A peer-reviewed version of this preprint was published in PeerJ on 24 February 2015.

Árnason E, Halldórsdóttir K. 2015. Nucleotide variation and balancing selection at the *Ckma* gene in Atlantic cod: analysis with multiple merger coalescent models. PeerJ 3:e786
https://doi.org/10.7717/peerj.786

Nucleotide variation and balancing selection at the *Ckma* gene in Atlantic cod: Analysis with multiple merger coalescent models

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

Nucleotide Variation and Balancing Selection at the *Ckma* gene in Atlantic cod: Analysis with multiple merger coalescent models

Einar Árnason[1] **and Katrín Halldórsdóttir**[2]

[1]**Institute of Biology, University of Iceland, Reykjavík, Iceland**
[2]**Institute of Biology, University of Iceland, Reykjavík, Iceland**

ABSTRACT

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

Keywords: Balancing Selection, *Ckma*, Atlantic cod, multiple-merger coalescent, time scales

INTRODUCTION

High fecundity translates into large excess reproductive capacity that would allow organisms to withstand substantial natural selection enabling them to bear the entailing high genetic load. Based on the concept of the cost of natural selection (Haldane, 1957) high-fecundity organisms relative to low-fecundity organisms should at any time be able to adapt a larger proportion of their genome to meet various environmental challenges. Trying to explain the paradox of sexual reproduction Williams (1975) in his

Sex and Evolution book argues that high-fecundity coupled with type III survivorship of heavy mortality of young may be able to pay the 50% fitness cost of meiosis. He developed several models, such as the Elm/Oyster and the Cod/Starfish models, which emphasize the importance of high-fecundity for selection. Williams also discussed the concept of reproductive sweepstakes. There is no heritability of fitness and sexual reproduction continuously assembles Sisyphean genotypes. The distribution of offspring numbers is highly skewed, heavy-tailed and with high variance (lognormal). That is Williams's fitness distribution. The environment factors are envisioned to act in a sequence of selective filters. With only a few factors (e.g. temperature, salinity, etc) there nevertheless can be an enormous number of different sequences of selective filters (environments) that do not recur. Hence a winning genotype is not permanent and must be continuously reassembled. Natural selection increases the variance in offspring number and thereby reduces effective population size genome-wide. Neutral variation will therefore drift faster under pervasive natural selection.

Coalescent theory (Kingman, 1982a,b) traces the genealogy of a sample and is very useful for making inference of molecular data. However, in an extreme case under a winner-take-all sweepstakes reproduction all samples would coalesce immediately in the previous generation (Árnason, 2004) and there would be no variation. The Kingman coalescent, which is derived from (Wrigth/Fisher) models of low fecundity non-skewed offspring distributions, assumes a bifurcating genealogy and is not appropriate for reproduction of this kind (Eldon and Wakeley, 2006; Schweinsberg, 2003; Wakeley, 2013). Some organisms may exhibit both high fecundity and highly skewed offspring distributions. For these organisms the Λ coalescent allowing multiple mergers of ancestral lineages (Pitman, 1999; Sagitov, 1999; Donnelly and Kurtz, 1999; Eldon and Wakeley, 2006; Schweinsberg, 2003; Sargsyan and Wakeley, 2008) or Ξ coalescent allowing simultaneous multiple mergers of ancestral lineages (Schweinsberg, 2000; Möhle and Sagitov, 2001) may be more appropriate. Wakeley (2013) gives an overview of the development of coalescent theory in new directions. There is also active development of statistical inference methods associated with multiple merger coalescents (e.g. Birkner et al., 2013b, 2014). Studies on the high fecundity organisms Pacific oyster *Crassostrea gigas* (Hedgecock and Pudovkin, 2011) and Atlantic cod *Gadus morhua* (Árnason and Pálsson, 1996; Árnason et al., 1998, 2000; Carr and Marshall, 1991a; Carr et al., 1995; Pepin and Carr, 1993; Árnason, 2004) have provided data for a number of tests of some of the new coalescent models (Eldon and Wakeley, 2006; Eldon, 2011; Eldon and Degnan, 2012; Steinrücken et al., 2013; Birkner et al., 2013b). Atlantic cod thus provides a model for studies of multiple merger coalescent. In this paper we apply some of these new methods for Λ coalescents in a study of balancing selection at a gene showing extreme spatial differentiation in Atlantic cod.

A dense genomic map of genetic variation in humans (and in model organisms) allows scanning the genome for signatures of natural selection (Voight et al., 2006; Sabeti et al., 2007; Storz, 2005). Asking what percentage of the human genome shows footprints of selection depends on the density of the maps and sensitivity of the various methods used (Voight et al., 2006; Sabeti et al., 2007; Storz, 2005). It is safe to say that only a small percentage of single nucleotide polymorphisms (SNPs) show footprints of selection in the low fecundity humans (Akey, 2009; Pickrell1 et al., 2009). For microsatellite loci 2% (13/624) were detected as outliers when African and non-African

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.528v1 | CC-BY 4.0 Open Access | rec: 9 Oct 2014, publ: 9 Oct 2014

human populations were compared (Storz et al., 2004). In contrast, comparable genome level studies in Atlantic cod find that 11% (26 out of 235) of independent SNPs (Moen et al., 2008) are F_{ST} outliers (by method of Beaumont and Nichols, 1996) and 4% SNPs (70 out of 1641 Bradbury et al., 2010) are Bayscan outliers (by method of Foll and Gaggiotti, 2008) likely undergoing selection. Similarly one fourth of microsatellite loci in Atlantic cod (Nielsen et al., 2006) are F_{ST} outliers. This supports our thesis that a considerable fraction of the Atlantic cod genome may be simultaneously under selection for different adaptations.

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

The Moen et al. (2008) study of differentiation among four Atlantic cod populations along the coast of Norway showed no differentiation among presumably neutral non-outliers loci with an average $\bar{F}_{ST} = 0.0012$. In contrast, the outlier loci, presumably under selection, the average $\bar{F}_{ST} = 0.27$ ranging from 0.08 to an extreme differention of 0.83, representing almost fixation of alternative alleles. We analyze nucleotide variation at a large fragment of the gene showing extreme spatial differention to understand the nature of selection. It is the *Ckma* gene encoding a muscle isoform A of creatine kinase.

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

Here we thus have a gene with a well defined and well understood function. The gene shows extreme spatial differentiation most likely due to selection considering the behavior of neutral non-outliers. We ask what a detailed analysis of nucleotide variation using methods of multiple merger Λ coalescents at the scale of the gene itself can tell us about the nature of selection.

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.528v1 | CC-BY 4.0 Open Access | rec: 9 Oct 2014, publ: 9 Oct 2014

MATERIALS AND METHODS

Population sampling

We randomly sampled 122 individual cod from various localities from the distributional range of Atlantic cod (Figure S1). The samples come from our large sample database. The localities are the waters around Newfoundland (New), Greenland (Gre), Iceland (Ice), Faroe Islands (Far), Norway (Nor), and the Barents Sea, North Sea (Nse), Celtic Sea (Cel), Irish Sea (Iri), Baltic Sea (Bal), and the White Sea (Whi).

For outgroup comparison we included samples of the sister taxa Arctic cod *Boreogadus saida* (Bsa) and Greenland cod *G. ogac* (Gog) both sampled in Greenland waters as well as Pacific cod *G. macrocephalus* (Gma) and Walleye pollock *Theragra chalcogramma* (Gch) sampled from the Pacific ocean. Carr et al. (1999) and Pogson and Mesa (2004) discuss the relationship and biogeography of these taxa. Coulson et al. (2006) provide the most comprehensive account based on mitochondrial genomics. They consider Arctic cod to be an outgroup for all these taxa. Atlantic cod and Walley pollock are sister taxa and Pacific cod slightly more distant. Pacific cod and Walleye pollock represent two separate but nearly simultaneous invasions of the Pacific with the Atlantic cod vs. Pacific cod split dated at 4 mya and the Atlantic cod vs. Walleye pollock split at 3.8 mya using conventional rates of mtDNA evolution (see time scales below). They suggest a nomenclature revision from *Theragra chalcogramma* to *Gadus chalcogrammus* for Walleye pollock. Greenland cod is a recent reinvasion of Pacific cod into the Arctic and Coulson et al. (2006) consider it to be a subspecies of Pacific cod.

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

(`http://www.wageningenur.nl/en/Expertise-Services/Research-Institutes/imares/Weblogs/Beam-Trawl-Survey.htm`)

of the Institute for Marine Resources & Ecosystem Studies (IMARES), Wageningen University, the Netherlands, which is approved by the IMARES Animal Care Committee and IMARES Board of Directors.

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.528v1 | CC-BY 4.0 Open Access | rec: 9 Oct 2014, publ: 9 Oct 2014

Molecular analysis

We used sequences associated with the Moen et al. (2008) high F_{ST} SNP's (Gm366-0514 with an $F_{ST} = 0.83$, Gm366-1022 with an $F_{ST} = 0.82$, and Gm366-1073 with an $F_{ST} = 0.82$) to make probes to search an Atlantic cod BAC library. We had positive clones 454 sequenced (Microsynth) and obtained a 34223 bp scaffold containing the gene of interest. From this sequence we generated primers (Table S1) for PCR amplifying a 4000 bp fragment for population studies. Our scaffold largely but not entirely aligned to GeneScaffold 4232 of the Atlantic cod genome sequence (Star et al., 2011) (`www.ensemble.org`).

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

For neutral locus comparisons we applied the same methods and sequenced 711 bp of the Hemoglobin α 2 (*HbA2*) locus (Halldórsdóttir and Árnason, 2009a,b; Borza et al., 2009) and 1021 bp of the myoglobin (*Myg*) locus (Lurman et al., 2007). The *HbA2* data were of 114 Atlantic cod individuals and 13 individuals of various sister taxa. The *Myg* data were from 45 Atlantic cod individuals and two individuals of Pacific cod. Other sister taxa did not amplify for *Myg*. The *HbA2* and *Myg* individuals covered much the same geographic localities as *Ckma*.

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

Statistical analysis

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

By PCR amplifying and cloning of fragments polymerase copy errors in the PCR reaction inevitably will be found in clones. The coalescent methods are especially sensitive to singleton variants and errors that would enter into the data as singleton variants should be removed. To remove PCR errors and ensure authenticity of natural variation among individuals we sequenced three clones from each individual. We claim that taking three clones is sufficient to eliminate PCR errors among clones of an individual and yield a consensus sequence of one allele from that individual. Two of the three clones will be of the same allele (the same chromosome). The third clone is expected to be of that same allele in one half cases and of the alternative allele from the other chromosome in one half cases. In the first case a consensus sequence will be a

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.528v1 | CC-BY 4.0 Open Access | rec: 9 Oct 2014, publ: 9 Oct 2014

true consensus of that allele. In the second case a consensus sequence will be a true consensus except at sites where the third clone (alternative allele) matches one of the other clones. That is when a naturally occurring site variant or a PCR error in the third clone matches a PCR error in one of the other two clones. This scenario is expected to be a rare event. The effect of such a rare event would be to generate variation that would look like recombination thus, if anything, reducing measures of linkage disequilibrium.

We thus got consensus sequences for a number of individuals. We visually inspected all variant sites using the above mentioned tools. To maximize the number of individuals and the size of the sequenced fragment we struck a balance between number of individuals and quality of sequence. We removed individuals with a short sequences and eliminated regions with a phred quality less than 30. We thus ended up with consensus sequences of three clones from each of 122 Atlantic cod and 10 individuals of sister taxa covering three fragments of the gene (Figure S2) concatenated to give a total sequence of 2500 bp.

We analyzed sequence variation for statistics of neutrality and selection using `DNAsp` (Rozas et al., 2003) and `R` functions. Site frequency spectra are a most important summary statistics for coalescent analysis of nucleotide data (Wakeley, 2009). We analyzed site frequency spectra using the Kingman coalescent (Kingman, 1982a) and statistical methods developed for multiple merger Λ coalescents (Birkner et al., 2013b).

RESULTS

Gene and protein

The gene is *Ckma* encoding creatin kinase muscle isoform a (CKMA). The locus is 3604 base pairs (bp) in GeneScaffold 4232 (coordinates 332764 to 336367, gene name ENSGMOG00000008778 in the cod genome, `www.ensemble.org` Star et al. (2011)). The gene has seven exons (Figure S2). Ensemble reports 382 amino acids (aa). However, both `genescan` (`http://genes.mit.edu/GENSCAN.html`) and `fgenesh` (`www.softberry.com`) predicted 381 aa and our analysis of our own data confirmed that. The `www.ensemble.org` sequence adds a Glycine (G) residue in position 323 apparently due to incorrect splicing at the junction of the last two exons.

For mapping the gene the SNP locus cgpGmo-S497 at position 19.5 in linkage group CGP16 is found in a partial cDNA mRNA sequence (Genbank accession number EX184243) (Hubert et al., 2010; Borza et al., 2010) matching the *Ckma* gene. We take that as the location of the gene.

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

Nucleotide variation and divergence

The variants of *Ckma* in Atlantic cod fell into two distinct and divergent groups which we refer to as *A* and *B* alleles (Figures 1 and S4). They were fixed for a C vs T at site 1732 in the concatenated sequence (Table S2). There also were nearly fixed differences between the alleles at 19 additional sites (Figure 2 and Table S2).

The divergence of the *A* and *B* alleles has arisen after the speciation between *Gadus*

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.528v1 | CC-BY 4.0 Open Access | rec: 9 Oct 2014, publ: 9 Oct 2014

morhua and its Pacific sister species *G. macrocephalus* or *G. chalcogrammus*. The gross and net nucleotide divergence between the *A* and *B* alleles was about one half that of the divergence between the sister taxa (Table S3). The *Ckma* and *HbA2* divergences between the sister taxa are very similar but the *Myg* divergence is about twice that (Figure S5 and Table S3). Contra Coulson et al. (2006) the maximum likelihood tree for *Ckma* (Figure 1) and divergence estimates (Table S3) imply that separation of *G. chalcogrammus* predates the separation of *G. macrocephalus* and *G. morhua*. Similarly, the *HbA2* locus showed the same pattern that *G. chalcogrammus* is outside of *G. macrocephalus* and *G. morhua* (Figure S6). Unfortunately the *Myg* locus did not yield sequences for *G. chalcogrammus*.

All summary statistics showed high variation for *Ckma* (Table 1). In particular nucleotide diversity $\hat{\pi}$ was high relative to the scaled population size $\hat{\theta}_S$ resulting in a non-significant Tajima's $\hat{D}$. This was due to the great number of high heterozygosity sites nearly fixed between the two alleles (Figure 2 and Table S2). Considering the North and South population and the *A* and *B* alleles separately there was much less variation. Although there were several polymorphic sites within both *A* and *B* alleles (Figure 2 and Table S2) nucleotide diversity was lower than for the entire sample and the relative difference of $\hat{\pi}$ and $\hat{\theta}_S$ for each allele was greater resulting in negative and significant Tajima's $\hat{D}$. The *HbA2* gene had a very low haplotype and nucleotide diversity but disparity with $\hat{\theta}_S$ gave overall a negative and significant Tajima's $\hat{D}$. In congruence with divergence measures the *Myg* locus had high haplotype and nucleotide diversity, albeit lower than *Cmka*, but overall a negative and significant Tajima's $\hat{D}$.

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

There was considerable linkage disequilibrium (LD) throughout the gene (Figure S7). The high heterozygosity sites nearly fixed between the alleles were influential in generating LD between sites throughout the gene.

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

Spatial differentiation

The variation was spatially patterned. The *A* allele was nearly fixed in an area that we call South (Faroe Islands, North Sea, Baltic Sea, Celtic Sea and Irish Sea) at a frequency of 97% (Table S6). Conversely the *B* allele was at a high frequency of 92% in an area that we call North ranging from the Northwest (Nova Scotia and Newfoundland in Canada) through Greenland, Iceland, Norway, Barents Sea and the White Sea. The differentiation was evident in interlocality F_{ST} values (Table S7). There was no significant differentiation among localities within either the North or the South but very high and significant differentiation between North and South localities.

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.528v1 | CC-BY 4.0 Open Access | rec: 9 Oct 2014, publ: 9 Oct 2014

Similarly, there was great differentiation between the *A* and *B* alleles (Table S8). This was in stark contrast to the lack of differentiation at the *HbA2* and *Myg* loci (Table S9).

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

There were also several high heterozygosity polymorphic sites within both the *A* and *B* alleles (Figure 2). This variation, however, did not show geographical patterns (Table S2). For example sites 1050 and 1428 mutated relative to outgroup within the *A* alleles were found among individuals from Iceland, White Sea, Celtic Sea, Faroe Islands and the Baltic. Similarly within the *B* alleles high heterozygosity sites 656, 691, 1340, and 1444, which were mutated relative to the outgroup, were all widespread among North localities ranging from the Northwest to the Northeast Atlantic (Figure S1).

Site frequency spectra

The unfolded site frequency spectrum for the *Ckma* gene was trimodal (Figure 3), with a mode at singleton sites, a mode at 43, and a mode at 79. The latter modes were at opposite frequencies out of a total of 122 and represented the *A* and *B* lineages of the genealogy. The Kingman coalescent did not fit the data well. Both the Beta$(2-\alpha,\alpha)$ and point-mass coalescent models gave a much better fit (Table S10) in particular by capturing the singleton class. None of the coalescent models captured the modes at 43 and 79.

In contrast the site frequency spectra for the *HbA2* and *Myg* genes were L shaped with a high peak at singleton sites (Figures S8 and S9). Again the Kingman coalescent did not fit well but both multiple merger coalescent models captured the high frequency of singleton sites.

The site frequency spectra of the *A* and *B* alleles alone were bimodal with a high singleton class and peaks around 40 and 78 respectively (Figure S10). The 40 and 78 modes came about because most of the high frequency and high heterozygosity sites that separate the two alleles were not fixed within each allele presumably due to recombination (Table S2).

Coalescent parameter estimates

Following Birkner et al. (2013b) we used the ℓ^2 distance, the sum of the squared differences between the observed and expected site frequency spectrum (scaled with the number of segregating sites), for estimating parameters of two Λ coalescent models, $\hat{\alpha}$ for the Beta$(2-\alpha,\alpha)$ and $\hat{\psi}$ for the point-mass coalescent (Table 2 and Figures S11 and S12). The Kingman coalescent, a null model for which $\alpha = 2.0$, had the highest ℓ^2 indicating worst fit among the models. The *HbA2* and *Myg* loci had an $\hat{\alpha} = 1.00$ and a $\hat{\psi} = 0.23$. The *Ckma* locus had overall a considerably higher α and lower ψ. The two alleles separately were in the direction of the presumed neutral loci *HbA2* and *Myg*.

For comparison we also estimated the parameters for the entire dataset of mtDNA variation in the North Atlantic (Árnason, 2004) and the various subsamples making up that total sample using the unfolded site frequency spectrum with *G. macrocephalus*

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.528v1 | CC-BY 4.0 Open Access | rec: 9 Oct 2014, publ: 9 Oct 2014

as outgroup (Table 2 and Figure S12). Previously these have been analysed using the folded site frequency spectrum (see for example Birkner et al., 2013b; Steinrücken et al., 2013). For the total sample, spanning a similar geographic range as the nuclear genes, the parameter estimates differed from the nuclear loci with $\hat{\alpha} = 1.53$ and $\hat{\psi} = 0.01$. The large samples from Newfoundland and Iceland and the sample from the Faroe Islands gave similar values. The values for Greenland, Norway, White Sea, and Baltic Sea were much closer to the results for the Kingman coalescent ($\alpha = 2.0$). For these localities homoplasies were relatively somewhat more frequent in the data than for the total and the large samples. Homoplasies will reduce the number of singletons and move such sites towards the right tail of the site frequency distribution. This explains the higher values for these localities.

DISCUSSION

Genes and proteins

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

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

Allele divergence and spatial differentiation

Three possible scenarios and explanations for the great divergence of the *A* alleles and *B* alleles, their spatial differentiation, and the trimodal site-frequency spectrum will now be considered.

First, there is the possibility of recent admixture of anciently separated and divergent gene pools that have come together in a hybrid zone of secondary contact (Bowcock et al., 1991; Bernardi et al., 1993; Guinand et al., 2004). The spatial patterns of genetic separation between the South (Faroe Islands, North Sea, Baltic Sea, Celtic Sea and Irish Sea) and the North (Nova Scotia and Newfoundland, Greenland, Iceland, Norway, Barents Sea, and White Sea) could be taken as evidence for this. The South is a shallow water environment whereas the the North has more diversity of depth ranging from shallow to deep waters. Differences in temperature, salinty and other environmental factors is correlated with the North South difference. The great nucleotide divergence

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.528v1 | CC-BY 4.0 Open Access | rec: 9 Oct 2014, publ: 9 Oct 2014

between the North and the South would imply either that this is an ancient divergence or even a not-so-ancient divergence driven by strong selection over a shorter time. If the time of separation of *G. morhua* and *G. macrocephalus* and *G. chalcogrammus* is taken at 3.8–4.0 Mya (Coulson et al., 2006) the time of separation of the *A* and *B* clades would then be 2 Mya based on the nucleotide divergence of the *A* and *B* clades which we show is one half that of the sister taxa. An even lower divergence time of 2.1 Mya has been suggested (Pogson and Mesa, 2004) that would still leave the divergence of the *A* and *B* clade at 1 Mya. These divergence times, however, are all based on the Kingman coalescent and time scales of the multiple merger coalescent are discussed below.

A counter argument is that isolation and admixture are part of the breeding structure of a population leaving genome-wide impacts (Wright, 1931). Under this scenario different genes should be concordant in their behavior (Bernardi et al., 1993). The *HbA2* and the *Myg* show no differentiation between the North and the South. Also the non-outlier SNPs in Moen et al. (2008) show no differentiation whereas three SNPs of the *Ckma* gene show high and extreme F_{ST}. Similarly, Bradbury et al. (2010) find that non-outlier SNPs show no differentiation although other SNPs show differentiation from parallel adaptation to temperature on the eastern and western side of the Atlantic Ocean. Nielsen et al. (2003) describe a pattern of microsatellite variation in a transition area between the Baltic and Danish Belt Sea which they interpret as a hybrid zone. There is no evidence for a hybrid zone at that location in the *Ckma* data. In fact, specific variants within the *A* allele are widely distributed among localities in the South including the Baltic Sea. This implies gene flow among localities in the South. Similar patterns within *B* alleles imply gene flow among localities in the North. If indeed there is a hybrid zone for the *Ckma* gene it would lie between the Faroe Islands on one hand and Iceland and north and middle Norway on the other hand. It is not a parsimonious explanation to consider there to be multiple hybrid zones of secondary contact within distribution of the species.

For comparison one can consider the *Pan* I locus (Fevolden and Pogson, 1995, 1997) that clearly is under selection (Pogson, 2001; Pogson and Mesa, 2004) related to depth and fisheries (Sarvas and Fevolden, 2005; Case et al., 2005; Árnason et al., 2009). At face value the locus shows similar differentiation between north and south (Sarvas and Fevolden, 2005) as the *Ckma* locus. However, the details differ. The *Pan* I *B* allele which is adapted to the deep (Pampoulie et al., 2007; Árnason et al., 2009) is largely absent from the South. However, there is no particular *Pan* I *A* allele that characterizes the South (Hernandez and Árnason, 2014). The *Pan* I *B* allele, which is found in the North and in deep water, is much less variable than the *Pan* I *A* alleles (Pogson, 2001; Hernandez and Árnason, 2014). This is opposite to what we find for the *Ckma A* alleles (the South allele) which has less variation than the *Ckma B* allele (Figure 1) although this is not seen in the summary statistics (Table 1) because of greater recombinational variation at the base of the *A* clade (Table S2). Also the *Pan* I locus variation is more related to depth than to geography (Árnason et al., 2009). Under the admixture scenario these two loci (and all loci showing genome wide effects) are expected to show the same pattern.

Overall, therefore, we find that the *Ckma* gene does not fit the scenario of ancient divergence of gene pools and admixture in secondary contact.

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.528v1 | CC-BY 4.0 Open Access | rec: 9 Oct 2014, publ: 9 Oct 2014

Site frequency spectra

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

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

Balancing selection

Balancing selection generates long branches in the genealogy and neutral variation accumulates on the branches. The balanced functional types (the *Ckma A* and *B* alleles in this case) act as they were separate and isolated populations accumulating neutral variation. Recombination can bring variation from one branch to another acting like migration that brings alleles from one population to another (Charlesworth et al., 1997, 2003; Charlesworth, 2006). However, the molecular signatures of balancing selection depend on many factors. Is it a long standing, even trans-species, polymorphism such as *MHC* in human and chimpanzee (Fan et al., 1989; Nei and Hughes, 1991) or *Cathelicidin* in gadids (Halldórsdóttir and Árnason, 2014)) or is it very recent? Examples of the latter are human glucose 6 phosphate dehydrogenase (G6PD) (Verrelli et al., 2002), and hemoglobin β *S* (Currat et al., 2002) and hemoglobin β *E* (Ohashi et al., 2004) and

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.528v1 | CC-BY 4.0 Open Access | rec: 9 Oct 2014, publ: 9 Oct 2014

spatially divergent selection of lactase persistence (Tishkoff et al., 2007; Ranciaro et al., 2014) in which a particular allele sweeps a chromosomal segment to an intermediate equilibrium frequency. In these instances recombination has not had time to break up LD which can extend over large regions. There is very little variation among the new alleles while the alternative chromosomes show much more variation in this region representing the standing variation in the population at the start of the partial sweep.

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

Pogson and Mesa (2004) further argue that the *Pan* I polymorphism is older than speciation of Atlantic cod and Walleye pollock, the closest relatives. The *Pan* I locus is in a "genomic island" (Bradbury et al., 2013; Hemmer-Hansen et al., 2013) a potential supergene of co-adapted complexes possibly locked together by structural variation. Looking in detail at variation at 12.5 kb region Hernandez and Árnason (2014) find large number of differences between the two functional *Pan* I types that are too extensive to be a partial sweep of a new allele. Such variation is likely to be built up over some time by selection (see time scales below). This is in face of considerable gene flow implied by lack of differentiation of neutral loci (Moen et al., 2009; Bradbury et al., 2010; Eiríksson and Árnason, 2013; Hemmer-Hansen et al., 2014). Similarly, the wide distribution of variants within both the *A* and *B* alleles of *Ckma* implies gene flow among localities within South and within North areas. The recombinant haplotypes between the *A* and *B* alleles of *Ckma* imply gene flow between the South and the North localities.

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

Similarly we consider the peaks in the site frequency spectrum of the *Ckma* gene to be evidence for balancing selection. The trimodal site frequency spectrum with two high frequency peaks at opposite frequencies that fold into one peak in a folded

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.528v1 | CC-BY 4.0 Open Access | rec: 9 Oct 2014, publ: 9 Oct 2014

site frequency spectrum points to the build-up of variation over time. Under a recent balanced polymorphism scenario, such as *G6PD* and β globins in humans, there would be one peak at a particular frequency in the site frequency spectrum representing all sites at which the new allele differs from the ancient alleles. There could be multiple peaks representing high frequency polymorphisms among the ancient alleles. However, they are not expected to be at opposite frequencies to the frequency of the new allele. We, therefore, argue that the pattern at *Ckma* represents a balanced polymorphism that has been built up over time.

Coalescent parameter estimates and time scales

The question of coalescent time scale, however, must be considered. Under the Kingman coalescent time is measured in terms of N/σ^2, population size scaled by the variance of family size (Sagitov, 1999; Árnason, 2004; Tavaré, 2004). With a Poisson distribution of family size $\sigma^2 = 1$ for a constant size haploid population so times scales with N. In an extreme winner-take-all sweepstakes $\sigma^2 = N$ and a sample would coalesce in the previous generation and there would be no variation (Árnason, 2004). In more realistic multiple merger coalescent models the time scale is the quantity $c_N = \frac{E(\nu_1-1)^2}{N-1}$ where c_N is the probability of two lineages coalescing in the previous generation in a haploid population of fixed size N and ν_1 is the random number of offspring of individual 1 (Sagitov, 1999). In general the time scale of multiple merger coalescent models can be much shorter than for Kingman coalescent. Under the Beta$(2-\alpha, \alpha)$ coalescent model time scales with $N^{\alpha-1}$ (Schweinsberg, 2003; Birkner et al., 2014). For this model our estimates of α for the nuclear genes are quite low which implies very short time scales. The neutral genes would seem to coalesce in the very recent past. The *A* and *B* alleles of *Ckma* run on very similar time scales to the neutral genes and the locus itself at a slower rate due to the balancing selection with a time scale approximately the cube root of the effective population size N_e. The mitochondrial DNA runs at yet another and slower time scale. For mtDNA time scales with approximately the square root of N. Predicted turnover of alleles is faster and ages of alleles shorter under multiple merger coalescent (Eldon, 2912). Different populations and species may run on different time scales (Eldon and Degnan, 2012) complicating divergence time estimates. Estimates based on Kingman coalescent of divergence times of Atlantic cod populations (Bigg et al., 2008) or divergence of gadid taxa (Coulson et al., 2006) may therefore be too high and may need revision.

Conclusion

The *Ckma* protein coding sequence is conserved between all but the most distantly related Arctic cod. The amino acid variants are all singletons in the sample. Based on these facts we conclude that the protein coding sequence is under purifying selection. At the same time silent and non-coding variation at the locus shows extreme spatial differentiation with an F_{ST} greater than 0.8 between the North and the South regions. The regulatory function of this variation is unclear. We argue that the high and locus-specific F_{ST}, the highest seen so far for any locus and any spatial comparison in Atlantic cod, indicates that selection and not admixture of anciently divergent gene pools is responsible. Selection is likely to be very strong. It follows that *Ckma* (or an extremely tightly linked locus) is the focus of selection because the highest F_{ST} indicates the site

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.528v1 | CC-BY 4.0 Open Access | rec: 9 Oct 2014, publ: 9 Oct 2014

of action of selection (Nielsen, 2005). Some of the variation may be neutral having risen in frequency within the balanced functional allele where it arose (Charlesworth, 2006). Alternatively some of the variation may be due to selection building co-adapted complexes (Thompson and Jiggins, 2014). In addition to a peak at singleton sites, characteristic of multiple-merger coalescent, the site frequency spectrum has two high-frequency modes at opposite but matching frequencies representing the two branches of the genealogy. This pattern is further support for balancing selection. Finally time scales faster under multiple-merger than the Kingman coalescent. Our estimates of parameters of multiple-merger Λ coalescent show that time-scales are fast.

ACKNOWLEDGMENTS

We thank Jarle Mork (Norwegian University of Science and Technology), Kristján Kristjánsson (Marine Research Institute in Reykjavik), Grant Pogson (University of California at Santa Cruz), Remment ter Hofstede (Institute for Marine Resources and Ecosystem Studies in the Netherlands), and Michael Canino (National Oceanic and Atmospheric Administration) for help in securing some of the samples. We thank Brenda Ciervo Adarna, Guðni Magnús Eiríksson, Lilja Stefánsdóttir, Ragnheiður Fossdal, Svava Ingimarsdóttir, Ubaldo Benitez Hernandez for help with some of the laboratory work. We thank Bjarki Eldon for programs and help with coalescent parameter estimation and for critical comments on the manuscript. Funding was provided by Icelandic Science Foundation grant of excellence (nr. 40303011), a University of Iceland Research Fund grant, and a SA Private Foundation grant to Einar Árnason and a doctoral grant from the University of Iceland Research Fund to Katrín Halldórsdóttir.

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.528v1 | CC-BY 4.0 Open Access | rec: 9 Oct 2014, publ: 9 Oct 2014

LITERATURE CITED

Akey, J. (2009). Constructing genomic maps of positive selection in humans: Where do we go from here? *Genome Res.*, 19:711–722.

Árnason, E. (2004). Mitochondrial cytochrome *b* DNA variation in the high-fecundity Atlantic cod: Trans-Atlantic clines and shallow gene genealogy. *Genetics*, 166(4):1871–1885.

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. (1996). Mitochondrial cytochrome *b* DNA sequence variation of Atlantic cod, *Gadus morhua*, from Norway. *Mol. Ecol.*, 5:715–724.

Árnason, E., Pálsson, S., and Petersen, P. H. (1998). Mitochondrial cytochrome *b* DNA sequence variation of Atlantic cod, *Gadus morhua*, from the Baltic- and the White Seas. *Hereditas*, 129:37–43.

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. *J. Fish Biol.*, 56:409–430.

Beaumont, M. A. and Nichols, R. A. (1996). Evaluating loci for use in the genetic analysis of population structure. *Proc. R. Soc. Ser. B.*, 263:1619–1626.

Bernardi, G., Sordino, P., and Powers, D. A. (1993). Concordant mitochondrial and nuclear DNA phylogenies for populations of the teleost fish *Fundulus heteroclitus*. *Proc. Natl. Acad. Sci.*, 90:9271–9274.

Bigg, G. R., Cunningham, C. W., Ottersen, G., Pogson, G. H., Wadley, M. R., and Williamson, P. (2008). Ice-age survival of Atlantic cod: agreement between palaeoecology models and genetics. *Proc. R. Soc. B.*, 275:163–172.

Birkner, M., Blath, J., and Eldon, B. (2013a). An ancestral recombination graph for diploid populations with skewed offspring distribution. *Genetics*, 193:255–290.

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

Birkner, M., Blath, J., Eldon, B., and Freund, F. (2014). Can the site-frequency spectrum distinguish exponential population growth from multiple-merger coalescents? *bioRxiv*.

Borza, T., Higgins, B., Simpson, G., and Bowman, S. (2010). Integrating the markers *Pan I* and Haemoglobin with the genetic linkage map of Atlantic cod (*Gadus morhua*). *BMC Res. Notes*, 3:261.

Borza, T., Stone, C., Gamperl, A. K., and Bowman, S. (2009). Atlantic cod (*Gadus morhua*) hemoglobin genes: multiplicity and polymorphism. *BMC Genet.*, 10:51.

Bowcock, A. M., Kidd, J. R., Mountain, J. L., Herbert, J. M., Carotenuto, L., Kidd, K. K., and Cavalli-Sforza, L. (1991). Drift, admixture, and selection in human evolution: A study with DNA polymorphisms. *Proc. Natl. Acad. Sci.*, 88:839–843.

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. *Proc. R. Soc. B.*, 277:3725–3734.

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.528v1 | CC-BY 4.0 Open Access | rec: 9 Oct 2014, publ: 9 Oct 2014

615 Bradbury, I. R., Hubert, S., Higgins, B., Bowman, S., Borza, T., Paterson, I. G., Snel-
616 grove, P. V. R., Morris, C. J., Gregory, R. S., Hardie, D., Hutchings, J. A., Ruzzante,
617 D. E., Taggart, C. T., and Bentzen, P. (2013). Genomic islands of divergence and their
618 consequences for the resolution of spatial structure in an exploited marine fish. *Evol.*
619 *Appl.*, 6:450–461.
620 Carr, S. M., Kivlichan, D. G. S., Pepin, P., and Crutcher, D. C. (1999). Molecular
621 phylogeny of Gadid fishes: Implications for the biogeographic origins of Pacific
622 species. *Can. J. Zool.*, 77:19–26.
623 Carr, S. M. and Marshall, H. D. (1991a). Detection of intraspecific DNA sequence
624 variation in the mitochondrial cytochrome *b* gene of Atlantic cod (*Gadus morhua*) by
625 the polymerase chain reaction. *Can. J. Fish. Aquat. Sci.*, 48:48–52.
626 Carr, S. M. and Marshall, H. D. (1991b). A direct approach to the measurement of
627 genetic variation in fish populations: Applications of the polymerase chain reaction to
628 studies of Atlantic cod (*Gadus morhua*). *J. Fish Biol.*, 39(Supplement A):101–107.
629 Carr, S. M., Snellen, A. J., Howse, K. A., and Wroblewski, J. S. (1995). Mitochondrial
630 DNA sequence variation and genetic stock structure of Atlantic cod (*Gadus morhua*)
631 from bay and offshore locations on the Newfoundland continental shelf. *Mol. Ecol.*,
632 4:79–88.
633 Case, R. A. J., Hutchinson, W. F., Hauser, L., Oosterhout, C. V., and Carvalho, G. R.
634 (2005). Macro- and micro-geographic variation in pantophysin (*Pan*I) allele frequen-
635 cies in NE Atlantic cod *Gadus morhua*. *Mar. Ecol. Prog. Ser.*, 301:267–278.
636 Charif, D. and Lobry, J. (2007). SeqinR 1.0-2: a contributed package to the R project for
637 statistical computing devoted to biological sequences retrieval and analysis. In Bas-
638 tolla, U., Porto, M., Roman, H., and Vendruscolo, M., editors, *Structural approaches*
639 *to sequence evolution: Molecules, networks, populations*, Biological and Medical
640 Physics, Biomedical Engineering, pages 207–232. Springer Verlag, New York. ISBN
641 : 978-3-540-35305-8.
642 Charlesworth, B., Charlesworth, D., and Barton, N. H. (2003). The effects of genetic
643 and geographic structure on neutral variation. *Ann. Rev. Ecol. Evol. Syst.*, 34:99–125.
644 Charlesworth, B., Nordborg, M., and Charlesworth, D. (1997). The effects of local
645 selection, balanced polymorphism and background selection on equilibrium patterns
646 of genetic diversity in subdivided populations. *Genet. Res.*, 70:155–174.
647 Charlesworth, D. (2006). Balancing selection and its effects on sequences in nearby
648 genome regions. *PLoS Genet*, 2(4):e64.
649 Coulson, M. W., Marshall, H. D., Pepin, P., and Carr, S. M. (2006). Mitochondrial
650 genomics of gadine fishes: Implications for taxonomy and biogeographic origins from
651 whole-genome data sets. *Genome*, 49:1115–1130.
652 Cunningham, K. M., Canino, M. F., Spies, I. B., and Hauser, L. (2009). Genetic isolation
653 by distance and localized fjord population structure in Pacific cod (*Gadus macro-*
654 *cephalus*: Limited effective dispersal in the northeastern Pacific Ocean. *Canadian*
655 *Journal of Fisheries and Aquatic Science*, 66:153–166.
656 Currat, M., Trabuchet, G., Rees, D., Perrin, P., Harding, R. M., Clegg, J. B., Langaney,
657 A., and Excoffier, L. (2002). Molecular analysis of the β-globin gene cluster in the
658 Niokholo Mandenka population reveals a recent origin of the β^S Senegal mutation.
659 *Am. J. Hum. Genet.*, 70:207–223.
660 Donnelly, P. and Kurtz, T. G. (1999). Particle representations for measure-valued

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.528v1 | CC-BY 4.0 Open Access | rec: 9 Oct 2014, publ: 9 Oct 2014

population models. *Ann. Prob.*, 27:166–205.
Dray, S. and Dufour, A. (2007). The ade4 package: Implementing the duality diagram for ecologists. *J. Stat. Soft.*, 22:1–20.
Edgar, R. C. (2004). Muscle: Multiple sequence alignment with high accuracy and high throughput. *Nucl. Acids Res.*, 32:1792–1797.
Eiríksson, G. M. and Árnason, E. (2013). Spatial and temporal microsatellite variation in spawning Atlantic cod, *Gadus morhua*, around Iceland. *Can. J. Fish. Aquat. Sci.*, 70:1151–1158.
Eldon, B. (2011). Estimation of parameters in large offspring number models and ratios of coalescence times. *Theor. Pop. Biol.*, 80:16–28.
Eldon, B. (2912). Age of an allele and gene genealogies of nested subsamples for populations admitting large offspring numbers. *arXiv*, 1212.1792v1.
Eldon, B. and Degnan, J. H. (2012). Multiple merger gene genealogies in two species: Monophyly, paraphyly, and polyphyly for two examples of lambda coalescents. *Theor. Pop. Biol.*, 82:117–130.
Eldon, B. and Wakeley, J. (2006). Coalescent processes when the distribution of offspring number among individuals is highly skewed. *Genetics*, 172:2621–2633.
Eldon, B. and Wakeley, J. (2008). Linkage disequilibrium under skewed offspring distribution among individuals in a population. *Genetics*, 178(3):1517–1532.
Ewing, B. and Green, P. (1998). Basecalling of automated sequencer traces using phred. II. error probabilities. *Genome Res.*, 8:186–194.
Ewing, B., Hillier, L., Wendl, M., and Green, P. (1998). Base-calling of automated sequencer traces using phred. I. accuracy assessment. *Genome Res.*, 8:175–185.
Fan, W., Kasahara, M., Gutknecht, J., Klein, D., Mayer, W. E., Jonker, M., and Klein, J. (1989). Shared class II MHC polymorphisms between humans and chimpanzees. *Hum. Immunol.*, 26(2):107–121.
Fevolden, S. E. and Pogson, G. H. (1995). Differences in nuclear DNA RFLPs between the Norwegian coastal and the Northeast Arctic populations of Atlantic cod. In Skjoldal, H. R., Hopkins, C., Eriksstad, K. E., and Leinaas, H. P., editors, *Ecology of Fjords and Coastal Waters*, pages 403–414, Amsterdam, The Netherlands. Elsevier Science Publishers.
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. *J. Fish Biol.*, 51:895–908.
Foll, M. and Gaggiotti, O. (2008). A genome-scan method to identify selected loci appropriate for both dominant and codominant markers: A Bayesian perspective. *Genetics*, 180:977–993.
Gordon, D., Abajian, C., and Green, P. (1998). Consed: A graphical tool for sequence finishing. *Genome Res.*, 8:195–202.
Gouy, M., Guindon, S., and Gascuel, O. (2009). SeaView version 4: A multiplatform graphical user interface for sequence alignment and phylogenetic tree building. *Molecular Biology and Evolution*, 27:221–224.
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.
Guindon, S. and Gascuel, O. (2003). A simple, fast and accurate algorithm to estimate

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.528v1 | CC-BY 4.0 Open Access | rec: 9 Oct 2014, publ: 9 Oct 2014

large phylogenies by maximum likelihood. *Syst. Biol.*, 52:696–704.
Haldane, J. B. S. (1957). The cost of natural selection. *J. Genet.*, 55:511–524.
Halldórsdóttir, K. and Árnason, E. (2009a). Multiple linked β and α globin genes in Atlantic cod: a PCR based strategy of genomic exploration. *Mar. Gen.*, 2:169–181.
Halldórsdóttir, K. and Árnason, E. (2009b). Organization of a β and α globin gene set in the teleost Atlantic cod, *Gadus morhua*. *Biochem. Genet.*, 47:817–830.
Halldórsdóttir, K. and Árnason, E. (2014). Trans-species balanced polymorphism at antimicrobial Cathelicidin genes of the innate immunity system. *In review*, 1.
Harrang, E., Lapegue, S., Morga, B., and Bierne, N. (2013). A high load of non-neutral amino-acid polymorphisms explains high protein diversity despite moderate effective population size in a marine bivalve with sweepstakes reproduction. *G3: Genes|Genomes|Genetics*, 3(2):333–341.
Hedgecock, D. and Pudovkin, A. I. (2011). Sweepstakes reproductive success in highly fecund marine fish and shellfish: A review and commentary. *Bull. Mar. Sci.*, 87:971–1002.
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. *Mol. Ecol.*, 22:2653–2667.
Hemmer-Hansen, J., Therkildsen, N. O., Meldrup, D., and Nielsen, E. E. (2014). Conserving marine biodiversity: Insights from life-history trait candidate genes in Atlantic cod (*Gadus morhua). Cons. Genet.*, 15:213–228.
Hernandez, U. B. and Árnason, E. (2014). DNA sequence variation at the (*Pan* I) locus and its peripheral regions in Atlantic cod (*Gadus morhua*): Analysis of linkage and selection. *In review*, 1.
Hubert, S., Higgins, B., Borza, T., and Bowman, S. (2010). Development of a SNP resource and a genetic linkage map for Atlantic cod (*Gadus morhua*). *BMC Genomics*, 11:191.
Hudson, R. R., Kreitman, M., and Aguadé, M. (1987). A test of neutral molecular evolution based on nucleotide data. *Genetics*, 116:153–159.
Johnson, M. S. and Wernham, J. (1999). Temporal variation of recruits as a basis of ephemeral genetic heterogeneity in the western rockc lobster *Panulirus cygnus*. *Mar. Biol.*, 135(1):133–139.
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.
Kimura, M. (1983). *The Neutral Theory of Molecular Evolution*. Cambridge University Press, Cambridge.
Kingman, J. F. C. (1982a). The coalescent. *Stoch. Proc. Appl.*, 13:235–248.
Kingman, J. F. C. (1982b). On the genealogy of large populations. *J. Appl. Prob.*, 19A:27–43.

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.528v1 | CC-BY 4.0 Open Access | rec: 9 Oct 2014, publ: 9 Oct 2014

753 Lurman, G. J., Koschnick, N., Pörtner, H.-O., and Lucassen, M. (2007). Molecular
754 characterisation and expression of Atlantic cod (*Gadus morhua*) myoglobin from two
755 populations held at two different acclimation temperatures. *Comp. Biochem. Physiol.*
756 *A.*, 148(3):681–689.
757 Moen, T., Delghandi, M., Wesmajervi, M. S., Westgaard, J.-I., and Fjalestad, K. T.
758 (2009). A snp/microsatellite genetic linkage map of the Atlantic cod (*Gadus morhua*).
759 *Anim. Genet.*, 40:993–996.
760 Moen, T., Hayes, B., Frank Nilsen and, M. D., Fjalestad, K. T., Fevolden, S., Berg,
761 P. R., and Lien, S. (2008). Identification and characterisation of novel SNP markers
762 in Atlantic cod: evidence for directional selection. *BMC Genet.*, 9:18.
763 Möhle, M. and Sagitov, S. (2001). A classification of coalescent processes for haploid
764 exchangeable population models. *Ann. Prob.*, 29:1547–1562.
765 Nei, M. and Hughes, A. L. (1991). Polymorphism and evolution of the major histo-
766 compatibility comples loci in mammals. In Selander, R., Clark, A., and Whittam,
767 T., editors, *Evolution at the Molecular Level*, chapter 11, pages 222–247. Sinauer
768 Associates, Inc., Sunderland, MA 01375.
769 Nielsen, E. E., Hansen, M. M., and Meldrup, D. (2006). Evidence of microsatellite
770 hitch-hiking selection in Atlantic cod (*Gadus morhua* L.): Implications for inferring
771 population structure in nonmodel organisms. *Mol. Ecol.*, 15:3219–3229.
772 Nielsen, E. E., Hansen, M. M., Ruzzante, D. E., Meldrup, D., and Grønkjær (2003).
773 Evidence of a hybrid-zone in Atlantic cod (*Gadus morhua*) in the Baltic and the
774 Danish Belt Sea revealed by individual admixture analysis. *Mol. Ecol.*, 12:1497–
775 1508.
776 Nielsen, R. (2005). Molecular signatures of natural selection. *Ann. Rev. Genet.*, 39:197–
777 218.
778 Ohashi, J., Naka, I., Patarapotikul, J., Hananantachai, H., Brittenham, G., Looareesuwan,
779 S., Clark, A. G., , and Tokunaga, K. (2004). Extended linkage disequilibrium
780 surrounding the Hemoglobin E variant due to malarial selection. *Am. J. Hum. Genet.*,
781 74:1198–1208.
782 Pampoulie, C., Jakobsdóttir, K. B., Marteinsdóttir, G., and Thorsteinsson, V. (2007).
783 Are vertical behaviour patterns related to the Pantophysin locus in the Atlantic cod
784 (*Gadus morhua* L.)? *Behav. Genet.*, 38:76–81.
785 Paradis, E. (2010). Pegas: an R package for population genetics with an integrated–
786 modular approach. *Bioinformatics*, 26:419–420.
787 Paradis, E., Claude, J., and Strimmer, K. (2004). APE: Analyses of phylogenetics and
788 evolution in R language. *Bioinformatics*, 20:289–290.
789 Pepin, P. and Carr, S. M. (1993). Morphological, meristic, and genetic analysis of stock
790 structure in juvenile Atlantic cod (*Gadus morhua*) from the Newfoundland shelf. *Can.*
791 *J. Fish. Aquat. Sci.*, 50:1924–1933.
792 Pickrell1, J. K., Coop, G., Novembre, J., andJun Z. Li, S. K., Absher, D., Srinivasan,
793 B. S., Barsh, G. S., Feldman, R. M. M. M. W., and Pritchard, J. K. (2009). Signals of
794 recent positive selection in a worldwide sample of human populations. *Genome Res.*,
795 19:922–933.
796 Pitman, J. (1999). Coalescents with multiple collisions. *Ann. Prob.*, 27(4):1870–1902.
797 Pogson, G. H. (2001). Nucleotide polymorphism and natural selection at the Pantophysin
798 (*Pan* I) locus in the Atlantic cod, *Gadus morhua* (L.). *Genetics*, 157:317–330.

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.528v1 | CC-BY 4.0 Open Access | rec: 9 Oct 2014, publ: 9 Oct 2014

799 Pogson, G. H. and Mesa, K. (2004). Positive Darwinian selection at the Pantophysin
800 (*Pan* I) locus in marine Gadid fishes. *Mol. Biol. Evol.*, 21:65–75.
801 Poulsen, N. A., Hemmer-Hansen, J., Loeschcke, V., Carvalho, G. R., and Nielsen, E. E.
802 (2011). Microgeographical population structure and adaptation in Atlantic cod *Gadus*
803 *morhua*: spatio-temporal insights from gene-associated DNA markers. *Mar. Ecol.*
804 *Prog. Ser.*, 436:231–243.
805 R Core Team (2013). *R A Language and Environment for Statistical Computing*. R
806 Foundation for Statistical Computing, Vienna, Austria.
807 Ranciaro, A., Campbell, M., Hirbo, J., Ko, W.-Y., Froment, A., Anagnostou, P., Kotze,
808 M., Ibrahim, M., Nyambo, T., Omar, S., and Tishkoff, S. (2014). Genetic origins
809 of lactase persistence and the spread of pastoralism in Africa. *Am. J. Hum. Genet.*,
810 94:496–510.
811 Rozas, J., Sánchez-DelBarrio, J. C., Messeguer, X., and Rozas, R. (2003). DnaSP,
812 DNA polymorphism analyses by the coalescent and other methods. *Bioinformatics*,
813 19:2496–2497.
814 Sabeti, P. C., Varilly, P., Fry, B., Lohmueller, J., Hostetter, E., Cotsapas, C., Xie, X.,
815 Byrne, E. H., McCarroll, S. A., Gaudet, R., Schaffner, S. F., Lander, E. S., and
816 Consortium, T. I. H. (2007). Genome-wide detection and characterization of positive
817 selection in human populations. *Nature*, 449:913–919.
818 Sagitov, S. (1999). The general coalescent with asynchronous mergers of ancestral lines.
819 *J. Appl. Prob.*, 36:1116–1125.
820 Sargsyan, O. and Wakeley, J. (2008). A coalescent process with simultaneous multiple
821 mergers for approximating the gene genealogies of many marine organisms. *Theor.*
822 *Pop. Biol.*, 74:104–114.
823 Sarvas, T. H. and Fevolden, S. E. (2005). Pantophysin (*Pan* I) locus divergence between
824 inshore v. offshore and northern v. southern populations of Atlantic cod in the north-
825 east Atlantic. *J. Fish Biol.*, 67:444–469.
826 Schweinsberg, J. (2000). Coalescents with simultaneous multiple collisions. *Elec. J.*
827 *Prob.*, 5:1–50.
828 Schweinsberg, J. (2003). Coalescent processes obtained from supercritical Galton-
829 Watson processes. *Stoch. Proc. Appl.*, 106:107–139.
830 Shin, J.-H., Blay, S., McNeney, B., and Graham, J. (2006). Ldheatmap: An R function
831 for graphical display of pairwise linkage disequilibria between single nucleotide
832 polymorphisms. *J. Stat. Soft.*, 16:Code Snippet 3.
833 Sigurgíslason, H. and Árnason, E. (2003). Extent of mitochondrial DNA sequence
834 variation in Atlantic cod from the Faroe Islands: A resolution of gene genealogy.
835 *Heredity*, 91:557–564.
836 Star, B., Nederbragt, A. J., Jentoft, S., Grimholt, U., Malmstrom, M., Gregers, T. F.,
837 Rounge, T. B., Paulsen, J., Solbakken, M. H., Sharma, A., Wetten, O. F., Lanzen, A.,
838 Winer, R., Knight, J., Vogel, J.-H., Aken, B., Andersen, Ø., Lagesen, K., Tooming-
839 Klunderud, A., Edvardsen, R. B., Tina, K. G., Espelund, M., Nepal, C., Previti, C.,
840 Karlsen, B. O., Moum, T., Skage, M., Berg, P. R., Gjoen, T., Kuhl, H., Thorsen, J.,
841 Malde, K., Reinhardt, R., Du, L., Johansen, S. D., Searle, S., Lien, S., Nilsen, F.,
842 Jonassen, I., Omholt, S. W., Stenseth, N. C., and Jakobsen, K. S. (2011). The genome
843 sequence of Atlantic cod reveals a unique immune system. *Nature*, 477:207–210.
844 Steinrücken, M., Birkner, M., and Blath, J. (2013). Analysis of DNA sequence variation

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.528v1 | CC-BY 4.0 Open Access | rec: 9 Oct 2014, publ: 9 Oct 2014

within marine species using Beta-coalescents. *Theor. Pop. Biol.*, 87:15–24.

Storz, J. F. (2005). Using genome scans of DNA polymorphisms to infer adaptive population divergence. *Mol. Ecol.*, 14(671–688).

Storz, J. F., Payseur, B. A., and Nachman, M. W. (2004). Genome scans of DNA variability in humans reveal evidence for selective sweeps outside of Africa. *Mol. Biol. Evol.*, 21:1800–1811.

Tavaré, S. (2004). Ancestral inference in population genetics. In Picard, J., editor, *Lectures on Probability Theory and Statistics. Ecole d'Eté de Probabilité de Saint-Flour XXXI–2001*, volume 1837 of *Lecture Notes in Mathematics*, pages 1–188. Springer Verlag, New York.

Therkildsen, N. O., Hemmer-Hansen, J., Hedeholm, R. B., Wisz, M. S., Pampoulie, C., Meldrup, D., Bonanomi, S., Retzel, A., Olsen, S. M., , and Nielsen, E. E. (2013). Spatiotemporal SNP analysis reveals pronounced biocomplexity at the northern range margin of Atlantic cod *Gadus morhua*. *Evol. Appl.*, 6:690–705.

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

Tishkoff, S. A., Reed, F. A., Ranciaro, A., Voight, B. F., Babbitt, C. C., Silverman, J. S., Powell, K., Mortensen, H. M., Hirbo, J. B., Osman, M., Ibrahim, M., Omar, S. A., Lema, G., Nyambo, T. B., Ghori, J., Bumpstead, S., Pritchard, J. K., Wray, G. A., and Deloukas, P. (2007). Convergent adaptation of human lactase persistence in Africa and Europe. *Nat. Genet.*, 39:31–40.

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*. *Am. J. Hum. Genet.*, 71:1112–1128.

Voight, B. F., Kudaravalli, S., Wen, X., and Pritchard, J. K. (2006). A map of recent positive selection in the human genome. *PLoS Biol.*, 4:e72.

Wakeley, J. (2009). *Coalescent Theory*. Roberts and Company Publishers, Greenwood Village, Colorado, USA.

Wakeley, J. (2013). Coalescent theory has many new branches. *Theor. Pop. Biol.*, 87:1–4.

Wallimann, T., Tokarska-Schlattner, M., and Schlattner, U. (2011). The creatine kinase system and pleiotropic effects of creatine. *Amino Acids*, 40:1271–1296.

Wallimann, T., Wyss, M., Brdiczka, D., Nicolay, K., and Eppenberger, H. (1992). Intracellular compartmentation, structure and function of creatine kinase isoenzymes in tissues with high and fluctuating energy demands: the 'phosphocreatine circuit' for cellular energy homeostasis. *Biochem. J.*, 281:21–40.

Williams, G. C. (1975). *Sex and Evolution*. Princeton University Press, Princeton, New Jersey.

Wiuf, C. and Hein, J. (1999). Recombination as a point process along sequences. *Theor. Pop. Biol.*, 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*, 168:1071–1076.

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.528v1 | CC-BY 4.0 Open Access | rec: 9 Oct 2014, publ: 9 Oct 2014

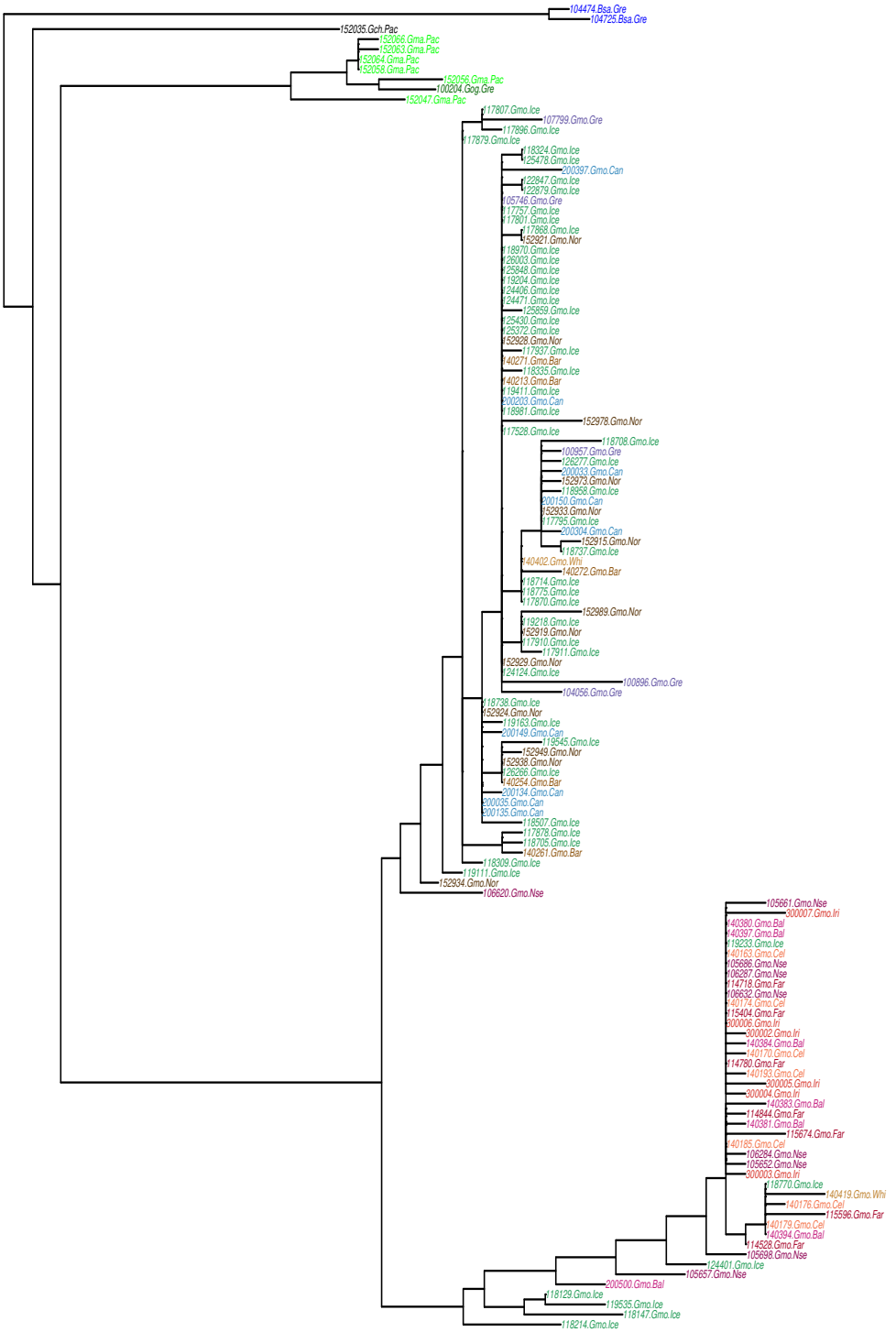

Figure 1. Maximum likelihood tree of *Ckma* variation among 122 individual Atlantic cod and 10 individuals of sister taxa. Localities and color codes as in Figure S1.

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.528v1 | CC-BY 4.0 Open Access | rec: 9 Oct 2014, publ: 9 Oct 2014

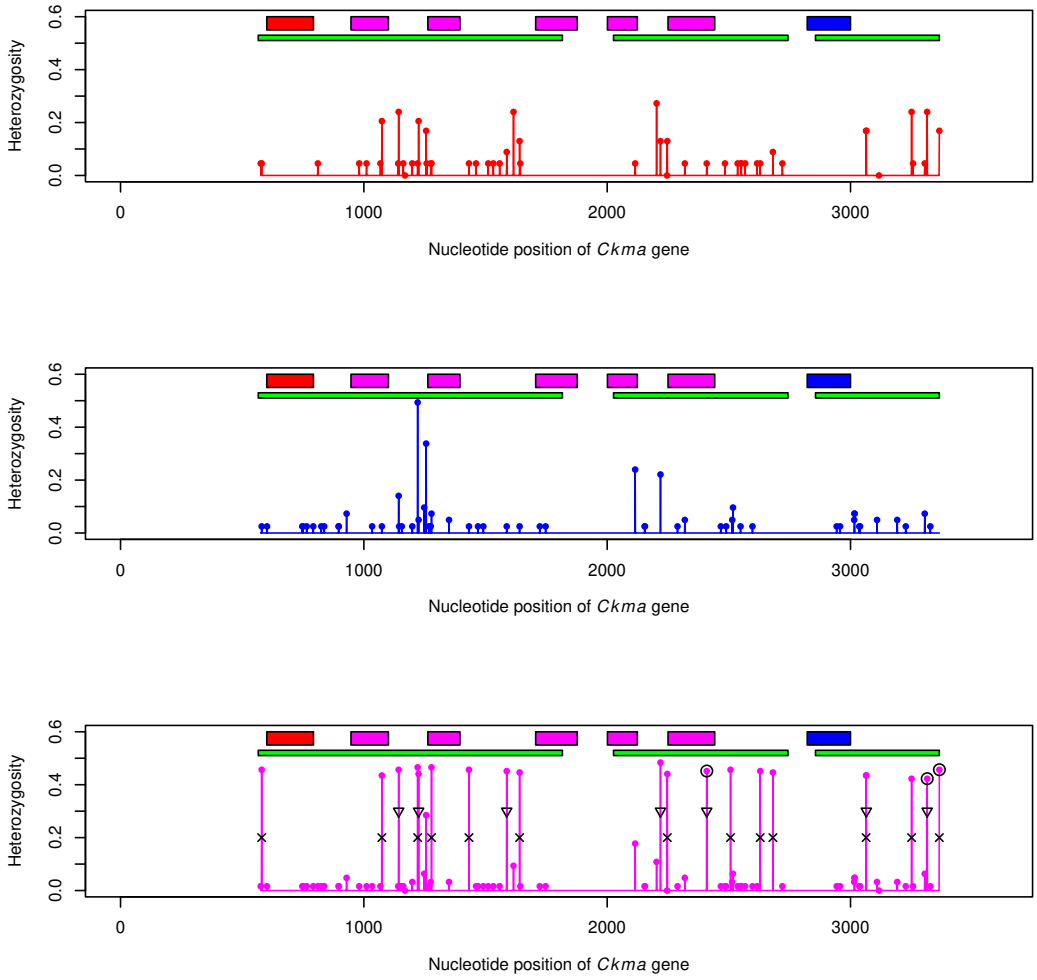

Figure 2. Heterozygosity per nucleotide site of *Ckma* locus among *A* alleles (red top panel, $n = 43$), *B* alleles (blue middle panel, $n = 79$), and all individuals combined (magenta bottom panel, $n = 122$). Boxes represent exons, start (red), internal (magenta) and terminal (blue). Green boxes represent sequenced fragments trimmed to Phred score of at least 30. The black circles mark the three SNPs of Moen et al. (2008), *Gm366-0514* locus with an $F_{ST} = 0.83$, *Gm366-1022* locus with an $F_{ST} = 0.82$, and *Gm366-1073* with an $F_{ST} = 0.82$ from left to right respectively. Crosses mark mutant sites relative to outgroup that were fixed or nearly fixed among *A* alleles. Triangles mark mutant sites relative to outgroup that have were fixed or nearly fixed among *B* alleles. *Gadus macrocephalus* individual 152047 was outgroup.

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.528v1 | CC-BY 4.0 Open Access | rec: 9 Oct 2014, publ: 9 Oct 2014

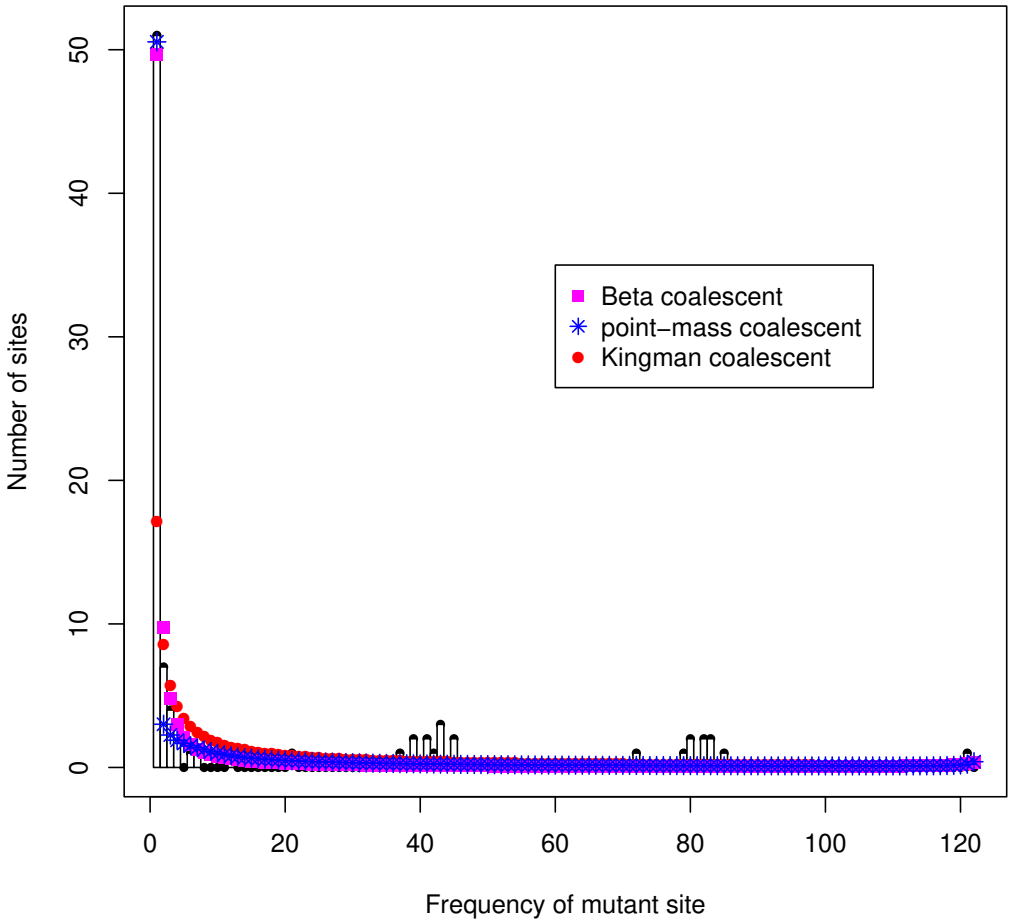

Figure 3. Unfolded site frequency spectrum of Atlantic cod *Ckma* gene. *Gadus macrocephalus* is outgroup. Number of individuals $n = 122$. Theroretical expectation under Kingman coalescent (red dots), Beta$(2-\alpha, \alpha)$ coalescent (magenta squares), and point-mass coalescent (blue stars).

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.528v1 | CC-BY 4.0 Open Access | rec: 9 Oct 2014, publ: 9 Oct 2014

Table 1. Summary statistics of polymorphism of 2500 bp fragment of the *Ckma* gene, 711 bp fragment of the *HbA2* gene and 1021 bp fragment of the *Myg* gene in Atlantic cod.

Group	n	S	H	$\hat{h}$	$\hat{K}$	$\hat{\theta}_S$	$\hat{\pi}$	$\hat{D}$
Ckma all	122	87	72	0.959	10.62	0.0067	0.0043	-1.13^{ns}
Ckma North	86	65	51	0.941	5.12	0.0054	0.0015	-1.97^{ns}
Ckma South	36	45	23	0.891	3.61	0.0045	0.0015	-2.43^{**}
Ckma A allele	43	49	28	0.907	4.37	0.0047	0.0018	-2.20^{**}
Ckma B allele	79	53	44	0.930	3.10	0.0044	0.0013	-2.33^{**}
HbA2 all	114	11	11	0.338	0.37	0.0030	0.0005	-2.09^{*}
HbA2 North	95	9	9	0.347	0.39	0.0025	0.0005	-1.95^{*}
HbA2 South	19	3	4	0.298	0.32	0.0016	0.0005	-0.95^{ns}
Myg all	45	30	24	0.901	2.74	0.0071	0.0028	-2.03^{*}
Myg North	36	28	20	0.894	2.65	0.0069	0.0027	-2.12^{*}
Myg South	9	10	7	0.944	3.22	0.0037	0.0033	-0.58^{ns}

Sample size n, number of segregating sites S, number of haplotypes H, haplotype diversity $\hat{h}$, average number of pairwise differences $\hat{K}$, scaled population size from S θ_S, nucleotide diversity $\hat{\pi}$, and Tajima's $\hat{D}$. ns is not significant, * represents $P < 0.05$, and ** represents $P < 0.01$.

Table 2. Parameter values minimizing the ℓ^2 distance (sum of squares) between observed and expected unfolded site frequency spectra for nuclear genes and for mtDNA variation of various localities.

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

Based on method of Birkner et al. (2013b). Parameters α of the Beta$(2-\alpha, \alpha)$, and ψ of the point-mass coalescent and their respective ℓ^2. The $\ell^2(0)$ is based on the Kingman coalescent for which $\alpha = 2$. For the mtDNA Carr *et al.* refers to Carr and Marshall (1991a,b); Carr et al. (1995); Pepin and Carr (1993), AP 1996 refers to Árnason and Pálsson (1996), and SA 2003 refers to Sigurgíslason and Árnason (2003).

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.528v1 | CC-BY 4.0 Open Access | rec: 9 Oct 2014, publ: 9 Oct 2014