Nucleotide Polymorphism and Natural Selection at the Pantophysin (*Pan* I) Locus in the Atlantic Cod, *Gadus morhua* (L.)

Grant H. Pogson

Department of Ecology and Evolutionary Biology and Institute of Marine Sciences, University of California, Santa Cruz, California 95064

ABSTRACT

Molecular studies of nucleotide sequence variation have rarely attempted to test hypotheses related to geographically varying patterns of natural selection. The present study tested the role of spatially varying selection in producing significant linkage disequilibrium and large differences in the frequencies of two common alleles at the pantophysin (*Pan* I) locus among five populations of the Atlantic cod, *Gadus morhua*. Nucleotide sequences of 124 *Pan* I alleles showed strong evidence for an unusual mix of balancing and directional selection but no evidence of stable geographically varying selection. The alleles were highly divergent at both the nucleotide level (differing on average by 19 mutations) and at amino acid level (each having experienced three amino acid substitutions since diverging from a common ancestral allele). All six amino acid substitutions occurred in a 56-residue intravesicular loop (IV1 domain) of the vesicle protein and each involved a radical change. An analysis of molecular variation revealed significant heterogeneity in the frequencies of recently derived mutations segregating within both allelic classes, suggesting that two selective sweeps may be presently occurring among populations. The dynamic nature of the *Pan* I polymorphism in *G. morhua* and clear departure from equilibrium conditions invalidate a simple model of spatially varying selection.

S TUDIES examining nucleotide sequence variation in natural populations have provided important insights into the role of natural selection in shaping the patterns of polymorphism within species and the patterns of divergence between species (Hudson 1990; Kreitman 1991; Kreitman and Akashi 1995). When combined with genealogical information, data on the existing levels and distribution of nucleotide sequence variation among populations can provide unparalleled information on the past and present selective forces that may be acting at a locus. Evidence accumulating from Drosophila has suggested that natural selection has played an important role in affecting the patterns of nucleotide variation at a substantial fraction of loci (Moriyama and Powell 1996; Hey 1999). However, the inability to reject the null hypothesis of no selection (*i.e.*, neutrality) is not uncommon (see Schaeffer and Miller 1992; Kliman and Hey 1993) and other factors, most notably the extent of recombination, exert strong effects on the standing levels of nucleotide variation (Begun and Aquadro 1992; Aquadro *et al.* 1994; Charlesworth 1998).

Few studies examining DNA sequence variation have attempted to test hypotheses related specifically to geographically varying patterns of selection. Most species are unlikely to experience similar selection pressures across their geographic ranges, and the extent to which selection can produce local adaptation at the molecular level, particularly in opposition to ongoing gene flow, remains poorly understood. The majority of studies that have examined spatial patterns of selection at the DNA level have focused on loci exhibiting clinal variation (*e.g.*, Berry and Kreitman 1993; Karotam *et al.* 1995; Katz and Harrison 1997; Schulte *et al.* 1997). However, only the detailed molecular dissection of the *Adh* cline in *Drosophila melanogaster* by Berry and Kreitman (1993) explicitly tested the role of selection in maintaining clinal variation in the frequencies of the fast/slow polymorphic site by examining patterns of silent polymorphisms segregating within and between allelic groups.

Unlike the situation for clines, localized selection favoring different alleles in different environments may produce heterogeneous patterns, and loci exhibiting unusually high levels of variation might indicate the possible action of selection (Cavalli-Sforza 1966). One such locus has been identified in the Atlantic cod, *Gadus morhua* (originally called GM798), that unlike other nuclear or mitochondrial markers exhibits significant differentiation among populations at large and small geographic scales (Pogson *et al.* 1995; Fevolden and Pogson 1997; Jonsdottir *et al.* 1999). This locus was also unusual in not showing a relationship between inferred levels of gene flow and geographic distance at large geographic scales (Pogson *et al.* 2001) and in exhibiting nearly complete linkage disequilibrium among three restriction site polymorphisms in the gene

Corresponding author: Grant H. Pogson, Department of Ecology and Evolutionary Biology, Earth and Marine Sciences Bldg., University of California, Santa Cruz, CA 95064. E-mail: pogson@darwin.ucsc.edu

region (Pogson and Fevolden 1998). The gene was originally identified as the cod synaptophysin (*Syp* I) locus by Fevolden and Pogson (1997) but is more likely to represent a recently discovered cellular isoform of synaptophysin called pantophysin (Haass *et al.* 1996). Pantophysin is an integral membrane protein found in microvesicles of both neuroendocrine and nonneuroendocrine tissues that function in a variety of shuttling, secretory, and endocytotic recycling pathways (Haass *et al.* 1996; Windoffer *et al.* 1999). Although the role of pantophysin in these pathways is poorly understood, its highly conserved structure of four transmembrane domains, two intravesicular loops, and two cytoplasmic tails allows all mutations identified at the molecular level to be localized to distinct domains.

The objective of the present study was to determine if geographically varying selection was acting at the pantophysin (*Pan* I) locus of *G. morhua*. To test this hypothesis, 124 *Pan* I alleles (1.85 kb in length) were sequenced from five populations distributed throughout the north Atlantic region. The levels of nucleotide polymorphism and spatial distribution of variable sites segregating within and among *Pan* I allelic classes were then compared among the populations. Three predictions of the variable selection hypothesis were tested. First, to account for the geographically varying selection, differences must exist between the two *Pan* I alleles at the nucleotide and/or amino acid levels. If a prolonged period of selection has favored different *Pan* I alleles in different regions then a genealogical signature of balancing selection may be present and statistical tests should reject neutral expectations. Second, if the selective regime has been stable over time and sufficient gene flow is occurring among populations, no differences should exist in the frequencies of neutral sites segregating within *Pan* I allelic classes because these would be invisible to selection (the "selective equivalence" test of Berry and Kreitman 1993). Third, greater peaks of diversity within the region(s) of the *Pan* I locus experiencing selection should be present between, rather than within, populations (Charlesworth *et al.* 1997). Although the nucleotide sequence data provide strong support for a long-lived polymorphism at the *Pan* I locus, selection on recently derived mutations in both allelic classes appears responsible for the heterogeneity among populations thus invalidating a model of spatially varying selection.

MATERIALS AND METHODS

Samples: Populations of *G. morhua* were sampled throughout the North Atlantic region and random subsamples were taken from these larger groups for sequencing. Subsamples from the NW Atlantic were randomly chosen from two large regional groups, Nova Scotia (NS) and Newfoundland (NF), as described in Pogson *et al.* (2001). Subsamples taken from the Iceland (IC), Balsfjord (BA), and the Barents Sea (BS)
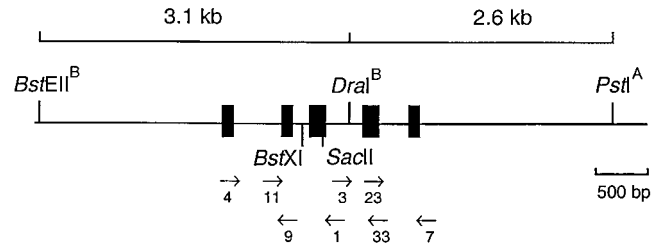

Figure 1.—Restriction map of the *Pan* I gene region showing the locations of polymorphic restriction sites. Exons are represented by solid boxes. The locations of the *Bst*XI and *Sac*II sites used to cut *Pan* I^A^I^B^ heterozygotes prior to PCR and the positions of primers used for sequencing are shown below the coding region.

populations are identical to those described in Pogson *et al.* (1995).

Southern blot analyses: Three restriction site polymorphisms in the vicinity of the *Pan* I locus (*Bst*EII, *Dra*I, and *Pst*I) were scored in 998 individuals on Southern blots as described in Pogson *et al.* (1995). A map of the *Pan* I gene region showing the locations of these polymorphic restriction sites is presented in Figure 1. The *Bst*EII^B, *Dra*I^B, and *Pst*I^A "alleles" refer to the presence of sites for each enzyme and alternate alleles refer to their absence. Frequencies of the three polymorphic restriction sites and the resulting haplotypes are listed in Table 1 and form the basis of the sampling scheme outlined below.

PCR and DNA sequencing: The cDNA clone representing the *Pan* I locus (GM798) was sequenced on an ABI Model 373 automated DNA sequencer. The full-length sequence was obtained from both strands using modified KS (5'-CGAGGTC GACGGTATCGATAAG-3') and SK primers (5'-TCTAGAACT AGTGGATCCCCCG-3') that flanked the *Eco*RI cloning site and two internal sequencing primers (B, 5'-TTGGTCCTCTAT CTGGGCTTCG-3'; G, 5'-GTGCTACTATGCTTGTGGGGC-3'). Two PCR primers were then designed from the clone that amplified a 1.94-kb fragment from genomic DNA (4, 5'-CTTCCATTCATCCGAGTTCTG-3'; 7, 5'-CGTAGCAGAAGA GTGACACAT-3'). PCR reactions were performed in 20 M Tris-HCl (pH 8.8 at 25°), 10 mM KCl, 10 mM (NH$_4$)$_2$SO$_4$, 2.5 mM MgSO$_4$, 0.1% Triton X-100, 100 ng/μl bovine serum albumin, 200 μM each dNTP, 0.25 μM forward and reverse primers, 0.4 units of *Taq* 2000 DNA polymerase (Stratagene, La Jolla, CA), 0.4 units *Taq* extender PCR additive (Stratagene), and 100 ng template DNA in 10-μl sealed glass capillary tubes in an Idaho Technology (Idaho Falls, ID) A1605 air thermal-cycler. After an initial denaturation step of 45 sec at 94° the tubes were exposed to 35 cycles of denaturation at 94° for 1 sec, primer annealing at 52° for 1 sec, and primer extension at 72° for 1 min and 40 sec followed by a hold at 72° for 2 min. PCR products were visualized on 1% agarose gels stained with ethidium bromide.

The 1.94-kb *Pan* I genomic fragment was sequenced from individuals known to be homozygous for the polymorphic *Dra*I restriction site located in the fourth intron of the gene (hereafter called the *Pan* I^A and *Pan* I^B alleles corresponding to the absence or presence of this site, respectively). Consensus restriction maps were then constructed from 4–5 homozygotes for both alleles, and mutations that produced unique restriction sites were identified. The presence of these sites allowed individual alleles to be amplified for sequencing from known *Pan* I^A I^B heterozygotes by digesting genomic DNA with the appropriate restriction enzyme before PCR. To amplify the *Pan* I^A allele, heterozygotes were digested with *Bst*XI (cutting

TABLE 1

Pan I restriction site and haplotype frequencies

	Population				
	Nova Scotia ($n = 412$)	Newfoundland ($n = 245$)	Iceland ($n = 84$)	Barents Sea ($n = 82$)	Balsfjord ($n = 87$)
Restriction sites					
*Bst*EII^A	0.177	0.429	0.298	0.884	0.408
*Bst*EII^B	0.796	0.545	0.690	0.067	0.557
*Dra*I^A	0.899	0.612	0.685	0.073	0.598
*Dra*I^B	0.100	0.386	0.310	0.921	0.402
*Pst*I^A	0.581	0.694	0.381	0.939	0.460
*Pst*I^B	0.410	0.294	0.613	0.055	0.523
Haplotypes					
1. *Bst*EII^A*Dra*I^B*Pst*I^A	0.094	0.373	0.304	0.926	0.411
2. *Bst*EII^B*Dra*I^A*Pst*I^B	0.334	0.237	0.608	0.061	0.544
3. *Bst*EII^B*Dra*I^A*Pst*I^A	0.487	0.320	0.089	0.014	0.044
4. *Bst*EII^A*Dra*I^B*Pst*I^B	0	0.002	0	0	0
5. *Bst*EII^A*Dra*I^A*Pst*I^B	0.081	0.057	0	0	0
6. *Bst*EII^A*Dra*I^A*Pst*I^A	0.004	0.011	0	0	0

the *Pan* I^B allele at nucleotide position 646) prior to PCR. To amplify the *Pan* I^B allele, digestions were performed with *Sac*II (cutting the *Pan* I^A allele at position 909) prior to PCR. Thirty-five cycles of PCR using *Bst*XI-digested DNA as template and the two flanking PCR primers (4 and 7) resulted in the amplification of the *Pan* I^A allele whereas *Sac*II-digested DNA allowed preferential amplification of the *Pan* I^B allele. To test the veracity of this method *Pan* I^A and *Pan* I^B alleles were amplified and sequenced in duplicate from two heterozygotes at two different dates. No differences among replicate sequences were detected.

Templates for sequencing were gel purified from 0.4% agarose gels and spun through spin columns containing 0.8 ml of Sephadex G-50. Complete sequences of both DNA strands were obtained from eight sequencing reactions per template. In addition to the two flanking primers, three additional forward (11, 5′-GCTGGATTTCCCGATGTTGATA-3′; 3, 5′-CGTT GGTCCTCTATCTGGGCTTC-3′; 23, 5′-GTTTCTCTGCAAGG ATCTGTTTG-3′) and reverse primers were used in sequencing (33, 5′-TCACAAATAGATCCTTGCAGAG-3′; 1, 5′-CGAAGAGT GGTTGCCAATAAGG-3′; 9, 5′-GCTGCATCAACCTAAAGTAG GAG-3′). Sequences were edited with SequenceNavigator, compiled into consensus sequences using AutoAssembler (both programs from Applied Biosystems, Foster City, CA), and aligned by eye. Nucleotide sequences have been deposited in GenBank under accession nos. AF288943–AF288977.

Statistical analyses: Samples of *Pan* I^A and *Pan* I^B alleles were obtained from five populations of *G. morhua* by randomly selecting 12 or 13 *Pan* I^{A}I^B heterozygotes previously identified from Southerns. For the *Pan* I^B alleles this involved sampling only one haplotype (numbered 1 in Table 1). Because *Pan* I^A alleles were distributed among three haplotypes (numbered 2, 3, and 5 in Table 1) samples of this allele were randomly selected from each population to ensure accurate representation of these haplotypes. Although this sampling protocol allows for statistical tests among *Pan* I allelic classes it is inappropriate for tests of neutrality that assume a random sampling of alleles. It also may not provide accurate estimates of nucleotide polymorphism in different populations because allele frequencies are extremely variable. To allow for comparisons of nucleotide variability among populations and to perform

Tajima's (1989) and Fu and Li's (1993) tests of neutrality, I followed the approach of Hudson *et al.* (1994) and assembled 50 constructed random samples (CRSs) from each population. These were made by randomly subsampling *Pan* I^A and *Pan* I^B alleles from each population in proportion to their known frequencies. The CRS sizes were identical to the total number of *Pan* I alleles sequenced per population (24 or 26). The CRSs generated from each population were pooled to create 50 "global" constructed random samples that would have been representative of sampling across the entire species range.

Heterogeneity of *Pan* I allele and haplotype frequencies among populations were tested using F_{ST} estimates obtained from BIOSYS-1 (Swofford and Selander 1989). An analysis of molecular variation (AMOVA) was used to test for differences in the patterns of nucleotide polymorphism segregating within the *Pan* I^A and *Pan* I^B allelic classes between populations (Excoffier *et al.* 1992). Phi$_{ST}$ statistics were estimated from *p*-distances among *Pan* I^A and *Pan* I^B haplotypes obtained from MEGA ver. 1.01 (Kumar *et al.* 1993) and were tested for significance by performing 5000 permutations of the null distributions of each variance component. Composite measures of linkage disequilibrium among restriction sites in the *Pan* I gene region were obtained using the LD86.FOR program of Weir (1990). *D* values were tested for significance by chi-square tests and standardized to D' values to allow comparisons among populations differing in allele frequencies. Mantel tests were done using the ISOLDE subprogram of GENEPOP ver. 1.2 (Raymond and Rousset 1995).

Phylogenetic analyses of *Pan* I alleles were performed using the neighbor-joining algorithm of Saitou and Nei (1987) implemented by MEGA and by maximum parsimony using PAUP ver. 3.1 (Swofford 1993). Both trees were rooted using the Greenland cod, *Gadus ogac*, as an outgroup. Estimates of nucleotide polymorphism (both π and θ) present within *Pan* I alleles and in the CRSs were obtained using DnaSP ver. 2.2 (Rozas and Rozas 1997). The DnaSP program was also used to perform Tajima's (1989) and Fu and Li's (1993) tests of neutrality (the latter using *G. ogac* as the outgroup sequence). An intraspecific McDonald and Kreitman (1991) test also was performed using DnaSP treating the two *Pan* I alleles as independent evolutionary lineages.
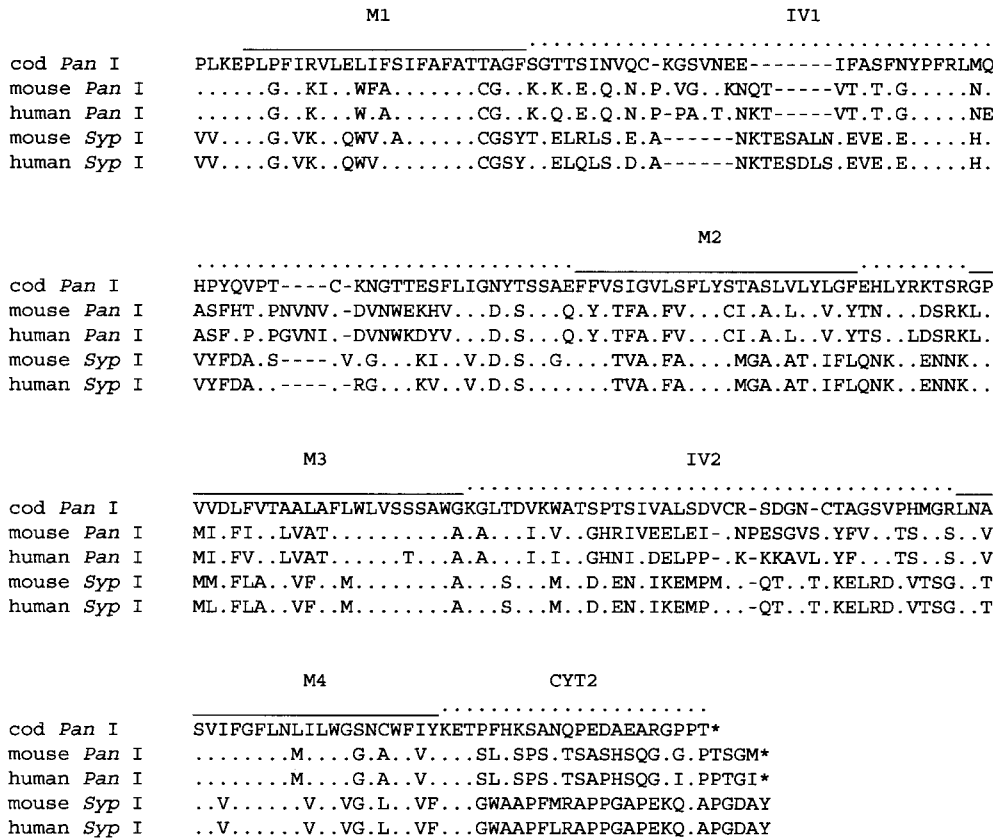
```
              M1                                        IV1
      _____ ...........................................
cod Pan I    PLKEPLPFIRVLELIFSIFAFATTAGFSGTTSINVQC-KGSVNEE-------IFASFNYPFRLMQ
mouse Pan I  ......G..KI..WFA.......CG..K.K.E.Q.N.P.VG..KNQT-----VT.T.G.....N.
human Pan I  ......G..K...W.A.......CG..K.Q.E.Q.N.P-PA.T.NKT-----VT.T.G.....NE
mouse Syp I  VV....G.VK..QWV.A......CGSYT.ELRLS.E.A------NKTESALN.EVE.E.....H.
human Syp I  VV....G.VK..QWV........CGSY..ELQLS.D.A------NKTESDLS.EVE.E.....H.
```

```
              ................................ _____ .........__
                                                          M2
cod Pan I    HPYQVPT----C-KNGTTESFLIGNYTSSAEFFVSIGVLSFLYSTASLVLYLGFEHLYRKTSRGP
mouse Pan I  ASFHT.PNVNV.-DVNWEKHV...D.S...Q.Y.TFA.FV...CI.A.L..V.YTN...DSRKL.
human Pan I  ASF.P.PGVNI.-DVNWKDYV...D.S...Q.Y.TFA.FV...CI.A.L..V.YTS..LDSRKL.
mouse Syp I  VYFDA.S-----.V.G...KI..V.D.S..G....TVA.FA....MGA.AT.IFLQNK..ENNK..
human Syp I  VYFDA..-----.-RG...KV..V.D.S.......TVA.FA....MGA.AT.IFLQNK..ENNK..
```

```
              M3                              IV2
      _____ ...........................................___
cod Pan I    VVDLFVTAALAFLWLVSSSAWGKGLTDVKWATSPTSIVALSDVCR-SDGN-CTAGSVPHMGRLNA
mouse Pan I  MI.FI..LVAT..........A.A...I.V..GHRIVEELEI-.NPESGVS.YFV..TS..S..V
human Pan I  MI.FV..LVAT......T...A.A...I.I..GHNI.DELPP-.K-KKAVL.YF...TS..S..V
mouse Syp I  MM.FLA..VF..M........A...S...M..D.EN.IKEMPM..-QT...T.KELRD.VTSG..T
human Syp I  ML.FLA..VF..M........A...S...M..D.EN.IKEMP...-QT..T.KELRD.VTSG..T
```

```
              M4                    CYT2
      _____ .....................
cod Pan I    SVIFGFLNLILWGSNCWFIYKETPFHKSANQPEDAEARGPPT*
mouse Pan I  ........M....G.A..V....SL.SPS.TSASHSQG.G.PTSGM*
human Pan I  ........M....G.A..V....SL.SPS.TSAPHSQG.I.PPTGI*
mouse Syp I  ..V......V..VG.L..VF...GWAAPFMRAPPGAPEKQ.APGDAY
human Syp I  ..V......V..VG.L..VF...GWAAPFLRAPPGAPEKQ.APGDAY
```

Figure 2.—Deduced amino acid structure of cod pantophysin (*Pan* I) aligned with mouse and human pantophysin and synaptophysin (*Syp* I) sequences. The complete 3′ cytoplasmic tails of the mouse and human synaptophysins are not presented. Solid lines indicate positions of the four membrane-spanning domains (labeled M1–M4) and the dotted lines show the positions of the two intravesicular loops (IV1 and IV2) and the cytoplasmic tail (CYT2).

RESULTS

Amino acid sequence and structure of cod pantophysin: cDNA clone GM798 had an open reading frame of 222 amino acids and a 186-bp translated but untranscribed 3′ tail. The gene was originally identified as the cod synaptophysin (*Syp* I) locus (Fevolden and Pogson 1997) but may represent the cod homologue of a recently discovered *Syp* I isoform called pantophysin

(Haass *et al.* 1996). Both physins belong to a growing family of integral membrane proteins found in synaptic or cytoplasmic vesicles that are characterized by four membrane-spanning domains, two intravesicular loops, and two cytoplasmic tails (Fernandez-Chacon and Südhof 1999). Figure 2 presents the deduced amino acid structure of the cod physin aligned with pantophysin sequences from mouse and human (both from Haass *et al.* 1996) and the closely related synaptophysin

TABLE 2

Linkage disequilibrium in the *Pan* I gene region

| Restriction sites | Statistic | Population | | | | |
		Nova Scotia (*n* = 412)	Newfoundland (*n* = 245)	Iceland (*n* = 84)	Barents Sea (*n* = 82)	Balsfjord (*n* = 87)
*Pst*I^A-*Bst*EII^B	D	0.0096	0.0680***	0.1821***	0.0479***	0.2066***
	SE	0.0061	0.0121	0.0144	0.0161	0.0111
	D'	0.129	0.524	0.989	0.935	0.919
*Pst*I^A-*Dra*I^B	D	−0.0406***	−0.1105***	−0.1787***	−0.0476***	−0.2064***
	SE	0.0040	0.0076	0.0146	0.0160	0.0108
	D'	0.991	0.973	0.933	0.934	0.970
*Bst*EII^B-*Dra*I^B	D	−0.0924***	−0.2021***	−0.2010***	−0.0612***	−0.2307***
	SE	0.0082	0.0070	0.0140	0.0169	0.0087
	D'	0.989	0.939	0.978	0.983	0.998

*** $P < 0.001$.

```
                  I2                E3  I3          E4                  I4            E5          I5        E6
 ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━■■━━━━━━■■■■■■■■━━━━━━━━━━━━━━━━━■■━━━━━━━━━■■━━━
                                                1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
                1 1 1 1 2 2 2 2 2 2 3 4 5 6 6 6 7 7 7 7 7 7 7 8 8 9 9 9 0 0 1 1 1 1 1 1 2 2 3 4 5 5 5 6 6 6 6 6 7 8
        2 4 8 0 3 6 9 3 4 5 5 6 6 3 3 4 7 1 4 9 1 4 4 5 5 7 9 3 3 0 2 6 3 6 3 4 5 5 7 8 1 2 0 0 5 8 9 2 2 2 5 5 2 3
Allele  1 9 6 2 8 8 6 6 4 2 4 5 9 1 2 1 0 6 6 9 4 5 7 4 5 5 9 7 8 9 9 8 7 2 9 0 0 3 3 6 0 0 5 7 6 0 5 0 6 9 0 4 0 0

NF11-A  A T G T T T - - A A - T T A C T A T C G C C A A C G A G T C C T A A T G C T C G C G G A A C G T G C G G T G
NF6-A   G . . . . . . . . . . . . T . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .
NF24-A  . . . . . . . . . . . . . . . . . . . . . . . . . A . . . . . . . . . . . . . . . . . T . . . . . . . . . .
NF36-A  . . . . C . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .
NS1-A   . A . . . . . . . . . . . . . . . . . . . . . . . A . . . . . . . . . . . . . . . . . . . . . . . . . . . .
NS12-A  . . . . . . . . . . A . . . . . . . . . . . . . . A . . . . . . . . . . . . . . . . . . . . . . C . . . . .
NS28-A  . . . . . . . . . . A . . . . . . . . . . . . . . A . . . . . . . . . . . . . . G . . . . . . . . . . . . .
NS58-A  . . . . . . . . . . . . . . . . . . . . . . . . . A . . . . . . A . . . . . . . . . . . . . . . . . . . . .
NS73-A  . . . . . . . . . . A . . . . . . . . . . . . . . A . . . . . . . . . . . . . . . . . . . . . . . . . . . .
NS74-A  . . . . . . . . . . . . . . . . . . . . . . . . . A . . . . . . . . . . . . . . . . . . . . . . . . . . . .
IC2-A   . . . . . . . . . . . . . . . . . . . . . . . . . A . . . . . . . . . . . . . . . . . . . . . A . . . . . .
IC41-A  . . . . . C . . . . . . . . . . . . . . . . . . . . . . . . . . A . . . . . A . . . . . . . . . . . . . . .
IC70-A  . . . . . . . . . . . . . . . . . . . . . . . . . . . T . . . . A C . . . . . . . . . . . . . . . . . C . .
IC80-A  . . . . . . . . . . . . . . . . . . . . . . . . . . . . A . . . . . . . . . . . . . . . . . . . . . . . . .
BA105-A . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . T . . . . . .
BA112-A . . . . . . . . . . . . T . . . . . . . . . . . . . . T . . . . . . . . . C . . . . . . . . . . . . . . . .
BA126-A . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . C . . . . . . . . . . . . . . . .
BA128-A . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . C . . . . . . . . . . . . . . . . T . . . . . .
BA138-A . . . . . . . . . . . . . . . . . . C . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .
BA140-A . . . . . . ∇1 . G . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .
BA143-A . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . T . . . C . . . . . . . . . . . . . . . .
BA149-A . . . . . . . . . G . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .
BS29-A  . . . . . . . . . . . . . . . . . . . . . . . . . . . T . . . . . . . . . C . . . . . . . . . . . . . C . .
BS31-A  . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . T
BS64-A  . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . T . . . . . . .
NF11-B  . . A . . . . ∇2 . T ∇3 G . . . G T . A . . A T G A . T . . T . . . C . A . G . . . . . G . T . . . . T . . .
NF17-B  . . A . . . . . . T ∇3 G . . . G T . A A . A T G A . T . . T . . . C . A . G . . . . . G . . . . . . . . . .
NS58-B  . . A . . . . . . T ∇3 G . . . G T . A . . A T G A . T . . T . . . C . A . G . . . . . G . . . . . . . . . .
NS70-B  . . A . . . . ∇2 . T ∇3 G . . . G T . A . . A T G A . T . . T . . . C . A . G . . . . . G . T . . . T A . . .
NS79-B  . . A . . . . . . T ∇3 G . . . G T . A . . A T G A . T . . T . . . C . A T G . . . . . G . . . . . . . . . .
IC74-B  . . A C . . . ∇2 . T ∇3 G . . . G T . A . . A T G A . T . . T . . . C . A . G . . T . . G . T . . . T . . . .
BA112-B . . A . . . . ∇2 . T ∇4 G . . . G T . A . . A T G A . T . . T . . . C . A . G . . . . . G . T . . . T . . . .
BS29-B  . . A . . . . ∇2 . T ∇3 G . . . G T . A . . A T G A . T A A T . . . C . A . G . . . . . G . T . . . T . . . .
BS72-B  . . A . . . . ∇2 . T ∇3 G . . . G T . A . . A T A T G A . T . . T . . . C . A . G . . . . . G . T . . . T . .
G. ogac A T G T T T - - A A - G T A C T A T C G C A A A A G T G T C C T A A T A C G C G - - G A A C G T G C G G T G
```

FIGURE 3.—Nucleotide polymorphism at the *Pan* I locus of *G. morhua*. Sequences are presented for the 34 unique *G. morhua* haplotypes and for the outgroup, *G. ogac*. Base positions of each variable site and the locations of introns (labeled I2–I5) and exons (labeled E3–E6) are indicated at the top of the figure. Four insertion/deletion mutations identified in the second intron are also shown: ∇1, a single CA insertion; ∇2, a 12-bp deletion of GCATAGTAAAAA; ∇3, a 6-bp insertion of TGTTTT; ∇4, a 6-bp insertion of TTTTTT. Amino acid replacement mutations have been underlined. BF, Balsfjord; BS, Barents Sea; IC, Iceland; NF, Newfoundland; NS, Nova Scotia.

sequences from both species (from SÜDHOF *et al.* 1987 and GAITANOU *et al.* 1997, respectively). Excluding the 27 amino acids missing from the amino terminus of the cod clone, amino acid identities between the *G. morhua* protein and pantophysin from mouse (49.8%) and human (50.5%) are only marginally higher than between the synaptophysins from both species (46.1 and 48.1%, respectively). Identities are highest in the four membrane-spanning domains (labeled M1–M4) and the charged residues that flank these regions as noted in previous studies (JOHNSTON *et al.* 1989; COWAN *et al.* 1990). The two intravesicular loops of the protein (called IV1 and IV2) and the short 3′ cytoplasmic tail were highly diverged and difficult to align. Although the *G. morhua* physin is almost equally related to the two mammalian proteins, it is more likely to be pantophysin on the basis of (i) its truncated carboxy terminus (22 amino acids in length), which is lacking the characteristic proline- and tyrosine-repeating motifs present in all synaptophysins characterized to date and (ii) its isolation from liver tissue where synaptophysin expression is expected to be absent (except in nerve fibers).

Linkage disequilibrium in *Pan* I gene region: Table 2 presents estimates of linkage disequilibrium between three polymorphic restriction sites that span a 5.7-kb region of the *Pan* I gene region (see Figure 1). Highly significant disequilibrium was detected between all pairs of sites in all populations with the exception of the flanking *Bst*EII[B] and *Pst*I[A] sites in Nova Scotia. This strong disequilibrium resulted in two common haplo-

TABLE 3

Nucleotide polymorphism in *Pan* I^A and *Pan* I^B alleles from different populations

Population	*Pan* I^A alleles					*Pan* I^B alleles				
	n	S	k	π	θ	n	S	k	π	θ
Nova Scotia	13	7	1.67	0.00090	0.00122	13	6	1.64	0.00089	0.00105
Newfoundland	13	6	1.69	0.00091	0.00104	13	4	1.56	0.00085	0.00070
Iceland	12	9	2.18	0.00118	0.00161	12	2	0.33	0.00018	0.00036
Balsfjord	12	9	2.02	0.00109	0.00161	12	1	0.30	0.00016	0.00018
Barents Sea	12	10	2.71	0.00147	0.00179	12	3	0.50	0.00027	0.00054
Pooled	62	26	2.26	0.00122	0.00299	62	11	1.02	0.00055	0.00127

S is the number of segregating sites, k is the average number of nucleotide differences, π is nucleotide diversity, and θ is theta per site.

types to predominate in most populations (numbered 1 and 2 in Table 1). *D* values were consistent in sign across all populations and the standardized coefficients approached their maximum theoretical limits everywhere except the two NW Atlantic populations. In Nova Scotia, this was caused by the high frequency of one haplotype (numbered 3) formed by a recombination event between the *Dra*I^A and *Pst*I^A sites that had the effect of uncoupling the two flanking sites. In Newfoundland, this recombinant haplotype was less frequent and the disequilibrium between the two flanking restriction sites was diminished but still significant.

Nucleotide polymorphism: A total of 62 *Pan* I^A and 62 *Pan* I^B alleles were sequenced from five different populations of *G. morhua*. The gene region sequenced contained four exons (208 amino acids) and four introns whose locations were identical to those described in mammalian pantophysin and synaptophysin genes by HAASS *et al.* (1996). Polymorphic sites are presented in Figure 3 (along with sequence from the outgroup *G. ogac*) and the levels of nucleotide polymorphism present within the *Pan* I^A and *Pan* I^B alleles in different populations are summarized in Table 3. All *Pan* I^A alleles were 1851 bp in length with the exception of one allele (BA140-A), which contained three rather than two copies of a CA repeat. The *Pan* I^B alleles were either 1845 or 1857 bp in length depending on the presence or

absence of a 12-bp deletion at position 236 in the second intron.

A total of 52 polymorphic nucleotide sites were identified in the total sample. Twenty-six segregating sites (and one insertion) were detected in the sample of *Pan* I^A alleles distributed among 25 haplotypes. In contrast, only 11 segregating sites (and one deletion) were found in the sample *Pan* I^B alleles represented among 9 haplotypes. In the pooled sample the *Pan* I^A alleles exhibited levels of nucleotide diversity (π) and θ that were more than twice that observed for the *Pan* I^B alleles (Table 3). Levels of nucleotide polymorphism varied considerably among populations from the NW and NE Atlantic. For the *Pan* I^A alleles variability was lowest in Nova Scotia and Newfoundland and highest in the three NE Atlantic populations. The *Pan* I^B alleles exhibited extremely low levels of polymorphism in the NE Atlantic but approached the levels of variability shown by the *Pan* I^A alleles in the NW Atlantic. A negative relationship was seen between the levels of nucleotide diversity and the population frequency of both *Pan* I^A alleles ($r = -0.873$, $P = 0.053$) and *Pan* I^B alleles ($r = -0.511$, $P = 0.379$) but neither correlation was significant.

In contrast to the minimal variation present within the *Pan* I^A and *Pan* I^B allelic groups, 15 nucleotide mutations and a 6-bp insertion were fixed between the two alleles (Figure 3). The average number of nucleotide

TABLE 4

Nucleotide polymorphism in the 50 constructed random samples

Population	No. of alleles						
	n	*Pan* I^A	*Pan* I^B	S	k	π	θ
Nova Scotia	26	24	2	22.9	3.96	0.00215	0.00326
Newfoundland	26	16	10	25.2	9.92	0.00540	0.00359
Iceland	24	16	8	26.7	9.85	0.00536	0.00389
Balsfjord	24	14	10	23.9	10.08	0.00548	0.00348
Barents Sea	24	2	22	23.4	3.53	0.00192	0.00341
Pooled	124	72	52	41.3	10.15	0.00552	0.00417

TABLE 5

Amino acid replacement mutations

Allele	Codon position	Nucleotide change	Amino acid change	Location	Classification	Distribution in sample
Pan I^A	61	AAA → CAA	K → Q	IV1	Radical	Fixed
	64	AAC → ACC	N → T	IV1	Radical	Fixed
	71	GAG → AAG	E → K	IV1	Very radical	Polymorphic
	79	TCT → ACT	S → T	IV1	Radical	Fixed
	214	GAC → TAC	D → Y	CYT2	Radical	Polymorphic
Pan I^B	43	GAG → GTG	E → V	IV1	Radical	Fixed
	61	AAA → AAT	K → N	IV1	Radical	Fixed
	64	AAC → GAC	N → D	IV1	Radical	Fixed
	92	TCC → ACC	S → T	M2	Radical	Polymorphic

Classification of amino acid changes is based on TAYLOR (1986).

IV1, first intravesicular domain; CYT2, carboxy-terminus cytoplasmic domain; M2, second transmembrane domain.

differences between any two randomly sampled *Pan* I^A and *Pan* I^B alleles (19.0) far exceeded that found within either allelic group (2.3 and 1.0, respectively). Because the majority of the variation was present between rather than within allelic classes, nucleotide diversity levels were strongly affected by the differences in allele frequencies among populations shown in Table 1. Estimates of nucleotide polymorphism in the five populations are presented in Table 4 from 50 constructed random samples that reflected *a priori* known differences in *Pan* I allele frequencies. Nucleotide diversity was highest in the three populations with intermediate frequencies of both alleles (Newfoundland, Iceland, and Balsfjord) and fell sharply in populations with high frequencies of either *Pan* I^A (Nova Scotia) or *Pan* I^B (Barents Sea). In contrast, θ was relatively invariant among populations because the number of segregating sites was largely determined by the presence of both alleles.

One-quarter of the polymorphisms detected in the study (13) fell within coding DNA and nine involved amino acid replacements (Table 5). Six of the nine replacement mutations were fixed between the two *Pan* I alleles (three within each allelic lineage) and all occurred within the first intravesicular (IV1) domain of the protein. Two codon positions (61 and 64) had each experienced two mutations so that at the protein level the two *Pan* I alleles differed by four amino acids. Based on the classification scheme of TAYLOR (1986) all nine amino acid replacement mutations were radical changes (six involving charged residues). Three replacement mutations were also detected segregating within *Pan* I^A and *Pan* I^B allelic groups. Two were singletons found in Norwegian waters (positions 92 and 214). However, the third mutation involved a very radical change (aspartic acid to lysine) in the IV1 domain of the protein and was detected in 22 of the 62 *Pan* I^A alleles sequenced. This mutation (hereafter the *Pan* I$^{A'}$ allele) was fixed in

the 19 haplotypes previously identified as recombinants between the *Dra*I^A and *Pst*I^A sites (haplotype 3 in Table 1) that were chosen for sequencing. It was also found in 3 of the 5 nonrecombinant *Pan* I^A haplotypes (haplotype 2 in Table 1) sampled from Nova Scotia but not in the same haplotype sampled from any other population.

The distribution of polymorphism across the *Pan* I gene region was examined by the sliding window approach of KREITMAN and HUDSON (1991). Nucleotide diversity exhibited little heterogeneity across the pantophysin gene region when the *Pan* I^A and *Pan* I^B alleles were analyzed separately (Figure 4). However, when both alleles were included in the analysis two peaks of polymorphism were identified. The first peak corresponded to a 30-bp region in the second intron (positions 236 to 265) that was capable of forming a stem-loop structure in *Pan* I^A alleles but had been disrupted by two insertion/deletion events in *Pan* I^B alleles (see Figure 3). The second peak of polymorphism occurred in the region of the IV1 domain of the protein in the fourth exon that was segregating for six amino acid replacement mutations (positions 745 to 799 in Figure 3). When only silent positions were included in the sliding window analysis, the latter peak of polymorphism disappeared (not shown).

Phylogenetic analyses: Genealogies of *Pan* I alleles were reconstructed by maximum parsimony and neighbor-joining approaches. A total of 32 parsimony-informative sites that produced a single most parsimonious tree of 74 steps with a consistency index of 0.987 were identified. The parsimony and neighbor-joining (NJ) trees were identical except for the position of a small subclade of *Pan* I^A alleles (not shown) and the NJ tree is presented in Figure 5. The *Pan* I^A and *Pan* I^B alleles formed two highly distinct clades of closely related sequences each having 100% bootstrap support. The *Pan* I^B clade was dominated by a group of 52 alleles that exhibited extremely low variability and 10 additional
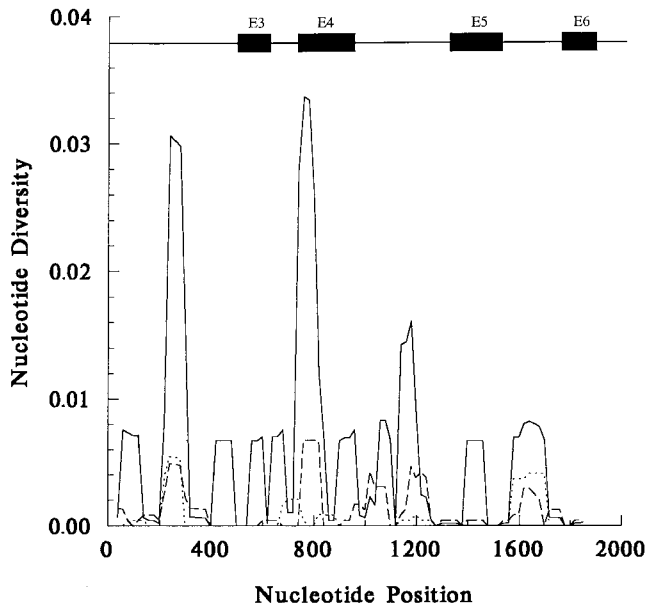
FIGURE 4.—Sliding window analysis of nucleotide polymorphism across the *Pan* I gene region. Insertion/deletions in the second intron have been included as single mutational events. Analyses are presented for the 62 *Pan* I^A alleles (dashed line), the 62 *Pan* I^B alleles (dotted line), and the complete data set (solid line). The window size was 75 bp and the step size was 20 bp. The positions of introns and exons (labeled E3–E6) are shown at the top of the figure.

alleles that were restricted to the NW Atlantic region. The former group (hereafter called $\nabla 2$ *Pan* I^B alleles) was characterized by a 12-bp deletion in the second intron (position 236 in Figure 3) and two mutations in the fifth intron (positions 1580 and 1650 in Figure 3). The ancestral subclade of 10 *Pan* I^B alleles from the NW Atlantic were identical to all *Pan* I^A alleles at these two positions. The clade of *Pan* I^A alleles was considerably more variable and possessed several subclades that exhibited limited geographic distribution. The most widely distributed subgroup was represented by the *Pan* I$^{A'}$ alleles characterized by the aspartic acid to lysine mutation in the IV1 domain. This mutation occurred at high frequencies in the NW Atlantic (0.687 in Nova Scotia and 0.320 in Newfoundland) but was rare in the NE Atlantic. Figure 5 also shows that the *Pan* I^B alleles have experienced a faster rate of evolution than the *Pan* I^A alleles. The genealogy underestimates the changes that have occurred in the lineage of *Pan* I^B alleles because it does not include the insertions/deletions shown in Figure 3.

Differentiation among populations: The frequencies of the three restriction sites scored in the vicinity of the

Pan I locus exhibit highly significant differences among populations of *G. morhua* (POGSON *et al.* 1995). F_{ST} values estimated for the *Bst*EII, *Dra*I, and *Pst*I site polymorphisms in the five populations included in the present study are 0.229, 0.300, and 0.157, respectively. If *Pan* I haplotypes are considered instead of individual restriction sites F_{ST} is 0.229. All F_{ST} values are highly significant ($P < 0.001$). To examine whether heterogeneity existed among *Pan* I^A and *Pan* I^B allelic classes from different populations, an AMOVA was performed using *p*-distances estimated among haplotypes. Table 6 shows that significant differentiation was observed among populations for the variable sites identified within *Pan* I^A and *Pan* I^B allelic classes even though Phi$_{ST}$ was low for both groups (0.119 and 0.152, respectively). This heterogeneity was caused by differences in the frequencies of the *Pan* I$^{A'}$ and $\nabla 2$ *Pan* I^B alleles described in the previous section.

Unlike the majority of nuclear restriction fragment length polymorphism (RFLP) loci examined in *G. morhua*, the individual restriction site polymorphisms scored in the *Pan* I gene region do not exhibit relationships between gene flow and geographic distance over the North Atlantic region (POGSON *et al.* 2001). The slope of the regression of log (gene flow) *vs.* log (geographic distance) for the *Pan* I locus was positive, suggesting that populations sampled at greater geographic distances are genetically more similar than populations sampled at shorter geographic distances. However, this conclusion derived from analyses performed on the frequencies of single restriction sites among populations, not the relatedness of the alleles themselves. To examine whether allelic similarity was related to geographic distance, the average number of nucleotide substitutions per site (d_{XY}) between *Pan* I^A and *Pan* I^B alleles sampled from different populations was regressed against the distance separating the populations. Strong positive correlations between (d_{XY}) and distance were present for both *Pan* I^A ($r = 0.629$, $P = 0.047$) and *Pan* I^B alleles ($r = 0.573$; $P = 0.066$) although Mantel tests indicated that the relationship was significant only for the former. The positive relationships observed between allelic similarity and geographic distance for both *Pan* I alleles considered individually contrasts with the patterns exhibited by their population frequencies.

Tests of neutrality: Results of Tajima's and Fu and Li's tests for neutrality on the 50 constructed random samples are presented in Table 7. Tajima's *D* statistic was negative in Nova Scotia and the Barents Sea (indicating an excess of low-frequency sites) but was not significant in 100 individual tests. Positive values of Tajima's *D*

FIGURE 5.—Neighbor-joining tree of 124 *Pan* I alleles. Numbers indicate the percentages of 100 bootstrap replicates supporting a specific clade. Bootstrap values below 60% are not shown. Clades corresponding to the *Pan* I$^{A'}$ and $\nabla 2$ *Pan* I^B alleles are marked. BF, Balsfjord; BS, Barents Sea; IC, Iceland; NF, Newfoundland; NS, Nova Scotia.
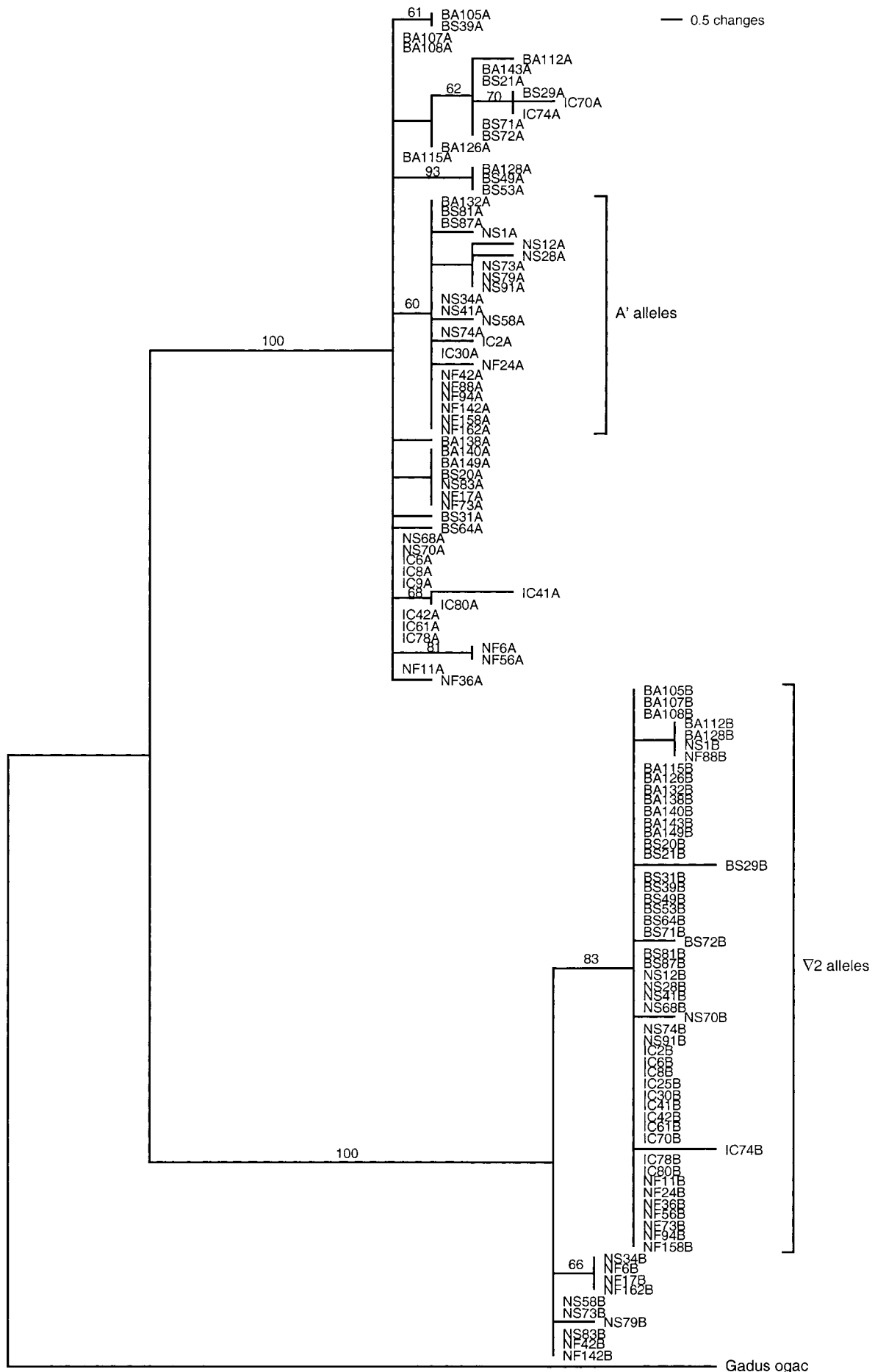
TABLE 6

AMOVA results

Allele	Variance component	Observed variance[a]	Phi$_{ST}$	P
Pan I^A	Among populations	0.07420	0.119	0.002
	Within populations	0.54812		
Pan I^B	Among populations	0.04264	0.152	0.012
	Within populations	0.23874		

[a] Values $\times 10^3$.

were found in the other populations but only Balsfjord produced a substantial number of significant test statistics (all 50 tests yielding $P < 0.10$ of which 32 were less than 0.05). Highly variable results were also observed for Fu and Li's D and F statistics. Some populations produced significant values for D but not F (Nova Scotia) and for F but not D (Balsfjord). The Iceland population produced a moderate number of positive tests for both statistics. Surprisingly, no significant test results were found in the 50 constructed random samples pooled from all five populations despite the fact that these samples were five times larger than the single population CRSs. Although the statistical meaning of these tests is unclear, the negative results obtained for the pooled CRS of 124 alleles was unexpected given the strong signal of selection in the data. An intraspecific McDonald and Kreitman test did, however, produce a significant result (Table 8) due to the proportion of fixed replacement differences between Pan I alleles (66.7%) being much higher than that of fixed silent differences (21.4%).

DISCUSSION

Spatial patterns of variation have commonly been used to identify genetic loci responding to some form of natural selection. In the present study, the pantophysin (Pan I) locus of $G.\ morhua$ was chosen for molecular characterization because of its exceptionally high differentiation among populations (POGSON $et\ al.$ 1995) and the highly significant linkage disequilibrium present among three restriction site polymorphisms spanning the gene region (Table 2). Examination of nucleotide sequences of 124 Pan I alleles sampled from five populations of $G.\ morhua$ has provided compelling evidence that this locus is indeed experiencing strong natural selection. The Pan I^A and Pan I^B alleles are highly differentiated at both the nucleotide level (differing on average by 19 mutations) and the protein level (each having undergone three amino acid substitutions since diverging from a common ancestral allele). All 6 replacement mutations cluster in the 56-amino-acid IV1 domain of the protein and each involves a radical substitution (Table 5). Recently derived mutations are also detected segregating with Pan I^A and Pan I^B allelic groups that exhibit significant heterogeneity among populations. Both mutations occur in regions of the gene exhibiting peaks of divergence between alleles, suggesting that historical and contemporary forms of selection acting at this locus are equivalent. The linkage disequilibrium and heterogeneity among populations thus do not appear to result from stable spatially varying selection but from the recent appearance and spread of selectively favored mutations in both allelic groups in different geographic areas.

There are two explanations for the large number of fixed differences detected between the two common Pan I alleles. One possibility is that the Pan I locus has experienced a prolonged period of balancing selection during which time recombination has played a minimal role in confounding the evolutionary histories of the two alleles. The other explanation is that the two alleles have spent most of their evolutionary histories in geographical isolation and have only recently been mixed together in extant populations. This "historical isolation" hypothesis can account for (i) the high divergence between alleles ($i.e.$, strong directional selection favoring different mutations in different regions) and (ii) the strong linkage disequilibrium in the Pan I gene region ($i.e.$, recombination has yet to break apart the historical associations). Although intuitively appealing, the historical isolation hypothesis makes two predictions

TABLE 7

Results of tests of neutrality on the 50 constructed random samples

	Tajima's test		Fu and Li's tests			
Population	Mean D	Proportion significant	Mean D	Proportion significant	Mean F	Proportion significant
Nova Scotia	−1.233	0/50	1.153	14/50	0.420	0/50
Newfoundland	1.864	13/50	1.032	7/50	1.627	22/50
Iceland	1.417	1/50	1.271	18/50	1.668	25/50
Barents Sea	−1.614	0/50	0.770	3/50	−0.018	0/50
Balsfjord	2.148	32/50	1.173	9/50	1.820	27/50
Pooled	1.005	0/50	−0.088	0/50	0.531	0/50

TABLE 8

McDonald and Kreitman test

	Replacement	Silent
Fixed between alleles	6	9
Polymorphic within alleles	3	33

G test with William's correction = 6.23, *P* < 0.0125

TABLE 9

Linkage disequilibrium in the GM842 and GM727 gene regions

Locus	Restriction site pair	*n*	*D*	*D'*
GM727	*Dra*I-*Bst*EII	998	−0.0028	0.072
GM842	*Pvu*II-*Dra*I	998	−0.0056	0.129
	*Pvu*II-*Sac*I	998	−0.0046	0.107
	*Dra*I-*Sac*I	998	−0.0064	0.081

that are not supported by the available data. First, it predicts that linkage disequilibrium should be common throughout the genome of *G. morhua* because all loci would have experienced similar histories of isolation. This prediction can be tested by examining linkage disequilibrium in the vicinity of two other nuclear loci (GM727 and GM842) scored for multiple restriction site polymorphisms by POGSON *et al.* (1995). Table 9 shows that no detectable linkage disequilibrium exists at either RFLP locus. The historical hypothesis also predicts that RFLP alleles at loci other than *Pan* I should exhibit differentiation at the nucleotide level (albeit at reduced levels) even if recombination has broken apart associations assessed at larger distances. Preliminary sequence data collected for a 960-bp region of the *G. morhua* GM842 locus has found no differences among RFLP alleles (G. H. POGSON, unpublished data). These observations fail to support the historical isolation hypothesis and suggest that the *Pan* I locus has experienced a very different evolutionary history from other genes.

Evidence that natural selection can act at the *Pan* I locus while both alleles coexist in the same population is provided by the distributions of recently derived mutations segregating within both allelic classes. These distributions suggest that two selective sweeps may be occurring among populations of *G. morhua*: the eastward movement of the *Pan* I$^{A'}$ allele and the westward spread of the ▽2 *Pan* I^B allele. The *Pan* I$^{A'}$ allele (having an aspartic acid to lysine mutation in the IV1 domain of the protein) probably originated in the Nova Scotia region where it is distributed among two haplotypes and occurs at high frequency (*P* = 0.687). The ▽2 *Pan* I^B allele (characterized by a 12-bp deletion in the second intron) is likely to have originated in the Barents Sea region where it is nearly fixed (*P* = 0.921). Two observations suggest that both alleles have recently displaced previously abundant alleles in their centers of origin. First, the *Pan* I$^{A'}$ and ▽2 *Pan* I^B alleles exhibit very low nucleotide diversities (π = 0.0049 and 0.00020, respectively) compared to the inclusive allelic groups summarized in Table 3. Second, in geographic regions where the *Pan* I$^{A'}$ and ▽2 *Pan* I^B alleles are most abundant, the alternate alleles exhibit their highest nucleotide diversities. *Pan* I^A alleles are most variable in the Barents Sea (π = 0.00147) where they occur at a frequency of only 0.073. Similarly, *Pan* I^B alleles are most polymorphic

in Nova Scotia (π = 0.00089) but are present at a frequency of only 0.100. These patterns are consistent with recent increases in the frequencies of selectively favored alleles at the expense of previously common alleles that had accumulated some silent polymorphism. Although this scenario hardly guarantees a stable balanced polymorphism, it suggests that evolutionary change can occur rapidly within both allelic groups without the need for geographic isolation.

A usual combination of balancing and directional selection is suggested from the genealogy of the two *Pan* I alleles shown in Figure 5. These two forms of selection are known to exert opposing effects on the predicted levels of silent polymorphism and the structures of allelic genealogies. Balancing selection is expected to significantly extend coalescence times (TAKAHATA 1990; TAKAHATA and NEI 1990) and lead to an accumulation of neutral polymorphism surrounding the site(s) affected by selection (STROBECK 1983; HUDSON and KAPLAN 1988). In contrast, directional selection is expected to shorten coalescence times and significantly reduce linked silent variation through hitchhiking effects (MAYNARD SMITH and HAIGH 1974; KAPLAN *et al.* 1989) that may extend large distances from the selected locus (see HUDSON *et al.* 1997). If balancing selection is responsible for the long lineages of *Pan* I alleles, it has clearly not led to an elevation of linked silent polymorphism. The two allelic groups differ, on average, by only a few mutations and exhibit levels of nucleotide diversity well below that typically found at autosomal loci in Drosophila (reviewed by MORIYAMA and POWELL 1996). The low within-allele diversity compared to the high between-allele divergence can only be explained by diversity-reducing processes like population bottlenecks, selective sweeps, or background selection (CHARLESWORTH *et al.* 1993). Some support exists for selective sweeps as the cause of the reduced diversity because the amino acid substitutions required to purge linked silent polymorphism have occurred within both allelic groups.

The molecular evidence to date indicates that long-lived balanced polymorphisms are rare. Notable exceptions include the *Mhc* class I and II loci in vertebrates and S alleles in plants both of which possess a high

number of alleles that commonly have long coalescent times that transcend species boundaries (reviewed by Klein *et al.* 1998). If a prolonged period of balancing selection has been acting at the *Pan* I locus in *G. morhua*, it differs from *Mhc* and S loci by apparently favoring only two alleles that have each undergone repeated amino acid substitutions. Therefore, the mechanisms favoring high allelic diversity at *Mhc* and S loci do not appear to be operating at the pantophysin locus of *G. morhua*. The rapid turnover of alleles at the *Pan* I locus of *G. morhua* is, however, consistent with data accumulating from Drosophila suggesting that many balanced polymorphisms may have evolved recently and not accumulated silent polymorphism around the selected site like the *Adh* locus in *D. melanogaster* (Kreitman and Hudson 1991). The list of allozyme polymorphisms studied at the molecular level in *D. melanogaster* that fail to show statistical support for persistent balancing selection include *Est*6 (Cooke and Oakeshott 1989), *Gpdh* (Takano *et al.* 1993), 6*Pgd* (Begun and Aquadro 1994), *G6pd* (Eanes *et al.* 1993, 1996), *Sod* (Hudson *et al.* 1994, 1997), and *Tpi* (Hasson *et al.* 1998).

A striking feature of the genealogy of *Pan* I alleles is the apparent absence of intragenic recombination. Although this gene may occur in a region of low recombination, this result is surprising because recombination was detected among three polymorphic restriction sites scored in the vicinity of the *Pan* I locus (especially in the NW Atlantic where four recombinant haplotypes were present). No intragenic recombination events were detected within either allelic group. However, 1 *Pan* I^A allele (NS28-A) had an A to G mutation at position 1407 that was fixed in all 62 *Pan* I^B alleles (Figure 3). A group of 10 *Pan* I^B alleles were also found in the NW Atlantic having mutations at positions 1580 and 1650 that were both fixed in *Pan* I^A alleles. Both may represent cases of interallelic recombination or gene conversion in the 3′ region of the gene. No recombination was detected at the 5′ end of the gene where the pattern of fixed differences between alleles suggests that such events could be disadvantageous.

The two *Pan* I alleles are most highly differentiated in a 30-bp region of the second intron and a 54-bp region in the fourth exon where five of the six amino acid replacements have occurred. The intron region is capable of forming a stem-loop structure and thus may affect pre-mRNA stability and/or processing. All intron insertion/deletion mutations have occurred within the *Pan* I^B lineage approximately 400 bp upstream from three amino acid changes. It is possible that epistatic natural selection is maintaining the association of the intron and amino acid mutations in the *Pan* I^B alleles thus generating linkage disequilibrium. A similar link between intron and amino acid polymorphisms has been made for the *Adh* locus of *D. melanogaster* where a compound insertion/deletion mutation (∇1) in the first intron exists in linkage disequilibrium with the fast allele and exhibits parallel clinal variation with the muta-

tion producing the fast/slow allozyme polymorphism in eastern North America (Berry and Kreitman 1993). If epistatic selection is acting at the *Pan* I locus between the second intron and fourth exon it is too restrictive to explain the disequilibrium detected over the entire gene region. Only selective sweeps driven by mutations in the second intron and/or fourth exon, or perhaps at closely linked genes, can account for the linkage disequilibrium present in the *Pan* I gene region.

The biochemical basis of how natural selection may be acting at the *Pan* I locus of *G. morhua* is unknown. Pantophysin is a recently discovered cellular isoform of the neuroendocrine integral membrane protein synaptophysin (Haass *et al.* 1996). Using immunoelectron microscopy, pantophysin has been localized to small (<100 nm) cytoplasmic microvesicles that likely function in various membrane-trafficking pathways of various cell types (see Haass *et al.* 1996). The tissue-specific expression of pantophysin appears variable and not closely paralleled by other vesicle-associated membrane proteins such as VAMPS and SCAMPS (Windoffer *et al.* 1999). Although nothing is known of the functioning of pantophysin in fishes, the differences detected between the *Pan* I alleles of *G. morhua* suggest that the polymorphism could be related to the differential expression and/or functioning of the protein in different tissues. This possibility can be tested by comparing the *in situ* levels and/or distribution of pantophysin in different tissues for different *Pan* I genotypes. The intravesicular loops of physins (notably synaptophysin) have not previously been identified as being important domains of the protein (see Johnston *et al.* 1989). However, the strong footprint of selection in the IV1 domain of pantophysin in *G. morhua* strongly suggests that it must be performing some critical function(s).

The form of balancing selection that could be operating at the *Pan* I locus is also unclear. The recent origin and spread of the *Pan* I$^{A'}$ and ∇2 *Pan* I^B alleles suggest that stable spatially varying selection is not favoring different alleles in different regions. Overdominance also appears unlikely because the *Pan* I locus was the only marker not to contribute to a correlation between DNA heterozygosity and growth rate in *G. morhua* (Pogson and Fevolden 1998) and because the near fixation of *Pan* I alleles in the Barents Sea, Nova Scotia, and the North Sea is inconsistent with a general fitness advantage expected for heterozygotes. Frequency-dependent selection may be operating at the *Pan* I locus but the mechanism(s) that would prevent complete fixation of alleles is unknown. The recent spread of selectively favored mutations in both allelic classes demonstrates the extremely dynamic nature of the *Pan* I polymorphism of *G. morhua*. Hitchhiking events that are not occurring uniformly across species ranges have also been described in *D. melanogaster* (Begun and Aquadro 1993) and *D. ananassae* (Stephan and Mitchell 1992).

Simulation studies have shown that Tajima's (1989) *D*

and Fu and Li's (1993) *F* and *D* statistics have reasonable power in detecting selective sweeps caused by the fixation of advantageous mutations (Braverman *et al.* 1995; Simonsen *et al.* 1995). Although the genealogies of both *Pan* I alleles show evidence for selective sweeps, both tests of neutrality failed to produce consistent results for the CRSs assembled from the five populations and neither test produced a single significant result for the pooled CRS of 124 alleles. This unexpected result was apparently caused by combining both *Pan* I alleles into the analyses, which had the effect of eliminating the signal from the data. When the alleles are considered separately, Tajima's test is significant for *Pan* I^A ($D = -1.883$, $P < 0.05$) and nearly significant for *Pan* I^B ($D = -1.583$, $0.10 > P > 0.05$) while Fu and Li's *D* and *F* statistics are significant for both *Pan* I^A alleles ($D = -2.954$ and $F = -3.045$, both $P < 0.02$) and *Pan* I^B alleles ($D = -2.952$ and $F = -2.916$, both $P < 0.02$). These results demonstrate that strong departures from neutral genealogies need not necessarily produce significant test statistics. The significant McDonald and Kreitman test (Table 8) was caused by the long independent evolutionary histories of the two alleles, which behaved as if they were sampled from different species. If recombination between alleles had been more extensive, this test would likely have not been significant.

In summary, nucleotide sequence variation at the *Pan* I locus in *G. morhua* has provided strong evidence for an unusual mixture of balancing and directional selection. The significant linkage disequilibrium and large differences in the frequencies of *Pan* I alleles among populations do not appear to be caused by stable spatially varying selection but by the recent appearance and spread of selectively favored mutations in both allelic groups in different geographic areas. Although the two *Pan* I alleles have had long evolutionary histories, they have not accumulated polymorphism at linked silent sites because of repeated amino acid substitutions within each allelic lineage. The type of balancing selection that could be acting at the *Pan* I locus is presently unknown. However, the discovery of this polymorphism at 1 of the 11 anonymous cDNA-based RFLP markers initially chosen for population studies by Pogson *et al.* (1995) suggests that long-lived polymorphisms may be more common than previously believed.

LITERATURE CITED

Aquadro, C. F., D. J. Begun and E. C. Kindahl, 1994 Selection, recombination and the levels of DNA polymorphism in Drosophila, pp. 45–56 in *Non-neutral Evolution: Theories and Molecular Data*, edited by G. B. Golding. Chapman & Hall, New York.

Begun, D. J., and C. F. Aquadro, 1992 Levels of naturally occurring DNA polymorphism correlate with recombination rates in *D. melanogaster*. Nature **356:** 519–520.

Begun, D. J., and C. F. Aquadro, 1993 African and North American populations of *Drosophila melanogaster* are very different at the DNA level. Nature **356:** 519–520.

Begun, D. J., and C. F. Aquadro, 1994 Evolutionary inferences from DNA variation at the 6-phosphogluconate dehydrogenase locus in natural populations of *Drosophila*: selection and geographic differentiation. Genetics **136:** 155–170.

Berry, A., and M. Kreitman, 1993 Molecular analysis of an allozyme cline: alcohol dehydrogenase in *Drosophila melanogaster* on the east coast of North America. Genetics **134:** 869–893.

Braverman, J. M., R. R. Hudson, N. L. Kaplan, C. H. Langley and W. Stephan, 1995 The hitchhiking effect on the site frequency spectrum of DNA polymorphisms. Genetics **140:** 783–796.

Cavalli-Sforza, L., 1966 Population structure and human evolution. Proc. R. Soc. Lond. Ser. B Biol. Sci. **164:** 362–379.

Charlesworth, B., 1998 Measures of divergence between populations and the effects of forces that reduce variability. Mol. Biol. Evol. **15:** 538–543.

Charlesworth, B., M. T. Morgan and D. Charlesworth, 1993 The effect of deleterious mutations on neutral molecular variation. Genetics **134:** 1289–1303.

Charlesworth, B., M. Nordborg and D. Charlesworth, 1997 The effects of local selection, balanced polymorphism and background selection on equilibrium patterns of genetic diversity in subdivided populations. Genet. Res. **70:** 155–174.

Cooke, P. H., and J. G. Oakeshott, 1989 Amino acid polymorphisms for *esterase-6* in *Drosophila melanogaster*. Proc. Natl. Acad. Sci. USA **86:** 1426–1430.

Cowan, D., M. Linial and R. H. Scheller, 1990 *Torpedo* synaptophysin: evolution of a synaptic vesicle protein. Brain Res. **509:** 1–7.

Eanes, W. F., M. Kirchner and J. Noon, 1993 Evidence for adaptive evolution of the *G6pd* gene in the *Drosophila melanogaster* and *Drosophila simulans* lineages. Proc. Natl. Acad. Sci. USA **90:** 7475–7479.

Eanes, W. F., M. Kirchner, J. Noon, C. H. Biermann, I.-N. Wang *et al.*, 1996 Historical selection, amino acid polymorphism and lineage-specific divergence at the *G6pd* locus in *Drosophila melanogaster* and *D. simulans*. Genetics **144:** 1027–1041.

Excoffier, L., P. E. Smouse and J. M. Quattro, 1992 Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. Genetics **131:** 479–491.

Fernandez-Chacon, R., and T. C. Südhof, 1999 Genetics of synaptic vesicle function: toward the complete functional anatomy of an organelle. Annu. Rev. Physiol. **61:** 753–776.

Fevolden, S. E., and G. H. Pogson, 1997 Genetic divergence at the synaptophysin (*Syp* I) locus among Norwegian coastal and northeast Arctic populations of Atlantic cod. J. Fish Biol. **51:** 895–908.

Fu, Y.-X., and W-H. Li, 1993 Statistical tests of neutrality of mutations. Genetics **133:** 693–709.

Gaitanou, M., A. Malaki, E. Merkouri and R. Matsas, 1997 Purification and cDNA cloning of mouse BM89 antigen shows that it is identical with the synaptic vesicle protein synaptophysin. J. Neurosci. Res. **48:** 507–514.

Haass, N. K., J. Kartenbeck and R. E. Leube, 1996 Pantophysin is a ubiquitously expressed synaptophysin homologue and defines constitutive transport vesicles. J. Cell Biol. **134:** 731–746.

Hasson, E., I.-N. Wang, L.-W. Zeng, M. Kreitman and W. F. Eanes, 1998 Nucleotide variation in the triosephosphate isomerase (*Tpi*) locus of *Drosophila melanogaster* and *Drosophila simulans*. Mol. Biol. Evol. **15:** 756–769.

Hey, J., 1999 The neutralist, the fly and the selectionist. Trends Ecol. Evol. **14:** 35–38.

Hudson, R. R., 1990 Gene genealogies and the coalescent process, pp. 1–44 in *Oxford Surveys in Evolutionary Biology*, edited by D. Futuyma and J. Antonovics. Oxford University Press, Oxford.

Hudson, R. R., and N. L. Kaplan, 1988 The coalescent process in models with selection and recombination. Genetics **120:** 831–840.

Hudson, R. R., K. Bailey, D. Skarecky, J. Kwiatowski and F. J. Ayala, 1994 Evidence for positive selection in the superoxide dismutase (*Sod*) region of *Drosophila melanogaster*. Genetics **136:** 1329–1340.

Hudson, R. R., A. G. Sáez and F. J. Ayala, 1997 DNA variation at the *Sod* locus of *Drosophila melanogaster*: an unfolding story of natural selection. Proc. Natl. Acad. Sci. USA **94:** 7725–7729.

Johnston, P. A., R. Jahn and T. C. Südhof, 1989 Transmembrane topography and evolutionary conservation of synaptophysin. J. Biol. Chem. **264:** 1268–1273.

Jonsdottir, O. D. B., A. K. Imsland, A. K. Danielsdottir, V. Thorsteinsson and G. Naevdal, 1999 Genetic differentiation among Atlantic cod in south and south-east Icelandic waters: synaptophysin (*Syp* I) and haemoglobin (*Hb*I) variation. J. Fish Biol. **54:** 1259–1274.

Kaplan, N. L., R. R. Hudson and C. F. Langley, 1989 The "hitchhiking effect" revisited. Genetics **123:** 887–899.

Karotam, J., T. M. Boyce and J. G. Oakeshott, 1995 Nucleotide variation at the hypervariable esterase 6 isozyme locus of *Drosophila simulans*. Mol. Biol. Evol. **12:** 113–122.

Katz, L. A., and R. G. Harrison, 1997 Balancing selection on electrophoretic variation of phosphoglucose isomerase in two species of field cricket: *Gryllus veletis* and *G. pennsylvanicus*. Genetics **147:** 609–620.

Klein, J., A. Sato, S. Nagl and C. O'Huigin, 1998 Molecular transspecies polymorphism. Annu. Rev. Ecol. Syst. **29:** 1–21.

Kliman, R. M., and J. Hey, 1993 DNA sequence variation at the period locus within and among species of the *Drosophila melanogaster* complex. Genetics **133:** 375–387.

Kreitman, M., 1991 Detecting selection at the level of DNA, pp. 204–221 in *Evolution at the Molecular Level*, edited by R. K. Selander, A. G. Clark and T. S. Whittam. Sinauer, Sunderland, MA.

Kreitman, M., and H. Akashi, 1995 Molecular evidence for natural selection. Annu. Rev. Ecol. Syst. **26:** 403–422.

Kreitman, M., and R. R. Hudson, 1991 Inferring the evolutionary history of the *Adh* and *Adh-dup* loci in *Drosophila melanogaster* from patterns of polymorphism and divergence. Genetics **127:** 565–582.

Kumar, S., K. Tamura and M. Nei, 1993 *MEGA: Molecular Evolutionary Genetic Analysis*, version 1.02. Pennsylvania State University, University Park, PA.

Maynard Smith, J., and J. Haigh, 1974 The hitch-hiking effect of a favorable gene. Genet. Res. **23:** 23–35.

McDonald, J. H., and M. Kreitman, 1991 Adaptive protein evolution at the *Adh* locus in *Drosophila*. Nature **351:** 652–654.

Moriyama, E. N., and J. R. Powell, 1996 Intraspecific nuclear DNA variation in Drosophila. Mol. Biol. Evol. **13:** 261–277.

Pogson, G. H., and S. E. Fevolden, 1998 DNA heterozygosity and growth rate in the Atlantic cod, *Gadus morhua* (L.). Evolution **52:** 915–920.

Pogson, G. H., K. A. Mesa and R. G. Boutilier, 1995 Genetic population structure and gene flow in the Atlantic cod *Gadus morhua*: a comparison of allozyme and nuclear RFLP loci. Genetics **139:** 375–385.

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

Raymond, M., and F. Rousset, 1995 GENEPOP (version 1.2): population genetics software for exact tests and ecumenicism. J. Hered. **86:** 248–249.

Rozas, J., and R. Rozas, 1997 DnaSP version 2.2: a novel software package for extensive molecular population genetics analysis. Comput. Appl. Biosci. **13:** 307–311.

Saitou, N., and M. Nei, 1987 The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. **4:** 406–425.

Schaeffer, S. W., and E. L. Miller, 1992 Estimates of gene flow in *Drosophila pseudoobscura* determined from nucleotide sequence analysis of the alcohol dehydrogenase region. Genetics **132:** 471–480.

Schulte, P. M., M. Gómez-Chiarri and D. A. Powers, 1997 Structural and functional differences in the promoter and 5′ flanking region of *Ldh-B* within and between populations of the teleost, *Fundulus heteroclitus*. Genetics **145:** 759–769.

Simonsen, K. L., G. A. Churchill and C. F. Aquadro, 1995 Properties of statistical tests of neutrality for DNA polymorphism data. Genetics **141:** 413–429.

Stephan, W., and S. J. Mitchell, 1992 Reduced levels of DNA polymorphism and fixed between-population differences in the centromeric region of *Drosophila ananassae*. Genetics **132:** 1039–1045.

Strobeck, C., 1983 Expected linkage disequilibrium for a neutral locus linked to a chromosomal arrangement. Genetics **103:** 545–555.

Südhof, T. C., F. Lottspeich, P. Greengard, E. Mehl and R. Jahn, 1987 A synaptic vesicle protein with a novel cytoplasmic domain and four transmembrane regions. Science **238:** 1142–1144.

Swofford, D. L., 1993 *PAUP: Phylogenetic Analysis Using Parsimony*, version 3.1. Illinois Natural History Survey, Champaign, IL.

Swofford, D. L., and R. B. Selander, 1989 *BIOSYS-1: A Computer Program for the Analysis of Allelic Variation in Population Genetics and Biochemical Systematics*. Illinois Natural History Survey, Champaign, IL.

Tajima, F., 1989 Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics **123:** 585–595.

Takahata, N., 1990 A simple genealogical structure of strongly balanced allelic lines and trans-species evolution of polymorphism. Proc. Natl. Acad. Sci. USA **87:** 2419–2423.

Takahata, N., and M. Nei, 1990 Allelic genealogy under overdominant and frequency-dependent selection and polymorphism of major histocompatibility complex loci. Genetics **124:** 967–978.

Takano, T. S., S. Kusakabe and T. Mukai, 1993 DNA polymorphisms and the origin of protein polymorphism at the *Gpdh* locus of *Drosophila melanogaster*, pp. 170–190 in *Mechanisms of Molecular Evolution*, edited by N. Takahata and A. G. Clark. Sinauer, Sunderland, MA.

Taylor, W. R., 1986 The classification of amino acid conservation. J. Theor. Biol. **119:** 205–218.

Weir, B. S., 1990 *Genetic Data Analysis*. Sinauer, Sunderland, MA.

Windoffer, R., M. Borchert-Stuhltrager, N. K. Haass, S. Thomas, M. Hergt *et al.*, 1999 Tissue expression of the vesicle protein pantophysin. Cell Tissue Res. **296:** 499–510.

Communicating editor: W. F. Eanes