Chapter 17

Use of Qualitative Environmental and Phenotypic Variables in the Context of Allele Distribution Models: Detecting Signatures of Selection in the Genome of Lake Victoria Cichlids

Stéphane Joost, Michael Kalbermatten, Etienne Bezault, and Ole Seehausen

Abstract

When searching for loci possibly under selection in the genome, an alternative to population genetics theoretical models is to establish allele distribution models (ADM) for each locus to directly correlate allelic frequencies and environmental variables such as precipitation, temperature, or sun radiation. Such an approach implementing multiple logistic regression models in parallel was implemented within a computing program named MATSAM. Recently, this application was improved in order to support qualitative environmental predictors as well as to permit the identification of associations between genomic variation and individual phenotypes, allowing the detection of loci involved in the genetic architecture of polymorphic characters. Here, we present the corresponding methodological developments and compare the results produced by software implementing population genetics theoretical models (DFDIST and BAYESCAN) and ADM (MATSAM) in an empirical context to detect signatures of genomic divergence associated with speciation in Lake Victoria cichlid fishes.

Key words: Genome scans, Signature of selection, Genotype × phenotype association, Environmental variables, Logistic regression, Cichlid fishes, Seascape genetics

1. Introduction

On the basis of data produced by genome scans, the main approach to identify loci under directional selection – or likely to be linked to genomic regions under directional selection – is to use population genetics theoretical models to detect outlier molecular markers showing a larger genetic differentiation than expected under the neutral hypothesis (1–3).

François Pompanon and Aurélie Bonin (eds.), *Data Production and Analysis in Population Genomics: Methods and Protocols*, Methods in Molecular Biology, vol. 888, DOI 10.1007/978-1-61779-870-2_17, © Springer Science+Business Media New York 2012

An alternative is to establish allele distribution models (ADM) for each examined locus to directly correlate allelic frequencies with the variation of explanatory variables of interest (e.g., environmental predictors such as precipitation, temperature, sun radiation, etc.) (4, 5). In this case, the geographic coordinates (spatial variables) of sampled individuals are used to link molecular data (presence or absence of a given allele at a genetic marker) with existing environmental variables (value of an environmental variable at the location where individuals were sampled). Of course, environmental variables can also be directly recorded with sensors in the field.

Mitton et al. (6) first had the idea to correlate the frequency of alleles with an environmental variable (elevation) to look for a signature of selection in ponderosa pine. They detected significant association between gene frequencies and slopes of different aspects. In another paper also dedicated to ponderosa pine, Mitton et al. (7) discovered that excess of heterozygosity was associated with xeric habitats. Then, Stutz and Mitton (8) applied the same approach to Engelmann spruce and showed that natural selection was varying with soil moisture. At the beginning of the 2000s, Joshi et al. (9) and Skøt et al. (10) implemented such association studies on a broad scale to study adaptation in common plant species. But until then, the number of loci considered remained very low, for instance six AFLP loci analyzed together with temperature data in Skøt et al. in 2002 (10). A few years later, Joost (11) contrasted a higher number of loci (and alleles) with ecoclimatic variables in goat, frog, and brown bear. For the purpose of running many simultaneous univariate logistic regressions, an application named Matsam was developed with Matlab (The MathWorks Inc.) (12). This software was successfully used to study adaptation in pine weevil and sheep (4), in common frog (13), in goat breeds (14), in fish (15), and in plants (16, 17). The results produced by Matsam were compared and/or validated by the application of theoretical population genetics approaches to the same data sets in all publications mentioned above.

In this chapter, we describe the principles of Matsam, its limits, and additional features implemented in a new version released in summer 2010. Then a case study applied to Lake Victoria cichlids illustrates the novel functions.

2. Matsam

In its first version, Matsam computes multiple simultaneous univariate logistic regressions to test for association between allelic frequencies at marker loci and quantitative ecoclimatic variables (12). To ensure the robustness of the method, two statistical tests (likelihood ratio G and Wald) assess the significance of coefficients calculated by the logistic regression function.

The molecular data sets used for analysis are in the form of matrices; each row of the matrix corresponds to a sampled individual, while columns are organized according to the sampled individual's geographic coordinates and contain binary information (1 or 0) related to the genetic information observed at each genetic marker. Dominant biallelic markers (e.g., AFLPs) can be used directly as they provide binomial information. Codominant multiallelic markers (e.g., microsatellites) need to be encoded as described in (4), and this is also the case for codominant biallelic markers (e.g., SNPs) (used in ref. 14).

The initial MATSAM stand-alone application comes with two Excel macros developed in Visual Basic able to (a) automatically process the large amount of results provided and highlight the most significant associations and (b) draw graphs of the logistic functions (sigmoids) corresponding to any pair of genetic markers vs. environmental variables constituting the models.

The second version of MATSAM released in 2010 also includes an upgrade allowing qualitative predictors to be correlated with the presence/absence of alleles. Indeed, many environmental databases available contain nominal or ordinal data that cannot be processed as quantitative variables by MATSAM (e.g., CORINE land cover or FAO soil map).

2.1. Design Variables

For continuous quantitative predictors, logistic regression models contain parameters (β_i) which represent the change in the response (y) according to a change of the predictor (x). For categorical predictors, parameters represent the different categories of a predictor. The predictor x is chosen to exclude or include a parameter for each observation. Hence, it is called a *design* or *dummy* variable. At least design variables need to be defined for categories, groups, or classes. Consequently, the model will in any case be multivariate with at least $m - 1$ parameters (to process m categories).

There are multiple ways to define design variables (18). Presently, MATSAM implements three of them: the *reference*, the *symmetrical*, and the *independent* parametrization. The difference between these different types of parametrization is highly dependent on the conceptual interpretation that is made of the categories of predictors.

2.1.1. Reference Design Variables

In this case, a group (or a category) always has to be set as the reference group. For groups, one is defined as the reference and the other groups are simply an increase of the expected value compared to the reference one. Each defined combination of design parameters will reflect the expected value of y. Thus, the type of design variable has to be carefully chosen regarding what the predictor conceptually represents (e.g., low–medium–high shrub density; see ref. 18).

If the first group is used as a reference, then the practical implementation regarding the expected value gives:

$$E(\Upsilon_1) = \mu,$$

$$E(\Upsilon_2) = \mu + \alpha_1,$$

$$\ldots$$

$$E(\Upsilon_m) = \mu + \alpha_1 + \cdots + \alpha_{m-1}.$$

Thus, the first group will represent the mean value of the groups, and all other groups will represent the mean ± a certain variation. Moreover, the predictor will have to be recoded according to the following scheme:

$$\text{Group 1: } \begin{bmatrix} 1 & 0 & \ldots & 0 \end{bmatrix}$$

$$\text{Group 2: } \begin{bmatrix} 1 & 1 & 0 & \ldots & 0 \end{bmatrix}$$

$$\ldots$$

$$\text{Group } m\text{: } \begin{bmatrix} 1 & \ldots & 1 \end{bmatrix}$$

Finally, all predictor values affiliated to group m will be recoded into a multivariate regression having an intercept and β_i ($i = 1,\ldots,m - 1$) (see an example in Subheading 2.2).

2.1.2. Symmetrical Design Variables

Here, the groups are treated symmetrically. That is to say, it is necessary to define a central group around which the symmetry is distributed:

$$E(\Upsilon_1) = \mu,$$

$$E(\Upsilon_2) = \mu + \alpha_1,$$

$$\ldots$$

$$E(\Upsilon_m) = \mu - \alpha_1 - \cdots - \alpha_u.$$

We need $[m / 2]$ variables to express the relationships between the groups. The recoding scheme is of size $m \times [m / 2]$:

$$\text{Group 1: } \begin{bmatrix} 1 & 0 & \ldots & 0 \end{bmatrix}$$

$$\text{Group 2: } \begin{bmatrix} 1 & 1 & 0 & \ldots & 0 \end{bmatrix}$$

$$\ldots$$

$$\text{Group } m\text{: } \begin{bmatrix} 1 & -1 & \ldots & -1 \end{bmatrix}$$

For example, let us imagine a categorical predictor composed of five groups with the following ordinal values: “very bad,” “bad,” "average," "good," and "very good." The "average" group is the central group in this case and the other groups are distributed around it:

$$E\left(\Upsilon_{very\ bad}\right) = \mu - \alpha_1 - \alpha_2,$$

$$E\left(\Upsilon_{bad}\right) = \mu - \alpha_1,$$

$$E\left(\Upsilon_{average}\right) = \mu,$$

$$E\left(\Upsilon_{good}\right) = \mu + \alpha_1,$$

$$E\left(\Upsilon_{very\ good}\right) = \mu + \alpha_1 + \alpha_2.$$

And the corresponding recoding scheme is:

Group "very bad"	$\begin{bmatrix} 1 & -1 & -1 \end{bmatrix}$
Group "bad"	$\begin{bmatrix} 1 & -1 & 0 \end{bmatrix}$
Group "average"	$\begin{bmatrix} 1 & 0 & 0 \end{bmatrix}$
Group "good"	$\begin{bmatrix} 1 & 1 & 0 \end{bmatrix}$
Group "very good"	$\begin{bmatrix} 1 & 1 & 1 \end{bmatrix}$

Conceptually, μ represents the overall average effect and α_i the group differences. Moreover, the sum of expected values has to be null (18):

$$\begin{aligned} &\left[E\left(\Upsilon_{very\ bad}\right) - \mu\right] + \left[E\left(\Upsilon_{bad}\right) - \mu\right] + \left[E\left(\Upsilon_{average}\right) - \mu\right] \\ &\quad + \left[E\left(\Upsilon_{good}\right) - \mu\right] + \left[E\left(\Upsilon_{very\ good}\right) - \mu\right] \\ &\quad = -\alpha_2 - \alpha_1 - \alpha_1 + \alpha_1 + \alpha_1 + \alpha_2 = 0. \end{aligned}$$

2.1.3. Independent Design Variables

In this last parametrization scheme, each group is independent of the other m groups (corresponding to nominal qualitative variables). It means that there is no intercept in the regression model:

$$E\left(\Upsilon_1\right) = \alpha_1,$$

$$E\left(\Upsilon_2\right) = \alpha_2,$$

$$\ldots$$

$$E\left(\Upsilon_m\right) = \alpha_m.$$

Table 1
Example of the recoding of the location predictor

Location	Location_1	Location_2	Location_3
Central Europe	1	0	0
Southern Europe	1	1	0
Alps	1	1	1

Design variables enable such an implementation by defining recoding values as follows:

$$\text{Group 1: } \begin{bmatrix} 1 & 0 & \dots & 0 \end{bmatrix}$$

$$\text{Group 2: } \begin{bmatrix} 0 & 1 & 0 & \dots & 0 \end{bmatrix}$$

$$\dots$$

$$\text{Group } m\text{: } \begin{bmatrix} 0 & \dots & 0 & 1 \end{bmatrix}$$

This recoding scheme is indeed similar to an identity matrix of size $m \times m$. Furthermore, as there is no intercept, the convergence of the model may become problematical. Indeed, a regression without an intercept becomes less stable (increase of the number of degrees of freedom of the model) and thus the maximum log-likelihood might not converge.

2.2. Example

We introduce here a small example to illustrate how design variables might be used. Let us consider a data set including a categorical predictor called "Location." It is a nominal predictor characterizing the habitat location of animals. It is made of three classes: "Alps," "Central Europe," and "Southern Europe." These location values are purely nominal and cannot be recoded into quantitative values or intervals. Furthermore, the class "Central Europe" will be used as a reference value. This consideration is completely conceptual, but it forces this predictor in a specific way. It implies that all location values will be recoded into three new location predictors (see Table 1) computed automatically by MATSAM. This clearly implies the computation of a multivariate regression:

$$y = \beta_1 \cdot \text{Location}_1 + \beta_2 \cdot \text{Location}_2 + \beta_3 \cdot \text{Location}_3 + \varepsilon.$$

As "Location_1" shows a value 1 everywhere, it substitutes and becomes the model intercept.

2.3. Limitations

Some limitations must be expressed regarding the use of design variables, in particular, the fact that if a predictor has many categories, the number of parameters in the model may be too high. This issue increases the number of degrees of freedom of the model and might result in overfitting problems. Moreover, this overfitting produces unstable estimated standard error (19), which is the consequence of an almost singular variance matrix. This happens most of the time when the maximum log-likelihood does not converge.

To limit the effect caused by the problem mentioned above, one should always try to reduce as much as possible the number of required design variables. To this end, the type of recoding scheme has to be carefully chosen, and the type of recoding should always be the one requiring the lowest number of design variables as possible. As a result, one should first choose the *symmetrical* recoding. If not possible, the second choice should be the *reference* recoding scheme, and at last the *independent* one.

2.4. Additional Improvements

In addition to its capacity to process qualitative variables, Matsam stand-alone application is able to:

(a) Generate a graph for each association model (without any complementary Excel Macro)

(b) Produce histograms to show the allelic frequency at each molecular marker for different values of environmental variables under investigation

(c) Produce the file containing the results with the names of genetic markers and of environmental variables defined by the user to create the input matrix

(d) Characterize the different types of errors that can be generated during the processing of the models (e.g., the model does not converge)

(e) Produce a matrix containing pseudo-R^2 (Efron, MacFadden, Cox & Snell, and Nagelkerke/Cragg & Uhler), Akaike information criterion (AIC) and Bayesian information criterion (BIC) goodness-of-fit indicators for each model

Another important change is that the different parameters to configure the application have to be indicated in a parameter file. The main parameters define the type of qualitative environmental variables (nominal, ordinal) in order to generate the adequate design variables.

Several improvements of Matsam are in progress and will mainly address spatial autocorrelation issues. They include, in particular, the processing and the mapping of Moran's I, of local indicators of spatial autocorrelation (LISA), and of geographically weighted regression (GWR). Moran's I and LISA are classical tools to measure spatial autocorrelation (see http://geoplan.asu.edu/anselin), while GWR is a family of regression models recently developed in which the β coefficients are allowed to vary spatially and therefore permit to reduce residuals (see http://ncg.nuim.ie/ncg/GWR/). The software will also permit to process multivariate models.

The logistic regression-based method developed here can be accurately used to detect statistical associations between genotype at any individual genomic locus with variations of environmental variables, in order to identify loci potentially playing a role in adaptation.

But the same approach can also be used to identify associations between genomic variation and individual phenotypes, to ultimately help reveal genomic regions involved in the genetic architecture of polymorphic characters. In the case of phenotypic traits known to be subject to divergent selection among study populations, this method can then be used to discover genomic signatures of selection associated with each of these traits specifically. The following case study applied to cichlid fishes will provide examples of both cases.

3. Signature of Genomic Divergence Associated with Speciation in Lake Victoria Cichlids

The Lake Victoria cichlid flock is one of the most explosive examples of adaptive radiation, with more than 500 species having evolved during the last 15,000 years. The repetitive occurrence of the same adaptively important traits in unrelated taxa makes the Lake Victoria flock an ideal model system for studying adaptive radiation in shape, ecology, and behavior.

Within the Lake Victoria cichlid radiation, *Pundamilia pundamilia* and *Pundamilia nyererei* are two sympatric sister species, inhabiting the shores of rocky islands and widely distributed in the lake (20). They differ not only in male nuptial coloration but also in other ecological characters, as feeding ecology, depth distribution, photic environment, visual pigment, and female mating preference for male nuptial coloration (21). However, such divergences appear only in near islands with high water transparency, whereas in near islands with low water transparency, genetic differentiation is reduced or absent and intermediate color phenotypes are common or even dominate (22–24) (Fig. 1).

Along the Mwanza Gulf in the Southern part of the lake, the rocky islands show a continuous gradient of water clarity, from turbid in the South to clear in the North, associated with an increased heterogeneity of the light environment. Following this gradient, populations of *Pundamilia* exhibit different stages of speciation, from a single polymorphic panmictic population to well-differentiated sibling species, constituting a "speciation transect" (21). Furthermore, pieces of evidence have been uncovered for divergent/disruptive selection acting on male breeding color and opsin gene variants, as well as on eco-morphological traits (23, 24). Finally, the global pattern of genetic differentiation among populations suggested a parallel divergence between divergent eco-morphs off the shore of each island along the Manza Gulf (23). The example presented here is taken from a larger population genomic study aiming at identifying the dynamics of genomic differentiation along the gradient of speciation in *Pundamilia* (25).

3.1. Method

We studied four replicate pairs of divergent *Pundamilia* populations along this speciation transect using an AFLP genome-scan

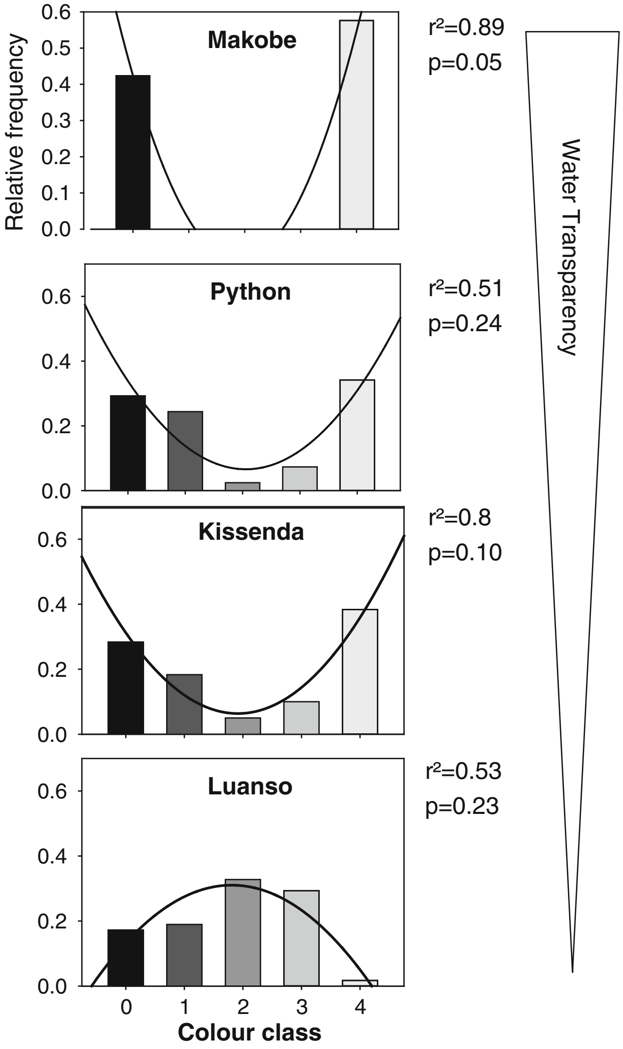


Fig. 1. Distribution of the color phenotypes within the analyzed samples of the two divergent *Pundamilia* species at each of the island studied along the speciation transect. *Fitted curves* are polynomial quadratic fits. The increase in relative frequency of class 2 (intermediate color phenotypes) with the decrease of water transparency and the collapse of populations into unimodal intermediate phenotypic distribution is notable from the quadratic fits.

based on 234 individuals and 520 loci to identify (a) signatures of divergent selection as well as (b) associations between genomic loci and eco-morphological traits under selection. For this purpose, we used a logistic regression approach combined with F_{ST}-outlier-based methods (see Note 1).

The investigation of genotype × phenotype association using logistic regression was conducted with the method implemented in the second version of Matsam v2 (http://www.econogene.eu/software/sam). In parallel, we ran two F_{ST}-outliers detection methods, Dfdist (2, 3) and BayeScan (26). To allow a comparison between the three methods, we conducted a similar analysis independently with each of these software programs, and considered two distinct levels of detection for the divergent loci: one stringent corresponding to "significant" loci and the second–less conservative–corresponding to "marginally significant" loci (see Notes 2 and 3). Here, it is important to stress the fact that the determination of the chosen significance thresholds used to identify "significant" and "marginally significant" loci for each method is not rigorous but based on what is generally admitted in the literature. These significance thresholds are summarized in Table 2 (see Note 3).

Pundamilia males were randomly sampled off the shore of each island. Then individual male nuptial coloration was assessed on the basis of a 5-point color-scale system (Figs. 1 and 2), reflecting the increase of redness of the dorsal color pattern of the individual, ranging from "0" for completely blue phenotype to "4" for completely red-dorsum phenotype (23, 27). Analyses were then carried out based on either (a) the individual male "color-score" or (b) the color-score grouping reflecting "species" morphotypes. First, the analysis based on individual male color-score allows testing the relationship between precise morphotype and genetic variation (with individual-based approach, i.e., "color" variable). Second, the definition of divergent "species" categories based on color-categories (i.e., individuals scored as 0 or 1 included in a "blue group" (*P. pundamilia*), individuals scored as 3 or 4 included in a "red group" (*P. nyererei*), and individuals scored as 2 included in an "intermediate group") allows testing association between morphotype and genotype simultaneously with individual-based and population-based approaches (i.e., "species" variable, Fig. 2). In this second case, only the two extreme phenotype groups, representing divergent species, were used in population-based approaches to quantify the differentiation between divergent eco-morph populations, whereas the "intermediate group" was also considered for the individual-based approach (see Note 4).

We focused on the identification of signatures of divergent selection between the two *Pundamilia* species or eco-morphs. Analyses were then conducted (a) independently within each replicate pair of divergent sympatric populations (i.e., at the island level), to detect outlier loci within each of the four study islands; as well as (b) across all island populations grouped by color-morph (i.e., blue *P. pundamillia* vs. red *P. nyererei*), to detect global outliers over the entire study area. This led to five comparison tests in total. Such pattern of independent replicate divergences across closely related population-pairs with a very low level of hierarchical genetic structure appears particularly suitable for the detection of signatures of selection within and across populations (28, 29) (see Note 5).

3.2. Dfdist and BayeScan Results

Over all five comparison tests, the combination of the two F_{ST}-outlier-based approaches allowed the identification of 49 loci potentially under selection with at least one method (Table 2). Among them, all loci detected with BayeScan were also detected with Dfdist (i.e., 15 loci, representing 31% of the detected outlier loci, Fig. 3),

Table 2
For the three analysis methods used in the *Pundamilia* genome-scan, comparison of (a) analysis parameters, (b) sample sets and grouping/tested variables, and (c) outlier loci detection results presented within and among methods, as well as the level of congruence between all pairs of methods

Methods	F_{ST}-outlier approach		Logistic regression
Software	**Dfdist**	**BayeScan**	**Matsam**
Analysis parameters			
Detection thresholds			
Significant	$P < 0.01$	$\log_{10}(\mathrm{BF}) \geq 1$ (equivalent $P < 0.24$)	$P < 0.05$
Marginally significant	$P < 0.05$	$\log_{10}(\mathrm{BF}) \geq 0.5$ (equivalent $P < 0.09$)	$P < 0.1$
Additional detection parameters	Sequential background F_{ST} estimate	F_{IS} was estimated from microsatellites	Detection with both Wald and G-tests
Sample set and test variables			
- Comparison tests	Separately within each island and across all islands populations ($n = 5$ tests in total)		
- Sample sets	The two groups of extreme morphotypes (excluding intermediate phenotypes)		All individuals (including intermediate)
- Analyzed variables	Species (based on color phenotype)		Species and color (habitat, depth, morphometrics)
Results			
Loci detection			
- per method (signif. + marg. signif.)	49 (17 + 32)	15 (8 + 7)	21 (11 + 10)
- with outlier methods	49		
- with the three methods	55		
Repeated detection			
- between pairs of populations	2	1	1
- across populations	11	5	5
- between Dfdist and BayeScan	15 (31%)		
- between Dfdist and Matsam	15 (27%)		
- between BayeScan and Matsam		11 (44%)	

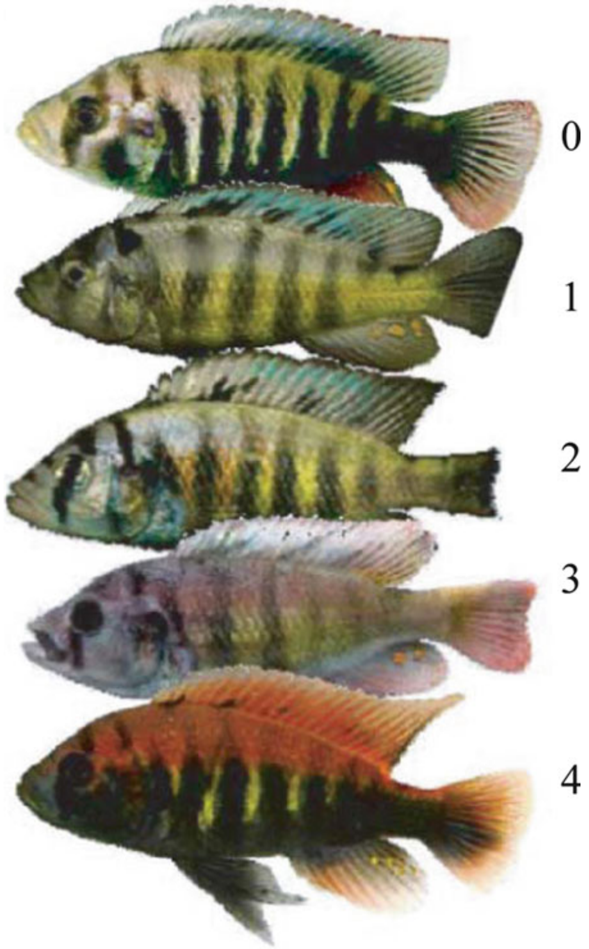
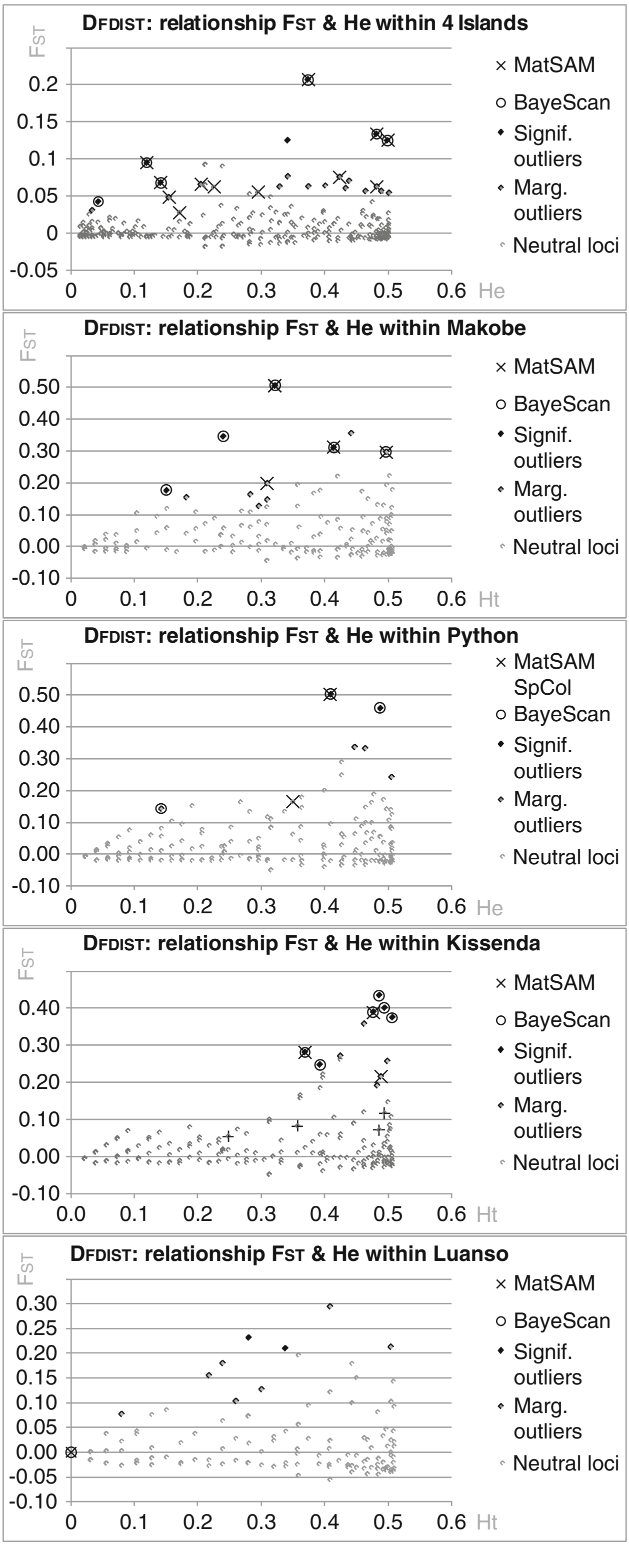


Fig. 2. Representation of the different male nuptial color phenotypes occurring in the two divergent *Pundamilia* species along the transect of speciation of the Mwanza Gulf; five discrete color-categories have been described relative to the increase of redness of the pattern; then the five color-scores correspond to (0) totally blue phenotype (absence of yellow and red colors), (1) yellow coloration on the flank (absence of red), (2) yellow flank with the presence of red along the lateral line, (3) yellow flank with a partially red dorsum, and (4) totally red dorsum. The color of the anal fin is not taken into account for the attribution of the color-score (for more details, see refs. 21 and 27).

Fig. 3. Relationship between heterozygosity (He) and locus-specific F_{ST} between the divergent *Pundamilia* populations computed using Dfdist, conducted independently overall four islands with divergent populations grouped by color-morphs across islands (**a**) as well as separately within each island (**b**–**e**). The loci detected by each of the three analysis methods are indicated: for Dfdist outliers at $P > 0.99$ and $P > 0.95$ are represented in *gray* and *black*, respectively; outlier loci detected by BayeScan with $\log_{10}$(BF) 0.05 are marked by a *circle*; and loci for which an association with a phenotypic variable has been detected using Matsam are marked by a *cross* (for Kissenda Island, association with depth and morphometric variables are represented separately).


DFDIST: relationship FST & He within 4 Islands
FST
0.2
0.15
0.1
0.05
0
-0.05
0
0.1
0.2
0.3
0.4
0.5
0.6
He
× MatSAM
○ BayeScan
◆ Signif. outliers
◆ Marg. outliers
◦ Neutral loci
DFDIST: relationship FST & He within Makobe
FST
0.50
0.40
0.30
0.20
0.10
0.00
-0.10
0
0.1
0.2
0.3
0.4
0.5
0.6
Ht
× MatSAM
○ BayeScan
◆ Signif. outliers
◆ Marg. outliers
◦ Neutral loci
DFDIST: relationship FST & He within Python
FST
0.50
0.40
0.30
0.20
0.10
0.00
-0.10
0
0.1
0.2
0.3
0.4
0.5
0.6
He
× MatSAM SpCol
○ BayeScan
◆ Signif. outliers
◆ Marg. outliers
◦ Neutral loci
DFDIST: relationship FST & He within Kissenda
FST
0.40
0.30
0.20
0.10
0.00
-0.10
0.0
0.1
0.2
0.3
0.4
0.5
0.6
Ht
× MatSAM
○ BayeScan
◆ Signif. outliers
◆ Marg. outliers
◦ Neutral loci
DFDIST: relationship FST & He within Luanso
FST
0.30
0.25
0.20
0.15
0.10
0.05
0.00
-0.05
-0.10
0
0.1
0.2
0.3
0.4
0.5
0.6
Ht
× MatSAM
○ BayeScan
◆ Signif. outliers
◆ Marg. outliers
◦ Neutral loci

reflecting the relatively lower stringency of Dfdist, especially when the background genomic differentiation between population is low (26, 30, 31). Then 7–12 loci were detected with either one or both methods at the island level and 21 loci across all islands (Table 3 and Fig. 3). Furthermore, 11 outlier loci were detected repeatedly within a specific island as well as across all islands, and only two loci (i.e., 2G225 and 1G117) were detected repeatedly in two different islands, representing, respectively, 22% and 4% of the detected outlier loci (see Table 2).

3.3. Matsam Results

Logistic regression was conducted over the four study islands to identify associations between genomic loci and male nuptial coloration or species belonging. Additionally, within Kissenda Island, logistic regression was also conducted to test for genotypic association with habitat depth and 12 morphological variables.

Morphometric characters were considered as strictly continuous variables, color and species variables were encoded and analyzed as categorical ordinal variables (symmetrical parametrization, i.e., distribution of categories around the intermediate phenotype). For the purpose of methodological testing, habitat depth was analyzed either as a quantitative continuous or as a categorical variable (in the latter case with the reference set to 0 m). This allowed comparing the statistical power of multivariate logistic regression models - implied when using categorical variables (see Subheading 2.1) - with univariate models commonly used for continuous variables and theoretically expected to provide a higher detection power. Among all variables and comparison tests, the observation of a high proportion of detected associations with both univariate and multivariate models (61%) - while in 30% of the cases association was only detected in the context of univariate models and 9% in the context of multivariate models only - is in accordance with the slight reduction of detection power when using multivariate models.

Over the five comparison tests, associations were detected with 21 loci at significant or marginally significant levels for at least one model (Table 2): 17 loci with species or color variables, two with depth, and five with morphometric characters (see Fig. 4).

Furthermore, even if the majority of the loci showed an association with only one category of variable, three loci simultaneously exhibited an association with the color phenotype and a morphometric trait, probably due to the statistic association of these characters in the populations. The absence of co-association with any other type of character for the two loci associated with habitat depth (and their lack of detection as outlier loci, see Fig. 3) suggests a relatively wider independence of individual habitat adaptation from species belonging, compared to other phenotypic characters.

Zero to six loci were detected within each of the island-specific analyses, and 12 loci were detected across islands (Table 3 and Fig. 3). Furthermore, five loci were detected repeatedly at the

Table 3
Summary of environmental characteristics at the four islands along the Mwanza gulf transect, number of sampled individuals for each island, estimators of genetic diversity and differentiation, and number of potentially divergent loci detected by each method as well as by all three methods, and finally estimate of the fraction of genomic loci under selection

Locality	Makobe	Python	Kissenda	Luanso	All_Islands	Total
Code	**Ma**	**Py**	**Ks**	**Lu**	**2Col**	**Loci**
Islands environmental characteristics						
Water transparency[a]	225 ± 67	96 ± 21	78 ± 24	50 ± 10	–	
Light slope[b]	8×10^{-3}	7.6×10^{-2}	7.9×10^{-2}	9.6×10^{-2}	–	
Sample sets (number of individuals)						
Pundamilia nyererei	34	30	29	26	119	
Pundamilia pundamilia	25	26	28	14	93	
Intermediate	0	1	3	18	22	
Island community	59	57	60	58	234	
Genetic diversity and differentiation						
Number of polymorphic loci ($P < 0.99$)	382	394	334	308	369	520
Detected divergent loci						
Dfdist	12	7	12	10	21	49
BayeScan	6	3	6	0	6	15
MATSAM	5	2	8 (+4)	0	12	21
Across all methods	12	8	13 (+4)	10	24	55
Percentage of divergent loci	2.31%	1.54%	2.5% (3.27%)	1.92%	4.62%	11%

[a] Secchi depth in centimeters

[b] The light slope is the steepness of the light gradient. It is calculated by regressing the transmittance orange ratio against the mean distance (inmeters) from the shore, measured along the lake floor in transects

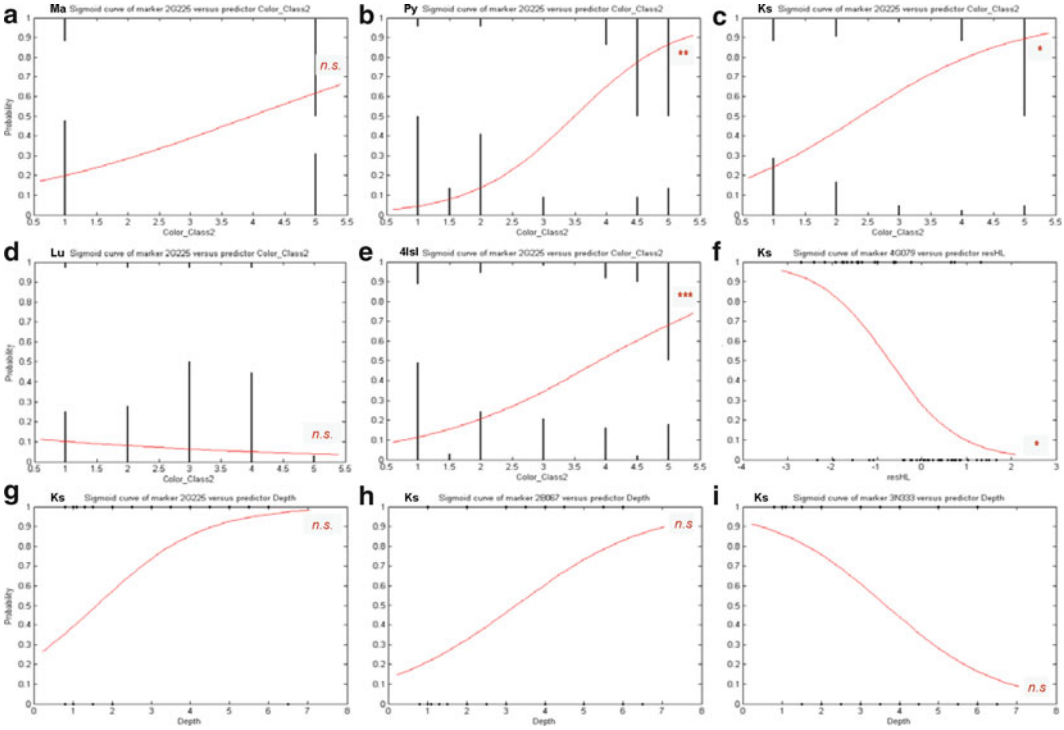


Fig. 4. Representation of different models of logistic regression to investigate association between locus genotype and phenotypic variables conducted with MATSAM; (**a**–**e**) test of association between genotype at locus 2G225 and individual color-score (considered as ordinal variable, ranging from 1 to 5) estimated independently within each island and across all islands; then within Kissenda Island populations, (**f**) test of association between genotype at locus 2G225 and a morphometric variable divergent between eco-morphs, the head length (HL), and (**g**–**i**) test of association between genotype at three loci (2G225, 2B067, and 3N333) and habitat depth. Respective levels of significance of the association are indicated for each model (****P*< 0.001, ***P*< 0.01, **P*< 0.05, applying Bonferroni correction; *n.s.* nonsignificant).

island level as well as across island comparison tests, and only one locus (i.e., 2G225) was detected repeatedly in two different island comparison tests, representing, respectively, 24% and 5% of the detected outlier loci.

3.4. Discussion

When comparing the F_{ST}-outlier and logistic regression approaches, a very high proportion of loci potentially under selection was also significantly associated with at least one phenotypic variable studied, representing 78.6% of the loci detected by both F_{ST}-outlier methods. All genotype × phenotype associations detected here strictly involved phenotypic variables previously identified to be under divergent selection between the two *Pundamilia* species (24). This supports the capacity of the logistic regression approach to (a) identify genomic loci under selection by association with characters targeted by selection, and then to (b) identify loci involved in the genetic architecture of these traits.

This high proportion of genomic loci exhibiting both a signature of selection and association with divergent color-morph characters (73% of significant and marginally significant loci) suggests a predominant action of selection on male nuptial coloration in the divergence between *P. pundamilia* and *P.nyererei*, as further demonstrated by detailed population genomics study (25).

4. Notes

1. Aside from F_{ST}-outlier-based methods, logistic regression can also be used to identify genomic signature of selection, when testing associations between genomic variation and traits identified previously as subjected to selection (e.g., by F_{ST}/Q_{ST} analysis (32, 33)). Rather than identifying the signature of selection based on the comparison of local (locus-specific) differentiation with background genomic differentiation, logistic regressions can identify genomic regions associated with the different targets of selection.

   On condition that traits under selection were identified among study populations or species, this provides two sets of methods based on different assumptions, which can be used complementarily to identify signatures of selection. The congruent detection with both methods can then be taken as a strong support factor (i.e., low probability to detect false positives by both methods).

2. The difference in detection power between F_{ST}-outlier methods and ADM (i.e., logistic regression) is primarily due to the fact that the former are population-based approaches, while the later are individual-based approaches. In F_{ST}-outlier methods, the number of discrete populations sampled is particularly important to depict the global pattern of differentiation within the system, while the number of individuals (total or per population) will mostly affect the accuracy (i.e., confidence interval) of the differentiation estimators. On the other hand, in logistic regression, the detection power is mostly affected by the total number of individuals analyzed and their continuous distribution among the range of the values of the tested variables. Indeed, sampled individuals have to be continuously and homogeneously distributed over the study area, in order to (a) ensure a maximum environmental representativeness and maximize the chance to encompass contrasted environments and (b) avoid superimposition of similar associations (see Note 5). Consequently, more accuracy and statistical power is expected in ADM when carrying out analysis on a global scale than on a regional or local one.

3. As significance is estimated from very different statistical tests in each of the three approaches used, it appears difficult to determine objectively detection thresholds a priori for each method to allow similar stringency among them. In the present study, we selected empirical significant thresholds generally used in similar studies.

   The method implemented in MATSAM implies the realization of numerous logistic regression tests between all possible loci and explanatory variables. Such a design then requires correcting the raw significance threshold for multiple testing. The Bonferroni correction was chosen due to its conservativeness, expected to minimize the false discovery rate. However, this correction stringency is also likely to reduce the true discovery rate, especially if numerous loci are analyzed. To compensate these two antagonist parameters, it is necessary to explore the results obtained for a large and systematic number of different significance thresholds and observe the number of significant associations well beyond the classic 95% and 99% significance thresholds, and also with lower significance thresholds (correction for multiple comparisons always included). Alternatively, other correction methods for multiple testing are available as referenced in (13) and could be tested in this context.

4. F_{ST}-outlier methods and logistic regressions allow us to detect signatures of selection according to different types of "grouping variables." On the one hand, F_{ST}-outlier methods are based on interpopulation divergence measurements and require the analysis of "grouping variables" allowing a strict assignment of individuals to produce discrete populations (i.e., absence of intermediate individuals). On the other hand, logistic regression allows the analysis of "grouping variables" (i.e., predictors), which could also have a continuous or even overlapping distribution among populations. This is reflected by two variables in our case study. First, considering the "species" variable, individuals were assigned to three categories based on their nuptial coloration, the two extreme ones representing the strictly alternative morph/species groups, whereas the third (minority) one represented intermediate morphs. All three categories were then included in the analysis with logistic regression, while only the two extreme ones were included in the F_{ST}-based analyses (i.e., the intermediate one being excluded). Second, the analysis of the "habitat depth" was only possible with the logistic regression approach, due to the fact that individuals show a fully continuous and a widely overlapping distribution between the two species under study (i.e., it is impossible to cluster individuals into unambiguous discrete categories).

5. The existence of a hierarchical genetic structure among studied populations can generate an increase of the false discovery

rate if it is not taken into account, as previously shown in the case of F_{ST}-based approaches (34). Similar patterns are expected with logistic regression when the variability along explanatory variables cannot be disentangled from a strong (hierarchical or not) population genetic structure. In our case study, we detected a lower genetic divergence between sympatric populations than between allopatric conspecific ones. This genetic pattern of differentiation suggests independent replicated island-specific divergence. In such a case, adaptive divergence and background genetic structure are not confounded even at the global level, and consequently permits to obtain reliable results.

Acknowledgments

We would like to thank Geoffrey Dheyongera, Salome Mwaiko, and Isabel Magalhaes for their participation to the project. We are also grateful to Martine Maan, Alan Hudson, Kay Lucek, and Mathieu Foll for their contribution to fruitful scientific discussions about the issues presented in this chapter.

References

1. Lewontin RC, Krakauer J (1973) Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms. Genetics 74:175–195
2. Beaumont MA, Nichols RA (1996) Evaluating loci for use in the genetic analysis of population structure. Proc R Soc Lond B 263: 1619–1626
3. Beaumont MA, Balding DJ (2004) Identifying adaptive genetic divergence among populations from genome scans. Mol Ecol 13:969–980
4. Joost S, Bonin A, Bruford MW et al (2007) A spatial analysis method (SAM) to detect candidate loci for selection: towards a landscape genomics approach to adaptation. Mol Ecol 16:3955–3969
5. Poncet BN, Herrmann D, Gugerli F et al (2010) Tracking genes of ecological relevance using a genome scan in two independent regional population samples of *Arabis alpina*. Mol Ecol 19:2896–2907
6. Mitton JB, Linhart YB, Hamrick JL, Beckman JS (1977) Observations on genetic structure and mating system of ponderosa pine in Colorado Front Range. Theor Appl Genet 51:5–13
7. Mitton JB, Sturgeon KB, Davis ML (1980) Genetic differentiation in ponderosa pine along a steep elevational gradient. Silvae Genet 29:100–103
8. Stutz HP, Mitton JB (1988) Genetic variation in *Engelmann spruce* associated with variation in soil moisture. Arctic Alpine Res 20: 461–465
9. Joshi J, Schmid B, Caldeira MC et al (2001) Local adaptation enhances performance of common plant species. Ecol Lett 4:536–544
10. Skøt L, Hamilton NRS, Mizen S, Chorlton KH, Thomas ID (2002) Molecular genecology of temperature response in *Lolium perenne*: 2. association of AFLP markers with ecogeography. Mol Ecol 11:1865–1876
11. Joost S (2006) The geographic dimension of genetic diversity: a GIScience contribution for the conservation of animal genetic resources. In: School of Civil and Environmental Engineering (ENAC), p. 178. No 3454, Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne
12. Joost S, Kalbermatten M, Bonin A (2008) Spatial analysis method (SAM): a software tool combining molecular and environmental data

to identify candidate loci for selection. Mol Ecol Resour 8:957–960

13. Joost S, Bonin A (2007) Quantitative geography and genomics: spatial analysis to detect signatures of selection along a gradient of altitude in the common frog (*Rana temporaria*). In: Institute of Geography (ed) 15th European colloquium on theoretical and quantitative geography. University of Lausanne, Montreux, Switzerland
14. Pariset L, Joost S, Marsan PA, Valentini A, Consortium E (2009) Landscape genomics and biased F_{ST} approaches reveal single nucleotide polymorphisms under selection in goat breeds of North-East Mediterranean. BMC Genet 10
15. Tonteri A, Vasemagi A, Lumme J, Primmer CR (2010) Beyond MHC: signals of elevated selection pressure on Atlantic salmon (*Salmo salar*) immune-relevant loci. Mol Ecol 19:1273–1282
16. Parisod C, Joost S (2010) Divergent selection in trailing- versus leading-edge populations of *Biscutella laevigata*. Ann Bot 105:655–660
17. Freeland JR, Biss P, Conrad KF, Silvertown J (2010) Selection pressures have caused genome-wide population differentiation of *Anthoxanthum odoratum* despite the potential for high gene flow. J Evol Biol 23:776–782
18. Dobson AJ, Barnett AG (2008) An introduction to generalized linear models, 3rd edn. CRC. 307p. Boca Raton, Florida 307p
19. Hosmer DW, Lemeshow S (2000) Applied logistic regression, 2nd edn. Wiley Series in Probability and Statistics. 375p. Hoboken, New Jersey
20. Seehausen O, van Alphen JM (1999) Can sympatric speciation by disruptive sexual selection explain rapid evolution of cichlid diversity in Lake Victoria? Ecol Lett 2:262–271
21. Seehausen O (2009) Progressive levels of trait divergence along a "speciation transect" in the Lake Victoria cichlid fish Pundamilia. In: Butlin RK, Schluter D, Bridle J (eds) Speciation and patterns of diversity. Cambridge University Press. Cambridge, UK
22. Seehausen O, van Alphen JJM, Witte F (1997) Cichlid fish diversity threatened by eutrophication that curbs sexual selection. Science 277:1808–1811
23. Seehausen O, Terai Y, Magalhaes IS et al (2008) Speciation through sensory drive in cichlid fish. Nature 455:620–623
24. Magalhaes IS, Mwaiko S, Schneider MV, Seehausen O (2009) Divergent selection and phenotypic plasticity during incipient speciation in Lake Victoria cichlid fish. J Evol Biol 22:260–274
25. Bezault E, Dheyongera G, Mwaiko S, Magalhaes I, Seehausen O. (in prep.) Genomic signature of divergent adaptation along replicated environmental gradients in Lake Victoria cichlid fish
26. Foll M, Gaggiotti O (2008) A genome-scan method to identify selected loci appropriate for both dominant and codominant markers: a Bayesian perspective. Genetics 180:977–993
27. Seehausen O (1997) Distribution of and reproductive isolation among color morphs of a rock-dwelling Lake Victoria cichlid (*Haplochromis nyererei*). Ecol Freshw Fish 6:57
28. Wilding CS, Butlin RK, Grahame J (2001) Differential gene exchange between parapatric morphs of *Littorina saxatilis* detected using AFLP markers. J Evol Biol 14:611–619
29. Campbell D, Bernatchez L (2004) Generic scan using AFLP markers as a means to assess the role of directional selection in the divergence of sympatric whitefish ecotypes. Mol Biol Evol 21:945–956
30. Caballero A, Quesada H, Rolan-Alvarez E (2008) Impact of amplified fragment length polymorphism size homoplasy on the estimation of population genetic diversity and the detection of selective loci. Genetics 179:539–554
31. Nielsen EE, Hemmer-Hansen J, Poulsen NA et al (2009) Genomic signatures of local directional selection in a high gene flow marine organism; the Atlantic cod (*Gadus morhua*). BMC Evol Biol 9 , 276p. doi:10.1186/1471-2148-9-276
32. Spitze K (1993) Population-structure in *Daphnia obtusa* – quantitative genetic and allozymic variation. Genetics 135:367–374
33. Whitlock MC (2008) Evolutionary inference from QST. Mol Ecol 17:1885–1896
34. Excoffier L, Hofer T, Foll M (2009) Detecting loci under selection in a hierarchically structured population. Heredity 103:285–298