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

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

[1] *Université Grenoble Alpes, Centre National de la Recherche Scientifique, Laboratoire TIMC-IMAG, UMR 5525, Grenoble, France.*

[2] *Université Grenoble Alpes, Centre National de la Recherche Scientifique, Laboratoire d'Ecologie Alpine UMR 5553, Grenoble, France*

Abstract

We introduce the R package *pcadapt* that performs genome scans to detect genes under selection based on population genomic data. The statistical method implemented in *pcadapt* assumes that markers excessively related with population structure are candidates for local adaptation. Because population structure is ascertained with principal component analysis (PCA), the package is fast and can handle large-scale data generated with next-generation technologies. It can also handle missing data as well as data obtained from pooled sequencing. By contrast to population-based approaches, the package can handle admixed individuals and does not require to group individuals into predefined populations. Using data simulated under an island model, a divergence model and range expansion, we compare *pcadapt* to other software performing genome scans (*BayeScan*, *hapflk*, *OutFLANK*, *sNMF*). For the different software, the average proportion of false discoveries is around the nominal false discovery rate set at 10% with the exception of *BayeScan* that generates 40% of false discoveries. When comparing statistical power for a realized percentage of false discoveries, we find that the power of *BayeScan* can be severely impacted by the presence of admixed individuals whereas *pcadapt* is not impacted. Last, we show that *pcadapt* is the most powerful method in a model of range expansion where population structure is continuous. Because *pcadapt* can handle molecular data generated with next sequencing technologies, we anticipate that it will be a valuable tool for modern analysis in molecular ecology.

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

Introduction

Looking for variants with unexpectedly large differences of allele frequencies between populations is a common approach to detect signals of natural selection (Lewontin and 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 a chi-square approximation of a F_{ST}-related test statistic (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 a hierarchical structure, when it occurs, generate a large excess of false positive loci (Bierne et al., 2013; Excoffier et al., 2009).

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 and Gaggiotti, 2006). Individual sampling may not be population-based but based on more continuous sampling schemes (Lotterhos and 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 there are based on related techniques of multivariate analysis including principal component analysis, factor models, and non-negative matrix factorization (Duforet-Frebourg et al., 2014; Hao et al., 2016; Galinsky et al., 2016; Duforet-Frebourg et al., 2016; Martins et al., 2016).

The last challenge arises from the nature of multilocus datasets generated from next generation sequencing platforms. Because datasets 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, data generated from next generation sequencing platforms 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 *pcadapt* `R` package. The statistical method implemented in *pcadapt* assumes that markers excessively related with population structure are candidates for local adaptation. Because *pcadapt* is based on Principal Component Analysis, it is fast and can handle large-scale data generated with next-generation technologies.

We use simulated data to compare software in terms of false discovery rate and statistical power. To perform comparisons, we included population-based software : *BayeScan* that implements a Bayesian algorithm to detect outliers (Foll and Gaggiotti, 2008), the F_{LK} statistic as implemented in the *hapflk* sotware that accounts for hierarchical population structure (Bonhomme et al., 2010), and *OutFLANK* that provides a robust estimation of the null distribution of a F_{ST} test statistic (Whitlock and Lotterhos, 2015). We additionally consider the *sNMF* software that implements an 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 that 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.

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.

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 principal component analysis (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 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 $X_1, \ldots, X_K$

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

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 $z_j = (z_{j1}, \ldots, z_{jK})$ 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 the

Mahalanobis distance D which is defined as

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

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. When $K > 1$, the covariance matrix is estimated with the Orthogonalized Gnanadesikan-Kettenring method that is a robust estimate of the covariance able to handle large-scale data (Maronna and 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, we found that Mahalanobis distances should rather be divided by a constant λ, to be approximated by a chi-square distribution with K degrees of freedom. This constant, called the *genomic inflation factor*, is estimated using the median of the Mahalanobis distances divided by the median of the chi-square distribution with K degrees of freedom λ (Devlin and 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 (François et al., 2016). 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 and Tibshirani, 2003). The *qvalue* R package transform 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

Principal component analysis 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 is the most computationally demanding part. To provide a fast routine, we compute the $n \times n$ covariance matrix instead of the much larger $p \times p$ covariance matrix. We compute the covariance 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

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

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 and 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 sequencing

When data have been sequenced in pool, the Mahalanobis distance is not based on the matrix of z-score anymore but on the matrix of allele frequency computed in each pool.

Materials and Methods

Simulated data

We generated simulated data under different demographic scenarios using the *ms* software and the Python package *simuPOP* (Hudson, 2002; Peng and Kimmel, 2005) (Fig SI1). All simulated data are composed of 3 populations, each of them containing 50 sampled diploid individuals. SNPs were simulated assuming no linkage disequilibrium. SNPs with minor allele frequencies lower than 5% were discarded from the datasets. The mean F_{ST} for each simulation is comprised between 5% and 10%. Using the simulations based on a island and a divergence model, we also created datasets composed of admixed individuals. We assume 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). In addition to these simulations, we also included simulations of range expansion in a continuous square landscape (Lotterhos and Whitlock, 2015). These simulations assume that there are two refugia from which range expansion occurred. The entire landscape was colonized during range expansion and adaptation is assumed to occur during the expansion process.

Island model

We used *ms* to create simulations under an island model (Fig SI1). We set a lower migration rate for the 50 adaptive SNPs compared to the 950 neutral ones to mimick diversifying selection (Bazin et al., 2010). For a given locus, migration from population i to j is 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 SNP data in the island model with different strengths of selection, where the strength of selection corresponds to the ratio $M_{\text{neutral}}/M_{\text{adaptive}}$, which varies here from 10 to $1,000$. 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

For the divergence models, we used the package *simuPOP*, which is an individual-based population genetic simulation environment (Peng and Kimmel, 2005), to simulate 6 data sets. An ancestral panmictic population evolves during 20 generations before splitting into two subpopulations. The second subpopulation then splits into another two subpopulations 2 and 3 at time $T > 20$, and all 3 subpopulations 1, 2 and 3 continue to evolve until 200 generations have been reached, without migration between them (Figure SI1). A total of 50 diploid individuals are sampled in each population. Selection only occurs in the branch associated with population 2 and selection is simulated by assuming an additive model (fitness is equal to $1 - 2s, 1 - s, 1$ depending on the genotypes). We simulated a total of $3,000$ SNPs comprising of 100 adaptive ones for which the selection coefficient is of $s = 0.1$.

Range expansion

We consider 6 simulations of range expansion with two glacial refugiees (Lotterhos and Whitlock, 2015). Adaptation occurs during the recolonization phase of the species range from the two refugia. We consider 6 different simulated data with 30 populations and a number of sampled individual per population that varies from 20 to 60.

Parameter settings for the different software

When using *hapflk*, we set $K = 1$ that corresponds to the computation of the FLK statistic. When using *BayeScan* and *OutFLANK*, we use the default parameter values. For *sNMF*, we use $K = 3$ 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 use the genomic inflation factor to recalibrate p-values. When using population-based methods with admixed individuals, we use known ancestries for each individual to make assignment using a Max rule.

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 (Figure 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 3 populations (Patterson et al., 2006). For the model of range expansion, applying Cattell's rule to choose K is more difficult (Figure 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 false discovery rate, 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 SI2).

An example of genome scans performed with *pcadapt*

To provide an example of results, we applied *pcadapt* with $K = 6$ in the model of range expansion. Population structure captured by the first 2 principal components is displayed in Figure 2. The p-values are well calibrated since there is a mixture of uniform distribution and of a peaky distribution around 0, which corresponds to outlier loci (Figure 2). Controlling 10% of false discoveries, 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 *pcadapt* and the *qvalue* packages can control a False Discovery rate set at 10% (Figure 3). All SNPs with a q-value smaller than 10% were considered as candidate SNPs for the simulations. For the island model, we find that the average proportion of false discoveries is of 8% and it increases to 10% when including admixture. For the divergence model, the average proportion of false discoveries is of 11% and it increases to 22% when including admixture. The largest average proportion of false discoveries is obtained under range expansion and is equal to 25%.

We then evaluate the average proportion of false discoveries obtained with *BayeScan*, *hapflk*, *OutFLANK*, and *sNMF* (Figure 3). We find that *hapflk* is the most conservative approach (FDR = 6%) followed by *OutFLANK* and *pcadapt* (FDR = 11%). The software *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 software, the range expansion scenario is the one that generates the largest proportion of false discoveries with expected proportion of false discoveries ranging from 22% (*OutFLANK*) to 93% (*BayeScan*).

Statistical power

To provide a fair comparison between software, we compare statistical power of the different software 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%.

For the simulations obtained with the island model where there is no hierarchical population structure, the statistical power is similar for all software (Figure SI3 and SI4). Including admixed individuals hardly changes their statistical power (Figure SI3).

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 software *pcadapt* and *hapflk*, which account for hierarchical population structure, as well as *BayeScan* are the most powerful in that setting (Figure 4 and Figure SI5). The values of averaged power in decreasing order are of 23% for *BayeScan*, of 20% for *pcadapt*, of 17% for *hapflk*, of 4% for *sNMF* and of 0% for *OutFLANK*. When including admixed individuals, the power of *hapflk* and of *pcadapt* hardly decreases whereas the power of *BayeScan* decreases to 0% (Figure 4).

The last model we consider is the model of range expansion. The package *pcadapt* is the most powerful approach in this setting but other software also discover many true positive loci with the exception of *BayeScan* that provides no true discovery (Figure 5 and SI6) The values of averaged 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 software

Last, we compared the software running times. 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. We discarded *BayeScan* as it is too time consuming compared to other software. For instance, running *BayeScan* on a genotype matrix containing 150 individuals and $1,500$ SNPs took 9 hours. The different software were run on genotype matrices containing 300 individuals and from 500 to $50,000$ SNPs. We find that *OutFLANK* is the software for which running time increases the most rapidly with the number of markers and it took around 25 minutes to analyse $50,000$ SNPs (Figure SI7). For the other 3 software (*hapflk*, *pcadapt*, *sNMF*), analyzing $50,000$ SNPs took less than 3 minutes.

Discussion

We have shown that the R package *pcadapt* implements a fast method to perform genome scans with next generation sequencing data. It can handle datasets where population structure is continuous or datasets containing admixed individuals. It can handle missing data as well as data generated with pooled sequencing techniques. We have shown with simulations that it compares favorably to other software of genome scans. In the future, we anticipate that statistical properties of genome-wide approaches should be investigated in settings that mimic more realistically NGS data including for instance pooled sequencing, linkage disequilibrium or a substantial proportion of missing data.

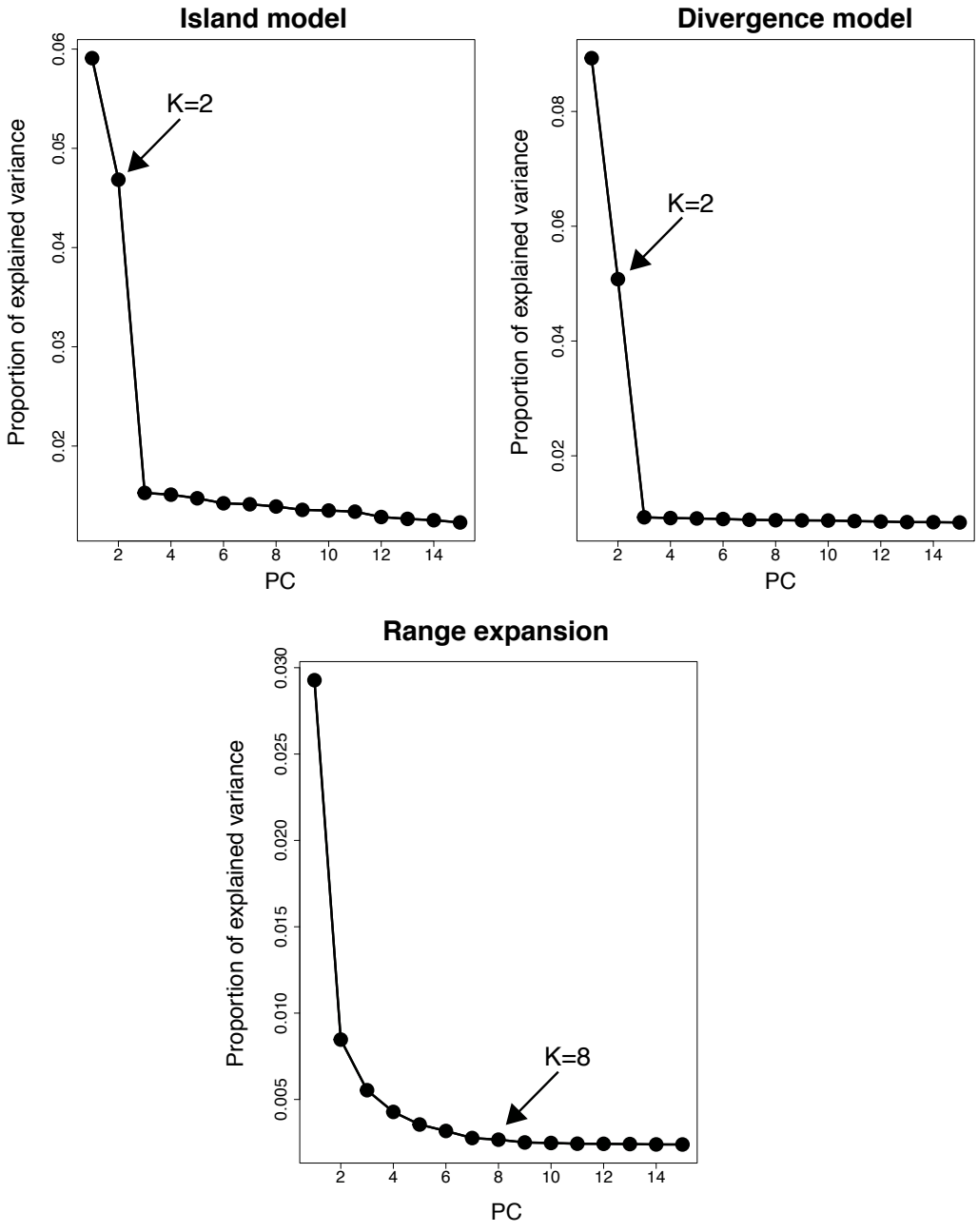

FIGURE 1 – Determining K with the screeplot. To choose K, we recommend to use Cattell's rule that states that components corresponding to eigenvalues to the left of the straight line should be kept. For the island and divergence model, the choice of K is evident. For the model or range expansion, a value of K between 5 and 8 is compatible with Cattell's rule.

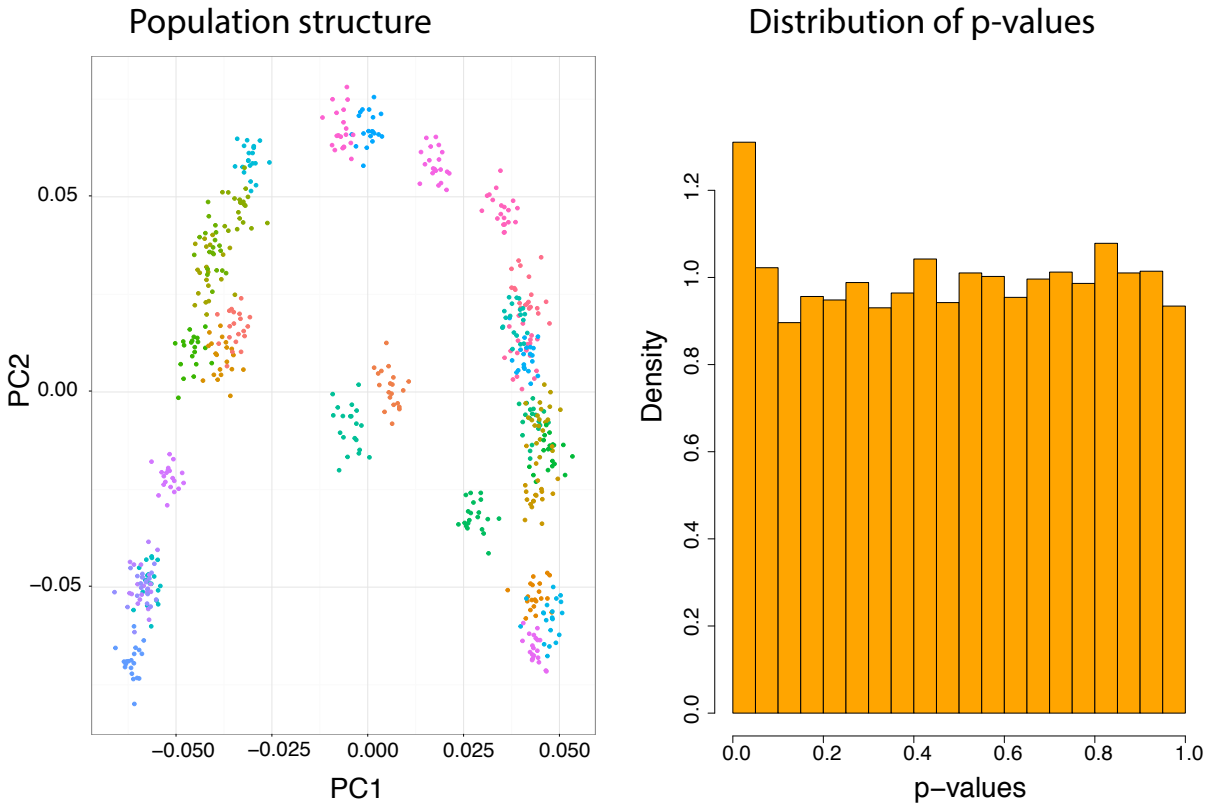

FIGURE 2 – Population structure (first 2 principal components) and distribution of p-value obtained with *pcadapt* for a simulation of range expansion. The p-values are well calibrated since there is a mixture of a uniform distribution and of a peaky distribution around 0, which corresponds to outlier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

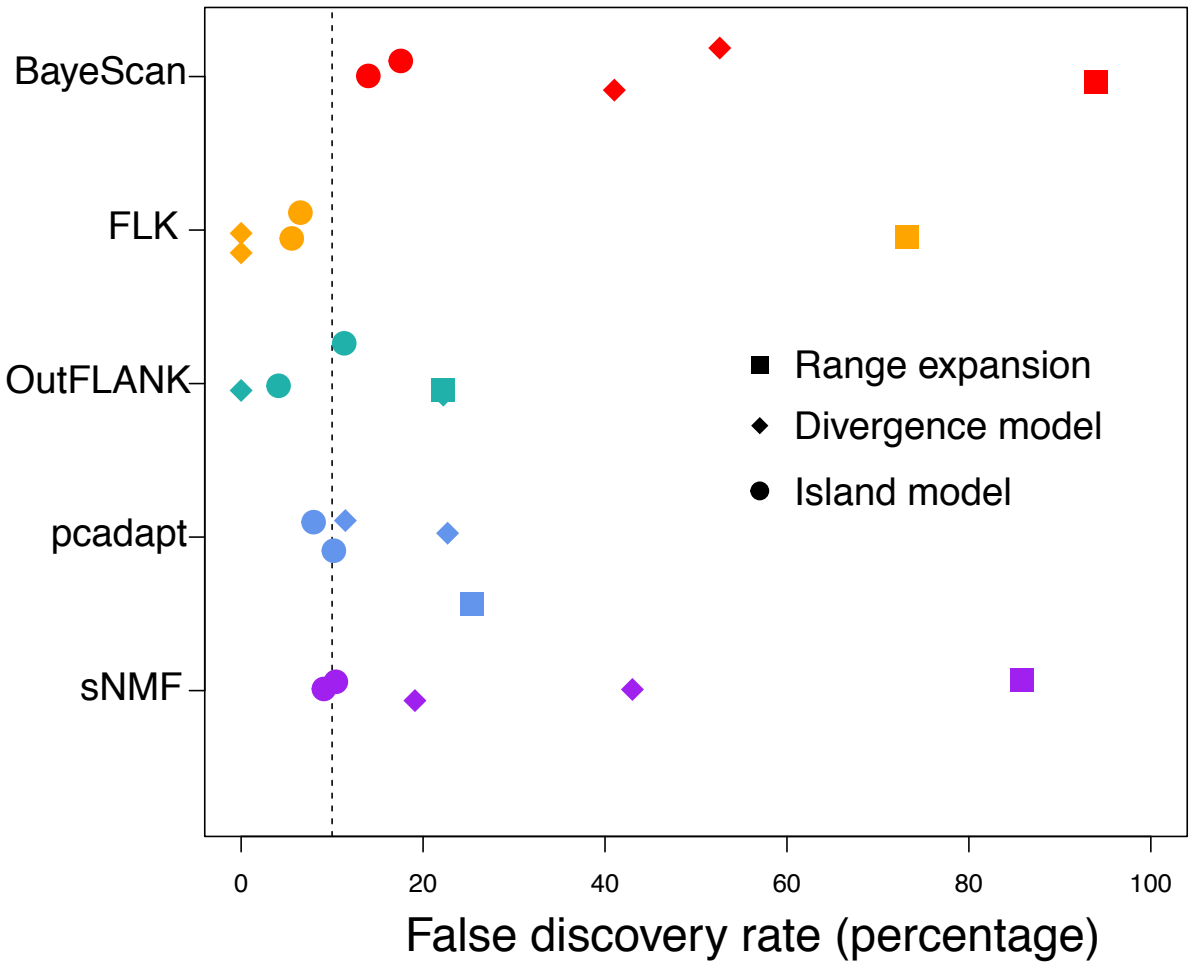

FIGURE 3 – Control of the False Discovery Rate for different software of genome scans. We find that the average proportion of false discoveries is around the nominal false discovery rate set at 10% (6% for *hapflk*, 11% for both *OutFLANK* and *pcadapt*, and 19% for *sNMF*) with the exception of *BayeScan*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

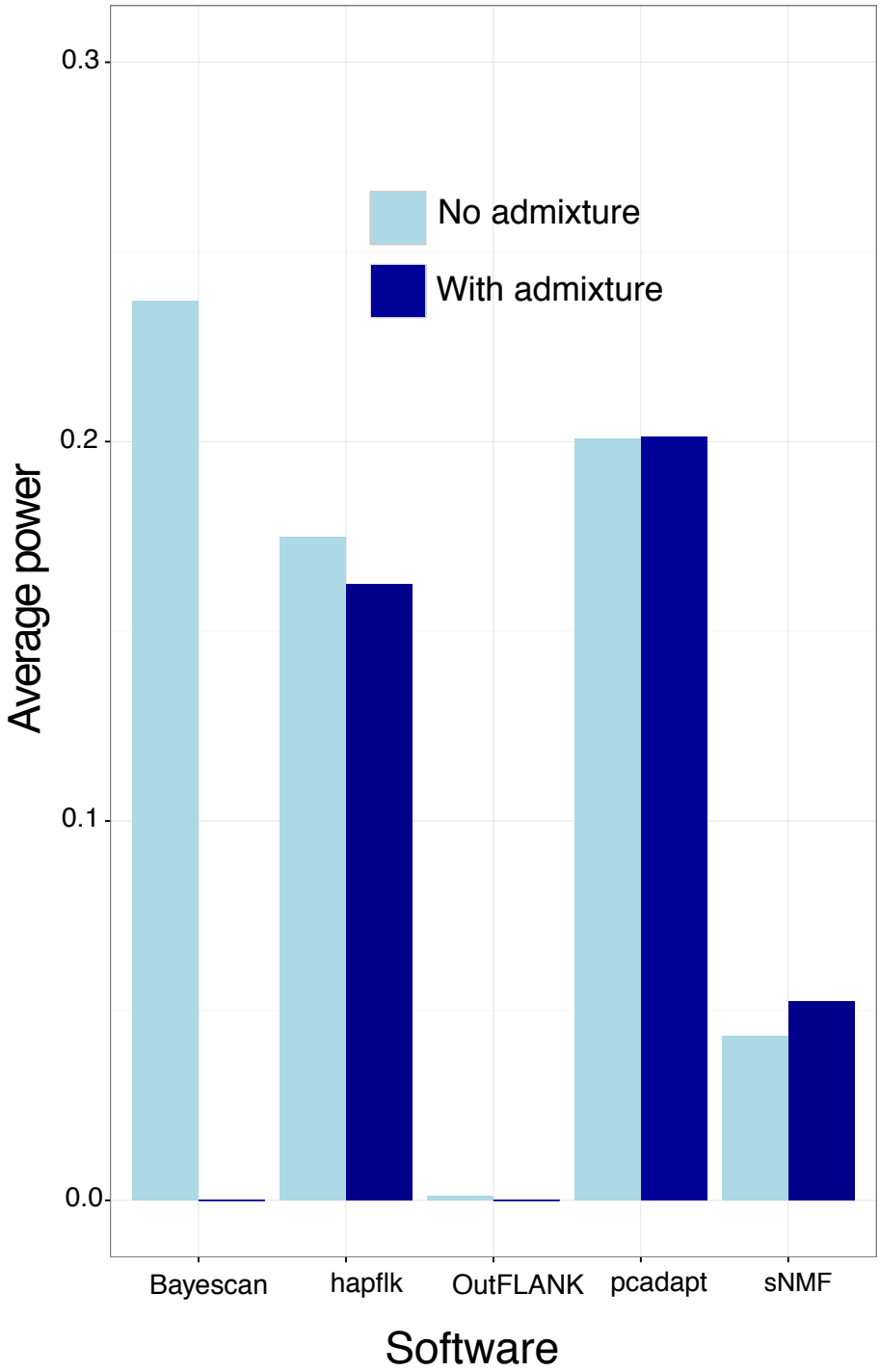

FIGURE 4 – Statistical power averaged over the expected proportion of false discoveries (ranging between 0% and 50%) for the divergence model with 3 populations. We assume that adaptation took place in an external branch that follows the most recent population divergence event.

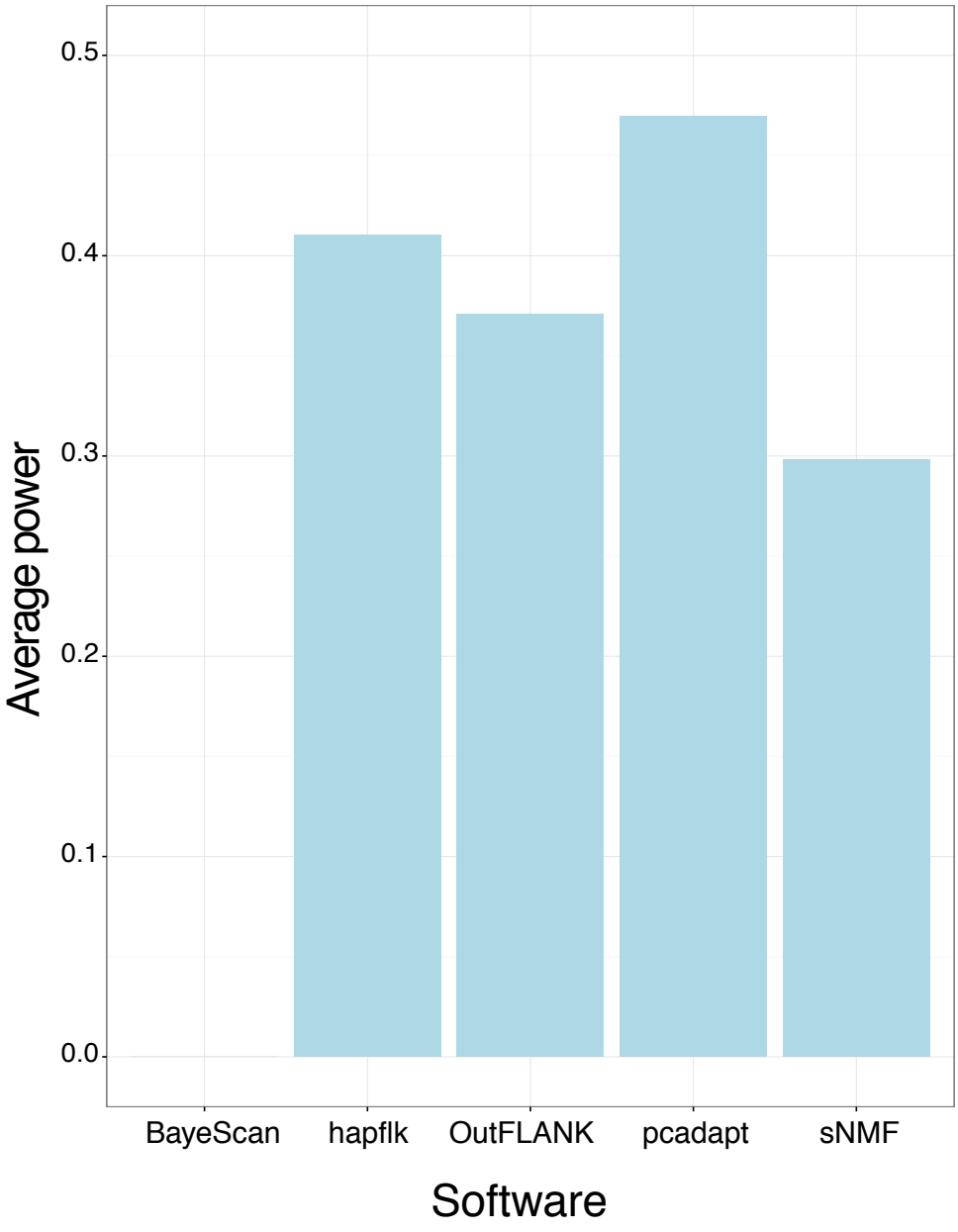

FIGURE 5 – Statistical power averaged over the expected proportion of false discoveries (ranging between 0% and 50%) for a range expansion model with two refugia. Adaptation took place during the recolonization event.

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).

References

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

Bazin, E., K. J. Dawson, and M. A. Beaumont, 2010. Likelihood-free inference of population structure and local adaptation in a bayesian hierarchical model. Genetics 185 :587–602.

Bierne, N., D. Roze, and J. J. Welch, 2013. Pervasive selection or is it... ? why are fst outliers sometimes so frequent ? Molecular ecology 22 :2061–2064.

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

Cattell, R. B., 1966. The scree test for the number of factors. Multivariate behavioral research 1 :245–276.

Devlin, B. and K. Roeder, 1999. Genomic control for association studies. Biometrics 55 :997–1004.

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

Duforet-Frebourg, N., E. Bazin, and M. G. B. Blum, 2014. Genome scans for detecting footprints of local adaptation using a bayesian factor model. Molecular biology and evolution 31 :2483–2495.

Duforet-Frebourg, N., K. Luu, G. Laval, E. Bazin, and M. 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 :1082–1093.

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

Foll, M. and O. Gaggiotti, 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., H. Martins, K. Caye, and S. D. Schoville, 2016. Controlling false discoveries in genome scans for selection. Molecular ecology 25 :454–469.

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

Galinsky, K. J., G. Bhatia, P.-R. Loh, S. Georgiev, S. Mukherjee, N. J. Patterson, and A. L. Price, 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., K. Gharbi, T. Cezard, J. Foucaud, C. Kerdelhué, P. Pudlo, J.-M. Cornuet, and A. Estoup, 2013. The effect of RAD allele dropout on the estimation of genetic variation within and between populations. Molecular Ecology 22 :3165–3178.

Hao, W., M. Song, and J. D. Storey, 2016. Probabilistic models of genetic variation in structured populations applied to global human studies. Bioinformatics 32 :713–721.

Hudson, R. R., 2002. Generating samples under a Wright–Fisher neutral model of genetic variation. Bioinformatics 18 :337–338.

Jackson, D. A., 1993. Stopping rules in principal components analysis : a comparison of heuristical and statistical approaches. Ecology Pp. 2204–2214.

Lange, K., J. C. Papp, J. S. Sinsheimer, and E. M. Sobel, 2014. Next generation statistical genetics : Modeling, penalization, and optimization in high-dimensional data. Annual review of statistics and its application 1 :279.

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

Lotterhos, K. E. and M. C. Whitlock, 2015. The relative power of genome scans to detect local adaptation depends on sampling design and statistical method. Molecular ecology 24 :1031–1046.

Maronna, R. A. and R. H. Zamar, 2012. Robust estimates of location and dispersion for high-dimensional datasets. Technometrics 44 :307–317.

Martins, H., K. Caye, K. Luu, M. G. Blum, and O. Francois, 2016. Identifying outlier loci in admixed and in continuous populations using ancestral population differentiation statistics. bioRxiv P. 054585.

Patterson, N., A. L. Price, and D. Reich, 2006. Population structure and eigenanalysis. PLoS genet 2 :e190.

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

Pritchard, J. K., M. Stephens, and P. Donnelly, 2000. Inference of population structure using multilocus genotype data. Genetics 155 :945–959.

Storey, J. D. and R. Tibshirani, 2003. Statistical significance for genomewide studies. Proceedings of the National Academy of Sciences 100 :9440–9445.

Waples, R. S. and O. Gaggiotti, 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, M. C. and K. E. Lotterhos, 2015. Reliable detection of loci responsible for local adaptation : Inference of a null model through trimming the distribution of fst. The American naturalist 186 :S24–36.

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

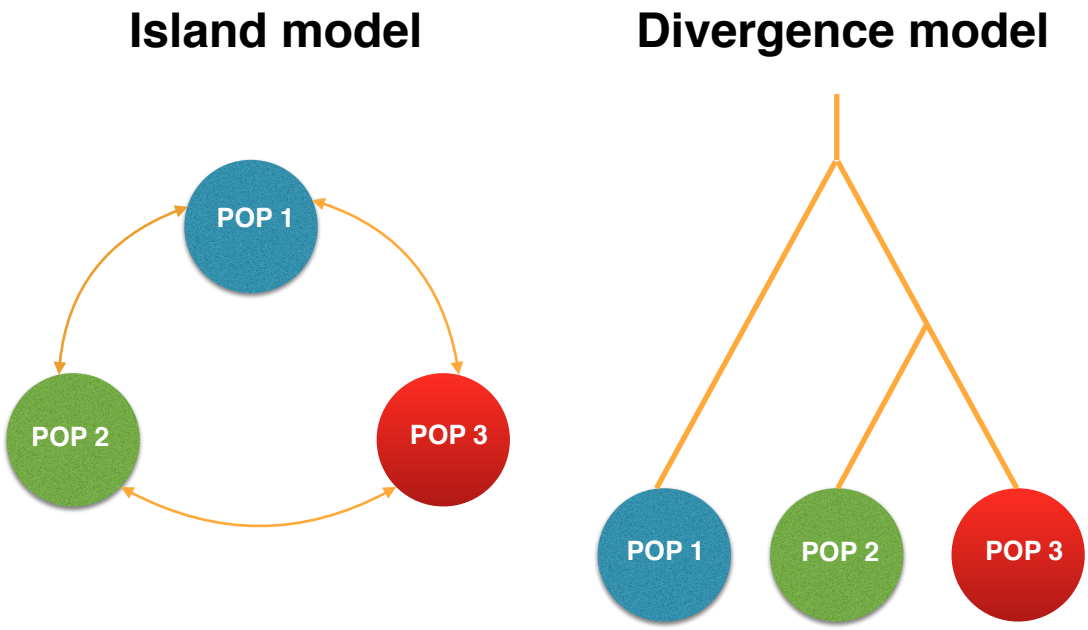

FIGURE SI1 – Schematic description of the island and divergence model. For the island model, adaptation occurs simultaneously in each population. For the island model, adaptation takes place in the branch leading to the second population.

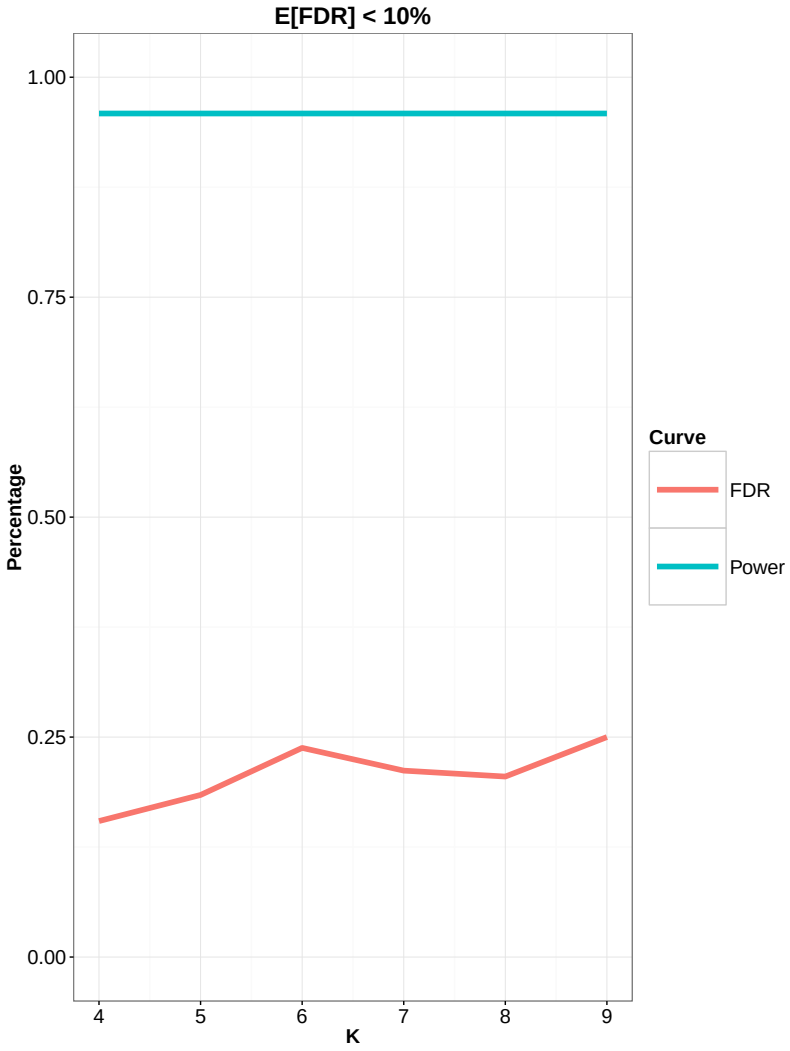

FIGURE SI2 – Proportion of false discoveries and statistical power as a function of the number of principal components in a model of range expansion.

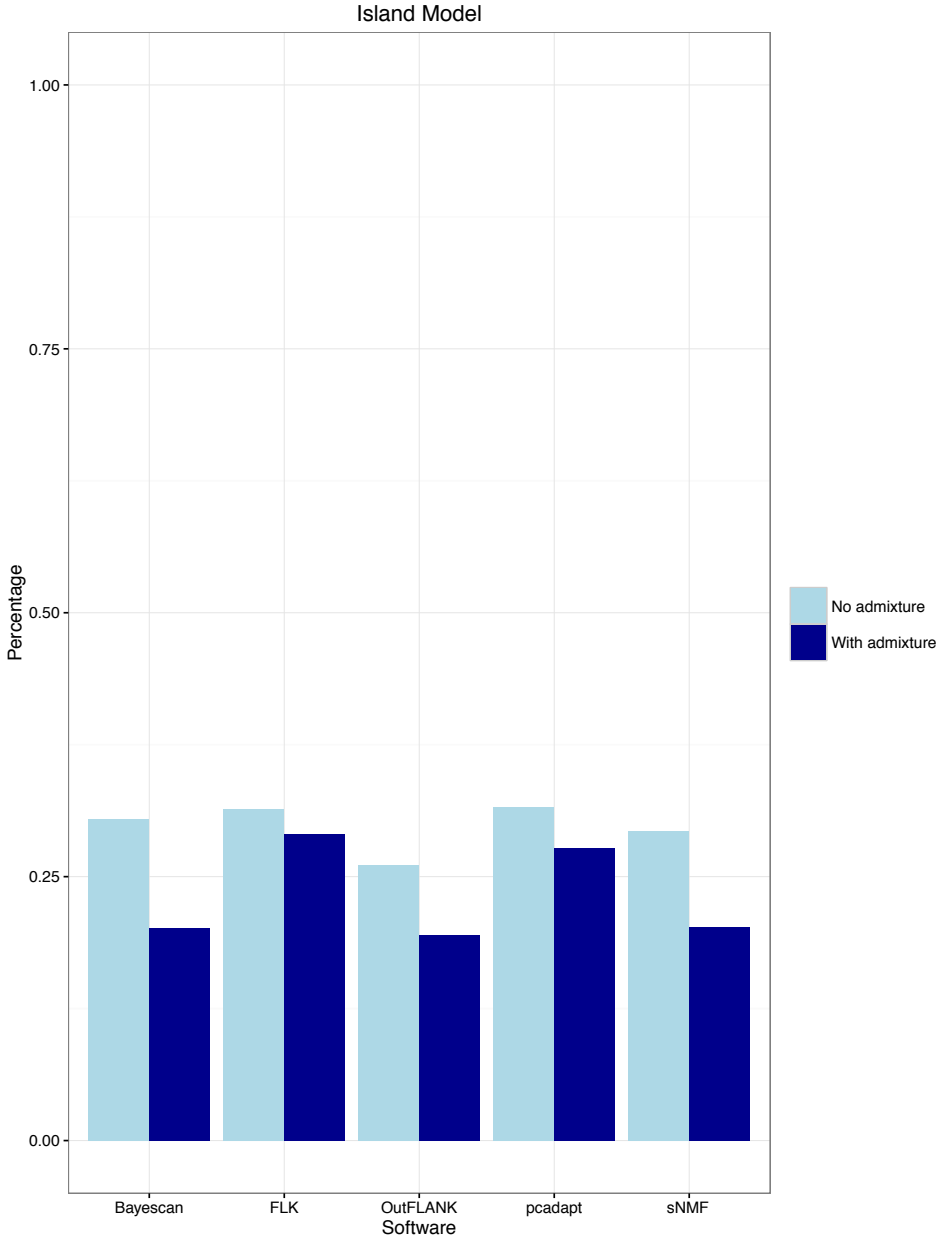

FIGURE SI3 – Statistical power averaged over the expected proportion of false discoveries (ranging between 0% and 50%) for the island model.

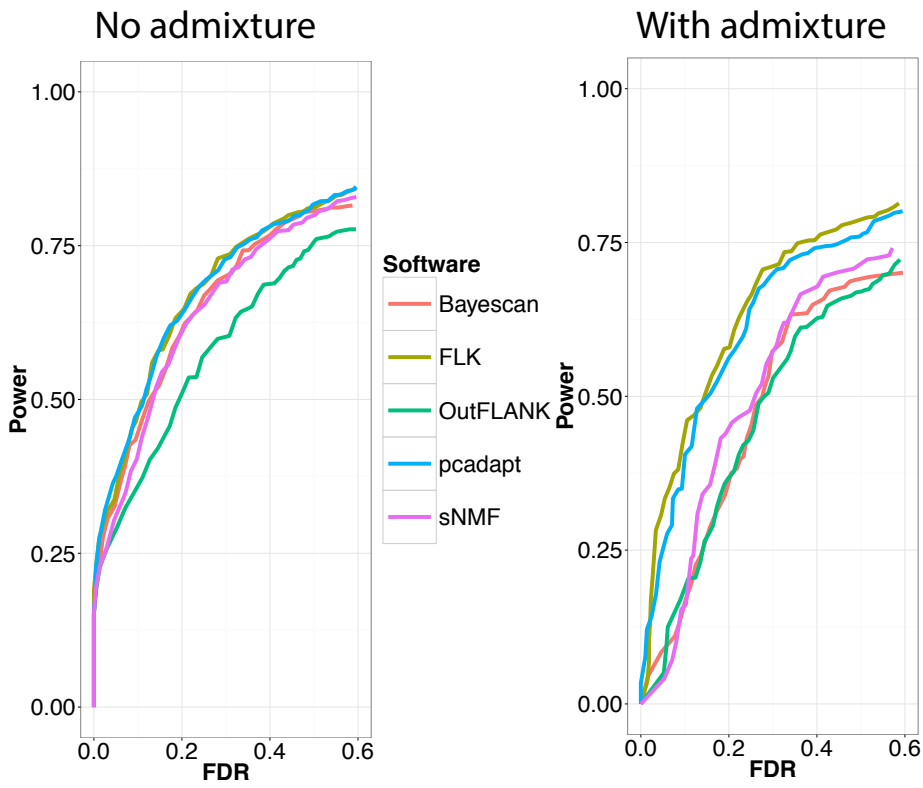

FIGURE SI4 – Statistical power as a function of the proportion of false discoveries for the island model.

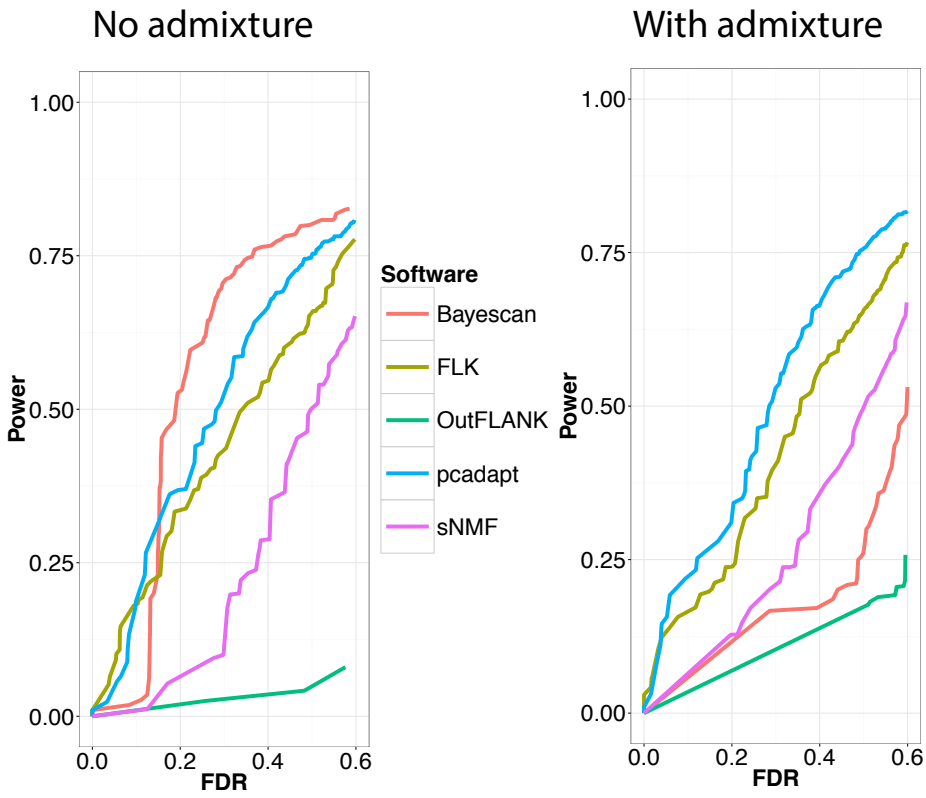

FIGURE SI5 – Statistical power as a function of the proportion of false discoveries for the divergence model.

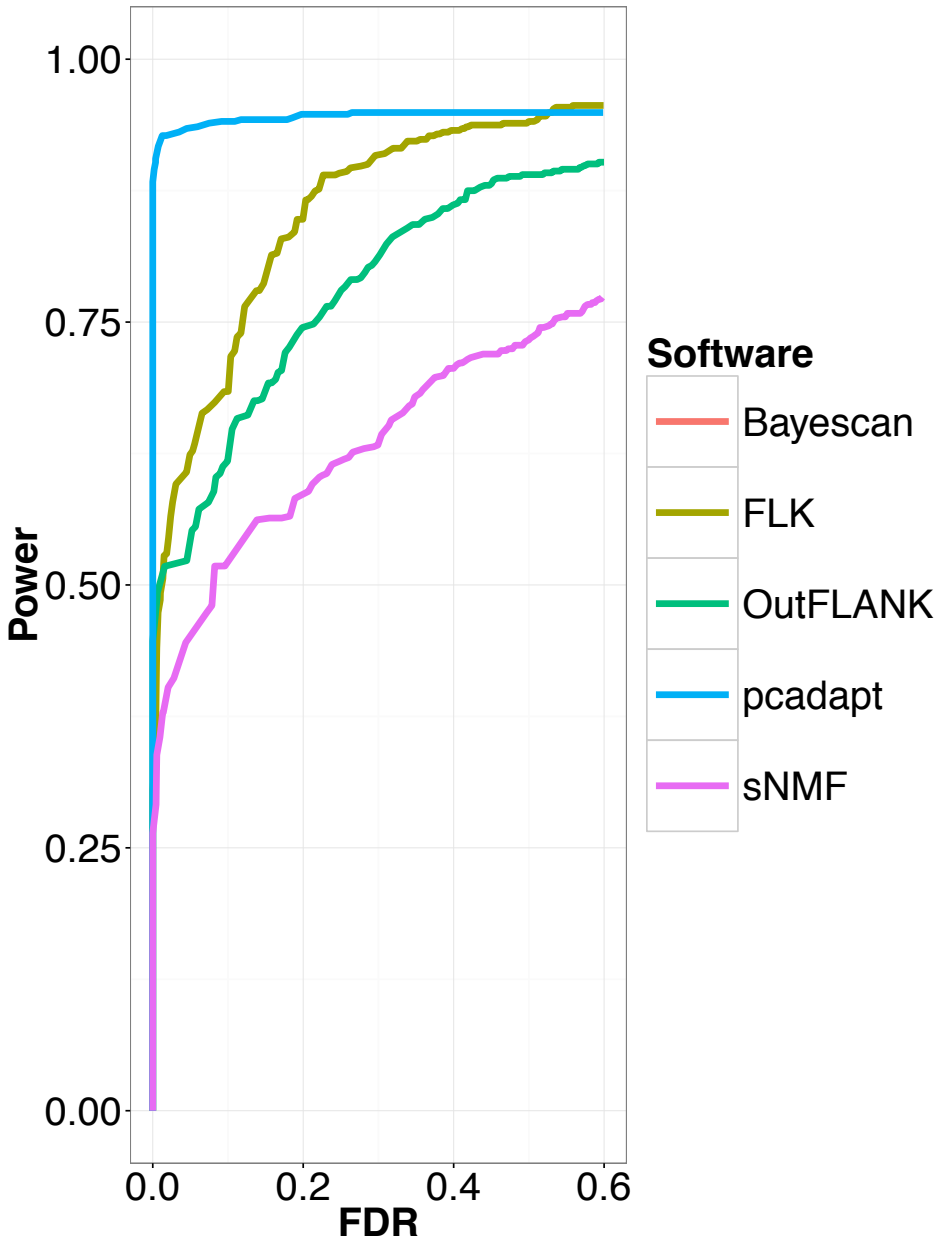

FIGURE SI6 – Statistical power as a function of the proportion of false discoveries for the model of range epxansion.

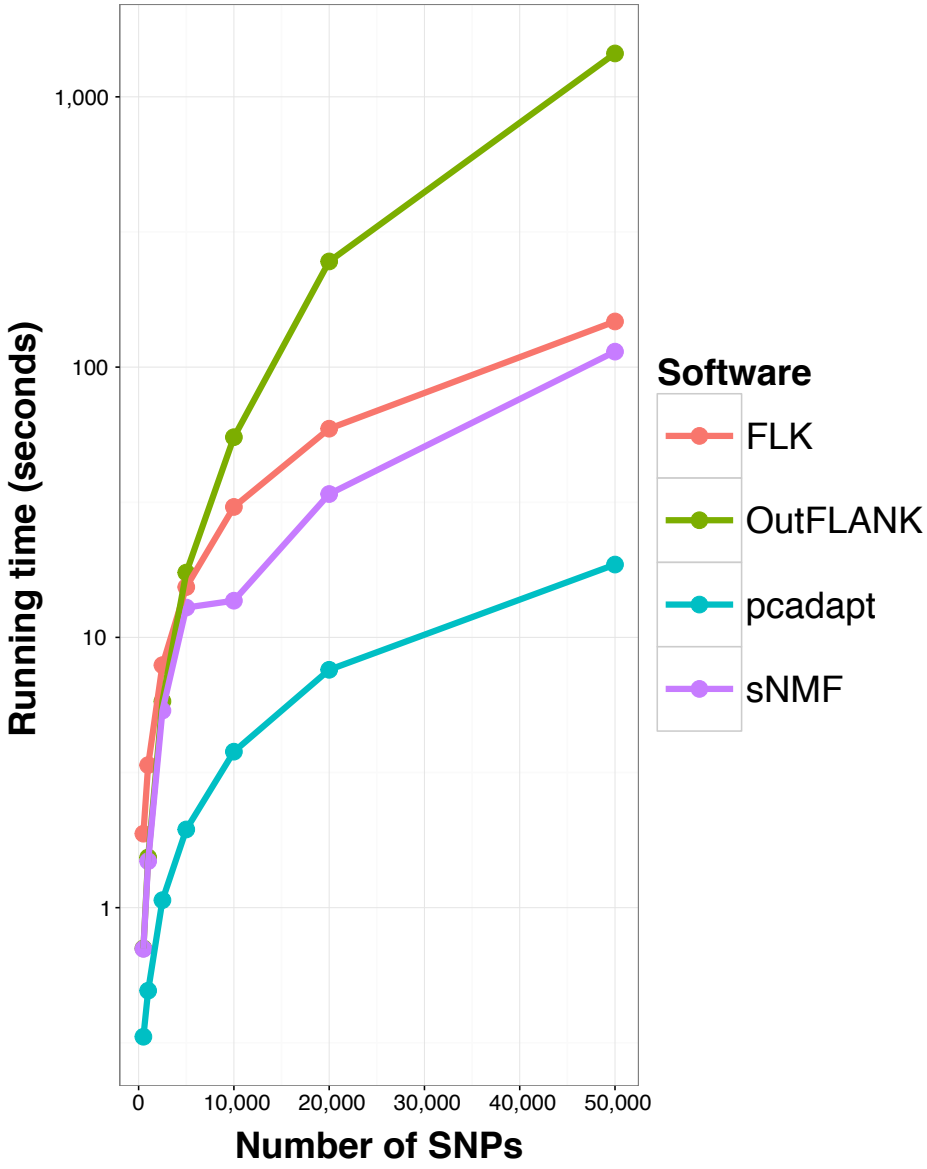

FIGURE SI7 – Running times of the different software. The different software were run on genotype matrices containing 300 individuals and from 500 to 50,000 SNPs. The characteristics of the computer we used to perform comparisons is the following : OSX El Capitan 10.11.3, 2,5 GHz Intel Core i5, 8 Go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