SUPPLEMENTARY MATERIAL

Quantifying genetic differentiation and population assignment among two contingents of Atlantic mackerel (*Scomber scombrus*) in the Northwest Atlantic

Audrey Bourret, Andrew Smith, Elisabeth Van Beveren, Stéphane Plourde, Kiersten L. Curti, Teunis Jansen, David E. Richardson, Martin Castonguay, Naiara Rodriguez-Ezpeleta, Geneviève J. Parent

TABLE S1

Number of loci and SNPs retained following filtration steps for both panels (NWA-NEA and NWA).

	NWA-NEA 0.1		NWA 0.1	
Steps	Loci	SNPs	loci	SNPs
gstacks module	1122204	-	1122204	-
population module	11167	16920	11067	16730
Filtration : Ho - F_{IS}	10888	16526	10759	16297
Filtration : Batch effect	10835	16462	10707	16232
Final panel	10832	10832	10703	10703

TABLE S2

AMOVA for 82 samples processed on HiSeq and NovaSeq illumina after the batch effect filtration step (see methods for more details) using the *poppr* R package (ade4 method). Sequencer type was used as the factor ("between" effect), and was not significant (Monte-Carlo test, P > 0.999, based on 999 replicates).

Variance	df	Sum Sq	Mean Sq	Sigma	%
Between samples	1	23	23	-19	-1
Within samples	162	255682	1578	1578	101
Total	163	255705	1569	1559	100

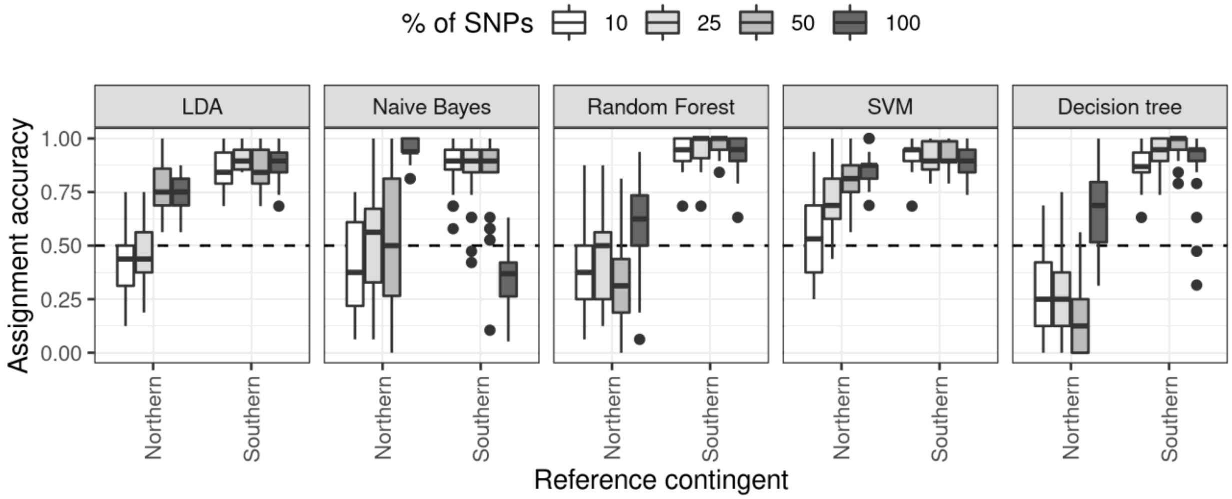

FIG. S1

Cross-validation results using northern and southern contingent reference samples, with five machine-learning algorithms (Linear discriminant function (LDA), naïve Bayes, random forest, support vector machine (SVM) and decision tree), and various proportions of SNPs (from 10% to 100%). SVM with all SNPs is the best predictive model.

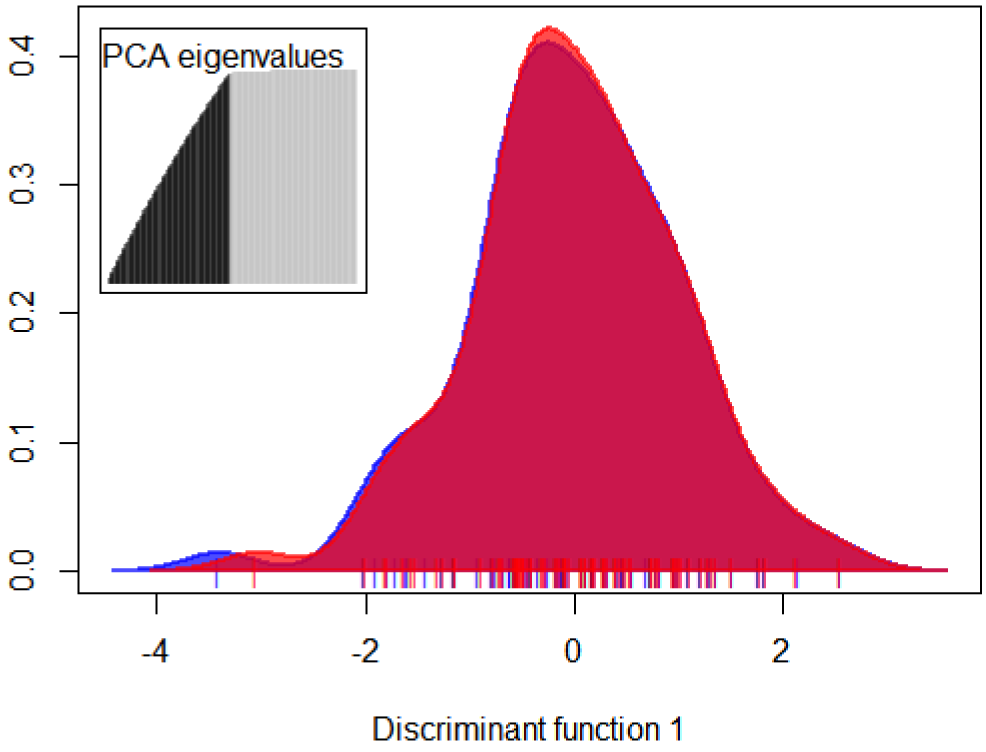

FIG. S2

Discriminant analysis of principal components (DAPC) results using the 82 samples processed on both sequencers (blue : HiSeq or HI; red = NovaSeq or NS).

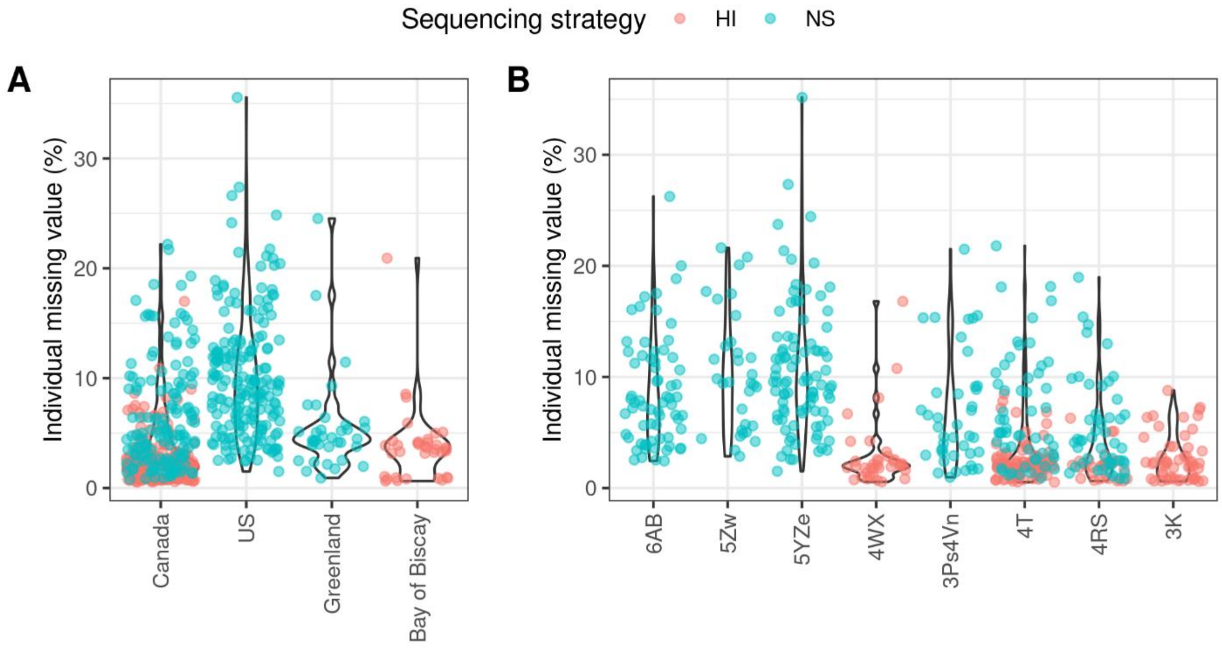

Fig. S3

Distribution of per individual missingness in the A) NWA-NEA panel and B) NWA panel. Violin plots represent the distribution by country or NAFO division, while each point a sample colored depending on the sequencer that was used (red = HiSeq (HI); blue = NovaSeq (NS)).

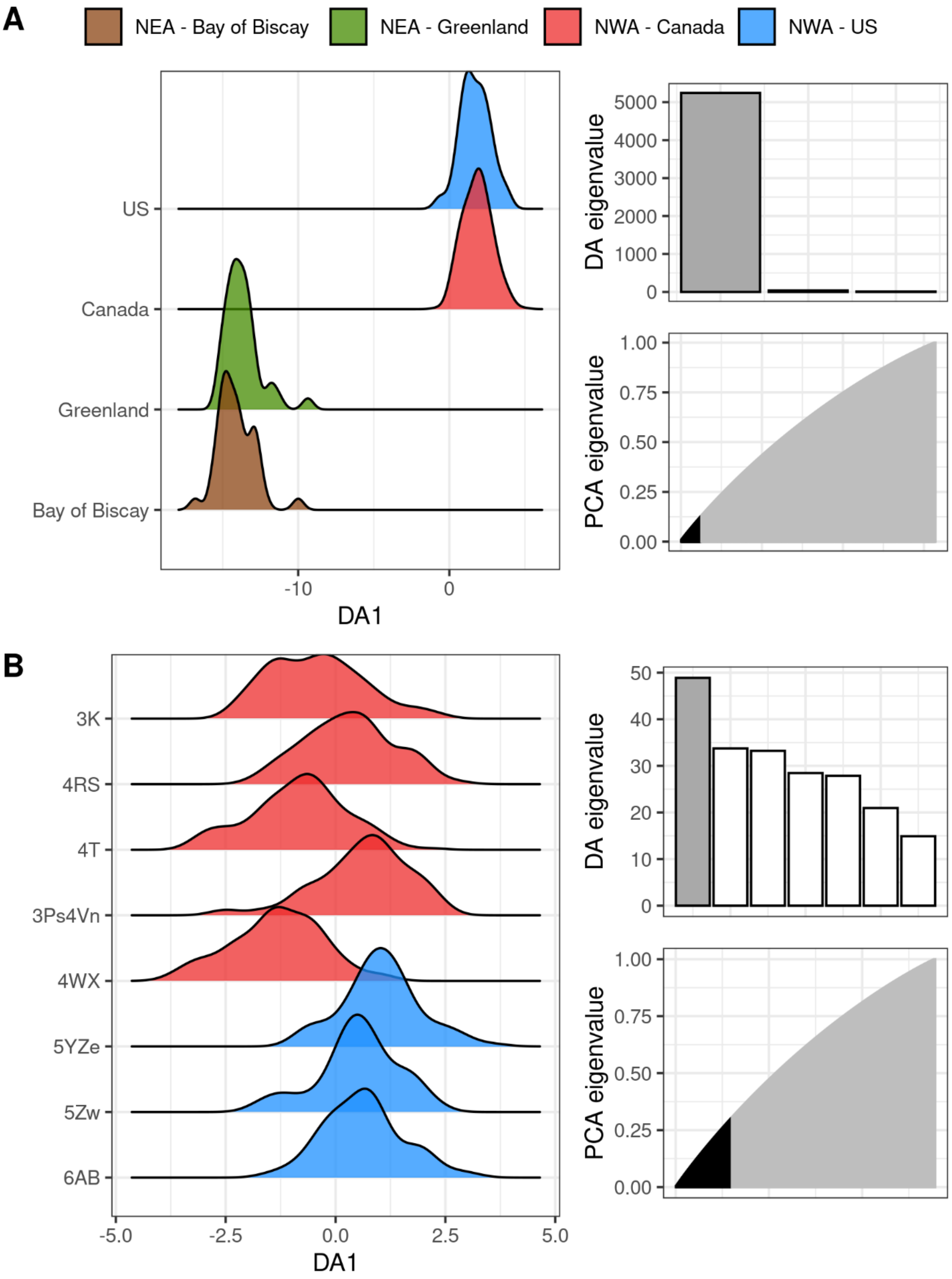

FIG. S4

Discriminant function analysis results (DAPC) for A) NEA-NWA SNPs panel with an a priori clustering by country and B) NWA SNPs panel with an a priori clustering by NAFO division. In both cases, only the first discriminant function (DA1) is presented as a density plot. Barplots of eigenvalues for the discriminant analysis and of cumulated variance explained by the PC are also presented. The optimal numbers of PCs retained were 50 and 120, respectively. Note that reference samples are included in their respective NAFO divisions (see Table 1).

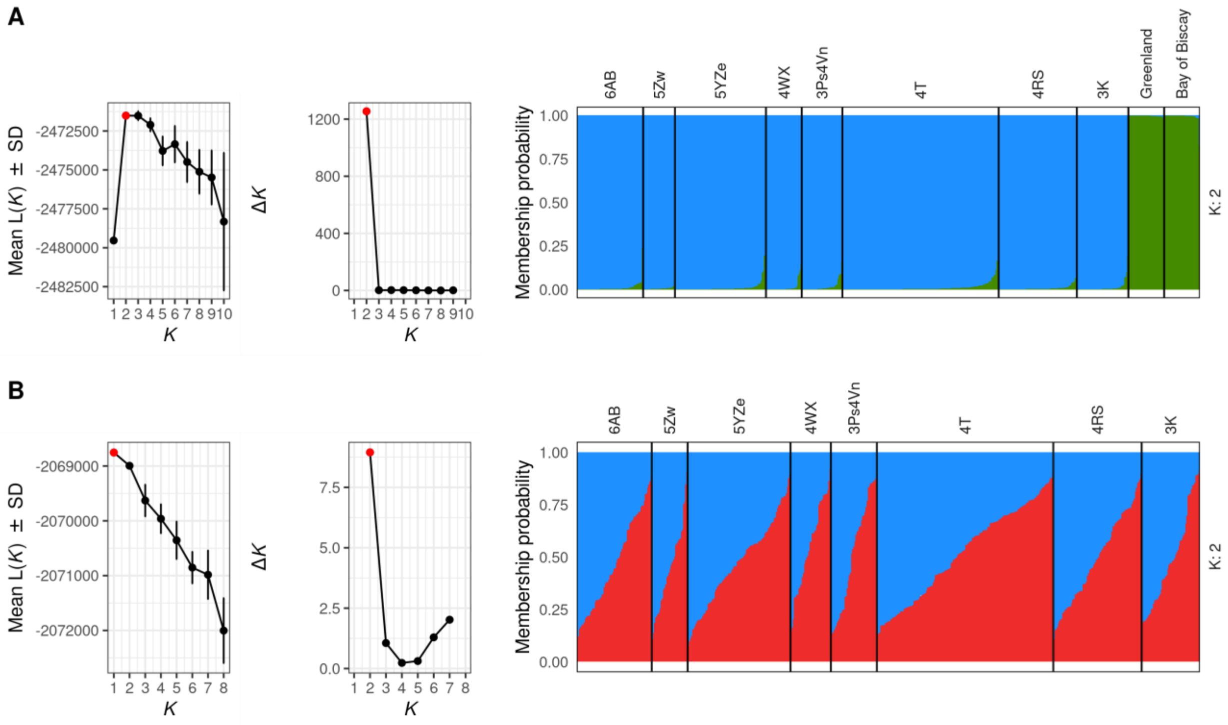

Fig. S5

Bayesian clustering result using Structure (Pritchard et al. 2000) for A) NEA-NWA SNPs panel for *K* 1 to 10 and B) NWA 0.1 SNPs panel for K 1 to 8. The Pritchard mean L(K) and Evanno ΔK critera are presented, with the best *K* value indicated in red. Note that *K* = 2 was the best K value under the Evanno criterion for the NWA panel, but that this method cannot reject *K* = 1 and that no relevant biological pattern was detected with *K* = 2.

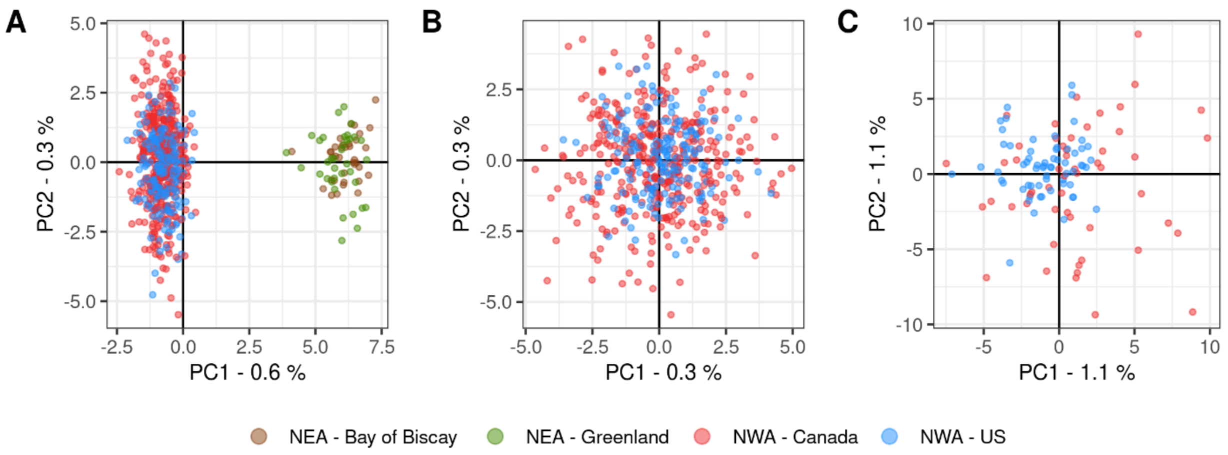

FIG. S6

Principal component analysis (PCAs) for A) all samples (NEA-NWA) with neutral SNPs only (10,608 SNPs), B) NWA samples with neutral SNPs only (10,628 SNPs) and C) reference samples with neutral SNPs only (10,628 SNPs). Each dot represents one sample and the color indicates the sampling country. The percentage of genetic variance explained by each axis is in parentheses.

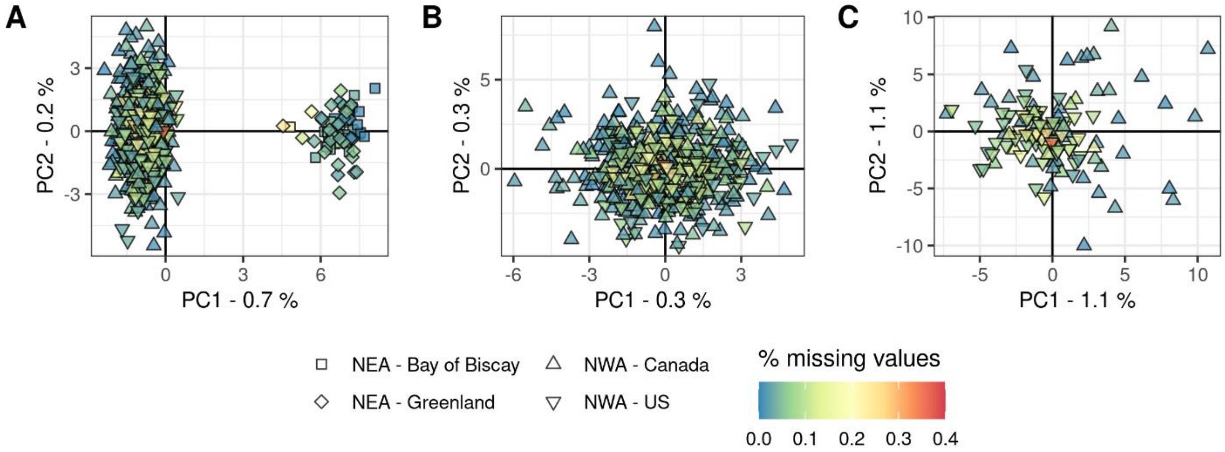

FIG. S7

Principal component analysis (PCAs) for A) all samples (NEA-NWA), B) NWA samples and C) reference samples only. Each dot represents one sample, with shape representing the sampling country and color the percentage of individual missingness. The percentage of genetic variance explained by each axis is in parentheses.

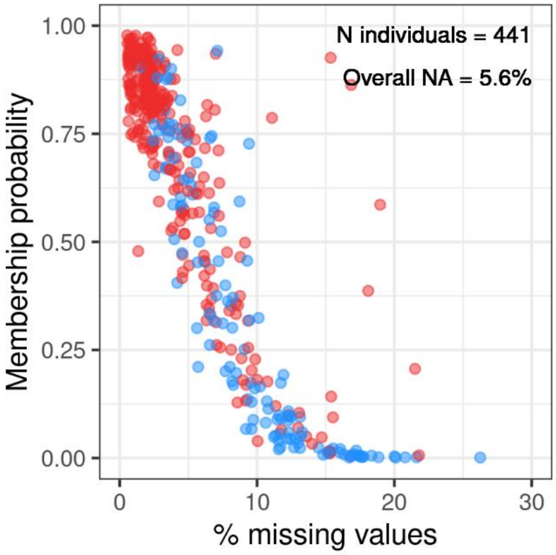

FIG. S8

Relationship between the estimated membership probability to the northern contingent for all non-reference samples and the percentage of missing values. Each dot represents one sample and the color indicates the sampling country (red = Canada, blue = U.S.).

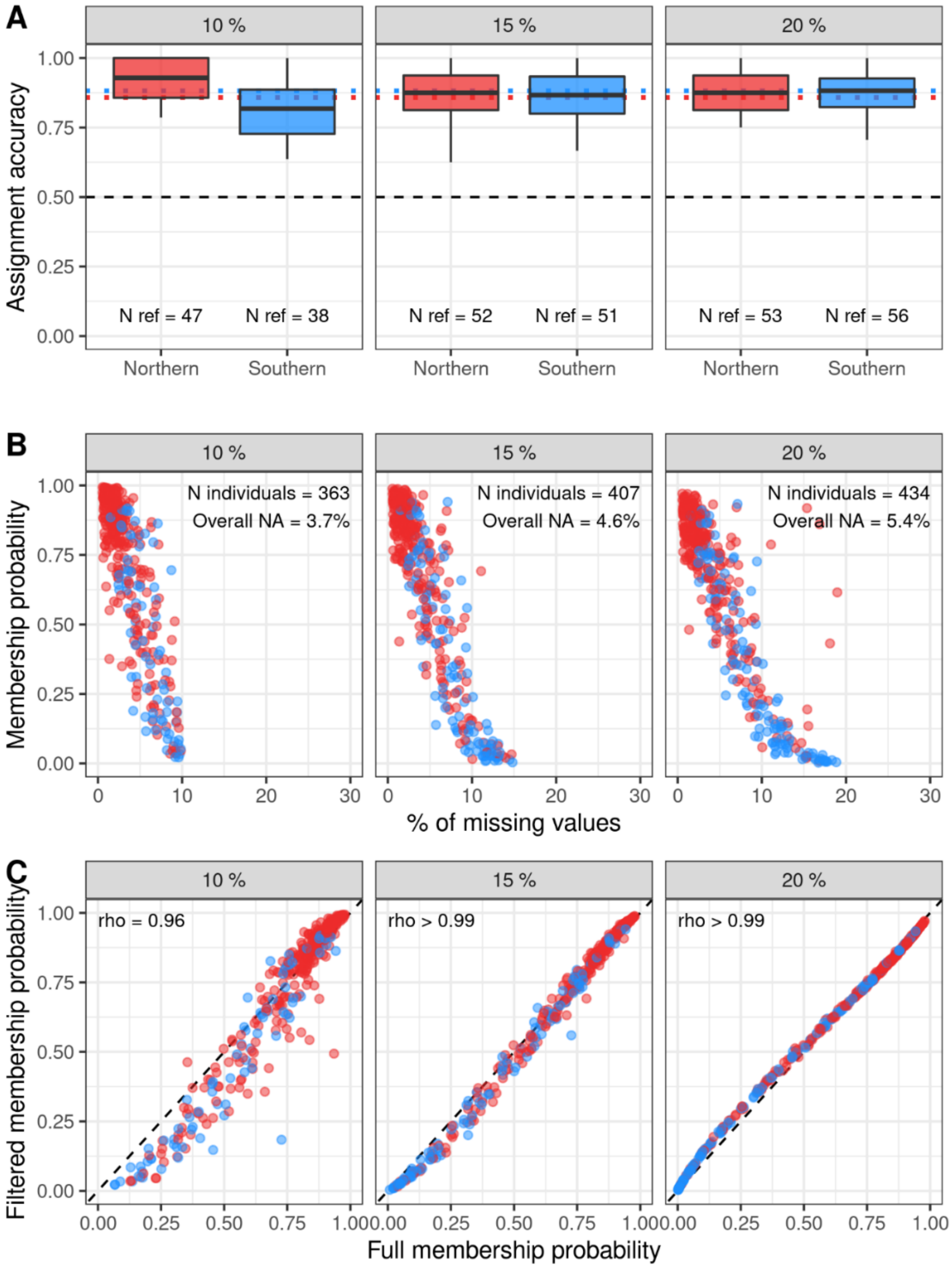

FIG. S9

Effect of missing data on genetic assignment with different filtering parameters (10%, 15% or 20% upper threshold of missing data per individual). A) Cross-validation results using truncated reference samples, using SVM and all SNPs. The dotted colored lines represent the observed assignment accuracy with the full dataset. B) Relationship between the estimated membership probability to the northern contingent for all non-reference samples and the percentage of missing values. C) Correlation between membership probability estimated with the full or filtered dataset. Each dot represents one sample and the color indicates the sampling country (red = Canada, blue = U.S.).

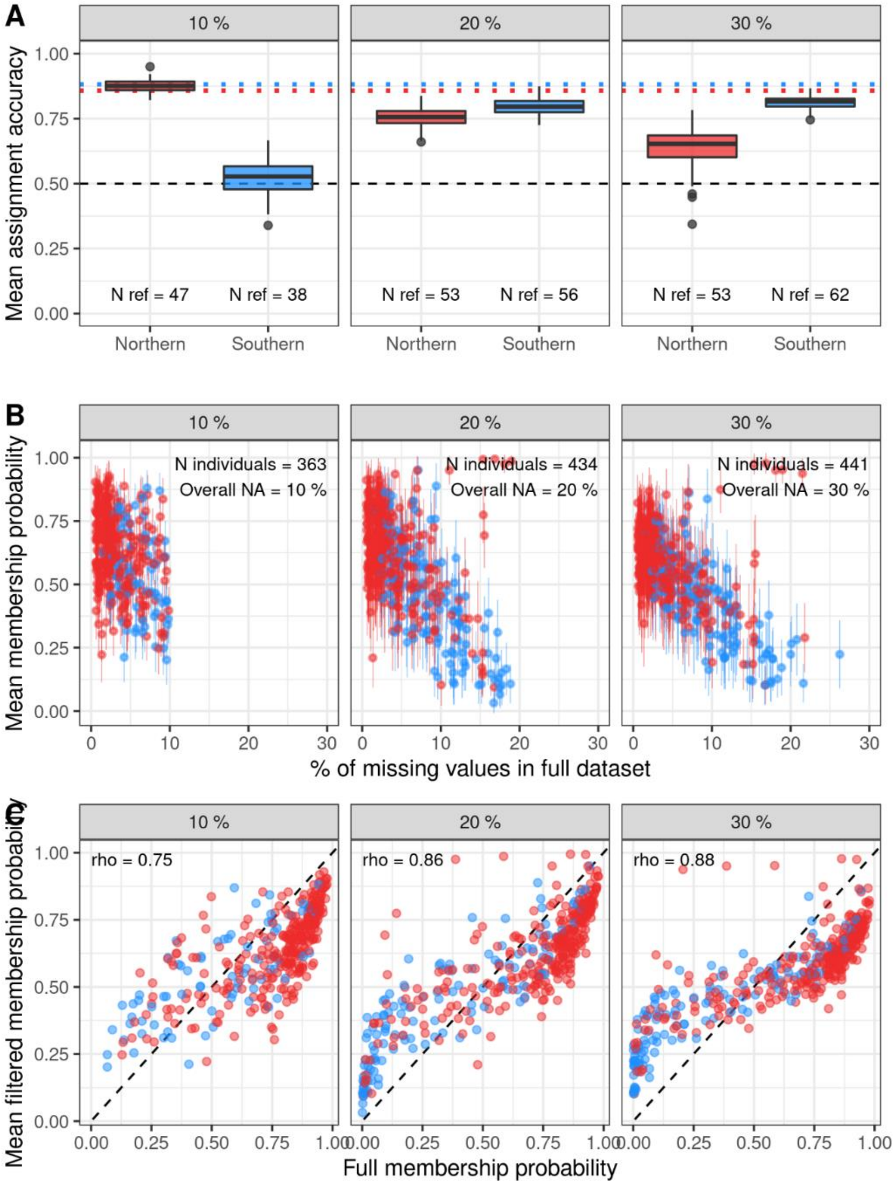

FIG. S10

Effect of missing data on genetic assignment with simulations of 100 dataset with both imputation of missing value to reach a higher threshold and removal of samples above this threshold (10, 20 or 30% threshold for per individual and overall missing data). A) Mean cross-validation results with the 100 simulated reference sample dataset, using SVM and all SNPs. The dotted colored lines represent the observed assignment accuracy with the full dataset. B) The relationship between the mean (± sd) estimated membership probability to the northern contingent for all non-reference samples and the percentage of missing values. C) Correlation between the mean membership probability to the northern contingent from the simulation datasets and the full dataset Each dot represents one sample and the color indicates the sampling country (red = Canada, blue = U.S.).

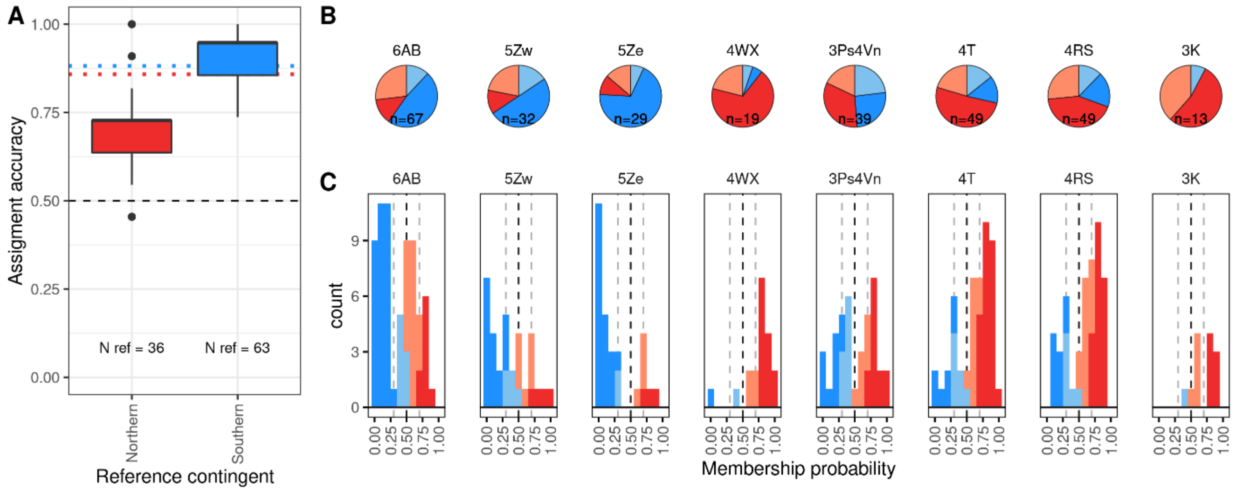

FIG. S11

Cross-validation and genetic assignment results using a reduced dataset including only samples sequenced under the NovaSeq technology (n = 396 individuals, including 36 and 63 northern and southern references, respectively). A) Cross-validation using northern and southern contingent reference samples and the best predictive model (support vector machine, SVM). The dotted colored lines represent the observed assignment accuracy with the full dataset. B) Composition estimates in each NAFO division derived from genetic assignment results for all non-reference samples C) Distribution of individual membership probabilities to the northern contingent, with dashed lines representing thresholds of membership probability. Individuals from the duplicated library between both sequencing technologies were added to cover NAFO division otherwise not represented (4WX and 3K, see Fig. S3). The color represents the membership to the northern (red) or southern (blue) contingents, with indication of membership probability to contingent (light colors: >50%; dark colors: >70%).