

1 PCADAPT : AN R PACKAGE TO PERFORM GENOME SCANS
2 FOR SELECTION BASED ON PRINCIPAL COMPONENT
3 ANALYSIS.

4 Keurcien Luu ¹ & Eric Bazin ² & Michael G. B. Blum ¹

5 ¹ *Université Grenoble Alpes, CNRS, Laboratoire TIMC-IMAG, UMR 5525, France.*

6 ² *Université Grenoble Alpes, CNRS, Laboratoire d'Ecologie Alpine UMR 5553, France*

7 **Abstract**

8 The R package *pcadapt* performs genome scans to detect genes under selection based
9 on population genomic data. It assumes that candidate markers are outliers with respect
10 to how they are related to population structure. Because population structure is ascer-
11 tained with principal component analysis, the package is fast and works with large-scale
12 data. It can handle missing data and pooled sequencing data. By contrast to population-
13 based approaches, the package handle admixed individuals and does not require grouping
14 individuals into populations. Since its first release, *pcadapt* has evolved both in terms of
15 statistical approach and software implementation. We present results obtained with robust
16 Mahalanobis distance, which is a new statistic for genome scans available in the 2.0 and
17 later versions of the package. When hierarchical population structure occurs, Mahalanobis
18 distance is more powerful than the communality statistic that was implemented in the
19 first version of the package. Using simulated data, we compare *pcadapt* to other software
20 for genome scans (*BayeScan*, *hapflk*, *OutFLANK*, *sNMF*). We find that the proportion of
21 false discoveries is around a nominal false discovery rate set at 10% with the exception of
22 *BayeScan* that generates 40% of false discoveries. We also find that the power of *BayeScan*
23 is severely impacted by the presence of admixed individuals whereas *pcadapt* is not im-
24 pacted. Last, we find that *pcadapt* and *hapflk* are the most powerful software in scenarios
25 of population divergence and range expansion. Because *pcadapt* handles next-generation
26 sequencing data, it is a valuable tool for data analysis in molecular ecology.

27 **Keywords.** population genetics, R package, outlier detection, Mahalanobis distance,
28 principal component analysis.

29 Introduction

30 Looking for variants with unexpectedly large differences of allele frequencies between
31 populations is a common approach to detect signals of natural selection (Lewontin and
32 Krakauer, 1973). When variants confer a selective advantage in the local environment,
33 allele frequency changes are triggered by natural selection leading to unexpectedly large
34 differences of allele frequencies between populations. To detect variants with large diffe-
35 rences of allele frequencies, numerous test statistics have been proposed, which are usually
36 based on chi-square approximations of F_{ST} -related test statistics (François et al., 2016).

37 Statistical approaches for detecting selection should address several challenges. The
38 first challenge is to account for hierarchical population structure that arises when genetic
39 differentiation between populations is not identical between all pairs of populations. Sta-
40 tistical tests based on F_{ST} that do not account for hierarchical structure, when it occurs,
41 generate a large excess of false positive loci (Bierne et al., 2013; Excoffier et al., 2009).

42 A second challenge arises because approaches based on F_{ST} -related measures require
43 to group individuals into populations, although defining populations is a difficult task
44 (Waples and Gaggiotti, 2006). Individual sampling may not be population-based but
45 based on more continuous sampling schemes (Lotterhos and Whitlock, 2015). Additionally
46 assigning an admixed individual to a single population involves some arbitrariness because
47 different regions of its genome might come from different populations (Pritchard et al.,
48 2000). Several individual-based methods of genome scans have already been proposed to
49 address this challenge and they are based on related techniques of multivariate analysis
50 including principal component analysis (PCA), factor models, and non-negative matrix
51 factorization (Duforet-Frebourg et al., 2014; Hao et al., 2016; Galinsky et al., 2016; Chen
52 et al., 2016; Duforet-Frebourg et al., 2016; Martins et al., 2016).

53 The last challenge arises from the nature of multilocus datasets generated from next
54 generation sequencing platforms. Because datasets are massive with a large number of
55 molecular markers, Monte Carlo methods usually implemented in Bayesian statistics may
56 be prohibitively slow (Lange et al., 2014). Additionally, next generation sequencing data
57 may contain a substantial proportion of missing data that should be accounted for (Arnold
58 et al., 2013; Gautier et al., 2013).

59 To address the aforementioned challenges, we have developed the software *pcadapt* and
60 the R package *pcadapt*. The software *pcadapt* is now deprecated and the R package only is

61 maintained. *pcadapt* assumes that markers excessively related with population structure
62 are candidates for local adaptation. Since its first release, *pcadapt* has substantially evolved
63 both in terms of statistical approach and software implementation (Table 1).

64 The first release of *pcadapt* was a command line C software. It implemented a Monte
65 Carlo approach based on a Bayesian factor model (Duforet-Frebourg et al., 2014). The
66 test statistic for outlier detection was a Bayes factor. Because Monte Carlo methods can
67 be computationally prohibitive with massive NGS data, we then developed an alternative
68 approach based on PCA. The first statistic based on PCA was the *communality* statistic,
69 which measures the percentage of variation of a SNP explained by the first K principal
70 components (Duforet-Frebourg et al., 2016). It was initially implemented with a command-
71 line C software (the *pcadapt fast* command) before being implemented in the *pcadapt* R
72 package. We do not maintain C versions of *pcadapt* anymore. The whole analysis that goes
73 from reading genotype files to detecting outlier SNPs can now be performed in R (R Core
74 Team, 2015).

75 The 2.0 and following versions of the R package implement a more powerful statistic
76 for genome scans. The test statistic is a robust Mahalanobis distance. A vector containing
77 K z -scores measures to what extent a SNP is related to the first K principal components.
78 The Mahalanobis distance is then computed for each SNP to detect outliers for which
79 the vector of z -scores do not follow the distribution of the main bulk of points. The
80 term robust refers to the fact that the estimators of the mean and of the covariance
81 matrix of z , which are required to compute Mahalanobis distances, are not sensitive to
82 the presence of outliers in the dataset (Maronna and Zamar, 2012). In the following, we
83 provide a comparison of statistical power that shows that Mahalanobis distance provides
84 more powerful genome scans compared to the communality statistic and to the Bayes
85 factor that were implemented in previous versions of *pcadapt*.

86 In addition to comparing the different test statistics that were implemented in *pcadapt*,
87 we compare statistic performance obtained with the 3.0 version of *pcadapt* and with other
88 software of genome scans. We use simulated data to compare software in terms of false
89 discovery rate (FDR) and statistical power. We consider data simulated under different
90 demographic models including island model, divergence model and range expansion. To
91 perform comparisons, we include software that require to group individuals into popu-
92 lations : *BayeScan* (Foll and Gaggiotti, 2008), the F_{LK} statistic as implemented in the
93 *hapflk* software (Bonhomme et al., 2010), and *OutFLANK* that provides a robust estima-

94 tion of the null distribution of a F_{ST} test statistic (Whitlock and Lotterhos, 2015). We
95 additionally consider the *sNMF* software that implements another individual-based test
96 statistic for genome scans (Frichot et al., 2014; Martins et al., 2016).

97 **Statistical and Computational approach**

98 **Input data**

99 The R package can handle different data formats for the genotype data matrix. In the
100 version 3.0 that is currently available on CRAN, the package can handle genotype data
101 files in the *vcf*, *ped* and *lfmm* formats. In addition, the package can also handle a *pcadapt*
102 format, which is a text file where each line contains the allele counts of all individuals
103 at a given locus. When reading a genotype data matrix with the *read.pcadapt* function, a
104 *.pcadapt* file is generated, which contains the genotype data in the *pcadapt* format.

105 **Choosing the number of principal components**

106 In the following, we denote by n the number of individuals, by p the number of genetic
107 markers, and by G the genotype matrix that is composed of n lines and p columns.
108 The genotypic information at locus j for individual i is encoded by the allele count G_{ij} ,
109 $1 \leq i \leq n$ and $1 \leq j \leq p$, which is a value in $0, 1$ for haploid species and in $0, 1, 2$ for
110 diploid species.

111 First, we normalize the genotype matrix columnwise. For diploid data, we consider the
112 usual normalization in population genomics where $\tilde{G}_{ij} = (G_{ij} - p_j) / (2 \times p_j(1 - p_j))^{1/2}$, and
113 p_j denotes the minor allele frequency for locus j (Patterson et al., 2006). The normalization
114 for haploid data is similar except that the denominator is given by $(p_j(1 - p_j))^{1/2}$.

115 Then, we use the normalized genotype matrix \tilde{G} to ascertain population structure
116 with PCA (Patterson et al., 2006). The number of principal components to consider is
117 denoted K and is a parameter that should be chosen by the user. In order to choose K , we
118 recommend to consider the graphical approach based on the scree plot (Jackson, 1993).
119 The scree plot displays the eigenvalues of the covariance matrix Ω in descending order.
120 Up to a constant, eigenvalues are proportional to the proportion of variance explained
121 by each principal component. The eigenvalues that correspond to random variation lie on
122 a straight line whereas the ones corresponding to population structure depart from the

123 line. We recommend to use Cattell’s rule that states that components corresponding to
124 eigenvalues to the left of the straight line should be kept (Cattell, 1966).

125 Test statistic

126 We now detail how the package computes the test statistic. We consider multiple linear
127 regressions by regressing each of the p SNPs by the K principal components X_1, \dots, X_K

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

128 where β_{jk} is the regression coefficient corresponding to the j -th SNP regressed by the
129 k -th principal component, and ϵ_j is the residuals vector. To summarize the result of the
130 regression analysis for the j -th SNP, we return a vector of z -scores $z_j = (z_{j1}, \dots, z_{jK})$
131 where z_{jk} corresponds to the z -score obtained when regressing the j -th SNP by the k -th
132 principal component.

133 The next step is to look for outliers based on the vector of z -scores. We consider a
134 classical approach in multivariate analysis for outlier detection. The test statistic is a
135 robust Mahalanobis distance D defined as

$$D_j^2 = (z_j - \bar{z})^T \Sigma^{-1} (z_j - \bar{z}), \quad (2)$$

136 where Σ is the $(K \times K)$ covariance matrix of the z -scores and \bar{z} is the vector of the
137 K z -score means (Maronna and Zamar, 2012). When $K > 1$, the covariance matrix Σ
138 is estimated with the Orthogonalized Gnanadesikan-Kettenring method that is a robust
139 estimate of the covariance able to handle large-scale data (Maronna and Zamar, 2012)
140 (*covRob* function of the *robust* R package). When $K = 1$, the variance is estimated with
141 another robust estimate (*cov.rob* function of the *MASS* R package).

142 Genomic Inflation Factor

143 To perform multiple hypothesis testing, Mahalanobis distances should be transformed
144 into p -values. If the z -scores were truly multivariate Gaussian, the Mahalanobis distances
145 D should be chi-square distributed with K degrees of freedom. However, as usual for
146 genome scans, there are confounding factors that inflate values of the test statistic and

147 that would lead to an excess of false positives (François et al., 2016). To account for the
148 inflation of test statistics, we divide Mahalanobis distances by a constant λ to obtain a
149 statistic that can be approximated by a chi-square distribution with K degrees of freedom.
150 This constant is estimated by the genomic inflation factor defined here as the median of
151 the Mahalanobis distances divided by the median of the chi-square distribution with K
152 degrees of freedom (Devlin and Roeder, 1999).

153 Control of the false discovery rate (FDR)

154 Once p -values are computed, there is a problem of decision-making related to the
155 choice of a threshold for p -values. We recommend to use the FDR approach where the
156 objective is to provide a list of candidate genes with an expected proportion of false
157 discoveries smaller than a specified value. For controlling the FDR, we consider the q -
158 value procedure as implemented in the *qvalue* R package that is less conservative than
159 Bonferroni or Benjamini-Hochberg correction (Storey and Tibshirani, 2003). The *qvalue*
160 R package transforms the p -values into q -values and the user can control a specified value
161 α of FDR by considering as candidates the SNPs with q -values smaller than α .

162 Numerical computations

163 PCA is performed using a C routine that allows to compute scores and eigenvalues
164 efficiently with minimum RAM access (Duforet-Frebourg et al., 2016). Computing the co-
165 variance matrix Ω is the most computationally demanding part. To provide a fast routine,
166 we compute the $n \times n$ covariance matrix Ω instead of the much larger $p \times p$ covariance
167 matrix. We compute the covariance Ω incrementally by adding small storable covariance
168 blocks successively. Multiple linear regression is then solved directly by computing an
169 explicit solution, written as a matrix product. Using the fact that the (n, K) score matrix
170 X is orthogonal, the (p, K) matrix $\hat{\beta}$ of regression coefficients is given by $G^T X$ and the
171 (n, p) matrix of residuals is given by $G - XX^T G$. The z -scores are then computed using
172 the standard formula for multiple regression

$$z_{jk} = \hat{\beta}_{jk} \sqrt{\frac{\sum_{i=1}^n x_{ik}^2}{\sigma_j^2}}, \quad (3)$$

173 where σ_j^2 is an estimate of the residual variance for the j^{th} SNP, and x_{ik} is the score of
174 k^{th} principal component for the i^{th} individual.

175 Missing data

176 Missing data should be accounted for when computing principal components and when
177 computing the matrix of z -scores. There are many methods to account for missing data in
178 PCA and we consider the pairwise covariance approach (Dray and Josse, 2015). It consists
179 in estimating the covariance between each pair of individuals using only the markers that
180 are available for both individuals. To compute z -scores, we account for missing data in
181 formula (3). The term in the numerator $\sum_{i=1}^n x_{ik}^2$ depends on the quantity of missing data.
182 If there are no missing data, it is equal to 1 by definition of the scores obtained with PCA.
183 As the quantity of missing data grows, this term and the z -score decrease such that it
184 becomes more difficult to detect outlier markers.

185 Pooled sequencing

186 When data are sequenced in pool, the Mahalanobis distance is based on the matrix of
187 allele frequency computed in each pool instead of the matrix of z -scores.

188 Materials and Methods

189 Simulated data

190 We simulated SNPs under an island model, under a divergence model and we downloa-
191 ded simulations of range expansion (Lotterhos and Whitlock, 2015). All data we simulated
192 were composed of 3 populations, each of them containing 50 sampled diploid individuals
193 (Table 2). SNPs were simulated assuming no linkage disequilibrium. SNPs with minor
194 allele frequencies lower than 5% were discarded from the datasets. The mean F_{ST} for each
195 simulation was comprised between 5% and 10%. Using the simulations based on a island
196 and a divergence model, we also created datasets composed of admixed individuals. We
197 assumed that an instantaneous admixture event occurs at the present time so that all
198 sampled individuals are the results of this admixture event. Admixed individuals were

199 generated by drawing randomly admixture proportions using a Dirichlet distribution of
200 parameter (α, α, α) (α ranging from 0.005 to 1 depending on the simulation).

201 **Island model**

202 We used *ms* to create simulations under an island model (Fig SI1). We set a lower
203 migration rate for the 50 adaptive SNPs compared to the 950 neutral ones to mimick
204 diversifying selection (Bazin et al., 2010). For a given locus, migration from population
205 i to j was specified by choosing a value of the effective migration rate that is set to
206 $M_{\text{neutral}} = 10$ for neutral SNPs and to M_{adaptive} for adaptive ones. We simulated 35 datasets
207 in the island model with different strengths of selection, where the strength of selection
208 corresponds to the ratio $M_{\text{neutral}}/M_{\text{adaptive}}$ that varies from 10 to 1,000. The *ms* command
209 lines for neutral and adaptive SNPs are given by ($M_{\text{adaptive}} = 0.01$ and $M_{\text{neutral}} = 10$)

```
210 ./ms 300 950 -s 1 -I 3 100 100 100 -ma x 10 10 10 x 10 10 10 x  
211 ./ms 300 50 -s 1 -I 3 100 100 100 -ma x 0.01 0.01 0.01 x 0.01 0.01 0.01 x
```

212 **Divergence model**

213 To perform simulations under a divergence model, we used the package *simuPOP*,
214 which is an individual-based population genetic simulation environment (Peng and Kim-
215 mel, 2005). We assumed that an ancestral panmictic population evolved during 20 gene-
216 rations before splitting into two subpopulations. The second subpopulation then split into
217 subpopulations 2 and 3 at time $T > 20$. All 3 subpopulations continued to evolve until
218 200 generations have been reached, without migration between them (Figure SI1). A total
219 of 50 diploid individuals were sampled in each population. Selection only occurred in the
220 branch associated with population 2 and selection was simulated by assuming an additive
221 model (fitness is equal to $1 - 2s, 1 - s, 1$ depending on the genotypes). We simulated a
222 total of 3,000 SNPs comprising of 100 adaptive ones for which the selection coefficient is
223 of $s = 0.1$.

224 **Range expansion**

225 We downloaded in the *Dryad Digital Repository* six simulations of range expansion
226 with two glacial refugia (Lotterhos and Whitlock, 2015). Adaptation occurred during the
227 recolonization phase of the species range from the two refugia. We considered six different

228 simulated data with 30 populations and a number of sampled individual per location that
229 varies from 20 to 60.

230 **Parameter settings for the different software**

231 When using *hapflk*, we set $K = 1$ that corresponds to the computation of the *FLK*
232 statistic. When using *BayeScan* and *OutFLANK*, we used the default parameter values.
233 For *sNMF*, we used $K = 3$ for the island and divergence model and $K = 5$ for range
234 expansion as indicated by the cross-entropy criterion. The regularization parameter of
235 *sNMF* was set to $\alpha = 1000$. For *sNMF* and *hapflk*, we used the genomic inflation factor
236 to recalibrate p -values. When using population-based methods with admixed individuals,
237 we assigned each individual to the population with maximum amount of ancestry.

238 **Results**

239 **Choosing the number of principal components**

240 We evaluate Cattell's graphical rule to choose the number of principal components.
241 For the island and divergence model, the choice of K is evident (Figure 1). For $K \geq 3$, the
242 eigenvalues follow a straight line. As a consequence, Cattell's rule indicates $K = 2$, which
243 is expected because there are 3 populations (Patterson et al., 2006). For the model of
244 range expansion, applying Cattell's rule to choose K is more difficult (Figure 1). Ideally,
245 the eigenvalues that correspond to random variation lie on a straight line whereas the
246 ones corresponding to population structure depart from the line. However, there is no
247 obvious point at which eigenvalues depart from the straight line. Choosing a value of K
248 between 5 and 8 is compatible with Cattell's rule. Using the package *qvalue* to control
249 10% of FDR, we find that the actual proportion of false discoveries as well as statistical
250 power is weakly impacted when varying the number of principal components from $K = 5$
251 to $K = 8$ (Figure SI2).

252 **An example of genome scans performed with *pcadapt***

253 To provide an example of results, we apply *pcadapt* with $K = 6$ in the model of range
254 expansion. Population structure captured by the first 2 principal components is displayed

255 in Figure 2. P -values are well calibrated because they are distributed as a mixture of a
256 uniform distribution and of a peaky distribution around 0, which corresponds to outlier
257 loci (Figure 2). Using a FDR threshold of 10% with the *qvalue* package, we find 122 outliers
258 among 10,000 SNPs, resulting in 23% actual false discoveries and a power of 95%.

259 Control of the false discovery rate

260 We evaluate to what extent using the packages *pcadapt* and *qvalue* control a FDR set
261 at 10% (Figure 3). All SNPs with a q -value smaller than 10% were considered as candidate
262 SNPs. For the island model, we find that the proportion of false discoveries is 8% and it
263 increases to 10% when including admixture. For the divergence model, the proportion of
264 false discoveries is 11% and it increases to 22% when including admixture. The largest
265 proportion of false discoveries is obtained under range expansion and is equal to 25%.

266 We then evaluate the proportion of false discoveries obtained with *BayeScan*, *hapflk*,
267 *OutFLANK*, and *sNMF* (Figure 3). We find that *hapflk* is the most conservative approach
268 (FDR = 6%) followed by *OutFLANK* and *pcadapt* (FDR = 11%). The software *sNMF*
269 is more liberal (FDR = 19%) and *BayeScan* generates the largest proportion of false
270 discoveries (FDR = 41%). When not recalibrating the p -values of *hapflk*, we find that the
271 test is even more conservative (results not shown). For all software, the range expansion
272 scenario is the one that generates the largest proportion of false discoveries. Proportion of
273 false discoveries under range expansion ranges from 22% (*OutFLANK*) to 93% (*BayeScan*).

274 Statistical power

275 To provide a fair comparison between methods and software, we compare statistical
276 power for equal values of the observed proportion of false discoveries. Then we compute
277 statistical power averaged over observed proportion of false discoveries ranging from 0%
278 to 50%.

279 We first compare statistical power obtained with the different statistical methods that
280 have been implemented in *pcadapt* (Table 1). For the island model, Bayes factor, commu-
281 nality statistic and Mahalanobis distance have similar power (Figure 4). For the divergence
282 model, the power obtained with Mahalanobis distance is 20% whereas the power obtai-
283 ned with the communality statistic and with the Bayes factor is respectively 4% and 2%
284 (Figure 4). Similarly, for range expansion, the power obtained with Mahalanobis distance

285 is 46% whereas the power obtained with the communality statistic and with the Bayes
286 factor is 34% and 13%. We additionally investigate to what extent increasing sample size
287 in each population from 20 to 60 individuals affects power. For range expansion, the power
288 obtained with the Mahalanobis distance hardly changes ranging from 44% to 47%. Ho-
289 wever, the power obtained with the other two statistics changes importantly. The power
290 obtained with the communality statistic increases from 27% to 39% when increasing the
291 sample size and the power obtained with the Bayes factor increases from 0% to 44%.

292 Then we describe our comparison of software for genome scans. For the simulations
293 obtained with the island model where there is no hierarchical population structure, the
294 statistical power is similar for all software (Figure SI3 and SI4). Including admixed indi-
295 viduals hardly changes their statistical power (Figure SI3).

296 Then, we compare statistical power in a divergence model where adaptation took place
297 in one of the external branches of the population divergence tree. The software *pcadapt*
298 and *hapflk*, which account for hierarchical population structure, as well as *BayeScan* are
299 the most powerful in that setting (Figure 5 and Figure SI5). The values of power in
300 decreasing order are of 23% for *BayeScan*, of 20% for *pcadapt*, of 17% for *hapflk*, of 7%
301 for *sNMF* and of 1% for *OutFLANK*. When including admixed individuals, the power of
302 *hapflk* and of *pcadapt* hardly decreases whereas the power of *BayeScan* decreases to 6%
303 (Figure 5).

304 The last model we consider is the model of range expansion. The package *pcadapt* is the
305 most powerful approach in this setting (Figure 6 and SI6). Other software also discover
306 many true positive loci with the exception of *BayeScan* that provides no true discovery
307 when the observed FDR is smaller than 50% (Figure 6 and SI6). The values of power in
308 decreasing order are of 46% for *pcadapt*, of 41% for *hapflk*, of 37% for *OutFLANK*, of 30%
309 for *sNMF* and of 0% for *BayeScan*.

310 Running time of the different software

311 Last, we compare the software running times. The characteristics of the computer we
312 used to perform comparisons are the following : OSX El Capitan 10.11.3, 2,5 GHz Intel
313 Core i5, 8 Go 1600 MHz DDR3. We discard *BayeScan* as it is too time consuming compared
314 to other software. For instance, running *BayeScan* on a genotype matrix containing 150
315 individuals and 3,000 SNPs takes 9 hours whereas it takes less than one second with

316 *pcadapt*. The different software were run on genotype matrices containing 300 individuals
317 and from 500 to 50,000 SNPs. *OutFLANK* is the software for which the runtime increases
318 the most rapidly with the number of markers. *OutFLANK* takes around 25 minutes to
319 analyse 50,000 SNPs (Figure SI7). For the other 3 software (*hapflk*, *pcadapt*, *sNMF*),
320 analyzing 50,000 SNPs takes less than 3 minutes.

321 Discussion

322 The R package *pcadapt* implements a fast method to perform genome scans with next
323 generation sequencing data. It can handle datasets where population structure is conti-
324 nuous or datasets containing admixed individuals. It can handle missing data as well as
325 pooled sequencing data. The 2.0 and later versions of the R package implements a robust
326 Mahalanobis distance as a test statistic. When hierarchical population structure occurs,
327 Mahalanobis distance provides more powerful genome scans compared to the communa-
328 lity statistic that was implemented in the first version of the package (Duforet-Frebourg
329 et al., 2016). In the divergence model, adaptation occurs along an external branch of the
330 divergence tree that corresponds to the second principal component. When outlier SNPs
331 are not related to the first principal component, the Mahalanobis distance provides a
332 better ranking of the SNPs compared to the communality statistic.

333 Simulations show that the R package *pcadapt* compares favorably to other software of
334 genome scans. When data were simulated under an island model, population structure
335 is not hierarchical because genetic differentiation is the same for all pairs of populations.
336 Statistical power and control of the FDR were similar for all software. In presence of
337 hierarchical population structure (divergence model) where genetic differentiation varies
338 between pairs of populations, the ranking of the SNPs is software dependent. The software
339 *pcadapt* and *hapflk* provide the most powerful scans whether or not simulations include
340 admixed individuals. *OutFLANK* implements a F_{ST} statistic and because adaptation does
341 not correspond to the most differentiated populations, it fails to capture adaptive SNPs
342 (Figure 5) (Bonhomme et al., 2010; Duforet-Frebourg et al., 2016). *BayeScan* does not
343 assume equal differentiation between all pairs of populations, which may explain why
344 it has a good statistical power for the divergence model. However its statistical power is
345 severely impacted by the presence of admixed individuals because its power decreases from
346 24% to 6% (Figure 5). Understanding why *BayeScan* is severely impacted by admixture is

347 out of the scope of this paper. In the range expansion model, *BayeScan* returns many null
348 q -values (between 376 and 809 SNPs out of 9,899 neutral and 100 adaptive SNPs) such
349 that the observed FDR is always larger than 50%. Overall, we find that *pcadapt* and *hapflk*
350 provides comparable statistical power. Compared to other software, they provide optimal
351 or near optimal ranking of the SNPs in different scenarios including hierarchical population
352 structure and admixed individuals. The main difference between the two software concerns
353 the control of the FDR because *hapflk* is found to be more conservative.

354 Because NGS data become more and more massive, careful numerical implementation
355 is crucial. There are different options to implement PCA and *pcadapt* uses a numerical
356 routine based on the computation of the covariance matrix Ω . The algorithmic complexity
357 to compute the covariance matrix is proportional to pn^2 where p is the number of markers
358 and n is the number of individuals. The computation of the first K eigenvectors of the
359 covariance matrix Ω has a complexity proportional to n^3 . This second step is usually
360 more rapid than the computation of the covariance because the number of markers is
361 usually large compared to the number of individuals. In brief, computing the covariance
362 matrix Ω is by far the most costly operation when computing principal components.
363 Although we have implemented PCA in *C* to obtain fast computations, an improvement
364 in speed could be envisioned for future versions. When the number of individuals becomes
365 large (e.g. $n \geq 10,000$), there are faster algorithms to compute principal components
366 (Halko et al., 2011; Abraham and Inouye, 2014). In addition to running time, numerical
367 implementations also impact the effect of missing data on principal components (Dray and
368 Josse, 2015). Achieving a good tradeoff between fast computations and accurate evaluation
369 of population structure in the face of large amount of missing data is a challenge for
370 modern numerical methods in molecular ecology.

| Test statistic | Pop. structure | Language | Command line | Versions of the R package | Ref. |
|-------------------|----------------|----------|--------------|---------------------------|--------------------------------|
| Bayes factor | Factor model | C | PCAdapt | NA | Duforet-Frebourg et al. (2014) |
| Communality | PCA | C and R | PCAdapt fast | 1.x | Duforet-Frebourg et al. (2016) |
| Mahalanobis dist. | PCA | R | NA | 2.x and 3.x | This paper |

TABLE 1 – Summary of the different statistical methods and implementations of *pcadapt*. Pop. structure stands for population structure and dist. stands for distance.

| | Individuals | SNPs | Adaptive SNPs | Simulations |
|----------------------------|-------------|------|---------------|-------------|
| Island model | 150 | 472 | 27 | 35 |
| Divergence model | 150 | 3000 | 100 | 6 |
| Island model (hybrids) | 150 | 472 | 30 | 27 |
| Divergence model (hybrids) | 150 | 3000 | 100 | 9 |
| Range expansion | 1200 | 9999 | 99 | 6 |

TABLE 2 – Summary of the simulations. The table above shows the average number of individuals, of SNPs, of adaptive markers and the total number of simulations per scenario.

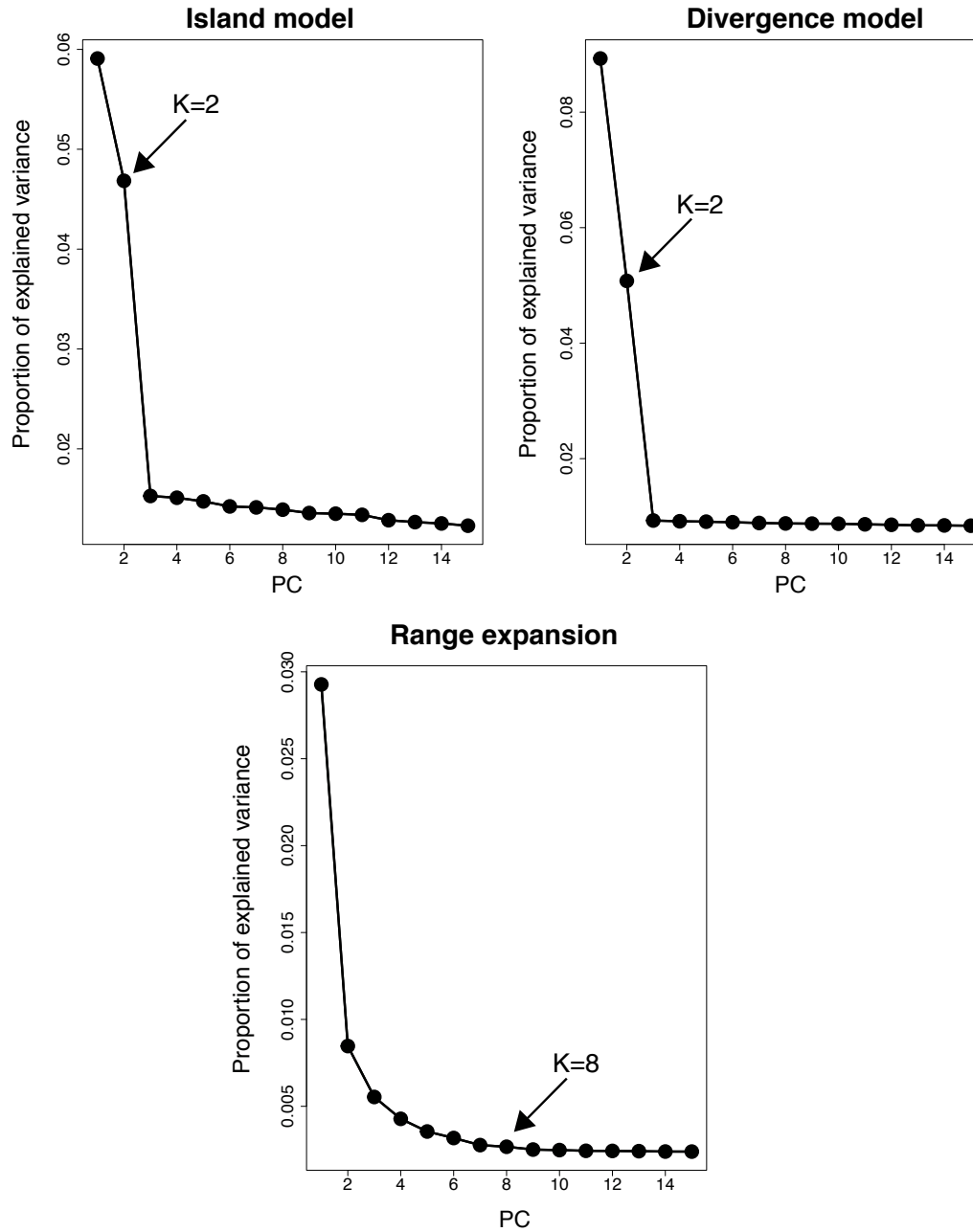


FIGURE 1 – Determining K with the screeplot. To choose K , we recommend to use Cattell’s rule that states that components corresponding to eigenvalues to the left of the straight line should be kept. According to Cattell’s rule, the eigenvalues that correspond to random variation lie on the straight line whereas the ones corresponding to population structure depart from the line. For the island and divergence model, the choice of K is evident. For the model or range expansion, a value of K between 5 and 8 is compatible with Cattell’s rule.

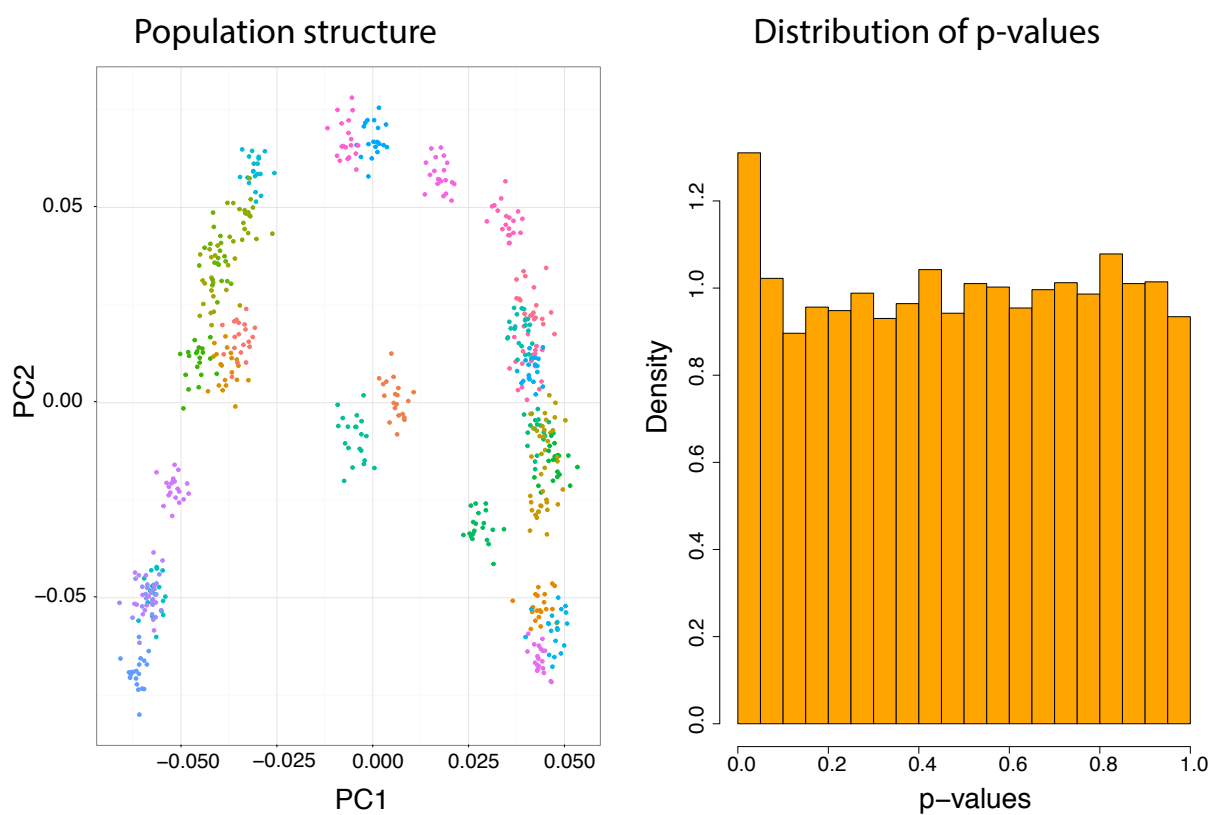


FIGURE 2 – Population structure (first 2 principal components) and distribution of p -value obtained with *pcadapt* for a simulation of range expansion. P -values are well calibrated because they are distributed as a mixture of a uniform distribution and of a peaky distribution around 0, which corresponds to outlier loci. In the left panel, each color corresponds to individuals sampled from the same population.

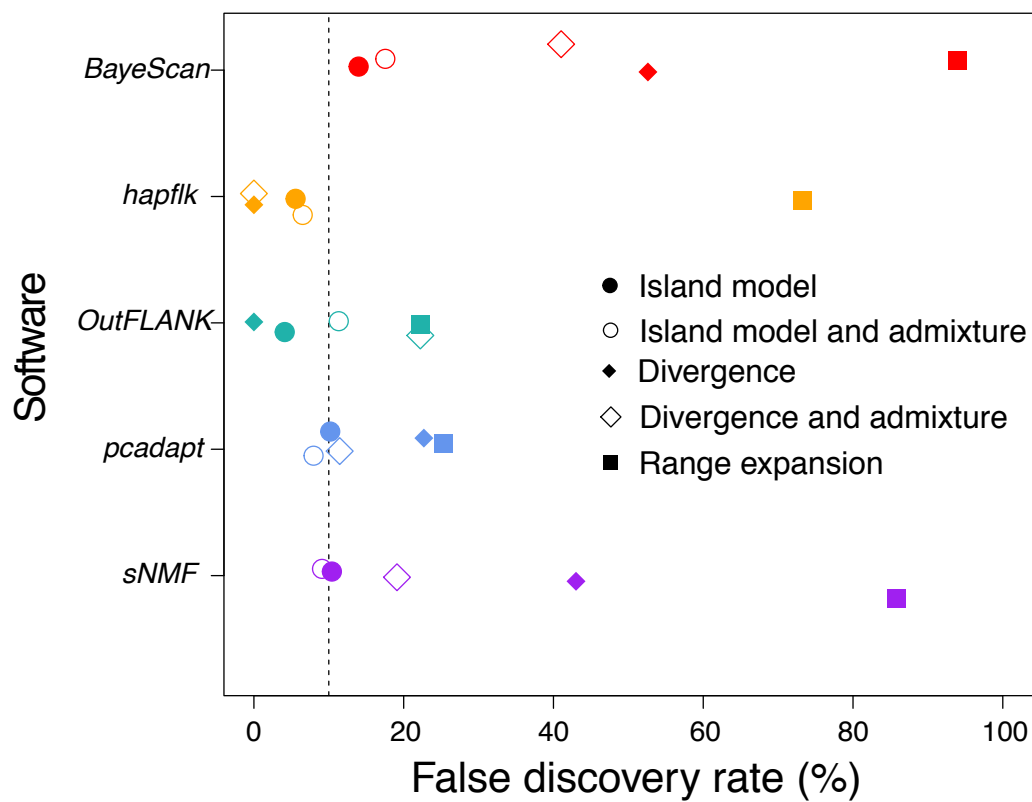


FIGURE 3 – Control of the FDR for different software of genome scans. We find that the median proportion of false discoveries is around the nominal FDR set at 10% (6% for *hapflk*, 11% for both *OutFLANK* and *pcadapt*, and 19% for *sNMF*) with the exception of *BayeScan* that generates 41% of false discoveries.

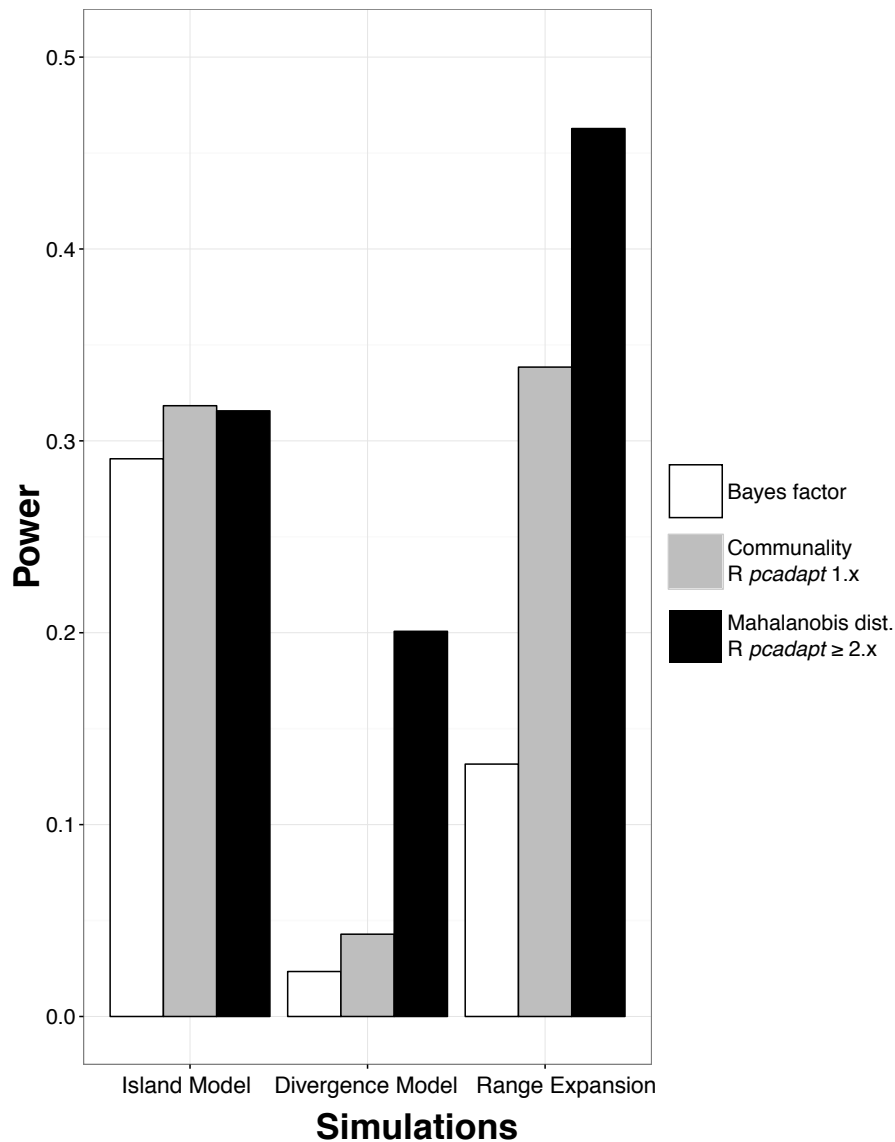


FIGURE 4 – Comparison of statistical power for the different test statistics that have been implemented in *pcadapt* (Table 1). Bayes factors corresponds to the test statistics implemented in the Bayesian version of *pcadapt* (Duforet-Frebourg et al., 2014); the communality statistic was the default statistic in version 1.x of the R package *pcadapt* (Duforet-Frebourg et al., 2016), and Mahalanobis distances are available since the release of the 2.0 version of the package. When there is hierarchical population structure (divergence model and range expansion), the Mahalanobis distance provides more powerful genome scans compared to the test statistic previously implemented in *pcadapt*. The abbreviation dist. stands for distance. Statistical power is averaged over the observed proportion of false discoveries (ranging between 0% and 50%).

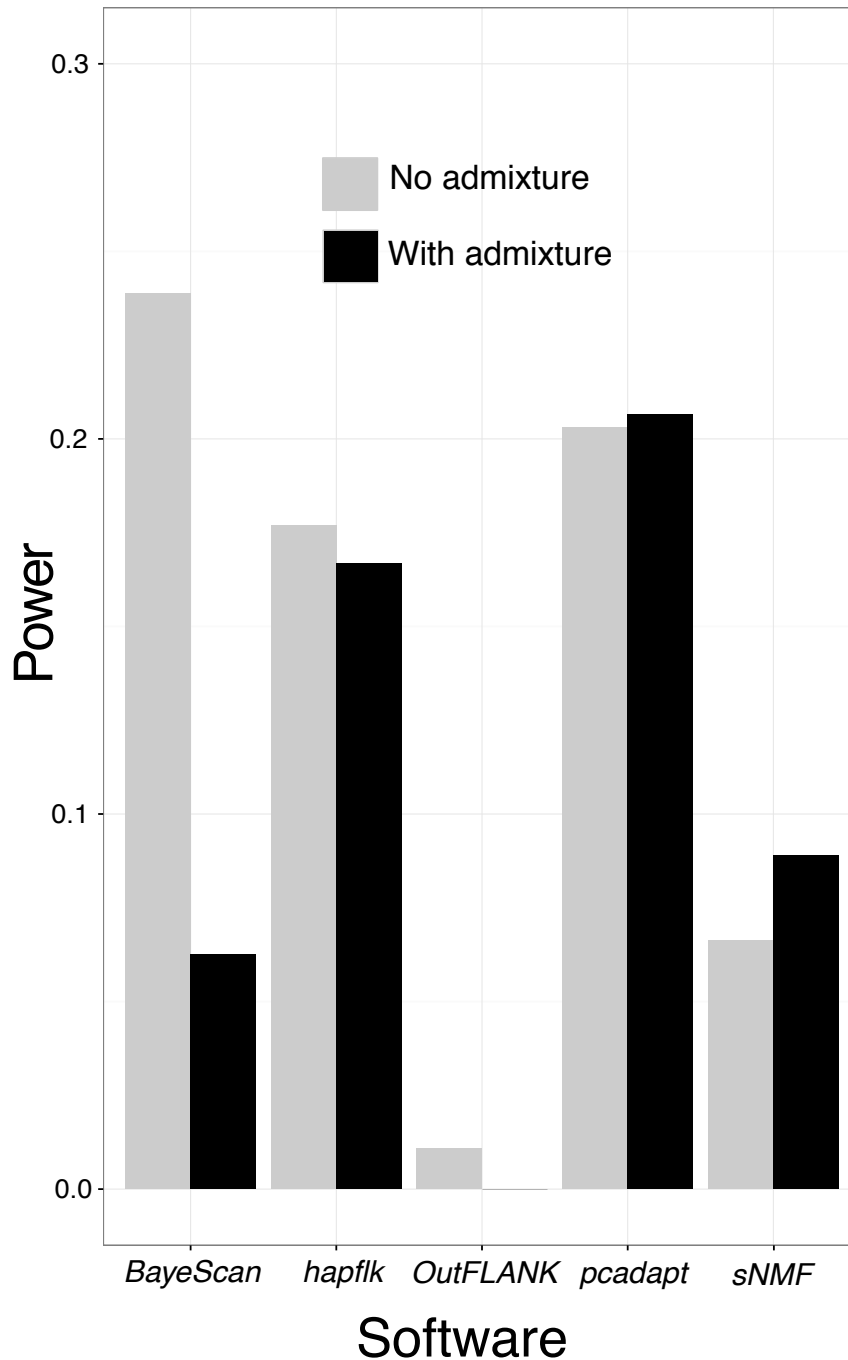


FIGURE 5 – Statistical power averaged over the expected proportion of false discoveries (ranging between 0% and 50%) for the divergence model with 3 populations. We assume that adaptation took place in an external branch that follows the most recent population divergence event.

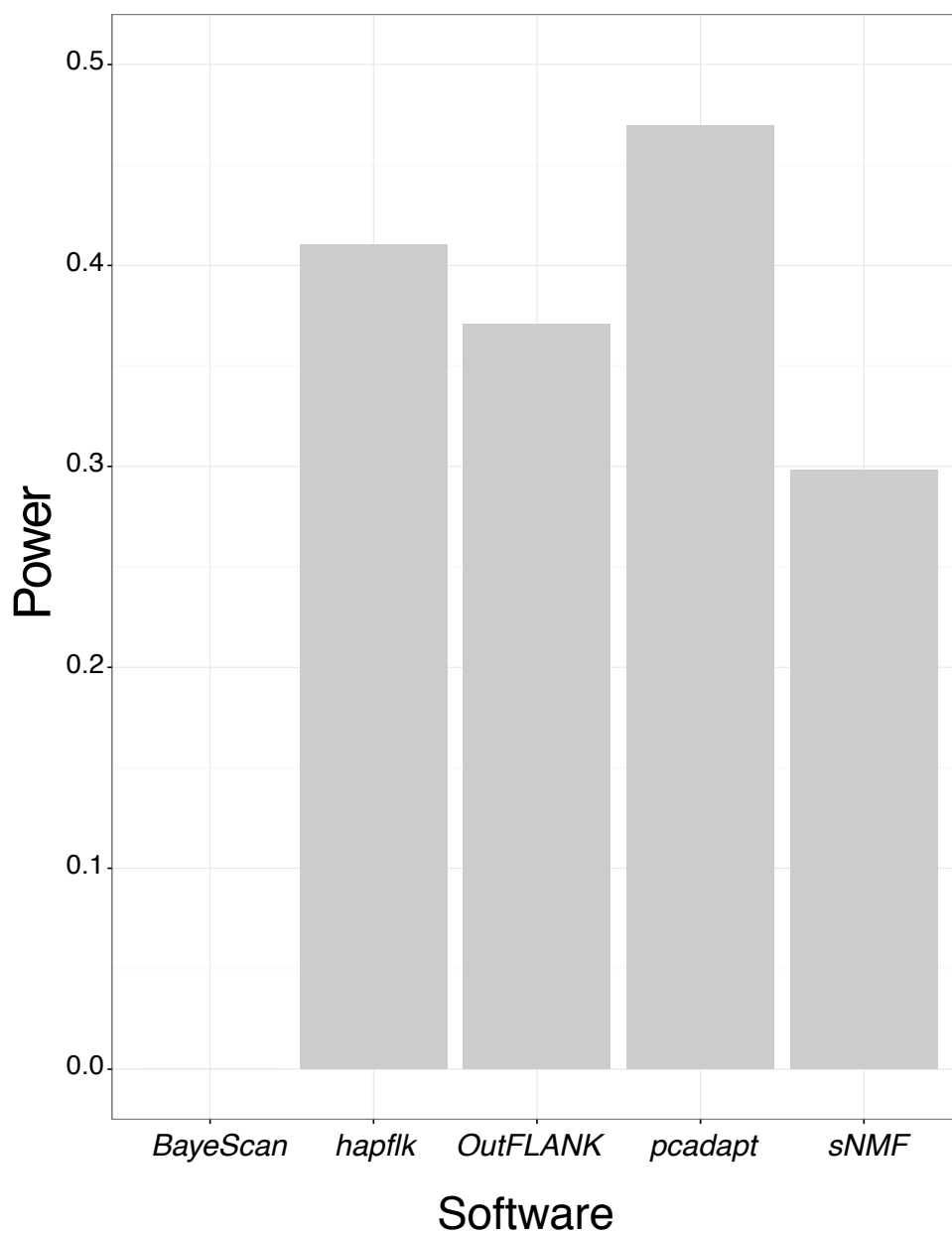


FIGURE 6 – Statistical power averaged over the expected proportion of false discoveries (ranging between 0% and 50%) for a range expansion model with two refugia. Adaptation took place during the recolonization event.

371 **Acknowledgements**

372 This work has been supported by the LabEx PERSYVAL-Lab (ANR-11-LABX-0025-
373 01) and the ANR AGRHUM project (ANR-14-CE02-0003-01).

374 References

- 375 Abraham, G. and M. Inouye, 2014. Fast principal component analysis of large-scale
376 genome-wide data. *PloS one* 9 :e93766.
- 377 Arnold, B., R. Corbett-Detig, D. Hartl, and K. Bomblies, 2013. RADseq underestimates
378 diversity and introduces genealogical biases due to nonrandom haplotype sampling.
379 *Molecular ecology* 22 :3179–3190.
- 380 Bazin, E., K. J. Dawson, and M. A. Beaumont, 2010. Likelihood-free inference of po-
381 pulation structure and local adaptation in a bayesian hierarchical model. *Genetics*
382 185 :587–602.
- 383 Bierne, N., D. Roze, and J. J. Welch, 2013. Pervasive selection or is it... ? why are fst
384 outliers sometimes so frequent ? *Molecular ecology* 22 :2061–2064.
- 385 Bonhomme, M., C. Chevalet, B. Servin, S. Boitard, J. Abdallah, S. Blott, and M. San-
386 Cristobal, 2010. Detecting selection in population trees : the Lewontin and Krakauer
387 test extended. *Genetics* 186 :241–262.
- 388 Cattell, R. B., 1966. The scree test for the number of factors. *Multivariate behavioral*
389 *research* 1 :245–276.
- 390 Chen, G.-B., S. H. Lee, Z.-X. Zhu, B. Benyamin, and M. R. Robinson, 2016. Eigengwas :
391 finding loci under selection through genome-wide association studies of eigenvectors in
392 structured populations. *Heredity* 117 :51–61.
- 393 Devlin, B. and K. Roeder, 1999. Genomic control for association studies. *Biometrics*
394 55 :997–1004.
- 395 Dray, S. and J. Josse, 2015. Principal component analysis with missing values : a compa-
396 rative survey of methods. *Plant Ecology* 216 :657–667.
- 397 Duforet-Frebourg, N., E. Bazin, and M. G. B. Blum, 2014. Genome scans for detecting
398 footprints of local adaptation using a bayesian factor model. *Molecular biology and*
399 *evolution* 31 :2483–2495.

- 400 Duforet-Frebourg, N., K. Luu, G. Laval, E. Bazin, and M. G. B. Blum, 2016. Detecting
401 genomic signatures of natural selection with principal component analysis : application
402 to the 1000 genomes data. *Molecular biology and evolution* 33 :1082–1093.
- 403 Excoffier, L., T. Hofer, and M. Foll, 2009. Detecting loci under selection in a hierarchically
404 structured population. *Heredity* 103 :285–298.
- 405 Foll, M. and O. Gaggiotti, 2008. A genome-scan method to identify selected loci appro-
406 priate for both dominant and codominant markers : a Bayesian perspective. *Genetics*
407 180 :977–993.
- 408 François, O., H. Martins, K. Caye, and S. D. Schoville, 2016. Controlling false discoveries
409 in genome scans for selection. *Molecular ecology* 25 :454–469.
- 410 Frichot, E., F. Mathieu, T. Trouillon, G. Bouchard, and O. François, 2014. Fast and
411 efficient estimation of individual ancestry coefficients. *Genetics* 196 :973–983.
- 412 Galinsky, K. J., G. Bhatia, P.-R. Loh, S. Georgiev, S. Mukherjee, N. J. Patterson, and
413 A. L. Price, 2016. Fast principal components analysis reveals independent evolution of
414 *adh1b* gene in Europe and East Asia. *American Journal of Human Genetics* 98 :456-
415 472 :018143.
- 416 Gautier, M., K. Gharbi, T. Cezard, J. Foucaud, C. Kerdelhué, P. Pudlo, J.-M. Cornuet,
417 and A. Estoup, 2013. The effect of RAD allele dropout on the estimation of genetic
418 variation within and between populations. *Molecular Ecology* 22 :3165–3178.
- 419 Halko, N., P.-G. Martinsson, and J. A. Tropp, 2011. Finding structure with randomness :
420 Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM*
421 *review* 53 :217–288.
- 422 Hao, W., M. Song, and J. D. Storey, 2016. Probabilistic models of genetic variation in
423 structured populations applied to global human studies. *Bioinformatics* 32 :713–721.
- 424 Jackson, D. A., 1993. Stopping rules in principal components analysis : a comparison of
425 heuristical and statistical approaches. *Ecology* Pp. 2204–2214.
- 426 Lange, K., J. C. Papp, J. S. Sinsheimer, and E. M. Sobel, 2014. Next generation statistical
427 genetics : Modeling, penalization, and optimization in high-dimensional data. *Annual*
428 *review of statistics and its application* 1 :279.

- 429 Lewontin, R. and J. Krakauer, 1973. Distribution of gene frequency as a test of the theory
430 of the selective neutrality of polymorphisms. *Genetics* 74 :175–195.
- 431 Lotterhos, K. E. and M. C. Whitlock, 2015. The relative power of genome scans to detect
432 local adaptation depends on sampling design and statistical method. *Molecular ecology*
433 24 :1031–1046.
- 434 Maronna, R. A. and R. H. Zamar, 2012. Robust estimates of location and dispersion for
435 high-dimensional datasets. *Technometrics* 44 :307–317.
- 436 Martins, H., K. Caye, K. Luu, M. G. Blum, and O. Francois, 2016. Identifying outlier loci
437 in admixed and in continuous populations using ancestral population differentiation
438 statistics. *bioRxiv* P. 054585.
- 439 Patterson, N., A. L. Price, and D. Reich, 2006. Population structure and eigenanalysis.
440 *PLoS genet* 2 :e190.
- 441 Peng, B. and M. Kimmel, 2005. simuPOP : a forward-time population genetics simulation
442 environment. *Bioinformatics* 21 :3686–3687.
- 443 Pritchard, J. K., M. Stephens, and P. Donnelly, 2000. Inference of population structure
444 using multilocus genotype data. *Genetics* 155 :945–959.
- 445 R Core Team, 2015. R : A Language and Environment for Statistical Computing. R Foun-
446 dation for Statistical Computing, Vienna, Austria. URL <https://www.R-project.org>.
- 447 Storey, J. D. and R. Tibshirani, 2003. Statistical significance for genomewide studies.
448 *Proceedings of the National Academy of Sciences* 100 :9440–9445.
- 449 Waples, R. S. and O. Gaggiotti, 2006. Invited review : What is a population ? an empirical
450 evaluation of some genetic methods for identifying the number of gene pools and their
451 degree of connectivity. *Molecular ecology* 15 :1419–1439.
- 452 Whitlock, M. C. and K. E. Lotterhos, 2015. Reliable detection of loci responsible for local
453 adaptation : Inference of a null model through trimming the distribution of *fst*. *The*
454 *American naturalist* 186 :S24–36.

455 Data Accessibility

456 Island and divergence model data : doi :10.5061/dryad.8290n

457 Range expansion simulated data : doi :10.5061/dryad.mh67v. Files :

458 2R_R30_1351142954_453_2_NumPops=30_NumInd=20

459 2R_R30_1351142954_453_2_NumPops=30_NumInd=60

460 2R_R30_1351142970_988_6_NumPops=30_NumInd=20

461 2R_R30_1351142970_988_6_NumPops=30_NumInd=60

462 2R_R30_1351142986_950_10_NumPops=30_NumInd=20

463 2R_R30_1351142986_950_10_NumPops=30_NumInd=60

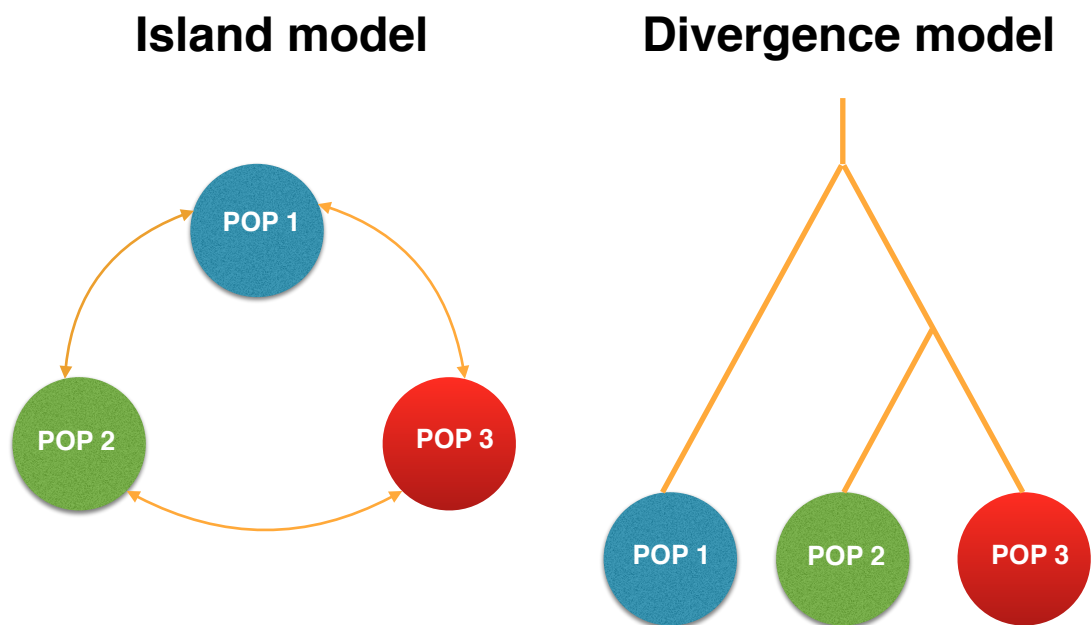


FIGURE SI1 – Schematic description of the island and divergence model. For the island model, adaptation occurs simultaneously in each population. For the island model, adaptation takes place in the branch leading to the second population.

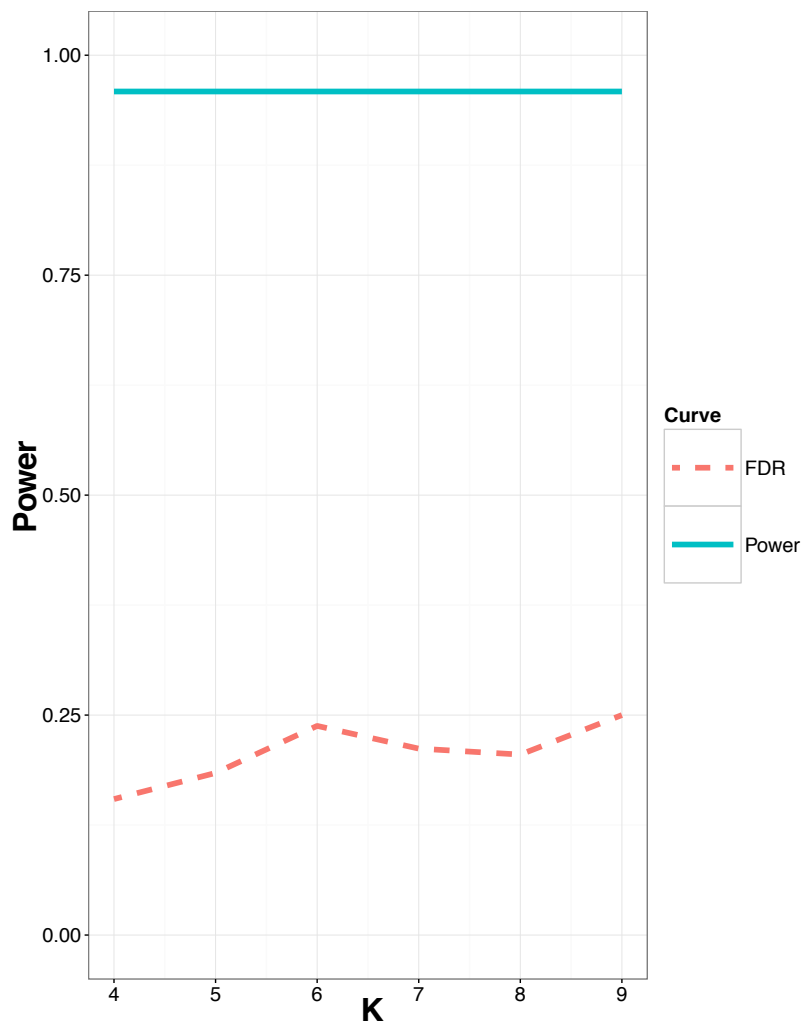


FIGURE SI2 – Proportion of false discoveries and statistical power as a function of the number of principal components in a model of range expansion.

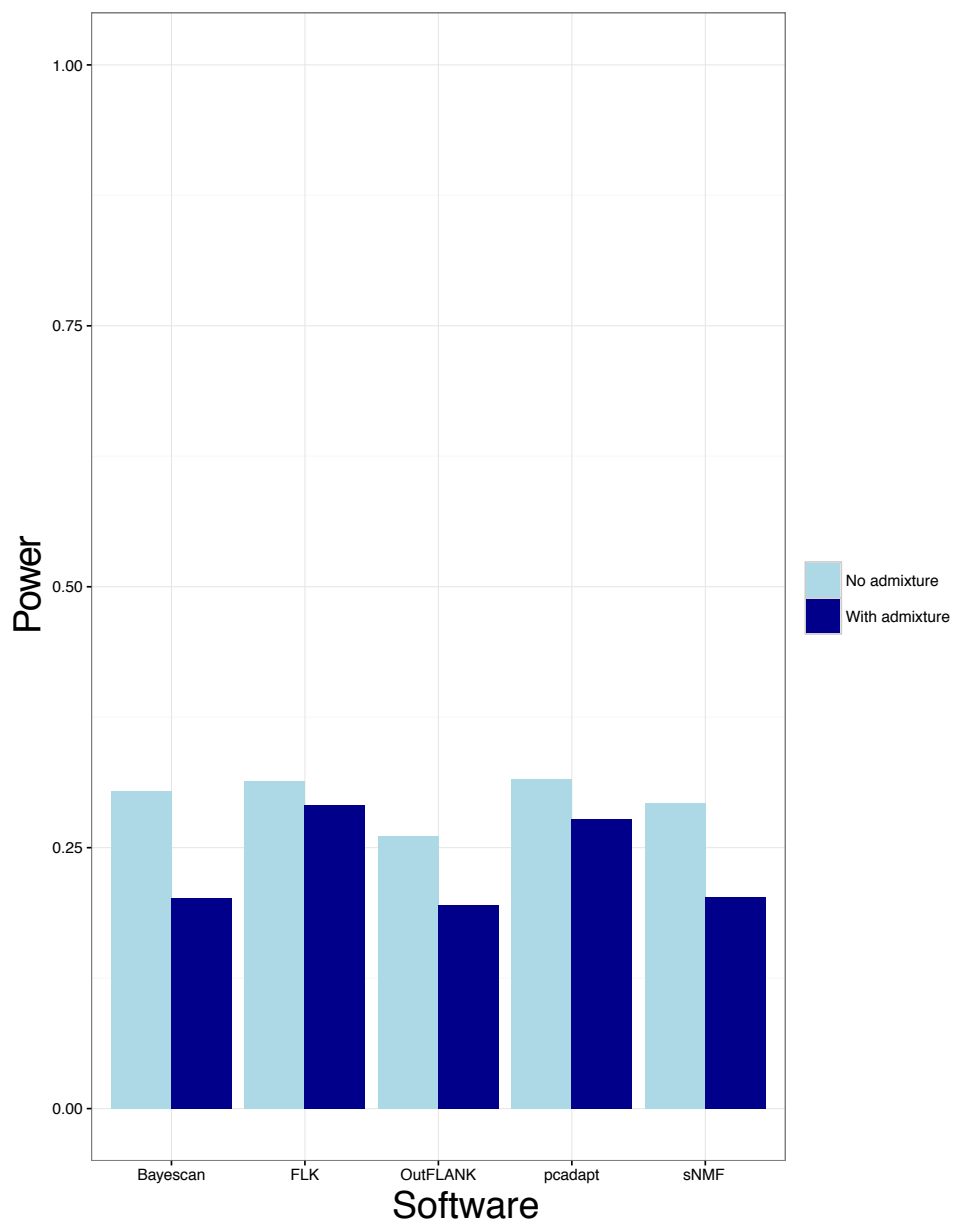


FIGURE SI3 – Statistical power averaged over the expected proportion of false discoveries (ranging between 0% and 50%) for the island model.

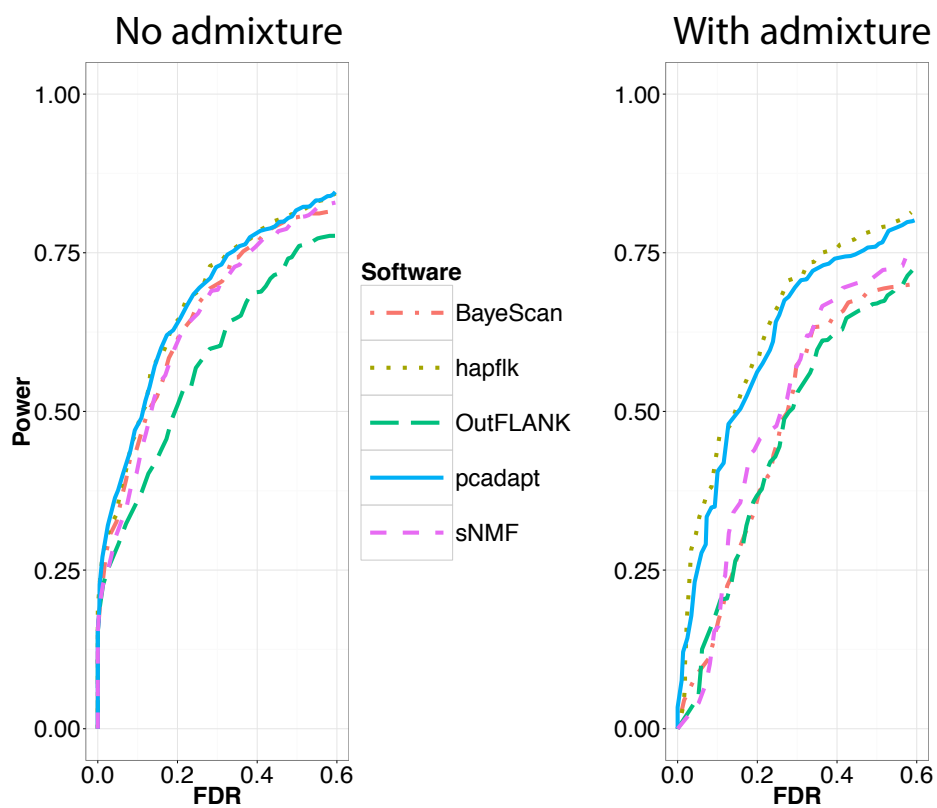


FIGURE SI4 – Statistical power as a function of the proportion of false discoveries for the island model.

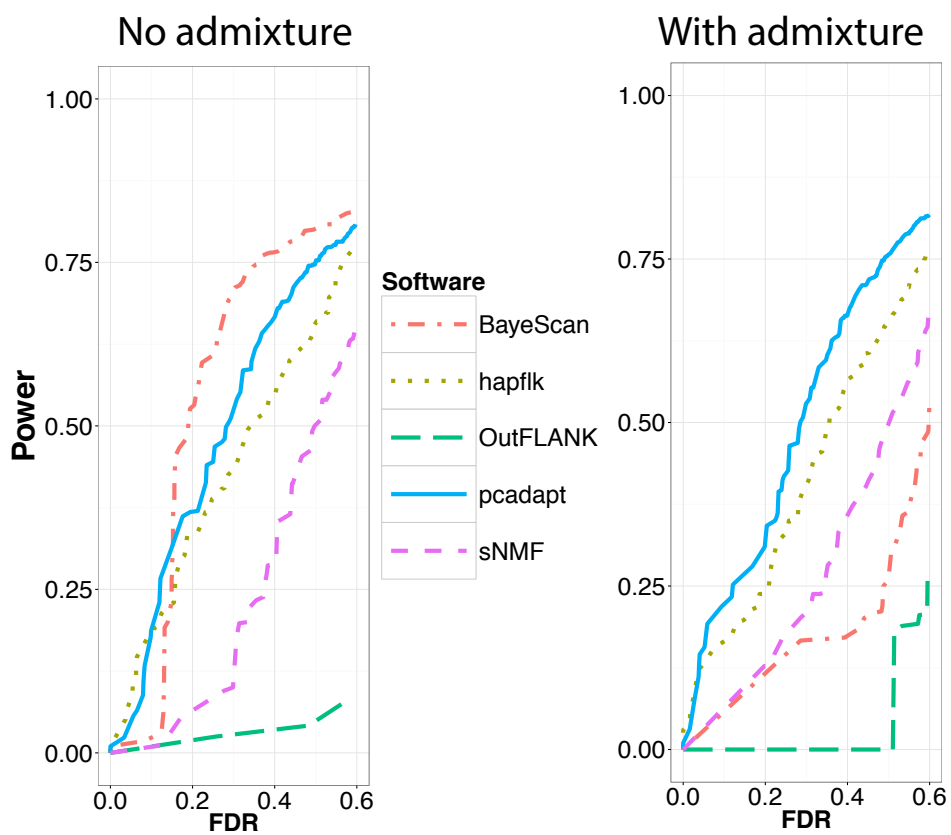


FIGURE SI5 – Statistical power as a function of the proportion of false discoveries for the divergence model.

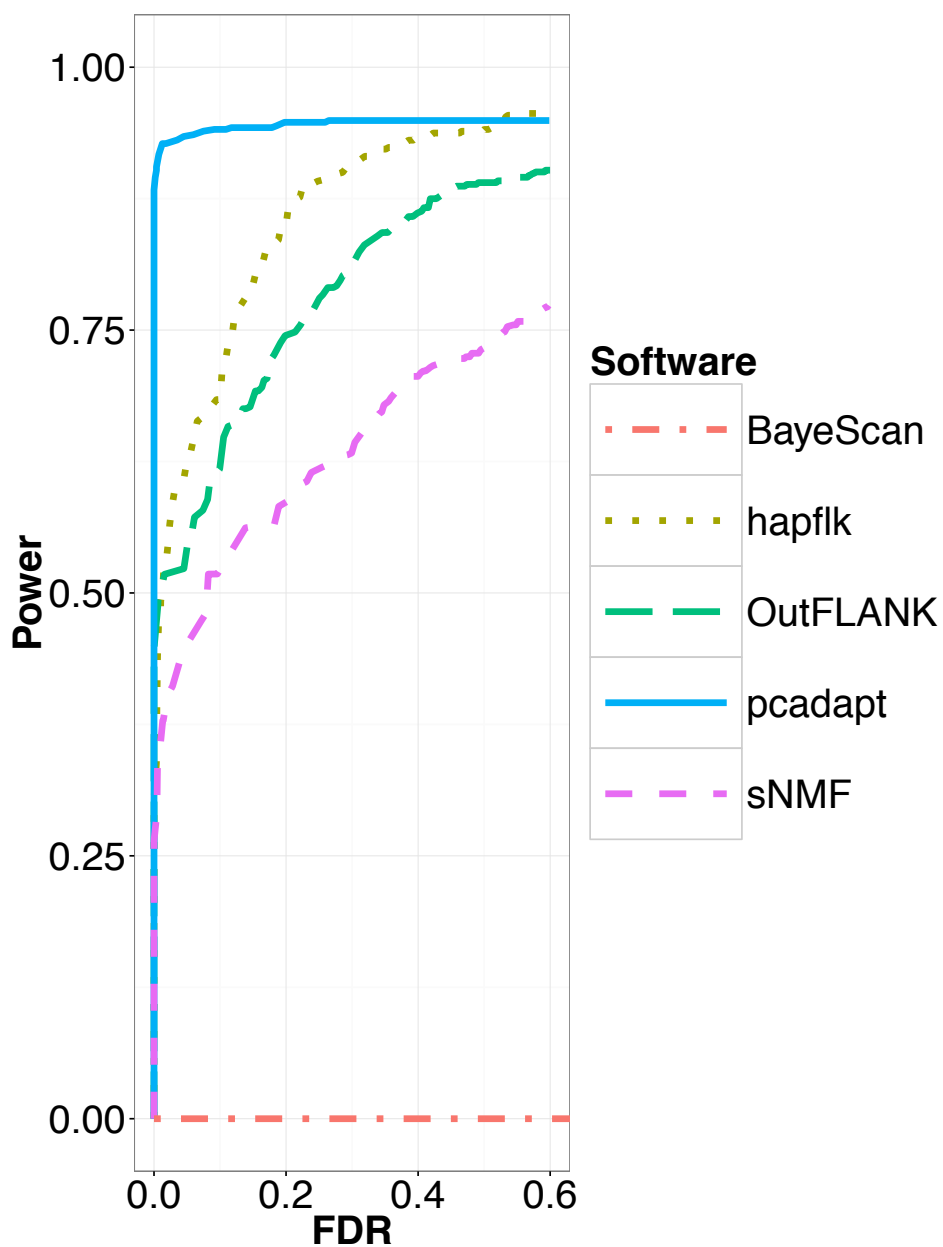


FIGURE SI6 – Statistical power as a function of the proportion of false discoveries for the model of range expansion.

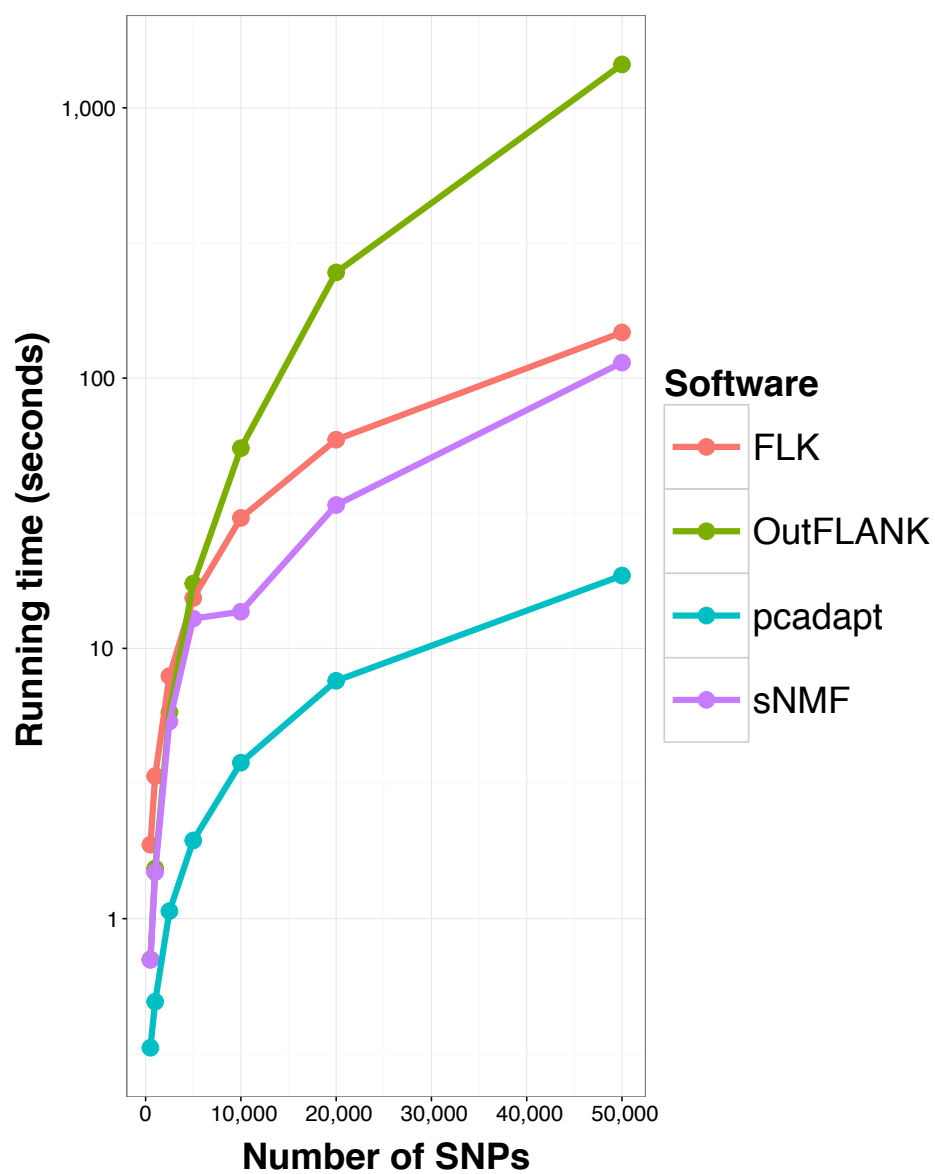


FIGURE SI7 – Running times of the different software. The different software were run on genotype matrices containing 300 individuals and from 500 to 50,000 SNPs. The characteristics of the computer we used to perform comparisons is the following : OSX El Capitan 10.11.3, 2,5 GHz Intel Core i5, 8 Go 1600 MHz DDR3.