En Homardant

exploring the population structure of *Homarus gammarus*

Simoens Kobe
International Master of Science in Marine Biological Resources
Numerical Ecology 4 November 2022

Abstract

Genetic markers are an excellent example of multivariate response variables. Multivariate tools have long been used in the analysis of genetic data. However, analyses such as Principal Component Analysis, Principal Coordinate Analysis and Redundancy Analysis are rarely directly applied to genetic data.

The onset of Single Nucleotide Polymorhpism identification via Restriction-site Associated DNA sequencing provides excellent multivariate data in large quantity and of high quality.
At the same time, excellent multivariate tools are used in ecology. Multiple attempts have been made to use these multivariate methods on SNP data, but their full potential has yet to be unleashed.

In this exercise, a thorough multivariate analysis is used with excellent genetic data on the European lobster (*Homarus gammarus*) presented in earlier studies. Advanced spatial tools such as distance-based Moran Eigenvector Maps and Asymmetric Eigenvector Maps are integrated in linear regression in order to capture the spatial structure in the SNP distributions.
Both F_{st} distances and allele frequencies are used as response variables.

The analysis shows promising results. Ordination clearly shows distinct meta-populations in the dataset. Variation partitioning after linear regression allows for the identification of local adaptation, isolation by distance and isolation by resistance. In particular the high differentiation of the Dutch Oosterschelde population could be attributed to directional neutral mechanisms such as dispersion via the dominant ocean currents.

These results show that the multivariate tools from ecology can offer an alternative way of analysing genetic SNP data. However, the results depend on the form of the response variables and this sensitivity should be further investigated.

Lastly, a sensitivity test is proposed by repeating the analysis after alternately removing SNPs from the dataset. This method can easily be applied to ecological species composition data as well.

"*Your scientists were so preoccupied with whether they could,*
they didn't stop to think if they should."

- Ian Malcolm in Jurassic Park -

Contents

1 Introduction

The European lobster or homard (*Homarus gammarus*) is a marine invertebrate of high economic importance. The species accounts for approximately 4,000 tons of annual landings in the European Union and the United Kingdom, representing 60 million euros of landing revenue (Figure (1)). Despite recorded collapses of local stocks (Agnalt et al., 1999), regional regulations on the management of lobster stocks are lacking.

The first step in proposing population conservation measures is understanding the genetic structure of the population under consideration (Rossi et al., 2021). A few previous projects have mapped the local population structure of *Homarus gammarus* (Watson et al., 2016; Ellis et al., 2017; Pavičić et al., 2020). However, all these studies used microsatellites as genetic marker. Jenkins et al. (2019) has been the first study to identify and use Single Nucleotide Polymorphisms (SNPs) in the exploration of the population structure of the European lobster. Due to their higher number and genome-wide distribution, SNPs have three main advantages over microsatellites: 1) the ability to distinguish population structure at higher resolutions; 2) a higher power to identify groups in clustering methods; and 3) the ability to discern local adaptation (Zimmerman et al., 2020). Consequently, in modern population structure studies, microsatellites are more and more replaced by SNPs as the genetic marker (Pavičić et al., 2020).

The population structure of the European lobster is highly dependent on larval dispersion. Larvae can drift with the currents for fifteen to 35 days, and water temperature greatly affects their development (Pavičić et al., 2020). Consequently, the lobster population structure is ideally suited for an analysis with advanced spatial tools such as distance-based Moran's Eigenvector Maps (dbMEMs) and Asymmetric Eigenvector Maps (AEMs) (Benestan et al., 2016).

In this exercise, the high-quality genetic data from Jenkins et al. (2019) is analysed using various multivariate methods. The goal is to provide alternative tools to explore the population structure of *Homarus gammarus*.

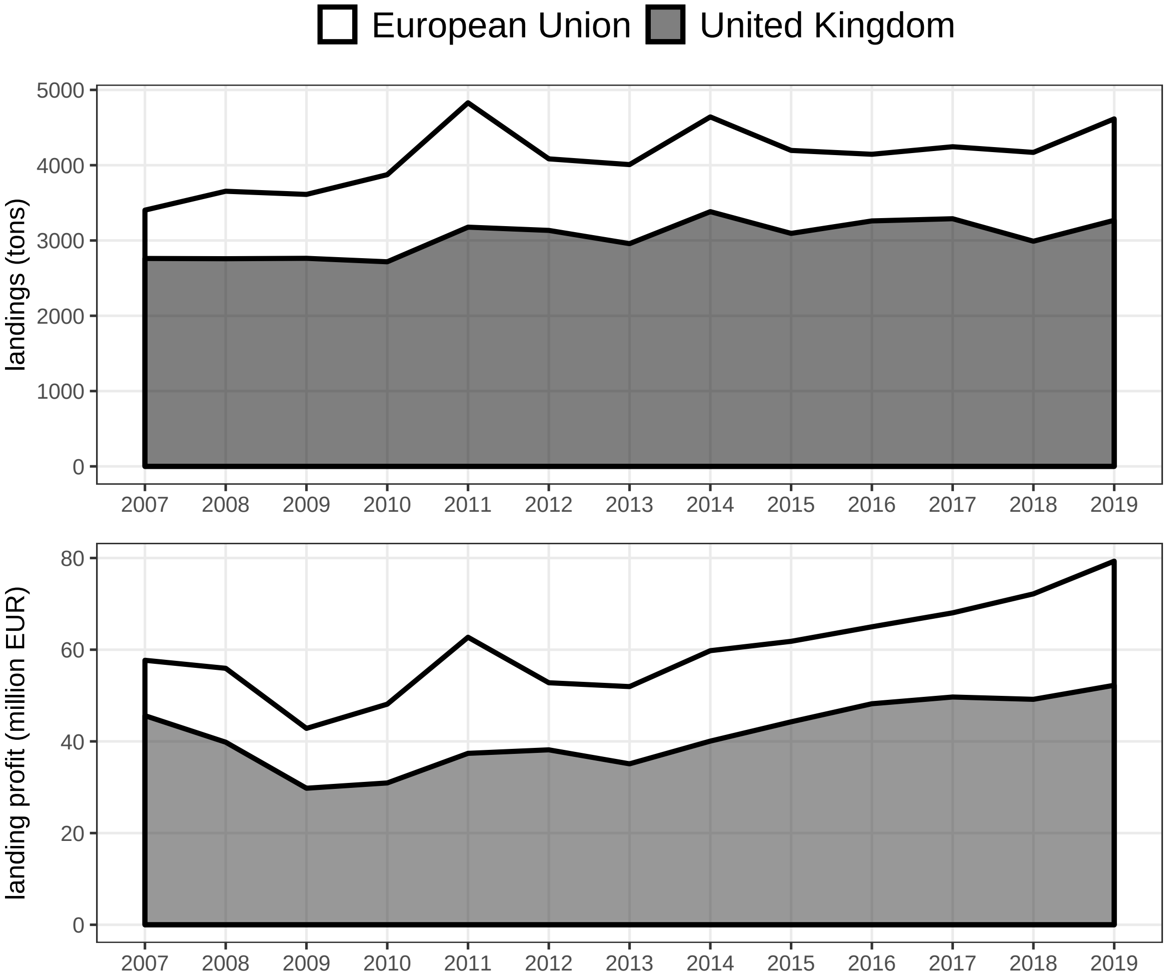


Figure 1: Economic value of lobster landings in the European Union; data from European Commission (2022)

2 Methods

All calculations and manipulations are done in R (R Core Team, 2022).
The `tidyverse` (Wickham et al., 2019) is used in making the graphics as well as the packages `ggforce` (Pedersen, 2022), `reshape2` (Wickham, 2007), `ggpubr` (Kassambara, 2020) and `ggnewscale` (Campitelli, 2022).

A list of all packages can be found in Appendix G.

2.1 Genetic data

Details on the collection and manipulation of the genetic data can be found in Jenkins et al. (2018) and Jenkins et al. (2019). Samples have been collected in 38 sites over the whole continent (Figure (2)), of which 31 sites represent the Atlantic sector and seven the Mediterranean.
Different temporal samples from the Île de Ré (Idr16 and Idr17) and Sardegna (Sar13 and Sar17) are combined in the following analysis.

Individual samples are sequenced using RAD sequencing (Davey & Blaxter, 2010) and SNPs are identified. After filtering, 79 biallelic SNPs are retained for a total of 1,278 individual samples.
Of these SNPs, 71 are found to be putatively neutral while 8 SNPs are considered outliers under selection.

Figure (3) shows the different manipulations of the original genetic data.

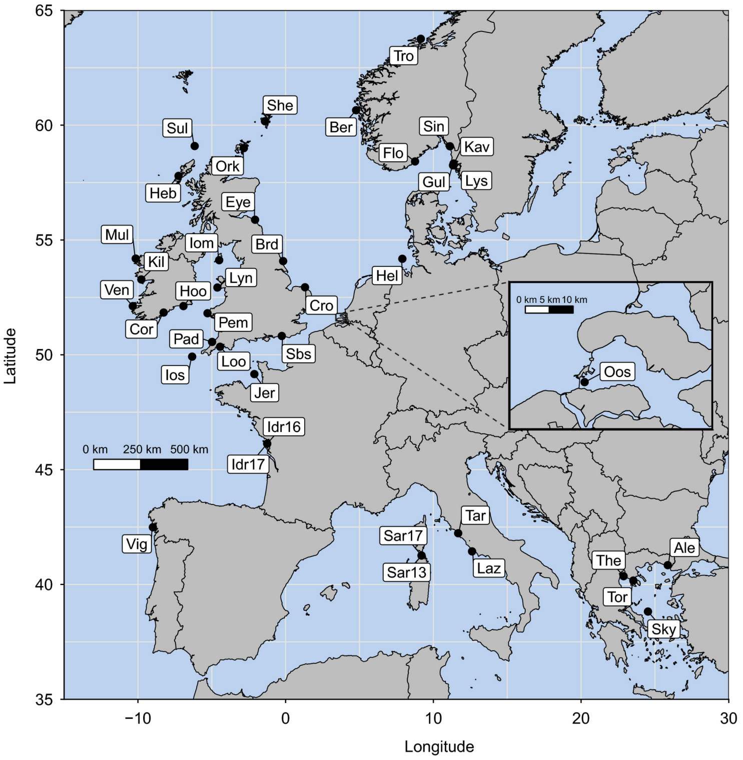


Figure 2: Sampling sites (Jenkins et al., 2019); names can be found in Table (1).

On the one hand, F_{st} values can be calculated between every pair of sampling sites based on any number of SNPs (Weir & Cockerham, 1984). The F_{st} value can be interpreted as the genetic distance between the two sites. Consequently, the result of this analysis is a genetic distance matrix. The package `adegenet` (Jombart & Ahmed, 2011) is used to read the genetic data and the F_{st} values are calculated using the package `hierfstat` (Goudet & Jombart, 2022). Next, this distance matrix is used in a Principal Coordinate Analysis (PCoA) (Legendre & Legendre, 2012).

On the other hand, allele frequencies can be calculated for each SNP. Only one allele is retained per SNP as the frequency of the other allele is unambiguously fixed by the frequency of the first allele ($q = 1 - p$). The package `poppr` version 2.9.3 (Kamvar et al., n.d.) is used in this calculation. While the F_{st} distance matrix loses all information on the original SNPs, the allele frequencies retain the identity of the SNPs, which can be used in a Principal Component Analysis (PCA). The PCA is performed with the package `vegan` (Oksanen et al., 2022). As a frequency of zero means total fixation of the other allele, a symmetrical distance measure is suitable for this type of data. Moreover, the problem of 'rare species' is entirely absent as all loci are present in every individual and population. If allele A is rare in the population, allele B will be ubiquitous. Consequently, the Euclidean distance should work fine for this type of data. No transformations are required in the PCA or RDA (Legendre & Legendre, 2012).

For the F_{st} distances, the mean or maximum F_{st} value can be used as a measure for the total β-diversity. For the allele frequencies, the total β-diversity can be calculated as the total variance of the allele frequency matrix (Legendre & De Cáceres, 2013).

2.2 Environmental data

Environmental data is collected from the Bio-ORACLE database (Assis et al., 2017). Raster layers for the annual mean and range of the Sea Surface Temperature (SST), salinity and Primary Productivity (PP) are used as well as the mean current velocity. Bathymetry data is collected from GEBCO (GEBCO Compilation Group, 2021).

Values for each environmental variable are extracted at the coordinates of the sampling sites using the package `raster` (Hijmans, 2022).

The distributions of the variables are shown in Appendix E.

Table 1: Sampling sites in WGS 84 (Jenkins et al., 2019)

code	**site**	**longitude**	**latitude**	**country**	**code**
Ale	Alexandroupoli	25.87	40.84	Greece	HEL
Ber	Bergen	4.77	60.65	Norway	NOR
Brd	Bridlington	-0.17	54.07	Great Britain	GRB
Cor	Cork	-8.26	51.84	Ireland	IRL
Cro	Cromer	1.31	52.94	Great Britain	GRB
Eye	Eyemouth	-2.07	55.88	Great Britain	GRB
Flo	Flødevigen	8.76	58.42	Norway	NOR
Gul	Gullmarfjord	11.33	58.25	Sweden	SVE
Heb	Outer Hebrides	-7.25	57.79	Great Britain	GRB
Hel	Helgoland	7.90	54.18	Germany	DEU
Hoo	Hook Peninsula	-6.92	52.12	Ireland	IRL
Idr	Île de Ré	-1.25	46.13	France	FRA
Iom	Isle of Man	-4.50	54.12	Great Britain	GRB
Ios	Isle of Scilly	-6.33	49.92	Great Britain	GRB
Jer	Jersey	-2.12	49.16	Channel Islands	CHA
Kav	Kavra	11.37	58.33	Sweden	SVE
Kil	Kilkieran Bay	-9.77	53.28	Ireland	IRL
Laz	Lazio	12.62	41.44	Italy	ITA
Loo	Looe Harbour	-4.44	50.35	Great Britain	GRB
Lyn	Llyn Peninsula	-4.62	52.93	Great Britain	GRB
Lys	Lysekil	11.37	58.26	Sweden	SVE
Mul	Mullet Peninsula	-10.15	54.19	Ireland	IRL
Oos	Oosterschelde	3.70	51.61	Netherlands	NDL
Ork	Orkney	-2.83	59.00	Great Britain	GRB
Pad	Padstow	-4.98	50.56	Great Britain	GRB
Pem	Pembrokeshire	-5.29	51.81	Great Britain	GRB
Sar	Sardegna	9.20	41.26	Italy	ITA
Sbs	Shoreham-By-Sea	-0.26	50.82	Great Britain	GRB
She	Shetland	-1.40	60.17	Great Britain	GRB
Sin	Singlefjord	11.12	59.08	Norway	NOR
Sky	Skyros	24.53	38.82	Greece	HEL
Sul	Sula Sgeir	-6.16	59.09	Great Britain	GRB
Tar	Tarquinia	11.68	42.23	Italy	ITA
The	Thermaikos Bay	22.88	40.36	Greece	HEL
Tor	Toronaios Bay	23.54	40.17	Greece	HEL
Tro	Trondheim	9.15	63.76	Norway	NOR
Ven	Ventry	-10.35	52.12	Ireland	IRL
Vig	Vigo	-8.99	42.49	Spain	ESP

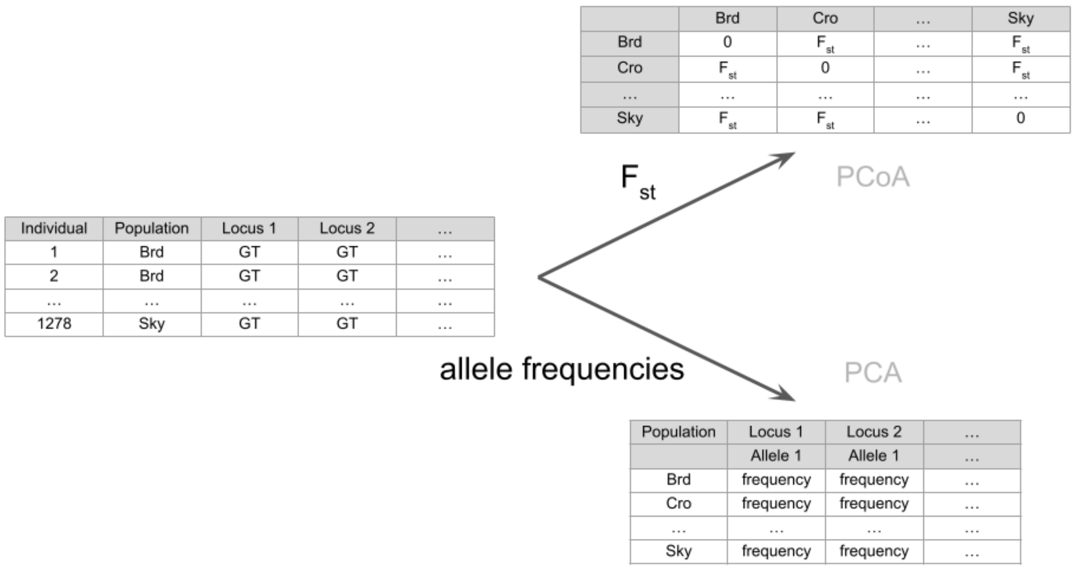
2.3 Spatial analysis

2.3.1 Geographical distances

Figure (2) clearly shows the complex geography of the study area. Populations at the different sampling sites are connected through dispersal of the planktonic larvae of *Homarus gammarus* (Benestan et al., 2016). As these larvae do not travel over land, the true distance between the sites is the in-water distance. Following the algorithm of Fetzer (2013), the packages `maptools` (Bivand & Lewin-Koh, 2022), `gdistance` (van Etten, 2017) and `geosphere` (Hijmans, 2021) are used to calculate the in-water distances between every pair of sampling sites (Figure (F.3)).

2.3.2 Linear component

Due to the complex geographical context of the study area, simple longitude and latitude coordinates cannot be used as the linear coordinates. Instead, a PCoA is conducted on the distance matrix in Section (2.3.1). The first two principal coordinates are used as the linear spatial coordinates.



Individual	Population	Locus 1	Locus 2	…
1	Brd	GT	GT	…
2	Brd	GT	GT	…
…	…	…	…	…
1278	Sky	GT	GT	…

	Brd	Cro	…	Sky
Brd	0	F_{st}	…	F_{st}
Cro	F_{st}	0	…	F_{st}
…	…	…	…	…
Sky	F_{st}	F_{st}	…	0

Population	Locus 1	Locus 2	…
	Allele 1	Allele 1	…
Brd	frequency	frequency	…
Cro	frequency	frequency	…
…	…	…	…
Sky	frequency	frequency	…

Figure 3: Schematic overview of the handling of the genetic data

2.3.3 distance-based Moran's Eigenvector Maps (dbMEMs)

dbMEMs are used to model more complex spatial structures than mere linear coordinates (Legendre & Legendre, 2012). The package `adespatial` (Dray et al., 2022) is used to calculate the dbMEMs from the in-water distances. Only the eigenvectors with positive eigenvalues are retained.

2.3.4 Asymmetric Eigenvector Maps (AEMs)

While dbMEMs model isotropic spatial structures, AEMs allow for clear directional preferences (Legendre & Legendre, 2012). While *isolation by distance* is a symmetric process, *isolation by resistance* is asymmetric. Dominant ocean currents will determine the travels of the planktonic larvae, thus creating a direction in the study area (Blanchet et al., 2011). Following the algorithm in Xuereb et al. (2018), the package `adespatial` (Dray et al., 2022) is used to translate the dominant ocean currents in asymmetric eigenvectors (Appendix A). The crude biophysical model only covers the 31 Atlantic sites.

2.4 Linear regression

The multivariate linear regression or RDA is performed with the package `vegan` (Oksanen et al., 2022). Four groups of variables are used in the regression: environment, linear coordinates, dbMEMs and AEMs. The geographic scope of the regression is limited to the 31 Atlantic sites.

Every group of explanatory variables is subjected to a forward-selection procedure from `adespatial` (Dray et al., 2022). Two selection criteria are used: $R^2_{adj} < R^2_{adj}(full)$ and $\alpha < 0.05$. A permutation test with 9999 permutations is used at each selection step. With the exception of the dbMEMs, a group of explanatory variable is only used if the full model is significant.

The four groups of explanatory variables allow for variation partitioning. Initially, the 'traditional' variation partitioning in `vegan` (Oksanen et al., 2022) is used. Every testable fraction is tested using a permutation test for the pseudo-F statistic with 9999 permutations.

Alternatively, the framework of hierarchical partitioning is used to interpret the different contributions of the groups of explanatory variables to explaining the variation in the genetic data. The package `rdacca.hp` is used and testable fractions are tested with a permutation test with a varying number of permutations (Lai et al., 2022).

When the F_{st} distance matrix is used as input, the principal coordinates represent the response variables. In contrast, the allele frequencies can be used directly as response variables.

Additional statistical tests are used to verify assumptions and model outcomes.

3 Results

3.1 PCoA versus PCA

3.1.1 PCoA versus PCA for all sites

Figure (4) shows the ordination of all the sites for both manipulations of the genetic data using all 79 SNPs. The PCA biplot is drawn in scaling 1 only showing loci with more than 90 percent of their length lying in the two plotted dimensions. The circle represents the equilibrium circle of descriptors (Legendre & Legendre, 2012).

The first observation is that both manipulations show remarkably similar patterns. The main difference appears to be that the PCA (70.18 %) captures more variation in its first two axes than the PCoA (35.91 %). As mentioned before, the PCoA with F_{st} distances loses all information on the locus descriptors while a biplot can easily be drawn for the PCA.

The distinction between the Mediterranean- and Atlantic sites is clear, representing the largest variation in the genetic data.

Secondary, the difference between the Italian and Greek sites in the Mediterranean and the British and Scandinavian sites in the Atlantic is clearly visible as well. Moreover, despite the different geographical context, these two distinctions appear to be determined by the same axis of variation.

Jersey is placed in the middle of the British sites while the Île de Ré and Vigo appear to be on the road between the Atlantic and the Mediterranean. Helgoland is placed in the middle of the Scandinavian sites.

Finally, the very peculiar ordination of the Oosterschelde has to be noted. This Dutch site appears to be differentiated from the other sites both along the first and the second axis.
Geographically, the Oosterschelde should be expected among the British sites (Figure (2)).

Clusters of loci could be made. The variation in loci 29889, 81462, 15581, 11291 and 58053 appears to determine the larger distinction between the Mediterranean and the Atlantic.
In contrast, the variation in 53314, 65064, 6157, 65576 and 15128 appears to determine the smaller distinction between Italy and Greece on the one hand and between Britain and Scandinavia on the other hand.

3.1.2 PCoA versus PCA for the Atlantic sites

Figure (5) shows the ordination of the Atlantic sites for both manipulations of the genetic data using all 79 SNPs. The PCA biplot is drawn in scaling 1 only showing loci with more than 70 percent of their length lying in the two plotted dimensions.

Yet again, very similar patterns emerge from both methods. The PCA (50.10 %) shows more variation in its first two axes than the PCoA (22.73 %).

The main axis now separates the British sites from the Scandinavian sites. Secondary, a faint distinction between the Irish and Great British sites is visible along the second axis.

Jersey is placed among the British sites while Helgoland is welcomed in the Scandinavian group. The Île de Ré lies more isolated from the other sites along the first axis and Vigo is clearly distinct.

However, the Oosterschelde anomaly is now clearly visible. This site is surprisingly different from all the other Atlantic sites.

Loci 6157, 65064, 39876, 65576 and 53935 form a distinct group of descriptors whose variation determines the difference between the British- and Scandinavian sites.

Similar graphs for the 71 putatively neutral SNPs and the eight outlier SNPs can be found in Appendix B.

3.2 Total β-diversity

The results of the global diversity analysis are represented in Table (2). The outlier dataset of 8 SNPs clearly displays a higher β-diversity. The diversity measures are approximately ten times higher than in the neutral dataset of 71 SNPs.

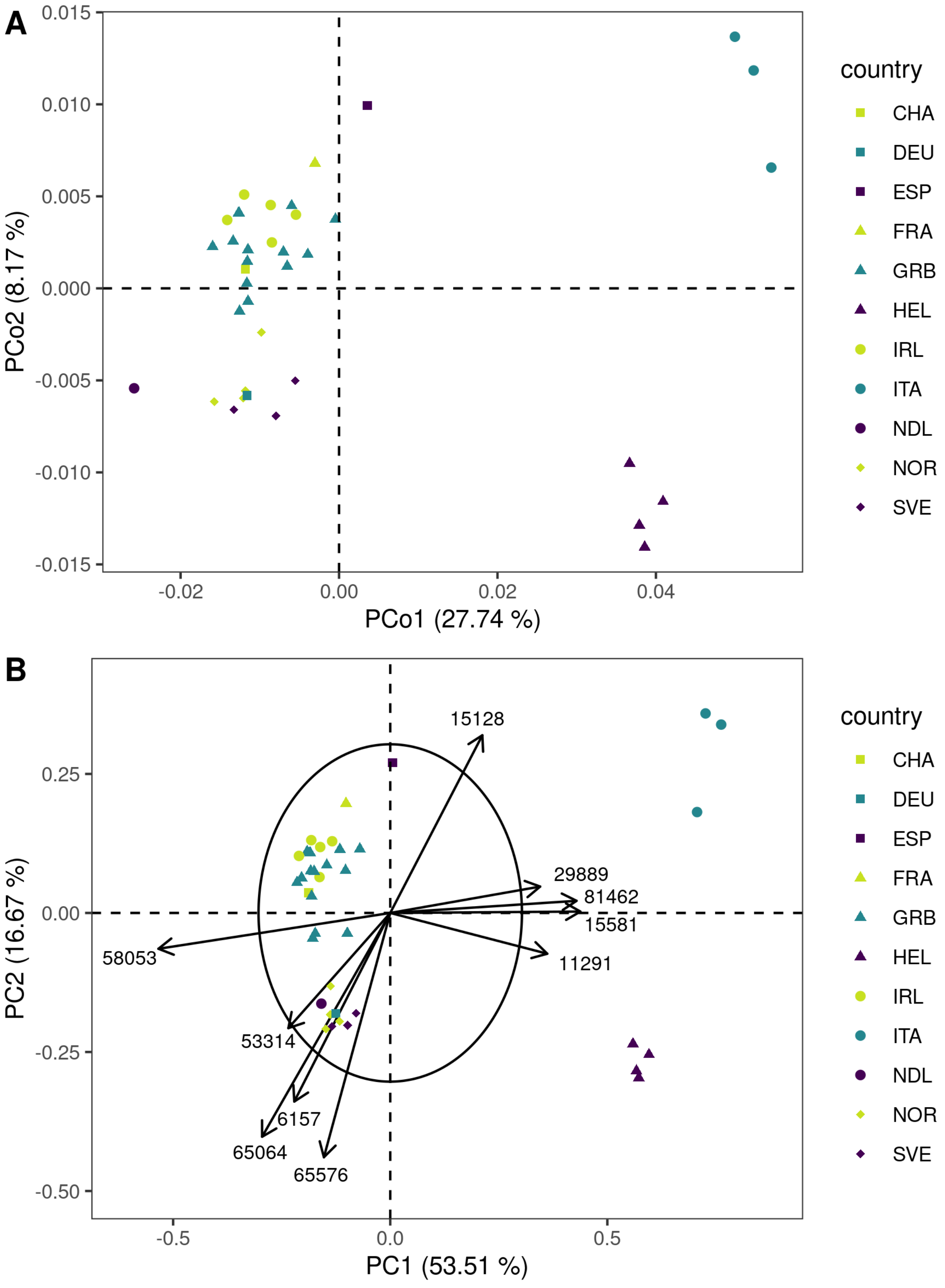


Figure 4: Ordination graphics for the (A) PCoA and (B) PCA for all sites; details can be found in Section (3.1.1).

Table 2: Total β-diversity measures for the different datasets

dataset	mean F_{st}	max F_{st}	total variance
8 SNPs	0.276	0.733	0.0249
71 SNPs	0.0297	0.138	0.00397

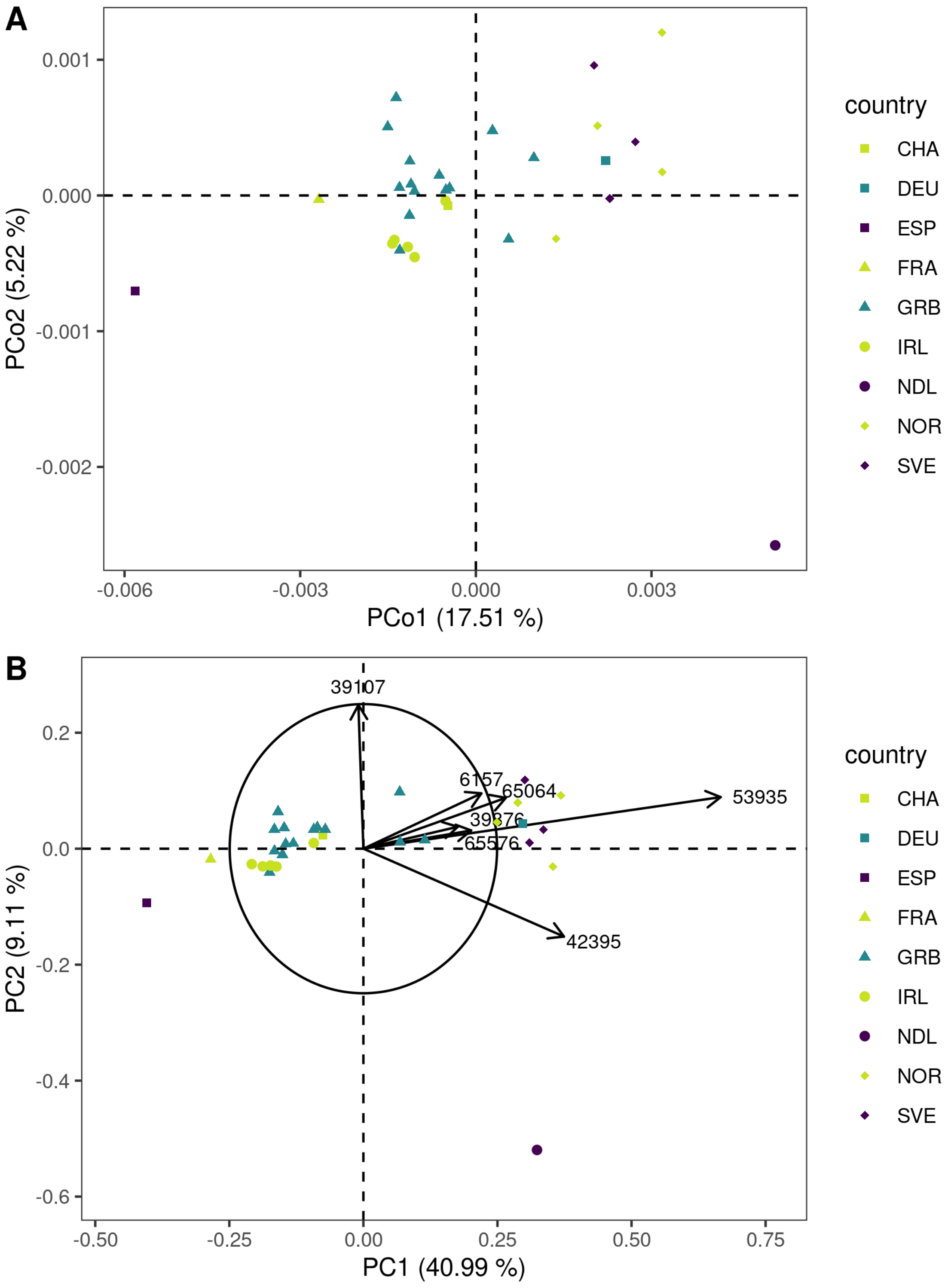


Figure 5: Ordination graphics for the (A) PCoA and (B) PCA for the Atlantic sites; details can be found in Section (3.1.2).

3.3 Redundancy Analysis (RDA)

3.3.1 8 outlier SNPs

The full models regressing the F_{st} principal coordinates to each of the groups of explanatory variables are all significant with the exception of the AEMs ($p_F = 0.3936$). Consequently, the AEMs are excluded from the further analysis.

Table (4) represents the remaining variables after the forward selection procedure.

The variation partitioning is represented in Figure (6). The shared fraction is substantial, which indicates that the different groups of variables explain the same part of the variation.

Table (3) displays the results of the permutation tests of every testable fraction in Figure (6). For example, the fraction [env] | [lin] + [MEM] corresponds to the fraction of variation that is explained by the environment after the fractions explained by the linear coordinates and the dbMEMs are taken into account. This is conditional regression. The environment costs three degrees of freedom (dof) for three environmental variables. The total remaining degrees of freedom after regression with the linear coordinates (two variables) and dbMEMs (three variables) is equal to the number of sites (31) minus one, minus two minus three (25).

The linear coordinates are the only variables that are able to explain a significant fraction ($p_F < 0.05$) of the variation after taking into account the fractions of the other groups.

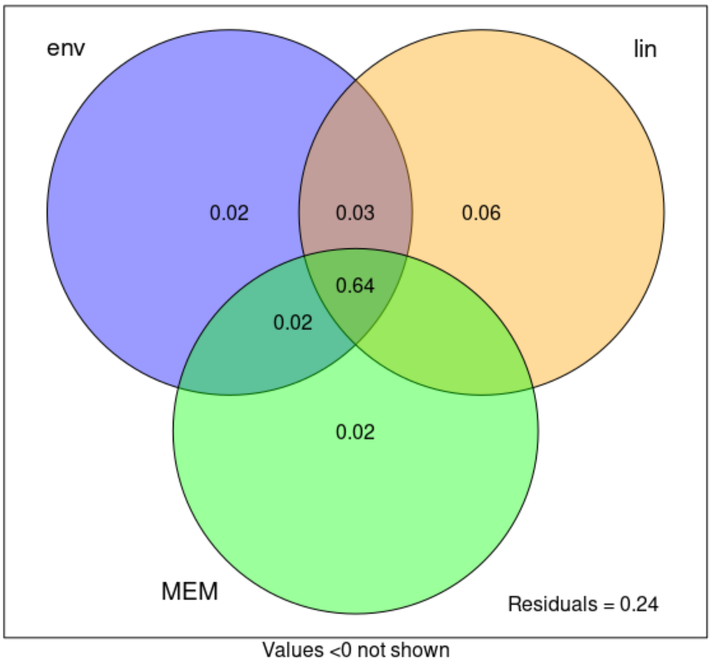


Figure 6: Venn diagram of the variation partitioning of the linear regression using the 8 outlier SNPs and F_{st} distances

Figure (7) shows the result of the RDA graphically.

The first axis represents an impressive 78.75 percent of the variation while the second axis barely adds anything to the analysis.

The first axis is determined by the difference between the British- and Scandinavian sites. In terms of the explanatory variables, the direction of greatest variation is determined by gradients in SST, MEM1 and PCo1. The variables SST mean, MEM1 and PCo1 appear to be highly correlated while they are anticorrelated with the SST range (Appendix F). These results are in good agreement with the large shared fraction in the variation partitioning (Figure (6)).

The variables PCo2, MEM3 and MEM5 mainly contribute to the second axis which contains only a very small part of the total variation.

The Oosterschelde is placed among the British sites, as expected geographically. Consequently, a linear regression of the F_{st} distances using the 8 outlier SNPs cannot reproduce the deviant placement of the Oosterschelde in the unconstrained ordination (Figure (5)). This last observation is also apparent in the ordination of the residuals after the regression (Figure (8)).

The results of the hierarchical variation partitioning are shown in Table (4). R^2_{adj} values are the sum of the unique variation fractions and a weighted average of the shared fractions. All three groups of variables have very similar individual fractions. Of the individual variables only those associated with the first RDA axis show significant fractions. Notably, PCo1 explains twenty percent of the total variation.

Table 3: Permutation test of the variation partitioning using the 8 outlier SNPs and F_{st} distances

fraction	R^2_{adj}	dof	p_F
[env]	0.707	3(30)	< 0.0001
[lin]	0.706	2(30)	< 0.0001
[MEM]	0.649	3(30)	< 0.0001
[env] + [lin]	0.736	5(30)	< 0.0001
[env] + [MEM]	0.699	6(30)	< 0.0001
[lin] + [MEM]	0.740	5(30)	< 0.0001
[env] + [lin] + [MEM]	0.756	8(30)	< 0.0001
[env] | [lin] + [MEM]	0.015	3(25)	0.11
[lin] | [env] + [MEM]	0.056	2(24)	0.0044
[MEM] | [env] + [lin]	0.020	3(25)	0.084
[env] | [lin]	0.031	3(28)	0.038
[env] | [MEM]	0.050	3(27)	0.027
[lin] | [env]	0.029	2(27)	0.028
[lin] | [MEM]	0.091	2(27)	0.0002
[MEM] | [env]	< 0	3(27)	0.64
[MEM] | [lin]	0.035	3(28)	0.024

Table 4: Remaining explanatory variables after forward selection for 8 SNPs and F_{st} distances; results of hierarchical variation partitioning with 9999 permutations

variable	R^2_{adj}	p_R
environment	0.2544	0.0021
SST mean	0.1254	0.0217
SST range	0.0987	0.0391
salinity mean	0.0890	0.0512
linear	0.2741	0.0005
PCo1	0.1992	0.0057
PCo2	0.0619	0.09
dbMEMs	0.2274	0.0093
MEM1	0.1445	0.0169
MEM3	0.0178	0.2296
MEM5	0.0193	0.2049

3.3.2 71 putatively neutral SNPs

An identical analysis can be performed, now using the 71 SNPs classified as putatively neutral.
As these SNPs are under control of neutral forces, the environment should not play a significant role in explaining their variation. The environment acts through selection, and selection is not a neutral process. Consequently, the fraction [env] | [lin] + [MEM] + [AEM] should not add significant explanatory power.

After forward selection, four environmental variables are retained: SST mean, SST range, PP mean and PP range. The fraction [env] | [lin] + [MEM] + [AEM] corresponds to $R^2_{adj} = 0.0037$ and is significant ($p_F = 0.0221$). Consequently, the four environmental factors appear to explain a significant part of the variation after the other explanatory variables are taken into account.

However, the environmental variables can be tested for collinearity. The Variance Inflation Factor (VIF) for each variable is shown in Table (5). Notably, the previously selected PP variables display high collinearity with the other variables (VIF $>$ 10). Consequently, removing the PP variables from the environment group is justified.

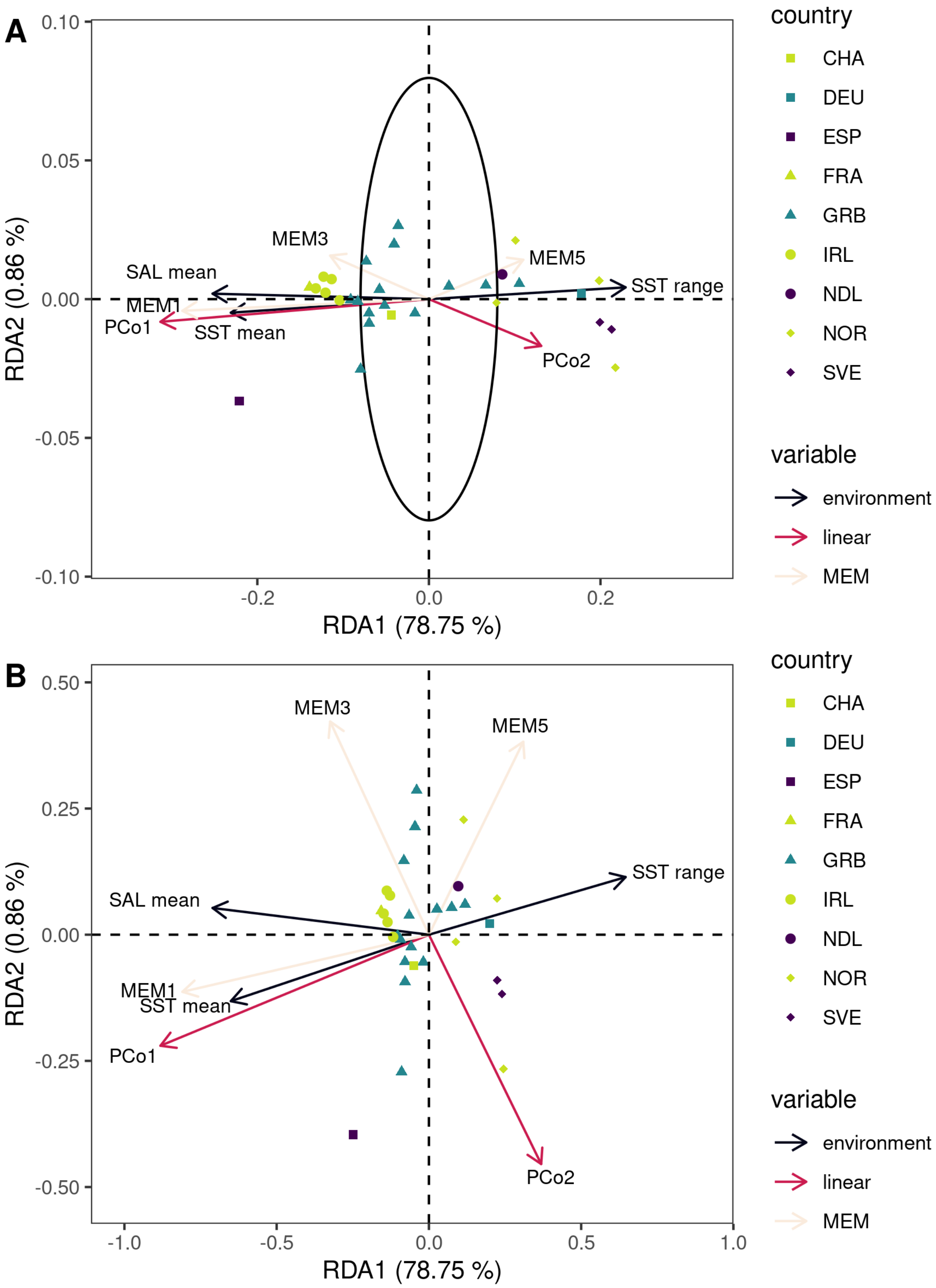


Figure 7: RDA with the F_{st} distances for the 8 outlier SNPs; (A) scaling 1, and (B) scaling 2 linear combinations of explanatory variables

Once the PP variables are removed, only SST range is retained after forward selection. Furthermore, the fraction [env] | [lin] + [MEM] + [AEM] is now highly non-significant ($R^2_{adj} < 0$; $p_F = 0.3758$).

In summary, the environmental variables are not considered in the further analysis.

The model with the linear coordinates is not significant ($p_F = 0.0878$). However, the linear coordinates are retained in order to separate linear trends from the more complex spatial structures.

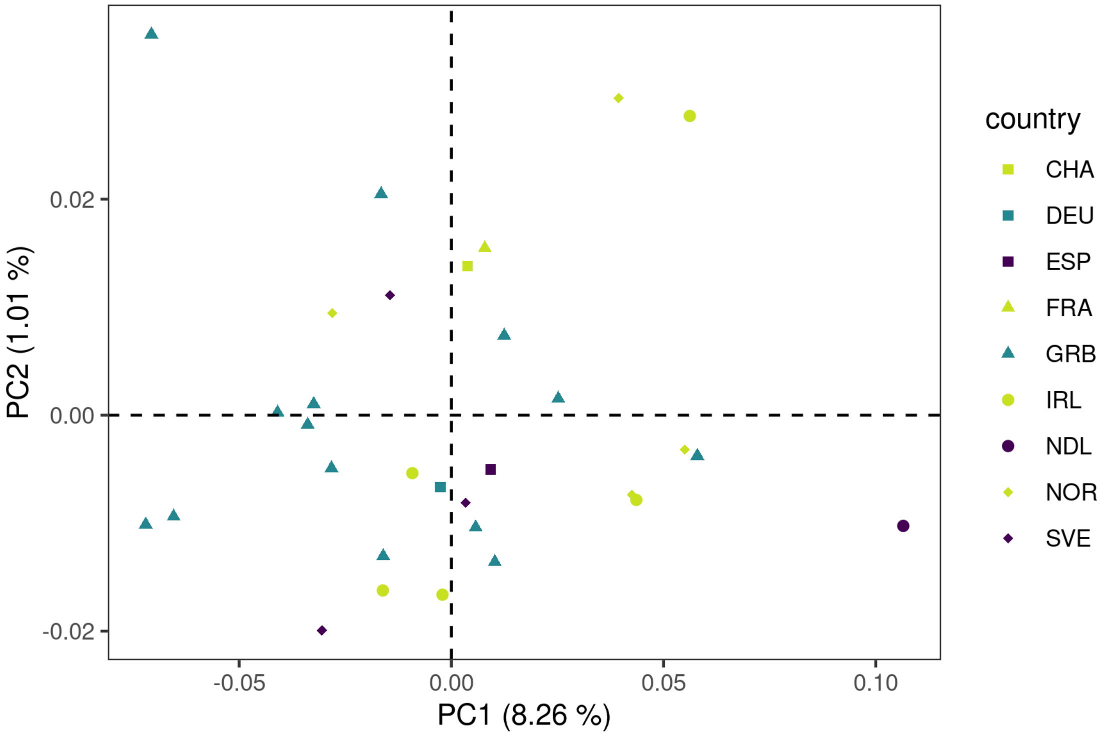


Figure 8: Residuals after RDA with F_{st} distances and 8 outlier SNPs

Table 5: Collinearity analysis of the environmental variables

variable	R^2	VIF
SST mean	0.6440	2.81
SST range	0.8249	5.71
salinity mean	0.8856	8.74
salinity range	0.8702	7.70
velocity mean	0.6052	2.53
PP mean	0.9592	24.50
PP range	0.9572	23.38
bathymetry	0.6014	2.51

The full model with the dbMEMs is not significant ($p_F = 0.1568$). However, MEMs should first be pruned using selection procedures.

Consequently, three groups of variables are considered: linear coordinates, dbMEMs and AEMs. The variables that are retained after forward selection can be found in Table (7).

Figure (9) shows the variation partitioning graphically. The shared fractions are now a lot smaller. The dbMEMs do not explain any additional variation. The most informative group of variables are the AEMs. The testable fractions are presented in Table (6). As mentioned before, the linear coordinates are not significant on their own. However, after the dbMEMs and AEMs are taken into account, the linear coordinates do explain a significant additional part of the variation.
The variation that is explained by the dbMEMs is also fully explained by the AEMs.
Finally, the AEMs clearly explain the largest part of the variation (± 25 %).

Figure (10) shows the RDA analysis graphically. Only 44.34 percent of the variation is captured by the first two constrained axes.

The neutral regression model accurately predicts the deviant position of the Oosterschelde. This is also visible in the residual unconstrained axes in Figure (11). No residual patterns are discernible.

The variable PCo1 determines the same NE-SW direction as in the analysis for the 8 outlier SNPs (Figure (F.2)). The differentiation of the Oosterschelde coincides with a host of AEM variables: AEM10, AEM17, AEM18, AEM19 and AEM22.

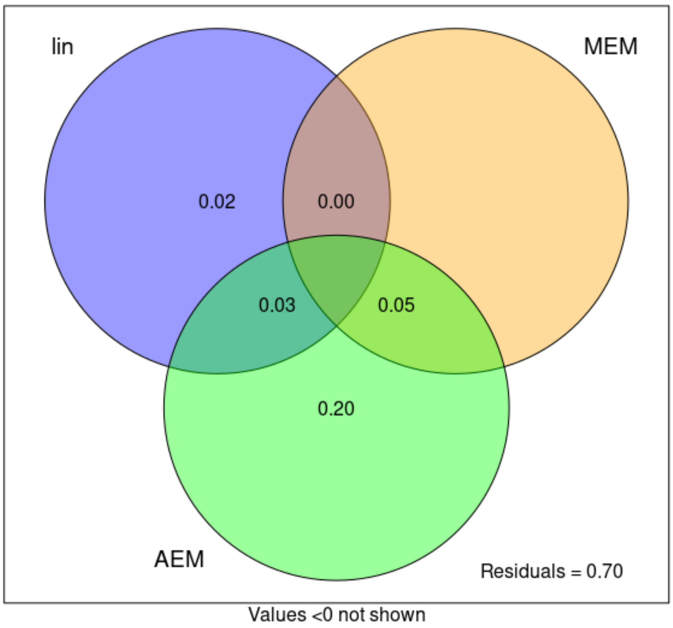


Figure 9: Venn diagram of the variation partitioning of the linear regression using the 71 neutral SNPs and F_{st} distances

Table 6: Permutation test of the variation partitioning using the 71 neutral SNPs and F_{st} distances

fraction	R^2_{adj}	**dof**	p_F
[lin]	0.045	2(30)	0.088
[MEM]	0.047	1(30)	0.025
[AEM]	0.281	9(30)	0.0015
[lin] + [MEM]	0.094	3(30)	0.023
[lin] + [AEM]	0.299	11(30)	0.0006
[MEM] + [AEM]	0.281	10(30)	0.0022
[lin] + [MEM] + [AEM]	0.296	12(30)	0.0031
[lin] | [MEM] + [AEM]	0.015	2(20)	0.017
[MEM] | [lin] + [AEM]	< 0	1(19)	0.28
[AEM] | [lin] + [MEM]	0.202	9(27)	< 0.0001
[lin] | [MEM]	0.047	2(29)	0.030
[lin] | [AEM]	0.018	2(21)	0.0093
[MEM] | [lin]	0.049	1(28)	0.012
[MEM] | [AEM]	< 0	1(21)	0.19
[AEM] | [lin]	0.253	9(28)	0.0002
[AEM] | [MEM]	0.234	9(29)	< 0.0001

The results of the hierarchical variation partitioning are shown in Table (7). Interestingly, only the variables PCo1 and AEM15 explain a significant fraction of the variation within the hierarchical partitioning framework. The AEMs, in contrast, are the only group of variables that explain a significant part of the variation.

Alternatively, the same analyses can be performed using the allele frequencies instead of the F_{st} distances (Appendix C). A Mantel test can be found in Appendix D.

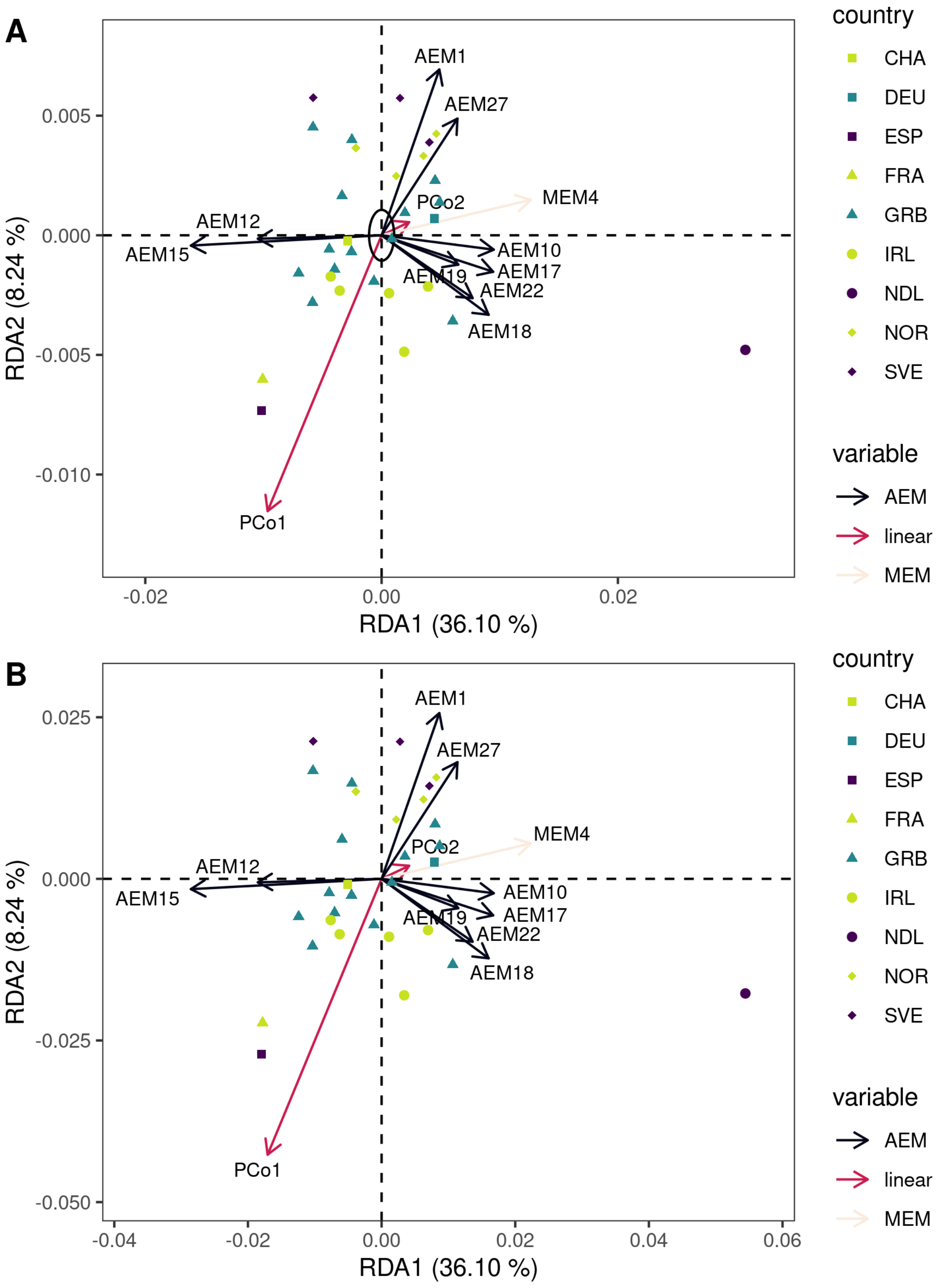


Figure 10: RDA with the F_{st} distances for the 71 neutral SNPs; (A) scaling 1, and (B) scaling 2 linear combinations of explanatory variables

3.4 Sensitivity test

In section (3.3.2) the neutral SNPs are analysed. Jenkins et al. (2019) elaborates on the statistical tests that are used to identify outlier SNPs. Three different methods are used: `BayeScan` (Foll & Gaggiotti, 2008), `OutFLANK` (Whitlock & Lotterhos, 2015) and `PCadapt` (Luu et al., 2016). Outliers are called only when at least two methods identify them as such ($\alpha < 0.05$).

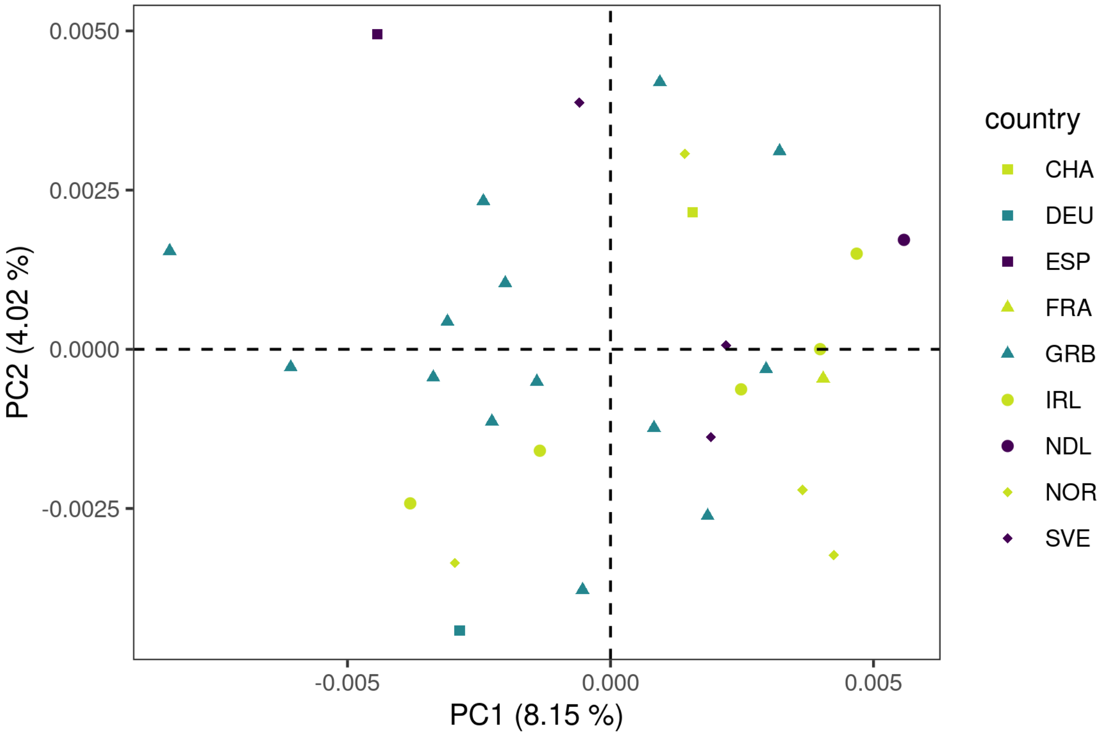


Figure 11: Residuals after RDA with F_{st} distances and 71 neutral SNPs

Table 7: Remaining explanatory variables after forward selection for 71 SNPs and F_{st} distances; results of hierarchical variation partitioning;
9999 permutations for variable groups and 999 permutations for individual variables

variable	R^2_{adj}	p_R
linear	0.0189	0.1222
PCo1	0.0336	0.043
PCo2	< 0	0.734
dbMEMs	0.0122	0.135
MEM4	0.0156	0.111
AEMs	0.2022	0.0004
AEM1	< 0	0.243
AEM10	0.0254	0.063
AEM12	0.0360	0.052
AEM15	0.0782	0.006
AEM17	0.0158	0.126
AEM18	0.0251	0.059
AEM19	0.0067	0.151
AEM22	0.0178	0.098
AEM27	< 0	0.230

The 8 outlier SNPs clearly show a correlation with the environmental variables. At the same time, the remaining variation in the 71 outlier SNPs cannot be explained by the environmental variables once the variation due to the spatial structure is taken into account.

The selection or filtering of SNPs is independent from the analyses presented here. In order to assess the sensitivity of the analysis to the SNP selection, all SNPs are removed from the neutral dataset in turn. After each removal, the significance of the fraction [env] | [lin] + [MEM] + [AEM] is recalculated. The result is shown in Figure (12).

Firstly, all fractions are clearly non-significant. Consequently, regardless of the removed SNP, the environment is not retained in the explanatory variables for the neutral SNPs.

Secondly, the p-values are considerably lower with the F_{st} distances. Consequently, the allele frequencies appear to behave more neutrally.

Thirdly, and most interestingly, the fluctuation of the p-values around the p-value for all SNPs is quite limited for the allele frequencies. In contrast, the p-values with the F_{st} distances show larger fluctuations. Certain SNPs are even missing from the F_{st} list as for some no environmental variables are selected during the forward-selection procedure or for others the environmental variables explain no additional variation at all.

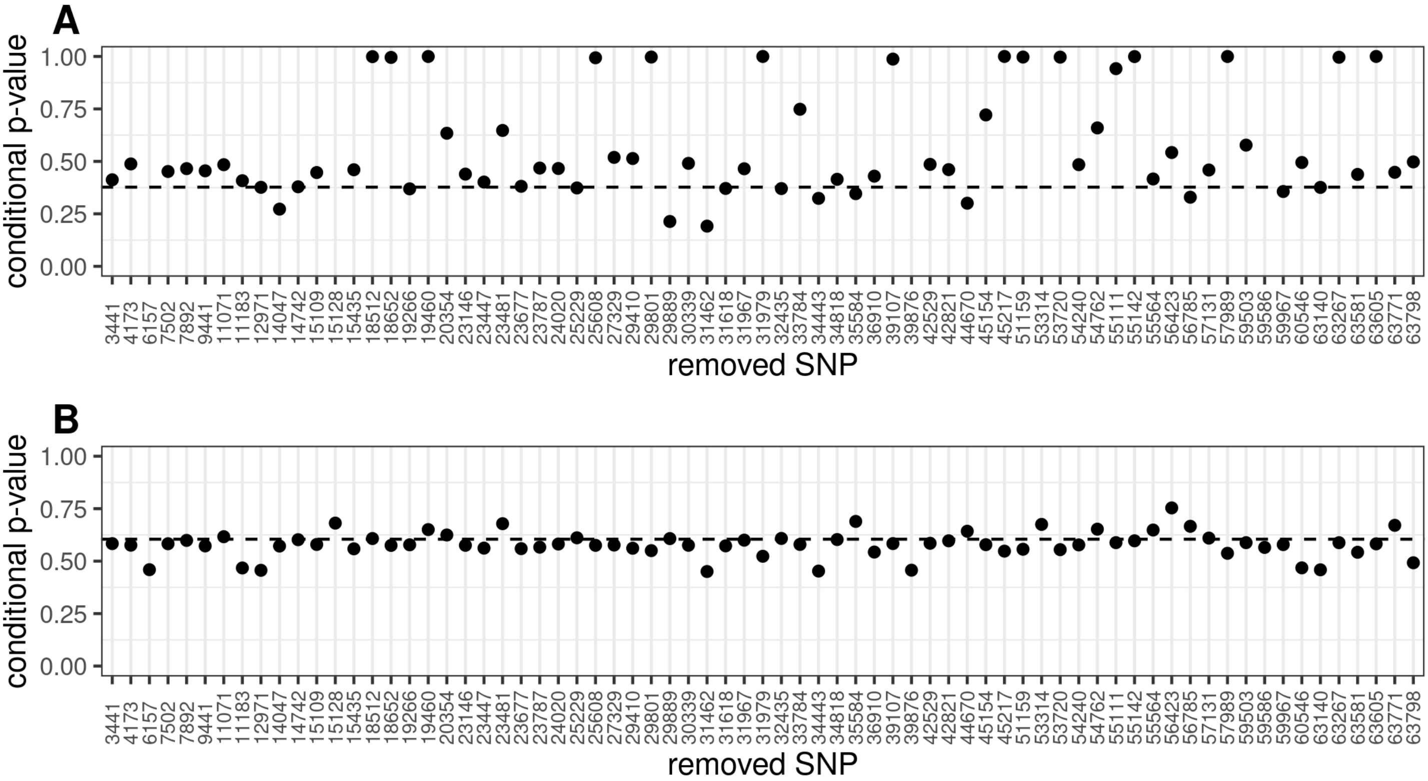


Figure 12: The significance level of the fraction [env] | [lin] + [MEM] + [AEM] for each removed neutral SNP (A) with F_{st} distances and (B) with allele frequencies; the horizontal dashed line represents the significance level with all SNPs.

4 Discussion

The use of multivariate methods in population genetics is not unheard of. PCA is used in outlier selection (Luu et al., 2016) and clustering (Jombart et al., 2010), for example. Examples of PCoA with F_{st} distances can also be found (García-Marín et al., 1999; Szczecińska et al., 2016; Ruppert et al., 2017). However, these multivariate methods are rarely used directly on the genetic data.
Population genetics aims to identify clusters of individuals or populations.
Consequently, the variation between the groups is more interesting than the variation within groups (Jombart et al., 2010). For example, Discriminant Analysis of Principal Components (DAPC) combines the multivariate character of PCA with the grouping potential of Discriminant Analysis (DA) (Jombart et al., 2010). However, DAPC either requires a priori assumptions on the number of groups in the dataset or uses algorithms to find this number automatically.

The prevalent methods in population genetics, such as STRUCTURE (Pritchard et al., 2000) and DAPC (Jombart et al., 2010), are hard to combine with linear regression. Consequently, local adaptation, *isolation by distance* and *isolation by resistance* can be detected with, for example, outlier selection (Feng et al., 2015) and the Mantel test (Diniz-Filho et al., 2013). However, further associations with environmental variables or spatial structures can only be done indirectly at best (Feng et al., 2015). Recently, some studies combined multivariate methods from ecology with high-resolution genetic data (Benestan et al., 2016; Xuereb et al., 2018; Ruiz Miñano et al., 2022). Forester et al. (2018) provides an excellent overview of multivariate linear regression for genetic data.

All of these studies use allele frequencies as response variables. In this exercise, an alternative is provided with the F_{st} distance matrix. While the results are generally very similar, important differences also emerge. For example, both methods clearly attribute the differentiation of the Oosterschelde lobster population to neutral processes (Jenkins et al., 2019). However, while the F_{st} distances point more to directional processes, the allele frequencies hint at more isotropic causes. Understanding this discrepancy could lead to further insights in the interpretation of regression results for genetic data.

Moreover, all of the above-mentioned studies (Benestan et al., 2016; Xuereb et al., 2018; Ruiz Miñano et al., 2022) use a Hellinger transformation on the allele frequencies (Legendre & Legendre, 2012). In this exercise, it is argued that such a transformation is not necessary and should therefore be avoided. Jombart et al. (2009) supports this decision although it makes a case for scaling the allele frequencies due to their inherent multinomial character.

Both F_{st} values and total variances can be used to estimate global β-diversity (Weir & Cockerham, 1984; Legendre & De Cáceres, 2013). The results of both methods are consistent with each other. The outlier SNP dataset shows a higher β-diversity than the neutral SNP dataset. This result is not surprising as selection processes will differentiate populations more from each other than neutral processes, although the history of the populations greatly influences the interpretation of F_{st} values and genetic β-diversity (Kitada et al., 2021).

The use of spatial analysis in the form of dbMEMs and AEMs is particularly suited for analyses of seascape genomics. Ocean currents clearly impose directionality upon the variation in genetic structure and community composition (Benestan et al., 2016; Bonifácio et al., 2018; Xuereb et al., 2018), and fine-scale structures can be very important in the management of species and populations (Vu et al., 2020). The results in this exercise clearly show the importance of complex spatial structures in explaining the genetic variation. Both for the 8 outlier SNPs and for the 71 neutral SNPs, the spatial structure explains the grunt of the variation.

Lastly, a sensitivity test is proposed. Sensitivity tests are customary in mechanistic and theoretical models (Pannell, 1997; Cariboni et al., 2007; Steenbeek et al., 2018; Zhou et al., 2022) but are rarely performed on statistical regression models (see for example Agler and De Boeck (2020)). In this exercise, SNPs are removed from the dataset in turn and certain model parameters are recalculated after each removal. In summary, when using F_{st} distances, the results appear to depend on the selected SNPs. When using allele frequencies, in contrast, the results are quite robust against the selection of SNPs. This conclusion will probably depend on the exact circumstances and the sensitivity of the analysis with F_{st} distances should be investigated further. However, similar sensitivity tests can be used for ecological data where the allele frequencies will be replaced by species. Removing species one by one can shine a light on the sensitivity of certain results as well as on the species most responsible.

Special thanks to María Fernanda Bayo, el ser de la oscuridad que me ha mostrado la luz...

Contact: kobe.simoens@imbrsea.eu
Full code available at: https://github.com/ksimoens/Numerical-Ecology-Project

5 References

Agler, R. A., & De Boeck, P. (2020). Factors associated with sensitive regression weights: A fungible parameter approach. *Behavior Research Methods*, *52*, 207–223. https://doi.org/10.3758/s13428-019-01220-6

Agnalt, A.-L., van der Meeren, G., Jørstad, K., Næss, H., Farestveit, E., Nøstvold, E., Svåsand, T., Korsøen, E., & Ydstebø, L. (1999). Stock enhancement and sea ranching. Blackwell Science.

Assis, J., Tyberghein, L., Bosch, S., Verbruggen, H., Serrão, E. A., & De Clerck, O. (2017). `Bio-ORACLE` v2.0: Extending marine data layers for bioclimatic modelling. *Global Ecology and Biogeography*, *27*(3), 277–284. https://doi.org/10.1111/geb.12693

Benestan, L., Quinn, B. K., Maaroufi, H., Laporte, M., Clark, F. K., Greenwood, S. J., Rochette, R., & Bernatchez, L. (2016). Seascape genomics provides evidence for thermal adaptation and current-mediated population structure in american lobster (Homarus americanus). *Molecular Ecology*, *25*(20), 5073–5092. https://doi.org/10.1111/mec.13811

Bivand, R., & Lewin-Koh, N. (2022). *`maptools`: Tools for handling spatial objects* [R package version 1.1-4]. https://CRAN.R-project.org/package=maptools

Blanchet, F. G., Legendre, P., Maranger, R., Monti, D., & Pepin, P. (2011). Modelling the effect of directional spatial ecological processes at different scales. *Oecologia*, *166*, 357–368. https://doi.org/10.1007/s00442-010-1867-y

Bonifácio, P., Grémare, A., Gauthier, O., Romero-Ramirez, A., Bichon, S., Amouroux, J.-M., & Labrune, C. (2018). Long-term (1998 vs. 2010) large-scale comparison of soft-bottom benthic macrofauna composition in the gulf of lions, nw mediterranean sea. *Journal of Sea Research*, *131*, 32–45. https://doi.org/10.1016/j.seares.2017.08.013

Campitelli, E. (2022). *`ggnewscale`: Multiple fill and colour scales in '`ggplot2`'* [R package version 0.4.8]. https://CRAN.R-project.org/package=ggnewscale

Cariboni, J., Gatelli, D., Liska, R., & Saltelli, A. (2007). The role of sensitivity analysis in ecological modelling. *Ecological Modelling*, *203*(1-2), 167–182. https://doi.org/10.1016/j.ecolmodel.2005.10.045

Davey, J., & Blaxter, M. (2010). RADSeq: next-generation population genetics. *Briefing in Functional Genomics*, *9*(5-6), 416–423. https://doi.org/10.1093/bfgp/elq031

Diniz-Filho, J. A. F., Soares, T. N., Lima, J. S., Dobrovolski, R., Landeiro, V. L., de Campos Telles, M. P., Rangel, T. F., & Bini, L. M. (2013). Mantel test in population genetics. *Genetics and Molecular Biology*, *36*(4), 475–485. https://doi.org/10.1590/S1415-47572013000400002

Dray, S., Bauman, D., Blanchet, G., Borcard, D., Clappe, S., Guénard, G., Jombart, T., Larocque, G., Legendre, P., Madi, N., & Wagner, H. H. (2022). *`adespatial`: Multivariate multiscale spatial analysis* [R package version 0.3-19]. https://CRAN.R-project.org/package=adespatial

Ellis, C., Hodgson, D., Daniels, C., Collins, M., & Griffiths, A. (2017). Population genetic structure in European lobsters: implications for connectivity, diversity and hatchery stocking. *Marine Ecology Progress Series*, *563*, 123–137. https://doi.org/10.3354/meps11957

European Commission. (2022). *Yearly simple tables.* https://eumofa.eu/en/landings-yearly

Feng, X.-J., Jiang, G.-F., & Fan, Z. (2015). Identification of outliers in a genomic scan for selection along environmental gradients in the bamboo locust, Ceracris kiangsu. *Scientific Reports*, *5*, 13758. https://doi.org/10.1038/srep13758

Fetzer, T. (2013). *Computing maritime routes in* R [R-bloggers]. https://www.r-bloggers.com/2013/03/computing-maritime-routes-in-r/

Foll, M., & Gaggiotti, O. (2008). A genome-scan method to identify selected loci appropriate for both dominant and codominant markers: A bayesian perspective. *Genetics*, *180*(2), 977–993. https://doi.org/10.1534/genetics.108.092221

Forester, B. R., Lasky, J. R., Wagner, H. H., & Urban, D. L. (2018). Comparing methods for detecting multilocus adaptation with multivariate genotype–environment associations. *Molecular Ecology*, *27*(9), 2215–2233. https://doi.org/10.1111/mec.14584

García-Marín, J. L., Sanz, N., & Pla, C. (1999). Erosion of the native genetic resources of brown trout in spain. *Ecology of Freshwater Fishes*, *8*(3), 151–158. https://doi.org/10.1111/j.1600-0633.1999.tb00066.x

GEBCO Compilation Group. (2021). *GEBCO 2021 grid*. https://doi.org/doi:10.5285/c6612cbe-50b3-0cff-e053-6c86abc09f8f

Goudet, J., & Jombart, T. (2022). *hierfstat: Estimation and tests of hierarchical F-statistics* [R package version 0.5-11]. https://CRAN.R-project.org/package=hierfstat

Hijmans, R. J. (2021). *geosphere: Spherical trigonometry* [R package version 1.5-14]. https://CRAN.R-project.org/package=geosphere

Hijmans, R. J. (2022). *raster: Geographic data analysis and modeling* [R package version 3.6-3]. https://CRAN.R-project.org/package=raster

Jenkins, T., Ellis, C., & Stevens, J. (2018). SNP discovery in European lobster (Homarus gammarus) using RAD sequencing. *Conservation Genetics Resources*, *11*, 253–257. https://doi.org/10.1007/s12686-018-1001-8

Jenkins, T., Ellis, C., Triantafyllidis, A., & Stevens, J. (2019). Single nucleotide polymorphisms reveal a genetic cline across the north-east Atlantic and enable powerful population assignment in the European lobster. *Evolutionary Applications*, *12*(10), 1881–1899. https://doi.org/10.1111/eva.12849

Jombart, T., & Ahmed, I. (2011). adegenet 1.3-1: new tools for the analysis of genome-wide SNP data. *Bioinformatics*, *27*(21), 3070–3071. https://doi.org/10.1093/bioinformatics/btr521

Jombart, T., Devillard, S., & Balloux, F. (2010). Discriminant analysis of principal components: A new method for the analysis of genetically structured populations. *BMC Genetics*, *11*, 94. https://doi.org/10.1186/1471-2156-11-94

Jombart, T., Pontier, D., & Dufour, A.-B. (2009). Genetic markers in the playground of multivariate analysis. *Heredity*, *102*, 330–341. https://doi.org/10.1038/hdy.2008.130

Kamvar, Z. N., Brooks, J. C., & Grünwald, N. J. (n.d.). Novel R tools for analysis of genome-wide population genetic data with emphasis on clonality. *Frontiers in Genetics*.

Kassambara, A. (2020). *ggpubr: 'ggplot2' based publication ready plots* [R package version 0.4.0]. https://CRAN.R-project.org/package=ggpubr

Kitada, S., Nakamichi, R., & Kishino, H. (2021). Understanding population structure in an evolutionary context: Population-specific F_{ST} and pairwise F_{ST}. *G3 Genes | Genomes | Genetics*, *11*(11), jkab316. https://doi.org/10.1093/g3journal/jkab316

Lai, J., Zou, Y., Zhang, J., & Peres-Neto, P. (2022). Generalizing hierarchical and variation partitioning in multiple regression and canonical analysis using the rdacca.hp R package. *Methods in Ecology and Evolution*, *13*(4), 782–788. https://doi.org/10.1111/2041-210X.13800

Legendre, P., & De Cáceres, M. (2013). Beta diversity as the variance of community data: Dissimilarity coefficients and partitioning. *Ecology Letters*, *16*(8), 951–963. https://doi.org/10.1111/ele.12141

Legendre, P., & Legendre, L. (2012). *Numerical ecology*. Elsevier.

Luu, K., Bazin, E., & Blum, M. G. B. (2016). `pcadapt`: An R package to perform genome scans for selection based on principal component analysis. *Molecular Ecology Resources*, *17*(1), 67–77. https://doi.org/10.1111/1755-0998.12592

Oksanen, J., Simpson, G. L., Blanchet, F. G., Kindt, R., Legendre, P., Minchin, P. R., O'Hara, R., Solymos, P., Stevens, M. H. H., Szoecs, E., Wagner, H., Barbour, M., Bedward, M., Bolker, B., Borcard, D., Carvalho, G., Chirico, M., De Caceres, M., Durand, S., ... Weedon, J. (2022). *`vegan`: Community Ecology Package* [R package version 2.6-4]. https://CRAN.R-project.org/package=vegan

Pannell, D. J. (1997). Sensitivity analysis of normative economic models: Theoretical framework and practical strategies. *Agricultural Economics*, *16*(2), 139–152. https://doi.org/10.1016/S0169-5150(96)01217-0

Pavičić, M., Žužul, I., Matić-Skoko, S., Triantafyllidis, A., Grati, F., Durieux, E., Celić, I., & Šegvić-Bubić, T. (2020). Population Genetic Structure and Connectivity of the European Lobster Homarus gammarus in the Adriatic and Mediterranean Seas. *Frontiers in Genetics*, *11*, 576023. https://doi.org/10.3389/fgene.2020.576023

Pedersen, T. L. (2022). *`ggforce`: Accelerating '`ggplot2`'* [R package version 0.4.1]. https://CRAN.R-project.org/package=ggforce

Pritchard, J. K., Stephens, M., & Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, *155*(2), 945–959. https://doi.org/10.1093/genetics/155.2.945

R Core Team. (2022). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/

Rossi, A., Petrosino, G., Crescenzo, S., Milana, V., Talarico, L., Martinoli, M., Rakaj, A., Lorenzoni, M., Carosi, A., Ciuffardi, L., & Tancioni, L. (2021). Phylogeography and population structure of Squalius lucumonis: A baseline for conservation of an Italian endangered freshwater fish. *Journal for Nature Conservation*, *64*, 126085. https://doi.org/10.1016/j.jnc.2021.126085

Ruiz Miñano, M., While, G. M., Yang, W., C P Burridge, D. S., & Uller, T. (2022). Population genetic differentiation and genomic signatures of adaptation to climate in an abundant lizard. *Heredity*, *128*, 271–278. https://doi.org/10.1038/s41437-022-00518-0

Ruppert, J. L. W., James, P. M. A., Taylor, E. B., Rudolfsen, T., Veillard, M., Davis, C. S., Watkinson, D., & Poesch, M. S. (2017). Riverscape genetic structure of a threatened and dispersal limited freshwater species, the rocky mountain sculpin (Cottus sp.) *Conservation Genetics*, *18*, 925–937. https://doi.org/10.1007/s10592-017-0938-6

Steenbeek, J., Corrales, X., Platts, M., & Coll, M. (2018). Ecosampler: A new approach to assessing parameter uncertainty in ecopath with ecosim. *SoftwareX*, *7*, 198–204. https://doi.org/10.1016/j.softx.2018.06.004

Szczecińska, M., Sramko, G., Wołosz, K., & Sawicki, J. (2016). Genetic diversity and population structure of the rare and endangered plant species Pulsatilla patens (L.) Mill in east central europe. *PLoS ONE*, *11*(3), e0151730. https://doi.org/10.1371/journal.pone.0151730

van Etten, J. (2017). R package `gdistance`: Distances and routes on geographical grids. *Journal of Statistical Software*, *76*(13), 21. https://doi.org/10.18637/jss.v076.i13

Vu, N. T. T., Zenger, K. R., Guppy, J. L., Sellars, M. J., Silva, C. N. S., Kjeldsen, S. R., & Jerry, D. R. (2020). Fine-scale population structure and evidence for local adaptation in australian giant black tiger shrimp (Penaeus monodon) using snp analysis. *BMC Genomics*, *21*, 669. https://doi.org/10.1186/s12864-020-07084-x

Watson, H., McKeown, N., Coscia, I., Wootton, E., & Ironside, J. (2016). Population genetic structure of the European lobster (Homarus gammarus) in the Irish Sea and implications for the effectiveness of the first British marine protected area. *Fisheries Research*, *183*, 287–293. https://doi.org/10.1016/j.fishres.2016.06.015

Weir, B., & Cockerham, C. (1984). Estimating F-Statistics for the Analysis of Population Structure. *Evolution*, *38*(6), 1358–1370. https://doi.org/10.2307/2408641

Whitlock, M. C., & Lotterhos, K. E. (2015). Reliable detection of loci responsible for local adaptation: Inference of a null model through trimming the distribution of F_{ST} (J. L. Bronstein, Ed.). *The American Naturalist*, *186*(S1), S24–S36. https://doi.org/10.1086/682949

Wickham, H. (2007). Reshaping data with the `reshape` package. *Journal of Statistical Software*, *21*(12), 1–20. https://doi.org/10.18637/jss.v021.i12

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., ... Yutani, H. (2019). Welcome to the `tidyverse`. *Journal of Open Source Software*, *4*(43), 1686. https://doi.org/10.21105/joss.01686

Xuereb, A., Benestan, L., Normandeau, É., Daigle, R. M., Curtis, J. M. R., Bernatchez, L., & Fortin, M.-J. (2018). Asymmetric oceanographic processes mediate connectivity and population genetic structure, as revealed by radseq, in a highly dispersive marine invertebrate (Parastichopus californicus). *Molecular Ecology*, *27*(10), 2347–2364. https://doi.org/10.1111/mec.14589

Zhou, S., Liu, Z., Ma, Q., Liu, Y., Zhang, L., Li, X.-D., Wang, Y., Wang, X., Yu, Y., Yu, H.-R., & Zheng, Y. (2022). Sensitivity tests of cosmic velocity fields to massive neutrinos. *Monthly Notices of the Royal Astronomical Society*, *512*(3), 3319–3330. https://doi.org/10.1093/mnras/stac529

Zimmerman, J., S., Aldridge, C., & Oyler-McCance, S. (2020). An empirical comparison of population genetic analyses using microsatellite and SNP data for a species of conservation concern. *BMC Genomics*, *21*, 382. https://doi.org/10.1186/s12864-020-06783-9