A model-based approach to characterize individual inbreeding at both global and local genomic scales

T. Druet* and M. Gautier§,†

*Unit of Animal Genomics, GIGA-R & Faculty of Veterinary Medicine, University of Liège, Quartier Hôpital, Avenue de l'Hôpital, 11, B-4000 Liège, Belgium

§ INRA, UMR CBGP (INRA – IRD – Cirad – Montpellier SupAgro), Campus international de Baillarguet, CS 30016, F-34988 Montferrier-sur-Lez, France

† Institut de Biologie Computationnelle, 95 rue de la Galera, 34095 Montpellier, France

Corresponding author:

Tom Druet

Unit of Animal Genomics, GIGA (B34 +1), Quartier Hôpital, Avenue de l'Hôpital, 11, B-4000 Liège, Belgium

Tel: +3243669172; Fax:+3243664151

E-mail: tom.druet@ulg.ac.be

Keywords: Inbreeding; Runs of Homozygosity (ROH); Hidden Markov Models (HMM); identity-by-descent (IBD); homozygosity-by-descent (HBD)

Running title: Individual genomic inbreeding characterization

Abstract

Inbreeding results from the mating of related individuals and may be associated with reduced fitness because it brings together deleterious variants in one individual. In general, inbreeding is estimated with respect to an arbitrary base population consisting of ancestors that are assumed unrelated. We herein propose a model-based approach to estimate and characterize individual inbreeding at both global and local genomic scales by assuming the individual genome is a mosaic of HBD (Homozygous-By-Descent) and non-HBD segments. The HBD segments may originate from ancestors tracing back to different periods in the past defining distinct age-related classes. The lengths of the HBD segments are exponentially distributed with class-specific parameters reflecting that inbreeding of older origin generates on average shorter stretches of observed homozygous markers. The model is implemented in a hidden Markov model framework that uses marker allele frequencies, genetic distances, genotyping error rates and the sequences of observed genotypes. Note that genotyping errors, low-fold sequencing or genotype-by-sequencing data are easily accommodated under this framework. Based on simulations under the inference model, we show that the genome-wide inbreeding coefficients and the parameters of the model are accurately estimated. In addition, when several inbreeding classes are simulated, the model captures them if their ages are sufficiently different. Complementary analyses, either on data sets simulated under more realistic models or on human, dog and sheep real data, illustrate the range of applications of the approach and how it can reveal recent demographic histories among populations (e.g., very recent bottlenecks or founder effects). The method also allows to clearly identify individuals resulting from extreme consanguineous matings.

Introduction

With his pioneering work on self-fertilization, Darwin early noticed that mating relatives generally leads to offspring with a reduced fitness (Darwin, 1876). This phenomenon now referred to as inbreeding depression may mostly result from an increased homozygosity for (recessive) deleterious variants although a lack of heterozygosity at loci displaying heterozygous advantage (overdominance) might also be involved (Charlesworth & Willis, 2009). Accordingly, populations displaying high levels of individual inbreeding show a higher prevalence of monogenic disorders (e.g., Charlier *et al*, 2008) or complex diseases (e.g., Rudan *et al*, 2003). Inbreeding depression can thus increase the risk of extinction by reducing the population growth rate (Hedrick & Kalinowski, 2000; Keller & Waller, 2002) although it may be conversely favorable in some conditions by purging deleterious variants from the population (Caballero *et al*, 2017; Estoup *et al*, 2016). Assessing individual inbreeding is then of paramount interest to improve the management of populations under conservation or selection, and from a more general evolutionary perspective to better understand the genetic architecture of inbreeding depression.

The first standard measure for the level of individual inbreeding was introduced by Wright (1922) as the coefficient of inbreeding (F) that he defined in terms of correlations between the parents uniting gametes. Further, Malécot (1948) proposed an alternative and more intuitive probabilistic interpretation of F as the probability that any two genes each randomly sampled in the parents gametes are identical by descent (IBD), i.e., are themselves derived from a common ancestor. In practice, estimation of F has long been only feasible using pedigree data and was hence limited to a few populations where such information had been recorded. Nevertheless, pedigrees remain usually limited to a few past generations leading to downward bias in the estimates of F since remote relationships are ignored (Keller *et al*, 2011), and they might also contain a non negligible proportion of errors even in well recorded domestic breeds (Leroy *et al*, 2012). In addition, whatever the pedigree depth and accuracy, pedigree-based estimates of F are only providing the expected proportion of individual genomic inbreeding which might departs from the actual genomic inbreeding due to mendelian sampling and linkage (Hill & Weir, 2011). With the advent of next generation sequencing and genotyping technologies, using genomic information to estimate the (realized) individual inbreeding proved particularly valuable (Wang, 2016) opening new avenues in the study of inbreeding in a wider range of populations including wild ones since genealogy is no more required (Hedrick & Garcia-Dorado, 2016; Kardos *et al*, 2016).

Genomic approaches to estimate F basically rely on the identity by state (IBS) status of genotyped markers and may be divided in two broad categories depending on whether or not they use linkage map information.

The first type of methods ranges from simple estimates of individual heterozygosities (e.g., Szulkin *et al*, 2010) or homozygosities (e.g., Bjelland *et al*, 2013) to more advanced approaches based on the estimation of the realized genomic relationship matrix (VanRaden, 2008; Yang *et al*, 2010) or moment-based estimators to correct for population-structure in the estimation of population allele frequencies (e.g., Manichaikul *et al*, 2010). Their accuracy depends strongly on the number and informativeness of the genotyped markers (Kardos *et al*, 2015) but they always remain global in the sense that they can only capture the total amount of individual inbreeding. With genetic map information, one may alternatively rely on the identification of stretches of homozygous markers also referred to Runs of Homozygosity (RoH) (e.g., McQuillan *et al*, 2008) to estimate individual inbreeding at both a local genome scale and genome-wide (as the proportion of the genome contained in locally inbred regions). RoH are indeed most often interpreted as homozygous by descent (HBD) or autozygous segments, i.e., made up of pairs of haplotypes that were inherited from a common ancestor without recombination (and mutation) in neither of them via two different genealogical paths. Assessing the distribution of RoH within individual genomes has thus become popular to characterize inbreeding in a wide range of model species including humans (Kirin *et al*, 2010; McQuillan *et al*, 2008; Pemberton *et al*, 2012), livestock (Bosse *et al*, 2012; Ferencakovic *et al*, 2013) or wild populations (Kardos *et al*, 2017). RoH also allows to distinguish between recent and more ancient inbreeding (Kirin *et al*, 2010; Pemberton *et al*, 2012; Purfield *et al*, 2012) since HBD segments tracing back to more remote ancestors are expected to be shorter because of a higher number of historical recombination events (Thompson, 2013).

Several approaches have been proposed to identify HBD segments from stretches of homozygous markers. First, empirical rule-based procedures aim at characterizing RoH over the genomes (as proxies for HBD segments) and thus rely on the prior definition of specific thresholds for their minimal number of homozygous markers and segment length together with the maximum proportion of allowed heterozygous markers (to allow for genotyping error). Broman & Weber (1999) proposed a formal statistical approach to assess the actual HBD status of the RoH they identified by accounting for population allele frequencies and genotyping error rates. Elaborating on this earlier work, likelihood-based approaches were further developed allowing in particular to compute a LOD-score to assess the strength of evidence in favor of autozygosity of genomic windows through the genome, the size of the window being previously optimized (e.g., Kardos *et al*, 2017; Pemberton *et al*, 2012; Wang *et al*, 2009). Alongside these window-based approaches, Leutenegger *et al* (2003) provided a full probabilistic modeling of the IBD process along the chromosomes by developing a Hidden Markov Model (HMM) to identify HBD segments. Such a HMM framework allows to make use efficiently of the available genetic information contained in the sequences

of both homozygous and heterozygous markers, and the linkage maps. It can also easily handle whole-genome sequence data (Narasimhan *et al*, 2016) including those obtained from low-fold sequencing experiments (Vieira *et al*, 2016). Although powerful, these full model-based approaches all rely on a two-states HMM considering that each marker either belongs to an non-HBD or an HBD segment. The transition probabilities between these two (hidden) states of successive markers then depend on i) their given genetic distances; ii) a parameter controlling the rate of changes per unit of genetic distance; and iii) the individual inbreeding coefficient. Considering only two states (HBD or non-HBD) actually amounts to assuming that all the HBD segments within a given individual have the same expected length. In other words, all the individual inbreeding is assumed to originate from one or several ancestors living in a single generation in the past (with genealogical paths of equal length). However, in both natural and domesticated populations, the sources of individual inbreeding are multiple, since they are all related to their usually complex past demographic history, making such a hypothesis of a single inbreeding event highly unrealistic. As a result, all individuals carry HBD chromosome segments from ancestors across a wide range of numbers of generations into the past (with genealogical paths of varying number of generations). Such HBD segments of different origins should be modeled with different transition probabilities.

We herein propose to extend previous two-states HMM by considering several classes for HBD segments. For each HBD-class, the length of HBD segments (in Morgan) is assumed exponentially distributed with a distinct rate that is related to the age of the inbreeding event (the higher the rate, the shorter the HBD segments and the older the inbreeding event). This new model that actually corresponds to an exponential mixture model allows to provide a better fit to individual genetic data (either genotyping or sequencing data) and to refine the genomic partitioning of inbreeding into stretches of HBD segments from possibly different ancestral origins. To evaluate the accuracy of the methods, we carried out comprehensive simulation studies. In addition, three real data sets from human, dog and sheep populations were analyzed in more detail to illustrate the range of application of the methods. As a by-product of this study, a freely available program, named ZooRoH was developed to implement inferences under the model.

The Models

In the following we describe our HMM to model individual genomes as mixtures of HBD and non-HBD segments. We first consider a model with only two states (one HBD or autozygous class and one non-HBD class) and then describe the extension of the model to combine several HBD classes with varying expected HBD segment lengths.

To deal with the specificities of Next-Generation Sequencing (NGS) data (whole genome sequencing, low-fold sequencing, genotype-by-sequencing) that may provide less accurate genotype call than SNP chip arrays, we also propose alternative emission probabilities functions that integrate over the uncertainties of each possible genotype.

As in previous similar studies (e.g., Leutenegger *et al*, 2003), it should be noticed that the genetic map is assumed to be known without error in the HMM specification. The model further relies on a one order Markov process to define the transition probabilities between successive hidden states. Such a model has been shown to represent a good approximation of the HBD process along the genome when there is no interference between recombination locations (Lander & Green, 1987; Leutenegger *et al*, 2003; Thompson, 2008).

The two–states model (1R model)

We start by describing a simplified HMM that models the transmission of chromosomes from ancestors present G generations in the past to an individual from the current generation (each having 2^G possible ancestors: two parents, four grand-parents, etc.). The paternal and maternal chromosomes of the individual each descend from a distinct set of $N_H = 2^G$ ancestor haplotypes and can hence be described as a mosaic of these haplotypes. To describe this process, we can follow Mott *et al* (2000) that proposed a HMM to model chromosomes of terminal lines as a mosaic of founder lines. The probability to descend from a given ancestor haplotype at the marker position M_{l-1} is $\frac{1}{N_H}$ and the number of recombinations on the path from the ancestors to the individual between two adjacent markers M_{l-1} and M_l separated by t_l Morgans ($l > 1$) is distributed as a Poisson random variable with mean Gt_l (Mott *et al*, 2000). In the context of HBD modeling, we are interested in the pair of inherited haplotypes and their IBD relationship (they either form HBD or non-HBD segments). In total, there are ${N_H}^2$ possible pairs of ancestor haplotypes and the number of recombinations on both paths between the two adjacent markers M_{l-1} and M_l is distributed as a Poisson random variable with mean $2Gt_l$. This means that in the current generation, the length of a diploid segment inherited by an individual without ancestry change (i.e., without recombination in both genealogical paths to the ancestor(s) living G generations ago) is exponentially distributed with rate $R = 2G$ (i.e., with expected mean equal to $\frac{1}{R}$ Morgans). R will be referred to as the rate of ancestry change in our model. Under this model, for a given (diploid) individual and chromosome, the maternally and paternally inherited haplotypes each consist of a mosaic of segments originating from a distinct set of N_H ancestor haplotypes that defines in turn a mosaic of either HBD (where maternally and paternally haplotype segments are IBD) or non-HBD segments. Over the whole individual genome, the proportion ρ of inherited haplotype pairs that are IBD is closely related to the individual inbreeding F defined as the probability that two genes randomly sampled in the paternal and maternal

gametes are IBD (i.e., that a randomly chosen position in the genome belongs to an HBD segment).

Capitalizing on these definitions, the 1R model now assumes that the genome is partitioned in either HBD and non-HBD tracts that actually correspond to the two hidden states ($K = 2$) of the HMM. Let S_l denote the (hidden) state of M_l with $S_l = 1$ and $S_l = K = 2$ if M_l lies within an HBD and a non-HBD segment respectively. The four transition probabilities between the hidden states of every pairs of consecutive markers are then defined as:

$$\begin{cases} \mathbb{P}\left[S_l = 1 \mid S_{l-1} = 1\right] & = e^{-Rt_l} + (1 - e^{-Rt_l})\rho \\ \mathbb{P}\left[S_l = 1 \mid S_{l-1} = 2\right] & = (1 - e^{-Rt_l})\rho \\ \mathbb{P}\left[S_l = 2 \mid S_{l-1} = 2\right] & = e^{-Rt_l} + (1 - e^{-Rt_l})(1 - \rho) \\ \mathbb{P}\left[S_l = 2 \mid S_{l-1} = 1\right] & = (1 - e^{-Rt_l})(1 - \rho) \end{cases} \quad (1)$$

The term e^{-Rt_l} represents the probability that there is no recombination on both genealogical paths between two consecutive markers M_{l-1} and M_l (i.e., the HBD status remains the same). Similarly, $1 - e^{-Rt_l}$ is the probability that the pair of inherited haplotypes changes between the two consecutive markers (as a result of recombination). In that case, the new pair of inherited haplotypes is either HBD (with probability ρ) or non-HBD (with probability $1 - \rho$) irrespective of the previous state. Because consecutive pairs of inherited haplotypes might belong to the same state (with probability ρ and $1-\rho$), the overall lengths of tracts of consecutive markers belonging to the HBD or to the non-HBD class have expected means equal to $\frac{1}{R(1-\rho)}$ and $\frac{1}{R\rho}$, respectively. This model is an approximation of the inheritance of HBD segments and real pedigrees are far more complex. In particular, transition probabilities are not so simple and depend on the position in the genealogy of the haplotypes inherited at marker M_{l-1} (e.g., Druet & Farnir, 2011). Consequently, R is not strictly identical to the size (in generations) of the inbreeding loop connecting the two haplotypes of a HBD segment (approximately equal to $2G$ for an ancestor living G generations ago).

The proposed transition probabilities are identical to those used by Leutenegger *et al* (2003) and Vieira *et al* (2016). Leutenegger *et al* (2003) showed that this HMM is a good approximation of the HBD process and that ρ can actually be interpreted as a measure of the individual inbreeding coefficient F (Leutenegger *et al*, 2003). It corresponds indeed to the marginal equilibrium HBD probability (Thompson, 2008). In these studies, the transition rate R determines the rate of change between the two states in units of genetic distance (Thompson, 2008) and is such that mean length of HBD and non-HBD segments are equal to $\frac{1}{R(1-\rho)}$ and $\frac{1}{R\rho}$, respectively (Leutenegger *et al*, 2003). Although this rate depends on time to common ancestor(s) (Vieira *et al*, 2016), it is not equal to

the generational age of HBD as illustrated by Leutenegger *et al* (2003) and Leutenegger *et al* (2011) for a few examples.

Extension to multi-states models (KR models)

With a unique HBD class, the 1R model described above considers that all the HBD segments have approximately the same age either because they originate from a single ancestor (one strong inbreeding event) or from multiple ancestors in the same generation (e.g., during a bottleneck). Population history might however lead to far more complex patterns and common ancestors tracing back to different generations are probably present in all finite populations (e.g., Kardos *et al*, 2017). This is probably frequent in small populations, in populations under strong selection or in endangered populations with declining size. We therefore propose to extend the model to K_{HBD} different HBD classes, each characterized by their own mixing coefficient ρ_c and rate R_c ($c \in (1, K_{\text{IBD}})$). For a given state c, the HBD segment length (in Morgan) is assumed exponentially distributed with a mean equal to $\frac{1}{R_c(1-\rho_c)}$. Hence, larger values of R_c are associated with smaller HBD tracks which might be interpreted as more ancient inbreeding events coming from more remote ancestors. For a constant mixing coefficient ρ_c, doubling the rate R_c of the HBD-class amounts to halve the expected HBD segment length (corresponding to approximately two times more generations of recombinations). As mentioned above, because the rates of HBD states (R_c) are related (but not equal) to the length of the inbreeding loop (in generations), this extension to multiple HBD states can be considered as a qualitative age-related classification of HBD segments.

For the sake of generality, we may include several non-HBD classes but in the present study we only used one non-HBD class labeled K (i.e., the total number of classes $K = K_{\text{HBD}} + 1$) with a mixing proportion ρ_K and a change rate R_K. The transition probabilities between the hidden states S_{l-1} and S_l of two adjacent loci M_{l-1} and M_l read:

$$\mathbb{P}\left[S_l = a \mid S_{l-1} = b\right] = \begin{cases} e^{-R_a t_l} + (1 - e^{-R_a t_l})\rho_a & \text{if } a = b \\ (1 - e^{-R_b t_l})\rho_a & \text{if } a \neq b \end{cases} \tag{2}$$

where $a \in (1, K)$ and $b \in (1, K)$ represents the identifier of the K different states (recalling that K also represents the non-HBD state). It is important to note that when $K = 2$, i.e. we only consider two states ($K_{\text{HBD}} = 1$ state and one non-HBD), the 2R model is slightly different than the 1R model since the two states are not constrained to have the same rate R.

Emission probabilities and extension to NGS data.

To complete the specification of the HMM we need to specify the emission probabilities, i.e., the probabilities of the data Y_l observed at each marker M_l given the underlying state S_l of the segment that might either be HBD ($S_l \neq K$) or non-HBD ($S_l = K$). Let I_l represent the number of alleles observed for marker M_l (in the rest of the study we only considered bi-allelic SNPs i.e., $I_l = 2$ for all l) and A_{li} the corresponding alleles ($i \in (1, I_l)$). Depending on the technology and the analyses performed, Y_l then either consists of i) a genotype $A_{li}A_{lj}$ (where $i \in (1, I_l)$ and $j \in (1, I_l)$) among the $J_l = \frac{I_l(I_l+1)}{2}$ possible genotypes; or ii) a vector of likelihoods $\mathbb{P}\left[Y_l \mid A_{li}A_{lj}\right]$ for each possible genotype as provided by a genotype calling model as implemented within standard and popular softwares such as `GATK` (McKenna *et al*, 2010) or `SAMTOOLS` (Li *et al*, 2009). This allows to account for the genotype uncertainty which is highly recommended when dealing with NGS, particularly with low-fold sequencing data.

Emission probabilities for genotyping data.

Let p_{li} be the population allele frequency of allele A_{li} which is assumed to be known. If M_l belongs to a HBD segment ($S_l \neq K$), we define the emission probabilities of the genotype $A_{li}A_{lj}$ as follows:

$$\mathbb{P}\left[A_{li}A_{lj} \mid S_l \neq K, p_{li}, \epsilon\right] = \begin{cases} (1-\epsilon)p_{li} & \text{if } i = j \\ \frac{2\epsilon}{I_l(I_l-1)} & \text{if } i \neq j \end{cases} \tag{3}$$

where ϵ is the probability (assumed to be known) to observe a heterozygous marker when M_l belongs to a HBD segment either resulting from a genotyping error or a recent mutation. In other words, we assume that the vast majority of the polymorphic markers were segregating in the population before the common ancestors of the HBD segments and thus interpret recent mutations as genotyping errors. For non-HBD segments (tracing back to much more ancient ancestors), each genotype emission probabilities are derived assuming Hardy-Weinberg equilibrium (HWE) and disregarding genotyping error (or mutation):

$$\mathbb{P}\left[A_{li}A_{lj} \mid S_l = K, p_{li}, p_{lj}\right] = \begin{cases} p_{li}^2 & \text{if } i = j \\ 2p_{li}p_{lj} & \text{if } i \neq j \end{cases} \tag{4}$$

Note that these emission probabilities slightly differ from those considered in Leutenegger *et al* (2003).

Emission probabilities for genotype likelihood data.

To account for genotype uncertainty, emission probabilities are obtained by integrating over all the possible genotypes:

$$\begin{cases} \mathbb{P}\left[Y_l \mid S_l \neq K\right] & = \sum\limits_{J_l} \mathbb{P}\left[Y_l \mid A_{li}A_{lj}\right] \mathbb{P}\left[A_{li}A_{lj} \mid S_l \neq K\right] \\ \mathbb{P}\left[Y_l \mid S_l = K\right] & = \sum\limits_{J_l} \mathbb{P}\left[Y_l \mid A_{li}A_{lj}\right] \mathbb{P}\left[A_{li}A_{lj} \mid S_l = K\right] \end{cases} \tag{5}$$

where $\mathbb{P}\left[A_{li}A_{lj} \mid S_l \neq K\right]$ and $\mathbb{P}\left[A_{li}A_{lj} \mid S_l = K\right]$ are as defined in equation 3 above (the error term ϵ then mostly capturing the effect of recent mutations). This modeling is similar to that recently proposed by Vieira *et al* (2016).

Materials and Methods

Inference

Estimation of model parameters.

Assuming the population allele frequencies (p_{li}) of each marker M_l and the error term ϵ are known, the set of parameters Θ that needs to be estimated for the defined HBD and non-HBD classes consists of their mixing proportions ρ and their rates R. Therefore, Θ consists of two parameters (ρ and one rate R) for the 1R model and $2K$ parameters for a multi-classes KR model (with $K_{\text{HBD}} = K - 1$ inbreeding classes). For multiple-HBD models, we alternatively consider reducing the parameter space by pre-defining the rates R_k of the K classes leading to only estimate the K mixing proportions ρ_k (hereafter called MixKR model). For all the models, parameter estimation was achieved with the Expectation-Maximization (EM) algorithm known as the Baum-Welch algorithm that is very popular in the HMM literature (Rabiner, 1989). The program ZooRoH implementing the algorithm for the different models is freely available at `https://github.com/tdruet/ZooRoH`. Unless otherwise stated, model parameters were estimated with 1000 iterations of the EM algorithm and setting $\epsilon = 0.001$ and $\epsilon = 0$ when analyzing real and simulated (without genotyping errors) data sets respectively. Marker allele frequencies were estimated by the program on the analyzed samples.

Estimation of the realized local (locus-specific) inbreeding (ϕ_l).

The Baum-Welch algorithm allows to estimate the local state probabilities that correspond in our case to the K probabilities $\mathbb{P}\left(S_l = c \mid \widehat{\Theta}, \mathbf{Y}\right)$ that the two chromosome segments belong to the HBD class c ($c \in (1, K_{\text{HBD}})$) or to

the non-HBD class ($c = K$) at the marker M_l position given the estimated parameter set $\widehat{\Theta}$ and the observed genetic data $\mathbf{Y}$. These probabilities can be used to estimate both the realized genome-wide (over all the markers) and local (for each and every marker) inbreeding. Indeed, genetic data allows to directly infer the realized IBD status between the maternal and paternal chromosomes from a given individual at each locus in the genome and over the whole genome as opposed to pedigree-based inbreeding estimates that only infer the corresponding expected IBD status. More precisely, the local estimate $\hat{\phi}_l$ of the realized inbreeding at marker M_l is defined as the probability that this marker lies in a HBD segment and may thus be computed by summing over all its local HBD state probabilities (i.e., excluding the non-HBD class):

$$\widehat{\phi_l} = \sum_{c=1}^{K_{\text{HBD}}} \mathbb{P}\left(S_l = c \mid \widehat{\Theta}, \mathbf{Y}\right) \tag{6}$$

Estimation of the realized inbreeding associated with each HBD class ($F_{\text{G}}^{(c)}$) and the genome-wide inbreeding (F_{G}).

As above, the inbreeding $\widehat{F_{\text{G}}^{(\text{c})}}$ associated to HBD class c ($c \in (1, K_{\text{HBD}})$) can be defined as the proportion of the genome belonging to the class c and is estimated as the average of the corresponding local state probabilities over all the L locus:

$$\widehat{F_{\text{G}}^{(\text{c})}} = \frac{1}{L} \sum_{l=1}^{L} \mathbb{P}\left(S_l = c \mid \widehat{\Theta}, \mathbf{Y}\right) \tag{7}$$

Finally, the genome-wide estimate of the realized individual inbreeding $\widehat{F_G}$ is simply the average over the genome of the local estimates obtained for the L markers:

$$\widehat{F_{\text{G}}} = \frac{1}{L} \sum_{l=1}^{L} \widehat{\phi_l} = \sum_{c=1}^{K_{\text{HBD}}} \widehat{F_{\text{G}}^{(\text{c})}} \tag{8}$$

Model assessment.

Because the optimal number of states (K_{HBD} or K) is usually unknown, we may be interested in characterizing, for a given data set, the strength of evidence for alternative number of states. To that end we relied on the Bayesian Information Criterion (**BIC**) which is a standard criterion for model selection among a finite set of models and was computed as:

$$\text{BIC} = -2ln\left(\mathbb{P}\left(\mathbf{Y} \mid \widehat{\Theta}\right)\right) + n_p ln(L) \tag{9}$$

where $\mathbb{P}\left(\mathbf{Y} \mid \widehat{\Theta}\right)$ is the maximum of the likelihood function obtained with the estimated parameters $\widehat{\Theta}$ (computed with the forward algorithm (Rabiner, 1989)), L is the number of markers and n_p is the number of independent parameters, i.e., $n_p = 2K - 1$ for a KR model (with K-1 HBD classes) and $n_p = K - 1$ for a MIXKR model (because the K mixing coefficients are constrained to sum to 1.0).

Simulated data sets

Simulation under the inference model.

The model was first tested by simulating data under the inference model. We simulated genotyping data at bi-allelic markers (SNPs) for 500 individuals considering a genome that consisted of 25 chromosomes of 100 cM length (i.e., 100 Mb length assuming a cM to Mb ratio of 1). The marker density was set to 10, 100 or 1,000 evenly spaced SNPs per cM (i.e., 25,000, 250,000 or 2,500,000 SNPs in total). When simulating data under the 1R inference model, the individual genome is a mosaic of either HBD or non-HBD segments whose length is exponentially distributed with the same rate equals to the simulated R. For each chromosome in turn, we successively generated consecutive segments by sampling their length in the corresponding exponential distribution and randomly declaring them as HBD or non-HBD with a probability ρ and $1 - \rho$ (where ρ represents the simulated mixing coefficients). The process stops when the cumulative length of the simulated segments was greater than 100 cM (the last simulated segment being trimmed to obtain a chromosome length exactly equal to 100 cM). Under the multi-states model with several HBD classes, simulations were performed sequentially with successive waves of inbreeding. We started by simulating the most ancient HBD class with the process described above. Then, each new HBD class was simulated similarly (with its own R_i and ρ_i) except that new inbreeding (HBD) masked previous classes whereas non-HBD segments did not change previously simulated states.

To simulate genotyping data, we first randomly sampled for each SNP the population frequency of an arbitrarily chosen reference allele either i) from an empirical distribution derived from real cattle genotyping SNP assay and WGS data (Figure S1), or ii) from a (U-Shaped) distribution $\beta(0.2, 0.2)$ that mimics NGS data (Figure S1). We further refer to these two different Allele Frequency Spectrum (AFS) as i) array-like AFS and ii) NGS-like AFS respectively. Given the simulated HBD status of the segments on which each SNP lie (see above), we used these sampled allele frequencies to simulate SNP genotypes as described for the emission probabilities above (eqs. 3 and 4) with $\epsilon = 0$ (without genotyping errors). Subsequently, we set either $\epsilon = 0.001$ or $\epsilon = 0.01$ to introduce random genotyping errors (changing one genotype to one of the two other genotypes) and to evaluate the robustness of the

models.

To simulate low-fold sequencing data (50 individuals) we sampled at each marker a number of reads t according to a Poisson distribution with mean λ (the average coverage). For homozygote genotypes (simulated as described above), the t sampled reads always carried the same allele (no sequencing error) and for heterozygotes, we used a binomial distributions (with parameters t and $\frac{1}{2}$) to sample the read counts for the two possible alleles. We then considered for each simulated SNP l, the read counts t_{l1} and t_{l2} observed for each of the two alleles to derive the three genotype likelihoods of the three genotypes $A_{l1}A_{l1}$, $A_{l1}A_{l2}$ and $A_{l2}A_{l2}$ following Li *et al* (2010) (with the per base sequencing error set to 0):

$$\begin{cases} \mathbb{P}\left[Y_l \mid A_{l1}A_{l1}\right] & = 1^{t_{l1}}0^{t_{l2}} \\ \mathbb{P}\left[Y_l \mid A_{l1}A_{l2}\right] & = \left(\frac{1}{2}\right)^{t_{l1}+t_{l2}} \\ \mathbb{P}\left[Y_l \mid A_{l2}A_{l2}\right] & = 1^{t_{l2}}0^{t_{l1}} \end{cases} \tag{10}$$

To assess the impact of variable local recombination rates τ (per Mb) that may typically be disregarded when converting physical to genetic distances with an average genome-wide cM to Mb ratio, we performed simulations where each 100 Mb chromosome (among the 25 simulated ones) was divided into small segments (10,000 of 10 kb or 1,000 of 100 kb) with varying τ values. In a first scenario, τ was set to 0.001, 0.002, 0.005, 0.010, 0.020, 0.050 and 0.100 for a proportion of 0.20, 0.20, 0.20, 0.24, 0.10, 0.04 and 0.02 of the segments (that were randomly assigned to their respective category). In other scenarios, the values of τ were randomly set to 0.001, 0.010 or 0.100 with probability equal to 0.40, 0.56 and 0.04. In all the cases, the value of τ varied over two orders of magnitude (from 0.001 to 0.100) but the overall average genome-wide recombination rate remained equal to 0.01 per Mb (1 Mb corresponding to 1 cM). We used the genetic map to simulate the alternation of HBD and non-HBD segments as described above. Parameters and inbreeding were then estimated using either the physical map (consisting of evenly spaced markers) as an approximation of the genetic map, or the actual genetic map.

Finally, to assess the accuracies of the model estimation, we computed the Mean Absolute Error (MAE) for each parameter α of interest as:

$$MAE(\alpha) = \frac{1}{N}\sum_{n=1}^{N}|\widehat{\alpha}_n - \alpha_n| \tag{11}$$

where N is the number of simulated individuals, $\hat{\alpha}_n$ is the estimated parameter value for individual n and α is the corresponding simulated value.

Simulations under a discrete time Wright-Fisher process.

The inference model we used is based on hypotheses (exponential distribution for HBD segment lengths, HWE in non-HBD states, etc.) commonly used and that have been proven to work well (e.g., Leutenegger *et al*, 2003; Vieira *et al*, 2016). Still, we performed additional simulations relying on population genetics models to obtain simulated data less dependent on these assumptions. To that end we used the program ARGON (Palamara, 2016) that simulates data under a discrete time Wright-Fisher process.

With constant and large effective population size N_e, inbreeding is expected to be low and to be spread over many generations. To concentrate inbreeding in specific age classes we simulated bottlenecks keeping large N_e outside these events to reduce the noise due to inbreeding coming from other generations. In the first scenario $WF1$ (Figure S2), we considered an ancestral population P_0 with a constant haploid effective population size equal to N_{e0}=20,000 that splits in two populations P_1 and P_2 at generation time T_s in the past with respective population sizes N_{e1}=10,000 or 100,000 (according to the scenario) and N_{e2}=10,000. During four generations centered around generation $T_b \ll T_s$ in the past, P_1 experienced a bottleneck with an (haploid) effective population size equal to N_{eb} and recovered its initial size. Population P_2 that always maintains a constant size is actually used to select markers that were also segregating in the ancestral population P_0 (markers segregating at MAF $\geqslant 0.05$ in both populations P_1 and P_2 were kept for further analyses). The different simulation parameters are expected to have various impacts on the distribution of inbreeding. For instance for larger T_s, inbreeding tends to accumulate after the two populations split and selected markers will have an older origin. Similarly, the larger N_{e1}, the less inbreeding is accumulating outside the bottleneck while with smaller N_{eb}, more inbreeding is created during the bottleneck. In total, 50 diploid individuals were simulated in both populations P_1 and P_2 considering a genome that consisted of a single chromosome of 250 cM length (i.e., 250 Mb assuming a cM to Mb ratio of 1). The mutation rate was set to $\mu = 10^{-8}$ and we use the functionalities of ARGON to identify all the HBD segments > 10 kb and to obtain their ages (generation time of the most recent common ancestor).

A second scenario $WF2$ (Figure S3) was also considered for simulations in which similar parameters were used but the bottleneck occurred at generation $T_b = 20$ and N_{e1} was kept constant for subsequent and more recent generations (instead of returning to its initial size as in scenario $WF1$). This scenario with a strong reduction of N_e was aimed at mimicking livestock populations for which inbreeding is expected to be mostly due to ancestors in the most recent generations.

In both scenarios, estimation of inbreeding was performed on the 50 diploid individuals from population P_1 and with a marker density of 100 SNPs per cM.

Human, dog and sheep real data sets

For illustration purposes, we used publicly available genotyping data from *i)* the Human Genome Diversity Panel (HGDP) (Li *et al*, 2008) as downloaded from `ftp://ftp.cephb.fr/hgdp_supp10/Harvard_HGDP-CEPH`; *ii)* the dog LUPA project (Vaysse *et al*, 2011) as downloaded from `http://dogs.genouest.org/SWEEP.dir/Supplemental.html`; and *iii)* the Sheep Diversity panel (Kijas *et al*, 2012) as downloaded from the WIDDE database (Sempere *et al*, 2015). We then used the software PLINK (Purcell *et al*, 2007) to process and filter the genotyping data by removing individuals with a genotyping call rate below 90%. As a result, the final data sets consisted 620,768, 164,064 and 47,365 SNPs in human, dog and sheep respectively. For each specie, we restricted our analysis to a subset of six populations corresponding to *i)* Karitiana (n=13), Pima (n=14), Melanesian (n=11), Papuan (n=17), French (n=28) and Yoruba (n=22) in humans; *ii)* Doberman Pinschers (n=25), Irish Wolfhounds (n=11), Jack Russell Terriers (n=12), English Bulldogs (n=13), Border Terriers (n=25) and Wolves (n=12) for the dog data set; and *iii)* Soay (n=110), Wiltshire (n=23), Dorset Horn (n=21), Milk Lacaune (n=103), Rasa Aragonesa (n=22) and Rambouillet (n=102) in sheep.

Results

Performance of the different models

Analyzing data simulated under the 1R inference model.

We first analyzed individual genomes of 2,500 cM (with a marker density of 10 SNPs per cM) that were simulated under the 1R inference model, i.e., the simplest model. Depending on the two chosen simulation parameters (rate parameter R and mixing proportion ρ), these individual genomes thus consisted of a mosaic of HBD and non-HBD segments (in proportions ρ and $1-\rho$ respectively) that both originated from the same ancestral generation. In total, we analyzed with the 1R, the 2R, the 3R and the 4R models, 500 individuals per simulated scenarios, considering in total 33 different scenarios representatives of a wide range of values for both R (from $R = 2$ to $R = 256$) and ρ (from $\rho = 0.0075$ to $\rho = 0.5$). As mentioned in the Model section above, under the 1R model that was used for these simulations, ρ is highly similar to the realized individual inbreeding F_{G}. Strictly speaking, ρ is the proportion of segments belonging to the HBD class (see Model section) and F_{G} is the proportion of markers lying in HBD segments. The results obtained from the analyses under the 1R model are detailed in Table 1 for 20 different scenarios. In addition, tables S1 and S2 give the results from the analyses under all the four models (1R, 2R, 3R

and 4R) for all the 33 different scenarios.

Overall, estimates of both model parameters ($\widehat{R}$ and $\widehat{\rho}$) and individual inbreeding F_{G} obtained under the 1R model (Table 1 and Table S1) were found virtually unbiased and quite accurate (small MAE) irrespective of the considered scenarios. As expected, the 1R model performed even better when the number of HBD segments was higher and these were longer (smaller R) since more SNPs are available for their identification. For instance, for a given simulated ρ (e.g., $\rho \simeq F_{\text{G}} = 0.100$), the MAE of $\widehat{F_{\text{G}}}$ increased with larger simulated R (e.g., from 1.1×10^{-3} when $R = 16$ to 4.6×10^{-3} when $R = 256$). The performance of the 1R model to estimate local inbreeding (ϕ_l) was further evaluated by computing the corresponding MAE either for all the SNPs ($\widehat{\phi_l}$) or for the SNPs lying within HBD segments only ($\widehat{\phi_{l^{\text{HBD}}}}$) (Table 1 and Table S1). Note that for every simulated SNP l, the actual ϕ_l value is known (i.e., $\phi_l = 0$ or $\phi_l = 1$ if the SNPs is within a non-HBD or a HBD segment respectively). Hence, if the model performs well and all the ϕ_l are accurately estimated (i.e., $\widehat{\phi_l}$ close to 0 or 1 for SNPs within a non-HBD or a HBD segment respectively), the MAE of $\widehat{\phi_l}$ should be close to 0. The MAE of $\widehat{\phi_l}$ are larger than 0 when SNPs lying in non-HBD segments have a non-zero probability to be HBD, or vice versa. Besides, inspecting the $\widehat{\phi_{l^{\text{HBD}}}}$ MAE allows to restrict attention to the prediction accuracy of truly HBD segments. As shown in Table 1, when inbreeding is recent ($R < 32$, .i.e. average length of HBD segments > 3 cM) MAE for both $\widehat{\phi_l}$ and $\widehat{\phi_{l^{\text{HBD}}}}$ are close to 0 indicating that both HBD and non-HBD positions are correctly identified with a high support. Also, at constant level of overall (simulated) inbreeding (e.g., $\rho \simeq F_{\text{G}} = 0.125$) the accuracy decreases with higher value of R (e.g., from 1.0×10^{-2} when $R = 4$ to 2.1×10^{-2} when $R = 8$ for the $\widehat{\phi_{l^{\text{HBD}}}}$ MAE). When considering more ancient (and/or) lower simulated inbreeding values, the $\widehat{\phi_{l^{\text{HBD}}}}$ MAE increased faster than the overall $\widehat{\phi_l}$ MAE. This indicates that there is not enough information (number of SNPs per HBD segments) to confidently classify some positions, in particular those within i) short HBD segments; ii) long stretches of markers homozygous by chance; or iii) segments boundaries. It is however important to notice that the local inbreeding estimates $\widehat{\phi_l}$ always remained very well calibrated, i.e., for any $p \in (0, 1)$, the proportion of SNPs truly lying within HBD segments among the SNPs with $\widehat{\phi_l} \simeq p$ was close to p (Figure S4). Accordingly, and as mentioned above, the global estimators of individual inbreeding (F_{G}) and the model parameters (ρ and R) remained accurate (Table 1).

[Table 1 about here.]

As shown in Table S1, the estimates of R for the HBD class under the 2R model started to be substantially biased for scenario with $R \geq 128$. More interestingly, the performances of the 2R model (Table S1) and both the 3R and 4R models (Table S2) were highly similar to those of the 1R model for the estimation of both genome-wide

(F_G) and local (ϕ_l) individual inbreeding.

Analyzing simulated data with several underlying HBD classes.

We further evaluated the performances of the different models on simulated data sets with more than one class for the underlying HBD segments, i.e. for which inbreeding originated from several sources of different ages and contributions to the overall inbreeding. We detail hereafter the analyses of individual genomes of 2,500 cM (with a marker density of 10 SNPs per cM) that were simulated under the 3R inference model, i.e., assuming two different classes for HBD segments and one non-HBD class. Each simulation scenario was thus defined by rates of HBD classes (R_1 and R_2) and the mixing proportions (ρ_1 and ρ_2) of the two classes of HBD segments. We remind that the simulated mixing proportions (ρ_1 and ρ_2) directly control (and are generally close to) the amount of inbreeding originating from the corresponding HBD class. However, due to the simulation procedure, some segments belonging to the first HBD class (with a more recent origin and a mixing proportion ρ_1) might overlap (and mask) HBD segments belonging to the second one leading to a reduction (by a factor $1 - \rho_1$ on average) of the actual contribution of the latter to the overall inbreeding. As shown in Table 2 for six different scenarios (and Tables S3 and S4 for a total of 23 different scenarios), estimates of the overall individual inbreeding (F_G), of the rates (R_1 and R_2) and of the inbreeding contributions ($F_G^{(1)}$ and $F_G^{(2)}$) for the two HBD classes were close (but slightly biased) to the simulated values providing the differences between the rates of the two HBD classes was large enough (e.g., $R_1/R_2 \geq 16$), i.e., the overlap between the distributions of the HBD segments lengths is reduced. As the difference between the ratio of successive R_i became smaller, all inbreeding tended to concentrate in the first HBD class that had an overestimated rate for small simulated R_1 (Table 2 and Table S3). For instance, for the scenario with $R_1 = 4$ ($\rho_1 = 0.125$) and $R_2 = 16$ ($\rho_1 = 0.100$), $med(\widehat{F_G^{(1)}}) = 0.195$ (*med* standing for median) and $med(\widehat{F_G^{(2)}}) = 0.004$ while $med(\widehat{R_1}) = 7.20$ and $med(\widehat{R_2}) = 391$ across the 500 simulated individuals (Table 2). Strikingly however, the overall individual inbreeding F_G always remained very well estimated with MAE$\leq$ 0.005 for all scenarios (Table 2 and Table S4). Finally, as for the simulations under the 1R model previously considered, accuracy in the estimation of local inbreeding was found to mostly depend on the rates R_1 and R_2 (Table 2 and Table S5), the MAE for both $\widehat{\phi_l}$ and $\widehat{\phi_{l^{\text{HBD}}}}$ lying in a similar range than the one observed previously on data simulated under the 1R model. More precisely, given the relatively sparse SNP density considered, MAE remained accurate (i.e., ≤ 0.05) while $R_1 < R_2 \leq 64$ but started to increase for higher values probably due to the inclusion of smaller HBD segments.

[Table 2 about here.]

To provide insights on the behavior of our model to a misspecification of the underlying number of HBD classes, we also analyzed these data simulated under the 3R model with the 1R, the 2R and the 4R models. As expected, when considering the 1R and 2R models, the estimated rate of the single assumed HBD class was intermediate between the two simulated R_1 and R_2 actual values (Table S3). In agreement with previous findings, the 1R and 2R lead to highly similar estimates except for large R_1 and R_2 for which the estimated R tended to be higher with the 2R than the 1R model (e.g., $med(\widehat{R}) = 181$ and $med(\widehat{R}) = 201$ respectively for the scenario with $R_1 = 128$ and $G_R = 256$). More interestingly, using the 1R and 2R models (i.e., with a single HBD class) to analyze these data resulted in an underestimation of F_{G} for scenarios with a marked differences between R_1 and R_2 (Table S4). Conversely, using an over-parameterized model (such as the 4R model) did not introduce any additional bias compare to the 3R model. For instance, for the scenario with $R_1 = 4$ ($\rho_1 = 0.125$) and $R_2 = 256$ ($\rho_1 = 0.100$) that lead to a median realized inbreeding equal to 0.211 across the 500 simulated individuals, the median estimated inbreeding was equal to 0.162 with both the 1R and 2R models while it was equal to 0.208 and 0.209 with the 3R and 4R models respectively (Table S4). This suggested that the 1R and 2R model failed to capture some inbreeding. Accordingly, when focusing on the estimation of local inbreeding (Table S5), although the 1R and 2R models displayed a lower MAE for $\widehat{\phi_l}$ (i.e., computed over all the SNPs), this was essentially driven by SNPs lying in non-HBD segments. Indeed, both the 3R and 4R resulted in a lower MAE for $\widehat{\phi_{l^{\mathrm{HBD}}}}$ (i.e., computed over SNPs lying within HBD segments) suggesting these models allowed to better capture HBD segments at the expense of a slightly higher misassignment of SNP lying in non-HBD segments.

Overall, similar conclusions about the performance of the models to estimate the simulated parameters could be drawn when considering data sets with more than two underlying HBD classes (see Table S6 for results on data sets simulated and analyzed under the 4R model). It should however be noticed that increasing the number of HBD classes in the model also increased misassignment of HBD segments towards incorrect HBD-classes (Figure S5). In other words, some HBD segments, although correctly identified as HBD, might display a non-zero probability to belong to an incorrect HBD class (most generally a neighboring one). As a result, when increasing the number of simulated HBD classes, higher deviations of the estimated inbreeding rate (R_c) and contribution ($F_{\mathrm{G}}^{(c)}$) of each classes from their actual values could be observed (e.g., Table S6). Nevertheless, for higher ratio between successive simulated class rates, these estimates remained fairly good. Importantly and as shown in previous simulations, the overall individual inbreeding (F_{G}) was accurately estimated in all scenarios and MAE for local inbreeding mostly depended on the length of the HBD segments.

Using a set of K pre-defined HBD-classes (the MIXKR model).

For a given model, instead of estimating the rates R_k of the different HBD classes, an alternative is to use a set of pre-defined age-related classes with fixed R_k and to only estimate the mixing proportions (ρ_k). To illustrate and evaluate this strategy we hereby considered models consisting of 9, 11 or 13 HBD-classes depending on the simulated marker density (see below) and one non-HBD class leading to the so-called MIX10R, MIX12R and MIX14R models according to our nomenclature. For each model, the pre-defined rates of the $K-1$ HBD-classes always ranged from 2 to 2^{K-1} (with $R_k = 2^k$ for each class $k \in (1, K-1)$) while the rate of the unique non-HBD class was the same as the most ancient HBD class (i.e., $R_K = R_{K-1} = 8192$). Application of these MIXKR models to the various data sets previously generated under the 1R, the 3R and the 4R inference models proved highly efficient (Table S7 and S8). For instance and in agreement with above results, the MIX10R model provided accurate estimation of the overall inbreeding F_G (MAE always lower than 0.005 irrespective of the simulated scenarios) but also of the local inbreeding as indicated by MAE's that were always as good as the best alternative model (e.g., compare Table S7 and Table S5). Moreover, such models with pre-defined rates for the HBD classes allowed to provide indications on the actual rates R_k used in simulations. We indeed observed that the estimated inbreeding contributions ($F_G^{(k)}$) for the $K-1$ HBD classes were mainly concentrated in those HBD-classes with pre-defined rates close to the true simulated ones as shown in Figure 1 for a dense SNP data sets (1000 SNPs per cM) analyzed under the MIX14R models and in Figures S6 to S10 for additional simulated data sets with smaller SNP density (either 10 or 100 SNPs per cM) that were analyzed under the MIX10R or MIX12R models.

[Figure 1 about here.]

Model comparisons and selection.

We finally evaluated the BIC criteria to compare the models. When comparing different KR models (from 1R to 6R) applied to various simulation scenarios (ranging from 1 to 4 simulated HBD-distributions), we observed that the BIC criterion tended to support the correct underlying models and never provided support for models with a number of classes K higher than the simulated ones (Tables S9 and S10). Nevertheless, for simulations involving HBD segments from several classes (i.e., simulated under the 3R to 5R inference models), BIC may favor a model with a smaller number of HBD classes than the actual ones when the rates between successive classes are too close, although increasing SNP density improves the BIC resolution (Table S10). It should also be noticed that the BIC criterion never provided a stronger support in favor of the MixKR model (as defined above) when compared to the

6 others models considered (from 1R to 6R), possibly due to its higher number of parameters (e.g., $n_p = 13$ for the Mix14R model against $n_p = 11$ for the 6R model) (Tables S11 and S12). Yet, for simulations with several HBD classes (Table S12), the BIC support was generally higher than for the 1R and 2R models.

Sensitivity of the models to genotyping error, marker informativeness and genetic map inaccuracy

As only partially investigated above, when analyzing data with different SNP density, we expected that SNP information content, both in terms of marker density and genotyping accuracy, might be a key determinant of the resolution of the models. As a matter of expedience, we investigated this further by focusing on the 1R model (for both simulation and analyses) and evaluated the effect of changing the marker density and the SNP informativeness (array-like or NGS-like AFS) on its overall performance. Results confirmed that both the estimation of the rate R and the identification of HBD positions associated to shorter HBD tracks (i.e., older inbreeding events) always improved when increasing marker density and informativeness (Table 3). For instance, when the simulated $R = 256$, the MAE for $\widehat{R}$ (respectively $\widehat{\phi_{l^{\text{HBD}}}}$) dropped from 36.9 (respectively 0.7313) with a marker density of 10 SNPs per cM and a $\beta(0.2, 0.2)$ AFS to 8.06 (respectively 0.1994) with a marker density of 100 SNPs per cM and to 5.79 (respectively 0.0824) if, in addition, AFS was array-like. We also observe a better assignation of HBD segment to the correct HBD class with higher marker density (Figure S5). It is interesting to note that, at least for the range of parameters considered, F_{G} was accurately estimated irrespective of the marker densities and informativeness.

[Table 3 about here.]

We also investigated the sensitivity of the 1R model to the quality of genotyping or sequencing data. As shown in Table S13, when considering genotyping data (analyzed by setting $\epsilon = 0$ for comparison purposes), we found that the presence of genotyping errors (with simulated $\epsilon = 0.01$ or $\epsilon = 0.001$) had little impact on the estimation of F_{G}, moderate effects on the estimation of local inbreeding ϕ_l but estimates of R were strongly affected with an upward bias and an increased MAE. The magnitude of these effects was actually a function of the number of incorrect genotypes per HBD segment that increased the probability of observing heterozygotes and thus to cut the HBD segment into smaller ROH. As a result, the impact of genotyping errors was stronger for more recent inbreeding, at higher marker density and for higher simulated error rate (Table S13). Interestingly, when analyzing the genotyping data with an appropriate error term i.e., setting $\epsilon = 0.01$ (respectively $\epsilon = 0.001$) for data simulated with a genotyping error rate of 0.01 (respectively 0.001), the estimates of R became unbiased (Table S13). The accuracy was similar than without error except in the case of data simulated with $\epsilon = 0.01$ and

higher rate (older inbreeding origin) where MAE remained larger. Note that in these limiting cases (e.g., simulated $G = 256$ and $\epsilon = 0.01$), the performance of the model when increasing SNP density (from 10 to 100 SNP per cM) was improved when including an error term in the analysis but decreased when analyzed without error (Table S13). More generally, including a small genotyping error term in the model ($\epsilon \neq 0$) had little influence in the analysis of data simulated without genotyping errors.

We further evaluated the sensitivity of the 1R model to various confidence levels in genotype calling by simulating data that mimic low-fold sequencing (or GBS) data for which several genotypes may have a non-zero probability. In these cases, read count data were simulated with a higher SNP density than above (1,000 SNP per cM) and variable coverage (from 1 to 10X). For each simulated SNP, the likelihood of the three possible genotypes were derived from the read count data as described in the Material and Methods section. The analyzed data sets then either consisted of i) the actual SNP genotypes (ideal situation) or ii) vectors of genotype likelihoods. As detailed in Table S14, we found that the model performed well in estimating the global parameters R and F_{G} with sequencing data. As expected, the performances improved with higher coverages and were similar than those obtained with the corresponding genotyping data as coverages $\geqslant$ 5X. Lowering sequencing coverages might indeed be viewed as decreasing SNP informativeness thereby leading to less accurate estimates for the different parameters (increased MAE), particularly for simulation in which inbreeding had an older origin (smaller HBD segments). For instance, for simulated $R \geqslant 512$ and 1X coverage, both F_{G} and R were slightly underestimated (and to a lesser extent with 2X coverage) while for $R \leqslant 256$, both global and local (ϕ_l) estimates were accurate even with coverage as low as 1X (Table S14).

We finally evaluated the impact of inaccurate genetic maps (i.e., correct marker order but incorrect genetic distances between markers) on the performances of our model. We first verified that if the all the genetic distances are multiplied by a same constant c, the estimated rate $\hat{R} \simeq \frac{1}{c}R$ (where R is the simulated rate) and the estimated inbreeding proportions remain identical (data not shown). This is expected from equation 1 since R is expressed on a genetic distance scale. In Table S15, we report the estimated rates R and F_{G} in various simulation scenarios in which the genome was divided in blocks of 10 kb (or 100 kb) with recombination rates per unit of physical distance ranging from 0.001 to 0.100 (see Material and Methods). Results indicate that analyzing the data with an inaccurate genetic map (e.g., using the physical map instead of the genetic map when local recombination rate is variable) might introduce a small downward bias in the estimates of R (Table S15). The effect is more pronounced when the simulated R is larger (older inbreeding) and the local recombination rate varies over longer distances (100 kb segments). The overall inbreeding F_{G} was slightly underestimated in the most extreme situations (Table S16).

In general, for more recent inbreeding, the average genetic length of HBD segments is higher and thus less affected by variable local recombination. Indeed, since the larger the HBD segment, the higher the number of (physical) blocks, for large HBD segments, genetic and physical length tend to coincide. Obviously, when the correct genetic maps were used, parameters and overall inbreeding were accurately estimated (Tables S15 and S16), confirming that the model can handle variable local recombination rate when the genetic map is known.

Comparison with other methods of inbreeding estimation

We compared the 1R model with other methods commonly used to estimate inbreeding on a subset of six scenarios previously simulated under the 1R inference model and without genotyping errors. We started by running FEstim (v 1.3.2) that implements the original HMM proposed by Leutenegger et al. (2003) to verify that it is indeed equivalent to our 1R model (Table S17). We regressed estimators obtained by both methods and obtained a perfect match between both estimated mixing proportions ρ and rates R. As expected our estimated rates R were equal to $100a$ (a being the rate estimated by FEstim with a map expressed in cM). Since both methods are identical, comparisons between FEstim and other methods are valid for our model too. For instance, Polasek *et al* (2010) found that FEstim was superior to estimators based on expected genome-wide homozygosity and locus-based homozygosity. Similarly, Narasimhan *et al* (2016) concluded that HMM based models outperformed rule-based ROH as implemented in plink (Purcell *et al*, 2007) or estimates obtained with Beagle (Browning & Browning, 2010). In addition, we computed the estimators based i) on the expected genome-wide homozygosity implemented in plink (Purcell *et al*, 2007); ii) the rule-based ROH (with 20 or 50 per ROH and no heterozygous SNP) and; iii) the likelihood-based ROH (Pemberton *et al*, 2012; Wang *et al*, 2009). The latter approaches compare the (LOD) ratio of the probabilities of the genotype data under hypotheses of autozygosity (HBD) and non-autozygosity (non-HBD) for sliding windows of n SNPs, n being chosen to obtain a clear bimodal distribution of the LOD scores. Here, this was achieved with $n = 60$, as in Pemberton *et al* (2012) and Kardos *et al* (2016), but we also considered windows of $n = 20$ SNPs that worked for most scenarios and allowed to capture smaller ROH. In addition, sliding windows were incremented by 1 SNP (we tested all windows of n snps) and the error term was set to 0.001. When the expected number of SNPs per HBD segment was large enough, all methods performed equally well (Tables S18-S21). Our model was able to identify smaller HBD segments (from more remote ancestors) than window-based approaches with 50 or 60 SNPs and had comparable behaviour with that respect as methods using 20 SNPs windows (it identified slightly less small segments). The 1R model proved the most accurate to estimate F_G, followed by the method based on excess of genome-wide homozygosity, particularly when the

expected number of SNPs per ROH was smaller (Tables S18-S19). In the most extreme case ($R = 256$ and with 10 SNPs per cM), we did not observe a clear bimodal distribution for the likelihood-based approach and could thus not apply the method. When the expected number of SNPs per ROH was limiting, approaches using SNP windows underestimated the number of HBD segments (this was more pronounced with larger SNP windows). As expected smaller SNP windows increased power to detect HBD segments (Tables S20-S21) but false positive rate too (increased MAE(ϕ_l)). In agreement with above results, in such limiting cases, the HMM approach is still able to provide an accurate estimation of the global inbreeding. The estimated probability for local inbreeding were less accurate (high MAE), particularly for SNPs lying in HBD segments (the model can not precisely determine which positions are HBD or not), but still remained well-calibrated.

Simulations under a discrete time Wright-Fisher process

To evaluate the robustness of the model to departure from model assumptions, we analyzed data simulated under a discrete-time Wright-Fisher process using the recently developed program ARGON (Palamara, 2016). For our purposes, a decisive advantage of ARGON is that it allowed to identify all the HBD segments (here we only considered those $\geqslant 0.01$ cM) and to obtain their age (i.e., time to most recent ancestor or TMRCA). Inbreeding was generated by assuming population histories with either i) a strong bottleneck in the recent past followed by a rapid expansion as might be observed in invasive populations (WF1 scenarios) or ii) a reduced effective population size in the last twenty generations as might be observed in some domestic populations (WF2 scenarios). In total we considered 12 different WF1 scenarios and two different WF2 scenarios (see Material and Methods) and analyzed 50 simulated diploid individuals from population P_1 per scenario with a marker density of 100 SNPs per cM. As illustrated in Figure 2A for one WF1 scenario (see Figures S11 and S12 for all the 12 WF1 and the 2 WF2 scenarios respectively), the simulated history leads as expected to an enrichment in HBD segments that trace back to the bottleneck period within the simulated individual genomes (about 20% on average in Figure 2A). Yet, in most scenarios, a substantial proportion of inbreeding was associated to more ancient classes that accumulate inbreeding over many more generations. Indeed, a segment was considered HBD if it traced back to an ancestor from a generation more recent than the split time ($T_s = 10^3$ or $T_s = 10^4$ generations depending on the scenarios) of two modeled populations (see Material and Methods). Accordingly, in WF1 scenarios, this proportion increased with lower effective population size (Ne_1), older split time (T_s) and to a lesser extent higher bottleneck population size (N_{eb}) and timing (T_b) (Figures S11 and S12).

We analyzed all these simulated data sets with a Mix14R model that consisted of 13 HBD-classes with pre-

defined rates ranging from 2 to 8192 (with $R_k = 2^k$ for each class k) and one non-HBD class that had the same rate as the older HBD class (i.e., $R_{14} = R_{13} = 8192$). The choice for a MIXKR model was motivated by our previous findings that demonstrated it was informative regarding the rates of the simulated inbreeding class(es) and performed as well as other models in estimating local and overall inbreeding. In addition, it allowed to compare all the simulated individuals according to the same age-related partitioning of inbreeding.

[Figure 2 about here.]

As shown in Figure 2B (see Figures S13 and S14 for all the 12 WF1 and the 2 WF2 scenarios respectively), our HMM always allowed to efficiently identify HBD segments tracing back to common ancestors with TMRCA smaller than 80 generations, since the underlying SNPs displayed an estimated local inbreeding probability (ϕ_l) close to one. In agreement with results obtained on simulations performed under the inference model (see above), the power to identify HBD segments of older origin gradually decreased (towards values almost always lower than 20% for TMRCA older than 5000 generations). Note that analyses of data sets simulated under the inference model showed that although the power was below one, overall inbreeding remained correctly estimated (see above). Figures S15 and S16 represent the same average local inbreeding probabilities for HBD-segments as a function of their length (instead of TMRCA). Theses probabilities were close to one for HBD-segments longer than 50 Kb, above 0.80 for HBD segments from 20 to 50 Kb long and dropped towards 0 for smaller HBD-segments. It is important to recall that with higher marker densities, it would be have been possible to identify older and smaller HBD segments.

Interestingly, we further observed that the HBD segments tracing back to the simulated bottleneck period were in their vast majority assigned to HBD classes whose pre-defined rates were close to twice the corresponding time (in generations). For instance, in the scenario with a bottleneck lasting from generations 17 to 14 in the past considered in Figure 2, the estimated proportions of the individual genomes assigned to HBD segments were concentrated in the HBD class with pre-defined rates equal to 32 ($R_k = 32$), 16 ($R_k = 16$) and to a lesser extent in an older HBD-class ($R_k \geqslant 2048$) (Figure 2C and Figures S17 and S18 for all the 12 WF1 and the 2 WF2 scenarios respectively). Moreover, in the simulated individuals, the HBD segments with a TMRCA $\approx$ 16 were mainly assigned ($> 70\%$) to the two neighboring HBD classes with $R_k = 32$ and $R_k = 16$ (Figure 2D). Similar patterns were observed in other simulations (Figures S19 and S20). Note that older HBD classes (with $R_k \geqslant 512$) also captured a small proportions of the HBD segments that traced back to the bottleneck period (Figures S19 and S20) together with those with an older TMRCA probably because these older HBD classes have high mixing

coefficients. This effect was stronger when the bottleneck contributed less to the overall inbreeding and when the bottleneck was older. HBD segments from an individual might also be smaller or larger than expected from the age of the bottleneck due to the stochastic nature of the Wright-Fisher process. In all cases however, we observed a peak of inbreeding in the HBD-class(es) with a rate close to twice the age corresponding to the period of reduced N_e or its neighbors (Figures S17 and S18). Finally, the vast majority of the non-HBD segments (with a TMRCA $>$ 10,000 generations) were correctly assigned to the non-HBD class, the remaining ones being assigned to most ancient contributing HBD-class (Figures S21 and S22). Overall, this simulation study thus confirmed that our model correctly identifies HBD segments and it also provided support in favor of an age-based interpretation of the HBD-class rates.

Note that likehood-based ROH methods with windows of 20 or 60 SNPs were also applied to these simulated data sets. The power to identify HBD segments according to the age of the TMRCA or to their length are reported in Figures S13-S16. As for simulations under the inference model, our model had comparable behavior than methods using 20 SNPs windows and identified smaller HBD segments (associated with more remote ancestors) than methods using 60 SNPs windows. The power and false positive rate would largely depend on the definition of an arbitrary base population making comparisons difficult. Indeed, at some time in the past, ancestors must be considered unrelated or all segments would be HBD. One of the benefits of a Mix14R model is to automatically estimate inbreeding relative to several base populations (at different time in the past), making the choice somewhat less arbitrary.

Application to human, dog and sheep real data

We applied our model to individuals from human, dog and sheep populations, i.e., species representative of a wide range of demographic histories. Individuals were genotyped, as part of previous experiments (see Material and Methods) with assays containing various number SNPs (ca. 600K, 150K and 50K for human, dog and sheep individuals respectively) leading to different SNP density (ca., 1 SNP per 5kb, per 20 kb and per 60 kb respectively). The genotyping data were further analyzed with a Mix14R model that consisted of 13 HBD-classes with pre-defined rates ranging from 2 to 8192 (with $R_k = 2^k$ for each class k) and one non-HBD class that had the same rate as the older HBD class (i.e., $R_{14} = R_{13} = 8192$). In all analyses, the estimated mixing proportions of HBD-classes with $R_k \leqslant 256$ were all extremely small (< 0.01) supporting an age-based interpretation of the R_k rates as the length of the inbreeding loop or approximately half the age of the underlying ancestor (both measured in generations). Indeed, the expected lengths of HBD tracks per class were consequently close to $\frac{1}{R_k}$ corresponding to the average

length for HBD segments transmitted by an ancestor living $G \approx 0.5R_k$ generations ago. It should however be stressed that this age-based interpretation is only approximated (see Discussion) and that populations have variable ratio between genetic and physical distances when averaged between sexes: 1.16 cM/Mb for human (Kong *et al*, 2010), 1.26 cM/Mb for sheep (Johnston *et al*, 2016) and 0.88 cM/Mb for dog (Campbell *et al*, 2016). Indeed, we used for the analyses the SNP position on the physical maps accompanying the respective data sets. Differences with real genetic maps together with variable local recombination rates might introduce some imprecisions in the assignment of actual HBD segments to their actual age-related HBD class (see above). The estimated contribution of each pre-defined HBD class (averaged over all the individuals) are detailed for each populations and each species in Figure 3.

[Figure 3 about here.]

Regarding humans, the six populations considered here (French, Yoruba, Melanesian, Papuan, Pima and Karitiana) have already been thoroughly analyzed (e.g., Jakobsson *et al*, 2008; Li *et al*, 2008) and in particular in studies aiming at characterizing inbreeding (Leutenegger *et al*, 2011; Pemberton *et al*, 2012) or providing a detailed assessment of the distribution of ROH of different lengths (Kirin *et al*, 2010). In each population, we observed some individuals with more than 1% of recent inbreeding ($R_k \leqslant 16$) but these were rare in Yoruba (1 out of 22) and French (2 out of 28) populations compared to Pima (12 out of 14) and Karitiana (13 out of 13). In these two latter populations, there is strong evidence for very recent inbreeding, some of the individuals having more than 10% of inbreeding in very young classes from $R_k = 2$ to $R_k = 8$ (Figure 4A and Figure S23). Oceanian populations displayed intermediate proportions of such individuals (1 out of 11 for Melanesian and 4 out of 17 for Papuan) but had higher proportions of inbreeding in intermediate HBD-classes ($32 \leqslant R_k \leqslant 128$) compared to French and Yoruba. Consistently, average cumulated inbreeding at $R_k = 16$ was high for Karitiana (4.1%) and Pima (2.8%) and low for other populations (< 0.5%). When cumulated up to HBD-class with $R_k = 128$, these values were still below 0.5% for Yoruba and French and larger than 1% in Melanesian (2.0%) and Papuan (3.2%) populations. These results are consistent with those reported by Leutenegger *et al* (2011) who concluded that Yoruba and French genotyped individuals were in a vast majority originating from unrelated matings (with the same outliers as in our study), that Melanesian and Papuan were associated to either unrelated or double-cousins (2C) matings (common ancestor 4 generations ago and expected inbreeding equal to 1.56%) and that Pima and Karitiana came from either first cousins (1C) (common ancestor 3 generations ago and expected inbreeding equal to 6.25%) and 2C matings (with two individuals presenting possibilities of avuncular or double 1C mating). With our model, children of un-

related matings presented no trace of recent inbreeding ($R_k \leqslant 16$), those from 2C and 1C mating had respectively 1.2% and 7.5% recent inbreeding (the two most extreme individuals having more than 10% inbreeding). Overall, as shown in Figure S24, the mean estimated inbreeding estimated by Leutenegger *et al* (2011) were highly correlated with our estimate of recent inbreeding (r = 0.945) defined as the sum of the contribution of the first four HBD-classes (from $R_k = 2$ to $R_k = 16$) but less with the overall inbreeding (r = 0.601). It should also be noticed that Leutenegger *et al* (2011) estimated inbreeding using LD-pruned maps of 6,500 SNPs (to get unbiased results) whereas we did not perform any LD-based filtering of the data and used more than 600,000 SNPs to partition inbreeding in the different classes of our model. Yet, in human populations, our results showed that the largest proportion of ROH were associated with the most ancient HBD-classes. Although interpretation of old inbreeding must be done with caution (see Discussion), it might be considered as associated with the background LD in the population and mostly influenced by the demographic characteristics of the populations (e.g., effective population size history). Accordingly, the amount of overall inbreeding increased from Africans to Europeans, Oceanians and Native Americans (from Central and Southern America) (Figure 3A,B). More precisely, the rates of the main contributing HBD-classes that were generally consistent within population were clearly related to their N_e. Hence, inbreeding concentrated in HBD-classes with $R_k = 512$ for Karitiana, with $R_k = 512$ and $R_k = 1024$ for Pima, with $R_k = 1024$ for Papuans and Melanesians, with $R_k = 1024$ for French and with $R_k = 2048$ for Yoruba. These results are qualitatively in agreement with previous findings by Kirin *et al* (2010) that suggested the presence of both recent (long ROH) and ancient (short ROH) inbreeding in Native Americans. Conversely, they found that individuals from Oceanian populations did not display long ROH (several Mb long) but had an excess of ROH of intermediate length (between 1 and 2 Mb) indicating a reduced N_e in the past. Finally, European and African populations mostly showed inbreeding arising from remote ancestors. One major difference between our results and the study by Kirin *et al* (2010) is that they only considered ROH > 500 kb leading to a lower estimated value (most probably downwardly biased) for the overall individual inbreeding. As previously mentioned, the power of all approaches to detect short HBD segments is a function of the available marker density which possibly leads to an underestimation of their proportions.

Modern dog breeds present large amounts of inbreeding and are known to have experienced strong bottlenecks associated with the recent breed creation from a small number of founders (e.g., Vaysse *et al*, 2011). In addition, strong artificial selection and matings in small closed populations further contributed to increase inbreeding in the last decades (Lewis *et al*, 2015). Accordingly, as shown in Figure 3C,D and Figure S25, we observed massive inbreeding (sometimes higher than 20%) in the HBD-class with $R_k = 16$ (a common ancestor approximately

8 generations ago) in all the five breeds we analyzed but the Jack Russell Terrier that has a larger N_e (Vaysse *et al*, 2011). As expected also, wolves that did not experienced domestication did not present such an excess of inbreeding in recent generations. In each population (including wolves), some individuals were found to be highly inbred with an $F_G \approx 50\%$ and approximately 25% of this inbreeding associated with the first two HBD-classes (i.e., a common ancestor living only one or two generations ago) (Figure 4B and Figure S25).

Finally, among the six sheep populations we investigated, three (the Rasa Aragonesa, Milk Lacaune and Rambouillet) displayed a large N_e (> 700) as described in Kijas *et al* (2012). Hence, individuals from the Rasa Aragonesa displayed almost no trace of recent inbreeding ($\leq 0.5\%$ when summing contributions of HBD-classes with $R_k \leq 8$) while the cumulative inbreeding remained lower than 6% on average for individuals from the Milk Lacaune and Rambouillet breeds up to classes $R_k = 32$ (Figure 3E,F and Figure S26). Yet, some Rambouillet individuals presented high levels ($> 20\%$) of recent inbreeding (Figure 4C and Figure S26). Conversely, the Wiltshire ($N_e = 100$) and Dorsethorn ($N_e = 137$) populations that went through a strong reduction in size in the early 1900's (Dorsethorn to a lesser extent) were both found to have a high level of recent inbreeding (Figure 3 and Figure S26). The main contributing HBD-class was the one with rate $R_k = 16$ for Wiltshire and $R_k = 4$ to $R_k = 32$ for Dorsethorn. Interestingly, the Wiltshire individuals were sampled from a New-Zealand flock that experienced several strong and successive bottlenecks in its recent history. Indeed, its founders were imported in 1974 from Australia where the breed had previously been introduced in 1952 and survived as a remnant population of as few as 12 ewes (O'Connell *et al*, 2012). Assuming a generation time of approximately 4 years in sheep, the distribution of the contribution of the most recent classes to the overall inbreeding is thus consistent with this demographic history. The sixth sheep population we investigated was the well known Soay sheep that had an estimated $N_e = 194$ (Kijas *et al*, 2012) and experienced a strong founder effect since the current population derives from a flock of 107 individuals that were transferred on the Hirta island in 1932 and then lived in complete isolation (Clutton-Brock & Pemberton, 2004). We observed for this population a small amount of recent inbreeding (for HBD classes with age $R_k \leq 16$), even lower than in Milk Lacaune or Rambouillet, but rather high levels of inbreeding associated with HBD classes of rates between between 32 and 64 (Figure 3E,F and Figure S26). Integrating over all the classes, the Soay sheep thus appeared on average even more inbred than Dorsethorn, which explains the small estimated N_e. However, despite this strong founder effect and the high resulting inbreeding level, we observed almost no individual with an inbreeding $F_G > 5\%$ in the most recent generations. The Soay breed represents an interesting example of a wild population resulting from a founder effect and in expansion. To summarize, our model allowed to provide deeper insights into the very different patterns of individual inbreeding observable in the

sheep breeds. Indeed, these inbreeding patterns ranged from small as in the Rasa Aragonesa or limited level (with a few overly and recently inbred individuals) as in the Rambouillet breed, to moderate to high inbreeding level that either originated from strong bottleneck in the very recent (Wiltshire) or recent (Soay) past, or that resulted from the cumulative effect of a less pronounced population size reduction over more generations (Dorsethorn).

[Figure 4 about here.]

Importantly, besides providing a global estimator of inbreeding for each individual, the model also informs on the partitioning of this individual inbreeding which is highly valuable. For instance, individuals born from extremely consanguineous marriages might be easily identified. As an illustration, Figure 4B showed three dogs (Dob_LU142, Dob_LU149 and BoT_LU45) that displayed approximately 25% inbreeding associated with the $R_k = 2$ or $R_k = 4$ HBD-class (ancestors living one or two generations ago) unlike other dogs from the same population (Dob_LU154 and BoT_LU70). These three individuals are likely resulting from matings between a sire and its daughter. This indicates that inbreeding is still present in these populations and is not only due to the breed creation event but to further management practices. High level of inbreeding associated to parents or grand-parents are also observed in sheep (19.2% for Rambouillet RMB63 in Figure 4C) and even in human (8.9% for Karitiana HGDP01019 in Figure 4A). For all these individuals, however, these recent events accounts only for a fraction of total inbreeding and a substantial proportion of inbreeding is due to more remote ancestors. More generally, by partitioning the total amount of inbreeding among ancestors from different generations, our model provides a better understanding of the origins of inbreeding in each individual. Hence, individuals with a similar overall inbreeding might display a quite different pattern of ancestral contributions captured by our model. For instance, for the three sheep individuals (Rambouillet RMB70, Wiltshire WIL2 and Soay SOA2172) represented in Figure 4C that all displayed an overall inbreeding of approximately 20%, the inbreeding is mostly associated to the HBD-class $R_k = 16$ for the Wiltshire WIL2, to the two HBD-classes $R_k = 32$ and $R_k = 64$ for the Soay SOA2172 whereas for the Rambouillet RMB70 individual, ancestors contributing to inbreeding trace back to a wide spectrum of generations (from $R_k = 4$ to $R_k = 256$). These observations are consistent with patterns at the population level. Interestingly, individuals with higher levels of inbreeding (Wiltshire WIL21 and Rambouillet RMB63) display comparable patterns with inbreeding concentrated in the HBD-class $R_k = 16$ for Wiltshire WIL21 and associated to several HBD classes for Rambouillet RMB63 (Figure 4C). In humans (Figure 4A), Native Americans from Central and Southern America were found to display different make-ups than Oceanians with similar levels of overall inbreeding (e.g., Karitiana HGDP01010 vs Melanesian HGDP01027 or Pima HGDP01044 vs Papuan HGDP00555). As expected

from previous results, Oceanians actually displayed little traces of very recent inbreeding but accumulated more inbreeding in distant generations.

Computational requirements

To assess the computational performances of our software, we ran ZooRoH on a cluster with Intel E5649 processors at 2.53 GHz to estimate inbreeding in populations of 500 individuals genotyped at 10 or 100 SNPs per cM with different models (1R, 4R and MIxKR). In total, 1000 iterations of the EM algorithm were realized. Running times range from less than 3 hours to process all 500 individuals genotyped with 25,000 SNPs under a 1R model to more than a day to process 50 individuals genotyped with 250,000 SNPs under a MIx12R (Table S22). Memory usage remained reasonable (below 200 MB) whereas running times were a function of the number of fitted classes and the marker density (e.g., 10 times slower to process an individual with 10 times more markers). We are currently working on a package working with optimization procedures (to reduce the number of iterations) and including parallelization of the analysis over individuals.

Discussion

In this study, we developed and evaluated HMM models that use genomic data to estimate and to partition individual inbreeding into classes of HBD segments with different lengths which might in turn be interpreted as originating from ancestors of different ages. There actually exists a wide variety of methods to estimate individual inbreeding and these have different properties. Pedigree-based methods rely on a genealogy (the inbreeding can only result from individuals within the genealogy) and predict the expected IBD status at a locus whereas genomic measures estimate realized inbreeding (the observed level of inbreeding) (Hill & Weir, 2011; Kardos *et al*, 2015, 2016). Genomic estimates can either be global, giving a unique measure per individual, or local. Obviously, these latter measures provide more information but require a higher marker density. Assessing the distribution of ROH within individual genome have recently become popular to characterize global and local inbreeding (Kirin *et al*, 2010; McQuillan *et al*, 2008; Pemberton *et al*, 2012). When definition of RoH is rule-based, many parameters must be defined and these need to be adapted to the characteristics of the population under study and the genotyping technology used. Alternatively, likelihood-based RoH classification (Broman & Weber, 1999; Pemberton *et al*, 2012; Wang *et al*, 2009) or HMM modeling (e.g., Leutenegger *et al*, 2003) make a better use of all the information since they take into account the marker allele frequencies and the genotyping error rates. Relying on

a full probabilistic HMM framework has several additional advantages. First it allows to directly account for the (genetic) map information. Second, as we showed in our study, HMM can be extended to account for uncertainties associated with NGS data (Narasimhan *et al*, 2016), including low-fold sequencing (Vieira *et al*, 2016) or GBS, whereas rule-based ROH are inappropriate in such conditions. Finally, when relying on the Forward-Backward algorithm (as in our study), HMM allows to integrate over all the available information to estimate the HBD probabilities at each marker position in opposition to a binary classification as obtained with window-based approaches or HMM methods that rely on the Viterbi algorithm (Narasimhan *et al*, 2016; Vieira *et al*, 2016). Overall, using a probabilistic model is particularly valuable when information is sparser and classification is more uncertain (e.g., for smaller and older HBD tracts, at lower marker density or informativeness, with higher genotyping error rates or with low-fold sequencing).

The most simple HMM we considered consists of a single HBD state (1R model) and is similar to several previously proposed ones (Leutenegger *et al*, 2003; Narasimhan *et al*, 2016; Vieira *et al*, 2016). This amounts to either assume that a single common ancestor is responsible for inbreeding or that the vast majority of HBD segments trace back to ancestors that lived in the same past generation. However, most populations have complex demographic histories, with varying N_e and common ancestors of HBD segments are thus expected to originate from many different generations in the past. As shown by our application in real data sets, even in domestic populations for which inbreeding might be expected to result from a limited number of founder individuals, individual inbreeding generally results from ancestors in different generations back in time probably due to the subsequent intense use of some key (selected) breeders. Hence, extending the model to several HBD-classes is highly valuable in such cases. The first benefit of a multiple HBD-classes model is to better fit the data and to obtain more accurate estimators of inbreeding both locally and globally. Indeed, our simulations under the inference model with several HBD classes clearly showed that the 1R (and 2R) model underestimated F_{G} as some HBD segments were missed while the power to detect HBD segments was decreased. In addition, in the presence of ancient inbreeding, the 1R model will tend to interpret recent (and thus longer) HBD segments as consecutive smaller segments of older origins because the estimated rate of the single HBD class would tend to be larger. Of course, in the absence of genotyping errors, the entire segment would then be correctly declared HBD and would appear as a long tract. However, at higher genotyping error rates (as with NGS data) such segments would be cut into smaller pieces. This would not happen when analyzing data with a model with multiple classes since recent HBD segments would then be associated to a class with a smaller rate and the penalty in the HMM to leave the HBD-class and start a new HBD segment would be too large. With two states HMM (Leutenegger *et al*, 2003), LD pruning is sometimes

used to get rid of background LD and then force the model to concentrate only on recent inbreeding. With multiple HBD-classes model ($> 2R$ models), ancient inbreeding associated with background population LD is assigned to the oldest HBD classes making LD pruning unnecessary for that purpose. This was illustrated by comparing inbreeding estimators obtained for human populations with a LD-pruned map or with a non-filtered map with respectively a 1R and Mix14R model. Also, HMM with multiple HBD classes allows to determine whether there is a single or multiple HBD distribution(s) with a major contribution to overall inbreeding. We can then clearly identify individuals from extreme consanguineous matings (sire x daughter, first cousins, etc) because inbreeding due to this recent ancestor is distinguished from the background inbreeding of remote origin (see examples with 25% inbreeding in class $R_k \leqslant 4$ in dog and sheep data analyses). Multi-HBD classes models allow in turn to obtain some information on the relatively recent demographic history of the population, high levels of inbreeding indicating that N_e was reduced at some recent time in the past such as in populations under conservation or invasive populations whereas an absence of inbreeding is indicative of a large N_e during the corresponding period. Application to real populations then demonstrated that the model can capture very different patterns including presence or absence of consanguineous matings, large N_e and low inbreeding, bottlenecks at varying time in the past, founder effects and reduced N_e due to isolation in the past ($R_k \geqslant 100$). Our HMM model actually explores more recent generations and can be considered as complementary to approaches that infer past N_e (Li & Durbin, 2011). It is however not intended to estimate N_e, other methods modeling IBD being better suited to that purpose (e.g., Browning & Browning, 2015).

Using the proposed HMM to obtain information on recent demographic history or to identify extreme consanguineous matings based on the estimated rates of the HBD-classes assumes that there is a link between the rate of the HBD-classes and the age of inbreeding. In our model, the transition rate per Morgan is not equal to the generational age of HBD but these quantities are related. Indeed, the length (in Morgans) of chromosomal segments inherited from ancestors living G generations ago is exponentially distributed with a mean $\frac{1}{G}$ (Thompson, 2013) and $\frac{1}{2G}$ for HBD segments that consist of a pair of IBD haplotypes inherited from the same ancestor ($2G$ representing the size of the inbreeding loop). Unfortunately, the lengths of HBD segments originating from a given ancestor are not directly observed because HBD tracts can be the result of the junction of several HBD segments (possibly inherited from distinct ancestors). If HBD segments inherited from ancestors G generations ago have a probability ω to be followed by another HBD segment inherited from an ancestor of the same age, then the length of the resulting HBD tract would be exponentially distributed with expected length $\frac{1}{2G(1-\omega)}$. In the present model, the length of HBD tracts is expected to be $\frac{1}{R(1-\rho)}$. When the difference between ρ and ω is small, R is approximately equal

to $2G$ and is related to the age of the HBD segments. Factors such as the pedigree structure, the distance between the markers or the size of the inbreeding loop determine the magnitude of this difference. In some specific mating types, ω is almost null (in a first-cousin 1C mating, HBD segments become non-HBD after a single recombination) whereas ρ is equal to the inbreeding coefficient (6.25% for a 1C mating). As an example, Leutenegger et al. (2003; 2011) estimated the expected rate R (named a in their study and expressed according to a genetic map in cM, i.e., $R = 100a$) for a few specific mating types such as 1C (R=6.3), double first cousins 2x1C (R=6.8), second-cousins 2C (R=8.0), avuncular AV (R=5.7) and 4 x 2C (R=8.4) matings. Even if R is different from $2G$ in these cases, both values remain close since the size of inbreeding loops ($2G$) corresponding to these five different mating types are equal to 5 (AV mating), 6 (1C and 2x1C matings) and 8 (2C and 4x2C matings). Simply setting ω to 0 for these matings (assuming HBD states are followed by non-HBD states after a recombination) and setting $\frac{1}{2G(1-\omega)} = \frac{1}{R(1-\rho)}$ would yield very similar estimates for R to those estimated above (respectively 6.4, 6.9, 8.1, 5.7 and 8.5) indicating that for these examples differences between ρ and ω account for a large part of the differences between $2G$ and R. Further using an approach similar to Leutenegger *et al* (2003), we estimated that the expected value of R to be equal to 12.01 and 32.02 for HBD segments originating from a common ancestor living 6 ($2G = 12$) and 16 ($2G = 32$) generations ago. In summary, although the rate R gives at least a qualitative indication and in some simple cases a good estimation of the inbreeding age, it should more generally only be viewed as an approximation of the true size of the inbreeding loop (in generations). Thompson (2013) stressed that estimating age of inbreeding from size of HBD segments (or RoH length) is very difficult due to the inherent stochastic nature of the underlying recombination process. As shown by our simulations, the estimation of R might further be influenced by other factors such as inaccuracies in the genetic map, genotyping errors (when not accounted for properly), presence of several HBD-classes with close rates and/or lower marker density and informativeness.

Some additional precautions must be taken regarding interpretation of the results because the model relies on three important assumptions. First, it assumes that no mutation occurred in HBD segments in the path between the individual and its ancestor. With standard mutation and recombination rates (e.g., as in human or cattle), few mutations per HBD segment are expected and their number is relatively constant regardless of the age since older segments are smaller but have more time for mutations. So, as long as enough SNPs are present per segment, the impact of mutations should be low and accounted for by the genotyping error rate parameter. In addition, favoring old SNPs (as in genotyping arrays or via MAF filtering) is advisable. The second assumption is that the marker allele frequencies in the base populations are known. A special attention must be taken when working with several very different populations and markers that have been selected based on their frequencies in only a subset

of these. When many markers are not segregating in one population (due to ascertainment bias) but frequencies are estimated across populations, they might generate spurious HBD signals. It is therefore important either to estimate the frequencies within population (which need a sample size large enough) or use markers segregating in a large number of representative populations. Finally, the model assumes that after conditioning on the HBD-state, adjacent markers are independent. This is obviously not the case in presence of LD and Polasek *et al* (2010) concluded that ignoring LD leads to upward biases in inbreeding estimates. Note that absence of background LD is also implicit in ROH-based methods and approaches using excess of homozygosity or the genomic relationship matrix. Ideally, HMM could be extended to explicitly account for background LD (e.g., Tang *et al*, 2006; Wang *et al*, 2006) but this would increase the complexity of the model (and computational costs). Simpler strategies relying on LD-pruning to remove markers in high LD have been proposed (e.g., Gazal *et al*, 2014; Leutenegger *et al*, 2011). Although applicable with any method, LD-pruning is however not systematically used since some authors consider that LD might be the result of the mating of (very distantly) related individuals (Broman & Weber, 1999) and of ancient coancestry (Thompson, 2013). In addition, from a practical point of view, reducing marker density might affect the power to identify the shortest HBD-segment (in particular for RoH-based approaches) and their boundaries. As the approach proposed by Pemberton *et al* (2012), our multiple HBD-classes model actually represent a valuable compromise between these two strategies to deal with LD. Indeed, it allows to partition inbreeding in different age-related classes so that short HBD segments (belonging to classes with the highest rate R_k) capture background LD (of ancient origin and thus of similar contribution across all individuals from the population) while long HBD segments capture inbreeding introduced by recent parental relatedness (displaying variation among individuals). Simulations under a Wright-Fisher process suggested that our model with multiple HBD classes was effective even in the presence of background LD. In addition, comparisons of our estimates with those obtained with LD-pruned maps for the analyzed human populations illustrated that the most recent HBD classes closely corresponded to the estimators obtained with the LD-pruned maps whereas short ROH associated with LD patterns were captured by the more ancient HBD-classes.

As other approaches identifying HBD-segments of different lengths, our model-based approach actually allows to explore inbreeding in several dimensions: the global (F_{G}), the local (ϕ_l) and age-variable ($F_{\text{G}}^{(k)}$). It has been suggested that more ancient inbreeding might be less detrimental since deleterious variants are expected to be purged from populations over time (e.g., Hinrichs *et al*, 2007; Leroy, 2014). Yet, the number of generations for this purging to complete depends on the population history (e.g., Hedrick & Garcia-Dorado, 2016). For instance, strong bottlenecks tend to reduce the intensity of selection against deleterious variants ("the cost of domestication")

and artificial selection might favor some breeders carrying deleterious variants. With our model we can estimate the inbreeding depression associated with different HBD classes. This requires appropriate data sets (individuals genotyped at high marker density to capture old inbreeding and with own fitness records) and sufficient variation in all HBD-classes. Alternatively, recent and old inbreeding can be compared by functional annotations of different segments. For instance, Szpiech *et al* (2013) showed that long ROH are enriched for deleterious variants in humans. We can also use our model to test for local inbreeding depression and identify regions or variants where homozygosity seems more deleterious (e.g., Leutenegger *et al*, 2006; Wang *et al*, 2009).

In practice, several strategies can be used to infer inbreeding in populations with our model. First, when using only one HBD class as in Leutenegger *et al* (2003), we can either estimate a single rate common to both HBD and non-HBD classes or a different value for both states. The first option results in a model similar to Leutenegger *et al* (2003) and Vieira *et al* (2016) (note that the model by Narasimhan *et al* (2016) does not estimate the rate but a single transition parameters combining R and the genetic distances) and results in better estimates of the rate. Next, we can select the best number of HBD-classes according to the BIC criterion to compare the different models. When evaluated under simulated data, the BIC appeared to be conservative since the selected values were smaller or equal to the simulated ones. Note that with this approach we select the number of classes that best fit the data (merging several close classes if necessary) and not the real number of classes. Finally, we can use a set of HBD (and non-HBD) classes with pre-defined rates (the so-called MIXKR models). It is then recommended to well separate these rates (e.g., using a ratio of two or more between successive rates to limit the overlap between the exponential distributions assumed for the HBD segment lengths) and cover a range of generations compatible with the available marker density. That strategy proved particularly efficient in most cases since it provided accurate estimates of the overall and local inbreeding while providing insights into the partitioning of inbreeding in the different HBD-classes and more easily comparable results across individuals from the same population. Such a model was only sub-optimal when a single and rare HBD class was simulated (which might not be usual in real populations) but required larger computational resources since more classes are simultaneously fitted.

Several direction might be followed to improve our model, for instance to better take into account the possibility of mutations or to estimate the allele frequencies. Another possible extension to capitalize on individual inbreeding for past demographic inference of the whole population would be to explicitly relate the contribution of each HBD class to each and every individual inbreeding to the corresponding past effective population size (see e.g., Browning & Browning, 2015) and further consider all the individuals jointly to estimate these (hyper–)parameters. Such a development might be viewed as an extension of the model from an individual-oriented framework towards

population parameter inference.

Acknowledgements

We are grateful to Martin Kardos and three other anonymous reviewers for their helpful comments. We thank the Human Genome Diversity Project, the LUPA consortium and the International Sheep Genomics consortium for data sharing. John McEwan helped us to obtain the sheep data set and shared his knowledge on history of different sheep populations. The ZooROH project and this work were supported by the Fonds de la Recherche Scientifique - FNRS (F.R.S.-FNRS) under Grant J.0134.16. Tom Druet is Research Associate from the F.R.S.-FNRS. We used the supercomputing facilities of the "Consortium d'Equipements en Calcul Intensif en Fédération Wallonie-Bruxelles" (CECI), funded by the F.R.S.-F.N.R.S.

References

Bjelland DW, Weigel KA, Vukasinovic N, Nkrumah JD (2013) Evaluation of inbreeding depression in holstein cattle using whole-genome snp markers and alternative measures of genomic inbreeding. *Journal of Dairy Science*, **96**, 4697–4706.

Bosse M, Megens HJ, Madsen O, Paudel Y, Frantz LAF, *et al* (2012) Regions of homozygosity in the porcine genome: consequence of demography and the recombination landscape. *PLoS genetics*, **8**, e1003100.

Broman KW, Weber JL (1999) Long homozygous chromosomal segments in reference families from the centre d'Etude du polymorphisme humain. *Am J Hum Genet*, **65**, 1493–500.

Browning SR, Browning BL (2010) High-resolution detection of identity by descent in unrelated individuals. *Am J Hum Genet*, **86**, 526–39.

Browning SR, Browning BL (2015) Accurate Non-parametric Estimation of Recent Effective Population Size from Segments of Identity by Descent. *Am J Hum Genet*, **97**, 404–18.

Caballero A, Bravo I, Wang J (2017) Inbreeding load and purging: implications for the short-term survival and the conservation management of small populations. *Heredity (Edinb)*, **118**, 177–185.

Campbell CL, Bharer C, Morrow BE, Boyko AR, Auton A (2016) A pedigree-based map of recombination in the domestic dog genome. *G3 (Bethesda)*.

975 Charlesworth D, Willis JH (2009) The genetics of inbreeding depression. *Nature Reviews Genetics*, **10**, 783–796.

976 Charlier C, Coppieters W, Rollin F, Desmecht D, Agerholm JS, *et al* (2008) Highly effective SNP-based association
977 mapping and management of recessive defects in livestock. *Nat Genet*, **40**, 449–54.

978 Clutton-Brock TH, Pemberton JM (2004) *Soay Sheep: Dynamics and Selection in an Island Population*. Cam-
979 bridge University Press.

980 Darwin C (1876) *The effects of cross and self fertilisation in the vegetable kingdom*. John Murray, London.

981 Druet T, Farnir FP (2011) Modeling of identity-by-descent processes along a chromosome between haplotypes and
982 their genotyped ancestors. *Genetics*, **188**, 409–19.

983 Estoup A, Ravigne V, Hufbauer R, Vitalis R, Gautier M, Facon B (2016) Is there a genetic paradox of biological
984 invasion? *Annual Review of Ecology, Evolution, and Systematics*, **47**, 51–72.

985 Ferencakovic M, Hamzic E, Gredler B, Solberg TR, Klemetsdal G, *et al* (2013) Estimates of autozygosity derived
986 from runs of homozygosity: empirical evidence from selected cattle populations. *Journal of Animal Breeding*
987 *and Genetics*, **130**, 286–293.

988 Gazal S, Sahbatou M, Perdry H, Letort S, Genin E, Leutenegger AL (2014) Inbreeding coefficient estimation with
989 dense SNP data: comparison of strategies and application to HapMap III. *Hum Hered*, **77**, 49–62.

990 Hedrick PW, Garcia-Dorado A (2016) Understanding inbreeding depression, purging, and genetic rescue. *Trends*
991 *in Ecology and Evolution*.

992 Hedrick PW, Kalinowski ST (2000) Inbreeding depression in conservation biology. *Annual Review of Ecology and*
993 *Systematics*, **31**, 139–162.

994 Hill WG, Weir BS (2011) Variation in actual relationship as a consequence of Mendelian sampling and linkage.
995 *Genet Res (Camb)*, **93**, 47–64.

996 Hinrichs D, Meuwissen TH, Odegard J, Holt M, Vangen O, Woolliams JA (2007) Analysis of inbreeding depression
997 in the first litter size of mice in a long-term selection experiment with respect to the age of the inbreeding.
998 *Heredity (Edinb)*, **99**, 81–8.

999 Jakobsson M, Scholz SW, Scheet P, Gibbs JR, VanLiere JM, *et al* (2008) Genotype, haplotype and copy-number
1000 variation in worldwide human populations. *Nature*, **451**, 998–1003.

1001 Johnston SE, Berenos C, Slate J, Pemberton JM (2016) Conserved genetic architecture underlying individual re-
1002 combination rate variation in a wild population of soay sheep (ovis aries). *Genetics*, **203**, 583–598.

1003 Kardos M, Luikart G, Allendorf FW (2015) Measuring individual inbreeding in the age of genomics: marker-based
1004 measures are better than pedigrees. *Heredity (Edinb)*, **115**, 63–72.

1005 Kardos M, Qvarnstrom A, Ellegren H (2017) Inferring Individual Inbreeding and Demographic History from Seg-
1006 ments of Identity by Descent in Ficedula Flycatcher Genome Sequences. *Genetics*, **205**, 1319–1334.

1007 Kardos M, Taylor HR, Ellegren H, Luikart G, Allendorf FW (2016) Genomics advances the study of inbreeding
1008 depression in the wild. *Evolutionary Applications*, **9**, 1205–1218.

1009 Keller LF, Waller DM (2002) Inbreeding effects in wild populations. *Trends in Ecology and Evolution*, **17**, 230–
1010 241.

1011 Keller MC, Visscher PM, Goddard ME (2011) Quantification of inbreeding due to distant ancestors and its detec-
1012 tion using dense single nucleotide polymorphism data. *Genetics*, **189**, 237–49.

1013 Kijas JW, Lenstra JA, Hayes B, Boitard S, Porto Neto LR, *et al* (2012) Genome-wide analysis of the world's sheep
1014 breeds reveals high levels of historic mixture and strong recent selection. *PLoS Biology*, **10**, e1001258.

1015 Kirin M, McQuillan R, Franklin CS, Campbell H, McKeigue PM, Wilson JF (2010) Genomic runs of homozygosity
1016 record population history and consanguinity. *PloS One*, **5**, e13996.

1017 Kong A, Thorleifsson G, Gudbjartsson DF, Masson G, Sigurdsson A, *et al* (2010) Fine-scale recombination rate
1018 differences between sexes, populations and individuals. *Nature*, **467**, 1099–1103.

1019 Lander ES, Green P (1987) Construction of multilocus genetic linkage maps in humans. *Proc Natl Acad Sci U S*
1020 *A*, **84**, 2363–7.

1021 Leroy G (2014) Inbreeding depression in livestock species: review and meta-analysis. *Anim Genet*, **45**, 618–28.

1022 Leroy G, Danchin-Burge C, Palhiere I, *et al* (2012) An abc estimate of pedigree error rate: application in dog,
1023 sheep and cattle breeds. *Animal Genetics*, **43**, 309–314.

1024 Leutenegger AL, Labalme A, Genin E, *et al* (2006) Using genomic inbreeding coefficient estimates for homozy-
1025 gosity mapping of rare recessive traits: application to Taybi-Linder syndrome. *Am J Hum Genet*, **79**, 62–6.

Leutenegger AL, Prum B, Genin E, *et al* (2003) Estimation of the inbreeding coefficient through use of genomic data. *American Journal of Human Genetics*, **73**, 516–23.

Leutenegger AL, Sahbatou M, Gazal S, Cann H, Genin E (2011) Consanguinity around the world: what do the genomic data of the HGDP-CEPH diversity panel tell us? *Eur J Hum Genet*, **19**, 583–7.

Lewis TW, Abhayaratne BM, Blott SC (2015) Trends in genetic diversity for all kennel club registered pedigree dog breeds. *Canine Genetics and Epidemiology*, **2**, 13.

Li H, Durbin R (2011) Inference of human population history from individual whole-genome sequences. *Nature*, **475**, 493–496.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, *et al* (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078–9.

Li JZ, Absher DM, Tang H, *et al* (2008) Worldwide human relationships inferred from genome-wide patterns of variation. *Science*, **319**, 1100–4.

Li Y, Willer CJ, Ding J, Scheet P, Abecasis GR (2010) MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genet Epidemiol*, **34**, 816–34.

Malécot G (1948) *Les Mathématiques de l'hérédité*. Masson et Cie.

Manichaikul A, Mychaleckyj JC, Rich SS, Daly K, Sale M, Chen WM (2010) Robust relationship inference in genome-wide association studies. *Bioinformatics*, **26**, 2867–2873.

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, *et al* (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res*, **20**, 1297–303.

McQuillan R, Leutenegger AL, Abdel-Rahman R, Franklin CS, Pericic M, *et al* (2008) Runs of homozygosity in european populations. *American Journal of Human Genetics*, **83**, 359–372.

Mott R, Talbot CJ, Turri MG, Collins AC, Flint J (2000) A method for fine mapping quantitative trait loci in outbred animal stocks. *Proc Natl Acad Sci U S A*, **97**, 12649–54.

Narasimhan V, Danecek P, Scally A, Xue Y, Tyler-Smith C, Durbin R (2016) Bcftools/roh: a hidden markov model approach for detecting autozygosity from next-generation sequencing data. *Bioinformatics*, **32**, 1749–1751.

1051 O'Connell D, Scobie D, Hickey S, Sumner R, Pearson A (2012) Selection for yearling fleece weight and its effect
1052 on fleece shedding in new zealand wiltshire sheep. *Animal Production Science*, **52**, 456–462.

1053 Palamara PF (2016) ARGON: fast, whole-genome simulation of the discrete time Wright-fisher process. *Bioinfor-*
1054 *matics*, **32**, 3032–4.

1055 Pemberton TJ, Absher D, Feldman MW, Myers RM, Rosenberg NA, Li JZ (2012) Genomic patterns of homozy-
1056 gosity in worldwide human populations. *American Journal of Human Genetics*, **91**, 275–292.

1057 Polasek O, Hayward C, Bellenguez C, *et al* (2010) Comparative assessment of methods for estimating individual
1058 genome-wide homozygosity-by-descent from human genomic data. *BMC Genomics*, **11**, 139.

1059 Purcell S, Neale B, Todd-Brown K, *et al* (2007) PLINK: a tool set for whole-genome association and population-
1060 based linkage analyses. *Am J Hum Genet*, **81**, 559–75.

1061 Purfield DC, Berry DP, McParland S, Bradley DG (2012) Runs of homozygosity and population history in cattle.
1062 *BMC Genetics*, **13**, 70.

1063 Rabiner LR (1989) A tutorial on hidden markov models and selected applications in speech recognition. In *PRO-*
1064 *CEEDINGS OF THE IEEE*, pp. 257–286.

1065 Rudan I, Smolej-Narancic N, Campbell H, *et al* (2003) Inbreeding and the genetic complexity of human hyperten-
1066 sion. *Genetics*, **163**, 1011–21.

1067 Sempere G, Moazami-Goudarzi K, Eggen A, Laloe D, Gautier M, Flori L (2015) WIDDE: a Web-Interfaced next
1068 generation database for genetic diversity exploration, with a first application in cattle. *BMC Genomics*, **16**, 940.

1069 Szpiech ZA, Xu J, Pemberton TJ, *et al* (2013) Long runs of homozygosity are enriched for deleterious variation.
1070 *American Journal of Human Genetics*, **93**, 90–102.

1071 Szulkin M, Bierne N, David P (2010) Heterozygosity-fitness correlations: a time for reappraisal. *Evolution*, **64**,
1072 1202–17.

1073 Tang H, Coram M, Wang P, Zhu X, Risch N (2006) Reconstructing genetic ancestry blocks in admixed individuals.
1074 *Am J Hum Genet*, **79**, 1–12.

1075 Thompson EA (2008) The IBD process along four chromosomes. *Theor Popul Biol*, **73**, 369–73.

Thompson EA (2013) Identity by descent: variation in meiosis, across genomes, and in populations. *Genetics*, **194**, 301–26.

VanRaden PM (2008) Efficient methods to compute genomic predictions. *J Dairy Sci*, **91**, 4414–23.

Vaysse A, Ratnakumar A, Derrien T, Axelsson E, Rosengren Pielberg G, *et al* (2011) Identification of genomic regions associated with phenotypic variation between dog breeds using selection mapping. *PLoS Genetics*, **7**, e1002316.

Vieira FG, Albrechtsen A, Nielsen R (2016) Estimating ibd tracts from low coverage ngs data. *Bioinformatics*, **32**, 2096–2102.

Wang H, Lin CH, Service S, Chen Y, Freimer N, Sabatti C (2006) Linkage disequilibrium and haplotype homozygosity in population samples genotyped at a high marker density. *Hum Hered*, **62**, 175–89.

Wang J (2016) Pedigrees or markers: Which are better in estimating relatedness and inbreeding coefficient? *Theor Popul Biol*, **107**, 4–13.

Wang S, Haynes C, Barany F, Ott J (2009) Genome-wide autozygosity mapping in human populations. *Genet Epidemiol*, **33**, 172–80.

Wright S (1922) Coefficients of inbreeding and relationship. *American Naturalist*, **56**, 330–338.

Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, *et al* (2010) Common SNPs explain a large proportion of the heritability for human height. *Nat Genet*, **42**, 565–9.

Data Accessibility

All data sets used in the present study are publicly available. the Human Genome Diversity Panel (HGDP) data was downloaded from `ftp://ftp.cephb.fr/hgdp_supp10/Harvard_HGDP-CEPH`, the dog LUPA project from `http://dogs.genouest.org/SWEEP.dir/Supplemental.html` and the Sheep Diversity panel from the WIDDE data base. The program ZooRoH implementing our model can be freely obtained at `https://github.com/tdruet/ZooRoH`.

List of Figures

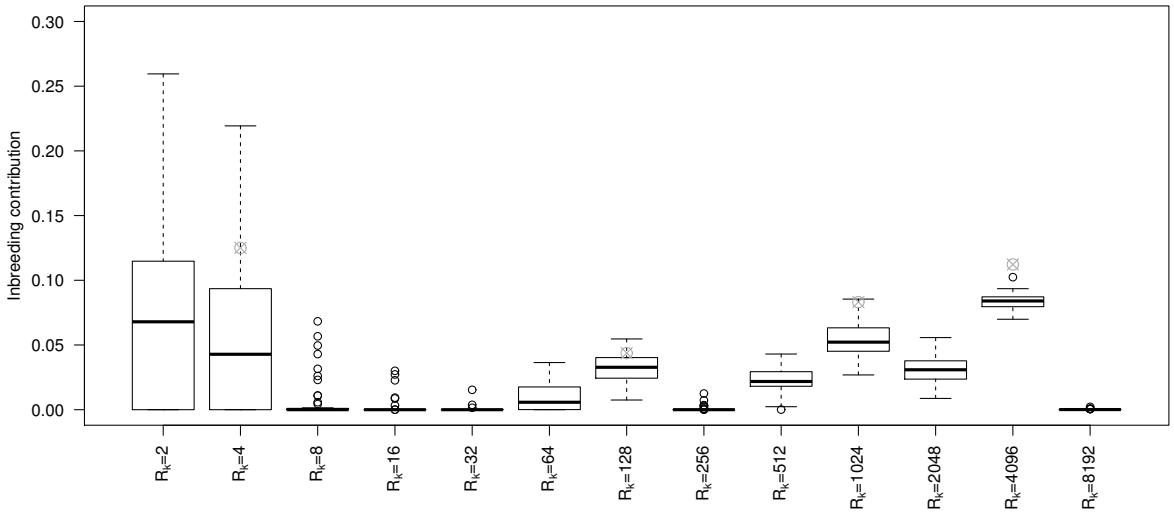

Figure 1. Estimated inbreeding contributions $F_{\mathrm{G}}^{(k)}$ for 13 HBD classes with pre-defined rates (MIX14R model) on data simulated under the 5R model (4 HBD classes). The simulated genome consisted of 25 chromosomes of 100 cM with a marker density of 1000 SNPs per cM. Genotyping data for 50 individuals were simulated under the 5R inference model i.e., with 4 HBD-classes with the following realized rates (inbreeding contributions) as indicated by a star in the plot: $R_1 = 4$ ($F_{\mathrm{G}}^{(1)} = 0.125$), $R_2 = 128$ ($F_{\mathrm{G}}^{(2)} = 0.08$), $R_3 = 1024$ ($F_{\mathrm{G}}^{(3)} = 0.04$) and $R_4 = 4096$ ($F_{\mathrm{G}}^{(4)} = 0.11$). The data were analyzed with the MIX14R that consisted of 13 HBD-classes with pre-defined rates ranging from 2 to 8192 (with $R_k = 2^k$ for each class k) and one non-HBD class that had the same rate as the older HBD class (i.e., $R_K = R_{K-1} = 8192$). For each of these 13 HBD classes, the boxplots give the distribution of the estimated inbreeding contribution ($\widehat{F_{\mathrm{G}}^{(k)}}$) over the 50 simulated individuals.

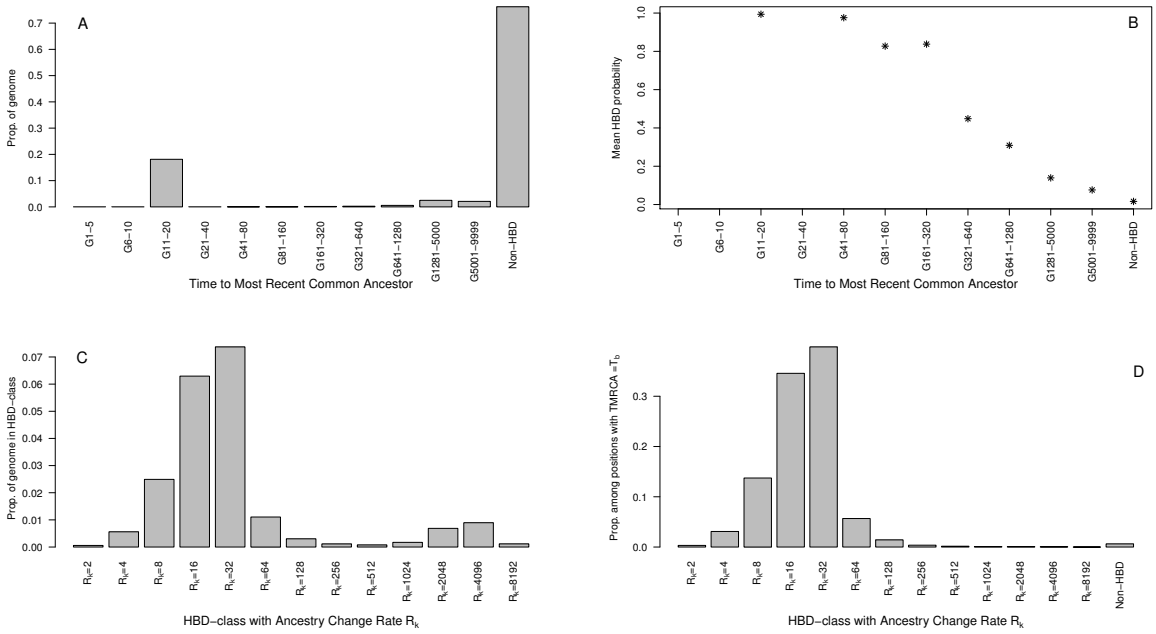

Figure 2. Evaluation of the MIX14R model on a data set consisting of 50 diploid individuals simulated under a Wright–Fisher demographic history with varying population sizes. The population evolved under a WF1 scenario (see the Material and Methods section) with $Ne_1 = 10^5$, $T_s = 10^4$ and a bottleneck lasting from generations 17 to 14 in the past and during which the population size was $N_{eb} = 20$. A) Realized distribution of the proportions of the simulated individual genomes lying within HBD segments as a function of their TMRCA (the interval G11-20 contains HBD segments tracing back to the bottleneck period, i.e., 14 to 17 generations backward in time) and within non-HBD segments (background). B) Estimated local inbreeding probabilities (ϕ_l) averaged over all the simulated individuals and markers as a function of the actual TMRCA of the underlying HBD segments. C) Distributions of the estimated proportion of the individual genomes assigned to each of the 13 pre-defined HBD classes (over the 50 simulated individuals). D) Proportion of the SNPs lying in HBD segments originating from the bottleneck period (i.e., 14 to 17 generations backward in time) that are assigned to the 14 different HBD and non-HBD classes of the MIX14R model (summed over all the 50 individuals).

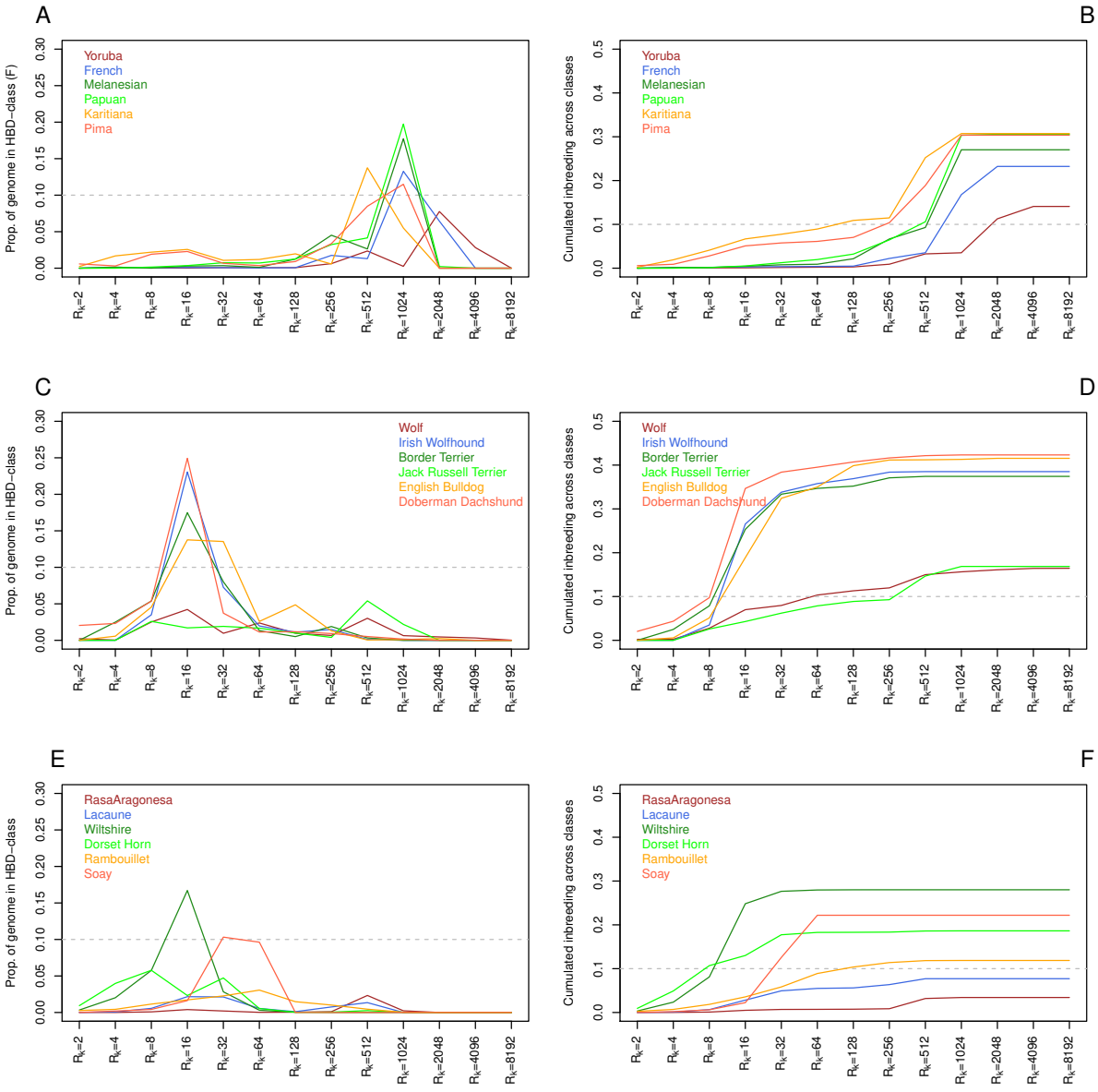

Figure 3. Average estimated proportions of inbreeding contribution of a set of 13 pre-defined HBD classes for human (A), dog (C) and sheep (E) populations and corresponding average cumulative inbreeding (B, D and F for human, dog and sheep populations respectively). These means were obtained by summarizing individual values from all individuals from a population / breed.

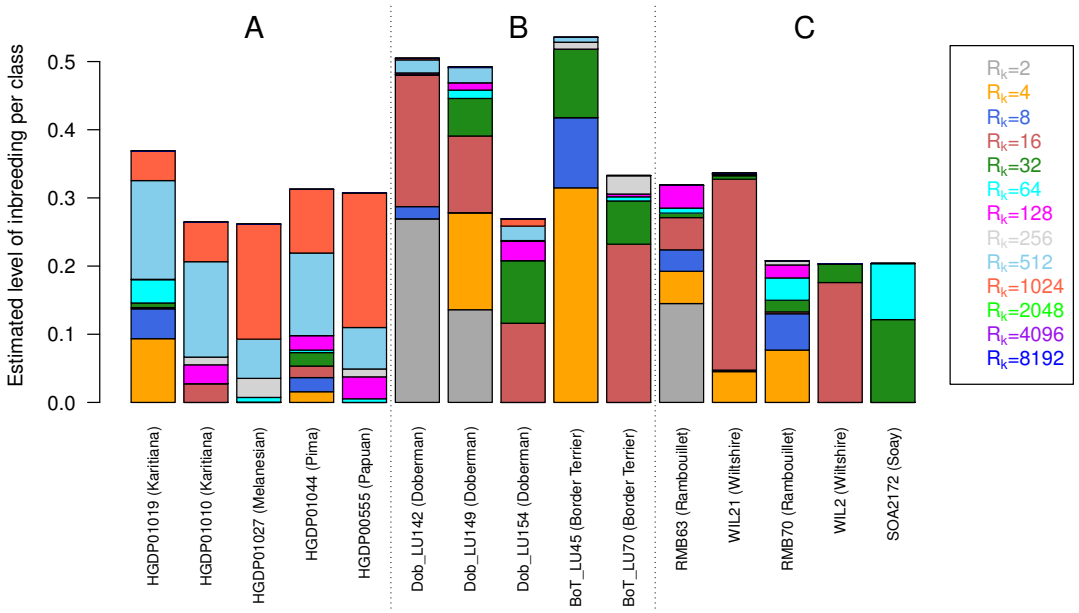

Figure 4. Estimated level of inbreeding per HBD class in five humans (A), five dogs (B) and five sheeps (C). Each color is associated with a distinct class (defined by its rate). The heights of each color bar represent the estimated level of inbreeding associated to the class and the total height represents the overall inbreeding.

List of Tables

Scenario		Realized median values				Median estimated values (1R model)			
R	ρ	R	ρ	F_{G}	$\#Tracts$	$\widehat{R}$ (MAE)	$\widehat{\rho}$ (MAE)	$\widehat{F_{\mathrm{G}}}$ (MAE)	MAE for $\widehat{\phi_l}$ ($\widehat{\phi_{l^{\mathrm{HBD}}}}$)
2	0.500	2.00	0.507	0.500	38.0	2.00 (0.34)	0.503 (0.0325)	0.500 (0.0005)	0.002 (0.002)
3	0.250	3.00	0.249	0.251	25.0	3.00 (0.43)	0.248 (0.0287)	0.251 (0.0005)	0.003 (0.006)
4	0.125	3.90	0.124	0.125	15.0	4.00 (0.57)	0.126 (0.0194)	0.124 (0.0005)	0.003 (0.010)
8	0.125	8.10	0.126	0.124	28.0	8.00 (0.82)	0.124 (0.0148)	0.124 (0.0008)	0.005 (0.021)
16	0.010	16.0	0.009	0.009	4.00	16.7 (10.1)	0.009 (0.0034)	0.009 (0.0005)	0.001 (0.065)
16	0.020	16.7	0.019	0.018	8.00	16.6 (4.02)	0.018 (0.0054)	0.018 (0.0007)	0.003 (0.062)
16	0.050	16.0	0.049	0.049	21.0	16.2 (1.99)	0.050 (0.0080)	0.048 (0.0009)	0.006 (0.055)
16	0.100	16.0	0.099	0.098	42.0	16.0 (1.35)	0.098 (0.0112)	0.097 (0.0011)	0.010 (0.050)
32	0.010	34.3	0.010	0.009	8.00	34.1 (11.9)	0.009 (0.0028)	0.009 (0.0009)	0.003 (0.160)
32	0.020	32.4	0.019	0.019	16.0	32.8 (6.13)	0.019 (0.0037)	0.019 (0.0011)	0.006 (0.141)
32	0.050	32.3	0.049	0.049	41.0	32.7 (3.62)	0.049 (0.0062)	0.049 (0.0014)	0.012 (0.123)
32	0.100	32.1	0.100	0.100	83.0	32.0 (2.26)	0.100 (0.0085)	0.100 (0.0017)	0.021 (0.103)
64	0.010	65.7	0.010	0.010	16.0	63.7 (17.6)	0.009 (0.0025)	0.009 (0.0016)	0.006 (0.326)
64	0.020	66.1	0.020	0.019	32.0	66.7 (11.2)	0.020 (0.0033)	0.020 (0.0017)	0.012 (0.291)
64	0.050	64.4	0.050	0.050	80.5	64.5 (6.17)	0.049 (0.0046)	0.049 (0.0021)	0.024 (0.243)
64	0.100	64.2	0.099	0.099	162	64.3 (4.06)	0.099 (0.0063)	0.099 (0.0024)	0.041 (0.206)
128	0.050	128	0.050	0.050	162	128 (11.8)	0.049 (0.0044)	0.049 (0.0030)	0.044 (0.439)
128	0.100	128	0.101	0.100	323	127 (8.03)	0.100 (0.0058)	0.100 (0.0037)	0.074 (0.368)
256	0.050	257	0.050	0.050	322	259 (26.7)	0.050 (0.0049)	0.050 (0.0043)	0.066 (0.669)
256	0.100	256	0.100	0.100	643	257 (16.7)	0.099 (0.0055)	0.099 (0.0046)	0.113 (0.569)

Table 1. Performance of the 1R model on data simulated under the 1R inference model. The simulated genome consisted of 25 chromosomes of 100 cM with a marker density of 10 SNPs per cM. Genotyping data for 500 individuals were simulated under the 1R inference model for each of 20 different scenarios defined by the simulated R and ρ values reported in the first two columns. The table reports the resulting median realized (true) values (across the 500 simulated individuals) for rate of co-ancestry change (R), the mixing proportions (ρ), the individual inbreeding (F_{G}) and the number of HBD tracks ($\#Tracks$). Similarly, the table gives the median estimated values and the Mean Absolute Errors (MAE) for the rate of co-ancestry change ($\widehat{R}$), the mixing proportions ($\widehat{\rho}$) and the individual inbreeding ($\widehat{F_{\mathrm{G}}}$). Finally, the table gives the MAE for the estimated local inbreeding (ϕ_l) either for all the SNPs ($\widehat{\phi_l}$) or for those actually lying within HBD segments ($\widehat{\phi_{l^{\mathrm{HBD}}}}$).

Scenario		Realized median values			Median estimated values (3R model)					
R_1 (ρ_1)	R_2 (ρ_2)	R_1 ($F_{\mathrm{G}}^{(1)}$)	R_2 ($F_{\mathrm{G}}^{(2)}$)	F_{G}	$\widehat{R_1}$ (MAE)	$\widehat{R_2}$ (MAE)	$\widehat{F_{\mathrm{G}}^{(1)}}$ (MAE)	$\widehat{F_{\mathrm{G}}^{(2)}}$ (MAE)	$\widehat{F_{\mathrm{G}}}$ (MAE)	MAE for $\widehat{\phi_l}$ ($\widehat{\phi_{l\mathrm{HBD}}}$)
4 (0.125)	16 (0.100)	4.1 (0.12)	17 (0.09)	0.210	7.20 (3.06)	391 (288)	0.195 (0.075)	0.004 (0.074)	0.210 (0.002)	0.012 (0.025)
4 (0.125)	64 (0.100)	4.1 (0.12)	64 (0.09)	0.211	3.60 (1.01)	64.6 (9.53)	0.123 (0.007)	0.086 (0.007)	0.211 (0.002)	0.038 (0.089)
4 (0.125)	256 (0.100)	4.0 (0.12)	257 (0.09)	0.211	3.60 (0.65)	275 (35.9)	0.120 (0.001)	0.087 (0.004)	0.208 (0.004)	0.101 (0.238)
8 (0.100)	128 (0.100)	8.2 (0.10)	128 (0.09)	0.189	7.20 (1.48)	126 (14.8)	0.098 (0.004)	0.090 (0.005)	0.189 (0.003)	0.069 (0.182)
32 (0.100)	64 (0.100)	32 (0.10)	67 (0.09)	0.190	33.9 (7.08)	102 (140)	0.157 (0.058)	0.030 (0.057)	0.192 (0.003)	0.051 (0.132)
32 (0.100)	256 (0.100)	32 (0.10)	260 (0.09)	0.188	29.6 (4.31)	265 (38.0)	0.097 (0.007)	0.089 (0.007)	0.188 (0.004)	0.114 (0.302)

Table 2. Performance of the 3R model on data simulated under the 3R inference model (i.e., two HBD classes and one non-HBD class). The simulated genome consisted of 25 chromosomes of 100 cM with a marker density of 10 SNPs per cM. Genotyping data for 500 individuals were simulated under the 3R inference model for each of 6 different scenarios defined by the simulated rates R_1 and R_2 (reported in the two first columns) and the corresponding mixing proportions ρ_1 and ρ_2 (reported in the third and fourth columns) of the two classes of HBD segments. The table reports the resulting median realized (true) values (across the 500 simulated individuals) for the rates of co-ancestry change (R_1 and R_2), the amount of inbreeding originating from each HBD class ($F_{\mathrm{G}}^{(1)}$ and $F_{\mathrm{G}}^{(2)}$) and the overall individual inbreeding (F_{G}). The table further gives the median (and their associated MAE) of the estimated values ($\widehat{R_1}$, $\widehat{R_2}$, $\widehat{F_{\mathrm{G}}^{(1)}}$, $\widehat{F_{\mathrm{G}}^{(2)}}$ and $\widehat{F_{\mathrm{G}}}$) obtained under the 3R model. The table also gives the MAE for the estimated local inbreeding (ϕ_l) either for all the SNPs ($\widehat{\phi_l}$) or for those actually lying within HBD segments only ($\widehat{\phi_{l\mathrm{HBD}}}$).

Simulation				Realized median value		Estimated median value		
R	ρ	SNP per cM	AFS	R	F_{G}	$\widehat{R}$ (MAE)	$\widehat{F_{\mathrm{G}}}$ (MAE)	MAE for $\widehat{\phi_l}$ ($\widehat{\phi_{l^{\mathrm{HBD}}}}$)
4	0.125	10	Array-like	3.90	0.125	4.00 (0.57)	0.124 (0.001)	0.0026 (0.0101)
4	0.125	100	Array-like	4.00	0.123	4.00 (0.51)	0.123 (0.000)	0.0002 (0.0009)
4	0.125	10	NGS-like	4.10	0.119	4.00 (0.64)	0.120 (0.002)	0.0068 (0.0272)
4	0.125	100	NGS-like	4.10	0.120	4.00 (0.55)	0.120 (0.000)	0.0006 (0.0023)
64	0.100	10	Array-like	64.2	0.099	64.3 (4.06)	0.099 (0.002)	0.0410 (0.2056)
64	0.100	100	Array-like	64.6	0.099	64.4 (2.00)	0.099 (0.000)	0.0035 (0.0181)
64	0.100	10	NGS-like	64.2	0.100	64.1 (6.26)	0.100 (0.006)	0.0807 (0.4032)
64	0.100	100	NGS-like	64.1	0.099	64.2 (2.50)	0.099 (0.000)	0.0095 (0.0482)
256	0.100	10	Array-like	256	0.100	257 (16.7)	0.099 (0.005)	0.1134 (0.5689)
256	0.100	100	Array-like	255	0.100	256 (5.79)	0.100 (0.000)	0.0164 (0.0824)
256	0.100	10	NGS-like	257	0.100	252 (36.9)	0.100 (0.008)	0.1462 (0.7313)
256	0.100	100	NGS-like	256	0.100	255 (8.06)	0.100 (0.001)	0.0398 (0.1994)

Table 3. Performance of the 1R model on simulated data sets with different SNP density and informativeness. The simulated genome consisted of 25 chromosomes of 100 cM with a marker density of either 10 or 100 SNPs per cM. Allele frequency spectrum (AFS) of each SNP reference allele were either sampled from an empirical distribution (array-like) derived from a real (cattle) genotyping assay (i.e., close to uniform) or from a $\beta(0.2, 0.2)$ distribution (U-shaped) that mimics NGS data (NGS-like). Genotyping data for 500 individuals were simulated under the 1R inference model for each of 3 different scenarios defined by the simulated R and ρ values reported in the first two columns. For each simulation, the table reports the resulting realized (true) median value (across the 500 simulated individuals) for the rate of co-ancestry change (R) and the individual inbreeding (F_{G}) together with the median of their estimated values $\widehat{R}$ and $\widehat{F_{\mathrm{G}}}$ and corresponding Mean Absolute Errors (MAE). Finally, the table gives the MAE for the estimated local inbreeding (ϕ_l) either for all the SNPs ($\widehat{\phi_l}$) or for those actually lying within HBD segments only ($\widehat{\phi_{l^{\mathrm{HBD}}}}$).