

1 Nucleotide Variation and Balancing 2 Selection at the *Ckma* gene in Atlantic cod: 3 Analysis with multiple merger coalescent 4 models

5 Einar Árnason¹ and Katrín Halldórsdóttir²

6 ¹Institute of Biology, University of Iceland, Reykjavík, Iceland

7 ²Institute of Biology, University of Iceland, Reykjavík, Iceland

8 ABSTRACT

9 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.

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

11 INTRODUCTION

12 High fecundity translates into large excess reproductive capacity that would allow
13 organisms to withstand substantial natural selection enabling them to bear the entailing
14 high genetic load. Based on the concept of the cost of natural selection (Haldane,
15 1957) high-fecundity organisms relative to low-fecundity organisms should at any time
16 be able to adapt a larger proportion of their genome to meet various environmental
17 challenges. Trying to explain the paradox of sexual reproduction Williams (1975) in his

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

32 Coalescent theory (Kingman, 1982a,b) traces the genealogy of a sample and is very
33 **useful for making inference** of molecular data. However, in an extreme case under a
34 winner-take-all sweepstakes reproduction all samples would coalesce immediately in
35 the previous generation (Árnason, 2004) and there would be no variation. The Kingman
36 coalescent, which is derived from (Wright/Fisher) models of low fecundity non-skewed
37 offspring distributions, assumes a bifurcating genealogy and is not appropriate for repro-
38 duction of this kind (Eldon and Wakeley, 2006; Schweinsberg, 2003; Wakeley, 2013).
39 Some organisms may exhibit both high fecundity and highly skewed offspring distri-
40 butions. For these organisms the Λ coalescent allowing multiple mergers of ancestral
41 lineages (Pitman, 1999; Sagitov, 1999; Donnelly and Kurtz, 1999; Eldon and Wakeley,
42 2006; Schweinsberg, 2003; Sargsyan and Wakeley, 2008) or Ξ coalescent allowing
43 simultaneous multiple mergers of ancestral lineages (Schweinsberg, 2000; Möhle and
44 Sagitov, 2001) may be more appropriate. Wakeley (2013) gives an overview of the
45 development of coalescent theory in new directions. There is also active development of
46 statistical inference methods associated with multiple merger coalescents (e.g. Birkner
47 et al., 2013b, 2014). Studies on the high fecundity organisms Pacific oyster *Crassostrea*
48 *gigas* (Hedgecock and Pudovkin, 2011) and Atlantic cod *Gadus morhua* (Árnason and
49 Pálsson, 1996; Árnason et al., 1998, 2000; Carr and Marshall, 1991a; Carr et al., 1995;
50 Pepin and Carr, 1993; Árnason, 2004) have provided data for a number of tests of
51 some of the new coalescent models (Eldon and Wakeley, 2006; Eldon, 2011; Eldon
52 and Degnan, 2012; Steinrücken et al., 2013; Birkner et al., 2013b). Atlantic cod thus
53 provides a model for **studies of multiple merger coalescent**. In this paper we apply
54 some of these new methods for Λ coalescents in a study of balancing selection at a gene
55 showing extreme spatial differentiation in Atlantic cod.

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

64 human populations were compared (Storz et al., 2004). In contrast, comparable genome
65 level studies in Atlantic cod find that 11% (26 out of 235) of independent SNPs (Moen
66 et al., 2008) are F_{ST} outliers (by method of Beaumont and Nichols, 1996) and 4% SNPs
67 (70 out of 1641 Bradbury et al., 2010) are Bayscan outliers (by method of Foll and
68 Gaggiotti, 2008) likely undergoing selection. Similarly one fourth of microsatellite loci
69 in Atlantic cod (Nielsen et al., 2006) are F_{ST} outliers. This supports our thesis that a
70 considerable fraction of the Atlantic cod genome may be simultaneously under selection
71 for different adaptations.

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

84 The Moen et al. (2008) study of differentiation among four Atlantic cod populations
85 along the coast of Norway showed no differentiation among presumably neutral non-
86 outliers loci with an average $\bar{F}_{ST} = 0.0012$. In contrast, the outlier loci, presumably
87 under selection, the average $\bar{F}_{ST} = 0.27$ ranging from 0.08 to an extreme differentiation of
88 0.83, representing almost fixation of alternative alleles. We analyze nucleotide variation
89 at a large fragment of the gene showing extreme spatial differentiation to understand the
90 nature of selection. It is the *Ckma* gene encoding a muscle isoform A of creatine kinase.

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

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

105 MATERIALS AND METHODS

106 Population sampling

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

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

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

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

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

148 **Molecular analysis**

149 We used sequences associated with the Moen et al. (2008) high F_{ST} SNP's (Gm366-
150 0514 with an $F_{ST} = 0.83$, Gm366-1022 with an $F_{ST} = 0.82$, and Gm366-1073 with an
151 $F_{ST} = 0.82$) to make probes to search an Atlantic cod BAC library. We had positive
152 clones 454 sequenced (Microsynth) and obtained a 34223 bp scaffold containing the gene
153 of interest. From this sequence we generated primers (Table S1) for PCR amplifying a
154 4000 bp fragment for population studies. Our scaffold largely but not entirely aligned
155 to GeneScaffold 4232 of the Atlantic cod genome sequence (Star et al., 2011) (www.ensemble.org).
156

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

161 For neutral locus comparisons we applied the same methods and sequenced 711 bp
162 of the Hemoglobin α 2 (*HbA2*) locus (Halldórsdóttir and Árnason, 2009a,b; Borza et al.,
163 2009) and 1021 bp of the myoglobin (*Myg*) locus (Lurman et al., 2007). The *HbA2* data
164 were of 114 Atlantic cod individuals and 13 individuals of various sister taxa. The *Myg*
165 data were from 45 Atlantic cod individuals and two individuals of Pacific cod. Other
166 sister taxa did not amplify for *Myg*. The *HbA2* and *Myg* individuals covered much the
167 same geographic localities as *Ckma*.

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

171 **Statistical analysis**

172 We base called, assembled and edited sequence reads using phred, phrap and
173 consed (Ewing et al., 1998; Ewing and Green, 1998; Gordon et al., 1998). We
174 aligned sequences using muscle (Edgar, 2004), inspected alignments using seaview
175 (version 4) (Gouy et al., 2009) and generated maximum likelihood trees with phym1
176 (Guindon and Gascuel, 2003) under seaview. We used R (R Core Team, 2013) and
177 the ape, pegas, seqinr, ade4, adegenet, and LDheatmap packages (Paradis
178 et al., 2004; Paradis, 2010; Charif and Lobry, 2007; Dray and Dufour, 2007; Jombart and
179 Ahmed, 2011; Shin et al., 2006) and various function written by us for managing, ana-
180 lyzing, and plotting the data. We used the MLHKA program (Wright and Charlesworth,
181 2004) for a maximum likelihood HKA test (Hudson et al., 1987) based on the Kingman
182 coalescent.

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

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

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

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

212 RESULTS

213 Gene and protein

214 The gene is *Ckma* encoding creatin kinase muscle isoform a (CKMA). The locus
215 is 3604 base pairs (bp) in GeneScaffold 4232 (coordinates 332764 to 336367, gene
216 name ENSGMOG00000008778 in the cod genome, www.ensembl.org Star et al.
217 (2011)). The gene has seven exons (Figure S2). Ensemble reports 382 amino acids
218 (aa). However, both genescan (<http://genes.mit.edu/GENSCAN.html>)
219 and fgenesh (www.softberry.com) predicted 381 aa and our analysis of our own
220 data confirmed that. The www.ensembl.org sequence adds a Glycine (G) residue
221 in position 323 apparently due to incorrect splicing at the junction of the last two exons.

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

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

230 Nucleotide variation and divergence

231 The variants of *Ckma* in Atlantic cod fell into two distinct and divergent groups which
232 we refer to as *A* and *B* alleles (Figures 1 and S4). They were fixed for a C vs T at site
233 1732 in the concatenated sequence (Table S2). There also were nearly fixed differences
234 between the alleles at 19 additional sites (Figure 2 and Table S2).

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

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

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

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

264 There was considerable linkage disequilibrium (LD) throughout the gene (Figure S7).
265 The high heterozygosity sites nearly fixed between the alleles were influential in gener-
266 ating LD between sites throughout the gene.

267 The results of a maximum likelihood HKA test of selection that is based on the
268 Kingman coalescent (Wright and Charlesworth, 2004) gave a selection parameter $k =$
269 2.12 in the direction of balancing selection (Table S5). However, the results were not
270 statistically significant possibly because of too high variation among the presumed
271 neutral loci (*HbA2* and *Myg*) used for comparison in the test.

272 **Spatial differentiation**

273 The variation was spatially patterned. The *A* allele was nearly fixed in an area that
274 we call South (Faroe Islands, North Sea, Baltic Sea, Celtic Sea and Irish Sea) at a
275 frequency of 97% (Table S6). Conversely the *B* allele was at a high frequency of
276 92% in an area that we call North ranging from the Northwest (Nova Scotia and
277 Newfoundland in Canada) through Greenland, Iceland, Norway, Barents Sea and the
278 White Sea. The differentiation was evident in interlocality F_{ST} values (Table S7).
279 There was no significant differentiation among localities within either the North or the
280 South but very high and significant differentiation between North and South localities.

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

283 The high differentiation was mostly due to the great number of fixed or nearly fixed
284 sites between the two alleles (Figure 2 and Table S2). Three of the sites were the SNPs
285 already found by Moen et al. (2008) with an $F_{ST} = 0.82$. The high frequency sites
286 showed indications of recombination between the *A* and *B* alleles (see for example
287 patterns of segregating sites for individuals 105698, 124401, 105657, 200500, 118129,
288 119535, 118147, and 106620 in Table S2).

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

296 Site frequency spectra

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

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

308 The site frequency spectra of the *A* and *B* alleles alone were bimodal with a high
309 singleton class and peaks around 40 and 78 respectively (Figure S10). The 40 and
310 78 modes came about because most of the high frequency and high heterozygosity
311 sites that separate the two alleles were not fixed within each allele presumably due to
312 recombination (Table S2).

313 Coalescent parameter estimates

314 Following Birkner et al. (2013b) we used the ℓ^2 distance, the sum of the squared
315 differences between the observed and expected site frequency spectrum (scaled with the
316 number of segregating sites), for estimating parameters of two Λ coalescent models, $\hat{\alpha}$
317 for the Beta($2 - \alpha, \alpha$) and $\hat{\psi}$ for the point-mass coalescent (Table 2 and Figures S11
318 and S12). The Kingman coalescent, a null model for which $\alpha = 2.0$, had the highest ℓ^2
319 indicating worst fit among the models. The *HbA2* and *Myg* loci had an $\hat{\alpha} = 1.00$ and a
320 $\hat{\psi} = 0.23$. The *Ckma* locus had overall a considerably higher α and lower ψ . The two
321 alleles separately were in the direction of the presumed neutral loci *HbA2* and *Myg*.

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

325 as outgroup (Table 2 and Figure S12). Previously these have been analysed using the
326 folded site frequency spectrum (see for example Birkner et al., 2013b; Steinrücken et al.,
327 2013). For the total sample, spanning a similar geographic range as the nuclear genes,
328 the parameter estimates differed from the nuclear loci with $\hat{\alpha} = 1.53$ and $\hat{\psi} = 0.01$. The
329 large samples from Newfoundland and Iceland and the sample from the Faroe Islands
330 gave similar values. The values for Greenland, Norway, White Sea, and Baltic Sea were
331 much closer to the results for the Kingman coalescent ($\alpha = 2.0$). For these localities
332 homoplasies were relatively somewhat more frequent in the data than for the total and
333 the large samples. Homoplasies will reduce the number of singletons and move such
334 sites towards the right tail of the site frequency distribution. This explains the higher
335 values for these localities.

336 DISCUSSION

337 Genes and proteins

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

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

355 Allele divergence and spatial differentiation

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

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

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

377 A counter argument is that isolation and admixture are part of the breeding structure
378 of a population leaving genome-wide impacts (Wright, 1931). Under this scenario
379 different genes should be concordant in their behavior (Bernardi et al., 1993). The
380 *Hba2* and the *Myg* show no differentiation between the North and the South. Also the
381 non-outlier SNPs in Moen et al. (2008) show no differentiation whereas three SNPs of
382 the *Ckma* gene show high and extreme F_{ST} . Similarly, Bradbury et al. (2010) find that
383 non-outlier SNPs show no differentiation although other SNPs show differentiation from
384 parallel adaptation to temperature on the eastern and western side of the Atlantic Ocean.
385 Nielsen et al. (2003) describe a pattern of microsatellite variation in a transition area
386 between the Baltic and Danish Belt Sea which they interpret as a hybrid zone. There is
387 no evidence for a hybrid zone at that location in the *Ckma* data. In fact, specific variants
388 within the *A* allele are widely distributed among localities in the South including the
389 Baltic Sea. This implies gene flow among localities in the South. Similar patterns within
390 *B* alleles imply gene flow among localities in the North. If indeed there is a hybrid zone
391 for the *Ckma* gene it would lie between the Faroe Islands on one hand and Iceland and
392 north and middle Norway on the other hand. It is not a parsimonious explanation to
393 consider there to be multiple hybrid zones of secondary contact within distribution of
394 the species.

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

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

413 **Site frequency spectra**

414 The trimodal site frequency spectrum is not predicted by any of the coalescent mod-
415 els considered here, the Kingman coalescent and the two Λ coalescent models, the
416 Beta($2 - \alpha, \alpha$) (Schweinsberg, 2003) and the point-mass coalescent (Eldon and Wake-
417 ley, 2006). Under the Λ coalescent at most a single multiple merger event occurs at any
418 one time. The distribution of family size is of interest and the parameter α influences
419 the probability of getting large families. Under the Beta($2 - \alpha, \alpha$) coalescent model the
420 probability of a family size of k or more viable offspring decays like $k^{-\alpha}$ (Schweinsberg,
421 2003) in the limit of a large k . The pool of viable offspring is then resampled to form the
422 next generation under the same conditions. For the Kingman coalescent $\alpha \geq 2$ and there
423 is little chance of seeing large families. For the Beta($2 - \alpha, \alpha$) coalescent $1 \leq \alpha < 2$ and
424 the lower α the greater is the chance of seeing a large family (Schweinsberg, 2003). The
425 ψ parameter of the point-mass coalescent (Eldon and Wakeley, 2006) similarly measures
426 the proportion of the population that is the offspring of a single individual and is thus an
427 indicator of reproductive sweepstakes. Our estimates of ψ indicate reproductive sweep-
428 stakes at the neutral loci and within the A and B alleles of *Ckma*. Balancing selection
429 at *Ckma* lessens the effects of sweepstakes reproduction. Sweepstakes reproduction
430 has been detected in other high fecundity organisms (Hedgecock and Pudovkin, 2011;
431 Harrang et al., 2013).

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

446 **Balancing selection**

447 Balancing selection generates long branches in the genealogy and neutral variation
448 accumulates on the branches. The balanced functional types (the *Ckma* A and B alleles
449 in this case) act as they were separate and isolated populations accumulating neutral
450 variation. Recombination can bring variation from one branch to another acting like
451 migration that brings alleles from one population to another (Charlesworth et al., 1997,
452 2003; Charlesworth, 2006). However, the molecular signatures of balancing selection
453 depend on many factors. Is it a long standing, even trans-species, polymorphism such as
454 *MHC* in human and chimpanzee (Fan et al., 1989; Nei and Hughes, 1991) or *Cathelicidin*
455 in gadids (Halldórsdóttir and Árnason, 2014) or is it very recent? Examples of the
456 latter are human glucose 6 phosphate dehydrogenase (G6PD) (Verrelli et al., 2002), and
457 hemoglobin β S (Currat et al., 2002) and hemoglobin β E (Ohashi et al., 2004) and

458 spatially divergent selection of lactase persistence (Tishkoff et al., 2007; Ranciaro et al.,
459 2014) in which a particular allele sweeps a chromosomal segment to an intermediate
460 equilibrium frequency. In these instances recombination has not had time to break up
461 LD which can extend over large regions. There is very little variation among the new
462 alleles while the alternative chromosomes show much more variation in this region
463 representing the standing variation in the population at the start of the partial sweep.

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

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

490 **The coalescent used here are models of neutrality.** One could argue that it is not
491 appropriate to apply such neutral models to the *Ckma* locus that is already suspected to
492 be under selection. However, understanding how the locus deviates from neutrality is
493 important for understanding the pattern of selection. Under the neutral theory (Kimura,
494 1983) polymorphism within species is the transient phase of molecular evolution that
495 leads to divergence between species. This is the rationale for the HKA test of selection or
496 neutrality (Hudson et al., 1987) that neutrally evolving genomic regions should have
497 the same proportion of polymorphism to divergence, Balancing selection would tend to
498 increase the level of polymorphism within species relative to divergence between them.
499 The results of HKA test are in the direction of balancing selection. The HKA test shows
500 a relative slowing down of divergence to rate of polymorphism at the *Ckma* locus.

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

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

512 **Coalescent parameter estimates and time scales**

513 The question of coalescent time scale, however, must be considered. Under the Kingman
514 coalescent time is measured in terms of N/σ^2 , population size scaled by the variance of
515 family size (Sagitov, 1999; Árnason, 2004; Tavaré, 2004). With a Poisson distribution
516 of family size $\sigma^2 = 1$ for a constant size haploid population so times scales with N . In
517 an extreme winner-take-all sweepstakes $\sigma^2 = N$ and a sample would coalesce in the
518 previous generation and there would be no variation (Árnason, 2004). In more realistic
519 multiple merger coalescent models the time scale is the quantity $c_N = \frac{E(v_1-1)^2}{N-1}$ where
520 c_N is the probability of two lineages coalescing in the previous generation in a haploid
521 population of fixed size N and v_1 is the random number of offspring of individual 1
522 (Sagitov, 1999). In general the time scale of multiple merger coalescent models can be
523 much shorter than for Kingman coalescent. Under the Beta($2 - \alpha$, α) coalescent model
524 time scales with $N^{\alpha-1}$ (Schweinsberg, 2003; Birkner et al., 2014). For this model our
525 estimates of α for the nuclear genes are quite low which implies very short time scales.
526 The neutral genes would seem to coalesce in the very recent past. The *A* and *B* alleles
527 of *Ckma* run on very similar time scales to the neutral genes and the locus itself at a
528 slower rate due to the balancing selection with a time scale approximately the cube root
529 of the effective population size N_e . The mitochondrial DNA runs at yet another and
530 slower time scale. For mtDNA time scales with approximately the square root of N .
531 Predicted turnover of alleles is faster and ages of alleles shorter under multiple merger
532 coalescent (Eldon, 2912). Different populations and species may run on different time
533 scales (Eldon and Degnan, 2012) complicating divergence time estimates. Estimates
534 based on Kingman coalescent of divergence times of Atlantic cod populations (Bigg
535 et al., 2008) or divergence of gadid taxa (Coulson et al., 2006) may therefore be too high
536 and may need revision.

537 **Conclusion**

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

548 of action of selection (Nielsen, 2005). Some of the variation may be neutral having
549 risen in frequency within the balanced functional allele where it arose (Charlesworth,
550 2006). Alternatively some of the variation may be due to selection building co-adapted
551 complexes (Thompson and Jiggins, 2014). In addition to a peak at singleton sites,
552 characteristic of multiple-merger coalescent, the site frequency spectrum has two high-
553 frequency modes at opposite but matching frequencies representing the two branches of
554 the genealogy. This pattern is further support for balancing selection. Finally time scales
555 faster under multiple-merger than the Kingman coalescent. Our estimates of parameters
556 of multiple-merger Λ coalescent show that time-scales are fast.

557 **ACKNOWLEDGMENTS**

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

570 **LITERATURE CITED**

- 571 Akey, J. (2009). Constructing genomic maps of positive selection in humans: Where do
 572 we go from here? *Genome Res.*, 19:711–722.
- 573 Árnason, E. (2004). Mitochondrial cytochrome *b* DNA variation in the high-fecundity At-
 574 lantic cod: Trans-Atlantic clines and shallow gene genealogy. *Genetics*, 166(4):1871–
 575 1885.
- 576 Árnason, E., Hernandez, U. B., and Kristinsson, K. (2009). Intense habitat-specific
 577 fisheries-induced selection at the molecular *Pan I* locus predicts imminent collapse of
 578 a major cod fishery. *PLoS ONE*, 4(5):e5529.
- 579 Árnason, E. and Pálsson, S. (1996). Mitochondrial cytochrome *b* DNA sequence
 580 variation of Atlantic cod, *Gadus morhua*, from Norway. *Mol. Ecol.*, 5:715–724.
- 581 Árnason, E., Pálsson, S., and Petersen, P. H. (1998). Mitochondrial cytochrome *b* DNA
 582 sequence variation of Atlantic cod, *Gadus morhua*, from the Baltic- and the White
 583 Seas. *Hereditas*, 129:37–43.
- 584 Árnason, E., Petersen, P. H., Kristinsson, K., Sigurgíslason, H., and Pálsson, S. (2000).
 585 Mitochondrial cytochrome *b* DNA sequence variation of Atlantic cod from Iceland
 586 and Greenland. *J. Fish Biol.*, 56:409–430.
- 587 Beaumont, M. A. and Nichols, R. A. (1996). Evaluating loci for use in the genetic
 588 analysis of population structure. *Proc. R. Soc. Ser. B.*, 263:1619–1626.
- 589 Bernardi, G., Sordino, P., and Powers, D. A. (1993). Concordant mitochondrial and
 590 nuclear DNA phylogenies for populations of the teleost fish *Fundulus heteroclitus*.
 591 *Proc. Natl. Acad. Sci.*, 90:9271–9274.
- 592 Bigg, G. R., Cunningham, C. W., Ottersen, G., Pogson, G. H., Wadley, M. R., and
 593 Williamson, P. (2008). Ice-age survival of Atlantic cod: agreement between palaeoe-
 594 cology models and genetics. *Proc. R. Soc. B.*, 275:163–172.
- 595 Birkner, M., Blath, J., and Eldon, B. (2013a). An ancestral recombination graph for
 596 diploid populations with skewed offspring distribution. *Genetics*, 193:255–290.
- 597 Birkner, M., Blath, J., and Eldon, B. (2013b). Statistical properties of the site-frequency
 598 spectrum associated with Λ -coalescents. *Genetics*, 195:1037–1053.
- 599 Birkner, M., Blath, J., Eldon, B., and Freund, F. (2014). Can the site-frequency spec-
 600 trum distinguish exponential population growth from multiple-merger coalescents?
 601 *bioRxiv*.
- 602 Borza, T., Higgins, B., Simpson, G., and Bowman, S. (2010). Integrating the markers
 603 *Pan I* and Haemoglobin with the genetic linkage map of Atlantic cod (*Gadus morhua*).
 604 *BMC Res. Notes*, 3:261.
- 605 Borza, T., Stone, C., Gamperl, A. K., and Bowman, S. (2009). Atlantic cod (*Gadus*
 606 *morhua*) hemoglobin genes: multiplicity and polymorphism. *BMC Genet.*, 10:51.
- 607 Bowcock, A. M., Kidd, J. R., Mountain, J. L., Herbert, J. M., Carotenuto, L., Kidd, K. K.,
 608 and Cavalli-Sforza, L. (1991). Drift, admixture, and selection in human evolution: A
 609 study with DNA polymorphisms. *Proc. Natl. Acad. Sci.*, 88:839–843.
- 610 Bradbury, I. R., Hubert, S., Higgins, B., Borza, T., Bowman, S., Paterson, I. G., Snel-
 611 grove, P. V. R., Morris, C. J., Gregory, R. S., Hardie, D. C., Hutchings, J. A., Ruzzante,
 612 D. E., Taggart, C. T., and Bentzen, P. (2010). Parallel adaptive evolution of Atlantic
 613 cod on both sides of the Atlantic Ocean in response to temperature. *Proc. R. Soc. B.*,
 614 277:3725–3734.

- 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 (*PanI*) 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

- 661 population models. *Ann. Prob.*, 27:166–205.
- 662 Dray, S. and Dufour, A. (2007). The ade4 package: Implementing the duality diagram
663 for ecologists. *J. Stat. Soft.*, 22:1–20.
- 664 Edgar, R. C. (2004). Muscle: Multiple sequence alignment with high accuracy and high
665 throughput. *Nucl. Acids Res.*, 32:1792–1797.
- 666 Eiríksson, G. M. and Árnason, E. (2013). Spatial and temporal microsatellite variation
667 in spawning Atlantic cod, *Gadus morhua*, around Iceland. *Can. J. Fish. Aquat. Sci.*,
668 70:1151–1158.
- 669 Eldon, B. (2011). Estimation of parameters in large offspring number models and ratios
670 of coalescence times. *Theor. Pop. Biol.*, 80:16–28.
- 671 Eldon, B. (2012). Age of an allele and gene genealogies of nested subsamples for
672 populations admitting large offspring numbers. *arXiv*, 1212.1792v1.
- 673 Eldon, B. and Degnan, J. H. (2012). Multiple merger gene genealogies in two species:
674 Monophyly, paraphyly, and polyphyly for two examples of lambda coalescents. *Theor.*
675 *Pop. Biol.*, 82:117–130.
- 676 Eldon, B. and Wakeley, J. (2006). Coalescent processes when the distribution of
677 offspring number among individuals is highly skewed. *Genetics*, 172:2621–2633.
- 678 Eldon, B. and Wakeley, J. (2008). Linkage disequilibrium under skewed offspring
679 distribution among individuals in a population. *Genetics*, 178(3):1517–1532.
- 680 Ewing, B. and Green, P. (1998). Basecalling of automated sequencer traces using phred.
681 II. error probabilities. *Genome Res.*, 8:186–194.
- 682 Ewing, B., Hillier, L., Wendl, M., and Green, P. (1998). Base-calling of automated
683 sequencer traces using phred. I. accuracy assessment. *Genome Res.*, 8:175–185.
- 684 Fan, W., Kasahara, M., Gutknecht, J., Klein, D., Mayer, W. E., Jonker, M., and Klein,
685 J. (1989). Shared class II MHC polymorphisms between humans and chimpanzees.
686 *Hum. Immunol.*, 26(2):107–121.
- 687 Fevolden, S. E. and Pogson, G. H. (1995). Differences in nuclear DNA RFLPs between
688 the Norwegian coastal and the Northeast Arctic populations of Atlantic cod. In
689 Skjoldal, H. R., Hopkins, C., Eriksstad, K. E., and Leinaas, H. P., editors, *Ecology of*
690 *Fjords and Coastal Waters*, pages 403–414, Amsterdam, The Netherlands. Elsevier
691 Science Publishers.
- 692 Fevolden, S. E. and Pogson, G. H. (1997). Genetic divergence at the Synaptophysin
693 (*Syp I*) locus among Norwegian coastal and north-east Arctic populations of Atlantic
694 cod. *J. Fish Biol.*, 51:895–908.
- 695 Foll, M. and Gaggiotti, O. (2008). A genome-scan method to identify selected loci
696 appropriate for both dominant and codominant markers: A Bayesian perspective.
697 *Genetics*, 180:977–993.
- 698 Gordon, D., Abajian, C., and Green, P. (1998). Consed: A graphical tool for sequence
699 finishing. *Genome Res.*, 8:195–202.
- 700 Gouy, M., Guindon, S., and Gascuel, O. (2009). SeaView version 4: A multiplat-
701 form graphical user interface for sequence alignment and phylogenetic tree building.
702 *Molecular Biology and Evolution*, 27:221–224.
- 703 Guinand, B., Lemaire, C., and Bonhomme, F. (2004). How to detect polymorphisms
704 undergoing selection in marine fishes? a review of methods and case studies, including
705 flatfishes. *Journal of Sea Research*, 51:167–182.
- 706 Guindon, S. and Gascuel, O. (2003). A simple, fast and accurate algorithm to estimate

- 707 large phylogenies by maximum likelihood. *Syst. Biol.*, 52:696–704.
- 708 Haldane, J. B. S. (1957). The cost of natural selection. *J. Genet.*, 55:511–524.
- 709 Halldórsdóttir, K. and Árnason, E. (2009a). Multiple linked β and α globin genes in
710 Atlantic cod: a PCR based strategy of genomic exploration. *Mar. Gen.*, 2:169–181.
- 711 Halldórsdóttir, K. and Árnason, E. (2009b). Organization of a β and α globin gene set
712 in the teleost Atlantic cod, *Gadus morhua*. *Biochem. Genet.*, 47:817–830.
- 713 Halldórsdóttir, K. and Árnason, E. (2014). Trans-species balanced polymorphism at
714 antimicrobial Cathelicidin genes of the innate immunity system. *In review*, 1.
- 715 Harrang, E., Lapegue, S., Morga, B., and Bierne, N. (2013). A high load of non-
716 neutral amino-acid polymorphisms explains high protein diversity despite moderate
717 effective population size in a marine bivalve with sweepstakes reproduction. *G3:
718 Genes|Genomes|Genetics*, 3(2):333–341.
- 719 Hedgecock, D. and Pudovkin, A. I. (2011). Sweepstakes reproductive success in highly
720 fecund marine fish and shellfish: A review and commentary. *Bull. Mar. Sci.*, 87:971–
721 1002.
- 722 Hemmer-Hansen, J., Nielsen, E. E., Therkildsen, N. O., Taylor, M. I., Ogden, R., Geffen,
723 A. J., Bekkevold, D., Helyar, S., Pampoulie, C., Johansen, T., Consortium, F., and
724 Carvalho, G. R. (2013). A genomic island linked to ecotype divergence in Atlantic
725 cod. *Mol. Ecol.*, 22:2653–2667.
- 726 Hemmer-Hansen, J., Therkildsen, N. O., Meldrup, D., and Nielsen, E. E. (2014). Con-
727 serving marine biodiversity: Insights from life-history trait candidate genes in Atlantic
728 cod (*Gadus morhua*). *Cons. Genet.*, 15:213–228.
- 729 Hernandez, U. B. and Árnason, E. (2014). DNA sequence variation at the (*Pan I*) locus
730 and its peripheral regions in Atlantic cod (*Gadus morhua*): Analysis of linkage and
731 selection. *In review*, 1.
- 732 Hubert, S., Higgins, B., Borza, T., and Bowman, S. (2010). Development of a SNP
733 resource and a genetic linkage map for Atlantic cod (*Gadus morhua*). *BMC Genomics*,
734 11:191.
- 735 Hudson, R. R., Kreitman, M., and Aguadé, M. (1987). A test of neutral molecular
736 evolution based on nucleotide data. *Genetics*, 116:153–159.
- 737 Johnson, M. S. and Wernham, J. (1999). Temporal variation of recruits as a basis of
738 ephemeral genetic heterogeneity in the western rock lobster *Panulirus cygnus*. *Mar.
739 Biol.*, 135(1):133–139.
- 740 Jombart, T. and Ahmed, I. (2011). ADEGENET 1.3-1: New tools for the analysis of
741 genome-wide SNP data. *Bioinformatics*, 27:3070–3071.
- 742 Joron, M., Frezal, L., Jones, R. T., Chamberlain, N. L., Lee, S. F., Haag, C. R., Whibley,
743 A., Becuwe, M., Baxter, S. W., Ferguson, L., Wilkinson, P. A., Salazar, C., Davidson,
744 C., Clark, R., Quail, M. A., Beasley, H., Glithero, R., Lloyd, C., Sims, S., Jones,
745 M. C., Rogers, J., Jiggins, C. D., and French Constant, R. H. (2011). Chromosomal
746 rearrangements maintain a polymorphic supergene controlling butterfly mimicry.
747 *Nature*, 477:203–206.
- 748 Kimura, M. (1983). *The Neutral Theory of Molecular Evolution*. Cambridge University
749 Press, Cambridge.
- 750 Kingman, J. F. C. (1982a). The coalescent. *Stoch. Proc. Appl.*, 13:235–248.
- 751 Kingman, J. F. C. (1982b). On the genealogy of large populations. *J. Appl. Prob.*,
752 19A:27–43.

- 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 complex 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., and Jun Z. Li, S. K., Absher, D., Srinivasan,
793 B. S., Barsh, G. S., Feldman, R. 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.

- 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

- 845 within marine species using Beta-coalescents. *Theor. Pop. Biol.*, 87:15–24.
- 846 Storz, J. F. (2005). Using genome scans of DNA polymorphisms to infer adaptive
847 population divergence. *Mol. Ecol.*, 14(671–688).
- 848 Storz, J. F., Payseur, B. A., and Nachman, M. W. (2004). Genome scans of DNA
849 variability in humans reveal evidence for selective sweeps outside of Africa. *Mol.*
850 *Biol. Evol.*, 21:1800–1811.
- 851 Tavaré, S. (2004). Ancestral inference in population genetics. In Picard, J., editor,
852 *Lectures on Probability Theory and Statistics. Ecole d’Eté de Probabilité de Saint-*
853 *Flour XXXI–2001*, volume 1837 of *Lecture Notes in Mathematics*, pages 1–188.
854 Springer Verlag, New York.
- 855 Therkildsen, N. O., Hemmer-Hansen, J., Hedeholm, R. B., Wisz, M. S., Pampoulie, C.,
856 Meldrup, D., Bonanomi, S., Retzel, A., Olsen, S. M., , and Nielsen, E. E. (2013).
857 Spatiotemporal SNP analysis reveals pronounced biocomplexity at the northern range
858 margin of Atlantic cod *Gadus morhua*. *Evol. Appl.*, 6:690–705.
- 859 Thompson, M. J. and Jiggins, C. D. (2014). Supergenes and their role in evolution.
860 *Heredity*, 113:1–8.
- 861 Tishkoff, S. A., Reed, F. A., Ranciaro, A., Voight, B. F., Babbitt, C. C., Silverman, J. S.,
862 Powell, K., Mortensen, H. M., Hirbo, J. B., Osman, M., Ibrahim, M., Omar, S. A.,
863 Lema, G., Nyambo, T. B., Ghorri, J., Bumpstead, S., Pritchard, J. K., Wray, G. A., and
864 Deloukas, P. (2007). Convergent adaptation of human lactase persistence in Africa
865 and Europe. *Nat. Genet.*, 39:31–40.
- 866 Verrelli, B. C., McDonald, J. H., Argyropoulos, G., Destro-Bisol, G., Froment, A.,
867 Drousiotou, A., Lefranc, G., Helal, A. N., Loiselet, J., and Tishkoff, S. A. (2002).
868 Evidence for balancing selection from nucleotide sequence analyses of human *G6PD*.
869 *Am. J. Hum. Genet.*, 71:1112–1128.
- 870 Voight, B. F., Kudaravalli, S., Wen, X., and Pritchard, J. K. (2006). A map of recent
871 positive selection in the human genome. *PLoS Biol.*, 4:e72.
- 872 Wakeley, J. (2009). *Coalescent Theory*. Roberts and Company Publishers, Greenwood
873 Village, Colorado, USA.
- 874 Wakeley, J. (2013). Coalescent theory has many new branches. *Theor. Pop. Biol.*,
875 87:1–4.
- 876 Wallimann, T., Tokarska-Schlattner, M., and Schlattner, U. (2011). The creatine kinase
877 system and pleiotropic effects of creatine. *Amino Acids*, 40:1271–1296.
- 878 Wallimann, T., Wyss, M., Brdiczka, D., Nicolay, K., and Eppenberger, H. (1992).
879 Intracellular compartmentation, structure and function of creatine kinase isoenzymes
880 in tissues with high and fluctuating energy demands: the 'phosphocreatine circuit' for
881 cellular energy homeostasis. *Biochem. J.*, 281:21–40.
- 882 Williams, G. C. (1975). *Sex and Evolution*. Princeton University Press, Princeton, New
883 Jersey.
- 884 Wiuf, C. and Hein, J. (1999). Recombination as a point process along sequences. *Theor.*
885 *Pop. Biol.*, 55:248–259.
- 886 Wright, S. (1931). Evolution in Mendelian populations. *Genetics*, 16:97–159.
- 887 Wright, S. I. and Charlesworth, B. (2004). The HKA test revisited: A maximum-
888 likelihood-ratio test of the standard neutral model. *Genetics*, 168:1071–1076.

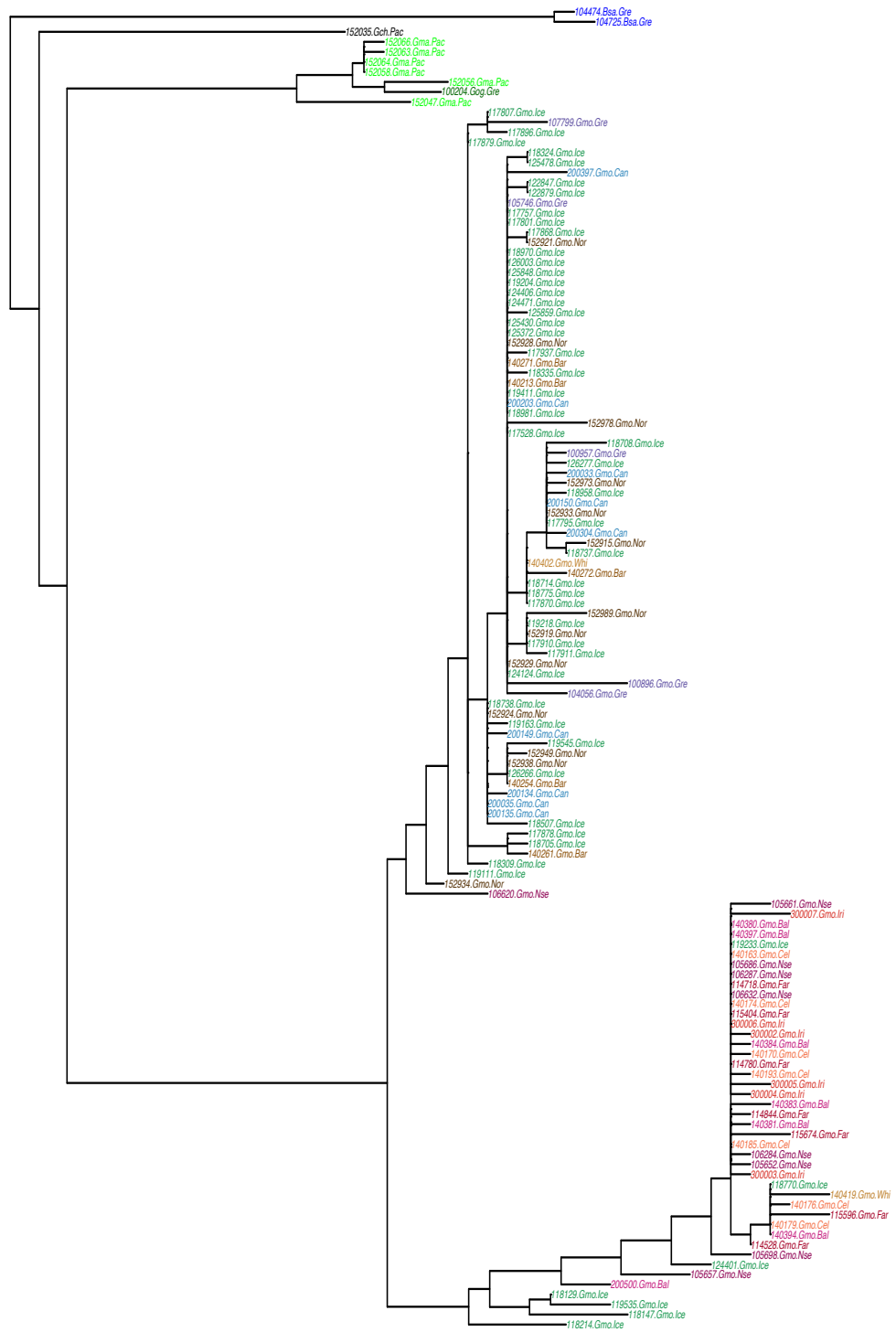


Figure 1. Maximum likelihood tree of *Ckma* variation among 122 individual Atlantic cod and 10 individuals of sister taxa. Localities and color codes as in Figure S1.

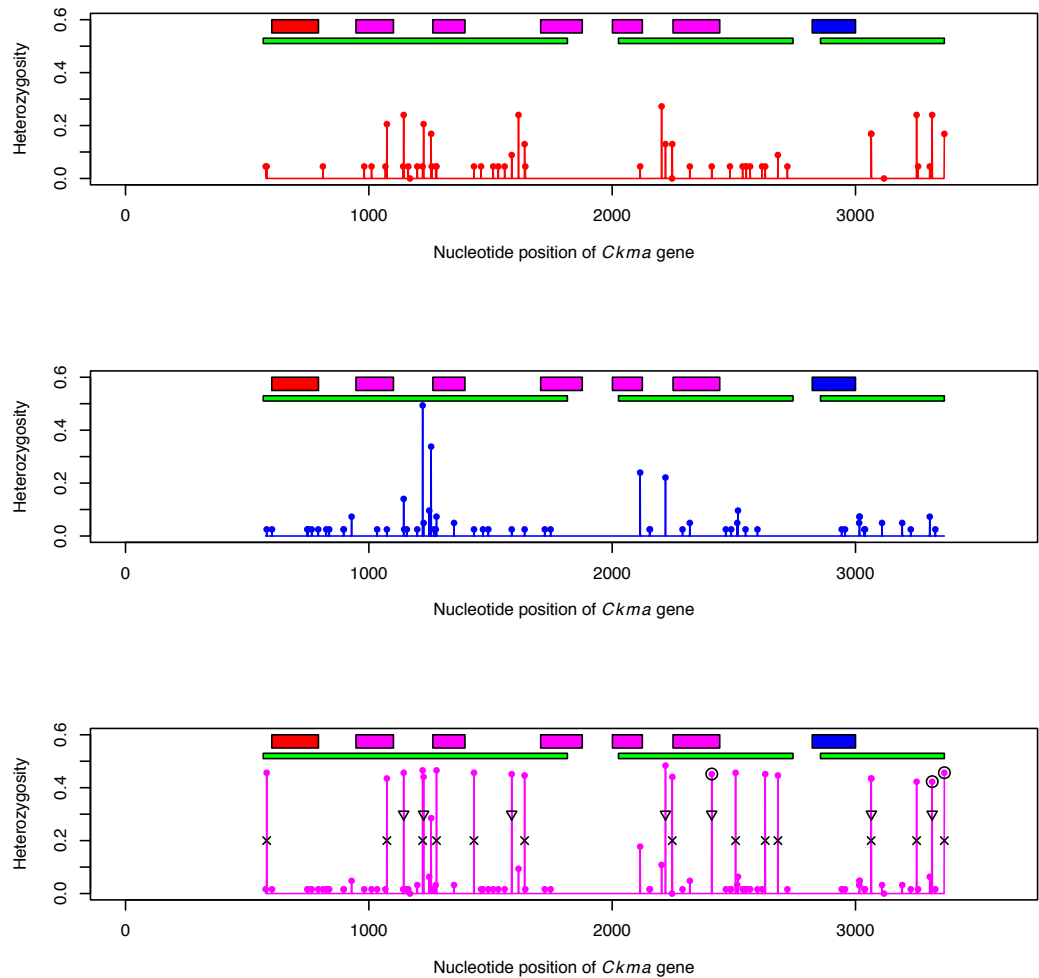


Figure 2. Heterozygosity per nucleotide site of *Ckma* locus among *A* alleles (red top panel, $n = 43$), *B* alleles (blue middle panel, $n = 79$), and all individuals combined (magenta bottom panel, $n = 122$). Boxes represent exons, start (red), internal (magenta) and terminal (blue). Green boxes 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.

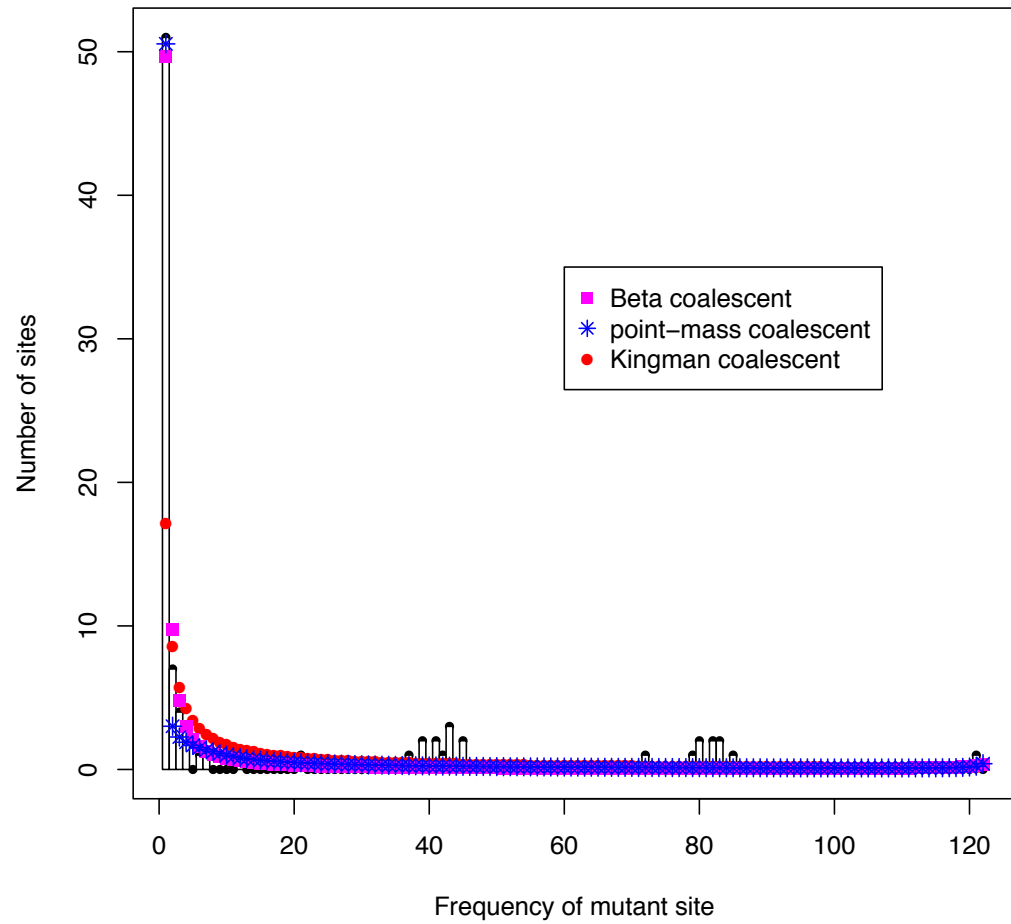


Figure 3. Unfolded site frequency spectrum of Atlantic cod *Ckma* gene. *Gadus macrocephalus* is outgroup. Number of individuals $n = 122$. Theoretical expectation under Kingman coalescent (red dots), Beta($2 - \alpha, \alpha$) coalescent (magenta squares), and point-mass coalescent (blue stars).

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}
<i>Ckma</i> all	122	87	72	0.959	10.62	0.0067	0.0043	-1.13 ^{ns}
<i>Ckma</i> North	86	65	51	0.941	5.12	0.0054	0.0015	-1.97 ^{ns}
<i>Ckma</i> South	36	45	23	0.891	3.61	0.0045	0.0015	-2.43 ^{**}
<i>Ckma</i> A allele	43	49	28	0.907	4.37	0.0047	0.0018	-2.20 ^{**}
<i>Ckma</i> B allele	79	53	44	0.930	3.10	0.0044	0.0013	-2.33 ^{**}
<i>HbA2</i> all	114	11	11	0.338	0.37	0.0030	0.0005	-2.09 [*]
<i>HbA2</i> North	95	9	9	0.347	0.39	0.0025	0.0005	-1.95 [*]
<i>HbA2</i> South	19	3	4	0.298	0.32	0.0016	0.0005	-0.95 ^{ns}
<i>Myg</i> all	45	30	24	0.901	2.74	0.0071	0.0028	-2.03 [*]
<i>Myg</i> North	36	28	20	0.894	2.65	0.0069	0.0027	-2.12 [*]
<i>Myg</i> 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 $\hat{\theta}_S$, 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							
<i>Hba2</i>	1.000	0.230	0.035	0.016	0.431	113	This study
<i>Myg</i>	1.000	0.225	0.010	0.018	0.230	45	This study
<i>Ckma</i>	1.280	0.070	0.006	0.007	0.141	122	This study
<i>Ckma</i> ^A	1.100	0.170	0.017	0.012	0.161	43	This study
<i>Ckma</i> ^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 <i>et al.</i>
Greenland	1.945	0.005	0.072	0.071	0.072	78	Árnason <i>et al.</i> (2000)
Iceland	1.550	0.010	0.006	0.050	0.078	519	Árnason <i>et al.</i> (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 <i>et al.</i> (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 <i>et al.</i> (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).