

Patterns of DNA Variation at the Pantophysin I (*Pan I*) Locus in Atlantic Cod (*Gadus morhua*): Inferences on Natural Selection.

Ubaldo Benitez Hernandez

A Dissertation submitted in partial satisfaction of the requirements for the
MS degree in Population Genetics

Advisor: Professor Einar Árnason



Department of Biology

Faculty of Science

UNIVERSITY OF ICELAND

October 2008

Patterns of DNA Variation at the Pantophysin I (*Pan I*) Locus in Atlantic Cod (*Gadus morhua*): Inferences on Natural Selection.

by

Ubaldo Benitez Hernandez

Licenciatura in Biochemical Engineering (Instituto Tecnológico y de Estudios Superiores de Monterrey, Campus Guaymas) 1996

A Dissertation submitted to Department of Biology of Faculty of Science
in partial satisfaction of the requirements for the
MS degree in Population Genetics

Committee in charge:

Professor Einar Árnason, Chair
Dr. Arnar Pálsson

External referee:

Dr. Pétur Henry Petersen



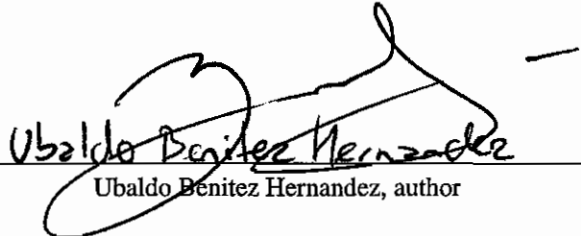
Department of Biology

Faculty of Science

UNIVERSITY OF ICELAND

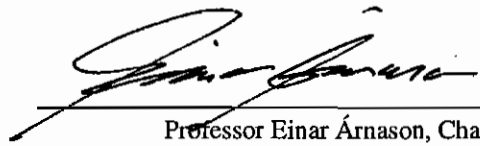
October 2008

I declare that this dissertation is based on my own observations, that it is written by myself,
and that it has not previously been submitted in part or in whole for a higher degree.

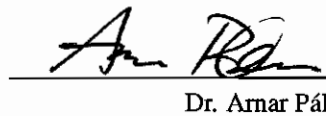


Ubaldo Benitez Hernandez, author
13. October 2008
Date


The MS dissertation of Ubaldo Benitez Hernandez is approved:



Professor Einar Arnason, Chair
13. October 2008
Date



Dr. Arnar Pálsson
10 October 2008
Date



Dr. Pétur Henry Petersen, External referee
13 Oct '08
Date

University of Iceland
October 2008

**Patterns of DNA Variation at the Pantophysin I (*Pan I*) Locus in Atlantic Cod
(*Gadus morhua*): Inferences on Natural Selection.**

Copyright © 2008
Ubaldo Benitez Hernandez

Abstract

Patterns of DNA Variation at the Pantophysin I (*Pan I*) Locus in Atlantic Cod (*Gadus morhua*): Inferences on Natural Selection.

by

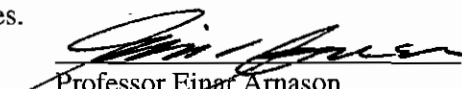
Ubaldo Benitez Hernandez

MS in Population Genetics

University of Iceland

Professor Einar Árnason, Chair

This study aims to enhance our understanding of the role of natural selection on the Pantophysin (*Pan I*) locus in Atlantic cod (*Gadus morhua*). The *Pan I* locus is distinctive in indicating high population differentiation in contrast with other, presumably neutral, loci. The evidence points to *Pan I* being under selection. Genotyping of the *Pan I* gene with two alleles *Pan I*^A and *Pan I*^B corresponding to absence or presence of a *DraI* restriction site was carried out for 8196 Atlantic cod individuals sampled around Iceland. Additionally, full sequences of the gene (1.85 kb) were obtained from 55 individuals. There is a clear correlation between *Pan I* allele frequencies and depth: There is an allele frequency gradient $\Delta p_A \approx -0.4\%/m$ which does not change between fall and spring although the genotypic composition changes between seasons. The clear correlation between *Pan I* allele frequency and depth, a proxy for other depth-related environmental variables that can act as selective agents, is indicative of selective effects. Several patterns of allelic frequency distribution are also evident between *Pan I* and age, year-class, length, and sex, indicating that selection is acting on *Pan I*. Variation in allele and genotypic frequencies among regions at Iceland, which has been taken as indications of geographic and historical population differentiation, is better explained by depth. Population structure of cod at Iceland estimated from the *Pan I* locus thus is due to the effects of selection at the locus. Comparative studies of DNA sequence variation of coding and non-coding parts of the *Pan I* gene revealed signals of ongoing selection at the gene. DNA sequence variation analysis did not show any relationship of gene-genealogy and either depth or locality. Thus there is no evidence from *Pan I* sequence variation of deep historical separation of cod at different depths or different localities.



Professor Einar Árnason
Chair, Committee in charge

Útdráttur

Mynstur DNA breytileika á Pantophysin I (*Pan I*) geninu hjá þorski (*Gadus morhua*):

Ályktanir um náttúrulegt val.

eftir

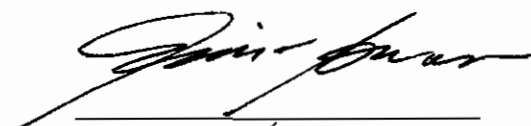
Ubaldo Benitez Hernandez

MS í Stofnerfðafræði

Háskóli Íslands

Prófessor Einar Árnason, formaður

Tilgangur þessa verkefnis er að efla skilning okkar á áhrifum náttúrulegs vals á Pantophysin (*Pan I*) genaset í þorski (*Gadus morhua*). Sérstaða *Pan I* setsins er að það gefur til kynna mikinn stofnaaðskilnað ólíkt öðrum setum sem talin eru hlutlaus. Vísbendingar eru um að *Pan I* sé undir vali. Arfgerðargreining á *Pan I* geninu var gerð á 8200 einstaklingum sem safnað var umhverfis Ísland. Skerðiset *DraI* skerðiensíms greimir á milli samsætu. Auk þess var genið (1,85 kb) raðgreint að fullu í 55 einstaklingum. Það eru skýr tengsl milli *Pan I* alleltíðni og dýpis: Stigull sést í samsætutíðni $\Delta p_A \approx -0.4\%/m$ sem er eins vor og haust þrátt fyrir breytingar á arfgerðasamsetningu milli árstíða. Hin skýru tengsl milli samsætutíðni *Pan I* og dýpis, sem hér er staðgengill fyrir aðrar dýpisháðar breytur sem geta stuðlað að vali, er til marks um áhrif vals. Tíðnidreifingu samsæta á *Pan I* sýnir tengsl við aldur, árgang, lengd og kyn, sem bendir til að val verki á *Pan I*. Breytileiki í tíðni *Pan I* samsæta og arfgerða meðal hafsvæða við Ísland, sem túlkaður hefur verið sem landfræðileg og söguleg uppskipting stofna, má frekar útskýra með dýpi. Stofngerð þorsks metin með *Pan I* geninu á því rætur að rekja til áhrifa vals á genið. Samanburðarrannsóknir á breytileika milli DNA raða *Pan I* gensins, jafnt í innröðum sem útröðum, sýndu merki vals innan gensins. Greining á raðgreindum röðum DNA sýndu engin tengsl genaættarsögu gensins við dýpi né við staðsetningu. Breytileiki í raðgreiningum gefur því engan vitnisburð um sögulega aðgreiningu þorsks á mismunandi dýpi eða mismunandi staðsetningum.



Professor Einar Árnason
Formaður umsjónarnefndar

To my parents,
Ubaldo and Maria Luisa,
the ones that truly sparked my interest in science.

Contents

List of Figures	viii
List of Tables	x
Introduction	1
Materials and Methods	7
Sampling	7
Genomic DNA extraction and Genotyping	10
Cloning and Sequencing	11
Data analysis	13
Results	16
Genotypic variation in relation to biological and environmental factors	16
<i>Pan</i> I in terms of depth (spring and fall surveys)	16
Depth and geographic region (spring and fall surveys)	20
Hierarchical <i>F</i> statistics (spring and fall surveys)	23
Associations of <i>Pan</i> I genotype with other biological variables (spring surveys)	23
DNA sequence variation	30
Quality of sequences	30
Nucleotide polymorphism	32
Tests of neutrality of DNA sequence variation	38
Gene genealogies, geography and depth	40
Genetic differentiation, considering geographic locality and depth	40
DNA divergence between populations	41
Discussion	44
Genotypic variation in relation to biological and environmental factors	44
DNA sequence variation	46
Conclusion	51
Appendix I. Depth and geographic region	59
Appendix II. <i>Pan</i> I genotype and length at age (fall surveys)	62
Appendix III. Quality of DNA sequences	65

Appendix IV. Gene genealogies, geography and depth	74
Appendix V. Various DNA sequence data analysis excluding singletons	77

List of Figures

1	Spring survey stations by MCdiv, years 2005, 2006, and 2007	8
2	Fall survey stations by MCdiv, years 2004, 2005, and 2006	9
3	An example gel image for scoring <i>Pan</i> I genotypes	11
4	Allele frequency vs depth, spring surveys 2005, 2006, and 2007 combined. . .	17
5	Sampling sites, latitude and longitude conditioned on depth; spring surveys 2005, 2006, and 2007	21
6	Sampling sites, latitude and longitude conditioned on depth; fall surveys 2004, 2005, and 2006	22
7	Interaction plot of mean length on genotype and age, for spring surveys 2005, 2006, and 2007.	28
8	Sequence quality of <i>Pan</i> I ^A alleles	31
9	Sliding window analysis (including singletons) of nucleotide diversity (π) throughout the <i>Pan</i> I gene region in <i>Pan</i> I ^A and <i>Pan</i> I ^B alleles	35
10	Sliding window analysis (including singletons) of nucleotide diversity (π) throughout the <i>Pan</i> I gene region in <i>Pan</i> I ^A alleles	36
11	Sliding window analysis (including singletons) of nucleotide diversity (π) throughout the <i>Pan</i> I gene region in <i>Pan</i> I ^B alleles	37

Figures in Appendices

1	Sampling sites by depth, latitude and longitude (3D view), for spring surveys 2005, 2006, and 2007	60
2	Sampling sites by depth, latitude and longitude (3D view), for fall surveys 2004, 2005, and 2006	61
3	Interaction plot of mean length on genotype and age, for fall surveys 2004, 2005, and 2006	63
4	Sequence quality of <i>Pan</i> I ^A alleles	66
5	Sequence quality of <i>Pan</i> I ^A alleles	67
6	Sequence quality of <i>Pan</i> I ^A alleles	68
7	Sequence quality of <i>Pan</i> I ^A alleles	69
8	Sequence quality of <i>Pan</i> I ^A alleles	70
9	Sequence quality of <i>Pan</i> I ^A alleles	71
10	Sequence quality of <i>Pan</i> I ^A alleles from a same clone	72
11	Sequence quality of <i>Pan</i> I ^B alleles	73

12	Gene tree from DNA sequences (including singletons) of <i>Pan I^A</i> alleles	75
13	Gene tree from DNA sequences (excluding singletons) of <i>Pan I^A</i> alleles	76
14	Sliding window analysis (excluding singletons) of nucleotide diversity (π) throughout the <i>Pan I</i> gene region in <i>Pan I^A</i> and <i>Pan I^B</i> alleles	83
15	Sliding window analysis (excluding singletons) of nucleotide diversity (π) throughout the <i>Pan I</i> gene region in <i>Pan I^A</i> alleles	84

List of Tables

1	Observed and expected <i>Pan</i> I genotypic and genic frequencies by depth among Atlantic cod at spring/spawning grounds	18
2	Observed and expected <i>Pan</i> I genotypic and genic frequencies by depth among Atlantic cod at fall-feeding grounds	19
3	Hierarchical <i>F</i> statistics for spring surveys 2005, 2006, and 2007 and fall surveys 2004, 2005, and 2006	23
4	Observed and expected <i>Pan</i> I genotypic and genic frequencies and X^2 statistic by age for spring surveys 2005, 2006, and 2007	25
5	Observed and expected <i>Pan</i> I genotypic and genic frequencies and X^2 statistic by year-class for spring surveys 2005, 2006, and 2007	26
6	Observed and expected <i>Pan</i> I genotypic and genic frequencies and X^2 statistic by sex for spring surveys 2005, 2006, and 2007	27
7	ANOVAs (one-way) of length on genotype at different ages, for spring surveys 2005, 2006, and 2007	29
8	Nucleotide polymorphism at <i>Pan</i> I, considering <i>Pan</i> I ^A and <i>Pan</i> I ^B together . .	33
9	Nucleotide polymorphism in <i>Pan</i> I ^A and <i>Pan</i> I ^B	34
10	Tajima's <i>D</i> test of Neutrality in <i>Pan</i> I	39
11	Fu and Li's tests of neutrality at <i>Pan</i> I	39
12	McDonald-Kreitman test for <i>Pan</i> I. Variation within and between <i>Pan</i> I ^A and <i>Pan</i> I ^B alleles	39
13	Genetic differentiation analysis of <i>Pan</i> I ^A from different MetaCod divisions (localities) in Iceland	40
14	Genetic differentiation analysis of <i>Pan</i> I ^A from different depth levels in Iceland	41
15	Extent of DNA divergence between <i>Pan</i> I ^A vs <i>Pan</i> I ^B alleles	42
16	Extent of DNA divergence between populations at different MetaCod divisions (localities), for <i>Pan</i> I ^A	42
17	Extent of DNA divergence between populations at different depth levels for <i>Pan</i> I ^A	43

Tables in Appendices

1	ANOVAs (one-way) of length on genotype at different ages, for fall surveys 2004, 2005, and 2006	64
2	Segregating sites in <i>Pan</i> I among single clones from 55 Atlantic cod from Iceland	78
3	Extent of DNA divergence between populations at different MetaCod divisions (localities), for <i>Pan</i> I ^A without singletons	85

4	Extent of DNA divergence between populations at different depth levels, for <i>Pan I^A</i> without singletons	85
---	--	----

Acknowledgements

I want to thank my supervisor, Professor Einar Árnasson, for guiding me in every step of this scientific journey, for his invaluable scientific advice, and for his enormous patience. It is always a pleasure to perform research under the guidance of such an enthusiastic and knowledgeable scientist. Thanks to Dr. Arnar Pálsson, the value of his advice on experimental methodology and numerical analysis can not be overestimated. Also, thanks to Dr. Snæbjörn Pálsson for his priceless advice on several statistical issues. Thanks to my laboratory colleagues, Svava Ingimarsdóttir, Katrín Halldórsdóttir, Guðni Magnús Eiríksson, and Hlynur Sigurgíslason for their support, practical help, and impartial criticism. Also thanks to Kristján Kristinsson for his assistance in providing samples for this study. Importantly, thanks to my son for being such a good kid while I was working at home, and for his truly interesting viewpoints regarding evolutionary issues. It reminded me that doing science requires imagination.

Introduction

Natural selection is differential reproduction of genotypes. It is a complex phenomenon of fitness variation spanning ontogenic stages and epigenetic relations. Genotypic variation among individuals in conjunction with environmental factors determines the phenotypic variation among individuals for the collection of traits expressed by a particular organism. In turn, the extent and nature of the phenotypic variation set the degree of variation in survival, fecundity, mating ability, and other aspects which determine fitness (the relative ability of different genotypes to pass on their alleles to future generations). Thus, natural selection operates on the phenotype (HEDRICK, 2005). It follows that natural selection is a consequential process by which genotypes with greater fitness leave more offspring (on average) than genotypes with less fitness (HARTL and CLARK, 1989). In other words, natural selection is an a posteriori principle born from the observations that, in a world of finite resources, organisms make use with differential efficiency of those resources in producing their progeny. The more efficient organisms will leave more descendants than their less efficient relatives (LEWONTIN, 1974, p. 3).

Most molecular variation is neutral, but for some loci the amount of genetic variation is determined by natural selection; i.e., they are loci under selection deviating from neutral expectations. In the neutral theory (KIMURA, 1968), alleles are selectively neutral with respect to each other. Although it makes no claims that allele substitutions underlying evolutionarily adaptive traits are neutral, it suggests that most allele substitutions at the molecular level have no selective advantage over those that they replace (KIMURA, 1991; HEDRICK, 2005). Near neutrality is an extension to the neutral theory: genes may be affected mostly by drift or mostly by selection depending on the effective size of a breeding population (OHTA, 2002). The selectionist/neutralist debate has evolved into acknowledging that much molecular variation is neutral, while endeavoring to estimate how selection acts and is distributed to affect genetic variation (KREITMAN, 2000; GUINAND *et al.*, 2004).

Natural selection, along with other factors, influences the extent of genetic differen-

tiation among the subdivisions of a population. Thus it can influence the ascertainment of the structure of a population. A population, a group of interbreeding individuals existing together in time and space (HEDRICK, 2005, p. 62), almost invariably exhibits differences in genotypic and allelic frequencies from one geographic area to another; i.e. there is a geographic population structure (HARTL and CLARK, 1989, p. 281). There are different approaches to estimating the extent of differentiation among the subdivisions of a population, the most important being hierarchical F statistics, where F_{ST} is a measure of genetic differentiation over subpopulations (HEDRICK, 2005, p. 488). Hierarchical F_{ST} statistics have bearings on the partition of genetic variation in a subdivided population (HEDRICK, 2005, p. 488). The extent and nature of genetic variation in turn are potentially affected by evolutionary factors, selection, inbreeding, genetic drift, gene flow, and mutation. Two or more of such factors combined can produce nearly any degree and pattern of genetic variation (HEDRICK, 2005, p. 30). It follows that natural selection and other factors can determine the degree of structure of a population. The current concept of a species' population structure amounts to the deviations exhibited by a species from the simplifying assumptions used by models of populations genetics. These assumption's are random mating (panmixia) and constant population size. However, such assumption's are very restrictive and no natural species fulfills them. Therefore, all species show some sort of a population structure (SLATKIN, 2005). Therefore, the main question is not whether some structure is observable in a population, but rather about the nature of the structure. An understanding of genetic population structure and gene flow is critical for conservation and management of exploited fish species. This is so for Atlantic cod (*Gadus morhua* L.), the most important commercial fish in the North Atlantic (ÁRNASON *et al.*, 2000).

Ever since variation at the Pantophysin I (*Pan* I) locus in Atlantic cod was first described by POGSON *et al.* (1995), compelling evidence has emerged about the distinctiveness of this gene with respect to among-population divergence (FEVOLDEN and POGSON, 1995, 1997; JÓNSDÓTTIR *et al.*, 1999, 2001; PAMPOULIE *et al.*, 2006; POGSON, 2001; POGSON and FEVOLDEN, 2003). POGSON *et al.* (1995, 2001) explored variation at supposedly neutral loci, *Pan* I among them, to study the genetic structure of Atlantic cod in the North Atlantic. This specific locus, *Pan* I (formerly labeled GM798), clearly stood out from other loci. It did not show an association between inferred levels of gene flow and geographic distance as observed with other loci. Instead, it showed high population differentiation (POGSON *et al.*, 2001). Furthermore, it displayed a high degree of linkage disequilibrium of three restriction site polymorphisms within *Pan* I (POGSON and FEVOLDEN, 1998). As it turned out, *Pan* I had two distinct alleles coexisting in populations, and two stretches of DNA within the *Pan* I

gene showed evidence of selection between the alleles. One stretch is the first intron of the gene, the other is the part of the gene coding for the intravesicular 1 (IV1) domain region of the protein (POGSON, 2001; GUINAND *et al.*, 2004). The *Pan I* gene has two alleles (1.85 kb in length), *Pan I^A* and *Pan I^B*, corresponding to the absence or presence of a *DraI* restriction site. The alleles are highly divergent at both nucleotide and amino acid level, with six amino acid substitutions separating the alleles (POGSON, 2001).

The *Pan I* locus has been used in attempts to separate populations of cod in the North-East Atlantic (SARVAS and FEVOLDEN, 2005). However, the evidence indicated that *Pan I* was under selection. Limited population structure over the whole geographic range of Atlantic cod has been reported in research that analyzed allozyme and mtDNA polymorphism (MORK *et al.*, 1985; ÁRNASON *et al.*, 1992). Subsequent studies (POGSON *et al.*, 1995; BENTZEN *et al.*, 1996; RUZZANTE *et al.*, 1996) that looked at nuclear DNA markers arrived at a different conclusion, bringing to view a significant, albeit weak, population structure at both large and small geographic scales. Among the nuclear markers was *Pan I*, which FEVOLDEN and POGSON (1997) sequenced and identified (by then as Synaptophysin [*Syp I*]). At this locus, they found considerable differentiation between Arctic and coastal populations of cod in northern Norway. Such considerable differentiation at *Pan I* was considered by FEVOLDEN and POGSON (1997) as evidence of a marked historical population structure, adding that the *Pan I* polymorphism discloses significant heterogeneity among cod populations at very localized scales. On the other hand, ÁRNASON and PÁLSSON (1996), analyzing the mitochondrial cytochrome *b* region among Norwegian cod, found no measurable population differentiation although they came across substantial levels of nucleotide polymorphism. They concluded that the cytochrome *b* variation does not support a theory of historical population structure among Norwegian cod stocks. In addition, ÁRNASON and PÁLSSON (1996) presented evidence that the patterns of cytochrome *b* sequence variation are consistent with neutral expectations. If such variation is indeed neutral, the striking differences detected among populations at *Pan I* are perhaps only explainable by the operation of selection acting at the locus or at a tightly linked locus (FEVOLDEN and POGSON, 1997). Further studies of the mitochondrial cytochrome *b* by ÁRNASON *et al.* (2000) conclude that there is no significant evolutionary difference between Greenland and Iceland cod, and that there is no substantiation for considering them to consist of separate evolutionary units. ÁRNASON *et al.* (2000) warn that although FEVOLDEN and POGSON (1997) found divergence between coastal and Arctic cod in Norway at the *Pan I* locus, which they interpret as evidence for historical population structure, natural selection is immediately implicated at this locus (FEVOLDEN and POGSON, 1997; POGSON and FEVOLDEN,

1998). Therefore, it is essential to consider the role of selection. KARLSSON and MORK (2003) conclude, from research conducted on Atlantic cod in Trondheimfjord, Norway, that mutation, genetic drift, and migration do not appear to contribute significantly to the observed genetic heterogeneity at *Pan I*. Thus, natural selection remains as the main explanatory factor for the Hardy Weinberg imbalance detected.

The biochemical mechanisms of operation of natural selection on *Pan I* are unknown since the precise physiological function of Pantophysin is poorly understood (POGSON, 2001). However, the structure of Pantophysin has been characterized (HAASS *et al.*, 1996; LEUBE, 1994; POGSON, 2001). The gene originally identified by Fevolden and Pogson (1997) as the Synaptophysin (*Syp I*) locus, more likely corresponds to a cellular isoform of Synaptophysin that LEUBE (1994) detected while attempting to identify Synaptophysin-related molecules, and which was referred to as Pantophysin. Its structure is characterized by four membrane-spanning domains, two intravesicular loops, and two cytoplasmic tails (HAASS *et al.*, 1996). Synaptophysin is one of the major integral membrane proteins of transmitter-containing synaptic vesicles in neurons and of similar vesicles in neuroendocrine (NE) cells, and has a central role for transmitter exocytosis (HAASS *et al.*, 1996). In addition, *Syp I* is transcribed in a cell-type specific manner (LEUBE, 1994). Pantophysin is also an integral membrane protein, whose gene closely resembles that of Synaptophysin. Pantophysin is localized in cytoplasmic microvesicles that function in a variety of shuttling, secretory, and endocytotic recycling pathways (HAASS *et al.*, 1996). In contrast to Synaptophysin, Pantophysin is expressed ubiquitously and found in both neuroendocrine and non-neuroendocrine tissues. It defines constitutive small cytoplasmic transport vesicles independent of their cargo in many cell types, and thus stands for a shared property of these different vesicles, suggesting that it carries out basic structural and house-keeping functions (HAASS *et al.*, 1996). Transport vesicles are the core operators of the endocytic and biosynthetic-secretory pathways (elaborate membrane systems that convey materials) of eukaryotic cells. These pathways are essential mechanisms that allow the cell to take up, process, and secrete a wide array of molecules. Vesicles constantly bud off from a membrane to fuse with another, carrying membrane components and soluble molecules. Such membrane traffic runs along highly organized directional routes, with the biosynthetic-secretory pathway flowing from the endoplasmic reticulum toward the Golgi apparatus and cell surface while the endocytic pathway flows inward from the plasma membrane (ALBERTS *et al.*, 2002).

The effects of selection can be detected with different approaches. Measurement of the effects of selection based solely on morphological or physiological features is an extremely hard task as it depends on using phenotypic traits that closely match the genotype. Since natu-

ral selection changes allele or haplotype frequencies and affects DNA sequences, other broad alternatives for the detection of effects of selection are allele frequency-based and nucleotide sequence-based approaches (GUINAND *et al.*, 2004).

Analysis of allele frequency distribution compared to the expected distribution under neutrality can indicate the effects of natural selection. Organisms are potentially affected by natural or artificial selective agents (GUINAND *et al.*, 2004), which change allele frequencies (HEDRICK, 2005). Thus, gradients of selective pressure due to varying magnitudes of selective agents may produce a corresponding gradient in allele frequency, which would be reflected in relations with defined form and direction (patterns) between the locus presumably under selection and the environmental or biological variable (or a proxy variable) acting as the selective agent. However, because of the complex nature of selection processes (HEDRICK, 2005), patterns may prove difficult to discern from noise.

Analysis of nucleotide sequence variation is an appropriate tool to examine the action of natural selection on *Pan I*. Since the arrival of DNA sequencing for population genetics research (started by KREITMAN, 1983), investigations looking at nucleotide sequence variation in natural populations have produced valuable insights into the role of natural selection in molding the patterns of nucleotide polymorphism within species and nucleotide divergence between species (HUDSON, 1990; KREITMAN, 1991; KREITMAN and AKASHI, 1995). Combining such an analysis with the theory of gene-genealogies renders a matchless device for analyzing the past and present selective forces operating on a genetic locus. This entails a comprehensive molecular dissection of variation (cf. BERRY and KREITMAN, 1993) that reveals the mechanisms of the evolutionary forces of natural selection, random drift, gene flow, and mutation. Tests of neutrality of DNA sequence variation are examples of nucleotide-based analytical tools that permit detection of deviations from neutrality, and recently the attention has centered on determining the presence and extent of selection on DNA sequence variation (HEDRICK, 2005). Nucleotide sequence variation at the *Pan I* locus in Atlantic cod strongly indicates an unusual mixture of balancing and directional selection (POGSON, 2001). The significant linkage disequilibrium and ample differences in *Pan I* allelic frequencies among populations of Atlantic cod do not seem to derive from stable spatially varying selection. Instead, they may be due to recent appearance and expansion of selectively favored mutations within both *Pan I*^A and *Pan I*^B allelic classes in different geographic regions. Furthermore, the nature of balancing selection that might be operating at the *Pan I* locus is not known (POGSON, 2001).

There are inherent difficulties in detecting the effects of natural selection. Popula-

tions are potentially affected by numerous selective pressures, and these selective pressures may also affect their demography (GUINAND *et al.*, 2004). Therefore, it is difficult to disentangle the effects of natural selection from those of demographic history (KREITMAN, 2000; WALL *et al.*, 2002). On the other hand, technical difficulties can also arise. Taq DNA polymerase induces DNA replication errors (KEOHAVONG and THILLY, 1989), which could introduce artifacts that may be confounded with nucleotide sequence variation.

This study aims to enhance the understanding of the role that natural selection has on the *Pan I* locus in Atlantic cod. This research tackles significant issues concerned with *Pan I*. Namely, identifying variables intimately connected to natural selection phenomena on the locus, and characterizing the effects of selection on *Pan I* with regards to allele frequency distribution and DNA sequence variation. One of the premises is that the effects of natural selection and variables intimately involved in selection phenomena can be detected in the form of relations with defined shape and direction between *Pan I* and other biological or environmental factors. Thus, the study evaluates the hypothesis that genotypic and genic variation in *Pan I* show spatial and temporal patterns of variation and the stability of such patterns. Another premise is that analysis of DNA sequence variation can give account of effects of natural selection. Thus, the study explores DNA sequence variation in *Pan I* to assess the effects of natural selection. Also in connection with effects of natural selection, this investigation evaluates population structure along with the nature of such structure, using genic and DNA sequence data. The analysis is based on population genetics theory, encompassing examination of genetic and genic data as well as comparative studies of DNA sequence variation of coding and non-coding parts of the *Pan I* gene. Altogether, the study seeks to advance our knowledge of natural selection acting on an ecologically and commercially important marine species, Atlantic cod.

Materials and Methods

Sampling

Samples were collected for the DNA based phylogeography and population genetics of North Atlantic fish (DNAfish) project during Marine Research Institute (MRI)/Hafrannsóknastofnunin spring-spawning ground surveys for the years 2005, 2006, and 2007 and fall-feeding ground surveys for the years 2004, 2005, and 2006.

Most of the samples from the spring-spawning ground surveys 2005, 2006 and 2007 were from inshore locations (Figure 1) while the majority of the samples from the fall-feeding ground surveys 2004, 2005, and 2006 were from offshore locations (Figure 2).

The MetaCod divisions (MCdiv) were a set of geographical regions surrounding Iceland, and the criteria for division was a composite of biological and environmental parameters such as ocean currents and larval drift (PAMPOULIE *et al.*, 2006). All MetaCod divisions, except division 6, were represented in the spring-spawning ground surveys 2005, 2006 and 2007 (Figure 1); all Metacod divisions were represented in the fall-feeding ground surveys 2004, 2005, and 2006 (Figure 2).

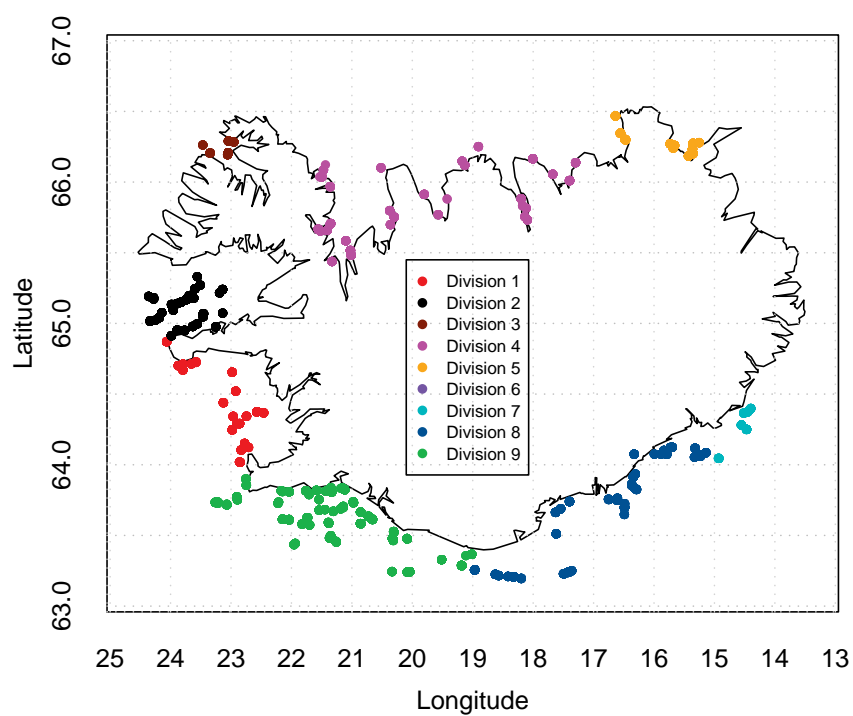


Figure 1: Spring survey stations by MCdiv, years 2005, 2006, and 2007. MCdiv is MetaCod division, a set of geographical regions surrounding Iceland

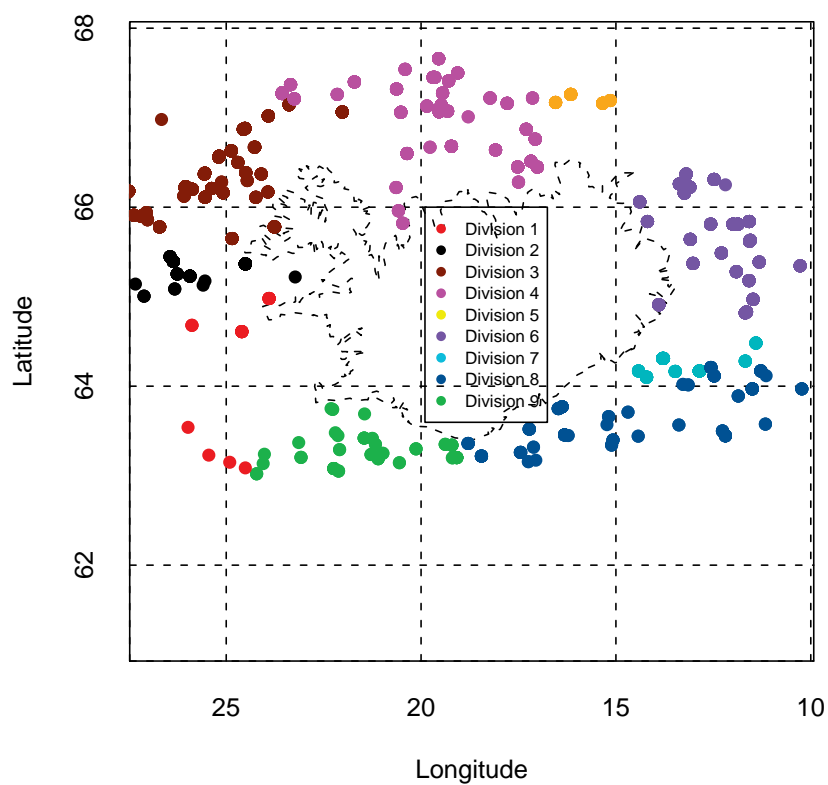


Figure 2: Fall survey stations by MCdiv, years 2004, 2005, and 2006. MCdiv is MetaCod division, a set of geographical regions surrounding Iceland

Genomic DNA extraction and Genotyping

The DNA extraction utilized a Chelex/Proteinase K methodology (WALSH *et al.*, 1991). Tissue fragments $\sim 1.5 \text{ mm}^2$ (0.3 mm^3) were cut with a knife from gill samples kept in 96% ethanol in barcoded 1.5 ml vials. Each tissue fragment was paper dried, put into 250 μl of a Chelex-Proteinase K solution (WALSH *et al.*, 1991) in a 96 deep well plate, and processed at 65°C and 950 rpm for 2–3 hours with a thermomixer. Three blank controls were used. To reduce errors, a digital record of the barcode of each sample and its location within the DNA extraction plate was entered into a spreadsheet simulating a 96 well plate using a barcode reader. The plate was placed in waterbath at 95°C for 5 minutes. It was centrifuged at 3000 rpm for 5 minutes, and $\sim 120 \mu\text{l}$ of the supernatant containing the DNA were transferred to a new 96 well plate. A 1:19 dilution was also made.

Genotyping was done using restriction fragment patterns (FEVOLDEN and POGSON, 1997) on nearly all Atlantic cod samples described in the sampling section. Altogether 8196 individuals were successfully genotyped: 5384 from the spring surveys 2005, 2006, and 2007 and 2812 from the fall surveys 2004, 2005, and 2006. Genomic DNA was used to amplify a 489 bp fragment representing a section of the *Pan I* locus (1.85 kb), using *Pan I* 3 forward primer (5' CGTTGGTCCTCTATCTGGGCTTC 3') and *Pan I* 20 Reverse primer (5' AAGACGAAAC-CAACCACAGGA 3') (POGSON and MESA, 2004). Amplifications were performed in 20 μl reaction volumes containing: 0.2 nmole/ μl dNTP, 0.16% Tween20, 1 \times Taq buffer, 0.105 $\mu\text{g}/\mu\text{l}$ BSA, 0.25 nmole/ μl *Pan I* 3 forward primer, 0.25 nmole/ μl *Pan I* 20 reverse primer, 0.0475 units/ μl Taq polymerase, 5 μl genomic DNA (1:19 dilution). PCR reactions used the following profile (POGSON, 2001): Initial denaturation step at 94°C for 45 seconds, and subsequently 35 cycles of 1 second denaturation at 94°C, 1 second annealing at 52°C and 1 minute elongation at 72°C; after cycling, there was a final elongation step of 2 minutes at 72°C. The 489 bp amplified fragment of *Pan I* was digested with restriction enzyme *DraI*. Within this fragment, *Pan I*^A lacked a *DraI* restriction site present in *Pan I*^B. The *DraI* digestion rendered one uncut (489 bp) fragment for *Pan I*^A and two fragments (209 bp and 280 bp) for *Pan I*^B. The *DraI* digestion reactions were performed in 27 μl reaction volumes with the following composition: 20 μl of the completed amplification solution, 0.045 units/ μl *DraI*. The reaction was incubated at 37°C for 3 hours. A total of 10 μl of completed digestion reaction was electrophoresed at 60 volts/cm for 70 minutes, with a 100 bp ladder, in a 2.5% agarose gel (0.5 \times TBE buffer) with ethidium bromide. The gel was placed on an UV table and a digital image of the restriction patterns was recorded. The products separated best with gels cast for 3 hours at least. The digital images of the restriction fragments were used to determine and score the genotypes.

Pan I^{AB} heterozygous individuals exhibited a pattern of 3 bands, 209 bp, 280 bp and 489 bp (Figure 3).

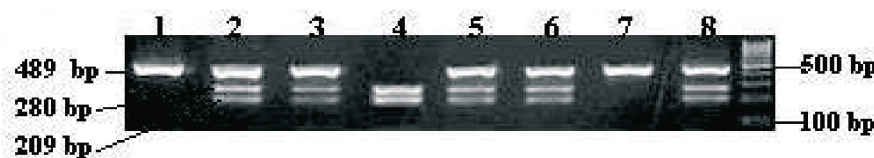


Figure 3: An example gel image for scoring *Pan I* genotypes. The genotypes are homozygous *Pan I*^{AA} (lanes 1 and 7) with 1 band: 489 bp; heterozygous *Pan I*^{AB} (lanes 2, 3, 5, 6, 8) with 3 bands: 209 bp, 280 bp, and 489 bp; homozygous *Pan I*^{BB} (lane 4) with 2 bands: 209 bp and 280 bp. A 100 bp ladder is shown for reference on the right.

Cloning and Sequencing

Utilizing the genotyping results, homozygous individuals were selected for cloning and sequencing by stratified random sampling at different depth strata and geographical regions/divisions surrounding Iceland. Genomic DNA was used to amplify a 1.94 kb fragment encompassing the *Pan I* gene (1.85 kb), using *Pan I* 4 forward primer (5' CTTCCATTCATCC-GAGTTCTG 3') and *Pan I* 7 Reverse primer (5' CGTAGCAGAAGAGTGACACAT 3') (POGSON, 2001). "Touchdown" PCR (DON *et al.*, 1991) was used to minimize the amplification of spurious DNA fragments. Amplifications were performed in 40 μ l reaction volumes containing: 0.2 nmole/ μ l dNTP, 0.16% Tween20, 1 \times Taq buffer, 0.105 μ g/ μ l BSA, 0.25 nmole/ μ l *Pan I* 4 forward primer, 0.25 nmole/ μ l *Pan I* 7 reverse primer, 0.0475 units/ μ l Taq polymerase, 10 μ l genomic DNA (undiluted). PCR reactions used a combination of the profile described by POGSON (2001) and the "touchdown" technique (DON *et al.*, 1991): Initial denaturation step at 94°C for 45 seconds, and subsequently 1 second denaturation at 94°C, 1 second annealing at 62°C (progressively decreasing 1°C every two cycles) and 1 minute 40 seconds elongation at 72°C; when the former cycling came to an annealing temperature of 52°C, 15 more cycles were repeated using 52°C as annealing temperature. The possibility exists that incomplete PCR products may prime DNA from the chromosome of a diploid. During a final extension step as is common in PCR a chimeric molecule may be produced. Therefore, no final elongation step was used here to reduce or minimize possibility of *in vitro* chimeric molecules. The amplified products were electrophoresed at 90 volts/cm for 45 minutes, with a 1 kb ladder, in a 1% agarose gel (1 \times TAE buffer) with ethidium bromide. The 1.94 kb fragments were identified

on a UV table and minute gel pieces containing them were excised from the larger agarose gel. The 1.94 kb fragments were recovered from the gel pieces with an Ultra-agarose-spin-kit (AB-gene) using the manufacturers instructions (except that 5 minutes were used for melting the gel pieces), rendering 50 μ l of agarose-gel-purified DNA. This was done to sort out the 1.94 kb amplified fragments from any spurious DNA fragments that would interfere with the cloning process.

The agarose-gel-purified 1.94 kb DNA fragments were TOPO TA cloned into a plasmid vector (pCR4-TOPO), using TOPO TA cloning kit for sequencing (Invitrogen) according to the manufacturer's instructions with the following alterations: The amount of vector per cloned sample was 1/8 μ l; the amount of DNA was 2 μ l; also, the incubation time (at room temperature) used for the cloning reaction was 30 minutes instead of 5. The cloning reaction was performed in 6 μ l reaction volumes. The plasmid vector (pCR4-TOPO) was kanamycin and ampicillin resistant with a cell cycle death (CCD) gene to facilitate selection. Chemically competent cells were transformed with the vector containing the 1.94 kb fragment using the manufacturer's instructions (Invitrogen) with the following alterations: The 6 μ l cloning reaction was carefully mixed by stirring with 100 μ l of in-house produced chemically competent *E. coli* cells (University of Iceland/Biology Department/Population Genetics). The incubation time on ice was 30 minutes, followed by a heat-shock of 30 seconds at 42°. After transformation, 250 μ l of SOC medium was added and the transformants spread on pre-warmed selective plates (composition 1.5% agar with standard LB medium, kanamycin 50 μ g/ml media). The plates were incubated overnight at 37°C. Three colonies were picked using tooth-picks and incubated overnight in 5 ml of standard LB media with kanamycin 50 μ g/ml at 37°C and 200 rpm. After overnight incubation, 5 ml of the liquid culture of bacterial transformants was centrifuged at 7000 rpm (Eppendorf centrifuge) for 3 minutes and vector with insert isolated using QIAprep-spin-miniprep kit (QIAGEN) according to manufacturer's instructions. The extracted DNA was retained in a 50 μ l volume of EB buffer (Qiagen). To see if a 1.94 kb DNA fragment had been successfully incorporated in the plasmid vector, the DNA was digested with *Eco*RI restriction enzyme. Two bands on electrophoresis were expected: One \sim 3.5 kb (vector) and the other \sim 2 kb (the desired 1.94 kb *Pan* I fragment). The digestion reactions were performed in 5 μ l reaction volumes with the following composition: 2 units *Eco*RI, 1 \times *Eco*RI buffer, 0.5 μ l plasmid DNA. Digestions were done at 37°C for at least one hour. The 5 μ l of completed digestion reaction were electrophoresed at 90 volts/cm for 45 minutes in a 1.5% agarose gel (0.5 \times TBE buffer) with ethidium bromide.

Clones verified to contain 1.94 kb inserts were sequenced. Separate sequencing re-

actions were performed using 6 different sequencing primers, which covered the *Pan I* gene in overlapping parts. The primers utilized were: *Pan I* 3 Forward (5' CGTTGGTCCTC-TATCTGGGCTTC 3'), *Pan I* 4 Forward (5' CTTCCATTCATCCGAGTTCTG 3'), *Pan I* 6 Forward (5' ACCTTTACTCTCTATCTCCCG 3'), *Pan I* 14 Forward (5' GACGCTTTCTTTGATTG-GCAG 3'), *Pan I* 7 Reverse (5' CGTAGCAGAAGAGTGACACAT 3'), *Pan I* 8 Reverse (5' CGAATGGGTCTCACTTGAAGTC 3') (POGSON and MESA, 2004). The sequencing reactions were performed in 10 μ l reaction volumes with the following composition: 1 μ l BigDye termination reaction reagent (TRR) (ABI), 1 \times BigDye buffer (ABI), 0.16 pmole/ μ l sequencing primer, 0.8 μ l plasmid DNA. Sequencing reactions used the following profile: Initial denaturation step at 96°C for 10 seconds and subsequently 25 cycles consisting of denaturation at 96°C for 10 seconds, annealing at 50°C for 10 seconds, and elongation at 60°C for 2 minutes. Upon completion of the thermal cycling, excess nucleotides, dye, and salts were removed with precipitation. To 10 μ l of sequencing reaction were added 50 μ l of a NaOAc-glycogen solution (0.3 M NaOAc, 0.1 μ g/ μ l) and 125 μ l of ice-cold (−20°C) ethanol (96%). The resulting solution was mixed by flipping back and forth the 96 well plate that contained it. The 96 well plate was centrifuged at 4000 rpm for 30 minutes at 4°C. Immediately after centrifugation the precipitation mix was poured off, the plates were inverted, wrapped on drying paper and centrifuged inverted at 300 rpm for 2 minutes at 4°C. Then, 250 μ l of ice-cold (−20°C) ethanol (70%) were added. The plate was centrifuged at 4000 rpm for 5 minutes at 4°C, and right upon completion the ethanol (70%) was discarded. The plates were once again wrapped in drying paper and centrifuged inverted at 300 rpm for 5 minutes at 4°C. The plates were air-dried for an hour in the dark and the DNA subsequently resuspended in 10 μ l of HiDi formamide (ABI). The samples were run on an ABI 3100 genetic analyzer.

The commercial sources of reagents utilized in all PCR reactions and enzymatic digestions were: dNTP (Fermentas), Tween20 (Pharmacia Biotech), Taq polymerase and Taq buffer (New England Biolabs), BSA (Fermentas), primers (TAG Denmark), *DraI* (Fermentas), *EcoRI* and *EcoRI* buffer (New England Biolabs). The apparatus used for all PCR reactions and enzymatic digestions was a DNA-Engine2Tetrad by MJ Research.

Data analysis

Quantitative analysis, computer simulations, and statistical and numerical analysis were performed on the genetic, environmental, biological, and DNA sequence data using R (R DEVELOPMENT CORE TEAM, 2008). Tests of association between *Pan I* genotype and biological and environmental variables were done using Pearsons Chi-squared statistic X^2 .

Analysis of Hardy Weinberg equilibrium and calculation of allele frequencies were performed at different values of depth, age, year-class, and sex. Hierarchical F statistics analysis was carried out with the package hierfstat (GOUDET, 2006). Multiple one way ANOVA's using *Pan* I genotype as a factor and means of length as response variable were calculated at each age group. A sequential Bonferroni correction was used to control the family-wise type I error (KEOUGH and QUINN, 2002) at a maximum of 0.05 (Significance Level, SL). This was done to know if there was a significant difference among groups of *Pan* I genotype in length, and also to know the range of ages at which such difference was significant.

Quality values for the DNA sequences were analyzed also with R using the APE package (PARADIS *et al.*, 2004). The DNA sequences were base-called and assembled into contigs to build consensus sequences per each clone using the program suite Phred-Phrap-Consed (EWING *et al.*, 1998; GORDON *et al.*, 1998; GREEN, 1994). Sequence alignment of consensus sequences was performed with Seaview, a graphical multiple sequence alignment editor (GALTIER *et al.*, 1996) using either the MUSCLE alignment program (EDGAR, 2004) or the ClustalW alignment program (THOMPSON *et al.*, 1994). Measures of nucleotide polymorphism (e.g. nucleotide diversity) and population genetic tests such as tests of neutrality of sequence data, analysis of genetic differentiation by permutations, pairwise analysis of genetic distance, were obtained using DNAsp (ROZAS *et al.*, 2003). Tests of neutrality performed included Tajima's (TAJIMA, 1983), Fu and Li's (FU and LI, 1993), and McDonald-Kreitman's (MCDONALD and KREITMAN, 1991) tests. Tajima's test looks at the difference between two estimates of nucleotide variation: θ_π , an estimate based on number of nucleotide differences and affected by frequency of sequences, minus θ_S , an estimate based on number of segregating sites and not affected by frequency of sequences. This difference is divided by its standard deviation, which constitutes the D test statistic (HEDRICK, 2005). Fu and Li's test of neutrality looks at the internal and external branches of the genealogy of a set of sequences. D^* statistic is based on the differences between η , the total number of mutations, and η_s , the number of singletons (FU and LI, 1993; ROZAS *et al.*, 2003). The F^* statistic is based on the differences between k , the average number of nucleotide differences between pairs of sequences, and η_s , the number of singletons (FU and LI, 1993; ROZAS *et al.*, 2003). The premise of the McDonald-Kreitman's test of neutrality is that, under neutrality, the ratio of non-synonymous to synonymous fixed differences between species is the same as the ratio of non-synonymous to synonymous polymorphisms within species (MCDONALD and KREITMAN, 1991; HEDRICK, 2005). Here the test is applied to a between alleles vs. within alleles comparison. Nucleotide polymorphism distribution was analyzed by a sliding window approach (KREITMAN and HUD-

SON, 1991) with window and step size of 100 bp and 25 bp, respectively. Genetic trees and a summary of segregating sites in *Pan* I were constructed using MEGA (KUMAR *et al.*, 2004).

Results

Genotypic variation in relation to biological and environmental factors

Pan I in terms of depth (spring and fall surveys)

Depth and *Pan* I allele frequencies (p_A) were correlated. At shallow waters there was a higher frequency of *Pan* I^A and at deeper waters there was a lower frequency of *Pan* I^A (Tables 1 and 2) with a regular allele frequency gradient $\Delta p_A \approx -0.4\%/m$. An outlier in allele frequency ($p_A = 0.55$) at the 175–200 m depth class of the spring surveys (Table 1) was not considered for the linear regression of allele frequency on depth above 200 m. This pattern was maintained during both spring and fall surveys (Figure 4). Nonetheless the genotypic composition differed between surveys.

The genotypic composition changed between spring and fall surveys. For the spring surveys, the samples from the three top layers in shallow waters were not in Hardy Weinberg (H.W.) equilibrium and showed a deficiency of heterozygotes (apparent Wahlund effects; Table 1). In contrast, a deficiency of heterozygotes in shallow waters was not seen for the fall surveys. Instead, an excess of heterozygotes was found at various depths (Table 2). When comparing the same depth classes, p_A 's were higher for fall surveys than for spring surveys in all instances but one, the 175–200 m depth class (Tables 1 and 2). However, there was a higher overall mean allele frequency for the spring surveys ($\bar{p}_A = 0.45$, Table 1) than for the fall surveys ($\bar{p}_A = 0.34$, Table 2). In general, this was due to sampling at different depths between the spring and fall surveys and the overall mean does not reflect differences in biology.

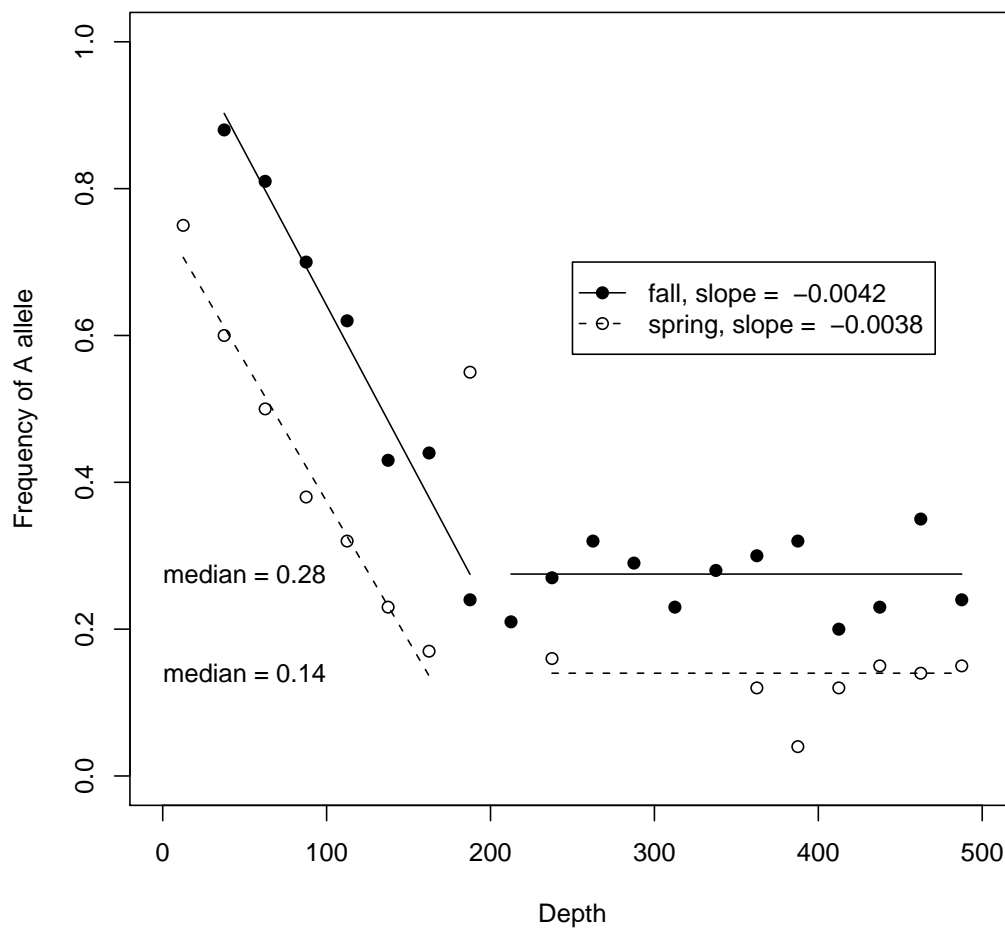


Figure 4: Allele frequency vs depth, spring surveys 2005, 2006, and 2007 combined and fall surveys 2005, 2006, and 2007 combined. The circles represent allele frequencies of *Pan I^A* (p_A) at depth intervals of 25 m, open circles for spring and filled circles for fall. The lines represent linear regressions of allele frequency (p_A) on depth above 200 m and medians of allele frequency p_A below 200 m depth, dashed lines for spring and solid lines for fall. The slopes of the linear regressions are shown in a box. Depth is in m.

Table 1: Observed and expected *Pan* I genotypic and genic frequencies by depth among Atlantic cod at spring-spawning grounds. Allele frequency of *Pan* I^A allele is represented by p_A . Significant deviations from Hardy Weinberg are marked in blue X^2 statistics and are heterozygote deficiency in all instances.

Depth	Observed			Sum	p_A	Expected			X^2
	AA	AB	BB			AA	AB	BB	
0–25	280	161	40	481	0.75	270.19	180.62	30.19	5.68
25–50	300	262	154	716	0.60	259.44	343.11	113.44	40.02
50–75	512	789	512	1813	0.50	453.25	906.50	453.25	30.46
75–100	157	442	394	993	0.38	143.89	468.22	380.89	3.11
100–125	48	238	241	527	0.32	52.92	228.16	245.92	0.98
125–150	7	64	96	167	0.23	9.11	59.78	98.11	0.83
150–175	4	31	83	118	0.17	3.22	32.56	82.22	0.27
175–200	6	10	4	20	0.55	6.05	9.90	4.05	0.00
225–250	1	12	31	44	0.16	1.11	11.77	31.11	0.02
350–375	0	16	49	65	0.12	0.98	14.03	49.98	1.28
375–400	0	2	21	23	0.04	0.04	1.91	21.04	0.05
400–425	0	17	54	71	0.12	1.02	14.96	55.02	1.31
425–450	5	39	117	161	0.15	3.73	41.54	115.73	0.60
450–475	0	12	32	44	0.14	0.82	10.36	32.82	1.10
475–500	0	26	63	89	0.15	1.90	22.20	64.90	2.60
Sum	1320	2121	1891	5332	0.45	1062.79	2635.43	1633.79	203.16

Table 2: Observed and expected *Pan* I genotypic and genic frequencies by depth among Atlantic cod at fall-feeding grounds. Allele frequency of *Pan* I^A allele is represented by p_A . Significant deviations from Hardy Weinberg are marked in red X^2 statistics and are heterozygote excess in all instances.

Depth	Observed			Sum	p_A	Expected			X^2
	AA	AB	BB			AA	AB	BB	
25–50	6	2	0	8	0.88	6.12	1.75	0.12	0.16
50–75	73	28	6	107	0.81	70.74	32.52	3.74	2.07
75–100	79	74	13	166	0.70	81.06	69.88	15.06	0.58
100–125	17	30	3	50	0.64	20.48	23.04	6.48	4.56
125–150	46	147	77	270	0.44	52.89	133.22	83.89	2.89
150–175	32	89	50	171	0.45	34.22	84.55	52.22	0.47
175–200	9	69	101	179	0.24	10.57	65.86	102.57	0.41
200–225	9	58	115	182	0.21	7.93	60.13	113.93	0.23
225–250	20	112	151	283	0.27	20.41	111.18	151.41	0.02
250–275	10	62	55	127	0.32	13.24	55.53	58.24	1.73
275–300	8	52	58	118	0.29	9.80	48.41	59.80	0.65
300–325	7	78	124	209	0.22	10.12	71.75	127.12	1.59
325–350	10	124	127	261	0.28	19.86	104.28	136.86	9.34
350–375	6	36	42	84	0.29	6.86	34.29	42.86	0.21
375–400	12	74	67	153	0.32	15.69	66.61	70.69	1.88
400–425	2	62	104	168	0.20	6.48	53.04	108.48	4.80
425–450	6	55	84	145	0.23	7.74	51.52	85.74	0.66
450–475	1	5	4	10	0.35	1.22	4.55	4.23	0.10
475–500	1	39	44	84	0.24	5.00	30.99	48.00	5.60
Sum	354	1196	1225	2775	0.34	326.60	1250.81	1197.60	5.33

Depth and geographic region (spring and fall surveys)

Depth and geographic region were confounded. Depths at the sampling locations were correlated with the geographic regions (Figures 5 and 6; also Figures 1 and 2 in the Appendix). The deepest locations coincided with the southernmost geographic regions for the spring surveys (Figure 5, and Figure 1 in the Appendix). Most locations deeper than 100 m were situated further south than 65° of latitude with very few exceptions in northern Icelandic regions (Figure 5, and Figure 1 in the Appendix). With regards to the fall surveys a similar trend was seen as well: The further offshore a locality, the greater was the depth (Figure 6, and Figure 2 in the Appendix). As for the inshore localities of the fall surveys, the same trend already described for the spring surveys was also found here: The further south the latitude, the greater was the depth (Figure 6). When contrasting the spring and fall surveys in terms of sampling depth (Figures 5 and 6; also Figures 1 and 2 in the Appendix) the sampling stations of the fall surveys (mean depth = 263 m, median depth = 250 m) were, in general, deeper than the sampling stations of the spring surveys (mean depth = 98 m, median depth = 73 m).

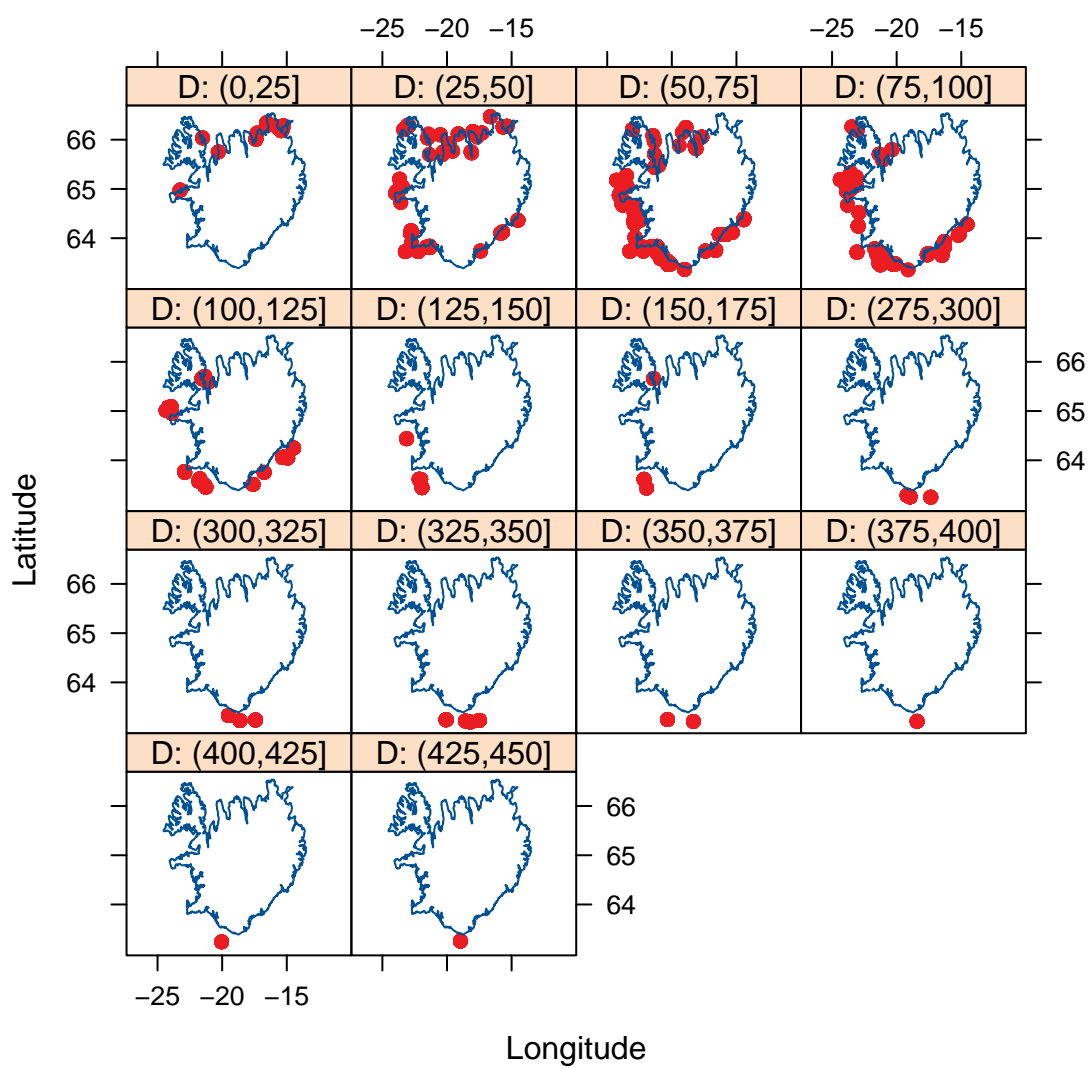


Figure 5: Sampling sites, latitude and longitude conditioned on depth; spring surveys 2005, 2006, and 2007.

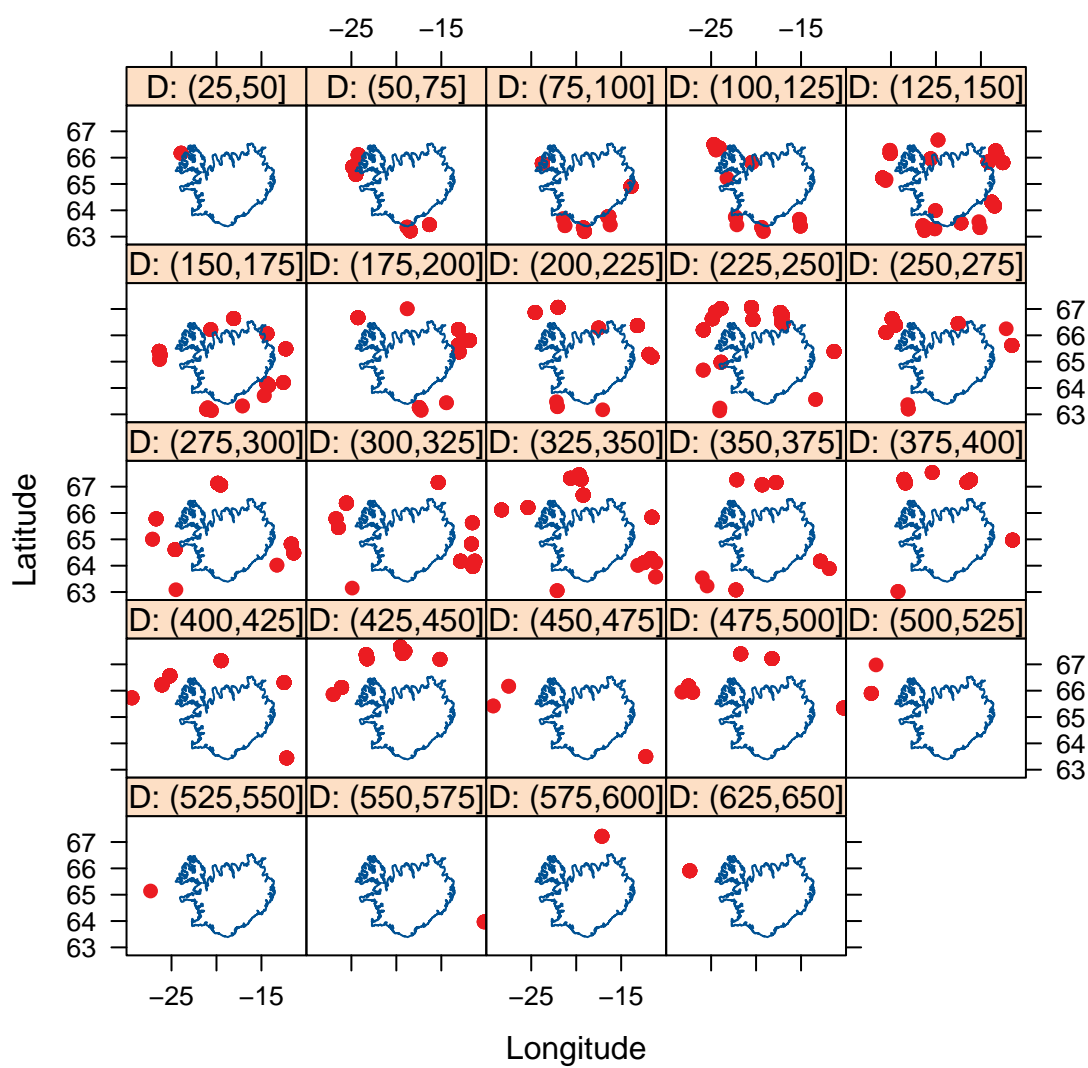


Figure 6: Sampling sites, latitude and longitude conditioned on depth; fall surveys 2004, 2005, and 2006.

Hierarchical F statistics (spring and fall surveys)

Genetic differentiation over subpopulations was significant in terms of geographic area and in terms of depth. Concerning the spring surveys, genetic differentiation over subpopulations, measured by F_{ST} , was significant when the subpopulations were defined by geographic area using MetaCod divisions, a set of geographic regions surrounding Iceland (Table 3, upper panel). Moreover, genetic differentiation over subpopulations was significant also when the subpopulations were defined in terms of depth (Table 3, upper panel). It is noteworthy that F_{ST} in terms of depth was larger than F_{ST} in terms of geographic area as seen in Table 3 (upper panel). Regarding the fall surveys, the hierarchical F analysis resembled that of the spring surveys: There was significant genetic differentiation in terms of both depth and geographic area (Table 3, lower panel). Again, the F_{ST} estimate was larger when the subpopulations were defined in terms of depth than in terms of geographic area (Table 3, lower panel).

Table 3: Hierarchical F statistics for spring surveys 2005, 2006, and 2007 and fall surveys 2004, 2005, and 2006. The upper panel shows F statistics for spring surveys and the lower panel for fall surveys. Fixation indices were calculated with one subdivision factor exclusively, MCdiv in one instance and Depth in another; MCdiv is MetaCod division, a set of geographical regions surrounding Iceland; $p_{F_{ST}}$ was calculated using 1000 permutations.

Comparison	Subdivision factor	F_{IS}	F_{ST}	F_{IT}	$p_{F_{ST}}$
Spring surveys	MCdiv only	0.142	0.077	0.208	0.001
Spring surveys	Depth only	0.095	0.135	0.217	0.001
Fall surveys	MCdiv only	-0.006	0.057	0.052	0.001
Fall surveys	Depth only	-0.085	0.123	0.048	0.001

Associations of *Pan* I genotype with other biological variables (spring surveys)

Pan I genotype in relation to age, year-class, and sex

There were significant associations between *Pan* I genotype and age, year-class, and sex. *Pan* I genotype was significantly associated with age (Pearson's Chi-squared test statistic $X^2 = 406.69$, $df = 18$, and $p < 2.2 \exp -16$). The interval of age used for this test of independence was 3 to 12 years (Table 4). There was a pattern with the highest allele frequency, p_A , corresponding to the youngest ages and p_A decreasing as age increased. Ages from 4 to 10 years were out of H.W. equilibrium ($X^2 > 3.84$ for those years), showing heterozygous deficiency and apparent Wahlund effect (Table 4). *Pan* I genotype was significantly associated

with year-class ($X^2 = 311.72$, $df = 18$, and $p < 2.2 \exp -16$). The interval of year-classes used for this test of independence was 1994 to 2003 (Table 5); other year-classes were excluded because of small sample sizes. There was a trend with the highest allele frequency, p_A , corresponding to the most recent year-classes and p_A decreasing with older year-class. Year-classes from 1995 to 2002 were out of H.W. equilibrium ($X^2 > 3.84$ for such year-classes), showing heterozygous deficiency and apparent Wahlund effect (Table 5). *Pan* I genotype was significantly associated with sex ($X^2 = 8.82$, $df = 2$, and $p = 0.01213$; Table 6). There was a higher overall p_A for females (0.46) than for males (0.43). Both, females and males, were out of H.W. equilibrium ($X^2 > 3.84$ for both sexes), showing a heterozygous deficiency and apparent Wahlund effect (Table 6). For the fall surveys, there were comparable associations between *Pan* I genotype and age, *Pan* I genotype and year-class, and *Pan* I genotype and sex (data not shown).

Table 4: Observed and expected *Pan* I genotypic and genic frequencies and X^2 statistic by age for spring surveys 2005, 2006, and 2007.

Age	Observed			Sum	p_A	Expected			X^2
	AA	AB	BB			AAexp	ABexp	BBexp	
2	0	1	1	2	0.25	0.12	0.75	1.12	0.22
3	24	13	3	40	0.76	23.26	14.49	2.26	0.42
4	63	37	12	112	0.73	59.31	44.39	8.31	3.10
5	169	135	53	357	0.66	156.67	159.65	40.67	8.51
6	289	380	234	903	0.53	254.09	449.83	199.09	21.76
7	292	576	428	1296	0.45	259.57	640.86	395.57	13.28
8	217	497	629	1343	0.35	161.35	608.30	573.35	44.96
9	109	224	299	632	0.35	77.28	287.44	267.28	30.79
10	74	118	131	323	0.41	54.76	156.47	111.76	19.53
11	37	56	38	131	0.50	32.25	65.50	33.25	2.75
12	14	25	23	62	0.43	11.33	30.35	20.33	1.92
13	5	10	10	25	0.40	4.00	12.00	9.00	0.69
14	2	3	2	7	0.50	1.75	3.50	1.75	0.14
15	0	2	2	4	0.25	0.25	1.50	2.25	0.44
16	0	1	0	1	0.50	0.25	0.50	0.25	1.00
17	1	1	0	2	0.75	1.12	0.75	0.12	0.22
18	2	1	0	3	0.83	2.08	0.83	0.08	0.12
19	0	1	0	1	0.50	0.25	0.50	0.25	1.00
Sum	1298	2081	1865	5244	0.45	1042.83	2591.35	1609.83	203.40

Table 5: Observed and expected *Pan* I genotypic and genic frequencies and X^2 statistic by year-class for spring surveys 2005, 2006, and 2007.

Year class	Observed			Sum	p_A	Expected			X^2
	AA	AB	BB			AAexp	ABexp	BBexp	
1987	2	1	0	3	0.83	2.08	0.83	0.08	0.12
1989	0	2	0	2	0.50	0.50	1.00	0.50	2.00
1990	1	1	1	3	0.50	0.75	1.50	0.75	0.33
1991	0	1	0	1	0.50	0.25	0.50	0.25	1.00
1992	4	8	4	16	0.50	4.00	8.00	4.00	0.00
1993	5	8	10	23	0.39	3.52	10.96	8.52	1.67
1994	12	39	19	70	0.45	14.18	34.65	21.18	1.10
1995	74	65	30	169	0.63	67.11	78.77	23.11	5.17
1996	38	78	103	219	0.35	27.07	99.85	92.07	10.49
1997	136	262	341	739	0.36	96.47	341.07	301.47	39.71
1998	216	434	459	1109	0.39	169.06	527.88	412.06	35.07
1999	291	583	539	1413	0.41	240.13	684.74	488.13	31.19
2000	302	446	278	1026	0.51	268.64	512.72	244.64	17.37
2001	105	84	42	231	0.64	93.55	106.91	30.55	10.61
2002	75	44	29	148	0.66	63.57	66.85	17.57	17.29
2003	31	20	8	59	0.69	28.49	25.02	5.49	2.37
2004	6	5	2	13	0.65	5.56	5.88	1.56	0.29
Sum	1298	2081	1865	5244	0.45	1042.83	2591.35	1609.83	203.40

Table 6: Observed and expected *Pan* I genotypic and genic frequencies and X^2 statistic by sex for spring surveys 2005, 2006, and 2007.

Sex	Observed			Sum	p_A	Expected			X^2
	<i>AA</i>	<i>AB</i>	<i>BB</i>			<i>AAexp</i>	<i>ABexp</i>	<i>BBexp</i>	
Male	636	1126	1028	2790	0.43	515.27	1367.46	907.27	86.99
Female	679	1012	897	2588	0.46	542.59	1284.82	760.59	116.69
Sum	1315	2138	1925	5378	0.44	1056.80	2654.41	1666.80	203.55

***Pan* I genotype and length at age (spring surveys): Multiple ANOVA's**

Pan I genotype was significantly associated to length at particular ages. There were differences among the *Pan* I genotypes in length at age (age dependent length). The three *Pan* I genotypes (*Pan* I^{AA}, *Pan* I^{AB}, *Pan* I^{BB}) showed a different mean length when compared at the same ages (Figure 7). The degree of the differences in length among *Pan* I genotypes varied for different ages, thus showing an interaction between genotype and age (Figure 7). However, the differences in length among genotypes was significant after a sequential Bonferroni correction only at ages 7 through 10 years (Table 7). For these ages, the length at age was lowest for the *Pan* I^{AA} genotype, intermediate for *Pan* I^{BB}, and highest for *Pan* I^{AB} (Figure 7). There were comparable differences among *Pan* I genotypes in length at age for the fall surveys (Figure 3 in the Appendix). Such differences were significant at ages 1, 3, 5, 6, and 10 (Table 1 in the Appendix). For these ages, the trend was that the length at age was lowest for the *Pan* I^{BB} genotype, intermediate for *Pan* I^{AB}, and highest for *Pan* I^{AA} (Figure 3 in the Appendix).

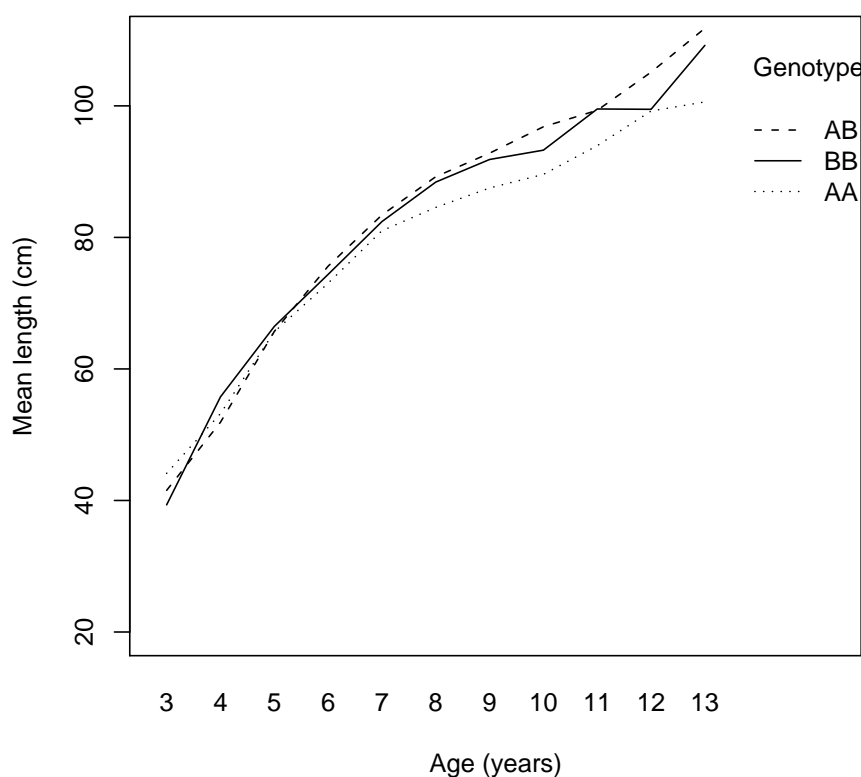


Figure 7: Interaction plot of mean length on genotype and age, for spring surveys 2005, 2006, and 2007.

Table 7: ANOVAs (one-way) of length on genotype at different ages, for spring surveys 2005, 2006, and 2007. SS is sum of squares, MS is means square, F is ratio of MS among genotypes and MS residuals, df is degrees of freedom, p is probability. SL is significance level of difference in length among genotypes; sequential Bonferroni correction was applied.

Age	Source of variance	df	SS	MS	F	p	SL
3	Genotype	2	98.25	49.13	0.75	0.4807	0.0250
	Residuals	37	2432.52	65.74			
4	Genotype	2	138.64	69.32	0.80	0.4527	0.0167
	Residuals	109	9464.14	86.83			
5	Genotype	2	28.19	14.10	0.10	0.9086	0.0500
	Residuals	354	52024.76	146.96			
6	Genotype	2	1106.35	553.18	4.50	0.0114	0.0071
	Residuals	900	110700.56	123.00			
7	Genotype	2	1127.05	563.52	5.11	0.0062*	0.0063
	Residuals	1293	142592.87	110.28			
8	Genotype	2	3415.73	1707.86	19.59	< 0.0001*	0.0045
	Residuals	1340	116827.19	87.18			
9	Genotype	2	2148.82	1074.41	10.71	< 0.0001*	0.0050
	Residuals	629	63078.90	100.28			
10	Genotype	2	2430.00	1215.00	8.56	0.0002*	0.0056
	Residuals	320	45442.93	142.01			
11	Genotype	2	791.95	395.97	2.69	0.0721	0.0100
	Residuals	128	18875.58	147.47			
12	Genotype	2	492.44	246.22	3.85	0.0268	0.0083
	Residuals	59	3770.74	63.91			
13	Genotype	2	425.84	212.92	2.00	0.1588	0.0125
	Residuals	22	2338.40	106.29			

*Statistically significant after Bonferroni correction

DNA sequence variation

Quality of sequences

All sequences utilized in this study exhibit high quality. In general, the quality values were above 40. In this context, an error is a wrongly called base (i.e. a misidentified base). A quality value of 10 meant 1 error in ten, a quality value of 20 meant 1 error in one hundred, and so forth. The sequence quality dropped to almost nil at the very beginning and end of each sequence. Otherwise, the quality values very rarely dropped below 40. A graphical representation of the quality of the sequences can be seen in Figure 4, and in Figure 5, 6, 7, 8, 9, 10, 11 in the Appendix. *Pan I* sequences were a total of 57, 51 *Pan I^A* sequences (3 from the same clone) and 6 *Pan I^B* sequences.

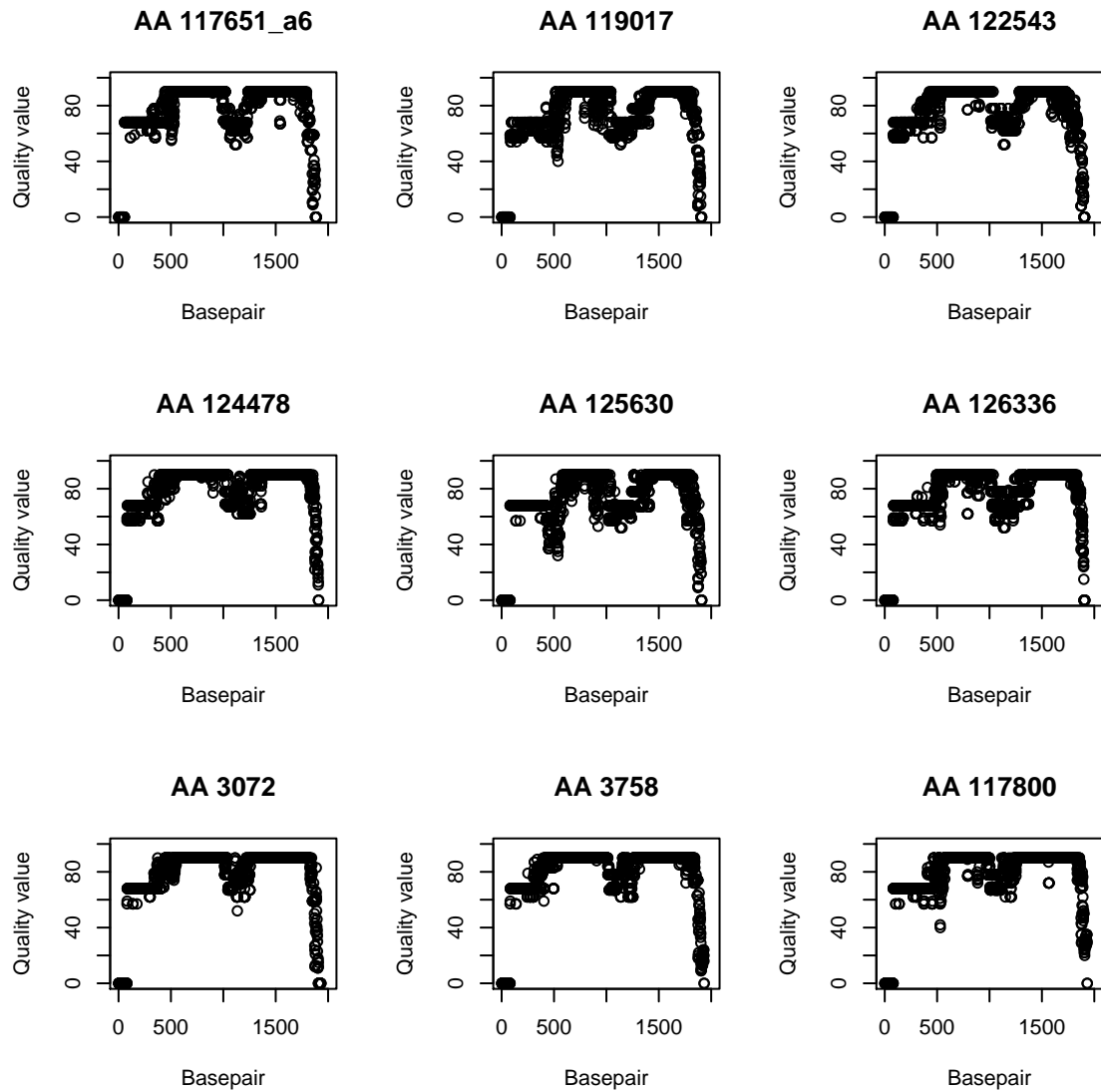


Figure 8: Sequence quality of *Pan I^A* alleles. The quality value of the sequence of each clone is plotted against the pertinent base pair. Above each panel is the genotype and the numeric name of the clone. A quality value of 10 means 1 error in ten to the 1.0 power, a quality value of 20 means 1 error in ten to the 2.0 power, and so forth. Samples are from spring surveys 2005, 2006, and 2007.

Nucleotide polymorphism

Measures of nucleotide polymorphism

Nucleotide polymorphism was largely affected by singletons, some of which doubtless were the product of polymerase errors. Taq polymerase is known to induce DNA replication errors (KEOHAVONG and THILLY, 1989). To study this I sequenced three clones from the same individual (1832 bp each). They differed by seven sites. Thus, the estimated polymerase error rate was $7/(3 \times 1832)$ or 1.27×10^{-3} average errors/site. Due to the DNA polymerase errors detected, all subsequent DNA sequence analysis was conducted with and without the consideration of singletons. I found, for *Pan I^A* and *Pan I^B* sequences pooled together ($n = 55$), that most of the measures of nucleotide polymorphism were considerably higher when singletons were considered than when they were not (Table 8). Nucleotide diversity (π), the average number of nucleotide differences (k), and the number of haplotypes (h) were roughly doubled when singletons were considered than when they were excluded. In the case of the number of segregating sites (S) and θ (per site) from S (Watterson's θ), the values were five-fold and four-fold, respectively, when singletons were considered. Haplotype diversity (Hd) was also different, but not considerably different.

Nucleotide polymorphism was similar between allelic types when considering nucleotide differences, but dissimilar when considering segregating sites. For the comparison between *Pan I^A* and *Pan I^B* including singletons (Table 9, upper panel), π and k were similar between both allelic types. However, S and θ were 8-fold and 4-fold larger for *Pan I^A* than for *Pan I^B*, respectively. Thus, there was an excess of segregating sites in *Pan I^A* relative to *Pan I^B* sequences that surpassed a correction factor for number of sequences. This was due to the difference in number of sequences of each allelic type. Hd was similar between *Pan I^A* and *Pan I^B*; however, h was 7-fold larger for *Pan I^A* than for *Pan I^B*. When contrasting *Pan I^A* and *Pan I^B* alleles (Table 9) in terms of nucleotide polymorphism, there was a considerably higher number of sequences (n) of *Pan I^A* than of *Pan I^B*. For the comparison between *Pan I^A* and *Pan I^B* excluding singletons (Table 9, lower panel), all the measures of nucleotide polymorphism utilized were considerably larger for *Pan I^A*, since for *Pan I^B* they were all null (except for h , which was 1). Therefore, all the variation in *Pan I^B* sequences was due to singletons. All the measures of nucleotide polymorphism used were, without exception, larger (for both allelic types) when singletons were included (Table 9, upper panel) than when singletons were excluded (Table 9, lower panel).

Sequence variation was higher between rather than within allelic types, and there were several amino acid differences between the *Pan I^A* and *Pan I^B* alleles. When singletons

were included, the average number of nucleotide differences k between any $Pan I^A$ and $Pan I^B$ sequences (Table 8, upper line) was larger than that observed either between $Pan I^A$ sequences only or $Pan I^B$ sequences only (Table 9, upper panel). This was consistent with results excluding singletons (Table 8, lower line; Table 9, lower panel). In a summary of segregating sites in $Pan I$ gene excluding singletons (Table 2 in the Appendix), there were a total of 43 variable sites, of which 18 were fixed between alleles, plus a group of indels (labeled ∇ 1 through ∇ 7, following the notation of POGSON, 2001). Indels ∇ 2 and ∇ 3 were fixed between alleles in this data set, but in data from POGSON (2001) ∇ 2 was not fixed. The 18 nucleotide polymorphisms fixed between alleles were at base numbers 76, 248, 255, 431, 560, 636, 736, 738, 745, 746, 790, 900, 1053, 1131, 1144, 1398, 1571, and 1641. Of these, there were 6 replacement mutations fixed between alleles (cf. POGSON, 2001, p. 323) that led to 4 amino acid differences (cf. POGSON and MESA, 2004, p. 71). They were all located in the intravesicular 1 (IV1) region of the gene. These 6 replacement mutations fixed between allelic types were at base number 560, 736, 738, 745, and 746. The 4 amino acid (in one letter code) differences between $Pan I^A$ and $Pan I^B$ were E and V, Q and N, T and D, and C and S. ∇ 1 was a CA indel, ∇ 2 was a 12 bp indel of two forms GCATAGTAAAAA or GCATAGTAGAAA, ∇ 3 was a 6 bp indel of TTTTGT, ∇ 4 was a variant of ∇ 3 and therefore it was not used, ∇ 5 was an indel of A, ∇ 6 was an indel of T or C. ∇ 7 identifies several different single base pair indels in different locations. From the 3' and 5' ends of all the sequences 10 bp and 9 bp were respectively cut out because sequence quality was deficient. Also, I found different indels than POGSON (2001). Therefore, the base numbers of the variable sites and indels are not the same as used by POGSON (2001).

Table 8: Nucleotide polymorphism in $Pan I$, considering $Pan I^A$ and $Pan I^B$ together. Statistics are shown for data including and excluding singletons. n is the number of sequences used, S is the number of segregating sites, k is the average number of nucleotide differences, π is nucleotide diversity and θ is Theta (per site) from S (Watterson's θ), h is the number of haplotypes and Hd is haplotype diversity.

<i>Pan I</i> , encompassing <i>Pan I^A</i> and <i>Pan I^B</i> alleles							
Singletons	n	S	π	θ	k	h	Hd
Included	55	207	0.0068	0.0250	12.33	51	0.997
Excluded	55	43	0.0034	0.0052	6.09	23	0.900

Table 9: Nucleotide polymorphism in *Pan I^A* and *Pan I^B*. The upper panel describes statistics for data including singletons, and the lower panel excluding singletons. n is the number of sequences used, S is the number of segregating sites, k is the average number of nucleotide differences, π is nucleotide diversity and θ is Theta (per site) from S (Theta-W), h is number of haplotypes, and Hd is haplotype diversity.

Singletons	Allelic Type	n	S	π	θ	k	h	Hd
Included	<i>Pan I^A</i> alleles	49	172	0.0049	0.0211	9.03	45	0.997
Included	<i>Pan I^B</i> alleles	6	22	0.0040	0.0053	7.33	6	1.000
Excluded	<i>Pan I^A</i> alleles	49	26	0.0016	0.0032	2.89	23	0.908
Excluded	<i>Pan I^B</i> alleles	6	0	0.00	0.00	0.00	1	0.000

Distribution of nucleotide polymorphism across the *Pan I* gene region in relation to signs of selection

A sliding window analysis of nucleotide diversity π throughout the *Pan I* gene revealed peaks of π when allelic types were analyzed together and separately. Throughout the *Pan I* region, including both *Pan I^A* and *Pan I^B* alleles, there were two distinctive peaks of nucleotide diversity (π): One in the intravesicular 1 region (IV1) and the other in the third intron (Intron 3; Figure 9, and Figure 14 in the Appendix). The peak in IV1 was not evident when intra-allelic sequence variation was examined for each allelic type (Figures 10 and 11; also Figure 15 in the Appendix). Thus, the π peak in IV1 represented a region of high sequence variation between allelic types. The π peak in Intron 3 was also evident when intra-allelic sequence variation was examined per allelic type. Intra-allelic sequence variation indicated segregation of alternative allelic variants within each allelic type. Throughout each one of the allelic types, *Pan I^A* and *Pan I^B* separately, there was a single distinctive peak of π at Intron 3 (Figures 10 and 11; also Figure 15 in the Appendix); i.e., there were alternative forms of each allele due to variation in Intron 3. This was observed for data including and excluding singletons. However, there was no observable distribution of nucleotide polymorphism in *Pan I^B* without singletons, since all the variation in *Pan I^B* was due to singletons.

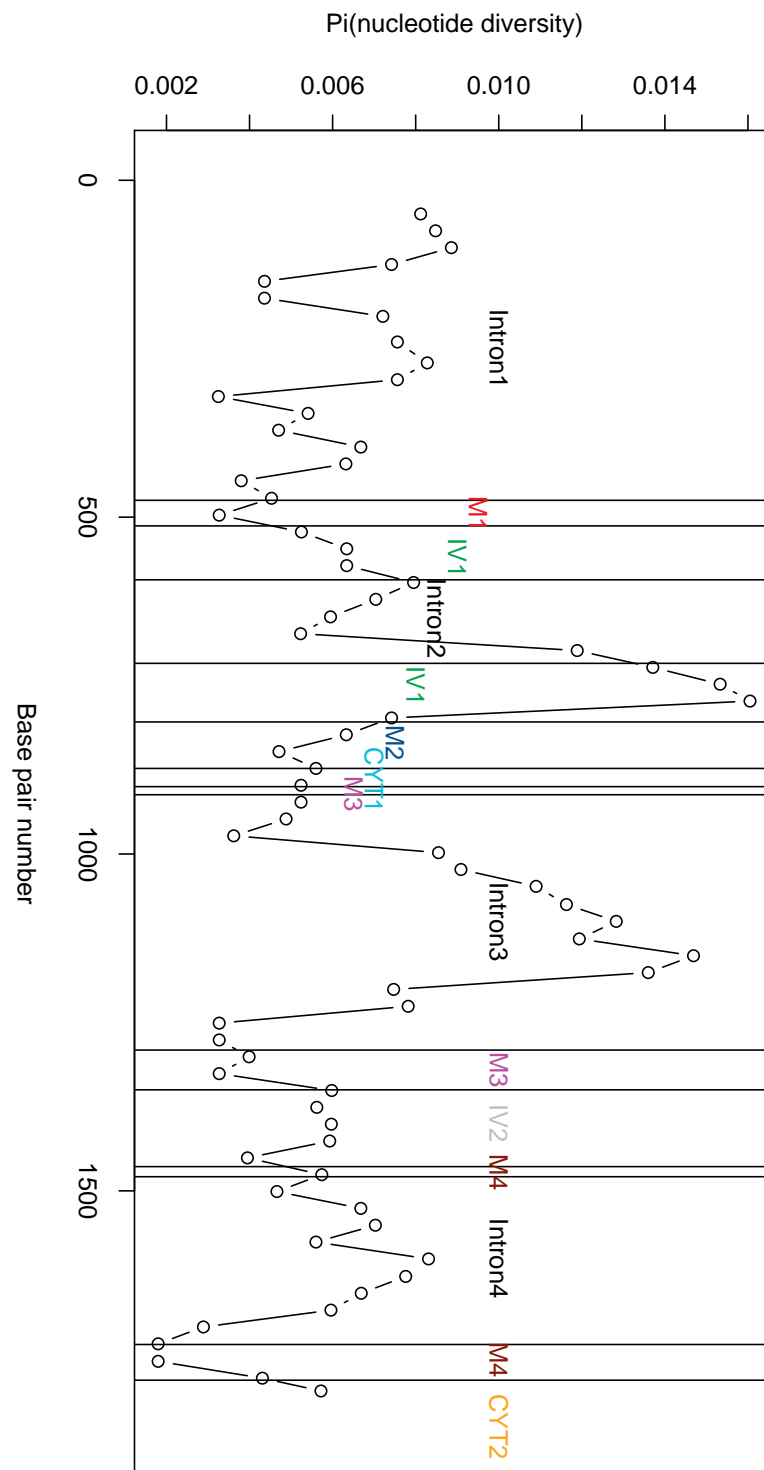


Figure 9: Sliding window analysis (including singletons) of nucleotide diversity (π) throughout the *Pan I* gene region in *Pan I^A* and *Pan I^B* alleles. Window size is 100 bp and step size is 25 bp. M1 through M4 are membrane spanning domains, CYT1 and CYT2 are cytoplasmic tails domains, IV1 and IV2 are intravesicular domains. Samples are from spring surveys 2005, 2006, and 2007.

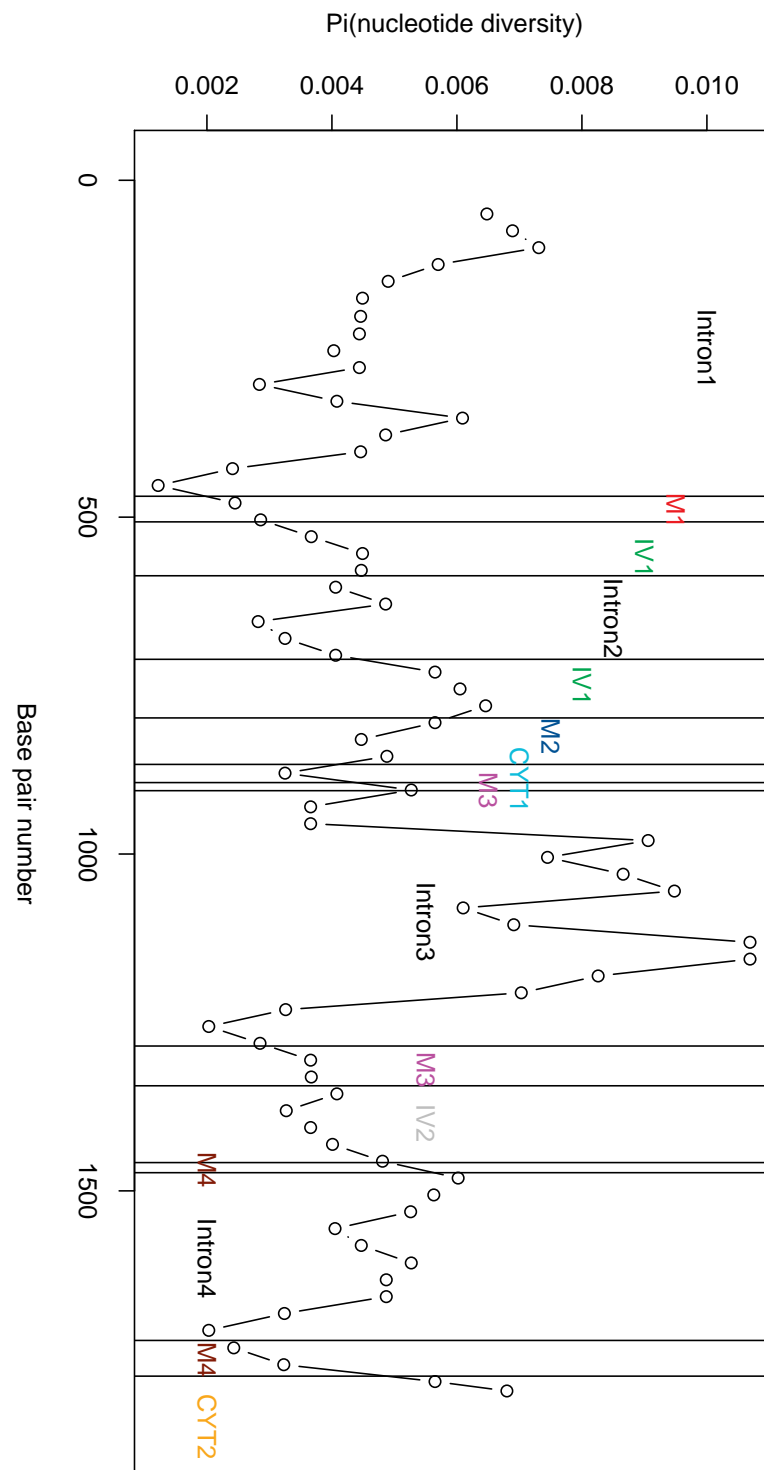


Figure 10: Sliding window analysis (including singletons) of nucleotide diversity (π) throughout the *Pan I* gene region in *Pan I^A* alleles. Window size is 100 bp and step size is 25 bp. M1 through M4 are membrane spanning domains, CYT1 and CYT2 are cytoplasmic tails domains, IV1 and IV2 are intravesicular domains. Samples are from spring surveys 2005, 2006, and 2007.

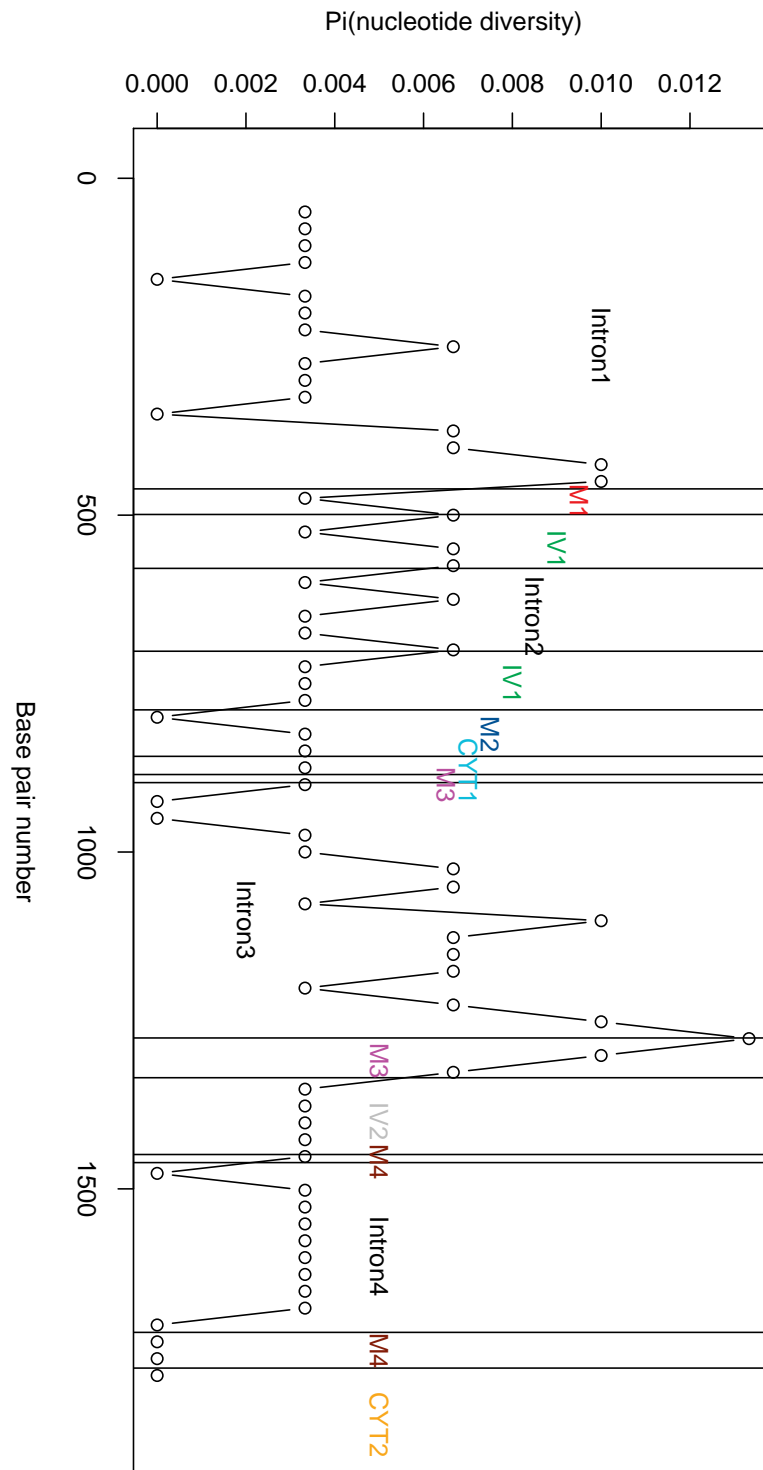


Figure 11: Sliding window analysis (including singletons) of nucleotide diversity (π) throughout the *Pan I* gene region in *Pan I^B* alleles. Window size is 100 bp and step size is 25 bp. M1 through M4 are membrane spanning domains, CYT1 and CYT2 are cytoplasmic tails domains, IV1 and IV2 are intravesicular domains. Samples are from spring surveys 2005, 2006, and 2007.

Tests of neutrality of DNA sequence variation

Overall, tests of neutrality of DNA sequence variation indicated that the *Pan I* locus deviated from neutrality when singletons were included, but not when they were excluded. Tajima's D statistic was significant for *Pan I*^A and *Pan I*^B alleles, separately and pooled, when singletons were included (Table 10). This indicated in all these instances that the *Pan I* locus deviated from neutrality. However, when singletons were excluded, Tajima's D was not significant and it could not be calculated for *Pan I*^B (all the polymorphisms found in *Pan I*^B came from singletons and thus I could not compute Tajima's test). The sign of Tajima's D was negative with and without singletons in all instances obtained (Table 10), indicating that there was an excess of segregating sites in relation to the number of nucleotide differences. Fu and Li's D^* statistic was significant for both allelic types, separately and pooled, including and excluding singletons (except for *Pan I*^B as noted above) (Table 11). Thus, it indicated that the *Pan I* locus deviated from neutrality. Also Fu and Li's F^* was significant for both allelic types, separately and pooled, when singletons were included (Table 11, upper panel), but was not significant when singletons were excluded (Table 11, lower panel). D^* and F^* showed negative signs when singletons were included. Thus, there was an excess of the number of singletons (η_s) with respect to the total number of mutations (η) and the average number of nucleotide differences between pairs of sequences (k) (Table 11, upper panel). This condition was reversed when singletons were excluded (Table 11, lower panel), except where D^* and F^* were not available because there were no polymorphisms other than singletons (which was the case of *Pan I*^B). The McDonald-Kreitman test was not significant in both cases, when including singletons ($G = 0.57$, $p = 0.45$, Table 12, upper panel) and when excluding singletons ($G = 0.29$, $p = 0.59$, Table 12, lower panel). Thus, this test did not indicate that the *Pan I* locus deviated from neutrality.

Table 10: Tajima's D test of Neutrality in *Pan* I. The upper panel describes statistics for data including singletons, and the lower panel excluding singletons. D is Tajima's D -value, n is the number of sequences of each allele, p is probability, NA is not available.

Singletons	Allelic type	D	p	n
Included	<i>Pan</i> I ^A	-2.790	< 0.001*	49
Included	<i>Pan</i> I ^B	-1.498	< 0.050*	6
Included	<i>Pan</i> I ^A and <i>Pan</i> I ^B	-2.627	< 0.001*	55
Excluded	<i>Pan</i> I ^A	-1.662	> 0.05	49
Excluded	<i>Pan</i> I ^B	NA	NA	6
Excluded	<i>Pan</i> I ^A and <i>Pan</i> I ^B	-1.187	> 0.10	55

*Statistically significant

Table 11: Fu and Li's tests of neutrality at *Pan* I. The upper panel describes statistics for data including singletons, and the lower panel excluding singletons. D^* and F^* are test statistics with no outgroup based on the genealogy of a set of sequences (Fu and Li, 1993). n is the number of sequences, p is probability, NA is not available (there were no polymorphisms and it was not possible to compute Fu and Li's test).

Singletons	Allelic Type	D^*	p	F^*	p	n
Included	<i>Pan</i> I ^A	-5.92	< 0.02*	-5.67	< 0.02*	49
Included	<i>Pan</i> I ^B	-1.54	< 0.05*	1.67	< 0.05*	6
Included	<i>Pan</i> I ^A and <i>Pan</i> I ^B	-5.72	< 0.02*	-5.42	< 0.02*	55
Excluded	<i>Pan</i> I ^A	1.78	< 0.02*	0.68	> 0.10	49
Excluded	<i>Pan</i> I ^B	NA	NA	NA	NA	6
Excluded	<i>Pan</i> I ^A and <i>Pan</i> I ^B	1.96	< 0.02*	0.96	> 0.10	55

*Statistically significant

Table 12: McDonald-Kreitman test for *Pan* I. Variation within and between *Pan* I^A and *Pan* I^B alleles. The upper panel describes substitutions in sequences including singletons, and the lower panel excluding singletons.

Singletons		Replacement	Silent
Included	Fixed between alleles	6	2
Included	Polymorphic within alleles	37	23
Excluded	Fixed between alleles	6	2
Excluded	Polymorphic within alleles	5	3

Gene genealogies, geography and depth

Clustering due to evolutionary relation was not correlated with either clustering by geographic area or by depth. When I considered only *Pan I^A*, the allelic type with the highest number of sequences, the topology of the gene genealogy did not show any aggregations by region or depth; i.e the clusters (albeit diffuse) produced by evolutionary relations were not at the same time clusters of geographic areas or depths (Figure 12 in the Appendix). For the same gene genealogy but excluding singletons (Figure 13 in the Appendix), the topology was less branched and some clusters were quite distinctive; however, I still did not detect correlation of genealogy with clustering either by geographical region or depth.

Genetic differentiation, considering geographic locality and depth

Globally, there was no genetic differentiation among subpopulations defined by geographic area or by depth. K_{ST} , the effect of subpopulations compared to the total population in terms of the average number of nucleotide differences between sequences K (HUDSON *et al.*, 1992), was not significant when subpopulations were defined either by geographic region (Table 13) or by depth (Table 14). The results were consistent when singletons were included (Tables 13 and 14, upper panels) and excluded (Tables 13 and 14, lower panels).

Table 13: Genetic differentiation analysis of *Pan I^A* from different MetaCod divisions (localities) in Iceland. The upper panel describes statistics for data including singletons, and the lower panel excluding singletons. K_{ST} is a nucleotide sequence-based test statistic based on K , the average number of nucleotide differences between sequences. n is the number of sequences used, S is the number of segregating sites.

	<i>Pan I^A</i>					
	n	S	K	K_{ST}	p	Singletons
MCdiv 1,2,3	11	50	10.98	0.004	0.19	Included
MCdiv 4,5	14	58	8.99			
MCdiv 7,8	14	48	7.91			
MCdiv9	10	40	8.31			
MCdiv 1,2,3	11	16	4.80	0.016	0.16	Excluded
MCdiv 4,5	14	14	2.70			
MCdiv 7,8	14	9	2.11			
MCdiv9	10	8	1.91			

Table 14: Genetic differentiation analysis of *Pan I^A* from different depth levels in Iceland. The upper panel describes statistics for data including singletons, and the lower panel excluding singletons. Depth levels are (intervals in m.) 1: 0–25; 2: 25–50; 3: 50–75; 4: 75–100; 5: 100–125; 6: 125–150; 7: 150–175; 8: 275–300. K_{ST} is a nucleotide sequence-based test statistic based on K , the average number of nucleotide differences between sequences. n is the number of sequences used, S is the number of segregating sites.

	<i>Pan I^A</i>					
	n	S	K	K_{ST}	p	Singletons
Depth level 1	7	33	10.05			
Depth level 2	11	36	7.18			
Depth level 3	8	42	10.86			
Depth level 4	6	21	7.20	−0.008	0.85	Included
Depth level 5	5	17	6.80			
Depth level 6	8	48	13.07			
Depth level 7	4	13	7.00			
Depth level 1	7	7	2.38			
Depth level 2	11	16	3.42			
Depth level 3	8	16	4.36			
Depth level 4	6	4	1.53	−0.029	0.89	Excluded
Depth level 5	5	6	2.40			
Depth level 6	8	10	3.57			
Depth level 7	4	3	1.67			

DNA divergence between populations

DNA divergence between *Pan I^A* and *Pan I^B* alleles

Pan I^A and *Pan I^B* alleles were divergent in accordance with previous findings, e.g. by POGSON and MESA (2004). The estimation of D_a , the number of net nucleotide substitutions per site between populations (here between allelic types), had a standard error SE_{D_a} such that a 95 % confidence interval of the test statistic D_a would not include 0 (no divergence; Table 15). The results were consistent whether singletons were included or excluded (Table 15).

Table 15: Extent of DNA divergence between *Pan I^A* vs *Pan I^B* alleles. Statistics are shown for data including and excluding singletons. D_{xy} with Jukes and Cantor (JC) is the average number of nucleotide substitutions per site between populations, D_a with Jukes and Cantor (JC) is the number of net nucleotide substitutions per site between populations, SE_{D_a} is the Standard Error of D_a (JC), n is the number of sequences of each allele.

Singletons	<i>Pan I^A</i> vs <i>Pan I^B</i>			<i>Pan I^A</i>	<i>Pan I^B</i>
	D_{xy} (JC)	D_a (JC)	SE_{D_a}	n	n
Included	0.01463	0.01013	0.00126	49	6
Excluded	0.01086	0.01009	0.00087	49	6

DNA divergence between populations considering geographic locality and depth

Pairwise divergence between populations was not observed whether defined by geographic area or by depth. The estimation of D_a for all pairwise comparisons between populations defined by geographic regions (Table 16) or by depth (Table 17) had SE_{D_a} such that a 95 % interval of D_a would at all times harbor 0. When singletons were excluded, the same results were obtained (Tables 3 and 4 in the Appendix).

Table 16: Extent of DNA divergence between populations at different MetaCod divisions (localities), for *Pan I^A*. MCdiv is MetaCod division, n is the number of sequences at each MetaCod division, D_a with Jukes and Cantor (JC) is the number of net nucleotide substitutions per site between populations, SE_{D_a} is the Standard Error of D_a (JC); D_a (JC) and SE_{D_a} appear above and below the diagonal, respectively.

<i>Pan I^A</i>					
	MCdiv 1,2,3	D_a (JC) $\times 1000$			n
		MCdiv 4,5	MCdiv 7,8	MCdiv 9	
MCdiv 1,2,3		0.10	0.10	0.07	11
MCdiv 4,5	0.87		-0.01	-0.04	14
MCdiv 7,8	0.80	0.66		-0.04	14
MCdiv 9	1.07	0.96	0.91		10
$SE_{D_a} \times 1000$					

Pan I ^A								
	$D_a(\text{JC}) \times 1000$							
	Depth	Depth	Depth	Depth	Depth	Depth	Depth	n
	level 1	level 2	level 3	level 4	level 5,6	level 7	level 8	
Depth level 1		-0.09	-0.01	-0.06	-0.09	-0.01	-0.11	7
Depth level 2	0.92		-0.07	-0.03	-0.13	-0.01	-0.07	11
Depth level 3	1.23	0.98		-0.04	-0.04	-0.01	0.06	8
Depth level 4	1.11	0.85	1.16		-0.11	-0.05	0.01	6
Depth level 5,6	1.31	1.06	1.38	1.22		-0.05	-0.05	5
Depth level 7	1.45	1.18	1.49	1.39	1.58		0.01	8
Depth level 8	1.35	1.07	1.44	1.28	1.44	1.65		4
$\text{SE}_{D_a} \times 1000$								

Discussion

Genotypic variation in relation to biological and environmental factors

The patterns established between *Pan* I and various biological and environmental variables with regard to allelic frequency distribution reflect the effects of selection on *Pan* I. Selection changes allele frequencies (HEDRICK, 2005) and may produce patterns in terms of allelic frequency distribution between *Pan* I and other variables intimately related with selection phenomena. Such patterns were chosen as the means of detecting the effects of selection. Thus, the establishment of several patterns of allelic frequency distribution between *Pan* I and various biological or environmental variables indicates the action of selection on *Pan* I and reveals that depth, age, year-class, sex, and length are variables intimately connected with selection phenomena.

The clear correlation between *Pan* I allele frequency p_A and depth (a proxy for other depth-related environmental variables) is indicative of selective effects. Variation in a number of environmental parameters (temperature, osmotic stress, radiation, etc.) are linked with differences in expression or function of genes and can act as selective agents (HEMMER-HANSEN *et al.*, 2007). Various environmental factors (temperature, salinity, pressure, oxygen content, light, etc.) change with depth (CHILDRESS, 1995). The correlation observed between *Pan* I allele frequency and depth is in agreement with previous findings described by FEVOLDEN and POGSON (1997), PAMPOULIE *et al.* (2006), and SARVAS and FEVOLDEN (2005). The Wahlund effects observed for the spring-spawning grounds at shallow waters could indicate a behavioral response (c.f. PAMPOULIE *et al.*, 2008), i.e. the individuals sampled are an amalgamation of differentiated groups, a fact that may be attributed to the fish going to shallow waters to reproduce. The allele frequency gradient with depth, along with changes in genotypic composition from spring to fall (i.e. heterozygote deficiency vs heterozygote excess) may be due to a combination of behavioral differences as well as differential reproduction and mortality,

namely the action of natural selection.

The confounding of depth and geographic region, along with the correlation between *Pan* I and depth, have profound implications on the nature of population structure of Atlantic cod around Iceland. Genetic differentiation over subpopulations, measured by F_{ST} , was significant in terms of both geographic region and depth, but F_{ST} was larger in terms of depth. Importantly, depth was correlated to geographic area. These findings indicate that the nature of the population structure of Atlantic cod around Iceland determined by F_{ST} at *Pan* I is rather due to depth than to geographic region. Moreover, as depth reflects effects of selection on *Pan* I, the findings suggest that the nature of the population structure of Atlantic cod around Iceland has roots in the effects of natural selection on *Pan* I rather than on a historical isolation of region-defined subpopulations as suggested by PAMPOULIE *et al.* (2006).

The associations established between *Pan* I and age, year-class, sex, and length, are indicative of effects of selection on *Pan* I. Correlation is not synonymous of causation (SOKAL and ROHLF, 1995). The correlation existing between *Pan* I and other biological variables, therefore, does not imply that *Pan* I necessarily determines such biological traits as longevity (in the case of age and year-class), sex, or body length. Instead, it is through these variables that the effects of selection on *Pan* I become evident.

Pan I^A allele frequency p_A decreased as age increased, and there were higher levels of p_A for the most recent year-classes. In other words, there were more *Pan* I^A types at younger ages, and there were more *Pan* I^B types that reached older ages. An explanation for this could be that the *Pan* I genotype determines the longevity of the organism, but another answer lies in the presence of a selective agent that systematically removes organisms that express a phenotypic or behavioral trait (HEDRICK, 2005) determined by *Pan* I or a locus closely related to it. Since more *Pan* I^{AA} genotypes are removed from the system (i.e. they do not reach older ages), and incidentally there is a higher *Pan* I^A allelic frequency p_A at shallow depths, the magnitude of the selective agent is higher at more shallow waters. Regarding sex, females and males were out of Hardy Weinberg equilibrium, showing heterozygous deficiency.

Length is a proxy for body size, a trait upon which natural selection may act (RIDLEY, 2003). Length is a phenotypic trait which has been previously analyzed in relation to *Pan* I during studies of growth in Atlantic cod (JÓNSDÓTTIR *et al.*, 2008). The differences in length among different *Pan* I genotypes were significant for particular ages. An explanation could be that growth (length at age) (JÓNSDÓTTIR *et al.*, 2008) is determined by *Pan* I. However, JÓNSDÓTTIR *et al.* (2008) concluded that the relation of growth with *Pan* I is rather complex and influenced by other factors, like size-selective fishing and food supply. In relation to that, an

alternative explanation is the presence of a selective agent which is size-selective and that systematically removes organisms that express a phenotypic trait or behavior determined by *Pan I* genotype or a closely associated locus. In other words, comparatively more *Pan I^{AA}* genotypes than *Pan I^{AB}* or *Pan I^{BB}* are being removed from the system upon reaching a certain length, and the smaller organisms of genotype *Pan I^{AA}* remain, which in average are then smaller than *Pan I^{AB}* or *Pan I^{BB}* genotypes.

(PÁLSSON and THORSTEINSSON, 2003) have described behavioral types with regards to depth preference and migrations for Atlantic cod which are related to *Pan I* genotype. Although PAMPOULIE *et al.* (2008) conclude that more research is needed to fully assess the potential relationship between the *Pan I* locus and the behavioral types described, their results revealed that *Pan I^{AA}* genotypes are likely to display a shallow water preference, while *Pan I^{BB}* prefer deeper waters and *Pan I^{AB}* shows an intermediate behavior. As the selective-size removal of individuals is strongest in the *Pan I^{AA}* genotypes, which prefer shallow waters, then the agent of selection is strongest at shallow waters. The selective pressures that create the correlations between *Pan I* and age, year-class, and length could be instigated by naturally occurring selective agents or by man-mediated selective agents, such as size-selective fisheries. Fishing pressures and practices have been described as selective agents whose effects can be seen in age-related variables (e.g. age and size at maturity) and in growth (length at age) (SWAIN *et al.*, 2007).

DNA sequence variation

Taq polymerase is known to induce DNA replication errors (KEOHAVONG and THILLY, 1989). Some of the singletons in this study were doubtless the product of such Taq polymerase-induced errors. The introduction of singletons due to polymerase errors amounts to the introduction of rare alleles (nucleomorphs). Measures of diversity based on the number of segregating sites S count all variable sites equally and thus may be strongly influenced by rare alleles. However, for nucleotide diversity π the frequency of alleles is considered and is not influenced much by rare alleles (TAJIMA, 1983). Measurements of nucleotide polymorphism without singletons cannot be used alone either, because in those instances true singletons have been withdrawn from the data. All results were presented with and without singletons and that gives a good perspective of nucleotide polymorphism measurements in the *Pan I* gene, but special attention must be given to π and k . Therefore, the issue of polymerase error-induced singletons biases the estimation of number of segregating sites and complicates the application and interpretation of tests of neutrality, since the latter make use of the number of segregating

sites or the number of singletons. However, this issue of singletons does not impair the utility of signal or pattern based sequence analysis like the nucleotide polymorphism distribution analysis by sliding window approach (KREITMAN and HUDSON, 1991), or gene genealogies coupled to geographic area and depth. These analysis are not impaired because the polymerase errors affect all sequences in the same proportion. Thus they simply represent noise. In the case of the sliding window analysis of π , such noise makes it more difficult to discern signals of selection (in this case peaks of π). If a signal is nevertheless detected it very likely represents a true signal. In the case of the genetic trees built with *Pan I^A* allele sequences and coupled to depth and geographic area, clustering is observed in the presence of noise. However, the question arises whether random noise can create clustering. Theoretically this would be possible. Random changes hitting the same site would be detected as clusters with the phylogenetic methods used. However, in this case the hypothesis that clusters are due to repeated errors is dismissed because (at least part of) the clustering observed in this study conforms to the clustering obtained by POGSON (2001) for sequences of Atlantic cod. Therefore, either geographic location or depth can be overlaid onto the genealogy of *Pan I^A* allele sequences. This approach allows testing a hypothesis of historical division of Atlantic cod populations around Iceland by either geographic region or depth. Under such hypothesis *Pan I^A* alleles from northern Iceland would be more similar to other *Pan I^A* alleles from northern Iceland, and *Pan I^A* alleles from southern Iceland would be more similar to other *Pan I^A* alleles from southern Iceland; in an analogous way, *Pan I^A* alleles from shallow waters would be more similar to other *Pan I^A* alleles from shallow waters, and *Pan I^A* alleles from deeper waters would be more similar to other *Pan I^A* alleles from deeper waters. The fact that there was no concurrent clustering of evolutionary relations and geographic area or depth means that these data do not show evidence of a historical division of Atlantic cod populations around Iceland by either geographic region or depth.

Regarding measurements of nucleotide polymorphism with *Pan I^A* and *Pan I^B* pooled together, the estimate of all the statistics, π , the average number of nucleotide differences k , the number of haplotypes h , haplotype diversity Hd , S , and θ (per site) from S (Watterson's θ), increases with the presence of singletons. The statistics most affected are S and Watterson's θ as already explained. To this respect, even though the statistics based on the number of segregating sites were the most affected, TAJIMA (1983) recommends the use of π and k to measure genetic variation within populations instead of S because the latter is dependent on sample size. This is also the reason why nucleotide polymorphism between *Pan I^A* and *Pan I^B* is so different when measured by S (the sample size of each allelic type was so different,

$n = 55$ for *Pan I^A* and $n = 6$ for *Pan I^B*), but not when measured by π or k in the presence of singletons. Using the perspective from the with-or-without singletons results, I can see that the variation in *Pan I^B* is completely erased by eliminating singletons. Therefore, going back to π and k in the presence of singletons, which are not affected much by rare alleles (TAJIMA, 1983) introduced by polymerase errors, I can deduce that the nucleotide diversity of *Pan I^A* and *Pan I^B* is very similar.

The nucleotide polymorphisms segregating within *Pan I* conform in its majority to those found by POGSON (2001), including indels, amino acid substitutions, fixed and polymorphic sites between the *A* and *B* alleles. The indels denoted by $\nabla 1$ to $\nabla 3$ are the same, although I found more indels than POGSON (2001) and they are denoted by $\nabla 5$ to $\nabla 7$.

The distribution of nucleotide polymorphism across the *Pan I* gene indicated the action of balancing selection maintaining both *Pan I^A* and *Pan I^B* alleles. When analyzing *Pan I^A* and *Pan I^B* together by a sliding window approach (MCDONALD and KREITMAN, 1991), a peak of π in the intravesicular 1 (IV1) region of the *Pan I* gene was revealed. The 6 fixed amino acid substitutions detected between *Pan I^A* and *Pan I^B* were located in the IV1 region. This region of high sequence variation between alleles created by the 6 amino acid substitutions represents a long-lived polymorphism described by POGSON (2001). Such peak at IV1 is consistent with balancing selection maintaining the two allelic types. The results also conform to findings on *Pan I* by CANINO and BENTZEN (2004) in Walleye pollock (*Theragra chalcogramma*). The peak of π on the intron 3, revealed by the same analytical approach on *Pan I^A* and *Pan I^B* separately and together is also indicative of selective effects. It could be due to lack of constraint in the region, or due to balancing selection, but may also be due to a selective sweep such as the one described by POGSON (2001). When an allele sweeping through reaches intermediate frequencies it generates a high π . Namely, in a process of positive selection where allelic variants are coexisting but some are increasing in frequency, there is a point of maximum diversity analogous to a peak of π .

Tests of neutrality indicated that the *Pan I* locus deviated from neutrality when singletons were included, but not when they were excluded. However, no conclusive inferences can be drawn from tests of neutrality applied in this study because of the presence of rare alleles induced by polymerase errors. They can deeply affect the number of segregating sites which is a statistic used by the tests. A with-or-without singletons approach is not very helpful in this instance because the workings of the tests of neutrality consider the number of segregating sites: Tajima's test (TAJIMA, 1983) makes use of θ_S , an estimate based on number of segregating sites which ignores the frequency of sequences, while the McDonald Kreitman's

test (MCDONALD and KREITMAN, 1991) makes use of the number of fixed and polymorphic sites which thus also involves the number of segregating sites. Fu and Li's tests (FU and LI, 1993) make use of the total number of singletons for both statistics D^* and F^* . Moreover, D^* uses the total number of mutations, and as rare mutations induce new segregating sites these measures are related. Thus, these tests are very sensitive to rare alleles and definite conclusions cannot be made using the outcome attained through tests of neutrality applied here.

Nucleotide based analysis do not show evidence of population structure in Atlantic cod around Iceland. The clusters in the gene genealogy are not related to depth or geographic region. This was so whether the clustering was diffuse or well defined due to presence or absence of singletons. This means that any aggregation of genetic similarity is not coincident with grouping defined by geographic area or depth. This inference is in agreement with a non-significant K_{ST} (HUDSON *et al.*, 1992) for both geographic region and depth. K_{ST} was used to measure genetic differentiation among subpopulations, in terms of average number of nucleotide differences between sequences. Also in agreement with such outcome is the lack of evidence of population divergence by pairwise comparisons of subpopulations in terms of D_a , the number of net nucleotide substitutions per site between populations defined by geographic area or by depth. The overall outcome of these nucleotide based analysis is in disagreement with the allelic frequency based results of hierarchical F statistics. However, the sample number used in the DNA sequence based analysis was far less than the sample number used in the hierarchical F statistics analysis.

With regards to future work, technical difficulties generated by polymerase-induced errors must be addressed, the number of *Pan* I sequencing primers increased, and knowledge of the physiological function of Pantophysin advanced. A feasible approach to resolving the matter of polymerase induced errors is to sequence several clones from each of the individuals that compose the study sample. By comparing clones from the same source, spurious polymorphisms can be detected and thus dismissed. It is advisable to increase the number of sequencing primers for building contigs of higher sequence density. This would provide improved coverage of the gene and raise sequence quality in areas of the gene that proved difficult to be read. In turn, more sequences could be considered as less of them would have to be dismissed due to insufficient sequence quality. Advancing knowledge on the physiological function of Pantophysin can be approached by a conjunction of venues: Garden experiments controlling *Pan* I genotypes and environmental conditions (e.g. depth-related variables such as temperature, pressure, oxygen content, salinity, light, etc.), differential gene expression analysis, immunological experiments of tissue distribution of Pantophysin, identification and characterization

of control regions of the *Pan* I locus, and studies of linkage disequilibrium between protein variants of the gene and control regions.

Conclusion

Selection is acting at *Pan I* as evidenced by several patterns of allelic frequency p_A distribution between the *Pan I* locus and other biological or environmental variables. Evidence indicates that the nature of the population structure of Atlantic cod around Iceland as revealed by *Pan I* derives from depth and has roots in the effects of selection on *Pan I*. DNA sequence variation is also indicative of the effects of selection acting on *Pan I*. Both, the particular mechanisms of the selection process and selective agents are unknown.

Bibliography

- B. ALBERTS, A. JOHNSON, J. LEWIS, M. RAFF, K. ROBERTS, and P. WALTER, 2002. *Molecular Biology of the Cell, Fourth Edition*. Garland Science, New York, New Yor.
- E. ÁRNASON and S. PÁLSSON, 1996. Mitochondrial cytochrome *b* DNA sequence variation of Atlantic cod *Gadus morhua*, from Norway. *Molecular Ecology* **5**: 715–724.
- E. ÁRNASON, S. PÁLSSON, and A. ARASON, 1992. Gene flow and lack of population differentiation in Atlantic cod, *Gadus morhua* L., from Iceland, and comparison of cod from Norway and Newfoundland. *Journal of Fish Biology* **40**: 751–770.
- E. ÁRNASON, P. H. PETERSEN, K. KRISTINSSON, H. SIGURGÍSLASON, and S. PÁLSSON, 2000. Mitochondrial cytochrome *b* DNA sequence variation of Atlantic cod from Iceland and Greenland. *Journal of Fish Biology* **56**: 409–430.
- P. BENTZEN, C. T. TAGGART, D. E. RUZZANTE, and D. COOK, 1996. Microsatellite polymorphism and the population structure of Atlantic cod (*Gadus morhua*) in the Northwest Atlantic. *Canadian Journal of Fisheries and Aquatic Sciences* **53**: 2706–2721.
- A. BERRY and M. KREITMAN, 1993. Molecular analysis of an allozyme cline: alcohol dehydrogenase in *Drosophila melanogaster* on the East Coast of North America. *Genetics* **134**: 869–893.
- M. F. CANINO and P. BENTZEN, 2004. Evidence for positive selection at the Pantophysin (*Pan* I) locus in walleye pollock, *Theragra chalcogramma*. *Molecular Biology and Evolution* **21**: 1391–1400.
- J. J. CHILDRESS, 1995. Are there physiological and biochemical adaptations of metabolism in deep-sea animals? *Trends in Ecology and Evolution* **10**: 30–36.
- R. H. DON, P. T. COX, B. J. WAINWRIGHT, K. BAKER, and J. S. MATTICK, 1991. "Touch-

- down" PCR to circumvent spurious priming during gene amplification. *Nucleic Acids Research* **19**: 1.
- R. C. EDGAR, 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research* **32**: 1792–1797.
- B. EWING, L. HILLIER, M. C. WENDL, and P. GREEN, 1998. Base-calling of automated sequencer traces using Phred. I. Accuracy assessment. *Genome Research* **8**: 175–185.
- S. E. FEVOLDEN and G. H. POGSON, 1997. Genetic divergence at the Synaptophysin (*Syp* I) locus among Norwegian coastal and North-east Arctic populations of Atlantic cod. *Journal of Fish Biology* **51**: 895–908.
- S.E. FEVOLDEN and G. H. POGSON, 1995. Differences in nuclear DNA RFLPs between the Norwegian coastal and the North-east Arctic population of Atlantic cod. In H.R. SKJOLDAL, C. HOPKINS, K.E. ERIKSTAD, and H.P. LEINAAS, eds., *Ecology of Fjords and Coastal Waters*, pp. 403–415. Elsevier, Amsterdam (Netherlands).
- Y. X. FU and W. H. LI, 1993. Statistical tests of neutrality of mutations. *Genetics* **133**: 693–709.
- N. GALTIER, M. GOUY, and C. GAUTIER, 1996. SEAVIEW and PHYLOWIN: two graphic tools for sequence alignment and molecular phylogeny. *Computer Applications in the Biosciences* **12**: 543–548.
- D. GORDON, C. ABAJIAN, and P. GREEN, 1998. Consed: a graphical tool for sequence finishing. *Genome Research* **8**: 195–202.
- J. GOUDET, 2006. *Hierfstat: estimation and tests of hierarchical F-statistics*. R package version 0.04-4.
- P. GREEN, 1994. Documentation for Phrap. <http://bozeman.mbt.washington.edu/phrap.docs/phrap.html>.
- B. GUINAND, C. LEMAIRE, and F. BONHOMME, 2004. How to detect polymorphisms undergoing selection in marine fishes? A review of methods and case studies, including flatfishes. *Journal of Sea Research* **51**: 167–182.
- N. K. HAASS, J. KARTENBECK, and R. E. LEUBE, 1996. Pantophysin is a ubiquitously expressed Synaptophysin homologue and defines constitutive transport vesicles. *Journal of Cell Biology* **134**: 731–746.

- D. L. HARTL and A. G. CLARK, 1989. *Principles of Population Genetics*. Sinauer Associates, Sunderland, Massachusetts, 2nd edn.
- P. W. HEDRICK, 2005. *Genetics of Populations*. Jones and Bartlett, Sudbury, Massachusetts, 3rd edn.
- J. HEMMER-HANSEN, E. E. NIELSEN, J. FRYDENBERG, and V. LOESCHKE, 2007. Adaptive divergence in a high gene flow environment: *Hsc 70* variation in the European flounder (*Platichthys flesus* L.). *Heredity* **99**: 592–600.
- R. R. HUDSON, 1990. Gene genealogies and the coalescent process. In D. FUTUYMA and J. ANTONOVICS, eds., *Oxford Surveys in Evolutionary Biology*, vol. 7, pp. 1–44. Oxford University Press, Oxford.
- R. R. HUDSON, D. D. BOOS, and N. L. KAPLAN, 1992. A statistical test for detecting geographic subdivision. *Molecular Biology and Evolution* **9**: 138–151.
- I. JÓNSDÓTTIR, G. MARTEINSDÓTTIR, and C. PAMPOULIE, 2008. Relation of growth and condition with the *Pan I* locus in Atlantic cod (*Gadus morhua* L.) around Iceland. *Marine Biology* **154**: 867–874.
- Ó.D.B. JÓNSDÓTTIR, A.K. DANÍELSDÓTTIR, and G. NÆVDAL, 2001. Genetic differentiation among Atlantic cod (*Gadus morhua* L.) in Icelandic waters: temporal stability. *International Council for the Exploration of the Sea Journal of Marine Science* **58**: 114–122.
- Ó.D.B. JÓNSDÓTTIR, A.K. IMSLAND, A.K. DANÍELSDÓTTIR, V. THORSTEINSSON, and G. NÆVDAL, 1999. Genetic differentiation among Atlantic cod in south and south-east Icelandic waters: Synaptophysin (*Syp I*) and Haemoglobin (*HbI*) variation. *Journal of Fish Biology* **54**: 1259–1274.
- S. KARLSSON and J. MORK, 2003. Selection-induced variation at the Synaptophysin locus (*Pan I*) in a Norwegian fjord population of cod (*Gadus morhua*). *Molecular Ecology* **12**: 3265–3274.
- P. KEOHAVONG and W. G. THILLY, 1989. Fidelity of DNA polymerases in DNA amplification. *Proceedings of the National Academy of Sciences of the United States of America* **86**: 9253–9257.
- M. J. KEOUGH and G. P. QUINN, 2002. *Experimental Design and Data Analysis for Biologists*. Cambridge University Press, New York, New York.

- M. KIMURA, 1968. Evolutionary rate at the molecular level. *Nature* **217**: 624–626.
- M. KIMURA, 1991. The neutral theory of molecular evolution: a review of recent evidence. *Japanese Journal of Genetics* **66**: 367–386.
- M. KREITMAN, 1983. Nucleotide polymorphism at the alcohol dehydrogenase locus of *Drosophila melanogaster*. *Nature* **304**: 2660–2674.
- M. KREITMAN, 1991. Detecting selection at the level of DNA. In R.K. SELANDER, A.G. CLARK, and T.S. WHITTAM, eds., *Evolution at the Molecular Level*, pp. 204–221. Sinauer Associates, Sunderland, Massachusetts.
- M. KREITMAN, 2000. Methods to detect selection in populations with applications to the human. *Annual Review of Genomics and Human Genetics* **1**: 539–559.
- M. KREITMAN and H. AKASHI, 1995. Molecular evidence for natural selection. *Annual Review of Ecology and Systematics* **26**: 403–422.
- M. KREITMAN and R. R. HUDSON, 1991. Inferring the evolutionary histories of the *Adh* and *Adh-dup* loci in *Drosophila melanogaster* from patterns of polymorphism and divergence. *Genetics* **127**: 565–582.
- S. KUMAR, K. TAMURA, and M. NEI, 2004. MEGA3: integrated software for molecular evolutionary genetics analysis and sequence alignment. *Briefings in Bioinformatics* **5**: 150–163.
- R. E. LEUBE, 1994. Expression of the Synaptophysin gene family is not restricted to neuronal and neuroendocrine differentiation in rat and human. *Differentiation* **56**: 163–171.
- R. LEWONTIN, 1974. *The Genetic Basis of Evolutionary Change*. Columbia University Press, New York, New York.
- J. H. McDONALD and M. KREITMAN, 1991. Adaptive protein evolution at the *Adh* locus in *Drosophila*. *Nature* **351**: 652–654.
- J. MORK, N. RYMAN, G. STAHL, F. UTTER, and G. SUNDNES, 1985. Genetic variation in Atlantic cod (*Gadus morhua*) throughout its range. *Canadian Journal of Fisheries and Aquatic Sciences* **42**(10): 1580–1587.
- T. OHTA, 2002. Near-neutrality in evolution of genes and gene regulation. *Proceedings of the National Academy of Sciences of the United States of America* **99**: 16,134–16,137.

- Ó. K. PÁLSSON and V. THORSTEINSSON, 2003. Migration patterns, ambient temperature, and growth of Icelandic cod (*Gadus morhua*): evidence from storage tag data. *Canadian Journal of Fisheries and Aquatic Sciences* **60**: 1409–1423.
- C. PAMPOULIE, K. B. JAKOBSDÓTTIR, G. MARTEINSDÓTTIR, and V. THORSTEINSSON, 2008. Are vertical behaviour patterns related to the Pantophysin locus in the Atlantic cod (*Gadus morhua* L.)? *Behavior Genetics* **38**: 76–81.
- C. PAMPOULIE, D.E. RUZZANTE, V. CHOSSON, T.D. JÖRUNSDÓTTIR, L. TAYLOR, V. THORSTEINSSON, A.K. DANÍELSDÓTTIR, and G. MARTEINSDÓTTIR, 2006. The genetic structure of Atlantic cod (*Gadus morhua*) around Iceland: insight from microsatellites, the *Pan I* locus, and tagging experiments. *Canadian Journal of Fisheries and Aquatic Science* **63**: 2660–2674.
- E. PARADIS, J. CLAUDE, and K. STRIMMER, 2004. APE: analyses of phylogenetics and evolution in R language. *Bioinformatics* **20**: 289–290.
- G. H. POGSON, 2001. Nucleotide polymorphism and natural selection at the Pantophysin (*Pan I*) locus in the Atlantic cod, *Gadus morhua* (L.). *Genetics* **157**: 317–330.
- G. H. POGSON and S. E. FEVOLDEN, 1998. DNA heterozygosity and growth rate in the Atlantic cod *Gadus morhua* (L.). *Evolution* **52**: 915–920.
- G. H. POGSON and S. E. FEVOLDEN, 2003. Natural selection and the genetic differentiation of coastal and Arctic populations of the Atlantic cod in northern Norway: a test involving nucleotide sequence variation at the Pantophysin (*Pan I*) locus. *Molecular Ecology* **12**: 63–74.
- G. H. POGSON and K. A. MESA, 2004. Positive Darwinian selection at the Pantophysin (*Pan I*) locus in marine gadid fishes. *Molecular Biology and Evolution* **21**: 65–75.
- G. H. POGSON, K. A. MESA, and R. G. BOUTILIER, 1995. Genetic population structure and gene flow in the Atlantic cod *Gadus morhua*: a comparison of allozyme and nuclear RFLP loci. *Genetics* **139**: 375–385.
- G. H. POGSON, C. T. TAGGART, K. A. MESA, and R. G. BOUTILIER, 2001. Isolation by distance in the Atlantic cod, *Gadus morhua*, at large and small geographic scales. *Evolution* **55**: 131–146.

- R DEVELOPMENT CORE TEAM, 2008. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
- M. RIDLEY, 2003. *Evolution*. Wiley-Blackwell, Malden, Massachusetts, 3rd edn.
- J. ROZAS, J. C. SÁNCHEZ-DELBARRIO, X. MESSEGUER, and R. ROZAS, 2003. DnaSP, DNA polymorphism analyses by the coalescent and other methods. *Bioinformatics* **19**: 2496–2497.
- D. E. RUZZANTE, C. T. TAGGART, C. COOK, and S. GODDARD, 1996. Genetic differentiation between inshore and offshore Atlantic cod (*Gadus morhua*) off Newfoundland: microsatellite DNA variation and antifreeze level. *Canadian Journal of Fisheries and Aquatic Sciences* **53** (3): 634–645.
- T. H. SARVAS and S. E. FEVOLDEN, 2005. Pantophysin (*Pan I*) locus divergence between inshore v. offshore and northern v. southern populations of Atlantic cod in the north-east Atlantic. *Journal of Fish Biology* **67**: 444–469.
- M. SLATKIN, 2005. A coalescent view of population structure. In R. SINGH and C. KRIMBAS, eds., *Evolutionary Genetics: From Molecules to Morphology*, chap. 21, pp. 418–429. New York: Columbia University Press.
- R. R. SOKAL and F. J. ROHLF, 1995. *Biometry: The Principles and Practice of Statistics in Biological Research*. W. H. Freeman and Co., New York, 3rd edn.
- D. P. SWAIN, A. F. SINCLAIR, and J. MARK HANSON, 2007. Evolutionary response to size-selective mortality in an exploited fish population. *Proceedings of the Royal Society B: Biological Sciences* **274**: 1015–1022.
- F. TAJIMA, 1983. Evolutionary relationship of DNA sequences in finite populations. *Genetics* **105**: 437–460.
- J. D. THOMPSON, D. G. HIGGINS, and T. J. GIBSON, 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research* **22**: 4673–4680.
- J. D. WALL, P. ANDOLFATTO, and M. PRZEWORSKI, 2002. Testing models of selection and demography in *Drosophila simulans*. *Genetics* **162**: 203–216.

P. S. WALSH, D. A. METZGER, and R. HIGUCHI, 1991. Chelex 100 as a medium for simple extraction of DNA for PCR-based typing from forensic material. *BioTechniques* **10**: 506–513.

Appendix I. Depth and geographic region

This appendix contains auxiliary images that illustrate the relation between depth and geographic regions.

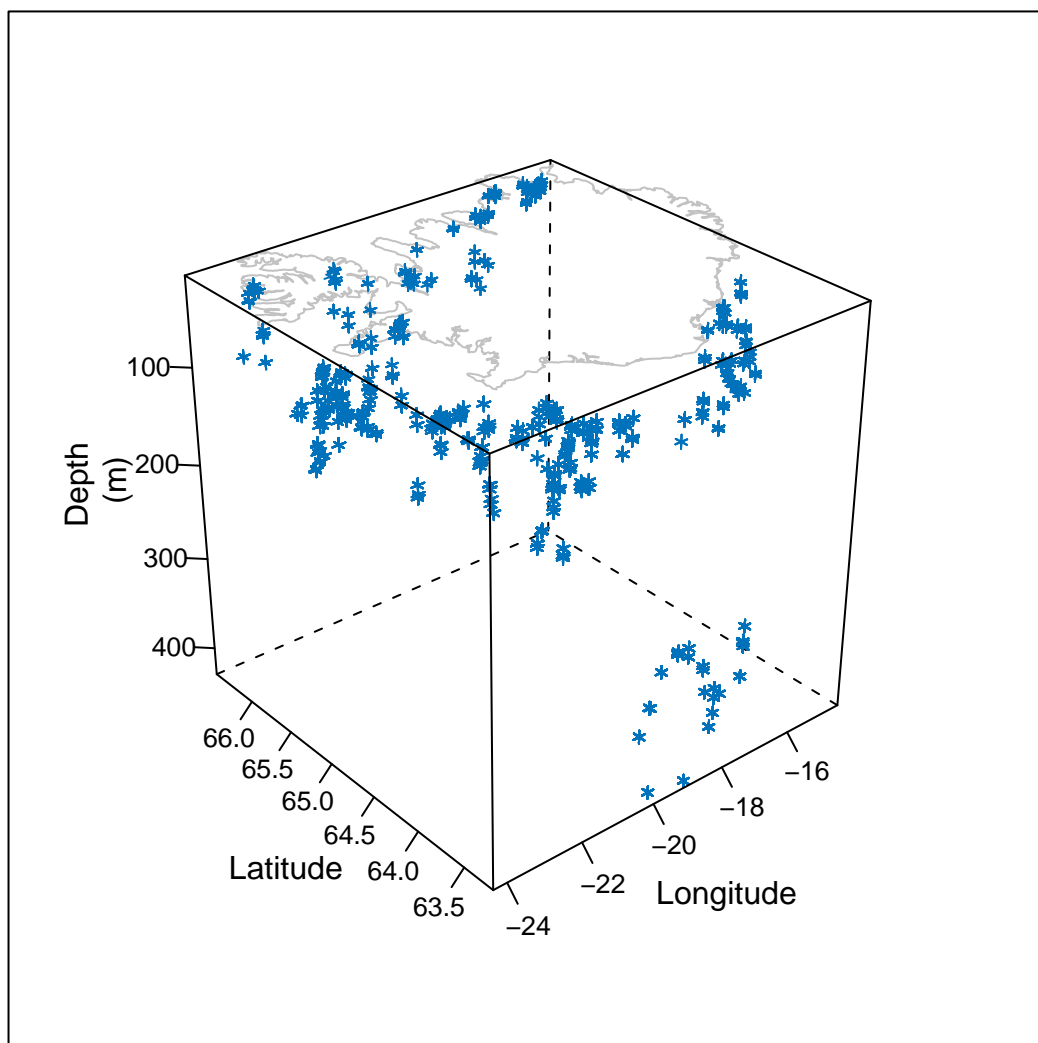


Figure 1: Sampling sites by depth, latitude and longitude (3D view), for spring surveys 2005, 2006, and 2007.

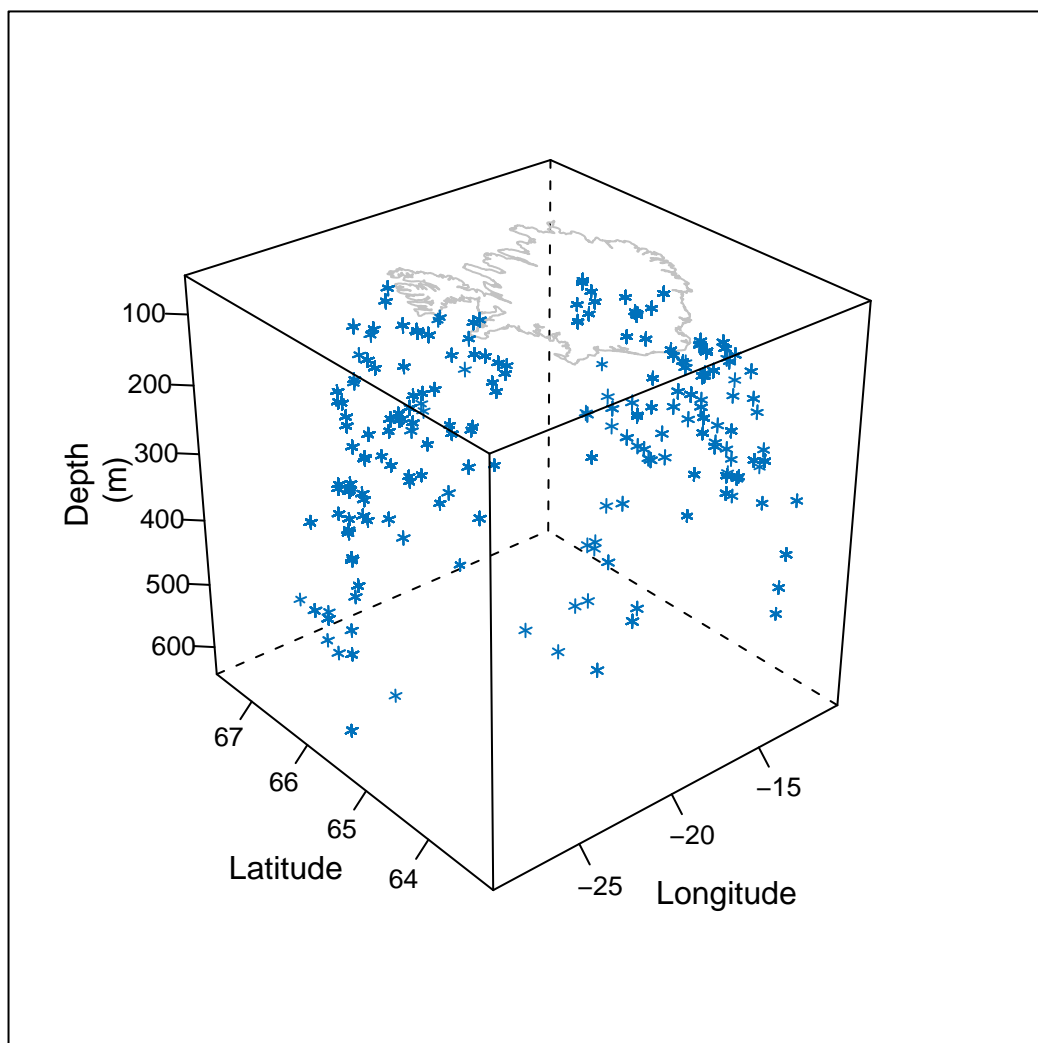


Figure 2: Sampling sites by depth, latitude and longitude (3D view), for fall surveys 2004, 2005, and 2006.

Appendix II. *Pan* I genotype and length at age (fall surveys)

This appendix contains an interaction plot and a table of multiple ANOVA's to illustrate correlation between *Pan* I genotype and length at age. The data is from fall surveys.

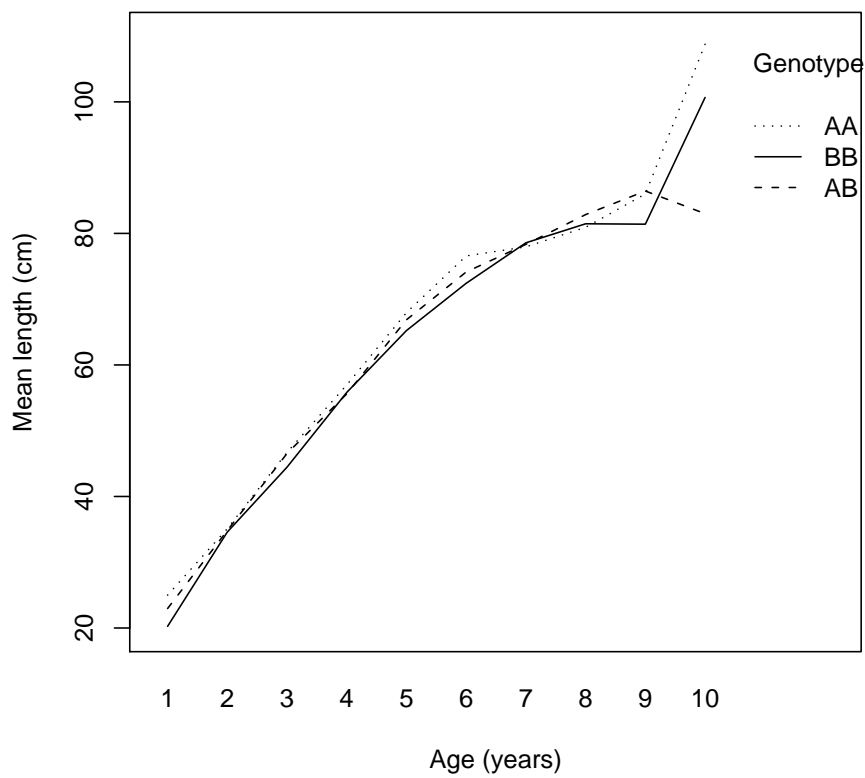


Figure 3: Interaction plot of mean length on genotype and age, for fall surveys 2004, 2005, and 2006.

Table 1: ANOVAs (one-way) of length on genotype at different ages, for fall surveys 2004, 2005, and 2006. SS is sum of squares, MS is means square, F is ratio of MS among genotypes and MS residuals, df is degrees of freedom, p is probability. SL is significance level of difference in length among genotypes; sequential Bonferroni correction was applied.

Age	Source of variance	df	SS	MS	F	p	SL
1	Genotype	2	208.67	104.33	9.11	0.0003*	0.0050
	Residuals	74	847.15	11.45			
2	Genotype	2	4.36	2.18	0.08	0.9277	0.0250
	Residuals	217	6294.24	29.01			
3	Genotype	2	425.45	212.72	6.46	0.0017*	0.0063
	Residuals	379	12479.54	32.93			
4	Genotype	2	87.62	43.81	0.92	0.4001	0.0100
	Residuals	543	25922.84	47.74			
5	Genotype	2	668.69	334.35	5.83	0.0031*	0.0071
	Residuals	658	37741.05	57.36			
6	Genotype	2	961.77	480.89	6.85	0.0012*	0.0056
	Residuals	506	35510.30	70.18			
7	Genotype	2	11.14	5.57	0.06	0.9461	0.0050
	Residuals	218	21891.04	100.42			
8	Genotype	2	50.25	25.13	0.24	0.7888	0.0167
	Residuals	73	7707.70	105.58			
9	Genotype	2	142.82	71.41	0.56	0.5812	0.0125
	Residuals	22	2824.62	128.39			
10	Genotype	2	1376.22	688.11	10.76	0.0054*	0.0083
	Residuals	8	511.42	63.93			

*Statistically significant after Bonferroni correction

Appendix III. Quality of DNA sequences

This appendix contains graphical representations of sequence quality, as determined by Phred/Phrap/Consed, of the *Pan I^A* and *Pan I^B* allele sequences utilized in the present study.

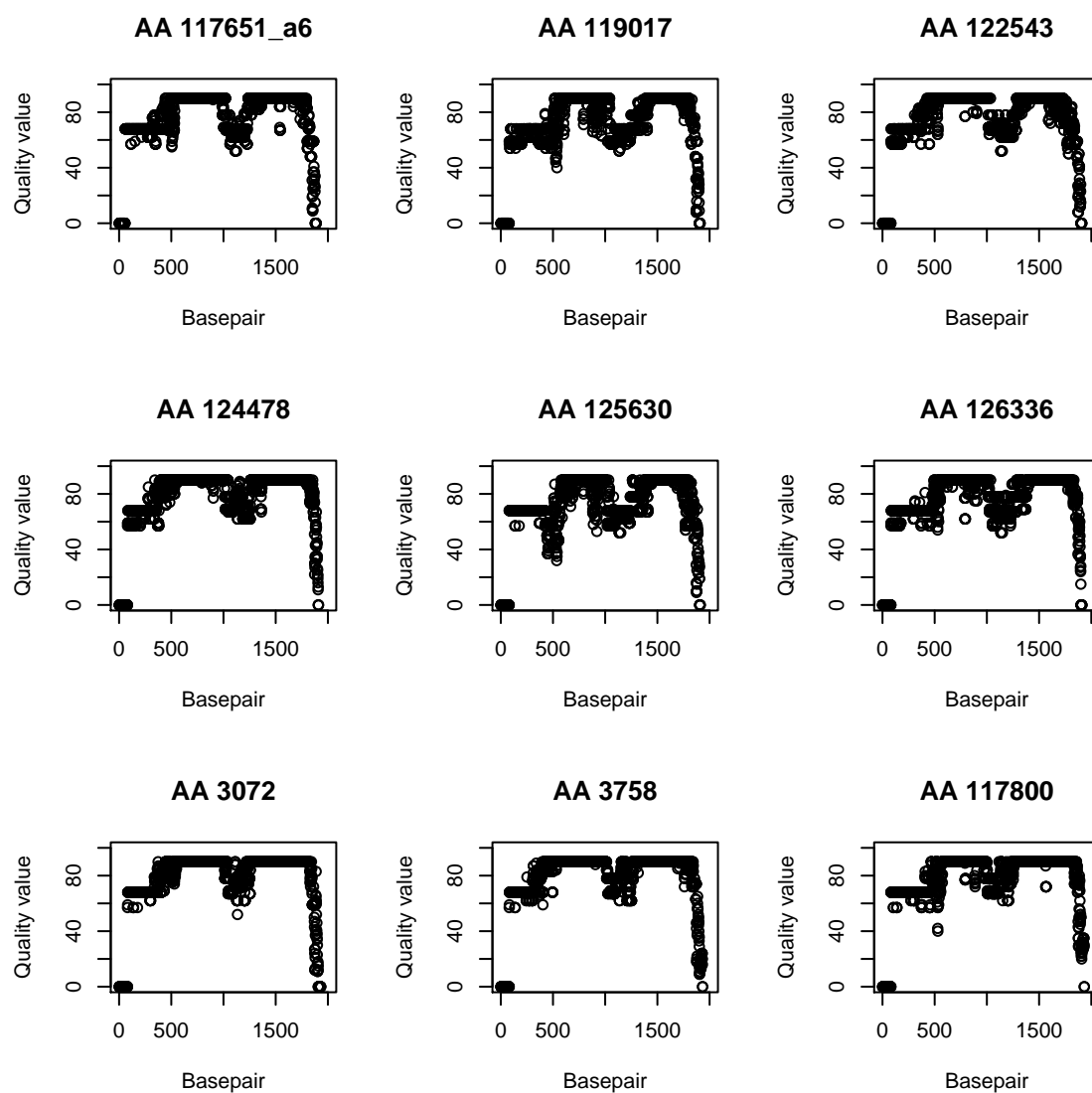


Figure 4: Sequence quality of *Pan I^A* alleles. The quality value of the sequence of each clone is plotted against the pertinent base pair. Above each panel is the genotype and the numeric name of the clone. A quality value of 10 means 1 error in ten to the 1.0 power, a quality value of 20 means 1 error in ten to the 2.0 power, and so forth. Samples are from spring surveys 2005, 2006, and 2007.

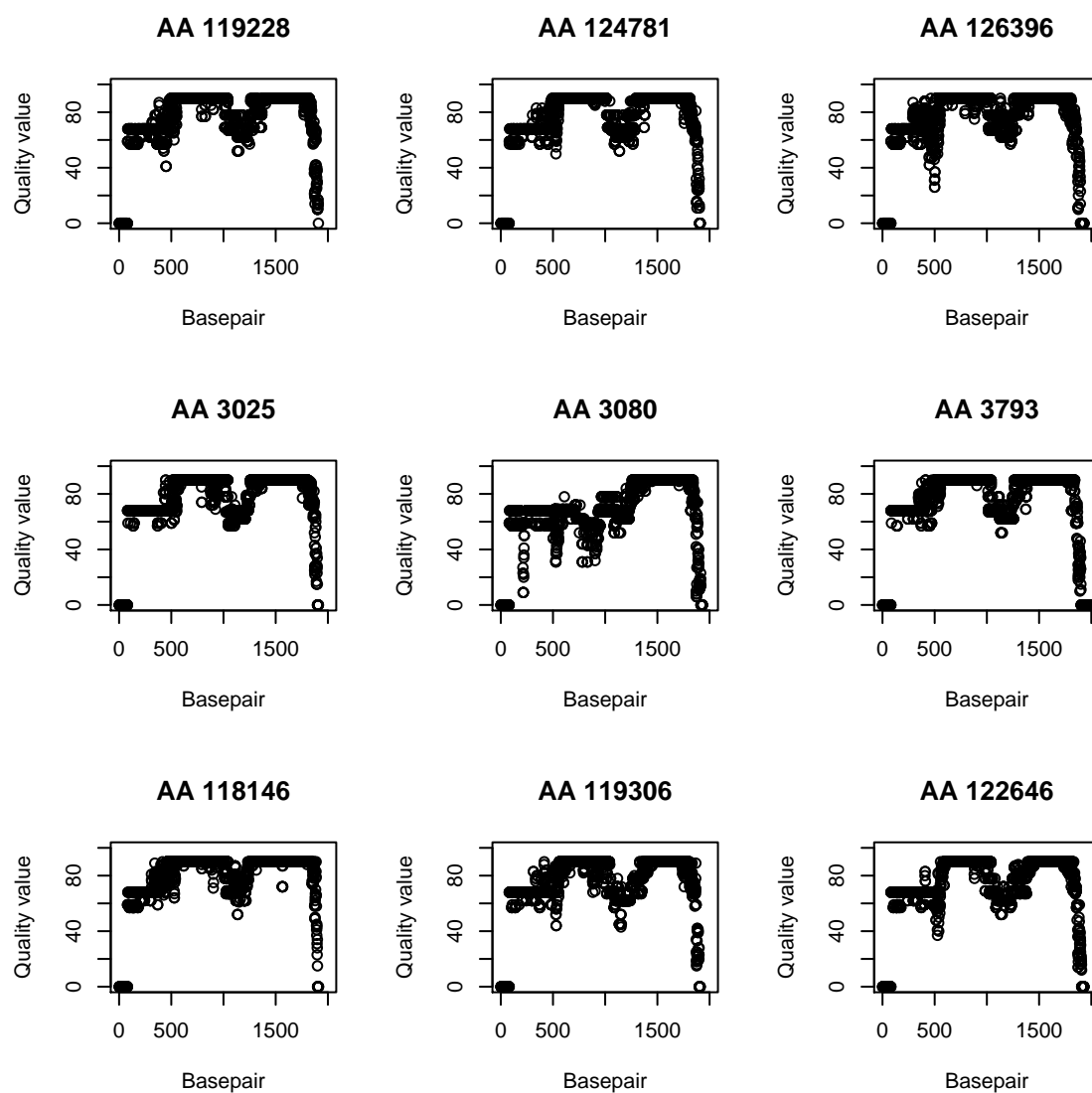


Figure 5: Sequence quality of *Pan I^A* alleles. The quality value of the sequence of each clone is plotted against the pertinent base pair. Above each panel is the genotype and the numeric name of the clone. A quality value of 10 means 1 error in ten to the 1.0 power, a quality value of 20 means 1 error in ten to the 2.0 power, and so forth. Samples are from spring surveys 2005, 2006, and 2007.

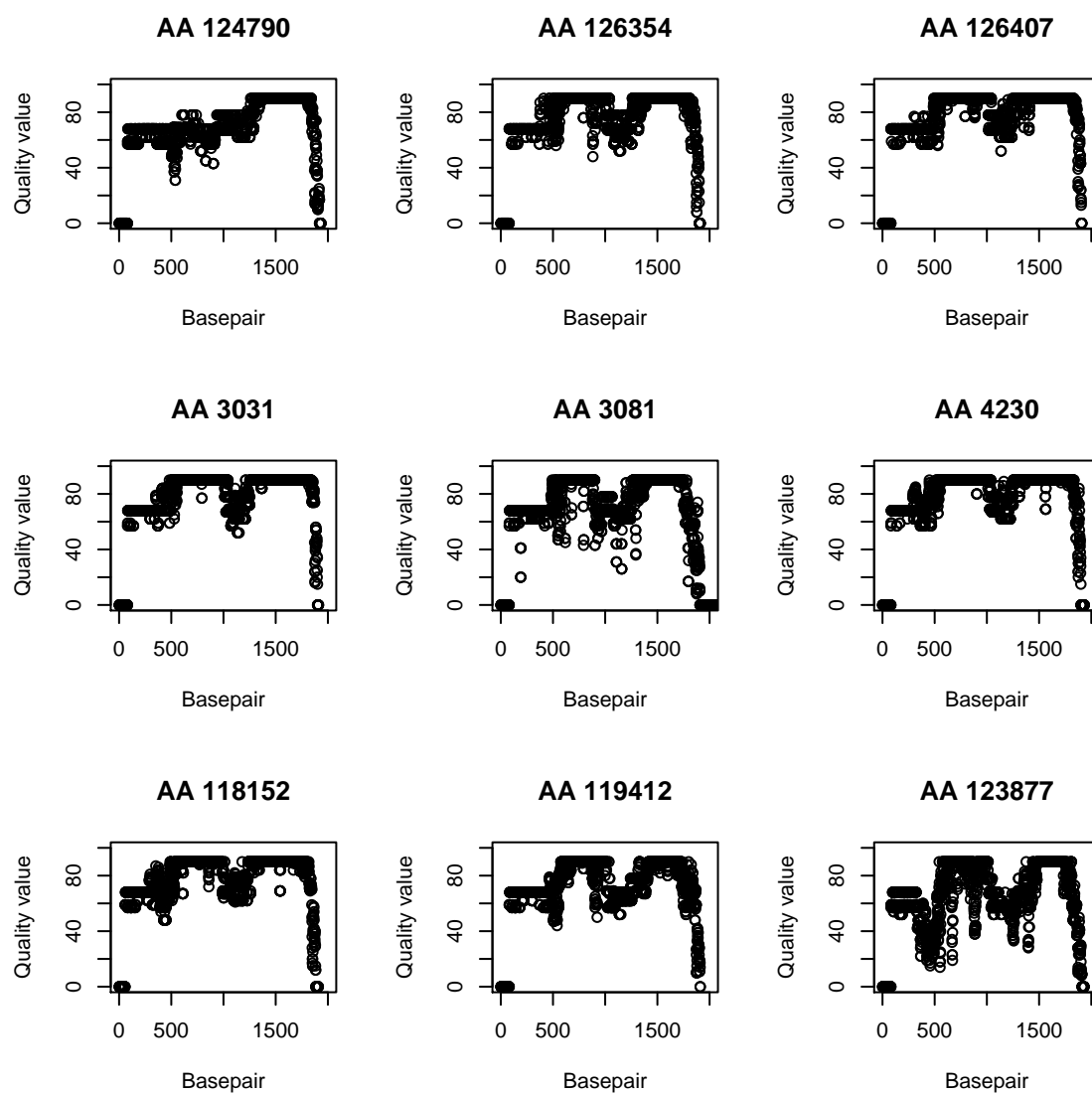


Figure 6: Sequence quality of *Pan I^A* alleles. The quality value of the sequence of each clone is plotted against the pertinent base pair. Above each panel is the genotype and the numeric name of the clone. A quality value of 10 means 1 error in ten to the 1.0 power, a quality value of 20 means 1 error in ten to the 2.0 power, and so forth. Samples are from spring surveys 2005, 2006, and 2007.

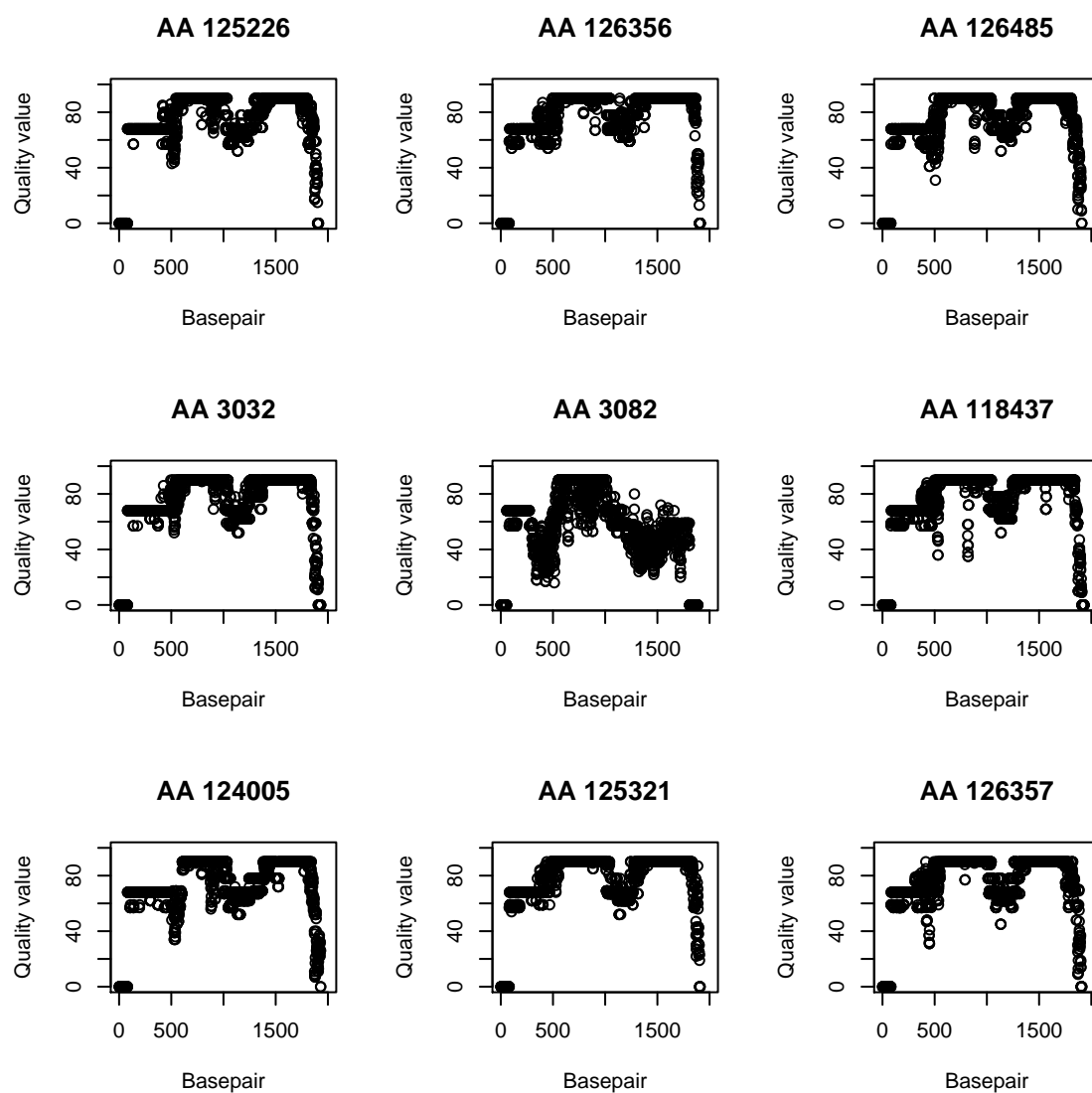


Figure 7: Sequence quality of *Pan I^A* alleles. The quality value of the sequence of each clone is plotted against the pertinent base pair. Above each panel is the genotype and the numeric name of the clone. A quality value of 10 means 1 error in ten to the 1.0 power, a quality value of 20 means 1 error in ten to the 2.0 power, and so forth. Samples are from spring surveys 2005, 2006, and 2007.

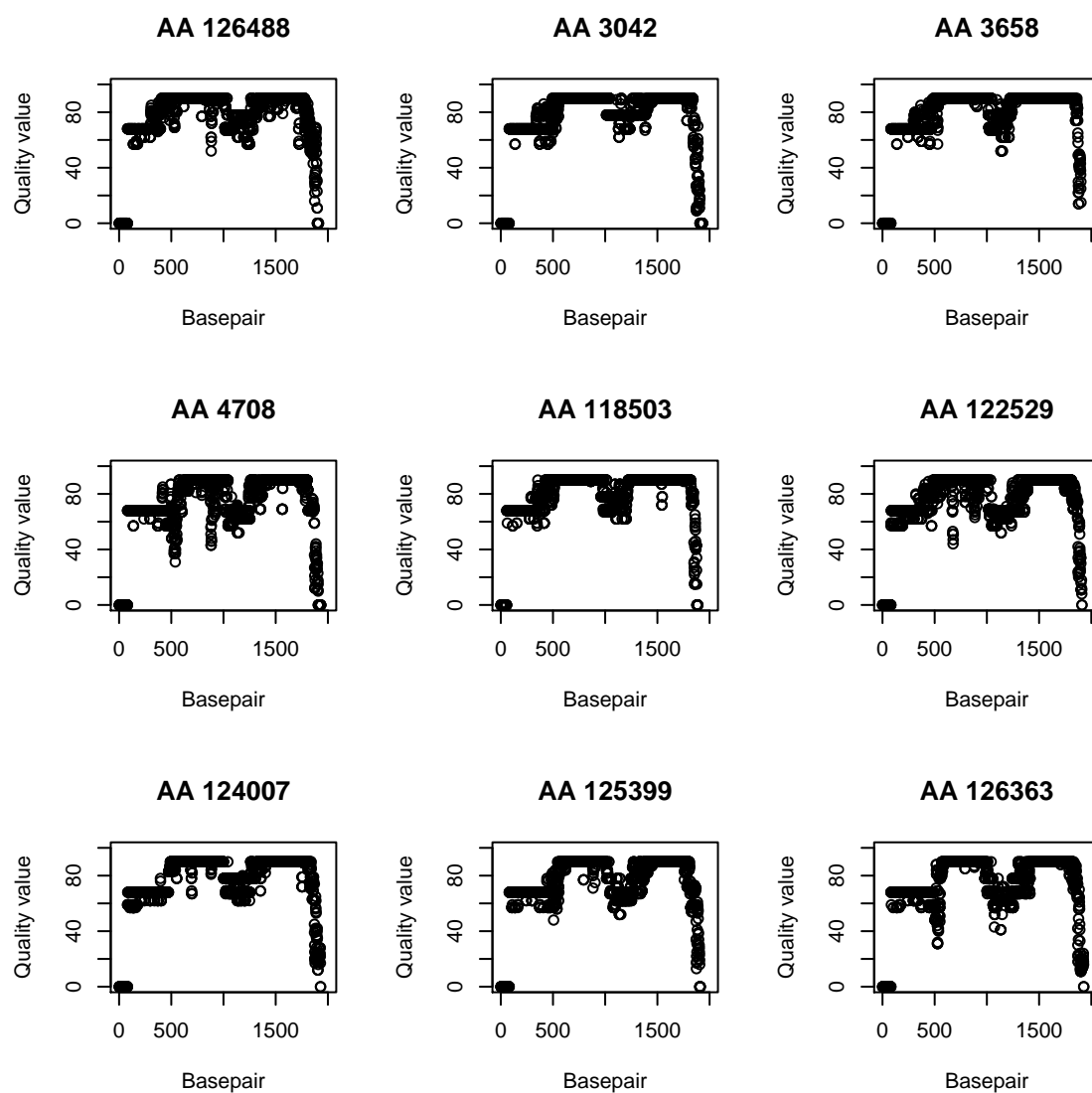


Figure 8: Sequence quality of *Pan* I^A alleles. The quality value of the sequence of each clone is plotted against the pertinent base pair. Above each panel is the genotype and the numeric name of the clone. A quality value of 10 means 1 error in ten to the 1.0 power, a quality value of 20 means 1 error in ten to the 2.0 power, and so forth. Samples are from spring surveys 2005, 2006, and 2007.

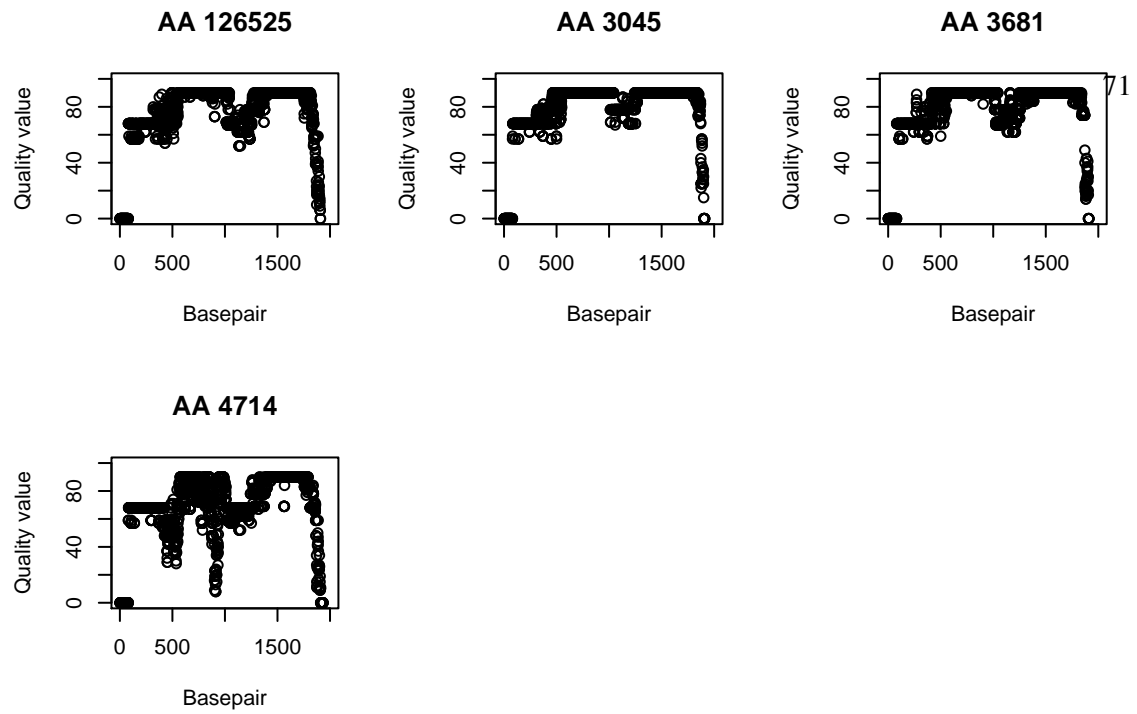


Figure 9: Sequence quality of *Pan I^A* alleles. The quality value of the sequence of each clone is plotted against the pertinent base pair. Above each panel is the genotype and the numeric name of the clone. A quality value of 10 means 1 error in ten to the 1.0 power, a quality value of 20 means 1 error in ten to the 2.0 power, and so forth. Samples are from spring surveys 2005, 2006, and 2007.

Quality of sequences of *Pan I^A* allele (re-sequenced clone)

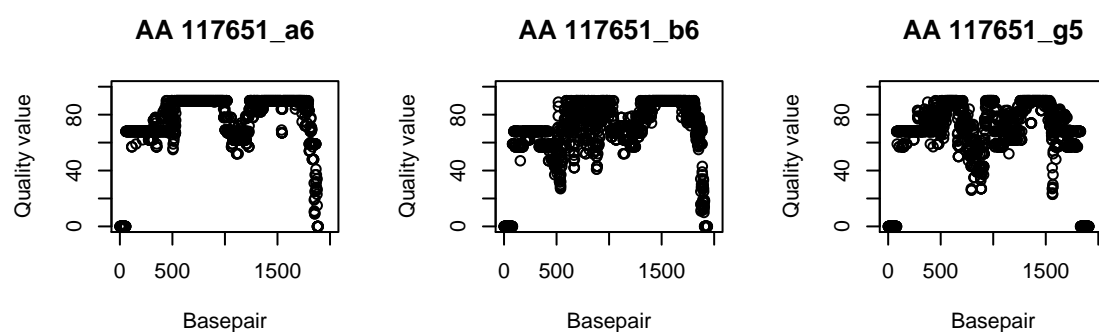


Figure 10: Sequence quality of *Pan I^A* alleles from a same clone. The quality value of the 3 sequences of the same (re-sequenced) clone are plotted against the pertinent base pair. Above each panel is the genotype, the numeric name of the clone, and a unique alphanumeric identifier for each sequence. A quality value of 10 means 1 error in ten to the 1.0 power, a quality value of 20 means 1 error in ten to the 2.0 power, and so forth. Samples are from spring surveys 2005, 2006, and 2007.

Quality of sequences of *Pan I^B* alleles

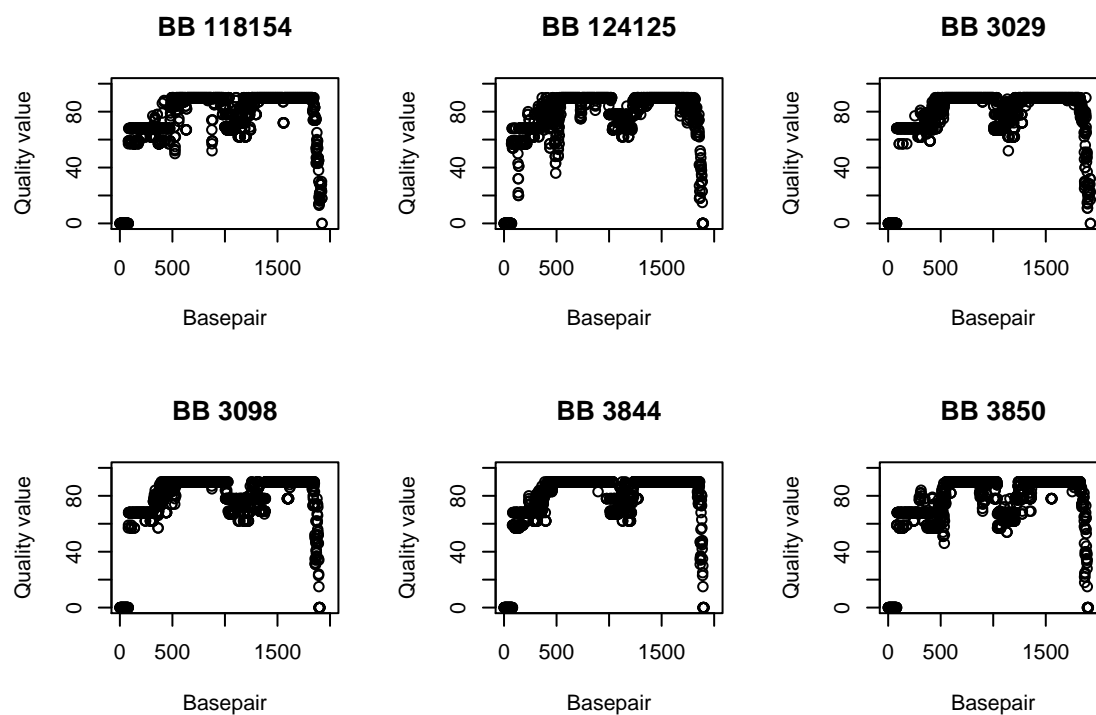


Figure 11: Sequence quality of *Pan I^B* alleles. The quality value of the sequence of each clone is plotted against the pertinent base pair. Above each panel is the genotype and the numeric name of the clone. A quality value of 10 means 1 error in ten to the 1.0 power, a quality value of 20 means 1 error in ten to the 2.0 power, and so forth. Samples are from spring surveys 2005, 2006, and 2007.

Appendix IV. Gene genealogies, geography and depth

This appendix contains gene genealogies of *Pan* I^A alleles with geographic location and depth overlaid. Data included and excluded singletons.

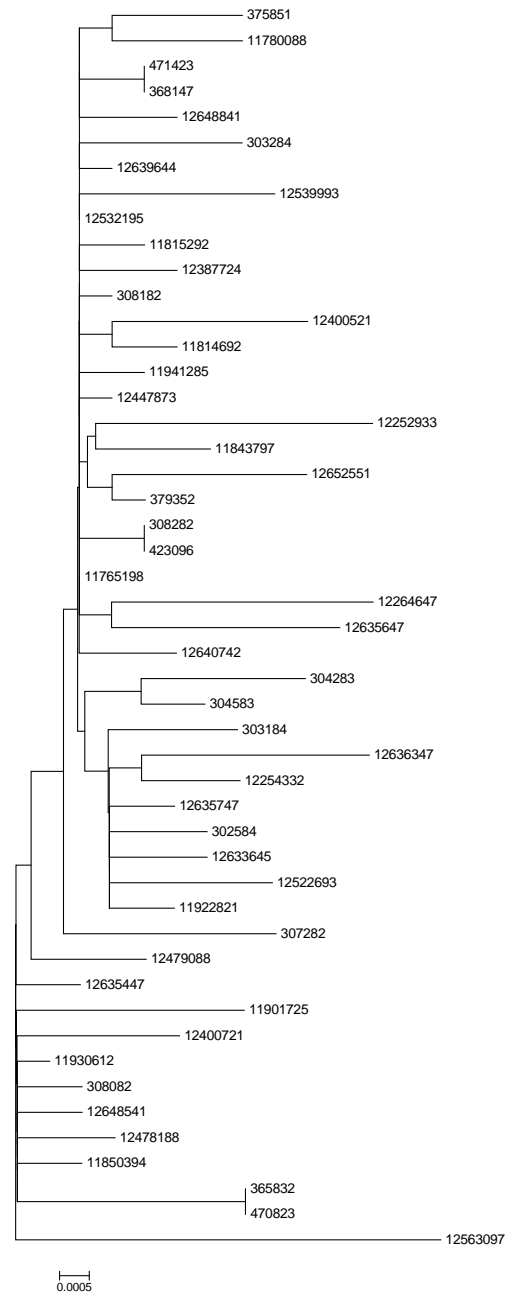


Figure 12: Gene tree from DNA sequences (including singletons) of *Pan I^A* alleles. The tree was built using the neighbor joining method. Samples are from spring surveys 2005, 2006, and 2007. The last two digits of the sequence names correspond to MetaCod division and depth respectively. The digits before that are individual identifiers. MetaCod division (1 to 9) is a set of geographic regions surrounding Iceland. The depth levels (in m.) are 1: 0–25; 2: 25–50; 3: 50–75; 4: 75–100; 5: 100–125; 6: 125–150; 7: 150–175; 8: 275–300.

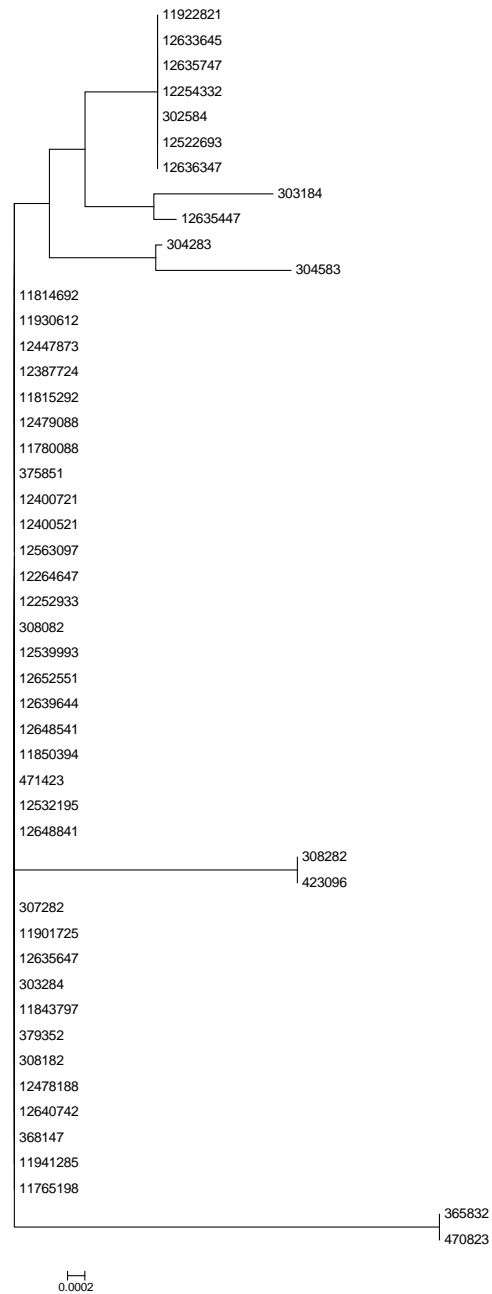


Figure 13: Gene tree from DNA sequences (excluding singletons) of *Pan I^A* alleles. The tree was built using the neighbor joining method. Samples are from spring surveys 2005, 2006, and 2007. The last two digits of the sequence names correspond to MetaCod division and depth respectively. The digits before that are individual identifiers. MetaCod division (1 to 9) is a set of geographic regions surrounding Iceland. The depth levels (in m.) are 1: 0–25; 2: 25–50; 3: 50–75; 4: 75–100; 5: 100–125; 6: 125–150; 7: 150–175; 8: 275–300.

Appendix V. Various DNA sequence data analysis excluding singletons

This appendix contains various tables and graphics that illustrate DNA sequence variation analysis excluding singletons.

Table 2: Segregating sites in *Pan I* among single clones from 55 Atlantic cod from Iceland. Singleton nucleotide substitutions are excluded. Individuals identified to MetaCod division (number after first dot) and depth level (number after second dot). MetaCod divisions are a set of geographic areas surrounding Iceland. The depth levels (in m.) are 1: 0–25; 2: 25–50; 3: 50–75; 4: 75–100; 5: 100–125; 6: 125–150; 7: 150–175; 8: 275–300. Indels (∇) identified with respective numbers. $\nabla 7$ identifies several different single base pair indels.

[illegible]

Continued on next page

Table 2 – Continued

		Variable nucleotide site															
Individual		1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
.division		3	5	7	9	8	2	2	2	3	3	3	3	3	3	3	3
.depth		0	7	6	7	2	3	1	6	7	8	9	0	1	2	3	4
4230.9.6	T...11	CTAA
4714.2.3	T...11	AA
3681.4.7	T...11	AA
126363.4.7	T...11	AA
3031.8.4	T...11	AA
122543.3.2	T...11	AA
125226.9.3	T...11	AA
3025.8.4	T...11	AA
119228.2.1	T...11	AA
126336.4.5	T...11	AA
126357.4.7	T...11	5	AA
119017.2.5	TA...11	AA
124790.8.8	T...11	AA
126354.4.7	T...11	AA

Continued on next page

[illegible]

80

Table 2 – Continued

		Variable nucleotide site															
Individual		1	1	2	2	2	2	2	2	2	2	2	2	2	2	2	1
.division		3	5	7	9	8	8	2	2	2	3	3	3	3	3	3	1
.depth		0	7	6	7	2	3	1	6	7	8	9	0	1	2	3	4
3793.5.2	T...	1	1	1
126407.4.2	TA...	1	1	1
122529.3.3	T...	1	1	1
3032.8.4	T...	1	1	1
125399.9.3	T...	1	1	1
118437.9.7	T...	1	1	1
126396.4.4	T...	1	1	1
118152.9.2	T...	1	1	1
3081.8.2	T...	1	1	1
123877.2.4	T...	1	1	1
124478.7.3	T...	1	1	1
125321.9.5	T...	1	1	1
119412.8.5	T...	1	1	1
126488.4.1	T...	1	1	1

Continued on next page

Table 2 – Continued

[illegible]

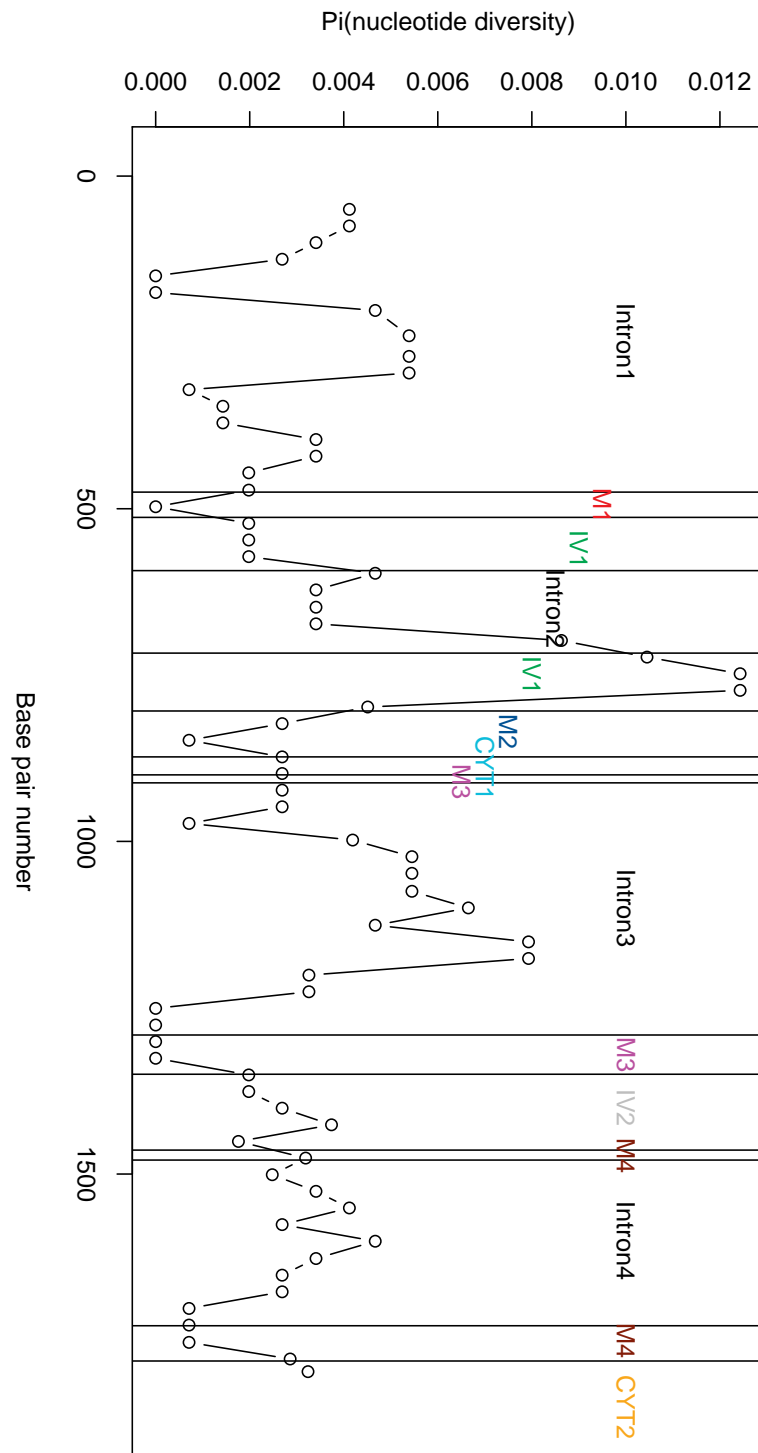


Figure 14: Sliding window analysis (excluding singletons) of nucleotide diversity (π) throughout the *Pan I* gene region in *Pan I^A* and *Pan I^B* alleles. Window size is 100 bp and step size is 25 bp. M1 through M4 are membrane spanning domains, CYT1 and CYT2 are cytoplasmic tails domains, IV1 and IV2 are intravesicular domains. Samples are from spring surveys 2005, 2006, and 2007.

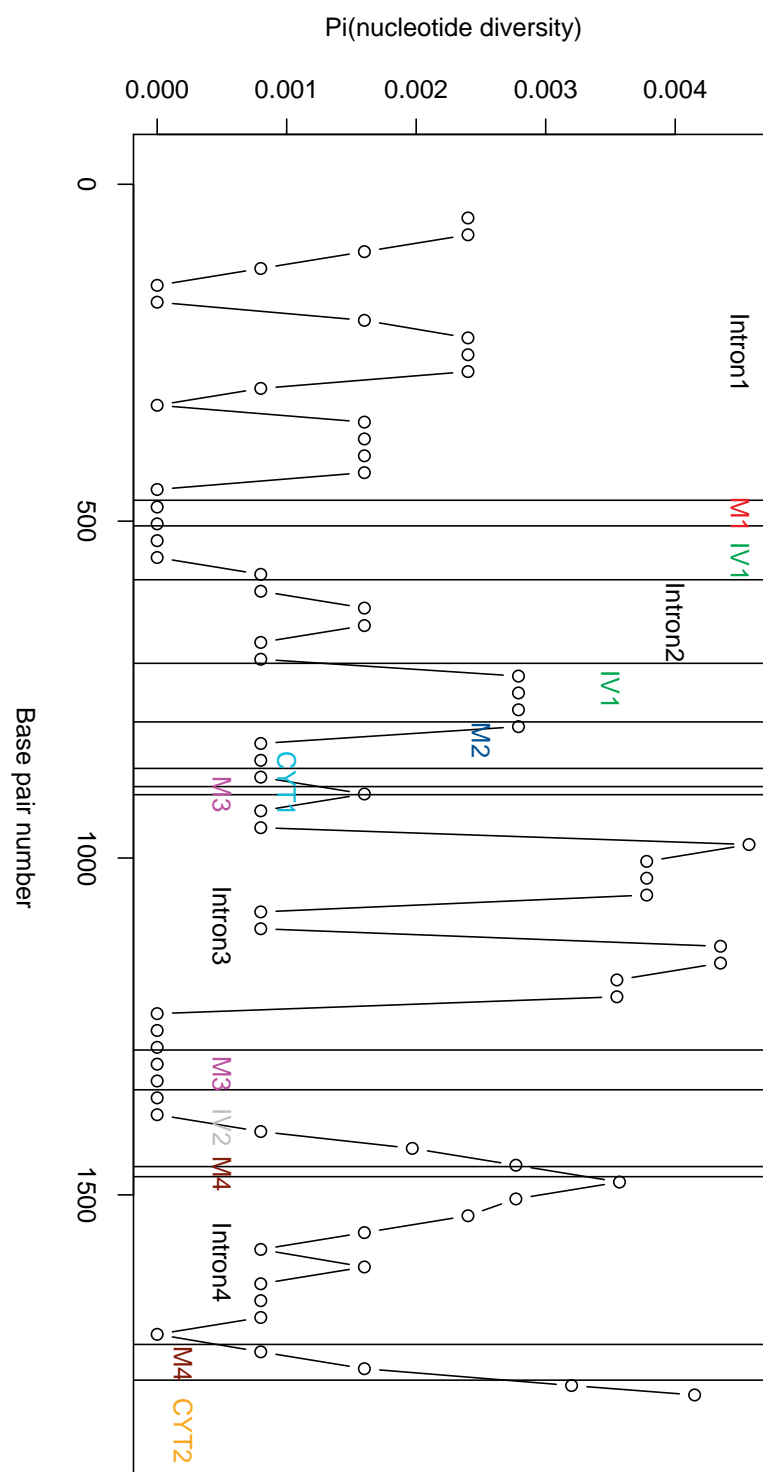


Figure 15: Sliding window analysis (excluding singletons) of nucleotide diversity (π) throughout the *Pan I* gene region in *Pan I^A* alleles. Window size is 100 bp and step size is 25 bp. M1 through M4 are membrane spanning domains, CYT1 and CYT2 are cytoplasmic tails domains, IV1 and IV2 are intravesicular domains. Samples are from spring surveys 2005, 2006, and 2007.

Table 3: Extent of DNA divergence between populations at different MetaCod divisions (localities), for *Pan* I^A without singletons. MCdiv is MetaCod division, n is the number of sequences at each MetaCod division, D_a with Jukes and Cantor (JC) is the number of net nucleotide substitutions per site between populations, SE_{D_a} is the Standard Error of D_a (JC); D_a (JC) and SE_{D_a} appear above below the diagonal, respectively.

<i>Pan I</i> ^A					
		<i>D_a</i> (JC)×1000			
	MCdiv 1,2,3	MCdiv 4,5	MCdiv7,8	MCdiv9	<i>n</i>
MCdiv 1,2,3		0.10	0.12	0.07	11
MCdiv 4,5	0.54		−0.02	−0.04	14
MCdiv 7,8	0.52	0.22		−0.04	14
MCdiv 9	0.55	0.27	0.26		10
		SE _{<i>D_a</i>} ×1000			

Table 4: Extent of DNA divergence between populations at different depth levels, for *Pan* I^A without singletons. Depth levels are (intervals in m.) 1: 0–25; 2: 25–50; 3: 50–75; 4: 75–100; 5: 100–125; 6: 125–150; 7: 150–175; 8: 275–300. n is the number of sequences at each depth level, D_a with Jukes and Cantor (JC) is the number of net nucleotide substitutions per site between populations; D_a (JC) and SE_{D_a} appear above and below the diagonal, respectively.

<i>Pan I</i> ^A								
<i>D_a</i> (JC)×1000								<i>n</i>
Depth	Depth	Depth	Depth	Depth	Depth	Depth		
level 1	level 2	level 3	level 4	level 5,6	level 7	level 8		
Depth level 1	−0.10	−0.03	−0.07	−0.10	−0.03	−0.13	7	
Depth level 2	0.46	−0.08	−0.04	−0.14	−0.01	−0.07	11	
Depth level 3	0.62	0.65	−0.04	−0.04	0.00	0.03	8	
Depth level 4	0.32	0.49	0.64	−0.11	−0.05	−0.01	6	
Depth level 5,6	0.41	0.54	0.72	0.42	−0.05	−0.07	5	
Depth level 7	0.39	0.51	0.67	0.38	0.48	0.00	8	
Depth level 8	0.34	0.50	0.67	0.37	0.44	0.43	4	
SE _{<i>D_a</i>} ×1000								