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

Ubaldo Benitez Hernandez and Einar Árnason

Institute of Life and Environmental Sciences, University of Iceland, Reykjavik, Iceland

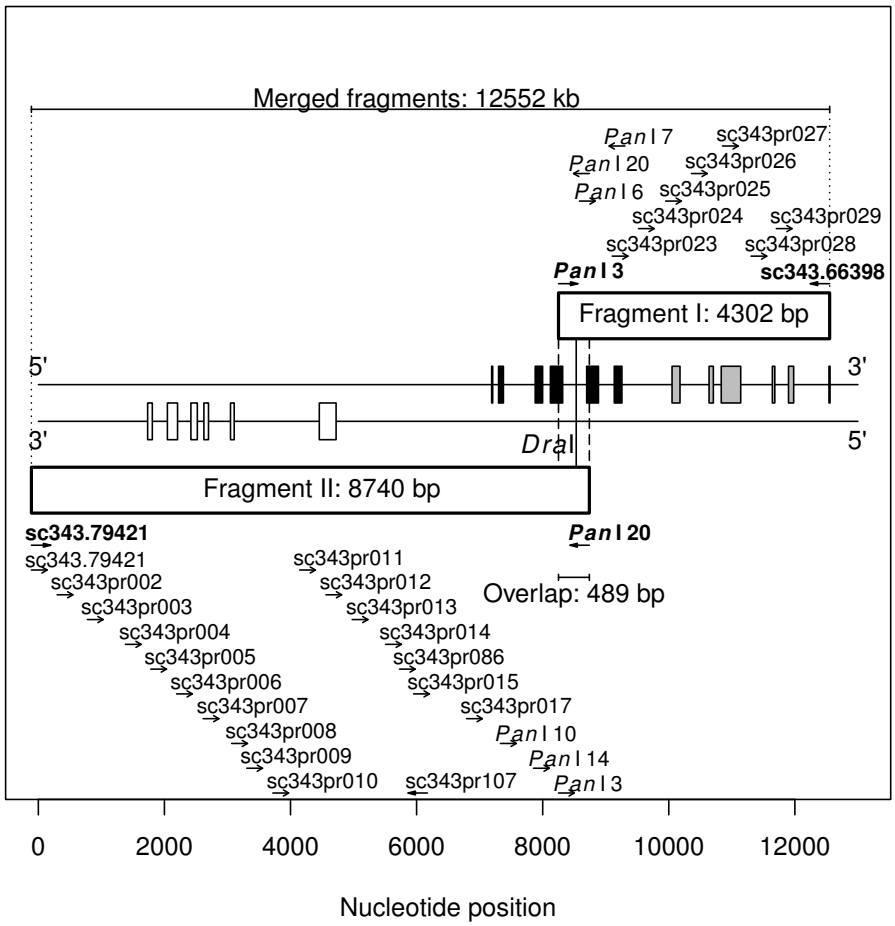

Figure S1. PCR fragment alignment and relative positions of the PCR and sequencing primers used in a 12.5 kb region containing the *Pan* I locus and its peripheric regions, the *Sort*1 and *Atxn7l2* loci (partial segments). All primers are identified by name and their position and direction are indicated by black arrows. PCR primers in bold. PCR fragments shown as rectangles and identified by name and length. The PCR fragments overlap by 489 bp demarcated by vertical dashed lines. Within this overlap lies the polymorphic *Dra*I restriction site (represented by a vertical solid line) defining the *A* and *B* alleles of the *Pan* I locus (Pogson, 2001). The start and end of the region obtained by sequence alignment is demarcated by vertical dotted lines. The length of the sequence alignment and PCR fragment overlap is indicated by horizontal solid truncated lines. Boxes represent the exons of *Sort*1 (partial segment), *Pan* I and *Atxn7l2* (partial segment), in white, black and gray, respectively. The solid black horizontal lines running through the boxes represent introns (between boxes of the same color) and intergenic space (between boxes of different color).

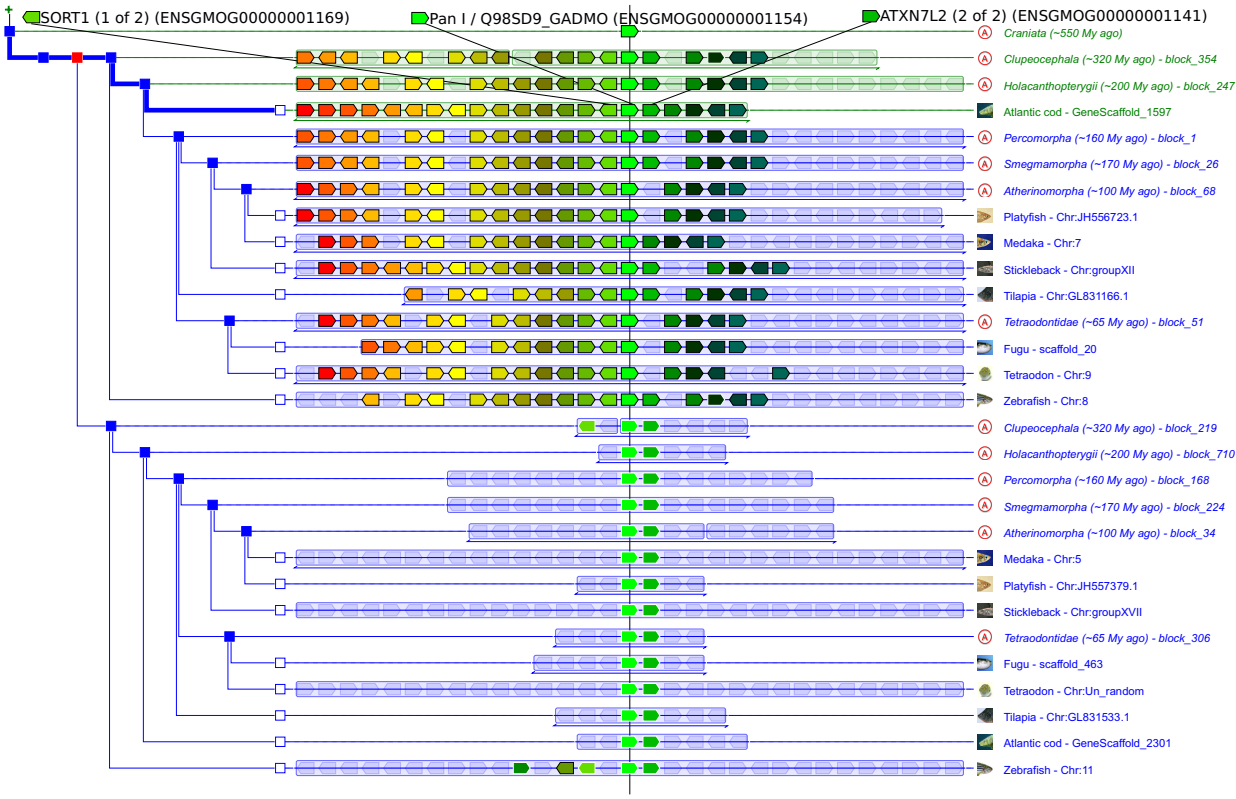

Figure S2. Gene order of the *Pan* I locus and its neighboring genes and its orthologs and paralogs across different species. The *Pan* I locus is the reference gene in the middle, flanked by the *Sort*1 and *Atxn7l2* loci. Orthologs in other species are shown in matching colors. The blues structure at the left is a phylogenetic tree for the *Pan* I locus. The image is a Phyloview diagram computed by Genomicus (Louis et al., 2013; Muffato et al., 2010) with version 70.01 and search name ENSGMOG00000001154. Phylogenetic tree computed by Ensembl v.70 Flicek et al. (2014)

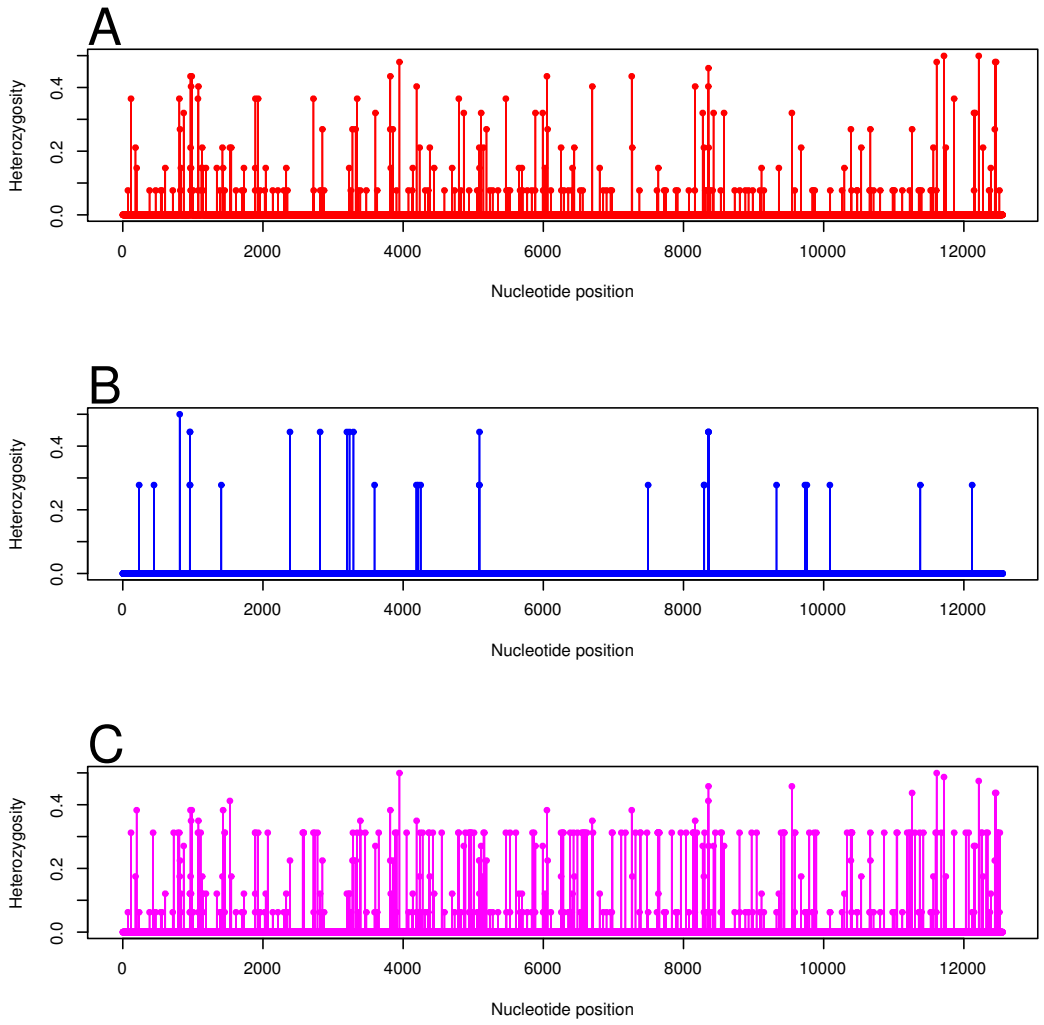

Figure S3. Heterozygosity on nucleotide position. Heterozygosity among *A* alleles (top panel, red), among *B* alleles (middle panel, blue), and among all sequences combined (bottom panel, magenta).

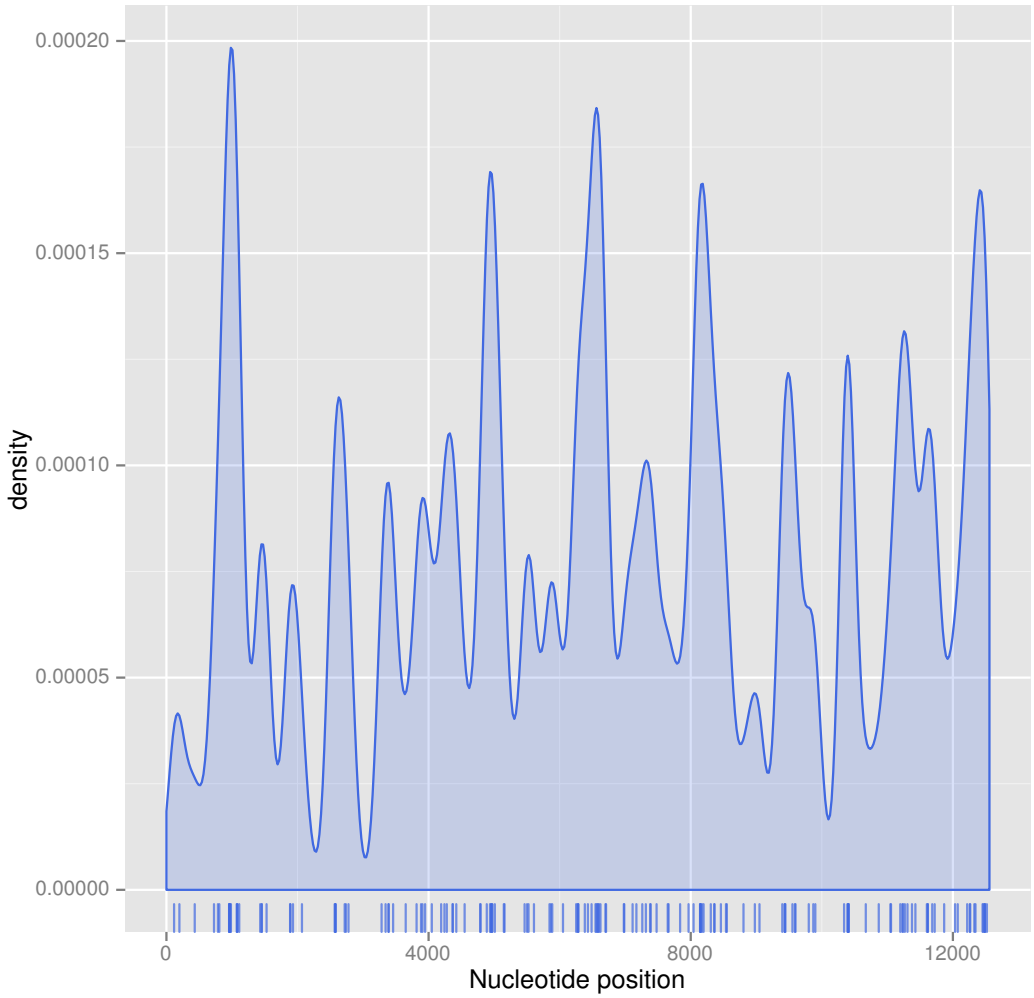

Figure S4. Density of high heterozygosity SNPs along the sequenced fragment. Minor allele frequency set at 6/31, the frequency of the *B Pan* I alleles among the 31 sequences.

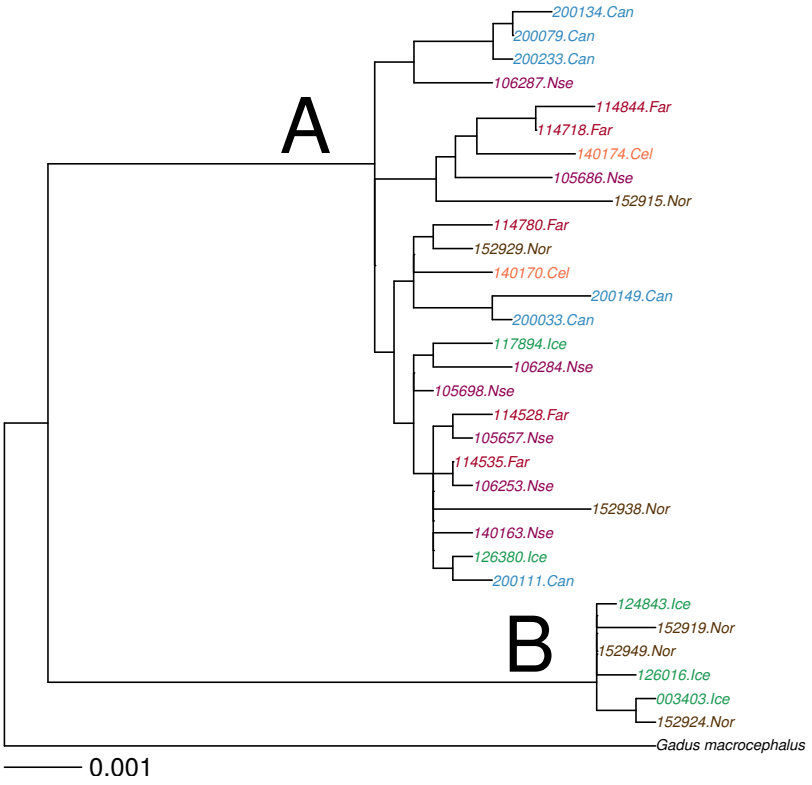

Figure S5. Phylogenetic tree of the *Pan* I and *Atxn7l2* loci in Atlantic cod. The tree was inferred by Maximum Likelihood with Tamura 3-parameter model with Gamma distribution and invariable sites, with a 4.2 kb alignment of 31 Atlantic cod DNA sequences (25 *Pan* I^A and 6 *Pan* I^B sequence variants) containing partial segments of the *Pan* I and *Atxn7l2* loci. Pacific cod *Gadus macrocephalus* was used as the outgroup. Branch lengths represent the number of substitutions per site. Taxa are labeled with a six digit individual barcode, country color and alphabetic code. Can=Canada (blue), Ice=Iceland (green), Far=Faroe Islands (red), Nor=Norway (brown), Nse=the North Sea (purple), and Cel=the Celtic Sea (orange). Clades A and B, respectively, encompass sequences with an absent or present *Dra*I restriction site that defines the *A* and *B* alleles of the *Pan* I locus.

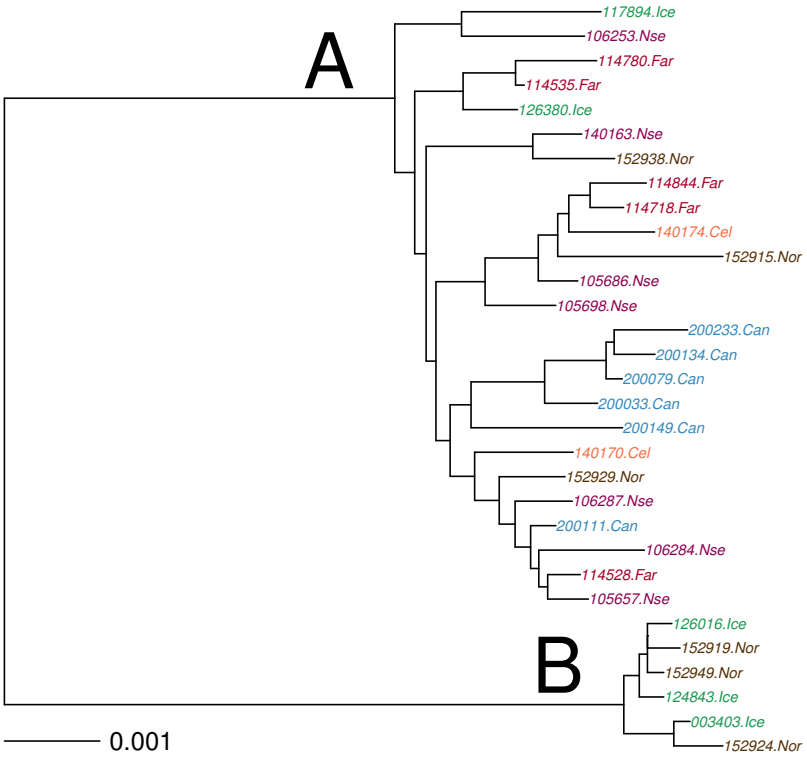

Figure S6. Phylogenetic tree of the *Pan* I, *Sort*1 and *Atxn*7*l*2 loci in Atlantic cod. The tree was inferred by Maximum Likelihood with Tamura 3-parameter model with Gamma distribution and invariable sites, with a 12.56 kb alignment of 31 Atlantic cod DNA sequences (25 *Pan* I^A and 6 *Pan* I^B sequence variants) containing the *Pan* I locus and its peripheric regions, the *Sort*1 and *Atxn*7*l*2 loci (partial sequences). Branch lengths as number of substitutions per site. Taxa are labeled with a six digit individual barcode, country color and alphabetic code. Can=Canada (blue), Ice=Iceland (green), Far=Faroe Islands (red), Nor=Norway (brown), Nse=the North Sea (purple), and Cel=the Celtic Sea (orange). Clades A and B, respectively, encompass sequences with an absent or present *Dra*I restriction site that defines the *A* and *B* alleles of the *Pan* I locus.

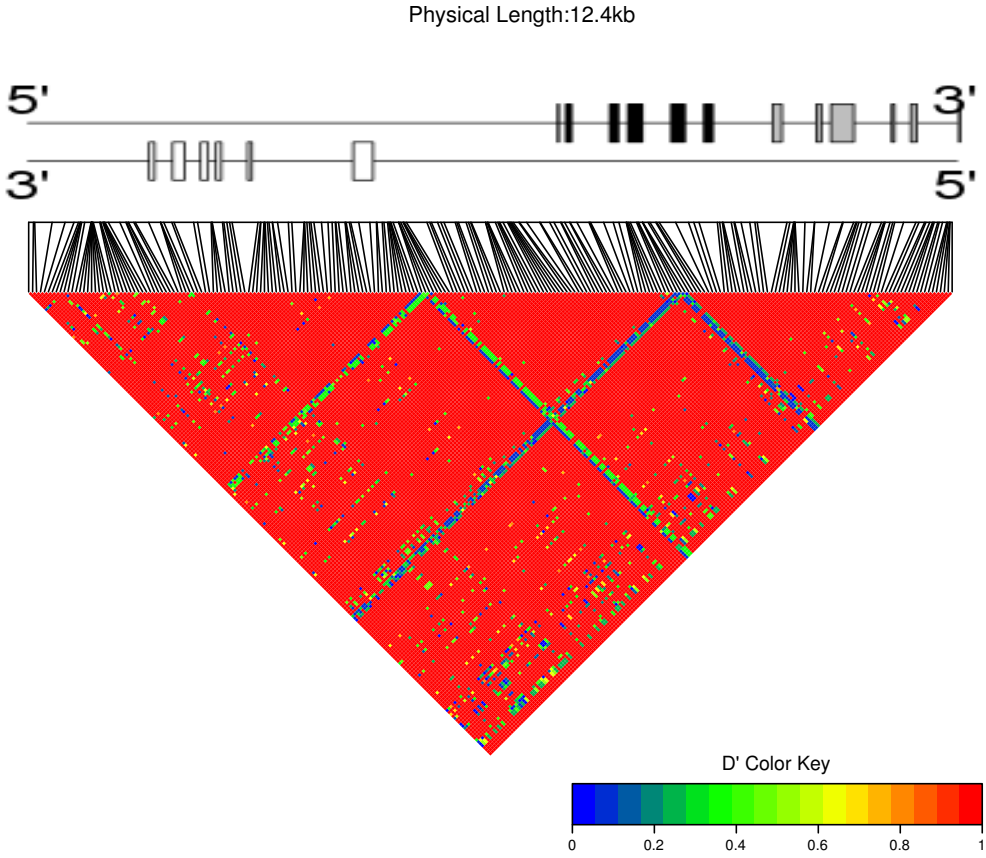

Figure S7. Linkage disequilibrium D' heatmap excluding singleton sites of the *Pan* I locus and its peripheric regions, the *Sort*1 and *Atxn7l2* loci. Minor allele frequency set at 2/31 to exclude singletons.

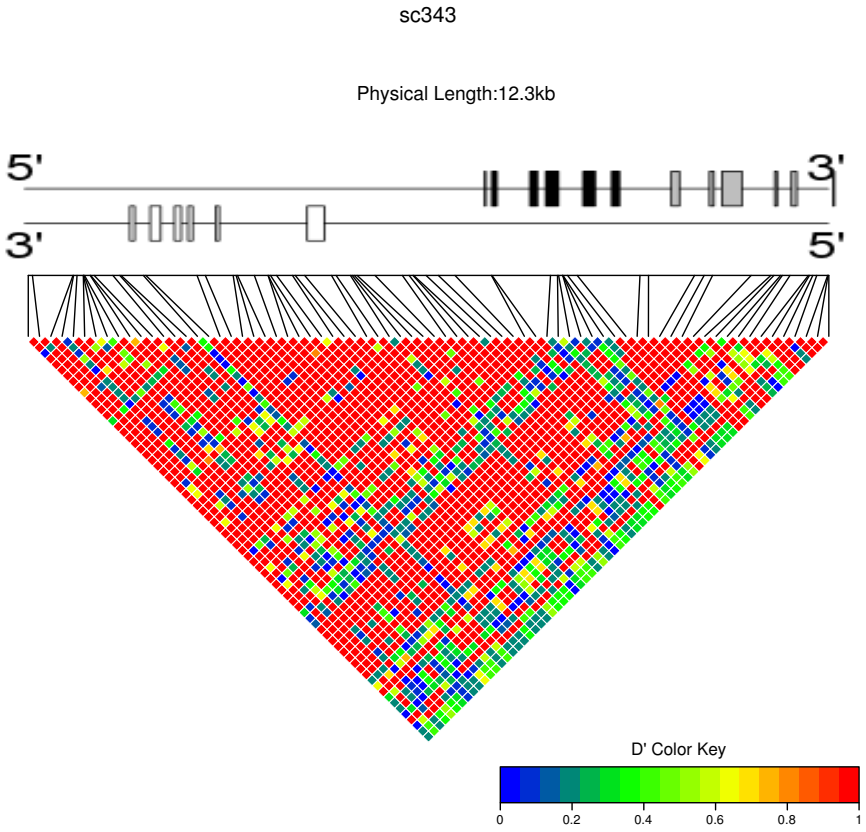

Figure S8. Linkage disequilibrium D' heatmap of high heterozygosity sites among *A* alleles. Minor allele frequency set at 3/25 excluding singletons and low frequency variants.

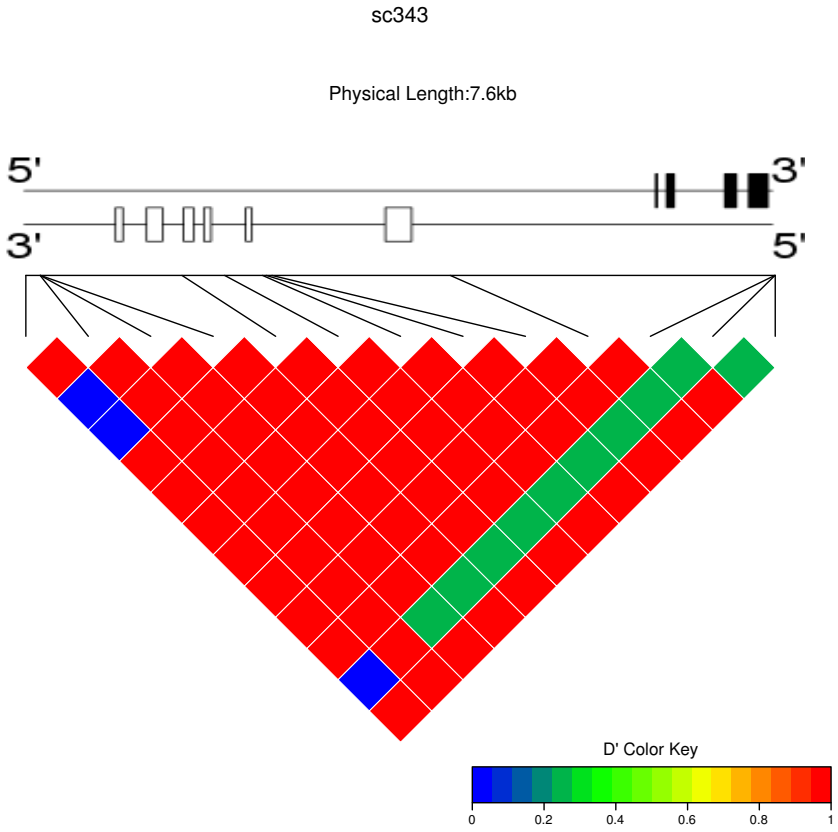

Figure S9. Linkage disequilibrium D' heatmap of high heterozygosity sites among *B* alleles. Minor allele frequency set at 2/6 to exclude singletons.

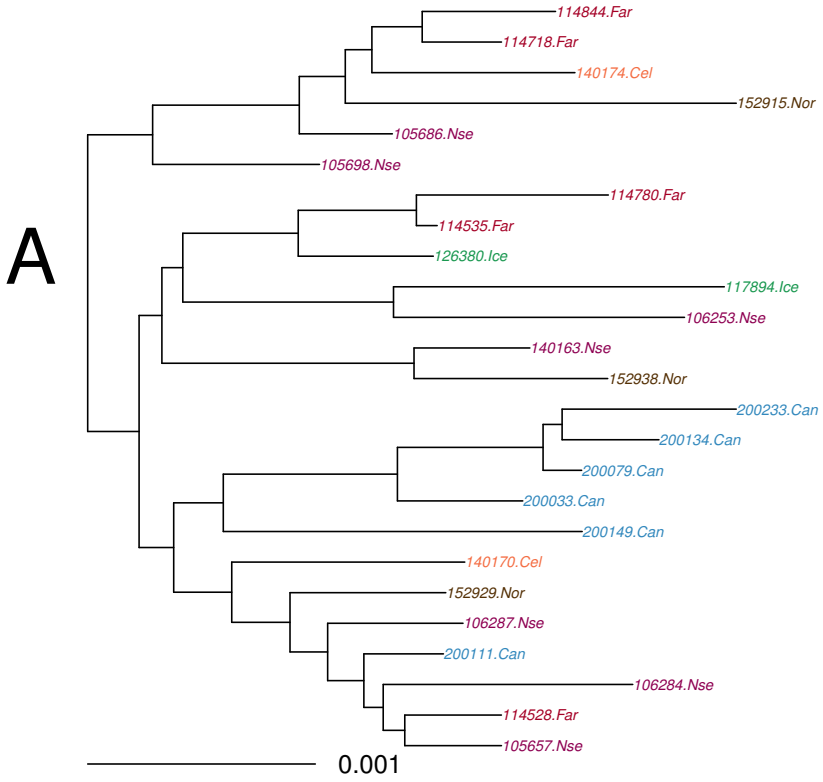

Figure S10. Phylogenetic tree of the *Pan* I, *Sort*1 and *Atxn*7*l*2 loci in Atlantic cod, from *Pan* I^A sequence variants. The tree was inferred by Maximum Likelihood with Tamura 3-parameter model with Gamma distribution and invariable sites, with a 12.56 kb alignment of 25 Atlantic cod DNA sequences (*Pan* I^A sequence variants) containing the *Pan* I locus and its peripheral regions, the *Sort*1 and *Atxn*7*l*2 loci (partial sequences). Branch lengths as number of substitutions per site. Taxa are labeled with a six digit individual barcode, country color and alphabetic code. Can=Canada (blue), Ice=Iceland (green), Far=Faroe Islands (red), Nor=Norway (brown), Nse=the North Sea (purple), and Cel=the Celtic Sea (orange). Clades A and B, respectively, encompass sequences with an absent or present *Dra*I restriction site that defines the *A* and *B* alleles of the *Pan* I locus.

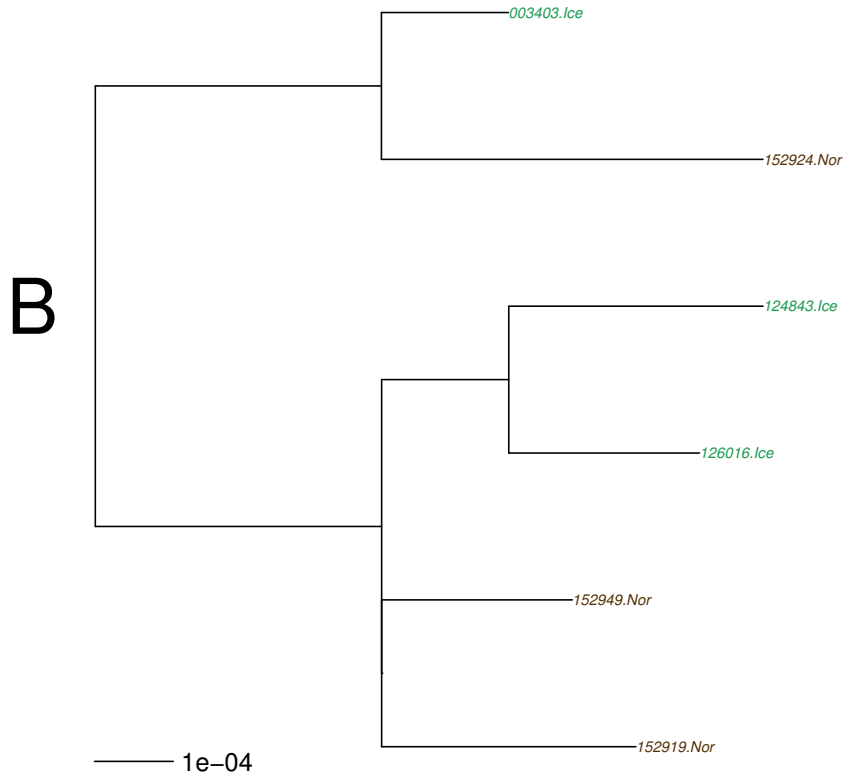

Figure S11. Phylogenetic tree of the *Pan* I, *Sort*1 and *Atxn*7*l*2 loci in Atlantic cod, from *Pan* I^B sequence variants. The tree was inferred by Maximum Likelihood with Tamura 3-parameter model, with a 12.56 kb alignment of six Atlantic cod DNA sequences (*Pan* I^B sequence variants) containing the *Pan* I locus and its peripheral regions, the *Sort*1 and *Atxn*7*l*2 loci (partial sequences). Branch lengths as number of substitutions per site. Taxa are labeled with a six digit individual barcode, country color and alphabetic code. Ice=Iceland (green) and Nor=Norway (brown).

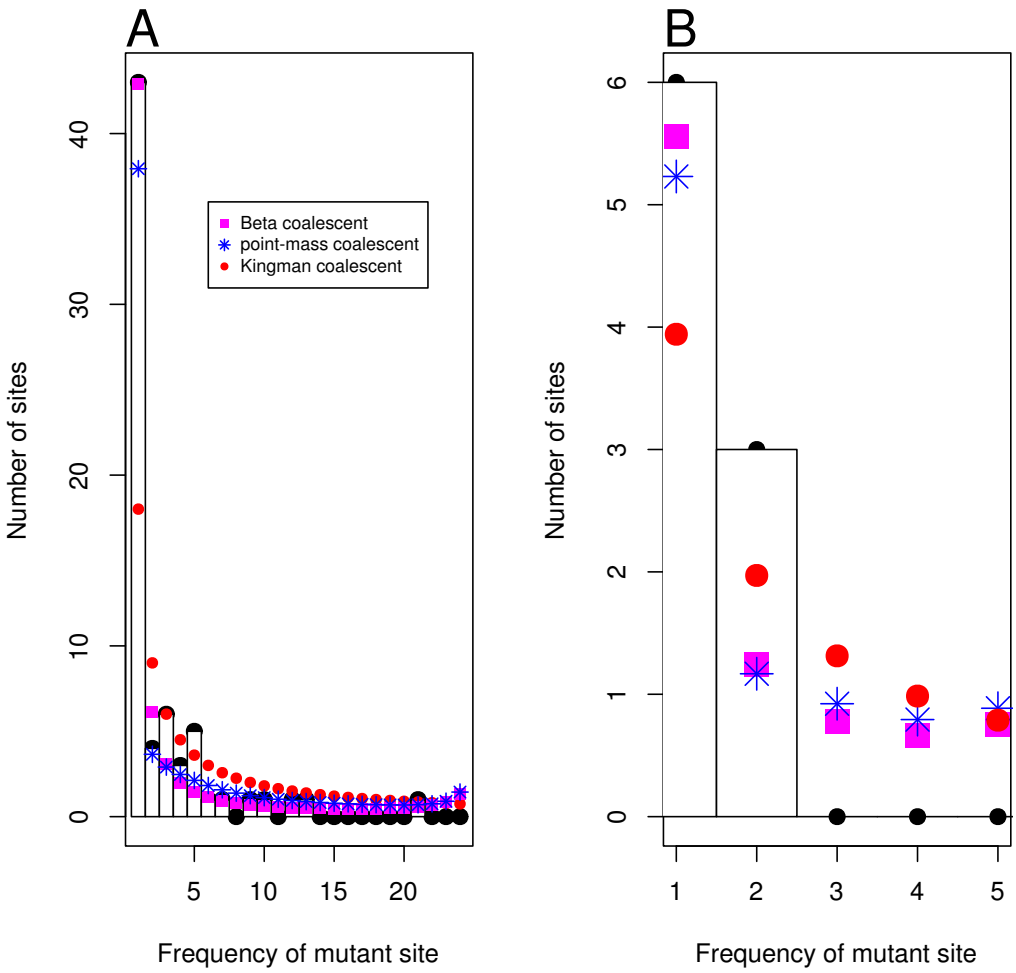

Figure S12. Unfolded site frequency spectrum of Atlantic cod *Pan I* and *Atxn7l2* genes classified by *Pan I A* alleles (**A** panel) and *B* alleles (**B** panel). *Gadus macrocephalus* was used as the outgroup. Number of individuals $n = 25$ and $n = 6$ respectively. Theoretical expectation under Kingman coalescent (solid red dots), Beta$(2-\alpha-\alpha)$ coalescent (magenta squares), and point-mass coalescent (blue stars).

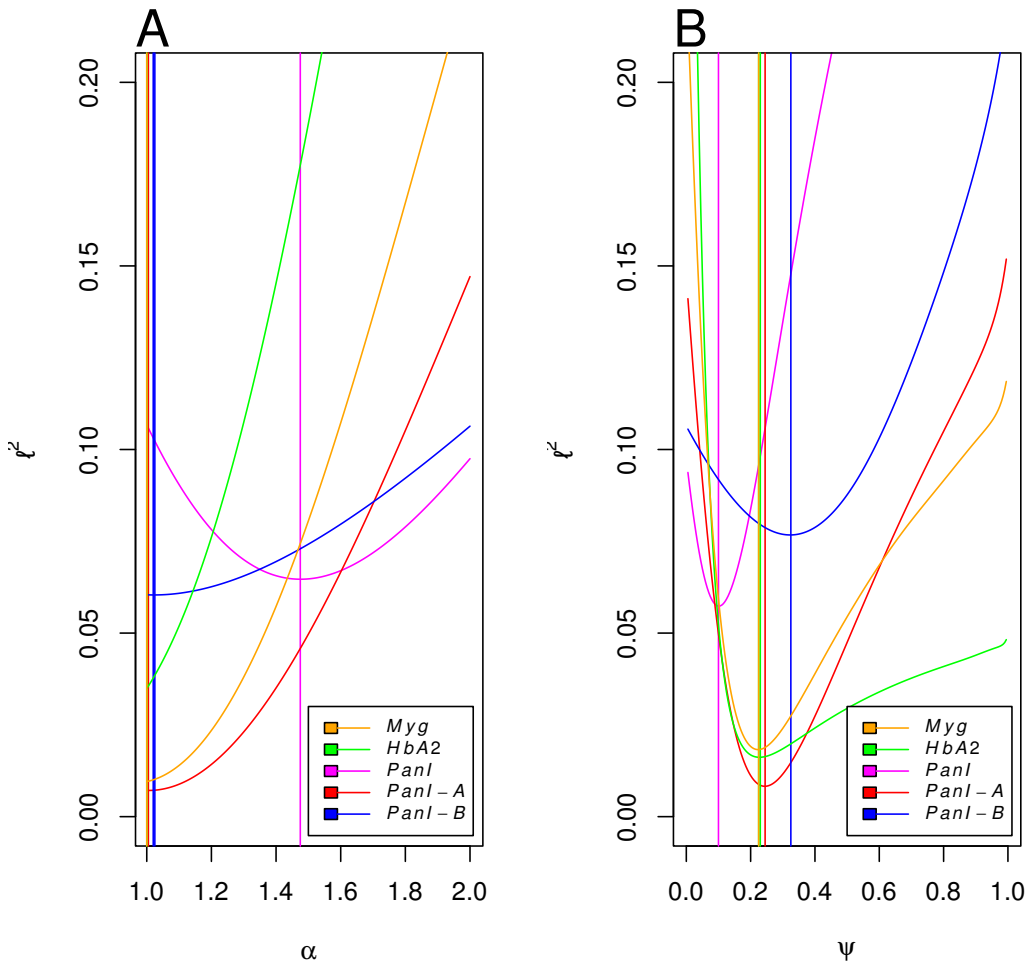

Figure S13. The ℓ^2 distance for the unfolded site frequency spectrum of the nuclear genes *Myg*, *Hb2A*, *Pan I* and *Atxn7l2* genes classified by the *Pan I-A* and *Pan I-B* alleles of *Pan I* on the α parameter of the Beta$(2-\alpha, \alpha)$ coalescent (**A** panel) and the ψ parameter of the Point-Mass coalescent (**B** panel). Data on *Myg* and *HaA2* from Árnason and Halldórsdóttir (2015).

Table S1. PCR and sequencing primers used in Atlantic and Pacific cod. All the primers were used for Atlantic cod for a 12.5 kb region containing the *Pan* I locus and its peripheral loci, the *Sort*1 and *Atxn*7*l*2 loci (partial segments). For Pacific cod (the outgroup), only the primers within positions 8251 bp and 12552 bp were used for a 4.3 kb region containing partial segments of the *Pan* I locus and its peripheral region the *Atxn*7*l*2 locus.

Assay	Primer Name	Primer Sequence 5′ - 3′	Position (in bp)
PCR and Sequencing	*Pan* I 3	CGTTGGTCCTCTATCTGGGCTTC	8251
Sequencing	*Pan* I 6	ACCTTTACTCTCTATCTCCCG	8583
Sequencing	*Pan* I 7	CGTAGCAGAAGAGTGACACAT	9304
Sequencing	*Pan* I 10	GCCATTGAAGGAGCCCCT	7330
Sequencing	*Pan* I 14	GACGCTTTCTTTGATTTGGCAG	7854
PCR and Sequencing	*Pan* I 20	AAGACGAAACCAACCACAGGA	8740
PCR	sc343.66398	GCTGGTGGATGGAGTGGAT	12552
PCR and Sequencing	sc343.79421	TGGCTGGTGAAGAAGATGGT	-110
Sequencing	sc343pr002	TGTAACACTGTGGCATGTAAACAG	298
Sequencing	sc343pr003	CGCTTACAGCTGTCATAGTC	778
Sequencing	sc343pr004	CCTCAAGTAGCGCAACATAGG	1381
Sequencing	sc343pr005	AAGTCTTTGGACAACCACAACTG	1783
Sequencing	sc343pr006	CTCCAGTCATGATGACCTTTGAG	2193
Sequencing	sc343pr007	ATGGGTACTTACCTCCGATAGAG	2614
Sequencing	sc343pr008	CAGGCCAGTGAAACAGATCC	3067
Sequencing	sc343pr009	TCTCTTTGTAACCTAGTAACGC	3306
Sequencing	sc343pr010	GAGAGGAGCAGAAAGTTGAG	3720
Sequencing	sc343pr011	GGTTTCAACTAAACTCTGTG	4145
Sequencing	sc343pr012	CAAGCCATGCAGGAAGAGAC	4561
Sequencing	sc343pr013	ACACAGTAGTCCTGACAGCG	4978
Sequencing	sc343pr014	ATTTGGACTTCTGTTACACG	5508
Sequencing	sc343pr015	CAGATTATATGGTTTGGTGGTG	5951
Sequencing	sc343pr017	GAGAGGTTACATCCAAATACC	6793
Sequencing	sc343pr023	CCCTGTCTCCTTATTTCTATTTGG	9103
Sequencing	sc343pr024	GTTGTGCCAACAGTGTTAAGTG	9519
Sequencing	sc343pr025	GTCGAGATATGGAAATATCTGC	9950
Sequencing	sc343pr026	CAAACTTAGTTTCTCGTGAC	10360
Sequencing	sc343pr027	GGCCCTTGACAACTTCTACC	10852
Sequencing	sc343pr028	CTCTGGTAACCCTTGCATCC	11304
Sequencing	sc343pr029	TTTGATTGTGCATGTCCTTGG	11706
Sequencing	sc343pr086	TCCTATCTTTACACTTAACCGAGC	5728
Sequencing	sc343pr107	GACCAAACCAGTCAGACCAG	6171

Table S2. Maximum likelihood analysis of a Kingman-coalescent HKA test of neutrality and selection at three genes in Atlantic cod.

					Hbα2		*Myg*		*Pan* I	
Description	$\ln L$	T	Test	*df*	θ	k	θ	k	θ	k
Neutral, all $k = 1$	-21.23	1.50			0.0039	1	0.0071	1	0.0051	1
Selection at *Pan* I	-17.99	3.83	6.48^{**}	1	0.0030	1	0.0055	1	0.0018	4.12

Test statistic is twice the $\ln L$ difference of the two models, neutrality and selection at *Pan* I. Three loci are under test: Hemoglobin α 2 (*Hbα2*), Myoglobin (*Myg*) (data from Árnason and Halldórsdóttir, 2015), and Pantophysin I (*Pan* I). θ is the scaled effective population size and the parameter k measures changes in diversity due to selection. Based on method of Wright and Charlesworth (2004). ** represents $P < 0.01$.

Table S3. Amova between North and South groups, within groups and among localities. The North (Can, Ice, Nor) vs South (Far, Nse, Cel) hypothesis was taken from Árnason and Halldórsdóttir (2015) based on patterns at the *Ckma* gene.

Source of variation	d.f.	Sum of squares	Variance components	Percentage of variation
Among groups	1	105.0	$V_a = 2.84$	5.56
Within groups	4	253.5	$V_b = 5.54$	10.87
Within populations	19	810.0	$V_c = 42.63$	83.57
Total	24	1168.5	51.01	

Fixation Indices $F_{SC} = 0.12$, $F_{ST} = 0.16$, $F_{CT} = 0.06$
Significance tests (1023 permutations): V_c and F_{ST}, $P = 0.021 \pm 0.004$; V_b and F_{SC}, $P = 0.095 \pm 0.007$; V_a and F_{CT}, $P = 0.289 \pm 0.013$

Table S4. Pairwise F_{ST} and associated probabilities among localities. Variation of the 12558 bp fragment among *Pan* I *A* alleles.

	Can	Ice	Nor	Nse	Cel	Far
Can		0.15 ± 0.04	$\mathbf{0.02 \pm 0.01}$	$\mathbf{0.02 \pm 0.0}1$	$\mathbf{0.05 \pm 0.02}$	$\mathbf{0.01 \pm 0.01}$
Ice	0.116		0.23 ± 0.04	$\mathbf{0.04 \pm 0.02}$	0.39 ± 0.05	0.19 ± 0.03
Nor	0.292	0.166		0.80 ± 0.04	0.40 ± 0.05	0.48 ± 0.04
Nse	0.257	0.195	-0.140		0.46 ± 0.05	0.22 ± 0.05
Cel	0.297	0.118	-0.000	0.021		0.73 ± 0.02
Far	0.280	0.147	-0.004	0.027	-0.173	

Table S5. Gross D_{xy} and net D_a nucleotide divergence per site and associated standard errors. For the 4.194 bp *Gadus macrocephalus* Gma was used as the outgroup. For the 12.558 bp fragment comparisons are made between the *A* and *B* alleles.

Fragment	Comparison	D_{xy}	$s_{D_{xy}}$	D_a	s_{D_a}
4.194 bp	*A* and *B* vs Gma	0.0144	0.0025	0.0115	0.0026
4.194 bp	*A* vs Gma	0.0139	0.0027	0.0125	0.0027
4.194 bp	*B* vs Gma	0.0155	0.0058	0.0151	0.0058
4.194 bp	*A* vs *B*	0.0123	0.0014	0.0104	0.0014
12.558 bp	*A* vs *B*	0.0122	0.0013	0.0103	0.0013

Divergence, *D*, and standard deviation, *s*, found using Jukes and Cantor correction.

Table S6. Likelihood ratio test statistics G for observed site frequency spectra and expectation according to different coalescent models.

Model	G	Comparison	ΔG	df
I. Kingman	477.10			
II. Beta$(2-\alpha,\alpha)$	455.92	I vs II	42.36	1
III. Point-Mass	451.79	I vs III	50.62	1