MOLECULAR ECOLOGY

Supplementary Information for

Landscape genomics of the American lobster (*Homarus americanus*)

Yann Dorant[1,2], Martin Laporte[1,3], Quentin Rougemont[1,4], Hugo Cayuela[1,5], Rémy Rochette[6] and Louis Bernatchez[1]

[1]Institut de Biologie Intégrative des Systèmes (IBIS), Université Laval, Québec, QC, Canada, G1V0A6
[2]IHPE, Univ. Montpellier, CNRS, Ifremer, Univ. Perpignan Via Domitia, Montpellier, France
[3]Ministère des Forêts de la Faune et des Parcs du Québec, Canada
[4]CEFE, Univ Montpellier, CNRS, EPHE, IRD, Montpellier, France
[5]Laboratoire de Biométrie et Biologie Évolutive, Université Lyon 1, CNRS, UMR 5558, Villeurbanne F-10769622, France
[6]Department of Biology, University of New Brunswick, P.O. Box 5050, Saint John, NB, Canada, E2L 4L5

This PDF file includes:

SI Material and Methods
SI References
Tables S2 to S5
Figures S1 to S13

Other supplementary material for this manuscript includes the following:

Table S1. Sampling information. LFA, Ns, Ho and He refer to Lobster Fishing Areas, number of individuals per sampling site, observed heterozygosity and expected heterozygosity respectively.

27 # Table of Contents:

SI Material and Methods

SNP filtering procedure

All scripts (code and description) used for the SNP data filtering procedure are available at https://github.com/enormandeau/stacks_workflow. First, we filtered the raw VCF file keeping only genotypes showing a minimum depth of four (parameter "m" hereafter) and called in at least 50% of the samples in each site (parameter "p" hereafter) and a minimum number of different individuals possessing the minor allele of two (parameter "S" hereafter) using the 05_filter_vcf_fast.py available in *stacks_workflow*. Here, the latter filter parameter applied is akin to minor allele frequency (MAF) filtering with the difference that it is not artificially boosted by genotyping errors which can occur where one heterozygous sample is erroneously genotyped as a rare-allele homozygote. We then removed individuals showing more than 15% of missing data. We also filtered out individuals showing putative DNA contamination using two parameters. First the relatedness between pairs of individuals was estimated following the equation proposed by Yang et al. (2010) and implemented in *vcftools*. While a relatedness coefficient of 0.5 is expected to represent full-siblings, high value of relatedness between two different individuals may represent identical twins or clones, which is not expected in the study species here. Hence, for each case where a pair of individuals exhibited a relatedness value > 0.9, the individual that showed the highest value of missing data was removed from the whole dataset. Second, the inbreeding coefficient (F_{IS}) was estimated for each individuals using the method of moment implemented in *vcftools*. Based on a graphical observation of individuals inbreeding, we defined a cutoff value (i.e. -0.25) to exclude outliers showing extreme values of F_{IS}. After removing individuals showing putative DNA contamination from the raw vcf file, we re-ran the 05_filter_vcf_fast.py from *stack_workflow*, keeping the same parameters previously used (i.e. m=4; p=50; S=2). The resulting filtered VCF file had a 98% genotype call rate across 4,190 individuals (with maximum allowed missing loci per individual of 15%) and a SNPs median read depth of 25X across all samples.

SNPs classification (non-duplicated vs. duplicated)

Following the low-filtering steps described above, we discriminated "non-duplicated" SNPs (i.e. non-paralogous) from "duplicated" SNPs using the same approach proposed in Dorant et al (2020). Briefly, this approach aims to distinct non-duplicated vs. duplicated SNPs using four parameters, the median of allele ratio in heterozygotes (MedRatio), the proportion of heterozygotes (PropHet), proportion of rare homozygotes (PropHomRare) and Inbreeding coefficient (F_{IS}). Each parameter was calculated from the filtered VCF file using the 08_extract_snp_duplication_info.py available in *stacks_workflow*. Individual values of the four parameters were plotted pairwise to visualize their distribution across all SNPs. Based on the graphical demonstration proposed by McKinney et al (2017) and Dorant et al., (2020), we considered different combinations of each parameter and graphically set the cut-off of the four categories of SNPs (i.e. non-duplicated, duplicated, high coverage and low confidence) (Fig. S1). Non-duplicated SNPs accompanied with duplicated or diverged SNP on the same 80 bp locus were not considered and removed. Finally, we only retained non-duplicated SNPs for downstream analyses. Finally, we retained all unlinked SNPs within each locus using the *11_extract_unlinked_snps.py* available in *stacks_workflow*. Briefly, the first SNP is kept and all remaining SNPs showing strong genotype correlation are pruned (i.e. two SNPs show strong genotype correlation if samples with the minor allele in one of the SNPs have the same

genotypes as samples with the minor allele in the other SNP more than 50% of the time). The procedure was repeated until all SNPs were either kept or pruned.

Correction for pattern of missingness (non-duplicated).

Missing genotypes can introduce patterns of similarity or differentiation that are able to be confounded with population structure. To detect such bias, we investigated the Identity-By-Missingness (IBM) distance across each pair of individuals, which represent the proportion of missing sites which are not shared between a pair of individuals. IBM was calculated using the program PLINK v.1.9 (Purcell et al., 2007). This information can be used to detect and correct for population stratification that could be shared across unrelated individuals due to identical missing data.

Pairwise IBM distances calculated across 4,190 individuals genotyped over 19,868 filtered SNPs were then visualized with a multidimensional scaling (hereafter MDS) approach. Graphical examination of the MDS scatter plot showed a dichotomic stratification among the 17 sequencing lanes (Fig. S2). Here, based on the second dimension of the MDS scatter plot, we defined two distinct groups of sequencing lanes, so-called sequencing batches hereafter (Fig. S2B). We then investigated the magnitude of the absolute difference of missing genotype proportion between the two sequencing batches for each SNP (i.e for a given SNP, the proportion of missing genotypes from the first batch minus the proportion of missing genotypes from the second batch). Herein, we expect that SNPs exhibiting the highest difference in terms of missing data between the two sequencing batches are those that mostly drive the pattern of missingness. Hence, removing them will enable us to correct any pattern of structuration caused by missing data.

Considering the distribution of absolute difference in terms of missing data between the two sequencing batches, we tested the effect of removing various sets of SNPs on the IBM pattern, according to four cut-off values defined by the quantiles 0.99, 0.95, 0.75 and 0.5 (Fig. S3). The number of SNPs filtered out was 200, 998, 4975 and 9945 for cut-off values of 0.99, 0.95, 0.75 and 0.50, respectively. The IBM was then calculated for each pruned dataset and visualized using the same MDS approach. We observed that pruning SNPs based on the quantile values of 0.75 and 0.50 were efficient to correct the missing pattern in our data (Fig. S4). Finally, we selected the quantile value of 0.75 to correct our dataset as it allows to minimize the SNP loosing.

SI References

DFO. (2016). Update of the stock status indicators for the American lobster (*Homarus americanus*) stocks in the southern Gulf of St. Lawrence. *DFO Can. Sci. Advis. Sec. Sci. Rep*, (2016/051).

Dorant, Y., Cayuela, H., Wellband, K., Laporte, M., Rougemont, Q., Mérot, C., … Bernatchez, L. (2020). Copy number variants outperform SNPs to reveal genotype–temperature association in a marine species. *Molecular Ecology*, (29), 4765–4782. doi: 10.1111/mec.15565

Lanteigne, M., Comeau, M., Mallet, M., Robichaud, G., & Savoie, F. (1998). *The American lobster, Homarus americanus, in the southern Gulf of St. Lawrence* (No. 1998/123). DFO Can. Sci. Advis. Sec. Res. Doc.

McKinney, G. J., Waples, R. K., Seeb, L. W., & Seeb, J. E. (2017). Paralogs are revealed by proportion of heterozygotes and deviations in read ratios in genotyping-by-sequencing data from natural populations. *Molecular Ecology Resources*, *17*(4), 656–669. doi: 10.1111/1755-0998.12613

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., … Sham, P. C. (2007). PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *The American Journal of Human Genetics*, *81*(3), 559–575. doi: 10.1086/519795

Yang, J., Benyamin, B., McEvoy, B. P., Gordon, S., Henders, A. K., Nyholt, D. R., … Visscher, P. M. (2010). Common SNPs explain a large proportion of the heritability for human height. *Nature Genetics*, *42*(7), 565–569. doi: 10.1038/ng.608

Supplementary tables

Table S2. Summary of data filtering procedure. After each filtering steps, remaing SNPs or individuals are provided. Eliminated SNPs and individuals are given between brackets. The filtering parameter MAS (*), which is analogous to MAF or MAC filters, refer to the minimum number of different individuals possessing the minor allele to retain the given SNP.

Filtering step	Number of remaining SNPs (eliminated SNPs)	Number of remaining individuals (elimitated individuals)
Stacks raw vcf	76,863	4,400
SNPs filter: - genotype >4X - SNP call in a least 50% of the samples in each site - MAS* ≥ 2	42,965 (33,898)	4,400 (0)
Missing data ≤ 15%	42,965 (0)	4,307 (93)
Putative DNA contamination - relatedness < 0.9 - FIS > -0.25	42,965 (0)	4,190 (117)
Non-duplicated SNPs	22,159 (20,806)	4,190 (0)
Unlinked SNPs	19,868 (938)	4,190 (0)
Patterns of identity-by-missingness	14,893 (4,975)	4,190 (0)

Table S3. SNPs classification results.

	non-duplicated	Duplicated	Low confidence	Diverged	MAS
SNPs	22,159	5,229	234	919	6,1112
Loci	6,236	1,204	119	309	3,759

Table S4. Hierarchical analysis of molecular variance (AMOVA) performed with the 12 sampling locations for which temporal replicates were available.

DF, Degree of freedom; *MSD*, mean squared deviation, ϕ provides the « Phi » population differentiation statistics. These are used to test hypotheses about population differentiation.

Source	DF	MSD	Variance component	% of variation	ϕ	Pvalue
Outliers						
Between regions (north vs. south)	1	413.71	0.693	1.187%	0.0118	0.01
Between sites within regions	10	61.51	0.035	0.060%	0.0006	0.16
Between years within sites	12	58.50	0.021	0.036%	0.0004	0.27
Within samples	994	57.62	57.623	98.716%	-	0.01
Total variations	1017	58.02	28.372	100%	0.0128	-
Neutral						
Between regions (north vs. south)	1	743.19	0.584	0.134%	0.0013	0.01
Between sites within regions	10	446.04	0.169	0.039%	0.0004	0.01
Between years within sites	12	431.79	0	0%	0	0.73
Within samples	994	434.37	434.372	99.848%	-	0.01
Total variations	1017	434.76	435.064	100%	0.0015	-

126

127

Table S5. Inferred parameter estimates of lobster demography under Isolation-with-migration (IMAG) model obtained from *dadi*.

Note: AIC, Akaike's information criterion; log likelihood, maximum likelihood; theta, effective mutation rate of the ancestral population; NRef, size of the ancestral population; Ne_south & Ne_north, effective population size of the compared pair just after the split event; m1 ← 2 and m2 ← 1, migration from population north to population south and vice versa; Tsplit, time of split of the ancestral population in the two species; T_ancestralpopChange = time of the start of ancestral population size change; Growth_south and Growth_north correspond to the efficient of growth of the two populations, which starts at the split time; Ne_south*growth represent the "contemporary" effective population size of the southern population at the end of model run, taking into account population dynamic (growth rate) across generations (same for north population).

		Model IMAG
Assumed μ		$1e^{-08}$
nLoci		4,340
Length		347,200
Optimized log-likelihood		-1,829.04
AIC		3,678.08
Untransformed parameters		
theta		243.388
NRef	**prior**	17,525.055
Ne_ancestral	[0.0001-100]	3.133
Growth_populationAncestral	[0.0001-100]	4.133
Ne_south	[0.0001-100]	1.073
Ne_north	[0.0001-100]	0.684
Growth_south	[0.0001-100]	14.997
Growth_north	[0.0001-100]	10.042
$M_{south \leftarrow north}$	[0.001-50]	46.967
$M_{north \leftarrow south}$	[0.001-50]	24.767
T_ancestralpopChange	[0.001-10]	4.198
Tsplit	[0.001-10]	0.052
Biological parameters		
Ne_ancestral		54,899
Ne_south		18,800
Ne_north		11,983
Ne_ancestral*growth		226,917
Ne_south*growth		281,937
Ne_north*growth		120,336
$m_{south \leftarrow north}$		0.00134
$m_{north \leftarrow south}$		0.00071
TimeOfpopulationChange		147,146
SplitTime		1,806

Supplementary Figures

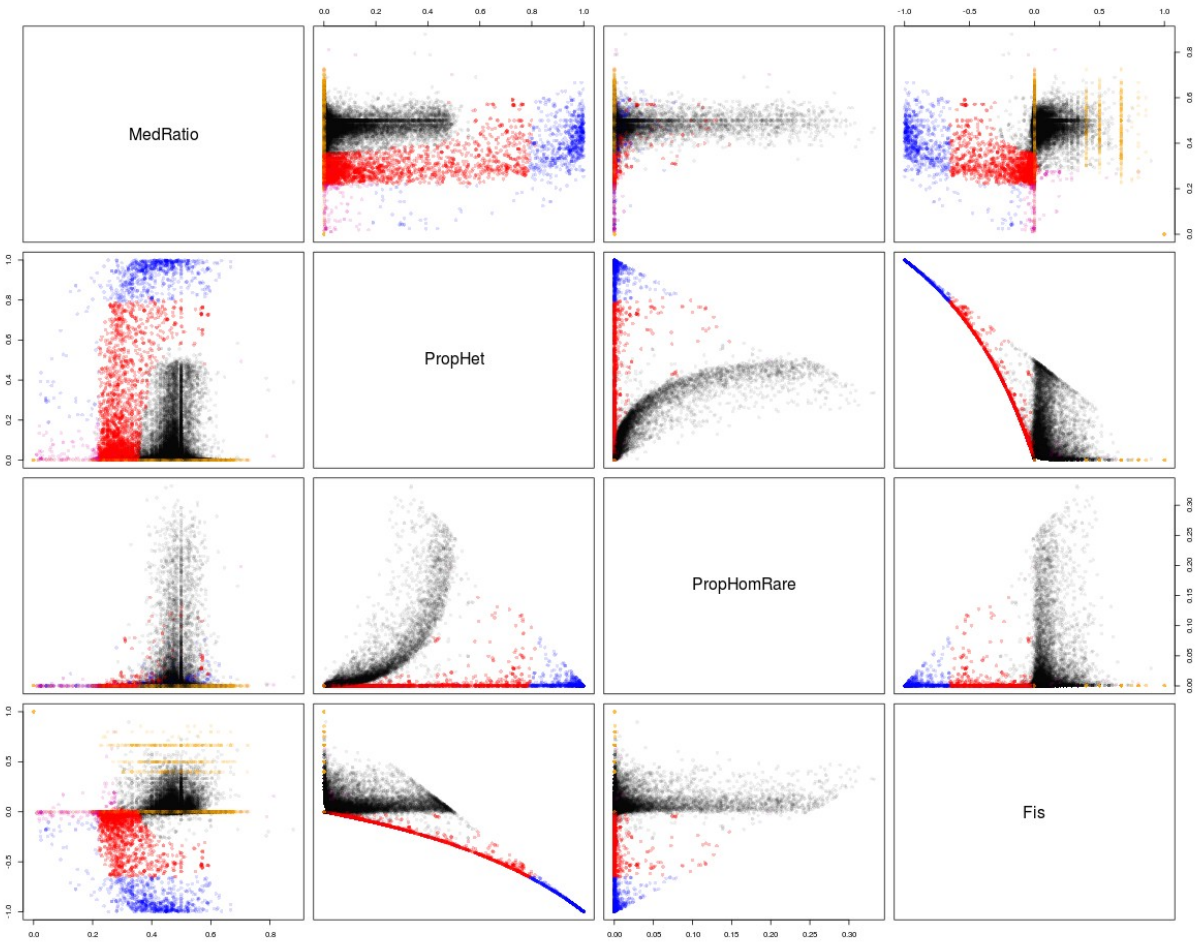

Figure S1. Characterization of duplication effect over the SNP dataset.

The bivariate scatter plots display the distribution of the 33,898 SNPs over four statistical parameters measured from the filtered VCF (i.e. (1) median of read allele ratio in heterozygotes (MedRatio), (2) proportion of heterozygotes (PropHet), (3) Proportion of rare homozygotes and (4) Fis). Based on the graphical patterns of SNPs categories (i.e. non-duplicated, duplicated, diverged) demonstrated by McKinney et al. (2017) with data simulations as well as empirical analyses, we fixed different cutoff values for each parameters displayed (detailed of the cut-off values are reported in an R script provided in the Dryad published data). Black, red, blue purple and orange points represent non-duplicated, duplicated, diverged, MAS and low confidence SNPs, respectively.

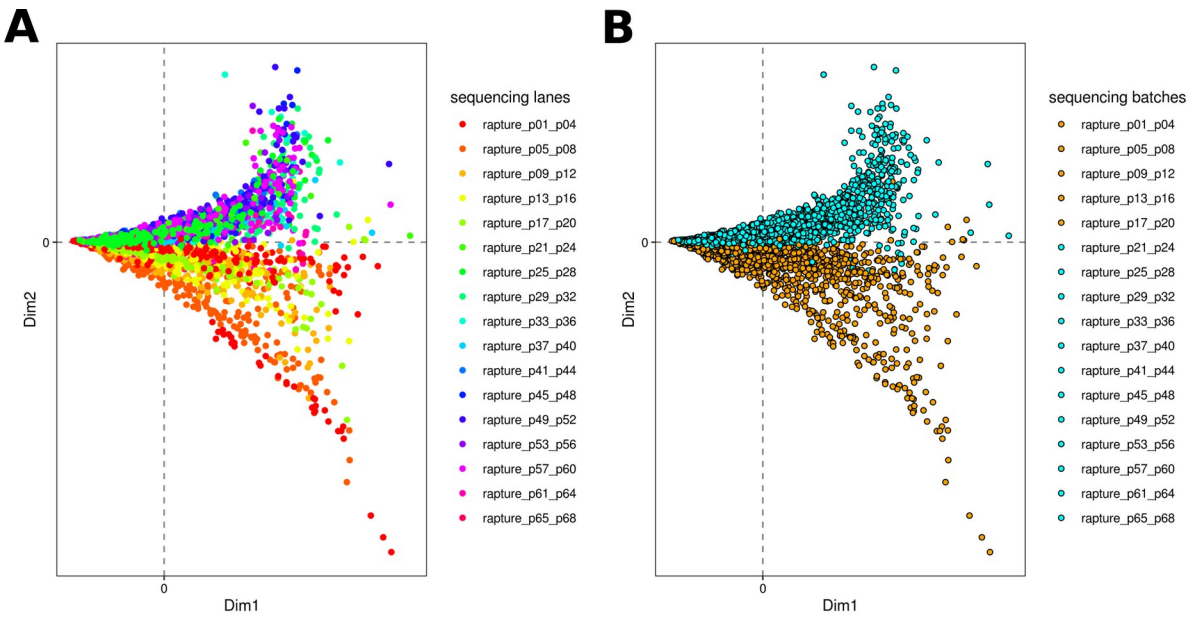

Figure S2. MDS analysis of identity-by-missingness (IBM) patterns calculated in *PLINK*.

First (x-axis) and second (y-axis) dimensions of 4,401 lobsters (96 sampling sites), distributed across 17 sequencing lanes and based on IBM analysis over 19,868 SNPs. (*A*) MDS scatter plot where each point represents a sample colored according to its sequencing lane membership. Two clusters of sequencing lanes where visually identified on the second dimension considering the tail direction of sample distribution for each sequencing lane (*B*) MDS scatter plot where each point represent a sample colored according to its sequencing batch membership.

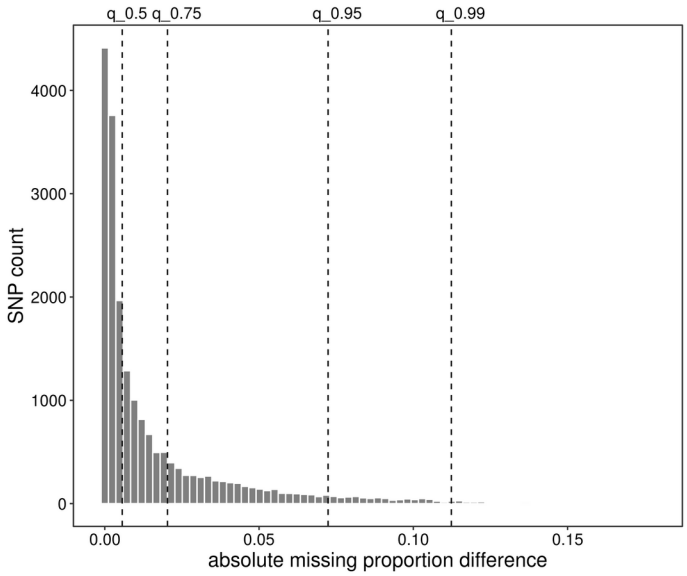

Figure S3. Distribution of the absolute difference in terms of missing data proportions between the two sequencing batches.

Vertical dotted lines represent the position of each cut-off value (i.e. quantiles 0.5, 0.75, 0.95 and 0.99), used to remove sets of SNPs showing extended degree of missing data between the two sequencing batches.

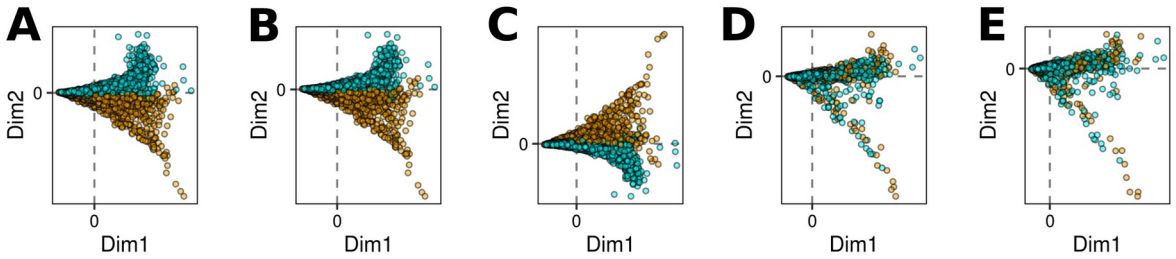

Figure S4. Correction of missing pattern.

Each scatter plot represents a *PLINK* MDS analysis of IBM pattern, where samples are coloured according to their missing-cluster membership. (*A*) MDS conducted across all SNPs (i.e. 19,868 SNPs). (*B*),(*C*),(*D*) and (*E*) represent MDS analyses conducted after removing SNPs based on the quantile filters 0.99, 0.95, 0.75 and 0.50, respectively.

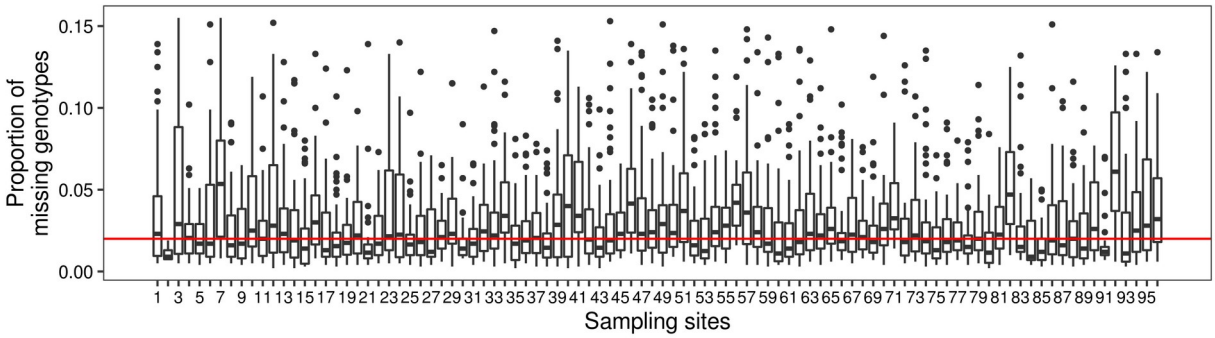

Figure S5. Proportion of missing data in the SNP non-duplicated filtered dataset.

Each boxplot represent the distribution of missing data within each sampling site where inbox horizontal line represent the average value of missing data per site. The horizontal red line represent the median value of missing data across all samples (i.e. 2%). Black dots represent outliers individuals. Note that only odd sites identification have been displayed for a better representation.

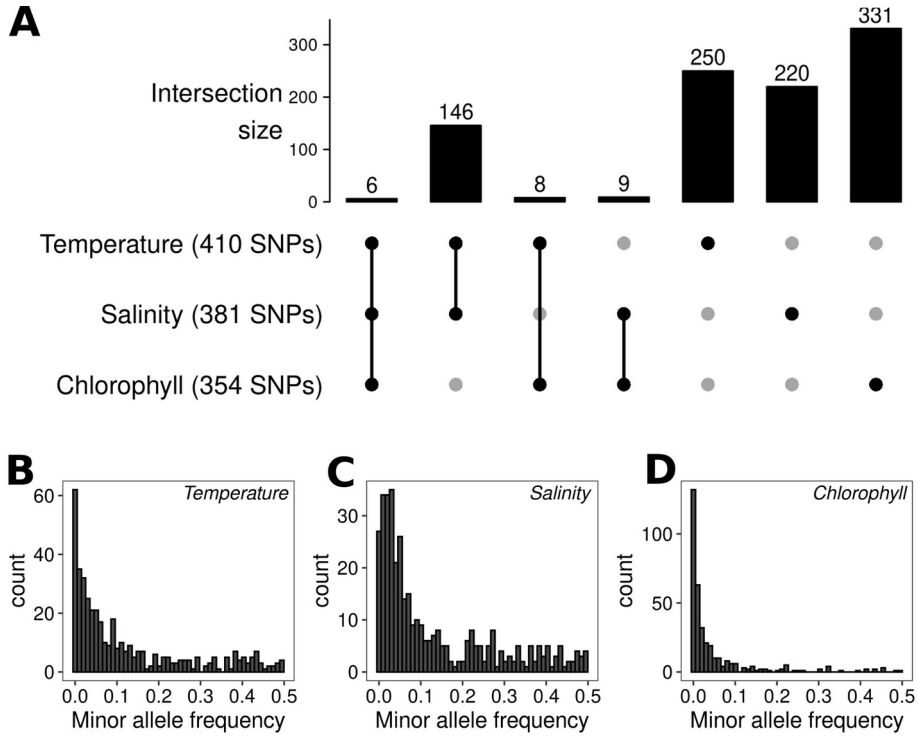

Figure S6. Summary of outlier detection associated with environmental variables.

(*A*) Upper plot summarizing RDA outlier detection (intersections and distinct sets). Total number of outliers detected for each environmental variable is given between brackets. (*B*, *C* and *D*). Minor allele frequency distributions of outlier sets for sea surface temperature, sea surface salinity and sea surface chlorophyll respectively.

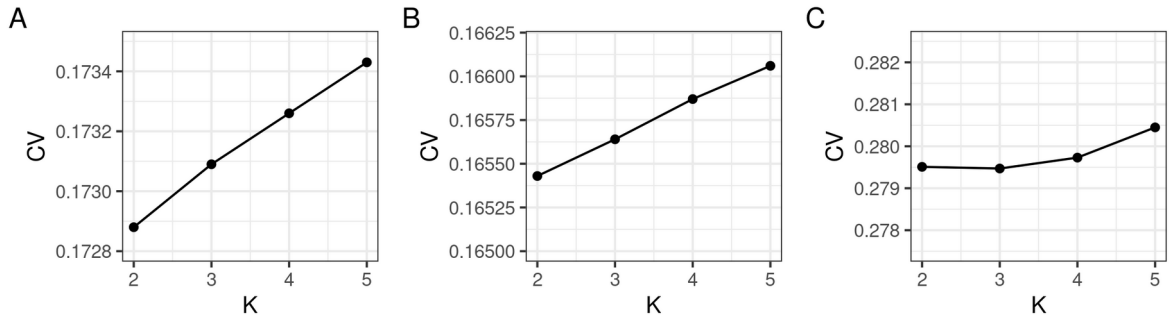

Figure S7. CV error for ADMIXTURE analysis of 96 lobster sampling sites.

K values ranged from 2 to 5. K=2 is best. (*A*) All SNP dataset, (*B*) Neutral SNP dataset and (C) Outliers SNPs dataset.

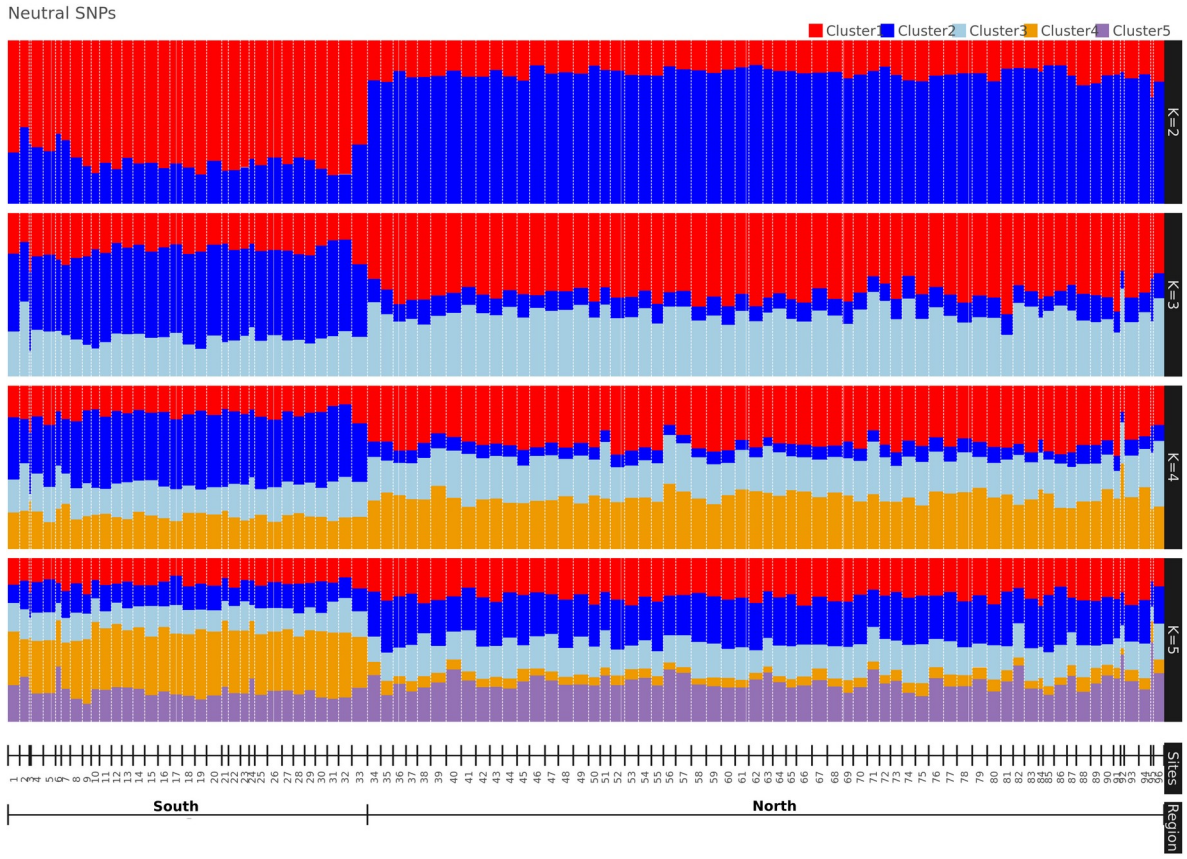

Figure S8. Population Admixture analysis of the 96 *H. americanus* populations based on 13,912 neutral SNPs. Each color bar represents the posterior estimates of each K cluster (K from 2 to 5) averaged by sampling site.

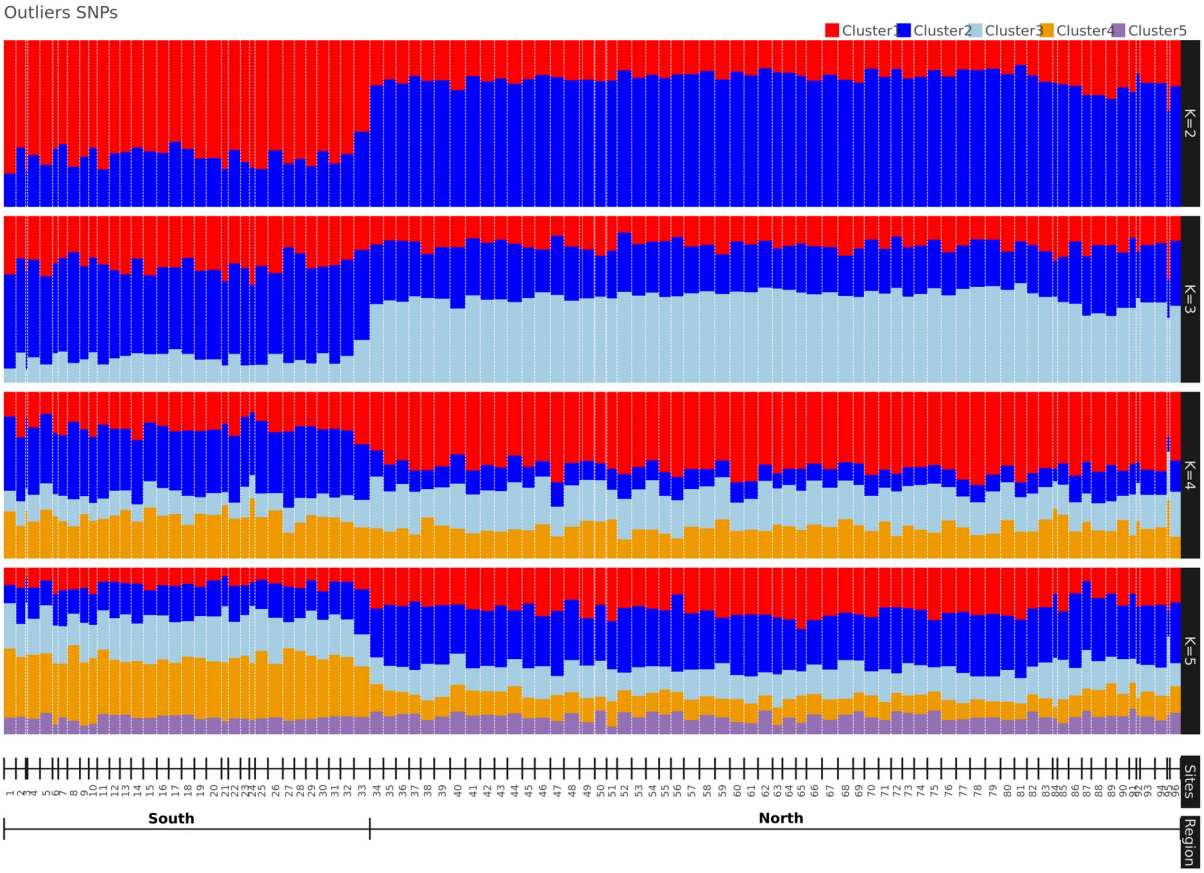

Figure S9. Population Admixture analysis of the 96 *H. americanus* populations based on 981 outliers SNPs. Each color bar represents the posterior estimates of each K cluster (K from 2 to 5) averaged by sampling site.

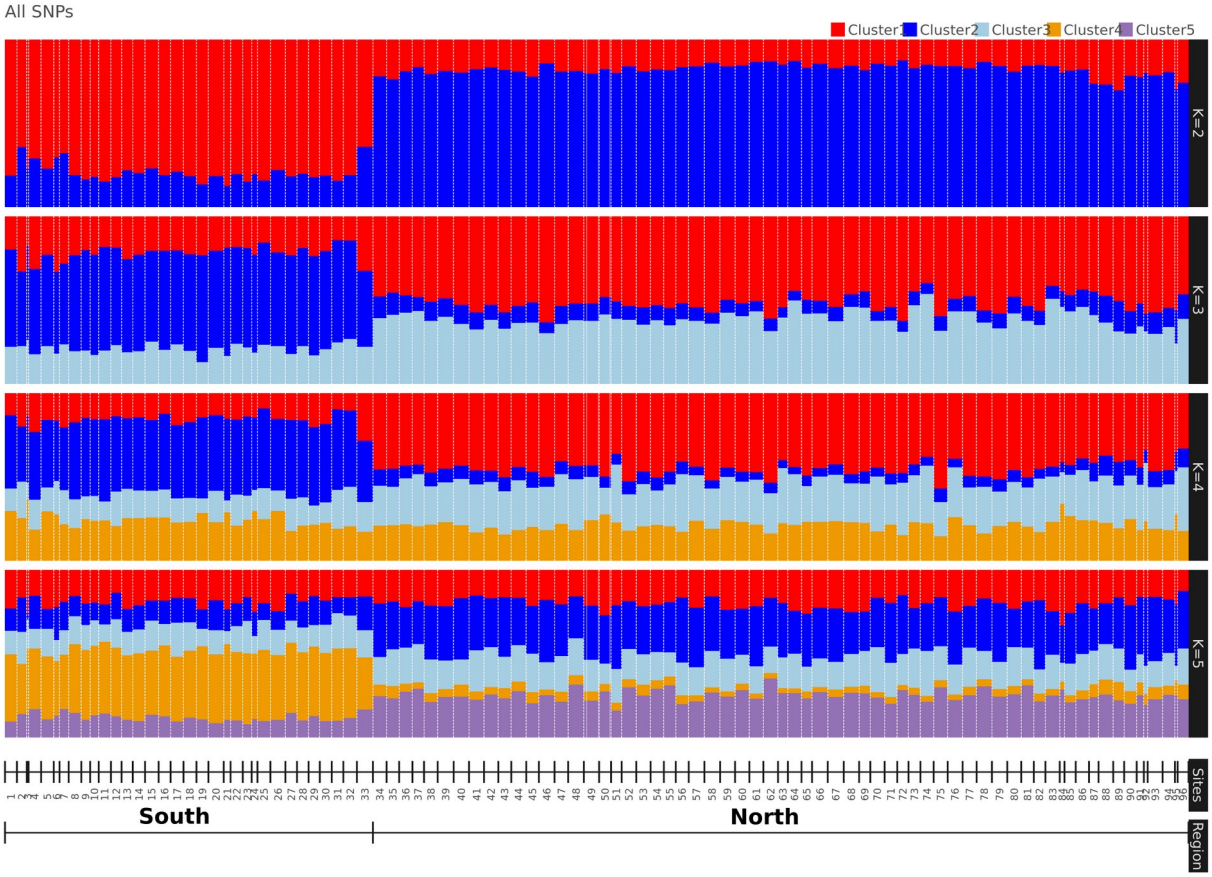

Figure S10. Population Admixture analysis of the 96 *H. americanus* populations based on the combined SNP dataset (14,893 SNPs). Each color bar represents the posterior estimates of each K cluster (K from 2 to 5) averaged by sampling site.

142

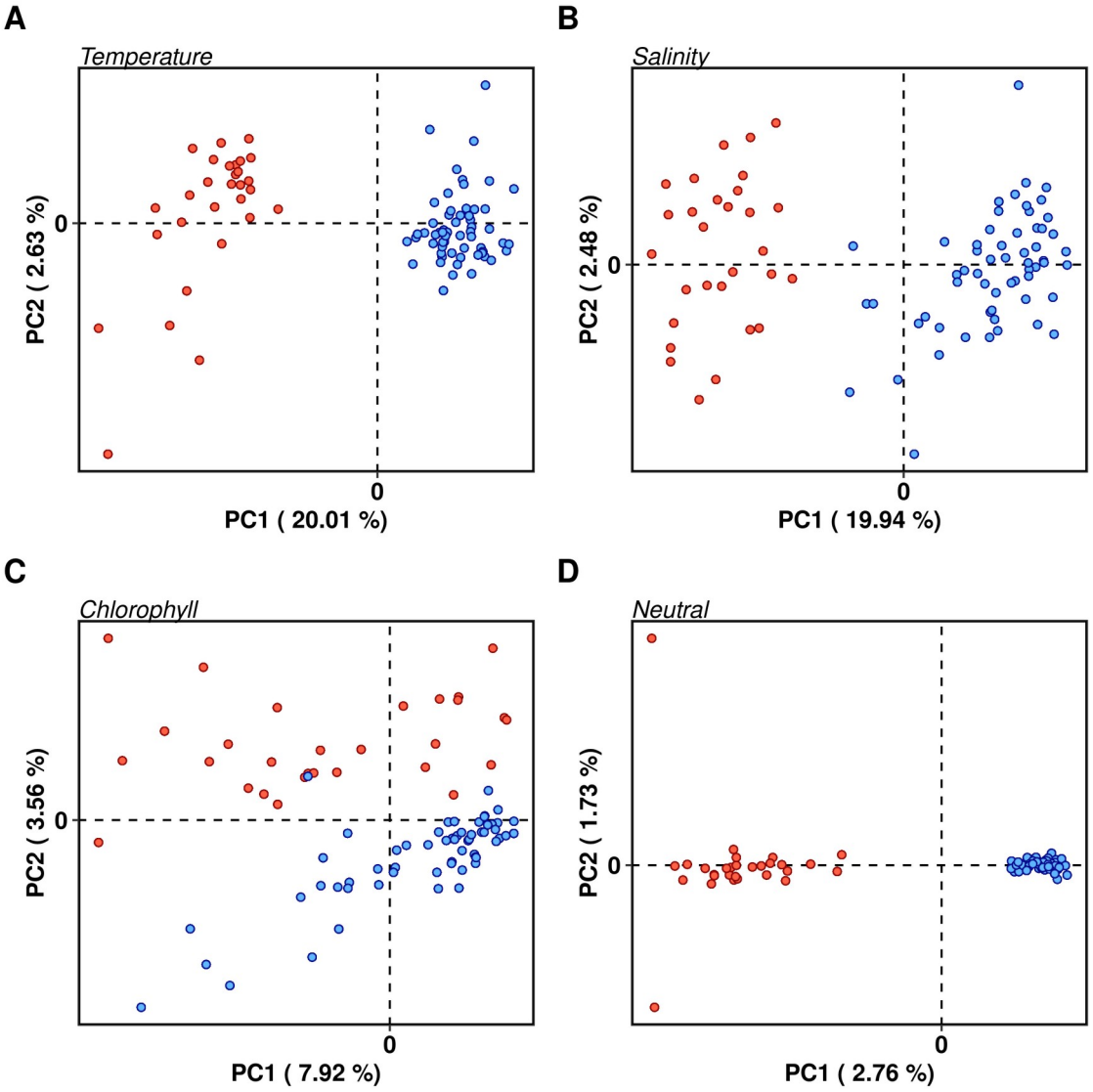

Figure S11. Large scale population clustering based on putative adaptive markers associated with environmental variables.

The scatterplots represent four different Principal Component Analysis (PCA) computed from allele frequencies using four sets of SNPs. (*A*) Candidate SNPs associated with sea surface salinity (424 SNPs). (*B*) Candidate SNPs associated with sea surface temperature (403 SNPs). (*C*) Candidate SNPs associated with sea surface chlorophyll concentration (376 SNPs). (*D*) Set of putative neutral SNPs (13,879 SNPs). Each dot represents one sampling site colored according to North (blue) and South (red) regions identified in Fig. 4.

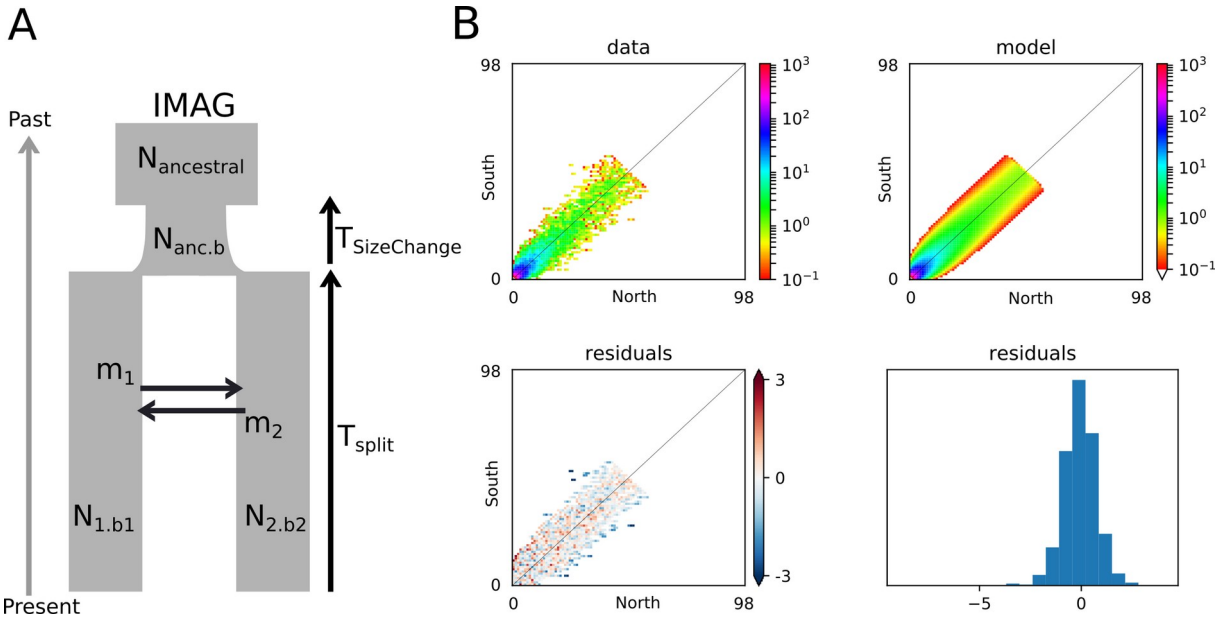

Figure S12. Demographic model with the highest log-likelihood obtained from the joint site frequency spectrum (jSFS), inferred by dadi for large scale genetic structure of the American lobster.

(*A*) Schema of the demographic model (IMAG) inferred for the northern and southern populations structure. (*B*) jSFS of lobster data (upper-left). jSFS of the demographic model (upper-right). Anscombe residuals between empirical data and model where colored cells inform about model prediction deviation (i.e. blue and red for reduced and increased polymorphism respectively)(bottom-left). Histogram of the residuals distribution (bottom-right).

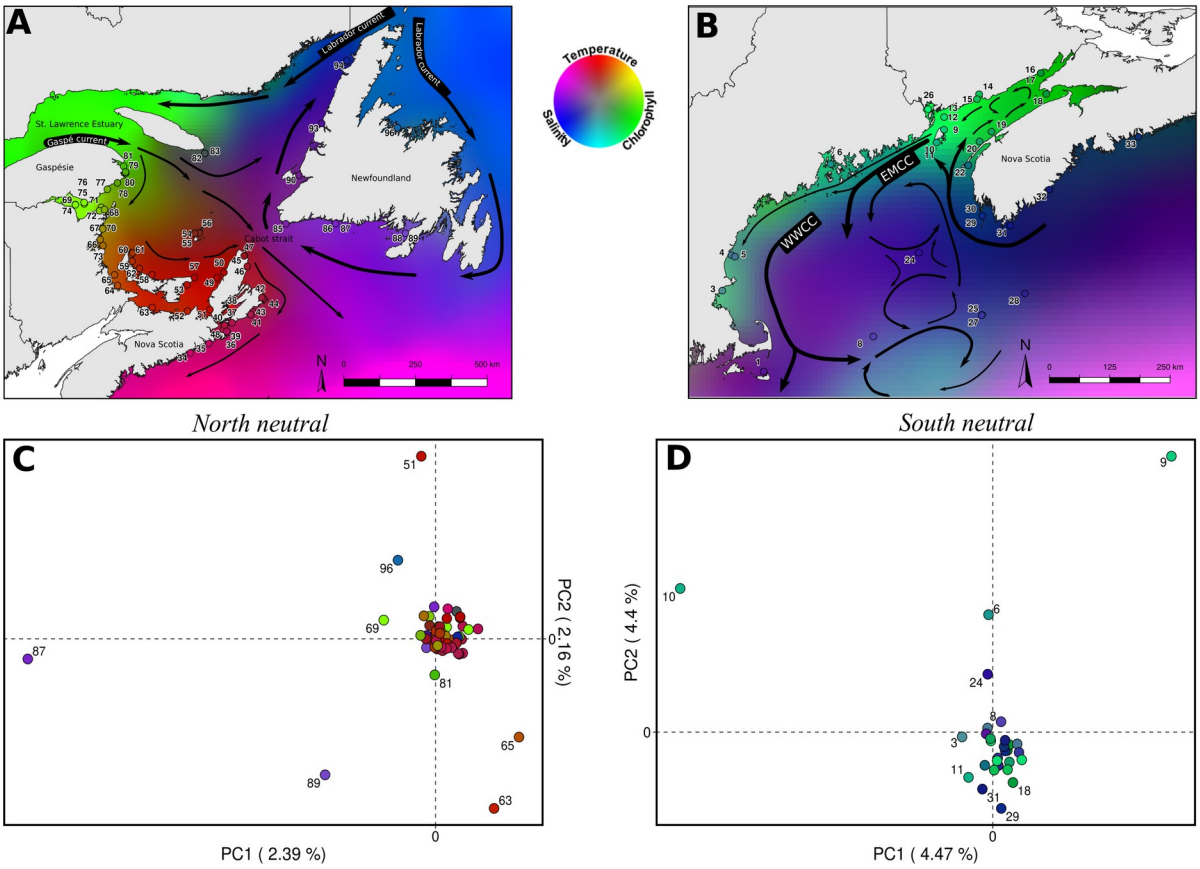

Figure S13. Fine-scale neutral population genetic structure.

(Upper panels) RGB composite habitat layer for the (*A*) northern and (*B*) southern genetic clusters. The red, green and blue color channels represent the intensities of the mean annual sea surface temperature (SST), sea surface chlorophyll concentration (SSC) and sea surface salinity (SSS) at each pixel, respectively. Environmental layers were normalized between 0 and 1 before RGB projection and normalized layers were "contrast stretched" to enhance visual clarity (min quantile = 0.05 and max quantile = 0.95). Sampling sites are represented by circles colored according to an RGB habitat value averaged over a buffer of two map units (2*5 arcmin ; ~18,4 km radius). Black arrows represent major current circulation within the two regions. EMCC and WMCC indicate the "Eastern Maine Coastal Current" and the "Western Maine Coastal Current", respectively. (bottom panels) PCA biplot based on allele frequencies of putative neutral loci for sampling sites in the (*C*) northern (13,543 SNPs) and (*D*) southern (12,944) study regions, where each circle represents a sampling site colored according to its RGB projection from their respective map above.