How to make use of ordination methods to identify local adaptation: a comparison of genome scans based on PCA and RDA.

Thibaut CAPBLANCQ[1*], Keurcien LUU[2], Michael G.B. BLUM[2] and Eric BAZIN[1,#]

[1] *Univ. Grenoble Alpes, CNRS, LECA UMR 5553, Grenoble, France*
[2] *Univ. Grenoble Alpes, CNRS, TIMC-IMAG UMR 5525, Grenoble, France*
[#]*Deceased*

* Corresponding author: Thibaut Capblancq, 2233 Rue de la Piscine, 38041 Grenoble Cedex, France. Tel: +33 (0)4 76 51 36 71, e-mail: thibaut.capblancq@univ-grenoble-alpes.fr

Date: January 30th, 2018

Running title: RDA-based procedure for genome scans

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, and the species turnover. Recently, ordination methods such as PCA have been used as a framework to detect adaptive variation based on genomic data. Taking advantage of environmental data in addition to genotype data, redundancy analysis (RDA) is another ordination approach useful to detect adaptive variation. These methodologies are specially used in Landscape Genomic where one wants to study the link between environmental variable and the distribution pattern of genome wide diversity, especially its adaptive component. This paper aims at proposing a test based on RDA approaches to search for genes under selection and to compare it to the PCA method implemented in the *pcadapt* package. Individual based simulations identify evolutionary scenarios where RDA genome scan has a greater statistical power than PCA genome scan. By constraining the analysis with environmental variables, RDA performs better than PCA on identifying adaptive variation when selection gradients are weakly correlated with population structure. Additionally, we show that RDA is efficient to identify the main selective gradients among a set of environmental variables. Finally, to give a concrete illustration of RDA in population genomics, we

apply this method to the detection of outliers and selective gradients on a SNP data set of *Populus trichocarpa* in North America (Geraldes et al, 2013). RDA-based approach identifies the main selective gradient that corresponds to a temperature gradient contrasting southern and coastal populations to northern and continental populations in North Western American coast.

KEYWORDS: Genome scans - Multivariate analysis - Redundancy Analysis – *pcadapt* - Selection.

INTRODUCTION

Natural selection is the result of a complex set of environmental pressures. Environment acts on several characters simultaneously and these adaptive characters can be determined by a large number of alleles of small effects (Pritchard and Di Rienzo, 2010). Patterns of adaptive variation within species are usually studied through physiological, morphometric and fitness comparisons in common gardens experiments. However, this is a difficult and constraining task and is even unrealistic on species with long generation time. The recent development of Next Generation Sequencing (NGS) technics has opened the possibility to access to a large amount of genetic variation across the genome. Therefore, relationships between genetic polymorphism, phenotypic variation and environmental variables can now

41 be studied and quantified by in situ approaches (Steane *et al.*, 2014, Vangestel *et al.*,
42 2018).
43
44 The search of adaptive genetic variation can be performed by genome scan
45 methods (Foll and Gaggiotti 2008, Frichot *et al.*, 2013, Vatsiou *et al.*, 2015, Duforet-
46 Frebourg *et al.*, 2016). Among the different genome scan procedures, ecological
47 association methods aim at detecting alleles that are associated with environmental
48 variables (e.g. temperature, drought) with the idea that these alleles may confer a
49 selective advantage in some environment (Coop *et al.*, 2010, Frichot *et al.*, 2013).
50 Finally, methods based on genetic differentiation assume that adaptive alleles have
51 increased genetic differentiation. A common statistic to perform genome scans
52 based on genetic differentiation is F_{st} (Lewontin and Krakauer, 1973). Many
53 likelihood or Bayesian methods use F_{st}-related statistics to scan genomes for local
54 adaptation (Bazin *et al.*, 2010, De Villemereuil and Gaggiotti, 2015, Foll and
55 Gaggiotti, 2008, Whitlock and Lotterhos, 2015).
56
57 F_{st}-based approaches require to define discrete populations. However, many
58 species are not structured in different populations but show a pattern of isolation by
59 distance and display continuous population structure (Martins *et al*. 2016). In these
60 cases, defining populations can be challenging and can directly interfere with the

61 result of the analysis. Instead of defining populations, multivariate methods such as
62 PCA can be used to ascertain population structure. Following this analysis, outlier
63 loci are detected if they have extremely high correlation with one or more ordination
64 axes (Duforet-Frebourg *et al.*, 2016). Simulations have shown that they have similar
65 power than model-based methods under island models and perform better when the
66 simulated demographic model drift from the island model, for instance under
67 hierarchical population structure or isolation by distance patterns (Luu *et al.*, 2017).
68 Nevertheless, one conundrum of such approaches is the difficulty to interpret
69 ordination axes in term of ecological meanings. These are usually tight to
70 geographical axes (latitudinal or longitudinal) but they are not necessary linked to
71 an environmental variable such as Temperature, drought, diet habit, etc. Therefore,
72 when environmental information exists, it has to be used a posteriori as a mean of
73 interpretation but it is not involved in the inference process.

74 However, it is possible to detect the adaptive fraction among genetic variation,
75 which is explained by environmental variation. Some authors (Lasky *et al.*, 2012,
76 DeKort *et al.*, 2014; Steane *et al.*, 2014, Hecht *et al.*, 2015) have proposed to use
77 ordination technics such as RDA (Redundancy Analysis) or CAPs (Canonical Analysis
78 of Principal coordinates) to subsequently align adaptive variation with
79 environmental data in order to be able to identify the environmental gradients that
80 are the most correlated with adaptive variation. In *Arabidopsis thaliana,* loci

81 involved in adaptation to climate have been found using RDA (Lasky *et al.*, 2012).
82 Outliers were identified as SNPs with the greatest squared loadings along the first
83 RDA axes (i.e. those in the 0.5 % tail). RDA can also be used to derive an adaptive
84 index that predicts the performance of individuals in different environmental
85 conditions. Steane *et al.* (2014) applied this approach to *Eucalyptus trichocarpa*
86 populations and showed that this in-situ estimation of adaptive variation is
87 correlated to what can be measured in common garden experiments. Finally, in two
88 recent studies, Forester and colleagues (2015 and 2017) tested the power of RDA
89 based method for detecting signature of local adaptation in a heterogeneous
90 landscape (Forester *et al*, 2015) and in comparison with other constrained
91 ordination method (Forester *et al*, 2017). Using simulations, they found that a RDA
92 based method shows a great potential to identify adaptive loci and, has a superior
93 combination of low false positive and high true positive rates when compared to
94 several other approaches.

95 In this paper, we build on these previous applications of RDA to develop a
96 statistical test for searching loci under selection. We use simulations and real data
97 sets to shed light on the conditions within which RDA is the most efficient and to
98 document the possibilities given by RDA when studying landscape genomics. We
99 also compare results of genome scans based on RDA and based on PCA as it is
100 implemented in the *pcadapt* package (Luu *et al.,* 2017).

101 First, we show that genome scans based on RDA allows to recover more
102 outliers than a genome scan based on PCA because it accounts for environmental
103 variation. Second, we show that RDA can identify important environmental
104 gradients that explain adaptive variation in the data. It is therefore a proof of
105 concept of the idea of using constrained ordination method as an environmental
106 genomic tool to identify relevant selective gradients in the environmental data
107 (DeKort *et al.*, 2014, Steane *et al.*, 2014). Finally, to illustrate the pertinence of such
108 method in concrete example, we apply it to a *Populus trichocarpa* SNP dataset
109 obtained from Geraldes et al (2013). Our results mostly concord with the ones
110 obtained with the *Fst* based methods used in their study but we also show some new
111 genetic markers specifics to the RDA-based approach. Additionally, RDA identifies
112 the main selective gradients along the natural populations of *Populus trichocarpa* in
113 the north-western coast of North America.
114

115 # MATERIALS AND METHODS

116 ## 1. Genome scan using a RAD based approach

117 Redundancy analysis (RDA) was first introduced by Rao (1964). It is the
118 extension of multiple regressions to the modelling of multivariate response data
119 (Legendre and Legendre 2012, section 11.1). The data are separated in two sets, a
120 response matrix Y of variable to be explained (e.g. species abundance in a set of

121 sites; m sites and n species) and an explanatory matrix X (e.g. a set of environmental
122 variable within each site; m sites and p environment). In the following analysis, loci
123 replace species abundances and individuals replace sites. RDA seeks to project on a
124 reduced space the genetic variation between individuals that is explained by
125 environmental data.

126 The analysis starts using a classical RDA procedure performed with the
127 function *rda* of the package R *"vegan"* (Oksanen *et al.*, 2015). A genetic dataset (e.g.
128 SNPs genotypes 0, 1, 2) is used as response matrix Y and a set of environmental
129 variables is used as explanatory matrix X. RDA amounts at constructing a matrix Y' of
130 predicted genotypes based on the environmental variables and at performing
131 principal component analysis on the matrix Y' (Legendre and Legendre 2012, section
132 11.1). After this constrained ordination step, we follow the methodology
133 implemented in *pcadapt* to find outliers loci (Luu *et al.,* 2017). First we recover from
134 the RDA analysis the z-scores obtained when regressing the SNPs by the RDA axes.
135 We obtained vectors of z-scores denoted $z_j = (z_{j1}, ..., z_{jK})$ where z_{jk} corresponds to the
136 z-score obtained when regressing the j-th SNP by the K-th ordination axis. Only the
137 z-scores of the most informative ordination axis are kept for the rest of the
138 procedure. This number (K) is determined by looking at the amount of information
139 retained on the different axes of the RDA. A robust Mahalanobis distance D is then
140 computed for each locus to identify outlier vectors of z-scores. Computation of the

141 robust Mahalanobis distance uses the *covRob* function of the "robust" R package
142 (Wang et al., 2014). Mahalanobis distances are distributed as Khi2 distribution with
143 K degrees of freedom after correcting with the genomic inflation factor (Luu et al.,
144 2017). We use the False-Discovery-Rate (FDR) approach to control for false
145 positives. Qvalue are computed with "qvalue" R package (Dabney and Storey, 2011)
146 and a locus is considered as an outlier if its qvalue is less than 10%. The complete
147 procedure and our script are available in supplementary material (Script S1).

148 To evaluate the capacity of the RDA based approach to detect loci underlying
149 environmental selective pressure we first test the method with a simulated dataset.
150 We look at its capacity to detect positive loci (loci simulated under selection) and to
151 return a controlled number of false positives (neutral loci detected as outliers by the
152 analysis). Then, to emphasize the utility of RDA, we compared it to the PCA-based
153 procedure implemented in *pcadapt* that does not use environmental variables (Luu
154 *et al.*, 2017).

155

156 **2. Simulated data**

157 To test for the efficiency of RDA in population genomics, we used simulations
158 performed with the simuPop python library (Peng *et al.*, 2005). A lattice of 8x8
159 populations is simulated (i.e. 64 populations in total). Each population is initialized
160 with 200 diploid individuals with random genotypes. Migration is set to 0.1 resulting

161 in an isolation-by-distance pattern across the species range. Loci are assumed to be
162 biallelic SNPs. Allele frequency of the whole population is initialized at 0.5. 1000 loci
163 are simulated in total and they are separated in 200 chunks of 5 SNPs in physical
164 linkage with recombination rate between adjacent loci fixed at 0.1. Three different
165 traits are coded by a group of 10 different loci. The first trait is coded by the loci 1,
166 11, 21, ..., 91. The trait value is simply the sum of the genotype values and therefore
167 can take value between 0 and 20. We add to each trait a random noise (non heritable
168 variation) drawn from a normal distribution $N(0,2)$. The second trait is coded by loci
169 101, 111, ..., 191 and the third is coded by loci 201, 211, ..., 291. Each trait is
170 therefore coded by 10 independent SNP loci resulting in a total of 30 causal SNPs
171 among 1000. Selection can have an effect on linked loci, for instance, loci 2, 3, 4 and 5
172 can be impacted by selection on locus 1. However, recombination is high enough
173 (0.1) to expect a limited linkage effect. We have defined 10 different environmental
174 variables. The first one determines the selective pressure on trait 1, the second one
175 on trait 2 and the third one on trait 3. The first environment variable is a quadratic
176 gradient coded by function $env1 = -(\cos(\theta)*(i-3.5))^2 - (\sin\theta)*(j-3.5))^2 + 18$, $\theta = \pi/2$, i
177 and j being the population indicator on the 8x8 lattice. The second one is a linear
178 plan gradient coded by function $env2 = h*\cos(\theta)*(i-1) + h*\sin(\theta)*(j-1) + k$ with h=2,
179 $\theta = \pi/4$ and k=3. The third environment variable simulates a coarse environment
180 with value $env3$ = 2 for all populations except population (i,j) = (2,2), (2,3), (3,2),

181 (3,3), (6,2), (6,3), (7,2), (7,3), (2,6), (2,7), (3,6), (3,7), (6,6), (6,7), (7,6), (7,7) for
182 which $env3$ = 18. Env4, env5 and env6 have exactly the same equation than env1,
183 env2 and env3 respectively except for a noise term explained below. The remaining
184 4 environment variables are similar to env2 but with different value of h and θ. Env7
185 has h=2, θ = 0 and k=3. Env8 has h=2, $\theta = \pi/4$ and k=0. Env9 has h=1, $\theta = \pi/4$ and
186 k=4. Env10 has h=0.5, $\theta = \pi/4$ and k=8. Graphical representation of the mean
187 environmental value is given in Fig. 1. Environmental variables 4, 5 and 6 have
188 respectively the same mean value spatial distribution than environmental variables
189 1, 2 and 3. Environmental equation gives a mean value of the environment variable.
190 To avoid colinearity between environmental variable, we added noise by drawing an
191 environment value within a normal distribution N(μ = env, σ = 1). Fitness for each
192 trait is set to be –exp(x-env)^2/(2*ω^2), x being the quantitative trait value, env the
193 environmental value and ω is defining selection strength and has been set to 20
194 which was found to be enough for loci to be often detected. To get the overall fitness
195 for a given individual, fitness associated to each trait are multiplied. Fitness values
196 are used to determine the number of offsprings during the simulations, which have
197 been run during 500 generations. At the end of simulation, we sample 10 individuals
198 per population resulting in 640 individuals with 1000 SNP-like loci.
199

200 **3. Real dataset**

201 The *Populus trichocarpa* dataset is a sample of 424 individuals genotyped on
202 33,070 SNPs from 25 drainages (i.e., topographic units separated by watershed
203 barriers) (Geraldes *et al.*, 2014). Genotyping of each accession was performed with a
204 34K Populus SNP array targeting 34,131 SNPs mostly within (plus 2kb upstream and
205 downstream) 3543 genes (Geraldes *et al.*, 2013). A total of 21 climatic variables are
206 available on each sampling site (Table S2). From the 33,070 SNPs, we removed the
207 SNPs with missing values and a minor allele frequency below 5% resulting in 17,224
208 SNPs.

209 We explored the adaptive genetic variability of *Populus trichocarpa* dataset
210 using both PCA-based (*pcadapt*) and RDA-based method using the 21 environmental
211 variables as explanatory variables (Table S2). We finally compared the list of outliers
212 found with these two ordination approaches to the ones found using *Fst* based
213 methods (Geraldes *et al.*, 2014).

214 We then performed a second RDA with only the loci previously found as
215 outliers (q.values < 0.1). This set of outliers gives an "adaptively enriched genetic
216 space" (Steane et al., 2014) and a second RDA on these specific loci allows to identify
217 the environmental variables that are the most correlated with "purely" adaptive
218 variation. In the same idea, we used the scores of the different *Populus trichocarpa*
219 populations along the first two RDA axes to build composite indexes that predict the
220 performance of individuals in the environment. The population's scores along one

221 axe of the adaptively enriched genetic RDA can be interpreted as a cursor of
222 population genetic adaptation to the environmental variables associated with this
223 axis. A map representation of these indexes allows for a better understanding of the
224 adaptive landscape in the sampling area.
225

226 # RESULTS

227 ## 1. Genome scan on simulated dataset using RDA or PCA based methods

228 The procedure was first performed on one simulation (sim1) with both PCA
229 and RDA based approaches. To perform RDA, we considered environmental
230 variables 1-10 as explanatory variables. For the two approaches, we retained the
231 first four ordination axes to compute Mahalanobis distances as indicated by the
232 scree plots (Fig. S1).

233 The software package *pcadapt* is successful at detecting QTL2 SNPs (Fig. 2A)
234 but fails entirely at detecting QTL1 and QTL3 SNPs. The RDA-based approach also
235 detects QTL2 SNPS but also detects most of the QTL1 SNPs (Power=75%) by
236 contrast to *pcadapt* (Fig. 2B). For the SNPs associated with QTL3, both methods have
237 negligible power to detect them (PCA Power=0%, RDA Power=25%).

238 When looking at the projection of each SNP in the RDA space we can detect
239 links between QTL SNPs and environmental variables (Fig. 3). RDA1 is strongly
240 correlated to env2 and env5 and the QTL2 SNP show extreme scores on this axis

241 comparing to the other SNPs, especially the neutral ones (Fig. 3A). RDA2 is also
242 associated to env7, env8 and env9, which had not been used as direct selective
243 drivers during the simulations. RDA3 is correlated to env1 and env4 corresponding
244 to QTL1 SNPs and there is also a weak association between RDA4 and the SNPs
245 associated with QTL3 (Fig. 3B & 3C).

246 When looking at the percentage of environmental variance explained by the
247 RDA and PCA axes we find that the RDA first four axes explain more variance of each
248 environmental variable than the first four axes of the PCA (Fig. 4). The difference is
249 especially true for env1 and env4, which are strongly associated with the first four
250 axes of the RDA whereas PCA axes show an almost null correlation. The variables
251 env7, env8, env9 and env10 are also correlated to the first fourth axes of the RDA
252 whereas PCs show far less correlation with these environments. Finally, both RDA
253 axes and PCs fail to correlate with the complete variability of env3 and env6 even if
254 RDA (R^2 = ~0.25) is slightly more efficient than PCA (R^2 = ~0). Furthermore, we can
255 notice in Fig. S2 that each RDA axis, from the first one to the fourth, is associated to
256 one or several environmental variables whereas only one PC (PC1) is associated to
257 env2 and env5.

258

259 Finally, over the 100 simulations, we have measured the average FDR and
260 power for both *pcadapt* and RDA (Fig. 5). The two methods identify correctly almost

261 all the QTL2 SNPs (Fig. 5A). RDA-based analysis is more powerful than *pcadapt* to
262 identify QTL1 SNPs with respectively 75% and less than 10 % of true identification.
263 And for QTL3, *pcadapt* is not able to find any of the SNPs under selection whereas
264 RDA reaches 24 % of true identification. Both method properly control for the false
265 discovery rate because the proportion of false discoveries obtained with *pcadapt*
266 and the RDA-based method are respectively of 5% and 2.3% when controlling for a
267 nominal error rate of 5%
268

269 **2. *Populus trichocarpa***

270 For the RDA and *pcadapt* analysis on *P. trichocarpa* dataset, we retained
271 respectively *K*=3 and *K*=6 axes as indicated by the screeplots (Fig. S2). Analysis with
272 a FDR of 0.1 leads to a list of 151 outliers for RDA and 117 for *pcadapt*. Interestingly,
273 both methods have 29 outliers in common so that 122 and 88 SNPs are outliers
274 specific to respectively RDA and *pcadapt*.

275 When we compared RDA genome scan results with the outliers found by
276 Geraldes *et al.* (2014) based on Fdist, bayescan and bayenv methods, we found a
277 substantial overlap between them (Fig. 6). Among the 151 outliers found with the
278 RDA-based method, 69 have already been picked up by *Fst*-based methods (see
279 Table S3). Interestingly, the loci showing the lowest p-values with the RDA analysis
280 are also pointed out as outliers by *Fst* and *pcadapt* methods (Fig. 6). However, RDA

281 based genome scan found 72 SNPs that no other methods including *pcadapt* had
282 detected as outlier.

283 When we performed a new RDA analysis on the set of 151 outliers loci as
284 detected by RDA based genome scan method (adaptive enriched genetic space), we
285 found that the two first axes keep most of the adaptive genetic variance among the
286 populations (Fig. 7A). These two axes correspond to the two principal selective
287 pressures in the sampling area. RDA1 is correlated with several environmental
288 variables used in the analysis (Fig. 7A) including DD_0, DD18, MAT, MCMT, EMT,
289 which correspond to temperature indexes. RDA1 is also correlated to moisture
290 variables (e. g. AHM or MAP), continentality (TD) and flowering period length (bFFP
291 and eFFP). This axis could summarize a gradient of global climatic variation
292 encountered in the sampling area with southern and coastal populations opposed to
293 northern and continental populations (Fig. 7B). On the other side RDA2 also
294 correlates with moisture and temperature variables such as AHM, MAP, TD, MCMT,
295 eFFP and contrasts individuals from coastal locations to individuals from inland
296 areas (Fig. 7C).

297

298 **DISCUSSION**

299 Genetic differentiation approaches, as implemented in *pcadap*t, detect true
300 positives when environmental gradients related to adaption are correlated to

301 population structure. For instance, QTL2 SNPs are adaptive along an environmental
302 gradient that is correlated with population structure (see Fig. 1) and they are
303 detected by *pcadapt* (Fig. 5). By contrast, when the environmental gradient is not
304 correlated to population structure (QTL3 and QTL1 to a lesser extent) it fails to
305 detect adaptive SNPs. In this case, PCA ordination is not able to orientate genetic
306 variation into the direction of environment 1 or 3. Figure 6 shows that RDA has a
307 larger statistical power than *pcadapt* to detect QTL1 SNPs by taking advantage of
308 information from environmental local conditions. It can be attributed to a better
309 alignment between the genetic space and the environmental variable improving the
310 power to detect true positives. This better alignment is easily visible when looking at
311 the PCA and RDA correlation with environmental variables (Fig. 4 & Fig. S2). In these
312 figures we notice that PCA is only capable of identify the preponderant genetic
313 variance correlated with geographic distance among populations (PC1 and PC2)
314 whereas RDA is able to also capture the genetic variability associated to other
315 environmental variables such as env1 (RDA3) and to a lesser extent env3 (RDA4).
316 However, RDA fails to detect QTL3 outliers in 75 % of the simulations (Fig. 5)
317 possibly because of the additional difficulty of detecting adaptive SNPs on a coarse
318 environment compared to a smooth environmental gradient as environment 1 and 2.
319 RDA-based genome scan relies on linear association between environment and
320 allelic abundance (0, 1 or 2). Patchy selective pressure like env3 is difficult to

321 identify with correlative approach. Forester and colleagues (2015) already found
322 that heterogeneous environmental selection decrease the capacity of RDA to identify
323 the loci under selection but they also showed that RDA remains the best method
324 among the ones they tested. The strength of selection and the level of allelic
325 dispersion during the simulations strongly influence the statistical power of RDA;
326 weak selective strength and strong dispersion being the less favourable conditions
327 (Forester *et al.*, 2015). Nonetheless, our simulations plead in favour of using
328 constrained ordination method when relevant environmental variable are available
329 in order to orientate ordination axis in the direction of environmental gradients.
330 This result concords with a very recent study of Forester and colleagues (2017) who
331 conclude that constrained ordination shows a superior combination of low false
332 positive and high true positives rates across all level of selection.

333 During our RDA-based procedure we use Mahalanobis distances as a test
334 statistic to identify genetic variation significantly associated with environmental
335 gradients. This distance is computed based on the scores of more than one RDA axis
336 and can thus condense in one statistic the information of several RDA axes. Some
337 other studies have used different RDA-based statistics to identify the loci under
338 putative selection. They often used a statistic estimated independently for the
339 different RDA axes such as the loadings of the loci along the axis (e. g. extreme
340 squared loadings for Latsky *et al.*, 2012) and considering a threshold such as ± 3 SD

341 (Forester *et al*, 2017). Mahalanobis distances can capture complex selection patterns
342 where allelic variation depends on more than one environmental gradient. This
343 statistic is appropriate to identify selective pressures on natural populations, which
344 usually encompass multi-factorial gradients of selection (Salmela 2014).
345
346 One application of ordination methods is to detect selective gradients. The
347 RDA-based approach succeeded very well to decompose the QTL SNPs-
348 environments associations along the different axes, at least in our simulations. In our
349 example, we succeeded to dissociate the outliers linked to the different
350 environmental variables and thus to precisely identify which set of variables drives
351 each part of the adaptive genetic variability (Fig. 3). RDA1 is strongly associated with
352 envir2, RDA2 with envir7 (collinear with isolation by distance pattern in the grid),
353 RDA3 with envir1 and RDA4 with envir3. As expected, the correlated environments
354 are also strongly associated with this respective axes. This is reflecting the fact that it
355 is difficult to distinguish among several collinear variables which environmental
356 variable has a causal effect on the individual fitness. However, it is often sufficient to
357 identify combination of environment variables having a strong association with
358 adaptive variation without knowing precisely the underlying mechanical process.
359 These simulations serve as a proof of concept to show that ordination axes can
360 provide holistic measures of genomic adaptation (Steane *et al.,* 2014). Under this

361 idea, scores of individuals on one RDA axis reflect their degree of adaptation along
362 this axis. It allows to produce useful spatial representations of the adaptive
363 constraint in the landscape (see Fig. 7) and to visualize a gradient of genetic
364 adaptation.
365
366 From the analysis of *Populus trichocarpa*, RDA found numerous outlier loci
367 (72) that are not detected by any other genome scan method, including *pcadapt* (Fig.
368 7 and table S2). These results support the simulation analysis conclusions that RDA
369 based genome scan is able to pick up more genes in relation to environmental
370 gradients than *pcadapt*. It is however noticeable that a proportion of outliers are
371 shared between PCA and RDA based methods (29), highlighting the fact that both
372 approaches are partially similar. The lack of interpretability of PCA method is
373 compensated by the fact that it is a blind method in regard to environmental data so
374 it can pick up genes that we will miss using RDA because some crucial
375 environmental data are lacking.
376 After focusing the RDA analysis on the loci showing significant association
377 with environmental variables (151 outliers), we identified the predominant
378 environmental gradients structuring the genetic adaptation of *Populus trichocarpa.*
379 The main one is a global climatic variation (RDA1 in Fig. 7), gathering precipitation,
380 temperature and seasonality variations. It clearly contrasts the populations coming

381 from the extreme northern part of the sampling (Alaska) and the population coming
382 from the inland locations (higher altitudes) to the populations close to the coast in
383 British Columbia (Canada) and the southern populations, near Portland (USA). This
384 results concord with Geraldes et al. (2014) conclusions. In their study they point out
385 the need of adaptation to seasonality, photoperiod and frost events for *Populus*
386 *trichocarpa* along its continuous distribution from 44 to almost 60 degree of latitude.
387 RDA2 is associated to moisture regimes and points out a gradient of adaptation
388 between coastal locations, strongly watered, and inland populations, receiving less
389 precipitation (Fig. 7). These two environmental gradients are the strongest selective
390 pressures driving *Populus trichocarpa* climatic adaptation in the North west of
391 America, considering the set of environmental variables available and the scale of
392 the analysis. To go further, these composite indexes could also serve to predict (i)
393 provenances that would perform well in common gardens, (ii) patterns of
394 adaptation to local climate across the geographic range or (iii) future adaptive
395 landscape in a context of climate change.
396

397 **ACKNOWLEDGEMENTS**

398 We would like to dedicate this paper to the memory of our colleague and
399 beloved friend Eric Bazin (1977-2017) who passed away too soon.

400 TC acknowledges support from the ANR project APPATS (ANR-15-CE02-

0004) and MB acknowledges Grenoble Alpes Data Institute, which is supported by the French National Research Agency under the "Investissements d'avenir" program (ANR-15-IDEX-02).

REFERENCES

Bazin, Eric, Kevin J. Dawson, and Mark A. Beaumont. 2010. "Likelihood-Free Inference of Population Structure and Local Adaptation in a Bayesian Hierarchical Model." *Genetics* 185 (2): 587–602.

Coop, Graham, David Witonsky, Anna Di Rienzo, and Jonathan K. Pritchard. 2010. "Using Environmental Correlations to Identify Loci Underlying Local Adaptation." *Genetics* 185 (4): 1411–23.

Dabney, Alan, and John Storey. 2011. "Bioconductor's Qvalue Package." *Gene*, no. 2001: 1–6.

Kort, Hanne De, Katrien Vandepitte, Hans Henrik Bruun, Déborah Closset-Kopp, Olivier Honnay, and Joachim Mergeay. 2014. "Landscape Genomics and a Common Garden Trial Reveal Adaptive Differentiation to Temperature across Europe in the Tree Species Alnus Glutinosa." *Molecular Ecology* 23 (19): 4709–21.

Villemereuil, Pierre de, and Oscar E. Gaggiotti. 2015. "A New FST-Based Method to Uncover Local Adaptation Using Environmental Variables." *Methods in Ecology and Evolution* 6 (11): 1248–58. doi:10.1111/2041-210X.12418.

Duforet-Frebourg, Nicolas, Keurcien Luu, Guillaume Laval, Eric Bazin, and Michael G B Blum. 2016. "Detecting Genomic Signatures of Natural Selection with Principal Component Analysis: Application to the 1000 Genomes Data." *Molecular Biology and Evolution* 33 (4): 1082–93.

Foll, Matthieu, and Oscar Gaggiotti. 2008. "A Genome-Scan Method to Identify Selected Loci Appropriate for Both Dominant and Codominant Markers: A Bayesian Perspective." *Genetics* 180 (2): 977–93.

Forester, Brenna R, Jesse R Lasky, Helene H Wagner, and Dean L Urban. 2017. "Comparing Methods for Detecting Multilocus Adaptation with Multivariate Genotype," no. 970: 1–42.

Forester, Brenna R., Matthew R. Jones, Stéphane Joost, Erin L. Landguth, and Jesse R. Lasky. 2016. “Detecting Spatial Genetic Signatures of Local Adaptation in Heterogeneous Landscapes.” *Molecular Ecology* 25 (1): 104–20.

Frichot, Eric, Sean D. Schoville, Guillaume Bouchard, and Olivier François. 2013. “Testing for Associations between Loci and Environmental Gradients Using Latent Factor Mixed Models.” *Molecular Biology and Evolution* 30 (7): 1687–99.

Geraldes, A., S. P. Difazio, G. T. Slavov, P. Ranjan, W. Muchero, J. Hannemann, L. E. Gunter, et al. 2013. “A 34K SNP Genotyping Array for Populus Trichocarpa: Design, Application to the Study of Natural Populations and Transferability to Other Populus Species.” *Molecular Ecology Resources* 13 (2): 306–23.

Geraldes, Armando, Nima Farzaneh, Christopher J. Grassa, Athena D. Mckown, Robert D. Guy, Shawn D. Mansfield, Carl J. Douglas, and Quentin C B Cronk. 2014. “Landscape Genomics of Populus Trichocarpa: The Role of Hybridization, Limited Gene Flow, and Natural Selection in Shaping Patterns of Population Structure.” *Evolution* 68 (11): 3260–80.

Hecht, Benjamin C., Andrew P. Matala, Jon E. Hess, and Shawn R. Narum. 2015. “Environmental Adaptation in Chinook Salmon (Oncorhynchus Tshawytscha) throughout Their North American Range.” *Molecular Ecology* 24 (22): 5573–95.

Lasky, Jesse R., David L. Des Marais, John K. McKay, James H. Richards, Thomas E. Juenger, and Timothy H. Keitt. 2012. “Characterizing Genomic Variation of Arabidopsis Thaliana: The Roles of Geography and Climate.” *Molecular Ecology* 21 (22): 5512–29.

Luu, Keurcien, Eric Bazin, and Michael G B Blum. 2017. “Pcadapt: An R Package to Perform Genome Scans for Selection Based on Principal Component Analysis.” *Molecular Ecology Resources* 17 (1): 67–77.

Martins, Helena, Kevin Caye, Keurcien Luu, Michael G.B. Blum, and Olivier François. 2016. “Identifying Outlier Loci in Admixed and in Continuous Populations Using Ancestral Population Differentiation Statistics.” *Molecular Ecology* 25 (20): 5029–42.

Oksanen, Jari, F Guillaume Blanchet, Michael Friendly, Roeland Kindt, Pierre Legendre, Dan Mcglinn, Peter R Minchin, et al. 2015. “Vegan: Community Ecology Package.” *R Package Version 2.3-2*, https://CRAN.R-project.org/package=vegan.

Peng, Bo, and Marek Kimmel. 2005. “simuPOP: A Forward-Time Population Genetics Simulation Environment.” *Bioinformatics* 21 (18): 3686–87.

Pritchard, Jonathan K., and Anna Di Rienzo. 2010. “Adaptation - Not by Sweeps Alone.” *Nature Reviews Genetics* 11 (10). Nature Publishing Group: 665–67.

Salmela, Matti J. 2014. “Rethinking Local Adaptation: Mind the Environment!” *Forest Ecology and Management* 312. Elsevier B.V.:271–81.

Savolainen, Outi, Martin Lascoux, and Juha Merilä. 2013. “Ecological Genomics of Local Adaptation.” *Nature Reviews Genetics* 14 (11): 807–20.

Steane, Dorothy A., Brad M. Potts, Elizabeth McLean, Suzanne M. Prober, William D. Stock, René E. Vaillancourt, and Margaret Byrne. 2014. “Genome-Wide Scans Detect Adaptation to Aridity in a Widespread Forest Tree Species.” *Molecular Ecology* 23 (10): 2500–2513.

Vangestel, Carl, Andrew J. Eckert, Jill L. Wegrzyn, J. Bradley St. Clair, and David B. Neale. 2018. “Linking Phenotype, Genotype and Environment to Unravel Genetic Components Underlying Cold Hardiness in Coastal Douglas-Fir (Pseudotsuga Menziesii Var. Menziesii).” *Tree Genetics & Genomes* 14 (1). Tree Genetics & Genomes:10.

Vatsiou, Alexandra I., Eric Bazin, and Oscar E. Gaggiotti. 2016. “Detection of Selective Sweeps in Structured Populations: A Comparison of Recent Methods.” *Molecular Ecology* 25 (1): 89–103.

Wang, Jiahui, Ruben Zamar, Alfio Marazzi, Victor Yohai, Matias Salibian-Barrera, Ricardo Maronna, Eric Zivot, et al. 2014. “Robust: Robust Library.” *R Package Version 0.4-16*. http://cran.r-project.org/package=robust.

Whitlock, Michael C., and Katie E. Lotterhos. 2015. “Reliable Detection of Loci Responsible for Local Adaptation: Inference of a Null Model through Trimming the Distribution of F_{ST}.” *The American Naturalist* 186 (S1): S24–36.

DATA ACCESSIBILITY STATEMENT

- Simulations datasets are available at Dryad Digital Repository at the following address: https://doi.org/10.5061/dryad.1s7v5

- Scripts used for simulations and *Populus trichocarpa* analysis are available on Github: https://github.com/Capblancq/RDA-genome-scan

- *Populus trichocarpa* SNP data are available on Dryad (http://datadryad.org/resource/doi:10.5061/dryad.1051d)

AUTHOR CONTRIBUTIONS

E. Bazin designed the study. E. B. and T. Capblancq performed the analysis and treatments. M. Blum, E. B. and T. C. wrote the manuscript, and all authors contributed substantially to revisions.

TABLES AND FIGURES

Figure 1: Graphical representations of mean environmental value for environment 1 to 10. Equation determining the environment is given in the main manuscript. Mean environmental value for environment 4, 5 and 6 are respectively equivalent to environments 1, 2 and 3.

Figure 2: Manhattan plots of the *pcadapt* (A) and RDA (B) results.

Figure 3: Projection of SNPs and environmental variables into the RDA space.

Figure 4: Proportion of environmental variance explained by the first four ordination axis using RDA or PCA. The values correspond to the r-squared of the regression between the environmental variable and the first four axes of the multivariate analysis.

Figure 5: Performance results of RDA and *pcadapt* methods. Each performance value is averaged over 100 simulated dataset (error bars are displayed). Power (A) is given separately for loci coding for quantitative traits 1, 2 and 3. False Discovery Rate (FDR) is given for the two methods (B).

Figure 6: Scatter plot of the p-values returned by RDA and *pcadapt* procedure for the *Populus trichocarpa* dataset. Black dots correspond to loci identified as outliers by *Fst* methods in Geraldes et al. (2014).

516 **Figure 7:** Projection of SNPs and environmental variables in the RDA performed on
517 the adaptively enriched genetic space (151 loci) of *P. trichocarpa* dataset (A) and
518 spatial representation of the population average scores on RDA1 (B) and RDA2 (C).
519

520 # SUPPORTING INFORMATION

521 **Script S1:** Detailed script of the RDA-based genome scan analysis.

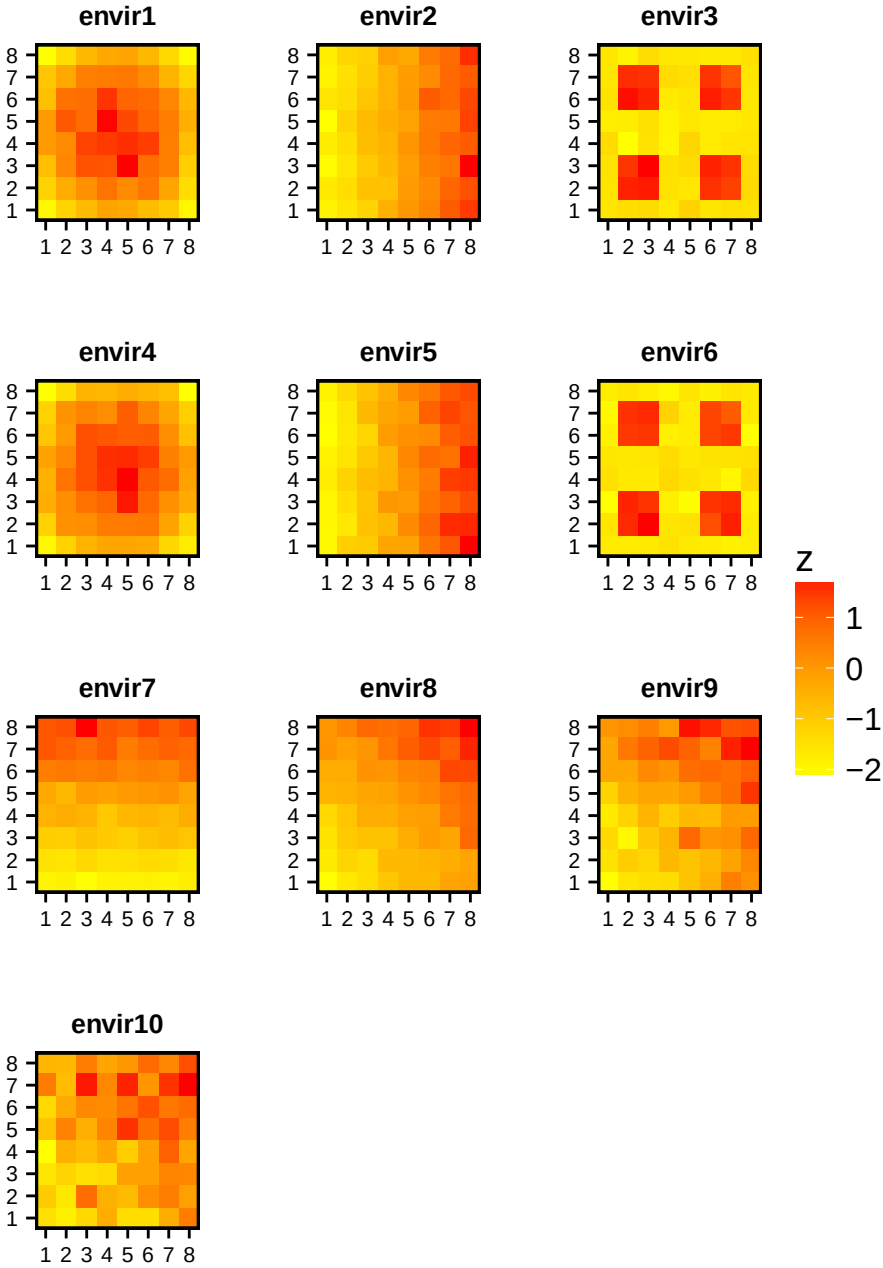
522 **Table S1:** Environmental variables used in the *Populus trichocarpa* analysis.

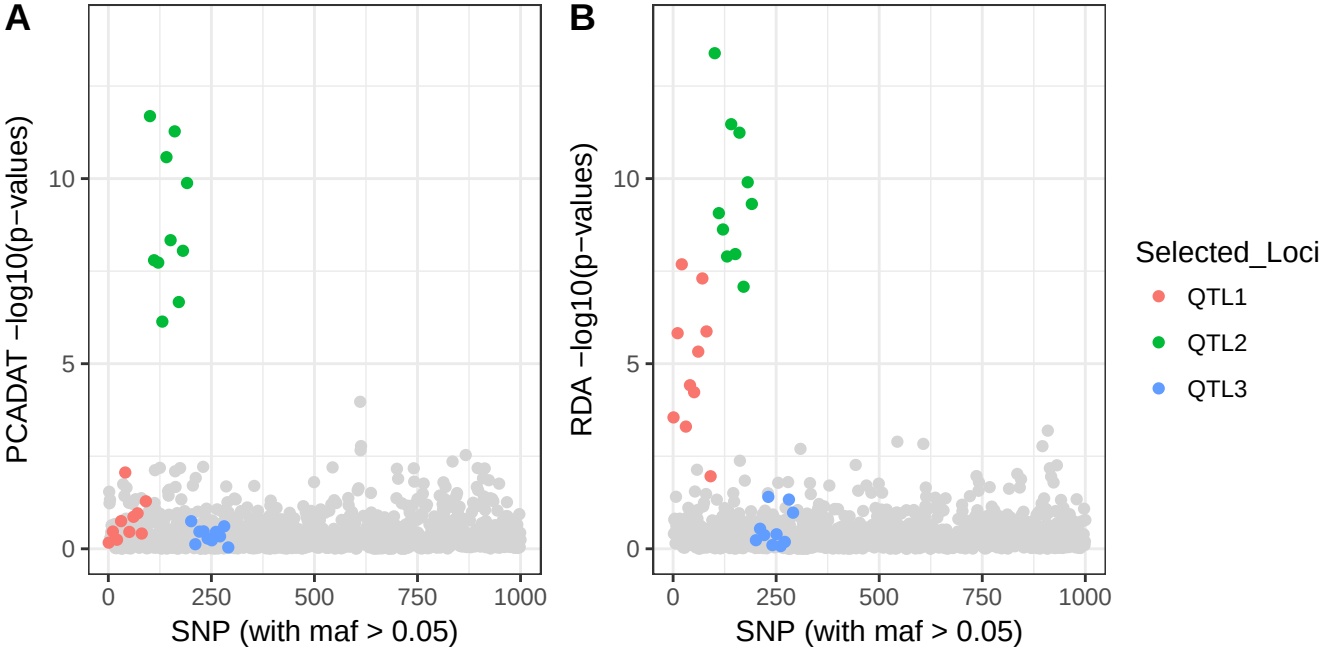
523 **Table S2:** Comparison between RDA, *pcadapt*, *fdist* and *bayescan* outliers loci.

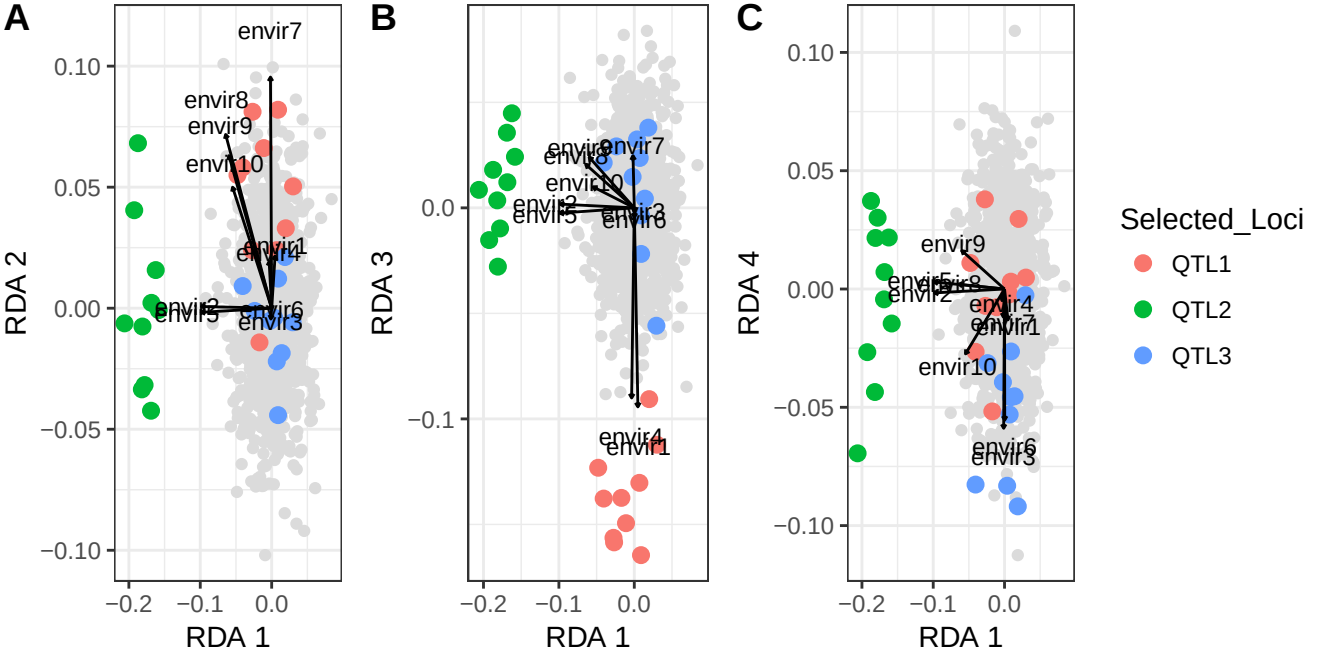
524 **Figure S1:** Scree plot for PCA and RDA analysis of simulation 1 dataset

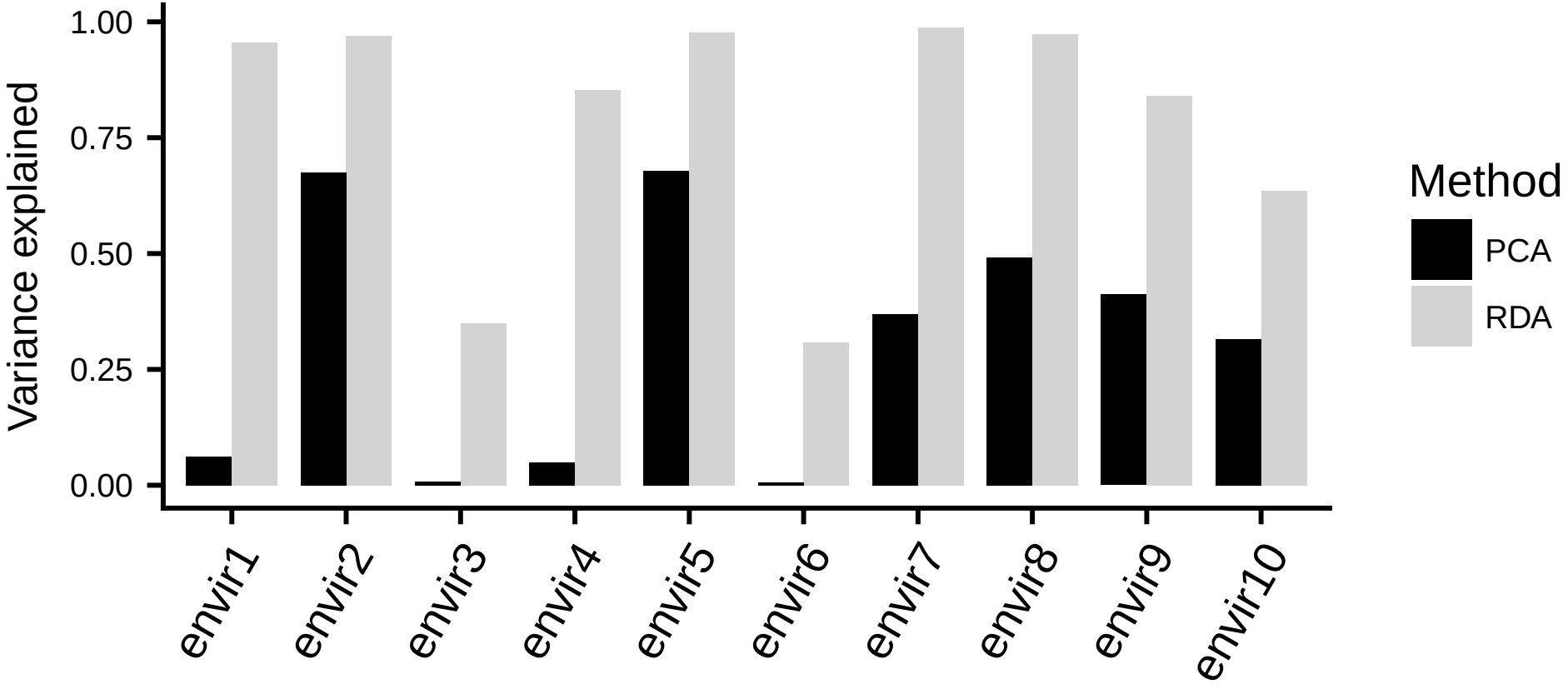
525 **Figure S2:** Heatmap of the R^2 resulting from linear regression between the fourth
526 first axes of PCA and RDA analysis and the environmental variables.

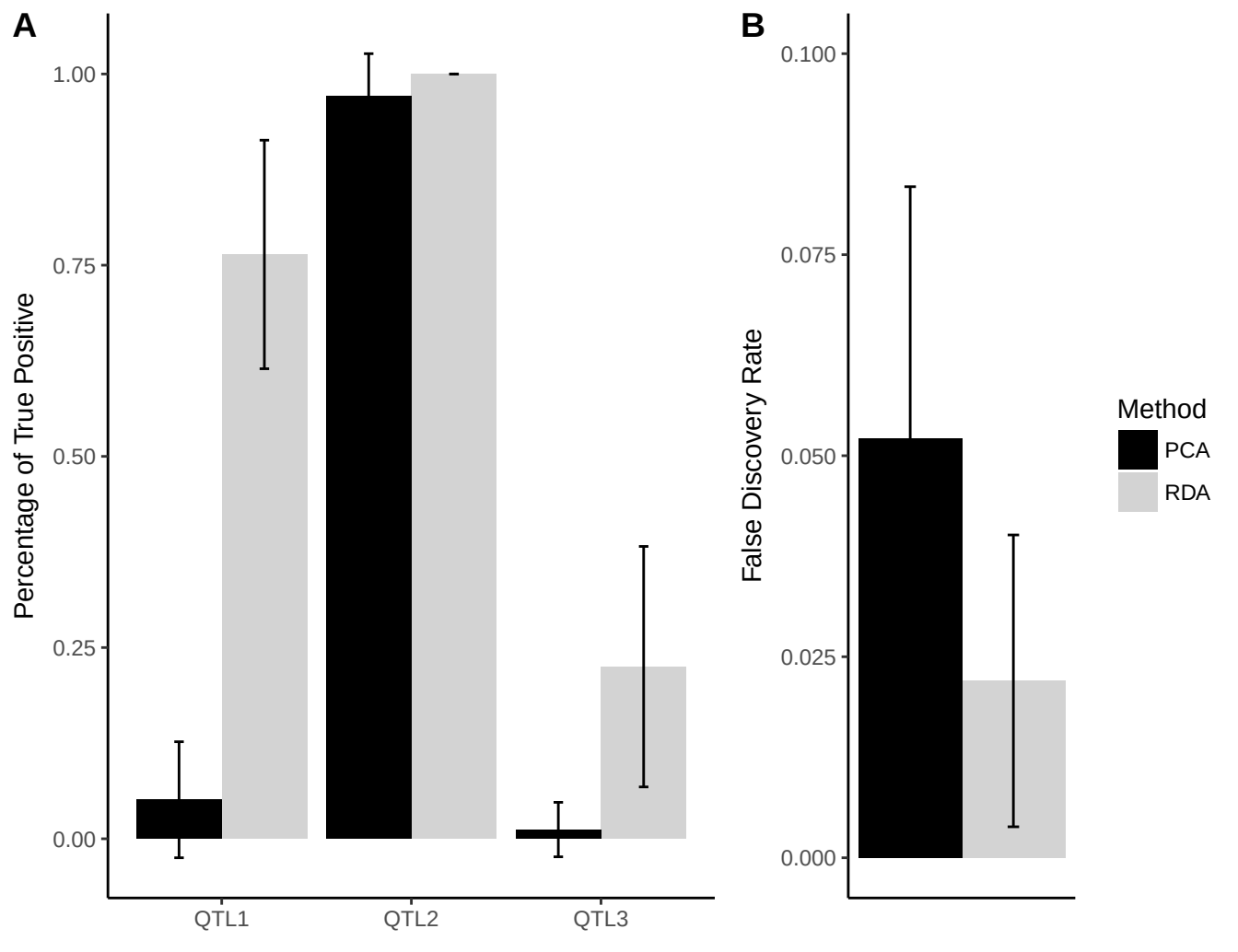
527 **Figure S3:** Scree plot for PCA and RDA analysis of *Populus trichocarpa* dataset.

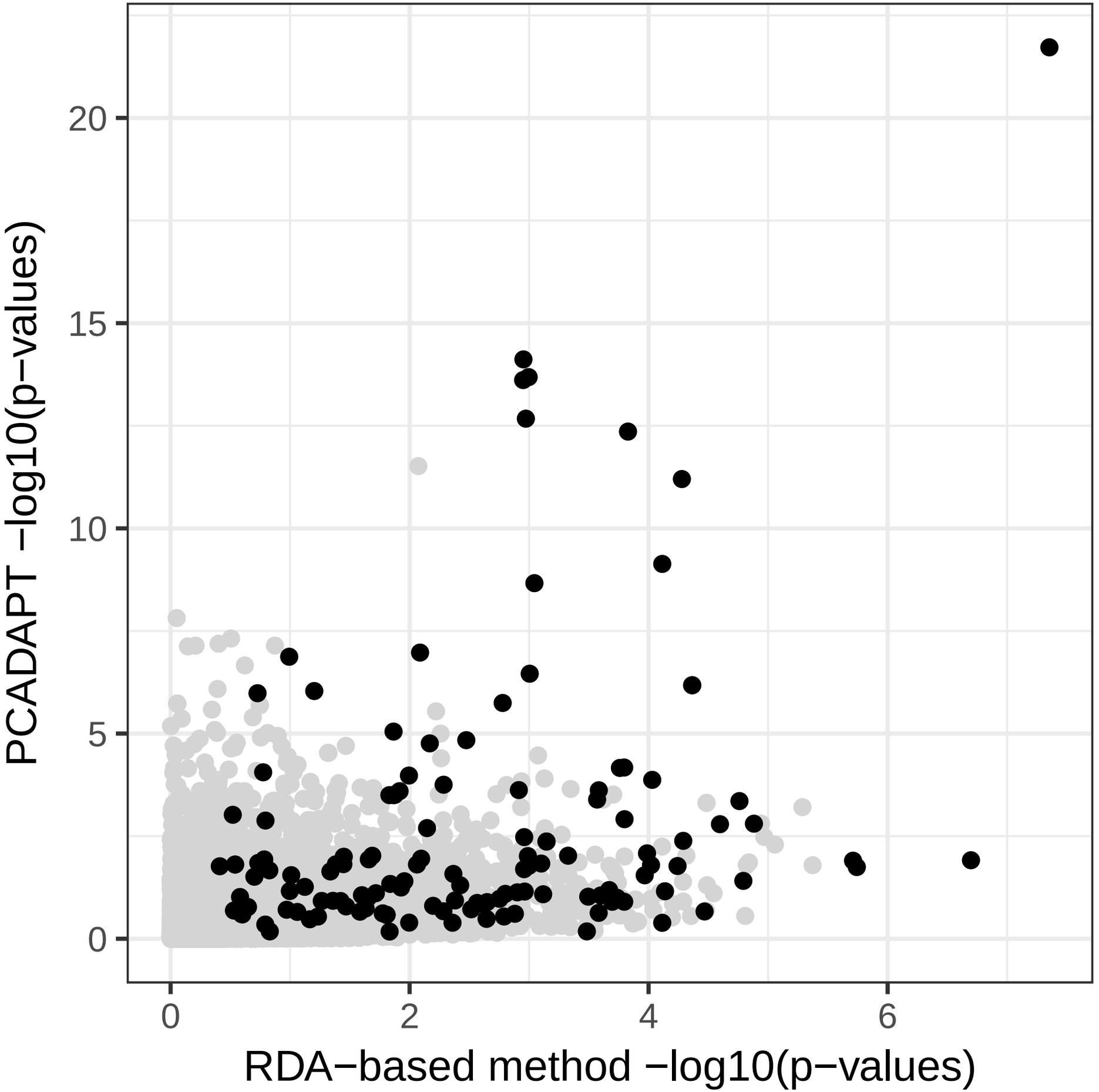

envir1
envir2
envir3
envir4
envir5
envir6
envir7
envir8
envir9
envir10
1 2 3 4 5 6 7 8
z
1
0
−1
−2

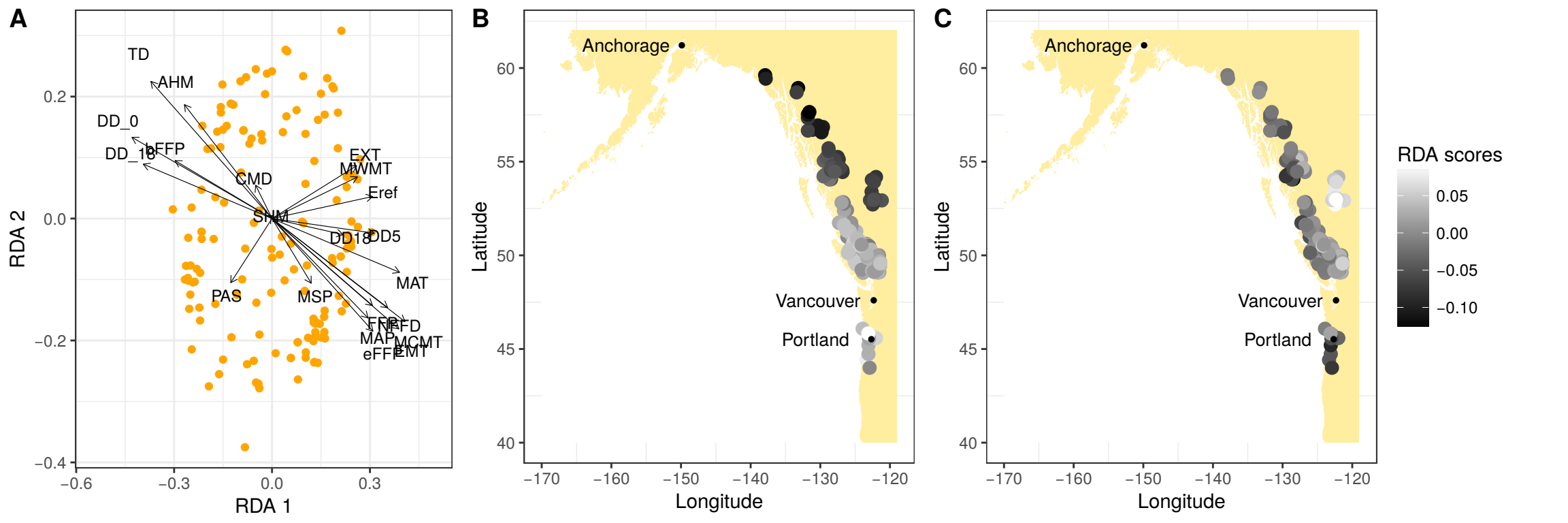

A
B
PCADAT −log10(p−values)
RDA −log10(p−values)
10
5
0
0
250
500
750
1000
SNP (with maf > 0.05)
Selected_Loci
QTL1
QTL2
QTL3


A
B
C
envir7
envir8
envir9
envir10
envir1
envir4
envir2
envir5
envir6
envir3
RDA 1
RDA 2
RDA 3
RDA 4
0.10
0.05
0.00
−0.05
−0.10
−0.2
−0.1
0.0
Selected_Loci
QTL1
QTL2
QTL3


Variance explained
1.00
0.75
0.50
0.25
0.00
envir1
envir2
envir3
envir4
envir5
envir6
envir7
envir8
envir9
envir10
Method
PCA
RDA


A
B
Percentage of True Positive
1.00
0.75
0.50
0.25
0.00
QTL1
QTL2
QTL3
False Discovery Rate
0.100
0.075
0.050
0.025
0.000
Method
PCA
RDA


PCADAPT −log10(p−values)
20
15
10
5
0
0
2
4
6
RDA−based method −log10(p−values)


A
TD
AHM
DD_0
DD_18
bFFP
CMD
EXT
MWMT
Eref
SHM
DD18
DD5
MAT
PAS
MSP
FFP
PFD
MAP
MCMT
eFFP
PMT
RDA 2
RDA 1
0.2
0.0
−0.2
−0.4
−0.6
−0.3
0.0
0.3
B
Anchorage
Vancouver
Portland
Latitude
Longitude
60
55
50
45
40
−170
−160
−150
−140
−130
−120
C
Anchorage
Vancouver
Portland
Latitude
Longitude
RDA scores
0.05
0.00
−0.05
−0.10