

Constrained ordination method as a tool for performing genome scan and environmental genomic studies

Eric Bazin,^{*,1} Michael G. B. Blum,²

¹LECA, Université de Grenoble

²TIMC, Université de Grenoble

***Corresponding author:** E-mail: eric.bazin@univ-grenoble-alpes.fr

Associate Editor:

Abstract

Key words:

1 Introduction

2 Performing genome scan in order to detect
3 genomic region of interest is a common task
4 in population genomic area (Foll and Gaggiotti,
5 2008; Frichot *et al.*, 2013; Luu *et al.*, 2016; Vatsiou
6 *et al.*, 2015). Some methods aim at detecting genes
7 that has suffered from a loss of genetic diversity
8 and increase of linkage disequilibrium following
9 the appearance of a beneficial allele and its spread
10 by the mean of selective sweep. Others aim at
11 picking up alleles with strong correlation with
12 some environmental variable (e.g. Temperature,
13 drought) with the idea that these alleles may
14 confer a selective advantage to the individuals
15 (Coop *et al.*, 2010; Frichot *et al.*, 2013). Finally,
16 other methods aim at detecting genomic region
17 involved in local adaptation process. These region
18 should have an increased differentiation between
19 population because different alleles tend to be

20 beneficial in each environment. Differentiation
21 between population is expected under the
22 hypothesis of geographical isolation. Therefore,
23 this region can be detected by quantifying the
24 level of differentiation using some statistics and
25 detecting the regions with unexpectedly high
26 values. A common statistic and very easily
27 comprehensible in population genetic is F_{st} . Many
28 methods use this parameter as a basis in many
29 different implementation of genome scans (Bazin
30 *et al.*, 2010; de Villemereuil and Gaggiotti, 2015;
31 Foll and Gaggiotti, 2008). These are model based
32 method where parameters such as F_{st} are usually
33 inferred using likelihood or Bayesian methods.
34 This mean that users must have some a priori
35 on their parameter value and the best model
36 that fits their data in order to expect the best
37 from their analysis. However, it is often difficult
38 to get a satisfactory a priori picture of the
39 demographic and population structure of the

40 species one is interested in. Indeed many species

© The Author 2013. Published by Oxford University Press on behalf of the Society for Molecular Biology and Evolution. All rights reserved.
For permissions, please email: journals.permissions@oup.com

are not clearly structured in different populations but more or less show a pattern of isolation by distance without clear geographical barrier to gene flow. One solution would be to use more complex model that better reflect reality but these are difficult to implement in Bayesian framework. Additionally, these latter methods are very time consuming and the increase of both model complexity and the amount of data to analyze in terms of the number of individuals and loci makes them more and more difficult to use. A new path has opened recently with the use of multivariate methods. The idea is to capture the whole genome geographic structure using an ordination method such as ACP. Following this analysis, outliers loci are detected if they have extremely high correlation with one or more ordination axis (Duforet-Frebourg *et al.*, 2014; Luu *et al.*, 2016). These are very efficient methods and simulations have shown that while they are very fast, they show similar efficiency than classical Bayesian method and sometimes perform better when the simulated demographic model drift from the model implemented in bayesian method, usually the island model. For instance Luu *et al.* (2016) have shown their method to be better when population are structured in hierarchical set or in isolation by distance pattern. Nevertheless, one conundrum of such approaches is the difficulty to interpret ordination axis in term of ecological meanings. These are usually tight to geographical axis (latitudinal or longitudinal) but they are not necessary linked to an environmental variable such as Temperature, drought, diet habit, etc. Therefore, when this information exists, it has to be a posteriori used as a mean of interpretation but are not involved in the inference process. It should be recalled that natural selection is the result of a complex set of environmental pressures and that it most often acts on several characters simultaneously and that these characters are encoded by several genes which generally have weak effects. In order to extract the maximum of all available information, it seems therefore necessary to use approaches that are able to compile all kind of variable (e.g. alleles, phenotypic measurement, biotic and abiotic variables). One natural way to overcome this limitation would be to use more sophisticated ordination method than ACP like methods. Constrained ordination methods (i.e. Redundancy Analysis, RDA, Canonical Correspondence Analysis, CCA) are well-known set of approaches in Ecology for instance to explain the species distribution pattern by the mean of environmental data. They have specifically been designed in order to deal with biological complexity. In the population genomic era, it seems that data amount, complexity and heterogeneity is often a limitation to the use of inference methods based on classical population genetic models. Although they are more difficult to interpret, such approaches would be complementary to the model based method

because of their long-term use in ecology and their efficiency on complex and large datasets. These method have sometimes been used in population genomic studies, not as a genome scan but in order to quantify multilocus adaptation to an environmental gradient (De Kort *et al.*, 2014; Hecht *et al.*, 2015; Lasky *et al.*, 2012; Steane *et al.*, 2014). These studies whereby relationships between environmental data and large multilocus data is explored are becoming more and more popular and are often coined as Ecological Genomics or Landscape Genomics studies. However the concept of using constrained

ordination methods to analyse genomic data has never been tested on simulated datasets. This paper aims at filling this gap. First, we show how one can make use of a constrained ordination method namely Redundancy Analysis (RDA) as an efficient and robust genome scan method. We discarded the other constraint ordination methods such as CCA since they are very similar in their principles. RDA has already been used for instance by Lasky *et al.* (2012) to perform genome scan in order to detect loci involved in the adaptation to climate in *Arabidopsis thaliana*. Outliers were identified as SNPs with the greatest squared scores along the first RDA axis (i.e. those in the 0.5 % tail). We build on this idea to develop a comprehensive and robust statistical test that allows to search for outliers on an arbitrary number of RDA axis simultaneously and allows to control precisely for the false discovery

rate. Using simulations, we show that it has better results than PCA-based method. Second, thanks to these simulations, we show that RDA can indeed help to identify important environmental gradient that better explain the adaptive variation in the data. It is therefore a proof of concept of the idea of using constrained ordination method as an environmental genomic tool to identify relevant selective gradient in the environmental data. Finally, to give a concrete illustration of RDA approach in population genomics, we apply this method to the detection of outliers on a real data set.

Material and method

Genome scan

Redundancy analysis (RDA) was first introduced by (Rao, 1964) and is clearly described in (Legendre and Legendre, 2012) section 11.1. It is the direct extension of multiple regression to the modeling of multivariate response data. Typically the data to be analysed are separated in two sets, a response matrix Y of variable to be explained (e.g. species abundance in a set of sites; m sites and n species) and an explanatory matrix X (e.g. a set of environmental variable within each site; m sites and p environment). In the following analysis, species are replaced by loci and sites by individuals. In other word, we wish to project on a reduced space the proportion of variance in genetic difference between individuals which is better explained by environmental data. After this ordination, we follow the Luu *et al.*

(2016) methodology to compute pvalues. First we compute the test statistic by regressing each of the p SNPs by the K ordination axis X_1, \dots, X_K .

$$G_j = \sum_{k=1}^K \beta_{jk} X_k + \epsilon_j, j=1, \dots, p$$

where β_{jk} is the regression coefficient corresponding to the j -th SNP regressed by the k -th ordination axis, and ϵ_j is the residuals vector. To summarize the result of the regression analysis for the j -th SNP, we return a vector of z-scores $z_j = (z_{j1}, \dots, z_{jK})$ where z_{jk} corresponds to the z-score obtained when regressing the j -th SNP by the k -th ordination axis. The test statistic is a robust Mahalanobis distance D computed using `covRob` function of the `robustR` package. D should be Khi^2 distributed after a correction with inflation factor (Luu et al., 2016). Pvalues are computed using K degree of freedom. We use the FDR approach to control for false positives. Qvalue are computed with `qvalueR` package and a loci is considered as an outlier if its qvalue is less than 10%. For the analysis of simulated dataset (see below), we retain the first four ordination axis to compute Mahalanobis distances as they seem to explain most of the variance in the data. To perform the ordination, we use the 10th environmental variables as input in the explanatory matrix. In the following example, we don't use phenotypic informations since these informations are often lacking in environmental genomics. Neither we use geographical coordinates (i, j) which is sometimes added to control for the geographical covariation in the differentiation pattern (Frichot et al., 2013).

To emphasize the utility of RDA, we compared to `pcadapt` from which the idea of using multivariate method for genome scan is based. On the simulated dataset, we retain $K=3$ axis to compute Mahalanobis distances as it seems to explain the main amount of variance in the data using scatter plots. To control for false positive, we used the same qvalue threshold (i.e. $q=10\%$).

Environmental genomic

Once outliers have been identified, we isolate them in a separate matrix A defining an "adaptively enriched genetic space" as coined by Steane et al. (2014). Following their methodology, we perform a second constrained ordination (RDA) on matrix A against environmental data. The rationale of this analysis is to remove neutral variation before performing ordination in order to have a better picture of which environmental gradients have the strongest association with the adaptive genetic space. On the simulated dataset, we report the R^2 statistics between `env1`, `env2` and `env3` and the first three ordination axis to have an idea of which they are better associated with and if the ordination space succeed in separating the environmental effect on different axis.

Simulations

To test for the efficiency of RDA in population genomic, we performed simulations using `simuPop` python library (Peng and Kimmel, 2005). We compared our approach to `PCAdapt` method

to perform genome scans. Both approach are is high enough (0.1) to expect a limited linkage
equivalent except their ordination method. effect. We have defined 10 different environmental
Finally we use these simulations to evaluate variables. The first one determines the selective
RDA approach as a mean to detect selective pressure on trait 1, the second one on trait 2 and
environmental gradient. A lattice of 8x8 the third one on trait 3. The first environment
populations is simulated (i.e. 64 populations variable is a quadratic gradient coded by function
in total). Each population is initialized with $env1 = -(\cos(\theta) * (i - 3.5))^2 - (\sin(\theta) * (j - 3.5))^2 +$
200 diploid individual with random genotypes. $18, \theta = \pi/2$, i and j being the population
Migration is set to 0.5 so that population indicator on the 8x8 lattice. The second one
structure must be very smooth and genetic is a linear plan gradient coded by function
differentiation must show an isolation by distance $env2 = h * \cos(\theta) * (i - 1) + h * \sin(\theta) * (j - 1) + k$
pattern over the 64 populations. This is where with $h = 2$, $\theta = \pi/4$ and $k = 3$. The third
pcadapt is best designed for. Loci are biallelic (0 environment variable simulates a coarse
or 1) like SNPs. Allele frequency of the whole environment with value $env3 = 2$ for all
population is initialized at 0.5. 1000 loci are populations except population (i,j) = (2,2),
defined. They are separated in 200 chunks of (2,3), (3,2), (3,3), (6,2), (6,3), (7,2), (7,3), (2,6),
5 SNPs in physical linkage with recombination (2,7), (3,6), (3,7), (6,6), (6,7), (7,6), (7,7) for
rate between adjacent loci fixed at 0.1. 3 different which $env3 = 18$. Env4, env5 and env6 have
Traits are coded by a group of 10 different loci. exactly the same equation than env1, env2 and
The first trait is coded by loci 1, 11, 21, ..., 91. env3 respectively. The remaining 4 environment
The trait value is simply the sum of genotype variable are similar to env2 but with different
value and therefore can take value between 0 and value of h and θ . Env7 has $h = 2$, $\theta = 0$ and
20. For the sake of realism, we add to each trait $k = 3$. Env8 has $h = 2$, $\theta = \pi/4$ and $k = 0$. Env9
a random noise (non heritable variation) drawn has $h = 1$, $\theta = \pi/4$ and $k = 4$. Env10 has $h = 0.5$,
from a normal distribution $N(0,2)$. The second $\theta = \pi/4$ and $k = 8$. Graphical representation of
trait is coded by loci 101, 111, ..., 191 and the mean environmental value for environment 1, 2
third is coded by loci 201, 211, ..., 291. Each trait and 3 is given in Fig. ?? . Environment 4, 5 and
is therefore coded by free recombining SNP loci. 6 have respectively the same mean value spatial
In other words, there are 30 coding SNPs among distribution. For a graphical representation of
1000. Selection can have an effect on linked loci, environment 7 to 10, see supplementary material.
for instance, loci 2, 3, 4 and 5 can be impacted Environmental equation gives a mean value of
by selection on locus 1. However, recombination the environment variable. To avoid colinearity

between environments variable, we added noise by drawing an environment value within a normal distribution $N(\mu=env, \sigma=1)$. Fitness for each trait is set to be $-e^{((x-env)^2/(2*\omega^2))}$, x being the quantitative trait value, env the environmental value and ω is defining selection strength and has been set to 10 which in our experience seems sufficient for loci to be often detected. To get the overall fitness for a given individual, fitness associated to each trait are multiplied. Fitness are relative and selection arises on parents and determine their number of offsprings. Simulations are made across 500 generations. At the end of simulation, we sample 10 individuals per population. Therefore, we have a sample of 640 individuals with 1000 SNP-like loci.

Real dataset

The Loblolly pine dataset is a sample of 682 individuals genotyped on 1,730 SNPs selected in ESTs (Eckert *et al.*, 2010). 60 climatic variables were available and summarized by the authors

in the five first axis of a PCA. The first axis, PC1 is mainly linked to latitude, longitude, temperature, and winter aridity. PC2 is linked to longitude, spring-fall aridity, and precipitation. We inputed the missing data using a very simple algorithm implement in function `sing.im` of the R package `linkim` (Lachenbruch, 2011). It imputes the missing value based on the observed data proportions. We used $K=4$ axis to compute Malahanobis distances.

The Chinook salmon consists of 19 703 SNP loci genotyped on 1956 total individuals pooled in 46 collections. Five variables (MigDistKM, StreamOrder, bio03, bio17 and bio18) have been used among 24 different climate and environmental variables because they have been tested as significantly associated with the SNP variation rangewide citepHecht2015. MigDistKM stands for Migration distance from collection site to ocean (km), StreamOrder for Stream Order of collection site using Strahler method, bio03 for Isothermality, bio17 for Precipitation of Driest Quarter (mm) and bio18 for Precipitation of Warmest Quarter (mm). We could have tested more variable but this is just an illustration and is by no mean an extensive study of this species. Since data are pooled, we have randomly created a sample of 100 individuals for each collection based on the allele frequencies to be able to analyze the data following our individual based pipeline. We used $K=4$ axis to compute Malahanobis distances.

Results

Genome scan

When looking at the analysis on one simulation, the `pcadapt` method seems successful at detecting QTL2 SNPs (Fig. 2) but fails at detecting QTL1 and QTL3 SNPs. On the other hand, RDA succeeds at detecting QTL2 SNPs and also some of the QTL1 and QTL2 SNPs (Fig. 3). The ordination seems to correctly detect environmental variable 1 and 3 as drivers of

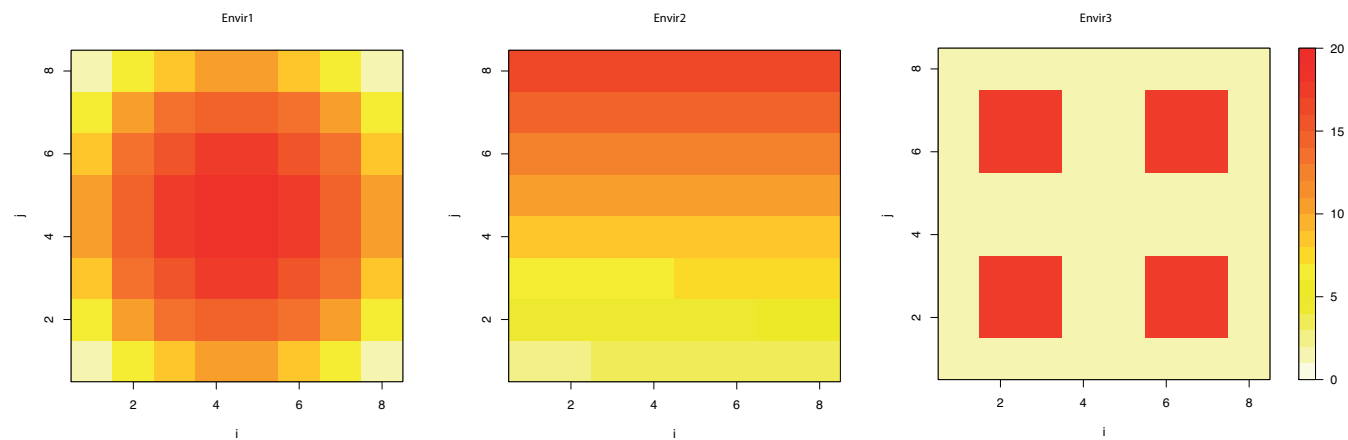


FIG. 1. Graphical representation of mean environmental value for environment 1, 2 and 3

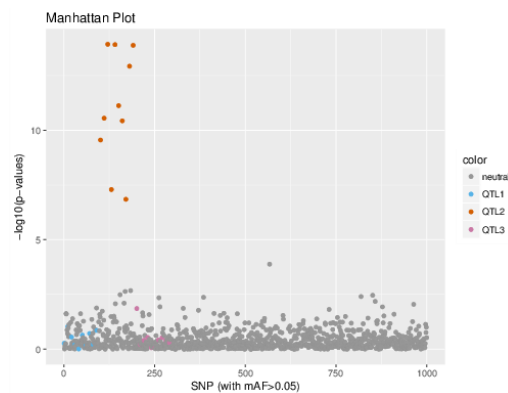


FIG. 2. Manhattan plot of the result of pcadapt on a simulated data set.

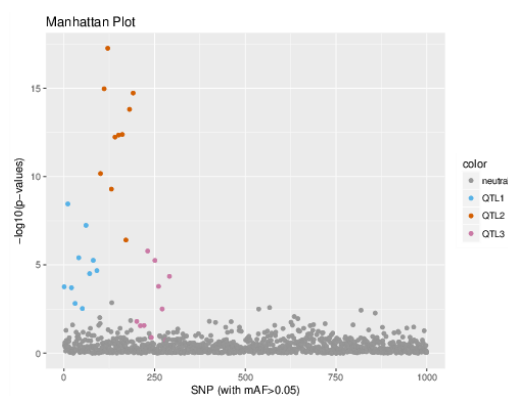


FIG. 3. Manhattan plot of the result of genome scan using RDA on a simulated data set.

360 genetical variance in the data. Over the 100
361 simulations, we have measured the average FDR
362 and power for both pcadapt and RDA (Fig 4).

363 Environmental genomics

364 We then performed a second RDA on the
365 ”adaptively enriched genetic space” as performed
366 by Steane *et al.* (2014) on the same simulated
367 dataset as in Fig. 2 and 3 and display its results on
368 Fig. 5. We did the same analysis and measured the
369 mean R^2 between env1, env2 and env3 and each of
370 the first three ordination axis. This is summarized
371 in Fig. 6.

372 Loblolly Pine

373 Discussion

374 Fig 2 shows that pcadapt approach works well
375 when the environmental gradient and the selective
376 pressures are acting in the same direction than
377 the geographical pattern of isolation by distance.
378 Whereas when the environmental gradient is
379 quadratic on the geographical range (QTL1) or
380 when it is a coarse environment (QTL3). Indeed,
381 we can hypothesize that the PCA ordination fails
382 at orienting the genetic space differentiation into
383 the direction of environment 1 and environment 3

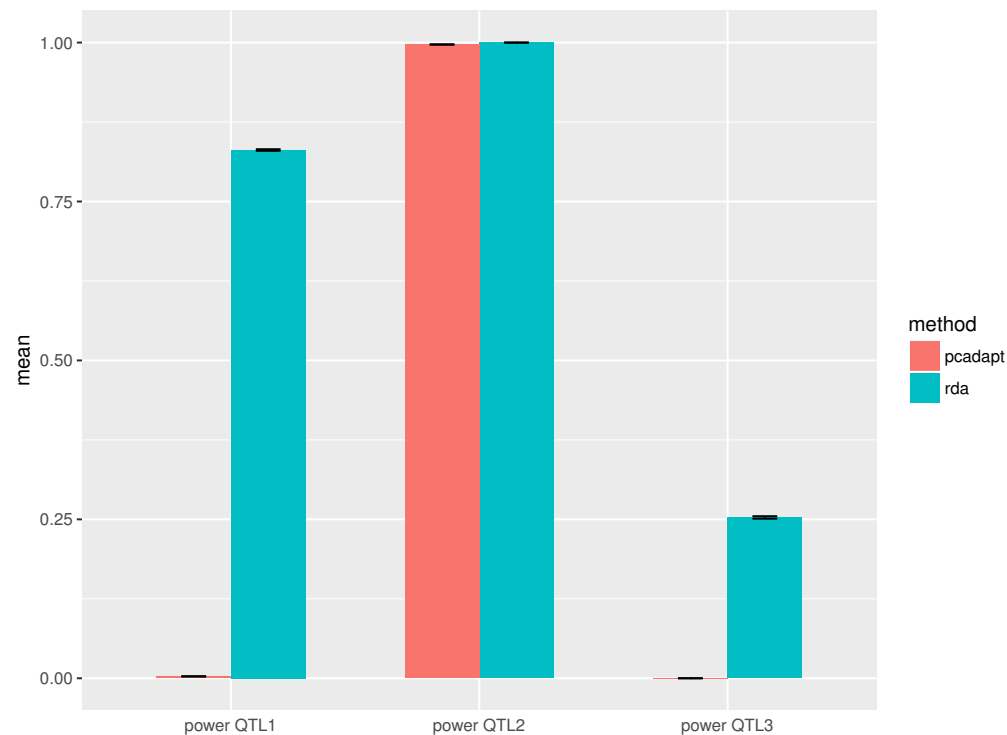


FIG. 4. Performance results of rda and pcadapt methods. Each performance value is averaged over 100 simulated dataset (error bars are displayed but hardly visible since they are very scarce). Power is given separately for loci coding for quantitative trait 1, 2 and 3.

therefore leaving no chance to detect any outliers on the QTL influenced by these environmental variables. Fig 3 shows that RDA has a much better behavior than pcadapt by taking advantage of using informations of environmental local conditions.

Results summarized on Fig 4 is confirming that both methods have a good control of false discovery rate (8.36×10^{-2} for pcadapt and 8.51×10^{-2} for RDA) and that overall RDA shows better performance at detecting true outliers since it succeeds to detect quite often QTL1 and QTL3 SNPs. It seems however less efficient at detecting QTL3 outliers but this might be due to the fact that local adaptation on a coarse environment is more difficult that adaptation on a smooth environmental gradient as environment 1 and

2. These simulations plead in favor of using constrained ordination method instead of PCA when non genetic data such as environmental variable are available in order to orientate the axis in the direction of informative gradients.

When performing an RDA on the "adaptively enriched genetic space", Fig. ?? and ?? show that the method succeed at detecting the relevant selective gradient and separating them on different axis at least on our simulations. This therefore serves as a proof of concept of Steane *et al.* (2014)'s approach to represent multilocus selective gradient and the possibility to use the ordination axis it to devise a metric that provides a holistic measure of genomic adaptation. Indeed, in RDA1 is strongly associated with envr2, RDA2 with envr1 and RDA3 with envr3 whereas poorly

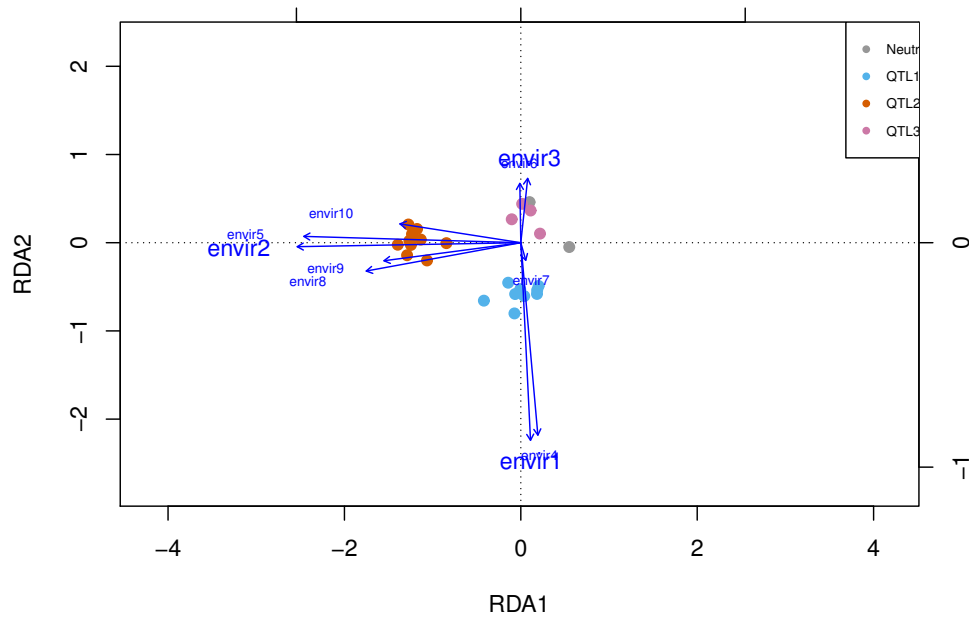


FIG. 5. RDA on the adaptively enriched genetic space. We discarded the individual points for readability. Dots represents outliers SNPs. R^2 of envir1 with the first, second and third axis is (0.02%, 77.5%, 14.5%), envir2 is (99.3%, 0.003%, 0001%) and envir3 is (0.009%, 0.82%, 64.7%)

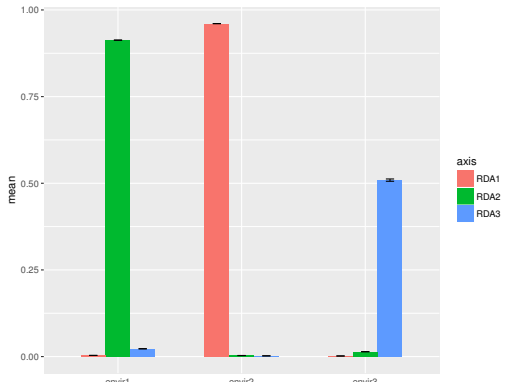


FIG. 6. R^2 between envir1, envir2 and envir3 and each of the first three ordination axis. Values are averaged across the 100 simulated datasets.

associated with the other axis. As expected, the correlated environment are also strongly associated with this respective axis. This is reflecting the fact that in reality it is difficult on an environmental gradient to distinguish among the covariable which one has a causal effect on the individual fitness. However, it is often sufficient for biologists when performing an

exploratory analysis to identify combination of environment variable having a strong association with adaptive variation without knowing precisely the underlying mechanical process.

Supplementary Material

Acknowledgments

References

- Bazin, E., Dawson, K. J., and Beaumont, M. A. 2010. Likelihood-Free Inference of Population Structure and Local Adaptation in a Bayesian Hierarchical Model. *Genetics*, 185(2): 587–602.
- Coop, G., Witonsky, D., Di Rienzo, A., and Pritchard, J. K. 2010. Using environmental correlations to identify loci underlying local adaptation. *Genetics*, 185(4): 1411–23.
- De Kort, H., Vandepitte, K., Bruun, H. H., Closset-Kopp, D., Honnay, O., and Mergeay, J. 2014. Landscape genomics and a common garden trial reveal adaptive differentiation to temperature across Europe in the

- tree species *Alnus glutinosa*. *Molecular ecology*, pages 4709–4721. scans for selection based on principal component analysis. *bioRxiv*, 33: 056135.
- de Villemereuil, P. and Gaggiotti, O. E. 2015. A new FST-based method to uncover local adaptation using environmental variables. *Methods in Ecology and Evolution*, 6(11): 1248–1258. Peng, B. and Kimmel, M. 2005. simuPOP: A forward-time population genetics simulation environment. *Bioinformatics*, 21(18): 3686–3687.
- Duforet-Frebourg, N., Bazin, E., and Blum, M. G. B. 2014. Genome scans for detecting footprints of local adaptation using a Bayesian factor model. *Molecular biology and evolution*, 31(9): 1–13. Rao, C. R. 1964. The Use and Interpretation of Principal Component Analysis in Applied Research. *Sankhy: The Indian Journal of Statistics, Series A*, 26: 329–358.
- Eckert, A. J., Bower, A. D., GonzÁlez-MartÍnez, S. C., Wegrzyn, J. L., Coop, G., and Neale, D. B. 2010. Back to nature: Ecological genomics of loblolly pine (*Pinus taeda*, Pinaceae). *Molecular Ecology*, 19(17): 3789–3805. Steane, D. a., Potts, B. M., McLean, E., Prober, S. M., Stock, W. D., Vaillancourt, R. E., and Byrne, M. 2014. Genome-wide scans detect adaptation to aridity in a widespread forest tree species. *Molecular ecology*, 23(10): 2500–13.
- Foll, M. and 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. Vatsiou, A. I., Bazin, E., and Gaggiotti, O. 2015. A comparison of recent methods for the detection of selective sweeps. *Mol Ecol*, Accepted.
- Frichot, E., Schoville, S. D., Bouchard, G., and François, O. 2013. Testing for Associations between Loci and Environmental Gradients Using Latent Factor Mixed Models. *Molecular biology and evolution*, 30(7): 1687–99.
- Hecht, B. C., Matala, A. P., Hess, J. E., and Narum, S. R. 2015. Environmental adaptation in Chinook salmon (*Oncorhynchus tshawytscha*) throughout their North American range. *Molecular Ecology*, 24(22): 5573–5595.
- Lachenbruch, P. A. 2011. Variable selection when missing values are present: a case study. *Statistical Methods in Medical Research*, 20(4): 429–444.
- Lasky, J. R., Des Marais, D. L., McKay, J. K., Richards, J. H., Juenger, T. E., and Keitt, T. H. 2012. Characterizing genomic variation of *Arabidopsis thaliana*: The roles of geography and climate. *Molecular Ecology*, 21(22): 5512–5529.
- Legendre, P. and Legendre, L. 2012. *Numerical ecology*. Elsevier.
- Luu, K., Bazin, E., Blum, M. G., Bazin, É., and Blum, M. G. 2016. pcadapt: an R package to perform genome