MOLECULAR ECOLOGY
RESOURCES
Molecular Ecology Resources (2017) **17**, 67–77 doi: 10.1111/1755-0998.12592

SPECIAL ISSUE: POPULATION GENOMICS WITH R

pcadapt: an R package to perform genome scans for selection based on principal component analysis

KEURCIEN LUU,* ERIC BAZIN† and MICHAEL G. B. BLUM*
Laboratoire TIMC-IMAG, UMR 5525, CNRS, Université Grenoble Alpes, Grenoble, France, †Laboratoire d'Ecologie Alpine UMR 5553, CNRS, Université Grenoble Alpes, Grenoble, France

Abstract

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

Keywords: R package, Mahalanobis distance, outlier detection, population genetics, principal component analysis

Received 31 May 2016; revision received 29 July 2016; accepted 1 August 2016

Introduction

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

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

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

Correspondence: Michael G. B. Blum, Fax: +33 (0) 4 56 52 00 55; E-mail: michael.blum@imag.fr

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

To address the aforementioned challenges, we have developed the computer program *pcadapt* and the R package *pcadapt*. The computer program *pcadapt* is now deprecated and the R package only is maintained. *pcadapt* assumes that markers excessively related to population structure are candidates for local adaptation. Since its first release, *pcadapt* has substantially evolved in terms of both statistical approach and implementation (Table 1).

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

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

In addition to comparing the different test statistics that were implemented in *pcadapt*, we compare statistic performance obtained with the 3.0 version of *pcadapt* and with other computer programs for genome scans. We use simulated data to compare computer programs in terms of false discovery rate (FDR) and statistical power. We consider data simulated under different demographic models including island model, divergence model and range expansion. To perform comparisons, we include programs that require to group individuals into populations: *BayeScan* (Foll & Gaggiotti 2008), the F_{LK} statistic as implemented in the *hapflk* computer program (Bonhomme *et al.* 2010), and *OutFLANK* that provides a robust estimation of the null distribution of a F_{ST} test statistic (Whitlock & Lotterhos 2015). We additionally consider the *sNMF* computer program that implements another individual-based test statistic for genome scans (Frichot *et al.* 2014; Martins *et al.* 2016).

Statistical and computational approach

Input data

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

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

Test statistic	Pop. structure	Language	Command line	Versions of the R package	References
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 study

Choosing the number of principal components

In the following, we denote by n the number of individuals, by p the number of genetic markers and by G the genotype matrix that is composed of n lines and p columns. The genotypic information at locus j for individual i is encoded by the allele count G_{ij}, $1 \leq i \leq n$ and $1 \leq j \leq p$, which is a value in 0,1 for haploid species and in 0,1,2 for diploid species. The current 3.0.2 version of the package can handle haploid and diploid data only.

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

Then, we use the normalized genotype matrix $\tilde{G}$ to ascertain population structure with PCA (Patterson *et al*. 2006). The number of principal components to consider is denoted K and is a parameter that should be chosen by the user. In order to choose K, we recommend to consider the graphical approach based on the scree plot (Jackson 1993). The scree plot displays the eigenvalues of the covariance matrix Ω in descending order. Up to a constant, eigenvalues are proportional to the proportion of variance explained by each principal component. The eigenvalues that correspond to random variation lie on a straight line whereas the ones corresponding to population structure depart from the line. We recommend to use Cattell's rule that states that components corresponding to eigenvalues to the left of the straight line should be kept (Cattell 1966).

Test statistic

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

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

where β_{jk} is the regression coefficient corresponding to the j-th SNP regressed by the k-th principal component, 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 $z_j = (z_{j1}, ..., z_{jK})$ where z_{jk} corresponds to the z-score obtained when regressing the j-th SNP by the k-th principal component.

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

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

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

Genomic inflation factor

To perform multiple hypothesis testing, Mahalanobis distances should be transformed into P-values. If the z-scores were truly multivariate Gaussian, the Mahalanobis distances D should be chi-square distributed with K degrees of freedom. However, as usual for genome scans, there are confounding factors that inflate values of the test statistic and that would lead to an excess of false positives (François *et al*. 2016). To account for the inflation of test statistics, we divide Mahalanobis distances by a constant λ to obtain a statistic that can be approximated by a chi-square distribution with K degrees of freedom. This constant is estimated by the genomic inflation factor defined here as the median of the Mahalanobis distances divided by the median of the chi-square distribution with K degrees of freedom (Devlin & Roeder 1999).

Control of the false discovery rate (FDR)

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

Numerical computations

PCA is performed using a C routine that allows to compute scores and eigenvalues efficiently with minimum RAM access (Duforet-Frebourg *et al*. 2016). Computing the covariance matrix Ω is the most computationally

demanding part. To provide a fast routine, we compute the $n \times n$ covariance matrix Ω instead of the much larger $p \times p$ covariance matrix. We compute the covariance Ω incrementally by adding small storable covariance blocks successively. Multiple linear regression is then solved directly by computing an explicit solution, written as a matrix product. Using the fact that the (n,K) score matrix X is orthogonal, the (p,K) matrix $\hat{\beta}$ of regression coefficients is given by $G^T X$ and the (n,p) matrix of residuals is given by $G - XX^T G$. The z-scores are then computed using the standard formula for multiple regression

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

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

Missing data

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

Pooled sequence data

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

Materials and methods

Simulated data

We simulated SNPs under an island model, under a divergence model and we downloaded simulations of range expansion (Lotterhos & Whitlock 2015). All data we simulated were composed of 3 populations, each of them containing 50 sampled diploid individuals (Table 2). SNPs were simulated assuming no linkage disequilibrium. SNPs with minor allele frequencies lower than 5% were discarded from the data sets. The mean F_{ST} for each simulation was comprised between 5% and 10%. Using the simulations based on an island and a divergence model, we also created data sets composed of admixed individuals. We assumed that an instantaneous admixture event occurs at the present time so that all sampled individuals are the results of this admixture event. Admixed individuals were generated by drawing randomly admixture proportions using a Dirichlet distribution of parameter (α,α,α) (α ranging from 0.005 to 1 depending on the simulation).

Island model

We used *ms* to create simulations under an island model (Fig. S1). We set a lower migration rate for the 50 adaptive SNPs compared with the 950 neutral ones to mimic diversifying selection (Bazin *et al.* 2010). For a given locus, migration from population i to j was specified by choosing a value of the effective migration rate that is set to $M_{\text{neutral}} = 10$ for neutral SNPs and to M_{adaptive} for adaptive ones. We simulated 35 data sets in the island model with different strengths of selection, where the strength of selection corresponds to the ratio $M_{\text{neutral}}/M_{\text{adaptive}}$ that varies from 10 to 1000. The *ms* command lines for neutral and adaptive SNPs are given by ($M_{\text{adaptive}} = 0.01$ and $M_{\text{neutral}} = 10$).

```
./ms 300 950 -s 1 -I 3 100 100 100
 -ma x 10 10 10 x 10 10 10 x
./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
```

Divergence model

To perform simulations under a divergence model, we used the package *simuPOP*, which is an individual-based population genetic simulation environment (Peng &

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

	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

Kimmel 2005). We assumed that an ancestral panmictic population evolved during 20 generations before splitting into two subpopulations. The second subpopulation then split into subpopulations 2 and 3 at time $T > 20$. All 3 subpopulations continued to evolve until 200 generations have been reached, without migration between them (Figure S1). A total of 50 diploid individuals were sampled in each population. Selection only occurred in the branch associated with population 2 and selection was simulated by assuming an additive model (fitness is equal to $1-2s, 1-s, 1$ depending on the genotypes). We simulated a total of 3000 SNPs comprising of 100 adaptive ones for which the selection coefficient is of $s = 0.1$.

Range expansion

We downloaded in the *Dryad Digital Repository* six simulations of range expansion with two glacial refugia (Lotterhos & Whitlock 2015). Adaptation occurred during the recolonization phase of the species range from the two refugia. We considered six different simulated data with 30 populations and a number of sampled individual per location that varies from 20 to 60.

Parameter settings for the different computer programs

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

Results

Choosing the number of principal components

We evaluate Cattell's graphical rule to choose the number of principal components. For the island and divergence model, the choice of K is evident (Fig. 1). For $K \geq 3$, the eigenvalues follow a straight line. As a consequence, Cattell's rule indicates K=2, which is expected because there are three populations (Patterson *et al.* 2006). For the model of range expansion, applying Cattell's rule to choose K is more difficult (Fig. 1). Ideally, the eigenvalues that correspond to random variation lie on a straight line whereas the ones corresponding to population structure depart from the line. However, there is no obvious point at which eigenvalues depart from the straight line. Choosing a value of K between 5 and 8 is compatible with Cattell's rule. Using the package *qvalue* to control 10% of FDR, we find that the actual proportion of false discoveries as well as statistical power is weakly impacted when varying the number of principal components from $K = 5$ to $K = 8$ (Figure S2).

An example of genome scans performed with pcadapt

To provide an example of results, we apply *pcadapt* with $K = 6$ in the model of range expansion. Population structure captured by the first two principal components is displayed in Fig. 2. *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 (Fig. 2). Using a FDR threshold of 10% with the *qvalue* package, we find 122 outliers among 10 000 SNPs, resulting in 23% actual false discoveries and a power of 95%.

Control of the false discovery rate

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

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

Statistical power

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

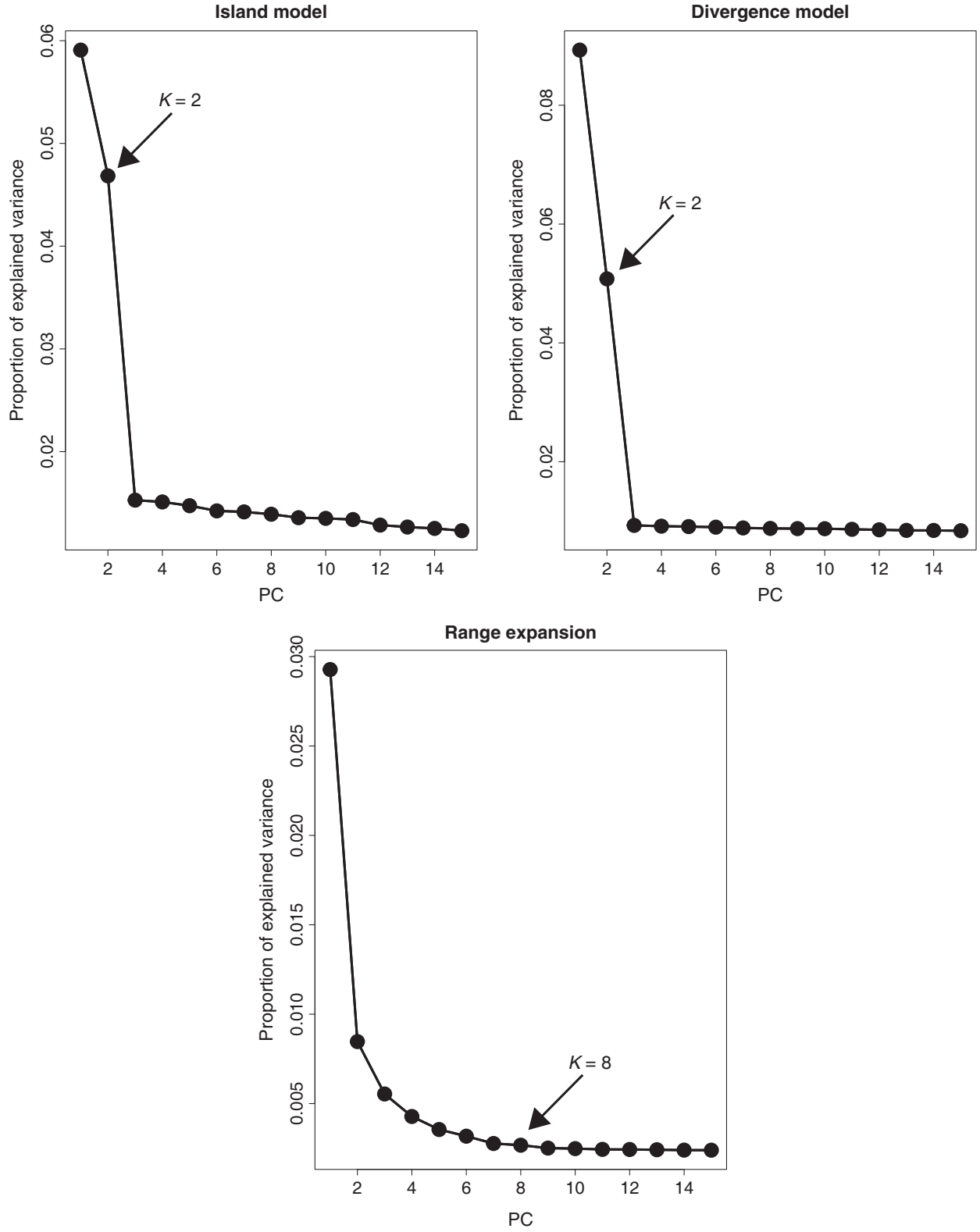

Fig. 1 Determining *K* with the scree plot. To choose *K*, we recommend to use Cattell's rule that states that components corresponding to eigenvalues to the left of the straight line should be kept. According to Cattell's rule, the eigenvalues that correspond to random variation lie on the straight 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.

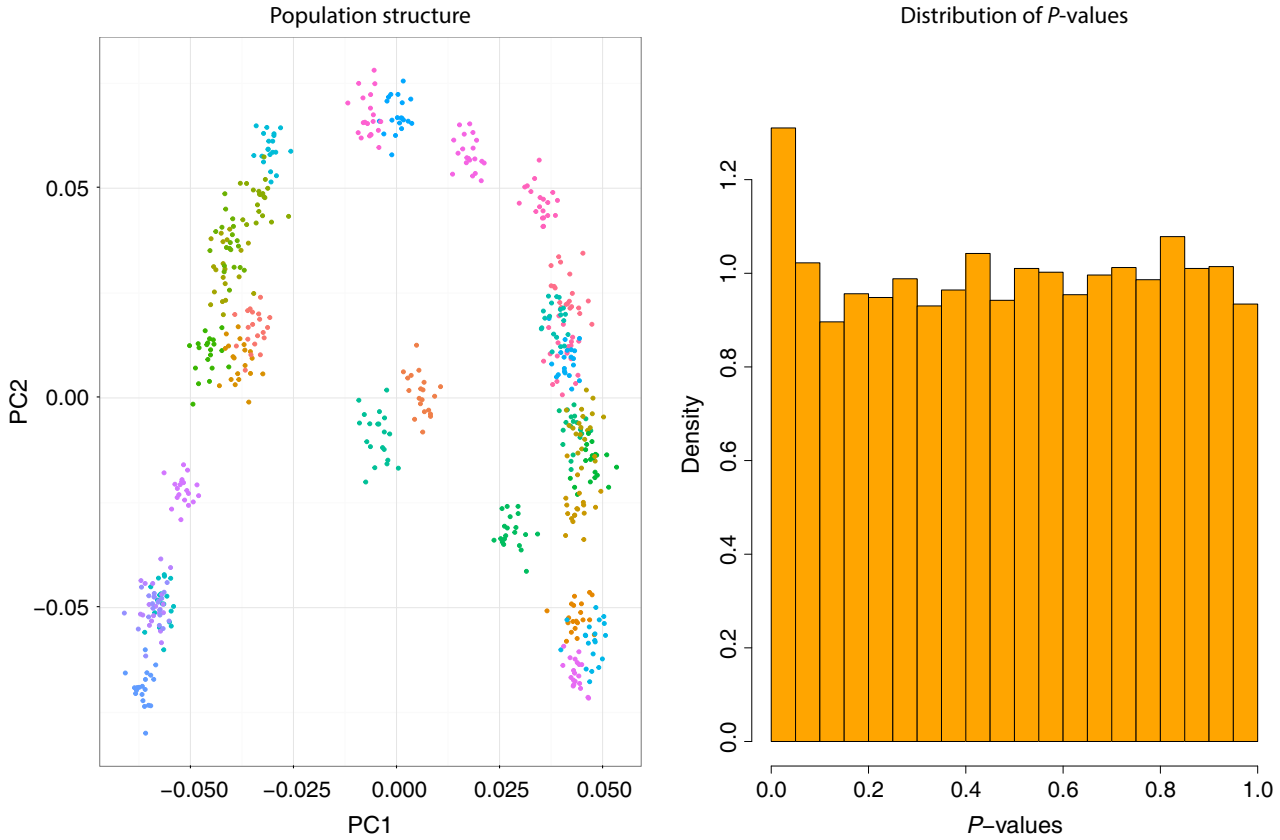

Fig. 2 Population structure (first 2 principal components) and distribution of *p*-value obtained with *pcadapt* for a simulation of range expansion. *P*-values are well calibrated because they are distributed as a mixture of a uniform distribution and of a peaky distribution around 0, which corresponds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

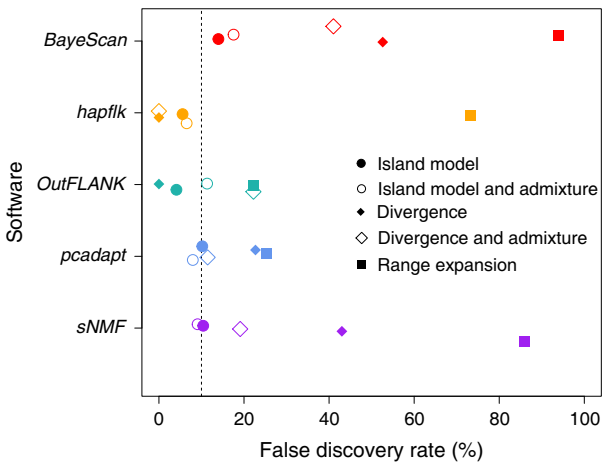

Fig. 3 Control of the FDR for different computer programs for genome scans. We find that the median proportion of false discoveries is around the nominal FDR set at 10% (6% for *hapflk*, 11% for both *OutFLANK* and *pcadapt* and 19% for *sNMF*) with the exception of *BayeScan* that generates 41% of false discoveries. Comparison of statistical power for the different test statistics that have been implemented in *pcadapt* (Table 1).

We first compare statistical power obtained with the different statistical methods that have been implemented in *pcadapt* (Table 1). For the island model, Bayes factor, communality statistic and Mahalanobis distance have similar power (Fig. 4). For the divergence model, the power obtained with Mahalanobis distance is 20% whereas the power obtained with the communality statistic and with the Bayes factor is, respectively, 4% and 2% (Fig. 4). Similarly, for range expansion, the power obtained with Mahalanobis distance is 46% whereas the power obtained with the communality statistic and with the Bayes factor is 34% and 13%. We additionally investigate to what extent increasing sample size in each population from 20 to 60 individuals affects power. For range expansion, the power obtained with the Mahalanobis distance hardly changes ranging from 44% to 47%. However, the power obtained with the other two statistics changes importantly. The power obtained with the communality statistic increases from 27% to 39% when increasing the sample size and the power obtained with the Bayes factor increases from 0% to 44%.

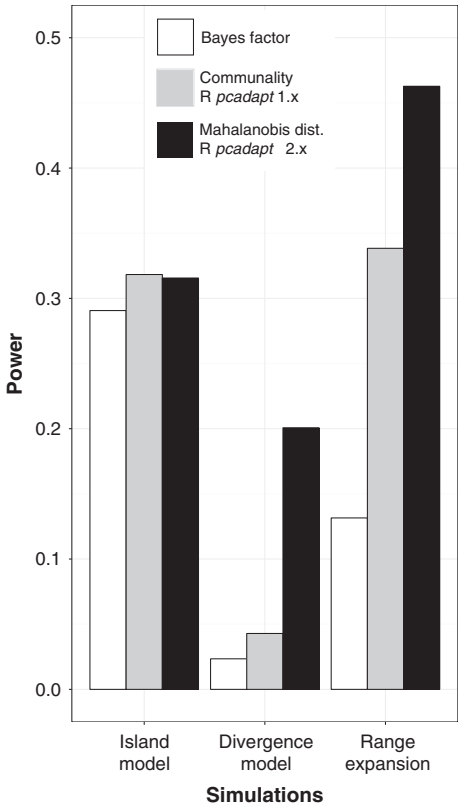

Fig. 4 Bayes factor corresponds to the test statistic implemented in the Bayesian version of *pcadapt* (Duforet-Frebourg *et al.* 2014); the communality statistic was the default statistic in version 1.x of the R package *pcadapt* (Duforet-Frebourg *et al.* 2016), and Mahalanobis distances are available 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 with 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%).

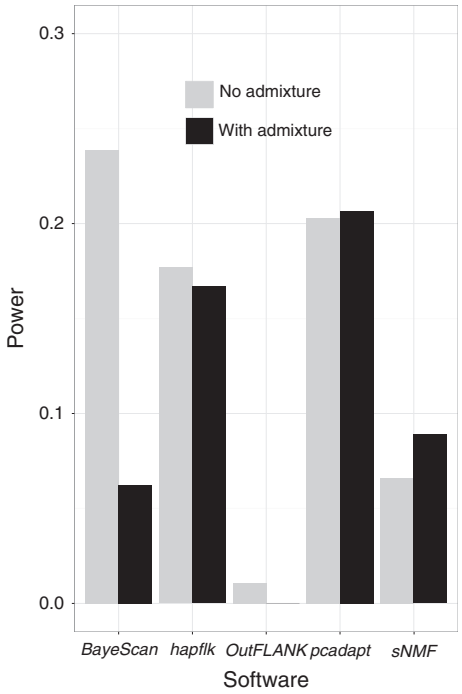

Fig. 5 Statistical power averaged over the expected proportion of false discoveries (ranging between 0% and 50%) for the divergence model with three populations. We assume that adaptation took place in an external branch that follows the most recent population divergence event.

Then we describe our comparison of computer programs for genome scans. For the simulations obtained with the island model where there is no hierarchical population structure, the statistical power is similar for all programs (Figure S3 and S4). Including admixed individuals hardly changes their statistical power (Figure S3).

Then, we compare statistical power in a divergence model where adaptation took place in one of the external branches of the population divergence tree. The programs *pcadapt* and *hapflk*, which account for hierarchical population structure, as well as *BayeScan* are the most powerful in that setting (Fig. 5 and Figure S5). The values of power in decreasing order are of 23% for *BayeScan*, of 20% for *pcadapt*, of 17% for *hapflk*, of 7% for *sNMF* and of 1% for *OutFLANK*. When including admixed individuals, the power of *hapflk* and of *pcadapt* hardly decreases whereas the power of *BayeScan* decreases to 6% (Fig. 5).

The last model we consider is the model of range expansion. The package *pcadapt* is the most powerful approach in this setting (Fig. 6 and S6). Other computer programs also discover many true-positive loci with the exception of *BayeScan* that provides no true discovery when the observed FDR is smaller than 50% (Fig. 6 and S6). The values of power in decreasing order are of 46% for *pcadapt*, of 41% for *hapflk*, of 37% for *OutFLANK*, of 30% for *sNMF* and of 0% for *BayeScan*.

Running time of the different computer programs

Last, we compare running times. The characteristics of the computer we used to perform comparisons are the following: OSX El Capitan 10.11.3, 2,5 GHz Intel Core i5, 8 Go 1600 MHz DDR3. We discard *BayeScan* as it is too time-consuming. For instance, running *BayeScan* on a genotype matrix containing 150 individuals and 3000 SNPs takes 9 h whereas it takes less than one second with *pcadapt*. The different programs were run on genotype matrices containing 300 individuals and from 500 to 50 000 SNPs. *OutFLANK* is the computer program for

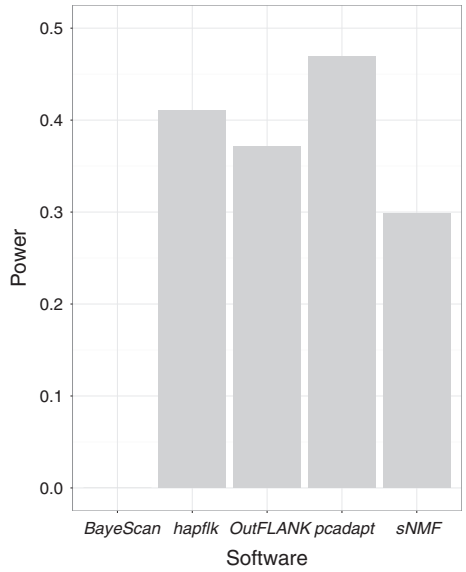

Fig. 6 Statistical power averaged over the expected proportion of false discoveries (ranging between 0% and 50%) for a range expansion model with two refugia. Adaptation took place during the recolonization event.

which the runtime increases the most rapidly with the number of markers. *OutFLANK* takes around 25 min to analyse 50 000 SNPs (Figure S7). For the other 3 computer programs (*hapflk*, *pcadapt*, *sNMF*), analysing 50 000 SNPs takes <3 min.

Discussion

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

Simulations show that the R package *pcadapt* compares favourably to other computer programs for genome scans. When data were simulated under an island model, population structure is not hierarchical because genetic differentiation is the same for all pairs of populations. Statistical power and control of the FDR were similar for all computer programs. In the presence of hierarchical population structure (divergence model) where genetic differentiation varies between pairs of populations, the ranking of the SNPs depends on the computer program. *pcadapt* and *hapflk* provide the most powerful scans whether or not simulations include admixed individuals. *OutFLANK* implements a F_{ST} statistic and because adaptation does not correspond to the most differentiated populations, it fails to capture adaptive SNPs (Fig. 5) (Bonhomme *et al.* 2010; Duforet-Frebourg *et al.* 2016). *BayeScan* does not assume equal differentiation between all pairs of populations, which may explain why it has a good statistical power for the divergence model. However, its statistical power is severely impacted by the presence of admixed individuals because its power decreases from 24% to 6% (Fig. 5). Understanding why *BayeScan* is severely impacted by admixture is out of the scope of this study. In the range expansion model, *BayeScan* returns many null q-values (between 376 and 809 SNPs of 9899 neutral and 100 adaptive SNPs) such that the observed FDR is always larger than 50%. Overall, we find that *pcadapt* and *hapflk* provides comparable statistical power. They provide optimal or near optimal ranking of the SNPs in different scenarios including hierarchical population structure and admixed individuals. The main difference between the two computer programs concerns the control of the FDR because *hapflk* is found to be more conservative.

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

structure in the face of large amount of missing data is a challenge for modern numerical methods in molecular ecology.

Acknowledgements

This work has been supported by the LabEx PERSYVAL-Lab (ANR-11-LABX-0025-01) and the ANR AGRHUM project (ANR-14-CE02-0003-01). We want to thank two anonymous reviewers and Stephane Dray for their critical reading of our manuscript.

References

Abraham G, Inouye M (2014) Fast principal component analysis of large-scale genome-wide data. *PloS One*, **9**, e93766.

Arnold B, Corbett-Detig R, Hartl D, Bomblies K (2013) RADseq underestimates diversity and introduces genealogical biases due to nonrandom haplotype sampling. *Molecular Ecology*, **22**, 3179–3190.

Bazin E, Dawson KJ, Beaumont MA (2010) Likelihood-free inference of population structure and local adaptation in a Bayesian hierarchical model. *Genetics*, **185**, 587–602.

Bierne N, Roze D, Welch JJ (2013) Pervasive selection or is it....? Why are FST outliers sometimes so frequent? *Molecular Ecology*, **22**, 2061–2064.

Bonhomme M, Chevalet C, Servin B, Boitard S, Abdallah J, Blott S, San-Cristobal M (2010) Detecting selection in population trees: the Lewontin and Krakauer test extended. *Genetics*, **186**, 241–262.

Cattell RB (1966) The scree test for the number of factors. *Multivariate Behavioral Research*, **1**, 245–276.

Chen G-B, Lee SH, Zhu Z-X, Benyamin B, Robinson MR (2016) EigenGWAS: finding loci under selection through genome-wide association studies of eigenvectors in structured populations. *Heredity*, **117**, 51–61.

Devlin B, Roeder K (1999) Genomic control for association studies. *Biometrics*, **55**, 997–1004.

Dray S, Josse J (2015) Principal component analysis with missing values: a comparative survey of methods. *Plant Ecology*, **216**, 657–667.

Duforet-Frebourg N, Bazin E, Blum MGB (2014) Genome scans for detecting footprints of local adaptation using a Bayesian factor model. *Molecular Biology and Evolution*, **31**, 2483–2495.

Duforet-Frebourg N, Luu K, Laval G, Bazin E, Blum MGB (2016) Detecting genomic signatures of natural selection with principal component analysis: application to the 1000 genomes data. *Molecular Biology and Evolution*, **33**, 1082–1093.

Excoffier L, Hofer T, Foll M (2009) Detecting loci under selection in a hierarchically structured population. *Heredity*, **103**, 285–298.

Foll M, Gaggiotti O (2008) A genome-scan method to identify selected loci appropriate for both dominant and codominant markers: a Bayesian perspective. *Genetics*, **180**, 977–993.

François O, Martins H, Caye K, Schoville SD (2016) Controlling false discoveries in genome scans for selection. *Molecular Ecology*, **25**, 454–469.

Frichot E, Mathieu F, Trouillon T, Bouchard G, François O (2014) Fast and efficient estimation of individual ancestry coefficients. *Genetics*, **196**, 973–983.

Galinsky KJ, Bhatia G, Loh P-R, Georgiev S, Mukherjee S, Patterson NJ, Price AL (2016) Fast principal components analysis reveals independent evolution of adh1b gene in Europe and East Asia. *American Journal of Human Genetics*, **98**, 456–472. 018143

Gautier M, Gharbi K, Cezard T, Foucaud J, Kerdelhué C, Pudlo P, Cornuet J-M, Estoup A (2013) The effect of RAD allele dropout on the estimation of genetic variation within and between populations. *Molecular Ecology*, **22**, 3165–3178.

Halko N, Martinsson P-G, Tropp JA (2011) Finding structure with randomness: probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Review*, **53**, 217–288.

Hao W, Song M, Storey JD (2016) Probabilistic models of genetic variation in structured populations applied to global human studies. *Bioinformatics*, **32**, 713–721.

Jackson DA (1993) Stopping rules in principal components analysis: a comparison of heuristical and statistical approaches. *Ecology*, **74**, 2204–2214.

Lange K, Papp JC, Sinsheimer JS, Sobel EM (2014) Next generation statistical genetics: modeling, penalization, and optimization in high-dimensional data. *Annual Review of Statistics and Its Application*, **1**, 279.

Lewontin R, Krakauer J (1973) Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms. *Genetics*, **74**, 175–195.

Lotterhos KE, Whitlock MC (2015) The relative power of genome scans to detect local adaptation depends on sampling design and statistical method. *Molecular Ecology*, **24**, 1031–1046.

Maronna RA, Zamar RH (2012) Robust estimates of location and dispersion for high-dimensional datasets. *Technometrics*, **44**, 307–317.

Martins H, Caye K, Luu K, Blum MG, François O (2016) Identifying outlier loci in admixed and in continuous populations using ancestral population differentiation statistics. *bioRxiv*, 054585.

Patterson N, Price AL, Reich D (2006) Population structure and eigenanalysis. *PLoS Genet*, **2**, e190.

Peng B, Kimmel M (2005) simuPOP: a forward-time population genetics simulation environment. *Bioinformatics*, **21**, 3686–3687.

Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. *Genetics*, **155**, 945–959.

R Core Team (2015) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org.

Storey JD, Tibshirani R (2003) Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences of the United States of America*, **100**, 9440–9445.

Waples RS, Gaggiotti O (2006) Invited review: what is a population? An empirical evaluation of some genetic methods for identifying the number of gene pools and their degree of connectivity. *Molecular Ecology*, **15**, 1419–1439.

Whitlock MC, Lotterhos KE (2015) Reliable detection of loci responsible for local adaptation: inference of a null model through trimming the distribution of FST. *The American Naturalist*, **186**, S24–S36.

K.L., E.B. and M.G.B.B. designed and performed the research.

Data accessibility

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

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

```
2R_R30_1351142954_453_2_NumPops=30_NumInd=20
2R_R30_1351142954_453_2_NumPops=30_NumInd=60
2R_R30_1351142970_988_6_NumPops=30_NumInd=20
2R_R30_1351142970_988_6_NumPops=30_NumInd=60
2R_R30_1351142986_950_10_NumPops=30_NumInd=20
2R_R30_1351142986_950_10_NumPops=30_NumInd=60
```

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Fig. S1 Schematic description of the island and divergence model.

Fig. S2 Proportion of false discoveries and statistical power as a function of the number of principal components in a model of range expansion.

Fig. S3 Statistical power averaged over the expected proportion of false discoveries (ranging between 0% and 50%) for the island model.

Fig. S4 Statistical power as a function of the proportion of false discoveries for the island model.

Fig. S5 Statistical power as a function of the proportion of false discoveries for the divergence model.

Fig. S6 Statistical power as a function of the proportion of false discoveries for the model of range expansion.

Fig. S7 Running times of the different computer programs.