Detecting Signatures of Selection Through Haplotype Differentiation Among Hierarchically Structured Populations

María Inés Fariello,^{*,†,‡,1} Simon Boitard,^{*} Hugo Naya,^{†,§} Magali SanCristobal,^{*} and Bertrand Servin^{*}

^{*}Laboratoire de Génétique Cellulaire, Institut National de la Recherche Agronomique, 31326 Toulouse, France, [†]Unidad de Bioinformática, Institut Pasteur, 11400 Montevideo, Uruguay, [‡]Facultad de Ingeniería, Universidad de la República, 11300 Montevideo, Uruguay, and [§]Facultad de Agronomía, Universidad de la República, 12900 Montevideo, Uruguay

ABSTRACT The detection of molecular signatures of selection is one of the major concerns of modern population genetics. A widely used strategy in this context is to compare samples from several populations and to look for genomic regions with outstanding genetic differentiation between these populations. Genetic differentiation is generally based on allele frequency differences between populations, which are measured by F_{ST} or related statistics. Here we introduce a new statistic, denoted hapFLK, which focuses instead on the differences of haplotype frequencies between populations. In contrast to most existing statistics, hapFLK accounts for the hierarchical structure of the sampled populations. Using computer simulations, we show that each of these two features—the use of haplotype information and of the hierarchical structure of populations—significantly improves the detection power of selected loci and that combining them in the hapFLK statistic provides even greater power. We also show that hapFLK is robust with respect to bottlenecks and migration and improves over existing approaches in many situations. Finally, we apply hapFLK to a set of six sheep breeds from Northern Europe and identify seven regions under selection, which include already reported regions but also several new ones. We propose a method to help identifying the population(s) under selection in a detected region, which reveals that in many of these regions selection most likely occurred in more than one population. Furthermore, several of the detected regions correspond to incomplete sweeps, where the favorable haplotype is only at intermediate frequency in the population(s) under selection.

THE detection of molecular signatures of selection is one of the major concerns of modern population genetics. It provides insight on the mechanisms leading to population divergence and differentiation. It has become crucial in biomedical sciences, where it can help to identify genes related to disease resistance (Tishkoff *et al.* 2001; Barreiro *et al.* 2008; Albrechtsen *et al.* 2010; Fumagalli *et al.* 2010; Cagliani *et al.* 2011), adaptation to climate (Lao *et al.* 2007; Sturm 2009; Rees and Harding 2012), or altitude (Bigham *et al.* 2010; Simonson *et al.* 2010). In livestock species, where artificial selection has been carried out by humans since domestication, it contributes to map traits of agronomical interest,

for instance, related to milk (Hayes *et al.* 2009) or meat (Kijas *et al.* 2012) production.

Efficiency of methods for detecting selection varies with the considered selection timescale (Sabeti *et al.* 2006). For the detection of selection within species (the ecological scale of time), methods can be classified into three groups: methods based on (i) the high frequency of derived alleles and other consequences of hitchhiking within population (Kim and Stephan 2002; Kim and Nielsen 2004; Nielsen *et al.* 2005; Boitard *et al.* 2009), (ii) the length and structure of haplotypes, measured by extended haplotype homozygosity (EHH) or EHH-derived statistics (Sabeti *et al.* 2002; Voight *et al.* 2006), and (iii) the genetic differentiation between populations, measured by F_{ST} or related statistics (Lewontin and Krakauer 1973; Beaumont and Balding 2004; Foll and Gaggiotti 2008; Riebler *et al.* 2008; Gautier *et al.* 2009; Bonhomme *et al.* 2010). Methods of the latter kind, which we focus on, are particularly suited to the study of species that are structured in well-defined populations, such

Copyright © 2013 by the Genetics Society of America
doi: 10.1534/genetics.112.147231

Manuscript received October 26, 2012; accepted for publication December 14, 2012
Supporting information is available online at <http://www.genetics.org/lookup/suppl/doi:10.1534/genetics.112.147231/-/DC1>.

¹Corresponding author: Auzeville, BP 52627 Chemin de Borde Rouge, 31326 Castanet Tolosan Cedex, France. E-mail: mfariello@toulouse.inra.fr

as most domesticated species. In contrast to methods based on the frequency spectrum (i) or the excess of long haplotypes (ii), they can detect a wider range of selection scenarios, including selection on standing variation or incomplete sweeps, albeit up to a given extent (Innan and Kim 2008; Yi *et al.* 2010).

The most widely used statistic with which to detect loci with outstanding genetic differentiation between populations is the F_{ST} statistic (Barreiro *et al.* 2008; Myles *et al.* 2008). The general application of the F_{ST} -based scan for selection is to identify outliers in the empirical distribution of the statistics computed genome-wide. One major concern with this approach is that it implicitly assumes that populations have the same effective size and derived independently from the same ancestral population. If this hypothesis does not hold, which is often the case, genome scans based on raw F_{ST} can suffer from bias and false positives, an effect that is similar to the well-known effects of cryptic structure in genome-wide association studies (Price *et al.* 2010). To cope with this problem several methods have been proposed to account for unequal population sizes (Beaumont and Balding 2004; Foll and Gaggiotti 2008; Riebler *et al.* 2008; Gautier *et al.* 2009); however, few solutions have been proposed to deal with the hierarchical structure of populations (Excoffier *et al.* 2009). Among them Bonhomme *et al.* (2010) proposed an extension of the classical Lewontin and Krakauer (LK) test (Lewontin and Krakauer 1973), where the hierarchical population structure is captured through a kinship matrix, which is used to model the covariance matrix of the population allele frequencies. A similar covariance matrix was also introduced in a related context to account for the correlation structure arising from population geography (Coop *et al.* 2010).

All F_{ST} -based approaches discussed above are single marker tests; *i.e.*, markers are analyzed independently from each other. As dense genotyping data and sequencing data are now common in population genetics, accounting for correlations between adjacent markers has become necessary. Furthermore, haplotype structure contains useful information for the detection of selected loci, as demonstrated by the within-population methods mentioned above (class ii). Several strategies for combining the use of multiple populations and of haplotype information have thus been proposed recently. These include the development of EHH-related statistics for the comparison of pairs of populations (Sabeti *et al.* 2007; Tang *et al.* 2007), the introduction of dependence between SNPs (single nucleotide polymorphisms) in F_{ST} -based approaches through autoregressive processes (Guo *et al.* 2009; Gompert and Buerkle 2011), or the computation of F_{ST} using local haplotype clusters that are considered as alleles (Browning and Weir 2010). However, none of these approaches accounts for the possibility that populations are hierarchically structured.

We present here an haplotype-based method for the detection of positive selection from multiple population data. This new statistic, hapFLK, builds upon the original FLK statistic (Bonhomme *et al.* 2010). As FLK, it incorporates hierarchical

structure of populations, but the test is extended to account for the haplotype structure in the sample. For this, it uses a multi-point linkage disequilibrium model (Scheet and Stephens 2006) that regroups individual chromosomes into local haplotype clusters. The principle is to exploit this clustering model to compute “haplotype frequencies,” which are then used to measure differentiation between populations. The idea of using localized haplotype clusters to study genetic data on multiple populations has been proposed before (Jakobsson *et al.* 2008; Browning and Weir 2010). Browning and Weir (2010) showed that using haplotype clusters rather than SNPs allowed circumvention, to some extent, of the problems arising from SNP ascertainment bias. They also showed that two genome regions known to have been under strong positive selection in particular human populations exhibited large population-specific haplotype-based F_{ST} . Jakobsson *et al.* (2008) showed by using fastPHASE that there was a predominance of a single cluster haplotype in the HapMap population of Utah residents with ancestry from northern and western Europe (CEU population) in the region of the LCT gene and interpreted this signal as a recent selective sweep.

In this article, we examined in detail the ability of statistics based on population differentiation at the haplotype level to capture selection signals. Using computer simulations, we study the power and robustness of our new haplotype-based method for different selection and sampling scenarios and compare it to single-marker [F_{ST} and FLK (Bonhomme *et al.* 2010)] and haplotype-based approaches (XP-EHH; Sabeti *et al.* 2007). To illustrate the interest of this approach, we provide a practical example on a set of six sheep breeds for which dense genotyping data have been recently released by the Sheep HapMap Project (International Sheep Genomics Consortium 2012). In this context, we propose a new strategy for the detection of outliers loci in genome scans for selection and describe a method for the identification of the populations that have experienced selection at a detected region.

Methods

Test statistics

F_{ST} and FLK tests for SNPs: Consider a set of n populations that evolved without migration from an ancestral population and a set of L SNPs in these populations. For a given SNP, let $p = (p_1, \dots, p_i, \dots, p_n)'$ be the vector of the reference allele frequency in all populations. Denoting $\bar{p}$ and s_p^2 the sample estimates of the mean and variance of the p_i s, F_{ST} at this SNP is given by $s_p^2/\bar{p}(1 - \bar{p})$. F_{ST} quantifies the genetic differentiation between populations and is commonly used to detect loci under selection. Loci with outstanding high (resp. low) values of F_{ST} can be declared as targets of positive (resp. balancing) selection.

However, if the sampled populations have unequal effective sizes or/and are hierarchically structured, genome scans based on raw F_{ST} values can bias inference. For instance, a given allele frequency difference between two closely related populations should provide more evidence for selection

than the same difference between two distantly related populations. To account for these drift and covariance effects when detecting loci under selection, Bonhomme *et al.* (2010) introduced the FLK statistic

$$T_{\text{FLK}} = (p - p_0 \mathbf{1}_n)' \text{Var}(p)^{-1} (p - p_0 \mathbf{1}_n), \quad (1)$$

where p_0 is the allele frequency in the ancestral population and $\text{Var}(p)$ is the expected covariance matrix of vector p , which they modeled as

$$\text{Var}(p) = \mathcal{F} p_0 (1 - p_0). \quad (2)$$

$\mathcal{F}_{ii}$ is the expected inbreeding coefficient in population i and $\mathcal{F}_{ij}$ is the expected inbreeding coefficient in the ancestral population common to populations i and j . The entries of the kinship matrix $\mathcal{F}$ represent the amount of drift accumulated on the different branches of the population tree. They can be derived as a function of the divergence times and the effective population sizes along the population tree, as described in [Supporting Information, File S1](#).

In practice, these demographic parameters are unknown and $\mathcal{F}$ must be estimated from genome-wide data. Here, it is done as follows: first, pairwise Reynolds' distances (Reynolds *et al.* 1983) between populations (including an outgroup) are computed for each SNP and averaged over the genome. Then, a phylogenetic tree is fitted from these distances using the neighbor-joining algorithm. The branch lengths of this tree are finally combined to compute $\mathcal{F}$ entries. More details on this procedure can be found in Bonhomme *et al.* (2010). Given the estimation of $\mathcal{F}$, the unbiased estimator of p_0 is obtained as

$$\hat{p}_0 = \frac{\mathbf{1}_n' \mathcal{F} - 1p}{\mathbf{1}_n' \mathcal{F}^{-1} \mathbf{1}_n} = w' p$$

and can be used in Equations 1 and 2 to obtain T_{FLK} .

Under the assumption that all populations diverged simultaneously from the same ancestral population (star-like evolution) and with the same population size, $\mathcal{F}$ is equal to $\bar{F}_{\text{ST}} I_n$, where $\bar{F}_{\text{ST}}$ is the average F_{ST} over all SNPs and I_n is the identity matrix of size n . In this case, T_{FLK} is equivalent to the LK statistic (Lewontin and Krakauer 1973):

$$T_{\text{LK}} = \frac{n-1}{\bar{F}_{\text{ST}}} F_{\text{ST}}.$$

FLK test for multiallelic markers: Considering haplotypes as multiallelic markers, an extension of the FLK statistic in the case where each locus presents more than two alleles is required. Letting A be the number of alleles at a given locus, the allele frequency vector becomes

$$P = \left(\underbrace{p_{11}, \dots, p_{1n}}_{\text{allele 1}}, \dots, \underbrace{p_{A1}, \dots, p_{An}}_{\text{allele A}} \right)' = (p_1, \dots, p_A)'$$

and a multiallelic version of the T_{FLK} statistic is provided by

$$T_{\text{FLK}} = (P - P_0 \otimes \mathbf{1}_n)' \text{Var}(P)^{-1} (P - P_0 \otimes \mathbf{1}_n), \quad (3)$$

where $\otimes$ denotes the Kronecker product and $P_0 = (p_{10}, \dots, p_{A0})'$ contains the allele frequencies of the A alleles in the ancestral population. $\text{Var}(P)$ is written

$$\begin{aligned} \text{Var}(P) &= \begin{pmatrix} \text{Var}(p_1) & \cdots & \text{Cov}(p_1, p_A) \\ \vdots & \text{Var}(p_A) & \vdots \\ \text{Cov}(p_A, p_1) & \cdots & \text{Var}(p_A) \end{pmatrix} \\ &= B_0 \otimes \mathcal{F}, \end{aligned} \quad (4)$$

with $B_0 = \text{diag}(P_0) - P_0 P_0'$. Each diagonal block of $\text{Var}(P)$ corresponds to the biallelic covariance matrix for one of the A alleles, while the extra-diagonal blocks arise from the covariance terms between different alleles. Similar to the biallelic case, P_0 is estimated by $\hat{P}_0 = (w p_1, \dots, w p_A)'$. $\text{Var}(P)$ is inverted using the Moore–Penrose generalized inverse.

FLK test for haplotypes: The Scheet and Stephens (2006) model summarizes local haplotype diversity in a sample through a reduction of dimension by clustering similar haplotypes together. These clusters can then be considered as alleles to compute the haplotype version of T_{FLK} statistic. Let g_i^ℓ be the genotype observed for individual i at marker ℓ . In the hidden markov model of Scheet and Stephens (2006), g_i^ℓ is associated to a hidden state $z_i^\ell = (z_i^{\ell 1}, z_i^{\ell 2})$, where $z_i^{\ell 1}$ and $z_i^{\ell 2}$ represent the pair of clusters giving rise to the (diploid) individual genotype. The Markov structure of $z_i = (z_i^1, \dots, z_i^L)$ along the genome implies that cluster memberships of close markers are correlated, which allows us to account for linkage disequilibrium effects. When this model is fitted to the whole genotype data g , it provides for each individual i , marker ℓ , and cluster k the posterior probabilities $\mathcal{P}(z_i^{\ell 1} = k | g, \Theta)$ and $\mathcal{P}(z_i^{\ell 2} = k | g, \Theta)$, where Θ is a vector of estimated model parameters (see Scheet and Stephens 2006 for more details). Cluster probabilities in each population j are obtained by averaging the probabilities of the n_j individuals of this population, i.e.,

$$p_{kj}^\ell = \frac{1}{2n_j} \sum_{i=1}^{n_j} (\mathcal{P}(z_i^{\ell 1} = k | g, \Theta) + \mathcal{P}(z_i^{\ell 2} = k | g, \Theta)). \quad (5)$$

Considering clusters as alleles and population-averaged probabilities as population frequencies, the allele frequency vector of a marker ℓ is

$$P^\ell = \left(\underbrace{p_{11}^\ell, \dots, p_{1n}^\ell}_{\text{cluster 1}}, \underbrace{p_{21}^\ell, \dots, p_{2n}^\ell}_{\text{cluster 2}}, \dots, \underbrace{p_{K1}^\ell, \dots, p_{Kn}^\ell}_{\text{cluster K}} \right)'.$$

For each marker ℓ , the multiallelic statistic T_{FLK} is computed according to Equation 3, with a small modification in the derivation of $\text{Var}(P)$. Clusters that are fitted in the present population cannot exactly be considered as real alleles that

already existed in the ancestral population, as assumed by the original T_{FLK} statistic. Moreover, the generalized inverse of $B_0 \otimes \mathcal{F}$ was found numerically unstable for small P_{a0} values, which are very common when the number of alleles is large. Consequently, the B_0 matrix is replaced by the identity matrix I_A in Equation 4, leading to the statistic

$$T_{FLK} = (P - P_0 \otimes \mathbf{1}_n)' (\mathcal{I} \otimes \mathcal{F})^{-1} (P - P_0 \otimes \mathbf{1}_n). \quad (6)$$

Simulations confirmed that this version of the test was more powerful than the one including B_0 (Figure S1).

For the model of Scheet and Stephens (2006), parameter estimates and cluster membership probabilities are obtained using an expectation maximization (EM) algorithm. Because this algorithm converges to a local maximum, it is useful to run it several times from different starting points. Applying the model to haplotype phasing, Guan and Stephens (2008) and Scheet and Stephens (2006) observed that averaging the results from these different runs was more efficient than keeping the maximum-likelihood run, which may be due to the fact that different runs are optimal in different genomic regions. Following their strategy, we averaged the statistics obtained using Equation 6 from different EM iterations to finally obtain the haplotype extension of FLK. We denote this extension hapFLK.

The haplotype extension of the F_{ST} test, denoted hap F_{ST} in the simulation study, was obtained by replacing $\mathcal{F}$ by I_n in Equation 6, therefore ignoring the hierarchical structure of populations.

Software and computational considerations: Software implementing the hapFLK calculations is available at <https://forge-dga.jouy.inra.fr/projects/hapflk>. hapFLK comes with an increased computational cost compared to F_{ST} and FLK arising from the need to estimate the LD model on the data. The computational cost of the LD model used here, applied on unphased genotype data, is in $o(K^2IL)$ with K the number of clusters, I the number of individuals, and L the number of loci on a chromosome. As an example, fitting the model on sheep chromosome 1 with 5284 SNPs for 40 clusters and 278 individuals takes about 1 hr on a single processor. In our implementation of the Scheet and Stephens (2006) model, we perform computations in a parallel fashion allowing the decrease of computational costs on multiprocessor computers.

Simulations

To evaluate the performance of hapFLK and compare it to that of other tests, we performed a set of simulations mimicking the data obtained from dense SNP genotyping or full sequencing of samples from multiple populations. In particular, we designed our simulation to match the data produced within the Sheep HapMap Project (International Sheep Genomics Consortium 2012) (analyzed below), in terms of population divergence and SNP density.

Scenarios with constant size and no migration: Two scenarios were simulated, one with two populations and the

other one with four populations (Figure 1). The two-population scenario was designed to be a subtree of the four-population scenario, which allows comparison of the detection power obtained by testing the four populations jointly, with the power obtained by testing all possible pairs of populations.

The ancestral population was simulated using *ms* (Hudson 2002), with mutation rate $\mu = 10^{-8}$, recombination rate $c = 10^{-8}$ (1 cM/Mb), and region length $L = 5$ Mb. The effective population size and the number of simulated haplotypes were $N_e = 1000$ and $n_h = 4000$ for the two-population case, and $N_e = 2000$ and $n_h = 8000$ for the four-population case. The generated haplotypes had ~ 200 SNPs/Mb. The first two populations (top branches in Figure 1) were created independently by sampling half of the individuals from the founder population. A forward evolution of the populations after their initial divergence was then simulated with the *simuPOP* Python library (Peng and Kimmal 2005), under the Wright–Fisher model. During forward simulations, recombination was allowed but mutation was not.

Simulations were performed with and without selection. For scenarios with selection, selection occurred at a single locus, in the red branch shown in Figure 1. The selected locus was chosen as the closest to the center of the simulated region, among the SNPs with minor allele frequency equal to a predefined value (0.1, 0.05, 0.10, 0.20, or 0.30). The less-frequent allele of this SNP was given fitness $1 + s$, with selection intensity $s = 0.05$ (leading to $\alpha = 2 \cdot N_e \cdot s = 100$). Individuals' fitness was 1 for homozygotes with the nonselected allele, $1 + s$ for heterozygotes, and $(1 + s)^2$ for homozygotes with the selected allele.

At the end of each simulation replicate 50 individuals were sampled from each of the final populations, and SNPs with a minor allele frequency (MAF) $> 5\%$ kept. Two different genotyping densities were considered: 20 SNPs/Mb (equivalent to that of 60K SNPs in sheep) and 100–125 SNPs/Mb (all remaining SNPs). The statistics $T_{F_{ST}}$, $T_{hapF_{ST}}$, T_{FLK} , and T_{hapFLK} were computed at each SNP, assuming that the kinship matrix $\mathcal{F}$ was known. The estimation of $\mathcal{F}$ is very accurate for evolution scenarios with constant population size and no migration (see Bonhomme *et al.* 2010 and Figure S2). Parameters used for running the test were $K = 5$ (number of clusters) and $em = 5$ (number of EM runs) for the two-population scenario and $K = 20$ and $em = 5$ for the four-population scenario. These values were chosen for maximizing the detection power. Greater values did not improve this power, but increased computation time. For the two-population scenario, the XP-EHH statistic (Sabeti *et al.* 2007) was also computed at each SNP, using software obtained from <http://hgdp.uchicago.edu/Software/>.

Power of the tests was computed as follows. Ten thousand data sets were simulated under the null (neutrality) and 3000 were simulated under the alternative (selection) hypotheses, for each scenario considered. In simulations under selection, only replicates where the final frequency of the selected allele was $> 60\%$ were kept. For each replicate

Figure 1 Population trees for the two simulated scenarios. The red branch indicates the selected population and time during which selection acts.

and statistic S , the maximum value $S^{\max}$ over the 5-Mb region was recorded. This provides the distribution of $S^{\max}$ under the null and the alternative hypotheses. The power of a test with statistic S , for a given type I error α , is the proportion of simulations under selection for which $S^{\max} > q_{\alpha}$, where q_{α} is the $(1-\alpha)$ th quantile of the null distribution of $S^{\max}$.

Scenarios with bottlenecks or migrations: To study the robustness of the approach, more complex demographic events were investigated through three scenarios. They derived from the two-population scenario described above, with the following modifications:

1. A bottleneck in a single population: the effective size in this population was set to $N_e = 100$ in the first five generations following the split and to $N_e = 1852$ in later generations.
2. Asymmetric migration: at generation 51, population 1 sent 10% of migrants to population 2.
3. Symmetric migration: at generation 51, population 1 sent 10% of migrants to population 2 and received 10% of migrants from population 2.

In terms of expected drift at a single SNP, these scenarios are equivalent to the constant size scenario (see SI section 1.1 for a proof). Hence, they can be used to evaluate the influence of the underlying demographic model on hapFLK, while conditioning on a fixed value of $\mathcal{F}$. To ensure that the $\mathcal{F}$ matrix used in hapFLK fits the one that would be estimated from real data, 100 artificial whole genome data sets were created for each of the scenarios i–iii and used to estimate $\mathcal{F}$. Each artificial whole genome data set was created by simulating 500 independent genome segments of 5 Mb.

Robustness of hapFLK and XP-EHH were evaluated by comparing quantiles of each statistic obtained under bottleneck or migration demography with those obtained under a constant size evolution. No selection was applied in these simulations.

Evaluation of the detection power of hapFLK and XP-EHH under bottleneck (or migration) with selection was performed as described above; *i.e.*, distributions obtained under neutrality provided quantiles used to calibrate type I

error. Because scenarios i and ii are asymmetric, each one provided two different simulation scenarios under selection, one with selection in population 1 and one with selection in population 2.

Sheep data analysis

A whole genome scan for selection in sheep was performed using the genotype data from the Sheep HapMap Project (available at <http://sheephapmap.org/download.php>). The sheep HapMap data set includes 2819 animals from 74 breeds, collected in such a way that it represents most of the worldwide genetic diversity in the sheep. Genotypes at 48703 autosomal SNPs (after quality filtering) are available for these animals. Focus was placed on the North European group, all populations with <20 individuals being removed. Populations resulting from a recent admixture were also excluded because they are not compatible with the population tree model assumed for our test. Finally, the following populations were included in the analysis (sample size in parentheses): Galway (49), Scottish Texel (80), New Zealand Texel (24), German Texel (46), Irish Suffolk (55), and New Zealand Romney (24). The Soay breed was used as an out-group for computing the $\mathcal{F}$ matrix.

Parameters of the hapFLK analysis: To determine the number of clusters to be used in the fastphase model, the cross-validation procedure of fastPHASE, which indicated an optimal number of 45 clusters, was used. As the computational cost increases quadratically with the number of clusters, and as the genome scans performed on one single chromosome for 40 and 45 clusters provided very similar results, 40 clusters were used for the rest of the analysis. A sensitivity analysis indicated that on this data set 45 EM runs were required to obtain a stable estimate of hapFLK.

Computation of P -values: In contrast to the simulated data sets, real data do not provide null distribution allowing computation of P -values from the hapFLK statistics. Also, due to ascertainment bias in the SNP panel, we believe that performing neutral simulations based on an estimation of $\mathcal{F}$ is not a good strategy for this particular data set (see the Discussion for more details). P -values were thus estimated using an empirical approach (described below) exploiting the fact that selected regions, at least those that can be captured with hapFLK, affect a small portion of the genome.

The genome-wide distribution of hapFLK appeared to be bimodal, with a large proportion of values showing a good fit to a normal distribution and a small proportion of extremely high values (Figure S3). Consequently, P -values were estimated as follows. First, robust estimators of the mean and variance of hapFLK were obtained, to reduce the influence of outliers. For this estimation the `rlm` function of the package MASS (Venables and Ripley 2002) in R was used. hapFLK values were then standardized using these estimates and corresponding P -values were computed from a standard normal distribution. The resulting distribution of P -values across the genome was

found to be close to uniform for large P -values, consistent with a good fit to the normal distribution apart from the outliers that exhibit small P -values. Using the approach of Storey and Tibshirani (2003), the FDR estimated when calling significant hypotheses with $P < 10^{-3}$ was 5%.

Pinpointing the selected population: Similar to all F_{ST} -related tests, hapFLK detects genomic regions in which genetic data are globally not consistent with a neutral evolution, but does not directly indicate where selection occurred in the population tree. To investigate this question, branch lengths of the population tree were reestimated for each significant region, using SNPs exceeding the significance threshold. The principle was to fit (using ordinary least squares) the branch lengths to the local values of Reynolds genetic distances. For each branch the P -value for the null hypothesis of no difference between the lengths estimated from data in the region and in the whole genome was computed. We did this local tree estimation using either SNP or haplotype clusters frequencies. Details on the procedure are provided in [File S1](#), section 1.3.

Results

Simulation results

We performed a set of simulations to evaluate and compare the performance of hapFLK and other tests (see *Methods* for more details). To present the results of these simulations, we begin with scenarios that fit the assumptions of our model: a population tree without migration and with constant size within each branch. We then move to more complex demographic scenarios, which are expected to be less favorable to our test.

Interest of using haplotypes over SNPs: We first simulated data from two populations of the same effective size (Figure 1, left). In this setting, the structure-aware tests (FLK and hapFLK) are equivalent to their unaware counterparts (F_{ST} and $hapF_{ST}$ resp.).

In simulations mimicking dense genotyping data, the use of haplotype information (hapFLK) provides more detection power than the use of single SNP tests (F_{ST}) (Figure 2). This holds for both hard sweeps ($p_0 = 0.01$) and soft sweeps detection (p_0 up to 0.3). XP-EHH, which also makes use of haplotype information, has more power than F_{ST} but less than hapFLK for hard sweeps detection. The decrease in power for soft sweeps is also more pronounced for XP-EHH (Figure 2), which is expected because XP-EHH is designed to detect the rise in frequency of one single haplotype.

Focusing on hapFLK, we further studied the evolution of the detection power as a function of the initial and final frequencies of the selected allele (Figure S5). Although soft sweeps are obviously harder to detect, there is still reasonable power to detect such events with hapFLK. For example, when the initial frequency is 20% and the final frequency is 90%, the detection power is >75%, for a type I error rate of 1%.

When selection acts on mutations at initial low frequency, the detection power is relatively high (around 60%) even for incomplete sweeps with a final frequency of 50–60%.

We also compared FLK, hapFLK, and XP-EHH in simulations mimicking data arising from full sequencing or imputation from a sequenced reference panel. This increase in marker density results in a greater power for all tests (Figure S6). In this setting, FLK is the most powerful. This comes from the fact that the selected SNP, where the allele frequency difference between populations is expected to be the largest, is always included in the sample in this simulation setting. In contrast, the selected SNP itself is often missing when analyzing genotyping data, and information concerning this SNP is then better captured by haplotypes than by single neighboring SNPs. All results below were obtained on simulations mimicking dense genotyping data.

Hierarchical structure of populations: We then considered a four-population sample, where populations are hierarchically structured (Figure 1, right). This allows to compare hapFLK with related tests accounting for population structure only (FLK), haplotype information only ($hapF_{ST}$), or none of these features (F_{ST}). As expected, the least powerful approach in this scenario is the classical F_{ST} . The gain in power provided by using a haplotype-based approach is of similar size to that provided by accounting for population structure. Finally, combining the two within a single statistic (hapFLK) results in an even greater power gain (Figure 3). This result holds for initial frequencies >1% although the difference between haplotype and single SNP tests tend to decrease with increasing initial frequencies (Figure S4).

A classical approach for selection scans based on more than two populations is to test pairs of populations. It is, for instance, the only possible option for selection scans based on XP-EHH. To evaluate the interest of this pairwise strategy, we compared the detection power obtained by applying hapFLK on pairs of populations or on the four populations jointly and found that testing all pairs of populations is always less powerful (Figure S7). Since XP-EHH also has less detection power than hapFLK in the two-population scenario, we can expect that applying hapFLK using the four populations jointly will be much more efficient than applying XP-EHH on pairs of populations.

Robustness and power of hapFLK in complex demographic scenarios: The model underlying hapFLK is that of pure drift evolution, with constant population size in each branch of a population tree with no admixture. These assumptions are made (i) when estimating the population covariance matrix $\mathcal{F}$ and (ii) when assuming allele frequency differences (either SNP or haplotype) are due only to $\mathcal{F}$. We studied the robustness of hapFLK in presence of admixture or bottleneck events by examining separately their consequences on (i) the estimation of the $\mathcal{F}$ matrix and (ii) the distribution of the hapFLK statistic. For this, we simulated the evolution of two populations with a bottleneck in one of the populations,

Figure 2 Power of hapFLK, F_{ST} , and XP-EHH as a function of the initial frequency of the selected allele. The power is evaluated at a type I error level of 5%.

migration from one population to the other, or migrations between both populations (see *Methods* for details).

The estimation of the $\mathcal{F}$ matrix is slightly affected by demographic events (Figure S2). When one of the populations has experienced a severe bottleneck (reduction in size by a factor 10), the estimated branch length for this population is increased by 10%. In the presence of migrations between populations, the two branches remain of the same length but the Reynolds genetic distance between the two populations is smaller than it should be (5% smaller in the one way migration case and 10% smaller in the two-way migration case).

Using this information we were able to perform simulations under pure drift evolution or bottleneck/migration evolution that led to the same *estimated* $\mathcal{F}$ matrix. As hapFLK is conditioned on this estimate, this approach allows evaluation of the effect of demographic events on the statistic, while integrating out their effect on $\mathcal{F}$. We found that the distribution of hapFLK was not greatly affected by deviations from pure drift evolution, on par with XP-EHH (Figure S8). Overall, these results show that while the estimate of $\mathcal{F}$ can be affected by deviation from the evolution model, and therefore coefficients in $\mathcal{F}$ must not be interpreted too literally, the distribution of hapFLK conditioned on this estimate is robust. In addition, the power of hapFLK is only slightly reduced under migration scenarios and unchanged under a bottleneck scenario (Figure S9).

Application to the sheep Hapmap data set

To provide an insight into the advantages and issues of using hapFLK on real data, we provide an example of application to a subset of the data from the Sheep HapMap Project. In sheep populations, drift accumulates rapidly, due to their

Figure 3 Power of F_{ST} , FLK, hap F_{ST} , and hapFLK in the four-population scenario as a function of the type I error rate. The initial frequency of the selected allele is 1%.

small effective size, typically a few hundred individuals (International Sheep Genomics Consortium 2012). As little power is expected from analyses based on genetic differentiation if populations are too distant, we focused on a group of relatively closely related breeds from Northern European origin. Six populations are included in this group, whose population tree is shown in Figure 5 (top left).

The genome scan performed with FLK provides little evidence for any sweep in these data, with P -values of the order of 10^{-4} , a hardly convincing figure, seen only on chromosomes 2 and 14. This is in great contrast (Figure 4) to the genome scan with hapFLK, which identifies seven genome-wide significant regions (Table 1), consistent with the additional power provided by hapFLK on simulated data sets. For each of these regions, we identified the population(s) under selection by reestimating the local population trees and comparing it to the tree estimated from whole genome data (see *Methods* for more details).

Figure 5 shows local trees for the two largest signals, on chromosome 2 and 14 (local trees for the other significant regions are provided in Figure S10).

The most significant selection signature (region 1 in Table 1) corresponds to a 17-Mb region in chromosome 2. Selection occurred in the three Texel breeds, most likely acting on the myostatin gene *GDF-8*, which is located in the middle of the region. Texel sheep carries a mutation in this gene, which contributes to muscle hypertrophy (Clop *et al.* 2006), a strongly selected trait in these populations. Although the mutation was discovered in Belgian Texels, our results imply that it must be present in these other Texel populations. SNPs within this region are almost fixed in the three Texel populations (Figure 6),

Figure 4 Genome scan for selection in Northern European sheep using a haplotype-based (hapFLK, top) or single SNP (FLK, bottom) test. x-axis: position on the genome. y-axis: $-\log_{10}(P\text{-value})$.

indicating a hard sweep signal. However, even in this “easy” case, using haplotype information makes the detection signal more interpretable: while *FLK* exhibits only moderate *P*-value decrease in the region, from which no clear conclusion concerning the selected site position can be drawn, hapFLK provides a continuous and strong signal covering the whole region and almost centered on the selected site. The local tree exhibits a large increase in branch length in the branch ancestral to the three Texel populations and reduced branch length between Texel populations (Figure 5). This is consistent with a shared selection event predating the split between populations. Finally, the example of region 1 also illustrates that our test can detect selection signatures that are shared by several populations, which we did not formally test in the simulations. In contrast, to detect this region with a F_{ST} genome scan, based on single SNP tests, International Sheep Genomics Consortium (2012) had to group the Texel breeds and test them against all other populations.

In contrast to the selection signature around *GDF-8*, the second most significant region (region 5, on chromosome 14) shows no evidence of a hard sweep (Figure 4) and cannot be identified using the single marker FLK test. The local tree (Figure 5) computed using SNP data exhibits slightly increased branch lengths, whereas the local tree computed using haplotype clusters presents very strong evidence for selection in two breeds: the New Zealand Texel and the New Zealand Romney, together with reduced haplotype diversity (Figure S14). These two breeds are not historically closely related (Figure 5, top left), but both have been imported in New Zealand (in 1843 and 1991, respectively). The selection signature could thus be due to a common recent

selection pressure on the two breeds in the last decades. This would be consistent with the relatively modest frequency of the selected clusters and the fact that these selected clusters are different in the two breeds, suggesting that selection started on different haplotype backgrounds. One possible underlying trait associated with this selection signal is resistance to nematode-like parasites, an important disease affecting sheep in New Zealand. Two studies (Hacariz *et al.* 2009; Matika *et al.* 2011) found evidence for association between genetic polymorphism and parasite resistance related traits in this region of the genome in Texel breeds. Matika *et al.* (2011) also found these polymorphisms associated with muscle depth. While the functional basis of these two effects is still unclear (pleiotropy, linkage disequilibrium with growth factors), it is possible that animal fitness in this region is related to multilocus haplotypes rather than to single SNPs.

We point the reader interested in details for all significant regions in Table 1 to the supporting information. In particular, allele and haplotype cluster frequencies are provided in Figure S11, Figure S12, Figure S13, Figure S14, Figure S15, and Figure S16 and local trees in Figure S10. An alternative approach for pinpointing the selected population(s) is also described (section 1.2, File S1) and applied to these regions (section 2.3, File S1; Figure S17, Figure S18, Figure S19, Figure S10, Figure S20, and Figure S21).

Discussion

Haplotype vs. single marker differentiation tests

For the analysis of dense genotyping data, where the selected site itself is generally not observed, we show that using haplotypes

Table 1 Selective sweeps detected by hapFLK within sheep populations from Northern Europe

Region	Chr	Position (Mb)	Max (Mb)	P-value	Population(s) (freqs)	Candidate genes
1	2	108.7–126.3	116.9	1.5×10^{-13}	STX (0.85), NTX (0.87), GTX (0.63)	GDF8
2	6	91.2–91.3	91.2	9.8×10^{-4}	ROM (0.36, 0.32)	
3	11	12.6–14.0	13.7	4.2×10^{-4}	ROM (0.75), GAL (0.45)	
4	14	12.2–14.6	13.9	4.5×10^{-4}	ISF (0.65)	
5	14	40.1–55.0	48.8	8.8×10^{-8}	ROM (0.29, 0.44), NTX (0.54)	TFGB1, IRF3
6	22	19.1–24.0	21.7	5.5×10^{-5}	GTX (0.62)	
7	22	38.5–38.8	38.6	8.6×10^{-4}	ROM (0.31, 0.35)	

For each significant region are listed: the chromosome region (in megabases on assembly OAR v. 2.0), the position of the maximum value for the statistic, the corresponding P-value, the suspected selected population(s) along with selected haplotypes frequencies, and potential candidate genes.

rather than single SNPs greatly improves the detection power of selection signatures. An intermediate approach between single SNP and haplotype-based tests consists in gathering multiple consecutive loci within sliding windows into a single *windowed* statistic (e.g., Browning and Weir 2010; Weir *et al.* 2005). However, in our simulation study, we found this approach to be less powerful at detecting selection (Figure S23). In the case of sequencing data, we found that a single SNP test was more powerful than hapFLK, consistent with previous results of Innan and Kim (2008), who found in a similar setting that an haplotype-based F_{ST} was less powerful than a single locus one. However, our simulations involved a single selected site and in many real situations, selection will act rather effectively on multilocus haplotypes (Pritchard *et al.* 2010), due, for instance, to recurrent mutations affecting the same gene, or to polygenic selection. We expect haplotype-based tests to be more powerful in such situations, which according to us justifies their use also for the analysis of sequencing data. In the particular case of low coverage resequencing, which is becoming a common experimental design in population genetics, this analysis will have to account for the additional uncertainty in genotype estimation, but we believe this can easily be tackled by the clustering algorithm used for hapFLK.

Different strategies for the inclusion of haplotype information in differentiation

To extend the single marker FLK and F_{ST} tests into haplotype based tests, we estimate local haplotype clusters from genotype data and consider these estimated clusters as alleles. Using a multipoint model for linkage disequilibrium (LD) in this context has several advantages. First, haplotypes are generally unknown and must be inferred from genotypes, which typically relies on a model for LD such as Scheet and Stephens (2006). Using directly the model parameters as we do has the advantage of allowing us to average over the uncertainty in the distribution of possible haplotypes rather than using a best guess that is known to include errors (Marchini *et al.* 2006). On a more practical side, hapFLK can be computed on unphased genotype data that are common in population genetics studies. Second, because the model of Scheet and Stephens (2006) is a hidden Markov model, it naturally accounts for variation in LD patterns along the chromosome and alleviates the need to use

windowing approaches, which have notorious difficulties accounting for this variation. Finally, several similar haplotypes may be associated to the same selected allele, and treating them independently should affect the detection power of the tests. In the Scheet and Stephens model, similar haplotypes are clustered together and will be considered as a single allele.

Other haplotype-clustering models, for instance, Beagle (Browning 2006), could certainly be used for constructing hapFLK. For example, the pattern of haplotype frequencies around the *LCT* gene in human populations was studied using either fastPHASE (Jakobsson *et al.* 2008) or Beagle (Browning and Weir 2010), and a strong evidence for selection in Europe was observed in both cases. However, to go beyond these observations and build a formal statistical test for selection, it is important to realize that the distribution of hapFLK (or $\text{hap}F_{ST}$) depends on the number of clusters used to model haplotype diversity. This number is fixed in fastPHASE but variable along the genome in Beagle. As this variation might be due to natural selection, but also to other effects such as variations in recombination or mutation rate, further studies would be required to evaluate the influence of using different clustering algorithms on the detection power.

Another important feature of hapFLK is its ability to account for the hierarchical structure of the sampled populations, arising from their evolutionary history within the species. FLK was already shown to be more powerful than the F_{ST} test in many simulated scenarios (Bonhomme *et al.* 2010). It was also compared to the Bayesian differentiation test of Foll and Gaggiotti (2008) in one simulated scenario with hierarchically structured populations and again provided more detection power. Consequently, we expect that hapFLK will also perform better than other haplotype-based differentiation tests (Guo *et al.* 2009; Browning and Weir 2010; Gompert and Buerkle 2011) for hierarchically structured populations.

To build tests that account for both the differentiation between populations and haplotype structure, all methods discussed above propose including haplotype information into single-marker differentiation tests. Another popular strategy, developed in the XP-EHH (Sabeti *et al.* 2007) and Rsb (Tang *et al.* 2007) statistics, is to compute a statistic quantifying the excess of long haplotypes within each population and to contrast this statistic among pairs of populations.

Figure 5 Local population trees estimated in two significant regions in the sheep data set. Population tree of the Northern European sheep populations from the Sheep HapMap Project (top left). Local population trees were estimated using Reynolds distance based on SNPs (left) or haplotype clusters (right). Abbreviations: Irish Suffolk (ISF), German Texel (GTX), New Zealand Texel (NTX), Scottish Texel (STX), Galway (GAL), New Zealand Romney (ROM).

Simulating a two-population sample, we found that XP-EHH and hapFLK had relatively similar power for hard sweep detection. However, one important difference was that hapFLK maintained some power for soft sweep detection, in contrast to XP-EHH. Interestingly, the combination of XP-EHH, FLK, and hapFLK allows a slight increase of power in the two-population simulations (File S1, section 1.5 and Figure S22), indicating that these different statistics do not capture exactly the same patterns in the data.

When more than two populations are sampled, comparing only pairs of populations raises a multiple testing issue leading to a significant decrease in power (Figure S7). Besides, computing a single test at the meta-population level seems more appropriate for several reasons. First, the signals we detected in sheep suggest that favorable alleles are often positively selected in several populations, either closely (region 1) or distantly related (region 5). Second, our ability to detect loci under selection depends on our ability to estimate the allele frequencies in this common ancestral population, which is clearly improved when using all populations simultaneously. One potential difficulty arising from our meta-population approach is the identification of the population (s) under selection, which is more difficult than when comparing pairs of populations. We proposed addressing this question using a local reestimation of the population tree, as illustrated in the sheep Hapmap data analysis. An alternative approach, which is based on a spectral decomposition of

hapFLK, is also described in the supporting information and applied to the sheep data.

Robustness of hapFLK and computation of P-values in a general situation

In many genome scans for selection, all loci above a given empirical quantile of the test statistic are considered as potential targets of selection. However, this so-called outlier approach does not allow control of the false-positive rate and can be inefficient in many situations (Teshima *et al.* 2006). To overcome this limitation and quantify the statistical significance of selection signatures, one must describe the expected distribution of the test statistic under neutral evolution, which depends on the demographic history of the sampled populations. In the case of hapFLK, this neutral distribution could be estimated by (i) fitting the kinship matrix $\mathcal{F}$ from genome-wide SNP data and (ii) simulating neutral samples conditional on $\mathcal{F}$, using a simple model with no migration and constant population size along each branch of the population tree. This approach avoids estimation of a full demographic model for the sampled populations and was found to be robust to bottlenecks or to intermediate levels of migration/admixture (Figure S8). For the analysis of samples involving stronger departures from the hierarchical population model assumed in this study (for instance, with hybrid populations), the expected covariance matrix of allele frequencies could also be modeled using relaxed hypotheses.

Figure 6 Allele (left) and haplotype cluster (right) frequencies in detected region 1 (chromosome 2) for each of the 6 sheep populations used in the test. Blue bars indicate the limits of the detected region and the position of maximum of the test. The reference allele used for the SNP frequency representation is arbitrary.

The strategies used in Bayenv (Coop *et al.* 2010) or TreeMix (Pickrell and Pritchard 2012) could, for instance, be adapted to the application of hapFLK.

However, in many situations (*e.g.*, in the sheep HapMap data), the neutral distribution of hapFLK is not only affected by demography, but also by SNP ascertainment bias. Simulating the ascertainment process is in general difficult: for example, in the sheep data it involves animals from a large panel of worldwide populations (International Sheep Genomics Consortium 2012). For single SNP tests such as FLK, this ascertainment issue can be circumvented by estimating a neutral distribution for several bins of the allele frequency in the ancestral population (Bonhomme *et al.* 2010), because we can assume that the only effect of SNP ascertainment is to bias the allele frequency distribution. But this strategy is not applicable to haplotype-based tests, for which the effect of SNP ascertainment is more complex. We consequently proposed a more empirical approach, in which the null distribution of hapFLK is directly estimated from the data using an estimator that is robust to outlier values. This empirical approach might be useful in future genome scans for selection, even if they are based on different test statistics than hapFLK, but its validity will depend on each particular data set and needs to be checked carefully by looking at the *P*-value distribution (see *Methods* for more details).

The most significant selection signatures detected in sheep using hapFLK exhibit extremely small *P*-values (down to 10^{-13}), while the smallest *P*-values obtained with FLK for the same data set were of order 10^{-4} . This difference of magnitude might be artificially inflated by the fact that we compute hapFLK *P*-values using a normal distribution, and FLK *P*-values using a chi-square distribution. However, we note that the choice of these distributions is supported by the data. Besides, we found that FLK *P*-values in simulated samples with selection using a chi-square distribution can go down to at least 10^{-11} (data not shown). We thus believe that the *P*-value difference observed in sheep reflects the fact that hapFLK is much more powerful than FLK, especially for SNP data where ascertainment bias leads to remove SNPs with extreme allele frequencies.

Soft or incomplete sweeps

While genome scans for selection have historically focused on hard sweeps, several recent studies have pointed out the importance of soft sweeps in the evolution of populations (Pritchard *et al.* 2010; Hernandez *et al.* 2011) and described the genomic signature of these selection scenarios (Hermisson and Pennings 2005). We tested hapFLK for initial frequencies of the favorable allele up to 30% and found that reasonable power could be achieved also in this situation. The detection

of incomplete sweeps is another important issue, which has not been much tackled in the literature. Detecting selected alleles at intermediate frequency is almost impossible with methods based on the allele frequency spectrum and very difficult with EHH- or F_{ST} -based existing approaches. In contrast, hapFLK is quite powerful in the case of incomplete sweeps, and several of the selection signatures detected in the sheep HapMap data correspond to intermediate frequencies of the selected haplotype (see Figure S11, Figure S15, and Figure S16).

Few hard sweeps were actually detected in the sheep data, although they are easier to detect than soft sweeps. This might be due to the short divergence time between these populations (a few hundred generations), which would limit the rise in frequency of favorable alleles. On the other hand, artificial selection has been associated with strong selection intensities, especially in the last decades, which should compensate for the short evolution time. One alternative explanation could be the variation of the selection intensity or direction over time, due to changes in agronomical objectives (e.g., in the sheep from wool to meat production) or importations of animals in a new environment (e.g., in the Texel and Romney breeds from Europe to New Zealand). The small number of hard sweeps can also be explained by the fact that artificial selection on quantitative traits is in general polygenic.

As a final and general remark on all methods aiming at discovering positive selection, selective constraints in functional and nonfunctional regions are probably more complex than what is usually simulated (with purifying and background selection, polygenic selection, balancing selection, etc). Definitely more research effort needs to be done on these aspects.

Conclusions

Overall, our study demonstrates that using haplotype information in F_{ST} -based tests for selection greatly increases their detection power. Consistent with several recent other studies (Excoffier *et al.* 2009; Bonhomme *et al.* 2010; Coop *et al.* 2010), it also confirms the importance of analyzing multiple populations jointly, while accounting for the hierarchical structure of these populations. The new hapFLK statistic, which combines these two features, can detect a wide range of selection events, including soft sweeps, incomplete sweeps, sweeps occurring in several populations, and selection acting directly on haplotypes.

Acknowledgments

We thank Michael Blum and Lucia Spangenberg for useful comments on the manuscript and Carole Moreno for earlier access to the sheep HapMap data. We thank the reviewers for their helpful comments, in particular the suggestion to construct local population trees in regions under selection. The ovine SNP50 HapMap data set used for the analyses described was provided by the International Sheep Genomics Consortium (ISGC) and obtained from <http://www.sheepmap.org> in

agreement with the ISGC Terms of Access. The simulations and data analysis were performed on the computer cluster of the bioinformatics platform Toulouse Midi-Pyrenees.

Literature Cited

- Albrechtsen, A., I. Moltke, and R. Nielsen, 2010 Natural selection and the distribution of identity-by-descent in the human genome. *Genetics* 186: 295–308.
- Barreiro, L. B., G. Laval, H. Quach, E. Patin, and L. Quintana-Murci, 2008 Natural selection has driven population differentiation in modern humans. *Nat. Genet.* 40: 340–345.
- Beaumont, M. A., and D. J. Balding, 2004 Identifying adaptive genetic divergence among populations from genome scans. *Mol. Ecol.* 13: 969–980.
- Bigham, A., M. Bauchet, D. Pinto, X. Mao, J. M. Akey *et al.*, 2010 Identifying signatures of natural selection in Tibetan and Andean populations using dense genome scan data. *PLoS Genet.* 6: 9.
- Boitard, S., C. Schloetterer, and A. Futschik, 2009 Detecting selective sweeps: a new approach based on hidden markov models. *Genetics* 181: 1567–1578.
- Bonhomme, M., C. Chevalet, B. Servin, S. Boitard, J. Abdallah *et al.*, 2010 Detecting selection in population trees: the Lewontin and Krakauer test extended. *Genetics* 186: 241–262.
- Browning, S. R., 2006 Multilocus association mapping using variable-length Markov chains. *Am. J. Hum. Genet.* 78: 903–913.
- Browning, S., and B. Weir, 2010 Population structure with localized haplotype clusters. *Genetics* 185: 1337–1344.
- Cagliani, R., S. Riva, M. Fumagalli, M. Biasin, S. L. Caputo *et al.*, 2011 A positively selected APOBEC3H haplotype is associated with natural resistance to HIV-1 infection. *Evolution* 65(11): 3311–3322.
- Clop, A., F. Marcq, H. Takeda, D. Pirotin, X. Tordoir *et al.*, 2006 A mutation creating a potential illegitimate microRNA target site in the myostatin gene affects muscularity in sheep. *Nat. Genet.* 38: 813–818.
- Coop, G., D. Witonsky, A. D. Rienzo, and J. K. Pritchard, 2010 Using environmental correlations to identify loci underlying local adaptation. *Genetics* 185: 1411–1423.
- 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.
- Fumagalli, M., R. Cagliani, S. Riva, U. Pozzoli, M. Biasin *et al.*, 2010 Population genetics of IFIH1: ancient population structure, local selection, and implications for susceptibility to type 1 diabetes. *Mol. Biol. Evol.* 27(11): 2555–2566.
- Gautier, M., L. Flori, A. Riebler, F. Jaffrezic, D. Laloe *et al.*, 2009 A whole genome Bayesian scan for adaptive genetic divergence in West African cattle. *BMC Genomics* 10: 550.
- Gompert, Z., and C. A. Buerkle, 2011 A hierarchical bayesian model for next-generation population genomics. *Genetics* 187: 903–917.
- Guan, Y., and M. Stephens, 2008 Practical issues in imputation-based association mapping. *PLoS Genet.* 4(12): e1000279.
- Guo, F., D. K. Dey, and K. E. Holsinger, 2009 A Bayesian hierarchical model for analysis of single-nucleotide polymorphisms diversity in multilocus, multipopulation samples. *J. Am. Stat. Assoc.* 104(485): 142–154.
- Hacariz, O., G. Sayers, R. J. Flynn, A. Lejeune, and G. Mulcahy, 2009 IL-10 and TGF-beta1 are associated with variations in fluke burdens following experimental fasciolosis in sheep. *Parasite Immunol.* 31(10): 613–622.

- Hayes, B. J., A. J. Chamberlain, S. Maceachern, K. Savin, H. McPartlan *et al.*, 2009 A genome map of divergent artificial selection between *Bos taurus* dairy cattle and *Bos taurus* beef cattle. *Anim. Genet.* 40(2): 176–184.
- Hermisson, J., and P. S. Pennings, 2005 Soft sweeps: molecular population genetics of adaptation from standing genetic variation. *Genetics* 169: 2335–2352.
- Hernandez, R. D., J. L. Kelley, E. Elyashiv, S. C. Melton, A. Auton *et al.*, 2011 Classic selective sweeps were rare in recent human evolution. *Science* 331: 920–924.
- Hudson, R., 2002 Generating samples under the Wright–Fisher neutral model of genetic variation. *Bioinformatics* 18(2): 337–338.
- Innan, H., and Y. Kim, 2008 Detecting local adaptation using the joint sampling of polymorphism data in the parental and derived populations. *Genetics* 179: 1713–1720.
- Kijas, J. W., J. A. Lenstra, B. Hayes, S. Boitard, L. R. Porto Neto *et al.*, 2012 Genome-wide analysis of the world's sheep breeds reveals high levels of historic mixture and strong recent selection. *PLoS Biol* 10(2): e1001258. doi:10.1371/journal.pbio.1001258.
- Jakobsson, M., S. W. Scholz, P. Scheet, J. R. Gibbs, J. M. VanLiere *et al.*, 2008 Genotype, haplotype and copy-number variation in worldwide human populations. *Nature* 451(7181): 998–1003.
- Kim, Y., and R. Nielsen, 2004 Linkage disequilibrium as a signature of selective sweeps. *Genetics* 167: 1513–1524.
- Kim, Y., and W. Stephan, 2002 Detecting a local signature of genetic hitchhiking along a recombining chromosome. *Genetics* 160: 765–777.
- Lao, O., J. M. de Gruijter, K. van Duijn, A. Navarro, and M. Kayser, 2007 Signatures of positive selection in genes associated with human skin pigmentation as revealed from analyses of single nucleotide polymorphisms. *Ann. Hum. Genet.* 71(3): 354–369.
- 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.
- Marchini, J., D. Cutler, N. Patterson, M. Stephens, E. Eskin *et al.*, 2006 A comparison of phasing algorithms for trios and unrelated individuals. *Am. J. Hum. Genet.* 78(3): 437–450.
- Matika, O., R. Pong-Wong, J. A. Woolliams, and S. C. Bishop, 2011 Confirmation of two quantitative trait loci regions for nematode resistance in commercial British terminal sire breeds. *Animal* 5(08): 1149–1156.
- Myles, S., K. Tang, M. Somel, R. E. Green, J. Kelso *et al.*, 2008 Identification and analysis of genomic regions with large between-population differentiation in humans. *Ann. Hum. Genet.* 72: 99–110.
- Nielsen, R., L. Williamson, Y. Kim, M. Hubisz, A. Clark *et al.*, 2005 Genomic scans for selective sweeps using SNP data. *Genome Res.* 15: 1566–1575.
- Peng, B., and M. Kimmel, 2005 simupop: a forward-time population genetics simulation environment. *Bioinformatics* 21(18): 3686–3687.
- Pickrell, J. K., and J. K. Pritchard, 2012 Inference of population splits and mixtures from genome-wide allele frequency data. *PLoS Genet* 8(11): e1002967. Available at: <http://hdl.handle.net/10101/npre.2012.6956.1>.
- Price, A. L., N. A. Zaitlen, D. Reich, and N. Patterson, 2010 New approaches to population stratification in genome-wide association studies. *Nat. Rev. Genet.* 11(7): 459–463.
- Pritchard, J. K., J. K. Pickrell, and G. Coop, 2010 The genetics of human adaptation: hard sweeps, soft sweeps, and polygenic adaptation. *Curr. Biol.* 20(4): R208–R215.
- Rees, J. L., and R. M. Harding, 2012 Understanding the evolution of human pigmentation: recent contributions from population genetics. *J. Invest. Dermatol.* 132(3): 846–853.
- Reynolds, J., B. S. Weir, and C. C. Cockerham, 1983 Estimation of the coancestry coefficient: basis for a short-term genetic distance. *Genetics* 105: 767–779.
- Riebler, A., L. Held, and W. Stephan, 2008 Bayesian variable selection for detecting adaptive genomic differences among populations. *Genetics* 178: 1817–1829.
- Sabeti, P. C., D. E. Reich, J. M. Higgins, H. Z. P. Levine, D. J. Richter *et al.*, 2002 Detecting recent positive selection in the human genome from haplotype structure. *Nature* 419: 832–837.
- Sabeti, P. C., S. F. Schaffner, B. Fry, J. Lohmueller, P. Varilly *et al.*, 2006 Positive natural selection in the human lineage. *Science* 312(5780): 1614–1620.
- Sabeti, P. C., P. Varilly, B. Fry, J. Lohmueller, E. Hostetter *et al.*, 2007 Genome-wide detection and characterization of positive selection in human populations. *Nature* 449: 913–918.
- Scheet, P., and M. Stephens, 2006 A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *Am. J. Hum. Genet.* 78: 629–644.
- Simonson, T. S., Y. Yang, C. D. Huff, H. Yun, G. Qin *et al.*, 2010 Genetic evidence for high-altitude adaptation in Tibet. *Science* 329(5987): 72–75.
- Storey, J. D., and R. Tibshirani, 2003 Statistical significance for genomewide studies. *Proc. Natl. Acad. Sci. USA* 100(16): 9440–9445.
- Sturm, R. A., 2009 Molecular genetics of human pigmentation diversity. *Hum. Mol. Genet.* 18(R1): 9–17.
- Tang, K., K. R. Thornton, and M. Stoneking, 2007 A new approach for using genome scans to detect recent positive selection in the human genome. *PLoS Biol.* 5: e171.
- Teshima, K. M., G. Coop, and M. Przeworski, 2006 How reliable are empirical genomic scans for selective sweeps? *Genome Res.* 16: 702–712.
- Tishkoff, S. A., R. Varkonyi, N. Cahinhinan, S. Abbes, G. Argyropoulos *et al.*, 2001 Haplotype diversity and linkage disequilibrium at human G6PD: recent origin of alleles that confer malarial resistance. *Science* 293(5529): 455–462.
- Venables, W. N., and B. D. Ripley, 2002 *Modern Applied Statistics with S*, Ed. 4. Springer, New York.
- Voight, B. F., S. Kudaravalli, X. Wen, and J. K. Pritchard, 2006 A map of recent positive selection in the human genome. *PLoS Biol.* 4: e72.
- Weir, B. S., L. R. Cardon, A. D. Anderson, D. M. Nielsen, and W. G. Hill, 2005 Measures of human population structure show heterogeneity among genomic regions. *Genome Res.* 15(11): 1468–1476.
- Yi, X., Y. Liang, E. Huerta-Sanchez, X. Jin, Z. X. P. Cuo *et al.*, 2010 Sequencing of 50 human exomes reveals adaptation to high altitude. *Science* 329(5987): 75–78.

Communicating editor: J. Hermisson

GENETICS

Supporting Information

<http://www.genetics.org/lookup/suppl/doi:10.1534/genetics.112.147231/-/DC1>

Detecting Signatures of Selection Through Haplotype Differentiation Among Hierarchically Structured Populations

María Inés Fariello, Simon Boitard, Hugo Naya, Magali SanCristobal, and Bertrand Servin

File S1

1 Methods

1.1 Deriving the kinship matrix $\mathcal{F}$ in complex evolutionary scenarios

In the two population scenario of Figure 1 (main paper), the kinship matrix is

$$\mathcal{F} = \begin{pmatrix} 0.05 & 0 \\ 0 & 0.05 \end{pmatrix}$$

because populations are independent conditional on the ancestral population, and

$$\mathcal{F}_{1,1} = \mathcal{F}_{2,2} = 1 - \left(1 - \frac{1}{2N_0}\right)^t,$$

with $N_0 = 1000$ and $t = 100$. Below we show that the kinship matrix in scenarios with migrations also have the same kinship matrix. Derivations for the bottleneck scenario are similar (and actually easier).

Asymmetric migration: This scenario is equivalent to the constant size scenario until generation 50. Then 10% of individuals from population 1 migrate to population 2. After generation 51 we resize both populations in order to match the genetic drift of the constant size scenario. To achieve this objective, the new population sizes are: $N_1 = 1003$ and $N_2 = 851$.

Let f_t be the drift accumulated in a population at generation t , N_t the population size at generation t and m the percentage of new individuals in population 1 after the migration event. For populations 1 and 2 we have

$$f_{50} = 1 - \left(1 - \frac{1}{2000}\right)^{50}, \quad (1)$$

Population 1 loses 100 individuals from generation 49 to generation 50, so

$$\begin{aligned} f_{51} &= \frac{1}{2N_{50}} + \left(1 - \frac{1}{2N_{50}}\right) f_{50} \\ &= \frac{1}{1800} + \left(1 - \frac{1}{1800}\right) \left(1 - \left(1 - \frac{1}{2000}\right)^{50}\right) \end{aligned}$$

The accumulated drift at generation 100 can then be calculated as

$$f_{100} = 1 - \left(1 - \frac{1}{2N_{51}}\right)^{49} (1 - f_{51}) \quad (2)$$

Setting $f_{100} = 0.05$ leads to $N_{51}^{(1)} = 1002.273$. Population 2 gains 100 individuals from generation 49 to generation 50, so using $m = \frac{100}{1100}$ and f_{50} from equation 1 we have

$$\begin{aligned} f_{51} &= \left[\frac{1}{2N_{50}} + \left(1 - \frac{1}{2N_{50}}\right) f_{50} \right] [(1-m)^2 + m^2] + \overbrace{\frac{1}{4N_{50}}(1-m)m}^{\approx 0} \\ &= \left[\frac{1}{2200} + \left(1 - \frac{1}{2200}\right) \left(1 - \left(1 - \frac{1}{2000}\right)^{50}\right) \right] [(1-m)^2 + m^2] \end{aligned} \quad (3)$$

Substituting f_{51} (3) in equation 2 we get $N_{50}^{(2)} = 851$.

Symmetric migration: Equations in this case are the same, but $m = \frac{100}{1000}$ and $N_{50} = 2000$, so we obtain $N_{51} = 840$ for both populations.

1.2 Pinpointing the selected population using a spectral decomposition of $\mathcal{F}$

As an alternative to the strategy described in the manuscript for pinpointing the selected population, we also propose to decompose T_{FLK} as the sum of n contributions arising from the different eigen-vectors of the matrix $\mathcal{F}$. For simplicity, we describe these derivations in the two allele case, but the generalization to the multiallelic case is immediate.

As $\mathcal{F}$ is positive definite, it can be decomposed as $Q'DQ$, where Q is an orthogonal matrix and D is a diagonal matrix with the eigen values of $\mathcal{F}$ on the diagonal. Denoting $\tilde{p} = p - \hat{p}_0 \mathbf{1}_n$ the vector of allele frequency variations from the ancestral population, we can thus re-write T_{FLK} as follows

$$\begin{aligned} T_{FLK} &= \tilde{p}' \text{Var}(\hat{p})^{-1} \tilde{p} \\ &= \frac{1}{p_0(1-p_0)} \tilde{p}' Q' D^{-1/2} D^{-1/2} Q \tilde{p} \\ &= \frac{1}{p_0(1-p_0)} \sum_{i=1}^n u_i^2 \end{aligned}$$

with $u = D^{-\frac{1}{2}} Q \tilde{p}$. u_i is the contribution to the test of the i -th eigenvector. Consequently, when a peak of T_{FLK} is detected, plotting the u_i 's independently indicates which eigenvector(s) most contributed to this peak. If the i -th eigenvector is found to have an important contribution, the exact populations responsible for the peak can then be recovered using the i -th line of Q . Indeed, $Q_{i,j}$ provides the loading of population j in eigenvector i , so large absolute values of $Q_{i,j}$ pinpoint the populations represented by eigenvector i .

Note that important contributions to T_{FLK} may well be found for populations without selected site in the region, for at least two reasons. First, eigenvectors often have several large loadings. This only depends on the population tree structure, which is reflected by $\mathcal{F}$, not on the population which is actually under selection in a given region. Second, as p_0 is estimated as a weighted average of the population frequencies, a very large frequency change in a single population may be interpreted as a moderately large frequency change in several populations.

Nevertheless, the population(s) under selection must be among the ones with a large contribution to T_{FLK} , so the decomposition described above generally permits to reduce considerably the set of potentially selected populations. Allele / haplotype frequency plots in each population can then be used to further dissect the observed signal.

1.3 Estimation of local trees

Here we describe how to re-estimate branch length of the population tree within a region of the genome. First note that, on the whole genome tree, the Reynolds distance between two populations i and j (d_{ij}) is :

$$d_{ij} = \sum_b x_{ij}(b) \beta_b + e_{ij} \quad (4)$$

where the sum is taken over all branches (indexed by b) in the tree and, $x_{ij}(b)$ equals 1 if the branch is in the path connecting population i and population j and 0 otherwise, β_b is the length of branch b and e_{ij} is a residual. Given the neighbor-joining tree, the x_{ij} elements are obtained using the `DesignTree` function of the `phangorn` R package (<http://cran.r-project.org/web/packages/phangorn/index.html>).

$\mu_{ij} = \sum_b x_{ij}(b)\beta_b$ is the expected length of the path between population i and j on the whole genome population tree. In practice $e_{ij} \approx 0$ so $\mu_{ij} \approx d_{ij}$. To measure deviation in branch length on a local tree we proceed by modeling the local Reynolds genetics distances δ_{ij} (computed using SNPs, or haplotype clusters, located in the significant region) as:

$$\delta_{ij} = \mu_{ij} + \sum_b x_{ij}(b)\beta'_b + e_{ij} \quad (5)$$

In this model, β'_b measures the difference in length between the whole genome tree and the local tree for branch b , *i.e.* if $\beta'_b > 0$ (resp. < 0) branch b is longer (resp. shorter) locally than genome wide. Model 5 is fitted using ordinary least squares, providing a p-value for the null hypothesis $\beta'_b = 0$. In rare situations we observed $\hat{\beta}_b + \hat{\beta}'_b < 0$, in which case the local branch length was set to 0 for the tree representation. In all these situations $\hat{\beta}_b + \hat{\beta}'_b$ was in fact very small.

Reynolds' distance: If $Q = (Q_{im})$, with $i \in \{1, \dots, n\}$ and $j \in \{1, \dots, L \cdot A\}$ is the matrix that contains all the allele frequencies for all populations. If $R = QQ'$, then the Reynolds' distance between populations i and j can be computed as:

$$D_{ij} = \frac{R_{ii} + R_{jj} - 2R_{ij}}{2(n \cdot A - R_{ij})}$$

In the SNP case $A = 2$ and in the haplotype case we consider as alleles the clusters from all the EM runs, so $A = K \cdot E$, where E is the number of EM runs. For example, if we have 4 clusters, and 15 EM runs, we will consider that there are 60 alleles at each locus.

For computing d_{ij} we include all loci available, both in haplotype and SNP cases. For computing δ_{ij} we used all the loci within the detected region for both cases.

1.4 Haplotype Cluster representation:

Plotting the cluster frequencies in the selected regions involves two preliminary steps. First, as the statistic is an average over different runs of the same algorithm, we identify a run where the statistic profile looks similar to the average. Second, cluster labels at each SNP are corrected. Indeed, cluster labels obtained from fastPHASE can sometimes switch from one SNP to the next one, which leads to an unnecessary switch in colors in the representation. This effect is similar to the label switching problem occurring across several fit of mixture models. In population genetics, a well-known case is that of models implemented in the STRUCTURE software (Pritchard et al., 2000), to the extent that another software, CLUMPP (Jakobsson and Rosenberg, 2007) was developed to reorder consistently labels across runs. To switch labels from one SNP to another, we use the greedy algorithm of CLUMPP. The cluster representations presented in this paper are similar to the representation that can be done using the haploscope software (San Lucas et al., 2011), except that we implement the label switching procedure. We provide the R code for performing these manipulations at the following address: <https://forge-dga.jouy.inra.fr/projects/hapflk>.

1.5 Combined test using hapFLK, FLK and XP-EHH

Here, we describe how to assess the power of a strategy combining multiple tests (here hapFLK, FLK and XP-EHH). Given a set of individual tests S_t^{max} for each statistic t in a simulated region, when combining the tests, we declare "significant" the region if any of the tests exceeds the individual type I error threshold of α . This leads to estimate the power as : $1 - \beta_c = \frac{TP_c}{N_s^1}$, where N_s^1 is the number of

simulations under the alternative and

$$TP_c = \# \left(\bigcup_t (S_t^{max} > q_{\alpha,t}) \right).$$

with $q_{\alpha,t}$ the quantile for statistic t associated with a type I error of α .

This procedure does not control the type error rate at a level α for the combined test. In the simulations we can evaluate the effective type I error α_c associated with this strategy by applying the same procedure on neutral simulations and counting the number of rejected null hypotheses. We thus estimate α_c as $\alpha_c = \frac{FP_c}{N_s^0}$, where N_s^0 is the number of simulations under the null and

$$FP_c = \# \left(\bigcup_t (S_t^{max} > q_{\alpha,t}) \right).$$

Finally we can represent power of the combined test as a function of α_c and compare it to tests based on each individual statistic. The resulting curve is presented in Figure S22.

2 Results

2.1 Detection power of *hapFLK* for soft and incomplete sweeps

We simulated different initial frequencies of the selected allele, ranging from 1% (hard-sweep scenario) to 30% (selection on standing variation / soft sweep scenario). We also considered different final frequencies of the selected allele, varying from its initial frequency up to fixation, as well as 50 and 100 generations after fixation (Figure S5).

In this case, power of *hapFLK* as a function of the favorable allele frequency was computed as follows. Simulations under neutrality were run for 450 generations after the last population split, and the 99% empirical quantile of T_{hapFLK}^{max} was computed every 25 generations. For each replicate simulated under selection, the statistic was computed when the selected allele reached the frequency of 5, 10, ..., 100%, as well as 50 and 100 generations after fixation. For each of these time points, T_{hapFLK}^{max} was compared with the quantile corresponding to the number of generations spent since the last population split and power computed as explained in the main text. When the number of generations was not a multiple of 25, the 99% quantile was computed by linear interpolation, using the two closest quantiles.

In the hard-sweep scenario, the power was maximal for an initial allele frequency of 5% , not 1% and conditional on the initial frequencies the power was maximal when the selected allele was in high frequency but not fixed in the population. This happens due to the fact that detection power is related to the time needed to reach a given frequency of the selected allele, because during this time drift accumulates in neutral populations. More precisely, due to the trajectories followed by the selected allele's frequencies (Figure S5.B), the time spent to go from a 1% to a 5% frequency, or to a 90% frequency to fixation, is not compensated by the increase of the statistic which ends up in a small loss of power.

2.2 Testing populations jointly vs in pairs

We compared *hapFLK* computed on all four populations with *hapFLK* computed on pairs of populations. Testing populations in pairs induces a multiple testing issue, which we accounted for in two ways : (i) assuming only two tests were performed, one for each two-population sub-tree or (ii) assuming all possible pairs of populations were tested (6 tests). We corrected the power of the two-population tests by Bonferroni in consequence. We found that using all four populations at the same time was more powerful than either of the approaches based on pairs of populations (Figure S7). The lower power with correction

(ii) was expected, since Bonferroni correction ignores the correlation between population pairs, and is thus conservative. The lower power with correction (i) is more surprising, but likely comes from the fact that p_0 estimation is less accurate using pairs of populations than using all populations jointly.

We also note that in a situation where selection has taken place in two populations from the same subtree, we can expect the four-population approach to maintain some power, while the tests based on comparing the two subtrees should have no power, because the situation is consistent with the null hypothesis of no differentiation between the compared populations.

2.3 Pinpointing the selected population using a spectral decomposition of $\mathcal{F}$: application to the Sheep data

In region 1, two eigenvectors contributed significantly to the selection signal: one is associated to the three Texel breeds, and the other to the Galway breed (Figure S17). Given that the Galway breed is the closest neighbor of the Texel breeds in the population tree (main text Figure 3), it can be considered that selection occurred somewhere in the Galway-Texel sub-tree. In this particular case, a closer look at the allele and haplotype clusters frequencies (main text Figure 6) clearly indicates that the selected populations are the three Texel breeds.

In region 5, the spectral decomposition of *hapFLK* (Figure S20) indicates that populations under selection in this region could be Scottish Texel, New Zealand Texel or New Zealand Romney. However, only the New Zealand Texel and New Zealand Romney showed evidence for reduced haplotypic diversity (Figure S14). Note that the Galway breed, which had also a frequent haplotype cluster in the region, was not associated to any component presenting an inflated value of the statistic. This illustrates that the identification of the breeds under selection can not be based only on the pattern of haplotype cluster frequencies, and thus justifies the use of the spectral decomposition proposed above.

References

- Browning, S. and Weir, B. (2010). Population structure with localized haplotype clusters. *Genetics*, 185(4):1337–1344.
- Jakobsson, M. and Rosenberg, N. A. (2007). Clumpp: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. *Bioinformatics*, 23(14):1801–1806.
- Pritchard, J. K., Stephens, M., and Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, 155(2):945–59.
- San Lucas, F. A., Rosenberg, N. A., and Scheet, P. (2011). Haploscope: a tool for the graphical display of haplotype structure in populations. *Genet. Epidemiol.*

Figure S1. Power of hapFLK using B_0 covariance matrix (black lines) or not (blue lines) for different initial frequency of the selected allele.

Figure S2. Estimate of difference in branch length (up) and Reynolds genetic distances (bottom) between two populations under different evolutionary scenarios. For each scenario, results are shown for whole SNP data (seq), only common SNPs (MAF > 5%) and genotyping array density (array).

Figure S3. Histogram of the distribution of *hapFLK* and of the fitted normal distribution (A) and histogram of the resulting p-values (B). The dotted line on the left corresponds to the significance p-value threshold of 10^{-3} , corresponding to an FDR of approximately 5%.

Figure S4. Power of *hapFLK*, *FLK*, *hapFST* and *FST* as a function of type I error. Initial frequencies for the selected allele: A:1%, B:5%, C:10%, D:20%.

Figure S5. (A) Power of hapFLK (type I error rate 1 %) in the two-population scenario for different initial frequencies of the selected allele and different frequency of the selected allele. *Fix + n* marks a time point corresponding to *n* generations after fixation of the selected allele. (B) Evolution of the selected allele frequency through time for different initial values. Points are values obtained from simulations, lines are logistic curves fitted to the observed values.

Figure S6. Power of *hapFLK*, F_{ST} and $XP - EHH$ with sequence data as a function of the initial frequency of the selected allele. The power is evaluated at a type I error level of 5%.

Figure S7. Power of different testing strategies for multiple population data Power of the *hapFLK* test performed on 4 populations jointly (solid line) on 2 independent sub-trees (dotted line) and on all pairs of populations (dashed lines). See details in the text section 2.2.

Figure S8. QQ-plots of the distribution of *hapFLK* and *XP – EHH* under different evolutionary scenarios against pure drift evolution. Points lying between the two gray lines are consistent with the effect of no effect of complex evolution on the distribution of the statistic.

Figure S9. Power of hapFLK under different evolutionary scenarios. Top row: selection in population 1, bottom row: selection in population 2. Note that for the two senses migration selection in population 1 is equivalent to selection in population 2.

Figure S10. Local trees estimated for significant regions in the Sheep HapMap data analysis. Only regions not discussed in the main text are shown

Figure S11. Chromosome 6 (Region 2) Allele (left) and haplotype cluster frequencies (right) in detected region 2 for each of the 6 Sheep populations used in the test. Blue bars indicate the limits of the detected region and the position of maximum of the test.

Figure S12. Chromosome 11 (Region 3) Allele (left) and haplotype cluster frequencies (right) in detected region 3 for each of the 6 Sheep populations used in the test. Blue bars indicate the limits of the detected region and the position of maximum of the test.

Figure S13. Chromosome 14 (Region 4) Allele (left) and haplotype cluster frequencies (right) in detected region 4 for each of the 6 Sheep populations used in the test. Blue bars indicate the limits of the detected region and the position of maximum of the test.

Figure S14. Chromosome 14 (Region 5) Allele (left) and haplotype cluster frequencies (right) in detected region 4 for each of the 6 Sheep populations used in the test. Blue bars indicate the limits of the detected region and the position of maximum of the test.

Figure S15. Chromosome 22 (Region 6) Allele (left) and haplotype cluster frequencies (right) in detected region 6 for each of the 6 Sheep populations used in the test. Blue bars indicate the limits of the detected region and the position of maximum of the test.

Figure S16. Chromosome 22 (Region 7) Allele (left) and haplotype cluster frequencies (right) in detected region 7 for each of the 6 Sheep populations used in the test. Red bars indicate the limits of the detected region and the position of maximum of the test.

Figure S17. Chromosome 2 Spectral decomposition of *hapFLK* for detected region 1. First line: *FLK* p-values for each SNP in the chromosome 2. Second line: *hapFLK* p-values. Each of the following lines corresponds to an orthogonal component of the population kinship matrix: For each component, population loadings are shown on the left and the projection of the test is plotted on the right. Dotted blue bars indicate the limits of the candidate region.

Figure S18. Chromosome 6 Spectral decomposition of *hapFLK* for detected region 2. See caption of Figure S17 for details.

Figure S19. Chromosome 11 Spectral decomposition of *hapFLK* for detected region 3. See caption of Figure S17 for details.

Figure S20. Chromosome 14. Spectral decomposition of *hapFLK* for detected regions 4 (blue) and 5 (red). See caption of Figure S17 for details.

Figure S21. Chromosome 22. Spectral decomposition of *hapFLK* for detected regions 6 and 7. See caption of Figure S17 for details.

Figure S22. Power of *hapFLK*, F_{ST} and *XP-EHH* and the their combination as a function of type I error. Initial frequencies for the selected allele: A:1%, B:5%, C:10%, D:20%, E:30 %.

Figure S23. Detection Power of *hapFLK*, F_{ST} and windowed F_{ST} . Windowed F_{ST} was computed as proposed by (Browning and Weir, 2010) for windows of 5, 10 and 20 SNPs. The power is evaluated at a type I error level of 5%.