal and deep-water migratory frontal ecotypes are associated with highly divergent genomic islands;⁶,⁷ however, intermediates remain an enigma.⁸ Here, we performed whole-genome sequencing of over 200 individuals using up to 33 million SNPs based on genotype likelihoods⁹ and showed that the evolutionary status of walleye pollock is a hybrid species: it is a hybrid between Arctic cod and Atlantic cod that transgresses the ecology of its parents. For the first time, we provide decisive evidence that the Atlantic cod coastal and frontal ecotypes are separate species that hybridized, leading to a true-breeding hybrid species that differs ecologically from its parents. We refute monophyly and dichotomous branching of these taxa, and stress the importance of looking beyond branching trees at admixture and hybridity. Our study demonstrates the power of whole-genome sequencing and population genomics in providing deep insights into fundamental processes of speciation. Our study was a starting point for further work aimed at examining the criteria of hybrid speciation,¹⁰ selection, sterility and structural chromosomal variation¹¹ among cod-fish, which are among the most important fish stocks in the world. The hybrid nature of both the walleye pollock and Atlantic cod raises the question concerning the extent to which very profitable fisheries¹²,¹³ depend on hybrid vigour. Our results have implications for management of marine resources in times of rapid climate change.¹⁴,¹⁵

Speciation often involves the splitting of a lineage and the adaptation of daughter lineages to different environments. It may also involve the merging of divergent lineages, thus creating a stable homoploid hybrid species¹ that constructs a new ecological niche by transgressing² the ecology of the parental types. Hybrid speciation, which is well known among plants,³,⁴ is found also among animals, such as Heliconius butterflies⁵ and swordfish.⁶,⁷ Models of homoploid hybrid speciation involve chromosomal rearrangements and reduced fertility of F₁ hybrids, which may produce novel balanced gametes. Inbreeding of the F₁ may lead to F₂ individuals with novel fertile and stable homokaryotypes that are, at least partially, reproductively isolated from the parental types.⁶,⁷ We suggest that hybrids among high-fecundity promiscuously mating organisms, such as cod-fish, could also pass through the F₁ barrier. Ecological selection is also very important for the climbing of a new adaptive peak by hybrid species.¹,⁶,⁷

Cryptic and sibling species, forms that are very similar morphologically, are common in the sea and may reflect adaptive divergence of habitat use, life-history and chemical recognition without morphological divergence.³,⁴ Marine populations often have high dispersal potential and the marine environment appears to have few barriers to gene flow. Thus allopatric divergence may be slow. Speciation in the marine environment frequently involves behavioural differences in spawning time and mate recognition, gametic incompatibility and habitat specialization such as salt tolerance.²² The role of hybrid speciation¹ contributing to enigmatic and cryptic biodiversity in the sea³,⁴ is unknown. The origin of morphologically distinct forms and niche shifts may present an enigma. Morphologically similar forms are often cryptic species that are genetically distinct as revealed molecular genetic studies.¹¹ Identifying cryptic species is important for evaluation of biodiversity. Identifying cryptic species complexes in commercially exploited organisms also is important for conservation and the protection and management of natural resources.

Cod-fish represent some of the most important commercial fisheries in the world. Among cod-fish, the enigmatic walleye pollock (Gadus chalcogrammus), which is not a pollock at all, is under the hypothesis of speciation by lineage-splitting of an Atlantic cod (Gadus morhua) that invaded the Pacific Ocean 3.8 Mya, according to mtDNA genomics.³ Pacific cod (Gadus macrocephalus) is a slightly older (4 Mya) invasion.² However, Pacific and Atlantic cod share more traits than either of them share with walleye pollock. The semi-pelagic schooling walleye pollock differs morphologically, ecologically and behaviourally from these presumed closest relatives. The specific traits of walleye pollock niche shift would then
have to have arisen by selective filtering during colonization or subsequent adaptation to Pacific environments. However, the Pacific cod that colonized the same habitat did not go through the same filtering. The Pacific cod is also thought to have re-invaded the Arctic and Atlantic Oceans at western Greenland and formed the Greenland cod *Gadus ogac*. The Greenland cod is morphologically similar to both the Pacific and Atlantic cod and thus no special filtering occurred at re-invasion under that hypothesis. The biogeography of these taxa makes walleye pollock stand out as an evolutionary enigma.

Among Atlantic cod, the shallow-water coastal and deep-water frontal ecological ecotypes, which are defined by storage tag data, correlate with *AA* and *BB* homozygotes, respectively, of the *Pan I* locus located in a highly divergent genomic island on linkage group LG01. However, there is a general heterozygote excess at this locus. Heterozygotes and some homozygotes are behaviourally atypical, being intermediate, for both the coastal and frontal types. A study of other genomic islands suggests cryptic speciation and extensive hybridization producing *F1* hybrids and few, if any, *F2* and back-crossed individuals. Thus, if the coastal and frontal types are reproductively isolated, as has been suggested, the intermediates are an enigma: most of the population composed of sterile hybrids, as implied by the lack of *F2* and back-crossed individuals. This is hardly a tenable proposition.

Population genomics promises to significantly advance our knowledge of enigmatic and cryptic forms and their speciation in the sea. Here, we performed whole-genome sequencing of over 200 individual cod fish using up to 33 million SNPs based on genotype likelihoods to elucidate evolutionary relationships and speciation among gadid taxa.

### Results and Discussion

An individual admixture analysis differentiated the Atlantic, Pacific and Arctic taxa with a model of $k = 2$ ancestral populations (Figure 1). However, the walleye pollock genome of all individuals was about 40% Atlantic cod and 60% Arctic cod (*Boreogadus saida*). As the number of ancestral populations in the model increased, the groups split up: western vs eastern Atlantic ($k = 3$), Pacific and Greenland cod vs Arctic cod, walleye pollock and frontal ecotype vs coastal ecotype ($k = 4$) and the coastal ecotype into north vs south ($k = 5$). However, walleye pollock was always about 40% Atlantic cod, except at $k = 4$, where it aligns with Arctic cod. High-coverage data showed similar patterns (although the proportions differed; Figure 2) that the pollock genome was 50%, 10% and 40% Atlantic, Pacific and Arctic cod, respectively. Walleye pollock was similarly admixed at all genomic regions assignable to linkage groups (Supplemental Figure 1). As a side-note, a single individual of Polar cod (*Arctogadus glacialis*) was similarly admixed, possibly from similar processes as those detected in walleye pollock. The first and second principal components (PCs) separated Pacific and Greenland cod on one linear cluster, Atlantic cod on another and Arctic cod on a third (Figure 3). Greenland cod clusters with Pacific cod (Supplemental Figure 2) but in general it is closer to the other taxa by the analysis of principal components (Figure 3). The linear behaviour represented geographic variation within each species. The walleye pollock (and the single Polar cod individual) lay at the nexus of the other species.

---

**Figure 1.** Individual admixture analysis of cod-fish. Ordinate is individual admixture proportion in a given model. The number of ancestral populations $k$ in the model range from 2 to 5. Results ordered based on species *Arctogadus glacialis* Agl, *Boreogadus saida* Bsa, *Gadus chalcogrammus* Gch, *Gadus macrocephalus* Gma, *Gadus ogac* Gog, and for *Gadus morhua* by localities from west to east and by ecotype: Cape Cod Cco, Western Bank Web, Sable Bank Sab, Trinity Bay Tri, Southern Grand Banks Sgb, Greenland frontal BBGre, Greenland intermediate ABGre, Iceland frontal BBIce, Iceland intermediate ABIce, Iceland coastal AAIce, Barents Sea frontal BBBar, White Sea Whi, Norway Nor, Faroe Islands Far, North Sea Nse, Celtic Sea Cel, western Baltic W-Bal, and eastern Baltic E-Bal. Based on 15 million variable sites from the entire genome after filtering.

Our interpretation of the evidence was that walleye pollock is a hybrid between Atlantic and Arctic cod. Walleye pollock shares two morphological traits with Arctic cod: an absent or much reduced chin barbel sensory organ and a forked tail, which define genera (*sensu* Svetovidov, 1948) within the Gadinae. Fisheries survey experts have difficulty in distinguishing between older slow-growing Arctic cod and younger fast-growing walleye pollock. These facts independently support our thesis that walleye pollock is a homoploid hybrid species.
Figure 2. Individual admixture analysis of Pacific, Arctic and Atlantic cod-fish. Based on whole genome sequencing with 20 – 30× coverage of entire genome. *Boreogadus saida* Bsa, *Gadus chalcogrammus* Gch, *Gadus macrocephalus* Gma, *Gadus ogac* Gog, and *Gadus morhua* Gmo individuals from Sable Bank Sab, Trinity Bay Tri, Iceland Ice, and North Sea Nse. Based on 32.9 million variable sites from the entire genome after filtering.

The minimum evolutionary tree of whole-genome genetic distances (Supplemental Figure 2) had the same topology as the mtDNA tree. However, this result was dependent on monophyly of speciation by lineage splitting. We refuted this hypothesis in the case of walleye pollock. Furthermore, the timing of the colonization of the Pacific by walleye pollock cannot be estimated from the dichotomous tree. Our analysis implies that it might have happened even as recently as 200 years ago, when the species was discovered and described. We also question the biogeographical hypothesis that Pacific cod is of an Atlantic cod invasion of the Pacific ocean and that Greenland cod is of a re-invasion of the Atlantic by Pacific cod. A more parsimonious single-invasion biogeographical hypothesis is that Greenland cod is a speciation from Atlantic cod at Greenland and that Pacific cod is of an invasion of Greenland cod into the Pacific ocean (see Figure 1 and Supplemental Figure 2).


Among Atlantic cod, an individual admixture analysis revealed western cod, eastern cod and the frontal ecotype as separate entities ($k = 3$) (Figure 4 and Supplemental Figure 1). All three remain distinct entities at all $k$ values, and there is a hybrid zone in the western Baltic. The coastal fish splits into Icelandic, White Sea, Norway and North Sea fish on the one hand, and the Faroe Islands and the Celtic Sea on the other, and the Baltic cod is clearly divergent (Supplemental Figure 3). The genomes of the intermediate (*Pan I AB*) of Greenland and Iceland are about 50% frontal and 50% coastal ecotypes (Figure 4, $k = 5$). However, the *Pan I* locus, which is a proxy for behavioural ecotypes, is located in a large genomic island of divergence on linkage group LG01. Thus, variation in LG01 may have overriding influence on the admixture patterns (Figure 4).

To address this question, we analysed separately the LG02 to LG23 linkage groups, and then added LG01. We also analysed genomic regions that did not map to linkage groups (see methods).

The assessment of population differentiation by sliding-window $F_{ST}$ between north and south in the eastern Atlantic (Extended Data Figure 4) revealed extensive islands of divergence in LG01, LG02, LG07 and LG12, as observed previously. However, the LG01 and LG07 islands were complex, as they were divided by subregions showing no differentiation, thus indicating the independence of the subregions. Most other linkage groups also showed smaller regions (mini islands) of high differentiation. The comparison of north and south with west revealed a higher level of differentiation throughout the genome, in accordance with the admixture analysis, which suggests that western cod is a distinct entity. The LG07 island appeared in north/south and west/south but
not in west/north comparisons. Therefore, it resulted in the south (coastal ecotype). Similarly, the LG01 island characterized north (frontal/intermediate) and the LG12 island differed between north and south. The fact that genomic islands differed between geographic regions has implications for the interpretation of long-distance linkage disequilibrium. 

Our interpretation of the totality of the genomic evidence (based on up to 8.6 million variable sites after filtering) was that the shallow-water coastal (some AA) and deep-sea frontal (some BB) ecotypes are reproductively isolated coastal and frontal species. They are cryptic species within the Atlantic cod complex that have adapted to environmental factors in shallow and deep waters and diverging at LG01. They hybridized and formed a new homoploid hybrid species. There is normal Mendelian segregation of variants within the true-breeding hybrid species (such as the Panl AA, AB, and BB on LG01). This addresses the contradictory results regarding intermediate forms and has implications for the interpretation of the relative importance of these groups (e.g., the contribution of the different Greenland ecotypes to the Icelandic population and vice versa).

The hybrid individuals defined by posterior membership probabilities shown in the genomic results (Supplemental Figure 6) were on average intermediate regarding phenotype and habitat use between the two pure types (Supplemental Figure 7). However, hybrids were also more variable and transgressed the phenotype of the parental forms. Considerable variation was observed regarding behaviour, as determined using storage-tag data, among individuals classified by the Pan I genotype. There were AA individuals that showed a frontal behaviour and BB individuals that exhibited a coastal behaviour (see appendix in ). This behavioural variation may reflect species differences within the Pan I genotypes. This heterogeneity of behavioural types provides independent support for our thesis.

It was difficult to reconcile the long-distance linkage disequilibrium and the behaviour of genomic islands observed here (for example with certain genomic islands specific to a particular location) with a model of speciation with gene flow. A modified model of divergence after speciation explains inter-chromosomal correlations and is applicable to this case of homoploid hybrid speciation. Under this model we assume rapid speciation by hybridization of already divergent lineages (Figure 4), possibly involving chromosomal rearrangements and further divergence after speciation. The large genomic islands are islands of ecological adaptation and/or sites of low recombination and not necessarily islands of speciation. Speciation genes may reside elsewhere in the genome.
Conclusions

Our study demonstrates the power of whole-genome sequencing and population genomics in providing deep insights into fundamental processes of speciation. The results raise many interesting hypotheses about the extent of hybrid speciation and its role in cryptic biodiversity in the sea. Our results also raise hypotheses about marine fish in general and about cod-fish biology in particular, which remain to be tested. For example, we need to examine the criteria of hybrid speciation and hybrid sterility, as well as structural chromosomal variation, of hybridization, biogeography and natural selection and adaptation to diverse environments such as deep and shallow or low salinity waters. Our results call for the re-evaluation of previous work on cod-fish with implications for resource management. The hybrid nature of both the walleye pollock and Atlantic cod raises the question concerning the extent to which very profitable fisheries depend on hybrid vigour. For example, did the decimation of the hybrid and frontal fish influence the collapse and non-recovery of the western fisheries? Ocean changes and their impact of ecosystems and fisheries have unpredictable consequences. The ongoing climate warming and northern hemisphere ice melting predict species interchange between the north Pacific and the Atlantic. The potential exists for further hybridization of Pacific, Arctic and Atlantic taxa, with unknown consequences for biodiversity.
Population sampling We randomly sampled over 200 individual cod-fish from our large sample collection of greater than 20,000 Atlantic, Arctic and Pacific cod-fish individuals. We stratified the sampling to cover the widest geographic range possible with our database. We sampled Pacific cod Gadus macrocephalus (Tilesius, 1810) (mnemonic: Gma) and walleye pollock Gadus chalcogrammus (Pallas, 1814) (Gch) from the Pacific Ocean. We sampled Polar cod Arctogadus glacialis (Dryagin, 1932) (Agl), Arctic cod Boreogadus saida (Lepechin, 1774) (Bsa), and Greenland cod (uvak) Gadus ogac (Richardson, 1836) (Gog) from the Arctic at western Greenland. We sampled Atlantic cod Gadus morhua (Linnaeus, 1758) (Gmo) from throughout its distribution in the North Atlantic Ocean. The localities range west to east from off Chatham on Cape cod (mnemonic Cco), the Western Bank (Web) and Sable Bank (Sab) off Nova Scotia, Trinity Bay (Tri) and the Southern Grand Banks (Sgb) of Newfoundland. Collectively we called these localities the West. We sampled from the west and east coast of Greenland (Gre), around Iceland (Ice), the Barents Sea (Bar), the White Sea (Whi), and coastal Norway (Nor). Collectively we called these localities the North. Furthermore we sampled from the Faroe Islands (Far), the North Sea (Nse), the Celtic Sea (Cel), and the Western and Eastern Baltic (Bal-W and Bal-E respectively). Collectively we called these localities the South. The sample from Iceland included a number of individuals of the three genotypes of the Pan I locus (AA, AB, and BB) located on linkage group LG01 which we used as a proxy to identify behaviourally consistent ecotypes of coastal, intermediate, and frontal cod as defined by results from storage-tags data.}

The molecular and morphological relationship and biogeography of these taxa have been discussed and the most comprehensive account is based on mitochondrial genomics. Coulson et al consider Arctic cod (Bsa) to be an outgroup for all these taxa. Atlantic cod (Gmo) and walleye pollock (Gch) are the most closely related taxa and Pacific cod (Gma) slightly more distant. They argue that Pacific cod and walleye pollock represent two separate but nearly simultaneous invasions of the Pacific Ocean. They date the Atlantic cod vs Pacific cod split at 4 Mya and the Atlantic cod vs walleye pollock split at 3.8 Mya using conventional rates of mtDNA evolution. They suggested a nomenclature revision from Theragra chalcogramma to Gadus chalcogrammus (Pallas, 1814) for walleye pollock that has been accepted by the American Fisheries Society. We follow the new nomenclature here. According to their view Greenland cod (Gog) is a recent reinvasion of Pacific cod (Gma) into the Arctic and they consider it to be a subspecies of Pacific cod. Their analysis is based on a hypothesis of speciation by lineage splitting. We designed our sampling partly to investigate these relationships.

Molecular analysis We prepared samples for high-coverage (20–30×) and for low-coverage (2×) sequencing on the Illumina HiSeq 2500 platform. We randomly selected one individual each of Arctic cod (Bsa), Pacific cod (Gma), walleye pollock (Gch), and Greenland cod (Gog) for a 20× sequencing coverage. We randomly selected six Atlantic cod (Gmo), two from the West, two from Iceland, and two from the North Sea for 30× sequencing coverage. For the low-coverage analysis we randomly selected about a dozen individuals of Arctic cod (Bsa), Pacific cod (Gma), walleye pollock (Gch), and Greenland cod (Gog), one individual of Polar cod (Agl), and 152 Atlantic cod (Gmo) individuals from various localities.

Our tissue collection is primarily gill tissue but also has fin clips and muscle tissue and DNA isolated from blood. Tissues are stored in 96% ethanol. We isolated genomic DNA from 10 individuals selected for high-coverage sequencing using the NucleoSpin® Tissue kit (Machery-Nagel, reference 740952.50) following the manufacturer’s protocol. We isolated genomic DNA from 200 individuals selected for low-coverage sequencing using the E.Z.N.A.® Tissue DNA Kit (Omega biotech) following the manufacturer’s protocol.

We quantified and estimated 260/280 and 260/230 quality cutoffs of the genomic DNA using a NanoDrop 2000 UV-Vis Spectrophotometer (Thermo Scientific). We also quantified genomic DNA using fluorescent detection with the Qubit® dsDNA HS Assay Kit (Life Technologies). Libraries for the high-coverage sequencing were made at the Bauer Core Facility at Harvard University. The facility used the Covaris S2200© (Covaris) to shear the genomic DNA to a target size of 550 bp. The Apollo 324© (Wafergen Biosystems) system was used to generate libraries for DNA sequencing. Each library had a single index. The size distribution of the libraries was determined with the Agilent Bioanalyzer (Agilent Technologies). Library sample concentration was determined with qPCR according to an Illumina protocol.

We prepared libraries for low coverage sequencing using the Nextera® DNA Library Preparation Kit (Illumina, FC-121-1031). We used the Nextera® Index Kit (FC-121-1031) with dual indices. Index N517 was used instead of N501. We followed the manufacturer’s Nextera protocol. We cleaned the fragmented DNA with a Zymo Purification Kit (ZR-96 DNA Clean & Concentrator-5, Zymo Research). We cleaned PCR products and size selected with Ampure XP beads (Beckman Coulter Genomics, reference A63881). We used the modification recommended for PCR clean-up for 2×250 runs on the MiSeq using 25µl of Ampure XP beads (instead of 30µl) for each well of the NAP2 plate. We quantified the individual libraries with fluorescent detection using a Quant-it® PicoGreen® dsDNA Assay Kit Quantit kit on a Spectramax i3x Multi-Mode Detection Platform (Molecular Devices). We determined the size distribution of 12 randomly chosen libraries. We then normalized the multiplexed DNA libraries to 2nM concentrations and pooled the libraries.
size distribution of the pooled libraries was determined by the Bauer Core Facility using the Agilent Bioanalyzer. Pooled library sample concentration was determined with qPCR.

Pooled libraries were sequenced on the HiSeq 2500 in rapid run mode (paired-end, 2 × 250 cycles) at the Bauer Core Facility at Harvard University.

Statistical analysis. The Bauer Core Facility through the department of Informatics and Scientific Applications returned the base-called data as de-multiplexed fastq files. Individuals were sequenced on two lanes of the HiSeq 2500. We merged both of the forward and the reverse reads for each individual. We fetched the Gadus_morhua.gadMor1.dna.toplevel.fa genomic reference sequence from www.ensemble.org and used it as a reference sequence. We aligned the merged fastq reads to the reference using bwa mem.46 We used samtools to generate sorted and indexed bam files from the sam files.

Next generation sequencing data of this kind has high error rates from multiple sources, including base-calling and alignment errors.5 The low coverage data suffer from these errors in particular. Such errors will affect downstream analysis that depend on calling SNPs and genotypes because errors will be compounded. We adopted the strategy of using methods based on genotype likelihoods, implemented in the ANGSD and related software.47–51 These methods yield SNP and genotype information and an associated uncertainty facilitating unbiased or low-biased statistical interpretation of low-coverage data. Using these tools allowed us to sequence a larger number of individuals.

We used NGSadmix44 to estimate individual admixture proportions based on a model of $k = 2 \ldots n$ ancestral populations. We used ngsDist50 to estimate pairwise genetic distances and used fastME (version 2.0.75) to make minimum evolution phylogeny from these distances. We used ngsCovar56 to estimate a covariance matrix for principal components (DAPC).

The analysis is based on not calling genotypes or alleles but instead using genotype likelihoods as implemented in the ANGSD software for fast analysis of large samples. We used realSFS of ANGSD to estimate site frequency spectra and $F_{ST}$. We used ANGSD and friends for quality filtering. We typically used a minimum mapping quality of 30 and minimum base quality of 20 (\(-\text{minMapQ} 30 \text{ and minQ} 20\)) and discarded bad reads (\(-\text{remove_bads} 1\)) filtered out sites with a minor allele frequency less than 0.05 (\(-\text{minMapQ} 0.05\)), filtered by number of individuals (e.g. \(-\text{minInd} 10\)), and filtered out sites that are very likely to be polymorphic with a P value less than $10^{-5}$ (\(-\text{SNP_pval} 1e-6\)). We did a pairwise sliding-window $F_{ST}$ analysis with a window size of 10,000 and a step size of 2,000.56 As an example of the effects of filtering the admixture analysis of Atlantic cod had 22.6 million sites that were reduced to 8.6 million sites after filtering.

We used R54 and various R scripts from the ANGSD and friends packages and our own functions and scripts to manipulate and plot the results. We did Principal Component Analysis (PCA) using the eigen function in R and plotted results with ggplot2.57 We used adegenet51 for Discriminant Analysis of Principal Components (DAPC). We performed cross-validation (CV) using the xvalDapc function52 to determine the number of principal components (PCs) to retain in the DAPC. We used the posterior membership probabilities returned by DAPC of the whole-genome sequence data to define “species” as priors for a DAPC of phenotypic and habitat use data.

Large genomic islands of divergence in the Atlantic cod genome46,72 are already known. The divergence of these islands may influence analysis of genome wide effects. To address these problems we separately analysed the parts of the genome that can be mapped to specific linkage groups as well as the parts that cannot be mapped to linkage groups. In order to map Atlantic cod genomic scaffolds to the Atlantic cod genome map, we used a local blast on the 120 base pairs (bp) surrounding the SNPs used for mapping onto the reference sequence and chose matches that were both greater than 100 bp and greater than 95% identity. The genetic map is sparse and about 235 megabase (Mb) (roughly 1/3 of the genome) can be thus assigned to specific linkage groups. We refer to the parts of the genome that map to linkage groups LG02 to LG23 as that part of the genome which definitively does not map to linkage group LG01. We then add LG01 data and refer to it as the part of the genome that is mapped to linkage groups. We also analyse the parts of the genome (about 2/3) that cannot be thus mapped to specific linkage groups.

The computations in this paper were run on the Odyssey cluster supported by the FAS Division of Science, Research Computing Group at Harvard University.
References


**Acknowledgements**

We thank Hopi Hoekstra for laboratory space, Kyle Turner for technical help, John Wakeley for help with computing resources, and Rasmus Nielsen for help on design. We thank Karsten Hartel, Andrew Williston and Ubaldo Benitez Hernandez for help with specimen transport. We thank Grant Pogson, Michael Canino, Jarle Mork and Kristján Kristinsdon for help with sampling. We also especially thank Dick Lewontin who made it all possible.

The work was supported by a grant from Svala Árnadóttir private fund, by a grant from the University of Iceland Research Fund, and by institutional funds from R.C. Lewontin.

**Author contributions statement**

E.Á. and K.H. conceived the experiments, E.Á. and K.H. conducted the experiment(s), E.Á. and K.H. analysed the results. Both authors wrote and reviewed the manuscript.

**Additional information**

Accession codes: SRP065670. We request that users of the data honor Fort Lauderdale data sharing principles of granting us first publication of results from our genomic data.

**Competing financial interests:** The authors declare no competing financial interests.
Supplementary data for Paper I

Supplemental Material for the Manuscript: 
Organization of a β and α Globin Gene Set in the Teleost Atlantic Cod, *Gadus morhua* 

Katrín Halldórsdóttir¹, Einar Árnason¹ 

Institute of Biology 
University of Iceland 
Sturlugata 7 
Reykjavik 
Iceland 

Received: 19 October 2008 / Accepted 8 May 2009 / Published online:
Further details on multiple linked $\beta$ and $\alpha$ globin genes will be published elsewhere (Halldórsdóttir & Árnason 2008).

### Table S1  PCR primers. A set of primers used to amplify $\beta$ gene (BL/BR in name) and $\alpha$ gene (AR/AL in name).

<table>
<thead>
<tr>
<th>Name</th>
<th>Oligonucleotide</th>
</tr>
</thead>
<tbody>
<tr>
<td>GmHBBBL29</td>
<td>5'-CCAACAACATCAGCAACC-3'</td>
</tr>
<tr>
<td>GmHBRR532</td>
<td>5'-TTGTTGACAGAATCTGCAAA-3'</td>
</tr>
<tr>
<td>GmHBAL26</td>
<td>5'-GAAAGCAACTCTCTGAAACGTCAA-3'</td>
</tr>
<tr>
<td>GmHBAR553</td>
<td>5'-ACCATTGAAACGGACCACAT-3'</td>
</tr>
<tr>
<td>GmHBBBL29_long</td>
<td>5'-CCAACAACATCAGCAACCATTGAGTGGACAGATAGTGAGC-3'</td>
</tr>
<tr>
<td>GmHBAR553_long</td>
<td>5'-ACCATTGAAACGGACCACATGATGGCCGGAGTCTTCA-3'</td>
</tr>
</tbody>
</table>

### Table S2  Walking primers for sequencing the 3000 base pair linked $\beta$ and $\alpha$ gene set. Primers are listed in sequential order of the walk and with direction.

<table>
<thead>
<tr>
<th>Name</th>
<th>Oligonucleotide</th>
<th>Direction</th>
</tr>
</thead>
<tbody>
<tr>
<td>M13Forwardlong</td>
<td>5'- CGTTGAAAACCCACGACCGCCAG</td>
<td>3'- Forward</td>
</tr>
<tr>
<td>GmHBBseq06</td>
<td>5'- CGTGAAGACATCGACGCTAAAC</td>
<td>3'- Forward</td>
</tr>
<tr>
<td>GmHBBlseq01</td>
<td>5'- TCTTCTCCCTCTCCCCCATAC</td>
<td>3'- Forward</td>
</tr>
<tr>
<td>GmHBSseq05</td>
<td>5'- GTCAACATCGTCAACACAACG</td>
<td>3'- Forward</td>
</tr>
<tr>
<td>GmHBBlseq02</td>
<td>5'- GCCCGTAAATCTCAGTCCCTT</td>
<td>3'- Forward</td>
</tr>
<tr>
<td>GmHBARseq02</td>
<td>5'- GGTCACGCAATTCAATAGGC</td>
<td>3'- Reverse</td>
</tr>
<tr>
<td>GmHBSseq04</td>
<td>5'- GCTTCTCTACCTACATGTATCCT</td>
<td>3'- Forward</td>
</tr>
<tr>
<td>GmHBARseq01</td>
<td>5'- TATATGGTGACACAGAAGC</td>
<td>3'- Reverse</td>
</tr>
<tr>
<td>GmHBSseq03</td>
<td>5'- CTGAACGTCAACATGAGTCTCT</td>
<td>3'- Forward</td>
</tr>
<tr>
<td>M13Reverselong</td>
<td>5'- CACACAGGAAACAGCTATGAC</td>
<td>3'- Reverse</td>
</tr>
</tbody>
</table>
### Table S3: Genscan output for sequence SS104.1. Predicted genes and their exons and control elements.

<table>
<thead>
<tr>
<th>Gn.Ex</th>
<th>Type</th>
<th>S</th>
<th>Begin</th>
<th>End</th>
<th>Len</th>
<th>Fr</th>
<th>Ph</th>
<th>I.Ac</th>
<th>Do.T</th>
<th>CodRg</th>
<th>P</th>
<th>Tscr</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.01</td>
<td>Intr+</td>
<td>10</td>
<td>101</td>
<td>92</td>
<td>0</td>
<td>2</td>
<td>101</td>
<td>46</td>
<td>153</td>
<td>1</td>
<td>12.36</td>
</tr>
<tr>
<td>2</td>
<td>1.02</td>
<td>Intr+</td>
<td>214</td>
<td>436</td>
<td>223</td>
<td>1</td>
<td>1</td>
<td>116</td>
<td>63</td>
<td>567</td>
<td>1</td>
<td>55.13</td>
</tr>
<tr>
<td>3</td>
<td>1.03</td>
<td>Term+</td>
<td>552</td>
<td>680</td>
<td>129</td>
<td>2</td>
<td>0</td>
<td>72</td>
<td>43</td>
<td>244</td>
<td>1</td>
<td>16.48</td>
</tr>
<tr>
<td>4</td>
<td>1.04</td>
<td>PlyA+</td>
<td>712</td>
<td>717</td>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1.05</td>
</tr>
<tr>
<td>5</td>
<td>2</td>
<td>Prom+</td>
<td>1096</td>
<td>1135</td>
<td>40</td>
<td>−</td>
<td></td>
<td></td>
<td></td>
<td>−6.36</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>2.01</td>
<td>Sngl+</td>
<td>1442</td>
<td>1579</td>
<td>138</td>
<td>1</td>
<td>0</td>
<td>19</td>
<td>42</td>
<td>521</td>
<td>0.55</td>
<td>34.20</td>
</tr>
<tr>
<td>7</td>
<td>2.02</td>
<td>PlyA+</td>
<td>1645</td>
<td>1650</td>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1.05</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>3</td>
<td>Prom+</td>
<td>2218</td>
<td>2257</td>
<td>40</td>
<td>−</td>
<td></td>
<td></td>
<td></td>
<td>−1.96</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>3.01</td>
<td>Init+</td>
<td>2303</td>
<td>2397</td>
<td>95</td>
<td>1</td>
<td>2</td>
<td>90</td>
<td>80</td>
<td>87</td>
<td>0.56</td>
<td>7.95</td>
</tr>
<tr>
<td>10</td>
<td>3.02</td>
<td>Intr+</td>
<td>2525</td>
<td>2732</td>
<td>208</td>
<td>2</td>
<td>1</td>
<td>102</td>
<td>102</td>
<td>279</td>
<td>0.89</td>
<td>29.78</td>
</tr>
<tr>
<td>11</td>
<td>3.03</td>
<td>Term+</td>
<td>2926</td>
<td>2954</td>
<td>129</td>
<td>2</td>
<td>0</td>
<td>86</td>
<td>49</td>
<td>178</td>
<td>0.89</td>
<td>11.88</td>
</tr>
</tbody>
</table>

---

### Table S4: FGENESH gene prediction output for concatenated sequence SS104.1 and λ1.6BamHIClone21. The 5' end of λ1.6BamHIClone21 defines start of concatenated sequence. TSS - Position of transcription start (TATA-box position and score).

<table>
<thead>
<tr>
<th>Feature</th>
<th>Start</th>
<th>End</th>
<th>Score</th>
<th>ORF Len</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>TSS</td>
<td>1667</td>
<td>−7.94</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>CDSei</td>
<td>1747</td>
<td>1838</td>
<td>14 1747</td>
</tr>
<tr>
<td>3</td>
<td>CDSei</td>
<td>1951</td>
<td>2173</td>
<td>44.10 1952</td>
</tr>
<tr>
<td>4</td>
<td>CDSei</td>
<td>2289</td>
<td>2417</td>
<td>18.20 2289</td>
</tr>
<tr>
<td>5</td>
<td>PolA</td>
<td>2449</td>
<td>1.25</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>TSS</td>
<td>2839</td>
<td>−10.14</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>CDSo</td>
<td>3179</td>
<td>3316</td>
<td>15.37 3179</td>
</tr>
<tr>
<td>8</td>
<td>PolA</td>
<td>3382</td>
<td>1.25</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>TSS</td>
<td>3959</td>
<td>−4.84</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>CDSei</td>
<td>4040</td>
<td>4134</td>
<td>13.59 4040</td>
</tr>
<tr>
<td>11</td>
<td>CDSei</td>
<td>4262</td>
<td>4469</td>
<td>29.20 4263</td>
</tr>
<tr>
<td>12</td>
<td>CDSei</td>
<td>4563</td>
<td>4691</td>
<td>22.12 4563</td>
</tr>
</tbody>
</table>
Supplemental Materials and Methods

Screening of λ Genomic Library and Subcloning

The probability of having any given DNA sequence in the library can be calculated from the equation

\[ N = \frac{\ln(1-p)}{\ln(1-f)} \]

where \( p \) = desired probability, \( f \) = fractional proportion of the genome in a single recombinant and \( N \) = necessary number of recombinants. To achieve 99% probability of having a given DNA sequence represented in a library of an average fragment size of 12 kb from the Atlantic cod genome (1.0 × 10^9 bp) requires

\[ N = \frac{\ln(1-0.99)}{\ln(1-(1.2 \times 10^5)/(1.0 \times 10^9))} = 3.8 \times 10^7 \text{ recombinants (Sambrook et al. 1989).} \]

We doubled this number and plated out 7.5 × 10^5 pfu in E. coli strain K802. Phage plaques were transferred to nylon filters (Hybond-N nylon membrane, Amersham Biosciences) and the DNA was crosslinked by UV exposure of 70,000 microjoules/cm² (UV crosslinker, Amersham Biotech). The filters were prehybridised at 65°C in 5× SSC (Saline Sodium Citrate). Hybridization was performed overnight at 65°C followed by stringency washes (2× SSC and 0.1% SDS; 1× SSC and 0.1% SDS; 0.1× SSC and 0.1% SDS). We made probes by pooling PCR amplifications of coding sequences of amplified β and α genes, radioactively labeled with rediprime™ II random prime labelling system (Amersham Biosciences) using Redivue 32P–dCTP 370 MBq/ml (Amersham Biosciences).

We picked positively hybridizing clones, replated and reprobed them. This procedure was repeated until all clones were positive. Positive clones were amplified in a liquid culture of E. coli strain K802. DNA was isolated with QIAGEN® Lambda Kit (Qiagen). We used SalI, BamHI and EcoRI endonuclease enzymes (Fermentas) to digest the DNA from phage clones to map fragments. We also plugged several restriction fragments from a 1% TAE agarose gel and subcloned into the pUC19 vector. The subcloned DNA from clone λ1.6BamHIClone21 was sequenced using two vector primers: M13F-long 5′–CGTTGTAAAACGACGGCCAG–3′ and revseq-48 5′–AGCGGATAACAATTTCACACAGGA–3′. From the sequence obtained we made two primers for walking: seq01L 5′–GTTCAGGCCCATACATT–3′ and seq02L 5′–ATTCACAAGAAGGGCTGCAC–3′. Together these primers yielded high quality sequences of the λ subclones reported in this paper.
Fig. S1. Phylogeny of β globins in Atlantic cod and model teleosts. The tree is based on amino acid sequences aligned with CLUSTAL W. The tree was build with Neighbor-joining method and Poisson corrected distances in the program MEGA4. Numbers above branches indicate Bootstrap values from 500 replicates. Myoglobin of Gadus morhua was used as outgroup. Sequences were taken from GenBank; accession numbers: Gadus morhua-Verde1 (P84610), Gadus morhua-Verde2 (P84611), Gadus morhua-myoglobin (ABL7386), Gadus morhua-Birley (CAA66903), Takifugu rubripes-adult (AAO61493), Gadus morhua-SS104.1 (EF644886), Gadus morhua-cDNA (unpublished KH), Oryzias latipes (ABO83077, ABO83078), Danio rerio-embryonic-2 (AAP93668), Danio rerio-embryonic-3 (AAP93667), Danio rerio-adult-1 (NP_571095), Danio rerio-adult-2 (NP_001005403).
Fig. S2 Phylogeny of α globins in Atlantic cod and model teleosts. The tree is based on amino acid sequences aligned with CLUSTAL W. The tree was built with Neighbor-joining method and Poisson corrected distances in the program MEGA4. Numbers above branches indicate Bootstrap values from 500 replicates. Myoglobin of Gadus morhua was used as outgroup. Sequences were taken from GenBank; accession numbers: Gadus morhua-Verde1 (P84609), Gadus morhua-Verde2 (O42425), Gadus morhua-myoglobin (ABL7386), Gadus morhua-Birley (CAA66866), Takifugu rubripes-adult-3 (AAO61494), Gadus morhua-cDNA (unpublished KH), Oryzias latipes (ABO83077, ABO83078), Danio rerio-embryonic-3 (NP_898889).  

References

Supplementary data for Paper II

Supplemental material for the manuscript:
Multiple Linked $\beta$ and $\alpha$ Globin Genes in Atlantic Cod: a PCR Based Strategy of Genomic Exploration

Katrín Halldórsdóttir$^a$, Einar Árnason$^a$

$^a$Institute of Biology
University of Iceland
Sturlugata 7
Reykjavík
Iceland

The nucleotide sequences analyzed in the present paper have the GenBank accession numbers EF644855 – EF644911.

*Corresponding author
Email addresses: katrinhalldorsdottir@gmail.com (Katrín Halldórsdóttir)

Preprint submitted to Marine Genomics September 25, 2009
Further details on the tail to head orientation of linked $\beta$ and $\alpha$ globin genes and their control regions are found in an accompanying paper Halldórsdóttir and Árnason (Halldórsdóttir and Árnason, 2009).

**Supplemental Table S1:** Names and numbers of 19 genotyped individuals and their clones. FF, FS and SS refer to HbI genotype. Number after that refers to individual and number after dot to clone from that individual.

<table>
<thead>
<tr>
<th>FF Individual</th>
<th>Clones</th>
</tr>
</thead>
<tbody>
<tr>
<td>FF127</td>
<td>FF127.1</td>
</tr>
<tr>
<td></td>
<td>FF127.2</td>
</tr>
<tr>
<td></td>
<td>FF127.3</td>
</tr>
<tr>
<td></td>
<td>FF127.4</td>
</tr>
<tr>
<td></td>
<td>FF127.5</td>
</tr>
<tr>
<td></td>
<td>FF127.7</td>
</tr>
<tr>
<td></td>
<td>FF127.8</td>
</tr>
<tr>
<td></td>
<td>FF127.9</td>
</tr>
<tr>
<td></td>
<td>FF127.10</td>
</tr>
<tr>
<td></td>
<td>FF127.11</td>
</tr>
<tr>
<td></td>
<td>FF127.12</td>
</tr>
<tr>
<td></td>
<td>FF127.13</td>
</tr>
<tr>
<td>FF4</td>
<td>FF4.1</td>
</tr>
<tr>
<td></td>
<td>FF4.2</td>
</tr>
<tr>
<td></td>
<td>FF4b.1</td>
</tr>
<tr>
<td></td>
<td>FF4b.2</td>
</tr>
<tr>
<td>FF10</td>
<td>FF10.1</td>
</tr>
<tr>
<td>FF16</td>
<td>FF16.1</td>
</tr>
<tr>
<td>FF20</td>
<td>FF20.1</td>
</tr>
<tr>
<td>FF23</td>
<td>FF23.1</td>
</tr>
<tr>
<td>FF24</td>
<td>FF24.1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>SS Individual</th>
<th>Clones</th>
</tr>
</thead>
<tbody>
<tr>
<td>SS103</td>
<td>SS103.1</td>
</tr>
<tr>
<td></td>
<td>SS103.3</td>
</tr>
<tr>
<td></td>
<td>SS103.4</td>
</tr>
<tr>
<td></td>
<td>SS103.5</td>
</tr>
<tr>
<td></td>
<td>SS103.6</td>
</tr>
<tr>
<td></td>
<td>SS103.7</td>
</tr>
<tr>
<td></td>
<td>SS103.8</td>
</tr>
<tr>
<td></td>
<td>SS103.10</td>
</tr>
<tr>
<td></td>
<td>SS103.11</td>
</tr>
<tr>
<td></td>
<td>SS103.12</td>
</tr>
<tr>
<td></td>
<td>SS103.13</td>
</tr>
<tr>
<td>SS51</td>
<td>SS51.1</td>
</tr>
<tr>
<td>SS104</td>
<td>SS104.1</td>
</tr>
<tr>
<td>SS129</td>
<td>SS129.1</td>
</tr>
<tr>
<td>SS130</td>
<td>SS130.1</td>
</tr>
<tr>
<td>SS131</td>
<td>SS131.1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>FS Individual</th>
<th>Clones</th>
</tr>
</thead>
<tbody>
<tr>
<td>FS113</td>
<td>FS113.1</td>
</tr>
<tr>
<td></td>
<td>FS113.2</td>
</tr>
<tr>
<td></td>
<td>FS113.3</td>
</tr>
<tr>
<td></td>
<td>FS113.4</td>
</tr>
<tr>
<td></td>
<td>FS113.5</td>
</tr>
<tr>
<td></td>
<td>FS113.6</td>
</tr>
<tr>
<td></td>
<td>FS113.7</td>
</tr>
<tr>
<td></td>
<td>FS113.9</td>
</tr>
<tr>
<td></td>
<td>FS113.10</td>
</tr>
<tr>
<td></td>
<td>FS113.11</td>
</tr>
<tr>
<td></td>
<td>FS113.12</td>
</tr>
<tr>
<td></td>
<td>FS113.13</td>
</tr>
<tr>
<td>FS14</td>
<td>FS14.1</td>
</tr>
<tr>
<td>FS36</td>
<td>FS36.1</td>
</tr>
<tr>
<td>FS43</td>
<td>FS43.1</td>
</tr>
<tr>
<td>FS54</td>
<td>FS54.1</td>
</tr>
<tr>
<td>FS82</td>
<td>FS82.1</td>
</tr>
</tbody>
</table>
**Supplemental Table S2**: Triplet phylogenetically informative sites of an approximately 3 kilobase \( \beta/\alpha \) globin gene region among 57 cloned contigs from genomic DNA of Atlantic cod. \( FF, FS, \) and \( SS \) refer to HbI genotypes; numbers after genotype refer to individual; numbers after dot refer to a clone from that individual. Phylogenetically informative sites are defined by a strict criterion of independence of being found in clones from three separate individuals, thus derived from three separate PCR and cloning events. Numbers in boxhead refer to nucleotide position in the 3 kilobase contigs. Nucleotide positions are read vertically, thus first site is number 47. Site numbers indicated starte at 5\(^{\prime} \) end.

<table>
<thead>
<tr>
<th>Segregating site</th>
<th>Clone</th>
<th>769249268123881594383389102372056365</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>FF127.1</td>
<td>CCACCTCGCCCACCTACGATGATCCAAGAAAAGACAA</td>
</tr>
<tr>
<td></td>
<td>FF4b.1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>FF127.2</td>
<td>. . . T . . . . . . . . . . . . . . . . .</td>
</tr>
<tr>
<td></td>
<td>FF127.7</td>
<td>. . . T . . . . . . . . . . . . . . . . .</td>
</tr>
<tr>
<td></td>
<td>FF16.1</td>
<td>. . . T . . . . . . . . . . . . . . . . .</td>
</tr>
<tr>
<td></td>
<td>FF127.10</td>
<td>. . . T . . . . . . . . . . . . . . . . .</td>
</tr>
<tr>
<td></td>
<td>FF127.13</td>
<td>. . . T . . . . . . . . . . . . . . . . .</td>
</tr>
<tr>
<td></td>
<td>FF127.3</td>
<td>. . . T . . . . . . . . . . . . . . . . .</td>
</tr>
<tr>
<td></td>
<td>FF127.12</td>
<td>. . . T . . . . . . . . . . . . . . . . .</td>
</tr>
<tr>
<td></td>
<td>FF127.4</td>
<td>. . . T . . . . . . . . . . . . . . . . .</td>
</tr>
</tbody>
</table>

Continued on next page
### Supplemental Table S2: Continuation

<table>
<thead>
<tr>
<th>Clone</th>
<th>Segregating site</th>
</tr>
</thead>
<tbody>
<tr>
<td>FF127.8</td>
<td>1111111111122222</td>
</tr>
<tr>
<td>SS131.1</td>
<td>34566666777788901223455555699003455</td>
</tr>
<tr>
<td>FF10.1</td>
<td>462501276777903748263700155037117738</td>
</tr>
<tr>
<td>FS113.10</td>
<td>769249268123881594383389102372056365</td>
</tr>
<tr>
<td>FF23.1</td>
<td>FF4.1</td>
</tr>
<tr>
<td>SS130.1</td>
<td>SS51.1</td>
</tr>
<tr>
<td>SS5.1</td>
<td>A.CT</td>
</tr>
<tr>
<td>FS54.1</td>
<td>CT</td>
</tr>
<tr>
<td>FS54.2</td>
<td>CT</td>
</tr>
<tr>
<td>FF4b.2</td>
<td>CT</td>
</tr>
<tr>
<td>FF4.2</td>
<td>CT</td>
</tr>
<tr>
<td>SS103.1</td>
<td>CT</td>
</tr>
<tr>
<td>SS103.13</td>
<td>CT</td>
</tr>
<tr>
<td>FS43.2</td>
<td>CT</td>
</tr>
<tr>
<td>SS103.4</td>
<td>CT</td>
</tr>
<tr>
<td>SS103.6</td>
<td>CT</td>
</tr>
<tr>
<td>SS103.5</td>
<td>CT</td>
</tr>
<tr>
<td>SS103.8</td>
<td>CT</td>
</tr>
<tr>
<td>SS103.11</td>
<td>CT</td>
</tr>
<tr>
<td>SS103.12</td>
<td>CT</td>
</tr>
</tbody>
</table>

Continued on next page
### Supplemental Table S2: Continuation

<table>
<thead>
<tr>
<th>Segregating site</th>
<th>Clone</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>FF20.1</td>
</tr>
<tr>
<td></td>
<td>SS103.7</td>
</tr>
<tr>
<td></td>
<td>SS103.3</td>
</tr>
<tr>
<td></td>
<td>FF127.5</td>
</tr>
<tr>
<td></td>
<td>FF127.11</td>
</tr>
<tr>
<td></td>
<td>SS103.10</td>
</tr>
<tr>
<td></td>
<td>FF127.9</td>
</tr>
<tr>
<td></td>
<td>FS129.6</td>
</tr>
<tr>
<td></td>
<td>SS129.1</td>
</tr>
<tr>
<td></td>
<td>FS43.1</td>
</tr>
<tr>
<td></td>
<td>FS113.7</td>
</tr>
<tr>
<td></td>
<td>FS113.2</td>
</tr>
</tbody>
</table>

Continued on next page
1. Amino acid variation

To analyse the polymorphism found in the coding sequences we did a comparable comparison among amino acid variants of predicted proteins. Supplemental Table S3 and supplemental Table S4 show the variable amino acids found between β and α proteins in 55 clones. Two clones, FF127.1 and FF4b.1, were left out of this comparison because their predicted proteins were clearly different. They are discussed separately below. In addition we compared our protein sequences to those of Tipping and Birley and Verde et al. (2006). Based on our strict criteria for phylogenetically informative sites, variable amino acids were not regarded informative unless found in at least two clones from separate individuals.

Of the 147 amino acids predicted in the β coding region, 41 amino acid residues were variable among the clones. However, based on our strict criterion of phylogenetically informative sites only two of them were informative in this sense. Of the 41 amino acids which were variable (Supplemental Table S3), nine were sequence specific for β globin 2 of Verde et al. (2006).

Of the 143 amino acid in the α coding region, 23 amino acid residues were variable (Supplemental Table S4). Using our strict doubleton criterion, four of them were phylogenetically
informative. Amino acid number 10 in exon 1, amino acid number 37 and 59 in exon 2, and amino acid number 102 in exon 3 (Table 2).

**Supplemental Table S3**: Variable amino acid sites in predicted β globins among 55 clones from Atlantic cod genomic DNA PCR amplification. Also included for reference is a sequence deposited in GenBank by Tipping and Birley and sequence 2 presented by Verde et al. (2006). They were excluded from analysis.

<table>
<thead>
<tr>
<th>Amino acid position</th>
<th>FF20.1</th>
<th>S A I N S I N E I G C I G N E L</th>
<th>CKLDNELHGNFVKGFDGW L VSRHQ</th>
</tr>
</thead>
<tbody>
<tr>
<td>SS51.1</td>
<td></td>
<td>. . . . . . . . . . . . . . . . .</td>
<td>. . . . . . . . . . . . . . . . .</td>
</tr>
<tr>
<td>FF127.11</td>
<td>T</td>
<td>. . . . . . . . . . . . . . . . .</td>
<td>. . . . . . . . . . . . . . . . .</td>
</tr>
<tr>
<td>FS113.5</td>
<td>T</td>
<td>. . . . . . . . . . . . . . . . .</td>
<td>. . . . . . . . . . . . . . . . .</td>
</tr>
<tr>
<td>FS113.4</td>
<td>T</td>
<td>. . . . . . . . . . . . . . . . .</td>
<td>. . . . . . . . . . . . . . . . .</td>
</tr>
<tr>
<td>SS103.10</td>
<td>T</td>
<td>. . . . . . . . . . . . . . . . .</td>
<td>. . . . . . . . . . . . . . . . .</td>
</tr>
<tr>
<td>FS113.1</td>
<td>T</td>
<td>. . . . . . . . . . . . . . . . .</td>
<td>. . . . . . . . . . . . . . . . .</td>
</tr>
<tr>
<td>FS113.7</td>
<td>T</td>
<td>. . . . . . . . . . . . . . . . .</td>
<td>. . . . . . . . . . . . . . . . .</td>
</tr>
<tr>
<td>FS113.12</td>
<td>T</td>
<td>. . . . . . . . . . . . . . . . .</td>
<td>. . . . . . . . . . . . . . . . .</td>
</tr>
<tr>
<td>FS36.1</td>
<td>T</td>
<td>. . . . . . . . . . . . . . . . .</td>
<td>. . . . . . . . . . . . . . . . .</td>
</tr>
<tr>
<td>FS113.2</td>
<td>T</td>
<td>. . . . . . . . . . . . . . . . .</td>
<td>. . . . . . . . . . . . . . . . .</td>
</tr>
<tr>
<td>FS113.6</td>
<td>T</td>
<td>. . . . . . . . . . . . . . . . .</td>
<td>. . . . . . . . . . . . . . . . .</td>
</tr>
<tr>
<td>FS113.3</td>
<td>T</td>
<td>. . . . . . . . . . . . . . . . .</td>
<td>. . . . . . . . . . . . . . . . .</td>
</tr>
<tr>
<td>SS104.1</td>
<td>T</td>
<td>. . . . . . . . . . . . . . . . .</td>
<td>. . . . . . . . . . . . . . . . .</td>
</tr>
</tbody>
</table>

Continued on next page
### Supplemental Table S3: Continuation

| Amino acid position | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
|                     | 7013458245245836776148135634380360 | 1570457 |

FF127.5 . . VT . . . . . . . D . . . E . . E
FF127.9 . . T . . . . . . . . D . . . LE .
FS113.9 . . T . . . . . . . . D . . . E . . . . P
SS103.13 . . T . . . . . . . . D . .
SS103.1 . . T . . . . . . . . D .
FF127.3 . . T . . . . . . . . D .
SS131.1 . . T . . . . . . . . D .
SS130.1 . . T . . . . . . . . D .
FS113.10 . . T . . . . . . . . D .
FF127.8 . . T . . . . . . . . D .
FF127.7 . . T . . . . . . . . D .
SS103.8 . . T . . . . . . . . D .
SS103.7 . . T . . . . . . . . D .
SS103.6 . . T . . . . . . . . D .
SS103.5 . . T . . . . . . . . D .

Continued on next page
### Supplemental Table S3: Continuation

| Amino acid position | SS103.4 | FF127.10 | FF10.1 | FF23.1 | FF16.1 | FF4.2 | FF4.1 | FS14.1 | FF127.2 | SS103.3 | FF4b.2 | SS103.11 | FF127.4 | FS43.2 | FS54.1 | FS54.2 | FF24.1 | FS43.1 | SS129.1 | Verde2 |
|---------------------|---------|----------|--------|--------|--------|-------|-------|--------|---------|---------|--------|----------|---------|--------|--------|--------|--------|--------|---------|---------|---------|
|                     | .       | .        | .      | .      | .      | .     | .     | .      | .       | .       | .      | .        | .       | .      | .      | .      | .      | .      | .       | .       | .       |
| SS103.4             | .       | .        | .      | .      | .      | .     | .     | .      | .       | .       | .      | .        | .       | .      | .      | .      | .      | .      | .       | .       | .       |
| FF127.10            | .       | .        | .      | .      | .      | .     | .     | .      | .       | .       | .      | .        | .       | .      | .      | .      | .      | .      | .       | .       | .       |
| FF10.1              | .       | .        | .      | .      | .      | .     | .     | .      | .       | .       | .      | .        | .       | .      | .      | .      | .      | .      | .       | .       | .       |
| FF23.1              | .       | .        | .      | .      | .      | .     | .     | .      | .       | .       | .      | .        | .       | .      | .      | .      | .      | .      | .       | .       | .       |
| FF16.1              | .       | .        | .      | .      | .      | .     | .     | .      | .       | .       | .      | .        | .       | .      | .      | .      | .      | .      | .       | .       | .       |
| FF4.2               | .       | .        | .      | .      | .      | .     | .     | .      | .       | .       | .      | .        | .       | .      | .      | .      | .      | .      | .       | .       | .       |
| FF4.1               | .       | .        | .      | .      | .      | .     | .     | .      | .       | .       | .      | .        | .       | .      | .      | .      | .      | .      | .       | .       | .       |
| FS14.1              | .       | .        | .      | .      | .      | .     | .     | .      | .       | .       | .      | .        | .       | .      | .      | .      | .      | .      | .       | .       | .       |
| FF127.2             | .       | .        | .      | .      | .      | .     | .     | .      | .       | .       | .      | .        | .       | .      | .      | .      | .      | .      | .       | .       | .       |
| SS103.3             | .       | .        | .      | .      | .      | .     | .     | .      | .       | .       | .      | .        | .       | .      | .      | .      | .      | .      | .       | .       | .       |
| FF4b.2              | .       | .        | .      | .      | .      | .     | .     | .      | .       | .       | .      | .        | .       | .      | .      | .      | .      | .      | .       | .       | .       |
| SS103.11            | .       | .        | .      | .      | .      | .     | .     | .      | .       | .       | .      | .        | .       | .      | .      | .      | .      | .      | .       | .       | .       |
| FF127.4             | .       | .        | .      | .      | .      | .     | .     | .      | .       | .       | .      | .        | .       | .      | .      | .      | .      | .      | .       | .       | .       |
| FS43.2              | .       | .        | .      | .      | .      | .     | .     | .      | .       | .       | .      | .        | .       | .      | .      | .      | .      | .      | .       | .       | .       |
| FS43.1              | .       | .        | .      | .      | .      | .     | .     | .      | .       | .       | .      | .        | .       | .      | .      | .      | .      | .      | .       | .       | .       |
| FS54.1              | .       | .        | .      | .      | .      | .     | .     | .      | .       | .       | .      | .        | .       | .      | .      | .      | .      | .      | .       | .       | .       |
| FS54.2              | .       | .        | .      | .      | .      | .     | .     | .      | .       | .       | .      | .        | .       | .      | .      | .      | .      | .      | .       | .       | .       |
| FS43.1              | .       | .        | .      | .      | .      | .     | .     | .      | .       | .       | .      | .        | .       | .      | .      | .      | .      | .      | .       | .       | .       |
| FS43.1              | .       | .        | .      | .      | .      | .     | .     | .      | .       | .       | .      | .        | .       | .      | .      | .      | .      | .      | .       | .       | .       |
| SS129.1             | .       | .        | .      | .      | .      | .     | .     | .      | .       | .       | .      | .        | .       | .      | .      | .      | .      | .      | .       | .       | .       |
| Verde2              | .       | .        | .      | .      | .      | .     | .     | .      | .       | .       | .      | .        | .       | .      | .      | .      | .      | .      | .       | .       | .       |
**Supplemental Table S4**: Variable amino acid sites in predicted α globins among 55 clones from Atlantic cod genomic DNA PCR amplification. Also included is a sequence deposited in GenBank by Tipping and Birley and sequence 2 presented by Verde et al Verde et al. (2006)

<table>
<thead>
<tr>
<th>Amino acid position</th>
<th>FF127.10</th>
<th>FF127.11</th>
<th>SS103.4</th>
<th>FS113.4</th>
<th>SS51.1</th>
<th>FF23.1</th>
<th>FF127.4</th>
<th>FF127.7</th>
<th>SS129.1</th>
<th>SS130.1</th>
<th>FF127.12</th>
<th>SS103.11</th>
<th>FF16.1</th>
<th>FF127.2</th>
<th>FF127.5</th>
<th>FS14.1</th>
<th>SS103.12</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>3</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>4</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>5</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>6</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>7</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>8</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>9</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>10</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
</tbody>
</table>

Continued on next page
### Supplemental Table S4: Continuation

<table>
<thead>
<tr>
<th>Amino acid position</th>
<th>1</th>
<th>1</th>
<th>1</th>
<th>1</th>
<th>1</th>
<th>1</th>
<th>1</th>
<th>1</th>
<th>1</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>8</td>
<td>9</td>
<td>10</td>
<td>11</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>0</td>
<td>6</td>
<td>9</td>
<td>0</td>
<td>7</td>
<td>0</td>
<td>4</td>
<td>8</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>7</td>
<td>5</td>
<td>9</td>
<td>0</td>
<td>2</td>
<td>6</td>
<td>4</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>4</td>
<td>5</td>
<td>9</td>
<td>0</td>
<td>9</td>
<td>0</td>
<td>2</td>
<td>6</td>
<td>4</td>
</tr>
<tr>
<td>FF127.9</td>
<td>.</td>
<td>.</td>
<td>F</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>FF20.1</td>
<td>.</td>
<td>.</td>
<td>P</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>FF4.2</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>G</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>FS36.1</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>T</td>
</tr>
<tr>
<td>Birley</td>
<td>.</td>
<td>.</td>
<td>T</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>Verde2</td>
<td>.</td>
<td>.</td>
<td>T</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>SS103.13</td>
<td>.</td>
<td>.</td>
<td>T</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>FF127.13</td>
<td>.</td>
<td>.</td>
<td>A</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>FS54.2</td>
<td>.</td>
<td>.</td>
<td>A</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>FF4b.2</td>
<td>.</td>
<td>.</td>
<td>S</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>PA</td>
</tr>
<tr>
<td>FF127.1</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>SS103.10</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>FS82.1</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>FS54.1</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>FS43.1</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>FS36.2</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>FS113.9</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>FS113.7</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>FS113.6</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>FS113.5</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>FS113.3</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>FS113.2</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
</tbody>
</table>

Continued on next page
### Supplemental Table S4: Continuation

| Amino acid position | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 2 | 2 | 3 | 3 | 4 | 4 | 5 | 6 | 8 | 9 | 0 | 0 | 1 | 2 | 2 | 2 | 3 | 3 |
| FS113.1             |   |   |   |   |   |   |   |   |   |   |   |   |   |   | P |   |   |   |   |   |   |   |   |   |   |   |   |
| FS113.13            |   |   |   |   |   |   |   |   |   |   |   |   |   |   | P |   |   |   |   |   |   |   |   |   |   |   |   |
| FS113.12            |   |   |   |   |   |   |   |   |   |   |   |   |   |   | P |   |   |   |   |   |   |   |   |   |   |   |   |
| FS113.11            |   |   |   |   |   |   |   |   |   |   |   |   |   |   | P |   |   |   |   |   |   |   |   |   |   |   |   |
| FS113.10            |   |   |   |   |   |   |   |   |   |   |   |   |   |   | P |   |   |   |   |   |   |   |   |   |   |   |   |
| FF127.8             |   |   |   |   |   |   |   |   |   |   |   |   |   |   | P |   |   |   |   |   |   |   |   |   |   |   |   |
| FF10.1              |   |   |   |   |   |   |   |   |   |   |   |   |   |   | P |   |   |   |   |   |   |   |   |   |   |   |   |
| SS103.3             |   |   |   |   |   |   |   |   |   |   |   |   |   |   | P |   |   |   |   |   |   |   |   |   |   |   |   |
| SS103.1             |   |   |   |   |   |   |   |   |   |   |   |   |   |   | P |   |   |   |   |   |   |   |   |   |   |   |   |
| FF4.1               |   |   |   |   |   |   |   |   |   |   |   |   |   |   | P |   |   |   |   |   |   |   |   |   |   |   |   |
| FF24.1              |   |   |   |   |   |   |   |   |   |   |   |   |   |   | P |   |   |   |   |   |   |   |   |   |   |   |   |
| FF127.3             |   |   |   |   |   |   |   |   |   |   |   |   |   |   | P |   |   |   |   |   |   |   |   |   |   |   |   |
| SS104.1             |   |   |   |   |   |   |   |   |   |   |   |   |   |   | P |   |   |   |   |   |   |   |   |   |   |   |   |
| SS103.8             |   |   |   |   |   |   |   |   |   |   |   |   |   |   | P |   |   |   |   |   |   |   |   |   |   |   |   |
| SS103.7             |   |   |   |   |   |   |   |   |   |   |   |   |   |   | P |   |   |   |   |   |   |   |   |   |   |   |   |
| SS131.1             |   |   |   |   |   |   |   |   |   |   |   |   |   |   | P |   |   |   |   |   |   |   |   |   |   |   |   |
| SS103.5             |   |   |   |   |   |   |   |   |   |   |   |   |   |   | P |   |   |   |   |   |   |   |   |   |   |   |   |
| SS103.6             |   |   |   |   |   |   |   |   |   |   |   |   |   |   | S |   |   |   |   |   |   |   |   |   |   |   |   | P | Q |

FS113.1, FS113.13, FS113.12, FS113.11, FS113.10, FF127.8, FF10.1, SS103.3, SS103.1, FF4.1, FF24.1, FF127.3, SS104.1, SS103.8, SS103.7, SS131.1, SS103.5, SS103.6.
Supplemental Figure S1: Segregating sites ($S$) on nucleotide position. (A) $S$ was estimated in a sliding window of 100 nucleotides with a 50 nucleotides step-size for all nucleotides of a 3000 base pairs linked $\beta$ and $\alpha$ fragment among all clones. (B) $S$ was estimated for the same fragment between pairs of original and repeat clone in the study of PCR errors. The diagram shows mean $S$ over the seven pairs. (C) A schematic of predicted exons, introns and intergenic regions of the 3000 base pairs fragment (Halldórsdóttir and Árnason, 2009).
Supplemental Figure S2: A maximum likelihood tree of triplet strictly phylogenetically informative sites among 57 clones of a 3000 base pairs βα gene set in Atlantic cod. Phylogenetically informative sites are defined by independence of being found in clones from three separate individuals, thus derived from three separate PCR and cloning events (Table S2).
Supplemental Figure S3: Maximum likelihood tree of doubleton strictly phylogenetically informative sites among 57 clones of a 3000 base pairs \(\beta/\alpha\) gene set in Atlantic cod.
Supplemental Figure S4: Most parsimonious tree of doubleton strictly phylogenetically informative sites among 57 clones of a 3000 base pairs β/α gene set in Atlantic cod.
Supplemental Figure S5: Neighbour joining tree of genetic distance of doubleton strictly phylogenetically informative sites among 57 clones of a 3000 base pairs β/α gene set in Atlantic cod.
2. Variable gene prediction

This section contains detailed analyses of various GENSCAN prediction among the clones. Most of the cloned contigs analysed showed the same pattern of gene structure in GENSCAN prediction. The structure of the region is extensively described in Halldorsdottir and Arnason (Halldórsdóttir and Árnason, 2009). In this study we analysed clones from several individuals and some variation from this main pattern was found. The FF4b.1 clone had a single base pair deletion in exon 2 in the α gene which caused a shift in reading frame in amino acid number 79. The translation continued to amino acid number 93 where an out-of-frame stop codon was found. GENSCAN predicted an α gene with two exons instead of three, an initial one and terminal one. The initial exon was similar to exons of most clones, however, the terminal exon was 24 base pairs shorter than the internal exon in α genes of other clones. Furthermore, exon 2 in the β gene of this clone was 10 amino acids shorter than exon 2 found in other clones. This clone was excluded from comparison of amino acid differences in the α gene (Supplemental Table S4) because of these differences.

Clone FF127.1 also had the same kind of short exon 2 in the β gene. However GENSCAN predicted a normal α gene. At the end of intron 1 in the β gene of these two clones a substitution was observed such that an –AG becomes –GG. An AG is a normal acceptor site for splicing. Another AG is found 30 base pairs further downstream after 10 amino acids of exon 2. Apparently the GENSCAN software takes this AG as terminating intron 2 and thus predicts a 10 amino acid indel in the protein.

The gene in the intergenic region predicted in both FF127.1 and FF4b.1 had the usual 38 amino acids, most commonly predicted for this gene. Both clones, however, FF127.1 and FF4b.1, were excluded from finding amino acid differences in the β gene (Supplemental Table S3) because of the large deletion predicted. There are the clones which define cluster 1 (Figure 2).

The β gene of the FS54.1 clone was not different from that predicted in the other clones. The single exon gene in the intergenic region, however, had a 27 base pairs longer exon than the common pattern. The exon was composed of atg tandem repeats which was translated to D (asparctic acid). The difference in exon size was due to nine more repeats in this clone compared...
to the other clones. Furthermore GENSCAN predicted an \( \alpha \) gene with only two exons. It had no predicted internal exon. Instead, GENSCAN took what was commonly defined as internal exon 2, joined with a translation of intron 2 and the commonly defined terminal exon 3 as one long terminal exon. Thus, GENSCAN predicted a 174 amino acid \( \alpha \) globin for clone FS54.1. This clone defines cluster 3a in Figure 2.

Clone FF20.1 was more similar to the sequence deposited in GenBank by Tipping and Birley (accession number 2154747, 2154750, 2154752) than the rest of the clones. They both had an insert in intron 1, an –AATG– at base pairs 36–39. Similarly base pairs number 3 and 10 in intron 1 also were similar and thus distinguished FF20.1 from our other clones. No other clones were similar with Tipping and Birley intron 1 for the first 39 base pairs. It is also seen in Supplemental Table S3 that clone FF20.1 and the Tipping/Birley sequence were identical in exon 1. The variation observed in FF20.1 thus also conforms to our strict criterion because the Tipping and Birley sequence is independent of ours.

In the GENSCAN prediction for clones FS43.2, FF127.9, SS103.5, 6, 7, 12, 13 (clusters 5 and 8 in Table 2 and Figure 2) the single exon in the intergenic region was joined to the \( \alpha \) gene, predicting a hybrid protein of 173 amino acids. The promoter for the gene in the intergenic region was predicted as the promoter of this gene. The promoter sequence for the \( \alpha \) gene was not detected by GENSCAN although the sequence was found in the clones. The asparctic acid coding sequence was predicted as the initial exon. Following that the normal first and second exons were added as internal exons followed by the terminal exon thus predicting a four exon \( \alpha \) gene. There were fewer than usual tandem repeats in these clones but there were also amino acid differences at the beginning and at the end of the sequence between these clones and the most common pattern (Supplemental Table S5). Further studies regarding this type of prediction was done and is discussed in the main text (Table 4).

In clones FF4b.2, FF4.1, 2, SS103.1, 4 in cluster 5 in Table 2 and Figure 2 GENSCAN did not predict a gene in the intergenic region. The promoter sequence, however, had a one base pair substitution relative to the common promoter sequence. The contigs of sequence data from these clones were of same length as the rest and the atg microsatellite were found as well. There were
Supplemental Table S5: Two different coding sequences for the gene in the intergenic region.

<table>
<thead>
<tr>
<th>Sequence</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>MNDDCNEGDDDDDDDDDDDDDNSTCSLESNYLNVN</td>
<td>The predicted intergenic exon in clones in which this exon forms the initial exon of the α gene.</td>
</tr>
<tr>
<td>MNDDCNDGDDDDDDDDDDDDDDDDDDDDDDDDDDNCSL</td>
<td>The most commonly predicted exon in intergenic region among clones.</td>
</tr>
</tbody>
</table>

no obvious sequence changes in these clones which would account for this prediction.

3. Genotypes and tandem repeats

Microsatellite tandem repeats were found in the intergenic region and also 5′ to β gene (Halldórsson and Árnason, 2009). The atg repeats showed some variation in numbers among the genotypes (Table S6). A hierarchical analysis of variance showed that differences among individuals were highly significant (data not shown). However, the difference in copy number among genotypes although suggestive were not significant, perhaps because of the high variation among individuals (Table S7).
Supplemental Table S6: Mean numbers of atg repeats among HbI genotypes.

<table>
<thead>
<tr>
<th>Genotype</th>
<th>FF</th>
<th>FS</th>
<th>SS</th>
</tr>
</thead>
<tbody>
<tr>
<td>atg</td>
<td>32.5</td>
<td>34.6</td>
<td>28.5</td>
</tr>
</tbody>
</table>

Supplemental Table S7: Hierarchical analysis of variance of the count of atg microsatellite tandem repeats.

<table>
<thead>
<tr>
<th>atg tandem repeat</th>
<th>Df</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>Among genotypes</td>
<td>2</td>
<td>336.10</td>
<td>168.05</td>
<td>3.45</td>
<td>0.055</td>
</tr>
<tr>
<td>Among individuals within genotypes</td>
<td>17</td>
<td>827.55</td>
<td>48.68</td>
<td>11.09</td>
<td>8.8e-10</td>
</tr>
<tr>
<td>Among clones within individuals</td>
<td>37</td>
<td>162.49</td>
<td>4.39</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Supplemental Figure S6: Phylogeny of β globins in Atlantic cod and model teleosts. The tree is based on amino acid sequences aligned with CLUSTAL W. The tree was build with Maximum likelihood method in the program PROTML from the package PHYLIP. Myoglobin of Gadus morhua was used as outgroup. Sequences were taken from GenBank; accession numbers: Gadus morhua-Verde1 (P84610), Gadus morhua-Verde2 (P84611), Gadus morhua-myoglobin (ABL7386), Gadus morhua-Birley (CAA66903), Takifugu rubripes-adult (AAO61493), Gadus morhua-cDNA (unpublished KH), Oryzias latipes-adult-1 (BAC06483), Oryzias latipes-adult-2 (BAC0291, BAC20293, BAC20296, BAC20297), Danio rerio-embryonic-1:10 (EF644855-EF644911), Gasterosteus aculeatus-adult (AAO27764.1), Boreogadus saida-1b (Q1AGS7), Boreogadus saida-2b (Q1AGS6), Gadus morhua-cluster1:10 (EF644855-EF644911).
Appendix II

Supplemental Figure S7: Phylogeny of α globins in Atlantic cod and model teleosts. The tree is based on amino acid sequences aligned with CLUSTAL W. The tree was build with Maximum likelihood method in the program PROTML from the package PHYLIP. Myoglobin of Gadus morhua was used as outgroup. Sequences were taken from GenBank; accession numbers: Gadus morhua-Verde1 (P84609), Gadus morhua-Verde2 (O42425), Gadus morhua-myoglobin (ABL7386), Gadus morhua-Birley (CAA66866), Takifugu rubripes-adult-4 (AAO61494), Takifugu rubripes-adult-3 (AAO61492), Gadus morhua-cDNA (unpublished KH), Oryzias latipes-adult-1 (BAC06482), Oryzias latipes-embryonic-0:4 (BAC20290, BAC20292, BAC20294, BAC20295, BAC20298), Danio rerio-embryonic-1 (AAH71550), Danio rerio-embryonic-3 (NP_898889), Danio rerio-adult-1 (NP_571332), Oncorhynchus mykiss-adult-1(BAA13533.1), Oncorhynchus mykiss-adult-4 (BAA13534.1), Oncorhynchus mykiss-embryonic-1 (NP_001117658.1), Oncorhynchus mykiss-embryonic-2(NP_001118054), Gadus morhua-Andersen-1 (FJ392681), Gadus morhua-Andersen-2 (FJ392682), Arctogadus glacialis-1 (Q1AGS5), Arctogadus glacialis-2 (Q1AGS4), Boreogadus saida-1 (Q1AGS9), Boreogadus saida-2 (Q1AGS8), Gadus morthua-cluster1:10 (EF644855-EF644911).
Supplemental Figure S8: Phylogeny of $\beta$ globins in Atlantic cod and model teleosts. The tree is based on amino acid sequences aligned with CLUSTAL W. The tree was build with Bayesian method using the program MrBayes according to best model predicted by ProtTest. Sequences were taken from GenBank; accession numbers: Gadus morhua-Andersen-3 (FJ392685), Boreogadus saida-2b (Q1AG56), Arctogadus glacialis-2 (P84604), Oryzias latipes-adult-1 (BAC06483), Oncorhynchus mykiss-embryonic-2 (NP_001117660.1), Danio rerio-embryonic-3 (AAP93667), Danio rerio-embryonic-2 (AAP93668), Oryzias latipes-embryonic-3 (BAC20296), Oryzias latipes-embryonic-1 (BAC20291), Gasterosteus aculeatus-adult (AAO27764.1), Gadus morhua-Andersen-1 (FJ392683), Boreogadus saida-1b (Q1AG57), Oryzias latipes-embryonic-2 (BAC20293), Danio rerio-adult-2 (NP_001005403), Oncorhynchus mykiss-adult (NP_001118017.1), Gadus morhua-cluster6 and 7 (consensus sequences from this study).
Supplemental Figure S9. Phylogeny of α globins in Atlantic cod and model teleosts. The tree is based on amino acid sequences aligned with CLUSTAL W. The tree was build with Bayesian method using the program MrBayes according to best model predicted by ProtTest. Sequences were taken from GenBank; accession numbers: Danio rerio-embryonic-1 (AAH71550), Oryzias latipes-embryonic-4 (BAC20298), Oryzias latipes-embryonic-2 (BAC20294), Oryzias latipes-adult-1 (BAC06482), Takifugu rubripes-adult-4 (AAO61494), Oncorhynchus mykiss-adult-4 (BAA13534.1), Danio rerio-adult-1 (NP_571332), Arctogadus glacialis-2 (Q1AGS4), Gadus morhua-Andersen-1 (FJ392681), Oncorhynchus mykiss-embryonic-2 (NP_001118054), Oncorhynchus mykiss-embryonic-1 (NP_001117658.1), Boreogadus saida-1 (Q1AGS9), Gadus morhua-Verde2 (O42425), Gadus morhua-Andersen-2 (FJ392682), Gadus morhua-cluster3a and 9 (consensus sequences from this study).
References


Appendix III

Supplementary data for Paper III

Katrín Halldórsdóttir and Einar Árnason. Trans-species polymorphism at antimicrobial innate immunity cathelicidin genes of Atlantic cod and related species. *PeerJ* **3**, e786
Supplementary data for the manuscript:

Trans-species Polymorphism at Antimicrobial Innate Immunity Cathelicidin Genes of Atlantic cod and Related Species

Katrín Halldórsdóttir* and Einar Árnason

Institute of Life and Environmental Sciences, University of Iceland,
Sturlugata 7, Reykjavík, Iceland

*Corresponding author. email: katrinhalldorsdottir@gmail.com
MTTQMRLLCFAAVTLLAEAQVMIPGDPLFIFPLKNFCRPLLDQLQRYDV
EATAGVYPEGVGDLS...73

↓↓↓↑↓
M
M
M

FG312333.1˙cath1
105746˙3.Gmo.Gre
103659˙2.Bsa.Gre
104931˙3.Gmo.Gre
152027˙3.Gch.Pac
152074˙1.Gma.Pac
Figure S1. Alignment of amino acid sequences of exons 1, 2, and 3 combined, the conserved part of cathelicidin, from clones of various individuals of Atlantic cod and four closely related taxa. Highly polymorphic sites are boxed. The four conserved cysteine residues characterizing cathelicidin are shaded. Up arrows represent positively selected sites and down arrows negatively selected sites from Tables 1 and 2. Bsa.Gre (B. roegneri), Gch.Pac (G. chalcogrammus), Gma.Pac (G. macrocephalus), Gog.Gre (G. ogac) and Gno (G. morhua) from various locations: Iceland (Gmo.Ice), Greenland (Gmo.Gre), Barents Sea (Gmo.Bar), Celtic Sea (Gmo.Cell), Baltic Sea (Gmo.Bal), Norway (Gmo.Nor), Faroe Islands (Gmo.Far), Canada (Gmo.Can).
Appendix III

Figure S2. Alignment of nucleotide sequences of cathelicidin among clones from various individuals of Atlantic cod and four closely related taxa. Bsa.Gre (Boreogadus saida), Gch.Pac (Gadus chalcogrammus), Gma.Pac (Gadus macrocephalus), Gog.Gre (Gadus ogac) and Gmo (Gadus morhua) from various locations; Iceland (Gmo.Ice), Greenland (Gmo.Gre), Barents Sea (Gmo.Bar), Celtic Sea (Gmo.Cel), Baltic Sea (Gmo.Bal), Norway (Gmo.Nor), Faroe Islands (Gmo.Far), Canada (Gmo.Can).
Figure S3. Maximum likelihood phylogenetic tree of Cathelicidin amino acid sequences of the conserved part. Phylogenetic tree built on amino acid sequences in exons 1, 2, and 3 combined, the conserved part of cathelicidin, of clones from various individuals of Atlantic cod and four closely related taxa. Bsa.Gre (*Boreogadus saida*), Gch.Pac (*Gadus chalcogrammus*), Gma.Pac (*Gadus macrocephalus*), Gog.Gre (*Gadus ogac*) and Gmo (*Gadus morhua*) from various locations; Iceland (Gmo.Ice), Greenland (Gmo.Gre), Barents Sea (Gmo.Bar), Celtic Sea (Gmo.Cel), Baltic Sea (Gmo.Bal), Norway (Gmo.Nor), Faroe Islands (Gmo.Far), Canada (Gmo.Can).
Figure S4. Maximum likelihood phylogenetic tree of nucleotide sequences. Phylogenetic tree of nucleotide sequences of the cathelicidin gene from 43 representative clones of various individuals of Atlantic cod and four sister taxa. Bsa.Gre (*Boreogadus saida*), Gch.Pac (*Gadus chalcogrammus*), Gma.Pac (*Gadus macrocephalus*), Gog.Gre (*Gadus ogac*) and Gmo (*Gadus morhua*) from various locations; Iceland (Gmo.Ice), Greenland (Gmo.Gre), Barents Sea (Gmo.Bar), Celtic Sea (Gmo.Cel), Baltic Sea (Gmo.Bal), Norway (Gmo.Nor), Faroe Islands (Gmo.Far), Canada (Gmo.Can).
Figure S5. Sliding window analysis of nucleotide diversity $\pi$ and the scaled mutation rate $\theta$ for Cath1. Window length was 100 bp with a 25 bp step size.
Figure S6. Sliding window Tajima’s $D$ for Cath1. Window length was 100 bp with a 25 bp step size.
Figure S7. Sliding window analysis of nucleotide diversity $\pi$ and the scaled mutation rate $\theta$ for Cath3. Window length was 100 bp with a 25 bp step size.
Figure S8. Sliding window Tajima’s $D$ for Cath3. Window length was 100 bp with a 25 bp step size.
Supplementary data for Paper IV

Supplement to:
Nucleotide Variation and Balancing Natural Selection at the \textit{Ckma} gene in Atlantic cod: Analysis with multiple merger coalescent models

Einar Árnason and Katrín Halldórsdóttir

Institute of Biology, University of Iceland,
Sturlugata 7, 101 Reykjavík, Iceland
Figure S1. Map of sampling localities for Atlantic cod in the North Atlantic. Sampling localities in the waters of Canada (Nova Scotia and Newfoundland) Can, Greenland Gre, Iceland Ice, Norway Nor, Faroe Islands Far, and from the Barents Sea Bar, White Sea Whi, North Sea Nor, Baltic Sea Bal, Celtic Sea Cel, and Irish Sea Iri.
**Figure S2.** Structure of the *Ckma* gene and sequenced parts. Boxes represent exons, start (red), internal (magenta) and terminal (blue). Green boxes represent sequenced fragments trimmed to Phred score of at least 30. Up and down arrows mark TATA box and poly A signal starts respectively.
Figure S3. Maximum likelihood tree of creatin kinase proteins of paralogous genes in the Atlantic cod genome, and \textit{Ckma} orthologs from this study the two alleles \textit{A} (\textit{Gmo.Nse A}) and \textit{B} (\textit{Gmo.Ice B}) in Atlantic cod and representatives from sister taxa. Predicted protein isoforms from mitochondria (CKMT), brain (CKB) and muscle (CKM). Sister taxa are \textit{Boreogadus saida} (Bsa), \textit{Gadus macrocephalus} (Gma), \textit{Gadus ogac} (Gog), and \textit{Gadus chalcogrammus} (Gch).
Figure S4. Maximum likelihood tree of variation of Ckma among 122 Atlantic cod individuals. Color codes for localities same as in Figure S1.
Figure S5. Maximum likelihood tree of nucleotide variation of Myg among 45 Atlantic cod and two Gadus macrocephalus individuals. Color codes for species and localities same as in Figure S1.
Figure S6. Maximum likelihood tree of nucleotide variation of HbA2 gene among 113 Atlantic cod and 14 individuals of sister taxa. Color codes for species and localities same as in Figure S1.
Figure S7. Linkage disequilibrium heatmap of \( D' \) and \( r^2 \) for 2500 bp fragment of the \textit{Ckma} gene of Atlantic cod.
Figure S8. Unfolded site frequency spectrum of Atlantic cod *HbA2* gene. *Gadus macrocephalus* is outgroup. Theoretical expectation under Kingman coalescent (red dots), Beta(2 − α, α) coalescent (magenta squares), and point-mass coalescent (blue stars).
Figure S9. Unfolded site frequency spectrum of Atlantic cod Myg gene. *Gadus macrocephalus* is outgroup. Theroretical expectation under Kingman coalescent (red dots), Beta(2 − α, α) coalescent (magenta squares), and point-mass coalescent (blue stars).
Figure S10. Unfolded site frequency spectrum of Atlantic cod CkmA alleles (left) and CkmB alleles (right). *Gadus macrocephalus* is outgroup. Number of individuals \( n = 43 \) and \( n = 79 \) respectively. Theroretical expectation under Kingman coalescent (red dots), Beta\((2 - \alpha, \alpha)\) coalescent (magenta squares), and point-mass coalescent (blue stars).
Figure S11. The $\ell^2$ distance for the unfolded site frequency spectrum of the nuclear genes $Myg$, $Hb2A$, $Ckma$, and the $Ckma-A$ and $Ckma-B$ alleles of $Ckma$ on the $\alpha$ parameter of the Beta$(2 - \alpha, \alpha)$ coalescent (left panel) and the $\psi$ parameter of the point-mass coalescent (right panel).
Figure S12. The $\ell_2$ distance for the unfolded site frequency spectrum of mtDNA from various localities of the North Atlantic on the $\alpha$ parameter of the Beta($2-\alpha$, $\alpha$) coalescent (left panel) and the $\psi$ parameter of the point-mass coalescent (right panel).

Table S1. Primer sequences for amplification and sequencing fragments of Ckma gene from Atlantic cod and sister taxa.

<table>
<thead>
<tr>
<th>Primer name</th>
<th>Use</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>creL8945</td>
<td>Amplification</td>
<td>5′-GTT TAG GAA TCT ACG CCC ATC CAG AGA CA-3′</td>
</tr>
<tr>
<td>creR12945</td>
<td>Amplification</td>
<td>5′-TGG CTA TCA TGC ATT CCC AAT GTT C-3′</td>
</tr>
<tr>
<td>cresseqR12388</td>
<td>Sequencing</td>
<td>5′-CAT GAC CGT TGG CTG ATG CGT TG-3′</td>
</tr>
<tr>
<td>cresseqL10486</td>
<td>Sequencing</td>
<td>5′-TGG AAC ACT CGG ACCG AGA-3′</td>
</tr>
<tr>
<td>cresseqR10602</td>
<td>Sequencing</td>
<td>5′-ACA GAT TTC GTC GGC CGA GA-3′</td>
</tr>
</tbody>
</table>
Appendix IV

Table S2. Segregating sites of the Ckma gene among 122 Atlantic cod individuals and 10 individuals of sister taxa (see separate file).

Table S3. Gross $D_{xy}$ and net $D_a$ nucleotide divergence per site between Gadus morhua Gmo and Gadus macrocephalus Gma and Gadus chalcogrammus Gch and between A and B alleles of Atlantic cod.

<table>
<thead>
<tr>
<th>Gene</th>
<th>Comparison</th>
<th>$D_{xy}$</th>
<th>$s_{D_{xy}}$</th>
<th>$D_a$</th>
<th>$s_{D_a}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>HbA2</td>
<td>Gmo vs Gma</td>
<td>0.013</td>
<td>0.004</td>
<td>0.012</td>
<td>0.004</td>
</tr>
<tr>
<td>HbA2</td>
<td>Gmo vs Tch</td>
<td>0.017</td>
<td>0.014</td>
<td>0.017</td>
<td>0.014</td>
</tr>
<tr>
<td>Myg</td>
<td>Gmo vs Gma</td>
<td>0.027</td>
<td>0.007</td>
<td>0.025</td>
<td>0.007</td>
</tr>
<tr>
<td>Ckma</td>
<td>Gmo vs Gma</td>
<td>0.014</td>
<td>0.001</td>
<td>0.011</td>
<td>0.001</td>
</tr>
<tr>
<td>Ckma</td>
<td>Gmo vs Tch</td>
<td>0.015</td>
<td>0.003</td>
<td>0.013</td>
<td>0.003</td>
</tr>
<tr>
<td>Ckma</td>
<td>A vs B</td>
<td>0.008</td>
<td>0.0005</td>
<td>0.006</td>
<td>0.0005</td>
</tr>
</tbody>
</table>

Divergence and standard deviation found using Jukes and Cantor correction.

Table S4. Non-synonymous changes within and between species.

<table>
<thead>
<tr>
<th>Individual</th>
<th>amino acids</th>
<th>position</th>
</tr>
</thead>
<tbody>
<tr>
<td>100896.Gmo.Gre</td>
<td>M \leftrightarrow T</td>
<td>37</td>
</tr>
<tr>
<td>117937.Gmo.Ice</td>
<td>I \leftrightarrow T</td>
<td>469</td>
</tr>
<tr>
<td>152915.Gmo.Nor</td>
<td>S \leftrightarrow G</td>
<td>1158</td>
</tr>
<tr>
<td>118708.Gmo.Ice</td>
<td>R \leftrightarrow G</td>
<td>1182</td>
</tr>
<tr>
<td>152978.Gmo.Nor</td>
<td>E \leftrightarrow G</td>
<td>2068</td>
</tr>
<tr>
<td>152066.Gma.Pac</td>
<td>I \leftrightarrow T</td>
<td>469</td>
</tr>
<tr>
<td>152047.Gma.Pac</td>
<td>V \leftrightarrow I</td>
<td>2094</td>
</tr>
<tr>
<td>104474.Bsa.Gre</td>
<td>R \leftrightarrow G</td>
<td>1182</td>
</tr>
<tr>
<td>104725.Bsa.Gre</td>
<td>Q \leftrightarrow G</td>
<td>1305, 1306, 1307</td>
</tr>
<tr>
<td>104474.Bsa.Gre</td>
<td>Q \leftrightarrow G</td>
<td>1305, 1306, 1307</td>
</tr>
</tbody>
</table>

Individuals with species and locality codes, first aa represents majority and the second the change, position refers to position in concatenated sequence in Table S2.
### Table S5. Maximum likelihood analysis of a Kingman-coalescent HKA test of neutrality and selection at three genes in Atlantic cod.

<table>
<thead>
<tr>
<th>Description</th>
<th>lnL</th>
<th>T</th>
<th>Test</th>
<th>df</th>
<th>θ</th>
<th>k</th>
<th>θ</th>
<th>k</th>
<th>θ</th>
<th>k</th>
</tr>
</thead>
<tbody>
<tr>
<td>Neutral, all $k = 1$</td>
<td>-18.66</td>
<td>2.46</td>
<td></td>
<td></td>
<td>0.0035</td>
<td>1</td>
<td>0.0068</td>
<td>1</td>
<td>0.0056</td>
<td>1</td>
</tr>
<tr>
<td>Selection at Ckma</td>
<td>-17.47</td>
<td>3.86</td>
<td>2.38</td>
<td>1</td>
<td>0.0029</td>
<td>1</td>
<td>0.0054</td>
<td>1</td>
<td>0.0032</td>
<td>2.12</td>
</tr>
</tbody>
</table>

Test is twice the lnL difference of the two models, neutrality and selection at Ckma. Three loci are under test: Hemoglobin α2 (Hbα2), Myoglobin (Myg), and Creatine Kinase Muscle (Ckma). θ is the scaled effective population size and the parameter $k$ measures changes in diversity due to selection. Based on method of Wright and Charlesworth (2004).

### Table S6. Frequency of A and B alleles in different localities.

<table>
<thead>
<tr>
<th></th>
<th>Can</th>
<th>Gre</th>
<th>Ice</th>
<th>Nor</th>
<th>Bar</th>
<th>Whi</th>
<th>Far</th>
<th>Nse</th>
<th>Bal</th>
<th>Cel</th>
<th>Iri</th>
<th>Sum</th>
</tr>
</thead>
<tbody>
<tr>
<td>A allele</td>
<td>0</td>
<td>0</td>
<td>7</td>
<td>0</td>
<td>1</td>
<td>7</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>7</td>
<td>6</td>
</tr>
<tr>
<td>B allele</td>
<td>9</td>
<td>5</td>
<td>45</td>
<td>13</td>
<td>5</td>
<td>1</td>
<td>7</td>
<td>8</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>6</td>
</tr>
<tr>
<td>Sum</td>
<td>9</td>
<td>5</td>
<td>52</td>
<td>13</td>
<td>5</td>
<td>2</td>
<td>7</td>
<td>9</td>
<td>7</td>
<td>7</td>
<td>6</td>
<td>122</td>
</tr>
</tbody>
</table>

### Table S7. Pairwise $F_{ST}$ values (lower triangular) of population differentiation among localities.

<table>
<thead>
<tr>
<th></th>
<th>Can</th>
<th>Gre</th>
<th>Ice</th>
<th>Nor</th>
<th>Bar</th>
<th>Far</th>
<th>Nse</th>
<th>Bal</th>
<th>Cel</th>
<th>ln</th>
</tr>
</thead>
<tbody>
<tr>
<td>Can</td>
<td>0.78</td>
<td>0.06</td>
<td>0.11</td>
<td>0.14</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Gre</td>
<td>0.04</td>
<td>0.22</td>
<td>0.55</td>
<td>0.12</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Ice</td>
<td>0.08</td>
<td>0.01</td>
<td>0.21</td>
<td>0.53</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Nor</td>
<td>0.03</td>
<td>0.00</td>
<td>0.02</td>
<td>0.13</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Bar</td>
<td>0.01</td>
<td>-0.08</td>
<td>-0.02</td>
<td>-0.06</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Far</td>
<td>0.88</td>
<td>0.81</td>
<td>0.77</td>
<td>0.84</td>
<td>0.84</td>
<td>0.41</td>
<td>0.88</td>
<td>0.21</td>
<td>0.22</td>
<td></td>
</tr>
<tr>
<td>Nse</td>
<td>0.78</td>
<td>0.71</td>
<td>0.65</td>
<td>0.74</td>
<td>0.74</td>
<td>0.04</td>
<td>0.93</td>
<td>0.48</td>
<td>0.16</td>
<td></td>
</tr>
<tr>
<td>Bal</td>
<td>0.82</td>
<td>0.76</td>
<td>0.71</td>
<td>0.78</td>
<td>0.78</td>
<td>-0.01</td>
<td>-0.03</td>
<td>0.36</td>
<td>0.82</td>
<td></td>
</tr>
<tr>
<td>Cel</td>
<td>0.91</td>
<td>0.84</td>
<td>0.80</td>
<td>0.87</td>
<td>0.87</td>
<td>-0.05</td>
<td>0.07</td>
<td>-0.01</td>
<td>0.09</td>
<td></td>
</tr>
<tr>
<td>Iri</td>
<td>0.88</td>
<td>0.82</td>
<td>0.78</td>
<td>0.84</td>
<td>0.84</td>
<td>0.02</td>
<td>0.04</td>
<td>0.01</td>
<td>0.04</td>
<td></td>
</tr>
</tbody>
</table>

Probabilities in black on upper triangular, boldface are significant P values. North (blue) and South (red) defined ad hoc by results.
Table S8. Pairwise $F_{ST}$ of $Ckma$ gene between North and South (lower left corner) and between $A$ and $B$ alleles (upper right corner).

<table>
<thead>
<tr>
<th></th>
<th>North</th>
<th>$B$ allele</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A$ allele</td>
<td>0.738</td>
<td>0.804</td>
</tr>
<tr>
<td>South</td>
<td>0.763</td>
<td>0.828</td>
</tr>
</tbody>
</table>

North and South populations defined according to differentiation at $Ckma$ locus.

Table S9. Pairwise $F_{ST}$ of neutral genes between North and South.

<table>
<thead>
<tr>
<th></th>
<th>North</th>
<th>South</th>
</tr>
</thead>
<tbody>
<tr>
<td>North</td>
<td>$-0.029$</td>
<td></td>
</tr>
<tr>
<td>South</td>
<td>0.004</td>
<td></td>
</tr>
</tbody>
</table>

$HbA2$ locus on lower left corner, $Myg$ locus on upper right corner. North and South populations defined according to differentiation at $Ckma$ locus.

Table S10. Likelihood ratio test statistics $G$ for observed site frequency spectra and expectation according to different coalescent models.

<table>
<thead>
<tr>
<th>Model</th>
<th>$G$</th>
<th>Comparison</th>
<th>$2\Delta G$</th>
<th>$df$</th>
</tr>
</thead>
<tbody>
<tr>
<td>I. Kingman</td>
<td>149.26</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>II. Beta($2 - \alpha, \alpha$)</td>
<td>116.21</td>
<td>I vs II</td>
<td>66.10</td>
<td>1</td>
</tr>
<tr>
<td>III. point-mass</td>
<td>114.42</td>
<td>I vs III</td>
<td>69.69</td>
<td>1</td>
</tr>
</tbody>
</table>

16/16
Appendix V

Supplementary data for Paper V

Whole genome sequencing uncovers cryptic and hybrid species among Atlantic and Pacific cod-fish. *In review*
Supplemental Information

Linkage Group LG01 admixture. 203 K sites.

Linkage Group LG02 admixture. 159 K sites.
Appendix V
Linkage Group LG05 admixture. 155 K sites.

Linkage Group LG06 admixture. 186 K sites.
Linkage Group LG07 admixture. 177 K sites.

Linkage Group LG08 admixture. 157 K sites.
Appendix V

Linkage Group LG09 admixture. 157 K sites.

Linkage Group LG10 admixture. 193 K sites.
Linkage Group LG11 admixture. 183 K sites.

Linkage Group LG12 admixture. 170 K sites.
Linkage Group LG15 admixture. 159 K sites.

Linkage Group LG16 admixture. 213 K sites.
Linkage Group LG17 admixture. 59 K sites.

Linkage Group LG18 admixture. 94 K sites.
Appendix V

Linkage Group LG19 admixture. 125K K sites.

Linkage Group LG20 admixture. 112 K sites.
Appendix V

Linkage Group LG21 admixture. 112 K sites.

Linkage Group LG22 admixture. 93 K sites.
Supplemental Figure 1. Individual admixture analysis of cod-fish by linkage group LG01 to LG23. Model-based analysis with the number of ancestral population $k$ ranging from 2 to 7. Ordinate gives admixture proportion of each individual. Results ordered based on species Arctogadus glacialis Agl, Boreogadus saida Bsa, Gadus chalcogrammus Gch, Gadus macrocephalus Gma, Gadus ogac Gog, and for Gadus morhua localities from west to east and by ecotype: Cape cod Cco, Western Bank Web, Sable Bank Sab, Trinity Bay Tri, Southern Grand Banks Sgb, Greenland frontal BBGre, Greenland intermediate ABGre, Iceland frontal BBIce, Iceland intermediate ABIce, Iceland coastal AAIce, Barents Sea frontal BBBar, White Sea Whi, Norway Nor, Faroe Islands Far, North Sea Nse, Celtic Sea Cel, western Baltic W-Bal, and eastern Baltic E-Bal. The number of variable sites in thousands (K) is given for each linkage group.
Supplemental Figure 2. Minimum evolution tree of genetic distances among cod-fish taxa. *Boreogadus saida* Bsa, *Gadus chalcogrammus* Gch, *Gadus macrocephalus* Gma, *Gadus ogac* Gog and six *Gadus morhua* Gmo, from Sable Bank Sab, Trinity Bay Tri, Iceland Ice, and the North Sea Nse. Based on whole genome sequencing with 20 – 30 × coverage of entire genome. Individuals are labelled with a six digit identifier plus a three letter code for species or localities for Atlantic cod. The numbers on the nodes are bootstrap supports.
Supplemental Figure 3. Discriminant analysis of principal components (DAPC) of variation among groups and localities of the eastern Atlantic. Black lines represent a minimum spanning tree. Variation of the part of the genome mapping to linkage groups LG01 to LG23. DAPC based on deep sea frontal fish from the Barents Sea (BBBar), Iceland (BBIce), and Greenland (BBGre), intermediate fish from Iceland (ABIce) and Greenland (ABGre) and finally shallow water coastal fish from the White Sea (Whi), Iceland (AAIce), Faroe Islands (Far), Norway (Nor), North Sea (Nse), Celtic Sea (Cel), and the western and eastern Baltic (W-Bal and E-Bal).
Supplemental Figure 4. Sliding-window $F_{ST}$ on window midpoint. Genomic data that map to linkage groups LG01 to LG23 among localities of Atlantic cod. Pairwise comparisons of West (Cape Cod, Western Bank, Sable Bank, Trinity Bay, and Southern Grand Banks pooled), North (Greenland, Iceland, Barents Sea, White Sea, and Norway pooled), and South (Faroe Islands, North Sea, Celtic Sea, and Baltic Sea pooled). Window size was 10,000 and step size was 2,000.
Supplemental Figure 5. Discriminant analysis of principal components (DAPC) of variation among groups and localities of Atlantic cod. Black lines represent a minimum spanning tree. Variation of the part of the genome mapping to linkage groups LG01 to LG23. DAPC based on localities from the western Atlantic (Cape Cod (Cco), Western Bank (Web), Sable Bank (Sab), Trinity Bay (Tri), Southern Grand Banks (Sgb)) and the eastern Atlantic, deep sea frontal fish from the Barents Sea (BBBar), Iceland (BBIce), and Greenland (BBGre), intermediate fish from Iceland (ABIce) and Greenland (ABGre) and finally shallow water coastal fish from the White Sea (Whi), Iceland (AAIce), Faroe Islands (Far), Norway (Nor), North Sea (Nse), Celtic Sea (Cel), and the western and eastern Baltic (W-Bal and E-Bal).
Supplemental Figure 6. Posterior membership probabilities from the discriminant analysis of principal components (DAPC) in Figure ??1. Genomic data that map to LG02 to LG23 (excluding LG01, top panel), data that map to all linkage groups LG01 to LG23 (middle panel), and data that do not map to linkage groups (bottom panel).
Supplemental Figure 7. Discriminant analysis of principal components (DAPC) of phenotypic and habitat variation. Individuals were classified as pure coastal, pure frontal, and hybrid species by posterior membership probabilities of the DAPC analysis of linkage groups LG02 to LG23, the genomic parts that definitively do not map to linkage group LG01 (Supplemental Figure 6 top panel). Phenotypic data are length, age, sex, ungutted and gutted weight, weight of liver, yearclass, and habitat data on depth.