Controlling False Discoveries in Genome Scans for Selection

Olivier François¹ Helena Martins¹ Kevin Caye¹ Sean D. Schoville²

¹Université Grenoble-Alpes, Centre National de la Recherche Scientifique, TIMC-IMAG UMR 5525, Grenoble, 38042, France.
²University of Wisconsin-Madison, Department of Entomology, Madison, WI, U.S.A.

Running Title: FDR control in genome scans for selection

Keywords: Genome Scans for Selection, Control of False Discovery Rates.

Corresponding Author: Olivier François
Université Grenoble-Alpes,
TIMC-IMAG, UMR CNRS 5525,
Grenoble, 38042, France.
+33 4 56 52 00 25 (Phone)
+33 4 56 52 00 55 (Fax)
olivier.francois@imag.fr
Abstract

Population differentiation (PD) and ecological association (EA) tests have recently emerged as prominent statistical methods to investigate signatures of local adaptation using population genomic data. Based on statistical models, these genome-wide testing procedures have attracted considerable attention as tools to identify loci potentially targeted by natural selection. An important issue with PD and EA tests is that incorrect model specification can generate large numbers of false positive associations. Spurious association may indeed arise when shared demographic history, patterns of isolation by distance, cryptic relatedness or genetic background are ignored. Recent works on PD and EA tests have widely focused on improvements of test corrections for those confounding effects. Despite significant algorithmic improvements, there is still a number of open questions on how to check that false discoveries are under control and implement test corrections, or how to combine statistical tests from multiple genome scan methods. This tutorial paper provides a detailed answer to these questions. It clarifies the relationships between traditional methods based on allele frequency differentiation and EA methods, and provides a unified framework for their underlying statistical tests. We demonstrate how techniques developed in the area of genome-wide association studies, such as inflation factors and linear mixed models, benefit genome scan methods, and provide guidelines for good practice while conducting statistical tests in landscape and population genomic applications. Finally, we highlight how the combination of several well-calibrated statistical tests can increase the power to reject neutrality, improving our ability to infer patterns of local adaptation in large population genomic datasets.
1 Introduction

Local adaptation often occurs among species occupying spatially heterogeneous environments, yet we know very little about the conditions in which it occurs or the particular genetic pathways involved, which would provide critical knowledge for how organisms will respond to environmental change (Davis & Shaw 2001, Davis et al. 2005, Jump & Penuelas 2005, Schoville et al. 2012, Savolainen et al. 2013, Fitzpatrick & Keller 2015). Adaptation to local environments triggers modifications in allele frequencies, and enables maintenance of genetic variation within and among populations by spatial and temporal variation in selection intensities (Nei 2005). By screening the genome for differences in allele frequencies among populations, genome scans for selection attempt to identify genomic regions that exhibit signatures of diversifying selection (Storz 2005, Vitti et al. 2013, Tiffin & Ross-Ibarra 2014, Haasl et al. 2015).

Genome scans for divergent selection typically focus on allele frequencies and can be divided into two main approaches: identifying 1) genomic loci that show unusual allele frequency differentiation among populations or 2) those loci with a strong association between allele frequencies and environmental variables (Savolainen et al. 2013). The first group of methods, population differentiation (PD) methods, compares single-locus estimates of a population differentiation statistic with their expectation from a null model of neutral evolution or with the genome-wide background (Beaumont & Nichols 1996, Ackey et al. 2002). If natural selection favors one allele at a particular locus in some populations, the test statistic at that locus will be large compared to loci in which among-population differences are the result of neutral demographic processes. Outliers in the genome-wide distribution of the test statistic correspond to loci potentially targeted by selection. One of the best examples of applying genome-wide scans is the discovery of several genomic regions containing genes involved in high-altitude adaptation in Tibetan populations (Simonson et al. 2010). Tibetans have lived at very high altitudes for thousands of years, and they share physiological traits that enable them to tolerate hypoxia. Contrasting lowland and highland populations, Peng et al. (2010) used genome scans to identify strong signals of selective sweeps in two hypoxia-related genes, EPAS1 and EGLN1, that were significantly associated with the body response to oxygen level in highland populations.
The second group of methods, genome-wide ecological association (EA) methods, estimate correlations between allele frequencies and one or more ecological variables (Hedrick et al. 1976, Joost et al. 2007, Hancock et al. 2008, Rellstab et al. 2015). EA methods rely on the common observation that selection along environmental gradients results in allele frequency clines in spatially distributed populations (Haldane 1948, Berry & Kreitman 1993). Thus EA methods are more likely to detect gradual changes in allele frequencies linked to spatially varying environments than PD methods (Hancock et al. 2010, Schoville et al. 2012). EA methods do not require population samples, but instead can be applied to an individual-based design that draws samples from discrete sites in geographic space. A compelling example of EA tests was provided by Hancock et al. (2010), who conducted a genome-wide scan to identify genetic loci associated with climate in the plant *Arabidopsis thaliana*. They found that non-synonymous variants were significantly enriched among the loci strongly correlated with climate, suggesting that adaptive alleles were effectively detected, and then used their results to predict relative fitness among a set of geographically diverse ecotypes of *A. thaliana*. Parallel to the development of PD methods, recent research efforts have been devoted to improve EA methods (surveyed by De Mita et al. 2013, De Villemereuil et al. 2014, Lotterhos & Whitlock 2015, Frichot et al. 2015), correcting for confounding effects due not only to population structure, but also to often unobserved additional factors (Günther & Coop 2013, Frichot et al. 2013, De Villemereuil & Gaggiotti 2015). Those unobserved factors include uneven sampling designs, genome sequencing biases, relatedness among individuals, gene interactions that affect phenotypic variation, and linkage disequilibrium within genomes, for example.

Despite significant algorithmic improvements, an overlooked question regarding PD and EA methods is how to make decisions about which loci to retain as candidates for further investigations. The same question has been asked for genome-wide association studies (GWAS), and led to several important improvements in the design and implementation of these studies. A general answer to this question is closely linked to the correct adjustment of tests for confounding factors, through the use of stringent corrections to the $p$-values across the many thousands of statistical tests performed in association studies (Balding 2006, Pearson & Manolio 2008, Korte & Farlow...
Displaying the empirical distribution of the significance values is a common way to show that confounding effects are removed from the analysis, and that false discoveries can be controlled (Storey & Tibshirani 2003, McCarthy et al. 2008). Assuming that confounding errors are removed, adjusting for multiple comparisons can then be achieved through the application of false discovery rate (FDR) control algorithms (Benjamini & Hochberg 1995, Storey & Tibshirani 2003). While corrections for population structure are often included in genome scans for selection, the assessment of test calibration and the correct application of FDR control procedures have received less attention than in GWAS.

In this study, we provide a brief overview of the recent literature on PD and EA methods, and we evaluate the capabilities of these methods to provide correct FDR control procedures. We emphasize that a unified framework is available for most hypothesis-testing methods, based on the application of the chi-squared distribution to various test statistics. We review two popular approaches to test calibration: 1) empirical null-hypothesis testing where test corrections are usually done on the basis of inflation factors, and 2) the inclusion of random effects to account for confounding errors in statistical models (Delvin & Roeder 1999, Efron 2004, Yu et al. 2006). We provide evidence that these approaches are useful for controlling false discoveries in genome scans for selection as well as for combining statistical testing methods. We provide tutorial examples showing how the two approaches can be applied in practice, and how they can be implemented using a few command lines in the R programming language (see Figure 1 for a global picture).

Finally, we illustrate the use of these calibration methods by providing a controlled list of selected loci from a 230k single locus polymorphism (SNP) dataset of Scandinavian lines of the model plant Arabidopsis thaliana (Atwell et al. 2010).

2 FDR control algorithms and genome scans for selection

FDR control algorithms. Given a list of genetic polymorphisms where the hypothesis of selective neutrality is rejected, the FDR is defined as the proportion of false discoveries among the positive tests. Controlling the FDR at level $\alpha$, for a given probability value $\alpha$, means that candidate lists are expected to contain less than a proportion $\alpha$ of false positives (Box 2). In genome scans, heuristic methods are often employed to minimize false discoveries, typically in
non-model organisms where there is much less known about the species evolutionary history and population structure, and there are fewer genetic markers. Heuristic approaches rely on identifying the most significant “top hits” in the list of putatively selected loci or comparing the output of multiple methods to identify loci that share significance across these approaches (Storz 2005, De Mita et al. 2013).

While minimizing the number of false positives is often the primary motivation in genome-wide studies, overly conservative tests can miss important associations. Approaches that result in overly conservative tests are potentially problematic because, as they inflate the Type II error rate (false negatives), they may lead to biases in the types of selective events identified and the interpretation of functional responses to selection (Williams & Haines 2011). In particular, selection on standing variation or polygenic traits is expected to result in moderate changes in allele frequencies (soft or partial selective sweeps) that are less likely to have the strong significance values and may be difficult to detect depending on the method employed (Teshima et al. 2006, De Mita et al. 2013). Downstream functional analyses of the significant hits from a pruned list, such as GO term enrichment analyses or KEGG pathway analyses, may then be biased towards a narrow set of genetic pathways (Huang et al. 2009). Though Type I errors are not desirable, they may be more easily identified in follow-up studies that involve different sampling strategies or functional validation. Thus, test calibration should aim to provide the largest lists satisfying a properly calibrated $\alpha$ level, a condition not guaranteed by the usual Bonferroni correction method which is overly conservative. When the $p$-values are calibrated, the FDR can be controlled using algorithms as described by Benjamini & Hochberg (1995), Storey & Tibshirani (2003) or Efron (2004) (Box 2).

A key condition for applying FDR control algorithms is that the test $p$-values behave as uniformly distributed random variables when the null hypothesis is correct. To check this condition, graphical approaches displaying histograms of $p$-values are very useful (Balding 2006). When performing multiple tests to evaluate loci under selection, a majority of loci are expected to be selectively neutral ($H_0$), whereas a minority of loci will deviate from the neutral distribution. In this case, the empirical distribution of $p$-values will consist of a mixture of a uniform distribution
and a peaky distribution showing a peak close to 0. The uniform distribution corresponds to neutral loci whereas the interesting loci are found under the peak (Efron 2004). Assessing whether the shape of the empirical distribution of $p$-values is correct is an essential step in GWAS and gene expression analysis (Balding 2006, Storey & Tibshirani 2003, Dudoit & Van der Laan 2007). This, however, is missing in the literature of genome scans for selection.

**Genome scans for selection.** In scans based on PD methods, the most frequently-used statistic to screen the genome is the fixation index, $F_{ST}$, which can be computed at each locus. The fixation index is directly related to the variance in allele frequency among populations, and to the degree of resemblance among individuals within populations (Holsinger & Weir 2009). According to standard population genetic theory, the definition of $F_{ST}$, which is based on Wright’s work, corresponds to the amount of variance in allele frequency that can be explained by population structure (Wright 1951). Thus, estimates of $F_{ST}$ correspond to estimates of correlation coefficients, and their computation can be based on an analysis of variance (ANOVA) of allele frequencies (Weir & Cockerham 1984, Weir 1996). A drawback of a direct application of ANOVA approaches to detecting selection is that they are likely to generate large numbers of false positive tests. To overcome this problem, Lewontin & Krakauer (1973) pioneered the development of the statistical theory of PD tests for the selective neutrality of polymorphisms by introducing the chi-squared distribution to evaluate the statistical significance of their test (Table 1). Beaumont & Nichols (1996) extended the Lewontin & Krakauer approach, and proposed to detect selected loci by using the distribution of neutral $F_{ST}$ conditioned on the expected heterozygosity, while assuming an island model of migration-drift equilibrium (program FDIST2). Similarly, the program DETSEL implemented a method of detecting selection that relies on a model of population divergence by pure random drift (Vitalis et al. 2001, Vitalis et al. 2003).

Other innovations to the Lewontin & Krakauer approach includes $F_{LK}$, a test which compares patterns of differences in allele frequencies among populations to the values expected under a scenario of neutral evolution (Bonhomme et al. 2010). To test selective neutrality, $F_{LK}$ reconstructs a topology modeling population divergence wherein the branch lengths correspond to the amount of genetic drift in each population. More recently, the program BAYENV2 considered a
new statistic, $X^T X$, which evaluates departure from selective neutrality by incorporating predictions from a population genetic model (Günther & Coop 2013). The model of BAYENV2 was improved and re-implemented in the program BAYPASS (Gautier 2015). In contrast to these model-based approaches, some model-free methods rely on the data to correct the effects of confounding factors empirically. The recently proposed OUTFLANK test used an empirical approach based on a trimming procedure to evaluate the test $p$-values (Withlock & Lotterhos 2015). An individual-based empirical method was implemented in TESS3 to perform tests for selection in continuous populations, based on ancestral allele frequency differentiation and spatially varying ancestry coefficients (Caye et al. 2015). Similarly, a fast version of PCADAPT implemented an empirical test for selection based on the eigenvalues of a principal component analysis and the *communality* statistic ($h^2$, Duforet-Frebourg et al. 2015).

The development of EA methods has been much more recent than the development of PD methods (Savolainen et al. 2013). Many studies have proposed that EA methods improve the ability to detect adaptation from standing variation (Pritchard et al. 2010). Most EA methods use hypothesis testing approaches to identify strong correlations between allele frequencies and an environmental variable (Table 1). Using a standard regression framework, EA methods amount to test the null hypothesis $H_0 : R^2 = 0$ against the alternative hypothesis $H_1 : R^2 > 0$, where $R^2$ is the proportion of the allele frequency variation explained by the environmental variable computed at each locus. A variety of EA statistical models have been developed (Rellstab et al. 2015), including generalized linear models (SAM, Joost et al. 2007), generalized linear mixed models (GLMM, Jones et al. 2013), or latent factor mixed models (LFMM, Frichot et al. 2013). At the exception of the SAM approach, a general feature of EA methods is to use information contained in the genotypic dataset to evaluate confounding effects, and eventually achieve test calibration empirically.

### 3 A unified testing framework for genome scans for selection

**Chi-squared distributions.** PD and EA methods have often been considered as two distinct approaches to genome scans for selection (Vitti et al. 2013, Savolainen et al. 2013). This section
argues that a common statistical framework is underlying PD and EA testing methods, and that this framework is useful in applying corrections for confounding errors, and in solving multiple test issues. The common statistical framework for PD and EA testing approaches is based on the use of the chi-squared distribution for rejecting the null-hypothesis of selective neutrality at a given locus (Table 1). The connection between PD and EA methods arises because PD methods can be viewed as evaluating the association between allele frequencies and categorical variables encoding population labels (factors) that represent the uncharacterized environment of each population.

However, it is important to note that the test statistic used to reject the null-hypothesis and the degrees of freedom differ in each method (Table 1). In PD scans, the degree of freedom is often estimated by the number of population samples (Lewontin & Krakauer 1973, Bonhomme et al. 2010, Günther & Coop 2013, Gautier 2015). Withlock & Lotterhos (2015) proposed estimates based on a maximum-likelihood principle, leading to degrees of freedom less than the actual number of populations, so that their approach to $F_{ST}$ tests accounts for the shared demographic history of the samples. Based on individual ancestry estimation methods, Duforet-Frebourg et al. (2015) and Caye et al. (2015) estimate the degree of freedom of their tests by the number of principal components or by the number of genetic clusters inferred from the data. The case of EA tests is simpler to describe as the degrees of freedom correspond to the number of environmental predictors (Frichot et al. 2015).

The ubiquity of the chi-squared distribution enables a unified treatment of test calibration and FDR control in genome scans for selection, which can be achieved by applying techniques developed for the analysis of GWAS and genome-wide gene expression analysis. Suppose that allele frequencies at a particular SNP are significantly associated with some disease. This may occur when a SNP is associated with a confounding factor, which correlates with the GWAS phenotype but is not in the same causal pathway. Just as in genome scans for selection, confounding variables include genetic ancestry, genotyping error, ascertainment bias, and epistatic effects (Vilhjálmssson & Nordborg 2013). Correcting the association tests for confounding effects is crucial to the control of false discovery rates, and this is usually based on the chi-squared distribution. A first correction strategy consists of modifying the null-hypothesis to match the levels of neutral genetic background
variation observed in the dataset. This technique is sometimes called genomic control in GWAS, and empirical null-hypothesis testing in studies of differential gene expression (Delvin & Roeder 1999, Efron 2004). A second strategy consists of modifying the regression equation in order to model the effect of various confounding factors directly (Price et al. 2006, Yu et al. 2006). Those two strategies are detailed in the next paragraphs.

**Test correction and inflation factors.** Genomic control relies on the introduction of inflation factors (Box 3). Inflation factors are constant values, λ, that are used to rescale the test statistic in order to limit inflation due to population structure and confounding factors. The goal of the rescaling procedure is to define a modified test statistic leading to a flat histogram for the significance values. Since the approaches listed in Table 1 rely on the chi-squared distribution, their test statistics will be designated as squared z-scores in the rest of our study. For chi-squared tests, the rescaled statistic is $z^2_\ell /\lambda$ where $z_\ell$ is the score computed at locus $\ell$, and the degree of freedom of the test is left unchanged. This technique has been called an empirical null-hypothesis testing approach by statisticians because it modifies the base-line null-hypothesis, $H_0 : z^2_\ell = 1$, and replaces it by a new null-hypothesis, $H_0 : z^2_\ell = \lambda$, in which $\lambda$ is estimated from the data.

For PD tests, empirical null-hypothesis tests correspond to testing the null-hypothesis $H_0 : F_{ST} = \theta$, where $\theta$ is the level of population differentiation expected at selectively neutral SNPs. For EA tests, the modified null-hypothesis corresponds to $H_0 : R^2 = \theta$, where $\theta$ is the proportion of the genetic variation explained by the environmental variable at selectively neutral SNPs (Wang et al. 2013). Using the correspondence between z-scores and correlation coefficients, one can show that the tested value, $\theta$, is linked to the inflation factor by a simple relationship

$$\lambda = (n - d) \frac{\theta}{1 - \theta},$$

or equivalently,

$$\theta = \frac{\lambda}{n - d + \lambda}.$$

Thus, estimates of $\lambda$ provide estimates of the level of population differentiation or the proportion of the genetic variation explained by the environmental variable expected at selectively
neutral SNPs. In the above equations, $d$ represents the number of populations in an ANOVA test, and $d - 1$ is equal to the number of environmental predictors in an EA test ($n$ is the number of individuals in the sample). For example, when the inflation factor is equal to $\lambda = 5$, then $F_{ST}$ can be estimated to be around 2.4% at a selectively neutral SNP in a two-population model where we have $n = 100$ individuals in each sample ($d = 2$). Following GWAS approaches, an estimate of $\lambda$ is commonly obtained after computing the genomic inflation factor, defined by the median of the squared $z$-scores divided by the median of a chi-squared distribution with $d - 1$ degrees of freedom (Devlin & Roeder 1999).

**Linear mixed models.** A second approach to adjusting for confounding factors consists of modifying the regression model while keeping the null-hypothesis unchanged. In principle, the modification introduces additional factors that represent the sources of error due to each confounding effect. For example, Price *et al.* (2006) have suggested the inclusion of principal components of neutral variation as fixed effects to correct for population stratification in GWAS. Yu *et al.* (2006) and Kang *et al.* (2010) considered mixed models in which random effects account for relatedness among individuals (program EMMAX). Those mixed models also perform well in presence of genetic linkage and epistasis (Platt *et al.* 2010, Vilhjálmsdóttir & Nordborg 2013). GWAS mixed models specify a similarity matrix for the covariance structure of random effects which is commonly based on kinship coefficients.

In general, corrections using fixed or random effects have proven useful to GWAS in which the proportion of SNPs associated with a particular trait is expected to be very small (Price *et al.* 2010), but they may be inappropriate when associations with ecological biotic or abiotic factors are investigated. For example, estimating principal components or a covariance matrix in a genome scan for selection would require a set of SNPs that are assumed to be truly neutral. This set of control SNPs can be difficult to define in EA analyses. Latent factor mixed models (LFMM) do not require any set of control SNPs, and they attempt to remove the effect of relatedness and genetic linkage when inferring ecological associations by using latent factors (Frichot *et al.* 2013). The principle is that $K$ random factors are included in the regression model. The number of factors, $K$, can be estimated by applying principal component analysis or ancestry estimation algorithms.
to the genotypic data. The factors are estimated from the full dataset at the same time as the
regression coefficients are computed. In addition, LFMM can be used as a PD test when the
environmental variable is defined as a categorical variable representing population membership for
each individual.

4 Tutorial examples and data analysis

In this section, we summarize our best practices for conducting statistical analysis of genome
scan tests and demonstrate these recommendations in supporting examples. First, scientists will
choose one or several testing procedures among those presented in Table 1. The application of a
particular testing procedure to the data produces test statistics corresponding to squared $z$-scores
for each locus. Equivalently, some computer programs result in significance values instead of
squared $z$-scores. For those programs, the significance values can be transformed into squared $z$-
scores by using the quantile function of the chi-squared distribution (see Supplementary Materials
for R code). The degrees of freedom of the chi-squared distribution are indicated in Table 1, fifth
column.

For each procedure, one then evaluates inflation factors from the test statistics. This is done by
computing the genomic inflation factor (Box 2) or by using more sophisticated approaches such as
Efron’s local FDR method. For appropriately calibrated methods, the inflation factor is expected
to be close to one, and the histogram of test significance values is expected to display a flat
shape (Box 2). Several procedures from Table 1 already include corrections for confounding errors
generated by population structure or other unobserved factors, and should be correctly calibrated.
If the inflation factor is not close to 1 and the significance values are not evenly distributed (i.e.
flat), then we can recalculate the test $p$-values by applying the formula given in Box 3. It should
be noted that assessing inflation factors is always subjective, as the proximity to the value of one
will vary, usually by increasing with the sample size and the background levels of $F_{ST}$ or $R^2$. To
be justified in recalibrating the significance values, it is advisable to report inflation factors prior
to adjustment and histograms following adjustment.

The final step of analysis could consist of combining well-calibrated significance values resulting
from distinct tests (example 1 below), or from distinct runs of a particular method (example 2
and 3 below). In both cases, our meta-analysis approach combines $z$-scores by using a robust variant of the Stouffer method (Box 3). In summary, corrections are applied at two stages, first to obtain well-calibrated significance values for each test, and then to account for correlation among tests resulting from distinct methods or program runs. A step-by-step description of the statistical analyses of example 1 and 2 and their corresponding R commands are provided as supplementary files (Supp. Mat. 1 and Supp. Mat. 2). Example 3 is computationally more intensive, and would take a few hours to reproduce. A short simulation study of the power of tests in described in Supp. Mat. 3.

**Combining statistical models (example 1): Two tests are better than one.** We considered a two-population model in which the populations evolved under migration-drift equilibrium (Wright’s 2-island model). We used the computer program ms to perform coalescent simulations of neutral and selected SNP loci (Hudson 2002). One hundred diploid organisms were genotyped in each population. The justification for the use of Wright’s models to simulate selection is that there is an overall migration rate of individuals, which, in principle, could be estimated using neutral markers, and an effective migration rate that reflects action of selection to filter out a fraction of migrants having maladapted genotypes (Bazin *et al.* 2010). In this simulation model, effective migration rates can vary across loci. Individual genotypes consisted of 1,000 unlinked SNPs, and the proportion of loci under selection was set to 10%. To create the data, we used an overall migration rate of $4Nm = 20$ at 900 truly neutral loci and $4Nm_s = 0.1$ at 100 truly adaptive loci. We performed statistical tests using two methods: the computer program FDIST2 and $F_{ST}$ computed from an ANOVA approach. For FDIST2, a target $F_{ST}$ value of 5% was used for the computation of significance values.

The FDIST tests were conservative whereas the ANOVA tests were liberal (Supp. Mat. 1). For a level of FDR of 10 percent, the Benjamini-Hochberg algorithm led to observed FDRs equal to 0.0 (FDIST2) and 0.70 (ANOVA). The candidate loci obtained from the FDIST2 tests were included in the ANOVA list, providing little additional insight into potentially selected loci, and the Venn diagram was highly unbalanced (Figure 2). We then applied corrections to the poorly calibrated tests in order to change the significance threshold of the null-hypothesis (Box 3). First,
we used genomic control to correct the ANOVA tests, and the histogram of significance values had
the desired shape (Supp. Mat. 1). After correction, the false positives were nearly all removed
from the list of ANOVA discoveries (observed FDR of 5.6%), and the power became superior to
FDIST2 (Figure 2). We then applied corrections to FDIST2 to recalibrate significance values.
After transforming the \( p \)-values into squared \( z \)-scores, we estimated an inflation factor equal to
\( \lambda = 0.4 \) (Supp. Mat. 1). Recalibration clearly increased the power of FDIST2 tests, and the Venn
diagram became balanced (Figure 2). This result was achieved at the cost of an increased level of
FDR (0.015), but this FDR remains below our nominal expected level of 10%.

While there is overlap in the two tests (Figure 2), each uniquely detects significant loci and
their combined results provide a longer list of candidate loci. By using a meta-analysis procedure
that combined the \( z \)-scores from both tests, our testing approach had power equal to 0.73 and the
observed FDR remained close to the expected level of 10 percent. In summary, this example shows
that: 1) statistical tests can only be compared when the null-hypothesis is correctly specified,
which could be achieved using simple calibration procedures, 2) the well-calibrated model-free
approach (ANOVA) outperformed the model-based approach (FDIST2), even though the data
were simulated under assumptions of the FDIST2 model, and 3) combining well-calibrated tests
increased to power to reject neutrality.

Adaptation to climate in European lines of \textit{A. thaliana} (example 2). We examined the
utility of the mixed model approach using EA methods in a study of adaptation to climate in 170
European inbred lines of the model plant species \textit{Arabidopsis thaliana} (Atwell \textit{et al.} 2010). \textit{A.}
\textit{thaliana} lines were genotyped using a SNP-chip containing 230k SNPs. In our example, the analysis
of genetic variation was restricted to the first chromosome with density of one SNP per 1,000 bp
(26k SNPs). For each of the 170 European lines in the dataset, eighteen bioclimatic variables were
extracted at 30 arcsecond (1km^2) resolution from the WorldClim database (Hijmans \textit{et al.} 2005).
To test for associations between loci and climate, we summarized the bioclimatic data with the
first principal component of the 18 variables.

Our genome scan for selection used 5 distinct statistical models fitted using the LFMM program
as implemented in the \texttt{R} package LEA (Frichot & François 2015). A prior analysis of population
genetic structure based on the ancestry estimation program sNMF (Frichot et al. 2014) suggested
that $K = 6$ clusters (and 5 principal components) could explain the observed genetic variation in
A. thaliana. In subsequent analysis of the data with LFMM, $K = 6$ latent factors were used to
perform genome scans for selection. The 5 models corresponded to 5 distinct runs of the program
with distinct initial values. Each model corresponded to a distinct local optimum of the likelihood
function, and was characterized by a specific set of $z$-scores (Supp. Mat. 2).

A meta-analysis of the 5 LFMM models was applied to the $z$-score matrix. The $z$-scores at each
locus were combined according to a robust variant of the Stouffer method (Brown 1975, Whitlock
2005). In this robust approach, the mean value of the $z$-scores was replaced by their median value
at each locus. The bias created by combining 5 distinct $z$-scores was corrected by the introduction
of an inflation factor ($\lambda = .78$, Supp. Mat. 2), and significance values were computed from the
chi-squared distribution with one degree of freedom. The histogram of $p$-values provided evidence
that the confounding effects were removed (Figure 3). A short list of candidate loci was proposed
on the basis of the Benjamini-Hochberg algorithm, applied at a level of FDR equal to 7%. The 3
top hits in the resulting list of SNPs identified a protein coding region corresponding to the gene
ARR3 (locus AT1G59940), a type A response regulator gene involved in circadian rhythm, and a
gene involved in the cytokinin-activated signaling pathway (Figure 3).

Local adaptation in Scandinavian lines of A. thaliana (example 3). Test calibration and
FDR control algorithms were also applied to address multiple testing in a genome scan for selection
in 52 Scandinavian lines of the model plant species Arabidopsis thaliana (Atwell et al. 2010, Huber
et al. 2014). We analyzed the 5 chromosomes with a density of one SNP per 500 bp. The 52
lines were grouped in 2 genetic clusters showing very low levels of shared ancestry. Each cluster
was restricted to a narrow geographic range, and corresponded to a geographic region to the
north or the south of Scandinavia (Huber et al. 2014). The Northern Sweden cluster included 14
individuals, and the Southern Sweden cluster included 38 individuals. Genome scans for selection
were performed with ANOVA and with LFMM. Here, LFMMs were used to run a PD test, by
considering population membership as a binary covariate taking the value 0 in the southern sample
and 1 in the northern sample. LFMM runs were based on two latent factors ($K = 2$), and resulted
in a correctly calibrated test (Figure S1).

For an expected FDR equal to 0.1 percent, the Benjamini-Hochberg algorithm resulted in a list of 167 chromosome positions, some of which span the same region due to genetic linkage. The estimate of the inflation factor from the ANOVA test was $\lambda \approx 4.5$, which provided an estimate of the neutral differentiation statistic, $F_{ST} \approx 8\%$. A simulation study comparing ANOVA and LFMM showed that LFMM had power similar to ANOVA tests when the ANOVA tests are optimally calibrated (Supp. Mat. 3). We then analyzed the top 50 hits from the LFMM runs (Table 2). Several genomic positions were reported in Huber et al. (2014, Table 4) as corresponding to regions undergoing selective sweeps. The top list contained polymorphisms in genes involved in photosynthesis, response to heat, response to UV, response to freezing, and response to light stimulus (AT1G03600, AT2G43130, AT5G07990, AT5G27540, AT5G27630). LFMM also detected mutations in genes involved in root, flower, meristem, and xylem growth or development (AT1G04240, AT1G04390, AT3G62160, AT5G07290). Several genes were involved in defense response, response to biotic stimulus, nematode resistance, or bacterial immunity (AT5G07390, AT5G11250, AT5G07220, AT3G09980, AT3G12100).

5 Discussion

**Summary of main points.** PD and EA methods are widely-used to detect signatures of natural selection from population genomic data. Our discussion of statistical tests focused on hypothesis testing methods, and compared approaches based on chi-squared tests. The methods studied in our examples ranged from the simplest to the most elaborate one among a longer list of methods available (Table 1). Thus the observations reported in our study are representative and applicable to a large category of statistical approaches. The methods rely on a two-stage procedure (Box 4). At the first stage, statistical tests are performed, and the tests return significance values for each locus. This first stage may include a combination of several methods. At the second stage, decisions are made about which loci to retain in a list of candidates potentially under selection. In this study, we argued that the decisions made at the second stage may be incorrect when the histogram of test significance values is not flat under $H_0$, and test recalibration at the first stage is often necessary.
The two-stage process described above is closely related to the methods employed in genome-wide association studies, for which a well-developed literature provides resources to improve FDR control in genome scans for selection (Delvin & Roeder 1999, Storey & Tibshirani 2003, Price et al. 2006, Dudoit & Van der Laan 2007). In the GWAS literature, statistical test calibration is the key to control the type I error or the FDR. In our study, two main calibration approaches were applied to genome scans for selection. The first calibration method was based on a technique called empirical null-hypothesis testing, and could be implemented by estimating inflation factors. The second calibration method adjusted for confounding errors using mixed models, and it was illustrated by the introduction of latent factors in regression models. In a simulation study, we observed that latent factor models outperformed tests including corrections on ANOVA or GLM for PD and EA methods (Supp. Mat. 3). When test calibration is applied, genome scan tests become complementary and can be combined to increase the power to reject neutrality.

**Extension to Bayesian methods.** Bayesian approaches to detecting selection have historically been considered as alternatives to hypothesis testing methods (Beaumont & Balding 2004, Foll & Gaggiotti 2008, Bazin et al. 2010, Coop et al. 2010, Günther & Coop 2013, Gautier 2015, De Villemereuil & Gaggiotti 2015). While hypothesis testing methods assess the null hypothesis of selective neutrality using significance values, Bayesian methods evaluate the probability of the null and alternative hypotheses given the data. Bayesian approaches can take advantage of population genetic model predictions, but these methods are not always robust to departure from their underlying model assumptions (Hermisson 2009, Narum & Hess 2011, Lotterhos & Whitlock 2014). Recalibration methods cannot be directly applied to the variety of Bayesian models available, unless these methods propose to compute significance values (Beaumont & Balding 2004, Günther & Coop 2013, Gautier 2015). When recalibration is not possible, a correct application of Bayesian methods requires that the null or the alternative model fit the data. In standard approaches, model fit is usually checked by using posterior predictive tests. While model-checking is commonly addressed in applications such as approximate Bayesian computation (Csilléry et al. 2010), it can be difficult to address for genome scan algorithms. The difficulty arises as model fit is performed internally, and observations on the fit of the model are not always available to
computer program users.

**Link to GWAS.** By considering genome scans for selection as a category of genome-wide association studies, one can draw on a much broader literature that has grappled with the problem of addressing calibration and multiple testing issues. Traditionally, classical GWAS put substantial efforts on avoiding false positives and on improving power to detect true associations. The eventual development of GWAS graphical methods, which display histograms of significance values and use genomic control to reduce the rate of false discoveries, provides a strong parallel to our approaches in this paper. However, there are some differences between regression models used in GWAS and genome scans for selection. First, the direct application of GWAS tests to genome scans for selection, for which polygenic effects can be considered to be the rule, leads to overly conservative tests (Frichot *et al.* 2013). Another difference can be explained in terms of generative models (Listgarden *et al.* 2010). Traditional GWAS approaches investigate the association between genetic polymorphisms and specific individual traits or phenotypes. Most phenotypic traits are heritable, and a part of the variation among individuals can be explained by genetic variance components. Most GWAS approaches correct for confounding effects by estimating a genetic similarity matrix, and use this matrix for the covariance of random effects in a mixed model (Yu *et al.* 2006). Unlike phenotypic traits, ecological variables do not follow any mode of inheritance, and generative GWAS models may be inappropriate in this context. Thus, direct applications of GWAS regression models to detecting ecological selection are not straightforward (Yoder *et al.* 2014). The same remarks apply to the closely related fields of phylogenetic regression methods (Grafen 1989, Harvey & Pagel 1991). We believe that statistical frameworks developed in GWAS (or phylogenetic regression methods) could largely benefit PD and EA tests if their generative models are reformulated for these new applications, for example by incorporating polygenic effects, or by modeling the covariance matrix of the random effects in a way that accounts for the spatial autocorrelation of ecological variables.

**Remarks on the power to reject neutrality.** Connections between genome scans for selection and chi-squared tests increase our understanding of which factors could influence the power of
neutrality tests. In $F_{ST}$-based tests, the power of tests is maximized when the sampled population sizes are equal, and having uneven sample sizes generally decreases the power to reject neutrality. This property is an immediate consequence of Fisher’s ANOVA $F$-statistic which has minimal variance when the sample sizes are equal. In EA tests, the power of tests increases when individuals are statistically uncorrelated. Thus, the tests have increased power when the geographic coverage of the study area is maximal, and when the sampling design does not cluster individuals into groups. These properties indicate that uneven sampling can decrease the power of EA and PD tests to reject neutrality (Lotterhos & Whitlock 2015). In simulations of EA tests, the tests have less power when the direction of the environmental gradient is parallel to the first principal component of genetic variation (Frichot et al. 2015, Supp. Mat. 3). This property is a very general feature of linear regression tests, and it explains why approaches testing correlation between genetic variation and population structure inherently lack power to detect weak selection. Typical cases include PD approaches based on $F_{ST}$ in which the first principal component of the genetic variation aligns to population groupings (McVean 2009). In this case, PD tests have limited power, and only the hardest sweeps can be detected. In EA tests, the direction of the ecological gradient does not necessarily align to the first principal axis of genetic variation, and EA tests can detect selection on markers having weaker effects on polygenic traits (Pritchard et al. 2010). Predicting which tests will lead to the maximal power in a particular model is, however, difficult to do. While test performance is case-specific, combining several well-calibrated tests can decrease the sensitivity to particular models, and lead to robust testing approaches with reasonable power to reject neutrality.

**Conclusions.** Genome scan methods that properly control for a false discovery rate are important to researchers identifying patterns of natural selection, and should be considered critically when researchers ascribe a functional interpretation to a list of candidate loci or choose to pursue experimental validation of 'selected' loci. Implementation of test calibration methods based on inflation factors offer an important validation step to ensure that genome scan tests have been properly calibrated. The main remaining challenges for the interpretation of results in genome scans for selection is to ensure that confounding variables, such as complex population structure,
uneven sample sizes, linkage disequilibrium, and sequencing biases, have been taken into account.

Corrections for these confounding effects and proper test calibration are essential before one takes the step to adjust significance values for multiple comparisons. One powerful approach to correct for confounding variables is to specify error structures in mixed models. With the advent of massive genomic datasets, regression methods that include random effects or latent factors may be the most robust way to provide calibration of test $p$-values and control false discoveries in genome scans for selection.

**Data accessibility**

The data used in our tutorial examples and in supplementary files 1 and 2 have been submitted to Dryad. Their entries are


The *A. thaliana* datasets used in this study are publicly available from the following link

https://github.com/Gregor-Mendel-Institute/atpolydb

**Acknowledgments**

We are grateful to Louis Bernatchez for his invitation to write this manuscript, and to Robin Waples for his useful comments. We are also grateful to three anonymous reviewers for their time and efforts in evaluating our manuscript. Olivier François acknowledges support from Grenoble INP, and from the “Agence Nationale de la Recherche” (project AFRICROP ANR-13-BSV7-0017). Helena Martins acknowledges support from the “Ciências sem Fronteiras” scholarship program from the Brazilian government. This work has been partially supported by the LabEx PERSYVAL-Lab (ANR-11-LABX-0025-01) funded by the French program Investissement d’Avenir.
References


identify loci underlying local adaptation. *Genetics* 185: 1411-1423.


in the genetic analysis of gene expression. *Proceedings of the National Academy of Sciences*
107(38):16465-16470.


Table 1. **Population differentiation and ecological association tests.** References for all the methods listed are provided in the text.

<table>
<thead>
<tr>
<th>Method</th>
<th>Pop.</th>
<th>Type</th>
<th>Test statistic</th>
<th>Null hypothesis</th>
<th>d</th>
<th>Correction method</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lewontin-Krakauer</td>
<td>Yes</td>
<td>PD</td>
<td>$(d - 1)F_{ST}^2/F_{ST}$</td>
<td>Chi$^2(d - 1)$</td>
<td>Number of population</td>
<td>Empirical null</td>
</tr>
<tr>
<td>ANOVA</td>
<td>Yes</td>
<td>PD</td>
<td>$(n - d)F_{ST}/(1 - F_{ST})$</td>
<td>Chi$^2(d - 1)$</td>
<td>Number of population</td>
<td>Empirical null</td>
</tr>
<tr>
<td>FDIST2</td>
<td>Yes</td>
<td>PD</td>
<td>$F_{ST}$</td>
<td>Monte Carlo</td>
<td>Number of population</td>
<td>Empirical</td>
</tr>
<tr>
<td>$F_{ILK}$</td>
<td>Yes</td>
<td>PD</td>
<td>$T_{F_{ILK}}^{'}$</td>
<td>Chi$^2(d - 1)$</td>
<td>Number of population</td>
<td>Kinship matrix</td>
</tr>
<tr>
<td>OUTFLANK</td>
<td>Yes</td>
<td>PD</td>
<td>$F_{ST}^{'}$</td>
<td>Chi$^2(d)$</td>
<td>Maximum Likelihood</td>
<td>Trimming</td>
</tr>
<tr>
<td>BAYENV2</td>
<td>Yes</td>
<td>PD/EA</td>
<td>$X^{'}X$</td>
<td>Chi$^2(d - 1)$</td>
<td>Number of population</td>
<td>Kinship matrix</td>
</tr>
<tr>
<td>BAYPASS</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PCADAPT</td>
<td>No</td>
<td>PD</td>
<td>$c'h^2$</td>
<td>Chi$^2(d)$</td>
<td>Number of PCs</td>
<td>Empirical null</td>
</tr>
<tr>
<td>TESS3</td>
<td>No</td>
<td>PD</td>
<td>$(n - d)f_{ST}/(1 - f_{ST})$</td>
<td>Chi$^2(d - 1)$</td>
<td>Number of clusters</td>
<td>Empirical null</td>
</tr>
<tr>
<td>SAM</td>
<td>No</td>
<td>EA</td>
<td>$z^2$-score</td>
<td>Chi$^2(d)$</td>
<td>Number of covariate</td>
<td>Empirical null</td>
</tr>
<tr>
<td>GLMM</td>
<td>No</td>
<td>EA</td>
<td>$z^2$-score</td>
<td>Chi$^2(d)$</td>
<td>Number of covariate</td>
<td>Covariance matrix</td>
</tr>
<tr>
<td>EMMAX</td>
<td>No</td>
<td>-</td>
<td>$z^2$-score</td>
<td>Chi$^2(d)$</td>
<td>Number of covariate</td>
<td>Kinship matrix</td>
</tr>
<tr>
<td>LFMM</td>
<td>No</td>
<td>EA</td>
<td>$z^2$-score</td>
<td>Chi$^2(d)$</td>
<td>Number of covariate</td>
<td>Latent factors</td>
</tr>
</tbody>
</table>

*Pop* indicates the use of population data (otherwise individual data). *Null hypothesis* describes the distribution of the test statistic under the null-hypothesis. *d* is the degree of freedom of the chi-squared distribution. *Correction method* summarizes the method used to obtain corrected significance values.
Table 2. **Local adaptation in Scandinavian lines of *A. thaliana***. List of loci with annotations among the fifty top “hits” (expected FDR level of 0.1%, 167 candidate loci).

<table>
<thead>
<tr>
<th>Chromosome</th>
<th>Position (kb)</th>
<th>Gene</th>
<th>Biological process</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>899</td>
<td>AT1G03600</td>
<td>Photosynthesis</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1,145</td>
<td>AT1G04280</td>
<td>Unknown</td>
<td>Huber et al.</td>
</tr>
<tr>
<td></td>
<td>1,182</td>
<td>AT1G04390</td>
<td>Flower morphogenesis</td>
<td>Huber et al.</td>
</tr>
<tr>
<td></td>
<td>20,009</td>
<td>Intergenic</td>
<td></td>
<td>Huber et al.</td>
</tr>
<tr>
<td></td>
<td>20,144</td>
<td>Intergenic</td>
<td></td>
<td>Huber et al.</td>
</tr>
<tr>
<td>2</td>
<td>9,608</td>
<td>AT2G22620</td>
<td>Carbohydrate metabolism</td>
<td></td>
</tr>
<tr>
<td></td>
<td>17,929</td>
<td>AT2G43130</td>
<td>Response to heat</td>
<td></td>
</tr>
<tr>
<td></td>
<td>18,256</td>
<td>AT2G44110</td>
<td>Response to biotic stimulus</td>
<td></td>
</tr>
<tr>
<td></td>
<td>18,679</td>
<td>AT2G45340</td>
<td>Regulation of meristem growth</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>2,265</td>
<td>AT5G07220</td>
<td>Regulation of abiotic stress</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2,298</td>
<td>AT5G07290</td>
<td>Meristem growth</td>
<td>Huber et al.</td>
</tr>
<tr>
<td></td>
<td>2,561</td>
<td>AT5G07990</td>
<td>Response to UV</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2,564</td>
<td>AT5G08000</td>
<td>Response to heat</td>
<td></td>
</tr>
<tr>
<td></td>
<td>3,591</td>
<td>AT5G11250</td>
<td>Defense response</td>
<td></td>
</tr>
<tr>
<td></td>
<td>6,779</td>
<td>AT5G20080</td>
<td>Unknown</td>
<td>Huber et al.</td>
</tr>
<tr>
<td></td>
<td>9,726</td>
<td>AT5G27540</td>
<td>Response to freezing</td>
<td></td>
</tr>
<tr>
<td></td>
<td>9,780</td>
<td>AT5G27630</td>
<td>Response to light stimulus</td>
<td></td>
</tr>
</tbody>
</table>

The last column indicates whether the SNP was reported to be under selection in (Huber et al. 2014).
Figure 1: *Main analytical steps in performing genome scans for selection.*
Figure 2: Summary of simulated data analysis for example 1. Venn diagrams, observed FDR and power for 4 genome scan tests: $F_{ST}$ (ANOVA), FDIST2, adjusted $F_{ST}$, adjusted FDIST2. Adjusted $F_{ST}$ and adjusted FDIST2 are tests for which the significance values were recalibrated using the genomic inflation factor. Observed FDR and power were computed for an expected level of FDR of 10 percent. The numbers in the circles represent the number of positive tests for each method. See Supp. Mat. 1 for an extended description of the data and all the processing steps that generated the Figure.
Figure 3: Adaptation to climate in European lines of A. thaliana. A) Histogram of test significance values resulting from the combination of 5 LFMM models (or runs) using $K = 6$ latent factors, B) Manhattan plot of minus log$_{10}$ (Log) $p$-values for the plant chromosome 1. Candidate loci resulting from the application of the FDR control algorithm are identified with circles (expected FDR level of 7%). See Supp. Mat. 2 for an extended description of this analysis and all the processing steps that generated the Figure.
Box 1. Glossary

1. **Calibration of \( p \)-values**: An algorithmic correction used to ensure that the histogram of \( p \)-values is flat when the null-hypothesis is true.

2. **Chi-squared (\( \chi^2 \)) distribution**: A probability distribution for most genome scan test statistics under selective neutrality.

3. **False Discovery Rate (FDR)**: The expected proportion of false positives among the list of positive tests.

4. **FDR control algorithm**: Any algorithm utilized to ensure that the expected value of the FDR is lower than a pre-specified level.

5. **Population differentiation (PD) tests**: Tests based on the \( F_{ST} \) statistic or variants of this statistic.

6. **Ecological association (EA) tests**: Tests based on the regression of allele frequencies on environmental variables.

7. **Genome-wide association studies (GWAS)**: Tests of association between phenotypic and genotypic frequencies or between phenotypes and gene expression levels.

8. **Genomic inflation factor (GIF)**: The median of squared \( z \)-scores divided by the median of the chi-squared distribution. Here, an estimate of the test statistic at a selectively neutral locus.

9. **Inflation factor**: Any constant used to rescale \( z \)-scores and recalibrate incorrect \( p \)-values.

10. **Linear mixed model**: A linear regression model for correlated responses that includes fixed and random effects.

11. **Latent factor mixed model (LFMM)**: A linear mixed model for which the environment is used as a fixed effect, and that includes latent factors.

12. **Power**: The proportion of tests that correctly reject the null hypothesis.
13. **Squared z-score**: Test statistic used in association studies ($z^2$ follows a chi-squared distribution).
Box 2. FDR control assumptions

The underlying principle of FDR control algorithms relies on the fact that significance values (p-values) corresponding to truly null hypotheses, i.e., selectively neutral loci, are uniformly distributed over the interval (0,1). To see why the uniform distribution assumption is critical here, let us recall that the FDR is the expected value of the ratio $F/S$ where $F$ is the number of false positive tests and $S$ is the number of significant (positive) tests (Storey & Tibshirani 2003). Let $L_0$ be the total number of truly null hypotheses. To provide control of the FDR at level $\alpha$, Benjamini & Hochberg (1995) proposed to sort the set of $p$-values and considered the largest value $k$ such that $p(k) \leq k\alpha/L$ ($L$ is the total number of tests). All tested items with $p$-values lower than $p(k)$ were included in the list of discoveries. To compute the expected value of the ratio $F/S$, let us assume that the random value $S$ is equal to $S = k$. According to the uniform distribution, the expected number of times a truly null hypothesis is rejected is equal to $E[F|S = k] = L_0 k\alpha/L$ in the Benjamini-Hochberg algorithm. Assuming $k \geq 1$, we have

$$E[F/S|S = k] = E[F/k|S = k] = L_0/L \alpha \leq \alpha.$$ 

Under these expectations, we obtain an FDR that is under control. To check that the uniform distribution assumption is correct, standard graphical approaches can be used. These methods display histograms of test $p$-values as in Figure 4.

Figure 4: Histograms of test significance values ($p$-values) prior to the application of FDR control algorithms (artificial data). GIF is the genomic inflation factor for each dataset.
Box 3. Recalibrating and combining significance values

False positives arise in statistical tests when the null-hypothesis ($H_0$) is misspecified. This phenomenon happens in genome scans because the tests ignore the proportion of variance explained by selectively neutral processes such as past demography, genetic relatedness, population structure and other confounding factors. The presence of confounding factors directly impacts the distribution of the test statistic under neutrality, and indirectly impacts multiple testing correction procedures that assume a uniform distribution of significance values under $H_0$.

Consider a statistical test based on the chi-squared distribution with $D$ degrees of freedom (Table 1), and denote by $z^2_\ell$ the test statistic used to evaluate significance at locus $\ell$. Test recalibration is the process by which one builds an empirical null-hypothesis from the data, and re-evaluates significance values in a way that accounts for confounding errors. The target of recalibration approaches is to evaluate the expected value of the test statistic at selectively neutral loci. Any estimate of this value, $\lambda$, is called an inflation factor. Given an inflation factor, significance values are computed for each locus $\ell$ as follows ($L$ is the number of loci)

$$p_\ell = P(\chi^2_D > \frac{z^2_\ell}{\lambda}), \quad \ell = 1, \ldots, L,$$

so that $\lambda$ corrects the inflation of the test statistic $z^2_\ell$ at each locus (see Supp. Mat. 1 for several examples of the use of this formula). A common approach to evaluate the constant $\lambda$ is a method called genomic control, that estimates the genomic inflation factor, obtained after computing the median of the squared $z$-scores and dividing this value by the median of the chi-square distribution with $D$ degrees of freedom (Devlin & Roeder 1999). A more general way to evaluate inflation factors is by using the local FDR method developed by Efron (2004), for example implemented in the R program fdrtool (Strimmer 2008).

Similar scaling approaches can be applied to calculate the significance values resulting from the combination of several methods, each testing the same null hypothesis. For example, a robust version of the Stouffer method is based on the median of $z$-scores obtained for each method at each locus (Whitlock 2005, see examples in Supp. Mat. 1 and 2). In the case of independent tests, the scaling factor for the median is equal to $\sqrt{\pi m/2}$, where $m$ is the number of methods used. In
the case of dependent tests, inflation factors can be used to determine the scaling factor, so that significance values resulting from the combination of tests have a flat distribution under $H_0$. 
Box 4. Summary points

1. Genome scans for selection are two-stage procedures: One first performs statistical tests that return locus significance values, and then makes decisions about which loci to retain as candidates for selection.

2. Based on the histogram of test significance values, statistical test calibration is the key to FDR control.

3. Two approaches are available: 1) empirical null-hypothesis testing, as illustrated by the estimation of inflation factors, and 2) modeling confounding errors, as illustrated by latent factor models.

4. Combining several well-calibrated statistical tests using the $z$-score method can increase power to reject neutrality.

5. Following test calibration, candidate loci can be selected on the basis of a classical FDR control algorithm.
Figure S1. Local adaptation in Scandinavian lines of A. thaliana. Top: Histogram of test $p$-values using five runs of LFMM with $K = 2$ latent factors. Bottom: Manhattan plot of minus log$_{10}$ $p$-values.