Living in each other's pockets: Nucleotide variation inside a genomic island harboring *Pan* I and its neighbors in Atlantic cod

Ubaldo Benitez Hernandez, Einar Árnason

The *Pan* I locus in Atlantic cod lies in a genomic island of divergence extending over a large genomic region. The locus has two divergent alleles, defined by a single *Dra*I restriction site, that have been related to behavioral differences of habitat selection by depth and temperature. The *Pan* I locus is known to be under an unusual mix of balancing selection and selective sweeps within the functional types. Here we study nucleotide variation in a 12.5 kb region inside the genomic island harboring Pan I and neighboring loci for sortilin 1 (*Sort*1) and ataxin 7-like 2 (*Atxn*7*l*2) which we partially covered. Variation of the 31 gene copies throughout the region falls into two divergent haplogroups that correlate with the 25 copies of *A* and six copies of *B* alleles of *Pan* I. The unfolded site frequency spectrum for the part with Pacific cod used as the outgroup is trimodal with a mode at singletons and two high frequency modes at 6/31 and 25/31 representing the two genealogical lineages. The folded site frequency spectrum for the entire region similarly has a high frequency mode of mutations that have accumulated on the two lineages. The high frequency of singletons is accounted for by multiple merger coalescent models. Parameter estimates using these models indicate sweepstakes reproduction. The high frequency modes of the spectrum is evidence for balancing selection. Analysis of non-synonymous changes shows that *Pan* I is at least one focus of selection within the genomic island. There may be multiple sites of selection and epistatic interactions. There is extensive linkage disequilibrium throughout the region. We suggest that the genomic island of divergence is a supergene of co-adapted complexes possibly locked together by structural variation.

Living in each other's pockets: Nucleotide variation inside a genomic island harboring *Pan* I and its neighbors in Atlantic cod

Ubaldo Benitez Hernandez[1] **and Einar Árnason**[2]

[1]**Institute of Life and Environmental Sciences, University of Iceland, Reykjavik, Iceland**
[2]**Institute of Life and Environmental Sciences, University of Iceland, Reykjavik, Iceland**

ABSTRACT

The *Pan* I locus in Atlantic cod lies in a genomic island of divergence extending over a large genomic region. The locus has two divergent alleles, defined by a single *Dra*I restriction site, that have been related to behavioral differences of habitat selection by depth and temperature. The *Pan* I locus is known to be under an unusual mix of balancing selection and selective sweeps within the functional types. Here we study nucleotide variation in a 12.5 kb region inside the genomic island harboring *Pan* I and neighboring loci for sortilin 1 (*Sort*1) and ataxin 7-like 2 (*Atxn7l2*) which we partially covered. Variation of the 31 gene copies throughout the region falls into two divergent haplogroups that correlate with the 25 copies of *A* and six copies of *B* alleles of *Pan* I. The unfolded site frequency spectrum for the part with Pacific cod used as the outgroup is trimodal with a mode at singletons and two high frequency modes at 6/31 and 25/31 representing the two genealogical lineages. The folded site frequency spectrum for the entire region similarly has a high frequency mode of mutations that have accumulated on the two lineages. The high frequency of singletons is accounted for by multiple merger coalescent models. Parameter estimates using these models indicate sweepstakes reproduction. The high frequency modes of the spectrum is evidence for balancing selection. Analysis of non-synonymous changes shows that *Pan* I is at least one focus of selection within the genomic island. There may be multiple sites of selection and epistatic interactions. There is extensive linkage disequilibrium throughout the region. We suggest that the genomic island of divergence is a supergene of co-adapted complexes possibly locked together by structural variation.

Keywords: Balancing selection, Linkage Disequilibrium, Pantophysin, Pan I, Genomic Island, Sortilin, Ataxin-7 like, Atlantic cod

INTRODUCTION

Selection, the differential fecundity and mortality of genotypes, is a most powerful evolutionary force. Organisms exploit finite resources with differential efficiency leading to fitness differences. They thereby pass on their alleles to future generations with differential efficiency and thus are selected for (Lewontin, 1974). The selective forces arise from diverse physical or biotic factors and can exist in different combinations, resulting also in diverse patterns of polymorphism. Various modes of selection exist. Under negative or purifying selection, detrimental mutations are purged. Under positive or advantageous selection variants sweep to fixation. Under balancing selection, however, alternative allelic forms exist at intermediate frequencies due to a tug of selective forces that may ensue because of spatially or temporally varying selective pressures favoring allelic forms differently, due to heterozygous advantage or overdominance, or due to inverse frequency-dependent selection.

The genealogical relationships among alleles are captured by the coalescent (Kingman, 1982) that is a retrospective model of the assignment of alleles of a sample to ancestors. Going back in time, alleles reach a point of common ancestry. Different present-day lineages of descent, in hind-sight fuse or coalesce in an ancestral form they originate from. Eventually all alleles of a sample merge in a single most recent common ancestor (Wakeley, 2009). The coalescent yields the theoretical site frequency spectrum of a population at neutrality that can be contrasted with the observed site frequency spectrum thus allowing for the detection of deviations from the theoretical expectations at neutrality. The classic coalescent model (Kingman, 1982), derived from Fisher/Wright model of reproduction among organisms

with non-skewed low fecundity offspring distributions, considers bifurcating coalescent events or binary mergers. The more generalized models such as the Beta $(2-\alpha, \alpha)$ (Schweinsberg, 2003) and point-mass coalescent (Eldon and Wakeley, 2006) that model multiple merger coalescent events, are a major research focus in new developments of coalescent theory (Wakeley, 2013). These models are appropriate for organisms exhibiting both high fecundity and highly-skewed heavy-tailed offspring distributions, and may be a better null model for many species (Eldon and Wakeley, 2006), such as the Atlantic cod (*Gadus morhua*, Linneaus, 1758) a marine organism with high-fecundity and opportunity for very high variance of offspring numbers due to its life history traits (Árnason, 2004).

The *Pan* I locus in Atlantic cod, which has been used widely as a marker for population genetics analysis, shows strong differentiation among populations at different geographic scales (Fevolden and Pogson, 1997; Pogson et al., 2001; Pogson and Fevolden, 2003; Karlsson and Mork, 2003; Árnason et al., 2009). The locus is acknowledged to be under selection or to be linked to selected loci (Fevolden and Pogson, 1997; Case et al., 2005). The continued use of this selected locus with the aim of identifying local populations at the very least begs advancing the knowledge on the evolutionary forces at work at the locus. Thus Nielsen et al. (2007) recommend that it not be used at microgeographical scale until more knowledge on evolutionary drivers is attained.

Pogson (2001) characterized a 1.85 kilo base (kb) region of the *Pan* I locus in Atlantic cod, and showed that it harbors an ancient polymorphic *Dra*I restriction site that through absence or presence defines alternative alleles or haplogroups, *A* and *B* respectively. The two alleles are maintained by balancing selection and are highly divergent at the nucleotide and amino acid level. Both alleles also show signs of selective sweeps within the functional types (Pogson, 2001). The two types differ by four amino acids representing six amino acid replacement mutations fixed between the lineages (three in each lineage). The amino acids differences reside in the first intra-vesicular loop domain (IV1) of the protein. The *Dra*I site defining the alleles and various other restriction sites in the region show high linkage disequilibrium (LD). There is LD with sites defining a 5.7 kb restriction fragment around the *Pan* I locus (Pogson, 2001).

There is a strong correlation between *Pan* I allele frequency and environmental settings (coastal vs offshore) of different depth. The *A* allele is found in higher proportions at coastal/shallow-water locales, and the *B* alleles at offshore/deep-water locales (Pogson and Fevolden, 2003; Case et al., 2005; Árnason et al., 2009). A very steep gradient is found of allele frequency, a change of 0.4% per meter of depth down to about 200 m (Árnason et al., 2009). This bears on the association found between the *Pan* I locus and Atlantic cod behavioral ecotypes defined using data storage tags (Pálsson and Thorsteinsson, 2003; Pampoulie et al., 2008). The ecotypes exhibit either a shallow-water behavior characterized by seasonal temperature trends (stationary cod), or deep-water behavior characterized by frequent vertical migrations and steep temperature changes possibly representing foraging at thermal fronts (migratory cod) (Pálsson and Thorsteinsson, 2003; Pampoulie et al., 2008; Thorsteinsson et al., 2012) . *Pan* I genotypes of Atlantic cod in relation to depth show that *AA* individuals have a shallow water behavior and *BB* a deep water behavior, however, seeking shallower waters during spawning. *AB* individuals show a mixed behavior (Pampoulie et al., 2008) somewhat intermediate between the homozygotes. The different cod ecotypes and their associated *Pan* I genotypes share their depth range during spawning (Pampoulie et al., 2008; Árnason et al., 2009). However, possible segregation may occur by the behavioral differences of the ecotypes in spawning habitat-selection (Grabowski et al., 2011). Thus the markedly divergent *A* and *B* alleles can be roughly classified as shallow and deep-water adapted types respectively. This presents a strong parallel between a genomic island and ecological divergence associated to *Pan* I.

Strong heterogeneity in levels of differentiation in different parts of the genome have revealed genomic islands of divergence (Wu, 2001; Renaut et al., 2013; Ruegg et al., 2014; Cruickshank and Hahn, 2014) in Atlantic cod. The *Pan* I locus is located within one of such genomic islands of divergence (Bradbury et al., 2013; Hemmer-Hansen et al., 2013; Karlsen et al., 2013). The genomic islands represent a non-random distribution of levels of divergence in the Atlantic cod genome, and are constituted by clusters of genomic regions of elevated divergence running within several linkage groups and with co-occurrence of loci most likely implicated in selective processes (Bradbury et al., 2013). The genomic island containing *Pan* I is in linkage group (LG) 1 and has been found to be linked to the aforementioned *Pan* I ecotypes (Hemmer-Hansen et al., 2013). Thus *Pan* I and loci co-occurring at the same genomic island are likely to be functionally related to the capability of the organism to thrive in different environments with complex differences based upon a discrete region of genomic divergence composed by multiple linked loci.

Different environments entail multidimensional differences that must be met by organisms inhabiting

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.956v1 | CC-BY 4.0 Open Access | rec: 2 Apr 2015, publ: 2 Apr 2015

those environments. When alternative forms of an organism inhabit different environments, divergent selection may be involved in building supergenes or switch-genes (see e.g. Thompson and Jiggins, 2014). Supergenes are genomic architectures of multiple, functional, co-adapted loci in tight linkage and little recombination by a variety of mechanisms. The polymorphic variants segregate together in particular combinations of alleles as if they were a single locus. Often those variants are kept at intermediate frequencies due to a balance of selective vectors (Thompson and Jiggins, 2014) as in butterfly mimicry. The latter particular combinations of alleles at co-adapted loci, which are reflected on complex phenotypes, allow alternative forms of an organism to meet the multidimensional challenges of particular habitats, with no maladaptive intermediate combinations thanks to little recombination and tight linkage among loci (Thompson and Jiggins, 2014). *Pan* I has shown highly divergent alleles likely maintained by balancing selection and the region shows considerable LD (Pogson, 2001). The question arises how far those influences extend into the neighboring loci of *Pan* I? How tight-knit is the LD within *Pan* I and its surrounding loci? High divergence, tight-knit LD among multilocus variants, the implication of balancing selection, and suggested functional correlation among loci would present conditions for the build up of a supergene structure. If the 20 cM genomic Island of divergence at LG1 (Hemmer-Hansen et al., 2013) represents a supergene of co-adapted complexes there may of course be multiple sites of epistatic interactions throughout this genomic island.

The *Pan* I locus has been linked to selective forces such as temperature and salinity (Case et al., 2005) and fisheries (Árnason et al., 2009). Although we do not have evidence of epistatic effects, the function of the proteins coded by *Pan* I and flanking loci suggests correlation at functional level. Knowledge of function is essential to understand the working of selection. The *Pan* I gene codes for pantophysin, a microvesicle membrane protein involved in transport events (Haass et al., 1996), specifically in the trafficking of the insulin-regulated glucose transporter GLUT4 (see reviews in Bradley et al., 2001; Larance et al., 2008). The loci on either side of *Pan* I are *Sort*1 and *Atxn7l2* (Star et al., 2011). *Sort*1 codes for sortilin, a protein that also is a major component of Glut4-containing microvesicles and that might be involved in the translocation or biogenesis of the Glut4-containing vesicles (Lin et al., 1997). Sortilin participates in trafficking processes at the Golgi apparatus and plasma membrane (Strong et al., 2012). *Atxn7l2* codes for ataxin 7-like 2, a protein that contains SCA7, a zinc-binding domain that binds with TFTC/STAGA subunits (Marchler-Bauer et al., 2012). TFTC/STAGA are histone acetyltransferase–containing coactivator complexes (Helmlinger et al., 2006) which are implicated in chromatin remodeling (Zhao et al., 2008). Chromatin remodeling is a response mechanism to environmental stressors such as temperature or salinity (de Nadal et al., 2011). From these considerations we speculate that the proteins of these genes may be important for burst energy metabolism (anaerobic glycolysis). Thus the nearest neighbors of *Pan* I may be functionally related to *Pan* I, with *Sort*1 products acting in the same transport vesicles and *Atxn7l2* products responding to similar environmental pressures.

If the *Pan* I locus is influenced by the dynamics of a selected linked locus the neighboring loci *Sort*1 and *Atxn7l2* are a reasonable location to consider for analysis of LD and selective effects. Yet, even if strong LD and signatures of selection were to be found at those loci, the question still remains what is the target of selection. A simple metric for elucidating the action of selective forces at a locus is the proportion of codons with multiple non-synonymous substitutions. Due to the nature of the genetic code, some amino acid substitutions are not accessible for certain codons. Instead a change from one amino acid to another may have to go through intermediate states, accumulating non-synonymous substitutions at the codons in the process. Environmental changes can create bursts of non-synonymous substitutions or evolutionary bursts (Gillespie, 1984). DNA sequence studies have pointed to areas suspected of being under such selective bursts (Bazykin et al., 2004, 2006). We apply this metric by comparing the proportion of codons hit by multiple non-synonymous mutations at *Pan* I and its flanking regions, segments of the (*Sort*1) and *Atxn7l2* loci.

In this paper we study detailed nucleotide variation of a 12.6 kb region within the genomic island containing the *Sort*1, *Pan* I, and *Atxn7l2* loci. The main questions we adress are: Do the signals of selective effects already known at *Pan* I extend to its neighboring genes or is there a peak signal at *Pan* I? How tightly knit is the LD in the region and what role does it have concerning selective effects? Can we detect a focal point of selection within this region?

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.956v1 | CC-BY 4.0 Open Access | rec: 2 Apr 2015, publ: 2 Apr 2015

MATERIALS AND METHODS

Sampling

We selected 31 individual Atlantic cod genomic DNA samples for genotyping, cloning and sequencing, by stratified random sampling on geographic regions of the species distribution spanning east to west and north to south. The samples come from our large sample laboratory database of greater than 20,000 individuals. All localities are represented with at least 100 individuals (except the White Sea with 24 individuals). The geographic regions were the waters of Newfoundland (New), Iceland (Ice), Faroe Islands (Far), Norway (Nor), the North Sea (Nse) and the Celtic Sea (Cel) (6, 5, 5, 6, 6, and 3 samples, respectively). We isolated DNA from gill tissue using a Chelex/Proteinase K digestion(Walsh et al., 1991). We included a specimen of the closely related Pacific cod *G. macrocephalus* (Tilesius, 1810) as an outgroup.

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

Molecular analysis

We genotyped the individuals at the *Pan* I locus in the manner described by Árnason et al. (2009). The locus has two alleles or haplogroups, *A* and *B*, corresponding to the absence or presence of a *Dra*I restriction site (Fevolden and Pogson, 1997). This site is in LD with several other sites of the *Pan* I locus (Pogson, 2001). If the individual was heterozygous *A/B* for the *Pan* I we chose one allele at random from that individual for this study.

We amplified two fragments 4.3 kb and 8.7 kb in length (fragment I and fragment II respectively) from genomic DNA using Long PCR Enzyme Mix (Fermentas) in a Tetrad2 (MJ Research). The fragments had a 489 base pair (bp) overlap that contains the polymorphic *Dra*I restriction site defining the *A* and *B* alleles of *Pan* I. The merged sequence of both fragments resulted in 12.5 kb sequence (Figure S1). We had previously 454-sequenced BAC clones with about a 150 kb insert which contained the *Pan* I gene as well as a number of other genes (and see Árnason and Halldórsdóttir, 2015). With this sequence at hand, we designed primers to PCR amplify and sequence the 12.5 kb fragment containing the full *Pan* I gene. The neighboring genes were sortilin 1 (*Sort*1) and ataxin 7-like 2 (*Atxn7l2*) which we partially covered. Subsequently, the cod genome was released (Star et al., 2011) and confirmed our BAC clone sequence. We have used the genomic sequence (`www.ensembl.org`) and comparative data from various species ranging from humans to fish to determine the exon/intron structure of our 12.5 kb fragment.

We used primer 3 (Pogson and Mesa, 2004) and primer sc343pr66398 to amplify fragment I, and primer 20 (Pogson and Mesa, 2004) and primer sc343pr79421 to amplify fragment II (Figure S1 and Table S1). The PCR was an initial denaturation step of 2 minutes at 94°C; followed by 10 cycles of 15 seconds denaturation at 94°C, 30 seconds annealing at 50.4°C, 9 minutes 18 seconds (for fragment I) and 4 minutes 36 seconds (for fragment II) elongation at 68°C. This was followed by 25 additional cycles, increasing the elongation time by 10 seconds every cycle.

We purified fragments for cloning by agarose gel purification. We loaded 40 μl of the PCR amplified products in a 0.8% agarose gel (1× TAE buffer) with crystal violet (1.6 μg/ml) and electrophoresed at 80 volts/cm for 56 minutes using a 1 kb DNA ladder (Fermentas) as reference. We excised the gel

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.956v1 | CC-BY 4.0 Open Access | rec: 2 Apr 2015, publ: 2 Apr 2015

pieces under visible light with the DNA bands of interest, froze and thawed, and used the resulting DNA suspension directly for cloning.

We TOPO-TA cloned fragments into vector pCR-XL-TOPO (Invitrogen) following the manufacturer's instructions except we used 1/7 of recommended amount of vector. After cloning, we isolated plasmid DNA using alkaline lysis minipreps (Birnboim and Doly, 1979). We miniprepped five clones per individual. We confirmed that the clones contained fragments of the size of interest, by EcoRI digestion and agarose gel electrophoresis.

We genotyped the clones for the *Pan* I *A* and *B* alleles. For each individual, we took three clones from the same allele for sequencing, thus sequencing 93 clones $\times$ 2 (three clones for both fragments per each of the 31 individuals). We sequenced with overlapping primers (Figure S1 and Table S1) using BigDye Terminator kit (ABI) and manufacturer's protocol except we used 1/16 of recommended amount of TRR. We purified reaction products with EtOH precipitation, resuspended in HiDi Formamide and ran on an ABI 3500xL genetic analyzer (ABI). The area we sequenced comprised 3 loci: the entire *Pan* I locus, and partial segments of *Atxn7l2* and *Sort*1. *Pan* I is located medially; *Atxn7l2* is located downstream and on the same strand as *Pan* I; and *Sort*1 is located upstream, and on the opposite strand of *Pan* I (Figure 1).

For comparative purposes, we sequenced three clones of a Pacific cod (*Gadus macrocephalus*) individual, covering the same region as fragment I (4.3kb) in Atlantic cod (partial segments of *Pan* I and *Atxn7l*). We used the same methods as with Atlantic cod. We could not get fragment II to amplify for Pacific cod.

Data analysis

We base-called, assembled and visually inspected 1836 DNA reads into 93 sequences 12.5 kb in length (a 12.5 kb sequence per clone) using the software suite `Phred-Phrap-Consed` (Ewing et al., 1998; Gordon et al., 1998; Green, 1994). For each of the 31 individuals, we aligned its three clone sequences using `MUSCLE` (Edgar, 2004) to edit and build a consensus sequence for each individual. We used the `R` language and environment (R Development Core Team, 2008) with the `APE` (Paradis et al., 2004), `Pegas` (Paradis, 2010b) and `SeqinR` (Charif and Lobry, 2007) packages and in-house functions to manipulate the clone sequence alignments and build 31 individual consensus sequences. These are phased haplotypic data. We applied the same procedures on 18 DNA reads from three clones of a Pacific cod individual to obtain a consensus sequence 4.3 kb in length.

PCR errors inevitably are found in clones by cloning PCR products. We consider that taking three clones is sufficient to eliminate PCR errors among the clones. We assume that the three clones do not share a PCR error site (Árnason and Halldórsdóttir, 2015). Two of the clones from each individual will be of the same chromosome. The third clone will be of the same chromosome with probability 1/2 and of the alternative chromosome with probability 1/2 (Árnason and Halldórsdóttir, 2015). The consensus sequence will eliminate PCR errors except in the rare cases in which PCR errors in one of the two clones from the same chromosome has hit a site which is polymorphic in the population and found in the third clone derived from the alternative chromosome. A small bias may be introduced by this. However, this will be seen as recombination and we would in such cases err on the conservative side in interpretation based on LD.

Since we sequenced two fragments for each individual and merged them (Figure S1), we had to address the possibility of inadvertently forming chimeras where the consensus sequence of one fragment corresponds to the alternative allele relative to the other fragment. To investigate if this was an issue we aligned and contrasted the 489 bp overlap in both fragments and checked that the polymorphisms in both fragments produced the same consensus sequence. Doing this we detected five possible chimeric individual sequences in our first overview of the data. For each of them, we replaced the 4.3 kb consensus sequence from fragment I with the sequence of the single clone in fragment I that was in phase with fragment II. This means that for these five individual sequences fragment I would have PCR errors in these cases that would appear as excess of singleton variable sites. We randomly eliminated the excess of singletons (corresponding to PCR errors) in these five clone sequences to obtain the same average number of singletons as in the 4.3 kb region of non-chimeric individuals. Therefore, the remaining singletons were scattered at random, in agreement with the nature of mutations.

We aligned the 31 individual consensus sequences with the Fast Statistical Alignment (`FSA`) program (Bradley et al., 2009), and visually inspected this alignment with `Seaview` (Galtier et al., 1996). We manually edited a few indel sites where `FSA` had made obvious errors. We used the alignment with

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.956v1 | CC-BY 4.0 Open Access | rec: 2 Apr 2015, publ: 2 Apr 2015

SNiplay (Dereeper et al., 2011) for SNP detection. We used the R ade4, adegenet LDheatmap, and popgen packages (Jombart and Ahmed, 2011; Shin et al., 2006; Marchini, 2013; Dray and Dufour, 2007) and various functions written by us for managing, analyzing, and plotting the data. We used Genomicus/Phyloview (Louis et al., 2013) to produce a multi-species comparative display in genomic context of our sequenced region and surrounding loci, showing ortholog and paralog genes (Figure S2). We used the snpposi functions of the adegenet package (Jombart and Ahmed, 2011) to plot and test the density of SNPs over the fragment. We used Arlequin v. 3.5 (Excoffier and Lischer, 2010) for analysis of molecular variance, AMOVA, and analysis of population differentiation with pairwise F_{ST}.

We generated folded and unfolded site frequency spectra using R with the package pegas Paradis (2010a) and with in-house functions. We compared observed spectra to expectation of Kingman coalescent (θ/i) (Kingman, 1982) and to multiple merger Beta $(2-\alpha, \alpha)$ (Schweinsberg, 2003) and point-mass coalescent (Eldon and Wakeley, 2006) using inference methods developed by Birkner et al. (2013). We used software from Bjarki Eldon (http://page.math.tu-berlin.de/~eldon/programs.html) to estimate various parameters of the multiple merger coalescents (and see Árnason and Halldórsdóttir, 2015). We carried out Tajima's *D*, Fu and Li's, and McDonald-Kreitman neutrality tests, analysis of recombination, and computed statistics of polymorphism and divergence with DNAsp (Rozas et al., 2003).

We performed Hudson-Kreitman-Aguadé (HKA) test (Hudson et al., 1987) in direct mode with DNAsp and also with a sliding window using DNA Slider (McDonald, 1998). For the HKA tests, we trimmed the ends of fragment I sequence alignment (resulting in 4.2 kb) of Atlantic and Pacific cod due to low phred scores of the Pacific cod sequences at the ends of the fragment. For the HKA test in direct mode we contrasted DNA sequences (31 and one, respectively) of Atlantic and Pacific cod at the miochondrial cytochrome *b* locus (data from Árnason, 2004) with the 4.2 kb region (fragment I) covering the *Pan* I and *Atxn7l2* loci partially. For the HKA test with sliding window we used the 4.2 kb alignment (fragment I) of Atlantic and Pacific cod sequences (31 and one, respectively) with silent polymorphisms and fixed differences, windows of 31 and 33 variable sites for the largest average and maximum sliding *G* value, and 100 replications. Also, for the HKA test we used 12.5 kb region to contrast all *Pan* I^A against*Pan* I^B as outgroup, and vice versa, using all polymorphisms and fixed differences, and 100 replications. We also performed the maximum likelihood HKA test using the MLHKA program (Wright and Charlesworth, 2004). For this test we used the data on Hemoglobin α 2 *HbA2* and Myoglobin *Myg* loci of Árnason and Halldórsdóttir (2015) with the 4.2 kb alignment (fragment I) of Atlantic and Pacific cod sequences for *Pan* I.

GenBank accession numbers for sequences reported in this paper are KR011783–KR011814.

RESULTS

Nucleotide variability

Extensive LD exists between various sites of the *A* and *B* alleles as already observed by Pogson (2001). We, therefore, decided to analyze and present our data with reference to the *Dra*I site defining the *A* (25 sequences) and *B* (six sequences) alleles of *Pan* I. Examination of the data showed that the *Dra*I site defining the *A* and *B* *Pan* I difference was tied to differences at a larger scale.

We found maximum haplotype diversity with every sequence representing a different haplotype. Haplotype diversity was thus not informative about differences at this level of sampling. The levels of polymorphism were notably higher for *Pan* I^A sequence variants compared to *Pan* I^B (Table 1). $\hat{\pi}$ and $\hat{\theta}$ values of *Pan* I^A sequence variants were roughly 3 times and 5 times larger, respectively, than the same statistics for *Pan* I^B. Combined the sequence variants showed much higher levels of polymorphism than each group of sequence variant separately (Table 1). The levels of divergence ($\hat{K}$) between the *Pan* I sequence variants were higher than the levels of nucleotide diversity ($\hat{\pi}$) within both (Figure 2), throughout the region.

Heterozygosity per site among sequences classified according to the *Pan* I *A* alleles, sequences classified according to the *B* alleles, and for all sequences combined are shown in Figure S3. High heterozygosity was found throughout the region for both the *A* and *B* alleles. Nevertheless there were concentrations of high heterozygosity sites in some parts (e.g. around 1000 and 8500 for *A* alleles and 3200 for B alleles). There was high heterozygosity throughout the region for the combined data. Heterozygosity of 0.31 represented the fixed differences between sequences classified by the *Pan* I *A* and *B* alleles (25 vs six respectively). Although high heterozygosity was found throughout the region there was significant clustering of SNPs ($P = 0.006$, snpposi.test, Jombart and Ahmed, 2011) (Figure S4).

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.956v1 | CC-BY 4.0 Open Access | rec: 2 Apr 2015, publ: 2 Apr 2015

The maximum likelihood tree of the alleles showed two distinct lineages with the variation grouped according to the two *Pan* I allelic variants (Figure S5). The *Pan* I^A lineages showed higher sequence variability than the *Pan* I^B lineage. Figure S5 also showed that with with respect to the outgroup the *Pan* I^B lineage had evolved further than the *Pan* I^A lineage (Figure S6).

Tajima's $D = -0.72989$ and Fu and Li's $D^\star = -0.93532$ and $F^\star = -1.01658$ were non-significant for the overall region ($P > 0.10$), as well as for each of the loci separately. With a sliding windows approach (100 bp and 25 bp window and step size, respectively), we found a region between 10914 bp and 11038 bp, in *Atxn7l2*, with Fu and Li's $D^\star = -2.7105$ and $F^\star = -2.8126$ that deviate significantly from neutrality ($P < 0.05$). The McDonald-Kreitman test did not show a significant deviation from neutrality for the overall region or for each of the loci separately ($P > 0.10$, Fisher's exact test).

The HKA test with sliding windows indicated a significant deviation from neutrality only at the *Pan* I locus when considering the Atlantic cod *Pan* I^A sequence variants as ingroup and *Pan* I^B sequence variants as outgroup, but not at the *Atxn7l2* locus or when considering *Pan* I^A and *Pan* I^B sequence variants of Atlantic cod as ingroup compared against Pacific cod as outgroup. The HKA test in direct mode comparing Atlantic and Pacific cod at segments of *Pan* I and *Atxn7l2*, and at cytochrome *b*, did not indicate a significant deviation from neutrality. However, the maximum likelihood HKA analysis showed a significant HKA test ($P < 0.01$) with a selection parameter $k = 4.12$ indicative of balancing selection (Table S2).

Linkage disequilibrium, LD

We observed very strong LD among most of the high heterozygosity polymorphic sites (those with minor allele frequency 6/31 or more) of the three analyzed loci over the 12.56 kb region (Figure 3). Virtually the whole 12.56 kb region, that harbors the *Pan* I locus surrounded by partial segments of *Sort1* and *Atxn7l2*, is one LD block with maximum LD (measured by D') throughout the whole region. Very few polymorphic sites had LD values lower than maximum. However, there were notable exceptions. Three adjacent sites (sites number 8360, 8362, and 8364) were in full linkage equilibrium. There are three possible explanations for this phenomenon. First they might be due to sequencing error. We have gone over the data and found no evidence for error. Second, these may hypermutable sites. In that case the variants at these sites are not identical by descent. Third, this may be a recombination tract with the blocks on either side of that tract being held together in full LD by epistatic interactions.

Measures of LD depend on allele frequencies (Hedrick, 1987) and in general no measure is independent of allele frequencies (Lewontin, 1988). Excluding only singleton sites the LD of sites with a minor allele frequency of 2/31 or more also showed large LD blocks (Figure S7). However, another recombination tract was observed having intermediate D' LD values in the intergenic region of *Sort1* and *Pan* I.

Considering the *A* allele sequences only the LD was much lower with evidence of extensive recombination among the *A* alleles (Figure S8, minor allele frequency set at 3/25). There was much less variation among *B* alleles and there were large blocks of LD but also recombination tracks with low LD (Figure S9, sites with minor allele frequency 2/6).

Population differentiation

Considering only variation for the 12558 bp region among the *Pan* I *A* alleles the AMOVA (Table S3) showed that most of the variation (84%) was within populations. On the basis of spatial patterns of variation at the *Ckma* gene Árnason and Halldórsdóttir (2015) observed a North (Canada, Iceland, and Norway) vs South (Faroe Islands, North Sea, and Celtic Sea) divide. Using this classification to group localities only 6% of the variation was among groups and 11% within groups. Only the within population variance component V_c and the associated F_{ST} fixation index was significant (Table S3). The lack of significance was probably to some extent due to small sample sizes but the size of the fragment counteracts that effect.

The pairwise F_{ST} of *A* allele variation between localities (Table S4) showed that Canada (Nova Scotia and Newfoundland combined) differed significantly from Norway, Faroe Islands, North Sea, and Celtic Sea. The differentiation of Canada and Iceland was a little over 1/3 that of Canada and the other localities but it was not significant. The only other significant difference was between Iceland and North Sea. There was no differentiation among any of the pairs of Norway, Faroe Islands, North Sea, and Celtic Sea with most F_{ST} having negative signs that are interpreted as nill. These patterns were also evident in the maximum likelihood tree of variation (Figure S10). One clade was confined to Canada but the Canadian samples were not, however, confined to that clade. In general individuals from most localities were widely

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.956v1 | CC-BY 4.0 Open Access | rec: 2 Apr 2015, publ: 2 Apr 2015

dispersed on branches of the tree. Overall the $F_{ST} = 0.09 \pm 0.02$ among *A* alleles for the North vs South areas defined in Árnason and Halldórsdóttir (2015). Thus differentiation at *Pan* I and peripheral regions could be described as an east vs west differentiation with Iceland intermediate. It did not fit the north vs south divide of Árnason and Halldórsdóttir (2015).

The sample contained only six *B* alleles, three from Iceland and three from Norway. All *B* carried the ∇_2 indel considered a sign of a selective sweep (Pogson, 2001). There were two clades among the *B* alleles that were defined by several sites in full LD (Figure S11). Both clades were found in both Iceland and Norway and were thus not geographically restricted. The $F_{ST} = -0.09$ ($P = 0.66 \pm 0.04$) for Iceland vs Norway comparison of *B* allele variants.

Allelic divergence

The $F_{ST} = 0.82$ ($P < 0.001$) between the *Pan* I *A* and *B* haplogroup variants. The average number of pairwise differences for the 12558 bp between the *A* and *B* alleles $D_{xy} = 442.7$, the average within allele difference $D_X = 97.4$, and the corrected pairwise difference $D_a = D_{xy} - (D_x + D_y)/2 = 384.2$. The net differentiation between the *A* and *B* alleles was thus 0.031 per nucleotide over the 12558 bp region (Table S5).

Considering the shorter 4194 bp fragment with *Gadus macrocephalus* as the outgroup the divergence was similar (Table S5).

Genomic aspects

The *Pan* I locus was comprised of 6 exons and 5 introns. We identified six exons in the segment of the *Atxn7l2* locus, and seven exons in the segment containing the *Sort*1 locus. .

There was a clear difference between sequences classified according to the two *Pan* I allelic variants. There were 121 fixed substitutions between *Pan* I allelic variants out of a total of 349 variable sites found in the entire region (Table 1 and Figure 1). Out of 121 fixed substitutions, eight were non-synonymous, seven were synonymous, and 106 were in non-coding regions. Six non-synonymous and three synonymous substitutions were at the *Pan* I locus, with two codons showing multiple (two) non-synonymous substitutions each and two other codons with one non-synonymous substitution each. Of those two codons with multiple non-synonymous substitutions (codons 61 and 64 in Table 5 of Pogson, 2001), one had C and A, and A and T nucleotides at the first and third position, respectively, for the *A* and *B* allelic sequence variants (codons CAA and AAT respectively). At this same codon most gadoid species sequenced by Pogson and Mesa (2004) including *G. macrocephalus/G. ogac* had A in both first and third position (codon AAA). The other codon, number 64, had A and G, and C and A nucleotides at the first and second codon position, respectively (codons ACC and GAC), for the *A* and *B* allelic sequence variants (Figure 1). At this same codon most gadoid species sequenced by Pogson and Mesa (2004) had G and A in the first and second position (codon GAC) whereas *G. ogac* (which Pogson, 2001, used as the outgroup for the *A* and *B* alleles of *Pan* I) and *G. macrocephalus* had an A in both the first and second position (codon AAC). Each of the *Sort*1 and *Atxn7l2* loci had one non-synonymous substitution and these loci had one and three synonymous substitutions, respectively (Figure 1). The fixed substitutions located furthest apart were 12088 bp apart (Figure 1).

We looked for and analyzed the *Dra*I, *BstE*II, *Bst*XI, *Pst*I, and *Sac*II restriction sites referred to by Pogson (2001). In his Figure 1 *BstE*II and *Pst*I are 5.7 kb apart on either side of the *Dra*I site and in strong LD with the *Dra*I site and with each other. The *Dra*I and *BstE*II sites were fixed substitutions between *Pan* I sequence variants and *Pst*I was polymorphic within *Pan* I^A sequence variants. However, we did not find a *BstE*II site 5′ to the *Dra*I site as observed by Pogson (2001). Instead, we found a *BstE*II site 3′ to the *Dra*I site at position 11308 in our sequence. It was also 3′ to the *Pst*I site at position 11257 in our sequence (Figure 1). This site was in perfect LD with the *A* and *B* *Pan* I alleles and thus behaves much like the *BstE*II site that Pogson (2001) observed.

Site frequency spectra and coalescent models

The unfolded site frequency spectrum of the 4.2 kb region of *Pan* I and *Atxn7l2* with Pacific cod used as the outgroup is in Figure 4. There were three peaks in the spectrum, singletons, sixtuplets, and twenty-five tuplets. The two latter peaks of 25 and 15 sites respectively represent the fixed differences between sequences classified according to the *B* and *A* *Pan* I alleles respectively. The Kingman coalescent model did not fit well. The Beta $(2-\alpha, \alpha)$ and the point-mass coalescent models gave a better fit, in particular

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.956v1 | CC-BY 4.0 Open Access | rec: 2 Apr 2015, publ: 2 Apr 2015

for the singletons. None of the coalescent models could account for the high frequency of sixtuplets and twenty-five tuplets.

The folded site frequency spectrum of the entire region (Figure 5; folded because we did not have an outgroup for the whole region) was bimodal, with peaks at singleton sites and combined sixtuplet and twenty-five tuplet in all sites (136 sites). This peak was almost as high as the singleton class. As was the case for the unfolded spectrum, the Kingman coalescent model gave the worst fit. Both the Beta $(2-\alpha,\alpha)$ and point mass coalescent models gave better fit to the data except for the high sixtuplet/twenty-five tuplet peak. None of the coalescent models of neutrality predicted the high peak at intermediate frequency.

The site frequency spectra of the *A* and *B* alleles separately (Figure S12) were unimodal. Again the Kingman coalescent model did not fit the data well whereas the Beta $(2-\alpha,\alpha)$ and point-mass coalescent gave significantly better fits (Table S6).

The parameter estimates for the Beta $(2-\alpha,\alpha)$ and point mass multiple-merger coalescent models are in Figure S13. The α parameter for the *A* and *B* alleles were similar to those for the *Myg* and *HbA2* genes (Árnason and Halldórsdóttir, 2015) slightly above 1.0. However, for the combined data $\alpha = 1.475$. A similar effect was observed for the ψ parameter which was similar ($\psi = 0.245$) for the sequences classified according to *Pan* I *A* alleles as for the *Myg* and *HbA2* loci. Sequences classified according to the *Pan* I *B* alleles had an even higher $\psi = 0.325$. For the combined data was considerably lower or $\psi = 0.100$.

DISCUSSION

Function of proteins

The *Pan* I codes for pantophysin, a protein whose function is involved in vesicle transport pathways in adipocytes, especially in the trafficking of insulin-regulated glucose transporter GLUT4 (reviewed in Bradley et al., 2001). Thus it is likely to be involved in energy metabolism, possibly burst activity. The allelic variants of *Pan* I have been associated to behavioral profiles with the *Pan* I^A allele connected to shallow waters and seasonal temperature changes while the *Pan* I^B allele is connected to deeper waters and steep temperature fluctuations (Pampoulie et al., 2008). The differences in *Pan* I^A allelic frequencies at different geographic scales have been connected to temperature and salinity Case et al. (2005). Sortilin is a major protein component of Glut4-containing microvesicles that might be involved in the translocation or biogenesis of the GLUT4-containing vesicles (Lin et al., 1997). Sortilin is also involved in trafficking processes at the Golgi apparatus and plasma membrane (Strong et al., 2012) whose expression is connected to hepatic reduction in triglycerides and to obesity (Ai et al., 2012). Thus it also seems involved in energy metabolism. Atxn7l2 codes for a protein involved in chromatin remodeling activities (Marchler-Bauer et al., 2012; Zhao et al., 2008). Chromatin dynamics have been documented to act as a control of gene expression and show a response to stress episodes mediated by e.g. temperature or salinity (de Nadal et al., 2011), the very same drivers that *Pan* I has been linked to Case et al. (2005). The attributes of the proteins thus suggest on one hand co-location and shared metabolic pathways of Glut4-containing vesicles for pantophysin and sortilin, and on the other hand shared correlations to steep fluctuations in temperature or depth-related environmental vectors among pantophysin and ataxin-l2.

Allelic divergence and spatial differentiation

There is a deep divergence of the *A* and *B* alleles only a little less than the divergence of Atlantic cod and Pacific cod. This is in line with results of Pogson and Mesa (2004) who found that the *A* and *B* split predated the divergence of Atlantic cod and Walleye pollock *Gadus chalcogrammus*. Using mitogenomic data Coulson et al. (2006) date the Atlantic cod vs. Pacific cod split at 4 mya and the Atlantic cod vs. Walleye pollock split at 3.8 mya. Accordingly the *A* vs *B* divergence in between those date, perhaps 3.9 mya. However, these dates are based on the Kingman coalescent. Times scales under the more appropriate multiple merger coalescents considered here may be considerably shorter (Árnason and Halldórsdóttir, 2015). Furthermore, if the *A* and *B* divergence is driven by repeated selective sweeps within each haplotype Pogson (2001) and strong selection time may be shorter.

Árnason and Halldórsdóttir (2015) considered as one possible explanation a historical hypothesis of ancient isolation and recent admixture for the *Ckma* gene in Atlantic cod. Their evidence did not support the historical hypothesis. We can use our results to shed further light on the issue. The *Ckma* gene shows large differentiation between a region that Árnason and Halldórsdóttir (2015) called South (Faroe Islands, North Sea, Baltic Sea, Celtic Sea, and Irish Sea) and North (Canada, Greenland, Iceland, Norway, Barents

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.956v1 | CC-BY 4.0 Open Access | rec: 2 Apr 2015, publ: 2 Apr 2015

Sea, and White Sea) with highly significant pairwise $F_{ST} \approx 0.8$ between North and South localities and no differentiation between localities within each region. Pogson and Fevolden (2003) devised a test of the historical vs selection hypothesis (Árnason and Pálsson, 1996b) of coastal vs North East Arctic cod in northern Norway. They stated that patterns of neutral variation within the *A* allelic class of *Pan* I would be a sensitive indicator of the historical hypothesis. Isolation and admixture are part of the breeding structure of a population with genome-wide effects (Wright, 1931). Different genomic regions should be concordant in their behavior (Bernardi et al., 1993) both neutral genes under random drift and selective genes. However, supposedly neutral variation within the *A* haplogroup of *Pan* I and neighboring loci is not congruent with the North vs South divide considered by Árnason and Halldórsdóttir (2015). Instead the differentiation is more east vs west. Thus spatial differentiation in Atlantic cod probably is primarily driven by selection (c.f. Bradbury et al., 2010) and not by history.

Balancing selection

Our evidence strongly suggests selective effects at the *Pan* I locus and its peripheral regions, partial segments of the *Sort*1 and *Atxn*7*l*2 loci. Our evidence also points to the *Pan* I locus as one target of selection.

The patterns of distribution of polymorphism at both site frequency spectra clearly indicate departure from neutrality and the action of balancing selection. The patterns that we find in these 4.2 kb and 12.5 kb regions are in agreement with the findings of Pogson (2001) at a 1.85 Kb region of *Pan* I. The *Pan* I locus contains an ancient polymorphism undergoing a mixture of directional and balancing selection that has maintained two highly differentiated alleles (Pogson, 2001). In the unfolded site frequency spectrum covering 4.2 kb, the signature of balancing selection is in the form of high frequency sixtuplets and twenty-five tuplets that do not fit any of the theoretical expectations of neutrality. The high frequency peaks of the spectrum are at opposite frequencies (6/31 and 25/31) and correspond to the differentiation between the six *Pan* I^A and the 25 *Pan* I^B sequence variants (i.e. six and 25 4.2 kb sequences classified according to the long-lived polymorphism maintained by balancing selection as proposed by Pogson, 2001). For the whole 12.5 kb region, the folded site frequency spectrum also exhibits the signature of balancing selection as the conflated sixtuples/twenty-five tuples peak clearly surpasses all theoretical expectations. It is almost as common as the singleton class. The unfolded and folded spectra accord well with each other. In the unfolded spectrum, the sixtuplets and twenty-five tuplets represent 40 fixed differences between *Pan* I^B and *Pan* I^A sequence variants (25 and 15 differences, respectively) over a 4.2 kb region in 31 sequences, i.e. 0.0095 fixed differences per site. In the folded one, the conflated sixtuples/twenty-five tuplets have a frequency of 136 differences over a 12.5 kb region in 31 sequences, i.e 0.0108 fixed differences per site. The site frequency spectra also match the phylogeny of alleles. The frequency of 25 and 15 sites of the sixtuplets and twenty-five tuplets respectively mean that the *Pan* I^B sequence variants are 10 sites further from the Pacific cod outgroup than are the *Pan* I^A sequence variants. Thus the *Pan* I^B have evolved further from the outgroup as seen in the phylogeny.

From the coalescent models, the better fit to the site frequency spectra that we observe with the Beta $(2-\alpha, \alpha)$ and point-mass coalescent than with the Kingman coalescent is in agreement with observations by Birkner et al. (2013) in mitochondrial DNA data (Árnason et al., 2000; Sigurgíslason and Árnason, 2003; Árnason, 2004) of Atlantic cod. The better fit to the site frequency spectra by the more generalized coalescent models than the Kingman coalescent is most likely because the last considers bifurcated coalescent events only, while the generalized models allow for multiple merger coalescent events and accommodate large variance in the number of offspring (Eldon and Wakeley, 2006). Thus, the better fit of the generalized coalescent models to the observed site frequency spectra, especially capturing the high frequency of singletons, may indicate large variance in offspring numbers in Atlantic cod and sweepstakes reproduction (Árnason and Halldórsdóttir, 2015). Such large variance is likely due to then the occurrence of frequent small and infrequent large offspring reproductive events as otherwise there would be no genetic variation (Árnason, 2004). The parameter estimate α represents the probability of large offspring events (i.e. large families), which is most likely as it approaches $\alpha = 1$. As α approaches $\alpha = 2$ the model behaves like the Kingman coalescent; the parameter ψ points to the proportion of reproduction that can be ascribed to an individual, where the model behaves like the Kingman coalescent when $\psi = 0$ and multiple merger coalescent events are predominant when $\psi = 1$ (Birkner et al., 2013). An indicator of balancing selection is that α is larger when we consider both allelic sequence variants *A* and *B* combined than when we consider each allelic variant separately; i.e when considering both *A* and *B* combined α

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.956v1 | CC-BY 4.0 Open Access | rec: 2 Apr 2015, publ: 2 Apr 2015

corresponds to a coalescent behavior that tends to longer coalescent times (with accumulated mutations) between alleles. In contrast, when we consider each allelic variant separately α has a lower value tending to faster coalescent times between alleles. From ψ, we also conclude that it indicates balancing selection as the combined data (both *A* and *B*) show ψ values that are lower and indicate longer coalescent times than when we consider each allelic sequence variant separately. These observations have a parallel on nucleotide variability in that the combined data have larger $\hat{\pi}$ and $\hat{\theta}$ than each allelic variant by itself. This is also an indication of balancing selection.

Concerning nucleotide variability, the signature of balancing selection is recognized in the patterns of silent divergence and polymorphism at *Pan* I and its neighboring genes. Balancing selection has been detected by looking at peaks of silent diversity among alleles that have experienced amino acid substitutions underlying adaptation to different environmental niches (Storz et al., 2007). Here, we see that the highly diverged sequence variants *A* and *B* (classified according to *Pan* I) have undergone amino acid replacements (six at *Pan* I and one at each of the neighboring regions) which strongly suggests functional differences in protein products. We have a series of peaks of silent divergence among sequence variants *A* and *B* larger than the polymorphism within each variant class. Thus the time to coalescence is longer and there are accumulated silent mutations among *A* and *B* sequence variants, and there is a shorter time to coalescence and less neutral mutations within each sequence variant class. This scenario fits in an iterative fashion that seen by Storz et al. (2007), who observed that the silent diversity among functional variants underlying altitude adaptation in deer mice was older and larger than the silent diversity segregating within such variants, signaling the works of diversifying selection(Storz et al., 2007). In this study, the larger levels of divergence among sequence variants *A* and *B* than the levels of polymorphism within each sequence variant separately is also evident in the larger measurements of $\hat{\theta}$ and $\hat{\pi}$ of the combined data than for each haplogroup separately. The same effect becomes apparent in the heterozygosity which is twofold and sixfold larger for the data combined than for the sequence variants classified according to the *Pan* I^A and *Pan* I^B alleles. Although these patterns point to selective effects at the *Pan* I gene and its neighboring genes, we can not rule out admixture (Bernardi et al., 1993).

The high LD observed throughout the entire 12.5 kb region is in agreement with the high LD detected by Therkildsen et al. (2013) in vast genomic tracts using genome scans with sets of gene-associated SNP genotyping. Our genomic horizon is narrower but at a higher resolution, revealing a very fine-grained LD at *Pan* I and flanking loci. The multiple peaks of divergence among sequence variants are a signal of balancing selection (Storz et al., 2007), and the iterative nature of the signal is strengthened by the fine-grained and high LD levels. As nearly all the variation is so tightly connected at *Pan* I and flanking loci, a signature of selection as peaks of divergence over relatively lower levels of polymorphism is seen repeatedly throughout the 12.5 kb region.

The summary statistics in general were non-significant, but some summary statistics for neutrality tests such as Tajima's *D* and Fu and Li's $D^\star$ and $F^\star$ are sensitive to low sample sizes (Guinand et al., 2004, revised in). Pogson (2001) has demonstrated that strong departures from neutrality are not necessarily reflected in test statistics.

The evidence indicates that there may have been bursts of non-synonymous substitutions at *Pan* I locus. This implies that *Pan* I locus is at least one focus of selection. The sites that contribute to the differentiation of the sequence variants (i.e. fixed differences among sequence variants) are more or less spread throughout the whole 12.5 kb region. However, a majority of the amino acid replacements sites that are fixed between the *A* and *B* haplogroups are located in the *Pan* I gene. They represent radical amino acid changes and probably lead to functional differences of the corresponding proteins (Pogson, 2001). The *A* and *B Pan* I alleles each have a codon that must have experienced multiple non-synonymous substitutions relative to the outgroup as already noted by Pogson (2001). Contrasting this codon to other gadoids sequenced by Pogson and Mesa (2004) suggests that for codon 61 (the codon numbers refer to Pogson and Mesa, 2004, sequence) *G. ogac* is the same as the ancestral allele (as Pogson, 2001, assumed) and that the *A* and *B* variants of *Pan* I in *G. morhua* both carry derived alleles. This implies adaptive substitutions in both the *A* and *B Pan* I Atlantic cod lineages. It also suggests that in codon 64 the *B* variant of *Pan* I in *G. morhua* carries the ancestral allele shared by most other gadoids while the *A* variant of *Pan* I and *G. ogac* both carry derived alleles at the first codon position, implying an older adaptive substitution, prior to the separation of *G. morhua* and the *G. macrocephalus/G. ogac* lineage (Coulson et al., 2006) . However, the *Pan* I *A* has an additional adaptive change in the second position of the codon. Thus *G. ogac* is an appropriate outgroup in codon 61 but not as a distant ancestor for codon 64.

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.956v1 | CC-BY 4.0 Open Access | rec: 2 Apr 2015, publ: 2 Apr 2015

These codons with multiple non-synonymous substitutions at *Pan* I constitute bursts of non-synonymous substitutions in the same lineage (Atlantic cod) and show that the differences between the *A* and *B* haplogroups occur as adaptive changes on both lineages. There is build-up of differences over time. This is an ancient balanced polymorphism (Charlesworth, 2006) and not simply a partial selective sweep bringing a particular chromosomal region to a high frequency (for example as seen in human G6PD and β globin polymorphisms Verrelli et al., 2002; Currat et al., 2002). This in conjunction with the number of other non-synonymous changes in the same region of the *Pan* I gene implies that the *Pan* I gene is most likely a focal point of selection. Due to the nature of the genetic code certain amino acid substitutions cannot occur without going through intervening amino acid states. Thus if selection favors an amino acid change of that type, there will be accumulation of non-synonymous mutations within respective codons. Thus codons with multiple non-synonymous mutations in the same lineage (what is referred to as evolutionary bursts Gillespie, 1984) are signals of focal points of selection (Bazykin et al., 2004, 2006). The intervening state for the codons we observe with multiple non-synonymous mutations are the corresponding codons observed in *G. ogac* by Pogson (2001), thus signaling the operation of balancing selection at *Pan* I in Atlantic cod. Hughes (2007) has criticized studies that look for the operation of selection under the criteria of concentration of amino acid replacements within a limited region. However, our main focus is not the concentration of amino acid replacements at a gene, but rather the occurrence of multiple codons with multiple non-synonymous substitutions at divergent haplogroups observed at a particular gene.

The entire 12.5 kb region of *Pan* I, *Sort*1), and *Atxn*7*l*2) is located is within a much larger region of a genomic island of divergence (Bradbury et al., 2013; Hemmer-Hansen et al., 2013). This genomic island of divergence has been connected (Hemmer-Hansen et al., 2007) to behavioral differences related to different habitat use with respect to temperature and depth regimes (Pampoulie et al., 2008). Recently, *Pan* I ecotypes in cod have been associated to polymorphisms at the Rhodopsin (*Rh*1) locus also located in LG1 as *Pan* I, with potential involvement of behavioral differences and visual capabilities as rhodopsin is a pigment involved in dim-light vision (Pampoulie et al., 2015). There is evidence for the build-up of the two haplogroups, two functionally balanced types, by selection as already stated. Neutral variation will accumulate on the two geneological lineages (Charlesworth et al., 2003). Utilization of different habitats with complex multidimensional differences may entail complex phenotypic differences with bearings on genomic structures known as supergenes (Thompson and Jiggins, 2014). The implication of balancing selection, the prevalence of divergence and of high levels of fine-grained LD, and a possible correlation in function suggested by protein function at the loci in the region, together hint the build up of a supergene inclusive of the region where *Pan* I and flanking segments of *Sort*1 and *Atxn*7*l*2 are located. The tightly knit LD throughout the 12.5 kb region is likely a product of the selective effects detected throughout the whole region, and seemingly little recombination among the *A* and *B* sequence variants classified by the *Pan* I alleles.

The effects of balancing selection at a single locus will extend only short distances from the selected sites with free recombination (Wiuf and Hein, 1999). Signs of a long standing balanced polymorphism therefore are the result of a build-up of co-adapted complexes of epistatic interactions among multiple sites or due to suppression of recombination (Wiuf and Hein, 1999). The very high LD observed here and the peculiar site frequency spectra with peaks at exactly opposite frequencies and no variation around the peaks (c.f. Árnason and Halldórsdóttir, 2015) imply very little recombination. We suggest that the genomic island of divergence is a supergene of co-adapted complexes possibly locked together by structural variation (Joron et al., 2011; Thompson and Jiggins, 2014). There may well be multiple selective sites within the genomic island. *Pan* I is very likely one such site.

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.956v1 | CC-BY 4.0 Open Access | rec: 2 Apr 2015, publ: 2 Apr 2015

REFERENCES

Ai, D., Baez, J. M., Jiang, H., Conlon, D. M., Hernandez-Ono, A., Frank-Kamenetsky, M., Milstein, S., Fitzgerald, K., Murphy, A. J., Woo, C. W., Strong, A., Ginsberg, H. N., Tabas, I., Rader, D. J., and Tall, A. R. (2012). Activation of ER stress and mTORC1 suppresses hepatic sortilin-1 levels in obese mice. *Journal of Clinical Investigation*, 122(5):1677–1687.

Árnason, E. (2004). Mitochondrial cytochrome *b* DNA variation in the high fecundity Atlantic cod: Trans-Atlantic clines and shallow gene-genealogy. *Genetics*, 166:1871–1885.

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

Árnason, E., Hernandez, U. B., and Kristinsson, K. (2009). Intense habitat-specific fisheries-induced selection at the molecular *Pan* I locus predicts imminent collapse of a major cod fishery. *PLoS ONE*, 4(5):e5529.

Árnason, E. and Pálsson, S. (1996a). Mitochondrial cytochrome *b* DNA sequence variation of Atlantic cod, *Gadus morhua*, from Norway. *Molecular Ecology*, 5:715–724.

Árnason, E. and Pálsson, S. (1996b). Mitochondrial cytochrome *b* DNA sequence variation of Atlantic cod *Gadus morhua*, from Norway. *Molecular Ecology*, 5(6):715–724.

Árnason, E., Petersen, P. H., Kristinsson, K., Sigurgíslason, H., and Pálsson, S. (2000). Mitochondrial cytochrome *b* DNA sequence variation of Atlantic cod from Iceland and Greenland. *Journal of Fish Biology*, 56(2):409–430.

Bazykin, G. A., Dushoff, J., Levin, S. A., and Kondrashov, A. S. (2006). Bursts of nonsynonymous substitutions in HIV-1 evolution reveal instances of positive selection at conservative protein sites. *Proceedings of the National Academy of Sciences of the United States of America*, 103(51):19396–19401.

Bazykin, G. A., Kondrashov, F. A., Ogurtsov, A. Y., Sunyaev, S., and Kondrashov, A. S. (2004). Positive selection at sites of multiple amino acid replacements since rat–mouse divergence. *Nature*, 429(6991):558–562.

Bernardi, G., Sordino, P., and Powers, D. A. (1993). Concordant mitochondrial and nuclear DNA phylogenies for populations of the teleost fish *Fundulus heteroclitus*. *Proceedings of the National Academy of Sciences of the United States of America*, 90(20):9271–9274.

Birkner, M., Blath, J., and Eldon, B. (2013). Statistical properties of the site-frequency spectrum associated with Λ-coalescents. *Genetics*, 195:1037–1053.

Birnboim, H. and Doly, J. (1979). A rapid alkaline extraction procedure for screening recombinant plasmid DNA. *Nucleic Acids Research*, 7(6):1513–1523.

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

Bradbury, I. R., Hubert, S., Higgins, B., Bowman, S., Borza, T., Paterson, I. G., Snelgrove, P. V. R., Morris, C. J., Gregory, R. S., Hardie, D., Hutchings, J. A., Ruzzante, D. E., Taggart, C. T., and Bentzen, P. (2013). Genomic islands of divergence and their consequences for the resolution of spatial structure in an exploited marine fish. *Evolutionary Applications*, 6:450–461.

Bradley, R., Cleveland, K., and Cheatham, B. (2001). The adipocyte as a secretory organ: mechanisms of vesicle transport and secretory pathways. *Recent progress in hormone research*, 56:329—358.

Bradley, R. K., Roberts, A., Smoot, M., Juvekar, S., Do, J., Dewey, C., Holmes, I., and Pachter, L. (2009). Fast statistical alignment. *PLoS Computational Biology*, 5(5):e1000392+.

Case, R. A. J., Hutchinson, W. F., Hauser, A. L., Van Oosterhout, A. C., and Carvalho, A. G. R. (2005). Macro- and micro-geographic variation in pantophysin (pan i) allele frequencies in NE Atlantic cod *Gadus morhua*. *Marine Ecology Progress Series*, 301:267–278.

Charif, D. and Lobry, J. R. (2007). SeqinR 1.0-2: a contributed package to the R project for statistical computing devoted to biological sequences retrieval and analysis. In Bastolla, U., Porto, M., Roman, H., and Vendruscolo, M., editors, *Structural approaches to sequence evolution: Molecules, networks, populations*, Biological and Medical Physics, Biomedical Engineering, pages 207–232. Springer Verlag, New York. ISBN : 978-3-540-35305-8.

Charlesworth, B., Charlesworth, D., and Barton, N. H. (2003). The effects of genetic and geographic structure on neutral variation. *Annual Review of Ecology, Evolution and Systematics*, 34:99–125.

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.956v1 | CC-BY 4.0 Open Access | rec: 2 Apr 2015, publ: 2 Apr 2015

Charlesworth, D. (2006). Balancing selection and its effects on sequences in nearby genome regions. *PLoS Genetics*, 2:e64.

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

Cruickshank, T. E. and Hahn, M. W. (2014). Reanalysis suggests that genomic islands of speciation are due to reduced diversity, not reduced gene flow. *Molecular Ecology*, 23:3133–3157.

Cunningham, K. M., Canino, M. F., Spies, I. B., and Hauser, L. (2009). Genetic isolation by distance and localized fjord population structure in Pacific cod (*Gadus macrocephalus*): Limited effective dispersal in the northeastern Pacific Ocean. *Canadian Journal of Fisheries and Aquatic Sciences*, 66:153–166.

Currat, M., Trabuchet, G., Rees, D., Perrin, P., Harding, R. M., Clegg, J. B., Langaney, A., and Excoffier, L. (2002). Molecular analysis of the β-globin gene cluster in the Niokholo Mandenka population reveals a recent origin of the β^S Senegal mutation. *American Journal of Human Genetics*, 70:207–223.

de Nadal, E., Ammerer, G., and Posas, F. (2011). Controlling gene expression in response to stress. *Nature Review Genetics*.

Dereeper, A., Nicolas, S., Le Cunff, L., Bacilieri, R., Doligez, A., Peros, J. P., Ruiz, M., and This, P. (2011). SNiPlay: a web-based tool for detection, management and analysis of SNPs. Application to grapevine diversity projects. *BMC Bioinformatics*, 12(1):134.

Dray, S. and Dufour, A. (2007). The Ade4 package: Implementing the duality diagram for ecologists. *Journal of Statistical Software*, 22:1–20.

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, 32(5):1792–1797.

Eldon, B. and Wakeley, J. (2006). Coalescent processes when the distribution of offspring number among individuals is highly skewed. *Genetics*, 172:2621–2633.

Ewing, B., Hillier, L., Wendl, M. C., and Green, P. (1998). Base-calling of automated sequencer traces using Phred. I. Accuracy assessment. *Genome Research*, 8(3):175–185.

Excoffier, L. and Lischer, H. E. L. (2010). Arlequin Suite Ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Molecular Ecology Resources*, 10(3):564–567.

Fevolden, S. E. and Pogson, G. H. (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.

Flicek, P., Amode, M. R., Barrell, D., Beal, K., Billis, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fitzgerald, S., Gil, L., Girón, C. G., Gordon, L., Hourlier, T., Hunt, S., Johnson, N., Juettemann, T., Kähäri, A. K., Keenan, S., Kulesha, E., Martin, F. J., Maurel, T., McLaren, W. M., Murphy, D. N., Nag, R., Overduin, B., Pignatelli, M., Pritchard, B., Pritchard, E., Riat, H. S., Ruffier, M., Sheppard, D., Taylor, K., Thormann, A., Trevanion, S. J., Vullo, A., Wilder, S. P., Wilson, M., Zadissa, A., Aken, B. L., Birney, E., Cunningham, F., Harrow, J., Herrero, J., Hubbard, T. J., Kinsella, R., Muffato, M., Parker, A., Spudich, G., Yates, A., Zerbino, D. R., and Searle, S. M. (2014). Ensembl 2014. *Nucleic Acids Research*, 42(D1):D749–D755.

Galtier, N., Gouy, M., and Gautier, C. (1996). SEAVIEW and PHYLOWIN: two graphic tools for sequence alignment and molecular phylogeny. *Computer Applications in the Biosciences*, 12(6):543–548.

Gillespie, J. H. (1984). Molecular evolution over the mutational landscape. *Evolution*, 38(5):1116.

Gordon, D., Abajian, C., and Green, P. (1998). Consed: a graphical tool for sequence finishing. *Genome Research*, 8(3):195–202.

Grabowski, T. B., Thorsteinsson, V., McAdam, B. J., and Marteinsdóttir, G. (2011). Evidence of segregated spawning in a single marine fish stock: Sympatric divergence of ecotypes in icelandic cod? *PLoS ONE*, 6(3):e17528.

Green, P. (1994). Documentation for Phrap. `http://bozeman.mbt.washington.edu/phrap.docs /phrap.html`.

Guinand, B., Lemaire, C., and Bonhomme, F. (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.

Haass, N. K., Kartenbeck, J., and Leube, R. E. (1996). Pantophysin is a ubiquitously expressed Synaptophysin homologue and defines constitutive transport vesicles. *Journal of Cell Biology*, 134:731–746.

Hedrick, P. W. (1987). Genetic disequilibrium measures: Proceed with caution. *Genetics*, 117:331–341.

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.956v1 | CC-BY 4.0 Open Access | rec: 2 Apr 2015, publ: 2 Apr 2015

Helmlinger, D., Hardy, S., Abou-Sleymane, G., Eberlin, A., Bowman, A. B., GansmÃŒller, A., Picaud, S., Zoghbi, H. Y., Trottier, Y., Tora, L., and Devys, D. (2006). Glutamine-expanded Ataxin-7 alters TFTC/STAGA recruitment and chromatin structure leading to photoreceptor dysfunction. *PLoS Biology*, 4(3):e67. 16494529[pmid].

Hemmer-Hansen, J., Nielsen, E. E., Frydenberg, J., and Loeschcke, V. (2007). Adaptive divergence in a high gene flow environment: *Hsc* 70 variation in the European flounder (*Platichthys flesus* L.). *Heredity*, 99:592–600.

Hemmer-Hansen, J., Nielsen, E. E., Therkildsen, N. O., Taylor, M. I., Ogden, R., Geffen, A. J., Bekkevold, D., Helyar, S., Pampoulie, C., Johansen, T., Consortium, F., and Carvalho, G. R. (2013). A genomic island linked to ecotype divergence in Atlantic cod. *Molecular Ecology*, 22:2653–2667.

Hudson, R. R., Kreitman, M., and Aguadé, M. (1987). A test of neutral molecular evolution based on nucleotide data. *Genetics*, 116(1):153–159.

Hughes, A. L. (2007). Looking for Darwin in all the wrong places: the misguided quest for positive selection at the nucleotide sequence level. *Heredity*, 99(4):364–373.

Jombart, T. and Ahmed, I. (2011). Adegenet 1.3-1: New tools for the analysis of genome-wide SNP data. *Bioinformatics*, 27:3070–3071.

Joron, M., Frezal, L., Jones, R. T., Chamberlain, N. L., Lee, S. F., Haag, C. R., Whibley, A., Becuwe, M., Baxter, S. W., Ferguson, L., Wilkinson, P. A., Salazar, C., Davidson, C., Clark, R., Quail, M. A., Beasley, H., Glithero, R., Lloyd, C., Sims, S., Jones, M. C., Rogers, J., Jiggins, C. D., and ffrench Constant, R. H. (2011). Chromosomal rearrangements maintain a polymorphic supergene controlling butterfly mimicry. *Nature*, 477:203–206.

Karlsen, B. O., Klingan, K., Emblem, Å., Jørgensen, T. E., Jueterbock, A., Furmanek, T., Hoarau, G., Johansen, S. D., Nordeide, J. T., and Moum, T. (2013). Genomic divergence between the migratory and stationary ecotypes of Atlantic cod. *Molecular Ecology*, 22(20):5098–5111.

Karlsson, S. and Mork, J. (2003). Selection-induced variation at the Pantophysin locus (*Pan* I) in a Norwegian fjord population of cod (*Gadus morhua*). *Molecular Ecology*, 12:3265–3274.

Kingman, J. F. C. (1982). The coalescent. *Stochastic Processes and their Applications*, 13:235–248.

Larance, M., Ramm, G., and James, D. E. (2008). The GLUT4 code. *Molecular Endocrinology*, 22(2):226–233.

Lewontin, R. (1974). *The Genetic Basis of Evolutionary Change*. Columbia University Press, New York, New York.

Lewontin, R. (1988). On measures of gametic disequilibrium. *Genetics*, 120:849–852.

Lin, B.-Z., Pilch, P. F., and Kandror, K. V. (1997). Sortilin is a major protein component of Glut4-containing vesicles. *Journal of Biological Chemistry*, 272(39):24145–24147.

Louis, A., Muffato, M., and Roest Crollius, H. (2013). Genomicus: five genome browsers for comparative genomics in eukaryota. *Nucleic Acids Research*, 41(D1):D700–D705.

Marchini, J. L. (2013). *popgen: Statistical and Population Genetics*. R package version 1.0-3.

Marchler-Bauer, A., Zheng, C., Chitsaz, F., Derbyshire, M. K., Geer, L. Y., Geer, R. C., Gonzales, N. R., Gwadz, M., Hurwitz, D. I., Lanczycki, C. J., Lu, F., Lu, S., Marchler, G. H., Song, J. S., Thanki, N., Yamashita, R. A., Zhang, D., and Bryant, S. H. (2012). CDD: conserved domains and protein three-dimensional structure. *Nucleic Acids Research*, 41(D1):D348–D352.

McDonald, J. H. (1998). Improved tests for heterogeneity across a region of DNA sequence in the ratio of polymorphism to divergence. *Molecular Biology and Evolution*, 15(4):377–384.

Muffato, M., Louis, A., Poisnel, C.-E., and Crollius, H. R. (2010). Genomicus: a database and a browser to study gene synteny in modern and ancestral genomes. *Bioinformatics*, 26(8):1119–1121.

Nielsen, E. E., MacKenzie, B. R., Magnussen, E., and Meldrup, D. (2007). Historical analysis of Pan i in Atlantic cod (*Gadus morhua*): temporal stability of allele frequencies in the southeastern part of the species distribution. *Canadian Journal of Fisheries and Aquatic Sciences*, 64(10):1448–1455.

Pálsson, Ó.. K. and Thorsteinsson, V. (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.

Pampoulie, C., Jakobsdóttir, K. B., Marteinsdóttir, G., and Thorsteinsson, V. (2008). Are vertical behaviour patterns related to the Pantophysin locus in the Atlantic cod (*Gadus morhua* L.)? *Behavior Genetics*, 38:76–81.

Pampoulie, C., Skirnisdóttir, S., Star, B., Jentoft, S., Jónsdóttir, I. G., Hj´orleifsson, E., Thorsteinsson,

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.956v1 | CC-BY 4.0 Open Access | rec: 2 Apr 2015, publ: 2 Apr 2015

V., Pálsson, Ó. K., Berg, P. R., Andersen, O., Magnusdóttir, S., Helyar, S. J., and Daníelsdóttir, A. K. (2015). Rhodopsin gene polymorphism associated with divergent light environments in atlantic cod. *Behavior Genetics*, 45(2):236–244.

Paradis, E. (2010a). Pegas: an R package for population genetics with an integrated–modular approach. *Bioinformatics*, 26:419–420.

Paradis, E. (2010b). pegas: an R package for population genetics with an integrated-modular approach. *Bioinformatics*, 26(3):419–420.

Paradis, E., Claude, J., and Strimmer, K. (2004). APE: analyses of phylogenetics and evolution in R language. *Bioinformatics*, 20:289–290.

Pogson, G. H. (2001). Nucleotide polymorphism and natural selection at the Pantophysin (*Pan* I) locus in the Atlantic cod, *Gadus morhua* (L.). *Genetics*, 157:317–330.

Pogson, G. H. and Fevolden, S. E. (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(1):63–74.

Pogson, G. H. and Mesa, K. A. (2004). Positive Darwinian selection at the Pantophysin (*Pan* I) locus in marine gadid fishes. *Molecular Biology and Evolution*, 21(1):65–75.

Pogson, G. H., Taggart, C. T., Mesa, K. A., and Boutilier, R. G. (2001). Isolation by distance in the Atlantic cod, *Gadus morhua*, at large and small geographic scales. *Evolution*, 55(1):131–146.

R Development Core Team (2008). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Renaut, S., Grassa, C. J., Yeaman, S., Moyers, B. T., Lai, Z., Kane, N. C., Bowers, J. E., Burke, J. M., and Rieseberg, L. H. (2013). Genomic islands of divergence are not affected by geography of speciation in sunflowers. *Nature Communications*, 4:1827.

Rozas, J., Sánchez-DelBarrio, J. C., Messeguer, X., and Rozas, R. (2003). DnaSP, DNA polymorphism analyses by the coalescent and other methods. *Bioinformatics*, 19:2496–2497.

Ruegg, K., Anderson, E. C., Boone, J., Pouls, J., and Smith, T. B. (2014). A role for migration-linked genes and genomic islands in divergence of a songbird. *Molecular Ecology*, 23(19):4757–4769.

Schweinsberg, J. (2003). Coalescent processes obtained from supercritical Galton-Watson processes. *Stochastic Processes and their Applications*, 106:107–139.

Shin, J.-H., Blay, S., McNeney, B., and Graham, J. (2006). LDheatmap: An R function for graphical display of pairwise linkage disequilibria between single nucleotide polymorphisms. *Journal of Statistical Software*, 16(c03):1–9.

Sigurgíslason, H. and Árnason, E. (2003). Extent of mitochondrial DNA sequence variation in Atlantic cod from the Faroe islands: a resolution of gene genealogy. *Heredity*, 91(6):557–564.

Star, B., Nederbragt, A. J., Jentoft, S., Grimholt, U., Malmstrom, M., Gregers, T. F., Rounge, T. B., Paulsen, J., Solbakken, M. H., Sharma, A., Wetten, O. F., Lanzen, A., Winer, R., Knight, J., Vogel, J.-H., Aken, B., Andersen, O., Lagesen, K., Tooming-Klunderud, A., Edvardsen, R. B., Tina, K. G., Espelund, M., Nepal, C., Previti, C., Karlsen, B. O., Moum, T., Skage, M., Berg, P. R., Gjoen, T., Kuhl, H., Thorsen, J., Malde, K., Reinhardt, R., Du, L., Johansen, S. D., Searle, S., Lien, S., Nilsen, F., Jonassen, I., Omholt, S. W., Stenseth, N. C., and Jakobsen, K. S. (2011). The genome sequence of Atlantic cod reveals a unique immune system. *Nature*, 477(7363):207–210.

Storz, J. F., Sabatino, S. J., Hoffmann, F. G., Gering, E. J., Moriyama, H., Ferrand, N., Monteiro, B., and Nachman, M. W. (2007). The molecular basis of high-altitude adaptation in deer mice. *PLoS Genetics*, 3(3):e45.

Strong, A., Ding, Q., Edmondson, A. C., Millar, J. S., Sachs, K. V., Li, X., Kumaravel, A., Wang, M. Y., Ai, D., Guo, L., Alexander, E. T., Nguyen, D., Lund-Katz, S., Phillips, M. C., Morales, C. R., Tall, A. R., Kathiresan, S., Fisher, E. A., Musunuru, K., and Rader, D. J. (2012). Hepatic sortilin regulates both apolipoprotein B secretion and LDL catabolism. *The Journal of Clinical Investigation*, 122(8):2807–2816.

Therkildsen, N. O., Hemmer-Hansen, J., Als, T. D., Swain, D. P., Morgan, M. J., Trippel, E. A., Palumbi, S. R., Meldrup, D., and Nielsen, E. E. (2013). Microevolution in time and space: SNP analysis of historical DNA reveals dynamic signatures of selection in Atlantic cod. *Molecular Ecology*, 22(9):2424–2440.

Thompson, M. J. and Jiggins, C. D. (2014). Supergenes and their role in evolution. *Heredity*, 113:1–8.

Thorsteinsson, V., Pálsson, Ó. K., Tómasson, G. G., Jónsdóttir, I. G., and Pampoulie, C. (2012). Consis-

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.956v1 | CC-BY 4.0 Open Access | rec: 2 Apr 2015, publ: 2 Apr 2015

tency in the behaviour types of the atlantic cod: repeatability, timing of migration and geo-location. *Marine Ecology Progress Series*, 462:251–260.

Verrelli, B. C., McDonald, J. H., Argyropoulos, G., Destro-Bisol, G., Froment, A., Drousiotou, A., Lefranc, G., Helal, A. N., Loiselet, J., and Tishkoff, S. A. (2002). Evidence for balancing selection from nucleotide sequence analyses of human *G6PD*. *American Journal of Human Genetics*, 71:1112–1128.

Wakeley, J. (2009). *Coalescent Theory: An Introduction*. Roberts & Company Publishers, Greenwood Village, Colorado, USA.

Wakeley, J. (2013). Coalescent theory has many new branches. *Theoretical Population Biology*, 87:1–4.

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

Wiuf, C. and Hein, J. (1999). Recombination as a point process along sequences. *Theoretical Population Biology*, 55:248–259.

Wright, S. (1931). Evolution in Mendelian populations. *Genetics*, 16:97–159.

Wright, S. I. and Charlesworth, B. (2004). The HKA test revisited: A maximum-likelihood-ratio test of the standard neutral model. *Genetics*, 168(2):1071–1076.

Wu, C.-I. (2001). The genic view of the process of speciation. *Journal of Evolutionary Biology*, 14(6):851–865.

Zhao, Y., Lang, G., Ito, S., Bonnet, J., Metzger, E., Sawatsubashi, S., Suzuki, E., Guezennec, X. L., Stunnenberg, H. G., Krasnov, A., Georgieva, S. G., Schüle, R., Takeyama, K.-I., Kato, S., LászlóTora, and Devys, D. (2008). A TFTC/STAGA module mediates histone H2A and H2B deubiquitination, coactivates nuclear receptors, and counteracts heterochromatin silencing. *Molecular Cell*, 29(1):92–101.

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.956v1 | CC-BY 4.0 Open Access | rec: 2 Apr 2015, publ: 2 Apr 2015

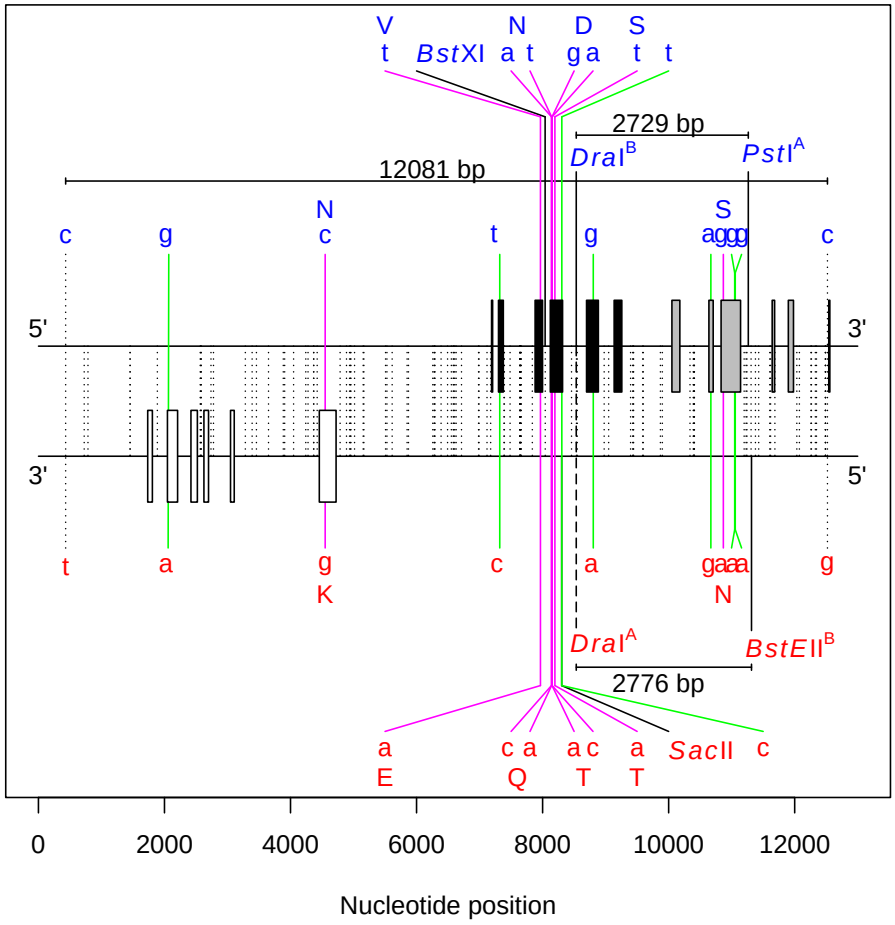

Figure 1. Map of polymorphism of the 12.5 kb region containing the *Pan* I locus and its peripheral regions, the *Sort*1 and *Atxn7l2* loci (partial segments). Boxes represent the exons of *Sort*1 (partial segment), *Pan* I and *Atxn7l2* (partial segment), in white, black and gray, respectively. Variation is displayed with respect to the *Dra*I site defining the *A* and *B* alleles of the *Pan* I locus (Pogson, 2001). The solid black horizontal lines running through the boxes represent introns (between boxes of the same color) and intergenic space (between boxes of different color). The polymorphic *Dra*I restriction site is represented with a solid and a dashed line for the *Dra*I^A and *Dra*I^B variants, corresponding to presence and absence of recognition site, respectively. *Pan* I^A and *Pan* I^B haplotypes are annotated in red and blue, respectively. Fixed non-synonymous and synonymous substitutions appear as solid vertical lines in magenta and green, respectively. Fixed substitutions in non-coding regions appear as vertical dotted lines between the solid horizontal lines except the outermost sites extend above and below. At the ends of the solid vertical lines, the substitution bases appear as lowercase letters, and the amino acid variants appear as uppercase letters (for non-synonymous substitutions). Restriction sites (cf. Pogson, 2001) appear as black vertical lines.

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.956v1 | CC-BY 4.0 Open Access | rec: 2 Apr 2015, publ: 2 Apr 2015

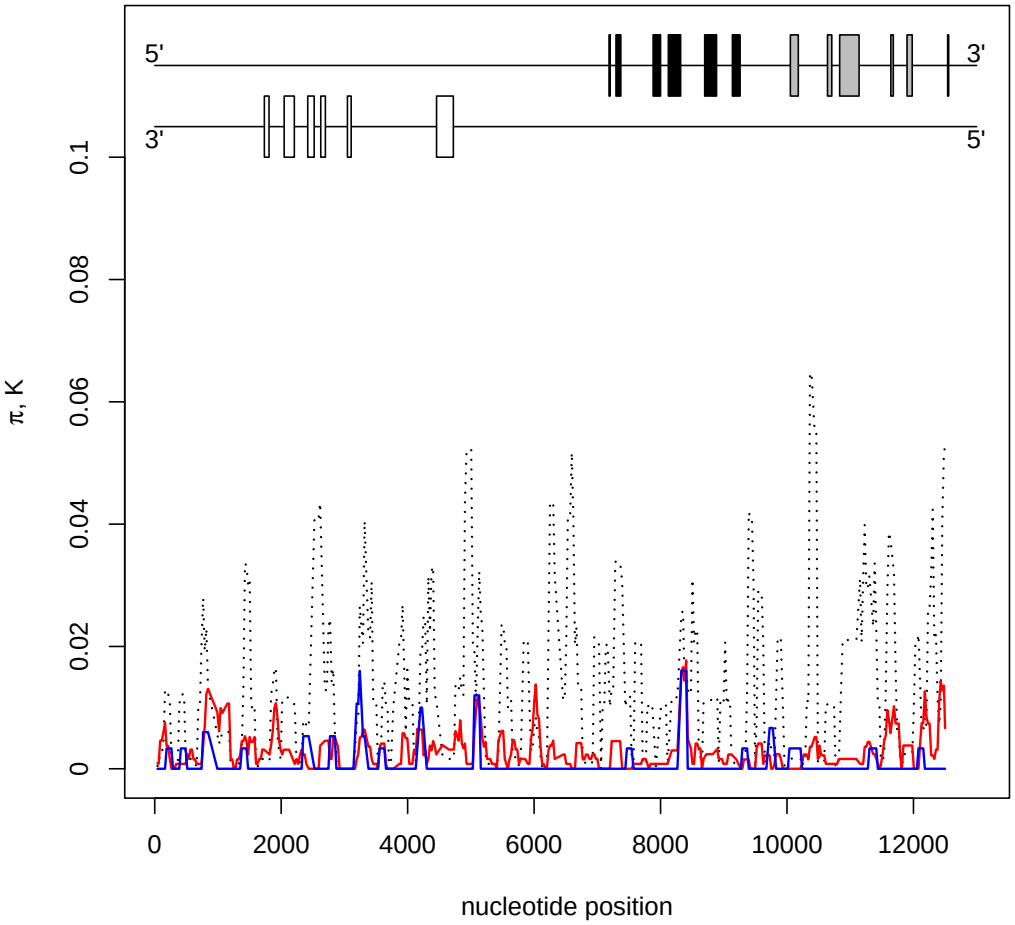

Figure 2. Polymorphism and divergence at the *Pan* I locus and its peripheral regions, the *Sort*1 and *Atxn7l2* loci (partial segments). Levels of polymorphism were calculated from silent, intronic, and intergenic sites, with a sliding window size of 100 bp and step size of 25 bp. Divergence ($\hat{K}$) between *Pan* I^A and *Pan* I^B allelic types is represented by a dotted line. Nucleotide diversity ($\hat{\pi}$) for *Pan* I^A and *Pan* I^B allelic types shown in red and blue, respectively. Boxes represent the exons of *Sort*1 (partial segment), *Pan* I and *Atxn7l2* (partial segment), in white, black and gray, respectively. The solid, black, horizontal lines running through the boxes represent introns (between boxes of the same color) and intergenic space (between boxes of different color).

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.956v1 | CC-BY 4.0 Open Access | rec: 2 Apr 2015, publ: 2 Apr 2015

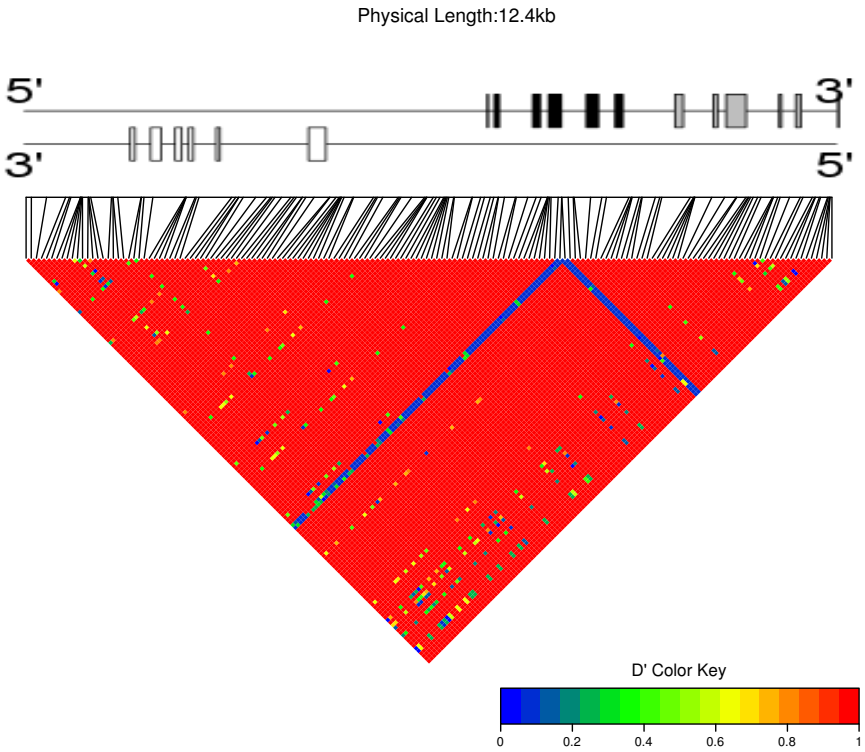

Figure 3. Linkage disequilibrium D' heatmap at high heterozygosity sites of the *Pan* I locus and its peripheral regions, the *Sort*1 and *Atxn*7*l*2 loci. Minor allele frequency set at 6/31, the frequency of the *B* alleles of the *Pan* I locus among the 31 samples.

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.956v1 | CC-BY 4.0 Open Access | rec: 2 Apr 2015, publ: 2 Apr 2015

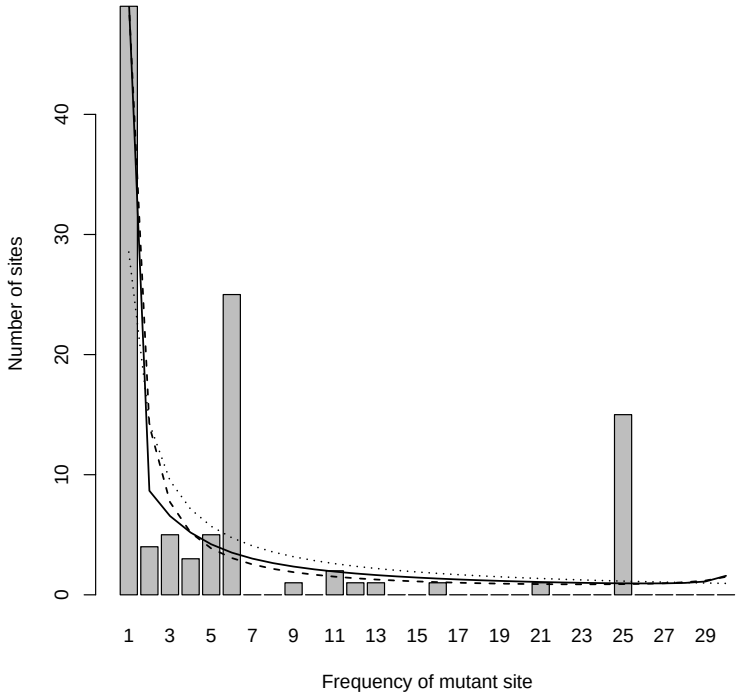

Figure 4. Unfolded site frequency spectrum of Atlantic cod *Pan I* and *Atxn7l2* genes. *Gadus macrocephalus* was used as the outgroup. Number of individuals $n = 31$. Theroretical expectation under Kingman coalescent (dotted line), Beta$(2-\alpha, \alpha$ coalescent (dashed line), and point-mass coalescent (solid line).The bars represent the observed spectrum. The spectrum represents the genetic variability from an alignment of 31 Atlantic cod sequences (25 *Pan* I^A and 6 *Pan* I^B) 4.2 kb long.

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.956v1 | CC-BY 4.0 Open Access | rec: 2 Apr 2015, publ: 2 Apr 2015

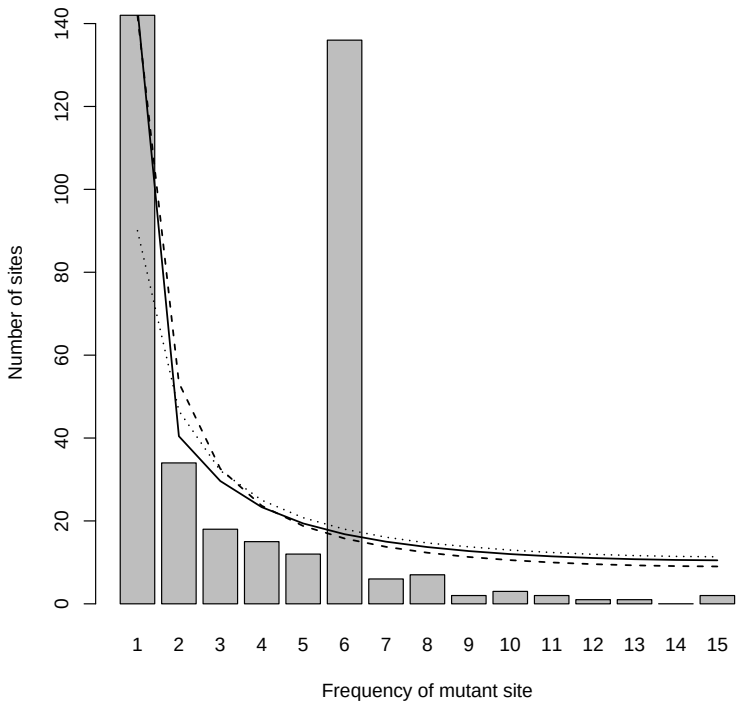

Figure 5. Folded site frequency spectrum of Atlantic cod *Pan* I, *Sort*1 and *Atxn7l2* genes. Number of individuals $n = 31$. Theoretical expectation under Kingman coalescent (dotted line), Beta$(2-\alpha, \alpha$ coalescent (dashed line), and point-mass coalescent (solid line). The bars represent the observed spectrum. The spectrum represents the genetic variability from an alignment of 31 Atlantic cod sequences (25 *Pan* I^A and 6 *Pan* I^B) 12.5 kb long.

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.956v1 | CC-BY 4.0 Open Access | rec: 2 Apr 2015, publ: 2 Apr 2015

Table 1. Summary statistics of nucleotide polymorphism at *Pan* I and its peripheric region, the *Sort*1 and *Atxn7l2* loci (partial segments). The region analyzed is 12.56 kb. n is the number of sequences used, $\hat{S}$ is the number of segregating sites, $\hat{k}$ is the average number of nucleotide differences, $\hat{\pi}$ is nucleotide diversity, $\hat{\theta}$ (per site) is based on S, $\hat{h}$ is number of haplotypes, and $\hat{Hd}$ is haplotype diversity.

Allelic Type	n	$\hat{S}$	$\hat{\pi}$	$\hat{\theta}$	$\hat{k}$	$\hat{h}$	$\hat{Hd}$
Pan I^A alleles only	25	209	0.00284	0.00455	34.580	25	1.000
Pan I^B alleles only	6	31	0.00103	0.00109	12.800	6	1.000
Pan I^A and *Pan* I^B combined	31	349	0.00593	0.00723	71.626	31	1.000

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.956v1 | CC-BY 4.0 Open Access | rec: 2 Apr 2015, publ: 2 Apr 2015