

How to improve PCA based methods of genome scan using ecological data: detecting selection using RDA.

Eric Bazin,^{*,1} Keurcien Luu,² 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

Ordination is a common tool in Ecology that aims at representing complex biological information on a reduced space. For instance, it is frequently used to study geographic distribution pattern of species diversity and to study the link between ecological variable such as temperature, drought, etc, on the species turnover. Recently, these methodologies are becoming quite popular in Landscape Genomic where one wants to study the link between environmental variable and the distribution pattern of genome wide diversity. However, it remains unclear what are the expected outcome of such approaches since genetic diversity has presumably a very different dynamic from species diversity. Simulations studies could help to shed light on this problem but they are still lacking whereas it tends to be broadly accepted as a pertinent approach. Furthermore, recent development have proposed to use ordination methods such as PCA to detect genes under selection. Simulations tend to support this idea as it seems to be quite robust to the underlying population structure and dynamic. Some authors have proposed to use other ordination approaches such as RDA, taking advantage of using environmental data. However no clear statistical framework have been developed to efficiently implement this idea in a robust and efficient test and once again, we don't know what is expected from the outcome of such approaches: which genes will be detected under which selective pressures? This paper aims at proposing a new test based on RDA approaches to search for genes under selection and to compare it to a classical PCA method. Thanks to individual based simulation, we compare both performance and robustness. Additionally, we test the efficiency of constrained ordination method such as RDA to detect relevant selective gradient since this was lacking in the Landscape Genomic literature. Finally, to illustrate the pertinence of such method in concrete example, we apply it to a real dataset.

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

and increase of linkage disequilibrium following
© 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

the appearance of a beneficial allele and its spread are not clearly structured in different populations by the mean of selective sweep. Others aim at but more or less show a pattern of isolation picking up alleles with strong correlation with by distance without clear geographical barrier some environmental variable (e.g. Temperature, to gene flow. One solution would be to use drought) with the idea that these alleles may more complex model that better reflect reality confer a selective advantage to the individuals but these are difficult to implement in Bayesian (Coop *et al.*, 2010; Fricot *et al.*, 2013). Finally, framework. Additionally, these latter methods are other methods aim at detecting genomic region very time consuming and the increase of both involved in local adaptation process. These region model complexity and the amount of data to should have an increased differentiation between analyze in terms of the number of individuals and population because different alleles tend to be loci makes them more and more difficult to use. beneficial in each environment. Differentiation A new path has opened recently with the use between population is excepted under the of multivariate methods. The idea is to capture hypothesis of geographical isolation. Therefore, the whole genome geographic structure using an this region can be detected by quantifying the ordination method such as ACP. Following this level of differentiation using some statistics and analysis, outliers loci are detected if they have detecting the regions with unexpectedly high extremely high correlation with one or more values. A common statistic and very easily ordination axis (Duforet-Frebourg *et al.*, 2014; comprehensible in population genetic is F_{st} . Many Luu *et al.*, 2016). These are very efficient methods methods use this parameter as a basis in many and simulations have shown that while they different implementation of genome scans (Bazin are very fast, they show similar efficiency than *et al.*, 2010; de Villemereuil and Gaggiotti, 2015; classical Bayesian method and sometimes perform Foll and Gaggiotti, 2008). These are model based better when the simulated demographic model method where parameters such as F_{st} are usually drift from the model implemented in bayesian method, usually the island model. For instance inferred using likelihood or Bayesian methods. Luu *et al.* (2016) have shown their method This mean that users must have some a priori to be better when population are structured on their parameter value and the best model in hierarchical set or in isolation by distance that fits their data in order to expect the best pattern. Nevertheless, one conundrum of such from their analysis. However, it is often difficult approaches is the difficulty to interpret ordination to get a satisfactory a priori picture of the axis in term of ecological meanings. These are demographic and population structure of the species one is interested in. Indeed many species usually tight to geographical axis (latitudinal or

longitudinal) but they are not necessary linked to because of their long-term use in ecology and
an environmental variable such as Temperature, their efficiency on complex and large datasets.
drought, diet habit, etc. Therefore, when this These method have sometimes been used in
information exists, it has to be a posteriori used population genomic studies, not as a genome scan
as a mean of interpretation but are not involved but in order to quantify multilocus adaptation
in the inference process. It should be recalled to an environmental gradient (De Kort *et al.*,
that natural selection is the result of a complex 2014; Hecht *et al.*, 2015; Lasky *et al.*, 2012;
set of environmental pressures and that it most Steane *et al.*, 2014). These studies whereby
often acts on several characters simultaneously relationships between environmental data and
and that these characters are encoded by several large multilocus data is explored are becoming
genes which generally have weak effects. In more and more popular and are often coined
order to extract the maximum of all available as Ecological Genomics or Landscape Genomics
information, it seems therefore necessary to use studies. However the concept of using constrained
approaches that are able to compile all kind of ordination methods to analyse genomic data has
variable (e.g. alleles, phenotypic measurement, never been tested on simulated datasets. This
biotic and abiotic variables). One natural way paper aims at filling this gap. First, we show
to overcome this limitation would be to use how one can make use of a constrained ordination
more sophisticated ordination method than ACP method namely Redundancy Analysis (RDA) as
like methods. Constrained ordination methods an efficient and robust genome scan method.
(i.e. Redundancy Analysis, RDA, Canonical We discarded the other constraint ordination
Correspondence Analysis, CCA) are well-known methods such as CCA since they are very similar
set of approaches in Ecology for instance to in their principles. RDA has already been used
explain the species distribution pattern by for instance by Lasky *et al.* (2012) to perform
the mean of environmental data. They have genome scan in order to detect loci involved in
specifically been designed in order to deal with the adaptation to climate in *Arabidopsis thaliana*.
biological complexity. In the population genomic Outliers were identified as SNPs with the greatest
era, it seems that data amount, complexity squared scores along the first RDA axis (i.e. those
and heterogeneity is often a limitation to the in the 0.5 % tail). We build on this idea to
use of inference methods based on classical develop a comprehensive and robust statistical
population genetic models. Although they are test that allows to search for outliers on an
more difficult to interpret, such approaches would arbitrary number of RDA axis simultaneously and
be complementary to the model based method 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 broadly 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.

We retain $K=5$ axis to compute Mahalanobis distances as it seems to explain most of the variance. 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 five 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

201 use geographical coordinates which is sometimes 231 Simulations
202 added to control for the geographical covariation 232 To test for the efficiency of RDA in population
203 in the differentiation pattern (Frichot *et al.*, 233 genomic, we performed simulations using simuPop
204 2013). 234 python library (Peng and Kimmel, 2005). We
205 To emphasize the utility of RDA, we compared 235 compared our approach to PCAdapt method
206 to pcadapt from which the idea of using 236 to perform genome scans. Both approach are
207 multivariate method for genome scan is based. On 237 equivalent except their ordination method.
208 the simulated dataset, we retain for both methods 238 Finally we use these simulations to evaluate
209 $K=5$ axis to compute Mahalanobis distances as 239 RDA approach as a mean to detect selective
210 it seems to explain the main amount of variance 240 environmental gradient. A lattice of 8x8
211 in the data using scatter plots. To control for false 241 populations is simulated (i.e. 64 populations
212 positive, we used the same qvalue threshold (i.e. 242 in total). Each population is initialized with
213 $q=10\%$). 243 200 diploid individual with random genotypes.
244 Migration is set to 0.1 so that population
245 structure must be very smooth and genetic
214 Environmental genomic 246 differentiation must show an isolation by distance
215 Once outliers have been identified, we isolate them 247 pattern over the 64 populations. This is where
216 in a separate matrix A defining an "adaptively 248 pcadapt is best designed for. Loci are biallelic (0
217 enriched genetic space" as coined by Steane *et al.* 249 or 1) like SNPs. Allele frequency of the whole
218 (2014). Following their methodology, we perform 250 population is initialized at 0.5. 1000 loci are
219 a second constrained ordination (RDA) on matrix 251 defined. They are separated in 200 chunks of
220 A against environmental data. The rational of 252 5 SNPs in physical linkage with recombination
221 this analysis is to remove neutral variation before 253 rate between adjacent loci fixed at 0.1. 3 different
222 performing ordination in order to have a better 254 Traits are coded by a group of 10 different loci.
223 picture of which environmental gradients have the 255 The first trait is coded by loci 1, 11, 21, ..., 91.
224 strongest association with the adaptive genetic 256 The trait value is simply the sum of genotype
225 space. On the simulated dataset, we report the 257 value and therefore can take value between 0 an
226 R^2 statistics between env1, env2 and env3 and 258 20. For the sake of realism, we add to each trait
227 the first three ordination axis to have an idea 259 a random noise (non heritable variation) drawn
228 of which they are better associated with and if 260 from a normal distribution $N(0,2)$. The second
229 the ordination space succeed in seperating the 261 trait is coded by loci 101, 111, ..., 191 and the
230 environmental effect on different axis. 262 third is coded by loci 201, 211, ..., 291. Each trait

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
is high enough (0.1) to expect a limited linkage between environments variable, we added noise
effect. We have defined 10 different environmental by drawing an environment value within a normal
variables. The first one determines the selective distribution $N(\mu=env, \sigma=1)$. Fitness for each
pressure on trait 1, the second one on trait 2 and trait is set to be $-e^{((x-env)^2/(2*\omega^2))}$, x being the
the third one on trait 3. The first environment quantitative trait value, env the environmental
variable is a quadratic gradient coded by function value and ω is defining selection strength and
 $env1 = -(\cos(\theta)*(i-3.5))^2 - (\sin(\theta)*(j-3.5))^2 +$ has been set to 20 which in our experience seems
18, $\theta = \pi/2$, i and j being the population sufficient for loci to be often detected. To get
indicator on the 8x8 lattice. The second one the overall fitness for a given individual, fitness
is a linear plan gradient coded by function associated to each trait are multiplied. Fitness
 $env2 = h*\cos(\theta)*(i-1) + h*\sin(\theta)*(j-1) + k$ are relative and selection arises on parents and
with $h=2$, $\theta = \pi/4$ and $k=3$. The third determine their number of offsprings. Simulations
environment variable simulates a coarse are made across 500 generations. At the end
environment with value $env3=2$ for all of simulation, we sample 10 individuals per
populations except population $(i,j) = (2,2)$, population. Therefore, we have a sample of 640
 $(2,3)$, $(3,2)$, $(3,3)$, $(6,2)$, $(6,3)$, $(7,2)$, $(7,3)$, $(2,6)$, individuals with 1000 SNP-like loci.
 $(2,7)$, $(3,6)$, $(3,7)$, $(6,6)$, $(6,7)$, $(7,6)$, $(7,7)$ for Real dataset
which $env3=18$. Env4, env5 and env6 have The *Populus trichocarpa* dataset is a sample of
exactly the same equation than env1, env2 and 424 individuals genotyped on 33,070 SNPs from
env3 respectively. The remaining 4 environment 25 drainages (i.e., topographic units separated
variable are similar to env2 but with different by watershed barriers) (Geraldes *et al.*, 2014).
value of h and θ . Env7 has $h=2$, $\theta=0$ and Genotyping of each accession was performed as
 $k=3$. Env8 has $h=2$, $\theta = \pi/4$ and $k=0$. Env9 described in (Geraldes *et al.*, 2013) using a 34K
has $h=1$, $\theta = \pi/4$ and $k=4$. Env10 has $h=0.5$, Populus SNP array targeting 34,131 SNPs mostly
 $\theta = \pi/4$ and $k=8$. Graphical representation of within (plus 2kb upstream and downstream) 3543
mean environmental value for environment 1, 2 genes. Details of SNP and gene selection can
and 3 is given in Fig. 1. Environment 4, 5 and be found in (Geraldes *et al.*, 2013). 21 climatic

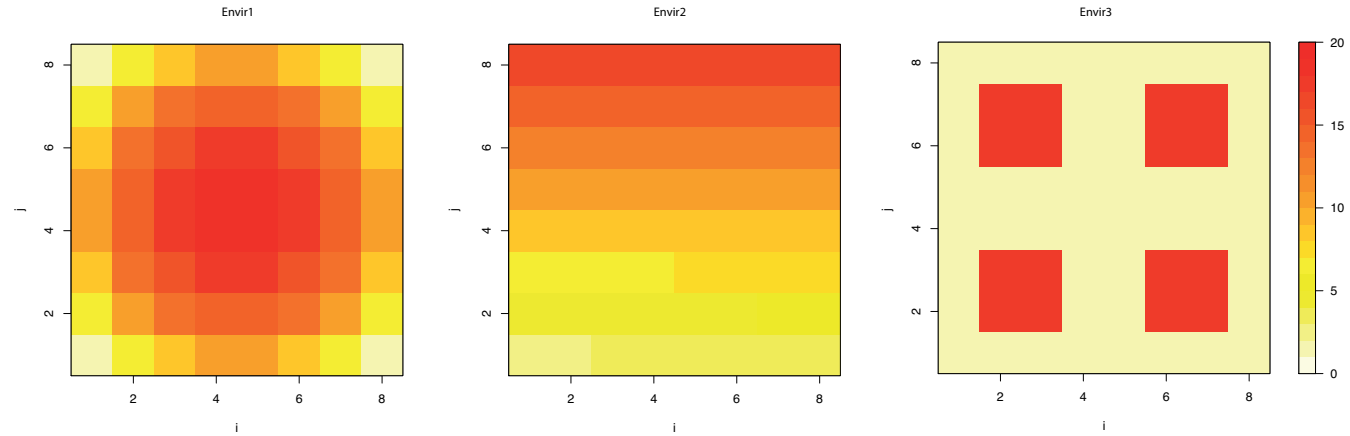


FIG. 1. Graphical representation of mean environmental value for environment 1, 2 and 3

variables are available on each sampling site (details in Table S1). For the RDA and pcadapt analysis, we retained respectively $K=6$ and $K=10$ axis as they seem to explain the majority of variance in the data. From the 33,079 SNPs, we removed the SNPs with missing values, leaving us 19,336 SNPs.

Results

Genome scan

When looking at the analysis on one simulation, the pcadapt method seems successful at detecting QTL2 SNPs (Fig. 2), is less performant at detecting QTL1 SNPs and fails entirely to pick-up QTL3 SNPs. On the other hand, RDA performs better at detecting QTL2 SNPs and retain a larger amount of the QTL1 and QTL2 SNPs (Fig. 3). Therefore, the ordination seems to correctly detect environmental variable 1, 2 and 3 as drivers of genetical variance in the data. Over the 100 simulations, we have measured the average FDR and power for both pcadapt and RDA (Fig 4).

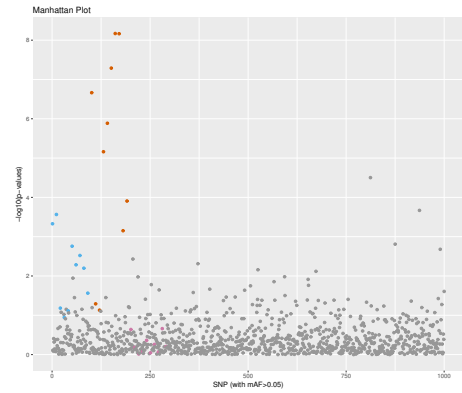


FIG. 2. Manhattan plot of the result of pcadapt on a simulated data set.

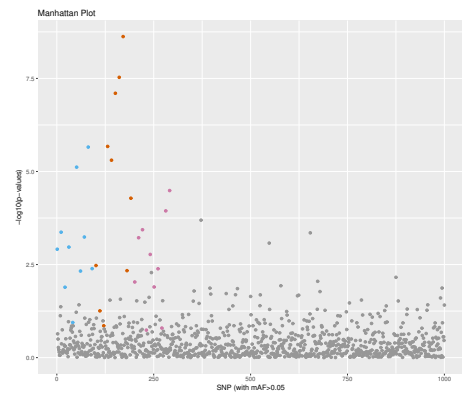


FIG. 3. Manhattan plot of the result of genome scan using RDA on a simulated data set.

Environmental genomics

We then performed a second RDA on the “adaptively enriched genetic space” as performed by Steane *et al.* (2014) on the same simulated

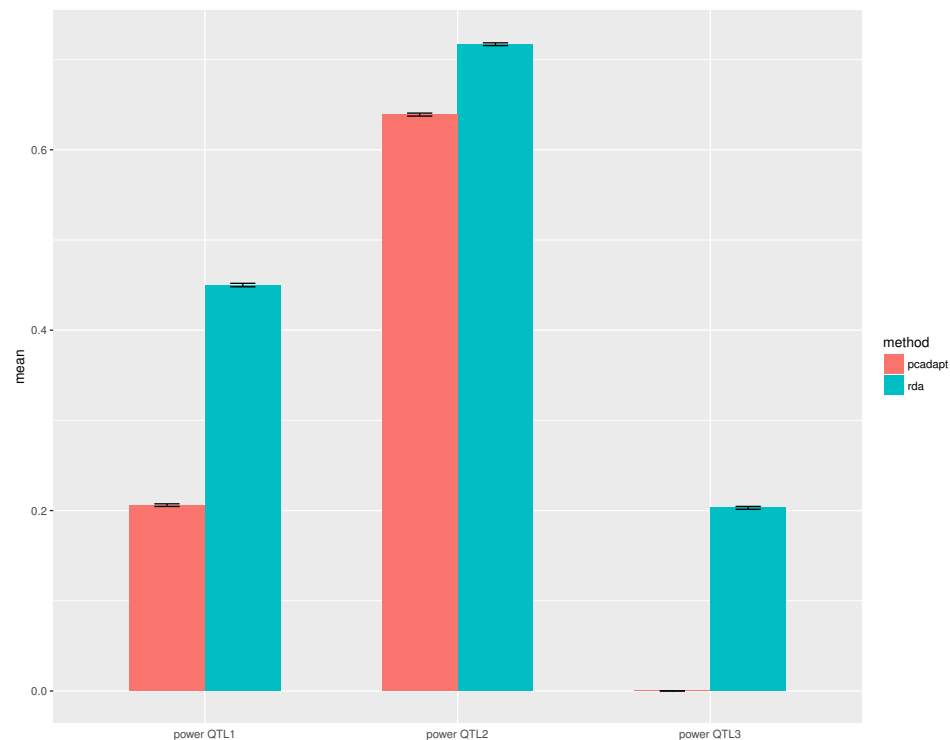


FIG. 4. Performance results of rda and pcadapt methods. Each performance value is averaged over 100 simulated dataset (error bars are displayed but hardly visible since they are very scarce). Power is given separately for loci coding for quantitative trait 1, 2 and 3.

dataset as in Fig. 2 and 3 and display its results on other methods (including pcadapt) had detected as outlier (see Table S2).

Fig. 5. We did the same analysis and measured the mean R^2 between env1, env2 and env3 and each of the first three ordination axis. This is summarized in Fig. 6.

Populus trichocarpa

Analysis of *P. trichocarpa* with fdr of 0.1 leads to a list of 105 outliers for RDA and 122 for pcadapt. Interestingly, both methods have 52 outliers in common so that 53 and 70 SNPs are outliers specific to respectively RDA and pcadapt. When we compared RDA genome scan results with the outliers found by Geraldès *et al.* (2014) based on Fdist, bayescan and bayenv methods, we found that substantial overlap between them. However, RDA based genome scan found 35 SNPs that no

Discussion

Fig 2 shows that pcadapt approach works well when the environmental gradient and the selective pressures are correlated to the geographical pattern of isolation by distance. Whereas when the environment is a coarse environment (QTL3) it fails dramatically. Indeed, we can hypothesize that the PCA ordination in this case is not able to orientate the genetic space differentiation into the direction of environment 3 therefore leaving no chance to detect any outliers on the QTL influenced by this environmental variable. Fig 3 shows that RDA has a much better behavior than pcadapt by taking advantage of using informations of environmental local conditions. It

can be attributed to a better alignment between the genetic space and the environmental variable improving the power to detect true positives.

Both methods have a good control of false discovery rate although slightly better for RDA (12.0×10^{-2} for pcadapt and 10.6×10^{-2} for RDA). Results summarized on Fig 4 is confirming 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 than 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. 5 and 6 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 env1, RDA2 with env2 and RDA3 with env3 whereas poorly 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.

From the analysis of *Populus trichocarpa*, we picked up some genes that can be interpreted regarding to the environmental variable. For instance, among the top SNPs, some of them are located within or nearby, POPTR_0015s00640 implicated in cellular response to cold, POPTR_0015s00440 involved in the regulation of circadian clock in plants. One outlier, POPTR_0017s06840 that is not detected by pcadapt is involved in amylase catabolism, an activity that is suspected to depend of climatic conditions. Interestingly, Table S2 shows that a substantial amount of RDA genome scan outliers (35) are not detected by any other genome scan method including pcadapt. For instance, four SNPs are assigned to POPTR_0007s04340, a dehydration-responsive protein-related which is similar to similar to early-responsive to dehydration stress ERD3 protein. However, its function remains mainly unknown. These results seem to support the

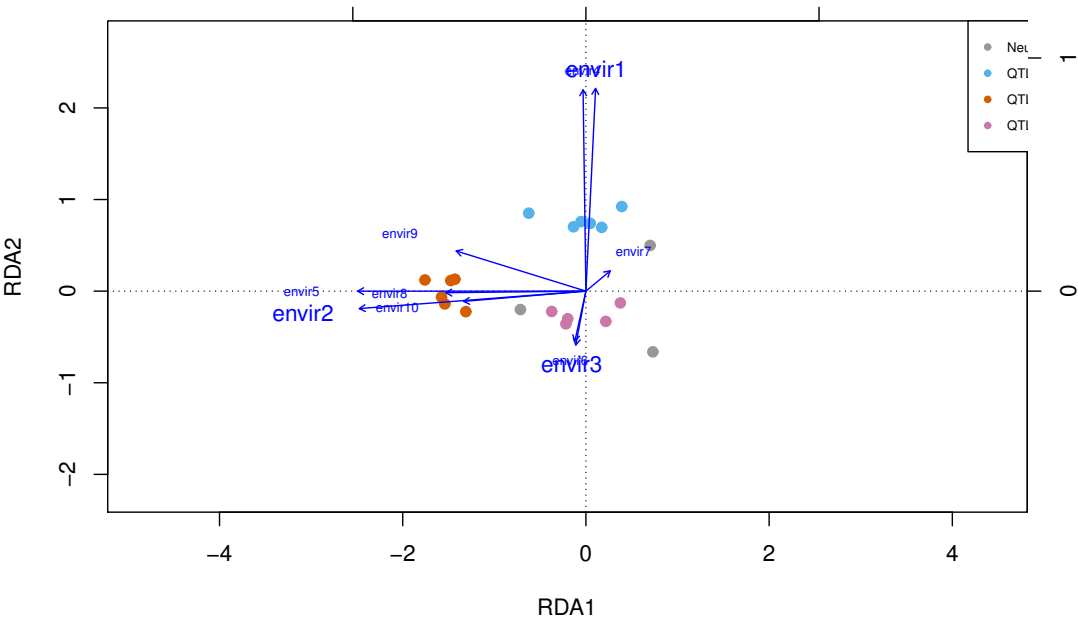


FIG. 5. RDA on the adaptively enriched genetic space. We discarded the individual points for readability. Dots represents outliers SNPs. R^2 of env1 with the first, second and third axis is (0.172%, 77.6%, 17.6%), env2 is (94.5%, 0.560%, 0.236%) and env3 is (0.189%, 5.34%, 90.6%)

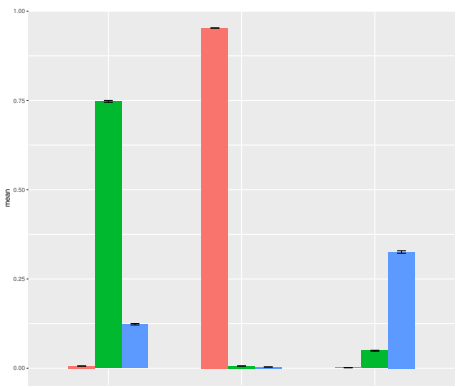


FIG. 6. R^2 between env1, env2 and env3 and each of the first three ordination axis. Values are averaged across the 100 simulated datasets.

simulation analysis conclusions that RDA based genome scan is able to pick up more genes in relation to some environmental gradient than pcadapt. It is however noticeable that a large proportion of outliers are shared between PCA and RDA based methods which highlight the fact that both approaches are very similar. The lack of interpretability of PCA method is compensated

by the fact that it is a blind method in regard to environmental data so it can pick up genes that we will miss using RDA because some crucial environmental data are lacking.

When performing an RDA analysis on the adaptively enriched genetic space (RDA and pcadapt outliers merged together), we noticed that first and second axis of RDA explains a large amount of constrained variance and can be used to identify important environmental gradient (Fig. S2). RDA1 is mainly explained by MAT, DD_0 and DD_18 (Fig. S3) which are temperature related climatic variables and RDA2 is mainly explained by SHM, Eref and CMD which are moisture related climatic variables. Although there is a large covariance between

climatic variable, this analysis proves that RDA can be used to identify and distinguish different environmental gradients that are responsible for the adaptive differentiation across the geographic distribution of *Populus trichocarpa*. This was of course expected since the ordination is constrained by environmental axis and is supported by simulations studies as explained previously.

In conclusion, this paper shows that both methods are complementary and should be used simultaneously on the same dataset. RDA will be more easily interpretable in terms of mechanism and should pick up more genes when some important selective gradient are identified but PCA might be able to pick up outliers where no environmental data are available. However this latter method will miss outliers when selective gradient are poorly correlated to the population structure (i.e. geographical distance between populations and individuals). It is therefore crucial when performing landscape genomics to clearly characterize the environmental condition by measuring important variable that are suspected to put selective pressures on the species of interest. This can be done through RDA as it has been validated in our simulations and in the real dataset on *Populus trichocarpa*.

A myriad of constrained ordination method exist that could be used and allows to control for confounding variable (i.e. partial RDA) for instance altitude, latitude and longitude or allow to perform none linear regression between genetic

and environmental data (i.e. LVM - Latent Variable Model). This is still to be tested and explored using both simulation and real data set validation.

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.
- 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.
- 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.
- 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.
- 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.

- 542 Geraldès, A., Difazio, S. P., Slavov, G. T., Ranjan, P., 581 *Indian Journal of Statistics, Series A*, 26: 329–358.
- 543 Muchero, W., Hannemann, J., Gunter, L. E., Wymore, 582 Steane, D. a., Potts, B. M., McLean, E., Prober, S. M.,
- 544 A. M., Grassa, C. J., Farzaneh, N., Porth, I., Mckown, 583 Stock, W. D., Vaillancourt, R. E., and Byrne, M.
- 545 A. D., Skyba, O., Li, E., Fujita, M., Klápště, J., 584 2014. Genome-wide scans detect adaptation to aridity
- 546 Martin, J., Schackwitz, W., Pennacchio, C., Rokhsar, 585 in a widespread forest tree species. *Molecular ecology*,
- 547 D., Friedmann, M. C., Wasteneys, G. O., Guy, R. D., 586 23(10): 2500–13.
- 548 El-Kassaby, Y. A., Mansfield, S. D., Cronk, Q. C. B., 587 Vatsiou, A. I., Bazin, E., and Gaggiotti, O. 2015. A
- 549 Ehlting, J., Douglas, C. J., and Tuskan, G. A. 2013. 588 comparison of recent methods for the detection of
- 550 A 34K SNP genotyping array for *Populus trichocarpa*: 589 selective sweeps. *Mol Ecol*, Accepted.
- 551 Design, application to the study of natural populations
- 552 and transferability to other *Populus* species. *Molecular*
- 553 *Ecology Resources*, 13(2): 306–323.
- 554 Geraldès, A., Farzaneh, N., Grassa, C. J., Mckown, A. D.,
- 555 Guy, R. D., Mansfield, S. D., Douglas, C. J., and
- 556 Cronk, Q. C. B. 2014. LANDSCAPE GENOMICS
- 557 OF *POPULUS TRICHOCARPA* : THE ROLE OF
- 558 HYBRIDIZATION , LIMITED GENE FLOW , AND
- 559 NATURAL SELECTION IN SHAPING PATTERNS
- 560 OF POPULATION STRUCTURE. pages 3260–3280.
- 561 Hecht, B. C., Matala, A. P., Hess, J. E., and Narum, S. R.
- 562 2015. Environmental adaptation in Chinook salmon
- 563 (*Oncorhynchus tshawytscha*) throughout their North
- 564 American range. *Molecular Ecology*, 24(22): 5573–5595.
- 565 Lasky, J. R., Des Marais, D. L., McKay, J. K.,
- 566 Richards, J. H., Juenger, T. E., and Keitt, T. H.
- 567 2012. Characterizing genomic variation of *Arabidopsis*
- 568 *thaliana*: The roles of geography and climate. *Molecular*
- 569 *Ecology*, 21(22): 5512–5529.
- 570 Legendre, P. and Legendre, L. 2012. *Numerical ecology*.
- 571 Elsevier.
- 572 Luu, K., Bazin, E., Blum, M. G., Bazin, É., and Blum,
- 573 M. G. 2016. pcadapt: an R package to perform genome
- 574 scans for selection based on principal component
- 575 analysis. *bioRxiv*, 33: 056135.
- 576 Peng, B. and Kimmel, M. 2005. simuPOP: A forward-
- 577 time population genetics simulation environment.
- 578 *Bioinformatics*, 21(18): 3686–3687.
- 579 Rao, C. R. 1964. The Use and Interpretation of Principal
- 580 Component Analysis in Applied Research. *Sankhy: The*