An hidden Markov model to estimate homozygous-by-descent probabilities associated with nested layers of ancestors

Tom Druet[1] and Mathieu Gautier[2]

November 15, 2021

[1]Unit of Animal Genomics, GIGA-R and Faculty of Veterinary Medicine, University of Liège, Liège, Belgium

[2]INRAE, UMR CBGP (INRAE—IRD—Cirad—Montpellier SupAgro), Montferrier-sur-Lez, France

Corresponding author: Tom Druet (tom.druet@uliege.be)

Abstract

Inbreeding results from the mating of related individuals and has negative consequences because it brings together deleterious variants in one individual. Genomic estimates of the inbreeding coefficients are preferred to pedigree-based estimators as they measure the realized inbreeding levels and they are more robust to pedigree errors. Several methods identifying homozygous-by-descent (HBD) segments with hidden Markov models (HMM) have been recently developed and are particularly valuable when the information is degraded or heterogeneous (e.g., low-fold sequencing, low marker density, heterogeneous genotype quality or variable marker spacing). We previously developed a multiple HBD class HMM where HBD segments are classified in different groups based on their length (e.g., recent versus old HBD segments) but we recently observed that for high inbreeding levels with many HBD segments, the estimated contributions might be biased towards more recent classes (i.e., associated with large HBD segments) although the overall estimated level of inbreeding remained unbiased. We herein propose a new model in which the HBD classification is modeled in successive nested levels with decreasing expected HBD segment lengths, the underlying exponential rates being directly related to the number of generations to the common ancestor. The non-HBD classes are now modeled as a mixture of HBD segments from later generations and shorter non-HBD segments (i.e., both with higher rates). The new model has improved statistical properties and performs better on simulated data compared to our previous version. We also show that the parameters of the model are easier to interpret and that the model is more robust to the choice of the number of classes. Overall, the new model results in an improved partitioning of inbreeding in different HBD classes and should be preferred.

Keywords: homozygous-by-descent; inbreeding; hidden Markov model; autozygosity; ROH

1 Introduction

In diploid species, offspring of related individuals can carry at autosomal loci a pair of DNA segments originating from the same common ancestor. These stretches of contiguous loci where the two DNA copies are identical-by-descent (IBD) are referred to as homozygous-by-descent (HBD) or autozygous segments. The length of these HBD segments is inversely related to the size of the so-called inbreeding loop that connects the individual to its common ancestor, since multiple generations of recombination will tend to reduce the size of each transmitted DNA copy. The inbreeding level of an individual can be defined as the proportion of its genome that lies in HBD segments. Genomic data may allow to directly estimate this proportion to provide an estimator of the realized inbreeding coefficient (Leutenegger *et al.*, 2003), whereas pedigree-based estimators, when available, can only provide expected values. Such estimates of inbreeding coefficients are highly valuable for the study of inbreeding depression and the management of livestock populations or those in conservation programs. In addition, detailed assessment of the distribution of HBD segments over the genomes can also be used in homozygosity mapping experiments (Abney *et al.*, 2002; Leutenegger *et al.*, 2006), to identify recessive alleles causing genetic defects or diseases, or for demographic inference purposes (Kirin *et al.*, 2010; Ceballos *et al.*, 2018).

In practice, HBD segments may be identified as runs-of-homozygosity (ROH) that correspond to long stretches of homozygous genotypes (Broman and Weber, 1999; McQuillan *et al.*, 2008). Such ROH can be empirically detected with rule-based approaches requiring the definition of parameters such as window size, minimum ROH length, marker density, maximum allowed spacing between successive markers and number of missing or heterozygous genotypes (Purcell *et al.*, 2007). More formally, likelihood-based ROH approaches allow to compare the likelihoods of segments to be allozygous versus autozygous regions based on marker allele frequencies and the genotyping error probabilities (Pemberton *et al.*, 2012; Wang *et al.*, 2009). These approaches still require the prior definition of fixed-length windows to scan the genome for ROH segments. Alternatively, several authors developed fully probabilistic approaches based on hidden Markov models (HMM) (Leutenegger *et al.*, 2003; Narasimhan *et al.*, 2016; Vieira *et al.*, 2016; Druet and Gautier, 2017). As likelihood-based approaches, they rely on genotype frequencies and genotyping error probabilities, but in addition, they take into account inter-marker genetic distances. Moreover, they do not require prior selection of some window size as HBD estimations are integrated over all possible window lengths. Uncertainty in genotype calling, as for low-fold sequencing data, can also be integrated over (Vieira *et al.*, 2016; Druet and Gautier, 2017). These two later characteristics thus make HMM methods particularly valuable for the analyses of data set with low marker density and/or heterogeneous genotype quality and/or heterogeneous marker spacing. For instance, they are the method of choice to work with ancient DNA (e.g., Renaud *et al.*, 2019; Ringbauer *et al.*, 2021), where genotype quality is particularly poor, and several HMM have been developed in the field. Similarly, they are particularly well suited to work with exome sequencing data (Magi *et al.*, 2014), low density marker array (e.g., Solé *et al.*, 2017; Druet *et al.*, 2020) or with low-fold sequencing data (Vieira *et al.*, 2016). Overall, fewer parameters need to be defined when using these tools.

In HMM based approaches, the length of HBD segments is generally assumed to be exponentially distributed. Modeling a single exponential distribution amounts to assume that all the autozygosity is associated to ancestors present in the same past generations. For complex population histories, this assumption may be too restrictive and Druet and Gautier (2017) proposed to use a mixture of exponential distributions to model HBD segment classes of different expected lengths, under a similar HMM framework. Such classification is actually an HMM counterpart to methods that were developed to automatically classify the identified ROH based on their observed length (Pemberton *et al.*, 2012; Szpiech *et al.*, 2017). With HMM, HBD classes can be viewed as group of ancestors present in different past generations. This model better accounts for complex demographic histories in which different ancestors from many different past generations may contribute to autozygosity. We showed that it improves the fit of individual genetic data and provides more accurate estimations of autozygosity levels. For instance, a single HBD class model might underestimate autozygosity when multiple generations contribute to it, and also tend to regress length of HBD segment towards intermediate values, cutting in particular the longest segments into shorter pieces (e.g., Solé *et al.*, 2017). An accurate estimation of HBD segment length distribution may however be critical to estimate the number of generations to the common ancestors. Likewise, the multiple HBD-class model provides insights into the past demographic history of populations by estimating the relative contributions of past generations to contemporary inbreeding levels (Druet and Gautier, 2017).

The properties of the multiple HBD-class model have already been studied in detail in some of our previous works and most particularly its robustness to i) HBD segment length (age of the ancestors); ii) marker density; iii) allele frequency spectrum; iv) sequencing depth; v) genotyping error; and vi) variable recombination rate (e.g., Solé *et al.*, 2017; Druet *et al.*, 2020). We also investigated aspects related to model selection, including the number of classes and their rates, in simulated (Druet and Gautier, 2017) and real data sets (Solé *et al.*, 2017). The behaviour of these models were also evaluated when multiple-HBD classes contributed to autozygosity, including comparisons when the underlying ancestors were separated by a few generations. A detailed discussion on these aspects is for instance available in Druet and Gautier (2017) and in the vignette from the RZooRoH package (Bertrand *et al.*, 2019). Importantly, we showed that the model was robust to background Linkage Disequilibrium (LD) that was well captured by the most ancient HBD classes (i.e., HBD segments) and thus, as desired, did not influence estimates of recent inbreeding levels (Druet and Gautier, 2017; Solé *et al.*, 2017). From a practical point of view, this made LD pruning of the analyzed data set unnecessary. In addition, the multiple-HBD class model has been compared to other methods, including rule-based ROH, likelihood-based ROH and the single HBD class HMM (Druet and Gautier, 2017; Solé *et al.*, 2017; Alemu *et al.*, 2021).

We recently observed that when the contribution of recent ancestors is extremely high, the multiple HBD classes model in its initial definition (as of Druet and Gautier, 2017) tended to underestimate the age of HBD segments by shifting HBD partitioning towards more recent classes (Druet *et al.*, 2020), although the overall estimated levels of inbreeding remained unbiased. To solve this issue we herein propose a modified model in which the HBD classification is modeled in successive nested levels, each level corresponding to a single HBD class model with a distinct rate. As a result, the non-HBD classes are now modeled as a mixture of HBD segments from later generations and shorter non-HBD segments (i.e., both from subsequent levels with higher rates). We carried out a detailed simulation study to show that the upgraded model had better statistical properties and performed better compared to our previous version. We also show that the parameters of the model are easier to interpret and that the model is more robust to the choice of the number of classes (e.g., the autozygosity partitioning remains more similar when additional classes are added). We also provide an illustration on genotyping data from European Bison that we previously analyzed with the original model (Druet *et al.*, 2020).

2 Models

2.1 Previous models

2.1.1 Single HBD-class model (1R model)

Leutenegger *et al.* (2003) proposed to describe the genome of an individual as a mosaic of HBD and non-

HBD segments with a HMM. In that model, the length of HBD segments inherited without recombination from a common ancestor is exponentially distributed with a rate R. This rate R is related to the number of generations of recombination along both paths connecting each of the two individual DNA copies (haplotype) to their common ancestor, and their frequency is a direct function of the mixing coefficient ρ. The HBD and non-HBD segments are not directly observed but their distribution can be inferred using genotype data available for a set of markers. In that case, the model can be represented as an HMM with two hidden states (state 1 = "HBD" and state 2 = "non-HBD") with the following transition probabilities between two consecutive markers m and $m+1$:

$$\begin{cases} \mathbb{P}\left[S_{m+1}=1 \mid S_m=1\right] & = e^{-Rd_m} + (1-e^{-Rd_m})\rho \\ \mathbb{P}\left[S_{m+1}=1 \mid S_m=2\right] & = (1-e^{-Rd_m})\rho \\ \mathbb{P}\left[S_{m+1}=2 \mid S_m=2\right] & = e^{-Rd_m} + (1-e^{-Rd_m})(1-\rho) \\ \mathbb{P}\left[S_{m+1}=2 \mid S_m=1\right] & = (1-e^{-Rd_m})(1-\rho) \end{cases} \tag{1}$$

where S_m is the state at position m, d_m is the genetic distance in Morgans between markers m and $m+1$. The term e^{-Rd_m} represents the probability that there is no recombination on both genealogical paths between two consecutive markers m and $m+1$ (i.e., the HBD status remains the same). We use the term 'coancestry changes' to refer to the presence of at least one recombination on these paths as in Leutenegger *et al.* (2003), and R will be called the rate of coancestry change accordingly. In this HMM, the equilibrium HBD probability is ρ, which has been shown to be an unbiased estimator of the inbreeding coefficient defined as the proportion of genome HBD (Leutenegger *et al.*, 2003). Note that the inbreeding coefficient may also be derived from the estimated posterior HBD probability at each marker (see eq. 21 below) leading to slightly different but highly correlated estimations.

It should be noticed that the expected length of HBD segments, that we define here and in the remainder of our work in a strict sense (i.e., without any coancestry change), is equal to $1/R$. Yet, in individual genomes, some HBD segments may actually be neighboring. For instance, in the case of a marriage between cousins a tract of IBD markers may consist of two consecutive HBD segments inherited from the grand-father and the grand-mother. More precisely, under the above 1R model, the number of consecutive HBD segments actually follows a geometric distribution with parameter $1-\rho$, the probability of entering a non-HBD segment after a coancestry change. As a result, the expected length of tracts of

IBD markers that may include one or several coancestry changes is equal to $1/R(1-\rho)$ as noticed by Leutenegger *et al.* (2003).

The emission probabilities are the probabilities to observe the marker genotypes conditionally on the underlying state. For non-HBD and HBD states, these emission probabilities are a function of expected genotype frequencies in non-HBD and HBD segments, respectively (Crow *et al.*, 1970; Broman and Weber, 1999; Leutenegger *et al.*, 2003). For the HBD state:

$$\mathbb{P}\left[A_{mi}A_{mj} \mid S_m = 1, p_{mi}\right] = \begin{cases} p_{mi} & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases} \tag{2}$$

where A_{mi} and A_{mj} are the two alleles observed at marker m, i and j representing the allele numbers, p_{mi} is the frequency of allele i at marker m. Ideally, these allele frequencies should be estimated from individuals in a reference population but they are generally computed from the sampled individuals. For the non-HBD state:

$$\mathbb{P}\left[A_{mi}A_{mj} \mid S_m = 2, p_{mi}, p_{mj}\right] = \begin{cases} p_{mi}^2 & \text{if } i = j \\ 2p_{mi}p_{mj} & \text{if } i \neq j \end{cases} \tag{3}$$

The expected frequencies in non-HBD segments (eqn. 3) correspond to Hardy-Weinberg proportions. These emission probabilities are similar to probabilities used in maximum likelihood estimators of the inbreeding coefficient (e.g., Weir *et al.*, 2006). As a result, when markers are considered independent (i.e., probability of coancestry change equal to 1), both approaches lead to very similar estimates (see Alemu *et al.*, 2021). The extension of these emission probabilities to incorporate genotyping error or mutation probability is straightforward (see Broman and Weber, 1999; Leutenegger *et al.*, 2003; Druet and Gautier, 2017). Similarly, the emission probabilities can also be modified to handle next-generation sequencing data (e.g., genotype likelihoods) allowing efficient analysis of shallow sequencing or GBS data (see Vieira *et al.*, 2016; Narasimhan *et al.*, 2016; Druet and Gautier, 2017).

2.1.2 Models with multiple HBD classes (KR and MixKR models)

In the single HBD class model, all HBD segments have the same expected length defined by the rate parameter R. Hence, ancestors contributing to HBD segments are assumed to have been present approximately in the same past generations. To model the contribution of different groups of ancestors to autozygosity (i.e., account for the difference in HBD segment lengths originating from ancestors living in different past generations), we introduced models with multiple HBD classes (Druet and Gautier, 2017). In these new models, each class correspond to a distinct state, with states 1 to $K-1$ for HBD segments originating from groups of ancestors living in different past generations and a Kth state for non-HBD positions. For each HBD class c ($c = 1, \dots, K-1$), HBD segment lengths are assumed exponentially distributed with rate R_c . The non-HBD state corresponds to positions that do not trace back to the same haplotype from a common ancestor up to the most remote HBD class, and has it own rate R_K. The transition probabilities from state b at marker m to state a at marker $m+1$ are:

$$\mathbb{P}\left[S_{m+1} = a \mid S_m = b\right] = \begin{cases} e^{-R_b d_m} + (1 - e^{-R_b d_m})\rho_a & \text{if } a = b \\ (1 - e^{-R_b d_m})\rho_a & \text{if } a \neq b \end{cases} \tag{4}$$

where ρ_c is the mixing coefficient associated with class c.

We previously called these multiple rate models, 'KR' models (e.g., 1R model corresponding to the single HBD-class model) where the number K refers to the total number of states (i.e., pertaining to $K-1$ HBD classes and 1 non-HBD class). We proposed to estimate for each individual either the K different rates R_c or to set these rates to pre-defined values (so-called MixKR model) (Druet and Gautier, 2017). In the latter case, the rate for the non-HBD class was set equal to the most remote HBD class (i.e., $R_K = R_{K-1}$). In practice, the MixKR modeling facilitates comparisons across different individuals and in the present work we only consider MixKR models. More importantly, the estimated ρ_c mixing coefficients associated to each HBD class c in KR models (with $K > 2$) can no longer be interpreted as inbreeding coefficients as in the single HBD class model. Indeed, although they correspond to the initial HMM state probabilities, the ρ_c values do not correspond to the marginal equilibrium proportions of genomes belonging to each HBD class c because these proportions also depends on the rates R_c that now differ between classes. Nevertheless, several measures related to individual inbreeding coefficients can be obtained from KR models as i) the genome-wide estimate of the realized individual inbreeding level $\widehat{F}_G$, corresponding to the proportion of the genome in HBD classes; ii) the inbreeding level $\widehat{F}_G^{(c)}$ associated with HBD class c defined as the proportion of the genome belonging to class c; and iii) the posterior HBD probability ϕ_m corresponding to the probability that a locus m lies in a HBD segment (Druet and Gautier, 2017).

In addition to the loss of interpretability of mixing coefficients, we previously showed that the MixKR model tended to assign HBD segments to more recent classes (i.e., with smaller R) when the overall inbreeding level of individuals was high (Druet *et al.*, 2020). Although the 1R model remained limited in its range of applications (because it models a single class of ancestors, see above), it provided both an unbiased estimate of R and an estimate of ρ that could be interpreted as an inbreeding coefficient.

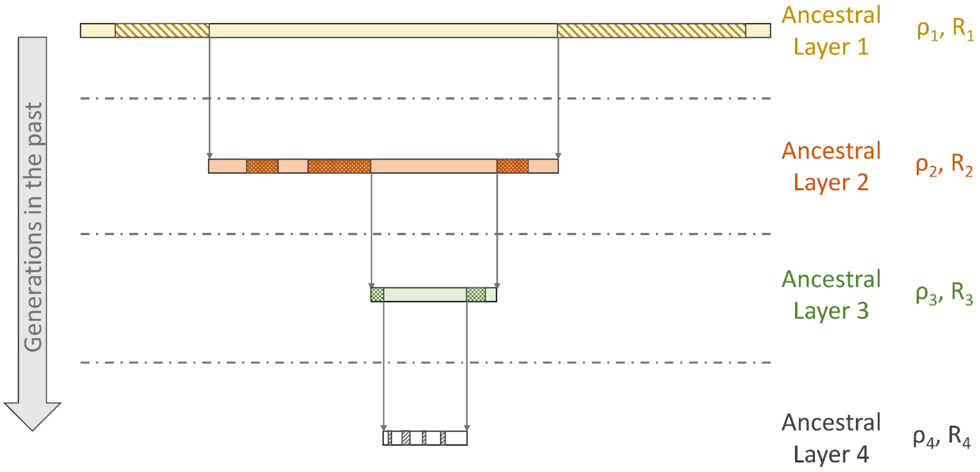

Figure 1. Graphical illustration of the Nested 1R model. Four layers of ancestors are represented. In each layer, the genome is represented as a mosaic of HBD and non-HBD segments with a 1R model with specific parameters ρ_c and R_c. Regions with motives correspond to HBD segments.

2.2 New model: the nested 1R model

Here were propose a modified multiple HBD classes model that preserves the desirable properties of the 1R model and allows for the contribution of multiple groups of ancestors to autozygosity (as in the MixKR model). As illustrated in Figure 1, we sequentially model multiple layers of ancestors (from the most recent to the oldest), each contributing to a distinct HBD class. More precisely, a 1R model is first used to describe the genome of an individual as a mosaic of HBD segments associated with the most recent layer of ancestors (first group of ancestors) and non-HBD segments (i.e., relative to these ancestors). Although these positions would be non-HBD with respect to this first layer, they could be inherited HBD from more remote ancestors. Therefore, we propose to model in turn the non-HBD positions in the first layer as a mosaic of HBD and non-HBD segments associated with a second layer of ancestors (see Figure 1). This would be achieved by fitting a second 1R model, nested in the first one, with different parameters, ρ_2 and R_2 (with $R_2 > R_1$). This approach can be repeated for several layers of ancestors (Figure 1).

Each layer c is thus described as a mosaic of HBD and non-HBD segments, labelled as HBD$_c$ and non-HBD$_c$ states (we use subscript c as layers match with HBD classes). The non-HBD class in layer c would be a mixture of HBD classes in subsequent layers and the non-HBD class in the last layer L (the total number of layers $L = K - 1$, where K is the number of hidden states). We assume that emission probabilities in HBD classes are the same in each layer, and identical to those used in the 1R model (eqn 2). Note that emission probabilities could be made layer dependent, e.g., to account for more generations of mutation or changes in allele frequencies through generations. Similarly, the emission probabilities for the non-HBD class in the last layer L matches those used in the 1R model (eq 3). However for non-HBD positions in layer $c = 1$ to $c = L-1$, the emission probabilities now also depend on the mixing coefficients ρ_c through the proportion $\pi_c = \prod_{i=c+1}^{L} (1 - \rho_i)$ of positions expected to ultimately lie in a non-HBD segment at the oldest layer L (i.e., not mapping to an HBD segment in any successive layers $c' > c$) as:

$$\pi_c \mathbb{P}\left[A_{mi}A_{mj} \mid S_m = 2, p_{mi}, p_{mj}\right] + (1 - \pi_c)\mathbb{P}\left[A_{mi}A_{mj} \mid S_m = 1, p_{mi}\right] \tag{5}$$

where $\mathbb{P}\left[A_{mi}A_{mj} \mid S_m = 2, p_{mi}, p_{mj}\right]$ and $\mathbb{P}\left[A_{mi}A_{mj} \mid S_m = 1, p_{mi}\right]$ are emission probabilities from the 1R model (eqns. 2 and 3).

As the parameters ρ_c for the different classes are required to obtain these emission probabilities, the implementation of this model is not trivial. A more convenient way to specify the Nested 1R model is to define L HBD states (one per layer) and a single non-HBD class associated to the Lth layer. This results in a parameterization very similar to a MixKR model with a number of hidden states $K = L + 1$ (Druet and Gautier, 2017) but with a modified transition probabilities matrix $\mathbf{T^m}$ between consecutive markers m and $m + 1$. More precisely, in the MixKR model, $\mathbf{T^m}$ can be decomposed in three parts i) a diagonal matrix $\mathbf{T}_0^m$ associated with the probability of absence of coancestry change within each of $K = L + 1$ hidden states; ii) a matrix $\mathbf{T}_{cc}^m$ associated with the probability of coancestry change within each state; and iii) a matrix $\mathbf{T}_{cs}$, that does not depend on the marker position, specifying the probability of entering each state after a coancestry change given the state of origin:

$$\mathbf{T}^m = \mathbf{T}_0^m + \mathbf{T}_{cc}^{m\prime}\mathbf{T}_{cs} \tag{6}$$

In the nested 1R model, the matrix $\mathbf{T^m}$ will have a similar structure but the matrices $\mathbf{T}_{cc}^m$ and $\mathbf{T}_{cs}$ in eq. 6 that are defined with respect to states (eq 4) are replaced by matrices $\mathbf{T}_{\chi}^m$ and $\mathbf{T}_C$ that are rather defined with respect to layers as we detail below. As a result, $\mathbf{T^m}$ is decomposed as:

$$\mathbf{T}^m = \mathbf{T}_0^m + \mathbf{T}_\chi^{m\prime}\mathbf{T}_C \tag{7}$$

2.2.1 Transition probabilities in nested 1R models

At marker position m, the genome can be associated with any state c from 1 to K. States from 1 to $K-1$ correspond to HBD segments in layers 1 to $K-1$ ($K-1$ being equal to L), respectively. HBD segments from state c are also non-HBD in layers 1 to $c-1$. The last state K is associated to non-HBD positions in the last layer L, and must also be non-HBD in layers 1 to $L-1$. To estimate the transition probabilities between the K different hidden states, we must consider several possible events:

1. the Markov chain remains in the same state c without any coancestry change. This requires no coancestry change between the two consecutive markers in all the generations included in both the genealogical paths to the ancestors from layer c;

2. the first coancestry change in time along the genealogy occurs within a given layer c (i.e., no coancestry change occurs before this layer, in previous generations). We must then account for both the probability of first coancestry change occurring in c and the conditional transition probabilities to the other states.

In this model, a coancestry change in layer c can be viewed as at least one recombination occurring in that specific layer of ancestors.

2.2.2 Absence of coancestry change from layers 1 to c

In the absence of coancestry change between the two consecutive markers, a HBD segment from a given layer c is simply extended. The same holds for non-HBD regions in layer L (i.e., for the state K). The probability of no coancestry change between markers m and $m+1$ from layers 1 to c is equal to $e^{-R_c d_m}$, as for a 1R model with rate R_c (eqn 1). These transitions can be summarized for all states as a diagonal matrix $\mathbf{T}_0^m$:

$$\mathbf{T}_0^m = \begin{pmatrix} e^{-R_1 d_m} & 0 & \cdots & 0 & 0 \\ 0 & e^{-R_2 d_m} & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & e^{-R_L d_m} & 0 \\ 0 & 0 & \cdots & 0 & e^{-R_L d_m} \end{pmatrix} \tag{8}$$

Note that the probabilities for the last two states ($K-1$ and K) are the same as they both belong to the last layer L.

2.2.3 Probability of first coancestry change occurring within a given layer c

From equation 8, the probability of at least one coancestry change occurring between two consecutive markers m and $m+1$ in the past generations covered by layers 1 to c is $1-e^{-R_c d_m}$. This is in agreement with eqn. 1 for a 1R model with rate R_c. The coancestry change may have occurred in any layers c' ($1 \leq c' \leq c$) but we are interested in the first coancestry change event since it implies the start of a new HBD or non-HBD segments in that layer (and thus also affects the status in subsequent layers).

The probability χ_m^c of a first coancestry change occurring within a specific layer c is equal to the probability of no coancestry change in earlier layers $c' < c$, $e^{-R_{c-1} d_m}$, multiplied by the probability of a coancestry change between layers $c-1$ and c which is equal to $1-e^{-(R_c - R_{c-1}) d_m}$:

$$\chi_m^c = e^{-R_{c-1} d_m}\left(1-e^{-(R_c-R_{c-1})d_m}\right) = e^{-R_{c-1} d_m} - e^{-R_c d_m} \tag{9}$$

Note that χ_m^c is also the probability of no coancestry change from layer 1 to $c-1$ minus the probability of no coancestry change from layer 1 to c. For notational convenience we set $R_0 = 0$ (i.e., the probability of no coancestry change before the first layer is equal to 1). We can further show that the sum of probabilities of first coancestry changes within each layer from 1 to c is equal to $1-e^{-R_c d_m}$ as expected:

$$\sum_{i=1}^{c} \chi_m^i = \sum_{i=1}^{c} \left(e^{-R_{i-1} d_m} - e^{-R_i d_m}\right) = e^{-R_0 d_m} - e^{-R_c d_m} = 1 - e^{-R_c d_m} \tag{10}$$

These probabilities can also be combined in a matrix $\mathbf{T}_\chi^m$, with K columns (for states) and L rows (for layers). The element $\mathbf{T}_\chi^m(c, c')$ represents the probability of first coancestry change within each layer c for a genomic position in an hidden state c' (which is an HBD segment if $c' \leq L$ and a non-HBD segment if $c' = K$):

$$\mathbf{T}_\chi^m(c,c') = \begin{cases} \chi_m^c = e^{-R_{c-1}d_m} - e^{-R_c d_m} & \text{if } c \leq c' \\ 0 & \text{if } c > c' \end{cases} \tag{11}$$

The two last columns of $\mathbf{T}_\chi^m$ both correspond to probabilities of first coancestry changes for genomic positions in states from the last layer, respectively HBD and non-HBD, and are thus identical. When $c > c'$, $\mathbf{T}_\chi^m(c,c')$ is 0 because HBD segments from layer c' are excluded from more ancestral layers in the modelling (as illustrated in Figure 1). In other words, for a HBD segment in layer c', coancestry changes can only occur from layers 1 to c' because historical crossovers in more remote generations cannot interrupt an HBD tract. Thus, $\mathbf{T}_\chi^m$ can be represented as:

$$\mathbf{T}_\chi^m = \begin{pmatrix} \chi_m^1 & \chi_m^1 & \chi_m^1 & \cdots & \chi_m^1 & \chi_m^1 \\ 0 & \chi_m^2 & \chi_m^2 & \cdots & \chi_m^2 & \chi_m^2 \\ 0 & 0 & \chi_m^3 & \cdots & \chi_m^3 & \chi_m^3 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \\ 0 & 0 & 0 & \cdots & \chi_m^L & \chi_m^L \end{pmatrix} \tag{12}$$

As indicated in Eq. 10, elements from the column c' of $\mathbf{T}_\chi^m$ sum to $1 - e^{-R_{c'}d_m}$ for $c' \leq L$. Each column corresponds to the marginal probability of a coancestry change when the marker m is in state c'.

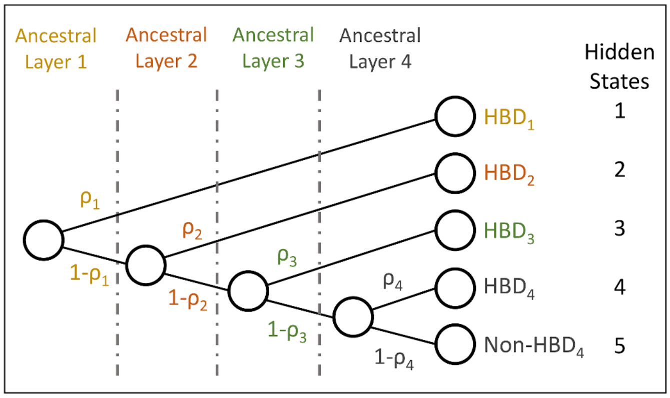

Figure 2. Representation of the transition probabilities in a Nested 1R model with $L = 4$ HBD states and one non-HBD states as a decision tree. In this representation the $K = 5$ states are the leaves of the tree. The tree allows one to estimate probabilities and conditional probabilities to reach a leaf.

2.2.4 Conditional transition probabilities after a coancestry change in layer c

If a (first) coancestry change occurs along the genealogy within a given layer c, the next position is either i) HBD of class c with probability ρ_c; or ii) non-HBD with probability $1-\rho_c$ (Figure 2). These latter non-HBD regions from layer c are also mixture of HBD and non-HBD segments of layer $c+1$ with probabilities ρ_{c+1} and $1-\rho_{c+1}$, respectively. The conditional transition probabilities towards the different HBD states $c' > c$ and the final non-HBD state from the last layer (i.e., state K of the HMM) can then be recursively obtained by following the decision tree represented in Figure 2 (see also Figure 3 for an example of a transition towards the fourth HBD state after a coancestry change in layer $c = 2$). Note that conditional transition probabilities to states $c' < c$ that are not a child of the corresponding node in the decision tree (Figure 2) are null. Thus, the conditional transition probabilities $\mathbf{T}_C(c, c')$ to reach state c' after a coancestry change occurring in layer c are:

$$\mathbf{T}_C(c,c') = \begin{cases} 0 & \text{if } c' < c \\ \rho_c & \text{if } c' = c \\ \left[\prod_{j=c}^{c'-1}(1-\rho_j)\right]\rho_{c'} & \text{if } c < c' \leq L \\ \prod_{j=c}^{L}(1-\rho_j) & \text{if } c' = L+1 = K \end{cases} \tag{13}$$

These conditional transition probabilities can be represented as a matrix $\mathbf{T}_C(c, c')$, independent of the marker position m, with L rows corresponding to layers, and K columns corresponding to the hidden states ($K-1$ HBD states and one non-HBD state):

$$\mathbf{T}_c = \begin{pmatrix} \rho_1 & (1-\rho_1)\,\rho_2 & (1-\rho_1)\,(1-\rho_2)\,\rho_3 & \dots & \left[\prod_{j=1}^{L-1}(1-\rho_j)\right]\rho_L & \prod_{j=1}^{L}(1-\rho_j) \\ 0 & \rho_2 & (1-\rho_2)\,\rho_3 & \dots & \left[\prod_{j=2}^{L-1}(1-\rho_j)\right]\rho_L & \prod_{j=2}^{L}(1-\rho_j) \\ 0 & 0 & \rho_3 & \dots & \left[\prod_{j=3}^{L-1}(1-\rho_j)\right]\rho_L & \prod_{j=3}^{L}(1-\rho_j) \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & \rho_L & 1-\rho_L \end{pmatrix} \tag{14}$$

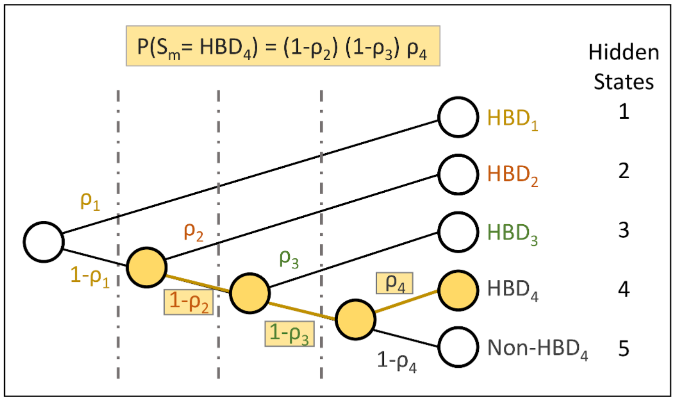

Figure 3. Illustration of conditional transition probabilities after a coancestry change. The illustration shows the conditional transition probability to reach the fourth HBD state after a coancestry change occurring within the second layer.

2.2.5 Initial state probabilities

The first row of $\mathbf{T}_C$ (eqns. 14 and 7) also corresponds to the vector of initial states probabilities $\boldsymbol{\delta} = \{\delta_c\}_{1,...,K}$ (i.e., δ_c representing the probability to start the chain in the hidden state c) which can obtained from the full decision tree (e.g., Figure 2). We have:

$$\delta_c = \begin{cases} \left[\prod_{j=1}^{c-1} (1-\rho_j)\right] \rho_c & \text{if } c < K \\ \prod_{j=1}^{L} (1-\rho_j) & \text{if } c = K \end{cases} \tag{15}$$

We show in Appendix that the product $\boldsymbol{\delta}\mathbf{T}^m = \boldsymbol{\delta}$, i.e., the Markov chain is stationary and the initial state distribution corresponds to the stationary distribution. These desired properties are also true for the 1R model, but not in our previous MixKR model. Note also that, if $L = 1$, the Nested 1R model reduces to the 1R model.

2.2.6 Parameter estimation

HMM are fully defined by their number of states, their initial state probabilities, their transition probabilities and their emission probabilities; elements that were described in previous sections. However, some of these probabilities depend on the model parameters (ρ_c, and R_c for KR models) that need to be estimated. In HMM, this parameter estimation can be performed by numerical maximization of the likelihood with respect to these parameters (Zucchini and MacDonald, 2009). The likelihood of the HMM for a set of parameters is obtained by applying the forward algorithm (Rabiner, 1989). The numerical maximization can be achieved with numerical optimization methods implemented in R packages such as the `optim` function from the R package `stats` (R Core Team, 2013). These methods require only a function that returns the likelihood of the HMM for a given set of parameters. In the RZooRoH package, the L-BFGS-B optimizer was selected for that purpose. Following our previous work (Bertrand *et al.*, 2019), we transformed the original parameters into new unconstrained parameters as advised by Zucchini and MacDonald (2009). For the rates R_c, the following transformation was applied:

$$\eta_c = \begin{cases} \log(R_c - R_{c-1}) & \text{if } 1 < c \leq L \\ \log(R_c) & \text{if } c = 1 \end{cases} \tag{16}$$

Back-transformation on the original scale ensures that rates are always positive and ordered (higher rates for older ancestral layers). Indeed, as the exponential function is always positive we obtain increasing rates:

$$R_c = \begin{cases} R_{c-1} + \exp(\eta_c) & \text{if } 1 < c \leq L \\ \exp(\eta_1) & \text{if } c = 1 \end{cases} \tag{17}$$

For the the mixing coefficients ρ_c, we applied the same transformation as for the MixKR model (Bertrand *et al.*, 2019), which here results in a logit transformation (i.e., $\tau_c = \log\left(\frac{\rho_c}{1-\rho_c}\right)$ for all $c \leq L$) and guarantees that all the estimated ρ_c are comprised between 0 and 1 (since $\rho_c = \frac{\exp(\tau_c)}{1+\exp(\tau_c)}$).

Note that it may also be possible to rely on numerical optimization with constraints instead of transforming the parameters but this is generally not recommended as the constraints might slow down convergence (Zucchini and MacDonald, 2009). We previously tested different optimization functions and obtained the best results with the parameter transformation and `optim` L-BFGS-B algorithm.

The N1R model is now implemented as the default model in the RZooRoH package (from version 0.3.1).

2.2.7 Estimation of the inbreeding coefficient

Following Leutenegger *et al.* (2003), the stationary distribution of the state probabilities $\boldsymbol{\delta}$ (eq. 15) can be used to estimate the inbreeding coefficient. We must first define a reference population by deciding which HBD classes are considered as truly autozygous. We could for instance consider that only layers with a rate R_c smaller than a threshold T contribute to autozygosity, and that ancestors in layers with $R_c > T$ are unrelated (see for instance in Solé *et al.*, 2017). The inbreeding coefficient with respect to that base population, set approximately 0.5×T generations in the past (Druet and Gautier, 2017), is:

$$F_{\delta-T} = \sum_{c=1}^{c'} \delta_c \tag{18}$$

where c' is the most ancient layer with rate $R_{c'} \leq T$. The inbreeding coefficient obtained with all layers is:

$$F_{\delta} = \sum_{c=1}^{L} \delta_c \tag{19}$$

In addition, as opposed to our previous MixKR models, the nested 1R model allows to estimate inbreeding coefficients within each layer. Indeed, the equilibrium probability ρ_c may directly be interpreted as the inbreeding coefficient of the progeny of individuals from the most recent generation of the layer c when individuals from the oldest generation of layer c are assumed unrelated. This coefficient may also be interpreted as the inbreeding accumulated within the time period covered by layer c and may thus be related to the effective population size over this same period. Contrary to the proportion of the genome associated to a specific HBD class, this measure is independent of inbreeding generated in more recent generations.

Metrics defined for the previous MixKR model (Druet and Gautier, 2017) and associated to the realized inbreeding have also their counterpart in the new Nested 1R model. First, the realized inbreeding $\widehat{F}_{\mathrm{G}}^{(\mathrm{c})}$ associated with each HBD class c ($c \in (1, K-1)$) 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 M loci:

$$\widehat{F}_{\mathrm{G}}^{(\mathrm{c})} = \frac{1}{M}\sum_{m=1}^{M} \mathbb{P}\left(S_m = c \mid \widehat{\Theta}, \mathbf{Y}\right) \tag{20}$$

where $\widehat{\Theta}$ and $\mathbf{Y}$ represent respectively the estimated parameters of the model and the data.

Next, 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 M markers:

$$\widehat{F}_{\mathrm{G}} = \frac{1}{M}\sum_{m=1}^{M} \widehat{\phi}_m = \sum_{c=1}^{K-1} \widehat{F}_{\mathrm{G}}^{(\mathrm{c})} \tag{21}$$

The realized inbreeding coefficients can also be estimated relative to different base populations by considering HBD classes with a rate $R_c \leq T$ as in Solé *et al.* (2017).

2.3 Evaluation based on simulated data sets

2.3.1 Simulations under the 1R model

To simulate data sets under the 1R model, we used the same approach as in our first study (Druet and Gautier, 2017). Briefly, we simulated individual genomes consisting of 25 chromosomes of 100 cM. Each individual genome is modeled as a mosaic of HBD and non-HBD segments modelled under the 1R model (Equation 1), where ρ represents the proportion of HBD segments (equivalent to F, the inbreeding coefficient). The length of HBD and non-HBD segments was exponentially distributed with rate R. The tested values for ρ were equal to 0.02, 0.05, 0.10, 0.20, 0.30 and 0.40, and those for R equal to 4, 8, 16, 32 and 64. Genotypes were simulated for 25,000 bi-allelic SNPs (10 per cM) using emission probabilities (Equations 2 and 3). For each set of parameters, we simulated 500 individuals. More details are available in Druet and Gautier (2017).

Individual inbreeding levels were estimated with MixKR and N1R models with 9 HBD classes with rates equal to {2, 4, 8, ..., 512}, and using the RZooRoH package (Bertrand *et al.*, 2019). The mean absolute error (MAE) for each parameter of interest α (F_G, F_δ, ϕ) was computed to evaluate the models as:

$$MAE(\alpha) = \frac{1}{N} \sum_{n=1}^{N} |\widehat{\alpha}_n - \alpha_n| \tag{22}$$

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

The partitioning of the autozygosity in different HBD classes was evaluated by assessing whether the autozygosity was concentrated in HBD classes with rates R_c close to the simulated rate R. Rates were compared on a $\log_2$ scale, resulting in a difference of -1, 0, 1 and 2 when R_c is equal to R multiplied by respectively 0.5, 1, 2 and 4. The associated MAE was estimated as follows:

$$MAE(\log_2(R)) = \frac{1}{N} \sum_{n=1}^{N} \sum_{c=1}^{K-1} \widehat{\Psi}_n^{(c)} |\log_2 R_c - \log_2 R| \tag{23}$$

where $\widehat{\Psi}_n^{(c)}$ is the contribution of HBD class c in individual n to its total HBD (computed as $\widehat{F}_{\mathrm{G}}^{(\mathrm{c})}/\widehat{F}_{\mathrm{G}}$

estimated at true HBD positions), and $K-1$ is the number of HBD classes. This criteria evaluates whether the identified HBD positions are assigned to the simulated HBD class.

2.3.2 Simulations under a discrete-time Wright–Fisher process

To simulate more realistic data relying on population genetic models, Druet and Gautier (2017) previously used the program ARGON (Palamara, 2016) that implements a discrete-time Wright-Fisher process. Here, we used the same simulated data sets. Bottlenecks were simulated to concentrate inbreeding in specific age classes (Druet and Gautier, 2017). Outside these events, N_e was kept large to reduce the noise due to inbreeding coming from other generations. The main simulation scenario is summarized in Supplementary Figure 1. The ancestral population P_0 had a constant haploid effective population size equal to 20,000 (N_{e0}). The time of population split T_s was set equal to 10,000 and the effective population size of the first population (P_1) outside the bottleneck was set to 100,000 (N_{e1}). Bottlenecks were simulated around generations T_b equal to either 16 or 64, and with effective population size (N_{eb}) equal to 20 or 50. A single chromosome of 250 cM length was simulated for 50 diploid individuals, and with a marker density of 100 SNPs per cM. More details about the simulation procedure are available in Druet and Gautier (2017). We further simulated data sets under three additional scenarios to evaluate more complex demographic history. The first two scenarios consisted of two successive bottlenecks (with N_{eb} being either equal to 20 or 50) either closely related (around generations 16 and 64) or more distant (around generations 16 and 128). The third scenario consisted of a continuous population expansion following a bottleneck simulated around generation 16, the final N_e being ten times larger than the bottleneck one.

Individual inbreeding levels were estimated with MixKR and N1R models with 13 HBD classes with rates equal to {2, 4, 8, ..., 8192} as implemented within the RZooRoH package (Bertrand *et al.*, 2019).

2.3.3 Application to estimation of inbreeding levels in the European bison

The N1R model was tested and compared to the MixKR model on a set of 183 genotyped European bison with high inbreeding levels (Druet *et al.*, 2020). These consisted of respectively 154 and 29 individuals from the Lowland and Lowland-Caucasian lines. Individuals from the first line experienced a stronger bottleneck as they trace back to fewer founders (see Druet *et al.*, 2020, for more details). After excluding monomorphic SNPs, those with a call-rate <0.95 or deviating from Hardy-Weinberg equilibrium (p ≤

0.001), each individual was genotyped for 22,602 autosomal SNPs (see Druet *et al.* (2020) for more details). This represents a marker-density below 10 SNPs per cM but we evaluated based on simulations and whole-genome sequence data that it allowed to capture recent autozygosity (Druet *et al.*, 2020). Partitioning of inbreeding levels in different HBD classes was first compared with MixKR and N1R models with five HBD classes with rates equal to {4, 8, 16, 32, 64}. In order to assess robustness of results to model specifications, we also applied models with 9 HBD classes ($R_c = \{4, 8, \ldots, 1024\}$). Analyses were carried out with the RZooRoH package (Bertrand *et al.*, 2019).

3 Results

3.1 Simulations under the 1R model

We begin by comparing results obtained from analyses of the data simulated under the 1R model with the MixKR and N1R models. We expected our new N1R model to perform better in partitioning inbreeding in different HBD classes most particularly when inbreeding levels are high. This is confirmed in Figure 4 that represents the MAE associated with R (eq. 23). With the N1R model, the MAE is higher when there are fewer segments to estimate parameters (small ρ and/or small R). When inbreeding levels are low (e.g., $\rho < 0.1$ in Figure 4), MAE are similar for both models whereas for large inbreeding levels, MAE starts to increase for the MixKR model whereas it continues to decrease for the N1R model. As a result, the proportions of true HBD positions associated with the class with R_c corresponding to the simulated value is higher with the N1R than with the former MixKR model when values of ρ are moderate to high (i.e., $\rho > 0.1$). In other words, a higher proportion of true autozygosity is correctly associated to its underlying HBD class of origin with the N1R model (see Supplementary Table 1). More precisely, these proportions range from 33% up to 84% and actually increased with ρ and R (i.e., the number of HBD segments available for parameter estimation). Note that in case of mis-classification, the HBD segments are most often assigned to neighboring classes as illustrated for instance in Supplementary Figure 2. As for the MAE, these proportions decrease for high values of ρ with the MixKR model whereas an opposite trend is observed for the N1R model, resulting in high differences. When ρ is high, the MixKR model tends to assign autozygosity to classes with smaller R_c rates, as we observed in real data sets. This is illustrated for four scenarios with $R = 8$ in Supplementary Figure 2. Similar patterns are obtained in simulations with two distributions of HBD segments (Supplementary Figure 3), with a shift towards more recent

HBD classes when using the MixKR model. Finally for comparison with ROH-based methods, we also identified ROH with likelihood-based approaches (Pemberton *et al.*, 2012; Szpiech *et al.*, 2017) and then assigned these ROH to different classes equivalent to our HBD-classes (see an example in Supplementary Figure 2). We found that the ROH-class that captured the largest proportion of the genome was not the ROH-class including segments of length L=100/2G (the expected length of ROH segments). With higher simulated inbreeding levels, the distribution in different ROH-classes changed and a larger proportion of the genome was found associated with even longer ROH. The inferred distribution of ROH segments was thus sensitive to the overall inbreeding levels as for our previous MixKR model.

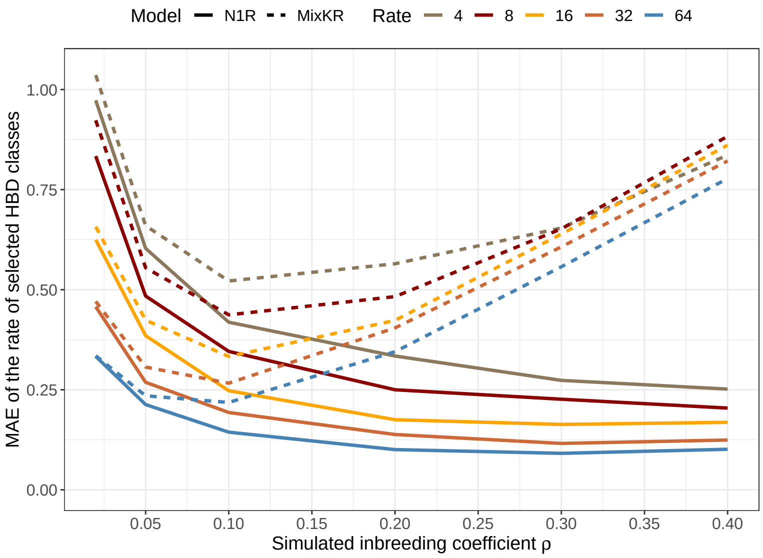

Figure 4. Concordance between simulated rates and partitioning in HBD classes on data sets simulated under the 1R model. The accuracy of partitioning is evaluated as the Mean Absolute Error between the $\log_2$ of the simulated rate and the $\log_2$ of the assigned HBD classes. This is equivalent to measuring the deviation from the simulated parameter in term of absolute value of $\log_2$ of the ratio between rates of simulated and estimated HBD classes. The comparisons are performed for different values of R and ρ.

In terms of estimation of realized inbreeding (F_G) and estimation of local HBD probabilities (ϕ_m), both the MixKR and N1R models showed very similar performances (Table 1). Hence, although these two models differ in their partitioning of inbreeding in different age classes, they remain equally accurate for the estimation of inbreeding levels. Finally, the inbreeding coefficient F_δ corresponding to the sum of initial state probabilities (the stationary distributions) and closely related to ρ displayed a low MAE, close to the values obtained with a $1R$ model in our previous study (Druet and Gautier, 2017). With the N1R model, the inbreeding coefficient F_δ represents an unbiased estimate of the simulated ρ (Supplementary Figure 4), as opposed to the sum of initial state probabilities of HBD classes from the MixKR models that were clearly not a proper estimator of ρ.

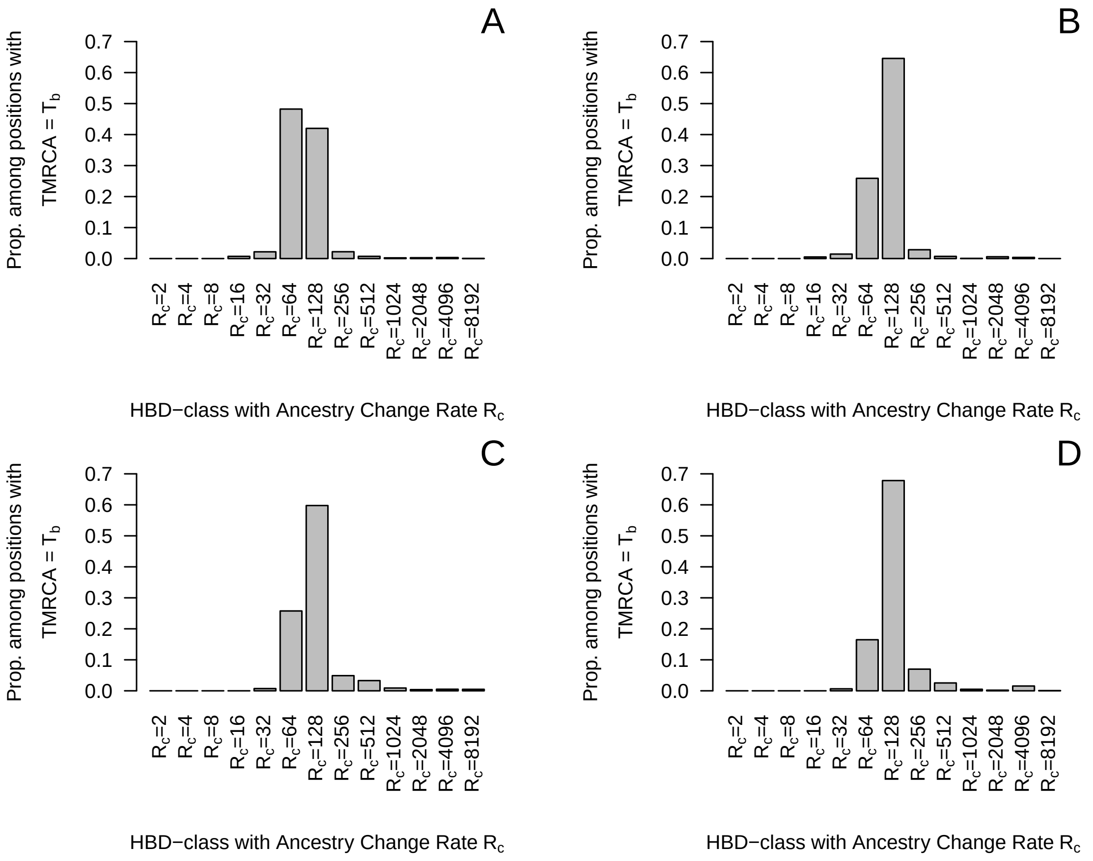

Figure 5. Partitioning of HBD segments related to the bottleneck in different HBD classes. Data were simulated with a Wright-Fisher process, with a bottleneck in generations 63 to 66 expected to be associated with the HBD class with $R_c = 128$. The simulated N_e during the bottleneck is equal to 20 (A & B) or 50 (C & D). The partitioning is realized with the MixKR (A & C) or N1R (B & D) models.

3.2 Simulations under Wright-Fisher process

Analyses realized on data sets simulated under a more realistic model confirm our first observations. For high inbreeding levels, the MixKR model captures a large fraction of the autozygosity generated by the bottleneck (when N_e drops to 20) into the more recent HBD class neighboring the class representative of the bottleneck period (e.g., class with $R_c = 64$ for a bottleneck pertaining to the class with $R_c = 128$,

Scenario		Mean estimated values with N1R model			MixKR model	
R	ρ	$\widehat{F_\delta}$ (MAE)	$\widehat{F_{\text{G}}}$ (MAE)	MAE for $\widehat{\phi_m}$ ($\widehat{\phi_{m^{\text{HBD}}}}$)	$\widehat{F_{\text{G}}}$ (MAE)	MAE for $\widehat{\phi_m}$ ($\widehat{\phi_{m^{\text{HBD}}}}$)
4	0.02	0.021 (0.007)	0.021 (0.002)	0.002 (0.012)	0.021 (0.002)	0.002 (0.013)
4	0.05	0.053 (0.012)	0.054 (0.002)	0.003 (0.011)	0.054 (0.002)	0.003 (0.013
4	0.10	0.103 (0.017)	0.103 (0.002)	0.004 (0.010)	0.103 (0.002)	0.004 (0.012)
4	0.20	0.200 (0.023)	0.200 (0.002)	0.005 (0.009)	0.200 (0.002)	0.005 (0.010)
4	0.30	0.302 (0.026)	0.301 (0.002)	0.006 (0.008)	0.301 (0.002)	0.007 (0.009)
4	0.40	0.401 (0.028)	0.402 (0.002)	0.007 (0.006)	0.402 (0.002)	0.007 (0.008)
8	0.02	0.023 (0.007)	0.022 (0.002)	0.003 (0.026)	0.022 (0.002)	0.003 (0.027)
8	0.05	0.051 (0.011)	0.052 (0.002)	0.004 (0.025)	0.052 (0.002)	0.004 (0.027)
8	0.10	0.101 (0.014)	0.101 (0.002)	0.006 (0.023)	0.101 (0.002)	0.006 (0.024)
8	0.20	0.204 (0.019)	0.204 (0.002)	0.009 (0.019)	0.204 (0.002)	0.010 (0.020)
8	0.30	0.299 (0.021)	0.299 (0.002)	0.011 (0.016)	0.299 (0.002)	0.012 (0.017)
8	0.40	0.404 (0.023)	0.404 (0.002)	0.012 (0.013)	0.403 (0.002)	0.013 (0.014)
16	0.02	0.023 (0.006)	0.022 (0.002)	0.004 (0.064)	0.022 (0.002)	0.004 (0.065)
16	0.05	0.052 (0.008)	0.052 (0.002)	0.007 (0.055)	0.052 (0.002)	0.007 (0.056)
16	0.10	0.102 (0.011)	0.102 (0.002)	0.012 (0.048)	0.102 (0.002)	0.012 (0.050)
16	0.20	0.201 (0.015)	0.201 (0.002)	0.017 (0.040)	0.201 (0.002)	0.018 (0.041)
16	0.30	0.302 (0.017)	0.302 (0.003)	0.022 (0.032)	0.302 (0.003)	0.022 (0.034)
16	0.40	0.403 (0.018)	0.403 (0.002)	0.023 (0.027)	0.403 (0.002)	0.024 (0.028)
32	0.02	0.022 (0.005)	0.021 (0.002)	0.007 (0.139)	0.021 (0.002)	0.007 (0.140)
32	0.05	0.052 (0.006)	0.052 (0.003)	0.014 (0.120)	0.052 (0.003)	0.014 (0.121)
32	0.10	0.101 (0.008)	0.101 (0.002)	0.022 (0.103)	0.101 (0.002)	0.023 (0.105)
32	0.20	0.202 (0.011)	0.202 (0.003)	0.034 (0.081)	0.202 (0.003)	0.035 (0.083)
32	0.30	0.302 (0.012)	0.302 (0.003)	0.042 (0.066)	0.302 (0.003)	0.042 (0.067)
32	0.40	0.402 (0.014)	0.402 (0.003)	0.045 (0.053)	0.402 (0.003)	0.046 (0.055)
64	0.02	0.022 (0.004)	0.022 (0.003)	0.013 (0.283)	0.022 (0.003)	0.013 (0.284)
64	0.05	0.052 (0.005)	0.052 (0.003)	0.026 (0.243)	0.052 (0.003)	0.026 (0.244)
64	0.10	0.102 (0.007)	0.102 (0.003)	0.043 (0.204)	0.102 (0.003)	0.043 (0.205)
64	0.20	0.202 (0.008)	0.202 (0.003)	0.066 (0.160)	0.202 (0.003)	0.066 (0.161)
64	0.30	0.301 (0.010)	0.302 (0.003)	0.079 (0.128)	0.302 (0.003)	0.080 (0.129)
64	0.40	0.402 (0.010)	0.402 (0.003)	0.084 (0.103)	0.403 (0.003)	0.086 (0.104)

Table 1. Performance of the two models on data simulated under the 1R 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 model for each of 30 different scenarios defined by the simulated R and ρ values reported in the first two columns. The table reports the mean estimated values and the Mean Absolute Errors (MAE) for the mixing proportions ρ estimated as $\widehat{F_\delta}$ and the individual realized inbreeding levels ($\widehat{F_{\text{G}}}$). The table gives also the MAE for the estimated local inbreeding (ϕ_m) either for all the genotypes ($\widehat{\phi_m}$) or for genotypes at HBD positions ($\widehat{\phi_{m^{\text{HBD}}}}$). These values are reported for both models, with the exception of $\widehat{F_\delta}$.

i.e., occurring 63 to 66 generations ago - Figure 5). This neighbouring class captures almost the same or even a larger fraction of autozygosity that the HBD class associated with the bottleneck. The pattern is less pronounced for milder bottleneck (N_e =50 in Figure 5). With the N1R model, the class $R_c = 128$ representative of the bottleneck period captures the majority of the HBD segments in both cases. Similar results were obtained for more recent bottlenecks (Supplementary Figure 5).

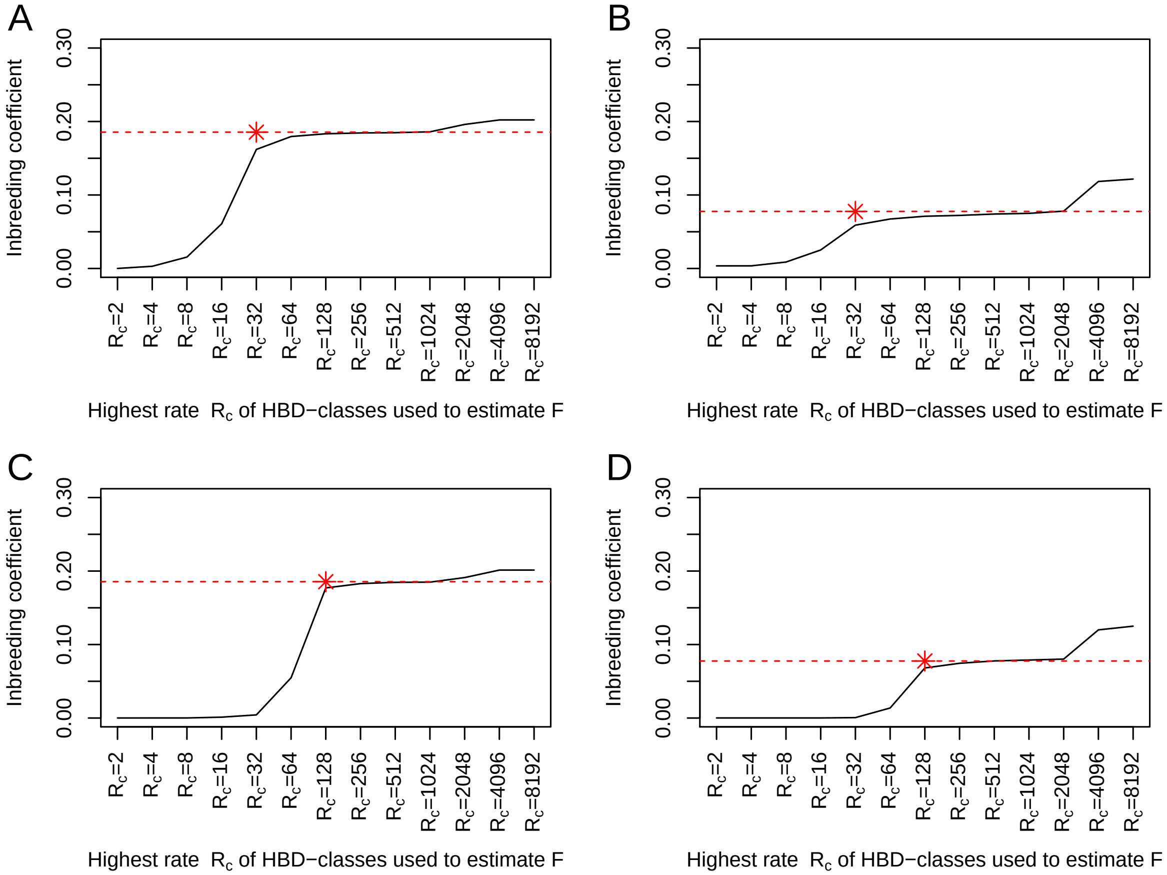

Figure 6. Inbreeding coefficients estimated as the equilibrium HBD distribution and for different base generations. The inbreeding coefficients $F_{\delta-T}$ are estimated from the equilibrium distributions from the different HBD classes and are obtained from their mixing coefficients ρ_c (see eq. 18)). Only HBD-classes with a rate $R_c \leq$ a threshold T are used to estimate $F_{\delta-T}$. This allows to set the reference population approximately $0.5 \times T$ generations in the past. Data were simulated with a Wright-Fisher process, with a bottleneck. The time of the bottleneck (T_b) and the effective population size during the bottleneck N_{eb} are A) $T_b = 16$ and $N_{eb} = 20$, B) $T_b = 16$ and $N_{eb} = 50$, C) $T_b = 64$ and $N_{eb} = 20$, D) $T_b = 64$ and $N_{eb} = 50$. The red star indicates the HBD-class associated to the bottleneck and the expected inbreeding levels generated during the bottleneck.

The global partitioning of the genome in HBD-classes presents similar patterns (Supplementary Figure

6). As the proportion of inbreeding in the HBD class associated with the bottleneck is always higher with the N1R model, the MAE associated with the rate of the selected HBD classes is lower than with the MixKR model (more so when the bottleneck was strong). With the N1R model, the MAE values are respectively equal to 0.546, 0.786, 0.386 and 0.426 for the four different scenarios ($\{N_{eb} = 20$, $T_b = 16\}$,$\{N_{eb} = 50$, $T_b = 16\}$,$\{N_{eb} = 20$, $T_b = 64\}$,$\{N_{eb} = 50$, $T_b = 64\}$), compared to 0.763, 0.793, 0.601 and 0.491 for the same scenarios with the MixKR model.

As for the simulations under the 1R model, the differences between models are mainly in the partitioning of autozygosity in HBD classes. For instance, the average local HBD probabilities for positions associated with ancestors present in different past generations are almost identical (Supplementary Figure 7). These local HBD probabilities indicate that HBD positions associated with ancestors up to 80 generations ago are identified with high confidence, and that for more remote ancestors (shorter segments), it is more difficult to identify unambiguously HBD positions. We also confirm in Figure 6 that mixing coefficients of the new model are interpretable and can be used to estimate the inbreeding coefficient F_δ. More precisely, we estimated $F_{\delta-T}$ by adding sequentially each HBD-class in the estimation. We estimated the expected inbreeding accumulated during the bottleneck as $1-(1-\frac{1}{2N_e})^t$, where N_e is the diploid effective population size (here, $N_e = 20$ or $N_e = 50$) and $t = 4$ is the number of generations of the bottleneck. We see that most of the inbreeding is captured by the HBD-class corresponding to the bottleneck and its close neighbours. As a result, $F_{\delta-T}$ remains relatively constant for generations before and after the bottleneck and increases sharply during the bottleneck. In addition, the estimated inbreeding levels match the expected values. We also observe inbreeding related to much more distant ancestors, accumulated over many more generations that was captured by the most remote HBD classes (e.g., $R_c \geq 2048$). Note that the ability of both models to capture ancient bottlenecks obviously depends on the marker density as we previously showed (Druet and Gautier, 2017). For instance, with a marker density of 10 SNPs per cM, a bottleneck occurring 16 generations ago would be fully captured whereas HBD-segments from a later bottleneck (G=64) would only be partially captured.

When autozygosity results from more complex demographic histories such as multiple bottlenecks or gradual population expansion after a bottleneck, it becomes more difficult to determine precisely which generations contribute to autozygosity and to classify HBD positions as illustrated in Supplementary Figures 8 to 13. For instance, in the presence of two relatively distant bottlenecks (in number of generations), the inferred contributions of the different HBD classes looks bimodal (Supplementary Figures 8 and 9), with the modes corresponding to the two classes representative of the bottleneck. However, the shortest segments of the recent bottleneck might sometimes be assigned to the older classes and vice versa. When the two bottleneck are closer, we no longer observe clear bimodality of the estimated distribution of HBD contributions, and the highest contribution is associated to a class located between the two classes representative of the bottleneck (Supplementary Figures 10 and 11). We previously and more comprehensively studied the properties of the MixKR model when several classes contributed to autozygosity and observed similar trends, with difficulties to disentangle contributions from close HBD classes (Druet and Gautier, 2017). As expected, in the scenario with population expansion following a bottleneck, the distribution of HBD classes contribution tends to be shifted towards classes representative of more recent times than the bottleneck, i.e., with $R_c < 32$ (Supplementary Figures 12 and 13). Likewise, the shift is more pronounced for smaller N_{eb} (i.e., for higher bottleneck intensity). This is likely related to the non negligible contribution to inbreeding of generations that immediately follow the bottlenecks (when N_e is still small) (e.g., those directly following the bottleneck which might have a substantial contribution). The classification was indeed better when all the autozygosity was associated to the bottleneck as shown for scenarios with a rapid expansion after the bottleneck (Figure 5 and Supplementary Figure 5). Nevertheless, both models still indicate periods of increased contribution to inbreeding, from the start of bottlenecks until the time period when the population has recovered. Interestingly, the estimated contributions from HBD classes to autozygosity tend to be less shifted towards more recent classes with the N1R than with the MixKR model.

3.3 Application to real data

Application of the two models on genotype data from two distinct lines of European bison, presenting high inbreeding levels, results in similar observations than applications to simulated data sets: partitioning of inbreeding in HBD-class is shifted towards more recent HBD-classes with the MixKR model compared to the N1R model (Figure 7A-B). Since for simulations the N1R performed better for the partitioning in HBD-classes, and since patterns are similar, the results from the N1R model fit probably better the reality. The shift is more pronounced when more HBD-classes are included in the MixKR model and the non-HBD class has consequently a higher rate R_K, and in the Lowland line where the inbreeding levels are higher. Higher shifts for higher inbreeding levels were also observed with simulated data. With the MixKR model, the partitioning in different HBD-classes and the estimated mixing coefficients

(Figure 7C-D) change according to the model specifications, whereas the N1R model proves robust to these changes (Figures 7A-D). Note that we also fitted HBD-classes corresponding to HBD segments shorter than the shortest HBD segments than could be captured with the available density. As a result, the contribution of these classes remains null. As for the simulated data sets, the overall inbreeding levels estimated by MixKR and N1R models remain highly similar (Figure 7E-F), the difference being essentially in the partitioning.

Analysis of real data with the N1R model confirms that mixing coefficients can now be interpreted, with levels close to estimated HBD proportions in different classes, contrary to those obtained with the MixKR model (Figure 7C-D). In addition, they can now be used to estimate the inbreeding coefficients, F_δ or $F_{\delta-T}$. These inbreeding coefficients based on the equilibrium distribution and on the number of HBD segments are close to values of the realized inbreeding coefficient, F_G and F_{G-T}, corresponding to the proportion of the genome in HBD classes (Figures 7E-F). The mixing coefficients estimate the proportion of HBD segments within a specific layer and provide an estimation of the inbreeding accumulated in that layer, which depends also on the number of generations included in the layer.

When inbreeding levels are lower, such as in cattle (see for instance in Solé *et al.* (2017)), differences are smaller. This is illustrated in Supplementary Figure 14 on a Holstein data set including 245 individuals genotyped for 30,000 markers (Alemu *et al.*, 2021).

4 Discussion

We herein proposed an improved model, we called the N1R model, for the characterization of individual genomic inbreeding levels and its partitioning into different HBD-classes. Compared to our previous MixKR model (Druet and Gautier, 2017), the main improvement relied on a new modelling of the transition probabilities which both resulted in better statistical properties in general, but also facilitated the interpretation of the mixing coefficients with initial state probabilities now corresponding to the stationary distributions. Although the estimation of both global and local inbreeding levels were almost identical between the N1R and the MixKR models, the partitioning of inbreeding into different HBD-classes was clearly improved and the N1R model provided more accurate estimation of the relative contribution of each group of ancestors.

Our main objective was indeed to improve this partitioning, in particular for high inbreeding levels

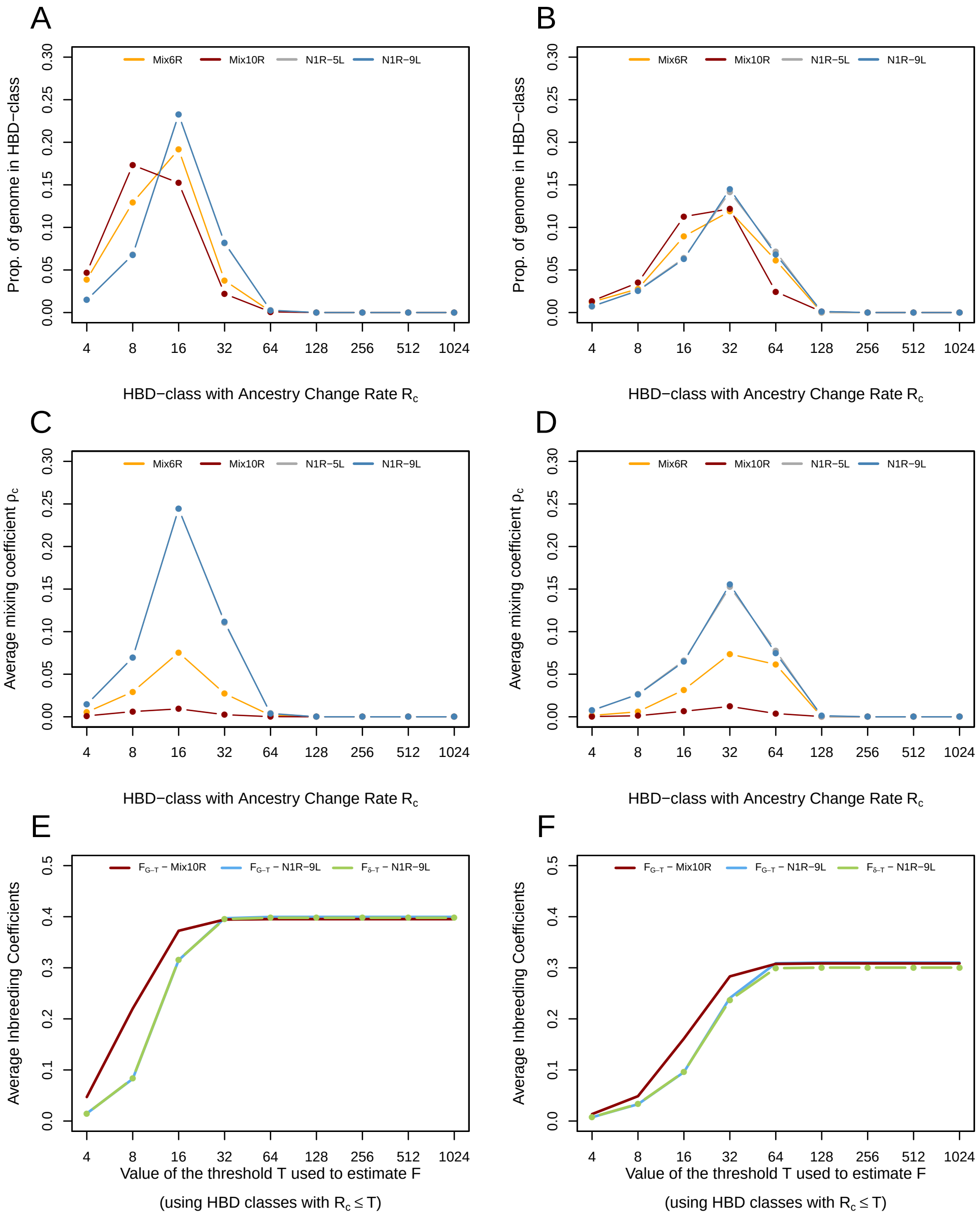

Figure 7. Estimation of inbreeding levels in the European bison. Inbreeding levels are estimated in 154 Lowland individuals (panels A-C-E) and 29 Lowland-Caucasian individuals (panels B-D-F). Estimation was performed with the MixKR and N1R models with 5 HBD-classes (Mix6R and N1R-5L) or with 9 HBD-classes (Mix10R and N1R-9L). A) and B) Proportion of the genome associated with different HBD-classes averaged over all individuals from a population (see eq. 20). C) and D) Estimated mixing coefficients ρ_c for each HBD class, averaged over all individuals. E) and F) Average estimated inbreeding levels. F_{G-T} is estimated as a the sum of contributions from the HBD classes with a rate $R_c \leq$ a threshold T, and $F_{\delta-T}$ as the sum of equilibrium distributions of the same HBD classes (see eq. 18). This allows to set the reference population approximately $0.5 \times T$ generations in the past. A cumulative curve is obtained by changing the values of T.

since we previously observed that in such cases, the partitioning could be shifted towards more recent HBD classes (Druet *et al.*, 2020). This problem was caused in our previous MixKR model by the difference of rates for HBD classes associated with recent ancestors (i.e., capturing large HBD segments) and the non-HBD class that resulted in high differences in their underlying mixing coefficients. More precisely, the non-HBD class had a very high mixing coefficient because it generally represented the main contribution to individual genomes and it was modelled with a large R_c. Conversely, mixing coefficients from recent HBD classes (long segments with low rates R_c) were very small as these segments were much less numerous than short HBD segments. Therefore, in the Markov chain, the probability to start a new recent HBD segment was extremely low and needed to be supported by long stretches of homozygous genotypes. In these conditions, two consecutive recent HBD segments were systematically modelled as a single long HBD segments because transitions to new recent HBD segments were heavily penalized, explaining the overestimation of segment length and the incorrect HBD partitioning (a shift towards more recent HBD classes). Yet, the strength of this problem was expected to be a function of the frequency of consecutive HBD segments, and was thus only observed in simulated and real data sets with high recent inbreeding levels (Druet *et al.*, 2020). We here showed that using the same rates for HBD and non-HBD states by modelling sequentially multiple nested 1R models in our new N1R model allowed to solve this issue. This property is important to better interpret the results by determining which generations of ancestors mostly contributed to autozygosity. Our improved N1R model should also allow better estimation of the number of generations to the common ancestor for an HBD position. Nevertheless, more work is required to quantify how precisely the age of individual HBD segments can be estimated with this or other similar approaches.

The new model is also more robust to the number and specifications (i.e., rates R_c) of the fitted classes in the sense that partitioning remains consistent when the rate of the non-HBD classes is modified. With our previous MixKR model, the choice of the rates associated with the non-HBD classes, often directly related to the number of fitted classes, might indeed influence the partitioning in HBD and non-HBD classes because higher rates (smaller segments) resulted in even higher mixing coefficients for the non-HBD class further penalizing the occurrence of two consecutive recent HBD segments (see above). The fact that the N1R model is less sensitive to model specifications is an important aspect because one of the advantages of methods relying on HMM (Leutenegger *et al.*, 2003; Vieira *et al.*, 2016; Narasimhan *et al.*, 2016; Druet and Gautier, 2017) is that fewer parameters need to be defined compared to rule-based

ROH approaches, where these definitions might sometimes be arbitrary. In general, there is less need to optimize parameters, HBD probabilities indicate whether the evidence for autozygosity is strong or not. In our model, the number of classes and their range must still be defined but it affects mainly interpretation in terms of age of ancestors. To this respect, the robustness of the N1R model is highly valuable since in the previous MixKR model partitioning could be affected by the definition of the last HBD class.

Our newly developed N1R model allows the definition of new inbreeding coefficients based on the initial state probabilities. These inbreeding coefficients fit closer to the original definition by Leutenegger *et al.* (2003) since under the 1R model, the mixing coefficient can be interpreted as both the frequency of HBD segment and the proportion of the genome that is HBD (i.e., the equilibrium distribution). Yet, this is slightly different from a direct estimation of the realized proportion of the genome in HBD segments (e.g., as obtained from the posterior HBD probability of each marker, see eq. 21), although both estimators are highly correlated. Interestingly, the mixing coefficients also provide direct estimators of the level of inbreeding associated with ancestors present in a specific period of time (corresponding to a layer in our model), independently on what happened in other more recent layers. In an ideal population, this inbreeding would directly be related to the number of generations and to the effective population size in the layer. These aspects must be further investigated and more work is required to understand which generations are captured by a specific layer, or the relationship with the underlying historical N_e. Indeed, generations do not map unambiguously to a single layer but are captured by several layers in a probabilistic framework. In practice, the variation of mixing coefficients across layers could be used to monitor whether inbreeding is increasing or not, for instance in a conservation program as suggested by Druet *et al.* (2020).

Comparisons of our previous MixKR and our new N1R models on genotyping data from European bison were in agreement with trends observed on simulated data. The overall inbreeding levels were similar with both models but the partitioning was different, shifted towards more recent HBD classes with the MixKR model. This shift was also more pronounced when inbreeding levels were higher and when the rate of the non-HBD class was higher, matching our predictions (see above). This suggests that the new partitioning is more accurate, strengthening our initial conclusions that the contribution from the most recent generations of ancestors to inbreeding is decreasing and that the restoration plan has been successful to control inbreeding in European bison (Druet *et al.*, 2020).

Finally, it is important to note that differences between our new N1R version of the model and the former MixKR one in terms of interpretation only concern the partitioning of inbreeding when inbreeding levels are high. For instance, differences would be minimal in most human populations. Even in cattle presenting moderate inbreeding levels, the impact on the partitioning remained limited.

5 Acknowledgements

We thank the three anonymous reviewers for their valuable and constructive comments that helped us improving the original version of the manuscript. This work was supported by the Fonds de la Recherche Scientifique–FNRS under grants J.0134.16 and J.0154.18. Tom Druet is Senior 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.–FNRS.

References

Abney, M., C. Ober, and M. S. McPeek, 2002 Quantitative-trait homozygosity and association mapping and empirical genomewide significance in large, complex pedigrees: fasting serum-insulin level in the hutterites. The American Journal of Human Genetics **70**: 920–934.

Alemu, S. W., N. K. Kadri, C. Harland, P. Faux, C. Charlier, A. Caballero, and T. Druet, 2021 An evaluation of inbreeding measures using a whole-genome sequenced cattle pedigree. Heredity **126**: 410–423.

Bertrand, A. R., N. K. Kadri, L. Flori, M. Gautier, and T. Druet, 2019 Rzooroh: An r package to characterize individual genomic autozygosity and identify homozygous-by-descent segments. Methods in Ecology and Evolution **10**: 860–866.

Broman, K. W. and J. L. Weber, 1999 Long homozygous chromosomal segments in reference families from the centre d'Etude du polymorphisme humain. Am J Hum Genet **65**: 1493–500.

Ceballos, F. C., P. K. Joshi, D. W. Clark, M. Ramsay, and J. F. Wilson, 2018 Runs of homozygosity: windows into population history and trait architecture. Nature Reviews Genetics **19**: 220.

Crow, J. F., M. Kimura, *et al.*, 1970 An introduction to population genetics theory. An introduction to population genetics theory. .

Druet, T. and M. Gautier, 2017 A model-based approach to characterize individual inbreeding at both global and local genomic scales. Molecular ecology **26**: 5820–5841.

Druet, T., K. Oleński, L. Flori, A. R. Bertrand, W. Olech, M. Tokarska, S. Kaminski, and M. Gautier, 2020 Genomic footprints of recovery in the european bison. Journal of Heredity **111**: 194–203.

Kirin, M., R. McQuillan, C. S. Franklin, H. Campbell, P. M. McKeigue, and J. F. Wilson, 2010 Genomic runs of homozygosity record population history and consanguinity. PloS One **5**: e13996.

Leutenegger, A.-L., A. Labalme, E. Génin, A. Toutain, E. Steichen, F. Clerget-Darpoux, and P. Edery, 2006 Using genomic inbreeding coefficient estimates for homozygosity mapping of rare recessive traits: application to taybi-linder syndrome. The American journal of human genetics **79**: 62–66.

Leutenegger, A. L., B. Prum, E. Genin, C. Verny, A. Lemainque, F. Clerget-Darpoux, and E. A. Thompson, 2003 Estimation of the inbreeding coefficient through use of genomic data. American Journal of Human Genetics **73**: 516–23.

Magi, A., L. Tattini, F. Palombo, M. Benelli, A. Gialluisi, B. Giusti, R. Abbate, M. Seri, G. F. Gensini, G. Romeo, *et al.*, 2014 H 3 m 2: detection of runs of homozygosity from whole-exome sequencing data. Bioinformatics **30**: 2852–2859.

McQuillan, R., A.-L. Leutenegger, R. Abdel-Rahman, C. S. Franklin, M. Pericic, *et al.*, 2008 Runs of homozygosity in european populations. American Journal of Human Genetics **83**: 359–372.

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

Palamara, P. F., 2016 ARGON: fast, whole-genome simulation of the discrete time Wright-fisher process. Bioinformatics **32**: 3032–4.

Pemberton, T. J., D. Absher, M. W. Feldman, R. M. Myers, N. A. Rosenberg, and J. Z. Li, 2012 Genomic patterns of homozygosity in worldwide human populations. American Journal of Human Genetics **91**: 275–292.

Purcell, S., B. Neale, K. Todd-Brown, L. Thomas, M. A. Ferreira, D. Bender, J. Maller, P. Sklar, P. I. de Bakker, M. J. Daly, and P. C. Sham, 2007 PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet **81**: 559–75.

R Core Team, 2013 *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0.

Rabiner, L. R., 1989 A tutorial on hidden markov models and selected applications in speech recognition. In *PROCEEDINGS OF THE IEEE*, pp. 257–286.

Renaud, G., K. Hanghøj, T. S. Korneliussen, E. Willerslev, and L. Orlando, 2019 Joint estimates of heterozygosity and runs of homozygosity for modern and ancient samples. Genetics **212**: 587–614.

Ringbauer, H., J. Novembre, and M. Steinrücken, 2021 Parental relatedness through time revealed by runs of homozygosity in ancient dna. Nature communications **12**: 1–11.

Solé, M., A.-S. Gori, P. Faux, A. Bertrand, F. Farnir, M. Gautier, and T. Druet, 2017 Age-based partitioning of individual genomic inbreeding levels in belgian blue cattle. Genetics Selection Evolution **49**: 1–18.

Szpiech, Z. A., A. Blant, and T. J. Pemberton, 2017 Garlic: genomic autozygosity regions likelihood-based inference and classification. Bioinformatics **33**: 2059–2062.

Vieira, F. G., A. Albrechtsen, and R. Nielsen, 2016 Estimating ibd tracts from low coverage ngs data. Bioinformatics **32**: 2096–2102.

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

Weir, B. S., A. D. Anderson, and A. B. Hepler, 2006 Genetic relatedness analysis: modern data and new challenges. Nature Reviews Genetics **7**: 771–780.

Zucchini, W. and I. MacDonald, 2009 Hidden markov models for time series, volume 110 of monographs on statistics and applied probability.

A Appendix

Here we show that in the N1R model, the Markov chain is stationary and the initial state distribution corresponds to the stationary distribution, i.e.:

$$\boldsymbol{\delta}\mathbf{T}^{m} = \boldsymbol{\delta}\left(\mathbf{T}_{0}^{m} + \mathbf{T}_{\chi}^{m\prime}\mathbf{T}_{C}\right) = \boldsymbol{\delta} \tag{24}$$

where $\boldsymbol{\delta}$ is a row vector of dimension K. Let the (row) vector $\zeta = \{\zeta_{\mathbf{c}}\}_{\mathbf{1},..,\mathbf{K}} = \boldsymbol{\delta}\mathbf{T^{m}}$. We want to show that $\zeta_c = \boldsymbol{\delta}\left(\boldsymbol{t^{m}_{0,c}} + \boldsymbol{t^{m}_{C\chi,c}}\right) = \delta_c$ for all c $\in (1, K)$, where $\boldsymbol{t^{m}_{0,c}}$ is the cth column vector of $\mathbf{T_0}^{m}$ and $\boldsymbol{t^{m}_{C\chi,c}}$ is the cth column vector of the matrix $\mathbf{T}_{\chi}^{m\prime}\mathbf{T}_{C}$:

$$\boldsymbol{t^{m}_{C\chi,c}} = \begin{pmatrix} \chi_m^1 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 \\ \chi_m^1 & \chi_m^2 & \cdots & 0 & 0 & 0 & \cdots & 0 \\ \chi_m^1 & \chi_m^2 & \cdots & 0 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \chi_m^1 & \chi_m^2 & \cdots & \chi_m^{c-1} & 0 & 0 & \cdots & 0 \\ \chi_m^1 & \chi_m^2 & \cdots & \chi_m^{c-1} & \chi_m^{c} & 0 & \cdots & 0 \\ \chi_m^1 & \chi_m^2 & \cdots & \chi_m^{c-1} & \chi_m^{c} & \chi_m^{c+1} & \cdots & 0 \\ \vdots & \vdots\ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \\ \chi_m^1 & \chi_m^2 & \cdots & \chi_m^{c-1} & \chi_m^{c} & \chi_m^{c+1} & \cdots & \chi_m^{L} \\ \chi_m^1 & \chi_m^2 & \cdots & \chi_m^{c-1} & \chi_m^{c} & \chi_m^{c+1} & \cdots & \chi_m^{L} \end{pmatrix} \times \begin{pmatrix} \left[\prod_{j=1}^{c-1}(1-\rho_j)\right]\rho_c \\ \left[\prod_{j=2}^{c-1}(1-\rho_j)\right]\rho_c \\ \vdots \\ (1-\rho_{c-1})\,\rho_c \\ \rho_c \\ 0 \\ \vdots \\ 0 \\ 0 \end{pmatrix} = \rho_c \begin{pmatrix} \sum_{i=1}^{1}\chi_m^i\left[\prod_{j=i}^{c-1}(1-\rho_j)\right] \\ \sum_{i=1}^{2}\chi_m^i\left[\prod_{j=i}^{c-1}(1-\rho_j)\right] \\ \vdots \\ \sum_{i=1}^{c-1}\chi_m^i\left[\prod_{j=i}^{c-1}(1-\rho_j)\right] \\ \sum_{i=1}^{c}\chi_m^i\left[\prod_{j=i}^{c-1}(1-\rho_j)\right] \\ \sum_{i=1}^{c}\chi_m^i\left[\prod_{j=i}^{c-1}(1-\rho_j)\right] \\ \vdots \\ \sum_{i=1}^{c}\chi_m^i\left[\prod_{j=i}^{c-1}(1-\rho_j)\right] \\ \sum_{i=1}^{c}\chi_m^i\left[\prod_{j=i}^{c-1}(1-\rho_j)\right] \end{pmatrix} \tag{25}$$

To simplify notations in the above equation, we assume that $\prod_{j=c}^{c-1}(1-\rho_j) = 1$. Still to keep notations general, for $c = K$ we define $\rho_K = 1 - \rho_{K-1}$. Note also that elements $c' \geq c$ of $\boldsymbol{t^{m}_{C\chi,c}}$ are all identical.

Hence,

$$
\begin{aligned}
\zeta_c &= \boldsymbol{\delta t^m_{0,c}} + \boldsymbol{\delta t^m_{C\chi,c}} \\
&= \delta_c e^{-R_c d_m} + \rho_c \sum_{c'=1}^{K} \left(\delta_c' \sum_{i=1}^{\min\left(c,c'\right)} \chi_m^i \left[\prod_{j=i}^{(c-1)} (1-\rho_j) \right] \right) \\
&= \delta_c e^{-R_c d_m} + \rho_c \sum_{i=1}^{c} \left(\chi_m^i \left[\prod_{j=i}^{c-1} (1-\rho_j) \right] \sum_{c'=i}^{K} \delta_c' \right) \\
&= \delta_c e^{-R_c d_m} + \rho_c \sum_{i=1}^{c} \left(\chi_m^i \left[\prod_{j=1}^{c-1} (1-\rho_j) \right] \right)
\end{aligned}
$$

The last equality follows from the nested model properties which consider each layer sequentially (see the main text and Figure 2). Hence, $\sum\limits_{c'=i}^{K} \delta_{c'}$ can be interpreted as the probability of starting a layer as old or older than i which is also the probability of not having entered any of the successive layers more recent than i i.e. $\sum\limits_{c'=i}^{K} \delta_{c'} = \prod\limits_{j=1}^{i-1} (1-\rho_j)$. Note also that $\sum\limits_{c'=1}^{K} \delta_{c'} = 1$. In addition, recalling that $\delta_c = \rho_c \prod\limits_{j=1}^{c-1} (1-\rho_j)$ (eq. 15) and $\sum\limits_{i=1}^{c} \chi_m^i = 1 - e^{R_c d_m}$ (eq. 10), we obtain:

$$
\begin{aligned}
\zeta_c &= \delta_c e^{-R_c d_m} + \rho_c \sum_{i=1}^{c} \left(\chi_m^i \left[\prod_{j=1}^{c-1} (1-\rho_j) \right] \right) \\
&= \delta_c e^{-R_c d_m} + \rho_c \left[\prod_{j=1}^{c-1} (1-\rho_j) \right] \sum_{i=1}^{c} \chi_m^i \\
&= \delta_c e^{-R_c d_m} + \delta_c \left(1 - e^{-R_c d_m} \right) \\
&= \delta_c
\end{aligned}
$$