Detection of positive selection from multi population samples using dense genome wide data : new multipoint methods and application to farm animal species

Maria Inès M. I. Fariello Rico

Université de Toulouse

THÈSE

En vue de l'obtention du

DOCTORAT DE L'UNIVERSITÉ DE TOULOUSE

Délivré par :
Université Toulouse 3 Paul Sabatier (UT3 Paul Sabatier)

Cotutelle internationale avec :
Universidad de la República (UdelaR)

Présentée et soutenue par :
María Inés FARIELLO RICO
Le 26/09/2013

Titre :
Détection pan génomique de locus sous sélection en présence de données multi-populationnelles en marquage dense : nouvelles méthodes multipoint et applications aux espèces animales d'élevage

---

École doctorale et discipline ou spécialité :
ED SEVAB : Écologie, biodiversité et évolution

Unité de recherche :
UMR444 Laboratoire de Génétique Cellulaire

Directeur(s) de Thèse :
Magali SAN CRISTOBAL
Simon BOITARD
Hugo NAYA

Rapporteurs :

Frédéric AUSTERLITZ
Renaud VITALIS

Autre(s) membre(s) du jury :

Christophe THÉBAUD
Xavier ROGNON

To Rami

N A

p

p

Acknowledgements

There was a lot of people that made this thesis possible. It is impossible to quantify their importance, so I will thank them in a chronological order. The way this thesis began was when Enrique Lessa "took" the Probability and Statistics course with my class in 2003 and in the last lecture he talked about the coalescent theory. I didn't know that mathematics and genetics were so close together and that we could actually combine both of them into a field called population genetics. Dr. Lessa helped me on my whole trajectory from the very beginning, and he is still around. I hope he'll be around for many years more. Mario Wshebor and Jean-Marc Azas were fundamental pieces in this way, because they helped me to come to France and find a lab that would let me do my thesis there. Finally, I landed at the LGC. Three years and a half later, I can say that I was really really lucky. I would like to thank Magali, Simon, Bertrand and Hugo for being my directors. I really learned a lot from them. They taught me a lot of statistics, biology, but each contributed in a different way to the future researcher I hope I will be. With all of them, individually or in group, we had really interesting conversations and it is thanks to them that this work was finished in time. Magali showed me how to face the research life, how to find a compromise between work and personal life and how it is to be in the middle of two separate sciences: mathematics and biology. She warned me of what is coming and she gave me some advise of how to survive in the science world. Simon tried to teach me to write scientific papers. He made the first correction of every single page of all my work and every slide. Even if the pages that he returned to me were red everywhere, he always found encouraging words and made me feel as my work was not that bad. He also took me to a lot of conferences and explained

to me every point that I missed from the talks. Bertrand taught me how to do beautiful graphics and pictures, even if there are some that are not so pretty in this work, it could we worse. He improve my programming knowledge, and was always there to correct my errors. I have to thank him for not killing me many times. Hugo gave me a place in the bioinformatics unit in the Pasteur Institute in Uruguay and helped me a lot with the Uruguayan bureaucracy. He taught me how to fill with biological hypothesis my mathematical results. I will be eternally grateful to all the four of them. I hope that the work we began together with my thesis will continue in future collaborations because it was a real pleasure to work with all of them!! I really learned a lot from you. In the way, Jean-Marc Azas, Guillaume Laval and Michael Blum in France and Enrique Lessa and Fernando Alvarez Valin in Uruguay where part of my thesis committees. In thank all of them for the time that they dedicated to my work, the discussions I have with them where really useful and interesting. I hope one day I will return this time to them and that we will continue to have this conversations. Last but not least in this academical way, I would like to thank to the jury of the thesis. I will remember the day of my defense as one of the best days in my life, and part of it is because of the jury that I had. I felt that moment as a pleasant scientific conversation, and not as I was being jugged. This let me really enjoy the moment and having a nice closure. I would like to thank Martine, Philipe and Denis Milan for letting me work at the LGC and everybody there. I had a wonderful these three years and a half, and I will never will forget "le plus beau de tous les labos du monde". I specially thank Pitou for receiving me in her office, Laure for supporting me every day and Katia for making me finally understand how we get the ACGTs in our computers. I would like to thank also the bioinformatics group for making me a place each time I went to Uruguay. I ate the best asados with you every year. As the time I spent doing my PhD. was almost the same time that I spent in France, I have to thank Laure, Claire, Manu and Mathieu who made an effort to show me the French culture and made a lot of efforts for teaching me French. You really made me feel like home. I hope seeing you in my other home, in Uruguay. I thank also the INRA rugby flag team for trying to teach me this wonderful sport, a part of the toulousain

culture. I also want to thank all the latin girls group for their support during these four years in Toulouse, specially Vale and Fer who were there every day at home. It was a hard time for them being there during the thesis redaction time, that was specially tough (for them also). And finally all the people that from distance helped me to stand in Toulouse, and don't miss Uruguay that much. Lula was always there from London. Luca, we shared the same way and it was really nice to have someone to talk with that could really understand the situation. All my family, specially my parents, supporting my changes of humor from skype, as it always was in my hole career. And a special thanks to Rami, these four years were really really tough, but we passed through. Thanks for everything!!

Contents

Chapter 1

General Introduction

It is impossible to think about adaptation without thinking about Charles Darwin and its book *On the Origin of Species* from 1859. Darwin and Wallace pointed out natural selection as the main force responsible for the vast diversity of species. Between 1920 and 1930, Fisher, Haldane and Wright integrated the Mendelian rules of inheritance within the Darwin's theory of adaptation, while until there the two theories were thought as mutually exclusive. The basis of this integration was the mathematical modeling of genetic inheritance from one generation to the other, giving birth to population genetics. Understanding the mechanisms of adaptation, and disentangling adaptation forces from neutral ones, has been since then one of the major concerns of population genetics.

The domestication of wild animals and plants by Humans is an outstanding example of the role of selection in species evolution and diversification. The first domesticated species were dogs (Vila et al., 1997), resulting in one of the most easily noticeable example of phenotypic variation shaped by humans. Domestication with agricultural purposes began 10,000 years ago with cattle, sheep, goats and pigs (Andersson, 2012). Horses and chickens were domesticated around 3,000 years later, and rabbits about 1,500 years ago (Carneiro et al., 2011). Domestication changed animal phenotypes through natural and artificial selection. For example, in early times of domestication, animals that were able to survive and reproduce in captivity were indirectly

favored. Later, humans intentionally selected animals based on specific traits of agronomic interest, like meat, milk or wool production (Diamond, 2002). Most animal domesticated species have their origins in the Eurasian continent, being naturally adapted to a certain clime, but then they were exported all over the world and were challenged by new clime, seasons and diseases, so new adaptations in this direction were also required (Andersson, 2012, Diamond, 2002). Nowadays, domesticated animals still continue being selected, and in the specific case of farm animals, intense artificial selection is being applied in order to achieve optimal animal production. As a consequence of their domestication history, farm animals represent a great model for evolutionary biology. Indeed, in *On the Origin of Species*, Darwin (1859) pointed out the importance of domestic species to understand the mechanisms of adaptation: " *It is, therefore, of the highest importance to gain a clear insight into the means of modification and co-adaptation. At the commencement of my observations it seemed to me probable that a careful study of domesticated animals and of cultivated plants would offer the best chance of making out this obscure problem. Nor have I been disappointed; in this and in all other perplexing cases I have invariably found that our knowledge, imperfect though it be, of variation under domestication, afforded the best and safest clue. I may venture to express my conviction of the high value of such studies,although they have been very commonly neglected by naturalists. From these considerations, I shall devote the first chapter of this Abstract to Variation under Domestication.*"

The detection of selection signatures in farm animals is not only a theoretical challenge and a model for natural species, as it can also have substantial agronomic outcomes. In the last 50 years, the scientific management of farm animals based on quantitative genetics has resulted in a spectacular increase of productivity, and more recently a lot of genome wide scans have highlighted genomic regions of agronomic importance. However, these scans generally focus on one specific production trait. On the opposite, genome scans for selection have the potential to pinpoint functionally important regions of the genome, which may be related to a large variety of traits.

In the last years, different techniques became available as dense genome

wide genotyping and next generation sequencing. The production of large amounts of data was facilitated, giving the chance to access to the genotypic information from large samples of individuals from different populations and parts of the world. Thanks to this type of data, we can look for signatures of natural and artificial selection, being able to identify, for example, genes of domestication by comparing domesticated species with their wild counterparts. Many tests for detecting footprints of selection in the genome, which are based on different mathematical and genetical models, have been proposed recently. These tests are challenged by the novel genotyping and sequencing technologies, essentially because of two features of the data.

First, with the novel technologies we can easily obtain sequences of 50,000 observed variables per individual, and up to millions of variables in the case of genome wide sequencing. For each individual, the observed variables are genotypes composed of genetic markers that are organized in chromosomes (whose number depends on the species). As the number of markers that we can observe increases, the average distance between observed markers decreases and the information provided by each marker is no more independent from that of other markers. Some tests take advantage of the fact that alleles at consecutive markers are generally transmitted together from one generation to the other, focusing on combinations of alleles at several consecutive markers. These type of tests are called haplotypic tests. Other tests simply try to account for correlations between adjacent markers and to exploit also this information. Another issue that arises with the huge amount of observed data is the computational cost associated to the tests.

Second, we get information from several populations at the same time. To profit from this, several tests intend to detect signatures of selection using the information that arises from comparing populations. Indeed, one specific trait is generally selected in only a subset of the sampled populations. Thus, comparing the genetic diversity observed in selected population(s) and non selected one(s) should help to detect genomic regions associated to the evolution of this trait. Many of the tests exploiting this idea have been designed to compare only pairs of populations. As the amount of data is increasing, it is very common to have genomic information from more than two popula-

tions so tests considering more than two populations simultaneously should be developed. On one hand, including more information in the model should give a clearer picture of the evolution scenario and thus increase the power to detect selected regions. On the other hand, this avoids the multiple testing issues that arise when performing one genome scan per each population pair.

Extending cross population tests (comparing 2 populations) to multi-population tests is however not trivial. When different populations are sampled, some populations are more closely related than others, because they derive from a more recent common ancestor. This leads to a data set with hierarchical structure, where close related populations will contribute with more correlated data, while more distant populations will be almost independent.

The aim of this thesis is to develop statistical tests for the detection of recent selection signatures using dense genetic data collected from multiple populations, and to apply them to several data sets collected from farm animal species. Although the initial motivation of these tests is the detection of selection in farm animal species, which explains some of the assumptions that I will make, I believe they should be also useful in many other types of species, for instance in humans where very large data sets are already available and where the detection of selection signatures has received much interest in the last years. Two different tests will be proposed in order to cope with different types of data: individual genotypes or population allele frequencies. The test that takes individual information will naturally be more computationally demanding than the test that considers just population information.

The thesis is organized as follows. In Chapter 2 I introduce basic notations and mathematical models related to population genetics theory within an "ecological" scale of time as well as to the type of data that I will test. I also describe important aspects of the genomic signatures left by positive selection. In chapter 3 I shortly review the detection methods that, in my opinion, are related to the tests I developed myself. In particular I will describe a method that is fundamental to the further developed tests: the

$\mathcal{F}$-LK test. We can consider the new tests almost as extensions of this one. The following chapters are presented under an article form. In Chapter 4 a new haplotypic test for detecting selection, denoted hapFLK, is presented. It is illustrated using real data from sheep, which was extracted from the recently released Sheep HapMap dataset. In Chapter (5) I present a more extensive analysis of the Sheep HapMap data set, based on hapFLK and its single SNP equivalent FLK. Populations were included in this genome scan for selection based on a preliminary study of population structure in the Sheep HapMap data set, so details about this analysis are also provided. Finally in Chapter 6 I consider the situation where genetic information is not available at the individual level but at the population level. This situation occurs for instance when the DNA of all sampled animals is sequenced in a single pool, which provides genome wide information at a much lower cost than individual sequencing. hapFLK can not be applied in this case, but I present another test, based on local score theory, which also accounts for the correlation between loci. I apply this test to several datasets, in particular one dataset resulting from the pooled sequencing of two divergent lines of quail.

Chapter 2

Introduction

J.Felsentstein begins his notes on theoretical evolutionary genetics saying: "Theoretical population genetics is arguably the area of biology in which mathematics has been most successfully applied." In this chapter I will introduce some standard mathematical models describing how the genetic material of a population evolves along generations. These models are essential to my work, because all the methodological developments presented in the following chapters are based on them. In the first section, I will describe how genetic information arising from biological measurements can be represented in a mathematical framework and introduce basic definitions. In the second and third sections I will focus on models describing the neutral evolution of allele frequencies, at a single locus or at several correlated loci. In the last section I will describe the effect of selection, in particular of positive directional selection.

2.1 A mathematical framework for genomic data

2.1.1 Definitions

All the genetic material of an organism is packaged in its genome. A fixed position on the genome is called a *locus* (pl. loci), which is arbitrarily

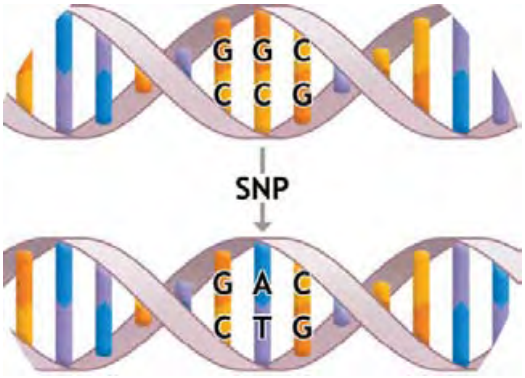


Figure 2.1: **SNP:** The alleles of this SNP are: G, the ancestral and A, the derived.

composed of one or several DNA nucleotides. If several variants of the DNA sequence exist at a locus, we call them *alleles*.

In the special case where the locus is a single nucleotide, we call it a *SNP*: Single Nucleotide Polymorphism (Figure 2.1 [1]). SNPs are currently the most commonly used genetic markers, because they are very common in the genome and because several recent technologies allow to measure them at a genome wide scale (see below).

Since the mutation rate per generation and per base pair is extremely low (approximately 10^{-8} for mammals), it is generally assumed that every single nucleotide in the genome can experience at most one mutation in the history of a species (this is called the infinite site model). As a consequence, SNPs are considered to have only two alleles in the genome: the ancestral allele, which existed before the mutation event, and the derived allele, which was created by the mutation. In practice, most SNPs have indeed only two alleles, and those with more than two alleles, being a really small subset of the total, can anyway be removed from the analysis.

SNPs represent 90% of the total variation in species Collins et al. (1998), but they are not the only source of variation in the genome. For instance, there are also microsatelites and copy number variations (CNVs). These type of markers evolve in a different way that SNPs do, for example they have different mutation rates, in general higher than nucleotide mutation rates, and the number of alleles is in general greater than 2. In CNVs the alleles

[1] This figure was taken from `http://www.ibbl.lu/personalised-medicine/what-is-personalised-medicine/dna-genes-snps`

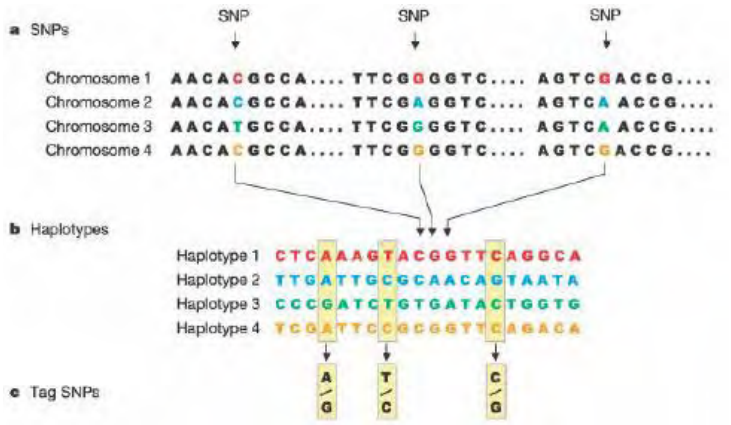


Figure 2.2: **Haplotypes: (a)** Aligned sequences. The SNPs of the set of chromosomes are highlighted. **(b)** Haplotypes containing all the SNPs present in the sample of the four chromosomes. **(c)** Depending on the technique that we use to extract the SNP information, we will recover a subset of the total sample of SNPs, named tag SNPs.

are the number of times that a sequence is repeated, from where comes its name. Thus, mathematical models differ from one type of marker to the other. This work focuses on SNP datasets.

We denote *haplotype* a combination of alleles at different loci carried on a same chromosome (Figure 2.2[2]). Diploid individuals as livestock animals or humans carry two copies of each chromosome, so they carry two haplotypes that form a *genotype*.

Knowing the genotypes of an individual at several loci, as provided by most available technologies (see below), does not imply that we know the two haplotypes. If haplotypes are known, we say that the data is phased. There are different ways to get phased data (both haplotypes) from unphased data (genotypes). I will briefly introduce some of the models and softwares available to do this further in this chapter (Section 2.3).

[2] This figure was taken from `http://www.brown.edu/Research/Istrail_Lab/proj_cmsh.php`

2.1.2 The datasets considered in this study

Different technologies are commonly used to produce genomic data from a sample of individuals. The most informative approach is *individual sequencing*, which provides the genotype of all sampled individuals for all the SNPs found in the species. With the recent advent of next generation sequencing technologies (see Mardis (2013) for a historical review), this experimental design has become realistic, for instance the genomes of 1000 humans from various populations or 1000 bulls from various breeds are currently being produced by international consortia (`http://www.1000genomes.org`, `http://www.1000bullgenomes.com`). However, this approach is still too expensive for lower scale projects, at least for "complex" species like mammals whose genome is very large.

One alternative is to focus on a smaller (but still large) set of SNPs distributed over the genome, and to obtain the genotype of each sampled individual at these SNPs by hybridization using a *genotyping array* (see `http://www.sheephapmap.org/genseq.php` for an example of how this can be done). This approach has been widely used in the last decade in human or animal genetics. The first dataset analyzed in this study (Chapters 4 and 5) has been obtained by this approach, analyzing about 3000 sheep with a chip of 50K SNPs in the context of the Sheep HapMap consortium (Kijas et al., 2012). A chip of higher resolution (700K) is currently being developed for sheep, and is already available in cattle. High density genotyping chips are also available in several other livestock species, for instance chicken, pig, horse among others.

A second cheaper alternative to individual sequencing is *pool sequencing*. In this approach, the DNA of sampled individuals is pooled before being sequenced. This provides allele frequencies at all SNPs in the genome, but individual information is lost. The second dataset analyzed in this study (Chapter 6) was obtained using this design. Two selectively divergent quail lines were considered, and one pooled sample of 10 quails was sequenced in each population.

For the detection of natural selection, we will see in the following chapters

that both approaches (individual data at a reduced set of SNPs obtained by genotyping, or population data at all SNPs obtained by pool sequencing) have their own advantages and drawbacks. Below we already illustrate in an example (2.1) one difference between genomic information obtained at the individual level and at the population level.

Example 2.1. *Let us suppose that we observe the genotypes for individuals at several biallelic loci, and that we denote A and a the two alleles that are present at one locus. As diploid individuals carry two alleles, individual genotypes can be AA, Aa or aa at this locus. We can recode the data by counting the number of A alleles that the individual carries: $aa \to 0$, $Aa \to 1$ and $AA \to 2$. Doing this at each locus each individual is represented by a sequence of 0s, 1s and 2s. If each individual sequence is a line of a matrix, we can do a Principal Components Analysis (PCA) and plot individuals in a two-dimensional plane to have an idea of the population structure. If instead, we had just the frequencies of each allele in each population (data coming from pool-sequencing), we could also build a matrix, but each line contains population information instead of individual information. A PCA can also be done with this matrix.*

In Figure 2.3 we see that at the population level (right) we can only have an idea of the genetic proximity between populations. For example, we observe that the Scottish and New Zealand Texel Breeds are the closest ones, and that the Irish Suffolk breed is the farthest away from the rest of the breeds.

At the individual level, we see that there are two Irish Suffolk individuals that are closer to other breeds than to the Irish Suffolk cluster. In fact, there is one Irish Suffolk individual that lies in the middle of the Texels cluster. This could come from a mislabeling, or signal a migrant individual from Scottish Texel to Irish Suffolk. We also see that in the Scottish Texel breed there are three distinct sub groups. The reason for that could be that the Scottish Texel samples come from three different rams, being a little bit genetically differentiated among them. We see also that the genetic diversity among German Texel individuals is larger than among other breeds.

As we see in this example, individual information can tell us a lot about the structure of the populations to whom individuals belong, for instance the

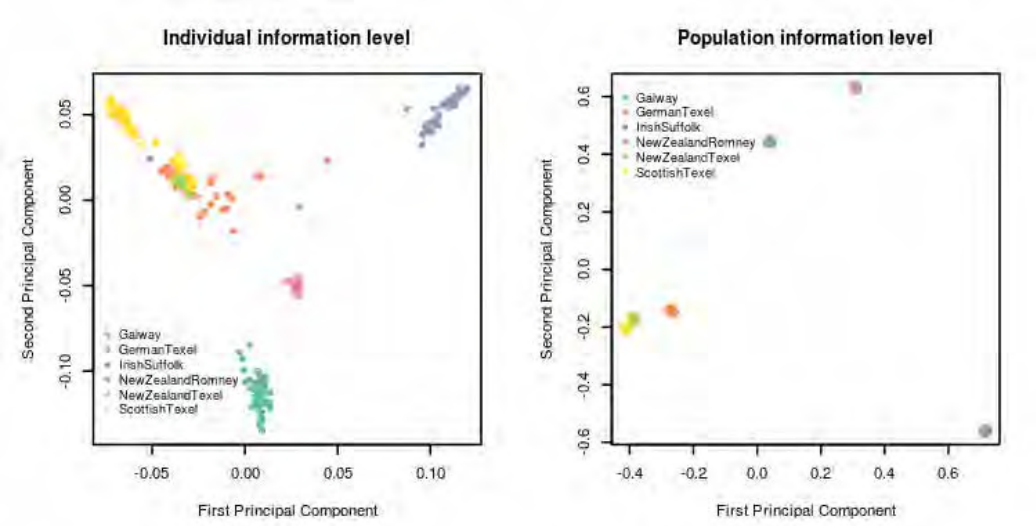


Figure 2.3: **First 2 principal components** of genetic data of sheep, considering individuals or populations.

genetic homogeneity of populations, genetic variation among populations and possible admixture (migrants from one population to another) and mislabeling. But, this is not the only aspect, we will see that when looking for signatures of selection, the type of data has a strong implication in the methods that we can use, and the type of signals that we can detect.

It is worth to note that, when the data is obtained by dense genotyping we have to be careful with the *ascertainment bias* INDEX. This is the bias produced in the allele frequency spectrum of the alleles, produced by the choice of the SNPs. In Figure 2.4[3] we see a simulation of the expected frequency spectrum depending on the amount of individuals used for detecting the SNPs in a sequence of DNA. To get more information about the uncommon SNPs we have to use more individuals. The bias can be also produced when using individuals from one continent to produce the chip and then genotyping individuals from other continent, because there can be some

[3] Figure taken from http://bioinformatics.bc.edu/ marth/BI820/pages/afsAnalysisComputerSession.html

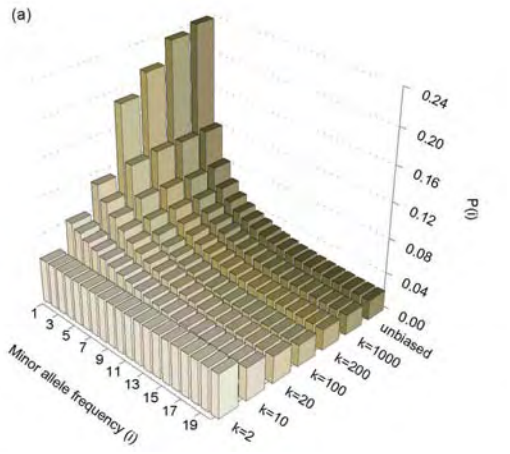


Figure 2.4: The expected folded spectrum under a simple, stationary history (constant effective population size), for a wide range of values of discovery sample size.

mutations that do not exist in one continent but have a high frequency in another.

In figure 2.1.2 we see a comparison between the site frequency spectra of Human genotypes depending on the consortium that produced the data. In the next chapters, we discuss the influence that the ascertainment bias can have when scanning genomes for selection.

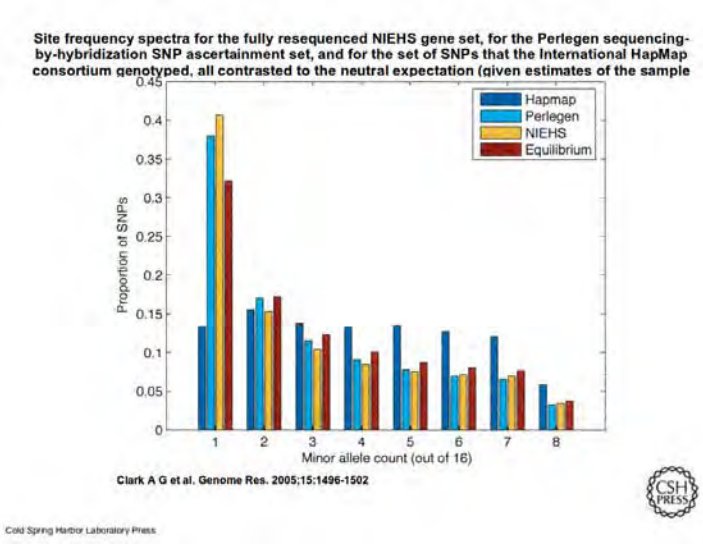


Figure 2.5: Ascertainment Bias

2.2 Modeling the neutral evolution of allele frequencies at a single locus under the "ecological" time scale

In this section I will introduce mathematical models that describe the evolution of allele frequencies along generations under neutral evolution. Although the focus on my work is on positive selection, describing neutral models is important, because the detection of loci under selection requires to know the distribution of genetic diversity that can be expected under neutrality. I will focus here on biallelic markers, first because the derivations are easier and thus more illustrative, second because the datasets I studied only included SNP data.

The models presented in this manuscript consider a population as a pool of genes that is carried by individuals and transmitted from one generation to the other. Individuals are assumed to produce an infinite quantity of gametes, and if an individual carries two different alleles, half of the gametes produced by this individual will be of each type. During the reproduction phase, two gametes are randomly and independently chosen from the population pool to produce a new diploid individual, which is called the random mating assumption.

Under these conditions, if the population size (i.e. the number of individuals) is infinite, with equal genotype frequencies in both sexes, no differential fertility or viability of the genotypes, no migration, no mutation and no selection, allele frequencies will remain constant along generations and genotype frequencies will be in *Hardy Weinberg Equilibrium*, i.e. for two alleles A and a with frequencies p and q, genotype frequencies will be p^2 for AA, $2pq$ for Aa and q^2 for aa. This principle is called the *Hardy-Weinberg Law*. Any departure of these assumptions will cause a change in the allele frequencies.

In the following sections, we will assume that there is no mutation, because we will work in a time scale, where there is a negligible probability that a new mutation arises, and if so, its frequency will be so low (under 1%) that we cannot distinguish it from a sequencing or genotyping error. Mutations

have of course occurred in the history of the species, otherwise there would be no SNPs at all, but we assume this was before the time period we consider. For simplicity, we also assume that the genotype frequencies in both sexes are equal, but this hypothesis could be relaxed, by conditioning on the allele that a child receives from each of his parents. In the model presented in the next section (2.2.1) the derivations are straightforward. But, if the allele frequencies among sexes are different, when considering more than one population, more than one locus at a time and/or selection, the derivations can be really complicated.

I will discuss here the situation where the infinite population size assumption is relaxed, either without or with migration. Differential fertility or viability, and in particular positive selection, will be discussed in the Section 2.4.

2.2.1 Genetic drift in a single population

The first assumption from the Hardy-Weinberg Equilibrium that does not hold when modeling real populations, is the infinite population size. Sampling a finite number of alleles ($2N$, where N is the population size) at each generation will affect the allele frequencies randomly, not remaining constant anymore. Sewall Wright (1931) and Ronald A. Fisher (1930) modeled the stochastic fluctuation of allele frequencies through generations. This process is called *genetic drift*, and their model (described below) is known as the *Wright-Fisher model* .

Let us assume that generations do not overlap and that the population has constant size N through generations. If we focus on a SNP that has alleles A and a, let $X(t)$ be the number of copies of the A allele at generation t and $p(t) = \frac{X(t)}{2N}$ the frequency of this allele in the population. To build the generation $t+1$ we will sample alleles with replacement from generation t, so the distribution of $X(t+1)$ conditional on $X(t)$ is a binomial distribution $\mathcal{B}(2N, p(t))$. For $i, j \in \{0, 1, \ldots, 2N\}$, the probability of having j copies of alleles A at generation $t+1$ given that there were i copies at generation t is:

$$p_{ij} = \mathbb{P}\left(X(t+1) = j | X(t) = i\right) = \binom{2N}{j} \left(\frac{i}{2N}\right)^j \left(1 - \frac{i}{2N}\right)^{(2N-j)} \quad (2.1)$$

Conditional on $X(t)$, $X(t+1)$ is independent on the value of $X(u)$ for earlier generations ($u < t$). Thus, the process $\{X(t)\}_{t\in\mathbb{N}}$ is a Markov Chain with probability transition matrix $P = (p_{ij})_{i,j=0,\ldots,2N}$ and initial state $X(0)$, where p_{ij} are given by Equation 2.1. This implies that :

$$\begin{aligned} \mathbb{E}(X(t+1)|X(t)) &= 2Np(t) = X(t) \\ Var(X(t+1)|X(t)) &= 2Np(t)(1-p(t)) \end{aligned}$$

For $t = 0$:

$$\begin{aligned} \mathbb{E}(p(1)) = &\quad \mathbb{E}(\mathbb{E}(p(1)|p_0)) = \tfrac{1}{2N}\mathbb{E}(2Np_0) &&= p_0 \\ Var(p(1)) = &\quad \mathbb{E}(Var(p(1)|p_0)) + Var(\mathbb{E}(p(1)|p_0)) &&= \frac{1}{2N}p_0(1-p_0) \end{aligned}$$

To get the variance and mean of $p(t+1)$ we have to condition and de-condition on $p(t)$, thus $\mathbb{E}(p(t+1)) = \mathbb{E}\left(\mathbb{E}(p(t+1)|p(t)) = \mathbb{E}(p(t)) = p_0\right.$ and

$$Var(p(t+1)) = \mathbb{E}(Var(p(t+1)|p(t))) + Var(\mathbb{E}(p(t+1)|p(t)))$$

If N remains constant, for each t we have:

$$p_0(1-p_0) - Var(p(t+1)) = \left(1 - \frac{1}{2N}\right)\left[p_0(1-p_0) - Var(p(t))\right]$$

By recursion after t generations we obtain:

$$\mathbb{E}(p(t)) = p_0 \tag{2.2}$$

$$Var(p(t)) = \left[1 - \left(1 - \tfrac{1}{2N}\right)^t\right] p_0(1 - p_0) \tag{2.3}$$

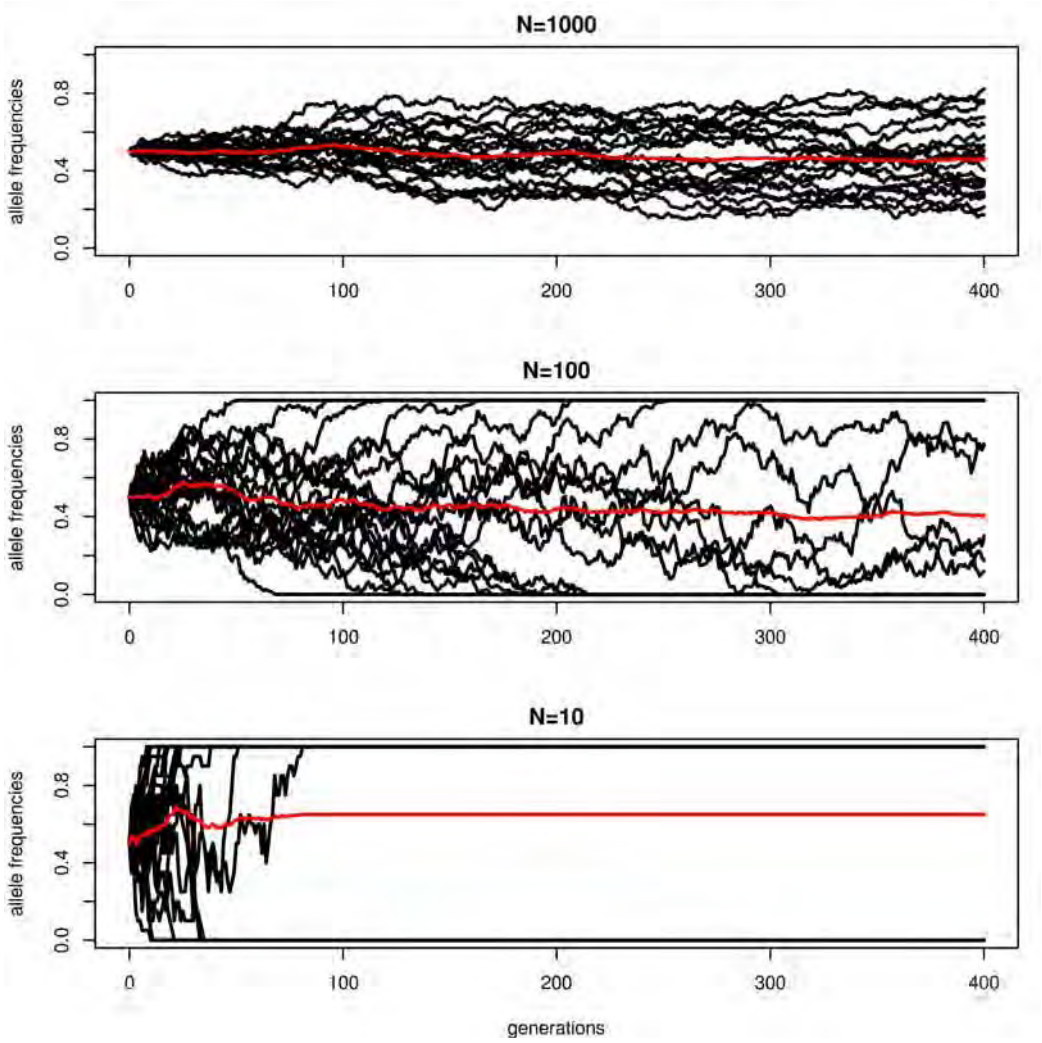


Figure 2.6: **Genetic drift:** effect of the population size N on the evolution of allele frequencies. 15 random trajectories are shown in black. The average of these trajectories is in red.

As t increases, the expected value of $p(t)$ remains constant but the vari-

ance of $p(t)$ increases. There is a inverse relation between the variance and the population size: the smaller is N, the larger is the variance, so the fluctuation of the frequencies is larger for smaller populations (Figure 2.6). One of the direct consequences of a small population size, is that alleles get fixed very fast, so it is easier to loose genetic variation in this cases.

Whatever the population size, eventually one of the two alleles will disappear and the other will get fixed. This means that for some sufficiently large t we will have $p(t) = 0$ or $p(t) = 1$, which are the absorbing states of the Markov Chain. Once an allele gets fixed or disappears, genetic drift cannot change its frequency anymore. The only sources of new variation at this locus are thus mutation (but we do not consider it here) or migration from another population.

The variance of $p(t)$ is related to the notion of identity by descent. We say that two identical alleles are *Identical By Descent* (IBD) if they descend from the same ancestral allele at generation 0. The probability of sampling two IBD alleles at generation t is called the *inbreeding coefficient* of a population. Under the previous assumptions (see Appendix A.1), this coefficient is equal to

$$F_t = \left[1 - \left(1 - \frac{1}{2N}\right)^t\right]$$

Thus, the expected variance of the allele frequency at time t can be written as

$$Var(p(t)) = F_t p_0(1 - p_0) \tag{2.4}$$

Actually, the fact that the variance at time t only depends on the initial frequency p_0 and on the inbreeding coefficient from time 0 to time t holds for much more general assumptions than those described here, and will be used in the next subsection. Then, conditional to the ancestral population, $p(t)$ can be modeled by a normal distribution with the computed mean and variance (Nicholson et al., 2002) (Figure 2.7):

$$p(t) \sim \mathcal{N}\left(p_0, F_t p_0 (1 - p_0)\right)$$

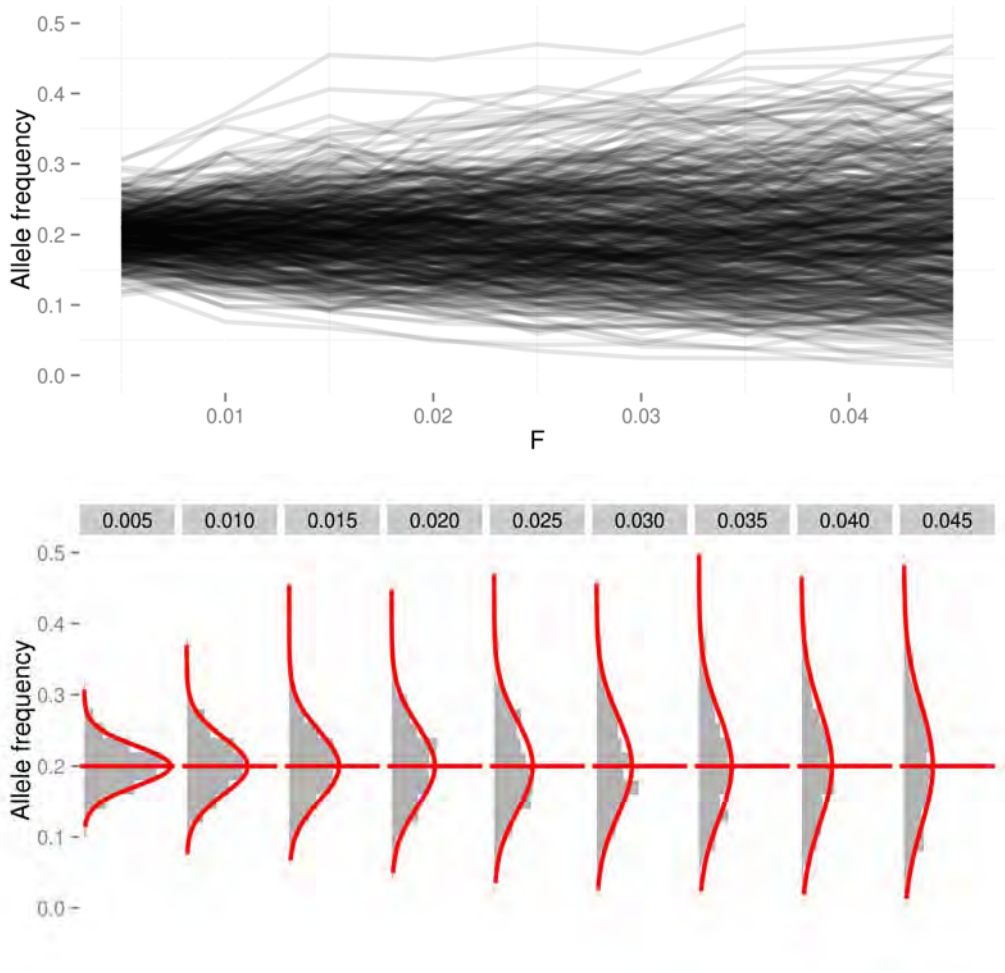


Figure 2.7: Top: effect of F_t on the variation of population allele frequencies ($p_0 = 0.2$). Bottom: Histogram of the allele frequencies for different F_t and estimated normal distribution conditional to p_0 and F_t (in red).

2.2.2 Genetic drift in several populations

In the previous subsection we considered one single population that descended from an ancestral population. We now consider the situation of n populations descending from the same ancestral population. This situation is central to my work, since my aim is to analyze genomic data sampled in several populations. In this model, populations can experience variations in population size, as bottlenecks or expansions, and can evolve independently or have a common evolution up to some point where they split. For the moment, we still assume that they do not exchange migrants.

Below I show that, under this model, the covariance matrix of the allele frequencies in the final populations can be expressed using a *kinship matrix*, which summarizes the history of the populations and is related to the inbreeding coefficients. To make derivations more illustrative, I consider an easy situation with only 3 populations, which is described in Figure 2.8

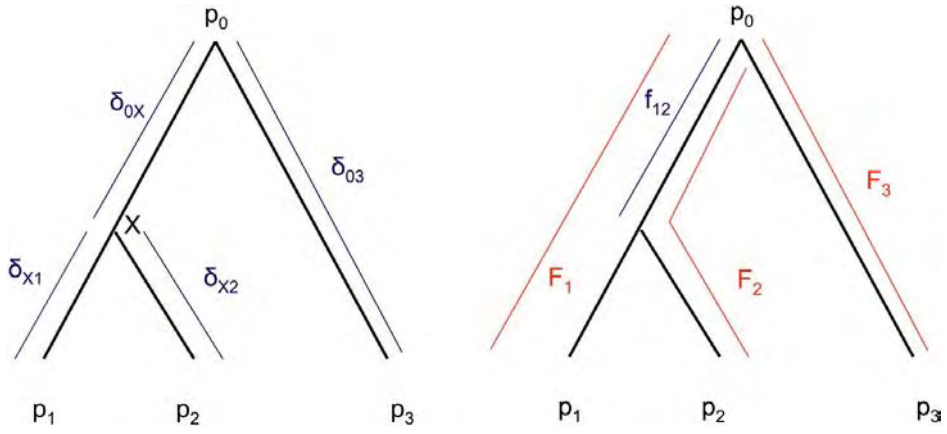


Figure 2.8: **Example of tree-like evolution**: construction of the kinship matrix

Consistent with Bonhomme et al. (2010) we denote:

p_i**:** the final allele frequency in population i (a leaf of the tree) with $i \in \{1, \ldots, n\}$

δ_{UV}**:** the variation of the inbreeding coefficient corresponding to the branch from U (an internal node or the root of the tree) to V (an internal node or a leaf of the tree)

f_{ij}**:** kinship coefficient between populations i and j

For $i = 3$, as already shown in the previous subsection (Equations 2.3, 2.4), we have:

$$\begin{aligned} \mathbb{E}(p_3) = &\ p_0 \\ Var(p_3) = &\ \delta_{03} \cdot p_0(1-p_0) \end{aligned}$$

For $i = 1, 2$ we have similarly:

$$\begin{aligned} \mathbb{E}(p_i|p_X) = &\ p_X \\ Var(p_i|p_X) = &\ \delta_{Xi} \cdot p_X(1-p_X) \end{aligned}$$

Deconditioning on p_X leads to:

$$\mathbb{E}(p_i) = p_0$$

and

$$\begin{aligned} Var(p_i) = &\ Var(\mathbb{E}(p_i|p_X)) + \mathbb{E}(Var(p_i|p_X)) \\ = &\ Var(p_X) + \mathbb{E}(\delta_{Xi} \cdot p_X(1-p_X)) \\ = &\ Var(p_X) + \delta_{Xi}(\mathbb{E}(p_X) - \mathbb{E}^2(p_X) - Var(p_X)) \\ = &\ \delta_{X0} \cdot p_0(1-p_0) + \delta_{Xi}[p_0(1-p_0) - \delta_{0X}p_0(1-p_0)] \\ = &\ p_0(1-p_0)[1-(1-\delta_{0X})(1-\delta_{Xi})] \end{aligned}$$

Besides, we have $cov(p_1, p_2|p_X) = 0$, because conditional on p_X (*i.e.* after splitting in two different populations), the two populations evolve independently, so:

$$\begin{aligned} cov(p_1, p_2) = &\ cov(\mathbb{E}(p_1|p_X), \mathbb{E}(p_2|p_X)) + \mathbb{E}(cov(p_1, p_2|p_X)) \\ = &\ cov(p_X, p_X) = Var(p_X) \\ = &\ \delta_{0X}p_0(1-p_0) \end{aligned}$$

In summary, we can write:

$$Cov(p_i, p_j) = f_{ij} p_0 (1 - p_0) \tag{2.5}$$

$$Var(p_i) = f_{ii} p_0 (1 - p_0) \tag{2.6}$$

where the f_{ij}s are given by:

$$\begin{aligned} f_{11} &= 1 - (1 - \delta_{X1})(1 - \delta_{0X}) \\ f_{22} &= 1 - (1 - \delta_{X2})(1 - \delta_{0X}) \\ f_{33} &= \delta_{03} \\ f_{12} &= \delta_{0X} \\ f_{13} &= 0 \\ f_{23} &= 0 \end{aligned} \tag{2.7}$$

The covariance matrix of $p = (p_1, p_2, p_3)$ is thus given by $\mathcal{F} p_0 (1 - p_0)$, where $\mathcal{F}$ is

$$\mathcal{F} = \begin{pmatrix} f_{11} & f_{12} & f_{13} \\ f_{12} & f_{22} & f_{23} \\ f_{13} & f_{23} & f_{33} \end{pmatrix} \tag{2.8}$$

Note that f_{ii} corresponds to the inbreeding coefficient of population i (2.3). We will denote it F_i in the rest of this manuscript.

Nicholson et al. (2002) say that F_i (written as c_j in their article) might be thought of as analogous to F_{ST} values but with one for each population. In the case that all F_i are identical, they are exactly equal to F_{ST} for the entire group of populations. F_{ST} is defined in Section 3.1.

2.2.3 Models with admixture

Different scenarios including migration can be imagined, and each of them leads to a different modification of the allele frequency distribution. Historically, a lot of studies in the field of population genetics have focused on *island models*, where populations (the islands) exchange migrants continuously within each others, at a constant rate along generations. These

models are considered to be at equilibrium, which means that populations have always existed and are not related by any philogenetical history.

Here we will rather concentrate on scenarios where only a few migration events have occurred in the recent history, which are also called *admixture* models. In livestock species, the typical example of an admixture event is when a breeder decides to cross his animals with those from another breed, so that the descendants inherit one interesting trait from this other breed. A drastic example is the Dorper sheep. This breed was born as a mixture of Dorset Horn and Blackhead Persian (of Animal Science Oklahoma State University, 1995). The Blackhead Persian breed is originally from Somalia, though adapted to arid clime and the Dorset breed is an easy care meat breed. By crossing these breeds, the South African Department of Agriculture created a meat sheep adapted to the more arid regions of the country.

Admixture models are also intensively studied in human genetics, where notorious examples of admixture events have been documented or at least discussed (Laval et al., 2010, Liu et al., 2006), for instance between Africans and Americans at slavery time (Tishkoff et al., 2009), between hunter-gatherers and pastoralists in the Neolithic, between Sapiens and Neandertals in the Palleolithic (Wall and Hammer, 2006).

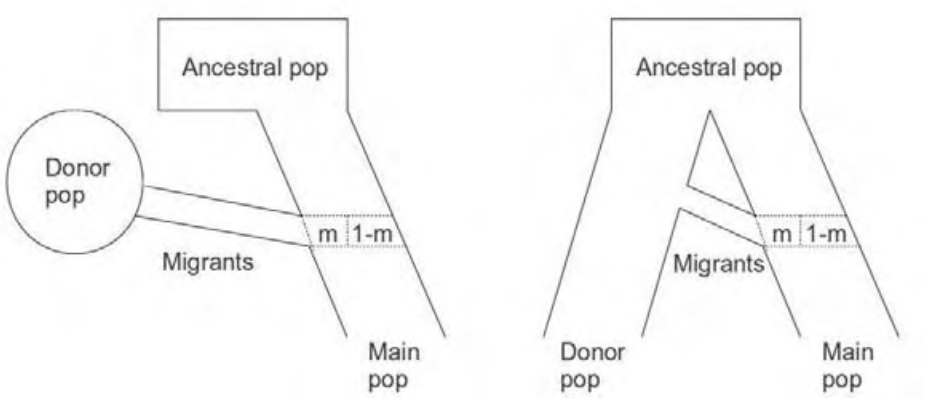


Figure 2.9: **Migration events.** The main population receives a proportion of m migrants from a donor population. The migrants can come from a population that shares the ancestral population (right) or not (left).

Let t' be the generation where admixture occurred. The population under study receives migrants from another population. In this section the former is called the main population and the latter the donor population. The donor population can share the ancestral population with the main one or not (Figure 2.9). Let also $p(t')$ be the frequency of allele A in the main population. In contrast with the model without migration described in Equation 2.1, $X(t'+1)$ does not depend only on $p(t')$, the frequency of the A-allele in the main population, but also on the proportion of alleles that are sampled from the migrant population, denoted m, and the frequency of allele A in the migrant population, denoted p_m. Indeed, $X(t'+1)|(p(t'), p_m, m)$ has a Binomial distribution $\mathcal{B}(2N, \pi(t'+1))$, where the probability $\pi(t'+1)$ of choosing allele A is given by:

$$\begin{aligned}
\pi(t'+1) = &\ \mathbb{P}(X(t'+1) = A) \\
= &\ \mathbb{P}(X(t'+1) = A | A\ comes\ from\ a\ migrant)\mathbb{P}(migrant) \\
&+\mathbb{P}(X(t'+1) = A | A\ comes\ from\ main\ pop)\mathbb{P}(main\ pop) \\
= &\ p_m m + p(t')(1-m)
\end{aligned}$$

Consequently, we have

$$E(X(t)) = \begin{cases} p_0 & if\ t \leq t' \\ (1-m)p_0 + mp_m & if\ t > t' \end{cases}$$

If the migrant population descends from the same ancestral population as the main population, then $\mathbb{E}(p_m(t) = p_0)$ and $E(X(t)) = p_0$ remain unchanged. But if the migrant comes from a population that does not share the ancestral population, depending on the proportion of migrants m and the frequency of the A allele in the migrants p_m, the mean can be shifted. If there is no information about the population that gives migrants, then the modeling of the frequencies depends from variants that we cannot control.

The variance of $X(t)$ also remains as in Equation 2.3 if $t \leq t'$, but for $t = t' + 1$ we are sampling from two different populations. As we sample from a larger population, the variance at this generation will be higher than

in a non-migration scenario.

(Pickrell and Pritchard, 2012) generalized the kinship matrix for admixed populations based on the same model that was presented above (Subsection 2.2.2), but with a slight modification to simplify the equations: when computing the variances of the populations as in Equation 2.7, they approximate $1-(1-\delta_{X1})(1-\delta_{0X})$ by $\delta_{X1}+\delta_{0X}$. When both quantities are small this approximation holds because $\delta_{X1}\cdot\delta_{0X}\approx 0$. They developed a software to compute these type of trees called *Treemix*.

2.3 Modeling the joint evolution of allele frequencies at several linked loci

Until now, we considered the evolution of allele frequency at only one biallelic locus. However, as two alleles that are located on the same chromosome tend to be transmitted together from one generation to another, the allele frequencies at distinct loci are generally not independent, and the information provided by haplotype frequencies is richer than that provided by the marginal allele frequencies.

Consider for example two loci with alleles A, a and B, b, and let us suppose that we have two populations, one with 50% of haplotypes AB and 50% of haplotypes ab, and the second with 50% of haplotypes Ab and 50% of haplotypes aB. The allele frequencies in both populations are $p_A=\frac{1}{2}$ and $p_B=\frac{1}{2}$, which would mean that populations are not differentiated. On the other hand, haplotype frequencies make it clear that these populations are very different.

This potential correlation between allele frequencies at two distinct loci is called **linkage disequilibrium**, and can be quantified by the measure

$$\begin{aligned} D_{AB} &= p_{AB}-p_A\cdot p_B \\ &= p_{AB}p_{ab}-p_{Ab}p_{aB} \end{aligned} \tag{2.9}$$

In this section, I will first briefly indicate how the derivations presented for single locus evolution might be extended to model haplotype evolution.

I will then introduce one approximative model that greatly simplifies statistical inferences based on haplotype frequencies, which will be central in the methodological development of Chapter 4.

2.3.1 Evolution of haplotype frequencies under the Hardy-Weinberg hypotheses

Let us come back to the ideal situation of a single population satisfying the Hardy Weinberg hypotheses, and consider two loci with allele frequencies p_A and p_B (necessarily constant over time), and with an initial linkage disequilibrium D_0 at generation 0. The recombination rate r between two loci is the probability for two alleles carried on a same chromosome to be inherited on different gametes during the meiosis due to a crossing over.

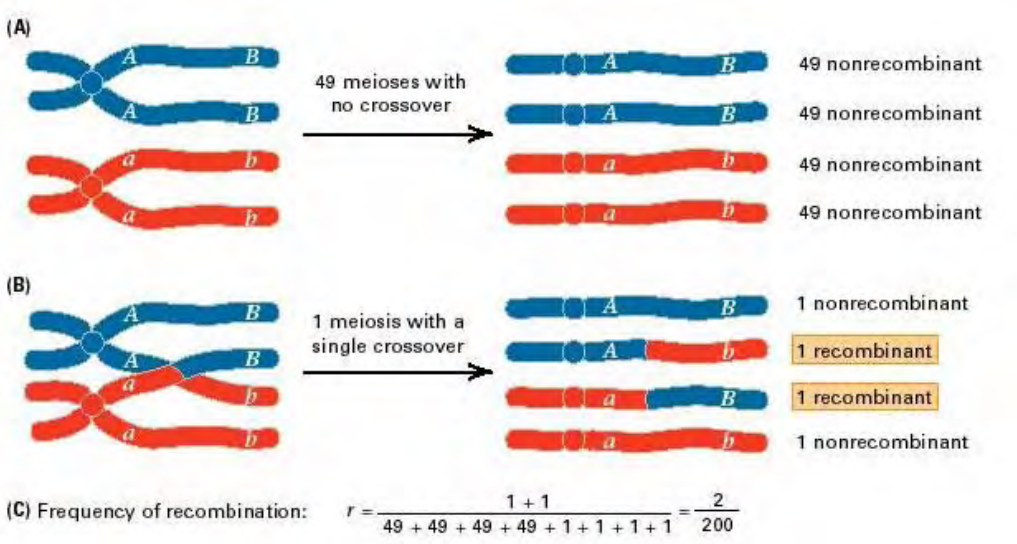


Figure 2.10: **Recombination event** The recombination rate is 1%

To sample one haplotype AB at generation t, we have to choose either one non recombining haplotype AB from generation $t-1$, or two independent alleles A and B that are put together by recombination. Consequently we have:

$$\begin{aligned} p_{AB}^{(t)} &= (1-r)p_{AB}^{(t-1)} + rp_Ap_B \\ \Rightarrow\ p_{AB}^{(t)} - p_Ap_B &= (1-r)(p_{AB}^{(t-1)} - p_Ap_B) \\ \Rightarrow\ D_t &= (1-r)D_{t-1} \end{aligned}$$

thus

$$D_t = (1-r)^t D_0 \tag{2.10}$$

and

$$p_{AB}(t) = p_Ap_B + (1-r)^t D_0 \tag{2.11}$$

The recombination rate varies from complete linkage ($r = 0$) to independent segregation ($r = \frac{1}{2}$). If $r = 0$, then $p_{AB}^{(t)} = p_{AB}^{(0)}\ \forall t$, so p_{AB} remains constant. If $r \neq 0$, D decreases exponentially with t down to 0 and allele frequencies at the two loci tend to linkage equilibrium, but the speed of this convergence can be very slow for close loci, the recombination rate per generation and per base pair being of order 10^{-8} in mammals.

Relaxing the Hardy-Weinberg hypotheses, even under models of genetic drift, the derivations become quickly more complex. Indeed, allele frequencies do not remain constant, and from the above equations we can see that $p_{AB}^{(t+1)}$ does not only depend on $p_{AB}^{(t)}$, but also on all other haplotype frequencies. Besides, the number of haplotypes increases exponentially with the number of loci (2^L haplotypes for L loci). To overcome this complexity and capture the haplotype diversity in a sampled population, several approximate models have been proposed, for instance Scheet and Stephens (2006), Stephens et al. (2001) and Browning (2006). In the following subsection I present two of them.

2.3.2 Multilocus models for linkage disequilibrium

The models described in this section were originally developed for inferring the haplotypes that segregate in a sample based on the observed genotypes in the sample, but they also provide a way to summarize the haplotype di-

versity. They capture two essential features of haplotype diversity. First, if we consider all homologous haplotypes found in a sample at a given set of loci, we will observe that locally many of them look very similar and differ only in a few sites, while others vary considerably in their sequence of allelic patterns. This is due to a different level of shared ancestry: in the former case, haplotypes have been carried on the same ancestral chromosome until a quite recent generation, while in the latter they have evolved independently for a much longer time and have become different due to successive recombinations and mutations. To account for these different differentiation levels, the models developed by Browning (2006), Scheet and Stephens (2006) aim to cluster similar haplotypes. Second, recombination implies that the ancestry of a sample changes continuously along the genome. Consequently, two haplotypes can be very similar in a given region and very different in another one. To capture this property, the cluster membership of each haplotype in the sample is modeled as a Markov Chain along the genome. As patterns of recombination change along the genome, Markov Chain based models, are more suited than the approaches based on "block-based" clustering, which allow the cluster membership to change only when changing from one block to the other and cluster membership is the same for each haplotype along the block. Small blocks can loose the structure given in large regions of strong linkage disequilibrium and too long blocks loose information because there will be too much noise if there are short regions of strong linkage disequilibrium (see Greenspan and Geiger (2006) for an example of a block-based algorithm).

Below I describe the model of (Scheet and Stephens, 2006) in more detail because it is the model used in Chapter 4. Then I describe shortly the model of Browning (2006), because it is close in spirit to the model of Scheet and Stephens (2006) and could also have been used. The former is the model on which the fastPHASE software is based, and the latter is for the BEAGLE software. The main difference between the models underlying fastPHASE and BEAGLE is that the model of fastPHASE is based on a Hidden Markov Model for a fixed number of clusters K, while BEAGLE is based on a Variable Length Markov Chain and the number of clusters varies

along the chromosome.

Hidden Markov Model

First we will see the model without recombination, allowing recombination in a second step. Then, I present the extension of the model to genotypes, where the haplotype phase is not known. This last one is the model used in Chapter 4.

A cluster model for haplotypes: Assume we have n haplotypes with M SNP markers each. Let $h_i = (h_{i1}, \ldots, h_{im}, \ldots, h_{iM})$ be the $i-th$ haplotype, where h_{im} is the allele carried by haplotype i at SNP m, with $i \in \{1, \ldots, n\},\ m \in \{1, \ldots, M\}$.

Let us assume that each haplotype comes from one of K clusters, labeled $k = 1, \ldots, K$. Let z_i be the cluster to which haplotype h_i belongs, α_k be the frequency of cluster k in the set of haplotypes, and θ_{km} be the frequency of allele 1 (allele A) at marker m in cluster k. The matrix $\theta = (\theta_{km})$ contains the frequencies of allele 1 in all clusters at all loci and the vector $\alpha = (\alpha_1, \ldots, \alpha_k, \ldots, \alpha_K)$, all the cluster frequencies in the haplotype set.

Conditional on the cluster to which each haplotype belongs, the alleles observed on each marker are independent Bernoulli variables, whose distribution is determined by the allele frequency matrix. Consequently,

$$\mathbb{P}(h_i|z_i = k, \theta) = \prod_{m=1}^{M} \theta_{km}^{h_{im}} (1 - \theta_{km})^{1-h_{im}} \tag{2.12}$$

However, haplotype cluster memberships are actually unknown (they are latent variables), we have to sum over the distribution of the z_i's. Thus:

$$\begin{aligned} \mathbb{P}(h_i|\alpha, \theta) = &\ \textstyle\sum_{k=1}^{K} \mathbb{P}(z_i = k|\alpha)\mathbb{P}(h_i|z_i = k, \theta) \\ = &\ \textstyle\sum_{k=1}^{K} \alpha_k \prod_{m=1}^{M} \theta_{km}^{h_{im}} (1 - \theta_{km})^{1-h_{im}} \end{aligned} \tag{2.13}$$

Ideally the coefficients of θ are close to 0 or 1, so clusters might be essentially seen as haplotypes, with a small uncertainty about the alleles at some

positions.

Local clustering of haplotypes If there is no recombination, each haplotype belongs to a single cluster. But if there is recombination, the cluster membership must be allowed to change along the genome. Due to linkage disequilibrium, cluster membership between close loci will however be correlated.

To account for this possible change, we now denote z_{im} the cluster membership at marker m for haplotype i. For each haplotype, the sequence of cluster memberships $z_i = \{z_{i1}, \dots z_{im}, \dots z_{iM}\}$ is modeled as a Markov Chain that takes values in $\{1, \dots, K\}$, with initial probabilities

$$\mathbb{P}(z_{i1} = k) = \alpha_{k1} \tag{2.14}$$

and transition probabilities at each marker

$$\begin{aligned} \mathbb{P}(k \to k') := \;& \mathbb{P}(z_{im} = k' | z_{i(m-1)} = k, \alpha, r) \\ = \;& \begin{cases} e^{-\beta_m} + (1 - e^{-\beta_m})\alpha_{k'm} & k' = k \\ (1 - e^{-\beta_m})\alpha_{k'm} & k' \neq k \end{cases} \end{aligned} \tag{2.15}$$

where, for $m = 2, \dots M$, β_m and $\alpha_m = \{\alpha_{m1}, \dots, \alpha_{km}, \dots \alpha_{Km}\}$ are parameters to be estimated.

The above transition probability (Equation 2.15) arises from the fact that the Markov Chain is a discretized version of a continuous Markov jump process, with jump rate β_m and transition probabilities $\alpha_{k'm}$ given that a jump occurred. So, when $k = k'$, the process either does not jump, or jumps but reaches again the same state. When $k \neq k'$ the process jumps and chooses the state k'. β_m depends on d_m, the physical distance between markers $m-1$ and m, and r_m, the recombination rate between these markers.

As in the previous model, the alleles observed at distinct loci are independent conditional on the cluster membership, so we have:

$$\mathbb{P}(h_i | z_i, \theta) = \prod_{m=1}^{M} \mathbb{P}(h_{im} | z_{im}, \theta), \tag{2.16}$$

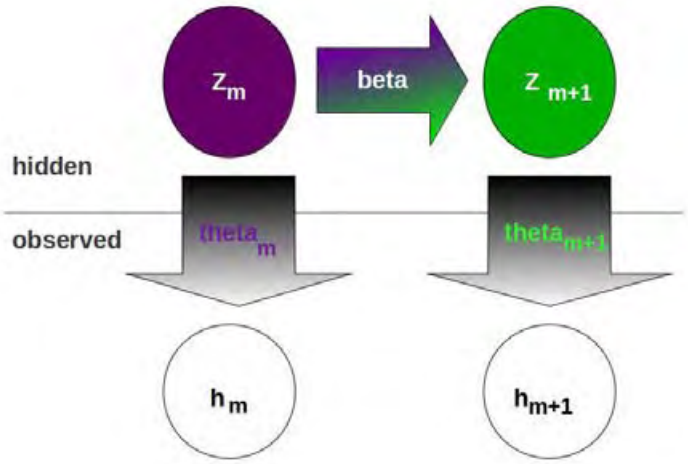


Figure 2.11: **Schematic representation of the Hidden Markov Model.** The subscript i is omitted for simplicity. z_m takes values in $\{1, \ldots, K\}$. The emission probability of the cluster θ_m, depends on the status of z_m, so when $z_m = k$, $\theta_m = \theta_{km}$.

where $\mathbb{P}(h_{im}|z_{im} = k) = \theta_{km}^{h_{im}}(1 - \theta_{km})^{h_{im}}$. Since z_i is unknown, we again get:

$$\mathbb{P}(h_i|\alpha, \theta, \beta) = \sum_{z_i} \mathbb{P}(h_i|z_i, \theta)\mathbb{P}(z_i|\alpha, \beta), \tag{2.17}$$

where $\mathbb{P}(z_i|\alpha, \beta)$ is determined by (2.14) and (2.15).

Extension to local clustering of genotypes For modeling haplotypes, phase data is needed. Here, we extend the model to genotypes to use it directly on unphased data. Actually, the genotype model is then used to phase data.

Let us note $g = (g_1, \ldots, g_n)$ the genotypes of n individuals, where $g_{im} = 2$ if individual i carries two A-allele copies at locus m, 1 if it carries one, and 0 if it carries two a. Under Hardy-Weinberg equilibrium, we can re-write the previous equations for unordered pairs of clusters $\dot{z}_{il}$, from which the genotype g_{im} originates. $\dot{z}_i = \{\dot{z}_{i1}, \ldots, \dot{z}_{iM}\}$ form a Markov chain with initial and transition probabilities as follows:

$$\mathbb{P}(\dot{z}_{i1} = \{k_1, k_2\}) = \begin{cases} (\alpha_{k_1 1})^2, & k_1 = k_2 \\ 2\alpha_{k_1 1}\alpha_{k_1 2}, & k_1 \neq k_2 \end{cases} \tag{2.18}$$

$\mathbb{P}(\{k_1, k_2\} \to \{k_1', k_2'\})$

$$= \begin{cases} \mathbb{P}(k_1 \to k_1')\mathbb{P}(k_2 \to k_2') + \mathbb{P}(k_1 \to k_2')\mathbb{P}(k_2 \to k_1'), & k_1 \neq k_2, k_1' \neq k_2' \\ \mathbb{P}(k_1 \to k_1')\mathbb{P}(k_2 \to k_2'), & otherwise \end{cases} \tag{2.19}$$

where $\mathbb{P}(k \to k')$ is defined as in Equation 2.15. As previously, the alleles are independent draws from the cluster allele frequencies, and we have:

$$\mathbb{P}(g_i|\dot{z}_i, \theta) = \prod_{m=1}^{M} \mathbb{P}(g_{im}|\dot{z}_{im}, \theta),$$

where

$\mathbb{P}(g_{im}| = \dot{z}_{im} = \{k_1, k_2\}, \theta)$

$$= \begin{cases} (1 - \theta_{k_1 m})(1 - \theta_{k_2 m}), & g_{im} = 0 \\ \theta_{k_1 m}(1 - \theta_{k_2 m}) + \theta_{k_2 m}(1 - \theta_{k_1 m}), & g_{im} = 1 \\ \theta_{k_1 m}\theta_{k_2 m}, & g_{im} = 2 \end{cases}$$

And again, since $\dot{z}_i$ is unknown, we sum over all possible values:

$$\mathbb{P}(g_i|\alpha, \theta, \beta) = \sum_{\dot{z}_i} \mathbb{P}(g_i|\dot{z}_i, \theta)\mathbb{P}(\dot{z}_i|\alpha, \beta), \tag{2.20}$$

and $\mathbb{P}(\dot{z}_i|\alpha, \beta)$ is determined by the initial and transition probabilities.

These models are called Hidden Markov Models (HMM), because the latent variables representing the cluster membership form a Markov Chain (see Figure 2.3.2). Standard estimation and prediction procedures have been developed for this class of models. Scheet (2006) developed fastPHASE, which is based on a Baum-Welsh expectation maximization algorithm, that estimates the parameters α, β and θ, and returns the cluster membership probabilities for each haplotype.

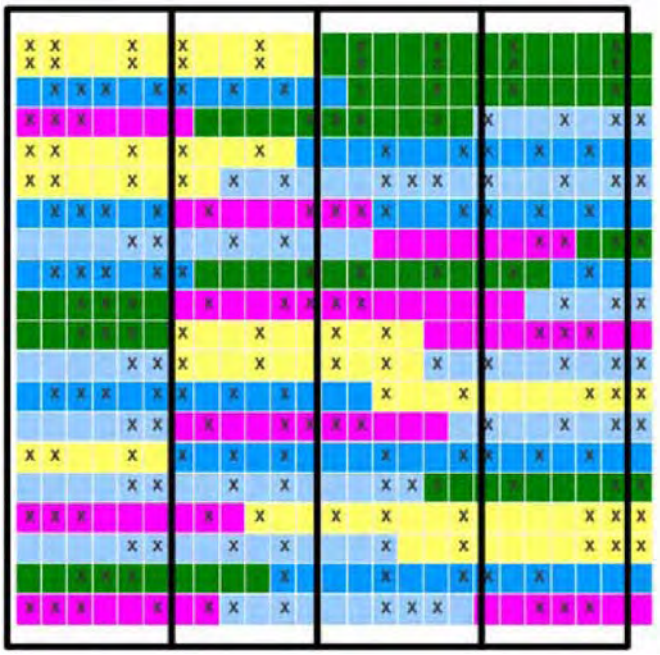


Figure 2.12: **A possible cluster classification for 20 haplotypes.** Each line corresponds to an haplotype and each column to a position on the genome. Clusters are represented by colors, and allele 1 (allele A) at each position is represented by a cross. The black vertical lines represent a possible choice of window size (6 SNPs). This shows that "block-based" models would not have the same flexibility that the local clustering as if the clusters were allowed to change just when changing from one window to another.

The cluster frequencies in each population give us an idea of the local haplotype diversity in each population. If there are just a few haplotypes in the population, we expect to have one or a few very frequent clusters, on the other hand if there is a lot of haplotype diversity, we expect to get a lot of more or less equally frequent clusters. In Figure 2.12 we present an example of local clustering of haplotypes.

Variable Length Markov Model

Based on the same idea as above, that locally haplotypes can be clustered together, Browning (2006) proposed a local clustering model based on an inhomogeneous Variable Length Markov Chain (VLMC). Here, the underlying idea is that in the zones where linkage disequilibrium is high the Markov Chain will have a long memory, but if the linkage disequilibrium is small, then the memory will be shorter, because recombination events break it. Browning (2006) claimed that VLMC do not require explicit modeling and are flexible enough to closely approximate HMM.

The VLMC can be represented by a directed acyclic graph, where each node has a level that corresponds to a locus (Figure 2.13). On level 1, there is just one node that contains no information, being just a starting node. For levels $m = 2, \ldots, M + 1$, each level m represents a history or a collection of possible haplotypes up to $m - 1$ and each edge going from a node at level m, to a node at level $m + 1$ represents an allele at locus m.

On Figure 2.13A we see a representation of the sample of haplotypes, without clustering. Figure 2.13B represents the same haplotypes after modeling. Two edges arriving to the same node, represent a loss of memory of the Markov Chain, which is caused by historical recombination. Two nodes are merged if the transition probabilities corresponding to all descendant nodes are sufficiently similar (Browning, 2006). Each haplotype is represented by a path from the first node, to the node at level $M + 1$.

(Browning and Browning, 2007a) defined a clustering method, that is locus dependent. Given an edge between two nodes at levels m and $m + 1$, a local cluster is the set of all haplotypes that trace their path through this edge. So local cluster membership changes depending on the locus that we take as reference, although it is expected that haplotypes that were in the same cluster when taking locus m as reference will be in the same cluster also when taking locus $m + 1$ as reference. But, anyway the number of clusters when considering locus m or locus $m + 1$ can change.

Other softwares have been proposed for phasing genotypes and summa-

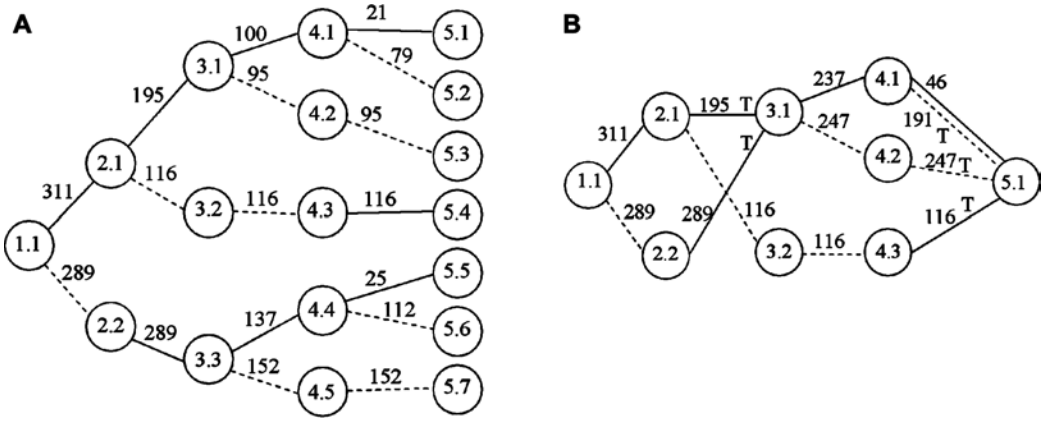


Figure 2.13: **Tree graph constructed using haplotype data. (A)** Circles represent nodes, and the values in them represent a level and a node identifier within level. For example, 3.2 denotes node 2 at level 3. A solid edge between nodes at levels m and $m+1$ represents allele 1 at SNP marker m; a dashed edge represents allele 0. Numbers above edges represent haplotype counts. Thus, 137 over the edge between 3.3 and 4.4 represents 137 haplotypes that have allele 0 at the first SNP, 1 at the second SNP, and 1 at the third SNP. Although directional arrows are not shown, a left-to-right direction is implied. **(B)** The graph from figure A after merging. Nodes 3.1 and 3.3 in figure A have been merged, as all nodes at level 5. Notation is as described for panel A. This figure was extracted from Browning (2006). The model was presented in the context of multilocus association mapping. So edges that are marked with T were tested in this context.

rizing haplotype diversity, as PHASE Stephens et al. (2001). PHASE software is based in a Product of Approximate Conditionals (PAC) model. The principle is to compute the probability of observing a new haplotype in the sample as a combination of existing ones (*i.e.* conditional to existing ones). When developing fastPHASE, Scheet and Stephens (2006) aimed to combine the flexibility to capture patterns of linkage disequilibrium of the PHASE model with the computational convenience of "block-based" models. Even if it could loose a little bit of accuracy, fastPHASE is much faster than PHASE, allowing to analyze much bigger datasets. Although, in terms of speed BEAGLE (Browning and Browning, 2007b) is anyway the fastest one, because its model is the easiest one to fit and thus largely used to phase data. In the next chapter we will see the advantages and drawbacks of using these models

when the aim is to detect selection.

2.4 The impact of selection on genetic diversity

I have focused so far on neutral evolution models, where all genotypes at a given locus have the same fertility and viability. A locus is considered under selection when this assumption does not hold, that is when some of the genotypes will have a higher or lower probability to be transmitted from one generation to the other. In this section I will briefly describe how allele frequencies evolve at a single locus under selection. I will then focus on positive selection, which is the type of selection I am mostly interested in in this work, and describe how it affects the genotype and haplotype diversity at neutral loci that are linked to the locus under selection.

2.4.1 Different types of selection

In order to illustrate the effect of natural selection on the evolution of allele frequencies at one single locus, I start again from the Wright-Fisher model of Section 2.2.1. But in contrast with the neutral situation, we now assume that the genotypes aa, aA and AA have different viabilities, which are denoted w_{aa}, w_{aA} and w_{AA}. Consequently, the probability of sampling one allele A from generation t is no longer equal to $p(t)$ but to

$$p^*(t) = \frac{p(t)\left[p(t)w_{AA} + (1-p(t))w_{aA}\right]}{\bar{w}},$$

where $\bar{w} = p^2(t)w_{AA} + 2p(t)\left(1-p(t)\right)w_{aA} + \left(1-p(t)\right)^2 w_{aa}$

Different types of evolution scenarios can be distinguished according to the viability values Balding and Bishop (2007):

Directional selection: $w_{AA} > w_{aA} > w_{aa}$ or $w_{AA} < w_{aA} < w_{aa}$. In the former situation the frequency of A (or a in the latter) will tend to increase until fixation (Figure 2.14). Within this category, one generally

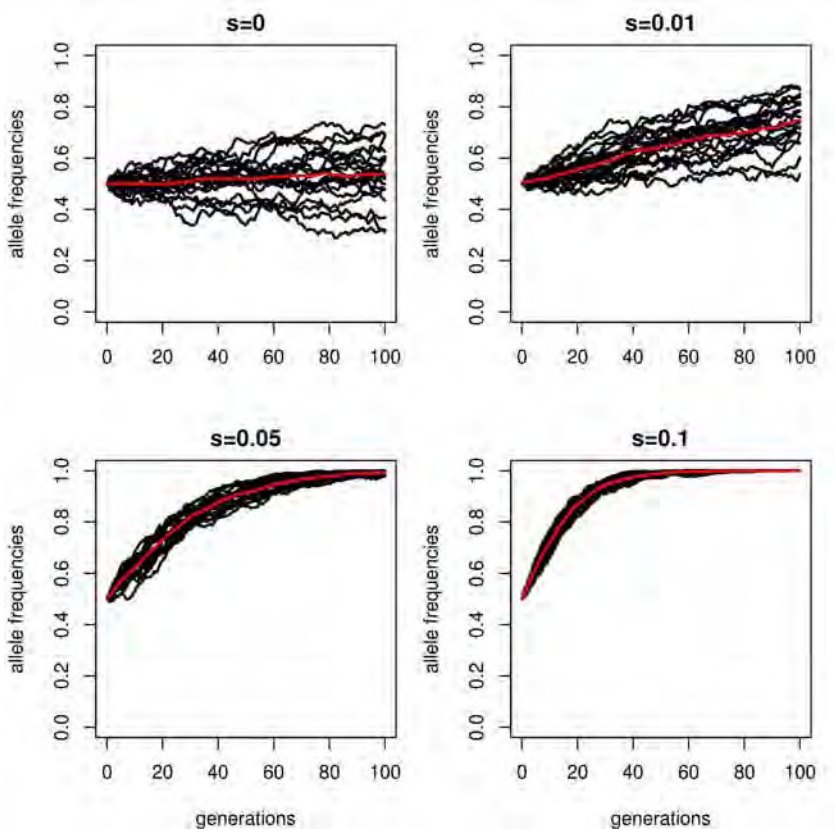


Figure 2.14: **Directional selection:** Effect of different selection intensities $s > 0$ on the evolution of allele frequencies. $w_{AA} = (1+s)^2, w_{Aa} = 1+s, w_{aa} = 1$. When $s = 0$ the frequencies evolve under genetic drift (the neutral model). N=1000. The red line is the mean frequency of the simulated trajectories.

makes a further distinction between *negative* (or *purifying*) selection, where the derived allele has a lower viability than the ancestral one and tends to be removed quite rapidly from the population, and *positive* (or *adaptative*) selection, where the derived allele has a higher viability and thus has a chance to maintain or even get fixed in the population.

Balancing selection: $w_{AA}, w_{aa} < w_{aA}$. In this situation heterozygotes are selected so the polymorphism tends to be stable at the locus (Figure 2.15).

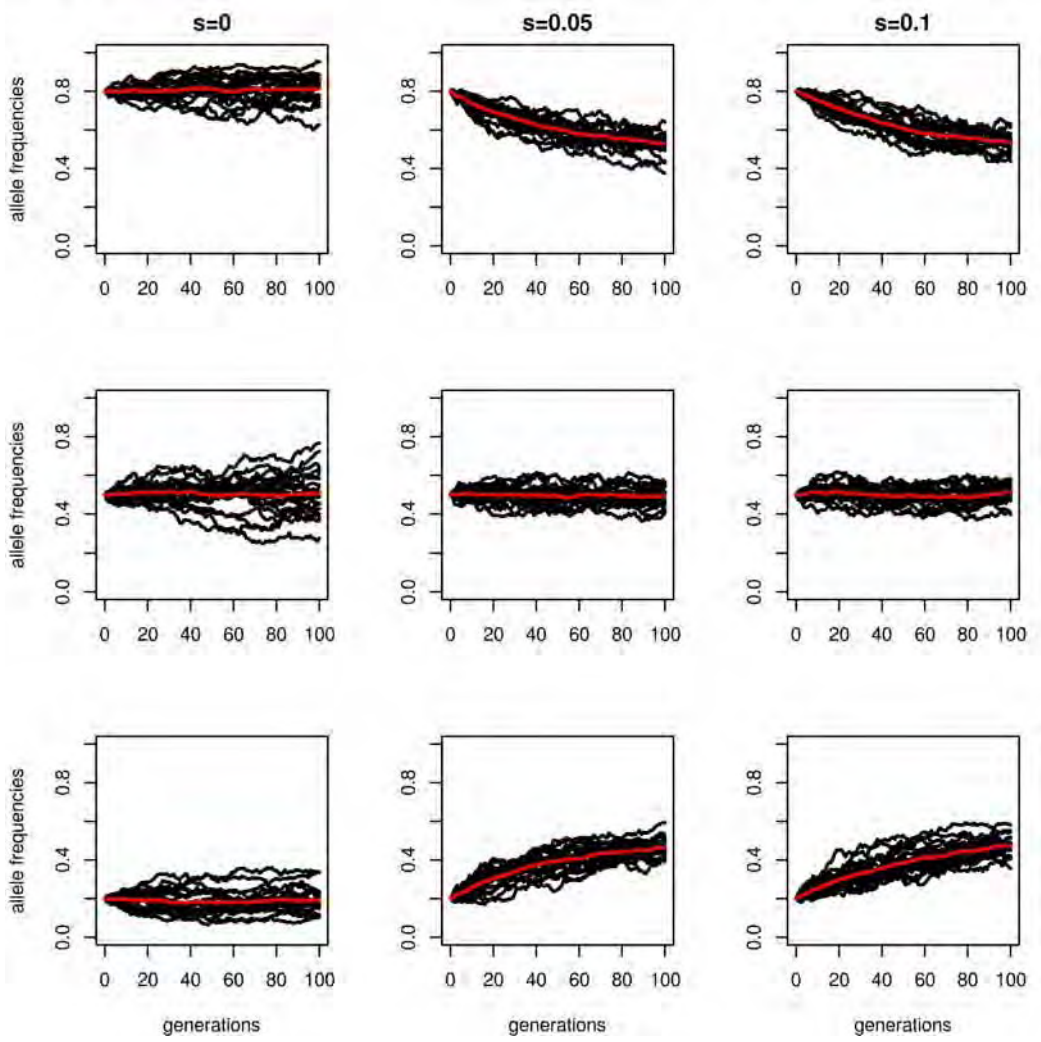


Figure 2.15: **Balancing selection:** Effect of different selection intensities $s \geq 0$ (columns) and initial allele frequencies (rows) on the evolution of allele frequencies. N=1000, $w_{AA} = 1, w_{Aa} = 1+s, w_{aa} = 1$ The red line is the mean frequency of the simulated trajectories.

$w_{AA}, w_{aa} > w_{Aa}$: In this case the polymorphism is unstable: either A or a eventually gets fixed, depending on the initial frequencies (2.16). When the initial frequency of A is higher, it is more probable that A gets fixed. If the initial frequency is 0.5, both alleles have the same probability of being fixed.

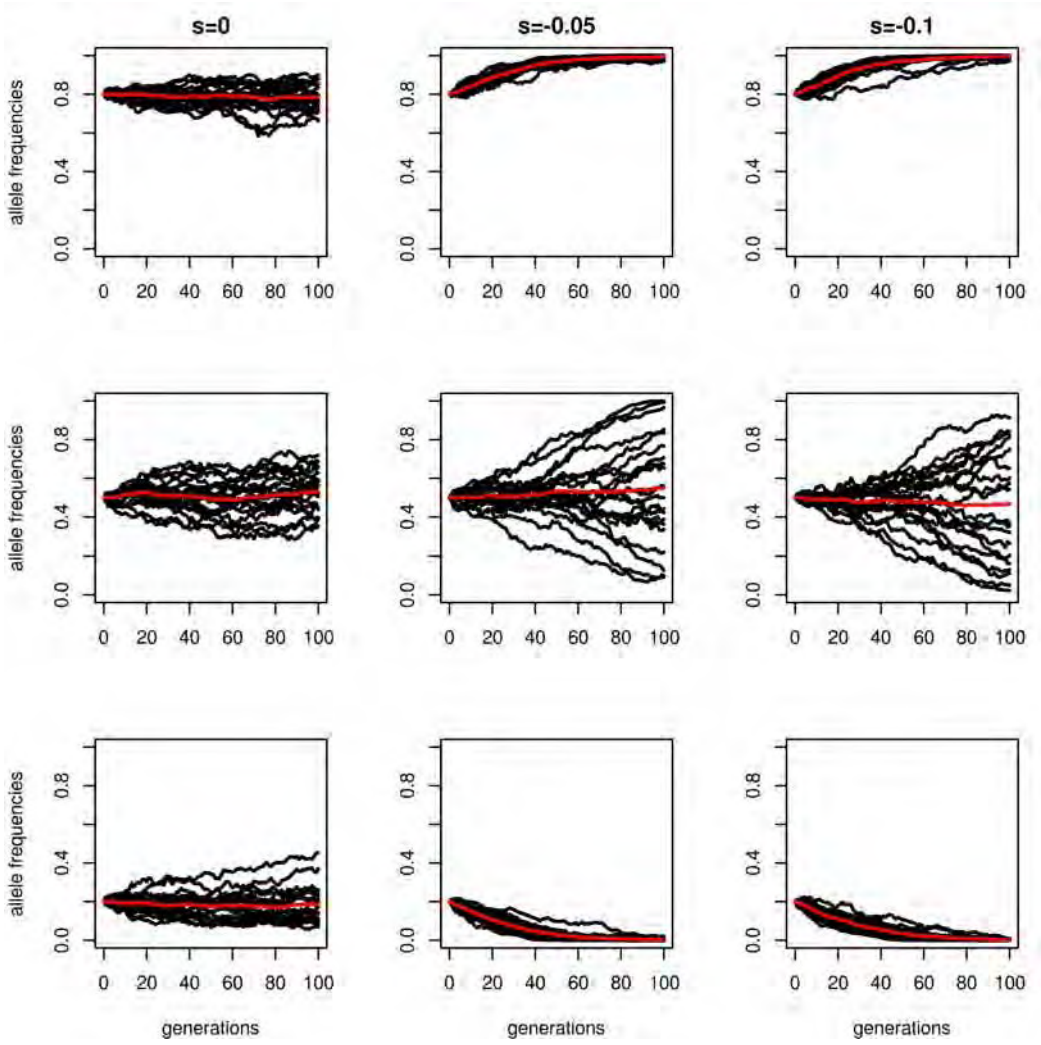


Figure 2.16: Effect of different selection intensities $s < 0$ (columns) and initial allele frequencies (rows) on the evolution of allele frequencies. N=1000, $w_{AA} = 1, w_{Aa} = 1 + s, w_{aa} = 1$ The red line is the mean frequency of the simulated trajectories.

2.4.2 Signatures of selection

In Section 2.3 we noted that under neutrality, allele frequencies at linked sites did not evolve independently from each other. Similarly, positive selection at one locus modifies the allele frequencies of neutral loci in a neighborhood of the selected locus, and the linkage disequilibrium pattern around it. Genetic diversity around the selected locus will thus have a particular spatial pattern.

Maynard Smith and Haigh (1974) first noted that a new positively selected mutation could modify the frequency of neutral alleles due to physical linkage and called this phenomenon the hitch-hiking effect. They described what is now known as a hard sweep scenario. After this pioneer study, a lot of work was devoted to describe the expected genetic pattern left by hitch-hiking, and to predict hard sweeps from genetic data, for instance Gillespie (1997), Kim and Nielsen (2004), Kim and Stephan (2002), Stephan et al. (2006).

More recently, a growing interest was put on selection from standing variation (Hermisson and Pennings (2005), Innan and Kim (2004), Pennings and Hermisson (2006a,b), Przeworski et al. (2005)), where selection starts acting on an allele that is already segregating in the population at a given frequency, and on other selection scenarios leading to different patterns than the classical hard sweep. Examples of such soft sweep scenarios are found in selection for malaria resistance (the Duffy blood group locus, Hamblin and Di Rienzo (2000), Hamblin et al. (2002) and G6PD gene Tishkoff et al. (2001)), or lactase persistence (Tishkoff et al. (2007)) in humans.

Separately, Hernandez et al. (2011) computed the probability of fixation of a single allele, depending if it was a new mutation or if it was already segregating in the population and (Pritchard et al., 2010) computed the probability that a sweep is from standing variation given that it is a sweep, depending on the size of the mutational target, *i.e.* the number of sites that can produce a phenotypic change, and the intensity of selection. Both concluded that the probability of fixation is larger for alleles coming from standing variation. The reason is that the probability for a signature of selection to arise from a hard sweep scenario is very small, because it implies that a new mutation occurred and provided an advantage to the individuals carrying it. Then, even if the selected advantage was huge, when the frequency of the mutation is really low, it can be easily lost right away due to the stochasticity in the sampling procedure for passing from one generation to the next one. On the other hand if selection from standing variation occurred, the allele was already available at a non negligible frequency, reducing the probability of stochastic loss compared to a single copy mutation. In addition, there is no waiting time for the mutation to appear, alleles are ready for being selected

(Hermisson and Pennings (2005)). When the reason of selection is either an environmental change, or selection for a beneficial trait by a breeder, selection from standing variation is the most probable scenario.

Innan and Kim (2004) pointed out that in domesticated species the proportion of soft sweeps compared to hard sweeps was very high, and described using simulations the expected pattern left by selection from standing variation (see more details below). Hernandez et al. (2011) analyzed 179 human sequences and pointed out that hard sweeps were rare in humans also. They found that aminoacid and putative regulatory sites were not significantly enriched in highly differentiated alleles between populations, and that diversity levels near exons and conserved noncoding regions decrease, in contrast to what could be expected under a hard sweep scenario.

Below the different possible sweep scenarios that have been described so far in the literature and the corresponding genetic signatures are described. All these scenarios involve positive selection, but they differ in the initial and final frequency of the selected allele or in the amount of selected alleles on a locus and have been named according to these differences (Price et al., 2010).

Hard sweep

In this scenario, a new advantageous mutation appears in the population (its initial frequency is $p_0 = \frac{1}{2N}$), increases in frequency and gets fixed in the population in a relatively short lapse of time. The neutral alleles that were present on the haplotype where the mutation appeared also raise in frequency. Those that are very close to the selected locus will also have a very high probability of getting fixed in the population, resulting in the elimination of almost all possible genetic variation in a neighborhood of the selected locus (Figure 2.4.2).

As the distance from the selected locus increases, neutral alleles that were initially associated to the advantageous allele are more likely to be separated from it by recombination, so the fixation probability is reduced and the expected genetic diversity converges again towards that of a neutral model. The expected size of the region with reduced genetic variation depends on

Figure 2.17: **Hard sweep:** One haplotype increases in frequency together with the selected mutation.

the strength of selection (the stronger the selection, the larger the region), on the recombination rate (the more recombination, the smaller the region) and on the demography of the population under selection (the smaller the population, the larger the region).

Hard sweep scenarios are relatively easy to detect using genomic data from the population under selection. Indeed, hard sweep regions are characterized by a deficit of segregating sites and an excess of low frequency and high frequency derived alleles, which can be detected even with single marker data. If genetic data in neutral related populations is available, the hard sweep signature should be even clearer. Indeed, the genetic differentiation between the selected population and the neutral populations is also increased in the sweep region, because some alleles that are rare in the neutral populations will show an important raise in frequency in the selected population, which can not be explained by drift alone.

Partial sweep

This scenario, which is also called ongoing sweep or incomplete sweep, is similar to a hard sweep scenario except that the selected allele has not yet been fixed. One haplotype is found at high frequency around the selected

locus, but other haplotypes are still segregating.

Genetic variation in the population is reduced, but not completely removed, around the selected locus, so such scenarios are generally difficult to detect based on single marker data from the selected population. Linkage disequilibrium around the selected locus has a specific structure, which has been studied by Stephan et al. (2006) modeling the evolution of three linked loci: two neutral and one under selection. Linkage disequilibrium across the selected locus (that is between one SNP upstream the locus and one SNP downstream the locus) increases until the selected allele reaches a frequency of 50%, but decreases back from this threshold. This counter-intuitive result comes from the fact that if there is a recombination event between two SNPs, the selected allele will be in a new haplotype, raising its frequency also, creating blocks of linkage disequilibrium. As recombination events are independent on each side of the selected locus, they do not create linkage disequilibrium across the locus. On the other hand, focusing on one side of the selected alleles, the linkage disequilibrium in the flanking regions of the selected locus (between two SNPs upstream or two SNPs downstream) remains very high until fixation of the selected allele. But the recombination events are not symmetric, so the patterns of linkage disequilibrium will not be symmetric either. Thus, there can be a lot of linkage disequilibrium on one side of the locus, but not too much on the other side. This means, that when detecting regions under selection, the selected locus does not need to be in the middle of the region.

As in hard sweep scenarios, genetic differentiation with related neutral populations is elevated around the selected locus. Although this increase is less pronounced than in hard sweep scenarios, it might still be distinguished from that expected under neutral evolution. Sometimes, this differentiation could be not really clear when looking at the frequencies of the alleles, but it should be clear when looking at haplotypes, because we expect that there is one long haplotype that rises its frequency in the selected population, while in the neutral, it could exist, but in a really low frequency.

When working with dense data (for example 50K chip), we do not consider that new mutations could hit the haplotypes, because new mutations

will not be considered in the chip. But, when working with higher density or sequencing data, new mutations could have hit selected haplotypes. Haplotypes are expected to be shorter, than when working with chip data, because these new mutations break long haplotypes down. The signals captured by tests for selection are going to be shorter, which could make easier to find the selected site, or at least the candidate gene.

Soft sweep

Almost all the remaining scenarios that imply positive selection at one locus are called a soft sweep. In contrast to the hard and partial sweep scenarios, in a soft sweep scenario the positively selected mutation has already drifted in the population when it becomes advantageous, and is thus found in several different haplotypes. These haplotypes can all raise in frequency when the mutation becomes advantageous.

There are mainly two situations under which a soft sweep scenario may occur. The first one is the single origin soft sweep scenario. As described in Hermisson and Pennings (2005), Innan and Kim (2004), in this scenario an allele that was neutral or slightly deleterious becomes positively selected due to a change of environment or of artificial selection objective. As the allele already drifted in the population, several copies of this allele exist, and are associated to several haplotypes, which were created by recombination and mutation. All these copies have a single common ancestor (they are identical by descent). Because of genetic drift, not all haplotypes carrying the positively selected allele will raise in frequency, but certainly more than one will (otherwise we are again, by chance, in the hard sweep scenario). This scenario is also called selection from standing variation.

The second scenario is the multiple origin soft sweep. In this case all the copies of the new allele are not identical by descendant, instead there was a collection of independent copies of the new allele. These independent copies can either arise before or after the allele began to be selected. The collection of haplotypes carrying the selected allele can be wider than in the single origin soft sweep because there are no longer identical by descent. This

difference between single origin and multiple origin sweep is mainly seen at the closest loci to the selected one. This scenario was studied by Pennings and Hermisson in two different situations (Pennings and Hermisson, 2006a,b). First, as recurrent mutations are needed, a high mutation rate and large effective population size are needed for mutation to hit several times the same locus. Then, once a mutation arises, as in the hard sweep scenario, there is little chance that it persists in the population, because even under a high selection rate, the new mutation can be lost just by stochasticity. So the chance of having a soft sweep with multiple origins, depends mostly on the probability of mutation per generation, and not that much on the selection intensity.

Soft sweep scenarios are more diverse than hard sweep scenarios, and the expected genetic pattern around the selected locus depends on several factors as the mutation rate before selection starts, or the frequency p_0 of the favorable allele when it becomes favorable. Consequently, the signature left by soft sweeps is not as marked as that left by hard sweeps.

The best situation for detection is a single origin scenario with a small p_0 (under 5% according to Innan and Kim, and under $\frac{1}{4Ns}$ according to Przeworski et al. (2005), which are equivalent in general). This second condition is for instance quite likely if the allele is slightly deleterious. In this case, provided mutation rate is not too high, there is little chance that mutations hit the initial haplotype carrying the favorable allele, so we expect that only one or a few haplotypes will carry this allele when selection starts. Since these haplotypes segregate in the population during a possibly long time before becoming selected, it will be shortened by recombination. We thus expect a signature that is similar to that of a hard sweep, but in a shorter region around the selected locus. This signature might thus be detected quite easily, provided we have sufficiently dense genomic data.

If the favorable allele was previously neutral, its frequency at the time when selection starts can vary from $\frac{1}{2Ns}$ to almost 1. Thus, depending on the mutation and recombination rates, there can be just a few or a lot of haplotypes carrying the selected allele and thus becoming selected. When looking only at allele frequencies, such scenarios will thus be very difficult to

distinguish from neutrality, and in the extreme case their genetic signature could even look more similar to that left by balancing selection than by directional selection. Haplotype frequencies in a neighborhood of the selected locus might still provide some specific signal, but this will depend on the amount of haplotypes whose frequency has raised.

In the particular case of artificial selection as modeled by Innan and Kim (2004), the breeder samples a part of the main population, creating a bottleneck and starts to select animals based on a specific trait. Then the population is multiplied and recovers the census that it had before the breeder began the selection process. In this particular case, an important part of the population genetic variation is lost because of the initial reduction in population effective size. If the selected allele was previously neutral, we noted above that there can be a lot haplotypes carrying it. But, because of the bottleneck, only a few of them are going to stay in the new population. If the favorable allele frequency before the bottleneck is not too high (typically $p_0 < 20\%$), Innan and Kim pointed out that the selective signature should be detected, because the polymorphism is significantly reduced around the selected locus. This assertion is made for tests based on one population and in allele frequencies. But, one also expects to find some haplotype structure in the population. On the other hand, if $p_0 > 50\%$, they observed that the probability of detecting the sweep is very low. Nevertheless Innan and Kim say that power might be improved if we can compare the selected population with its wild progenitor using the shared polymorphisms. Pennings and Hermisson made the same assertion but considering a close related neutral population instead of the wild progenitor. The reason of this, is that new mutations (after the population split) that hit the haplotypes carrying the mutations noise the signal.

In conclusion, under a soft sweep scenario we generally expect a reduction of variability around the selected locus, but not as pronounced as in the hard sweep scenario. On the opposite, high linkage disequilibrium is expected across the selected locus. There might be several haplotypes carrying the favorable allele, but still with a higher frequency than expected under neutral evolution, so the differentiation between the selected population and

related neutral populations will be higher than in the rest of the genome. When looking at the allele frequencies there might be shifts on their frequencies higher than expected under the neutral model, but for capturing these frequencies, good information about the ancestral frequencies might be needed.

Polygenic adaptation

Previously, I presented the patterns that selection on one locus can leave. In polygenic selection (Price et al., 2010) there are several loci that can be selected because all of them contribute to a single phenotype. Thus, selection towards an optimal value of this phenotype leads to small allele frequency changes at all or several of these loci, and no allele reaches fixation. This type of selection is also known as canalising selection (SanCristobal-Gaudy et al., 1998). This concept is a bit more general, because the responsible of the phenotypic optimum could be just one locus, that does never reach fixation.

In breeding populations, the experience from several decades of quantitative trait locus detection has shown that production traits (milk quantity or quality for instance) are mostly polygenic. Since recent artificial selection in breeding populations has essentially focused on such traits, polygenic selection appears as an important model for the detection of selection signatures in this field. A textbook example is human height. In Europeans, several genome wide scans showed that 50 different loci contribute to this phenotype ((Aulchenko et al., 2009, Gudbjartsson et al., 2008, Lettre et al., 2008, Weedon et al., 2008)) in increasing about with 3-6 mm the height of the individual. When selection occurs in traits like this, small shifts on the frequencies of the different loci (and not all at the same time) modify the phenotype very quickly allowing a very rapid adaptation.

Depending on the amount of loci involved in the trait, and if there are close together or dispersed along the genome, they could create a soft-sweep type of signal, with some haplotypes raising in frequency (if they are close) or several low frequency partial sweeps (if they are far away). But, as the

changes can be very small, this type of selection is certainly really difficult to detect at least with the methods proposed till now. Because of the complexity of this type of selection and of all possible scenarios under which it can happen, there is no estimation about the proportion of selective signatures that could have been created by polygenic selection.

2.4.3 Conclusion

The majority of tests for selection have been created under the models of hard sweep detection, but as shown by Hernandez et al. (2011), Innan and Kim (2004) hard sweeps are not the most common recent selection events. The motivation for using differentiation tests, is that the pattern left by soft sweeps can be confused with a neutral pattern when looking at the selected population alone. The chances of confounding is even higher when looking just at allele frequencies, instead of haplotypes. Innan and Kim (2004) suggested that we could gain some power by comparing this population with the wild ancestral one. In the absence of ancestral population information, if a parental population is large enough and if its divergence from the target population is not too old, we expect that the genetic patterns observed in this population will be sufficiently close to those that were present in the ancestral population. Innan and Kim (2008) showed through simulations that "selection causes a drastic change in the pattern of polymorphism in the derived population, but not in the parental population". Sometimes, such ideal parental populations are not available, therefore adding more related populations could help to infer the patterns that existed in the ancestral population. Even if most of domesticated populations have somehow been selected for one or a few specific traits in a specific window of time, a population that has been selected for a different trait than the tested population, can serve as a parental population for the comparison.

Chapter 3

Detection Methods

In this chapter I will shortly review several approaches for detecting signatures of adaptive selection from genomic data, and some specific methods that are representative of these approaches. The number of existing approaches is extremely large, so I will concentrate on those that are related to some extent to the tests I introduce later in the manuscript. These approaches include differentiation tests, which compare the genetic diversity in several related populations, but also those single population tests that attempt to account for linkage disequilibrium information rather than just single marker patterns. Indeed, comparing several related populations should provide a great gain of power for detecting soft sweeps, as discussed in the previous chapter. On the other hand, single population tests that account for linkage disequilibrium information may be extended in order to use data from several populations at the same time, so they are also of interest for us.

I remind, that in this work we always consider an ecological scale of time (Sabeti et al., 2006), so tests between different species are not presented here.

3.1 Detecting selection from single marker allele frequencies in multiple populations

Genetic differentiation between a group of n populations can be measured using the statistic F_{ST} (Wright, 1951), defined as: "the correlation between

random gametes, drawn from the same subpopulation [population here], relative to the total [the group of n populations, here]". Different statistics have been proposed to estimate F_{ST} from the data (Weir and Cockerham, 1984, Weir and Hill, 2002) The sampling distribution of F_{ST} can be estimated genome wide as follows.

Considering one SNP with alleles A and a, let $p = (p_1, \ldots, p_i, \ldots, p_n)'$ be the vector of the A-allele frequencies per population, as in section 2.2.2 and $\overline{p}$ and s_p^2 the sampling estimates of the mean and variance of p.

Genetic differentiation between populations at this locus l can be measured by (Weir and Cockerham, 1984):

$$F_{ST}^l = \frac{s_p^2}{\overline{p}(1-\overline{p})} = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(p_i - \overline{p})^2}{\overline{p}(1-\overline{p})} \tag{3.1}$$

This means that F_{ST}^l is defined as the ratio between the variance of the allele frequency between populations and the maximum possible variance that can be reached when alleles have gone to fixation in all populations.

F_{ST} is the most widely used statistic to measure population differentiation and is the basis to detect selection in many tests. Cavalli-Sforza (1966) pointed out, that while demographic changes such as expansions, bottlenecks or migration will affect the genome entirely, selection would affect it locally. Therefore, neutral loci should have the same F_{ST} distribution along the genome, and loci with excessively high (small) F_{ST} , compared to the rest of the genome, may be subject to directional (balancing) selection (see Beaumont (2005) for a review).

Lewontin and Krakauer (1973) proposed a formal statistical test to detect selection at a locus based on F_{ST}. If genotypes at L SNPs are available, the LK test statistic at locus l is defined as:

$$T_{LK}^{(l)} = \frac{n-1}{\overline{F}_{ST}} F_{ST}^{(l)}, \tag{3.2}$$

where $\overline{F}_{ST}$ is the average of $F_{ST}^{(l)}$ over the L loci. Lewontin and Krakauer (1973) showed that, under neutral evolution, this statistic approximately follows a chi-square distribution with $(n-1)$ degrees of freedom, assuming

that the p_i's are *i.i.d.* and normally distributed. They consequently proposed to reject neutrality based on the quantiles of this chi-square distribution.

The LK test was rapidly criticized (Robertson (1975b)) because of its too strong assumptions. In particular, these assumptions do not hold as soon as the populations do not have a star-like tree, or have unequal effective sizes. High mutation rates or migration would also cause a departure from the chi-square distribution (Lewontin and Krakauer (1975), Nei and Chakravarti (1977), Nei et al. (1977), Nei and Maruyama (1975), Robertson (1975a), Tsakas and Krimbas (1976)).

Indeed, as described in Section 2.2.2, under genetic drift the first two moments of allele frequencies at a single locus are given by:

$$\begin{aligned} \mathbb{E}(p) = &\quad p_0 \mathbf{1}_n \\ Var(p) = &\quad \mathcal{F} p_0 (1 - p_0), \end{aligned} \tag{3.3}$$

where p_0 is the allele frequency in the ancestral population, $\mathbf{1}_n$ is a n-vector of 1's and $\mathcal{F}$ is the kinship matrix defined in (2.8). From this formula we can see that the p_i's are independent only if the non diagonal elements of $\mathcal{F}$ are 0, which implies a star-like evolution. Besides, the p_i's can only be identically distributed if the diagonal terms (the F_i's) are all equal, implying that the tree has equal branch lengths. To relax these hypotheses, Bonhomme et al. (2010) proposed an extension of the LK test accounting for any possible $\mathcal{F}$ matrix. More details about this method are provided in the next section.

Besides, assuming that allele frequencies are normally distributed is reasonable for intermediate p_0 values, but not for extreme ones (close to 0 or 1) where the fixation probability becomes important and stops the random evolution due to drift. Beaumont and Nichols (1996) showed for instance that the distribution of F_{ST} also depends on the heterozygosity. They consequently proposed to account for heterozygosity, rather than for allele frequency when looking for locus with extreme F_{ST} values. In the same spirit, Vitalis et al. (2001) proposed a robust way to evaluate the genetic differentiation between a pair of populations, accounting for the heterozygosity variation among locus. They then conduct simulations, conditioning on the population branch lengths and on the meta population allele counts, to get

the expected joint distributions of the branch lengths and detect outlier loci from this distribution.

3.1.1 Two step methods

As we saw in the previous section, demographic effects affect the distribution of F_{ST} values. But, once they are controlled, we can hope to detect the loci under selection using an outlier approach. The demographic parameters can be either estimated from independent markers assumed to be neutral, like neutral microsatelites, or using all available SNPs. This second strategy is based on the assumption that only a small proportion of these SNPs are affected by selection, and that these few SNPs should not bias too much the estimation of demographic effects, which in contrast leave a genome-wide signature.

Li et al. (2012) reviewed the potential difficulties that can be encountered when trying to disentangle demographic effects from selective ones. For example, in drosophila it seems that selection acts continuously, so it is expected that due to hitch-hiking and interferences between the selected regions, a neutral demography would not be easy to infer. Based on previous studies, Hahn (2008) claimed that "anywhere between 30% and 94% of all amino acid substitutions were fixed by adaptive natural selection". On the other hand, Wright et al. (2005) compared a domesticated modern maize population with its wild ancestor and estimated that around $4 - 10\%$ of the genome has been selected since domestication. This last scenario is closer to that of domesticated animal species, where we look for recent selection events.

$\mathcal{F}$-LK test

A natural two step extension to the LK-test, when testing hierarchically structured populations, is the $\mathcal{F}$-LK test (Bonhomme et al., 2010). The $\mathcal{F}$ in this name comes from the kinship matrix $\mathcal{F}$, which is estimated in a first step, to account for the population structure. Once $\mathcal{F}$ is estimated, it can be included in the statistic used for detecting selection in the data.

In this test $\mathcal{F}$ is computed using the Reynolds distance.

Reynolds distance

An estimation of $\mathcal{F}$ is based on the Reynolds distance matrix DR.

Let M be the matrix shown in Table 3.1.1, whose rows are the populations, and the columns are the frequencies of all alleles at all loci.

$$M = \begin{array}{c} \\ pop\ 1 \\ \vdots \\ pop\ n \end{array} \begin{pmatrix} \overbrace{p_{11}^{(1)} \dots p_{1A}^{(1)}}^{locus\ 1} & \cdots & \overbrace{p_{L1}^{(1)} \dots p_{LA}^{(1)}}^{locus\ L} \\ \vdots & \vdots & \vdots \\ p_{11}^{(n)} \dots p_{1A}^{(n)} & \cdots & p_{L1}^{(n)} \dots p_{LA}^{(n)} \end{pmatrix}$$

Table 3.1: Data organization

If l denotes the locus and a the alleles, the Reynolds distance can be calculated as:

$$d_{ij} = \frac{1}{2} \frac{\sum_l \sum_a (p_{la}^{(i)} - p_{la}^{(j)})^2}{\sum_l (1 - \sum_a p_{la}^{(i)} p_{la}^{(j)})}$$

where L is the number of loci. The matrix DR is defined as the matrix whose elements are the d_{ij}, $DR = (d_{ij})$

To build the $\mathcal{F}$ matrix the lengths of the branches of the phylogenetic tree built using the neighbor joining method (Saitou and Nei (1987)) are used. This method is based on clustering principles and requires knowledge of the distance between each pair of populations (d_{ij}'s). In order to be able to root the tree, an outgroup is also needed.

The test

If the p_i's are normally distributed, it follows immediately from Equation (3.3) that the quadratic form in p:

$$T_{\mathcal{F}-LK}(p_0) = (p - p_0 \mathbf{1}_n)' Var(p)^{-1} (p - p_0 \mathbf{1}_n)$$

follows a chi-square distribution with n degrees of freedom. If the ancestral allele frequency p_0 was known, this would be an interesting statistic for testing if the p_i's evolved under neutrality.

However, p_0 is generally unknown, so it has to be estimated from the data. Since the p_i's are not *i.i.d.*, the average frequency $\overline{p}$ is no longer an optimal estimator of p_0, so Bonhomme et al. (2010) used the linear unbiased estimator with minimal variance:

$$\hat{p}_0 = \frac{\mathbf{1}_n' \mathcal{F}^{-1} p}{\mathbf{1}_n' \mathcal{F}^{-1} \mathbf{1}_n} \tag{3.4}$$

Note that this estimator is not the maximum likelihood estimator, even under the normality assumption.

Replacing p_0 by $\hat{p}_0$, Bonhomme et al. (2010) defined the $T_{\mathcal{F}-LK}$ statistic as :

$$T_{\mathcal{F}-LK} = (p - \hat{p}_0 \mathbf{1}_n)' \hat{Var}(p)^{-1} (p - \hat{p}_0 \mathbf{1}_n) \tag{3.5}$$

and showed that it follows approximately a chi-square distribution with $n-1$ degrees of freedom. Below I briefly report some important steps of this proof.

Denoting:

$$\hat{p}_0 = w' p, \tag{3.6}$$

with:

$$w = \frac{\mathcal{F}^{-1} \mathbf{1}_n}{\mathbf{1}_n' \mathcal{F}^{-1} \mathbf{1}_n} \tag{3.7}$$

the first two moments of $\hat{p}_0$ are:

$$\begin{aligned} \mathbb{E}(\hat{p}_0) = &\quad w' \mathbb{E}(p) = p_0 \\ Var(\hat{p}_0) = &\quad w' Var(p) w \\ = &\quad \frac{p_0(1-p_0)}{\mathbf{1}_n' \mathcal{F}^{-1} \mathbf{1}_n} \end{aligned}$$

and it follows that

$$\mathbb{E}(\hat{p}_0(1-\hat{p}_0)) = p_0(1-p_0)\left(1 - \frac{1}{\mathbf{1}_n'\mathcal{F}^{-1}\mathbf{1}_n}\right) \tag{3.8}$$

As $Var(p) = \mathcal{F}p_0(1-p_0)$, to estimate this variance one needs to estimate $p_0(1-p_0)$. From equation (3.8), we see that this quantity can be estimated without bias by $\hat{p}_0(1-\hat{p}_0)\left(1 - \frac{1}{\mathbf{1}_n'\mathcal{F}^{-1}\mathbf{1}_n}\right)^{-1}$. Consequently, one can re-write the statistic as:

$$T_{\mathcal{F}-LK} = \frac{(p-\hat{p}_0\mathbf{1}_n)'\mathcal{F}^{-1}(p-\hat{p}_0\mathbf{1}_n)}{\hat{p}_0(1-\hat{p}_0)\left(1 - \frac{1}{\mathbf{1}_n'\mathcal{F}^{-1}\mathbf{1}_n}\right)^{-1}} \tag{3.9}$$

It can be shown that:

$$\begin{aligned}
\mathbb{E}(T_{\mathcal{F}-LK}) &\approx \frac{\mathbb{E}((p-\hat{p}_0\mathbf{1}_n)'\mathcal{F}^{-1}(p-\hat{p}_0\mathbf{1}_n))}{\mathbb{E}\left[\hat{p}_0(1-\hat{p}_0)\left(1 - \frac{1}{\mathbf{1}_n'\mathcal{F}^{-1}\mathbf{1}_n}\right)^{-1}\right]} = n-1 \\
Var(T_{\mathcal{F}-LK}) &\approx \frac{Var((p-\hat{p}_0\mathbf{1}_n)'\mathcal{F}^{-1}(p-\hat{p}_0\mathbf{1}_n))}{\mathbb{E}^2\left[\hat{p}_0(1-\hat{p}_0)\left(1 - \frac{1}{\mathbf{1}_n'\mathcal{F}^{-1}\mathbf{1}_n}\right)^{-1}\right]} = 2(n-1)
\end{aligned} \tag{3.10}$$

The normality assumption further implies that $T_{\mathcal{F}-LK}$ follows approximately a χ^2_{n-1} distribution under neutral evolution.

Consequently, if the $T_{\mathcal{F}-LK}$ observed at one locus shows a significant departure from this distribution, the allele frequencies at this locus are not compatible with a neutral evolution whose demography is summarized by the kinship matrix. As this matrix was computed from the data, the most likely hypothesis is that selection has been acting on that locus.

The main improvement of the $\mathcal{F}$-LK test compared to other existing tests for detecting selection was to account for hierarchically structured populations. Accounting for unequal branch lengths is also important, but several alternative methods already did that, instead, there are not too much that account for correlations between frequencies, due to common evolution. Below I illustrate on a small example why accounting for the hierarchical structure of populations is very important in this context.

Example 3.1. *I simulated the allele frequency at one locus under a 3 population model, starting from a fixed initial frequency p_0 in the ancestral population. Four scenarios were simulated, with and without hierarchical structure and with and without selection. In the star-like evolution tree scenario (without structure), 3 populations evolved independently from the ancestral one during 200 generations. In the hierarchically structured scenario, 2 populations first evolved independently during 100 generations. Then, one of these split into two populations, leading to a total of 3 populations which further evolved during 100 generations. For both scenarios, when selection was simulated it was introduced in population 1, from the 100th generation to the final generation.*

Frequencies were computed for all generations and in all populations. As the evolution history is known here, I could derive the kinship matrix theoretically and did not need to compute it from the data. Similarly, p_0 was fixed and did not need to be estimated.

To illustrate the importance of considering or not the population structure when computing the test for selection, I computed the p-values of the $\mathcal{F}$-LK test using either the kinship matrix $\mathcal{F}$ or the diagonal matrix $F_{ST}\mathcal{I}$. Note that this second strategy is equivalent to the Lewontin and Krakauer test. In the star-like tree scenario simulated here, the 3 branch lengths were equal. In this case the LK and $\mathcal{F}$-LK statistics are equal, so one single p-value was computed.

On Figure 3.1, we observe that when populations evolve under a star-like topology, their allele frequencies evolve completely independently. But when populations evolve under a hierarchically structured topology (Figure 3.2), the allele frequencies in populations 1 and 2 begin to differentiate only after 100 generations, so the differentiation level that they can reach is smaller than in the star-like scenario. On the other hand, we expect a much larger differentiation between population 3 and populations 1 or 2 than between populations 1 and 2. If we take this structure into account in the test ($\mathcal{F}$-LK), the p-value that we get would not lead to reject the neutral hypothesis, but if we do not consider the structure then population 3 is too differentiated from populations 1 and 2 and we obtain a relatively small p-value (0.07),

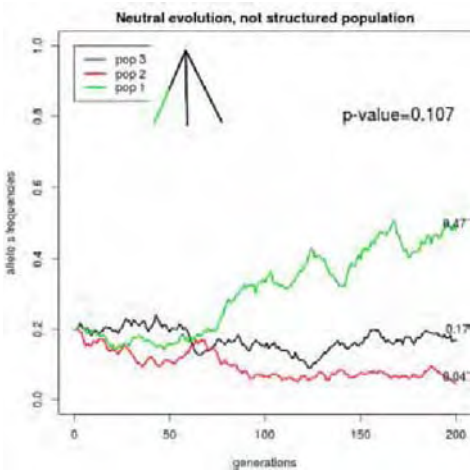


Figure 3.1: **Star-like topology, neutral evolution:** The p-values were computed for the $\mathcal{F}$-LK statistic using the kinship matrix $\mathcal{F}$, which is diagonal in this case. Final allele frequencies in each population are written at the end of the frequency evolution curve.

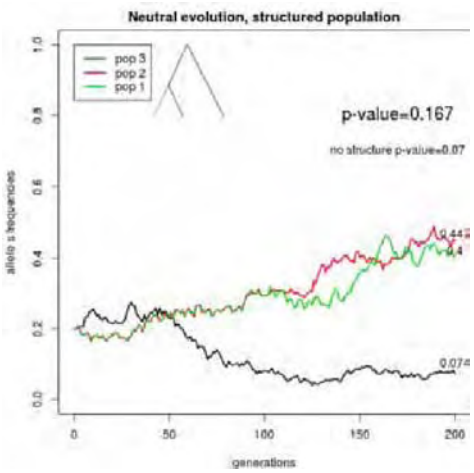


Figure 3.2: **Hierarchically structured topology, neutral evolution:** The p-values were computed for the $\mathcal{F}$-LK statistic using either the kinship matrix $\mathcal{F}$ (above) or the matrix $F_{ST}\mathcal{I}$ (which is equal to the diagonal of $\mathcal{F}$). Final allele frequencies in each population are written at the end of the frequency evolution curve.

which could lead to reject the neutral evolution hypothesis, creating a false positive.

When selection occurs in population 1 under a star-like topology (Figure 3.3), the differentiation has to be large to be detected, because we expect a lot of variation just because of the independent evolution between population. In the structured scenario selection occurs in population 1 (Figure 3.4), which is less differentiated from population 2, so smaller variation could already indicate the presence of selection. In this case the p-value accounting for structure is 0.048, so we probably reject the neutral evolution hypothesis and we have a true positive. Again, if we do not consider the structure, the differentiation does not seem too high and the neutral hypothesis is not rejected (p-value=0.241), leading to a false negative.

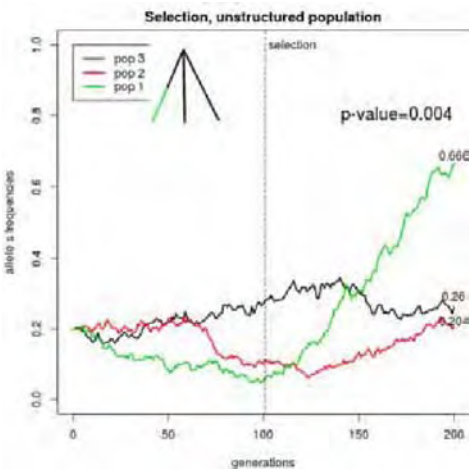


Figure 3.3: **Star-like topology, evolution with selection in population 1:** Evolution under selection is indicated in green in the tree. See Figure 3.1 for other details.

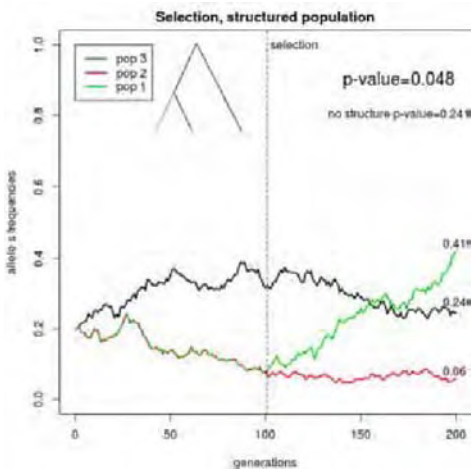


Figure 3.4: **Hierarchically structured topology, evolution with selection in one branch:** Evolution under selection is indicated in green in the tree. See Figure 3.2 for other details.

In conclusion, not considering the structure underlying the populations evolution, could both lead to an excess of false positives and false negatives.

Tree-based likelihood ratio test

Bhatia et al. (2011) proposed a test that is similar in spirit to the $\mathcal{F}$-LK test, but focuses on datasets with 3 populations and considers the unrooted tree formed by these 3 populations. As in $\mathcal{F}$-LK, they first estimate the branch lengths of the unrooted tree using the pairwise differences of allele frequencies at all loci, and then detect the loci where allele frequencies are not compatible with these branch lengths. One important difference is that they consider the branch lengths from the central node and the allele frequency $\hat{p}_c$ in this central node, while $\mathcal{F}$-LK considers the branch lengths from the root of the tree (i.e. the ancestral population) and the allele frequency in this root. But the allele frequency in the central node is estimated by weighting the alleles frequencies of the sampled populations according to the branch lengths, similar to what is done in $\mathcal{F}$-LK.

In contrast with $\mathcal{F}$-LK, Bhatia et al. (2011) test for selection at one locus independently in each population, using the likelihood ratio based statistic:

$$2ln(LRT) = \frac{D^2_{i_{SEL}}}{\sigma^2_{D_{i_{SEL}}}},$$

where the subscript $i_{SEL} \in \{1, 2, 3\}$ refers to the tested population and $D_{i_{SEL}} = p_i - \hat{p}_c$. In this formula $\sigma^2_{D_{i_{SEL}}}$ is the expected variance of $D_{i_{SEL}}$ and is equal to $\hat{p}_c(1 - \hat{p}_c)(2F^i_{ST} + 1/n_i)$, where $(2F^i_{ST} + 1/n_i)$ is the branch length according to Bhatia et al. (2011), n_i is the sample size and F^i_{ST} is the re-estimate of F_{ST} between population i and the central population.

A Bonferroni correction for multiple testing is applied to the p-values, to account for the fact that 3 tests are computed at each locus.

Extending this test to more that 3 populations requires to find out the topology of the unrooted tree and to decide where to place the central node. Although the number of possible unrooted tree topologies explodes when the number of populations increases (there are $(2n-5)!/[2]^{n-3}(n-3)!]$ topologies for n populations, as pointed out by the authors), the first problem could be solved using standard clustering or phylogeny heuristics. On the other hand, the second problem is more fundamental and can only be solved using rooted

trees as in $\mathcal{F}$-LK, where the central node is the ancestral node. However, this approach requires an outgroup.

Note also that in a 3-population rooted tree with topology $((1,2),3)$, if the branch connecting populations 1 and 2 with the root is long, using an unrooted tree could penalize population 3. Indeed, the branch from the central node to population 3 will be much longer than the branch connecting the ancestral population to population 3, so the expected allele frequency variance under the unrooted tree model will be too high. But this effect should be small in cases where F_{ST} is not too high, and the authors proposed the test for closely related populations, with typically $F_{ST} < 0.01$.

The authors also proposed to account for admixture between populations, recalculating the allele frequencies in the admixed population by subtracting a weighted frequency from the original frequency.

3.2 Bayesian methods

Several Bayesian methods have been proposed for detecting selection based on the genetic differentiation between populations. The concept of F_{ST} is also central in these methods, where it is generally decomposed in a population-specific effect and a locus-specific effect. Similar to the two step methods described above, Bayesian methods are also suited to test several populations simultaneously, but the difference is that they also estimate population specific effects (the F_is) and locus specific effects simultaneously.

A seminal work on this approach was proposed by Beaumont and Balding (2004), based on the work of Beaumont and Nichols (1996). Based on this work Foll and Gaggiotti (2008) implemented BayeScan, a widely used software to detect loci under selection. Gautier et al. (2009), Riebler et al. (2008), Excoffier et al. (2009) and Foll and Gaggiotti (2008) developed detection methods based on the same ideas, but using different prior distributions for the parameters, or different methods to estimate the posterior distributions like Monte-Carlo Markov Chains (MCMC) or reverse-jumping MCMC. Gompert and Buerkle (2011) took advantage of the uncertainty that Bayesian approaches permit to account for possible errors arising from next-generation

sequencing. To illustrate the spirit of all these methods I will follow Gautier et al. (2009).

Let x_{li} be the allele count of the A allele at locus l and in population i. The conditional distribution of x_{li} given the true allele frequency α_{li} is assumed to be binomial with parameters $2n_i$ and α_{li}, where n_i is the number of genotyped individuals:

$$x_{li}|\alpha_{li}, n_i \sim_{iid} \mathcal{B}(2n_i, \alpha_{li})$$

Following Nicholson et al. (2002), α_{li} is assumed to be sampled from a truncated normal distribution on the (0,1) segment, that is:

$$\alpha_{li}|c_i, \pi_l \sim_{iid} \mathcal{N}_T\left(\pi_l, c_i\pi_l(1-\pi_l)\right)$$

where $\pi_l \sim_{iid} Beta(a_\pi, b_\pi), \quad c_i \sim_{iid} Beta(a_c, b_c)$

To test for selection, Gautier et al. (2009) proposed to compute the Posterior Predictive P-values (PPP-values), that are the Bayesian counterparts of the frequentist p-values. In the same sense as p-values, small (large) PPP-values, correspond to positive (balancing) selection.

Note that the distribution of the true sample allele frequencies (α) in this model, is almost the same as in Bonhomme et al. (2010), with π_l being equivalent to the ancestral frequency p_0 at each locus, and c_i being equivalent to the inbreeding coefficient F_i in each population. The distribution proposed in Nicholson et al. (2002) assumes a star-like tree (the c_is are *i.i.d.*), so the test suffers from the same problems as LK does when looking for selection in a hierarchically structured population. Indeed, Bonhomme et al. (2010) showed using simulations that the $\mathcal{F}$-LK test was more powerful than the test proposed by Foll and Gaggiotti (2008) when the populations were hierarchically structured.

3.3 LD methods

In the previous section I described several differentiation tests and pointed out the importance of considering the dependencies between populations, which are generated by their hierarchical structure. All the tests I considered were based on allele frequency data at single loci, and used the loci independently of each other. However, as discussed in Section 2.4, selection at one locus impacts genetic diversity in the whole neighborhood of the locus, and generally leaves specific haplotypic patterns. In this section I will review several existing methods that try to account for linkage disequilibrium and haplotype structure when looking for signatures of adaptive selection. These methods can be divided in two quite different approaches. The first one consists in cumulating tests that have been first obtained using single locus allele frequency data, while the second is directly based on haplotype data. The type of observed data is very important here. Indeed, methods that cumulate single locus statistics can be applied to data at the population level, as that obtained from pool sequencing, while methods that are based on haplotype lengths or haplotype counts can not. More generally, all the methods described in this section require relatively dense genomic data, otherwise the linkage disequilibrium between adjacent SNPs is negligible and can not be exploited.

3.3.1 Smoothing methods

Genome scans based on single locus statistics show a very high variability of the signal between adjacent markers, whatever the statistic used, due to the high stochasticity in the evolution of allele frequencies around their expected value, for instance F_{ST} ((Weir et al., 2005), Figure 3.5) and $\mathcal{F}$-LK (Figure 3.6).

Weir et al. (2005) proposed to reduce the noise in F_{ST} scans by combining data from several adjacent markers. They studied the HapMap (Consortium, 2005) and Perlengen (Hinds et al., 2005) datasets with 3 and 4 populations each, and observed that the F_{ST} followed approximately a χ^2 distribution with 2 or 3 degrees of freedom. To clarify graphical representations of the

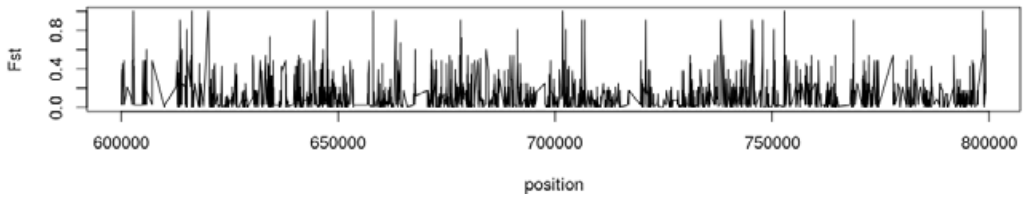


Figure 3.5: F_{ST} **between two divergent lines of quail:** F_{ST} between positions 0.5 and 0.8Mb of chromosme 1 between two divergent quail lines that were pool-sequenced. Each pool contained 10 indiviuals. F_{ST} shows high stochasticity, even in small regions.

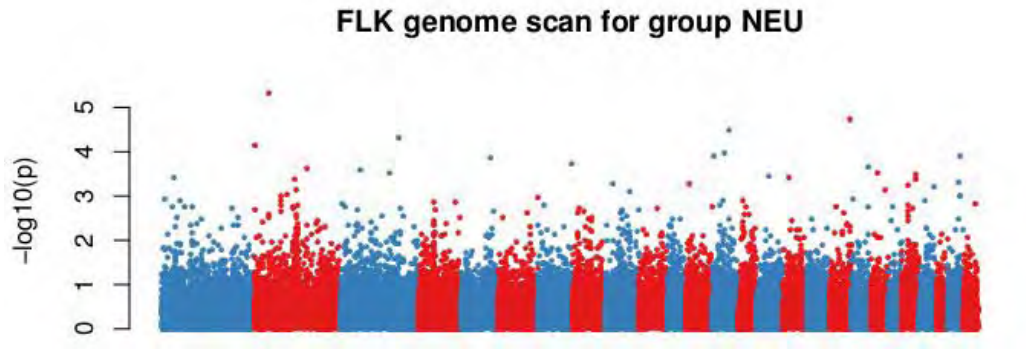


Figure 3.6: **Genome Scan using $\mathcal{F}$-LK of North-European sheep**

statistic genome wide, they proposed to average the F_{ST} values over 5Mb windows. The distribution of the new averaged statistic followed approximately a normal distribution, with reduced variance compared with that of the initial χ^2. The window size was chosen in a completely subjective way. The authors mentioned that this choice should actually depend on the recombination rate in the region, but also acknowledged that it was not clear, how to do it in this context.

Weir et al. (2005) also computed the expected correlation between the variations of allele frequencies at close loci. They showed that this correlation depends on the initial linkage disequilibrium (in the ancestral population), which is not observed and can only be predicted from the linkage disequilibrium in the present generation. These predictions generally underestimate

the true correlations. They acknowledge that the estimated correlations and the correlations of F_{ST} values, are very similar. So when looking for clusters of high F_{ST} values to detect selection, one has to be careful, because this clusters can be originated just by linkage disequilibrium. This point will be discussed further in Chapter 6.

Weir et al. (2005) proposed to consider a window as exceptionally extreme if its average F_{ST} value differs by more than three empirical standard deviations from the chromosome mean window value. They acknowledged that this procedure has no specific statistical significance, but expect that such extreme values are beyond the values that can be reached under neutral evolution, and should be explained by selection.

Several authors have followed the averaging approach of Weir et al. (2005) (Oleksyk et al. (2008)), with subjective variations in the statistic used, the window size or the use of overlapping or non overlapping windows.

XP-CLR

Rather than averaging over windows, Chen et al. (2010) proposed a composite likelihood approach called XP-CLR, where they compute the marginal likelihood of the allele frequencies at each SNP under a given evolution model, and then multiply these likelihoods over windows of k consecutive SNPs. While computing the average or the maximum F_{ST} over a window reduces the variance of F_{ST} values and smooths their distribution over the genome, but does not really take advantage of the correlated evolution between adjacent SNPs, XP-CLR is meant to take this information into account.

Similar to $\mathcal{F}$-LK and several other tests described in this chapter, XP-CLR is based on the assumption that under neutrality the frequency p_i of an allele in population i follows a normal distribution $\mathbb{N}(p_0, \omega_i p_0(1-p_0))$, where p_0 is the frequency of an ancestral population and ω_i is the inbreeding coefficient starting from this ancestral population. The ω_i of (Chen et al., 2010) is equal to F_t in Equation (2.4).

Considering two populations that diverged from a common ancestral population, the allele frequencies p_1 and p_2 in these populations are both nor-

mally distributed with mean p_0, where p_0 is the allele frequency in the ancestral population. One key idea is then to use the fact that the allele frequency evolution process can be reversed : we can think that it began in population 2, with an initial frequency p_2, went back to the ancestral population, where it reached a frequency p_0, and then continued until population 1 to reach the frequency p_1. Besides, as p_0 is unknown, we can actually skip this evolution step and just model the process of going from p_2 to p_1, assuming that $p_1 \sim \mathbb{N}(p_2, (\omega_1 + \omega_2)p_2(1 - p_2))$.

Following Maynard Smith and Haigh (1974) and Durrett and Schweinsberg (2004), Chen et al. (2010) also extend this neutral model for a neutral locus that is located in the vicinity of a locus under selection.

They derive an approximate density for p_1 conditional on p_2, on the recombination rate r between the two loci and on the selection coefficient s (this density is denoted $f(p_1|r, s, p_2, w)$). For any window of k loci, they compute the k marginal densities, assuming that a selected locus with selection coefficient s was located in the middle of the window, and multiply these densities to obtain the composite likelihood. They finally test for selection using a likelihood ratio between the alternative hypothesis ($s \neq 0$) and the null hypothesis ($s = 0$). They claimed that the test does not depend on the choice of k, provided the size of the window is large enough.

Chen et al. (2010) analyzed a human dataset with $XP - CLR$, CLR (similar to $XP - CLR$ but for single populations tests), $XP - EHH$ and iHS (both explained in next section). They found several overlapping signals, but commented on two specific signals that were detected by $XP - CLR$ and not by $XP - EHH$. In the first signal 349 of 918 alleles were fixed in one population, so they hypothesized that the selection signature should be really ancient. This could explain why $XP - EHH$ did not detect it, because, as we will see, this test is rather designed to detect recent selection events. There was no information about the genes in the detected region. The second signal, which contains the $NRXN3$ gene, is likely a selection signature from standing variation, because the authors found two haplotypes with outstanding high frequency in the region. Again, it is not surprising that $XP - EHH$ did not detect it, because it is designed to detect hard-sweeps.

Following the Bayesian approach described in Subsection 3.2, Guo et al. (2009) proposed to account for linkage disequilibrium by introducing an autoregressive model when modeling the distribution of allele frequencies. This model conditions the allele frequency at a locus on the allele frequencies at all other loci located in the chromosome, but the correlation between loci are weighted by a function that accounts for the distance between loci.

On one hand, when testing structured populations, this model also suffers from the same problems as the other Bayesian methods do, because they model the inbreeding coefficients independently. On the other hand, it is computationally demanding. Authors did not analyze datasets larger than 3000 SNPs in human data. In simulations they analyzed 1000 SNPs at the same time, but it is not totally clear, how the SNPs are related. They where able to classify if each SNP was neutral or under selection, which is quite surprising, because as they account for linkage disequilibrium, one would expect that they detect zones, and not just one SNP.

3.3.2 Haplotype tests

As described in Section 2.4, we expect selection to leave a certain amount of relatively long haplotypes. Depending on the type of selective sweep we expect to observe in the population either one single long haplotype at very high frequency (hard sweep), or a few long haplotypes at high frequency (soft sweep). Here I will describe several tests that aim to capture these signals, essentially those left by hard sweeps.

Single population tests

Single population haplotype tests are based on the principle that one can distinguish the ancestral allele and the derived allele at each SNP, which is indeed possible if data from outgroup species are available. Under neutral evolution, derived alleles that segregate at high frequency in a population have to be old, because drifting from a frequency close to 0 to a high frequency requires a lot of time. Consequently, the haplotypes carrying high frequency derived alleles are expected to be very short, due to the action of

recombination during a long time period. On the other hand, alleles under strong positive selection can reach the same frequency in a much shorter time, so they can be carried by long haplotypes (Figure 3.7[1]). Finally, derived alleles with low to moderate frequency could be young or old, so the associated haplotypes can be either long or short.

Based on this idea, Sabeti et al. (2002) developed a Long Range Haplotype test (LRH) which involves the notion of Extended Haplotype Homozygosity (EHH). EHH is defined relative to a core region and the derived allele. For a given test locus, EHH is the probability that two extended haplotypes around a given locus are the same, given that they have the same allele at the locus. The support of EHH is thus $[0, 1]$, from no homozygosity at all to complete homozygosity. Assume that M haplotypes are found in the core region, each with C_i chromosomes ($i \in 1 \dots M$). Denoting EHH_i the EHH for haplotype i, we can also define the Relative EHH for haplotype i as the ratio between EHH_i and the average EHH for other haplotypes, i.e.

$$REHH_i = EHH_i / \left[\frac{\sum_{\substack{j=1 \\ j \neq i}}^{M} \binom{C_j}{2} EHH_j}{\sum_{\substack{j=1 \\ j \neq i}}^{M} \binom{C_j}{2}} \right]$$

In this formula the EHH of each haplotype is weighted by the probability that two chromosomes randomly chosen from the whole sample carry this haplotype. The support of REHH is $[0, \infty]$.

Sabeti et al. (2002) proposed to test for selection at one core locus by computing REHH at different distances for this locus and comparing the resulting values with those obtained under a wide range of demographic scenarios. They applied this strategy to only two genes. Performing a genome wide scan using LRH would be difficult, because it would require to identify all core haplotypes on the genome.

As the distance between the core locus and the test locus increases, the

[1] This figure was taken from Voight et al. (2006)

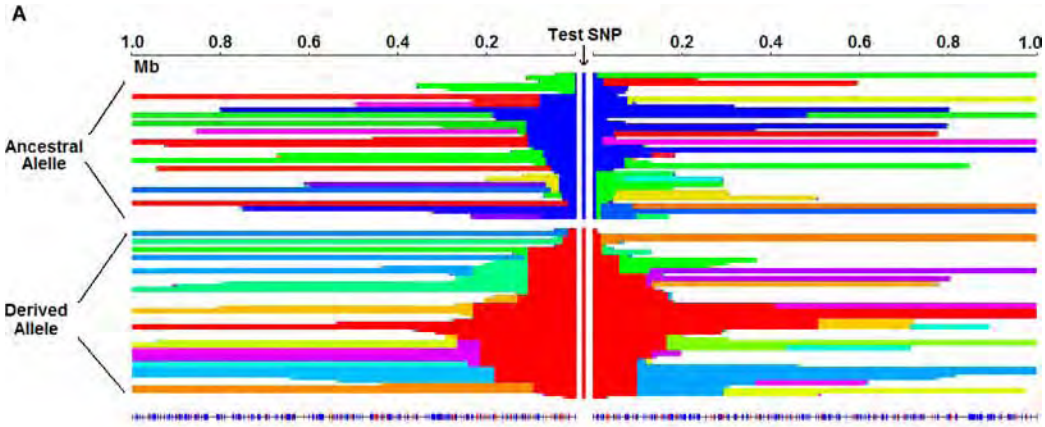


Figure 3.7: **Decay of EHH in Simulated Data for an Allele at Frequency 0.5** Decay of haplotypes in a single region in which a new selected allele (red, center column) is sweeping to fixation, replacing the ancestral allele (blue). Horizontal lines are haplotypes; SNP positions are marked below the haplotype plot using blue for SNPs with intermediate allele frequencies (minor allele > 0.2), and red otherwise. For a given SNP, adjacent haplotypes with the same color carry identical genotypes everywhere between that SNP and the central (selected) site. The left- and right-hand sides are sorted separately. Haplotypes are no longer plotted beyond the points at which they become unique.(This figure and legend were taken from Voight et al. (2006))

probability of sampling two IBD segments between those loci decreases due to possible recombination and mutation events. Consequently, EHH is expected to decrease from 1 (at the core SNP) to 0 (at a sufficiently distant locus). To summarize the information contained in this EHH trajectory, Voight et al. (2006) proposed to integrate it against the distance from the core SNP. They defined iHH as the sum of the two integrals (one from each side of the SNP) from a specific core allele to an EHH threshold of 0.05. Denoting iHH_A (iHH_D) the iHH computed for the ancestral (derived) allele, they further introduced the unstandardized iHS

$$ln\left(\frac{iHH_A}{iHH_D}\right)$$

If EHH has approximately the same value for both alleles, then $\frac{iHH_A}{iHH_D} \approx 1$ and $iHH \approx 0$. Large positive values mean that the ancestral allele is carried

by a long range haplotype, while large negative values mean that the derived allele is carried by a long range haplotype.

Since in general low frequency alleles are expected to be younger as high frequency alleles, and thus to be associated to longer haplotypes, the authors proposed to bin the core SNPs according to their derived allele frequency, and to standardize the statistics within each bin. For a SNP with derived allele frequency p, the standardized statistic is thus given by :

$$iHS = \frac{ln\left(\frac{iHH_A}{iHH_D}\right) - \mathbb{E}_p\left[ln\left(\frac{iHH_A}{iHH_D}\right)\right]}{SD_p\left[ln\left(\frac{iHH_A}{iHH_D}\right)\right]} \tag{3.11}$$

The authors noted that iHS is not a statistical test, but a measure of how unusual the haplotypes are around a given SNP, and advised to look for windows including several extreme iHS values.

One of the advantages of iHS is that it is robust to the heterogeneity of the recombination rate. This comes from the two following features. First, the integral is computed using the genetic distance, so cold spots are naturally down weighted, and hot spots up weighted. Second, it is based on a ratio between two alleles, which serve as internal control of each other (this second argument also applies to LRH).

However, one limitation of iHS is that it would not be able to capture selected alleles at low frequency, or near to fixation. To improve the power of these tests, some authors (Sabeti et al. (2007), Tang et al. (2007)) proposed to compare EHH values between populations. These extensions are described below.

Cross populations tests

A direct extension of iHS in the case of two populations was proposed by Sabeti et al. (2007). In each population, instead of computing the ratio of iHH between the ancestral and derived alleles, they integrate the EHH profiles for the derived allele as in iHH for both populations and then compute the ratio of the quantities obtained for each population. Positive (negative) scores indicate that there was selection in the population corresponding to

the numerator (denominator). Sabeti et al. (2007) called this test XP-EHH, because it is a cross population test.

Tang et al. (2007) proposed an alternative way of measuring haplotype homozygosity between two sites i and j, which in contrast with EHH is not related to the derived allele at the core locus. They defined $EHHS_{i,j}$ as the ratio between the homozygosity between sites i and j, and the homozygosity at site i, i.e.:

$$EHHS_{i,j} = \frac{\mathbb{E}(H_{Oi,j})}{\mathbb{E}(H_{Oi})}$$

To avoid phasing the data, they also proposed to estimate $EHHS_{i,j}$ as the proportion of individuals that remain homozygous for intervals starting at i in both directions. They again integrate the $EHHS_{i,j}$ values against physical distance, and denote iES the resulting integral. Finally, as in (Sabeti et al., 2007), they detect selection using the standardized log-ratio between populations:

$$ln(Rsb_i) = ln\left(\frac{iES_{pop1,i}}{iES_{pop2,i}}\right)$$

which in contrast to the other approaches are standardized independently of the frequency bin. Note also that they integrate against physical distance instead of genetic distance, arguing that using a population ratio is enough, since each population serves as internal control (in regions with low recombination, we will find long haplotypes in both populations so the effects will cancel out).

Cross population tests represent a first step to multiple population tests accounting for haplotype information rather than just allele frequency information. However, because the differentiation measures are defined as ratios between pairs of populations, it seems difficult to naturally extend these tests to more than two populations. The authors of the cross population tests observed that they allow to detect a wider range of selective sweep scenarios compared to related single population tests. Still, these tests detect mostly ongoing or hard sweeps. Only Rsb seems to be able to detect sweeps from standing variation (Tang et al., 2007). Indeed, it does not look for

conserved haplotypes, but for IBD individuals, so the signals provided by distinct extended haplotypes can sum up. On the other hand, two different long haplotypes that would be carried by the same individual would not be counted, because this individual would be heterozygote, so the power for detecting selection from standing variation is still not very high.

3.4 Need for new two step methods

The EHH based tests presented above are mostly designed to detect high frequ

Chapter 4

A new haplotype-based test for detecting signatures of selection

In the previous chapter I pointed out that despite of the variety of available methods for detecting positive selection, there is still a great need for new methods accounting for haplotype information, in particular when testing more than two populations. In this chapter I present a new test, called hapFLK , contributing to filling this gap. It is presented under the form of an article, which has been published in march 2013.

In general, when testing more than two populations, some populations are more closely related than the others. The population structure is hierarchical and can be represented with a kinship matrix. As we saw in the previous chapter, not accounting for this structure can lead to a loss of power and to a higher rate of false positives. To avoid this issue, we therefore decided to build a new haplotype test by including haplotype information in the $\mathcal{F}$-LK test (Section 3.1.1), which to our knowledge is almost the only existing test that accounts for the hierarchical structure of sampled populations. To extend FLK into hapFLK, we took advantage of the clustering model of Scheet and Stephens (2006) (presented in Section 2.3.2), which infers local haplotype clusters for all sampled individuals at each observed SNP position, and included this individual haplotype information in a new multiallelic version of $\mathcal{F}$-LK.

Through simulations we showed that this haplotypic extension results in an increased detection power and in the possibility to detect a wider range of selection scenarios. In a scenario with two populations, we showed that hapFLK also had more power than $XP - EHH$ (Section 3.3.2) especially for detecting selection from standing variation. Similar to FLK, hapFLK assumes a model without migration or admixture, but simulations proved that it was robust to moderate levels of migration, as well as to eventual bottlenecks.

Besides the detection step, we also proposed in this article two strategies the population(s) under selection. One of them is based on a local estimation of the kinship matrix. It is given more emphasis in the article because it seems to be the most user friendly method. But the second approach, which is based on a spectral decomposition of the hapFLK statistic and is described as supplementary information, may also be helpful, for instance to elucidate whether there is interference between several close selection signals. We also present a cluster frequency representation of the signals, that gives a picture of the frequency of the selected haplotype(s) in the sampled populations.

4.1 Article: Detecting Signatures of Selection Through Haplotype Differentiation Among Hierarchically Structured Populations

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

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.
[1]Corresponding 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 multipoint 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 $p = (p_1,\ldots,p_i,\ldots,p_n)'$ 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_is, 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

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

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

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

$\mathcal{F}_{ii}$ 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

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

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 $\bar{F}_{\mathrm{ST}} I_n$, where $\bar{F}_{\mathrm{ST}}$ 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):

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

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

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

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

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

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

$$\begin{aligned} \mathrm{Var}(P) &= \begin{pmatrix} \mathrm{Var}(p_1) & \cdots & \mathrm{Cov}(p_1, p_A) \\ \vdots & \mathrm{Var}(p_a) & \vdots \\ \mathrm{Cov}(p_A, p_1) & \cdots & \mathrm{Var}(p_A) \end{pmatrix} \\ &= \mathcal{B}_0 \otimes \mathcal{F}, \end{aligned} \tag{4}$$

with $\mathcal{B}_0 = \mathrm{diag}(P_0) - P_0 P_0'$. Each diagonal block of Var(P) corresponds to the biallelic covariance matrix for one of he 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 $\hat{P}_0 = (wp_1, \ldots, wp_A)'$. 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 $z_i = (z_i^1, \ldots, z_i^L)$ 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.*,

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

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

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

For each marker ℓ, the multiallelic statistic T_{FLK} is computed according to Equation 3, with a small modification in the derivation of 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 $\mathcal{B}_0 \otimes \mathcal{F}$ was found numerically unstable for small P_{a0} values, which are very common when the number of alleles is large. Consequently, the $\mathcal{B}_0$ matrix is replaced by the identity matrix I_A in Equation 4, leading to the statistic

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

Simulations confirmed that this version of the test was more powerful than the one including $\mathcal{B}_0$ (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 hapF_{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 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 $N_{\mathrm{e}} = 1000$ and $n_{\mathrm{h}} = 4000$ for the two-population case, and $N_{\mathrm{e}} = 2000$ and $n_{\mathrm{h}} = 8000$ 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 $\alpha = 2 \cdot N_{\mathrm{e}} \cdot s = 100$). 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 $T_{F_{\mathrm{ST}}}$, $T_{\mathrm{hap}F_{\mathrm{ST}}}$, 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

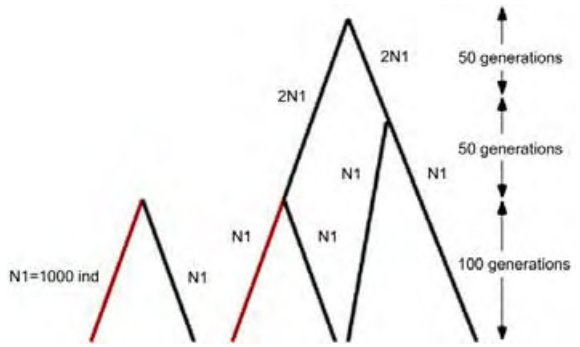


Figure 1 Population trees for the two simulated scenarios. The red branch indicates the selected population and time during which selection acts.

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 recieved 10% of migrants from population 2.

In terms of expected drift at a single SNP, these scenario 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 outgroup 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,

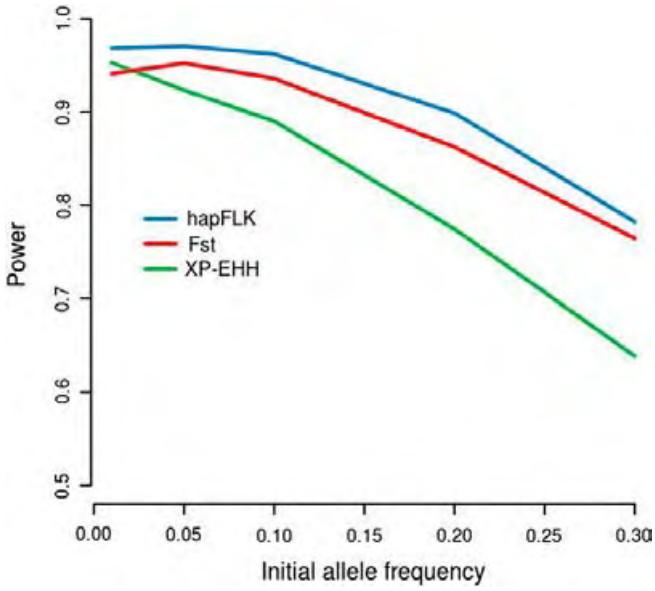


Figure 2 Power of hapFLK, F_{ST}, and XP–EHH as a function of the initial frequency of the selected allele. The power is evaluated at a type I error level of 5%.

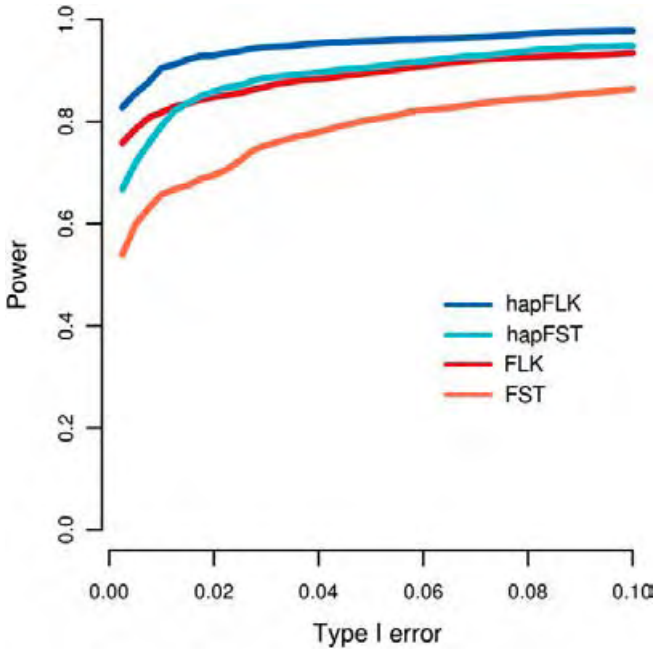


Figure 3 Power of F_{ST}, FLK, hapF_{ST}, and hapFLK in the four-population scenario as a function of the type I error rate. The initial frequency of the selected allele is 1%.

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 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),

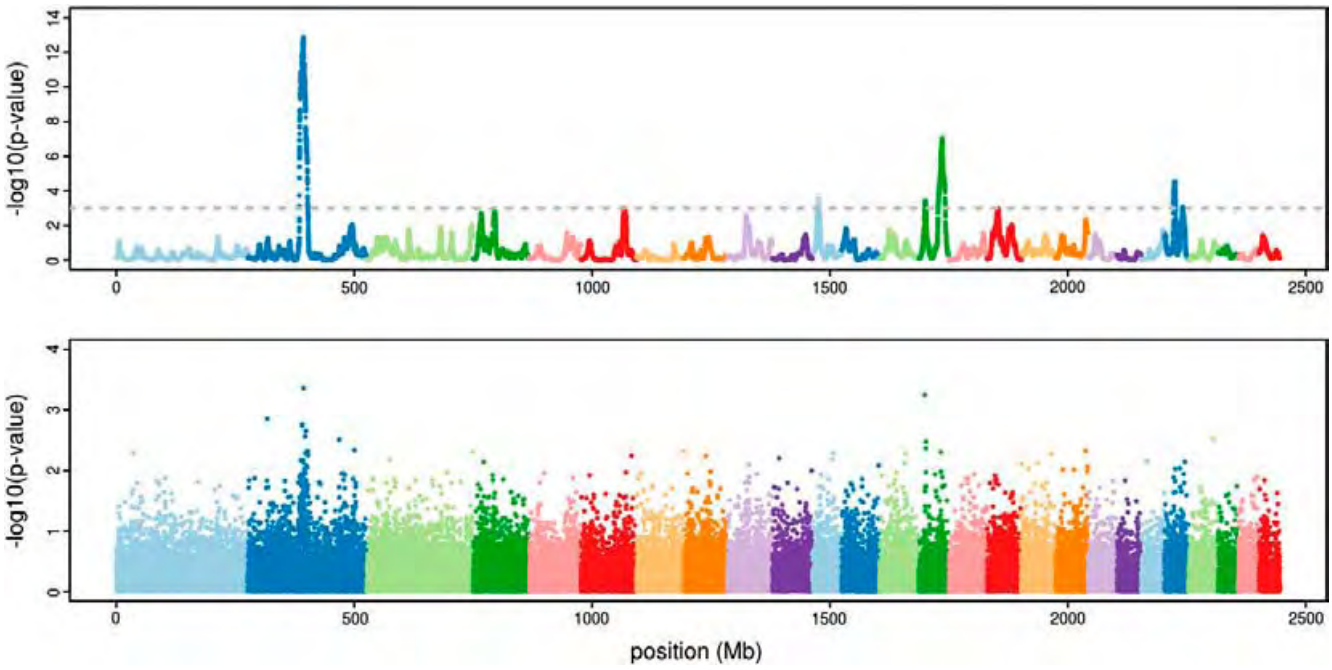


Figure 4 Genome scan for selection in Northern European sheep using a haplotype-based (hapFLK, top) or single SNP (FLK, bottom) test. *x*-axis: position on the genome. *y*-axis: −log10(*P*-value).

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 Rommey, 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 uncertainity 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 hapF_{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.

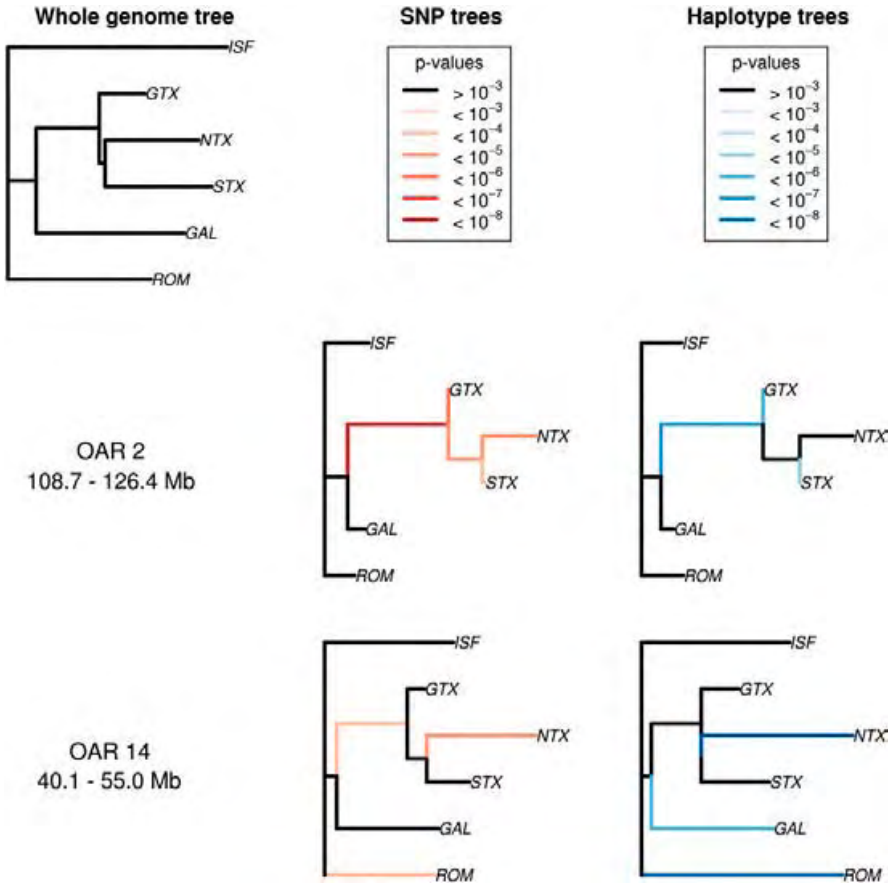


Figure 5 Local population trees estimated in two significant regions in the sheep data set. Population tree of the Northern European sheep populations from the Sheep HapMap Project (top left). Local population trees were estimated using Reynolds distance based on SNPs (left) or haplotype clusters (right). 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.

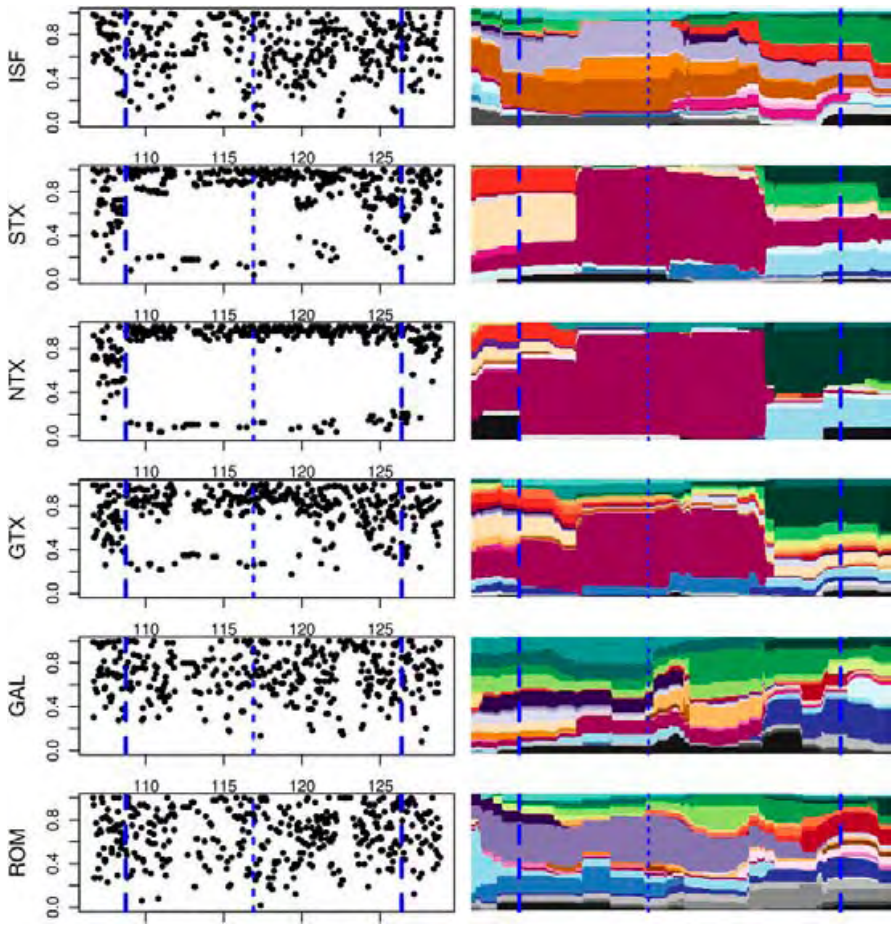


Figure 6 Allele (left) and haplotype cluster (right) frequencies in detected region 1 (chromosome 2) for each of the 6 sheep populations used in the test. Blue bars indicate the limits of the detected region and the position of maximum of the test. The reference allele used for the SNP frequency representation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

DOI: 10.1534/genetics.112.147231

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 indivuals 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

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

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

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

The accumulated drift at generation 100 can then be calculated as

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

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

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

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 decompise 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

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

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 dissecate 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 :

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

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:

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

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 $i \in \{1, \ldots, n\}$ and $j \in \{1, \ldots, L \cdot A\}$ 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 occuring 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_α, 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 has 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 $5, 10, \ldots, 100\%$, 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 Rommey 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.*

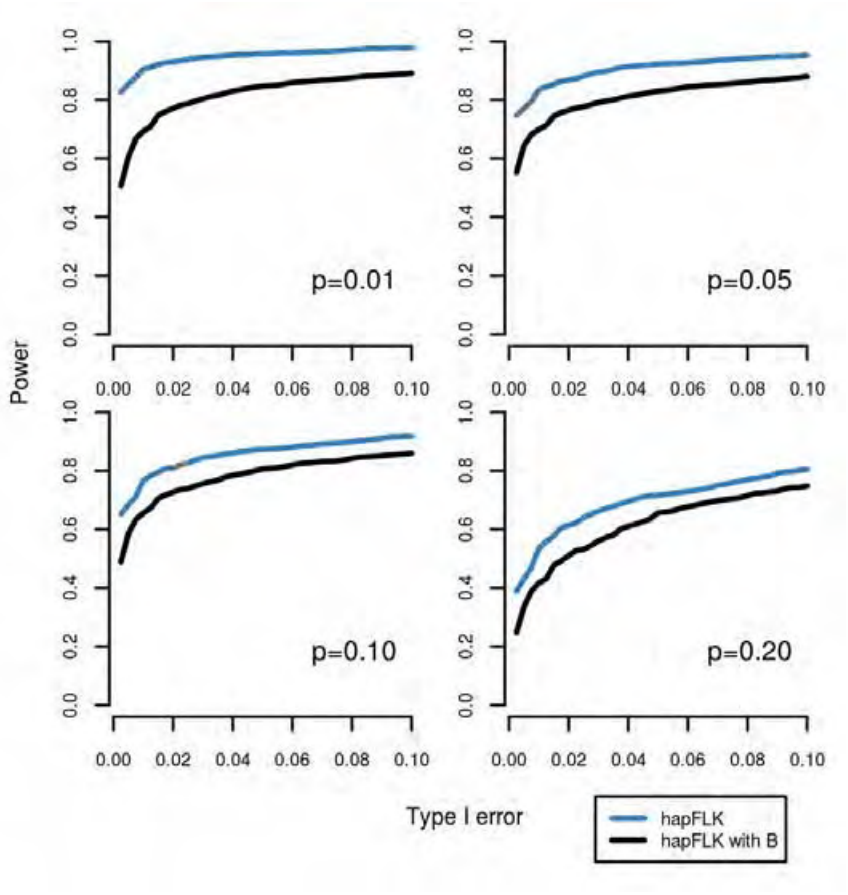


Figure S1. Power of hapFLK using $\mathcal{B}_0$ covariance matrix (black lines) or not (blue lines) for different initial frequency of the selected allele.

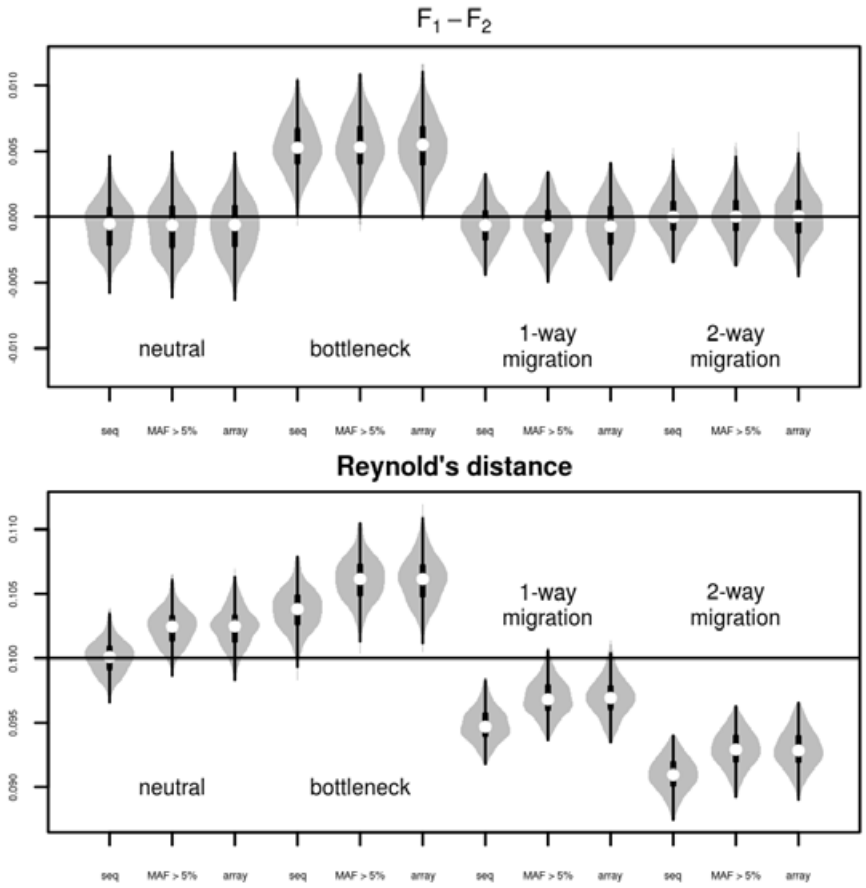


Figure S2. Estimate of difference in branch length (up) and Reynolds genetic distances (bottom) between two populations under different evolutionary scenarios. For each scenario, results are shown for whole SNP data (seq), only common SNPs (MAF $> 5\%$) and genotyping array density (array).

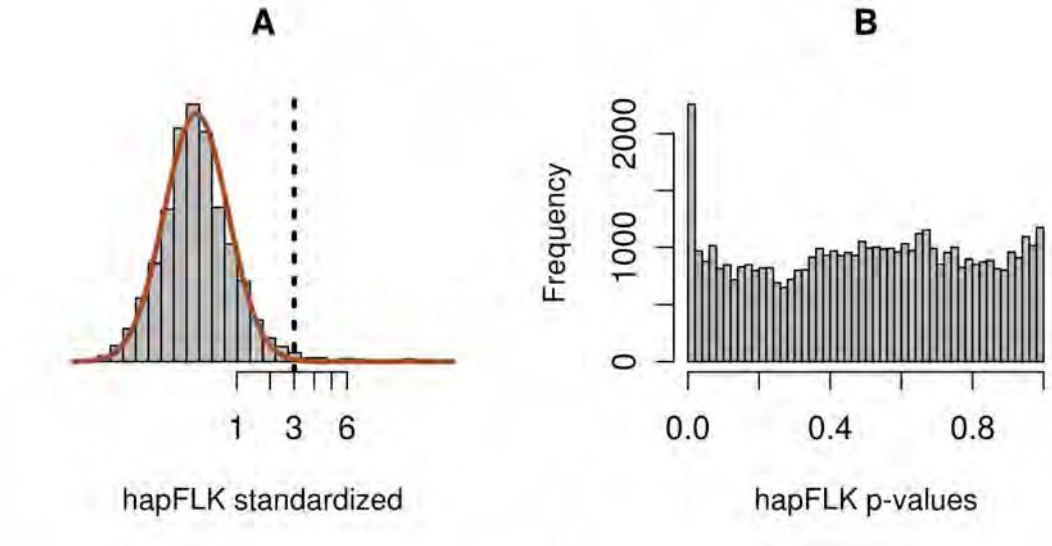


Figure S3. Histogram of the distribution of *hapFLK* and of the fitted normal distribution (A) and histogram of the resulting p-values (B). The dotted line on the left corresponds to the significance p-value threshold of 10^{-3}, corresponding to an FDR of approximately 5%.

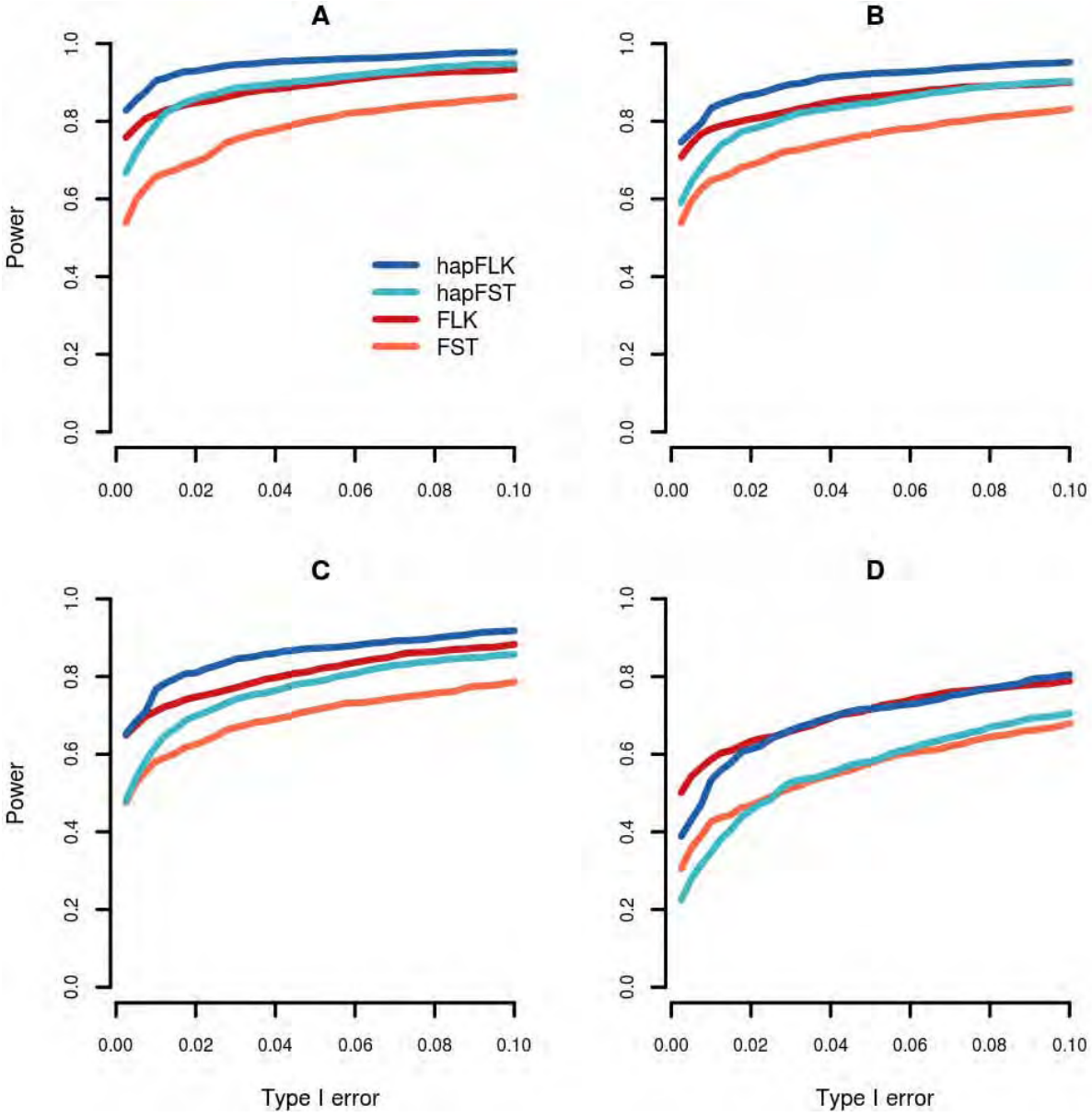


Figure S4. Power of $hapFLK, FLK, hapF_{ST}$ and F_{ST} as a function of type I error. Initial frequencies for the selected allele: A:1%, B:5%, C:10%, D:20%.

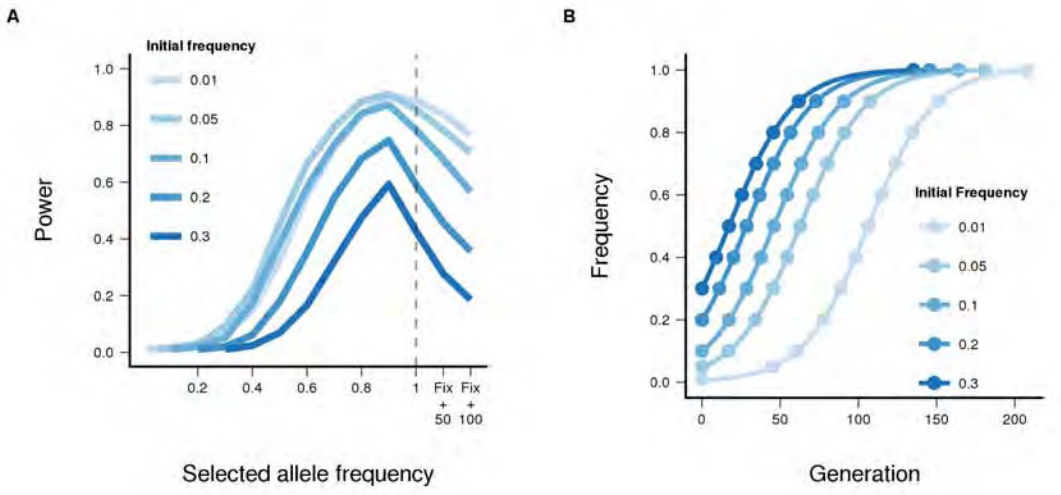


Figure S5. **(A)** Power of hapFLK (type I error rate 1 %) in the two-population scenario for different initial frequencies of the selected allele and different frequency of the selected allele. $Fix + n$ marks a time point corresponding to n generations after fixation of the selected allele. **(B)** 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.

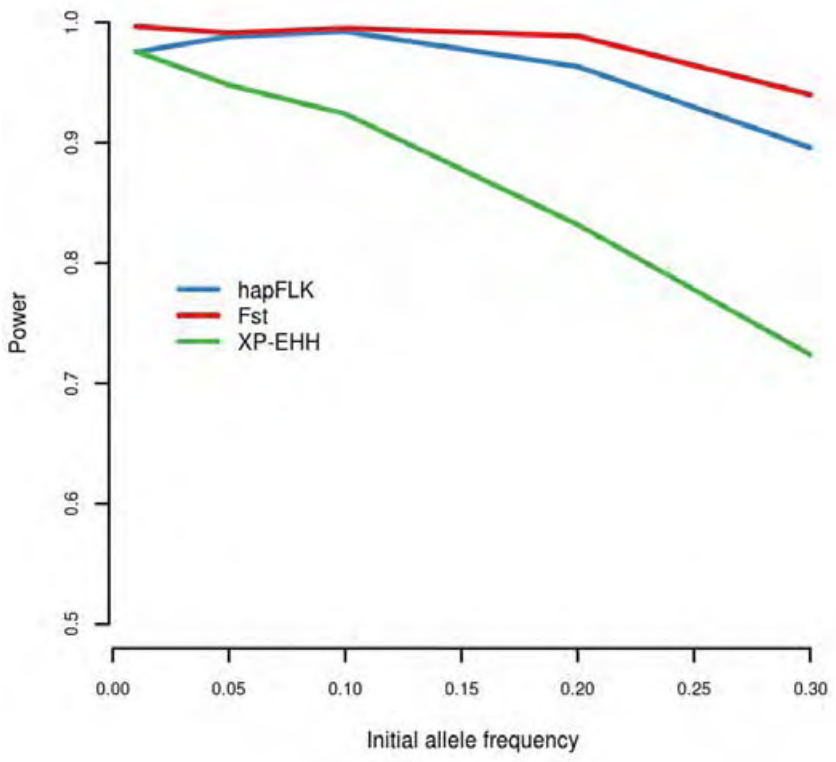


Figure S6. Power of $hapFLK, F_{ST}$ and $XP-EHH$ with sequence data as a function of the initial frequency of the selected allele.
The power is evaluated at a type I error level of 5%.

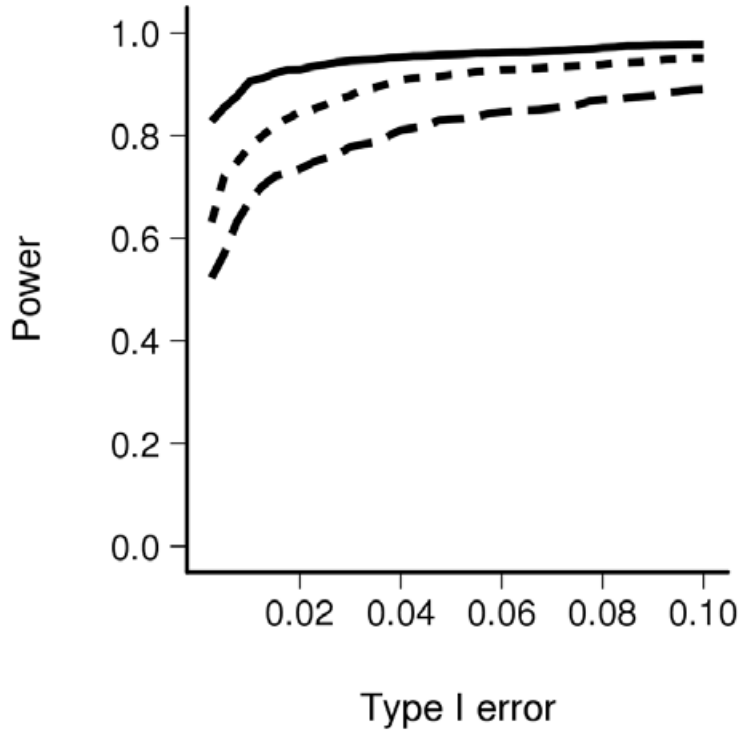


Figure S7. Power of different testing strategies for multiple population data Power of the *hapFLK* test performed on 4 populations jointly (solid line) on 2 independent sub-trees (dotted line) and on all pairs of populations (dashed lines). See details in the text section 2.2.

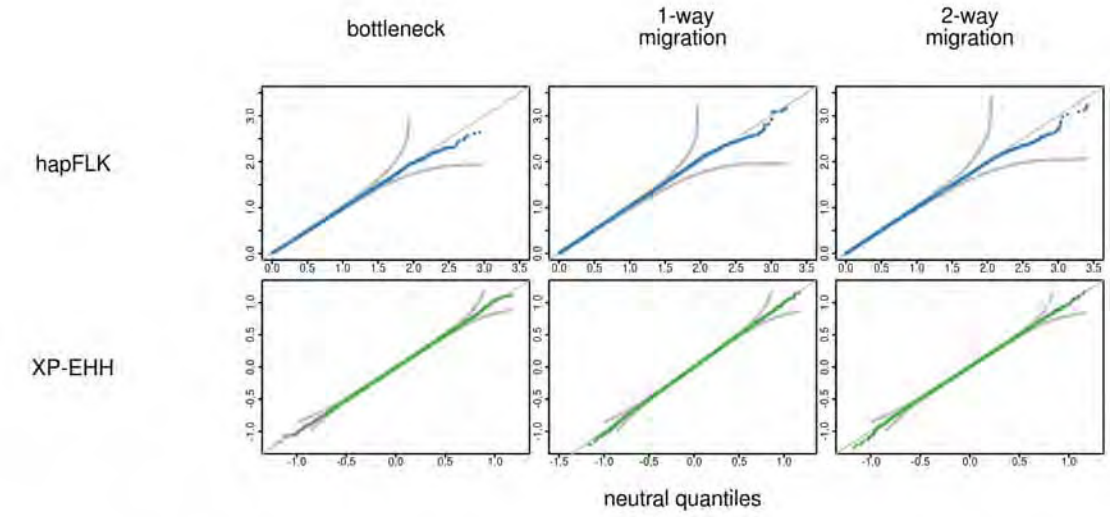


Figure S8. QQ-plots of the distribution of $hapFLK$ and $XP-EHH$ under different evolutionary scenarios against pure drift evolution. Points lying between the two gray lines are consistent with the effect of no effect of complex evolution on the distribution of the statistic.

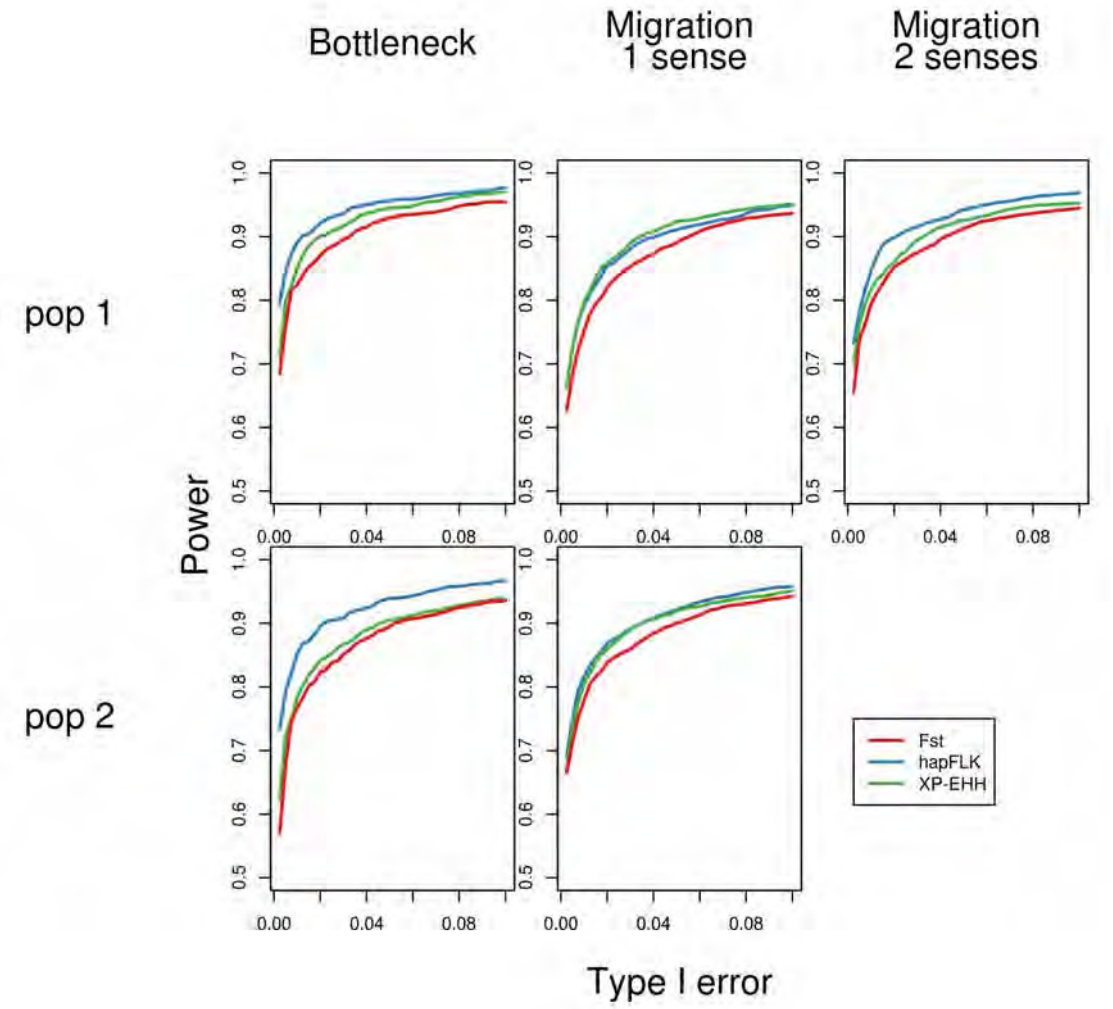


Figure S9. Power of hapFLK under different evolutionary scenarios. Top row: selection in population 1, bottom row: selection in population 2. Note that for the two senses migration selection in population 1 is equivalent to selection in population 2.

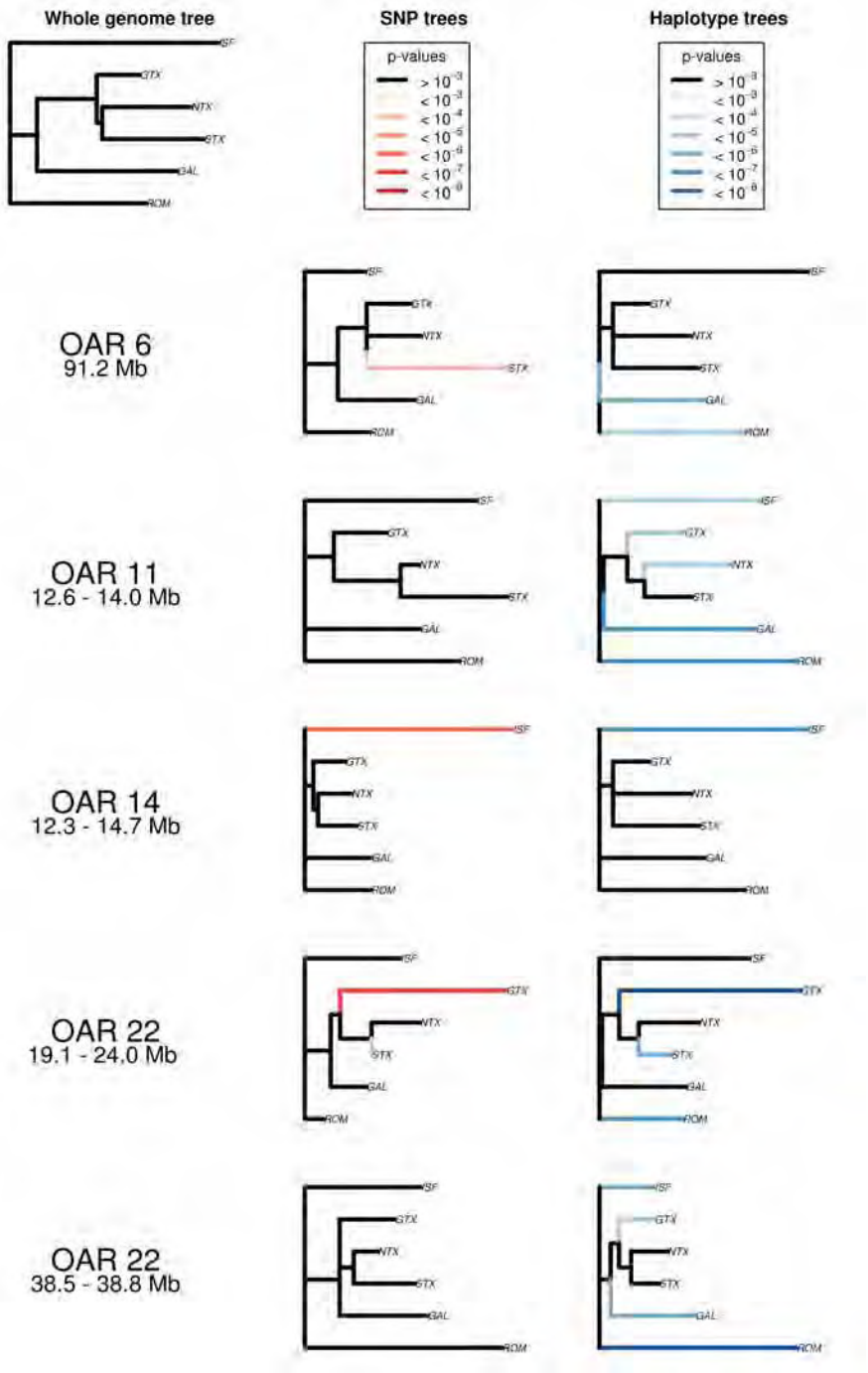


Figure S10. Local trees estimated for significant regions in the Sheep HapMap data analysis. Only regions not discussed in the main text are shown

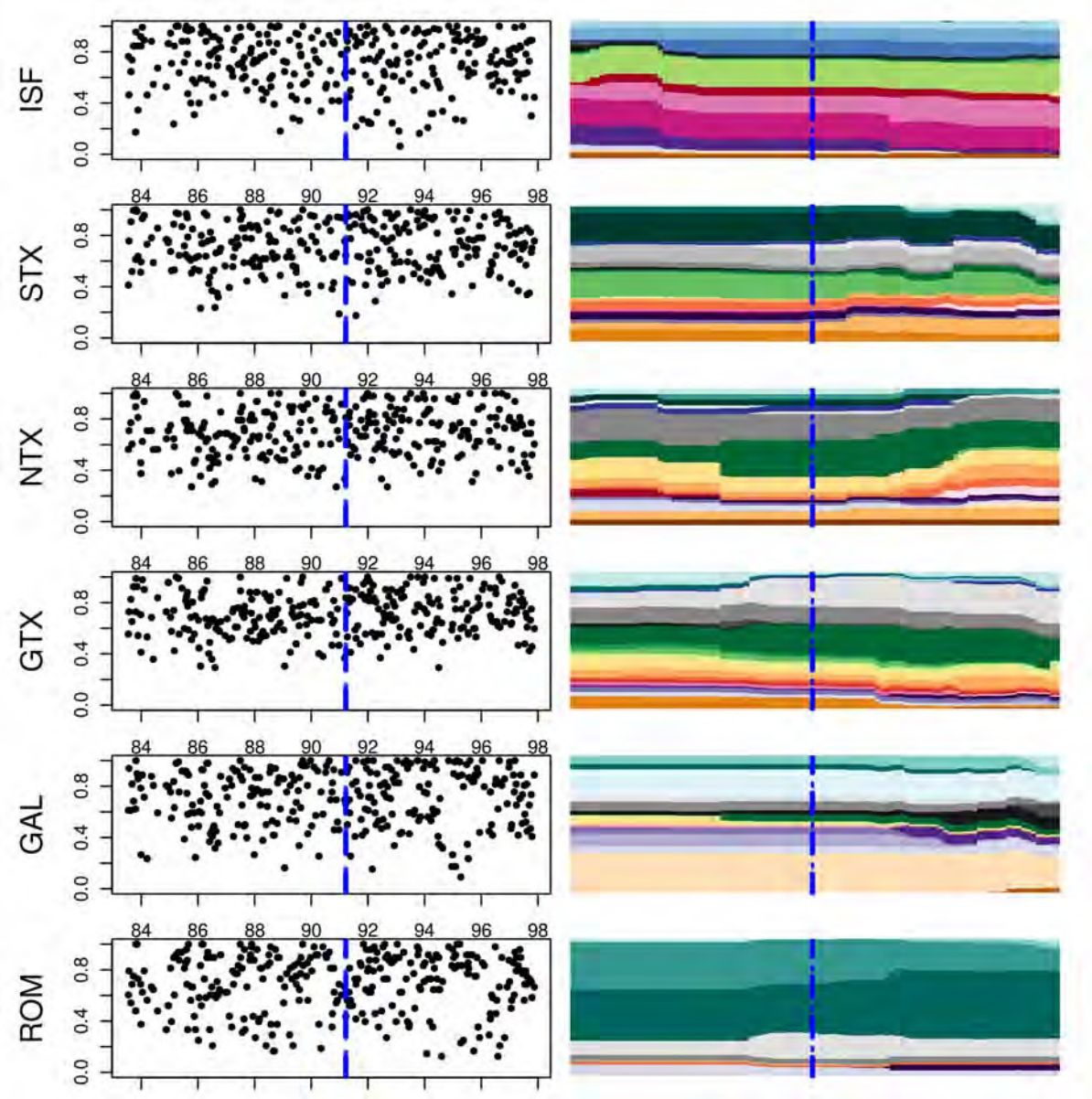


Figure S11. Chromosome 6 (Region 2) Allele (left) and haplotype cluster frequencies (right) in detected region 2 for each of the 6 Sheep populations used in the test. Blue bars indicate the limits of the detected region and the position of maximum of the test.

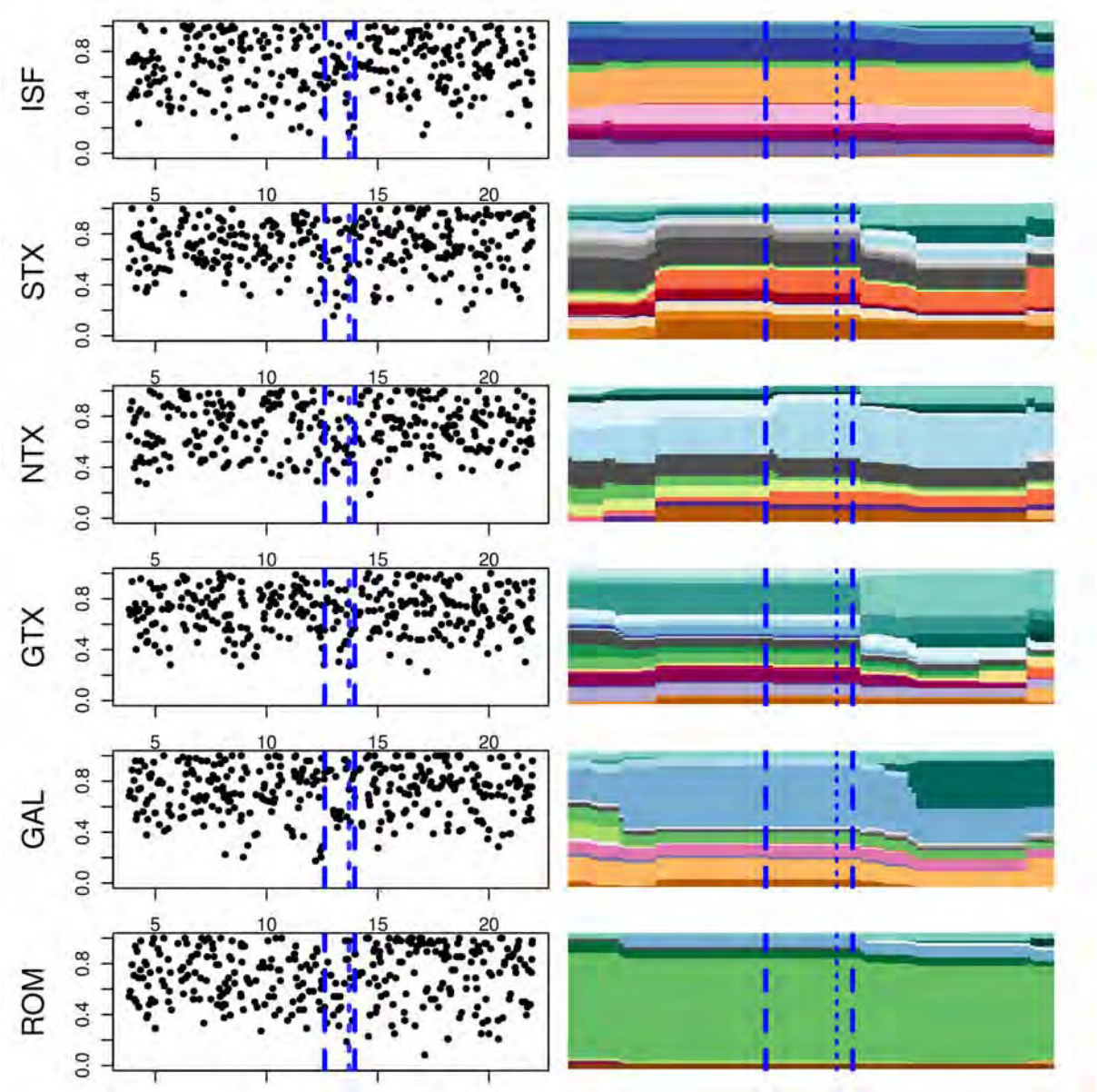


Figure S12. Chromosome 11 (Region 3) Allele (left) and haplotype cluster frequencies (right) in detected region 3 for each of the 6 Sheep populations used in the test. Blue bars indicate the limits of the detected region and the position of maximum of the test.

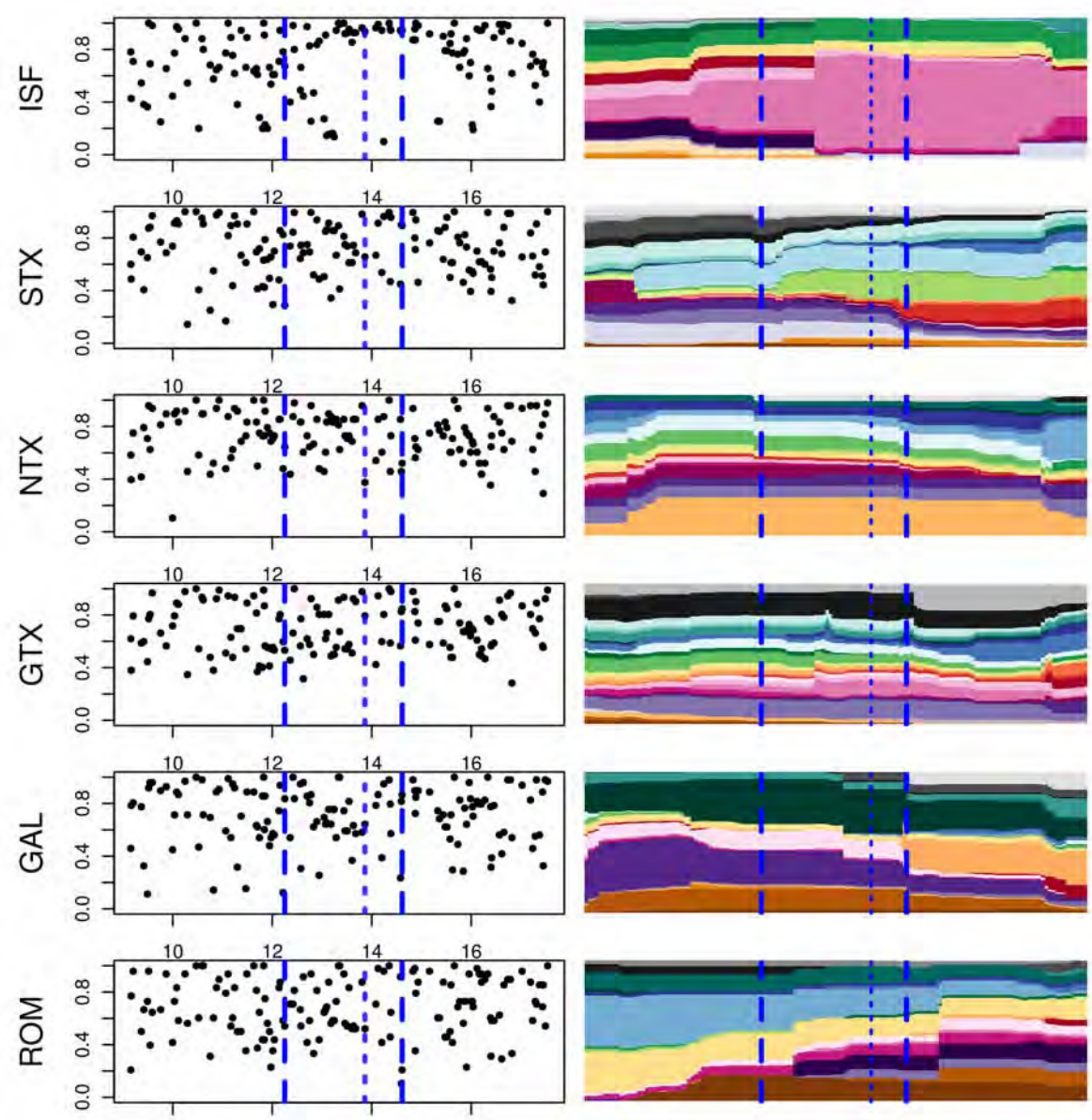


Figure S13. Chromosome 14 (Region 4) Allele (left) and haplotype cluster frequencies (right) in detected region 4 for each of the 6 Sheep populations used in the test. Blue bars indicate the limits of the detected region and the position of maximum of the test.

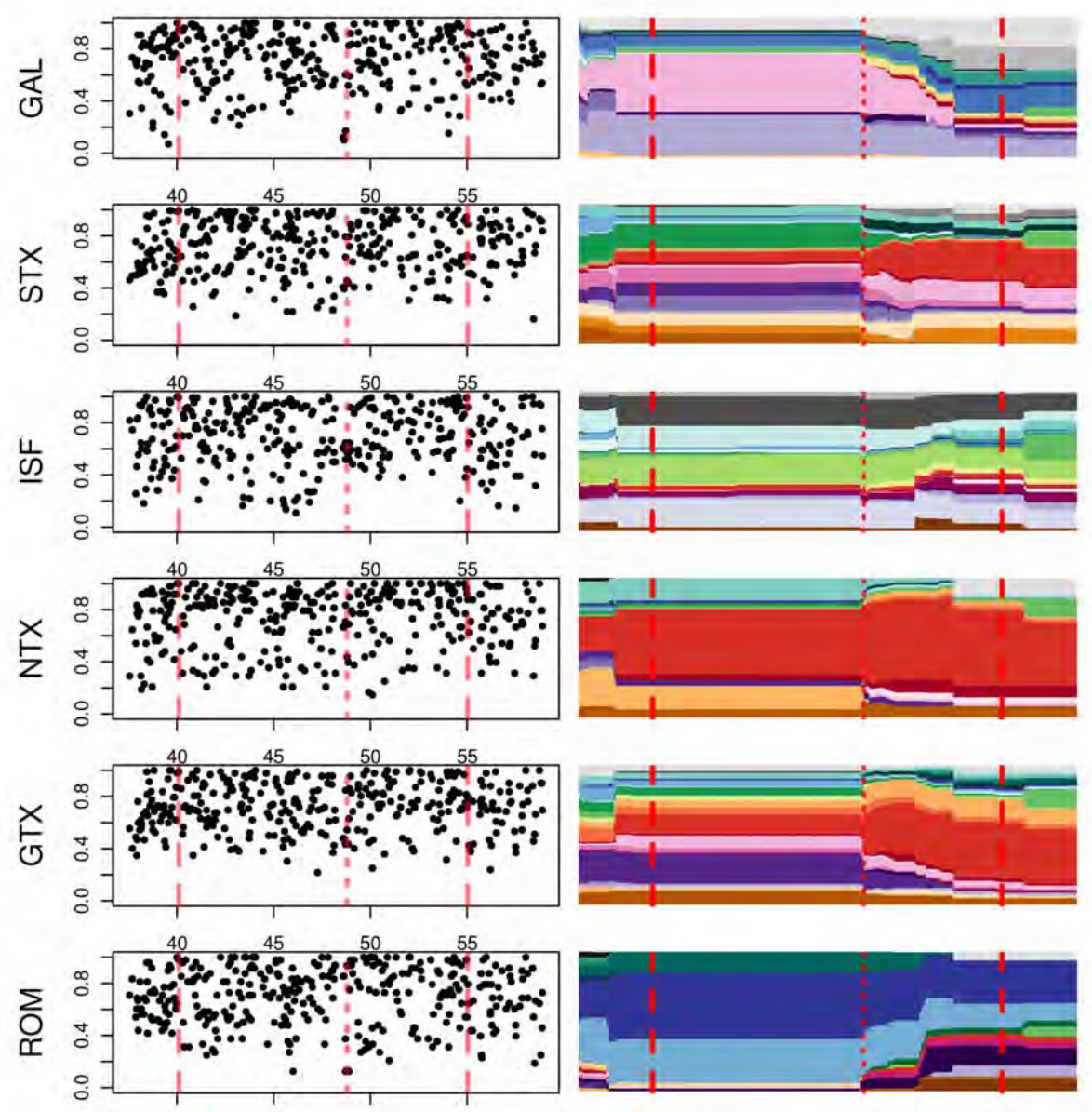


Figure S14. Chromosome 14 (Region 5) Allele (left) and haplotype cluster frequencies (right) in detected region 4 for each of the 6 Sheep populations used in the test. Blue bars indicate the limits of the detected region and the position of maximum of the test.

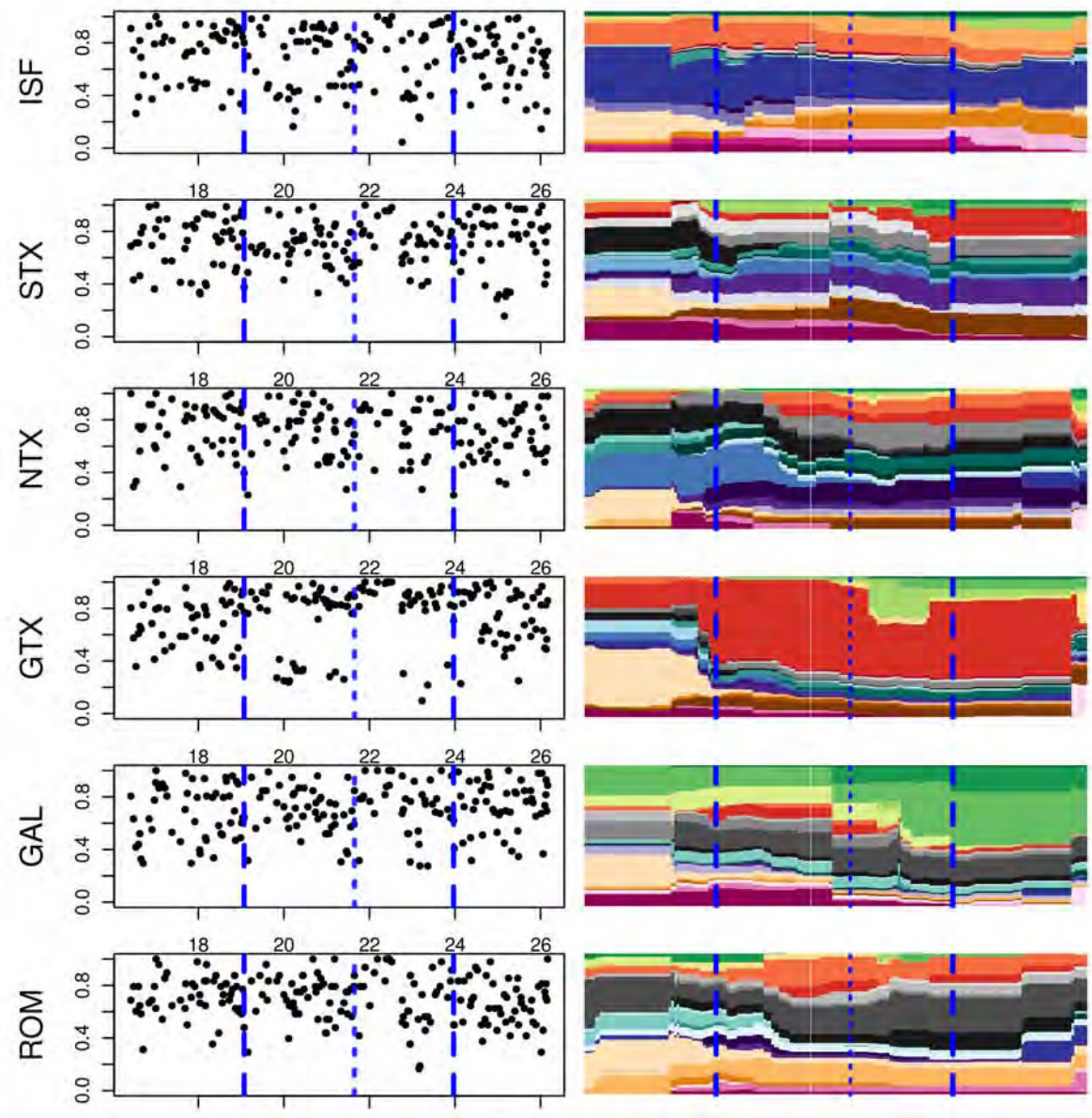


Figure S15. Chromosome 22 (Region 6) Allele (left) and haplotype cluster frequencies (right) in detected region 6 for each of the 6 Sheep populations used in the test. Blue bars indicate the limits of the detected region and the position of maximum of the test.

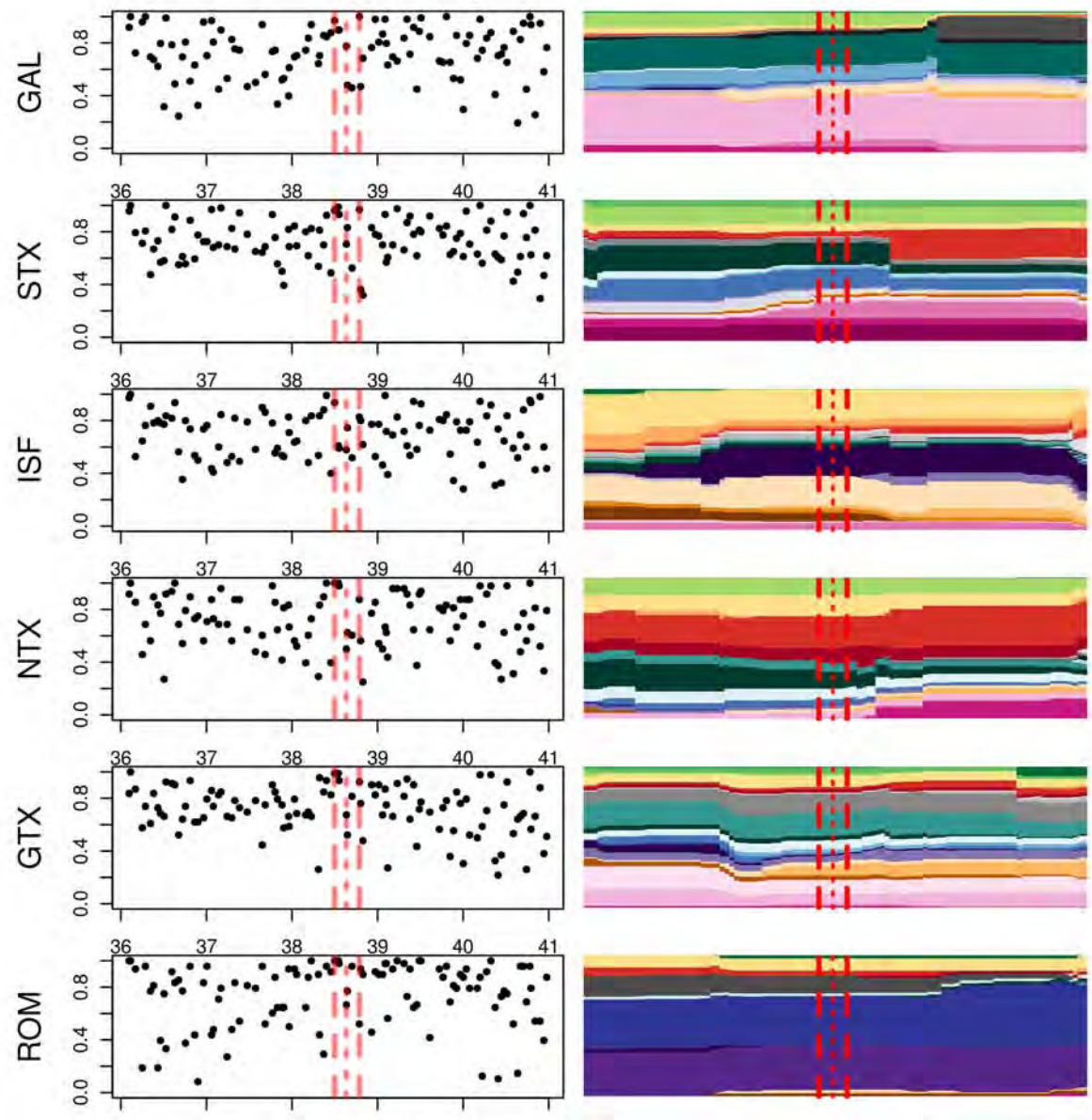


Figure S16. Chromosome 22 (Region 7) Allele (left) and haplotype cluster frequencies (right) in detected region 7 for each of the 6 Sheep populations used in the test. Red bars indicate the limits of the detected region and the position of maximum of the test.

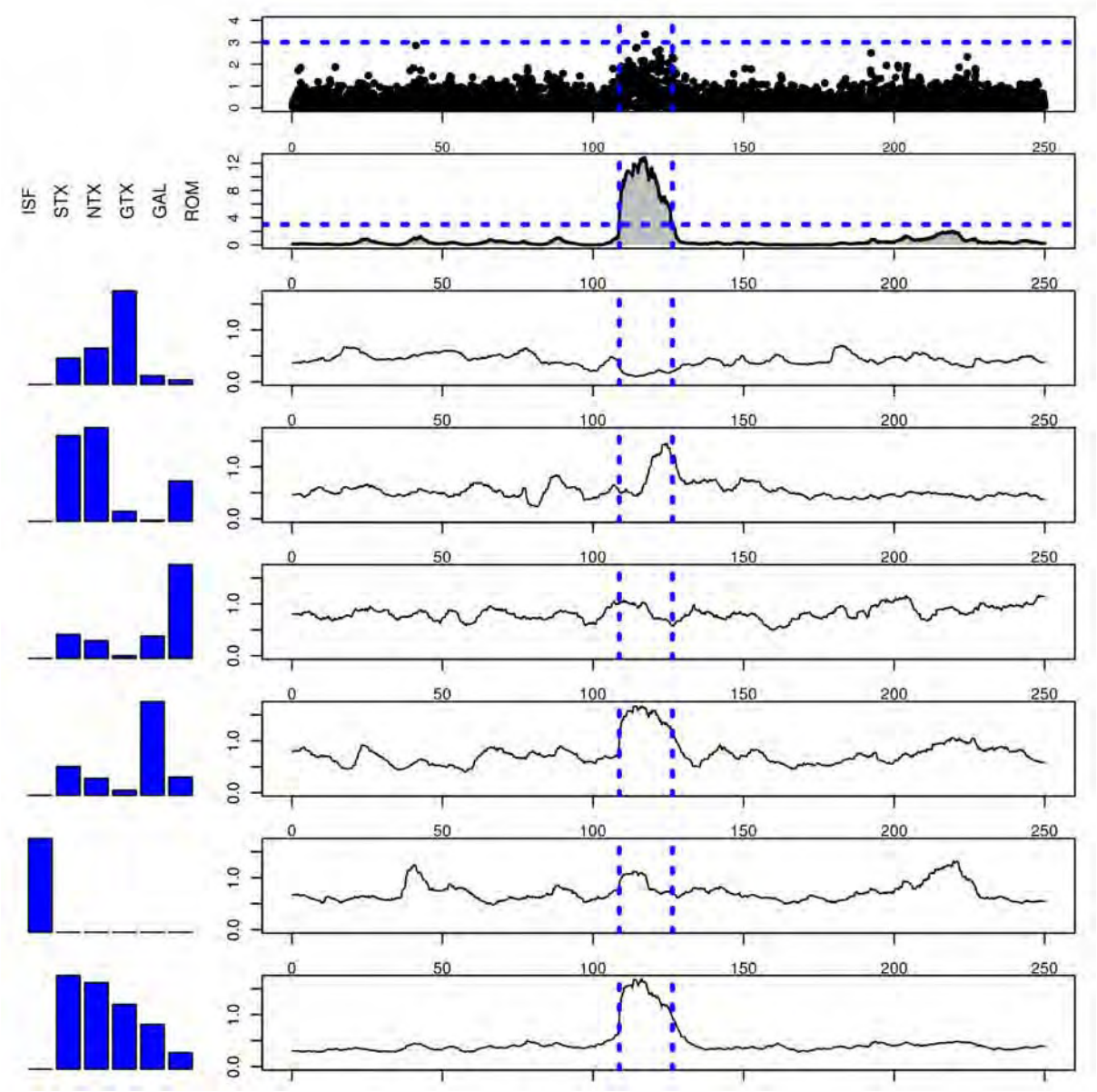


Figure S17. Chromosome 2 Spectral decomposition of *hapFLK* for detected region 1. First line: FLK p-values for each SNP in the chromosome 2. Second line: *hapFLK* p-values. Each of the following lines acorresponds to an orthogonal component of the population kinship matrix: For each component, population 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.

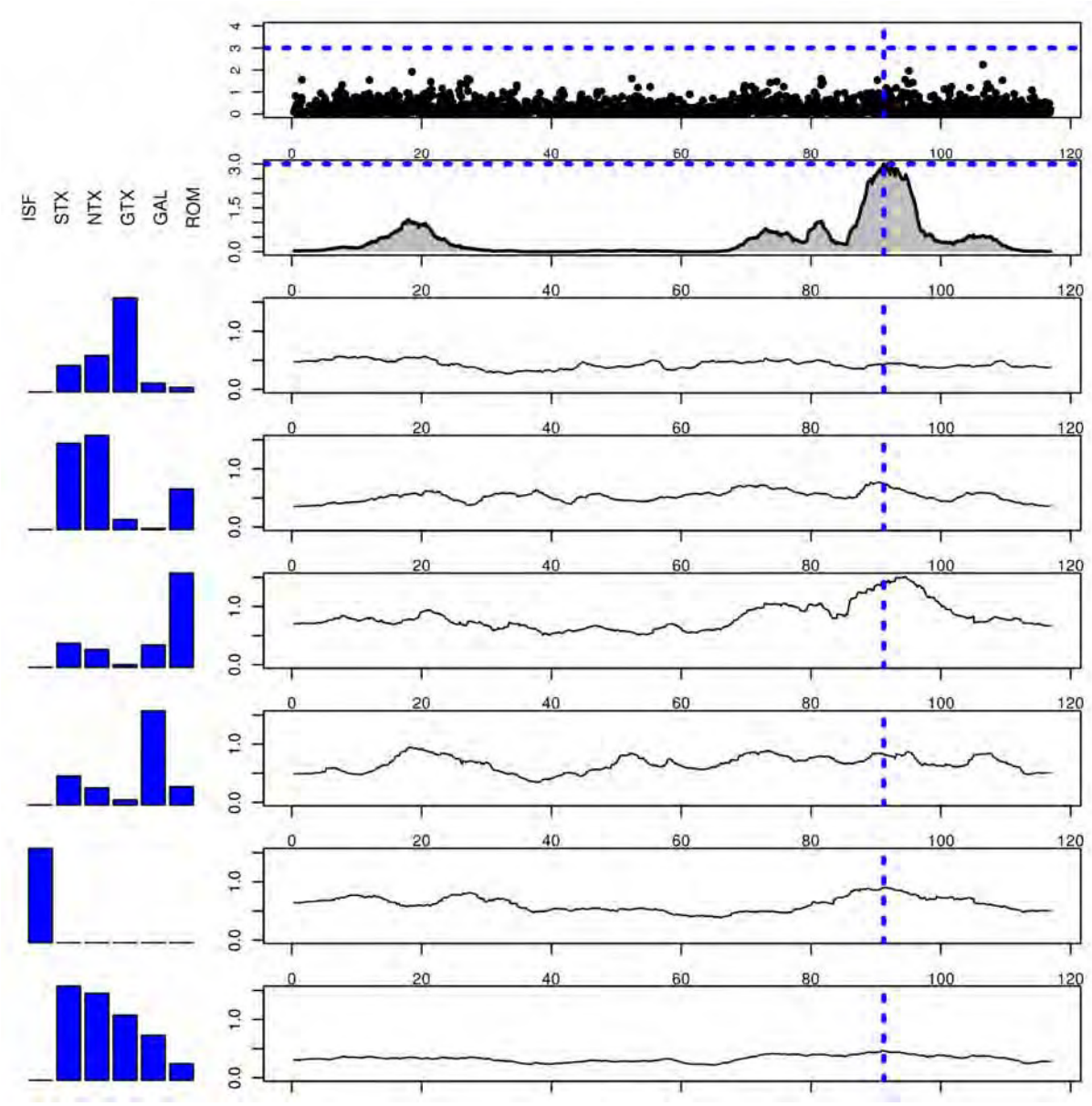


Figure S18. Chromosome 6 Spectral decomposition of *hapFLK* for detected region 2. See caption of Figure S17 for details.

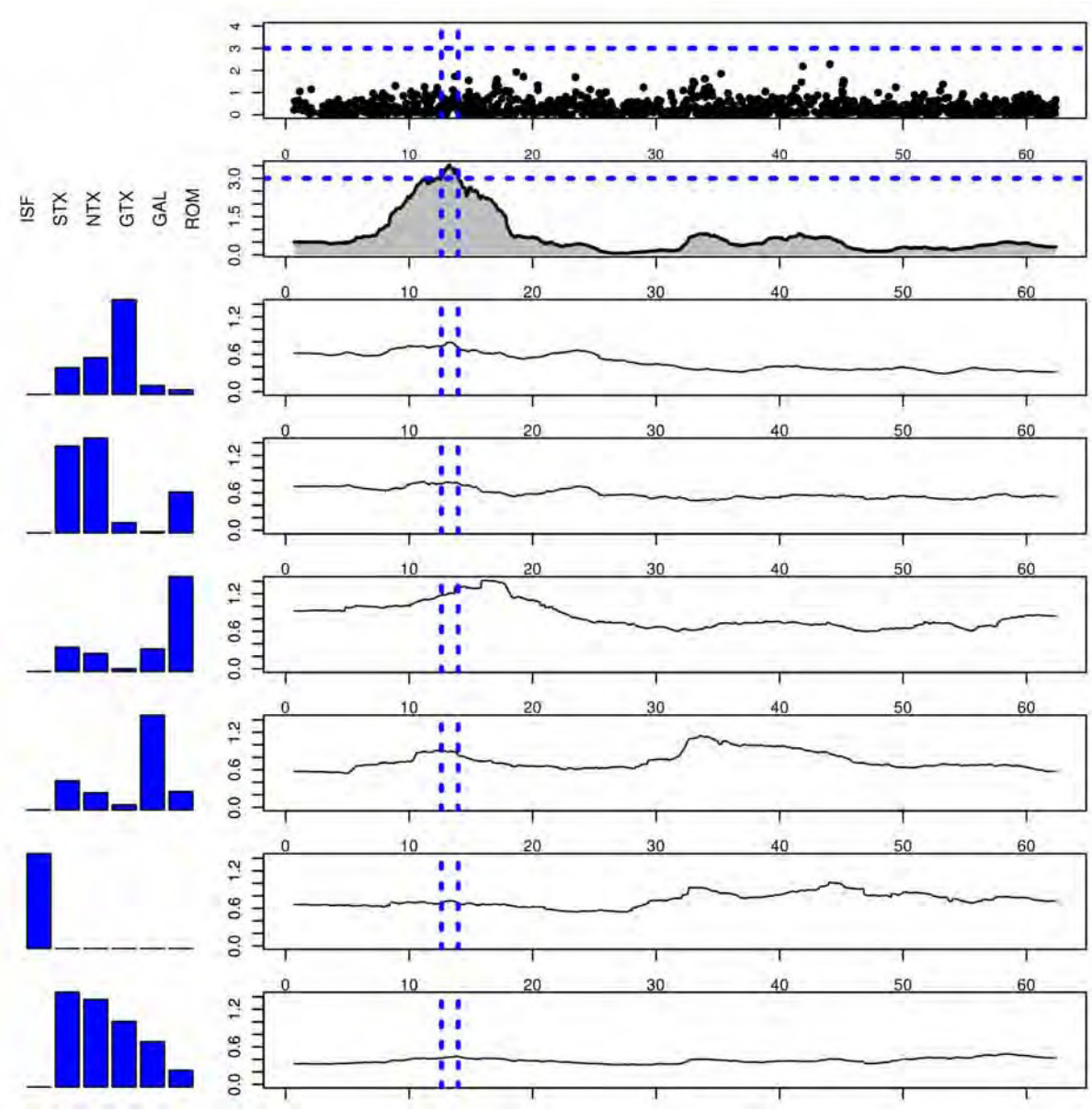


Figure S19. Chromosome 11 Spectral decomposition of *hapFLK* for detected region 3. See caption of Figure S17 for details.

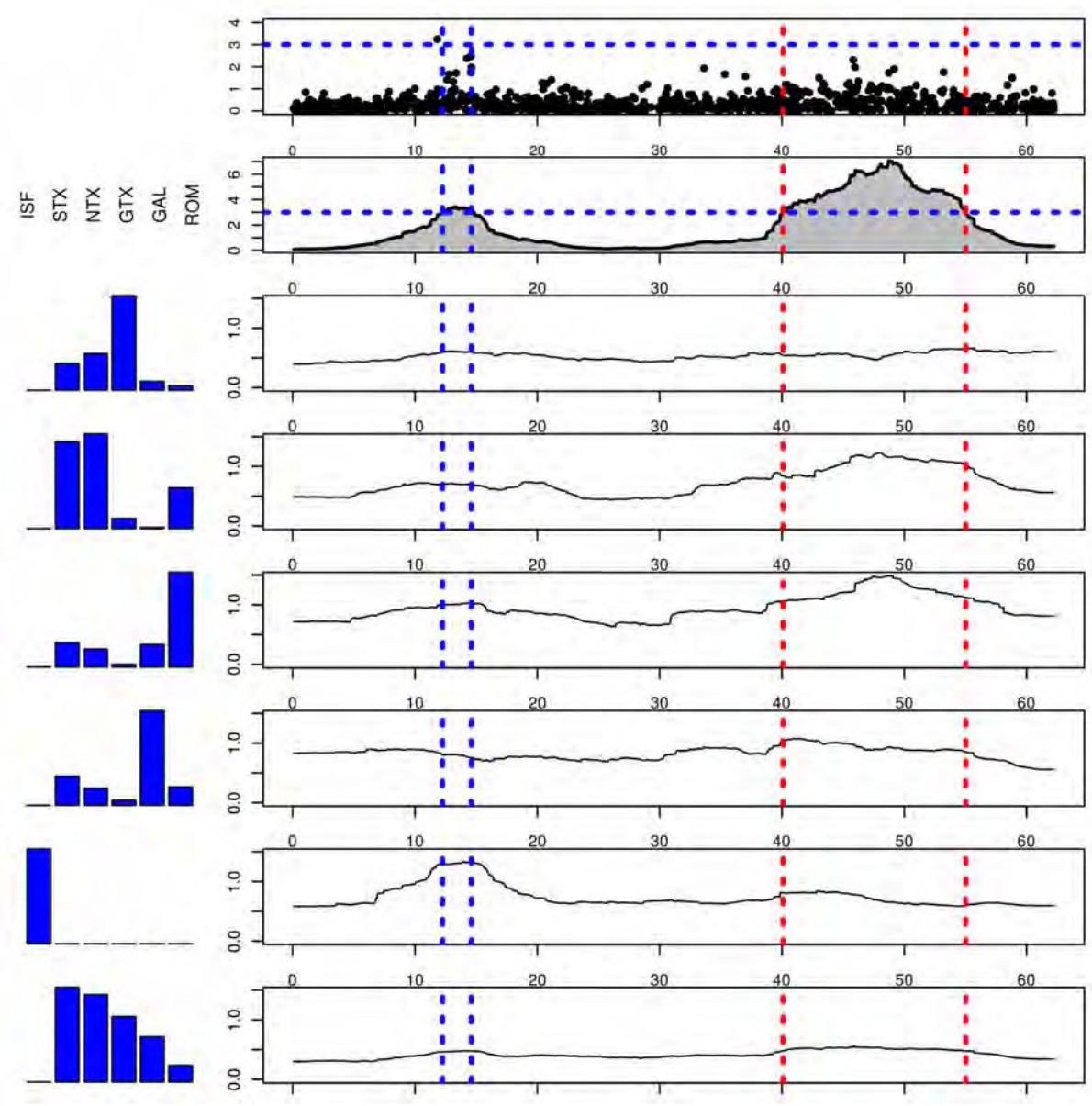


Figure S20. Chromosome 14. Spectral decomposition of *hapFLK* for detected regions 4 (blue) and 5 (red). See caption of Figure S17 for details.

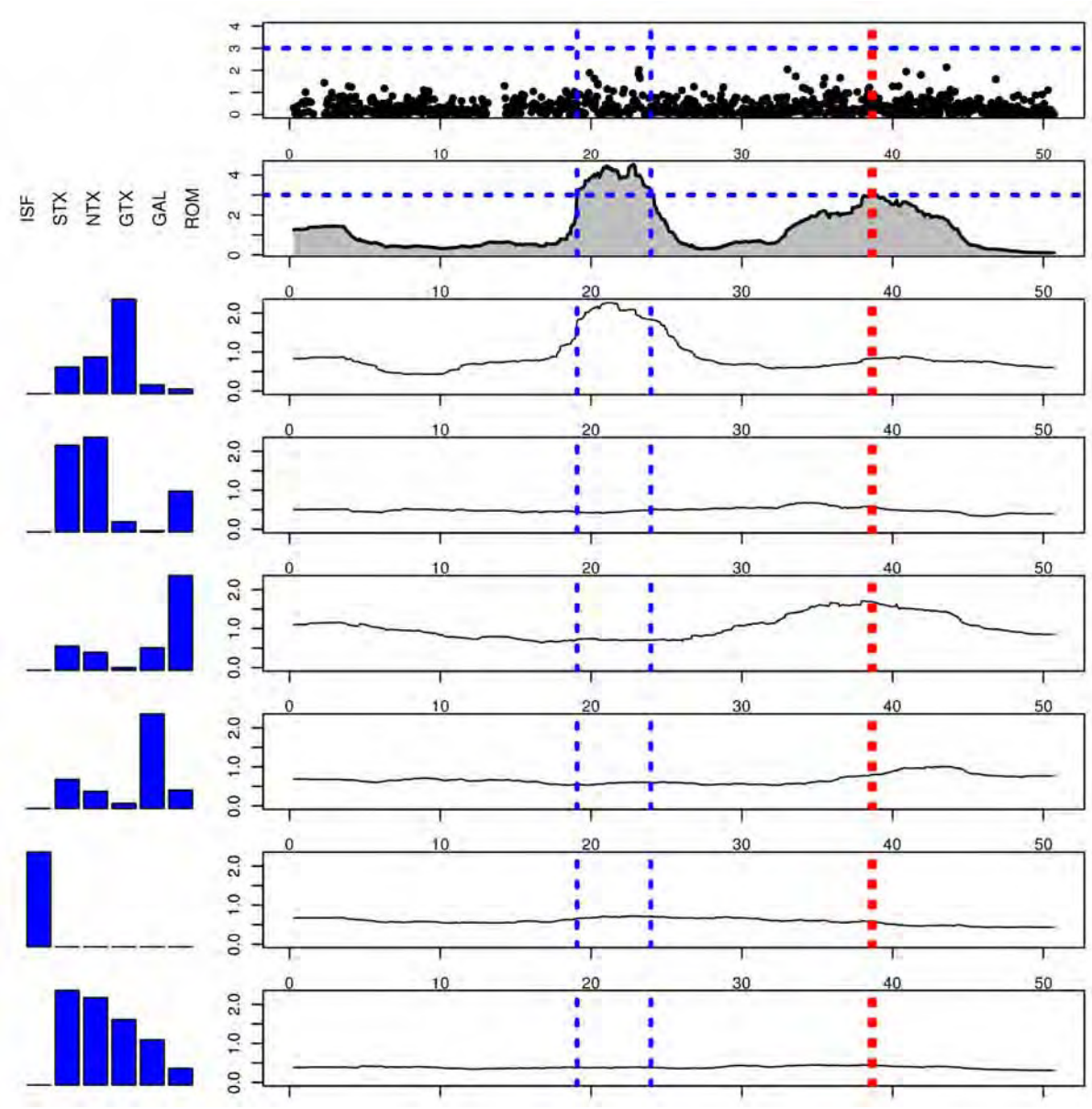


Figure S21. Chromosome 22. Spectral decomposition of *hapFLK* for detected regions 6 and 7. See caption of Figure S17 for details.

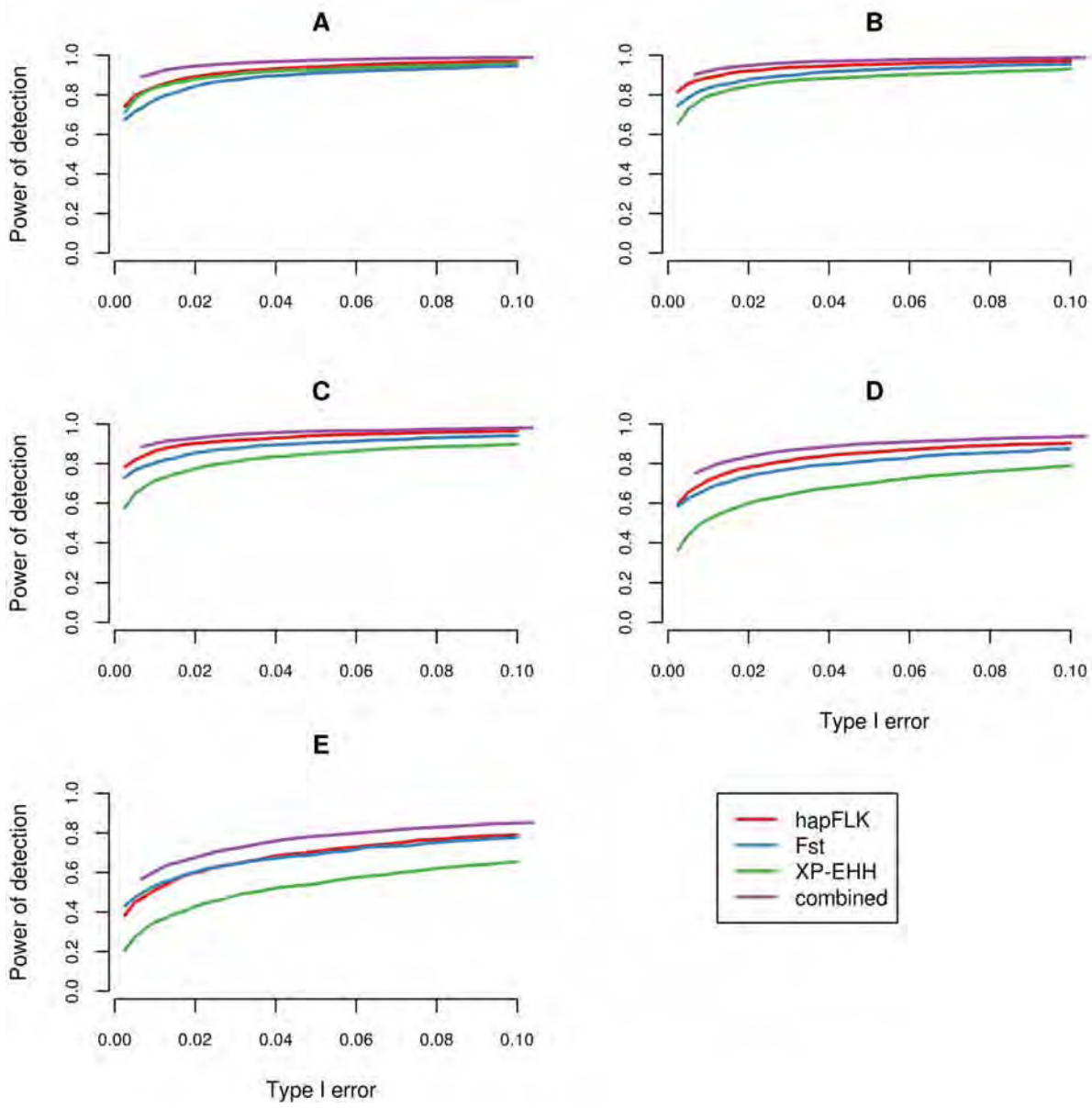


Figure S22. Power of $hapFLK, F_{ST}$ and $XP-EHH$ and the their combination as a function of type I error. Initial frequencies for the selected allele: A:1%, B:5%, C:10%, D:20%, E:30 %.

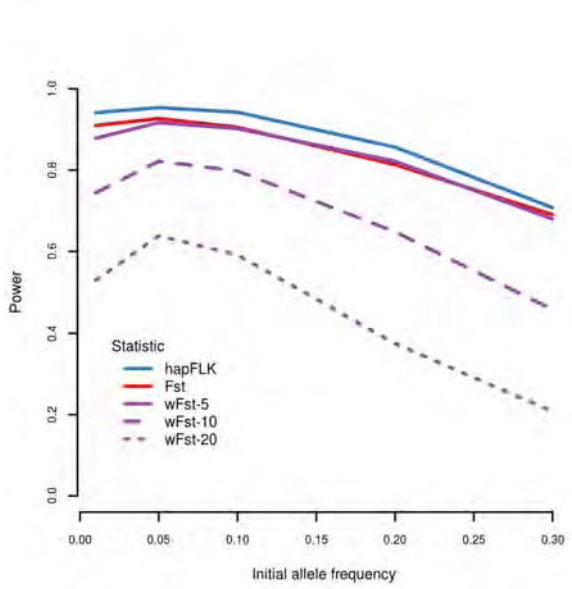


Figure S23. Detection Power of $hapFLK, F_{ST}$ and windowed F_{ST}. Windowed F_{ST} was computed as proposed by (Browning and Weir, 2010) for windows of 5, 10 and 20 SNPs. The power is evaluated at a type I error level of 5%.

Chapter 5

Analysis of the Sheep HapMap dataset

In the previous chapter the hapFLK test was applied to a group of breeds of North European origin from the Sheep HapMap dataset. The Sheep HapMap dataset is one of the largest genomic datasets of domestic animals, comprising 74 breeds from almost all over the world. In this chapter I present, under the form of a manuscript in preparation, a deeper analysis of the Sheep HapMap dataset: 7 geographic groups of breeds, and a group including the 7 putative ancestral populations of these groups, are studied. This dataset was already analyzed in Kijas et al. (2012) using a global F_{ST} approach. As already explained, this might not be the best strategy to test several populations simultaneously. Applying $\mathcal{F}$-LK and hapFLK to this dataset thus provides a great opportunity to evaluate the interest of these approches in a large scale real study, and to see to which extent their results overlap with those of a standard approach.

Studying this large real dataset also lead me to come up with a strategy for cleaning the data before performing the genome scan in itself. First, I removed populations with small sample size (< 20) or very long tree branches. Second, I explored the population structure within each geographic group using Principal Components Analysis and the software ADMIXTURE (Alexander et al., 2009), in order to detect and remove admixed individuals or pop-

ulations. The reason for this second step is that the model underlying $\mathcal{F}$-LK and hapFLK assumes no migration. Although hapFLK appeared to be robust to moderate levels of admixture, it should provide more reliable results if strongly admixed populations are cleared out from the data set. The detailed application of this strategy is provided for all geographic groups, and I believe these examples will be helpful for future studies based on $\mathcal{F}$-LK or hapFLK.

Results from this genome scan indicate that $\mathcal{F}$-LK and hapFLK provide very meaningfull regions, as many of them include candidate genes that are known to be under selection in sheep or in other farm animal species. Twenty out of the 31 regions detected by Kijas et al. (2012) are recovered, including those with the most obvious candidate genes. Many new interesting regions are also pointed out. Quite little overlap is found between the regions detected by $\mathcal{F}$-LK and hapFLK, confirming that these tests tend to detect different kinds of signals.

This study is also very informative from a biological point of view, as detected regions include many genes implied in coat pigmentation, morphology or meat, milk and wool production.

Selection signatures in worldwide Sheep populations

Maria-Ines Fariello[1], Bertrand Servin[1], Gwenola Tosser-Klopp[1], Rachel Rupp[2], Carole Moreno[2], Magali San Cristobal[1], Simon Boitard[3,4]

1 Laboratoire de Génétique Cellulaire, INRA, Castanet-Tolosan, France

2 Station d'Amélioration Génétique des Animaux, INRA, Castanet-Tolosan, France

3 Génétique Animale et Biologie Intégrative, INRA & AgroParisTech, Jouy-en-Josas, France

4 Origine, Structure et Evolution de la Biodiversité, Museum National d'Histoire Naturelle, Paris, France

Abstract

The diversity of populations in domestic species offer great opportunities to study genome response to selection. The recently published Sheep Hapmap dataset is a great example of characterization of the world wide genetic diversity in the Sheep. In this study, we re-analyzed the Sheep Hapmap dataset to identify selection signatures in worldwide Sheep populations. Compared to previous analyses, we make use of statistical methods that (i) take account of the hierarchical structure of sheep populations and (ii) make use of Linkage Disequilibrium information. We show that this allows to pinpoint several new selection signatures in the sheep genome. The newly identified regions, together with the one previously identified, reveal the extensive genome response to selection on morphology, color and adaptation to new environments.

Introduction

Domestication of animal and plants played a major role in human history. With the advance of high-throughput genotyping and sequencing technologies, the analysis of large datasets in domesticated species offer great opportunities to study genome evolution in response to phenotypic selection [1]. Sheep was the first grazing animal to be domesticated [2] in part due to its manageable size and an ability to adapt to different climates and poor nutrition diets. A large variety of breeds with distinct morphology, coat color or specialized production (meat, milk or wool) were subsequently shaped by artificial selection. Since the realease of the 50K SNP array [3], it is now possible to scan the genetic diversity in Sheep in order to detect loci that have been involved in these various adaptative selection events. The Sheep HapMap dataset, which includes 50K genotypes for 3000 animals from 74 breeds with diverse world-wide origins, provides a considerable ressource for deciphering the genetic bases of phenotype diversification in Sheep. In the first analysis of this data set, [4] looked for selection by computing a global F_{ST} among the 74 breeds at all SNPs in the genome. They identified 31 genomic regions with extreme differentiation between breeds, which included candidate genes related to coat pigmentation, skeletal morphology, body size, growth, and reproduction. Further studies took advantage of the Sheep HapMap ressource to detect genetic variants associated with pigmentation [5], fat deposition [2], or microphtalmia disease [6]. [7] performed a genome scan for selection focused on American synthetic breeds, using an F_{ST} approach similar to that in [4].

The 74 breeds of the Sheep HapMap dataset have a strong hierarchical structure, with at least 3 distinct differentiation levels: an inter-continental level (e.g. European breeds vs Asian breeds), an intra-continental level (e.g. Texel vs Suffolk European breeds), and an intra-breed level (e.g. German Texel vs Scottish Texel flocks). Recent studies [8, 9] showed that, when applied to hierarchically structured data sets, F_{ST} based genome scans for selection may lead to a large proportion of false positives (neutral loci wrongly detected as under selection) and false negatives (undetected loci under selection). This statistical issue is also compounded by the heterogeneity of effective population size among breeds, implying that some breeds are more prone to contribute large locus-specific F_{ST} values than others [9]. Apart from these statistical considerations, merging populations with various degrees of shared ancestry can limit our understanding of the selective process at detected loci. Indeed, the regions pointed out by [4] can be

related to either ancient selection, as the poll locus which has likely been selected for thousands of years, or fairly recent selection, as the myostatin locus which has been specifically selected in the Texel breed. But in most situations the time scale of adaptation can not be easily determined.

Another limit of genome scans for selection based on single SNP F_{ST} computations is that they do not sufficiently account for the very rich linkage disequilibrium information, even when the single SNP statistics are combined into windowed statistics. Recently, we proposed a new strategy to evaluate the haplotypic differentiation between populations [10]. We showed that using this approach greatly increases the detection power of selective sweeps from SNP chip data, and enables to detect also soft or incomplete sweeps. These latter selection scenarios are particularly relevant in the case of breeding populations, where selection objectives have likely varied along time and where the traits under selection are often polygenic.

In this study we provide a new genome scan for selection based on the Sheep HapMap data set, where we distinguish selective sweeps between and within 7 broad geographical groups. The within group analysis aims at detecting recent selection events related to the diversification of modern breeds. It is based on the single marker FLK test [9] and on its haplotypic extension [10], that both account for population size heterogeneity and for the hierarchical structure between populations. The between group analysis focuses on older selection events and is only based on FLK. Overall, we confirm 19 of the 31 sweeps discovered by [4], while providing more details about the past selection process at these locus. We also identify 68 new regions under selection, with candidate genes related to coloration, morphology or production traits.

Results and discussion

We detected selection signatures using methods that aim at identifying regions of outstanding genetic differentiation between populations, based either on single SNP, FLK [9], or haplotype, hapFLK [10], information. These methods have optimal power when working on closely related populations so we analyzed separately seven groups of breeds, previously identified as sharing recent common ancestry [4] and corresponding to geographical origins of breeds. Before performing genome scans for selection signatures, we studied the population structure of each group to identify outlier animals as well as admixed and strongly bottelnecked populations, using both PCA and model-based approaches [11, 12]. hapFLK was found robust to bottlenecks or moderate levels of admixture, but these phenomena may affect the detection power so we preferred to minimize their influence by removing suspect animals or populations. Details of these corrections are provided in the methods section. The final composition of populations groups are given in table S1.

Overview of selected regions

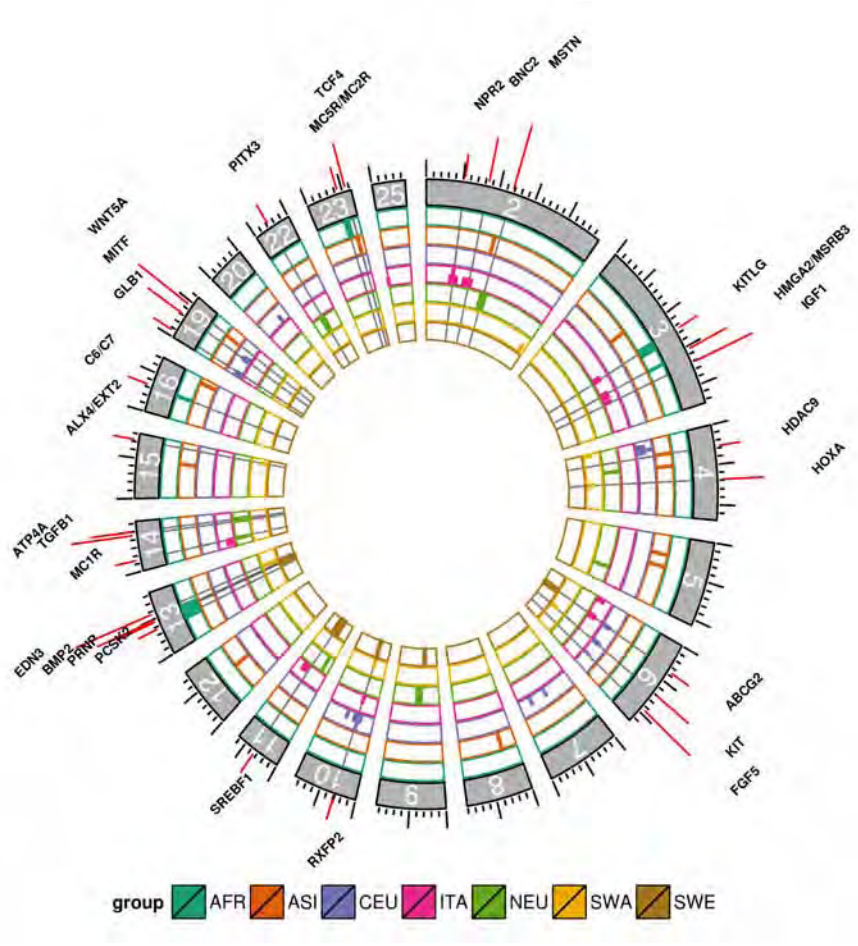


Figure 1. Localisation of selection signatures identified in 7 groups of populations. Candidate genes are indicated above their genomic localisation. Only chromosomes harboring selection signatures are plotted.

An overview of selection signatures on the genome across the different groups is plotted on Figure 1 and Table 1 provides their detailed description. We found 40 selection signatures with hapFLK and 24 with FLK, although we allowed a slightly higher false discovery rate for FLK than hapFLK (10% vs 5%). This result is consistent with a higher power for hapFLK than FLK, as was shown before [10]. Four regions are found with both the single SNP and the haplotype test and harbor strong functional candidate genes: NPR2, KIT, RXFP2 and EDN3 (see below). The overlap is thus small, illustrating that the two tests tend to capture different signals. In particular, hapFLK will fail to detect ancient selective sweeps where the mutation-carrying haplotype is small, and not associated with many SNPs on the chip. On the other hand, single SNP tests will fail to capture selective sweeps when a single SNP is not in high LD with the causal mutation. Six regions were detected in more than one group of breeds. They all contain strong candidate genes. Three of these genes are related to coat color (KIT, KITLG and MC1R), and could correspond to independent selection events (see discussion below). One region

harbors a gene (RXFP2) for which polymorphisms have been shown to affect horn size and polledness in the Soay [13] and Autralian Merino [14]. The signatures of selection in this region exhibit different patterns among groups. The signal is very narrow in the SWE and SWA groups, and is in fact not detected by the hapFLK test, whereas it affects a large genomic region in the CEU group where it is detected by hapFLK. In the ITA group, the FLK statistics do not reach significance, and the hapFLK signal is not high (minimum qvalue of 0.04). Together, the selection signatures suggest that selection on RXFP2, most likely due to selection on horn phenotypes, was carried out worldwide at different times and intensities. The last two regions harbor the HMGA2 gene, involved in selection for stature in dogs [15], and ABCG2, a strong QTL for milk production in cattle [16]. Populations selected for ABCG2 variants belong to different European regions (SWE, ITA and CEU).

Table 1. List of genome regions corresponding to selection signatures. Regions identified with the hapFLK and FLK test, with the corresponding population group and most differentiated populations (except for the AFR group). Overlapping regions in different groups or with different tests are grouped by background color. †: signatures of selection previously identified [4]. ‡: this outlying region is not due to evolutionary processes (see details in the main text). Full names of groups and populations are given in Table S1.

OAR	Begin (Mbp)	End (Mbp)	P-value	Q-value	Group	Test	Cand. gene	Diff. pop.
2	46.65	57.99	6.3e-10	7.1e-07	ITA	hapFLK	NPR2†	COM
2	51.41	53.44	4.1e-09	1.6e-04	ITA	FLK		COM
2	74	74.86	7.4e-04	3.7e-02	ITA	hapFLK		COM
2	81.27	87.32	4.1e-09	2.3e-06	ITA	hapFLK	BNC2	COM
2	110.08	112.08	1.5e-05	6.7e-02	ASI	FLK		SUM TIB GUR
2	113.36	122.24	7.0e-06	3.3e-03	NEU	hapFLK	MSTN†	GTX NTX STX
2	239.76	241.76	2.9e-05	9.3e-02	SWA	FLK	RH locus	AFS
3	84.4	86.4	2.5e-05	9.1e-02	ASI	FLK		–
3	120.91	125.49	5.3e-04	3.0e-02	ITA	hapFLK	KITLG	COM
3	122.07	130.85	6.8e-08	4.2e-04	AFR	hapFLK		
3	151.42	156.93	3.3e-16	3.1e-12	ITA	hapFLK	HMGA2†	COM SAB
3	154.79	154.93	5.9e-04	4.3e-02	AFR	hapFLK		
3	159.64	161.6	6.1e-04	3.3e-02	ITA	hapFLK		COM
3	167.85	171.67	1.5e-04	1.3e-02	ITA	hapFLK	IGF1	COM ALT SAB
4	4.61	6.61	5.3e-06	2.1e-02	SWA	FLK		MOG
4	8.5	19.66	4.2e-06	1.1e-03	CEU	hapFLK		VBS VRS
4	15.11	17.11	8.4e-07	1.5e-02	CEU	FLK		VBS
4	26.46	28.46	2.4e-05	9.1e-02	ASI	FLK	HDAC9	GUR IDC SUM
4	44.49	45.76	2.7e-04	3.4e-02	NEU	hapFLK		NZR
4	45.57	47.57	1.8e-06	2.4e-02	ASI	FLK		SUM
4	67.75	69.8	3.5e-07	2.3e-03	SWA	FLK	HOXA	MOG

Table 1 – continued from previous page

5	29.4	31.4	1.1e-05	6.7e-02	ASI	FLK	BMPR1B	GAR
5	47.35	49.35	1.4e-05	6.7e-02	ASI	FLK		BGA
5	78.16	78.76	4.2e-04	4.2e-02	NEU	hapFLK		NZT
6	5.62	7.62	3.1e-06	6.0e-02	ITA	FLK		SAB
6	33.22	41.02	3.4e-08	8.0e-05	SWE	hapFLK	ABCG2†	LAC LAM
6	34.71	39.12	1.6e-07	4.1e-05	ITA	hapFLK		COM
6	35.94	38.31	2.1e-04	1.9e-02	CEU	hapFLK		VRS VBS
6	67.98	70.36	4.3e-06	1.1e-03	CEU	hapFLK	KIT†	VBS
6	68.9	70.95	9.6e-07	5.3e-03	SWA	FLK		
6	93.3	94.39	3.8e-04	2.7e-02	CEU	hapFLK	FGF5†	(VRS&VBS) or (ERS&BOS)
7	49.15	51.15	1.1e-05	9.7e-02	CEU	FLK		VRS
7	78.31	80.31	8.1e-07	1.5e-02	CEU	FLK		VRS ERS
8	23.97	25.97	2.9e-05	9.6e-02	ASI	FLK		TIB
9	29.46	31.55	3.7e-04	3.4e-02	SWE	hapFLK		CHU MER
9	37.79	46.03	1.9e-05	6.2e-03	NEU	hapFLK		NZT ISF
10	24.02	34.91	1.4e-14	1.1e-10	CEU	hapFLK	RXFP2†	BOS ERS VRS
10	29.42	29.71	9.6e-04	4.4e-02	ITA	hapFLK		COM ALT
10	28.5	30.5	6.3e-06	7.5e-02	CEU	FLK		BOS ERS
10	28.5	30.5	3.2e-05	9.7e-02	SWA	FLK		NDZ
10	28.5	30.5	1.3e-06	5.4e-02	SWE	FLK		MER
10	48.9	49.59	5.2e-04	3.1e-02	CEU	hapFLK		–
11	12.55	14.12	1.4e-04	2.2e-02	NEU	hapFLK		
11	24.18	38.74	9.8e-09	8.0e-05	SWE	hapFLK	SREBP1	LAC MER
11	40.31	46.7	3.3e-06	5.5e-04	ITA	hapFLK		SAB
12	42.66	44.66	3.4e-07	7.6e-03	ASI	FLK		SUM
13	33.1	40.02	5.7e-06	1.8e-03	AFR	hapFLK	PCSK2	
13	40.6	50.3	4.9e-07	4.9e-04	AFR	hapFLK	BMP2†	
13	43.34	51.28	2.7e-07	1.7e-04	SWE	hapFLK	PRNP	LAC LAM
13	56.11	57.17	2.5e-08	4.8e-04	SWA	hapFLK	EDN3	MOG
13	55.33	57.43	8.4e-11	1.1e-06	SWA	FLK		MOG
14	6.37	13.6	1.6e-04	1.4e-02	ITA	hapFLK		SAB
14	13.64	13.7	5.3e-04	4.9e-02	NEU	hapFLK	MC1R	ISF
14	13.7	16.46	1.2e-04	1.1e-02	ITA	hapFLK		SAB
14	45.49	50.09	1.6e-04	2.5e-02	NEU	hapFLK	TGFB1	NTX NZR
15	48.87	50.87	1.5e-05	6.7e-02	ASI	FLK		GAR IDC
15	71.71	73.71	3.8e-06	1.6e-02	SWA	FLK	ALX4 EXT2	MOG

Table 1 – continued from previous page

16	33.2	35.1	1.8e-04	1.8e-02	AFR	hapFLK	C6/C7	
16	63.97	65.97	1.1e-05	6.7e-02	ASI	FLK		GAR IDC
19	4.42	7.43	2.2e-04	1.9e-02	CEU	hapFLK	GLB1†	VRS BOS
19	30.42	35.09	3.2e-05	4.2e-03	CEU	hapFLK	MITF†	VBS BOS ERS
19	44.6	46.6	3.9e-06	3.9e-02	ASI	FLK	WNT5A	GAR BGA
20	36.74	38.52	2.8e-04	2.3e-02	CEU	hapFLK		VRS
22	18.9	24.36	1.5e-11	7.4e-08	NEU	hapFLK	PITX3‡	GTX
23	42.5	46.96	2.2e-05	5.4e-03	AFR	hapFLK	MC5R MC2R	
23	54.14	56.14	3.8e-07	7.6e-03	ASI	FLK		GAR
25	0.08	3.08	3.7e-04	2.4e-02	ITA	hapFLK		SAB

In the paper presenting the sheephapmap dataset [4], 31 selection signatures where found, corresponding to the 0.1% highest single SNP F_{ST}. Using FLK and hapFLK, we confirm signatures of selection for 11 of these regions. Considering the two analyses were performed on the same dataset, this overlap can be considered as rather small. Tow reasons can explain it.

First, the previous analysis was based on the F_{ST} statistic. Although this statistic is commonly used for selection scans, it is prone to produce false positives when the history of populations is characterized by population trees with unequal branch lengths (*i.e.* variation in the amount of drift experienced by different populations) [9]. In particular, strongly bottlenecked breeds will contribute high F_{ST} values preferentially, even under neutral evolution. With FLK and $hapFLK$, this varying amount of drift is accounted for, and populations with long branch lengths will not contribute to the signal more than others [10]. In fact they will tend to contribute less as it is harder to rule out the effect of drift alone in such populations.

Second, the previous analysis was performed using all breeds at the same time. It is therefore possible that some of these regions correspond to differentiation between groups of breeds rather than within groups. To investigate this question, we performed a genome scan for selection between the ancestors of the seven population groups using the FLK statistic computed on their estimated allele frequencies [9]. We did not include SNPs lying in regions detected within groups as selection biases their estimated ancestral allele frequencies. The population tree was reconstructed using SNPs for which we have unambiguous ancestral allele information (Figure 2). The tree is decomposed into two main lineages, one for European breeds and one for Asian and African breeds. The African group exhibits a slightly higher branch length. We note however that this could be due to ascertainment bias of SNPs on the SNP array.

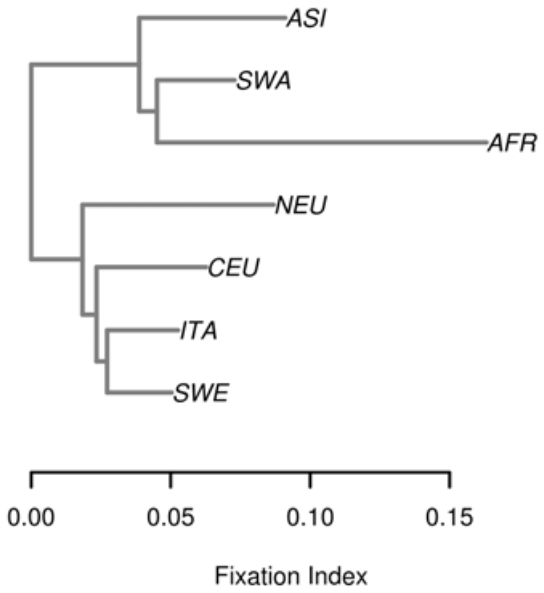


Figure 2. Phylogenetic tree of the ancestral populations of geographical groups.

This led to the identification of 23 new selection signatures (figure 3 and table 2), 9 of them being common to the previous analysis. Overall, we fail to replicate with this analysis 12 of the regions in [4].

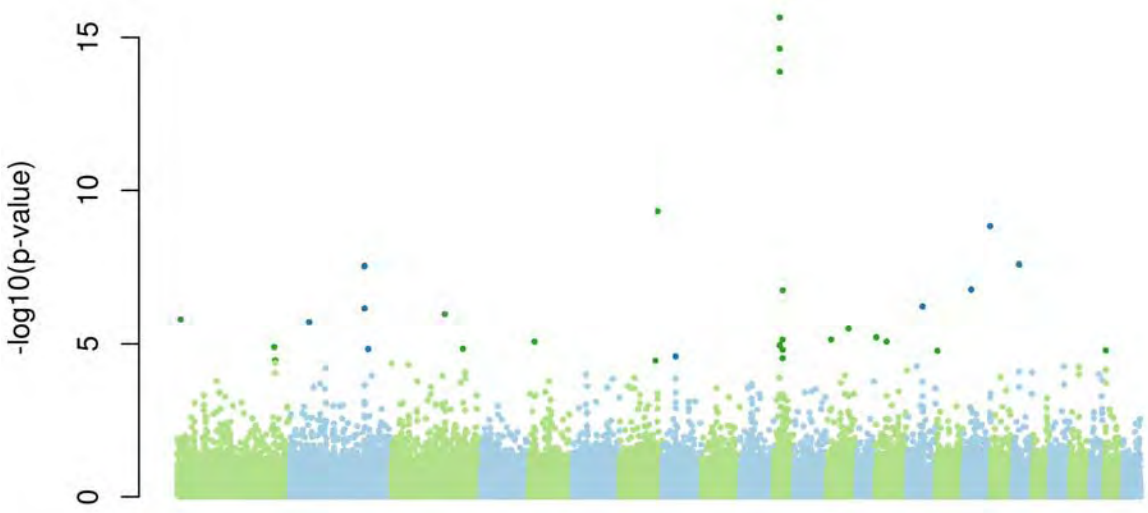


Figure 3. Genome scan for selection signature in ancestral populations of the geographical groups. Significant SNPs at the 5% FDR level are plotted in darker color.

Selection Signatures within population groups

Coloration Many selection signatures are located around genes that have been shown to be involved in hair, eye or skin color. In particular many genes underlying selection signatures are involved in the

development and migration of melanocyte and in pigmentation: EDN3, KIT, KITLG, MC1R and MITF. We can add to this list SOX10 and ASIP that show some evidence of selection: in the ITA group, the q-value of hapFLK near SOX10 is 6.2%, while the closest SNP to ASIP (s66432 and s12884) present suggestive FLK p-values of respectively $7.5\,10^{-4}$ and $6.8\,10^{-5}$ in the ASI group, and is significantly differentiated between the ancestral groups. All these genes have previously been reported as being likely selection targets and/or associated to color patterns in different mammalian species. Finally we found a signal for selection around the BNC2 gene, that has recently been associated with skin pigmentation in Humans [17]. All population groups present at least a selection signature on one of these genes, reflecting the widespread importance of color patterns to define sheep breeds. Inferring a precise history of underlying causal mutations for color patterns in this dataset is hard for several reasons: the precise phenotypic characterizations of coat color patterns in the SheepHapMap breeds are not available; the 50K SNP array used does not offer sufficient density to associate a given selection signature to a specific set of polymorphisms; finally, from the litterature, it appears that coat color is a complex trait, with high genetic heterogeneity. In particular, mutations in different genes can give rise to the same phenotype (*e.g.* in Horse [18]). Also, within a gene different mutations can give rise to different phenotypes, *e.g* mutations in the MC1R gene (also named the extensions locus) have been associated to a large panel of skin or coat colors [19–21]. Studying more precisely selection signatures related to coat color and the underlying selected mutations will likely require further sequencing experiments targeted at these genes. This in turn will help understand the evolutionary history of the breeds and the effect of selection [22]. To potentially help in this task, in table S2 we list, for each "color gene", the populations that have likely been selected for.

Morphology Another group of genes that are found in selection signatures have known effects on body morphology and development. NPR2, HMGA2 and BMP2 were identified previously [4], but we also found selection signatures around IGF1, ALX4 or EXT2, WNT5A and two Hox gene clusters (HOXA and HOXC). IGF1 has been shown to be a major determinant of small body size in dogs [23]. WNT5A and ALX4 are two genes involved in the development of the limbs and skeleton. ALX4 loss of function mutations cause polydactily in the mouse, through disregulation of the sonic hedgehog (SHH) signaling factor [24, 25]. EXT2 is responsible for the development of exostose in the mouse [26], it is located just besides ALX4 and corresponds to the same selection signature. Mutations in WNT5A are causing the dominant Human Robinow syndrome, *characterized by short stature, limb shortening, genital hypoplasia and craniofacial abnormalities* [27]. An ancestral selection signature is found near the ACAN gene. Mutations in the ACAN gene have been shown to induce osteochondrosis [28] and skeletal dysplasia [29]. The ACAN region has also been shown to be associated with Human adult height [30]. Two selection signatures are localized close to *Hox* genes clusters. *Hox* genes are responsible for antero-posterior development and skeletal morphology along the anterior-posterior axis in vertebrates. One is a recent selection signature in the SWA group near the HOXA gene cluster and the other is an ancestral signature near the HOXC gene cluster, with a high differentiation of the ASI ancestor compared to AFR and SWA at the most significant SNP (OAR3_141586525).

Traits of agronomical importance Sheeps have been raised for meat, milk and wool production. Under selection signatures, we found several genes associated with these production traits. Apart from the selection signature in Texels on the MSTN gene for increased muscularity [31], discussed in [10], selection on HDAC9 could also be linked to muscling. HDAC9 is a known transcriptional repressor of myogenesis. Its expression has been shown to be affected by the callypige mutation in the sheep at the DLK1-DIO3 locus [32]. The HDAC9 signal corresponds to a selection signature in the Garut breed from Indonesia, a breed used in ram fights. Two selection signatures contain genes shown to be underlying QTLs with large effects on milk production (yield and composition) in cattle: ABCG2 [16] and SREBP1 [33]. The SREBP1 gene is also found in a genome region associated with milk composition in the Lacaune breed

(unpublished data). Also, one of the ancestral selection signatures lies close to the INSIG2 gene, in the SREBP1 signaling pathway and recently shown to be associated with milk fatty acid composition in Holstein cattle [34]. Two selection signatures relate to wool characteristics, one in the CEU group near the FGF5 gene, partly responsible for hair type in the domestic dog [35], and an ancestral selection signature on chromosome 25 in a QTL region associated to wool quality traits in the sheep [36, 37].

One of the strong outlying regions in the selection scan contains the PITX3 gene. Further analysis revealed that this signature was due to the German Texel population haplotype diversity differing from the other Texel samples (results not shown). It turns out that the German Texel sample consisted of a case/control study for microphtalmia [6], although the case/control status information in this sample is not given in the Sheep Hapmap dataset. The consequence of such a recruitment is to bias haplotype frequencies in the region associated with the disease, which provokes a very strong differentiation signal between the German Texel and the other Texel populations. This illustrates that our method for detecting selection has the potential to identify causal variants in case/control studies, while using haplotype information.

Ancestral signatures of selection

It is difficult to estimate how far back in time signatures of selection found in the ancestral tree appended. In particular, it would be interesting to place this population tree with respect to sheep domestication. Two genes lying close to ancestral selection signatures might indicate that the selection signatures captured could be rather old. First, we found selection near the TRPM8 gene, which has been shown to be a major determinant of cold perception in the mouse [38]. The pattern of allele frequency at the significant SNP (OAR1_6722309) is consistent with the climate in the geographical origins of the population groups. AFR, ASI and ITA, living in warm climates, have low frequency (0.04-0.16) of the A allele, while NEU and CEU, from colder regions, have higher frequencies (0.55-0.7), the SWE group having an intermediate frequency of 0.38. Overall, this selection signature might be due to an adaptation to cold climate through selection on a TRPM8 variant. Another selection signature lies close to a potential chicken domestication gene, TSHR [39], which signaling regulates photoperiodic control of reproduction [40]. This selection signature was identified before [4] and our analysis indicates that it happened in the ancestral population tree, consistent with an early selection event. Given its role, we can speculate that selection on TSHR gene is related to seasonality of reproduction. Under temperate climates, sheep experience a reproductive cycle under photoperiodic control. Furthermore, there is evidence that this control was altered during domestication [41] so our analysis suggests genetic mutations in TSHR may have contributed to this alteration.

As discussed above, some of the genes found underlying ancestral selection signatures can be related to production or morphological traits (*e.g.* ASIP, INSIG2, ACAN, wool QTL), indicating that these traits have likely been important at the beginning of the sheep history. The other genes that we could identify as likely selection targets in the ancestral population tree relate to immune response (GATA3) and in particular to antirival response (TMEM154 [42], TRAF3 [43]). The most significant ancestral selection signature coincides with the NF1 gene, encoding neurofibromin. This gene is a negative regulator of the ras signal transduction pathway, therefore involved in cell proliferation and cancer, in particular neurofibromatosis. Due to this central role in intra-cellular signaling, mutations affecting this gene can have many phenotypic consequences so that its role in the adaptation of sheep breeds remains unclear.

		Estimated ancestral allele frequencies									
chr	pos	AFR	ASI	SWA	NEU	CEU	ITA	SWE	P-value	Q-value	candidate gene
1	7192190	0.15	0.08	0.16	0.55	0.69	0.04	0.38	1.7e-06	5.3e-03	TRPM8
1	237070498	0.87	0.95	0.91	0.48	0.24	0.77	0.35	1.4e-05	2.5e-02	GYG1
1	239424807	0.46	0.68	0.06	0.21	0.15	0.11	0.17	3.4e-05	4.8e-02	
1	239491620	0.53	0.41	0.94	0.86	0.93	0.93	0.88	4.3e-05	5.6e-02	
2	45500785	0.43	0.91	0.23	0.76	0.87	0.87	0.93	2.2e-06	6.4e-03	LPL
2	182607165	0.99	0.97	0.18	0.64	0.73	0.83	0.64	3.4e-08	1.8e-04	INSIG2
2	182672296	0.99	0.94	0.32	0.9	0.86	0.89	0.81	7.7e-07	2.8e-03	
2	192231314	0.59	0.93	0.36	0.96	0.89	0.81	0.95	1.6e-05	2.8e-02	
3	132478420	0.24	0.89	0.18	0.93	0.81	0.84	0.82	1.2e-06	3.9e-03	HOXC †
3	180860403	0.71	0.53	0.28	0.82	0.31	0.12	0.13	1.7e-05	2.8e-02	
5	15522700	0.68	0.63	0.92	0.27	0.76	0.99	0.78	9.8e-06	2.0e-02	
7	89519883	0.63	0.61	0.19	0.89	0.18	0.6	0.95	6.1e-10	5.2e-06	TSHR †
8	31748642	0.84	0.93	0.94	0.16	0.63	0.47	0.19	2.8e-05	4.1e-02	PREP/BVES †
11	18248852	0.35	0.32	0.82	0.64	0.94	0.96	0.92	1.3e-05	2.5e-02	NF1 †
11	18325488	0.87	0.93	0	0.35	0.04	0.03	0.04	3.3e-16	7.2e-12	
11	18335747	0.87	0.93	0	0.35	0.04	0.03	0.04	3.3e-16	7.2e-12	
11	18433474	0.87	0.93	0.02	0.35	0.07	0.02	0.05	3.8e-15	5.4e-11	
11	18440783	0.78	0.93	0.02	0.34	0.07	0.02	0.05	2.0e-14	2.2e-10	
11	25704651	0.97	0.96	0.97	0.42	0.94	0.94	0.96	8.5e-06	1.9e-02	
11	26284826	0.99	0.97	0.94	0.38	0.93	0.95	0.79	3.2e-05	4.6e-02	
11	26571629	0.92	0.94	0.98	0.29	0.89	0.88	0.86	1.8e-05	2.8e-02	
11	26872280	0.78	0.71	0.93	0.15	0.89	0.9	0.9	2.2e-07	9.5e-04	
13	12120674	0.29	0.84	0.97	0.91	0.97	0.92	0.84	7.7e-06	1.8e-02	GATA3
13	62857560	0.52	0.62	0.65	0.98	0.67	0.92	0.36	3.6e-06	9.7e-03	ASIP †
15	3706790	0.71	0.22	0.96	0.28	0.27	0.34	0.21	6.8e-06	1.7e-02	
15	29856310	0.98	0.99	0.99	0.47	0.92	0.95	0.96	9.8e-06	2.0e-02	
16	38696505	0.95	0.98	0.95	0.99	0.68	0.31	0.3	6.8e-07	2.7e-03	PRLR †
17	4867509	0.91	0.95	0.85	0.54	0.18	0.58	0.17	1.8e-05	2.8e-02	TMEM154
18	19342316	0.9	0.79	0.67	0.35	0.75	0.1	0.09	1.9e-07	9.3e-04	ACAN †
18	66470371	0.99	0.97	0.9	0.9	0.18	0.04	0.08	1.9e-09	1.3e-05	TRAF3
20	17381047	0.24	0.61	0.97	0.98	0.93	0.99	0.91	3.1e-08	1.8e-04	VEGFA †
25	7517270	0.95	0.94	0.93	0.14	0.27	0.57	0.19	1.8e-05	2.8e-02	wool QTL †

Table 2. Selection signatures in ancestral populations. Significant SNPs at the 5% FDR level. †: signatures of selection previously identified [4].

Conclusions

We conducted a genome scan for selection in a large worldwide set of breeds from the Sheep Hapmap dataset. Using recently developed methods, we were able to detect a very large number of selection signatures in different geographical groups. We also found selection signatures that most likely predate the formation of contemporary breeds. This analysis reveals strong response of the genome diversity in sheep populations with respect to selection on morphology and color, and the influence of recent selection on production traits. We also pinpoint two strong candidate genes (TRPM8 and TSHR) most likely involved in selection response during the early history of domestic sheep.

Elucidating causal variation underlying these selection signatures will most likely require large scale

sequencing projects, together with phenotypic characterization of individuals or populations. This study can help in targeting specific breeds and traits to be studied in priority in such projects.

Methods

Selecting populations and animals Seventy four breeds are represented in the Sheep HapMap data set, but we only used a subset of these breeds in our genome scan. We removed the breeds with small sample size (< 20 animals), for which haplotype diversity can not be determined with sufficient precision. Based on historical information, we also removed all breeds resulting from a recent admixture or having experienced a severe recent bottleneck. Focusing on the remaining breeds, we then studied the genetic structure within each population group, in order to detect further admixture events. We performed a standardized PCA of individual based genotype data and applied the admixture software [12].

In two population groups (AFR and NEU) the different breeds were clearly separated into distinct clusters of the PCA and showed no evidence of recent admixture. These samples were left unchanged for the genome scan for selection. A similar pattern was observed in three other groups (ITA, SWA, ASI), except for a few outlier animals that had to be re-attributed to a different breed or simply removed (Figures S1, S2 and S3). In the two last groups (CEU and SWE), several admixed breeds were found and were consequently removed from the genome scan analysis (Figures S4, S5).

We performed a genome scan within each group of populations listed in table S1, with a single SNP statistic FLK [9] and its haplotype version hapFLK [10].

Population trees Both statistics require estimating the population tree, with a procedure described in details in [9]. Briefly, we built a population tree for each group by first calculating Reynolds' distances between each population, and then applying the Neighbour Joining algorithm on the distance matrix. For each group, we rooted the tree using the Soay sheep as an outgroup. This breed has been isolated on an Island for many generations and exhibits a very strong differentiation with all the breeds of the Sheep hapmap dataset, making it well suited to be used as an outgroup.

FLK and hapFLK genome scans The FLK statistic was computed for each SNP within each group. The evolutionary model underlying the FLK statistic assumes that the mutation was present in the ancestral population. To consider only loci that most likely match this hypothesis, we restricted our analysis within each group to SNPs which estimated ancestral minor allele frequency p_0 was above 5%. Under neutrality, the FLK statistic should follow a χ^2 distribution with $n-1$ degrees of freedom (DF), where n is the number of populations in the group. Overall, the fit of the theoretical distribution to the observed distribution was very good (supporting information Text S1) with the mean of the observed distribution ($\overline{FLK}$) being very close to $n-1$ (table S4). Using $\overline{FLK}$ as DF for the χ^2 distribution provided a better fit to the observed data than the $n-1$ theoretical value. We thus computed FLK p-values using the $\chi^2(\overline{FLK})$ distribution. To compute the hapFLK statistic, we make use of the Scheet and Stephens LD model [44], a mixture model for haplotypes which requires specifying a number of haplotype clusters to be used. To choose this number, for each group, we used the fastPHASE cross-validation based estimation of the optimal number of clusters. Results of this estimation are given in table S3. The LD model was estimated on unphased genotype data. The hapFLK statistic is computed as an average over 20 runs of the EM algorithm to fit the LD model. As in [10], we found that the hapFLK distribution could be modelled relatively well with a normal distribution (corresponding to non outlying regions) and a few outliers; we used robust estimation of the mean and standard deviation of the hapFLK statistic to eliminate the influence of outlying (*i.e.* potentially selected) regions. This procedure was done within each group, the resulting mean and standard deviation obtained are given in table S3. Finally, we computed at each SNP a p-value for the null hypothesis from the normal distribution.

Selection in ancestral groups The within-group FLK analysis provides for each SNP an estimation of the allele frequency p_0 in the population ancestral to all populations of the group. We used this information to test SNP for selection using between groups differentiation, with some adjustments. First, the FLK model assumes tested polymorphisms are present in the ancestral population. SNPs for which the alternate allele has been seen in only one population group are likely to have appeared after divergence (within the ancestral tree) and were therefore removed of the analysis. Second, regions selected within groups affect allele frequency in some breeds and therefore bias our estimation of the ancestral allele frequency in this group. We therefore removed all SNPs that were included in within-group selection signatures. Finally, the FLK test requires a rooted population tree. For the within group analysis, we could use a very distant population to the current breeds (the Soay sheep). For the ancestral tree, we created an outgroup homozygous for ancestral alleles at all SNPs.

Identifying selected regions and candidate genes We defined significant regions for each statistic and within each group of populations. Using the neutral distribution (χ^2 for FLK and Normal for hapFLK), we computed the p-value of each statistic at each SNP. To identify selected regions, we estimated their q-value [45] to control the FDR. For FLK, we called significant SNPs with q-values less than 0.1 (therefore controlling the FDR at the 10% level). As the power of hapFLK is greater than that of FLK [10], we used an FDR threshold of 5%. For the FLK analysis in ancestral populations, we used an FDR threshold of 5%.

We then aimed at identifying genes that seem good candidates for explaining selection signatures. We proceeded differently for the single SNP FLK and hapFLK. For FLK, we considered that significant SNP less than 500Kb apart were capturing the same selection signal. Then, we considered as potential candidate genes any gene that lie less than 500Kb of any significant SNP. For hapFLK, the genome signal is much more continuous than single SNP tests, because the statistic captures multipoint LD with the selected mutations. A consequence is that the significant regions can span large chromosome intervals. To restrict the list of potential candidate genes, and target only the ones closest to the most significant SNP, we restricted our search to the part of the signal where the difference in hapFLK value with the most significant SNP was less than 0.5σ. This allowed to take into consideration the profile of the hapFLK signal, *i.e.* if the profile ressembles a plateau, the candidate region will be rather broad while very sharp hapFLK peaks will provide a narrower candidate region. We listed all the genes present in the significant regions using the OAR3.1 genome browser at `http://www.livestockgenomics.csiro.au/cgi-bin/gbrowse/oarv3.1/`.

Some very likely candidate genes for selection were found in many of the significant regions. This is for example the case of the MSTN (GDF8) gene on chromosome 2 in the NEU group. In these cases, we did not list any other candidates in the region, *i.e.* we made a strong prior assumption of selection for these genes. Note however that we provide the position of the selected regions for the reader interested in knowing all the genes present in significant regions.

References

1. Andersson L (2012) How selective sweeps in domestic animals provide new insight into biological mechanisms. J Intern Med 271: 1-14.

2. Moradi MH, Nejati-Javaremi A, Moradi-Shahrbabak M, Dodds KG, McEwan JC (2012) Genomic scan of selective sweeps in thin and fat tail sheep breeds for identifying of candidate regions associated with fat deposition. BMC Genet 13: 10.

3. Kijas JW, Townley D, Dalrymple BP, Heaton MP, Maddox JF, et al. (2009) A genome wide survey of SNP variation reveals the genetic structure of sheep breeds. PLoS One 4: e4668.

4. Kijas JW, Lenstra JA, Hayes B, Boitard S, Porto Neto LR, 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: e1001258.

5. Garcia-Gámez E, Reverter A, Whan V, McWilliam SM, Arranz JJA, et al. (2011) Using regulatory and epistatic networks to extend the findings of a genome scan: identifying the gene drivers of pigmentation in merino sheep. PLoS ONE 6: e21158.

6. Becker D, Tetens J, Brunner A, Burstel D, Ganter M, et al. (2010) Microphthalmia in Texel sheep is associated with a missense mutation in the paired-like homeodomain 3 (PITX3) gene. PLoS One 5: e8689.

7. Zhang L, Mousel MR, Wu X, Michal JJ, Zhou X, et al. (2013) Genome-Wide Genetic Diversity and Differentially Selected Regions among Suffolk, Rambouillet, Columbia, Polypay, and Targhee Sheep. PLoS One 8: e65942.

8. Excoffier L, Hofer T, Foll M (2009) Detecting loci under selection in a hierarchically structured population. Heredity (Edinb) 103: 285-98.

9. Bonhomme M, Chevalet C, Servin B, Boitard S, Abdallah J, et al. (2010) Detecting selection in population trees: the Lewontin and Krakauer test extended. Genetics 186: 241-62.

10. Fariello MI, Boitard S, Naya H, SanCristobal M, Servin B (2013) Detecting signatures of selection through haplotype differentiation among hierarchically structured populations. Genetics 193: 929-41.

11. Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. Genetics 155: 945-59.

12. Alexander DH, Novembre J, Lange K (2009) Fast model-based estimation of ancestry in unrelated individuals. Genome Res 19: 1655-64.

13. Johnston SE, McEwan JC, Pickering NK, Kijas JW, Beraldi D, et al. (2011) Genome-wide association mapping identifies the genetic basis of discrete and quantitative variation in sexual weaponry in a wild sheep population. Mol Ecol 20: 2555-66.

14. Dominik S, Henshall JM, Hayes BJ (2012) A single nucleotide polymorphism on chromosome 10 is highly predictive for the polled phenotype in Australian Merino sheep. Anim Genet 43: 468-70.

15. Akey JM, Ruhe AL, Akey DT, Wong AK, Connelly CF, et al. (2010) Tracking footprints of artificial selection in the dog genome. Proc Natl Acad Sci U S A 107: 1160-5.

16. Cohen-Zinder M, Seroussi E, Larkin DM, Loor JJ, Everts-van der Wind A, et al. (2005) Identification of a missense mutation in the bovine ABCG2 gene with a major effect on the QTL on chromosome 6 affecting milk yield and composition in Holstein cattle. Genome Res 15: 936-44.

17. Jacobs LC, Wollstein A, Lao O, Hofman A, Klaver CC, et al. (2013) Comprehensive candidate gene study highlights UGT1A and BNC2 as new genes determining continuous skin color variation in Europeans. Hum Genet 132: 147-58.

18. Hauswirth R, Haase B, Blatter M, Brooks Sa, Burger D, et al. (2012) Mutations in MITF and PAX3 cause "splashed white" and other white spotting phenotypes in horses. PLoS genetics 8: e1002653.

19. Lin JY, Fisher DE (2007) Melanocyte biology and skin pigmentation. Nature 445: 843–50.

20. Klungland H, Vå ge DI, Gomez-Raya L, Adalsteinsson S, Lien S (1995) The role of melanocyte-stimulating hormone (MSH) receptor in bovine coat color determination. Mammalian genome 6: 636–9.

21. Joerg H, Fries HR, Meijerink E, Stranzinger GF (1996) Red coat color in Holstein cattle is associated with a deletion in the MSHR gene. Mammalian genome 7: 317–8.

22. Linnen CR, Poh YP, Peterson BK, Barrett RDH, Larson JG, et al. (2013) Adaptive evolution of multiple traits through multiple mutations at a single gene. Science 339: 1312–6.

23. Sutter NB, Bustamante CD, Chase K, Gray MM, Zhao K, et al. (2007) A single IGF1 allele is a major determinant of small size in dogs. Science 316: 112-5.

24. Kuijper S, Feitsma H, Sheth R, Korving J, Reijnen M, et al. (2005) Function and regulation of Alx4 in limb development: complex genetic interactions with Gli3 and Shh. Developmental biology 285: 533–44.

25. Qu S, Tucker SC, Ehrlich JS, Levorse JM, Flaherty LA, et al. (1998) Mutations in mouse Aristaless-like4 cause Strong's luxoid polydactyly. Development 125: 2711-21.

26. Stickens D, Zak BM, Rougier N, Esko JD, Werb Z (2005) Mice deficient in Ext2 lack heparan sulfate and develop exostoses. Development 132: 5055–68.

27. Person AD, Beiraghi S, Sieben CM, Hermanson S, Neumann AN, et al. (2010) WNT5A mutations in patients with autosomal dominant Robinow syndrome. Developmental dynamics 239: 327–37.

28. Stattin EL, Wiklund F, Lindblom K, Onnerfjord P, Jonsson BA, et al. (2010) A missense mutation in the aggrecan C-type lectin domain disrupts extracellular matrix interactions and causes dominant familial osteochondritis dissecans. American journal of human genetics 86: 126–37.

29. Tompson SW, Merriman B, Funari Va, Fresquet M, Lachman RS, et al. (2009) A recessive skeletal dysplasia, SEMD aggrecan type, results from a missense mutation affecting the C-type lectin domain of aggrecan. American journal of human genetics 84: 72–9.

30. Weedon MN, Lango H, Lindgren CM, Wallace C, Evans DM, et al. (2008) Genome-wide association analysis identifies 20 loci that influence adult height. Nature genetics 40: 575–83.

31. Clop A, Marcq F, Takeda H, Pirottin D, Tordoir X, et al. (2006) A mutation creating a potential illegitimate microRNA target site in the myostatin gene affects muscularity in sheep. Nat Genet 38: 813-8.

32. Fleming-Waddell JN, Olbricht GR, Taxis TM, White JD, Vuocolo T, et al. (2009) Effect of DLK1 and RTL1 but not MEG3 or MEG8 on muscle gene expression in Callipyge lambs. PloS one 4: e7399.

33. Bouwman AC, Bovenhuis H, Visker MH, van Arendonk JA (2011) Genome-wide association of milk fatty acids in Dutch dairy cattle. BMC Genet 12: 43.

34. Rincon G, Islas-Trejo A, Castillo AR, Bauman DE, German BJ, et al. (2012) Polymorphisms in genes in the SREBP1 signalling pathway and SCD are associated with milk fatty acid composition in Holstein cattle. J Dairy Res 79: 66-75.

35. Cadieu E, Neff MW, Quignon P, Walsh K, Chase K, et al. (2009) Coat variation in the domestic dog is governed by variants in three genes. Science 326: 150-3.

36. Ponz R, Moreno C, Allain D, Elsen JM, Lantier F, et al. (2001) Assessment of genetic variation explained by markers for wool traits in sheep via a segment mapping approach. Mamm Genome 12: 569-72.

37. Bidinost F, Roldan D, Dodero A, Cano E, Taddeo H, et al. (2007) Wool quantitative trait loci in merino sheep. Small Ruminant Research 74: 113 - 118.

38. Bautista DM, Siemens J, Glazer JM, Tsuruda PR, Basbaum AI, et al. (2007) The menthol receptor TRPM8 is the principal detector of environmental cold. Nature 448: 204–8.

39. Rubin CJ, Zody MC, Eriksson J, Meadows JRS, Sherwood E, et al. (2010) Whole-genome resequencing reveals loci under selection during chicken domestication. Nature 464: 587–91.

40. Nakao N, Ono H, Yamamura T, Anraku T, Takagi T, et al. (2008) Thyrotrophin in the pars tuberalis triggers photoperiodic response. Nature 452: 317-22.

41. Balasse M, Tresset A (2007) Environmental constraints on the reproductive activity of domestic sheep and cattle : what latitude for the herder ? Anthropozoologica 42: 71–88.

42. Heaton MP, Clawson ML, Chitko-Mckown CG, Leymaster Ka, Smith TPL, et al. (2012) Reduced lentivirus susceptibility in sheep with TMEM154 mutations. PLoS genetics 8: e1002467.

43. Oganesyan G, Saha SK, Guo B, He JQ, Shahangian A, et al. (2006) Critical role of TRAF3 in the Toll-like receptor-dependent and -independent antiviral response. Nature 439: 208–11.

44. Scheet P, Stephens M (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: 629-44.

45. Storey JD, Tibshirani R (2003) Statistical significance for genomewide studies. Proc Natl Acad Sci U S A 100: 9440-5.

Supplementary material

Group	Abbreviation	Size	Populations (Abbrev.)
Africa	AFR	2	Red Maasai (RMA)
			Ethiopian Menz (EMZ)
Asia	ASI	8	Bangladeshi BGE (BGE)
			Bangladeshi Garole (BGA)
			Changthangi (CHA)
			Deccani (IDC)
			Garut (GUR)
			Indian Garole (GAR)
			Sumatra (SUM)
			Tibetan (TIB)
Central Europe	CEU	4	Bundner Oberlander (BOS)
			Engadine Red (ERS)
			Valais Blacknose (VBS)
			Valais Red (VRS)
Italy	ITA	4	Altamurana (ALT)
			Comisana (COM)
			Leccese (LEC)
			Sardinian Ancestral Black (SAB)
Northern Europe	NEU	6	Galway (GAL)
			German Texel (GTX)
			Irish Suffolk (ISF)
			New Zealand Texel (NTX)
			New Zealand Romney (NZR)
			Scottish Texel (STX)
South West Europe	SWE	4	Autralian Merino (MER)
			Churra (CHU)
			Meat Lacaune (LAM)
			Milk Lacaune (LAC)
South West Asia	SWA	4	Afshari (AFS)
			Moghani (MOG)
			Norduz (NDZ)
			Qezel (QEZ)

Table S1. Population groups used for the detection of selection signatures

Candidate Gene	Populations
BNC2	COM
KITLG	COM, EMZ
KIT	VBS
EDN3	MOG
MC1R	SUF, SAB, GAL
ASIP	MER
MITF	VBS, ERS, BOS

Table S2. List of genes potentially associated with coloration patterns found under selection signatures and the most likely candidate populations for selection.

group	K	μ	σ
AFR	5	1.97	0.53
ASI	25	7.78	0.92
CEU	10	3.41	0.62
ITA	15	2.83	0.44
NEU	40	4.09	0.45
SWA	10	3.28	0.55
SWE	25	3.32	0.39

Table S3. Parameters for the hapFLK genome scan. K : Number of haplotype clusters used in the LD model for each group, as determined by the fastPHASE cross-validation procedure. μ,σ mean and standard deviation of the normal distribution used to model the hapFLK neutral distribution

group	DF	$\overline{FLK}$
AFR	1	0.97
ASI	7	6.65
CEU	3	2.91
ITA	3	2.93
NEU	5	4.78
SWA	3	2.92
SWE	3	2.96

Table S4. Theoretical number of degrees of freedom (DF) and observed mean of the FLK statistic ($\overline{FLK}$) in each group of populations.

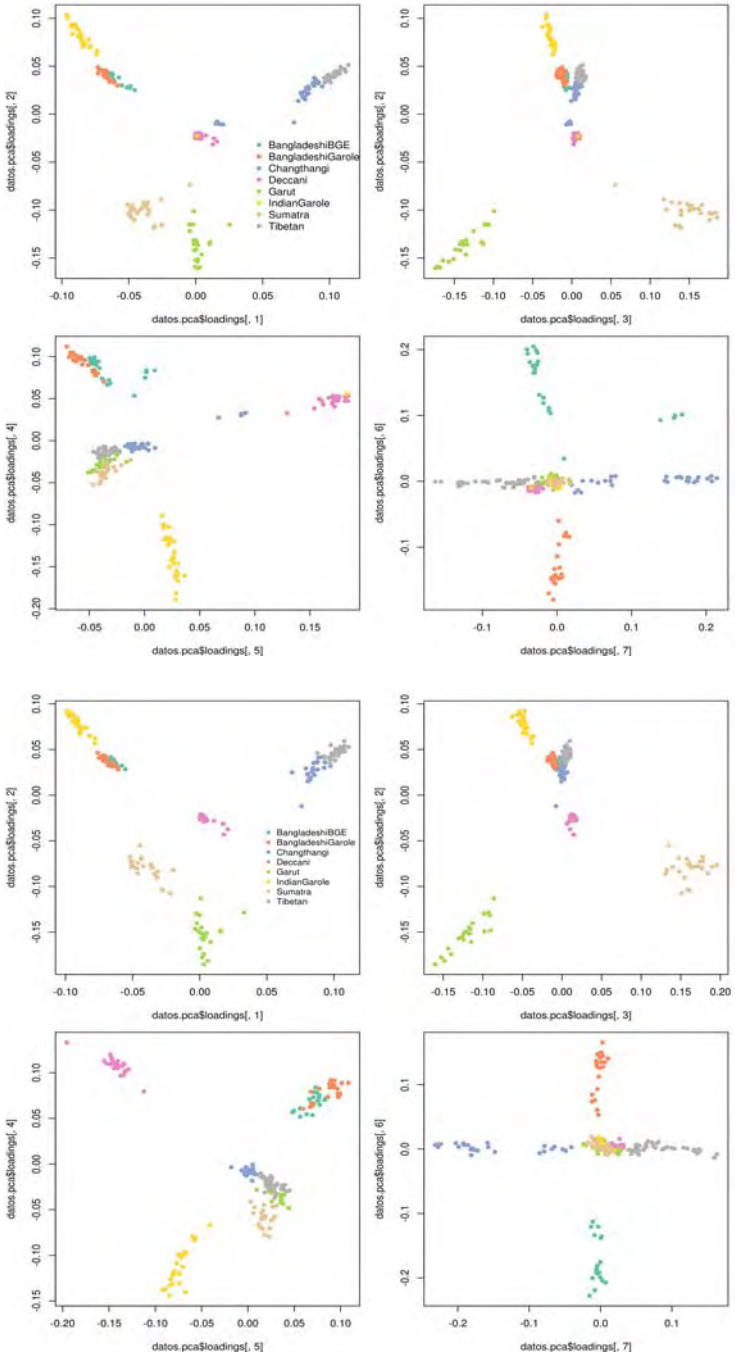


Figure S1. Projection of animals from the Asian group on the first 8 principal components, before (top) and after (bottom) correction. One Indian Garole animal located in the Deccani cluster was attributed to this breed. Five Bangladeshi BGE animals clustering away from the rest of the breed were removed. Four Changthangi abimals, which clustered away from the rest of the breed and appeared admixed with the Deccani breed, were removed. One outlier Sumatra animal was also removed.

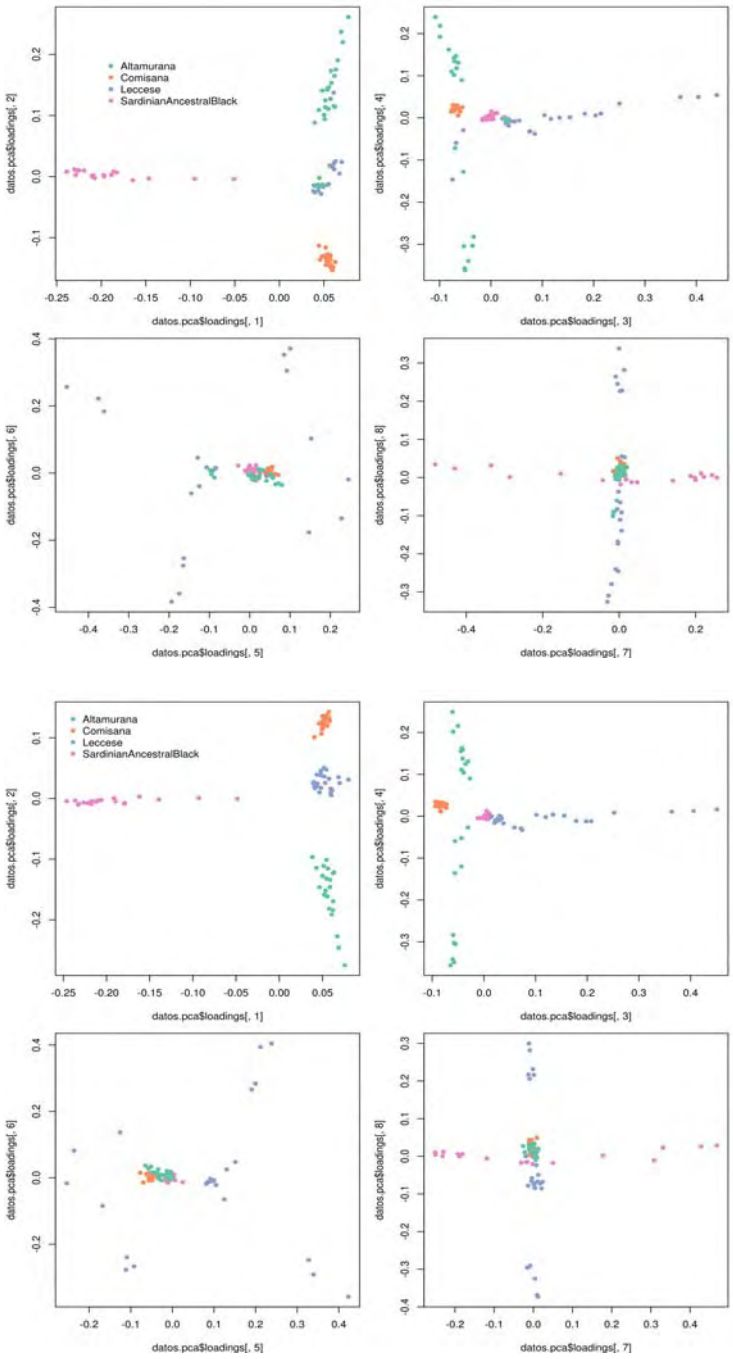


Figure S2. Projection of animals from the Italian group on the first 8 principal components, before (top) and after (bottom) correction. Four Altamurana animals located in the Leccese cluster were attirubted to the Leccese breed. Similarly, three Lecese animals located in the Altamurana cluster was attributed to the Altamurana cluster.

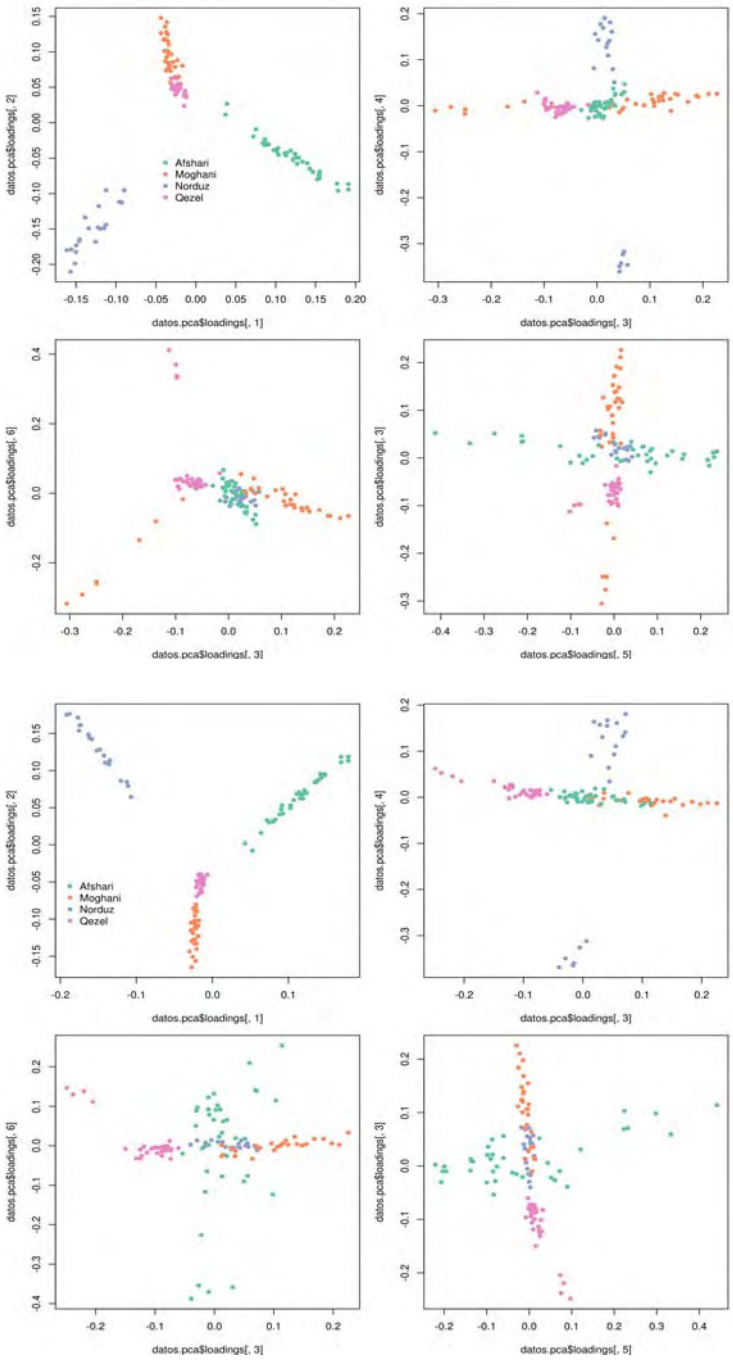


Figure S3. Projection of animals from the South West Asian group on the first 8 principal components, before (top) and after (bottom) correction. Seven Moghani animals clustering away from the rest of the breed, and possibly admixed with the Qezel breed, were removed.

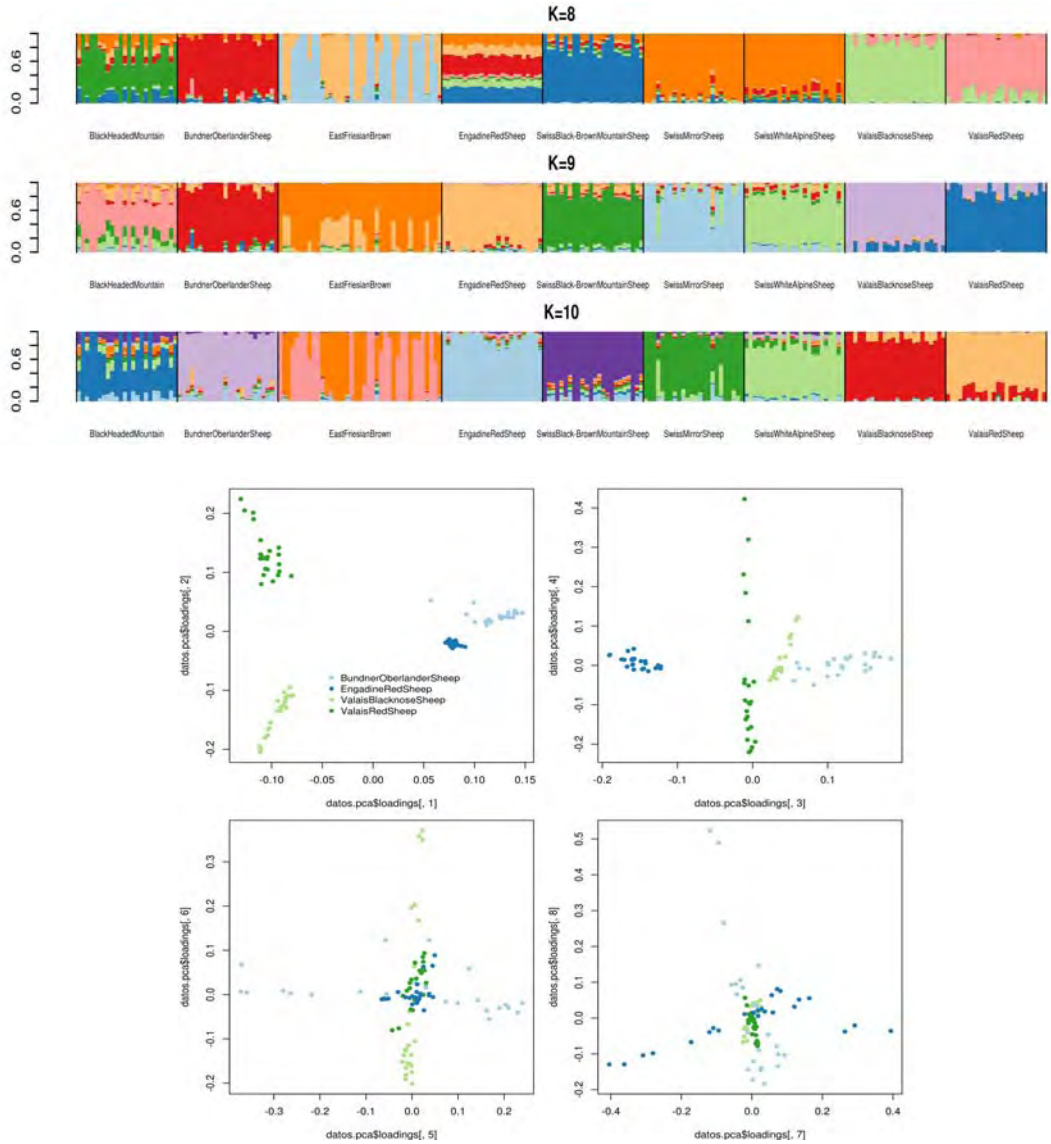


Figure S4. Admixture analysis for animals of the Central European group (top), and projection of the selected animals on the first 8 principal components (bottom). Three breeds (Swiss Mirror Sheep, Swiss Alpine White Sheep and Swiss Black-Brown Mountain Sheep) were reported as admixed breeds in [4] and were consequently removed, although they appeared relatively homogeneous in the admixture analysis. Two further breeds (East Friesian Brown and Black Headed Mountain) were removed based on the admixture analysis.

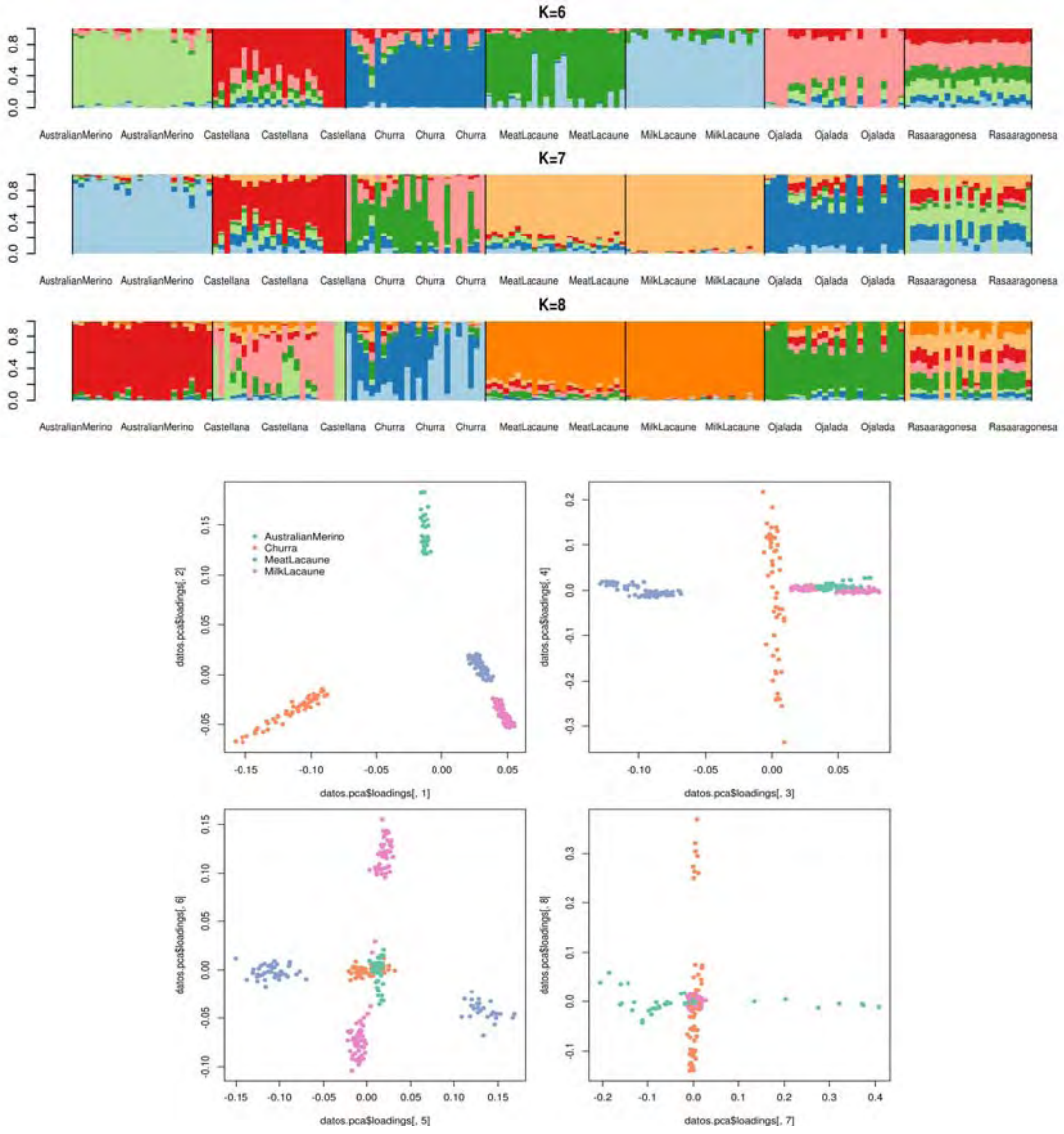
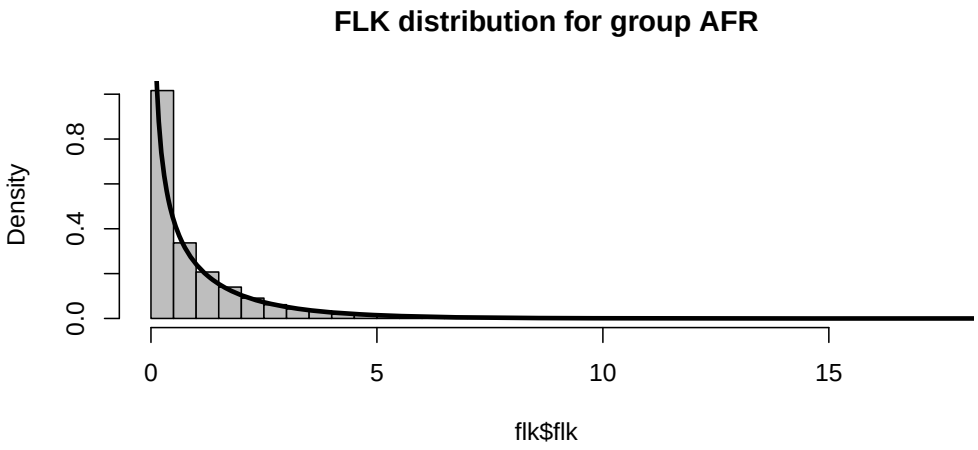


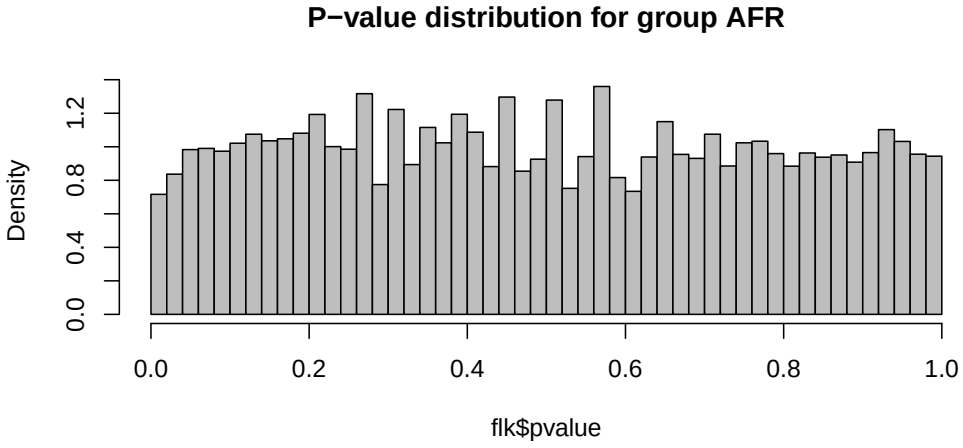
Figure S5. Admixture analysis for animals of the South West European group (top), and projection of the selected animals on the first 8 principal components (bottom). For the admixture analysis, a subset of 24 animals was sampled at random within each breed in order to obtain balanced sample sizes. Three breeds (Castellana, Ojalada and Rasaaragonesa) were removed based on the admixture analysis, as they were clearly admixed. Historical records concerning those 3 breeds are ambiguous, some of them reporting that they may result from a cross between Merino and Churra animals. The Churra breed also looked admixed in this analysis, but this was no longer the case after removing the three other problematic breeds.

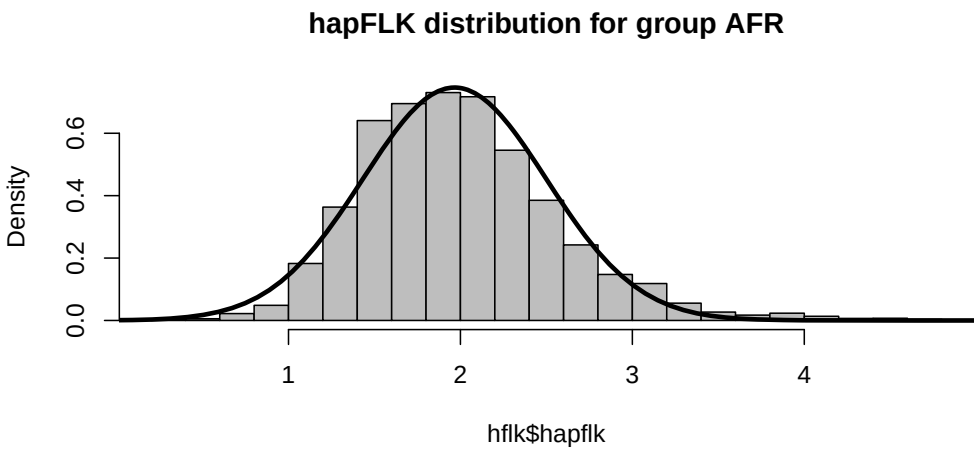
Text S1. FLK and hapFLK genome scans within groups of populations

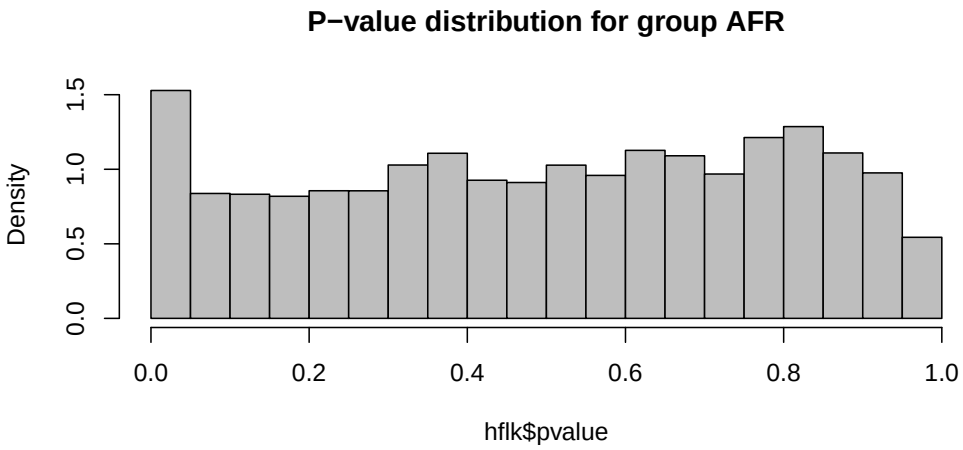
For each population group, we show, in this order

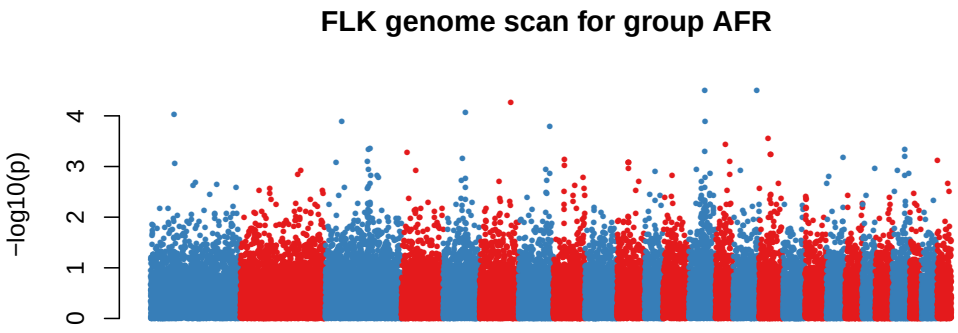
1. histogram of the observed FLK distribution, and corresponding theoretical χ^2 distribution
2. the FLK p-value distribution
3. histogram of the hapFLK distribution and corresponding estimated normal distribution (see details in Methods)
4. the hapFLK p-value distribution
5. Manhattan plot of the FLK p-values
6. Manhattan plot of the hapFLK p-values

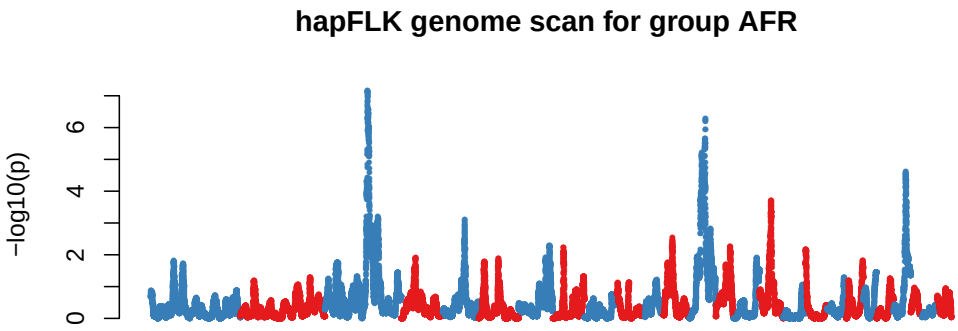

FLK distribution for group AFR
Density
0.0
0.4
0.8
0
5
10
15
flk$flk

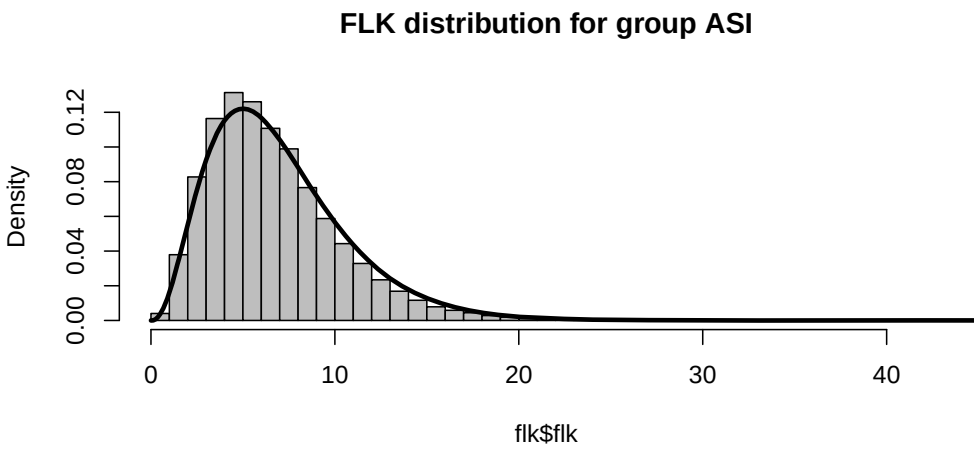

P−value distribution for group AFR
Density
0.0
0.4
0.8
1.2
0.0
0.2
0.4
0.6
0.8
1.0
flk$pvalue

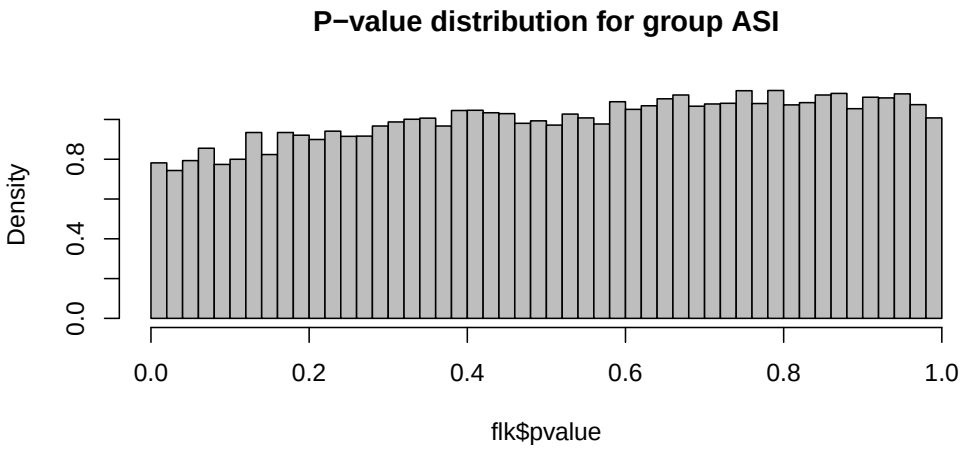

hapFLK distribution for group AFR
Density
0.0
0.2
0.4
0.6
1
2
3
4
hflk$hapflk

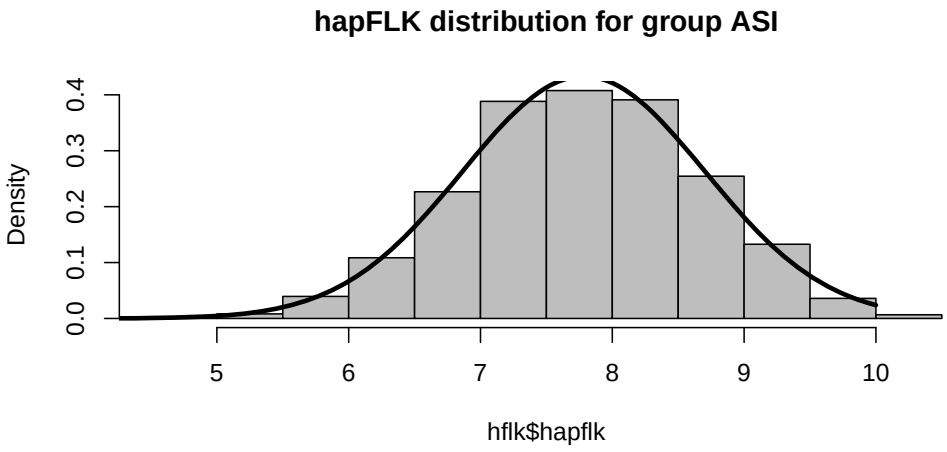

P−value distribution for group AFR
Density
0.0
0.5
1.0
1.5
0.0
0.2
0.4
0.6
0.8
1.0
hflk$pvalue

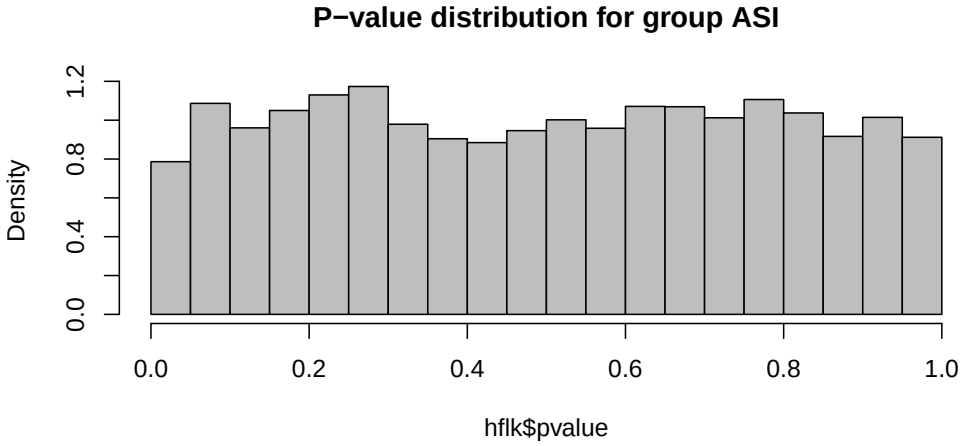

FLK genome scan for group AFR
−log10(p)
0
1
2
3
4

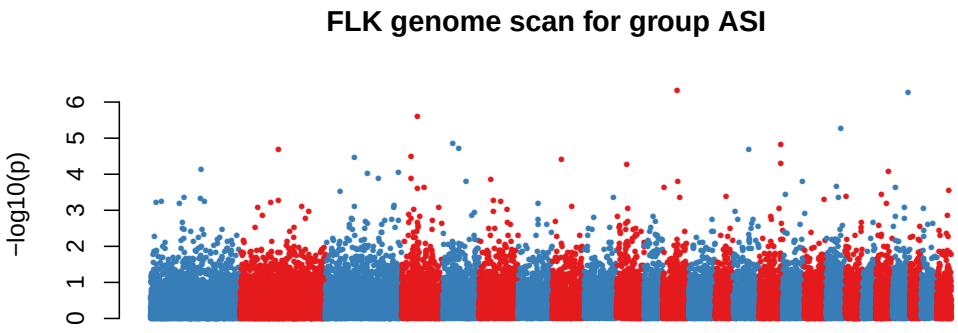

hapFLK genome scan for group AFR
−log10(p)
0
2
4
6

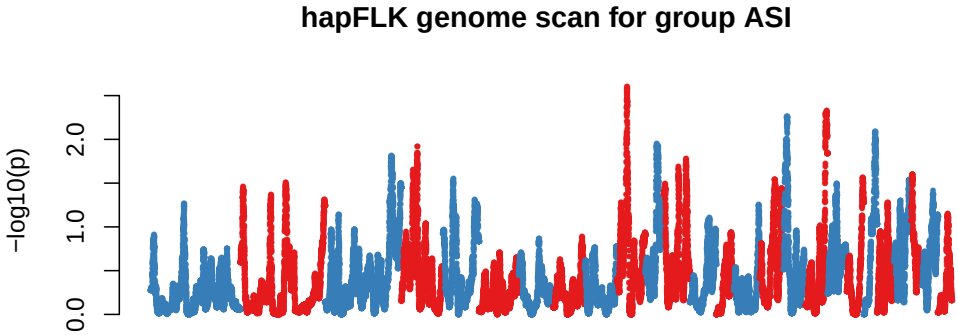

FLK distribution for group ASI
Density
0.00
0.04
0.08
0.12
0
10
20
30
40
flk$flk

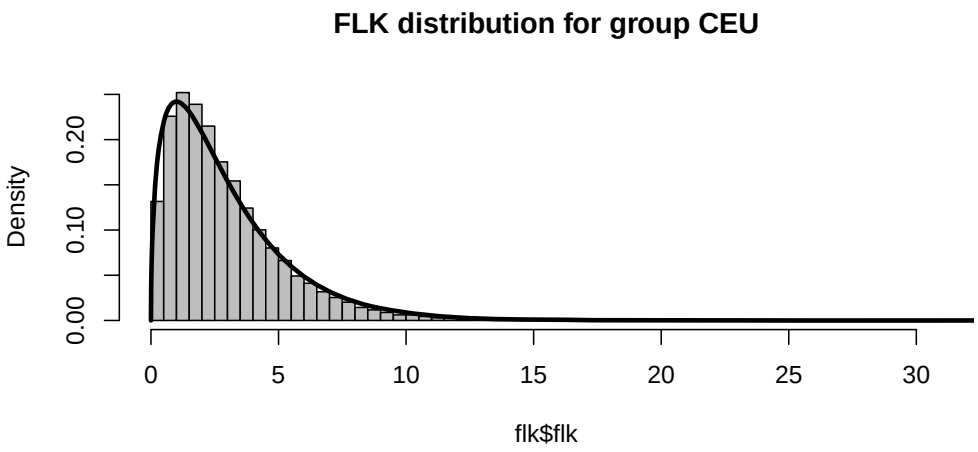

P−value distribution for group ASI
Density
0.0
0.4
0.8
0.0
0.2
0.4
0.6
0.8
1.0
flk$pvalue

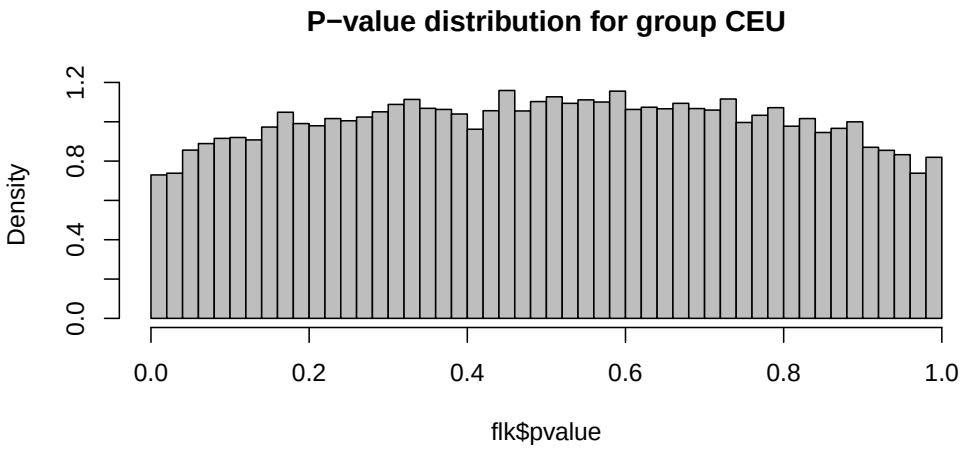

hapFLK distribution for group ASI
Density
0.0
0.1
0.2
0.3
0.4
5
6
7
8
9
10
hflk$hapflk

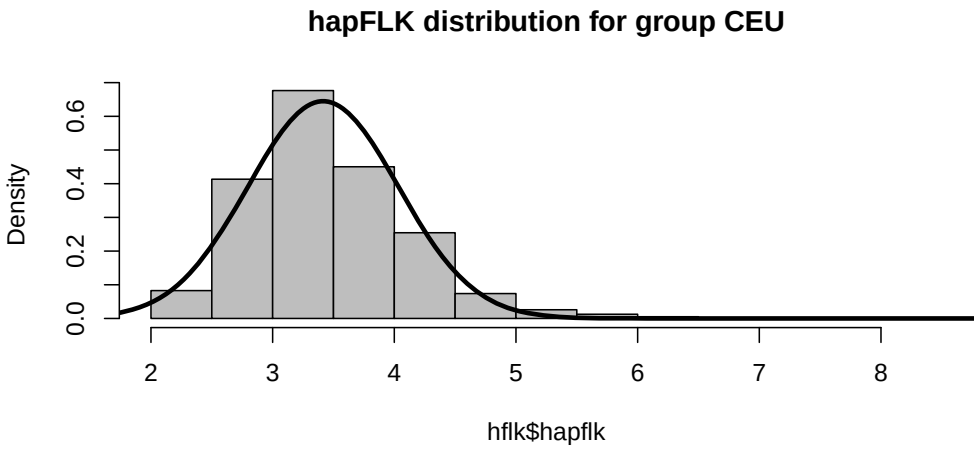

P−value distribution for group ASI
Density
0.0
0.4
0.8
1.2
0.0
0.2
0.4
0.6
0.8
1.0
hflk$pvalue

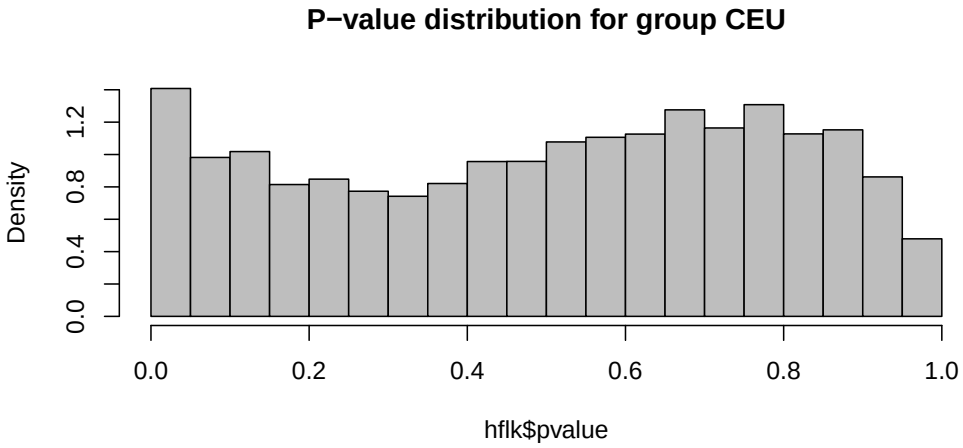

FLK genome scan for group ASI
−log10(p)
0
1
2
3
4
5
6

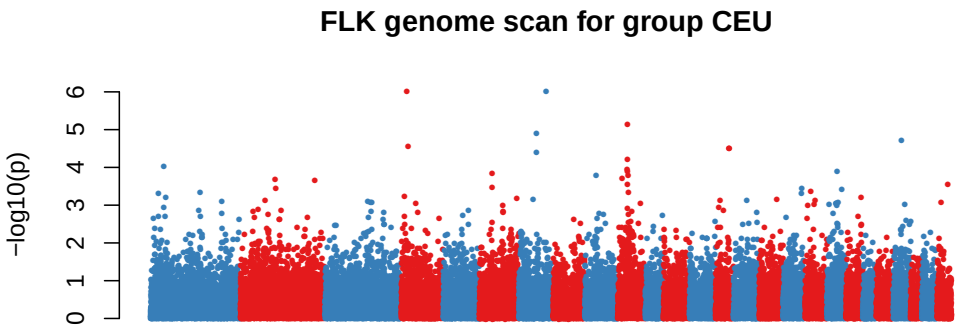

hapFLK genome scan for group ASI
−log10(p)
0.0
1.0
2.0

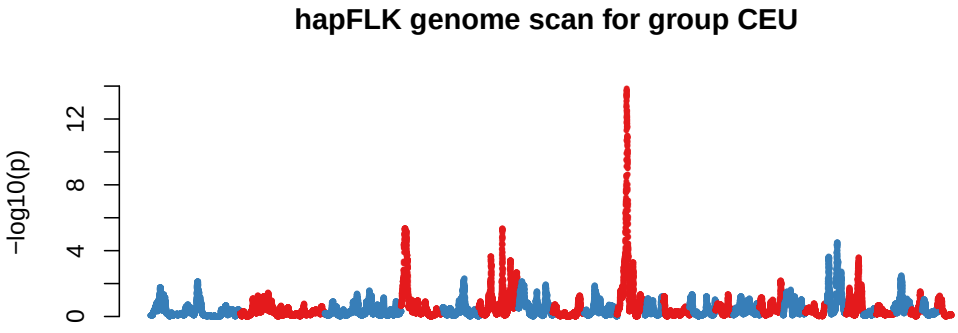

FLK distribution for group CEU
Density
0.00
0.10
0.20
0
5
10
15
20
25
30
flk$flk

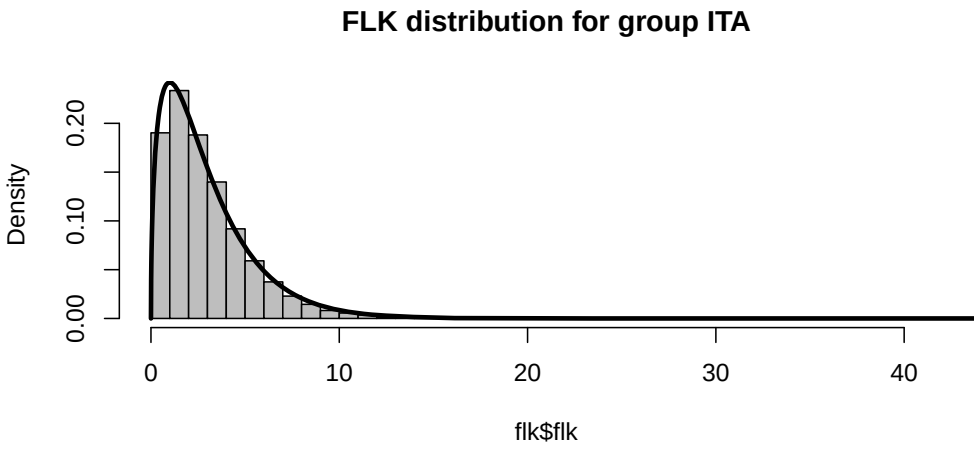

P−value distribution for group CEU
Density
0.0
0.4
0.8
1.2
0.0
0.2
0.4
0.6
0.8
1.0
flk$pvalue

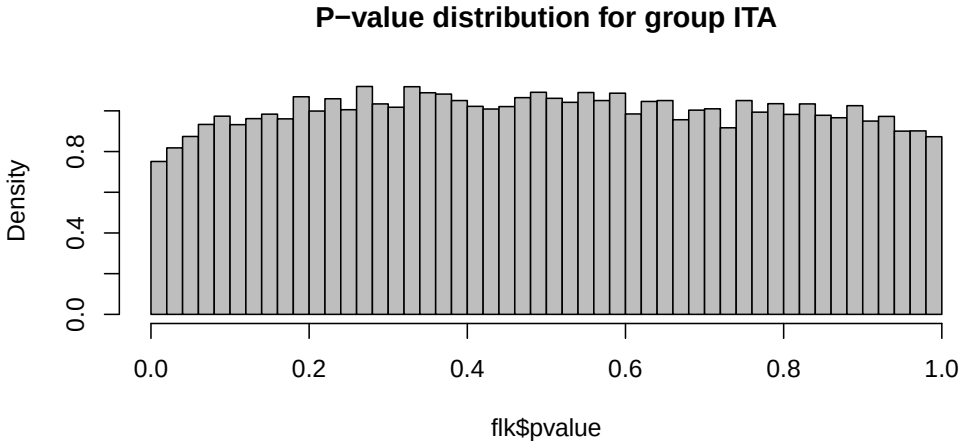

hapFLK distribution for group CEU
Density
0.0
0.2
0.4
0.6
2
3
4
5
6
7
8
hflk$hapflk

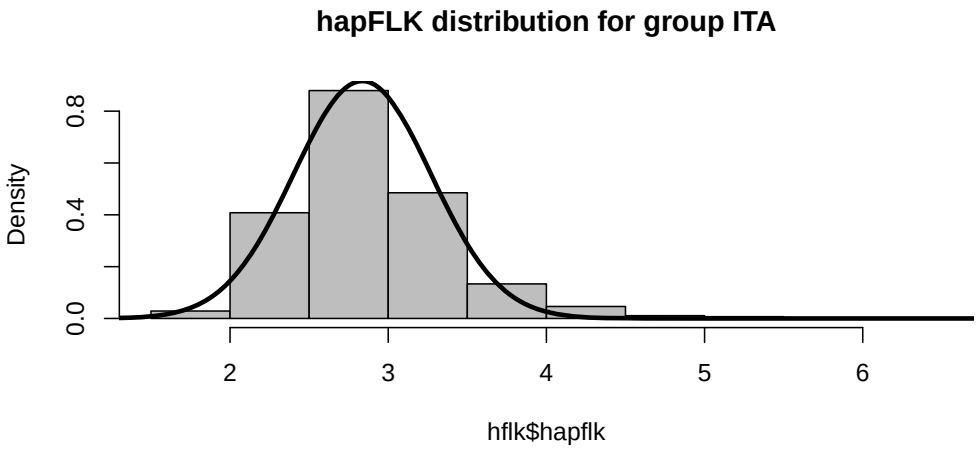

P−value distribution for group CEU
Density
0.0
0.4
0.8
1.2
0.0
0.2
0.4
0.6
0.8
1.0
hflk$pvalue

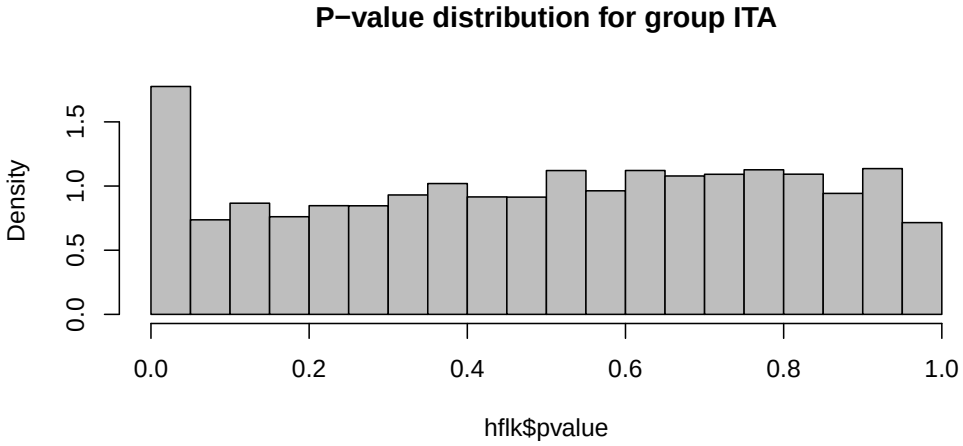

FLK genome scan for group CEU
−log10(p)
0
1
2
3
4
5
6

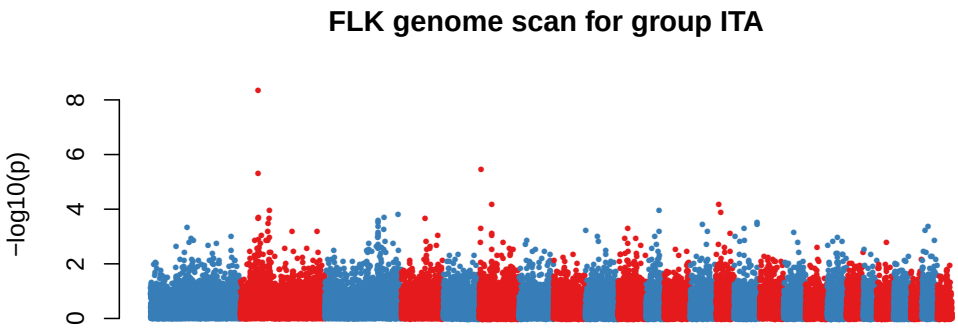

hapFLK genome scan for group CEU
−log10(p)
0
4
8
12

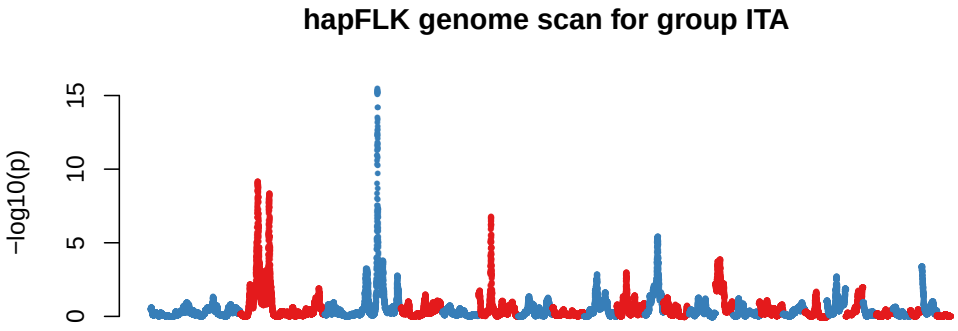

FLK distribution for group ITA
Density
0.00
0.10
0.20
0
10
20
30
40
flk$flk

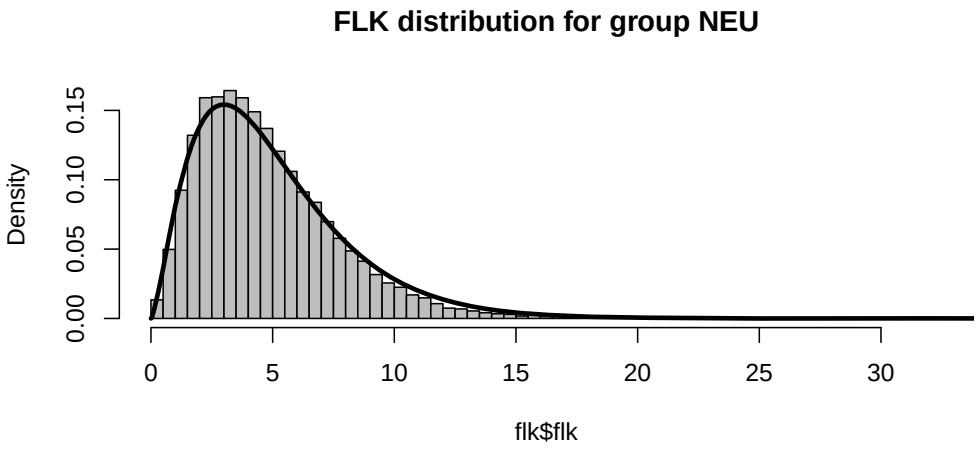

P−value distribution for group ITA
Density
0.0
0.4
0.8
0.0
0.2
0.4
0.6
0.8
1.0
flk$pvalue

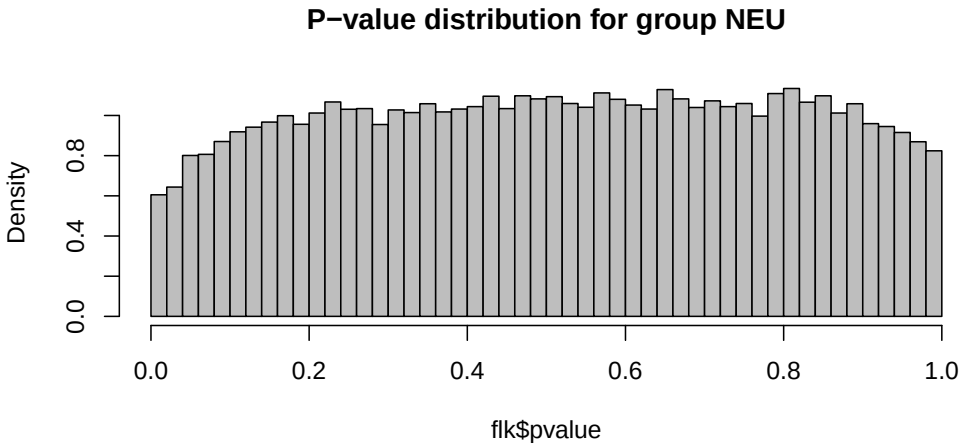

hapFLK distribution for group ITA
Density
0.0
0.4
0.8
2
3
4
5
6
hflk$hapflk

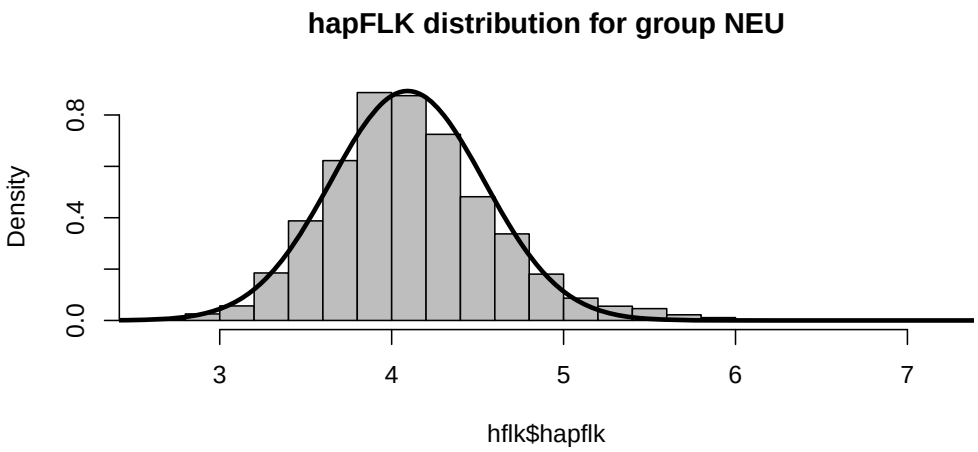

P−value distribution for group ITA
Density
0.0
0.5
1.0
1.5
0.0
0.2
0.4
0.6
0.8
1.0
hflk$pvalue

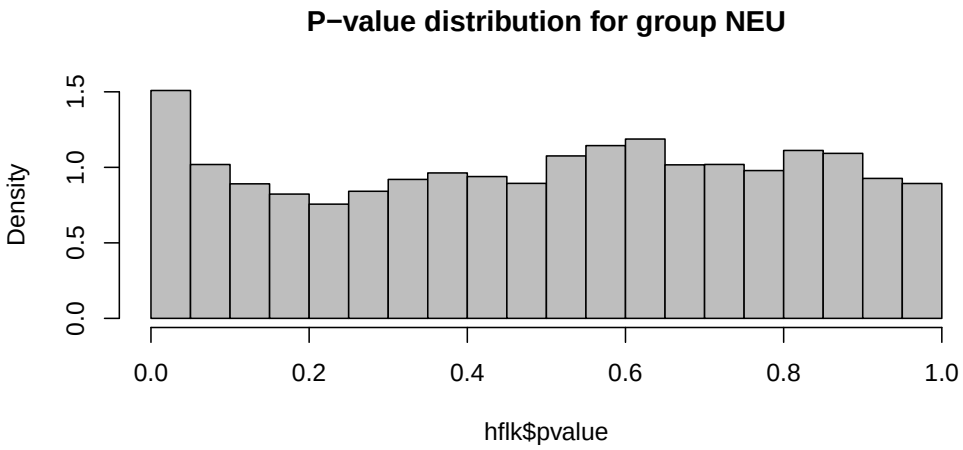

FLK genome scan for group ITA
−log10(p)
0
2
4
6
8

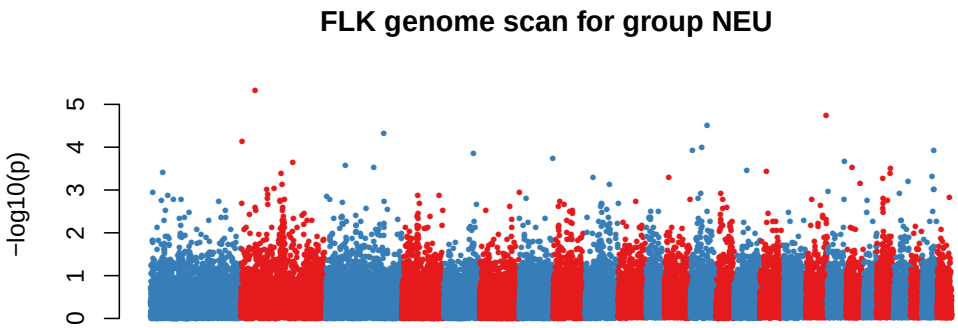

hapFLK genome scan for group ITA
−log10(p)
0
5
10
15

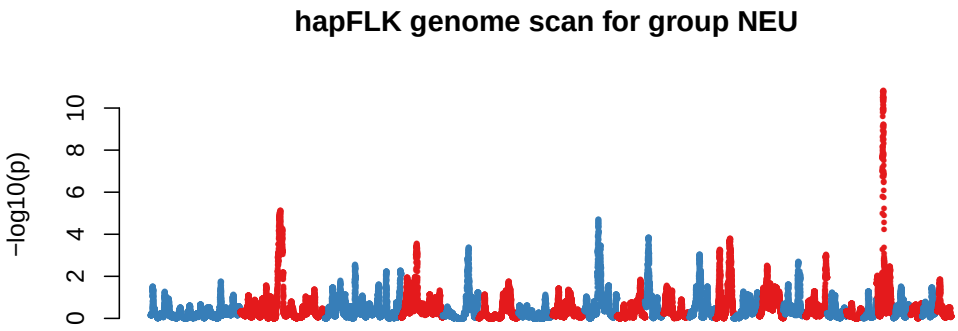

FLK distribution for group NEU
Density
0.00
0.05
0.10
0.15
0
5
10
15
20
25
30
flk$flk

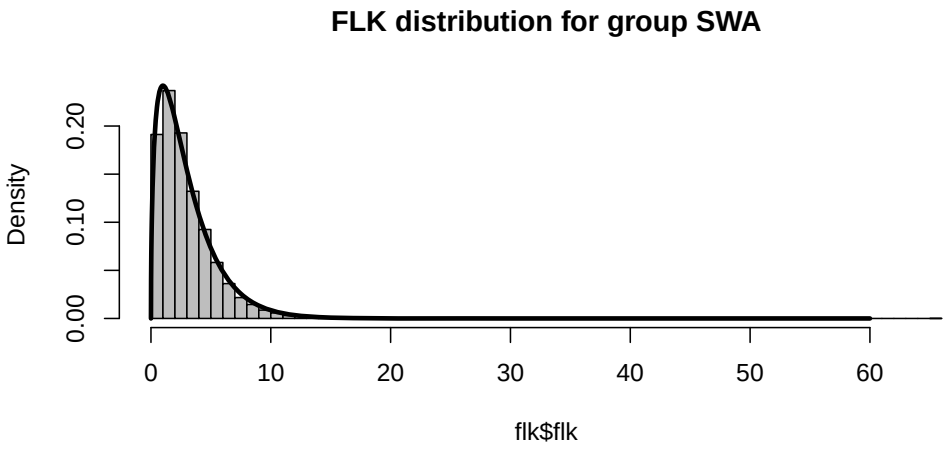

P−value distribution for group NEU
Density
0.0
0.4
0.8
0.0
0.2
0.4
0.6
0.8
1.0
flk$pvalue

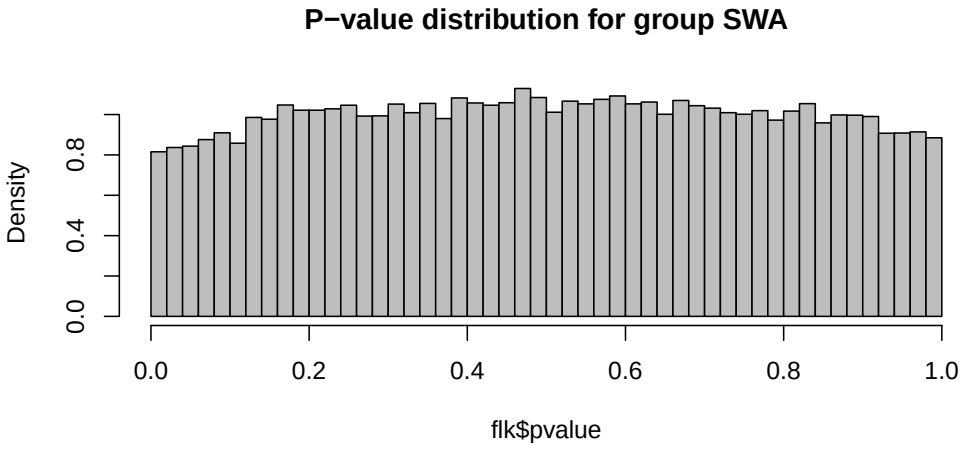

hapFLK distribution for group NEU
Density
0.0
0.4
0.8
3
4
5
6
7
hflk$hapflk

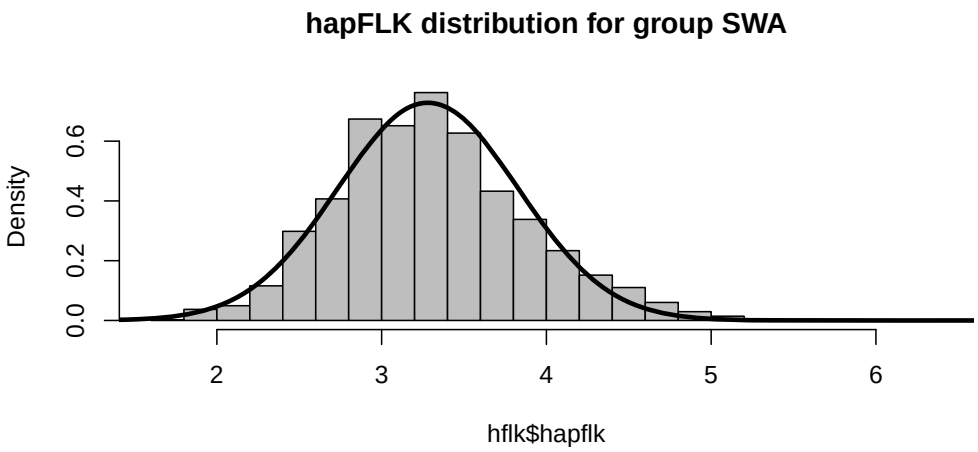

P−value distribution for group NEU
Density
0.0
0.5
1.0
1.5
0.0
0.2
0.4
0.6
0.8
1.0
hflk$pvalue

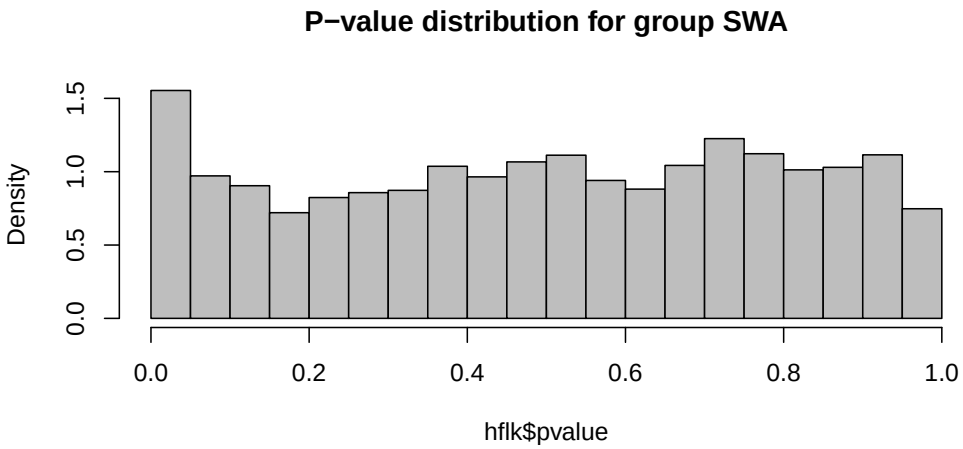

FLK genome scan for group NEU
−log10(p)
0
1
2
3
4
5

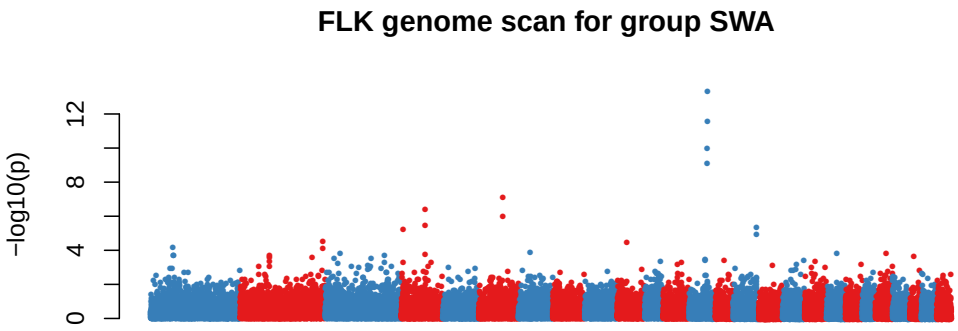

hapFLK genome scan for group NEU
−log10(p)
0
2
4
6
8
10

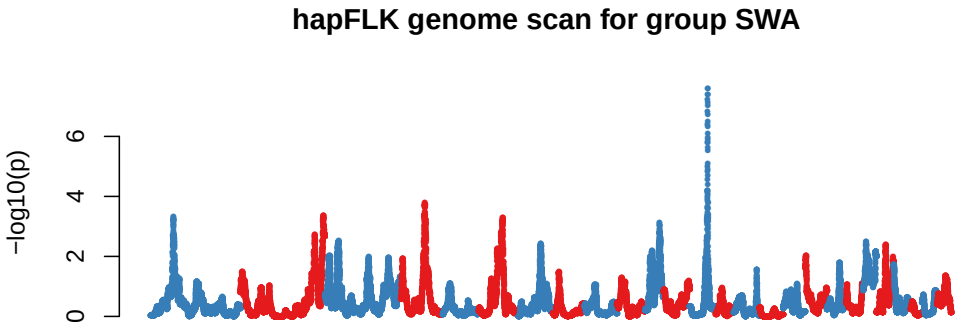

FLK distribution for group SWA
Density
0.00
0.10
0.20
0
10
20
30
40
50
60
flk$flk

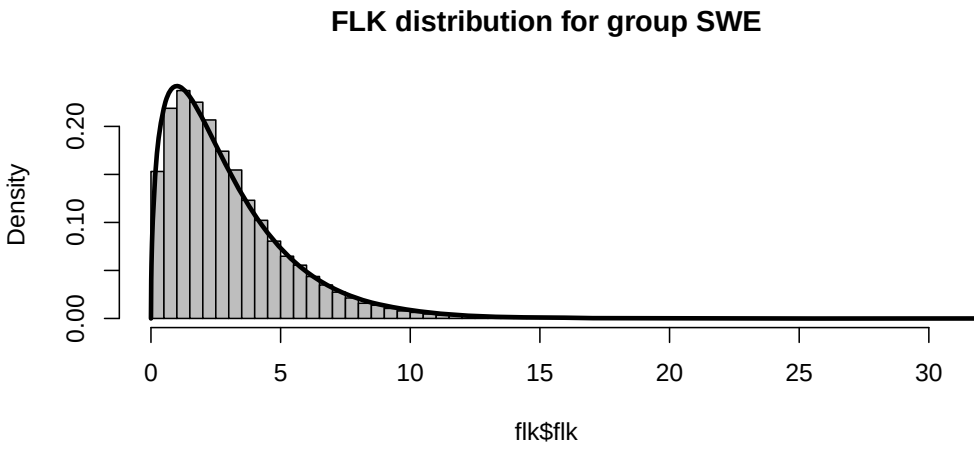

P−value distribution for group SWA
Density
0.0
0.4
0.8
0.0
0.2
0.4
0.6
0.8
1.0
flk$pvalue

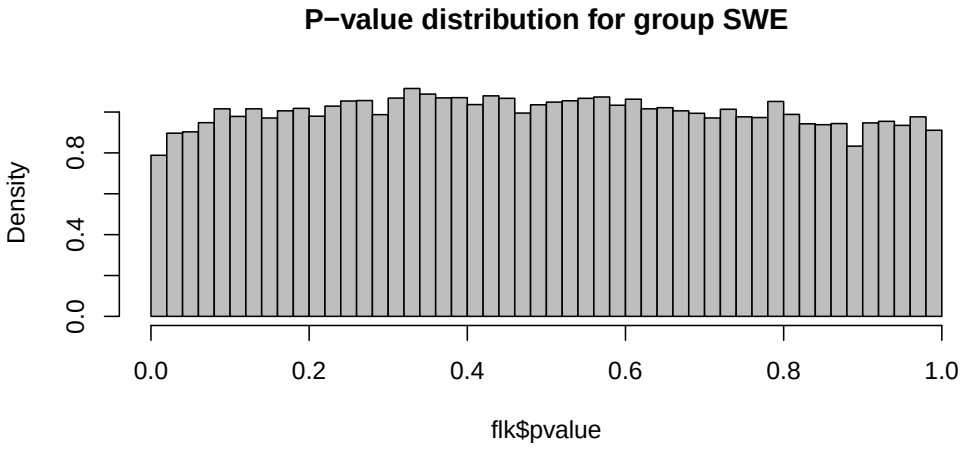

hapFLK distribution for group SWA
Density
0.0
0.2
0.4
0.6
2
3
4
5
6
hflk$hapflk

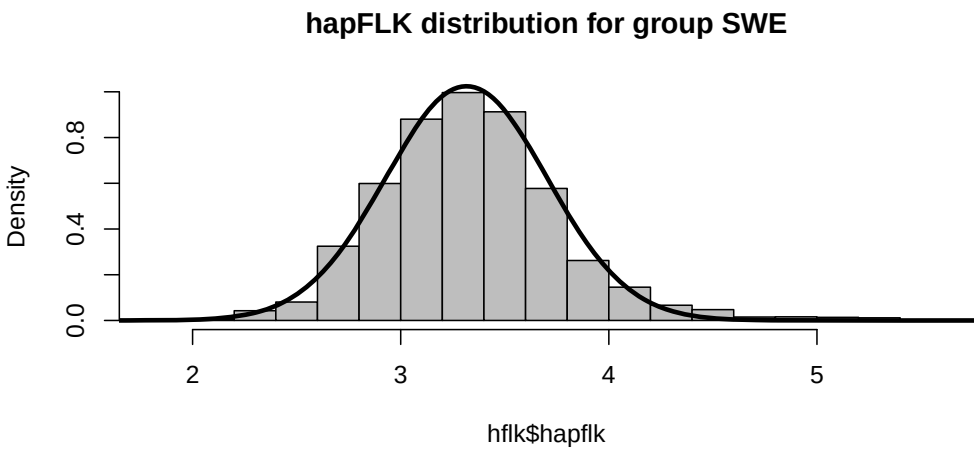

P−value distribution for group SWA
Density
0.0
0.5
1.0
1.5
0.0
0.2
0.4
0.6
0.8
1.0
hflk$pvalue

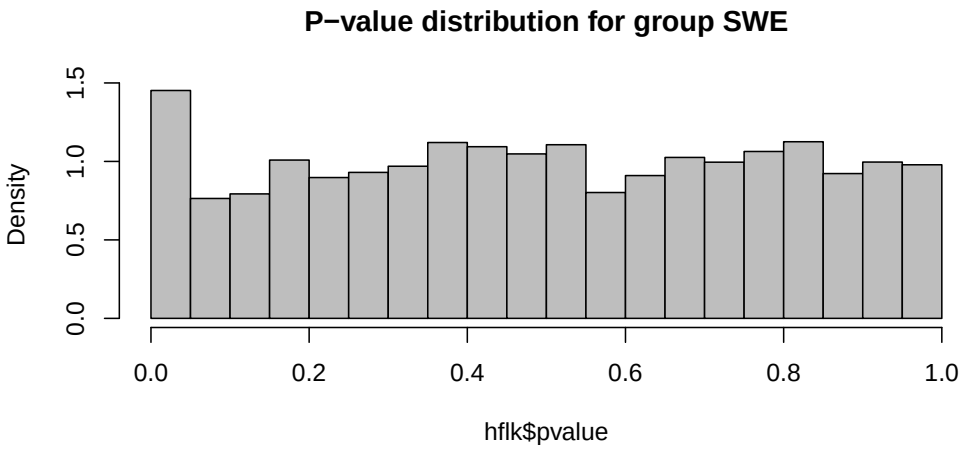

FLK genome scan for group SWA
−log10(p)
0
4
8
12

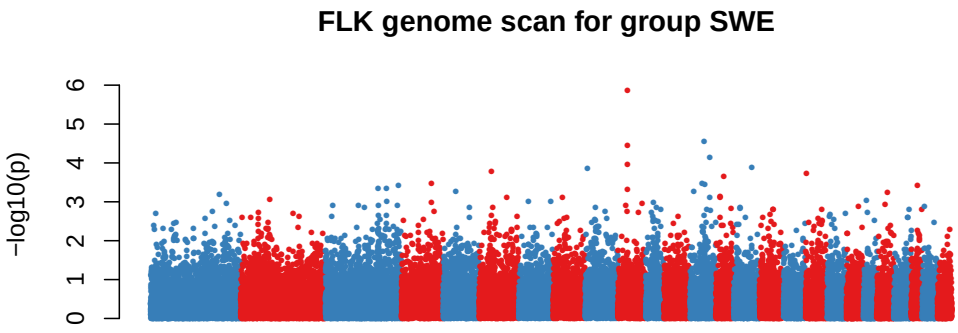

hapFLK genome scan for group SWA
−log10(p)
0
2
4
6

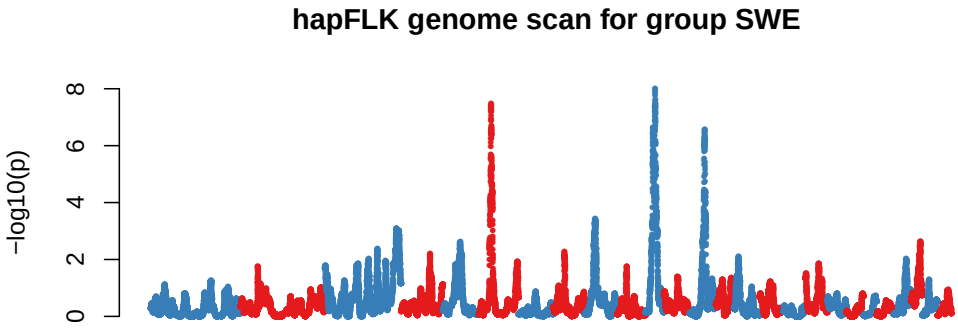

FLK distribution for group SWE
Density
0.00
0.10
0.20
0
5
10
15
20
25
30
flk$flk


P−value distribution for group SWE
Density
0.0
0.4
0.8
0.0
0.2
0.4
0.6
0.8
1.0
flk$pvalue


hapFLK distribution for group SWE
Density
0.0
0.4
0.8
2
3
4
5
hflk$hapflk


P−value distribution for group SWE
Density
0.0
0.5
1.0
1.5
0.0
0.2
0.4
0.6
0.8
1.0
hflk$pvalue


FLK genome scan for group SWE
−log10(p)
0
1
2
3
4
5
6


hapFLK genome scan for group SWE
−log10(p)
0
2
4
6
8

Chapter 6

Local Score-based test

In the previous chapters I presented and applied the hapFLK test. This method provided interesting results, but it requires individual genotypes to be performed. In this chapter, I will consider a different type of data, that is population allele frequencies obtained from the next generation sequencing of pools of individuals. This experimental design has been very popular recently, as it gives a picture of the genetic diversity of a population at a very large number of SNPs genome wide, for a much lower cost than individual sequencing.

As already discussed, single marker tests are very erratic and may result in a high rate of false positives, because high statistic values can be obtained just from genetic drift. A common solution is to compute windowed statistics by averaging single marker statistics, or to count the number of single marker statistics exceeding a given threshold, within genomic windows. This approach has several drawbacks, in particular the window size and other tunning parameters have to be chosen arbitrarily.

Other approaches, as the Bayesian approaches described in Section 3.2 or the XP-CLR mrthod (Section 3.3.1), try to model the correlation between contiguous SNPs, but they are computationally demanding and do not scale easily to dense genome wide data. In addition, XP-CLR is not suited for more than two populations, and the existing Bayesian approaches do not account for hierarchical population structure. Thus, there is a need for multi-

population tests that account for allele frequency correlations between both contiguous markers and related populations.

Given a single marker test, I propose here to find selected regions by cumulating the effects of extreme p-values of the statistic. The principal idea is that sometimes series of low single p-values are observed in a region. Individually, these statistics are not in the most extreme part of the distribution (for example between 10^{-1} and 10^{-3}) so the markers would not be considered under selection, but the fact that these statistics cluster in the same region points out this region as a potential selection target.

I tested this local score approach on the Sheep HapMap dataset and on the lactase region of the human HapMap dataset, where it provided similar results as hapFLK. I then applied it to allele frequency data from a pair of behavior divergent quail lines. It pointed out clear candidate regions, while single marker statistics or windowed approaches could not detect any particular region.

In what follows I present all these results under the form of a manuscript in preparation

Local score based method on pool-sequenced behaviour-divergent quail lines precisely detected selection signatures related to autism.

Fariello María Inés[1,2,3],
and (alphabetic order):
Arnould Cécile[6], Boitard Simon[7,8,9], Dehais Patrice[10],
Faraut Thomas[1], Lebihan Elisabeth[6], Leterrier Christine[6],
Mercier Sabine[5], Pitel Frédérique[1], Recoquillay Julien[6],
Salin Gérald[1,11], SanCristobal Magali[1,4]

[1] Laboratoire de Génétique Cellulaire, INRA, Toulouse, France
[2] Unidad de Bioinformática, Institut Pasteur, Montevideo, Uruguay
[3] Facultad de Ingeniería, Universidad de la República, Montevideo, Uruguay
[4] Département Génie Mathématiques et Modélisation, INSA Toulouse, France
[5] Université Toulouse Le Mirail
[6] Tours
[7] INRA, UMR1313 Génétique animale et biologie intégrative, 78352 Jouy-en-Josas, France
[8] AgroParisTech, UMR1313 Génétique animale et biologie intégrative, 75231 Paris 05, France
[9] UMR 7205 Origine, Structure et Evolution de la Biodiversité (MNHN / CNRS / EPHE / UPMC), Muséum National d'Histoire Naturelle, 75005 Paris 05, France
[10] SIGENAE, INRA Toulouse, France
[11] INRA - GeT-PlaGe Genotoul, F-31326 Castanet-Tolosan, France
[12] ENVT, UMR444 Laboratoire de Génétique Cellulaire, F-31076 Toulouse, France

Running title: Footprints of selection in quail
Key-words: Footprint of selection, local score, quail, NGS, pool sequencing

1 Abstract

Detecting genomic footprints of selection might help deciphering traits that underwent selection in some populations. Accounting for haplotype information in genome scans for selection allows to increase the detection power, but haplotype-based methods require individual genotypes and are not applicable when only allele frequencies are available. We propose here to take advantage of the local score approach to cumulate (possibly small) signals from single markers over a genomic segment, to clearly pinpoint a selection signal. This method gave results similar to haplotype-based methods on two benchmark data sets with individual genotypes. Results obtained for a divergent selection experiment on behavior in quail, where two lines were sequenced in pools, are precise and biologically coherent. This local score approach is general and can be applied to other genome-wide analyses such as GWAS or genome scans for selection.

2 Introduction

Detecting genomic regions that have evolved under selection has received much interest these last years. In this context, a common hypothesis is that selection targets only one gene. This gene generally includes several polymorphic markers, but the proportion of these markers that are genotyped depends on the analyzed data set. Linkage disequilibrium (LD) leads to the persistence of a footprint of selection around positively selected mutations, so the selection signature is not limited to a single causal mutation, but generally extends to a wider genomic interval including this mutation. Consequently, detection power is expected to increase when searching for such intervals rather than considering markers independently of each others. Indeed, markers in the neighborhood of the selection target also show a departure from the neutral evolution null hypothesis [1]. Several detection methods taking advantage of LD in a single population [2, 3, 4, 5] or in a pair of populations [6, 7, 8] have been proposed in the literature. Recently, Fariello et al. (2013) [9] developed the hapFLK approach for detecting locus showing extreme haplotypic differentiation between populations. This approach was proved powerful in most situations, and robust to departures from population demography. In simulation scenarios with one single site under positive selection, it outperformed single markers tests such as F_{ST} [10] or $\mathcal{F}$-LK [11], as well as their windowed versions [12], except for sequencing data where all polymorphic sites (including the one under selection) were observed.

Haplotypic methods [2, 3, 6, 9] require individual genotype data (for inferring haplotypes) and are rather computationally demanding. On the other hand, single marker statistics have a lot of variability [12] and high values of these statistics can be reached just by chance. This variability is due to genetic drift : when populations have evolved with a lot of genetic drift, the probability of false positives when considering single markers tests is particularly high. But, as mentioned above, selection should result in high values of the chosen statistic at several close markers, not only at a single one. Here we will focus on methods that try to detect regions of elevated differentiation between populations, which can be measured for instance using the statistics F_{ST} or $\mathcal{F}$-LK. In order to cumulate F_{ST} signals from single markers and to smooth the profile of the statistic genome wide, Weir et al. (2005) [12] proposed to average single marker F_{ST} within a sliding window along the genome.

Sliding window approaches do not require individual genotypes and run faster than haplotype-based methods. However, they are also less powerful and imply to choose a window size, which is usually done arbitrarily. To overcome this problem alternative approaches to find clusters of high F_{ST} values were proposed by Myles et al. (2008) [13] and Johansson et al. (2010) [14]. Myles et al. (2008) [13] proposed an algorithm to find clusters of markers with F_{ST} values in the top 1% of the genome-wide distribution (they called them the $T1$ SNPs). For each SNP $t_i \in T1$, they counted the number of non$-T1$ SNPs located between t_i and t_{i+9}. The resulting number was defined as the clustering coefficient K, with low values corresponding to regions of high density of $T1$ SNPs. Johansson et al. (2010) [14] proposed a different algorithm in the same spirit. They considered that two SNPs are in the same cluster if the distance between them is shorter than 1Mb. Here, instead of taking the top 1% of a distribution, they considered the SNPs that were fixed in both populations for different alleles. Although the two above approaches avoid to define fixed windows, they involve tunning parameters whose value is generally fixed arbitrarily. In [13] windows with 5, 10, 15 and 20 $T1$ SNPs were considered, and in [14] regions with 2 or 5 SNPs in a Mb were detected, but the authors acknowledged that this parameter choice was subjective.

The objective of this study is to present a new strategy for detecting footprints of selection, which runs very fast and is suited to the case where only allele frequencies are available. This strategy aims at cumulating selection signals via small p-values of single-marker tests, (or equivalently large values of $-log10(p-value)$) in an automatic manner, using the statistical theory of local scores. The local score approach plays an important role in bioinformatics, where it is used for computing sequence alignment scores. The null distribution of the local score can be computed exactly for small sequences, assuming a Markovian model. For very long sequences, this exact computation is computationally too demanding, but approximations have been proposed under the assumption of independence between markers [15, 16], and improved by Mercier et al (2003) [17]. All these results are valid for integer scores. Here we use a real score function related to $-log10(p-value)$, as was done by Guedj et al (2006) [18] in epidemiology or Teyssèdre et al (2012) [19] for a GWAS study, but we show that quantiles of the local score under the null hypothesis can easily be obtained by simulations.

We demonstrate that this novel approach performs well by comparing it with an haplotype based approach on 2 benchmark data sets. We also apply it to detect genes associated to social reinstatement behavior in quail, using unpublished pooled NGS data from two quail lines that have been divergently selected for this trait.

3 Material and Methods

3.1 Data

HapMap data We tested a 4Mb region (134-138 Mb) on Human chromosome 2 containing the LCT gene, because a known causal mutation for the lactase persistent phenotype in the CEU population is located in chromosome 2 at position 136,325,116. Data was taken from the HapMap Phase III dataset and consisted in the genotypes of 370 founder individuals from the CEU, TSI, CHB and JPT populations. Only 25% of the available SNPs (that is 497 SNPs) were included in the analysis.

Sheep HapMap data The Sheep HapMap dataset includes individual genotypes at 60K SNPs for 2819 animals from 74 worldwide sheep breeds [20]. Kijas et al. (2012) [20] scanned this dataset for selection using simple marker F_{ST} . We considered two subsets of this dataset : a group of 6 breeds originating from Northern Europe (278 animals), and a group of 4 breeds (two Spanish breeds and two French breeds) originating from South-western Europe (256 sheep). Genome scans for selection using the single marker test $\mathcal{F}$-LK and the haplotypic test hapFLK have already been performed in these two groups, by Fariello et al. (2013) [9] and Fariello et al. (in preparation) respectively. Interestingly, in the South-western European group, genome scans with $\mathcal{F}$-LK and hapFLK lead to distinct detected regions.

Quail data Two divergent lines produced and maintained at the INRA experimental unit 1295 (UE PEAT, F-37380 Nouzilly, France) were used in the experiment. These lines with high social reinstatement behavior (HSR) and low social reinstatement behaviour (LSR) have been divergently selected on their propensity to rejoin a group of conspecifics when 10-day old [21]. They differed consistently on several aspects of their social behavior [22, 23] and also notably on the characteristics of the social bond they developed [24, 25].

A total of 10 individuals from generation 50 of each quail line were used: 3 males and 7 females, chosen as unrelated as possible. Genomic DNA was obtained from blood samples of these 20 animals through a high-salt extraction method [26]. Sequencing was performed on 1 DNA pool per line, consisting of an equimolar mix of the ten samples. Two libraries, one for each pool, with an insert size of 300 bp, were prepared following Illumina instructions for genomic DNA sequencing (TruSeq DNA sample v2). Samples were then sequenced (paired-ends, 100 bp) on a HiSeq 2000 sequencer (Illumina), by using one lane per line (TruSeq SBS kit v3).

In the absence of an available genome sequence for the quail, the reads of the two divergent lines (190,159,084 and 230,805,732 reads respectively) were mapped to the chicken genome assembly (GallusWU2.58). To achieve good sensitivity, the reads were aligned using the glint aligner (http://lipm-bioinfo.toulouse.inra.fr/download/) with default parameters. The glint program, a general purpose nucleic sequence aligner, was designed specifically to align medium-divergent sequences, characteristic of interspecific genome comparison. 54.6% and 55.4% of the reads were aligned in a proper pair to the chicken genome (mapping quality of at least 20) for the S5 and S6 lines respectively, corresponding to 8 and 10X genome coverage. In contrast, the bwa aligner [27] was only able to align 10.2% and 10.1% of the reads respectively (less than 2X coverage). The alignments were first converted into the pileup format using the mpileup command of samtools with options -B, -q 20 and -f. Within each line, the frequency of the reference chicken allele was estimated for all SNPs that were covered by at least 5 reads, using Pool-HMM [28] with the options –estim, –a reference and -c 5. Pool-HMM accounts for the sampling effects and the sequencing error probabilities inherent to pooled NGS experiments, when estimating allele frequencies. $\mathcal{F}$-LK values were finally computed at all SNPs for which allele frequency data had been obtained in the two lines, using private python scripts.

3.2 Motivation and the local score approach

Using computer simulations, Fariello et al. (2013) [9] showed that the haplotypic test hapFLK had a higher detection power than the equivalent single marker test $\mathcal{F}$-LK. This was also con-

firmed by the analysis of the Northern European group from the Sheep HapMap dataset (see their Figure 4). Our motivation here is to also take advantage of the additional information provided by linkage disequilibrium, but in a situation where individual haplotypes can not be inferred. We thus propose to highlight segments of adjacent loci with small p-values, starting from results of single marker tests along the genome. More precisely, our strategy is to consider high values of the "score" $-log10(p-value)$ and to cumulate them over adjacent loci.

Local score and corresponding interval Assume we observe data at L consecutive positions along a sequence, and we have chosen a score function that transforms the data at position ℓ into a real number X_ℓ. We denote $X = (X_1, ...X_\ell, ...X_L)$ the resulting sequence of scores. In this study, ℓ refers to a fixed position on a reference chromosome, and the observed data at position ℓ is a vector of allele frequencies estimated in several samples at this position. We further introduce the Lindley process $h = (h_1, \ldots, h_\ell, \ldots, h_L)$, with $h_\ell = max(0, h_{\ell-1} + X_\ell)$ and $h_0 = 0$. Based on these notations, the local score of X is finally defined as

$$H_L(X) = max_{\ell=1,...L} h_\ell. \quad (1)$$

This local score is associated with an interval of interest $[\ell_{start}(X), \ell_{stop}(X)]$, which is enriched in high values of X. The end of this interval corresponds to the locus where the local score is realized, i.e. $\ell_{stop}(X) = argmax_\ell h_\ell$, and the start of this interval corresponds to the last locus before $\ell_{stop}(X)$ where the Lindley process is equal to 0, i.e. $\ell_{start}(X) = max\{\ell \leq \ell_{stop}, h_\ell = 0\}$. Note that the interval of interest is unchanged if we read the sequence in the opposite direction. Indeed, denoting $\bar{X} = (X_L, ...X_\ell, ...X_1)$ (i.e. $X_i = \bar{X}_{L-i+1}$), It can be shown that $\ell_{start}(X) = \ell_{stop}(\bar{X})$ and $\ell_{stop}(X) = \ell_{start}(\bar{X})$. An example of a Local Score, the interval and the local maximum H_L is shown in Figure 1.

Distribution of the local score under the null hypothesis For independent sites (i.i.d. model), it is possible to obtain approximate p-values by discretizing the score function, taking advantages of known theoretical results [15, 16, 17] Details are given in Supporting information 1 and 2.

For a non zero correlation between sites, we propose to obtain empirical p-values for the local score H_L on a chromosome of length L. For quail data for instance, correlation of score function values of adjacent sites ranged from 0.81 to 0.91 for the 28 quail chromosomes, with a mean of 0.85. The maximum value 0.91 was reached for the "shortest" chromosome, GGA 16, which encountered alignment problems, so was not further considered. Simulations (10,000 runs) were performed to obtain the empirical distribution under the null hypothesis of neutrality of the local score H_L, with a correlation of 0.85 for the score function between adjacent sites. More precisely, each run consisted in randomly drawing L correlated values from a uniform distribution (what is expected for p-values of single marker tests under the null hypothesis), then computing the Lindley process and taking the maximum value of it as the local score. This was done for the shortest chromosome (GGA 28, 40,000 bp long), since the computational time was prohibitive for the largest one (GGA1, 2,631,000 bp long). We chose 48 as a proxy for a chromosome wise threshold of the local score H_{40000}. This value corresponded to the 0.1% threshold of GGA 28, so using Bonferroni correction it provides a conservative 5% threshold even for $H_{2631000}$ (GGA1).

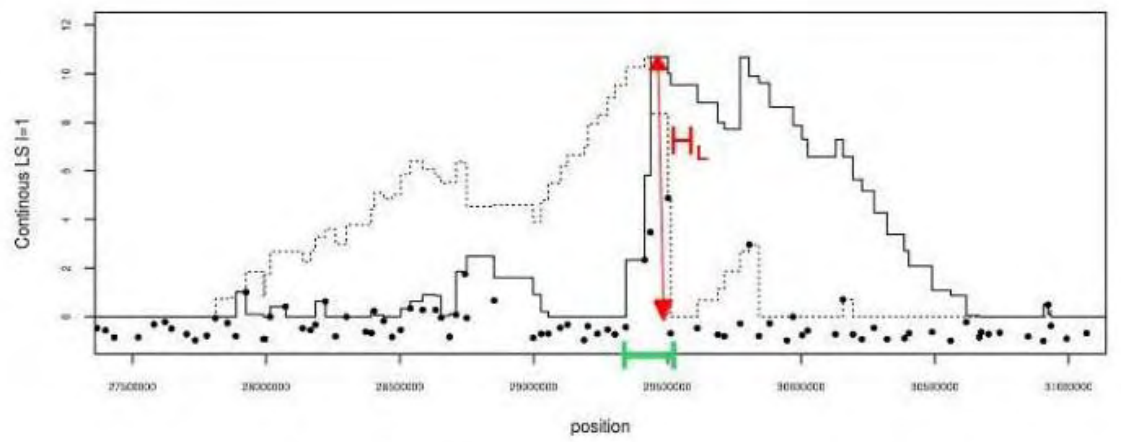


Figure 1. Example of local score on a chromosomal segment in sheep: region related to the presence / absence of horns. Single marker significance is displayed by black points, representing $-log10(pvalue_{\mathcal{F}-LK})$ for a $\mathcal{F}$-LK test of neutrality. Three consecutive SNPs with high scores have been associated to the presence/absence of horns [29]. The Lindley process of score function $-log10(pvalue_{\mathcal{F}-LK}) - 1$ is drawn with a solid line going from the left to the right of the chromosome, and with a dashed line from right to left. The local score H_L is achieved at the red arrow, and the corresponding segment, materialized in green, contains exactly the 3 SNPs.

4 Results

Lactase region of HapMap data We compared results obtained from the single locus approach $\mathcal{F}$-LK, the local score and the haplotypic approach hapLK, when analyzing Human data from the Lactase region (Figure 2. The top markers for $\mathcal{F}$-LK were quite far from the Lactase gene. The intervals given by hapLK and the local score were close to the Lactase gene, but the local score provided a smaller interval. Here the local score was based on the continuous score function $-log10(pvalue_{\mathcal{F}-LK}) - 1$, the p-value being the one of the $\mathcal{F}$-LK test. This means that p-values of single marker tests that were greater than $10^{-1} = 0.1$ were cumulated to find an interval achieving the local score. On this benchmark region, the local score clearly highlighted a well known target of selection in Human, thus performing as well as the hapFLK test.

SheepHapMap data SheepHapMap data first analyzed by Kijas et al (2011) [20] and re-analyzed with hapFLK by Fariello et al (2013) [9] and Fariello et al (in preparation), were taken as a second benchmark. The genome wide scans using various local scores are given in Figure 3 for breeds of Northern Europe, and in Figure S5 for breeds of South-western Europe. In both cases, the approaches taking account of the dependence between adjacent markers gave a clearer picture of selection signatures.

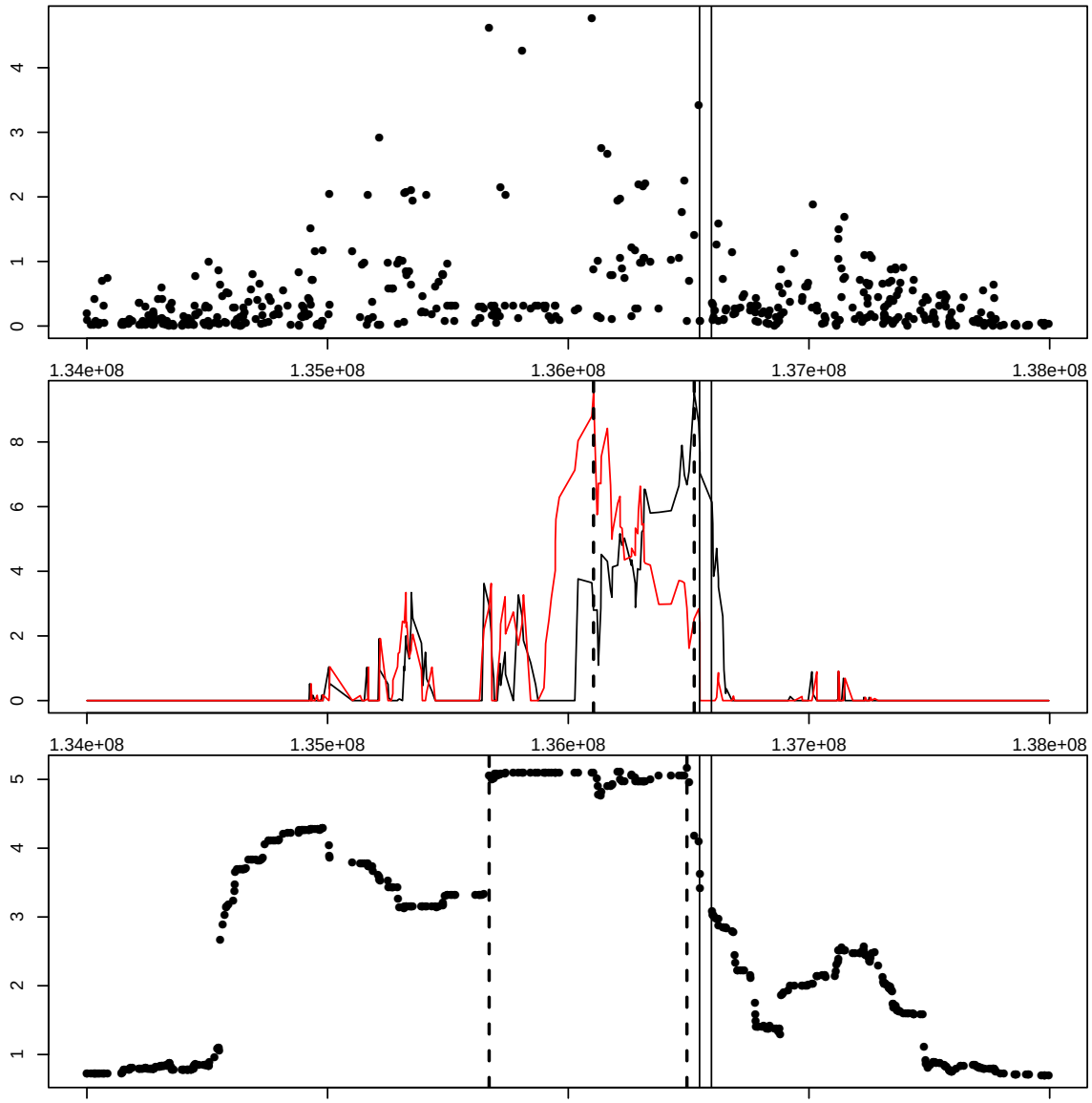


Figure 2. Selection footprints for HapMap data: focus on Lactase region. The lactase gene is located within the 2 vertical solid lines (in the 3 plots). The top graph displays $-log10(pvalue_{\mathcal{F}-LK})$ of the $\mathcal{F}$-LK test. The Lindley process based on the score function $-log10(pvalue_{\mathcal{F}-LK}) - 1$ is plotted in the middle graph, starting from the left (black curve) and from the right (red curve). The detected interval (achieving the local score) ranges between the dotted vertical lines. The bottom graph shows the hapFLK values. The detected interval ranges between the dotted vertical lines.

For the South-western Europe group, the local scores highlighted a region on OAR10 that was significant for $\mathcal{F}$-LK but not for hapFLK (Figure 1). In this case, no particular haplotype was selected, but some SNPs on several haplotypic backgrounds were. This corresponded to polymorphisms linked with the presence/absence of horns among breeds. On the other hand, hapFLK detected a region on OAR11, that the local scores failed to detect. This shows that the use of multipoint approaches based either on haplotype or allele frequency data, can be complementary.

The CPU time for a whole genome scan with hapFLK was about a few hours, while it took only a few minutes to compute $\mathcal{F}$-LK and a few more minutes to compute the local score.

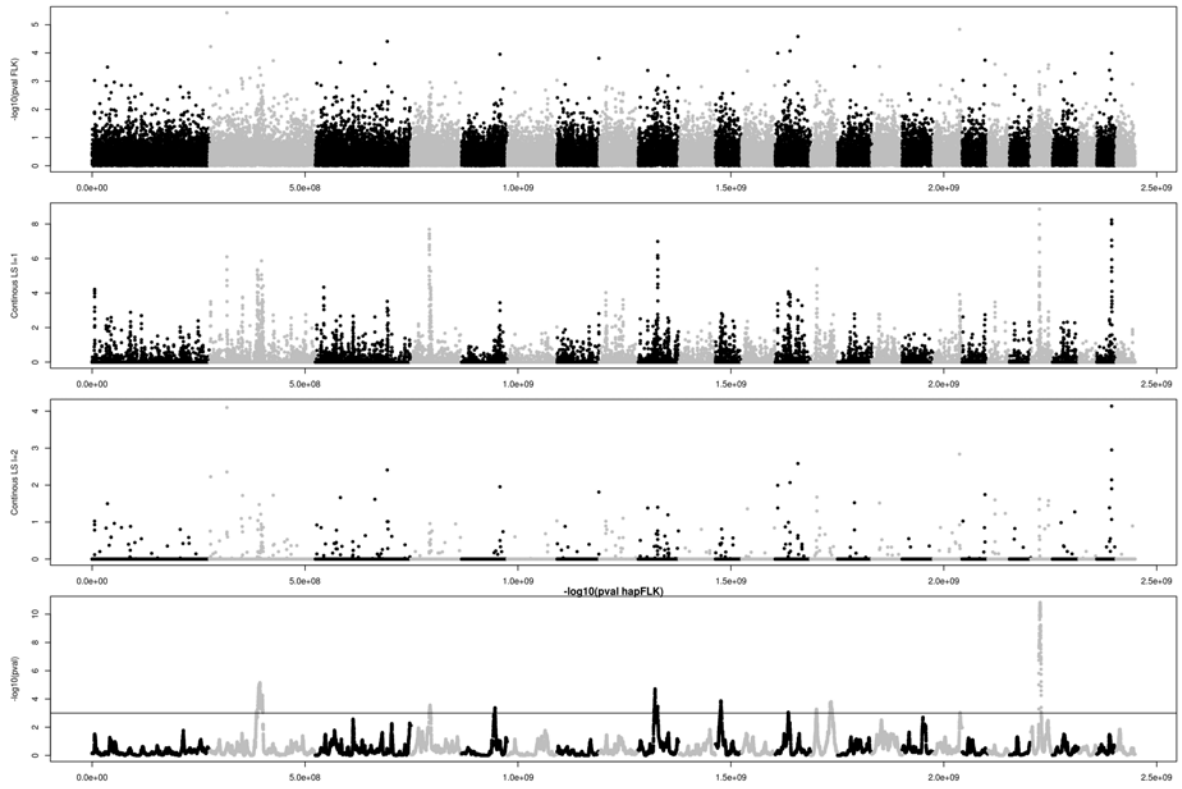


Figure 3. Genome wide selection footprints for SheepHapMap data, breeds from Northern Europe. A change in color represents a change in chromosome. From top to bottom: $-log10(pvalue_{\mathcal{F}-LK})$ from the $\mathcal{F}$-LK test, local score (Lindley process) based on the $-log10(pvalue_{\mathcal{F}-LK}) - 1$ score function, on the $-log10(pvalue_{\mathcal{F}-LK}) - 2$ score function, and significance results from the hapFLK test ($-log10(pvalue)$).

Quail data We looked for selection signatures genome wide in two divergent Quail lines using the local score associated to score function $-log10(pvalue_{F_{ST}}) - 1$, as well as several sliding windows statistics: F_{ST} (equivalent to $\mathcal{F}$-LK here), proportion of "significant" SNPs, i.e. such

that $-log10(pvalue_{F_{ST}})$ is greater than 1, and proportion of fixed SNPs. We call a fixed SNP, a SNP that has reached fixation in at least one line, but still displays differences in allele frequencies between lines. Windows of 10kb were considered. These windows included 132 SNPs on average. To estimate the statistics described above, windows with less than 32 SNPs were discarded (660 out of 19,842).

The local score (maximum value of the Lindley process) was in the range 15-50 for almost all chromosomes, except chromosomes GGA1 (with a local score equal to 250), GGA2 (local score 150), GGA3 (local score 150) and GGA6 (local score 250). A very clear peak was observed in these cases, in contrast with windowed F_{ST} (Figures S6 and S7).

We took a deeply look on the detected zone on chromosome 1. The detected region on chromosome 1 was 219.23 kb long, and contained 2646 SNPs. The SNP density in this region was 74.8 SNPs per 10kb, less than the whole genome average, so we can dismiss the possibility of having a larger local score because of a higher density of markers. In addition, we found that there were approximately 23.8% SNPs that were fixed for alternative alleles, while it was just about 1.5% chromosome wide. We further look to on the amount of SNPs fixed on each line. While in the S^- line there was almost no difference between the detected region and the whole chromosome 0.30% and 0.27%, the S^+ line had a little bit more than 2-fold the chromosome wide value of alleles fixed (0.86% against 0.38$). This means that the social motivated line is probably the one on which selection left its signature. The region displayed a slightly decreased $\mathcal{F}$-LK pvalue, and even ther was a larger proportion of "significant" SNPs or fixed SNPs per window (Figure 4), this did not appear very clear on the graphs, in contrast with the local score signal.

The detected regions, mapped on the chicken genomic sequence, are listed on Table 1.

	GGA	Region	Gene name	Description
1	1	92,963,200-93,182,431	NSUN3	putative methyltransferase NSUN3
			ARL13B	ADP-ribosylation factor-like protein 13B
2	2	1,583,808-1,688,282	VIPR1	vasoactive intestinal polypeptide receptor 1 precursor
3	3	61,585,151-61,604,462	ECHDC1	ethylmalonyl-CoA decarboxylase
			RNF146	ring finger protein 146
4	3	75,088,236-75,170,475	MMS22L	protein MMS22-like
5	4	11,412,108-11,452,505	GLOD5	Glyoxalase domain-containing protein 5
6	4	90,952,530-91,008,189	CTNNA2	catenin (cadherin-associated protein) α-2
7	6	35,234,541-35,336,717	FOXI2	forkhead box protein I3
			PTPRE	receptor-type tyrosine-protein phosphatase ϵ precursor
8	6	6,311,686-6,644,350	UBE2D1	ubiquitin-conjugating enzyme E2 D1
			CISD1	CDGSH iron sulfur domain-containing protein 1
			IPMK	inositol polyphosphate multikinase

Table 1. Footprints of selection in the quail experiment including two divergent lines for social behavior, and the genes that are comprised in each detected region.

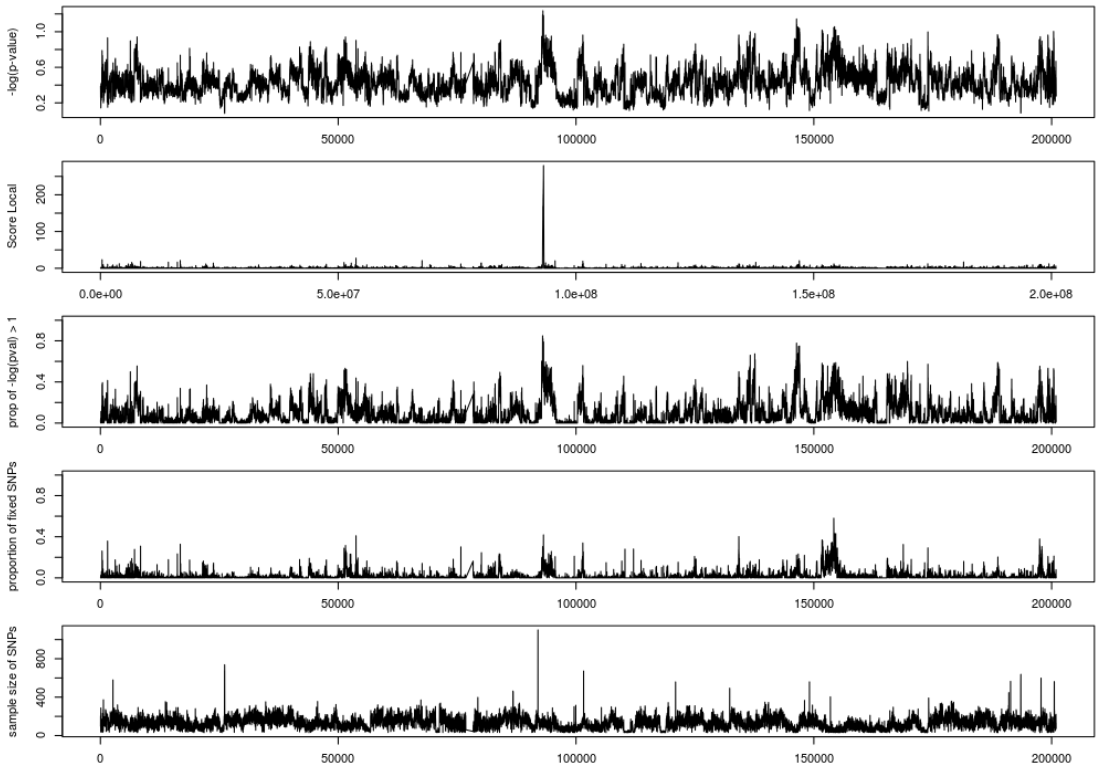


Figure 4. Selection footprints for Quail data on GGA1. From top to bottom: $-log10(pvalue)$ for windowed F_{ST} (sliding window of 10kb), local score (Lindley process) of score function $-log10(pvalue_{F_{ST}}) - 1$, proportion of SNPs with a F_{ST} p-value greater than 0.1 (sliding window of 10kb, with at least 32 SNPS in the window), proportion of fixed SNPs in each window, and number of SNPs in the window.

5 Discussion

The local score is a widely used method in sequence analysis, for example sequence alignment scores arecomputed from this method. Thus it should be natural to use it as a general tool for genome-wide statistics. In this paper, it was used to cumulate information of single marker significance over the genome, to detect an interval of highest cumulated significance. With this strategy, only a significance or importance value (the score) is required for each marker. We proposed to build the score from the single marker p-value. This implies that the local score distribution only depends on the distribution of single marker test p-values, which under the null hypothesis can be assumed to be uniform. Other studies, like Genome Wide Association Studies (GWAS), could also gain in applying this local score strategy, in particular when p-values are not extreme and when marker density is high [14]. In genome scans for selection, the local score approach can be applied to data with or without individual genotypic information. This approach gave a clear picture of the selection targets on 2 benchmark data sets (HapMap and SheepHapMap) including individual genotypes obtained from SNP chips, and on quail data obtained by the NGS of population pools.

Single marker statistics show an erratic picture along the genome, making clearer the need of smoothing methods [12]. Although we accounted for uncertainty using the pool-HMM approach to estimate the allele counts [28], pooling samples increases the risk of errors in data. Therefore, there might be a mis-estimation in allele frequencies with the technology of pool sequencing, which could contribute even more to the erratic profile of single marker tests. The length of sliding windows is still an open problem in practice. One of the problems is that as linkage disequilibrium changes all over the genome the size of the windows should change accordingly. In this work, several attempts were made to highlight outlier regions from sliding window statistics, but no really clear outliers were extracted from these approaches.

The quail experiment perfectly illustrates the difficulty to distinguish selective processes from neutral processes, and the importance to cumulate signals from multiple markers to overcome this difficulty. Indeed, a lot of drift has been accumulated in the two quail lines, because only 60 individuals were kept at each generation, so that markers with high F_{ST} value are very likely even under the null hypothesis of neutrality. For instance, 1.5% of the markers on GGA 1 were fixed differentially in the two lines and thus achieved the maximal F_{ST} value. A large proportion of these differential fixations might be due to drift. Consequently, considering all markers in the top 1% of the F_{ST} distribution as selection candidates, which is a common practice in single marker based genome scans, would result here in a large proportion of false positives.

Fortunately, as selection does not only pull up the frequency of the selected allele, but also of other alleles in its neighborhood, we expect to find clusters of markers with high F_{ST} value around a selected mutation. Johansson et al. (2010) [14] exploited this property by looking for clusters of alleles that were fixed for alternative alleles in the divergent lines. However, in this type of experiments, selection generally does not act on new variants (divergence time is probably too small for new advantageous mutations to appear) but on variants that are already segregating in the founder population. Because of linkage disequilibrium, we expect that some of the alleles in the neighborhood of the selected variant will be fixed differentially in the two lines. But at many other sites we just expect an increased allele frequency difference between the two lines, without differential fixation. Consequently, alleles showing a large frequency difference

between the two lines are also informative, even though a bit less than differentially fixed ones. The local score approach tries to take advantage of this information, as alleles that are not fixed but have a high F_{ST} value also contribute to the local score. Actually, the local score proposed here can be seen as a generalization of the clustering method of Johansson et al. (2010) [14], where the markers get a positive score if they are differentially fixed, and a negative or zero score otherwise. When analyzing the quail data, we tried a windowing approach based on the number of fixed SNPs in each window. Although our definition of fixed SNPs was a bit more liberal than that of [14] (the allele had to fix in only one line), this approach was quite similar in spirit to that of [14]. No clear signal could be detected with this approach, in contrast to the local score, confirming the fact that cumulating information from all allele frequency differences is very important.

Sequencing the founder population of the two quail lines would be very helpful to decipher the selective processes that have been acting in these lines [**?**]. This would both increase our power to detect the regions under selection, and help us understanding which line has been selected in each detected region. Indeed, sequencing the ancestral population provides a precise estimation of founder allele frequencies. Without this information, founder allele frequencies are estimated by the mean of the allele frequencies in the two divergent lines. This estimation procedure is the best we can imagine in this situation, but it is clearly biased for sites under selection. Imagine for instance that an allele with initial frequency 0.1 increases in frequency in line 1 due to selection, and is lost in line 2 due to drift. The allele frequency trajectory in line 1 cleary suggests the influence of positive selection. But from the final allele frequencies (0 and 1) we will assume that founder allele frequency was 0.5,so that allele frequency trajectories in the two populations are less informative. For future similar studies, we therefore recommend to also sequence individuals from the founder population or, if this not possible, to sequence a higher number of present populations in order to better estimate ancestral allele frequencies.

Interetingly, several of the genes comprised or partially localized in the detected regions under selection in Quail, have been associated with autistic disorders (`http://genome.ucsc.edu/cgi-bin/hgTracks`). PTPRE (receptor-type tyrosine-protein phosphatase epsilon precursor, in region 7) is one of the candidate genes present on human chromosome 10q26 and has been shown to be involved in autism spectrum disorder [30]. Similarly, ARL13B (in region 1) is one of the genes involved in the Joubert syndrome [31], a psychiatric disorder with possible autistics symptoms [32]). Finally, IMPK (inositol polyphosphate multikinase, in region 8) maps to a position homologous to a region of human chromosome 10 (10q21.1) which shows a male-only signal of linkage with a social responsiveness trait [33]. This study detected social responsiveness quantitative trait loci in multiplex autism families. As the linkage region in human is quite large, more work has to be done before considering this gene as a functional candidate.

Autistic spectrum disorders are observed in a number of disorders that have very different aetiology, including fragile X Syndrome, Rett Syndrome or Foetal Anticonvulsant Syndrome. While these disorders have very different underlying etiologies, they share common qualitative behavioral abnormalities in domains particularly relevant for social behaviors such as language, communication and social interaction ([34, 35]. In line with this, a number of experiments conducted on HSR and LSR quails indicate that the selection program carried out with these lines is not limited to selection on a single response, social reinstatement, but affect more generally the ability of the quail to process social information. Differences in social motivation, but also

individual recognition have been described between LSR and HSR quail [36, 24]. Inter-individual distances are longer in LSR quail [36] and LSR young quails have decreased interest in unfamiliar birds [37] and lower isolation distress than HSR ones ([22] for review). A last interesting candidate gene is CTNNA2 (catenin alpha 2, in region 6). This gene, involved in synaptic plasticity, has been shown to be implicated in several behavioral traits, and has recently been associated with excitement-seeking, partly related to social behavior [38]. These genes may thus represent particularly relevant candidates to explain the difference wteween two quail lines that diverge on many aspects of social behaviour. Further experiments will be required to examine the possible functional link between the selected genes and the divergent phenotype observed in these lines.

6 Conclusion

This work enhanced the added value of a divergent selection experiment on a behavior trait, pool sequencing of two divergent lines, and an appropriate statistical approach. All this combined lead to the discovery of four small genomic regions exhibiting for candidate genes related to autism.

Acknowledgements

Sequencing was performed at GeT-PlaGe Genotoul platform. This work is supported by the french ANR SNP-BB grant. We thank Bertrand Servin for stimulating discussions.

References

[1] Pritchard JK, Pickrell JK, Coop G (2010) The genetics of human adaptation: hard sweeps, soft sweeps, and polygenic adaptation. Curr Biol 20: R208–215.

[2] Sabeti PC, Reich DE, Higgins JM, Levine HZP, Richter DJ, et al. (2002) Detecting recent positive selection in the human genome from haplotype structure. Nature 419: 832-837.

[3] Voight BF, Kudaravalli S, Wen X, Pritchard JK (2006) A map of recent positive selection in the human genome. PLoS Biol 4: e72.

[4] Boitard S, Schloetterer C, Futschik A (2009) Detecting selective sweeps: a new approach based on hidden markov models. Genetics 181: 1567-1578.

[5] Nielsen R, Williamson L, Kim Y, Hubisz M, Clark A, et al. (2005) Genomic scans for selective sweeps using SNP data. Genome Research 15: 1566-1575.

[6] Sabeti PC, Varilly P, Fry B, Lohmueller J, Hostetter E, et al. (2007) Genome-wide detection and characterization of positive selection in human populations. Nature 449: 913-918.

[7] Bhatia G, Patterson N, Pasaniuc B, Zaitlen N, Genovese G, et al. (2011) Genome-wide comparison of African-ancestry populations from CARe and other cohorts reveals signals of natural selection. Am J Hum Genet 89: 368–381.

[8] Chen H, Patterson N, Reich D (2010) Population differentiation as a test for selective sweeps. Genome Res 20: 393–402.

[9] Fariello MI, Boitard S, Naya H, SanCristobal M, Servin B (2013) Detecting signatures of selection through haplotype differentiation among hierarchically structured populations. Genetics 193: 929–941.

[10] Weir B, Cockerham C (1984) Estimating f-statistics for the analysis of population structure. Evolution 38: 1358-1370.

[11] Bonhomme M, Chevalet C, Servin B, Boitard S, Abdallah J, et al. (2010) Detecting selection in population trees: The Lewontin and Krakauer test extended. Genetics 186(1): 241-262.

[12] Weir BS, Cardon LR, Anderson AD, Nielsen DM, Hill WG (2005) Measures of human population structure show heterogeneity among genomic regions. Genome Res 15: 1468-76.

[13] Myles S, Tang K, Somel M, Green RE, Kelso J, et al. (2008) Identification and analysis of genomic regions with large between-population differentiation in humans. Ann Hum Genet 72: 99–110.

[14] Johansson AM, Pettersson ME, Siegel PB, Carlborg O (2010) Genome-wide effects of long-term divergent selection. PLoS Genet 6: e1001188.

[15] Karlin S, Altschul SF (1990) Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. PNAS 87: 2264–2268.

[16] Karlin S, Dembo A (1992) Limit distributions of maximal segmental score among Markov-dependent partial sums. AdAP 24: 113–140.

[17] Mercier S, Cellier D, Charlot D (2003) An improved approximation for assessing the statistical significance of molecular sequence features. Applied Probability Trust .

[18] Guedj M, Robelin D, Hoebeke M, Lamarine M, Wojcik J, et al. (2006) Detecting local high-scoring segments: a first-stage approach for genome-wide association studies. Stat Appl Genet Mol Biol 5: Article22.

[19] Teyssedre S, Dupuis MC, Guerin G, Schibler L, Denoix JM, et al. (2012) Genome-wide association studies for osteochondrosis in French Trotter horses. J Anim Sci 90: 45–53.

[20] Kijas JW, Lenstra JA, Hayes B, Boitard S, Porto Neto LR, 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: e1001258.

[21] Mills A, Faure J (1991) Divergent selection for duration of tonic immobility and social reinstatement behavior in Japanese Quail (Coturnix coturnix japonica) chicks. Journal of Comparative Psychology 105: 25-38.

[22] Jones R, Mills A (1999) Divergent selection for social reinstatement behaviour in Japanese quail: Effects on sociality and social discrimination. Poultry and Avian Biology Reviews 10: 213-223.

[23] Richard S, Arnould C, Guemene D, Leterrier C, Mignon-Grasteau S, et al. (2008) Emotional reactivity in the quail: an integrated approach to animal welfare. INRA Productions Animales 21: 71–77.

[24] Schweitzer C, Houdelier C, Lumineau S, Levy F, Arnould C (2010) Social motivation does not go hand in hand with social bonding between two familiar japanese quail chicks, coturnix japonica. Animal Behaviour 79: 571-578.

[25] Schweitzer C, Levy F, Arnould C (2011) Increasing group size decreases social bonding in young japanese quail, coturnix japonica. Animal Behaviour 81: 535-542.

[26] Roussot O, Feve K, et al (2003) AFLP linkage map of the Japanese quail Coturnix japonica. Genet Sel Evol 35: 559-572.

[27] Li H, Durbin R (2009) Fast and accurate short read alignment with burrows-wheeler transform. Bioinformatics 25: 1754-60.

[28] Boitard S, Kofler R, Francoise P, Robelin D, Schlotterer C, et al. (2013) Pool-hmm: a Python program for estimating the allele frequency spectrum and detecting selective sweeps from next generation sequencing of pooled samples. Mol Ecol Resour 13: 337–340.

[29] Johnston SE, McEwan JC, Pickering NK, Kijas JW, Beraldi D, et al. (2011) Genome-wide association mapping identifies the genetic basis of discrete and quantitative variation in sexual weaponry in a wild sheep population. Mol Ecol 20: 2555–2566.

[30] Tonk VS WG (2011) Autism spectrum disorder with microdeletion 10q26 by subtelomere fish. Pediatric Health, Medicine and Therapeutics 2: 49-53.

[31] Cantagrel V, Silhavy JL, Bielas SL, Swistun D, Marsh SE, et al. (2008) Mutations in the cilia gene ARL13B lead to the classical form of Joubert syndrome. Am J Hum Genet 83: 170–179.

[32] Doherty D (2009) Joubert syndrome: insights into brain development, cilium biology, and complex disease. Semin Pediatr Neurol 16: 143-154.

[33] Duvall JA, Lu A, Cantor RM, Todd RD, Constantino JN, et al. (2007) A quantitative trait locus analysis of social responsiveness in multiplex autism families. Am J Psychiatry 164: 656–662.

[34] Rutter M (1978) Diagnosis and definition of childhood autism. J Autism Child Schizophr 8: 139-161.

[35] Association AP (2000) Diagnostic and statistical manual of mental disorders: DSM IV. American Psychiatric Association.

[36] François N, Mills AD, et al (2000) Inter-individual distances during open-field tests in Japanese quail (Coturnix japonica) selected for high or low levels of social reinstatement behaviour. Behavioural Processes 47: 73-80.

[37] François N, Decros S, et al (2000) Effect of group disruption on social behaviour in lines of Japanese quail (Coturnix japonica) selected for high or low levels of social reinstatement behaviour. Behavioural Processes 48: 171-181.

[38] Terracciano A, Esko T, et al (2011) Meta-analysis of genome-wide association studies identifies common variants in ctnna2 associated with excitement-seeking. Transl Psychiatry 1: e49.

[39] Asmussen (1995) Advances in Queuing : Models, Methods and Problems. Jewgeni H. Dshalov, CRS Press.

Supporting Information

1. Details of the calculations of the local score p-value

The score function is observed at each locus ℓ, taking values as a natural integer in the interval $\{-v, ..., u\}$. In the following, only symmetric scores will be considered ($v = u$).

Let use the following notation:

$$p_i = P(X = +i) \text{ for } i = 0, ...u \quad (2)$$
$$q_j = p_{-j} = P(X = -j) \text{ for } j = 1, ...v. \quad (3)$$

In the following, only symmetric scores will be considered ($v = u$).

The exact distribution of the local score can be obtained under the IID model (see Daudin and Mercier, 2001, and Hassenforder and Mercier, 2007 for the Markovian case) using

$$P(H_L < a) = 1 - P_0 \Lambda^L P_a' \quad (4)$$

where $P_0 = (1, 0, \ldots, 0)$, $P_a = (0, \ldots, 0, 1)$, and Λ is the $((a+1) \times (a+1))$ matrix:

$$\Lambda = \left(\begin{array}{c|ccc|c} P(X \leq 0) & P(X = 1) & \ldots & P(X = a-1) & 1 - P(X \leq a-1) \\ \vdots & & & & \\ P(X \leq -h) & \ldots & P(X = l-h) & \ldots & 1 - P(X \leq a-h-1) \\ \vdots & & & & \\ P(X \leq 1-a) & P(X = 2-a) & \ldots & P(X = 0) & 1 - P(X \leq 0) \\ \hline 0 & 0 & \ldots & 0 & 1 \end{array} \right)$$

This method is accurate for not too long sequences and is independent of the sign of the expected score $E[X]$.

When the expectation of X is stricly negative, the distribution of the local score H_L can be approximated by the Karlin et al. formulae :

$$\lim_{L \to \infty} P(H_L \leq \frac{\log L}{\lambda} + x) = \exp(-K^* e^{-\lambda x}) \quad (5)$$

where K^* and λ depend on the model (distribution of X). We have $\lambda = 1/R$ where R is the only real root in $]0, 1[$ of the equation $E(e^{\lambda X}) = 1$ and

$$K^* = \frac{\left(1 - E(e^{\lambda S^-})\right)^2}{(e^\lambda - 1)\mu^2 E(X e^{\lambda X})}$$

with $\mu = E(S^-)/E(X)$ and S^- the first negative partial sum.

These two methods, the exact one and the approximation, are complementary ones. An improvement of this approximation was obtained by Mercier et al. (2003) [17]. But for small sequences exact method must be preferred.

Let used these two methods on a simple example.

Let us consider the following score $X^{(2)}$, function of allele frequencies through the p-value of the single-marker test $\mathcal{F}$-LK:

$$\begin{array}{lll} X=2 & \text{when } pval_{\mathcal{F}-LK} \leq 10^{-3} & \text{so } p_2 = 10^{-3}, \\ X=1 & \text{when } 10^{-3} < pval_{\mathcal{F}-LK} \leq 10^{-2} & \text{so } p_1 = 10^{-2} - 10^{-3}, \\ X=0 & \text{when } 10^{-2} < pval_{\mathcal{F}-LK} \leq 10^{-1} & \text{so } p_0 = 10^{-1} - 10^{-2}, \\ X=-1 & \text{when } 10^{-1} < pval_{\mathcal{F}-LK} \leq 2.10^{-1} & \text{so } q_1 = p_{-1} = 10^{-1}, \\ X=-2 & \text{when } 2.10^{-1} < pval_{\mathcal{F}-LK} & \text{so } q_2 = p_{-2} = 1 - 2.10^{-1}. \end{array} \tag{6}$$

Thus we have

$$p_2 = 0.001 \quad p_1 = 0.009 \quad p_0 = 0.09 \quad q_1 = 0.1 \quad q_2 = 0.8 \tag{7}$$

The expectation of X is negative. We have

$$E(X) = \sum_{i=-2,\ldots,2} iP(X=i) = -1.689$$

and so both method can be used.

Let consider the observed local score equal to 3 and two different lengths $L = 500$ and $L = 5000$. We are interested by calculating $P[H_L \geq 3]$.

First we have for the exact method $P[H_L \geq 3] = 1 - P[H_L < 3]$ so we have $a = 3$ in (8) and

$$\Lambda = \left(\begin{array}{c|cc|c} p_0+q_1+q_2 & p_1 & p_2 & 0 \\ q_1+q_2 & p_0 & p_1 & p_2 \\ q_2 & q_1 & p_0 & p_1+p_2 \\ \hline 0 & 0 & 0 & 1 \end{array}\right) = \left(\begin{array}{c|ccc} 0.99 & 0.009 & 0.001 & 0 \\ 0.9 & 0.09 & 0.009 & 0.001 \\ 0.8 & 0.1 & 0.09 & 0.01 \\ \hline 0 & 0 & 0 & 1 \end{array}\right)$$

and we can clearly verify here that the sum of each row gives 1 as expected. We also have

$$P_0 = (1,0,0,0) \quad \text{and} \quad P_a = (0,0,0,1).$$

For $n = 500$ we get using the R library expm

$$\Lambda^{500} = \left(\begin{array}{c|ccc} 0.97822 & 0.00980 & 0.001171965 & 0.01079 \\ 0.97707 & 0.00979 & 0.00117 & 0.01197 \\ 0.96737 & 0.00969 & 0.00116 & 0.02177 \\ 0.00000 & 0.00000 & 0.00000 & 1.00000 \end{array}\right)$$

and for $n = 5000$

$$\Lambda^{5000} = \left(\begin{array}{c|ccc} 0.88698 & 0.00889 & 0.00106 & 0.10306 \\ 0.88593 & 0.00888 & 0.00106 & 0.10412 \\ 0.87714 & 0.00879 & 0.00105 & 0.11301 \\ 0.00000 & 0.00000 & 0.00000 & 1.000 \end{array}\right)$$

and we can also verify that the sum of each row still gives 1 as expected. Thus we obtain

$$P[H_{500} \geq 3] = 1.08 \cdot 10^{-2} \text{ and } P[H_{5000} \geq 3] = 1.03 \cdot 10^{-1}.$$

For the asymptotic method (see Karlin and Altschul 1990, Karlin and Dembo 1992), we have in equation (9), $\lambda = 1/R$ where R is the only real root in $]0,1[$ of the equation $E[e^{\lambda X}] = 1$ and

$$K^* = \frac{\left(1 - E(e^{\lambda S^-})\right)^2}{(e^\lambda - 1)\mu^2 E(Xe^{\lambda X})}$$

with $\mu = E(S^-)/E(X)$. Resolving the equation $E(e^{\lambda X}) = 1$ and taking the variable transformation $x = e^{-\lambda}$, leads to the problem of finding the roots of the following polynom: $P(x) = \sum_{i=1,...u} p_i x^{u-i} + (p_0 - 1)x^u + \sum_{j=1,...v} q_j x^{v+j}$.

For our example, we get

$$P(x) = p_2 + p_1 x + (p_0 - 1)x^2 + q_1 x^3 + q_2 x^4.$$

The real roots of P with module lower than 1, are 1, $R = R_1 = 0.03858$, and $R_2 = -0.02854$. The other real root is $R_3 = -1.13504$. Hence $\lambda = -\log R = 3.2549$. The distribution of S^- can be established following [39], who showed that any root z of $E(z^X) = 1$ such that $|z| \leq 1$, verifies also $E(z^{S^-}) = 1$. It can also be shown that z_i is solution of $E(z_i^X) = 1$ iff $R_i = 1/z_i$ is root of polynom P. Moreover, there exists u roots such that $|z| \leq 1$, so the elements of the S^- distribution are obtained by solving a linear system of u equations $E(R_i^{-S^-}) = 1$ for $i = 1, ...u$ and u unknowns $Q_j = P(S^- = -j)$, for $j = 1, ...u$: $A.Q = 1_u$, where 1_u is the vector of dimension u composed with 1's, and A is the Vandermonde matrix

$$A = \begin{pmatrix} 1 & 1 & ... & 1 \\ R_{u+2} & R_{u+2}^2 & ... & R_{u+2}^u \\ \vdots & \vdots & & \vdots \\ R_{2u} & R_{2u}^2 & ... & R_{2u}^u \end{pmatrix}$$

and $(R_i)_{i=u+2,...u}$ are the roots of P such that $|R_i| \geq 1$. In our case, the linear system is

$$\begin{pmatrix} 1 & 1 \\ R_3 & R_3^2 \end{pmatrix} \begin{pmatrix} Q_1 \\ Q_2 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

so that $Q_1 = P(S^- = -1) = 1 + 1/R_3$ and $Q_2 = P(S^- = -2) = -1/R_3$. It follows that $E(S^-) = -Q_1 - 2Q_2 = -1 + 1/R_3$, and $\mu = E(S^-)/E(X) = 1.113691$. Hence $K^* = 0.02036$. That brings

$$\lim_{L\to\infty} P(H_L \leq \frac{\log L}{3.2549} + x) = \exp(-0.02036e^{-3.2549x}).$$

As we want to calculate $P[H_L \geq 3] = 1 - P[H_L < 3] = 1 - P[H_L \leq 2]$ we have

$$\lim_{L\to\infty} P(H_L \geq 3) = 1 - \exp(-0.02036 \times L \times e^{-3.2549\times 2}).$$

For $L = 5000$, we get

$$P[H_{5000} \geq 3] \simeq 1.41 \cdot 10^{-1}.$$

For $L = 500$, we get

$$P[H_{500} \geq 3] \simeq 1.50 \cdot 10^{-2}$$

and we can check that obviously we have $P[H_{5000} \geq 3] > P[H_{500} \geq 3]$.

For $P[H_L \geq 4]$ we have

L/p_{value}	Exact	Appx
500	$7.05 \cdot 10^{-4}$	$5.85 \cdot 10^{-4}$
5000	$7.04 \cdot 10^{-3}$	$5.83 \cdot 10^{-3}$
10000	$1.40 \cdot 10^{-2}$	$1.16 \cdot 10^{-2}$

We can observe that actually Karlin's approximation becomes more accurate for long sequences.

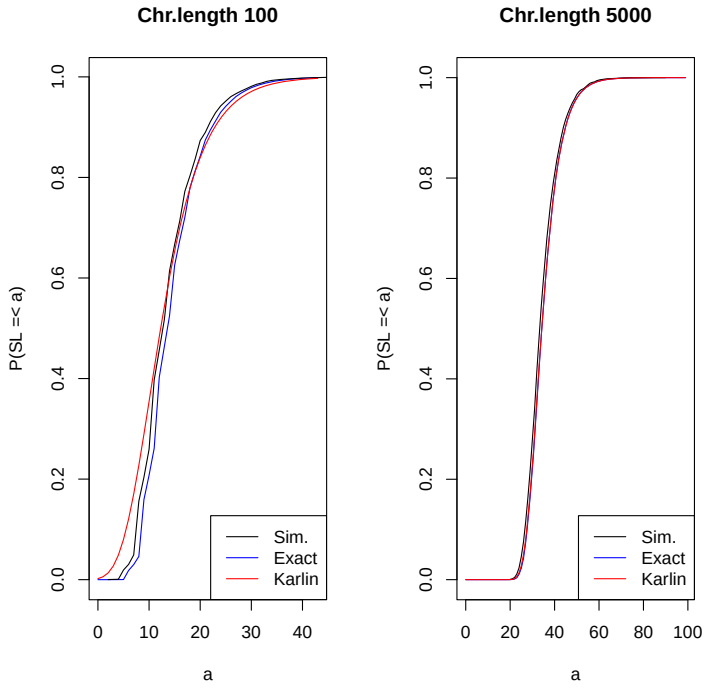


Figure S1. Cumulative density function of the local score. Comparison of simulations, exact distribution and Karlin's approximation for a segment of length $L = 100$. The score function is taking integer values in $\{-3, \ldots, 3\}$ with probabilities given in equation (13).

2. From integer scores to continuous scores

We focus here on score functions that are integer-valued and symmetric, i.e. that take integer values in $\{-u, ..., u\}$ for $u \in \mathbf{N}$.

If the scores X_ℓ are independent and identically distributed (i.i.d.), the exact distribution of the local score can be obtained by (Mercier and Daudin, 2001):

$$P(H_L < a) = 1 - P_0 \Lambda^L P'_a \tag{8}$$

where $P_0 = (1, 0, \ldots, 0)$ and $P_a = (0, \ldots, 0, 1)$ are two vectors of length $a + 1$, and Λ a squared matrix of dimension $(a + 1)$ that can be deduced from the distribution of X_ℓ. This method is accurate for not too long sequences.

For very long sequences, and still assuming i.i.d. scores, the distribution can be approximated by (Karlin et al. 1990, 1992):

$$\lim_{L\to\infty} P(H_L \leq \frac{\log L}{\lambda} + x) = \exp(-K^* e^{-\lambda x}) \tag{9}$$

where K^* and λ depend on the distribution of X. In contrast with the above exact formula, this approximation is only valid for scores with stricly negative expected value.

An improvement of this approximation was obtained by Mercier et al. (2003).

Details about the computation of the local score distribution using these two formula are provided in Appendix 1 on a small example.

From integer scores to the real score of interest

We decided to base our analysis on the continuous score $X^{(\infty)_\xi} = -log10(pval_{\mathcal{F}-LK}) - \xi$ for some constant ξ. This constant that must be stricly positive and strictly lower than the maximum value of $X^{(\infty)_0} = -log10(pval_{\mathcal{F}-LK})$, which will be denoted m. This translation is motivated by the two following reasons. First, a proper choice of ξ will ensure that the expectation of $X^{(\infty)_\xi}$ is strictly negative, so Karlin's approximation will be valid for any discrete score approaching $X^{(\infty)_\xi}$. Second, the basic idea is to cumulate low single marker p-values, so low (resp. high) p-values will be transformed into positive (resp. negative) scores. Typically, ξ must be such that $10^{-\xi}$ can be considered as a rough threshold of single marker p-values, under which the markers are under the null. We suggest to choose $\xi = 1$ or $\xi = 2$ in practice, which correspond respectively to whole genome $\mathcal{F}$-LK p-values of $10^{-1} = 0.1$ and $10^{-2} = 0.01$. For instance, single marker p-values obtained from the Sheep HapMap data (Figure 3) may invite the analyst to cumulate p-values $pval_{\mathcal{F}-LK}$ below 10^{-1}, or equivalently $-log10(pval_{\mathcal{F}-LK})$ above 1, since values below 1 seem to be only noise or background signal. To anticipate the discussion, Figure 3 shows that the segments that realize local scores are located at the same positions for $\xi = 1$ and $\xi = 2$.

The approximations of the local score distribution, which have been summarized in the preceding subsection, are only valid for integer and symmetric score functions. Hence, we consider a series of integer and symmetric score functions $X^{(1)}, X^{(2)}, ..., X^{(u)}..., $ that approximate $X^{(\infty)}$. The underlying idea is to approximate a continuous function by step functions with smaller and smaller steps. For each u, the integer score $X^{(u)}$ takes $2u + 1$ values in $\{-u, ... - 1, 0, 1, ...u\}$. To be consistent with our objective of cumulating p-values below ξ, we propose to cut the interval from $-\xi$ to 0 into u intervals of same length, and the interval from 0 to $m - \xi$

into u other intervals of same length (Figure S2). The distribution of a score $X^{(u)}$ is thus:

$$P(X^{(u)}=i) = P\left(\frac{i(m-\xi)}{u+1} \leq X_\xi^{(\infty)} < \frac{(i+1)(m-\xi)}{u+1}\right) \quad (10)$$

$$= \frac{10^{-i.(m-\xi)/(u+1)-\xi} - 10^{-(i+1)(m-\xi)/(u+1)-\xi}}{1-10^{-m}} \text{ for } i=0,...u \quad (11)$$

$$P(X^{(u)}=-i) = P\left(\frac{i\xi}{u} \leq X_\xi^{(\infty)} < \frac{(i-1)\xi)}{u}\right) \quad (12)$$

$$= \frac{10^{\xi i/u-\xi} - 10^{\xi(i-1)/u-\xi}}{1-10^{-m}} \text{ for } i=1,...u. \quad (13)$$

Denoting $\lceil\rceil$ the ceiling of a real number, an equivalent definition for $X^{(u)}$ is

$$X^{(u)} = \left\lceil \frac{u+1}{m-\xi} X_\xi^{(\infty)} \right\rceil \text{ for } X_\xi^{(\infty)} \geq 0 \quad (14)$$

$$X^{(u)} = \left\lceil \frac{u}{\xi} X_\xi^{(\infty)} + 1 \right\rceil \text{ for } X_\xi^{(\infty)} < 0 \quad (15)$$

Simulations We considered several simulation scenarios, which all involved 10,000 independent segments of $L = 100$ or $L = 5,000$ scores. To simulate each score, fictive p-values of the $\mathcal{F}$-LK test under the null hypothesis of neutrality were randomly drawn from a uniform distribution in the interval $[0, 1]$. Depending on the scenario, the p-values within a segment were drawn independently, or with an autocorrelation between adjacent sites equal to 0.5 or 0.9. The maximum value $m - \xi$ of the continuous score $X_\xi^{(\infty)}$ was computed. Consecutive integer scores $X^{(u)}$ were calculated for u in $1, ...50$ as described in the previous section. The local scores associated to the continuous and the dscrete scores were computed for each segment. The empirical cumulative density function of each local score was evaluated based on the 10,000 replictaes. in cutting the interval $[-\xi, 0]$ into intervals of length $1/u$, and the interval $[0, m - \xi]$ into intervals of length $1/(u + 1)$. Negative values of $X_\xi^{(\infty)}$ were transformed into negative integer values $-u, ..., -1$: $X^{(u)} = -i$ when $X_\xi^{(\infty)}$ falls into the i^{th} interval of $[-\xi, 0]$. Similarly, $X^{(u)} = i$ when $X_\xi^{(\infty)}$ falls into the i^{th} interval of $[0, m - \xi]$. The empirical cumulative density functions of the local scores were compared, and a plot is given in Figure S3 for $u = 3$ and $u = 20$, for "standardized" local scores H: $H(X^{(\infty)})/(m-\xi)$ and $H(X^{(u)})/u$. The distributions of the local scores $H(X_\xi^{(\infty)})/(m - \xi)$ and $H(X^{(u)})/u$ for large u appeared to be the same. This was to be expected with regard to equation 14 for positive scores, since segments that realize local score are enriched in positive scores. This implies that p-values of local scores based on the continuous score $X_\xi^{(\infty)}$ can be approximated with the threshold values of an integer score $X^{(u)}$ for a large u (i.e. 20), either with Karlin's formula for large chromosome length (of the order of 5,000 in practice), or with the exact formulae for smaller chromosomes.

Similar results were obtained for correlated loci (autocorrelation between adjacent sites equal to 0.5 or 0.9): the distribution of the local score $H(X_\xi^{(\infty)})/(m - \xi)$ was well approximated by the one of $H(X^{(u)})/u$ for u large enough (greater than 10 or 20), as illustrated in Figure S3.

However, local scores are expected to reach higher values for greater correlations between sites. This problem is still the same for sequence analyses: softwares of sequence alignment give Karlin's approximation under the independence model. One solution when sequences are too long to run simulations for the empicical distribution of the local score, is to use approximations under the independence model but with a more stringent threshold.

The effect of considering integer or continuous scores is exemplified on Figure S4. Visually, the local scores were clearer than the single point approach ($\mathcal{F}$-LK), and the continuous score outperformed all the discrete scores.

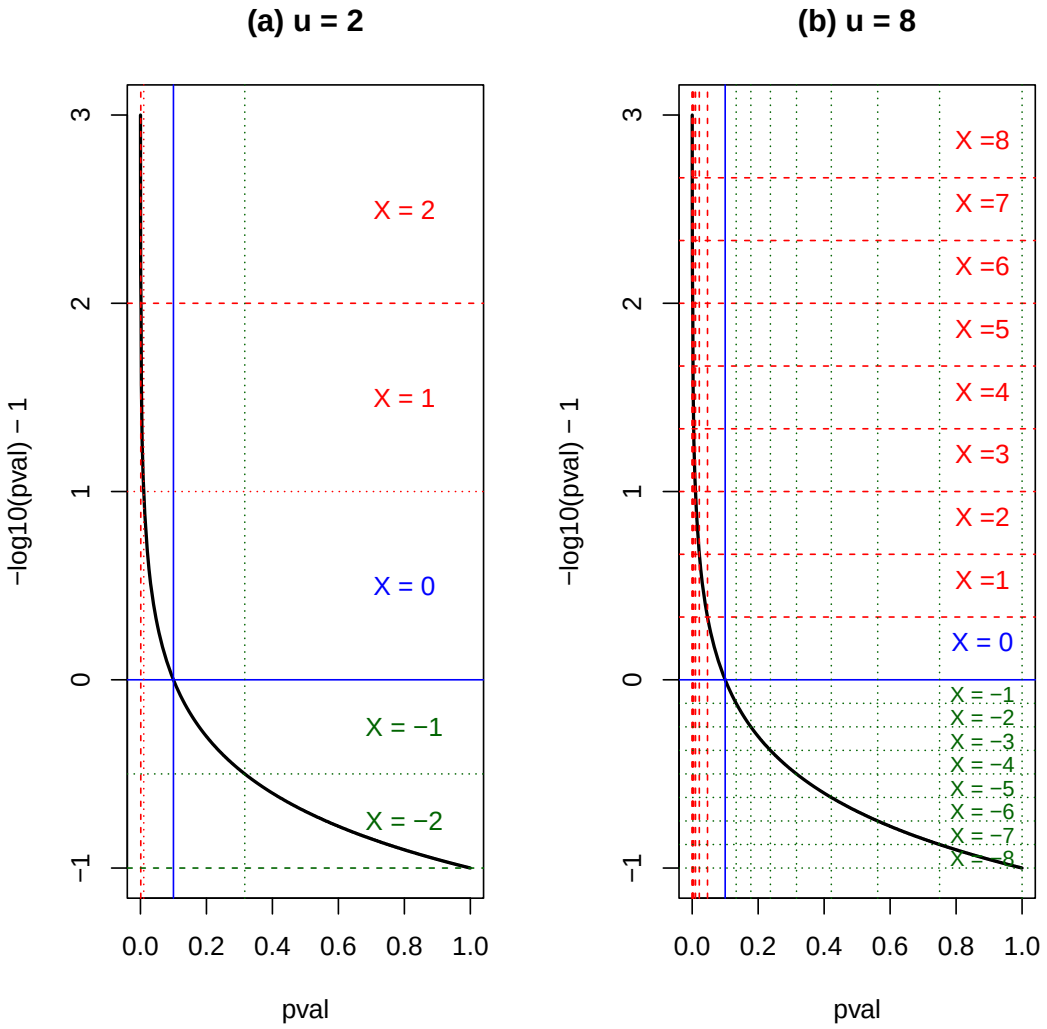


Figure S2. Building of a series of integer score functions $X^{(u)}$ approximating the continuous score function $X^{(\infty)} = -log10(pval_{\mathcal{F}-LK}) - 1$, with maximum value $m = 4$: (a) for $u = 2$, (b) for $u = 8$.

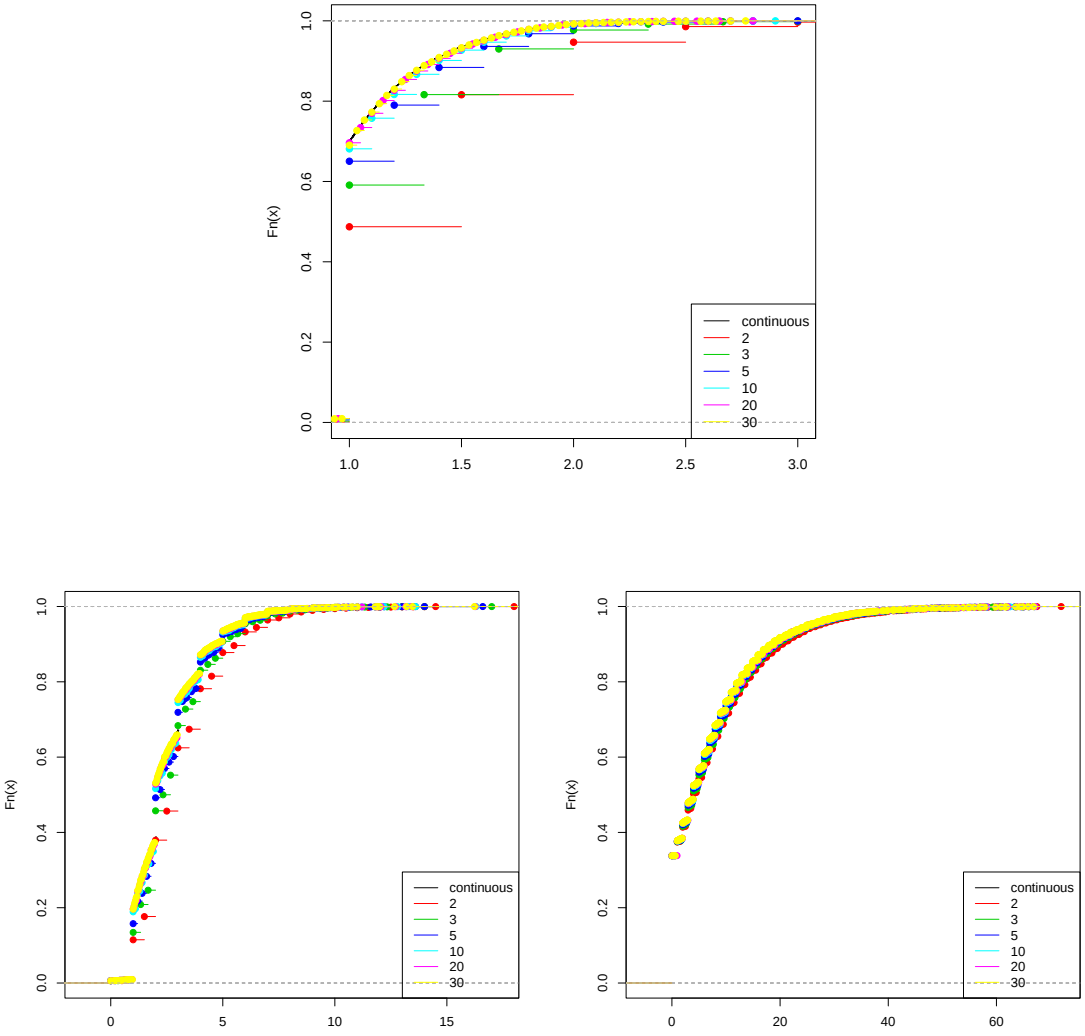


Figure S3. Empirical cumulative density function of local scores H for a chromosome segment of length $L = 100$ under the independence model (top), with an autocorrelation equal to 0.5 (bottom left) and with an autocorrelation equal to 0.9 (bottom right). The local scores were based on the continuous score function $X^{(\infty)} = -log10(x) - 1$ and on discrete score functions $H(X^{(u)})$, for u in $\{2, 3, 5, 10, 20, 30\}$. The local scores were standardized by $m - 1$ (the maximum value of $X^{(\infty)}$) or by u.

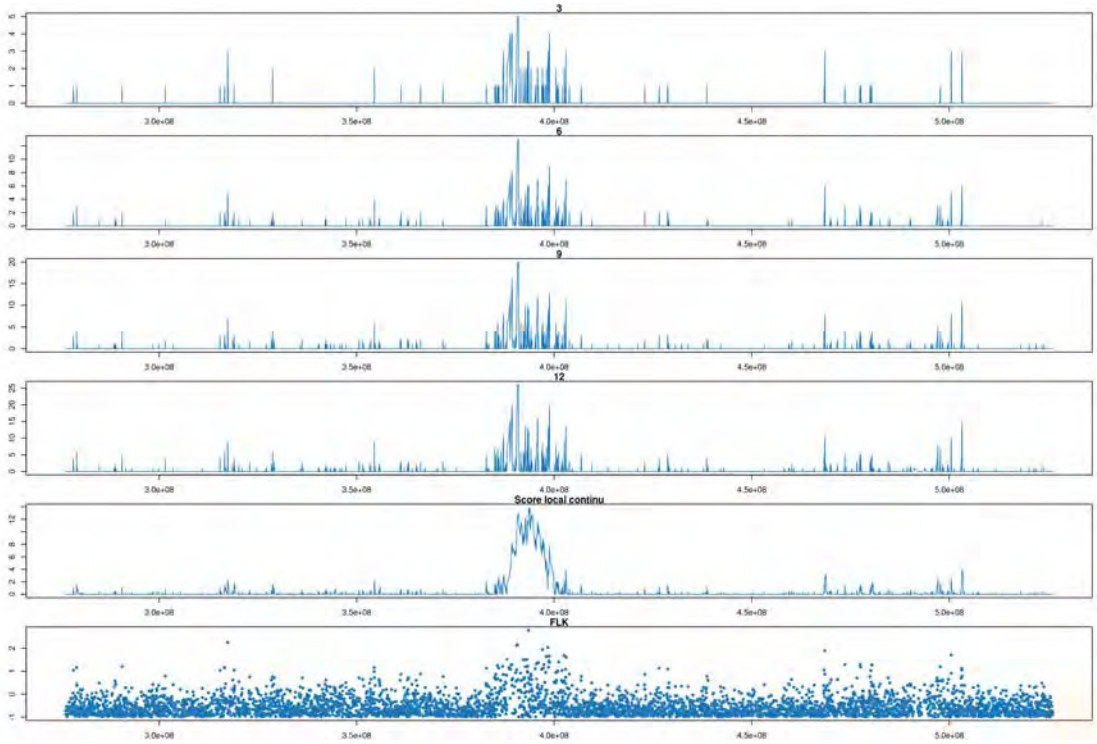


Figure S4. Selection footprints for SheepHapMap data (Northern Europe). Focus on OAR2. Top to bottom: local score (Lindley process) of discretized score functions for $u = 3$, $u = 6$, $u = 9$, $u = 12$, local score for the continuous score function $-log10(pvalue_{\mathcal{F}-LK}) - 1$, and significance result of single point $\mathcal{F}$-LK test ($-log10(pvalue_{\mathcal{F}-LK})$).

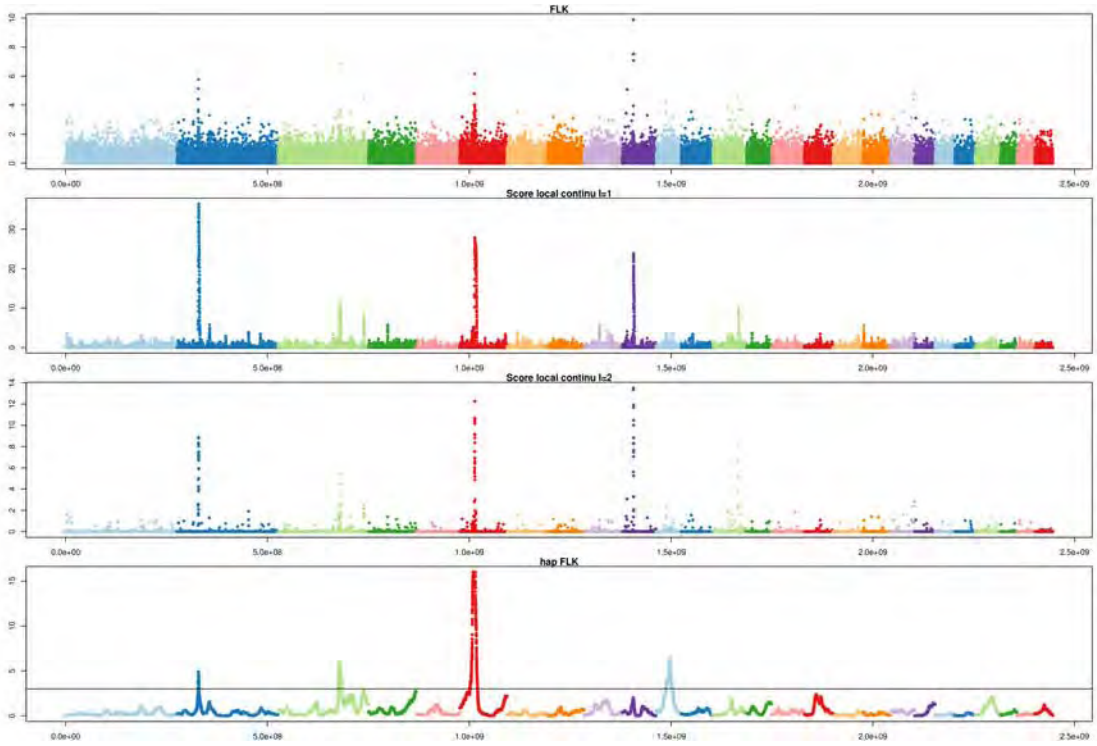


Figure S5. Genome wide selection footprints for SheepHapMap data, breeds from South-western Europe. Same legend as for Figure 3

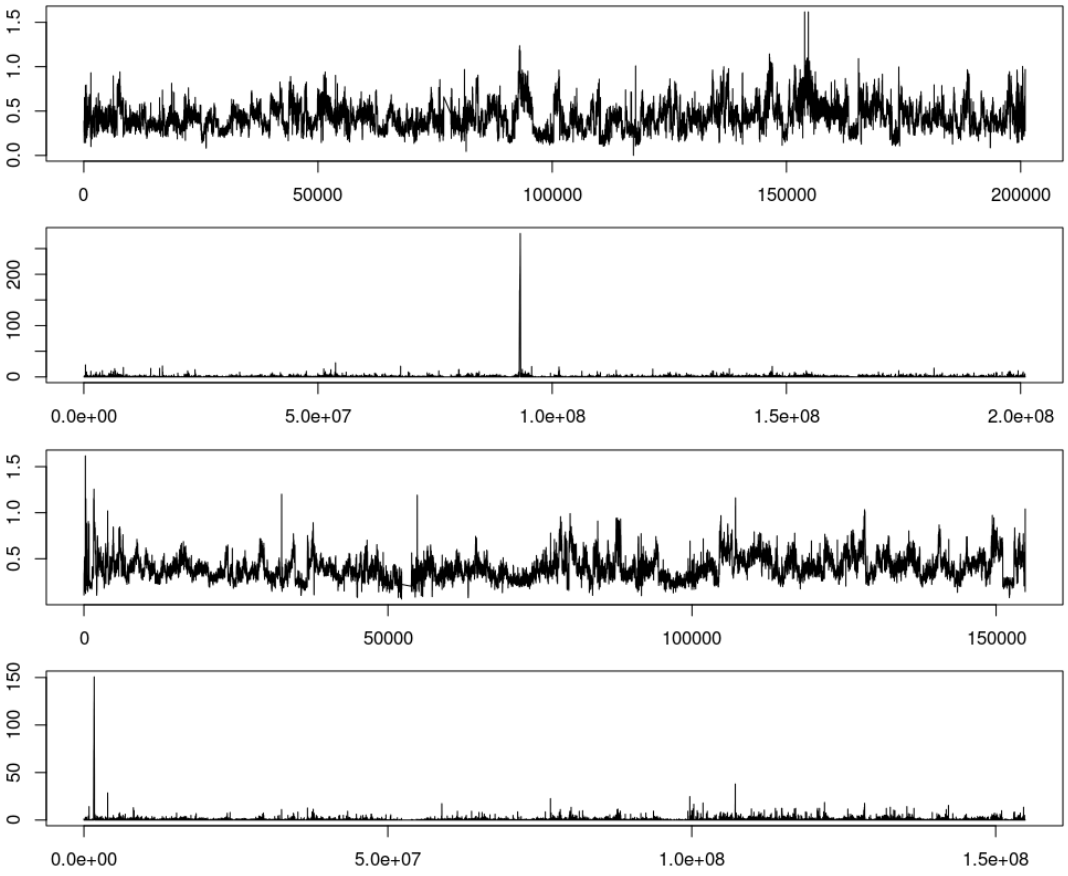


Figure S6. Selection footprints for Quail data on GGA1 and GGA2. From top to bottom: windowed F_{ST} (sliding window fo 10kb) on GGA1, local score (Lindley process) of score function $-log10(pvalue_{FST}) - 1$ on GGA1, windowed F_{ST} (sliding window of 10kb) on GGA2, local score of score function $-log10(pvalue_{FST}) - 1$ on GGA2

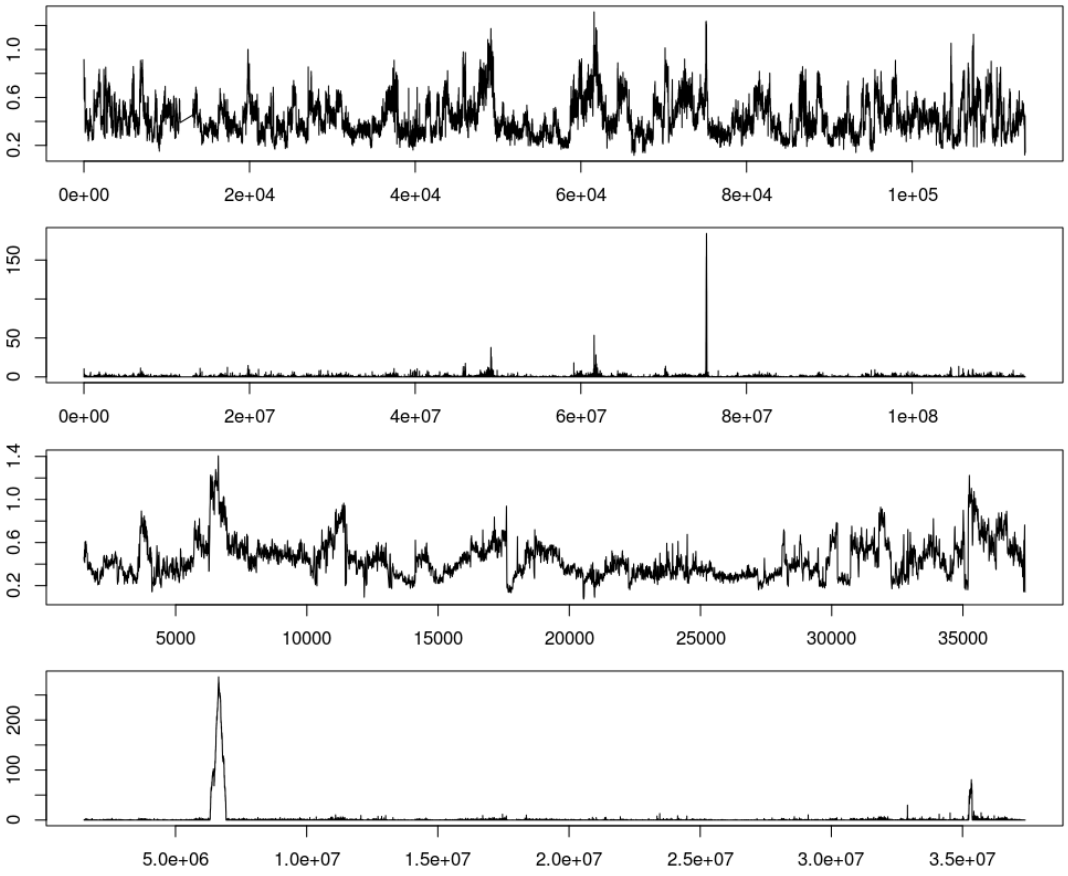


Figure S7. Selection footprints for Quail data on GGA3 and GGA6. From top to bottom: windowed F_{ST} (sliding window fo 10kb) on GGA3, local score (Lindley process) of score function $-log10(pvalue_{FST}) - 1$ on GGA3, windowed F_{ST} (slidding window of 10kb) on GGA6, local score of score function $-log10(pvalue_{FST}) - 1$ on GGA6

Chapter 7

Conclusion

Main findings In this work I proposed two different tests for detecting positive selection using dense genetic data sampled from multiple populations. The first one is directly based on haplotype frequencies, while the second one cumulates allele frequency signals at several loci following the local score theory, taking advantage of linkage disequilibrium. The development of these tests was motivated by the lack of detection methods accounting for linkage disequilibrium in a multiple population context in the literature. Indeed, several haplotype based tests had previously been proposed (reviewed in section 3.3.2), but they are not suited to study more than two populations simultaneously and extending them in this direction does not seem obvious. This is why we proposed hapFLK. On the other hand, several tests accounting for correlations between loci in the absence of individual genotype information had also been proposed (Chen et al., 2010, Guo et al., 2009). These tests are highly computationally demanding. In addition, XP-CLR (Chen et al., 2010) can just be applied to pairs of population, while the Bayesian test of Guo et al. (2009) does not account for the possibly hierarchical structure of sampled populations.

Analyzing a large number of populations simultaneously is very important for the detection of selection signatures, because this provides a more accurate estimation of the ancestral allele frequency p_0 at each locus. When analyzing the quail dataset, I explained why the large uncertainty about p_0,

resulting from the fact that only two populations were observed, could decrease our detection power. In this particular situation, and in other similar experimental designs, the best solution to estimate p_0 is clearly to sequence or genotype individuals from the founder population. However, for genome scans based on outbred populations, DNA from the founder population is generally not available, so the best alternative strategy is to compare as many populations as possible.

Assuming we have data from multiple populations, another important point to consider when estimating p_0 is the structure of these populations. In $\mathcal{F}$-LK and hapFLK, the estimation of p_0 is not a simple average of the allele frequencies in all sampled populations, but a weighted average of these frequencies accounting for the drift between the ancestral population and each sampled population: if a population experienced a lot of drift, its frequency will be down weighted when estimating p_0. Correlated populations will also be down weighted, because they contribute redundant information.

A recurrent problem of the tests designed for detecting selection is the threshold used for classifying a signal as significant or not. A common practice is to consider an arbitrary proportion ($1, 5, 10\%$ depending on the test) of the tail of the genome wide statistic distribution, as being under selection. This procedure provides no control of the false positive rate. I tried to propose objective statistical procedures to declare a locus as significantly not neutral. For the hapFLK test, we proposed to fit the neutral distribution of the statistic with a normal distribution, estimating the mean and the variance of the distribution using a robust linear model in order to reduce the influence of the locus under selection. In all sheep groups we could check that this normal approximation was reasonable, but users should also verify that it holds when testing other datasets. We further proposed to determine the detection threshold based on q-values rather than on p-values. In Chapter 5 we saw that this statistical procedure lead to different amounts of signals for the different groups, for instance no selection signature was detected in the Asian group. Considering the locus in the upper tail of hapFLK genome wide distribution as selection targets, would give the same quantity of signals in every group, and the regions detected in Asia would most likely be false

positives.

In the local score framework, the p-value is not associated to a single marker, but to a whole sequence. It corresponds to the probability of obtaining a local score greater or equal than the observed one, given the length of the sequence and the expected correlation between markers. Theoretical formulas allow to approximate the p-value of a segment assuming independence between markers, but we saw in Chapter 6 that using these formulas when markers are highly correlated leads to underestimated p-values (which implies an excess of false positives). We therefore decided to evaluate p-values by simulating sequences of correlated markers. The q-values could not be computed for this approach, as their computation requires a p-value at every SNP, not a single p-value for the whole segment.

Finally, I described two methods allowing to determine the population(s) under selection in a region that has been previously detected. The local tree representation seems to be the most user friendly method, in particular it provides more interpretable results when the number of populations is large. But the spectral decomposition is also interesting. For instance it can help to elucidate whether there is interference between several selection signals in a same region. Such interferences are quite unlikely when the observed data is very dense, but in the case of 50K marker genotypes they could easily occur, because the signals often cover more than 1Mb. These two procedures could also be applied to genome scans based on the local score approach, but they require more than two populations so I did not use them when analyzing the quail dataset. In a multi-population context, local trees could be computed using the allele frequencies within the interval associated to the local score, and a local score could be computed independently for each of the contributions arising from the spectral decomposition of $\mathcal{F}$-LK.

In the case of hapFLK, I also presented a cluster frequency representation of the signals, which can help to find out the population(s) under selection. However, as we saw in Chapter 4 with the signal on chromosome 14, clusters that seem highly frequent in a population are not necessarily the ones causing the signal. I consequently advise to always use one of the two methods mentioned above, in addition to the cluster frequency representation.

Some perspectives Simulations showed that hapFLK is robust to some bottleneck or migration scenarios. Bottlenecks are accounted for in the tree model underlying $\mathcal{F}$-LK and hapFLK, and result in very long branches for the populations that experienced a bottleneck. Selection signatures have to be very strong to be detected in bottlenecked populations, otherwise the signal can be confounded with that of the bottleneck. If the bottleneck was too strong, the best thing is probably to leave the population out of the analysis in order to increase the detection power for the other populations in the group. On the other hand, migration is not accounted for in the current implementation of $\mathcal{F}$-LK and hapFLK. Punctual admixture events between populations that belong to the tree could easily be modeled by replacing the kinship matrix proposed in $\mathcal{F}$-LK by the kinship matrix proposed by Pickrell and Pritchard (2012). Admixture events from populations outside the tree are more problematic. Indeed, they do not only imply a modification of the kinship matrix, but also of the ancestral population, which could be much more distant than assumed by the model. In this work populations that were admixed with populations from either inside (e.g. the Rasaaragonesa) or outside (e.g. the Swiss crossbreeds) the tree were left out of the analysis. Further simulations will be required to evaluate the performance of $\mathcal{F}$-LK, hapFLK and the local score in different migration scenarios, using either the current kinship matrix or a new one accounting for migration.

hapFLK and $\mathcal{F}$-LK could also be extended to include response values possibly related to the selection pressure, as latitudes or coat colors. Such information was not available for the sheep that were genotyped in the Sheep HapMap dataset, so I did not investigate this point. However, the example of microphatlmia shows that we are able to detect a genomic region related to a specific response variable, even without observed data from this variable, which is an encouraging fact for further research in this direction. Actually, a Bayesian method allowing to correlate allele frequencies at a single marker with environmental variables has recently been proposed Coop et al. (2010). In this method, called Bayesenv, allele frequencies are modeled in a very similar way as in $\mathcal{F}$-LK, in particular population structure is also accounted for. Using simulations, Gunther and Coop (2013) found that a

version of Bayesenv without environmental variables performed similarly as $\mathcal{F}$-LK, while adding the environmental information increased the power to detect local adaptation. These results motivate the inclusion of environmental variables within an haplotype test like hapFLK, which might provide an even greater power for the detection of local adaptation events.

Types of detected sweep When looking for selection signatures, we usually do not know if selection intensity was strong or moderate, if selection started from a new mutation (hard sweep) or from standing variation (soft sweep) and if the selected mutation already fixed in the population. We saw in the first chapter that ongoing or recent hard sweeps are generally the easiest to detect, because they leave a recognizable pattern on both the haplotype structure and the allele frequencies. For starting hard sweeps, the signal is clearer on haplotypes than on the frequency spectrum. Power can be raised in this case by comparing related populations. Accounting for correlations between markers using hapFLK or the local score should also increase the detection power. Indeed, even if the allele frequency difference between populations is not very high, this difference extends over a long region because recombination did not have enough time to break the selected haplotype. hapFLK should be able to detect the elevated frequency of such a long haplotype, and the local score will cumulate a large number of small signals into a globally significant region. If the hard sweep is old, we have exactly the opposite situation. Haplotypes will be very short, because of recombination and eventually new mutations around the selected locus, but the selected haplotypes and the allele frequencies should have very different frequencies between populations. In this case dense data is needed to detect the selection signature, and a single marker test could be more powerful than hapFLK (in Chapter 5 we saw that some signals were detected by $\mathcal{F}$-LK but not by hapFLK). The local score approach should be suited for detecting these type of signals also, because a few very large scores will be cumulated.

It is more difficult to predict how the methods I developed will behave in the case of soft sweeps, because the origins and types of soft sweeps are much more diverse. Besides, I think we still need to better understand the signals

we detected, to know what kind of soft sweeps they are. Nevertheless, several of the results presented in this work indicate that the methods I developed can detect soft sweeps, which is generally difficult with existing tests.

Through simulations I showed that hapFLK has some power to detect sweeps on standing variation. As already discussed, this is actually the scenario where the difference in detection power with XP-EHH is the largest. In addition, some of the signals detected by hapFLK in the sheep are clearly soft sweeps. A good example is the signal detected on chromosome 14 in the New Zealand breeds, where selection seems to have started after the arrival of the breeds in New Zealand. Three distinct haplotypes were selected in this region, one in the Texels and two other in the Romneys, so clearly selection can not have started from one single haplotype. We do not have enough information yet to determine if selection started from a single standing mutation, or if a same gene was affected by several independent mutations.

In the case of the quail dataset, sweeps are expected to be from standing variation, because evolution time is too short for a new functional mutation to appear. On the other hand, several mutations related to quail behavior were likely present in the founder population, and became selected when the experiment started. Selection was then very strong, so these selected mutations experienced a large increase in frequency, and in many cases probably fixed, during the 53 generations of the experiment. Incomplete hard sweeps are thus also quite unlikely in this case. Besides, the fact that we observed a relatively small increase of the number of fixed SNPs in detected regions supports the soft sweep hypothesis. Indeed, assuming a couple of haplotypes are initially selected, SNPs where a single allele is carried by chance by all these haplotypes will become fixed, while other SNPs will likely maintain their two alleles. Under a complete hard sweep scenario, only one haplotype is selected so we would expect a small region with almost no variation. And under an incomplete hard sweep, we would also expect a small region with reduced heterozygosity, but with almost no increase in the number of fixed SNPs. Consequently, the fact that the local score performed very well in the quail experiment indicates that it has good power to detect soft sweeps. This is due to the fact that it cumulates signals from alleles that have fixed, but

also from alleles that have raised in frequency without fixing.

Designing a genome scan for selection With the two tests developed in this work, genome scans for selection accounting for multiple population information and for linkage disequilibrium, can be performed for a large variety of genomic datasets. Datasets including only population allele frequencies can be analyzed with the local score approach, while datasets including individual genotypes can be analyzed using both hapFLK and the local score approach. I conclude this study by discussing some aspects of the experimental design that will allow to take maximum advantage of these detection methods.

I did not study explicitly the influence of the sample size in each population, but it seems to me that that we do not need a large amount of individuals per population, because the footprints that we aim to detect are based on large frequency differences (either haplotype frequencies or allele frequencies). In the quail dataset, only 10 individuals (20 alleles) were sequenced in each line, and clear signatures were however detected. More individuals are probably required to account for haplotype information, first because haplotypes are multi allelic markers whose diversity is more difficult to capture, second because haplotypes generally have to be inferred from genotype data, and this phasing step requires relatively large sample sizes. When analyzing the Sheep HapMap data, we considered only populations with more than 20 individuals (40 haplotypes).

In contrast, the number of sampled populations seems to be a very important parameter and should be a priority, unless an ancestral population is available. As discussed above, observing more populations helps to better estimate the ancestral allele and haplotype frequencies. If possible, at least more than 3 populations should be sampled, so that even in the regions under selection we can hopefully have two neutral populations allowing to estimate allele frequencies. The choice of the populations is also very important. Genotyping closely related populations will maximize the detection power of recent sweeps. If populations are more distant from each others, older sweeps will possibly be detected, but there is a risk of loosing detection

power because some frequency patterns will become as likely under genetic drift as under selection.

The marker density and the type of data provided by a given technology are other important parameters, which often have to be balanced due to financial constraints. Clearly, a minimum marker density is required in order to take advantage of linkage disequilibrium information, otherwise the allele frequencies observed at consecutive locus become independent and can not be cumulated by the local score, or combined into meaningful haplotypes. In the Sheep HapMap dataset, phasing the animals was sometimes difficult, so decreasing marker density below that of the 50K chip might be for instance problematic. But up to a given marker density (which will depend on the extent of linkage disequilibrium in the sampled populations), one may wonder if, for a similar cost, it is better to maximize the number of observed markers while loosing individual information, which can be done by sequencing DNA pools, or to have individual genotypes for only a subset of SNPs, using for instance dense SNP chips or RAD sequencing.

When studying diversity in a group of populations and trying to understand the mechanisms of adaptation, like in the worldwide sheep dataset, genotypes should be more informative. This design already provides a high detection power, and provides a clear picture of the variations of haplotype frequencies in the sampled populations, allowing to get an idea of the evolution of these frequencies along the population tree. Also, while some soft sweeps were apparently detected in the quail experiment using the local score approach, we can intuitively guess that complex soft sweeps patterns will be better captured using haplotype data, which represents a richer source of information. However, if the goal is to detect ancient selective sweeps, then marker density should be a priority, because linkage disequilibrium decays with time, so close markers are needed. With the 50K chip, we detected a selection signature around the horns locus, which must be quite ancient, but we could maybe detect more traits like this one with a higher density.

Finally, if the objective is to identify causal mutations, as for instance in the quail experiment or other similar experiments where we already know that selection has occurred, we need to delimit the selected regions as pre-

cisely as possible. In this case, pool sequencing seems to be a better choice than dense genotyping. Indeed, the regions under selection should be detected despite the loss of individual information, and thanks to the almost exhaustive genome coverage causal mutations will likely be among the most differentiated observed markers. The risk of pool-sequencing would be to miss some complicated sweeps signals, but from the quail experiment I believe this risk is rather small.

Appendix A

Appendix

A.1 Inbreeding coefficient

Definition A.1. *We call* inbreeding coefficient *the probability of sampling two identical alleles that descend from the same ancestral allele at a given generation. We call such alleles Identical By Descent (IBD).*

We can decompose the probability of sampling two IBD alleles by (1) sampling the same allele twice ($prob = \frac{1}{2N}$) or (2) sampling two different alleles IBD ($prob = (1 - \frac{1}{2N})IBD^{(t-1)}$), with $IBD^{(t)}$ the probability of being IBD at generation t. Under genetic drift, we have:

$$\begin{aligned} IBD^{(t)} &= \tfrac{1}{2N} + \left(1 - \tfrac{1}{2N}\right) IBD^{(t-1)} \\ &= 1 - \left(1 - \tfrac{1}{2N}\right)\left(1 - IBD^{(t-1)}\right) \end{aligned}$$

By recurrence,

$$IBD^{(t)} = 1 - \left(1 - \frac{1}{2N}\right)^t \left(1 - IBD^{(0)}\right)$$

and by definition $IBD^{(0)} = 0$.

The notion of identity by descent is always related to the founder generation. For simplicity we note F_t instead of $IBD^{(t)}$ and call it the *inbreeding coefficient*.

Replacing $1 - \left(1 - \frac{1}{2N}\right)^t$ by F_t in equation 2.3, the variance of the fre-

quency $p(t)$ can be written in terms of the inbreeding coefficient as:

$$Var(p_t) = F_t p_0(1 - p_0) \tag{A.1}$$

List of Figures

Bibliography

Alexander, D. H., Novembre, J., and Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.*, 19(9):1655–1664.

Andersson, L. (2012). How selective sweeps in domestic animals provide new insight into biological mechanisms. *J. Intern. Med.*, 271(1):1–14.

Aulchenko, Y. S., Struchalin, M. V., Belonogova, N. M., Axenovich, T. I., Weedon, M. N., Hofman, A., Uitterlinden, A. G., Kayser, M., Oostra, B. A., van Duijn, C. M., Janssens, A. C., and Borodin, P. M. (2009). Predicting human height by Victorian and genomic methods. *Eur. J. Hum. Genet.*, 17(8):1070–1075.

Balding, D. and Bishop, M.and Cannings, C. (2007). *Handbook of Statistical Genetics*, volume 2. Wiley, 3 edition.

Beaumont, M. A. (2005). Adaptation and speciation: what can fst tell us? *Trends in Ecology and Evolution*, 20(8):435 – 440.

Beaumont, M. A. and Balding, D. J. (2004). Identifying adaptive genetic divergence among populations from genome scans. *Mol. Ecol.*, 13:969 – 980.

Beaumont, M. A. and Nichols, R. A. (1996). Evaluating loci for use in the genetic analysis of population structure. 263(1377):1619–1626.

Bhatia, G., Patterson, N., Pasaniuc, B., Zaitlen, N., Genovese, G., Pollack, S., Mallick, S., Myers, S., Tandon, A., Spencer, C., Palmer, C. D.,

Adeyemo, A. A., Akylbekova, E. L., Cupples, L. A., Divers, J., Fornage, M., Kao, W. H., Lange, L., Li, M., Musani, S., Mychaleckyj, J. C., Ogunniyi, A., Papanicolaou, G., Rotimi, C. N., Rotter, J. I., Ruczinski, I., Salako, B., Siscovick, D. S., Tayo, B. O., Yang, Q., McCarroll, S., Sabeti, P., Lettre, G., De Jager, P., Hirschhorn, J., Zhu, X., Cooper, R., Reich, D., Wilson, J. G., and Price, A. L. (2011). Genome-wide comparison of African-ancestry populations from CARe and other cohorts reveals signals of natural selection. *Am. J. Hum. Genet.*, 89(3):368–381.

Bonhomme, M., Chevalet, C., Servin, B., Boitard, S., Abdallah, J., Blott, S., and San Cristobal, M. (2010). Detecting selection in population trees: The Lewontin and Krakauer test extended. *Genetics*, 186(1):241–262.

Browning, B. L. and Browning, S. R. (2007a). Efficient multilocus association testing for whole genome association studies using localized haplotype clustering. *Genet. Epidemiol.*, 31(5):365–375.

Browning, S. R. (2006). Multilocus association mapping using variable-length Markov chains. *Am. J. Hum. Genet.*, 78:903–913.

Browning, S. R. and Browning, B. L. (2007b). Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am. J. Hum. Genet.*, 81(5):1084–1097.

Carneiro, M., Afonso, S., Geraldes, A., Garreau, H., Bolet, G., Boucher, S., Tircazes, A., Queney, G., Nachman, M. W., and Ferrand, N. (2011). The genetic structure of domestic rabbits. *Mol. Biol. Evol.*, 28(6):1801–1816.

Cavalli-Sforza, L. L. (1966). Some old and new data on the genetics of human populations. *Ala J Med Sci*, 3(4):376–381.

Chen, H., Patterson, N., and Reich, D. (2010). Population differentiation as a test for selective sweeps. *Genome Res.*, 20(3):393–402.

Collins, F. S., Brooks, L. D., and Chakravarti, A. (1998). A DNA polymorphism discovery resource for research on human genetic variation. *Genome Res.*, 8(12):1229–1231.

Consortium, T. I. H. (2005). A haplotype map of the human genome. *Nature*, 437(7063):1299–1320.

Coop, G., Witonsky, D., Di Rienzo, A., and Pritchard, J. (2010). Using environmental correlations to identify loci underlying local adaptation. *Genetics*, 185(4):1411–1423.

Darwin, C. (1859). *On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in he Struggle for Life*. John Murray, London, 1 edition.

Diamond, J. (2002). Evolution, consequences and future of plant and animal domestication. *Nature*, 418(6898):700–707.

Durrett, R. and Schweinsberg, J. (2004). Approximating selective sweeps. *Theor Popul Biol*, 66(2):129–138.

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

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

Gautier, M., Flori, L., Riebler, A., Jaffrezic, F., Laloe, D., Gut, I., Moazami-Goudarzi, K., and Foulley, J. L. (2009). A whole genome Bayesian scan for adaptive genetic divergence in West African cattle. *BMC Genomics*, 10:550.

Gillespie, J. H. (1997). Junk ain't what junk does: neutral alleles in a selected context. *Gene*, 205(1-2):291–299.

Gompert, Z. and Buerkle, C. A. (2011). A hierarchical bayesian model for next-generation population genomics. *Genetics*, 187(3):903–917.

Greenspan, G. and Geiger, D. (2006). Modeling haplotype block variation using Markov chains. *Genetics*, 172(4):2583–2599.

Gudbjartsson, D. F., Walters, G. B., Thorleifsson, G., Stefansson, H., Halldorsson, B. V., Zusmanovich, P., Sulem, P., Thorlacius, S., Gylfason, A., Steinberg, S., Helgadottir, A., Ingason, A., Steinthorsdottir, V., Olafsdottir, E. J., Olafsdottir, G. H., Jonsson, T., Borch-Johnsen, K., Hansen, T., Andersen, G., Jorgensen, T., Pedersen, O., Aben, K. K., Witjes, J. A., Swinkels, D. W., den Heijer, M., Franke, B., Verbeek, A. L., Becker, D. M., Yanek, L. R., Becker, L. C., Tryggvadottir, L., Rafnar, T., Gulcher, J., Kiemeney, L. A., Kong, A., Thorsteinsdottir, U., and Stefansson, K. (2008). Many sequence variants affecting diversity of adult human height. *Nat. Genet.*, 40(5):609–615.

Gunther, T. and Coop, G. (2013). Robust identification of local adaptation from allele frequencies.

Guo, F., Dey, D. K., and Holsinger, K. E. (2009). A bayesian hierarchical model for analysis of single-nucleotide polymorphisms diversity in multilocus, multipopulation samples. *Journal of the American Statistical Association*, 104(485):142–154.

Hahn, M. W. (2008). Toward a selection theory of molecular evolution. *Evolution*, 62(2):255–265.

Hamblin, M. T. and Di Rienzo, A. (2000). Detection of the signature of natural selection in humans: evidence from the Duffy blood group locus. *Am. J. Hum. Genet.*, 66(5):1669–1679.

Hamblin, M. T., Thompson, E. E., and Di Rienzo, A. (2002). Complex signatures of natural selection at the Duffy blood group locus. *Am. J. Hum. Genet.*, 70(2):369–383.

Hermisson, J. and Pennings, P. S. (2005). Soft sweeps: molecular population genetics of adaptation from standing genetic variation. *Genetics*, 169:2335–2352.

Hernandez, R. D., Kelley, J. L., Elyashiv, E., Melton, S. C., Auton, A., McVean, G., Sella, G., and Przeworski, M. (2011). Classic selective sweeps were rare in recent human evolution. *Science*, 331:920–924.

Hinds, D. A., Stuve, L. L., Nilsen, G. B., Halperin, E., Eskin, E., Ballinger, D. G., Frazer, K. A., and Cox, D. R. (2005). Whole-genome patterns of common dna variation in three human populations. *Science*, 307(5712):1072–1079.

Innan, H. and Kim, Y. (2004). Pattern of polymorphism after strong artificial selection in a domestication event. *Proc. Natl. Acad. Sci. U.S.A.*, 101(29):10667–10672.

Innan, H. and Kim, Y. (2008). Detecting local adaptation using the joint sampling of polymorphism data in the parental and derived populations. *Genetics*, 179(3):1713–1720.

Kijas, J. W., Lenstra, J. A., Hayes, B., Boitard, S., Porto Neto, L. R., San Cristobal, M., Servin, B., McCulloch, R., Whan, V., Gietzen, K., Paiva, S., Barendse, W., Ciani, E., Raadsma, H., McEwan, J., Dalrymple, B., and other members of the International Sheep Genomics Consortium (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.

Kim, Y. and Nielsen, R. (2004). Linkage disequilibrium as a signature of selective sweeps. *Genetics*, 167:1513–1524.

Kim, Y. and Stephan, W. (2002). Detecting a local signature of genetic hitchhiking along a recombining chromosome. *Genetics*, 160:765–777.

Laval, G., Patin, E., Barreiro, L. B., and Quintana-Murci, L. (2010). Formulating a historical and demographic model of recent human evolution based on resequencing data from noncoding regions. *PLoS ONE*, 5(4):e10284.

Lettre, G., Jackson, A. U., Gieger, C., Schumacher, F. R., Berndt, S. I., Sanna, S., Eyheramendy, S., Voight, B. F., Butler, J. L., Guiducci, C.,

Illig, T., Hackett, R., Heid, I. M., Jacobs, K. B., Lyssenko, V., Uda, M., Boehnke, M., Chanock, S. J., Groop, L. C., Hu, F. B., Isomaa, B., Kraft, P., Peltonen, L., Salomaa, V., Schlessinger, D., Hunter, D. J., Hayes, R. B., Abecasis, G. R., Wichmann, H. E., Mohlke, K. L., and Hirschhorn, J. N. (2008). Identification of ten loci associated with height highlights new biological pathways in human growth. *Nat. Genet.*, 40(5):584–591.

Lewontin, R. C. and Krakauer, J. (1973). Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms. *Genetics*, 74(1):175–195.

Lewontin, R. C. and Krakauer, J. (1975). Letters to the editors: Testing the heterogeneity of F values. *Genetics*, 80(2):397–398.

Li, J., Li, H., Jakobsson, M., Li, S., Sjodin, P., and Lascoux, M. (2012). Joint analysis of demography and selection in population genetics: where do we stand and where could we go? *Mol. Ecol.*, 21(1):28–44.

Liu, H., Prugnolle, F., Manica, A., and Balloux, F. (2006). A geographically explicit genetic model of worldwide human-settlement history. *The American Journal of Human Genetics*, 79(2):230 – 237.

Mardis, E. R. (2013). Next-Generation Sequencing Platforms. *Annu Rev Anal Chem (Palo Alto Calif)*.

Maynard Smith, J. and Haigh, J. (1974). The hitch-hiking effect of a favourable gene. *Genet. Res.*, 23(1):23–35.

Nei, M. and Chakravarti, A. (1977). Drift variances of FST and GST statistics obtained from a finite number of isolated populations. *Theor Popul Biol*, 11(3):307–325.

Nei, M., Chakravarti, A., and Tateno, Y. (1977). Mean and variance of FST in a finite number of incompletely isolated populations. *Theor Popul Biol*, 11(3):291–306.

Nei, M. and Maruyama, T. (1975). Letters to the editors: Lewontin-Krakauer test for neutral genes. *Genetics*, 80(2):395.

Nicholson, G., Smith, A. V., Jonsson, F., Gustafsson, O., Stefnsson, K., and Donnelly, P. (2002). Assessing population differentiation and isolation from single-nucleotide polymorphism data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(4):695–715.

of Animal Science Oklahoma State University, D. (1995). Dorper sheep.

Oleksyk, T. K., Zhao, K., De La Vega, F. M., Gilbert, D. A., O'Brien, S. J., and Smith, M. W. (2008). Identifying selected regions from heterozygosity and divergence using a light-coverage genomic dataset from two human populations. *PLoS ONE*, 3(3):e1712.

Pennings, P. S. and Hermisson, J. (2006a). Soft sweeps II–molecular population genetics of adaptation from recurrent mutation or migration. *Mol. Biol. Evol.*, 23(5):1076–1084.

Pennings, P. S. and Hermisson, J. (2006b). Soft sweeps III: the signature of positive selection from recurrent mutation. *PLoS Genet.*, 2(12):e186.

Pickrell, J. K. and Pritchard, J. K. (2012). Inference of population splits and mixtures from genome-wide allele frequency data. *PLoS Genet.*, 8(11):e1002967.

Price, A. L., Zaitlen, N. A., Reich, D., and Patterson, N. (2010). New approaches to population stratification in genome-wide association studies. *Nat. Rev. Genet.*, 11(7):459–463.

Pritchard, J. K., Pickrell, J. K., and Coop, G. (2010). The genetics of human adaptation: hard sweeps, soft sweeps, and polygenic adaptation. *Curr. Biol.*, 20(4):R208–215.

Przeworski, M., Coop, G., and Wall, J. D. (2005). The signature of positive selection on standing genetic variation. *Evolution*, 59(11):2312–2323.

Riebler, A., Held, L., and Stephan, W. (2008). Bayesian variable selection for detecting adaptive genomic differences among populations. *Genetics*, 178:1817–1829.

Robertson, A. (1975a). Gene frequency distributions as a test of selective neutrality. 81(4):775–785.

Robertson, A. (1975b). Remarks on the lewontin-krakauer test. 80(2):396.

Sabeti, P. C., Reich, D. E., Higgins, J. M., Levine, H. Z. P., Richter, D. J., Schaffner, S. F., Gabriel, S. B., Platko, J. V., Patterson, N. J., McDonald, G. J., Ackerman, H. C., Campbell, S. J., Altshuler, D., Cooper, R., Kwiatkowski, D., Ward, R., and Lander, E. S. (2002). Detecting recent positive selection in the human genome from haplotype structure. *Nature*, 419:832–837.

Sabeti, P. C., Schaffner, S. F., Fry, B., Lohmueller, J., Varilly, P., Shamovsky, O., Palma, A., Mikkelsen, T. S., Altshuler, D., and Lander, E. S. (2006). Positive natural selection in the human lineage. *Science*, 312(5780):1614–1620.

Sabeti, P. C., Varilly, P., Fry, B., Lohmueller, J., Hostetter, E., Cotsapas, C., Xie, X., Byrne, E. H., McCarroll, S. A., Gaudet, R., Schaffner, S. F., and Lander, E. S. (2007). Genome-wide detection and characterization of positive selection in human populations. *Nature*, 449:913–918.

Saitou, N. and Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, 4:406–425.

SanCristobal-Gaudy, M., Elsen, J., Bodin, L., and Chevalet, C. (1998). Prediction of the response to a canalising selection of a continuous trait in animal breeding. . *Genet. Sel. Evol.*, 25:3–30.

Scheet, P. (2006). *A flexible computationally tractable model for patterns of population genetic variation*. PhD thesis, Universitiy of Washington, Graduate School.

Scheet, P. and Stephens, M. (2006). A fast and flexible statistical model for large-scale population genotype data: Applications to inferring miss-

ing genotypes and haplotypic phase. *The American Journal of Human Genetics*, 78:629–644.

Stephan, W., Song, Y. S., and Langley, C. H. (2006). The hitchhiking effect on linkage disequilibrium between linked neutral loci. *Genetics*, 172(4):2647–2663.

Stephens, M., Smith, N. J., and Donnelly, P. (2001). A new statistical method for haplotype reconstruction from population data. *Am. J. Hum. Genet.*, 68(4):978–989.

Tang, K., Thornton, K. R., and Stoneking, M. (2007). A new approach for using genome scans to detect recent positive selection in the human genome. *PLoS Biol*, 5:e171.

Tishkoff, S. A., Reed, F. A., Friedlaender, F. R., Ehret, C., Ranciaro, A., Froment, A., Hirbo, J. B., Awomoyi, A. A., Bodo, J.-M., Doumbo, O., Ibrahim, M., Juma, A. T., Kotze, M. J., Lema, G., Moore, J. H., Mortensen, H., Nyambo, T. B., Omar, S. A., Powell, K., Pretorius, G. S., Smith, M. W., Thera, M. A., Wambebe, C., Weber, J. L., and Williams, S. M. (2009). The genetic structure and history of africans and african americans. *Science*, 324(5930):1035–1044.

Tishkoff, S. A., Reed, F. A., Ranciaro, A., Voight, B. F., Babbitt, C. C., Silverman, J. S., Powell, K., Mortensen, H. M., Hirbo, J. B., Osman, M., Ibrahim, M., Omar, S. A., Lema, G., Nyambo, T. B., Ghori, J., Bumpstead, S., Pritchard, J. K., Wray, G. A., and Deloukas, P. (2007). Convergent adaptation of human lactase persistence in Africa and Europe. *Nat. Genet.*, 39(1):31–40.

Tishkoff, S. A., Varkonyi, R., Cahinhinan, N., Abbes, S., Argyropoulos, G., Destro-Bisol, G., Drousiotou, A., Dangerfield, B., Lefranc, G., Loiselet, J., Piro, A., Stoneking, M., Tagarelli, A., Tagarelli, G., Touma, E. H., Williams, S. M., and Clark, A. G. (2001). Haplotype diversity and linkage disequilibrium at human G6PD: recent origin of alleles that confer malarial resistance. *Science*, 293(5529):455–462.

Tsakas, S. and Krimbas, C. B. (1976). Testing the heterogeneity of f values: A suggestion and a correction. *Genetics*, 84(2):399–401.

Vila, C., Savolainen, P., Maldonado, J. E., Amorim, I. R., Rice, J. E., Honeycutt, R. L., Crandall, K. A., Lundeberg, J., and Wayne, R. K. (1997). Multiple and ancient origins of the domestic dog. *Science*, 276(5319):1687–1689.

Vitalis, R., Dawson, K., and Boursota, P. (2001). Interpretation of variation across marker loci as evidence of selection. *Genetics*, 158:1811–1823.

Voight, B. F., Kudaravalli, S., Wen, X., and Pritchard, J. K. (2006). A map of recent positive selection in the human genome. *PLoS Biol*, 4:e72.

Wall, J. D. and Hammer, M. F. (2006). Archaic admixture in the human genome. *Current Opinion in Genetics and Development*, 16(6):606 – 610.

Weedon, M. N., Lango, H., Lindgren, C. M., Wallace, C., Evans, D. M., Mangino, M., Freathy, R. M., Perry, J. R., Stevens, S., Hall, A. S., Samani, N. J., Shields, B., Prokopenko, I., Farrall, M., Dominiczak, A., Johnson, T., Bergmann, S., Beckmann, J. S., Vollenweider, P., Waterworth, D. M., Mooser, V., Palmer, C. N., Morris, A. D., Ouwehand, W. H., Zhao, J. H., Li, S., Loos, R. J., Barroso, I., Deloukas, P., Sandhu, M. S., Wheeler, E., Soranzo, N., Inouye, M., Wareham, N. J., Caulfield, M., Munroe, P. B., Hattersley, A. T., McCarthy, M. I., and Frayling, T. M. (2008). Genome-wide association analysis identifies 20 loci that influence adult height. *Nat. Genet.*, 40(5):575–583.

Weir, B. and Cockerham, C. (1984). Estimating f-statistics for the analysis of population structure. *Evolution*, 38:1358–1370.

Weir, B. S., Cardon, L. R., Anderson, A. D., Nielsen, D. M., and Hill, W. G. (2005). Measures of human population structure show heterogeneity among genomic regions. *Genome Res.*, 15(11):1468–1476.

Weir, B. S. and Hill, W. G. (2002). Estimating F-statistics. *Annu. Rev. Genet.*, 36:721–750.

Wright, S. (1951). The genetical structure of populations. *Ann Eugen*, 15:323–354.

Wright, S. I., Bi, I. V., Schroeder, S. G., Yamasaki, M., Doebley, J. F., McMullen, M. D., and Gaut, B. S. (2005). The effects of artificial selection on the maize genome. *Science*, 308(5726):1310–1314.

Index

Detection of positive selection from multi population samples using dense genome wide data: new multipoint methods and application to farm animal species.

Since initial domestication by Humans, farm animal species have experienced great phenotype diversification and thus represent an interesting model for the study of natural and artificial selection. Besides, the detection of selection signatures in these species can have substantial agronomic outcomes, pointing out genomic regions related to production traits or resistance to diseases.

Many datasets giving access to genome wide genotypic information have become available enabling to scan entire genomes for signatures of selection. Many tests designed to detect positive selection are challenged by at least one of the following problems. First, as the number of available markers increases, so do their correlations which need to be taken into account. Second, many tests have been designed to compare populations only pairwise. Considering more than two populations simultaneously should increase the power to detect selected regions, but necessitates to account for correlations between them arising from their shared history.

I proposed two statistical tests for the detection of positive selection signatures using dense genetic data collected from multiple populations. One is based on haplotypic differentiation between populations, requiring genetic data at the individual level. The second one consists in cumulating single marker signals using local score theory, requiring data at the population level. Through simulated and real datasets, I showed that these tests increased the detection power compared to other existing tests in many cases. Applied to two data sets in sheep and quail, they also point out biologically relevant candidate genes under selection.

FARIELLO RICO, Maria Inés

Détection pan génomique de locus sous sélection en présence de données multi-populationnelles en marquage dense : nouvelles méthodes multipoint et applications aux espèces animales d'élevage

SAN CRISTOBAL Magali, BOITARD Simon, NAYA Hugo
Toulouse, 26/09/2013

Depuis leur domestication, les animaux de ferme ont montré une grande diversification phénotypique. Ils représentent ainsi un modèle pour l'étude de la sélection. De plus, la détection des traces de sélection dans ce type d'espèce peut donner des résultats importants pour l'agronomie, en identifiant des régions du génome associées aux caractères agronomiques ou à la résistance aux maladies.

Des données donnant accès à l'information génotypique de populations permettent d'effectuer des études de détection de traces de sélection pan génomiques. Dans ce contexte, les tests de détection de sélection existants doivent relever deux nouveaux défis : avec le nombre croissant de marqueurs typés, la corrélation entre eux augmente, ce qui doit être pris en compte. D'autre part, les tests utilisés se basent sur la comparaison de deux populations. Considérer plus de deux populations devrait permettre d'augmenter la puissance de détection, mais nécessite que les corrélations entre les populations générées par leur évolution conjointe soient prises en compte.

J'ai proposé deux tests statistiques pour détecter la sélection en utilisant des données génétiques denses collectées dans plusieurs populations. Le premier est basé sur la différentiation haplotypique entre populations et utilise des données individuelles. Le deuxième cumule des signaux de tests simple marqueur en utilisant la théorie du score local et ne nécessite que des données populationnelles. Par simulations et application à des données réelles, j'ai montré que la puissance de détection augmente par rapport à d'autres tests. L'analyse de jeux de données chez le mouton et la caille permet de proposer des gènes candidats.

Détection de sélection – mouton – caille – génétique des populations - scan génomique

ED SEVAB : Ecologie, biodiversité et évolution

UMR444 Laboratoire de Génétique Cellulaire

24 chemin de Bordes-Rouge 31326 Auzeville, France