

## **Genomic prediction of individual inbreeding levels for the management of genetic diversity in populations with small effective size**

Natalia Soledad Forneris<sup>1†</sup>, Mirte Bosse<sup>2,3</sup>, Mathieu Gautier<sup>4</sup> and Tom Druet<sup>1</sup>

<sup>1</sup>Unit of Animal Genomics, GIGA-R & Faculty of Veterinary Medicine, University of Liège, Liège, Belgium

<sup>2</sup>Animal Breeding and Genomics, Wageningen University & Research, Wageningen, The Netherlands

<sup>3</sup>Amsterdam Institute for Life and Environment (A-LIFE), Section Ecology and Evolution, Vrije Universiteit Amsterdam, Amsterdam, The Netherlands

<sup>4</sup>CBGP, INRAE, CIRAD, IRD, L'institut Agro, Université de Montpellier, Montpellier, France

†Corresponding author: [nforneris@uliege.be](mailto:nforneris@uliege.be)

## Abstract

In populations of small effective size ( $N_e$ ), such as those in conservation programs, companion animals or livestock species, management of diversity and inbreeding is essential. Homozygosity-by-descent (HBD) segments provide relevant information in that context, as they allow efficient estimation of the inbreeding coefficient, provide locus-specific information and their length is informative about the “age” of inbreeding. Therefore, our objective was to evaluate tools for predicting HBD in future offspring based on parental genotypes, a problem equivalent to identifying segments identical-by-descent (IBD) among the four parental chromosomes. In total, we reviewed and evaluated 16 approaches using simulated and real data with small  $N_e$ . The methods included model-based approaches, mostly hidden Markov models (HMMs), which considered up to 15 IBD configurations among the four parental chromosomes, as well as more computationally efficient rule-based approaches. The accuracy of the methods was then evaluated, including with low-density marker panels, genotyping-by-sequencing data and small groups of individuals, typical features in such populations. Two HMMs performed consistently well, while two rule-based approaches proved efficient for genome-wide predictions. The model-based approaches were particularly efficient when information was reduced (low marker density, locus-specific estimation). Methods using phased data proved to be more efficient, while some approaches relying on unphased genotype data proved to be sensitive to the allele frequencies used. In some settings, pedigree information was competitive in predicting recent inbreeding levels. Finally, we showed that our evaluation is also informative about the accuracy of the methods for estimating relatedness and identifying IBD segments between pairs of individuals.

## Introduction

In conservation genetics, companion animals and livestock species, effective population sizes ( $N_e$ ) are often small and optimal management strategies are implemented to maintain genetic diversity and limit levels of inbreeding, often relying on reproducer selection and mating advice. The expected level of inbreeding of future offspring is an important criterion for such mating advice, as it allows the risk of inbreeding depression (ID) or genetic defects to be reduced.

The inbreeding coefficient ( $F$ ) of an individual has been defined as the correlation between the uniting gametes (Wright, 1922) and the probability that the two alleles present at a locus are identical-by-descent (IBD), i.e. inherited twice from a common ancestor (Malécot, 1948). In that case, the neighboring loci will also be IBD because a whole segment has been inherited IBD from the founder. In the absence of mutations, these segments are homozygous and are therefore called homozygous-by-descent (HBD) (Schäffer, 1999). The length of these HBD segments depends on the number of generations  $G$  to the common ancestor, with more generations providing more opportunities for the recombination process to cut the transmitted segment. With genotyping data, these HBD segments will appear as long stretches of homozygous genotypes called runs-of-homozygosity (ROH), which are often used as a proxy for HBD and to estimate  $F$  (Broman and Weber, 1999; McQuillan et al., 2008). Model-based approaches have also been implemented to describe individual genomes as a mosaic of HBD and non-HBD segments (Leutenegger et al., 2003), allowing identification of HBD segments, calculation of HBD probabilities at each marker position, and estimation of realized inbreeding levels. Such model-based approaches have been shown to be more efficient than rule-based methods when information is degraded or marker density is low (Druet and Gautier, 2017; Lavanchy and Goudet, 2023).

ROH-based ( $F_{ROH}$ ) and HBD-based ( $F_{HBD}$ ) estimators of  $F$  present several advantages over other marker-based estimators. Indeed, studies from Nietlisbach et al. (2019) and Caballero et al. (2021) showed that  $F_{ROH}$  performs well in populations with low  $N_e$ . In agreement, Alemu et al. (2021) concluded that HBD-based methods are efficient in livestock species where deleterious alleles can reach high frequencies (e.g., Keller and Waller, 2002; Bosse et al., 2019). In addition, these methods provide locus-specific estimates, which can be used to manage recessive alleles that cause genetic defects or have a large contribution to ID. They are also informative about the age of HBD segments (Kirin et al., 2010; Pemberton et al., 2012), thus allowing to estimate the recent inbreeding that is expected to be deleterious (Hinrichs et al., 2007; Szpiech et al., 2013; Stoffel et al., 2021; Naji et al., 2024) and therefore more relevant for management strategies. They are also more robust to the used allele frequencies (AF), that might introduce biases (Keller et al., 2011; Caballero et al., 2022; Naji et al., 2024). Finally, these estimators are more interpretable as they range between 0 and 1 as the pedigree-based estimators, and allow to define a base population comparable to the pedigree (Solé et al., 2017). In summary, the use of measures based on HBD segments offers several advantages for managing diversity and inbreeding in populations with small  $N_e$ , as in conservation genetics, some wildlife species, companion animals and livestock populations. Actually, the benefits of using segment-based measures to maintain diversity and fitness have been demonstrated in similar populations (de Cara et al., 2013; Bosse et al., 2015; Gómez-Romano et al., 2016), for example by minimizing the occurrence of long HBD segments in offspring (Bosse et al., 2015). Similarly, Meuwissen et al. (2020) concluded that the use of IBD-based approaches performed well in optimal contribution selection schemes.

In the present study, our objective was to evaluate different methods for predicting inbreeding levels in future offspring based on genotypes from their parents, as these are an important component in mating plans implemented to manage diversity and inbreeding in populations with small  $N_e$ . More specifically, our aim was to predict future HBD levels, as these metrics have been proven to be efficient

and offer several advantages (see above). Therefore, we reviewed methods for identifying IBD segments between parental chromosomes and predicting HBD levels in offspring, and selected state-of-the-art model- and rule-based approaches to perform a comprehensive evaluation study. Importantly, this evaluation was performed on real data sets from a large sequenced cattle pedigree and a population of Mexican wolves, similar to those typically found in the field of conservation genetics, animal breeding or molecular ecology. It is indeed essential to measure their accuracy in such settings because the genomic structure of the population has been shown to affect their performance (e.g. Caballero et al., 2021), while most of the methods have only been evaluated on human datasets characterized by large  $N_e$ . Simulated data with similar population structures were also generated to consolidate our results. Accuracy was also assessed at lower marker density or with genotyping-by-sequencing (GBS) data, genotyping strategies that are frequently used in these populations. We start by a description of the different evaluated methods.

## Material and methods

### Overview of approaches to predict homozygosity-by-descent in offspring

Our objective is to predict the level of inbreeding of an individual based on the genotypes of its parents (Figure 1A). More specifically, the aim is to predict the proportion of HBD segments and to estimate the probability that a locus will be HBD. The expected inbreeding coefficient of an individual is equal to the kinship coefficient between its two parents (Malécot, 1967). Similarly, the prediction of HBD can be obtained from the probability that the haplotypes transmitted by the parents are IBD. Each parent can transmit two haplotypes, so there are four possible combinations of pairs of parental haplotypes at a locus, and the HBD probability can be computed from the IBD probabilities of these four combinations (Figure 1). Note that due to recombination, the transmitted haplotypes might differ from the parental haplotypes. Nevertheless, the expected HBD level can still be obtained as the average IBD of the four parental combinations (recombination rates and interference levels do not affect the expected inbreeding levels but their variance).

The methods we will evaluate are therefore based on the modeling of the IBD relationships between the four chromosomes of two individuals (here, the four parental homologues) that can take on fifteen different configurations (Figure 1B), called the identity states (Jacquard, 1974). If the parental origins of the haplotypes within the two individuals are unknown, symmetric configurations are equivalent and the 15 identity-states can be grouped into 9 condensed identity-states (Jacquard, 1974). The configurations can be further reduced to three states if the inbreeding of the individuals is ignored. In this case, the three states simply correspond to the sharing of 0, 1 and 2 IBD haplotypes between the two individuals. Accordingly, several methods model the observed genotypes or haplotypes conditionally on these possible configurations (referred to as 15, 9 or 3-STATES approaches). Another approach is to model chromosomes individually, ignoring which chromosomes belong to the same individual. In this case, methods can either model the IBD relationship between each possible pair of haplotypes sequentially (PAIR approach – Figure 1C), leading to the analysis of four pairs when working with two individuals, or simultaneously model all haplotypes from all individuals in the analyzed sample and efficiently identify IBD sharing in this panel of haplotypes (PANEL approach – Figure 1D).

Prediction methods can be further classified according to whether they use genotypes or haplotypes. The use of genotypes is only possible in the 9 or 3-STATES approaches by modeling the genotype probabilities conditionally on the underlying state. Haplotypes require phasing of the data, a procedure that provides additional information, but that can also introduce errors.

Finally, we can also distinguish between rule-based and model-based approaches. Rule-based approaches use a set of rules to determine whether two individuals share IBD haplotypes based on their genotypes or whether two haplotypes are IBD. These rules are typically based on the number of identical-by-state (IBS) alleles or genotypes, the length of the segments compared, or the number of mismatches. Model-based approaches compute the probability of different configurations based on the likelihood of observing the data conditional on the configuration. The possible configurations could, for example, correspond to the 9 identity states in a 9-STATES model, or to IBD versus non-IBD for a PAIR approach. Estimation of these likelihoods may include parameters such as the AF, the probability of genotyping error, the genetic distance between successive markers, etc. Model-based approaches can also handle genotype probabilities to account for genotype uncertainty, e.g. with low-fold sequencing data.

## Description of the evaluated IBD estimation methods

**STATES approaches.** IBD\_Haplo (Thompson, 2008; 2009), GIBDLD (Han and Abney, 2011; 2013) and LocalNgsRelate (Severson et al., 2022) model the IBD process along the four parental chromosomes based on continuous time Markov chains that can be implemented as HMM. In these three methods, the emission probabilities are based on AF and on genotyping error probabilities or genotyping uncertainty, and the transition probabilities are a function of the genetic distances. Through local decoding, they estimate the posterior state probabilities,  $P(X_i=k)$ , at each locus  $i$ , where  $X_i$  is the unknown state at locus  $i$  and  $k$  is one of the modeled states. With genotyping data, these hidden states correspond to the 9 condensed identity states with IBD\_Haplo and GIBDLD, and to the sharing of 0, 1 or 2 IBD with LocalNgsRelate (3-STATES model). With haplotype data, IBD\_Haplo fits a model with 15 hidden states corresponding to the 15 identity-states. At each locus, the offspring HBD can be predicted from the locus-specific state probabilities  $P(X_i=k)$  using the rules shown in Figure 1. The three methods define the transition matrices differently. For example, in IBD\_Haplo they are based on two parameters related to  $N_e$  and the number of generations to the common ancestors. The model assumes that only one additional IBD relationship can be gained or lost between two successive markers. These parameters are not estimated and are identical for all pairs of individuals. Conversely, GIBDLD and LocalNgsRelate estimate the parameters for each pair of individuals. These parameters include the proportions of the genome associated with the different condensed identity-states, which also provide a genome-wide estimate of relatedness. They also present differences in the emission probabilities. GIBDLD accounts for LD by allowing the emission probabilities at a given locus to depend on the genotypes at a defined number of preceding loci. LocalNgsRelate is designed to work with low depth sequencing data, a feature made possible by using genotype likelihoods to define emission probabilities. Like LocalNgsRelate, TRUFFLE (Dimitromanolakis *et al.*, 2019) fits a 3-STATES model but uses a rule-based approach.

**PAIR approaches.** When modeling IBD for pairs of haplotypes, it is possible to use methods designed to identify ROH or HBD segments within individual genomes. Although this is somewhat artificial, since the two haplotypes don't really belong to the same individual, it has the advantage that they behave like the HBD measures we want to predict in the offspring. This strategy has been used, for example, with ROH in several studies (Pryce et al., 2012; Bosse et al., 2015; de Cara et al., 2013). The ZooRoH model (Druet and Gautier, 2022) is an alternative to rule-based approaches to identify ROH. It is an HMM that describes the genome as a mosaic of HBD and non-HBD segments. A specific feature of ZooRoH is that it fits several HBD classes, each class  $c$  having its own rate parameter  $R_c$  that defines the expected length of HBD segments (equal to  $1/R_c$  Morgans). Each class is therefore associated with a different set of ancestors present in different past generations. This makes it possible to estimate the level of inbreeding with respect to different base populations (Solé et al., 2017). Note that when a single HBD class is fitted, we refer to a ZooRoH-1R model. This model is identical to that of Leutenegger et al. (2003) and has only two parameters, a rate and a mixing coefficient, both estimated for each individual.

**PANEL approaches.** GERMLINE (Gusev et al., 2009), hap-IBD (Zhou et al., 2020) and phasedibd (Freyman et al., 2021) belong to the PANEL approach and use rule-based methods to find long segments shared IBS between two haplotypes. These three methods avoid comparing each pair of haplotypes sequentially to improve their computational efficiency. For this purpose, GERMLINE is based on hash tables and libraries of haplotypes, while hap-IBD uses the Positional Burrows-Wheeler Transform (PBWT; Durbin, 2014) and phasedibd relies on an extension of the PBWT called the Templated PBWT, which allows to mask errors and expand haplotype matches. Although they share a

common principle, these methods differ in their implementation and their handling of genotyping and haplotyping errors. For example, hap-IBD and phasedibd can account for genotyping errors and even correct for some so-called switch errors resulting from the phasing process. Finally, Refined-IBD (Browning and Browning, 2013) first relies on GERMLINE to identify long segments shared IBS, and then uses a LOD score to determine whether the haplotypes are IBD or not.

**Pedigree-based and genomic relationship matrices.** The genome-wide inbreeding coefficient can also be predicted as half the relationship between the parents obtained using either genealogical information or genotyping data, SNP-by-SNP, and the rules described in VanRaden (2008). Depending on the selected rules, using the elements from the genomic relationship matrix (GRM) is equivalent to predicting  $F_{GRM}$  (VanRaden, 2008) or the correlation between the uniting gametes,  $F_{UNI}$  (Li and Horvitz, 1953; Yang et al., 2010).

## Evaluation design

The objective of the present study is to evaluate the accuracy of predicting HBD in future offspring based on parental genotypes in populations with small  $N_e$ . For this evaluation we used both simulated and real data sets. With simulations, the true HBD levels are known, whereas with real data, the methods are evaluated in more realistic data structures and conditions. Therefore, we used a design that can be applied to both datasets, taking advantage of the genotyped trios available for the real data. In this design, we use the parental genotypes to perform the predictions and the offspring genotypes to evaluate their accuracy (Figure S1). The genotypes of the offspring provide the possibility to estimate its realized inbreeding. In particular, with sequence data, genotypes are available for all variants, allowing accurate estimation of whole-genome heterozygosity (Kardos et al., 2016; Alemu et al., 2021).

The accuracy was then assessed using the correlations between the predicted and reference genome-wide levels of HBD (defined as the average locus-specific HBD levels). For locus-specific predictions, the correlations between predicted and reference locus-specific levels of HBD were computed, and ROC curves associated with these predictions were derived.

## Software and parameters used for prediction of homozygosity-by-descent

For evaluation, offspring genotyping data were first removed from the analysis. Parental genotypes were then phased using Beagle 5.4 (Browning et al., 2021) and default parameters. The software used to predict global and locus-specific HBD levels in offspring based on parental genotypes or haplotypes are listed in Table 1 and Table S1. For IBD\_Haplo, GIBDLD, LocalNGsRelate and TRUFFLE, we essentially used default parameters (see Table S1). IBD\_Haplo (Thompson, 2008) was run either with genotypes (IBD\_Haplo9c), either with haplotypes (IBD\_Haplo15c) and the transition matrix described in Brown et al. (2011). For RZooRoH (Bertrand et al., 2019), we used a 'layer' model (Druet and Gautier, 2022) with 6 HBD classes with predefined rates  $R_c = \{5, 25, 125, \dots, 15,625\}$ . We then estimated IBD between each pair of parental haplotypes using posterior probabilities from HBD classes with  $R_c \leq 25$  (recent IBD – ZooRoH-25) or from HBD classes with  $R_c \leq 125$  (total IBD or ZooRoH-125). The last HBD classes were not included as they are less reliable (only a few SNPs per segment) and more dependent on AF like maximum likelihood estimators (Alemu et al., 2021). In addition, we also ran a model estimating the rate of a single HBD class (ZooRoH-1R). ROH were identified using PLINK (Purcell et al., 2007) with 40 SNP windows and minimum length of 2 Mb. We also used default parameters for Refined-IBD, GERMLINE and hap-IBD, while the minimum segment length was set to 2 Mb for phasedibd (Table S1).



By default, methods use AF estimated from the sample, although ideally the AF of the base population should be used. Therefore, we also tested methods that accept external AF as input (IBD\_Haplo, ZooRoH, LocNgsRelate, GRM and UNI) with founder AF estimated using the gene content approach of Gengler et al. (2007). SNPs with a founder MAF less than 0.005 were discarded.

## Data

**Simulation study.** We simulated two populations with small  $N_e$  in the most recent generations. These demographic histories, characterized by successive reductions in  $N_e$ , were chosen to obtain populations with similar levels of inbreeding and genomic structure to the two real datasets (see below). The scenarios resulted in moderate and high levels of inbreeding, termed MODF and HIGHF, corresponding to the levels observed in a typical livestock population (dairy cattle) and an endangered wild population (the Mexican wolf).

Simulations were performed with SLiM 4.0.1 (Haller and Messer, 2023) and msprime 1.2.0 (Baumdicker et al., 2022) using forward-time simulation and a “recapitation” technique (Haller et al., 2018). For both populations, individual genomes consisted of 25 chromosome pairs of 100 cM each. Recombination and mutation rates were set to  $10^{-8}$  per bp. The initial population consisted of 10,000 diploid individuals, mating was assumed to be random and the sex ratio was set to 1. Subsequent evolution of demographic parameters is described in Figure 2. In total, 100 and 20 offspring were simulated in the last generation in the MODF and HIGHF scenarios respectively. The simulated data consisted of the trios formed by these offspring and their parents. In each replicate, we randomly selected a subset of 5,000 or 25,000 evenly spaced bi-allelic markers with a MAF  $\geq 0.01$  (corresponding to low and medium-density arrays with 1 SNP / cM and 10 SNPs / cM). In addition, we kept track of the pedigree for the last 15 generations. Each scenario was repeated 100 times.

We used the tree sequence recording feature to keep track of the true local ancestry at each marker position for each individual. At each position, HBD was then declared if the time to the most recent common ancestor (TMRCA) was less than the age of the defined base population expressed in generations. We used three different base populations to define the reference levels of inbreeding, corresponding respectively to recent inbreeding only (young base population set to 15 generations -  $F_{\text{YOUNG}}$ ), to 50 generations of inbreeding (intermediate base population -  $F_{\text{MID}}$ ), and to a large number of generations including ancient inbreeding (ancient base population set to 500 generations -  $F_{\text{TOT}}$ ).

**Dutch Holstein data set.** The first real data set represents a population typical of livestock species, with a small  $N_e$  and under intense selection. We used whole-genome sequence data (8,417,679 SNPs) available for 264 Dutch Holstein individuals from the DAMONA cattle pedigree previously used by Oget-Ebrad et al. (2022). This pedigree can be divided into 127 parents, corresponding to 84 different pairs, and their 98 offspring. Offspring belonging to a sequenced trio that were also parents in another trio, were excluded from the list of sequenced parents. The pedigree available for these individuals consisted in 12,238 individuals.

The reference inbreeding measures were computed using ZooRoH and the whole-genome sequence data because inbreeding estimation with ZooRoH proved highly accurate on the simulated data (Table S2). In addition, the model has been proven efficient and robust to genotyping errors in previous studies. We ran a ‘layer’ model (Druet and Gautier, 2022) with 6 HBD classes with fixed rates  $R_c = \{5, 25, 125, 625, 3125, 15,625\}$  on the sequence data from the offspring only. Recent inbreeding  $F_{\text{YOUNG}}$  was defined using only posterior HBD probabilities of classes with  $R_c \leq 25$ , while  $F_{\text{MID}}$  and  $F_{\text{TOT}}$  were obtained



using HBD classes with  $R_c \leq 125$  and  $R_c \leq 3125$ , respectively. These correspond approximately to base populations set at 12.5, 62.5 and 1500 generations in the past.

**Mexican Wolf data set.** We also evaluated the accuracy of the methods in a population with higher levels of inbreeding and under a conservation program. It consisted of 13 trios from the endangered Mexican Wolf (MW, *Canis lupus baileyi*) population genotyped using the Illumina CanineHD BeadChip array (Illumina, Inc., San Diego, CA). The MW population was reintroduced into the wild in the 1990s. It is derived from three unrelated captive lineages, called McBride (MB), Aragón (AR) and Ghost Ranch (GR), each descended from 2 or 3 founders. The trios were extracted from a data set (Fitak *et al.*, 2018) including 88 MW genotyped for 118,287 SNPs. One individual was excluded for having more than 10% Mendelian conflicts. In addition to the original filtering, we kept only markers with call rate  $\geq 0.90$ , MAF  $\geq 0.01$ , without Mendelian conflicts and in Hardy-Weinberg equilibrium ( $p \geq 0.05$ ). The final set included 33 MB, 2 AR, 6 GR and 46 crossbred (i.e. crosses between two lineages) individuals genotyped for 54,037 SNPs. Of these crossbred individuals, 13 individuals and their genotyped parents formed the trios used in our validation experiment. The reference inbreeding measures were computed using all the selected SNPs and the same approach as for the DAMONA data set.

### Impact of marker density and genotyping method

The methods were evaluated with different marker panels, such as low and medium genotyping densities (LD and MD arrays). For the simulated data sets, these corresponded to densities of 1 and 10 SNPs per Mb. For the DAMONA data set, we selected markers in common with the Illumina BovineLD and BovineSNP50 commercial arrays. This resulted in a selection of 5,388 and 29,375 SNPs respectively (approximately 2 and 10 SNPs per Mb). In addition, the DAMONA sequence data allowed us to mimic genotyping-by-sequencing (GBS) data. To do this, we performed an *in silico* digestion of the bovine reference genome using the *Pst*I restriction enzyme using the GBSX package (Herten *et al.*, 2015). We selected 72,828 fragments out of 1,503,470 fragments with a length between 200 and 300 bp. The distance between fragments is shown in Figure S2. A total of 60,842 SNPs from the WGS data were located in the selected fragments. To account for allelic dropout (Gautier *et al.*, 2013), when a SNP was located near the restriction site ( $\pm 3$ bp), heterozygous individuals were considered homozygous in the associated fragment for the haplotype carrying the reference alleles (the other haplotype was not amplified). If individuals were homozygous for the alternate alleles of the SNP in the restriction site, all genotypes in the associated fragment were set to missing. We then filtered out SNPs surrounding the restriction site ( $\pm 3$ bp), those with more than 5% missing genotypes or MAF  $< 0.01$ . This resulted in a GBS panel of 56,098 SNPs (GBS-50K). To further reduce the number of markers, we first selected fragments with a length ranging between 250 and 300 bp, resulting in a panel of 31,339 SNPs (GBS-30K). A smaller panel of 15,493 SNPs (GBS-15K) was obtained by randomly sampling half of these fragments. Finally, we used only one marker density for the MW (54,037 SNPs, corresponding to about 20-25 SNPs per Mb). In addition to their variable marker densities, the different marker panels differ in the distribution of their AF and of their marker spacing (Figure S3). These elements could also influence the properties of the prediction methods.

## Results

The simulated data closely matched the inbreeding levels and partitioning of HBD observed in the corresponding real data (Figure S4). Comparisons also illustrate that dairy cattle and Mexican wolf populations, and the corresponding MODF and HIGH simulations, provide complementary scenarios.

### Accuracy of predicted genome-wide HBD levels

We first evaluated the predictions obtained using MD genotyping arrays. The correlations between predicted and reference levels of HBD for the 16 methods compared and the four scenarios are shown in Figure 3 and Figure S5 (significance levels of pairwise comparisons are available in Table S3). The reference levels of HBD were defined with respect to different base populations,  $F_{\text{YOUNG}}$ ,  $F_{\text{MID}}$  and  $F_{\text{TOT}}$ , corresponding to approximately 15, 50 and 500 generations. In the simulations, the correlations were high, ranging from 0.78 to 0.86 and from 0.74 to 0.83 in the MODF and HIGHF scenarios, respectively. The ranking of the methods was comparable in the two simulation scenarios and for the different base populations. Several methods, including IBD\_Haplo15c, ZooRoH, phasedibd, PLINK ROH and hap-IBD, were consistently among the best methods. Using unphased data and a 9-STATES model with IBD\_Haplo (IBD\_Haplo9c) also performed well, but was generally less accurate than using phased data and a 15-STATES model. For ZooRoH, only minor differences were observed between the three models tested. Among the three best rule-based approaches, the correlations obtained with hap-IBD were generally slightly lower than those obtained with phasedibd or PLINK-ROH. On the other hand, GIBDLD and LocalNgsRelate achieved lower correlations in the model-based group, while this was the case for GERMLINE and TRUFFLE in the rule-based group. Refined-IBD had one of the lowest correlations in most scenarios. As with GIBDLD and LocalNgsRelate, the relative performance of the SNP-by-SNP approaches was rather variable. Finally, the vast majority of genotype-based predictors outperformed the pedigree-based predictions.

Compared to the simulations, the correlations between predicted and reference HBD levels were lower in the DAMONA data set, which contains about 100 trios for evaluation, and higher in the MW population, where we have only 13 trios. Regarding the relative performance of the evaluated methods, some trends observed in these two real data sets were similar to those highlighted with simulated data. IBD\_Haplo15c, ZooRoH, PLINK ROH and phasedibd performed well, closely followed by Hap-IBD. Conversely, Refined-IBD, TRUFFLE and LocalNgsRelate had poor relative performance in at least one scenario. GERMLINE and GIBDLD were closer to the best methods, but still achieved systematically lower correlations than, for example, IBD\_Haplo15c or phasedibd. As in the simulated data set, IBD\_Haplo15c performed better than the 9-STATES model in the cattle data, while the performances of different ZooRoH models were close. The SNP-by-SNP approaches still had variable performances, achieving the highest correlations in the MW population but almost the worst prediction in the DAMONA data set. Interestingly, the correlations obtained with the pedigree-based approach were quite high, among the best methods, especially for recent HBD levels.

### Accuracy of predicted locus-specific HBD levels

We then evaluated the methods in terms of the accuracy of locus-specific correlations (Figure 4, Figures S6-7). The SNP-by-SNP-based and pedigree-based methods were not included in these comparisons because they only provide genome-wide predictions. As expected, prediction accuracies decreased compared to genome-wide predictions averaged over many more loci. Overall, the best methods were almost the same as for genome-wide predictions. However, the best model-based approaches (IBD\_Haplo15c and ZooRoH models) now outperformed the best rule-based methods (phasedibd and PLINK ROH). Although hap-IBD was still competitive in some scenarios, it often achieved less accurate predictions than phasedibd and PLINK ROH. With regard to the model-based

approaches, the disadvantage of using unphased data (IBD\_Haplo9c versus IBD\_Haplo15c) was more pronounced, while the three ZooRoH models still had similar performances. Interestingly, the prediction of recent HBD (ZooRoH-25) was more accurate at predicting HBD levels defined with respect to a recent base population ( $F_{\text{YOUNG}}$  or  $F_{\text{MID}}$ ), which was not clearly observed with genome-wide predictions. As before, the other two model-based approaches, GIBDL and LocalNgsRelate, continued to show variable performance. The other evaluated methods achieved lower accuracies, for instance systematically lower than hap-IBD. These observations were consistent across all evaluated scenarios with both simulated and real data sets, with a few rare exceptions.

When the locus-specific accuracy was assessed using ROC curves (Figure 4, Figure S8) and the associated AUC (Figure S9-10), these trends were confirmed. The advantages of the best model-based approaches were even clearer. ZooRoH and IBD\_Haplo15c were indeed associated with the highest AUC across all scenarios. Most often, ZooRoH models were the best, especially when the base population was more ancient. For more recent reference HBD levels, predictions using ZooRoH-25 (recent HBD only) were better than those using ZooRoH-125.

### **Accuracy of HBD predictions with low-density marker panels and genotyping-by-sequencing data**

To reduce genotyping costs, low-marker density panels or genotyping-by-sequencing (GBS) data are commonly used in genomic studies of wild populations, conservation genetics or livestock species. To evaluate HBD predictions in such data, we relied mainly on the DAMONA data, for which low-marker density panels are available (Boichard et al., 2012) and for which we could use the reference genome to generate *in-silico* GBS panels. In addition, we also defined a LD density panel in the simulations.

For genome-wide HBD predictions using an LD array, accuracy was lower compared to values obtained using an MD genotyping array (Figure 5, Figure S11-12). However, some methods were more robust to this change and their loss of accuracy was limited. This was the case for the best model-based approaches, IBD\_Haplo15c and ZooRoH. In the DAMONA data set, some rule-based methods performed even better, including phasedibd, hap-IBD, GERMLINE (with modified parameters) and Refined-IBD. However, this was only confirmed for hap-IBD and GERMLINE in the simulated data. With our parameter setting, no ROH were identified with LD panels (even when reducing the number of SNPs per window and per ROH to 25 and setting the number of heterozygous SNPs to 0). The overall ranking was not much affected, with IBD\_Haplo and ZooRoH still among the best methods in all scenarios, while among the rule-based methods, hap-IBD was more reliable. Interestingly, the pedigree-based predictions ranked better and were closer to the best methods at this marker density. Overall, the best methods were still efficient when using a LD marker panel, resulting in correlations close to those obtained using a MD marker panel, especially for recent inbreeding.

When HBD levels were predicted for specific loci using low-density marker arrays, correlations with reference values decreased for all methods, more so when the reference base population was more distant. LocalNgsRelate and GERMLINE were two exceptions, as their correlations increased for several scenarios. Conversely, methods such as phasedibd, TRUFFLE or GIBDL were more affected, with greater reductions in accuracy, whereas hap-IBD was relatively robust. Despite these variations, IBD\_Haplo15c and ZooRoH performed best in all scenarios. Although hap-IBD was now the best rule-based method, it systematically achieved lower accuracy than the best model-based approaches.

For genome-wide predictions using GBS data, the accuracy increased with the number of fragments and their size, both of which affect the number of markers available. However, there were some exceptions. For example, the correlations were constant for ZooRoH-25, while they were identical for GBS-30K and GBS-50K when using IBD\_Haplo15c. The multiple HBD class ZooRoH-125 performed better than these two approaches (ZooRoH-25 was better only when predicting recent HBD levels

using GBS-15K), whereas the performance of the single HBD class ZooRoH model was severely compromised when using GBS data. The performance of the model-based approaches using GBS data was inferior to that using the MD array. Similarly, more accurate predictions were obtained using an LD array rather than a GBS-15K panel. On the contrary, rule-based methods such as phasedibd, hap-IBD and GERMLINE performed better with GBS-30K or GBS-50K panel than with MD arrays. These three methods are somewhat peculiar in that the highest accuracies were obtained using the LD and GBS-50K panels. Importantly, phasedibd was the most accurate method in almost all evaluated configurations (i.e. with the GBS marker densities and with different reference HBD levels), while hap-IBD was systematically better than model-based approaches with GBS-30K and GBS-50K panels. PLINK ROH had poor accuracies with GBS-15K, but its performance was good with GBS-50K. As a result, three rule-based methods achieved the highest genome-wide prediction accuracies when using the GBS-50K panel. Despite their lowest accuracies with GBS data, ZooRoH (multiple HBD class models) and IBD\_Haplo15c remained among the best methods, still outperforming other model-based approaches such as LocalNgsRelate and GIBDL. Interestingly, the ranking of the pedigree-based predictor increased, especially when fewer markers were available (GBS-15K or GBS-30K) or when predicting recent HBD levels. It even became the best or second best in some configurations.

When using GBS data for locus-specific predictions, accuracy increased systematically with the number of markers in the panel. It is important to note that the locus-specific accuracy was evaluated at the marker positions. For GBS data, this corresponds to small fragments with high marker density, and accuracy in less marker dense regions is not evaluated. Overall, ZooRoH models with multiple HBD classes achieved the highest accuracies, often followed by IBD\_Haplo15c. Phasedibd or hap-IBD were regularly the next best methods, sometimes better than IBD\_Haplo15c. PLINK-ROH performed poorly with GBS-15K or GBS-30K, but like phasedibd and hap-IBD, its ranking increased when more markers were available. At the locus-specific level, ZooRoH-1R was less affected by the use of GBS data. It was still significantly less efficient than the other ZooRoH models or IBD\_Haplo15c when a young or intermediate reference base population was used, but it became the best approach for comparisons with the most distant base populations. Interestingly, ZooRoH-125 was best and ZooRoH-25 second best with the intermediate base population, whereas the opposite was observed with the recent base population.

When locus-specific prediction accuracy was assessed using AUC (Figure S9-10), ZooRoH with multiple HBD classes and IBD\_Haplo15c were consistently the best across different scenarios. The only exception was the prediction of recent HBD levels using GBS-30K or GBS-50K panels, where hap-IBD resulted better than IBD\_Haplo15c. Interestingly, for distant base populations, the highest AUC values were obtained with GBS-50K, whereas for recent HBD levels, higher values were obtained with the LD marker array.

### **Accuracy of predictions when founder allele frequencies are used**

Ideally, methods that use estimated AF in their model should work with founder AF, but these are rarely available. The DAMONA data set provides the opportunity to estimate these founder AF and to assess the impact of using founder AF versus sample AF in these methods.

For genome-wide prediction of HBD levels (Figure 6), the use of founder AF dramatically increased the correlations obtained with SNP-by-SNP approaches. Significant improvements were also observed for IBD\_Haplo9c (using unphased data), which achieved similar accuracy to IBD\_Haplo15c (using phased data). When using commercial genotyping arrays, the performance of the ZooRoH models, IBD\_Haplo15c and LocalNgsRelate did not change with founder AF (we only observed slight increases with ZooRoH-1R). With GBS data, the multiple HBD classes ZooRoH model and LocalNgsRelate were not

affected by the change in AF, whereas large and modest improvements were observed for ZooRoH-1R and IBD\_Haplo15c, respectively. Nevertheless, ZooRoH-1R was still less accurate than the other ZooRoH models. For locus-specific predictions, the AF used had little effect, and we did not observe any changes in performance (Figure 6).

***Direct estimation of relatedness between the parents.*** The expected inbreeding coefficient of an individual is equal to the relatedness coefficient between its parents. Similarly, the prediction of HBD segments in the offspring is closely related to the identification of IBD segments between the parents. Thus, our evaluation procedure also provides indirect information on the accuracy of estimation of IBD segments and relatedness between parents. In simulated data, it is possible to directly assess the efficiency of the same methods for estimating relatedness. We therefore took advantage of our simulations to compare the accuracy of relatedness measures (see Figure S13). The ranking of the methods was the same as for the prediction of HBD levels, but with higher levels of accuracy and smaller differences between methods. The higher accuracy may be due to the fact that there is unpredictable random variation associated with Mendelian sampling in predicting HBD levels, a feature not present in kinship estimation. Interestingly, the pedigree-based approach was also less accurate for kinship because it can only estimate expected relatedness, whereas molecular-based estimators estimate realized relatedness.

## Discussion

**Evaluation with real and simulated data.** We herein used simulations and empirical comparisons with two real data sets, corresponding to a typical livestock species and a wild population in a conservation program, to evaluate the accuracy of different approaches used to predict HBD levels in future offspring of a pair of genotyped parents, an important information for the management of these populations. The two types of data set we used are complementary. The true HBD levels are known in the simulations, whereas more realistic conditions are obtained with real data sets. These include levels of relatedness in the sample, past demographic history, genome and LD structure, distribution of AF, effects of past and ongoing selection, levels of genotyping and phasing errors, etc. Evaluations of prediction methods on both types of data showed the same trends in terms of ranking the methods, indicating that the empirical analysis on real data was informative. The two real data sets had different structures. MW had higher inbreeding levels and a smaller sample size, that might affect the phasing efficiency and the accuracy of estimation of AFs for example. These would be the typical field conditions in conservation genetics or studies of wild fauna. The use of small samples might also increase the role of random variation in method comparisons. The cattle population had the advantage to contain more individuals that were whole-genome sequenced, providing more information on the true inbreeding levels. In general, important information is available in livestock populations, including deep and accurate pedigrees, a good reference genome and genetic maps, that are important for methods identifying IBD segments (see also Bosse et al., 2015). Genotyping marker arrays are also available at different densities and the reference genome allowed us to simulate GBS data, although these were probably cleaner than real GBS data. This illustrates that data available from livestock populations might be useful to study techniques that will be applied in molecular ecology. Here, we had access to a relatively large sequenced pedigree including many trios, as needed for our study.

**Relative performance of evaluated methods.** Using these data sets, we evaluated 16 methods for predicting HBD levels in the offspring of genotyped parents. Importantly, we showed that this evaluation is also informative about the accuracy of the methods for estimating relatedness and identifying IBD segments. Such estimators are useful in more applications than the prediction of HBD levels, including, for example, diversity management, selection, control for population structure, etc. In terms of accuracy, IBD\_Haplo15c and multiple HBD class ZooRoH models (ZooRoH-MixKL) were consistently among the most efficient when evaluated across different scenarios, on simulated and real datasets. IBD\_Haplo15c is the most complex model-based approach as it models the 15 identity states, while ZooRoH-MixKL, designed to estimate HBD levels within individuals, conceptually has two main states (HBD and non-HBD) but adds an additional layer of complexity by partitioning HBD into different length-based classes. IBD\_Haplo15c was generally better at predicting recent HBD levels, whereas ZooRoH-MixKL was more efficient when the base population was more distant. Still, it was possible to shift ZooRoH towards predicting recent HBD levels by selecting the appropriate HBD classes, this approach being more efficient for locus-specific prediction. Both methods outperformed the best rule-based methods, especially when probabilities were useful (i.e. with ROC curves) and information was reduced, at lower marker density and for locus-specific predictions. This behavior has previously been described for reduced marker panels in Lavanchy and Goudet (2023), at lower marker densities (Solé et al. (2017); Druet et al. (2020)) and for locus-specific estimation of HBD levels (Alemu et al., 2021). Such locus-specific predictions could be useful for managing recessive deleterious alleles that cause genetic defects, loci that are major contributors to inbreeding depression, predicted harmful mutations, or for maintaining heterozygosity at specific loci that should have high diversity, such as the



major histocompatibility complex. This could be particularly valuable in managing populations derived from a few founders with fully characterized genomes, including at deleterious loci. This information could also be used to give more weight to regions predicted to contribute to genetic load (Bertorelle et al., 2022), as neutral diversity may not always be representative of extinction risk (Teixeira and Huber, 2021). Locus-specific HBD predictions could also be used to manage inbreeding and diversity at all loci simultaneously, using algorithms that optimize all loci together. For genome-wide HBD predictions, phasedibd, PLINK-ROH and, to a lesser extent, hap-IBD, were also very accurate. Nevertheless, the performance of PLINK-ROH and hap-IBD decreased in some configurations, for example with LD and GBS-15K panels for PLINK-ROH which may be even less accurate in the presence of genotyping errors. Phasedibd performed well in all scenarios and thus represents an excellent option for genome-wide predictions, combining accuracy and computational efficiency. Here we have focused on small samples of interest in conservation genetics and wildlife, where model-based approaches are still applicable, but this may not be the case for larger data sets or at higher marker densities (biobanks and large livestock populations). However, these rule-based approaches were less efficient for locus-specific predictions, especially with the genotyping arrays, while the efficiency improved at higher marker densities such as with the GBS-50K panel. The other methods evaluated were less accurate in most scenarios, although they could perform well in some rare exceptions. In addition, many of them had high variability and therefore performed particularly poorly in some configurations. The SNP-by-SNP approaches did not perform well in all configurations, especially with sample AF, and only provide genome-wide predictions, but they have the advantage of being applicable with sub-optimal genome assemblies or without a genetic map.

**Method features affecting predictions accuracy.** From these comparisons we learned several properties of the prediction methods. First, the use of phased data achieved higher accuracy than the use of genotypes, despite possible errors introduced during the phasing process, consistent with the findings of Gómez-Romano et al. (2016). This was also the case for (real) small data sets, where phasing errors are expected to be more common. Interestingly, we observed no differences in accuracy using true versus estimated haplotypes at the genome-wide level, and only minor differences for locus-specific predictions (Figure S14). However, accurate phasing is not always possible, e.g. when physical or genetic marker maps are not available or with low-fold sequencing data (genotypes are not unambiguously known). Among the methods evaluated, LocalNgsRelate was the only one to handle low-fold sequencing data. This feature could also be added to other model-based methods using unphased data (e.g. IBD\_Haplo9c). Second, several methods, including the SNP-by-SNP approaches, were sensitive to the AF used. In agreement with theoretical expectations and the study by Caballero et al. (2022), better estimators were obtained with founder AF. Interestingly, the performance of IBD\_Haplo9c (using unphased data) increased close to that of IBHaplo15c (using phased data). This suggests that when phasing accuracy is compromised, such as with smaller datasets or low-fold sequencing data, methods that work with unphased data may be a good option if founder AF are available, but this is unfortunately rarely the case. Here we used the gene content approach, which requires a relatively large genotyped sample and a deep pedigree, which are rarely available in conservation genetics or wildlife. Conversely, it also shows that when founder AFs are not known, it is better to rely on a method using phased data. Overall, several methods were robust to the AF used, including obviously the rule-based approaches, but also IBD\_Haplo15c and ZooRoH-MixKL. It should be noted that the AF used had only a marginal effect on the accuracy of locus-specific HBD predictions, which depend more on the homozygosity of the markers around the target position. Thirdly, the 3-



STATES models were not among the best methods. This suggests that the assumption that the parents were non-inbred is not optimal in populations with small  $N_e$ . Locus-specific coancestry and predicted HBD levels are indeed reduced if we ignore that the haplotypes from one or both parents are IBD (e.g. the maximum value drops from 1 to 0.5), and this can have a significant impact in inbred populations. Fourth, we also learned that the use of fixed parameters defining the frequency of IBD (as in IBD\_Haplo) or the length of IBD segments (as in IBD\_Haplo or ZooRoH-MixKL) did not result in worse performance than the estimation of these parameters (as in GIBDLD or LocalNgsRelate). Finally, we observed that multiple HBD classes ZooRoH models were more robust to marker density or used AF than a ZooRoH-1R model. This is in agreement with the observation that pruning strategies are recommended with single HBD-class HMM (Leutenegger et al., 2003; Narasimhan et al., 2016; Vieira et al. 2016), whereas this is not necessary with a ZooRoH-MixKL model, as we have previously shown (Druet and Gautier, 2017). Nevertheless, the ZooRoH-1R performed well with the LD and MD genotyping arrays.

**Performance of pedigree-based predictions.** We observed that in some configurations, pedigree-based predictions of HBD levels performed well compared to methods using molecular data. This is different from estimating individual levels of inbreeding, where genomic estimators proved superior (Keller et al., 2011; Wang, 2016). In this situation, molecular data allow estimation of realized HBD levels, whereas the pedigree-based approach is limited to expected levels because it can not predict Mendelian sampling. In the prediction context, both approaches are unable to predict Mendelian sampling and therefore achieve more similar accuracies. The pedigree-based approach will be more competitive for predicting recent HBD levels (since pedigrees only capture recent generations), and when the accuracy of molecular-based approaches decreases, such as with LD marker arrays, some GBS panels, unavailable founder AF (for some methods) and inaccurate phasing. These results are in agreement with Woolliams et al. (2022), who suggested that the pedigree-based approaches could be a good option for genetic diversity management. However, the pedigree-based estimators require that the genealogy is accurately and deeply recorded, which is not always possible in wildlife populations for example. This argues for recording pedigrees as well as possible for population management, even when molecular data are available (Galla et al., 2021).

## Acknowledgements

Tom Druet is Research Director from the Fonds de la Recherche Scientifique - FNRS (F.R.S-FNRS). Computation were carried out using the supercomputing facilities of the “*Consortium d’Equipements en Calcul Intensif en Fédération Wallonie-Bruxelles*” (CECI), funded by the F.R.S-FNRS.

## References

- Alemu, S. W., Kadri, N. K., Harland, C., Faux, P., Charlier, C., Caballero, A., & Druet, T. (2021). An evaluation of inbreeding measures using a whole-genome sequenced cattle pedigree. *Heredity*, **126**(3), 410–423. <https://doi.org/10.1038/s41437-020-00383-9>
- Baumdicker, F., Bisschop, G., Goldstein, D., Gower, G., Ragsdale, A. P., Tsambos, G., Zhu, S., Eldon, B., Ellerman, E. C., Galloway, J. G., Gladstein, A. L., Gorjanc, G., Guo, B., Jeffery, B., Kretschumar, W. W., Lohse, K., Matschiner, M., Nelson, D., Pope, N. S., ... Kelleher, J. (2022). Efficient ancestry and mutation simulation with msprime 1.0. *Genetics*, **220**(3), iyab229. <https://doi.org/10.1093/genetics/iyab229>
- Bertorelle, G., Raffini, F., Bosse, M., Bortoluzzi, C., Iannucci, A., Trucchi, E., Morales, H. E., & van Oosterhout, C. (2022). Genetic load: Genomic estimates and applications in non-model animals. *Nature Reviews Genetics*, **23**(8), 492–503. <https://doi.org/10.1038/s41576-022-00448-x>
- Bertrand, A. R., Kadri, N. K., Flori, L., Gautier, M., & Druet, T. (2019). RZooRoH: An R package to characterize individual genomic autozygosity and identify homozygous-by-descent segments. *Methods in Ecology and Evolution*, **10**(6), 860–866. <https://doi.org/10.1111/2041-210X.13167>
- Boichard, D., Chung, H., Dasonneville, R., David, X., Eggen, A., Fritz, S., Gietzen, K. J., Hayes, B. J., Lawley, C. T., Sonstegard, T. S., Van Tassell, C. P., VanRaden, P. M., Viaud-Martinez, K. A., Wiggans, G. R., & Bovine LD Consortium. (2012). Design of a bovine low-density SNP array optimized for imputation. *PloS One*, **7**(3), e34130. <https://doi.org/10.1371/journal.pone.0034130>
- Bosse, M., Megens, H.-J., Madsen, O., Crooijmans, R. P. M. A., Ryder, O. A., Austerlitz, F., Groenen, M. A. M., & de Cara, M. A. R. (2015). Using genome-wide measures of coancestry to maintain diversity and fitness in endangered and domestic pig populations. *Genome Research*, **25**(7), 970–981. <https://doi.org/10.1101/gr.187039.114>
- Bosse, M., Megens, H.-J., Derks, M. F. L., de Cara, Á. M. R., & Groenen, M. A. M. (2019). Deleterious alleles in the context of domestication, inbreeding, and selection. *Evolutionary Applications*, **12**(1), 6–17. <https://doi.org/10.1111/eva.12691>
- Broman, K. W., & Weber, J. L. (1999). Long homozygous chromosomal segments in reference families from the centre d’Etude du polymorphisme humain. *American Journal of Human Genetics*, **65**(6), 1493–1500. <https://doi.org/10.1086/302661>
- Browning, B. L., & Browning, S. R. (2013). Improving the accuracy and efficiency of identity-by-descent detection in population data. *Genetics*, **194**(2), 459–471. <https://doi.org/10.1534/genetics.113.150029>

- Browning, B. L., Tian, X., Zhou, Y., & Browning, S. R. (2021). Fast two-stage phasing of large-scale sequence data. *American Journal of Human Genetics*, **108**(10), 1880–1890. <https://doi.org/10.1016/j.ajhg.2021.08.005>
- Caballero, A., Villanueva, B., & Druet, T. (2021). On the estimation of inbreeding depression using different measures of inbreeding from molecular markers. *Evolutionary Applications*, **14**(2), 416–428. <https://doi.org/10.1111/eva.13126>
- Caballero, A., Fernández, A., Villanueva, B., & Toro, M. A. (2022). A comparison of marker-based estimators of inbreeding and inbreeding depression. *Genetics, Selection, Evolution: GSE*, **54**(1), 82. <https://doi.org/10.1186/s12711-022-00772-0>
- de Cara, M. Á. R., Villanueva, B., Toro, M. Á., & Fernández, J. (2013). Using genomic tools to maintain diversity and fitness in conservation programmes. *Molecular Ecology*, **22**(24), 6091–6099. <https://doi.org/10.1111/mec.12560>
- Dimitromanolakis, A., Paterson, A. D., & Sun, L. (2019). Fast and accurate shared segment detection and relatedness estimation in un-phased genetic data via TRUFFLE. *American Journal of Human Genetics*, **105**(1), 78–88. <https://doi.org/10.1016/j.ajhg.2019.05.007>
- Druet, T., & Gautier, M. (2017). A model-based approach to characterize individual inbreeding at both global and local genomic scales. *Molecular Ecology*, **26**(20), 5820–5841. <https://doi.org/10.1111/mec.14324>
- Druet, T., Oleński, K., Flori, L., Bertrand, A. R., Olech, W., Tokarska, M., Kaminski, S., & Gautier, M. (2020). Genomic footprints of recovery in the European bison. *The Journal of Heredity*, **111**(2), 194–203. <https://doi.org/10.1093/jhered/esaa002>
- Druet, T., & Gautier, M. (2022). A hidden Markov model to estimate homozygous-by-descent probabilities associated with nested layers of ancestors. *Theoretical Population Biology*, **145**, 38–51. <https://doi.org/10.1016/j.tpb.2022.03.001>
- Durbin, R. (2014). Efficient haplotype matching and storage using the positional Burrows-Wheeler transform (PBWT). *Bioinformatics (Oxford, England)*, **30**(9), 1266–1272. <https://doi.org/10.1093/bioinformatics/btu014>
- Fitak, R. R., Rinkevich, S. E., & Culver, M. (2018). Genome-wide analysis of SNPs is consistent with no domestic dog ancestry in the endangered Mexican Wolf (*Canis lupus baileyi*). *The Journal of Heredity*, **109**(4), 372–383. <https://doi.org/10.1093/jhered/esy009>
- [dataset] Fitak, R. R., Rinkevich, S.E. & Culver, M. (2018). Data from: Genome-wide analysis of SNPs is consistent with no domestic dog ancestry in the endangered Mexican wolf (*Canis lupus baileyi*). Dryad. <https://doi.org/10.5061/dryad.g68k008>
- Freyman, W. A., McManus, K. F., Shringarpure, S. S., Jewett, E. M., Bryc, K., The 23 and Me Research Team, & Auton, A. (2021). Fast and robust identity-by-descent inference with the templated positional Burrows-Wheeler transform. *Molecular Biology and Evolution*, **38**(5), 2131–2151. <https://doi.org/10.1093/molbev/msaa328>
- Galla, S. J., Brown, L., Couch-Lewis (Ngāi Tahu: Te Hapū o Ngāti Wheke, Y., Ngāti Waewae), Cubrinovska, I., Eason, D., Gooley, R. M., Hamilton, J. A., Heath, J. A., Hauser, S. S., Latch, E. K., Matocq, M. D., Richardson, A., Wold, J. R., Hogg, C. J., Santure, A. W., & Steeves, T. E. (2022).

- The relevance of pedigrees in the conservation genomics era. *Molecular Ecology*, **31**(1), 41–54. <https://doi.org/10.1111/mec.16192>
- Gautier, M., Gharbi, K., Cezard, T., Foucaud, J., Kerdelhué, C., Pudlo, P., Cornuet, J.-M., & Estoup, A. (2013). The effect of RAD allele dropout on the estimation of genetic variation within and between populations. *Molecular Ecology*, **22**(11), 3165–3178. <https://doi.org/10.1111/mec.12089>
- Gengler, N., Mayeres, P., & Szydlowski, M. (2007). A simple method to approximate gene content in large pedigree populations: Application to the myostatin gene in dual-purpose Belgian Blue cattle. *Animal: An International Journal of Animal Bioscience*, **1**(1), 21–28. <https://doi.org/10.1017/S1751731107392628>
- Gómez-Romano, F., Villanueva, B., Fernández, J., Woolliams, J. A., & Pong-Wong, R. (2016). The use of genomic coancestry matrices in the optimisation of contributions to maintain genetic diversity at specific regions of the genome. *Genetics, Selection, Evolution: GSE*, **48**, 2. <https://doi.org/10.1186/s12711-015-0172-y>
- Gusev, A., Lowe, J. K., Stoffel, M., Daly, M. J., Altshuler, D., Breslow, J. L., Friedman, J. M., & Pe'er, I. (2009). Whole population, genome-wide mapping of hidden relatedness. *Genome Research*, **19**(2), 318–326. <https://doi.org/10.1101/gr.081398.108>
- Haller, B. C., Galloway, J., Kelleher, J., Messer, P. W., & Ralph, P. L. (2018). Tree-sequence recording in SLiM opens new horizons for forward-time simulation of whole genomes. *Molecular Ecology Resources*, **19**(2), 552–566. <https://doi.org/10.1111/1755-0998.12968>
- Haller, B. C., & Messer, P. W. (2023). SLiM 4: Multispecies eco-evolutionary modeling. *The American Naturalist*, **201**(5), E127–E139. <https://doi.org/10.1086/723601>
- Han, L., & Abney, M. (2011). Identity by descent estimation with dense genome-wide genotype data. *Genetic Epidemiology*, **35**(6), 557–567. <https://doi.org/10.1002/gepi.20606>
- Han, L., & Abney, M. (2013). Using identity by descent estimation with dense genotype data to detect positive selection. *European Journal of Human Genetics*, **21**(2), 205–211. <https://doi.org/10.1038/ejhg.2012.148>
- Herten, K., Hestand, M. S., Vermeesch, J. R., & Van Houdt, J. K. J. (2015). GBSX: A toolkit for experimental design and demultiplexing genotyping by sequencing experiments. *BMC Bioinformatics*, **16**(1), 73. <https://doi.org/10.1186/s12859-015-0514-3>
- Hinrichs, D., Meuwissen, T. H. E., Ødegard, J., Holt, M., Vangen, O., & Woolliams, J. A. (2007). Analysis of inbreeding depression in the first litter size of mice in a long-term selection experiment with respect to the age of the inbreeding. *Heredity*, **99**(1), 81–88. <https://doi.org/10.1038/sj.hdy.6800968>
- Jacquard, A. (1974). *The Genetic Structure of Populations*. Springer-Verlag, New-York, NY.
- Kardos, M., Taylor, H. R., Ellegren, H., Luikart, G., & Allendorf, F. W. (2016). Genomics advances the study of inbreeding depression in the wild. *Evolutionary Applications*, **9**(10), 1205–1218. <https://doi.org/10.1111/eva.12414>
- Keller, L. F., & Waller, D. M. (2002). Inbreeding effects in wild populations. *Trends in Ecology & Evolution*, **17**(5), 230–241. [https://doi.org/10.1016/S0169-5347\(02\)02489-8](https://doi.org/10.1016/S0169-5347(02)02489-8)

- Keller, M. C., Visscher, P. M., & Goddard, M. E. (2011). Quantification of inbreeding due to distant ancestors and its detection using dense single nucleotide polymorphism data. *Genetics*, **189**(1), 237–249. <https://doi.org/10.1534/genetics.111.130922>
- Kirin, M., McQuillan, R., Franklin, C. S., Campbell, H., McKeigue, P. M., & Wilson, J. F. (2010). Genomic runs of homozygosity record population history and consanguinity. *PLoS ONE*, **5**(11), e13996. <https://doi.org/10.1371/journal.pone.0013996>
- Lavanchy, E., & Goudet, J. (2023). Effect of reduced genomic representation on using runs of homozygosity for inbreeding characterization. *Molecular Ecology Resources*, **23**(4), 787–802. <https://doi.org/10.1111/1755-0998.13755>
- Leutenegger, A.-L., Prum, B., Génin, E., Verny, C., Lemaître, A., Clerget-Darpoux, F., & Thompson, E. A. (2003). Estimation of the inbreeding coefficient through use of genomic data. *American Journal of Human Genetics*, **73**(3), 516–523. <https://doi.org/10.1086/378207>
- Lee, Y. L., Bouwman, A. C., Harland, C., Bosse, M., Costa Monteiro Moreira, G., Veerkamp, R. F., Mullaart, E., Cambisano, N., Groenen, M. A. M., Karim, L., Coppieters, W., Georges, M., & Charlier, C. (2023). The rate of de novo structural variation is increased in in vitro-produced offspring and preferentially affects the paternal genome. *Genome research*, **33**(9), 1455–1464. <https://doi.org/10.1101/gr.277884.123>
- Li, C. C., & Horvitz, D. G. (1953). Some methods of estimating the inbreeding coefficient. *American Journal of Human Genetics*, **5**(2), 107–117.
- Malécot, G. (1948). *Les Mathématiques de l'hérédité*. Masson et Cie, Paris. <https://cir.nii.ac.jp/crid/1130000797475730048>
- Malécot, G. (1967). Identical loci and relationship. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 4 (eds LM Le Cam, J Neyman), 317–332. University of California Press, Berkeley, California. <https://projecteuclid.org/accountAjax/Download?urlId=bsmsp%2F1200513803&downloadTy pe=presschapter&isResultClick=True>
- McQuillan, R., Leutenegger, A.-L., Abdel-Rahman, R., Franklin, C. S., Pericic, M., Barac-Lauc, L., Smolej-Narancic, N., Janicijevic, B., Polasek, O., Tenesa, A., Macleod, A. K., Farrington, S. M., Rudan, P., Hayward, C., Vitart, V., Rudan, I., Wild, S. H., Dunlop, M. G., Wright, A. F., Campbell, H., & Wilson, J. F. (2008). Runs of homozygosity in European populations. *American Journal of Human Genetics*, **83**(3), 359–372. <https://doi.org/10.1016/j.ajhg.2008.08.007>
- Meuwissen, T. H. E., Sonesson, A. K., Gebregiorgis, G., & Woolliams, J. A. (2020). Management of genetic diversity in the era of genomics. *Frontiers in Genetics*, **11**, 880. <https://doi.org/10.3389/fgene.2020.00880>
- Naji, M. M., Gualdrón Duarte, J. L., Forneris, N. S., & Druet, T. (2024). Inbreeding depression is associated with recent homozygous-by-descent segments in Belgian Blue beef cattle. *Genetics Selection Evolution*, **56**(1), 10. <https://doi.org/10.1186/s12711-024-00878-7>
- 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**(11), 1749–1751. <https://doi.org/10.1093/bioinformatics/btw044>

- Nietlisbach, P., Muff, S., Reid, J. M., Whitlock, M. C., & Keller, L. F. (2019). Nonequivalent lethal equivalents: Models and inbreeding metrics for unbiased estimation of inbreeding load. *Evolutionary Applications*, **12**(2), 266–279. <https://doi.org/10.1111/eva.12713>
- Oget-Ebrad, C., Kadri, N. K., Moreira, G. C. M., Karim, L., Coppieters, W., Georges, M., & Druet, T. (2022). Benchmarking phasing software with a whole-genome sequenced cattle pedigree. *BMC Genomics*, **23**, 130. <https://doi.org/10.1186/s12864-022-08354-6>
- Pemberton, T. J., Absher, D., Feldman, M. W., Myers, R. M., Rosenberg, N. A., & Li, J. Z. (2012). Genomic patterns of homozygosity in worldwide human populations. *American Journal of Human Genetics*, **91**(2), 275–292. <https://doi.org/10.1016/j.ajhg.2012.06.014>
- Pryce, J. E., Hayes, B. J., & Goddard, M. E. (2012). Novel strategies to minimize progeny inbreeding while maximizing genetic gain using genomic information. *Journal of Dairy Science*, **95**(1), 377–388. <https://doi.org/10.3168/jds.2011-4254>
- Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., Maller, J., Sklar, P., de Bakker, P. I. W., Daly, M. J., & Sham, P. C. (2007). PLINK: A tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics*, **81**(3), 559–575. <https://doi.org/10.1086/519795>
- Schäffer, A. A. (1999). Computing probabilities of homozygosity by descent. *Genetic Epidemiology*, **16**(2), 135–149. [https://doi.org/10.1002/\(SICI\)1098-2272\(1999\)16:2<135::AID-GEPI2>3.0.CO;2-W](https://doi.org/10.1002/(SICI)1098-2272(1999)16:2<135::AID-GEPI2>3.0.CO;2-W)
- Severson, A. L., Korneliussen, T. S., & Moltke, I. (2022). LocalNgsRelate: A software tool for inferring IBD sharing along the genome between pairs of individuals from low-depth NGS data. *Bioinformatics*, **38**(4), 1159–1161. <https://doi.org/10.1093/bioinformatics/btab732>
- Solé, M., Gori, A.-S., Faux, P., Bertrand, A., Farnir, F., Gautier, M., & Druet, T. (2017). Age-based partitioning of individual genomic inbreeding levels in Belgian Blue cattle. *Genetics, Selection, Evolution: GSE*, **49**(1), 92. <https://doi.org/10.1186/s12711-017-0370-x>
- Stoffel, M. A., Johnston, S. E., Pilkington, J. G., & Pemberton, J. M. (2021). Mutation load decreases with haplotype age in wild Soay sheep. *Evolution Letters*, **5**(3), 187–195. <https://doi.org/10.1002/evl3.229>
- Szpiech, Z. A., Xu, J., Pemberton, T. J., Peng, W., Zöllner, S., Rosenberg, N. A., & Li, J. Z. (2013). Long runs of homozygosity are enriched for deleterious variation. *American Journal of Human Genetics*, **93**(1), 90–102. <https://doi.org/10.1016/j.ajhg.2013.05.003>
- Teixeira, J. C., & Huber, C. D. (2021). The inflated significance of neutral genetic diversity in conservation genetics. *Proceedings of the National Academy of Sciences of the United States of America*, **118**(10), e2015096118. <https://doi.org/10.1073/pnas.2015096118>
- Thompson, E. A. (2008). The IBD process along four chromosomes. *Theoretical Population Biology*, **73**(3), 369–373. <https://doi.org/10.1016/j.tpb.2007.11.011>
- Thompson, E. A. (2009). Inferring coancestry of genome segments in populations. *Invited Proceedings of the 57<sup>th</sup> Session of the International Statistical Institute*, IPM13, Paper 0325. Durban, South Africa. [http://faculty.washington.edu/eathomp/Anonftp/Papers/ipm13\\_thompson.pdf](http://faculty.washington.edu/eathomp/Anonftp/Papers/ipm13_thompson.pdf)
- VanRaden, P. M. (2008). Efficient methods to compute genomic predictions. *Journal of Dairy Science*, **91**(11), 4414–4423. <https://doi.org/10.3168/jds.2007-0980>



- Vieira, F. G., Albrechtsen, A., & Nielsen, R. (2016). Estimating IBD tracts from low coverage NGS data. *Bioinformatics (Oxford, England)*, **32**(14), 2096–2102. <https://doi.org/10.1093/bioinformatics/btw212>
- Wang, J. (2016). Pedigrees or markers: Which are better in estimating relatedness and inbreeding coefficient? *Theoretical Population Biology*, **107**, 4–13. <https://doi.org/10.1016/j.tpb.2015.08.006>
- Woolliams, J. A., & Meuwissen, T. H. E. (2022). Genetic management meets genomics. *Proceedings of the 12th world congress of genetics applied to livestock production*, Rotterdam, Netherlands. [https://www.wageningenacademic.com/pb-assets/wagen/WCGALP2022/61\\_015.pdf](https://www.wageningenacademic.com/pb-assets/wagen/WCGALP2022/61_015.pdf)
- Wright, S. (1922). Coefficients of inbreeding and relationship. *The American Naturalist*, **56**(645), 330–338. <https://doi.org/10.1086/279872>
- Yang, J., Benyamin, B., McEvoy, B. P., Gordon, S., Henders, A. K., Nyholt, D. R., Madden, P. A., Heath, A. C., Martin, N. G., Montgomery, G. W., Goddard, M. E., & Visscher, P. M. (2010). Common SNPs explain a large proportion of the heritability for human height. *Nature Genetics*, **42**(7), 565–569. <https://doi.org/10.1038/ng.608>
- Zhou, Y., Browning, S. R., & Browning, B. L. (2020). A fast and simple method for detecting identity-by-descent segments in large-scale data. *The American Journal of Human Genetics*, **106**(4), 426–437. <https://doi.org/10.1016/j.ajhg.2020.02.010>



## Tables

**Table 1.** Main properties of the methods used to predict levels of homozygosity-by-descent in a future offspring of a genotyped pair of parents. The table also indicates whether the methods use genotype (GEN) or haplotype (HAP) data and the names of the corresponding software. The STATES, PAIR and PANEL approaches are described in the main text and in Figure 1.

Method Name	Data	Approach		Software
IBD_Haplo15c	HAP	15-STATES	Model-based	IBD_Haplo
IBD_Haplo9c	GEN	9-STATES	Model-based	IBD_Haplo
GIBDL	GEN	9-STATES	Model-based	IBDL
LocalNgsRelate	GEN	3-STATES	Model-based	LocalNgsRelate
TRUFFLE	GEN	3-STATES	Rule-based	TRUFFLE
ZooRoH	HAP	PAIR	Model-based	RZooRoH
PLINK ROH	HAP	PAIR	Rule-based	PLINK
phasedibd	HAP	PANEL	Rule-based	phasedibd
hap-IBD	HAP	PANEL	Rule-based	hap-IBD
GERMLINE	HAP	PANEL	Rule-based	GERMLINE
Refined IBD	HAP	PANEL	Hybrid	Refined IBD
UNI	GEN	/	/	GCTA – algo0
GRM	GEN	/	/	GCTA – algo1
Pedigree	Pedigree	/	/	In-house script

## Figures

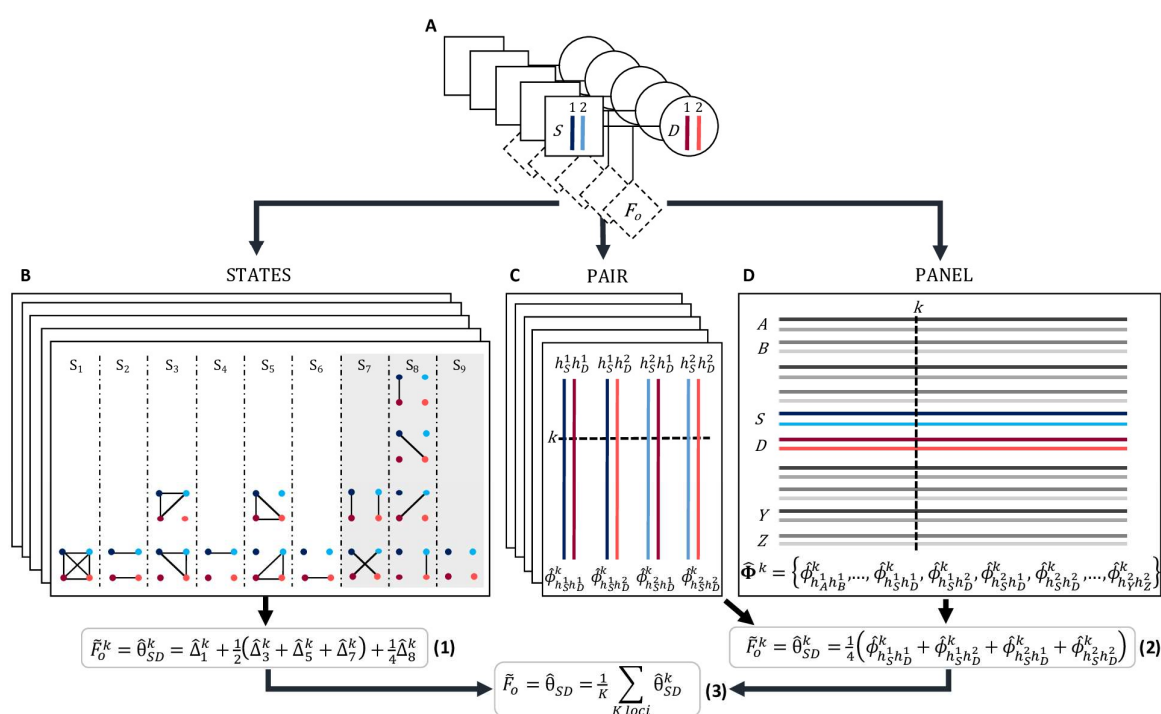


Figure 1. Approaches used to predict the homozygous-by-descent (HBD) level of an individual ( $F_O$ ) based on the genotypes of its parents. A) Predictions were made in several trios, each including a sire (S) and a dam (D), having each a paternal and maternal haplotype (labelled 1 and 2 respectively), shown in blue for the sire and red for the dam, and the future offspring (dashed diamond). B) STATES approach: for each pair of parents, the four parental haplotypes can take fifteen possible identity-by-descent (IBD) configurations at a locus (Jacquard, 1974). The four parental haplotypes are represented by dots using the same colors as in A) while solid lines indicate IBD between linked haplotypes. The 15 IBD states can be grouped into 9 condensed IBD states  $\{S_1, S_2, \dots, S_9\}$  if the parental origins of the haplotypes within the two individuals are considered unknown. The number of states are reduced to three if the parents are assumed to be non-inbred (grey background). The STATES approaches model the observed genotypes or haplotypes conditional on these configurations. At locus  $k$ , the estimated probabilities of the 9 IBD modes  $\{\hat{\Delta}_1^k, \hat{\Delta}_2^k, \dots, \hat{\Delta}_9^k\}$  can be used to estimate the locus specific coancestry between the parents ( $\hat{\theta}_{SD}^k$ ), which corresponds to the predicted locus-specific HBD level in the offspring  $\bar{F}_O^k$  (Eq. 1). C) PAIR approach: IBD is modelled sequentially for each of the four possible combinations of parental haplotypes ( $h_S^1 h_D^1, h_S^1 h_D^2, h_S^2 h_D^1, h_S^2 h_D^2$ ), where  $h_I^1$  and  $h_I^2$  denote the paternal and maternal haplotypes of individual  $I$ , respectively. This analysis estimates the IBD probability  $\hat{\phi}_{h_S^i h_D^j}^k$  between two haplotypes  $h_S^i$  and  $h_D^j$  at locus  $k$  for the four possible pairs. D) PANEL approach: a large panel of haplotypes from a large number of individuals are analyzed jointly to detect IBD segments. At locus  $k$ , a vector  $\hat{\Phi}^k$  containing the IBD probabilities for each pair of haplotypes is estimated (note that with rule-based approaches, IBD probabilities are either 0 or 1). In the PAIR and PANEL approaches, the locus specific coancestry between the parents is obtained as the average over the four possible pairs of parental haplotypes (Eq. 2). Finally, the genome-wide HBD levels  $\bar{F}_O$  can be predicted as the average locus-specific values at the  $K$  loci (Eq. 3).

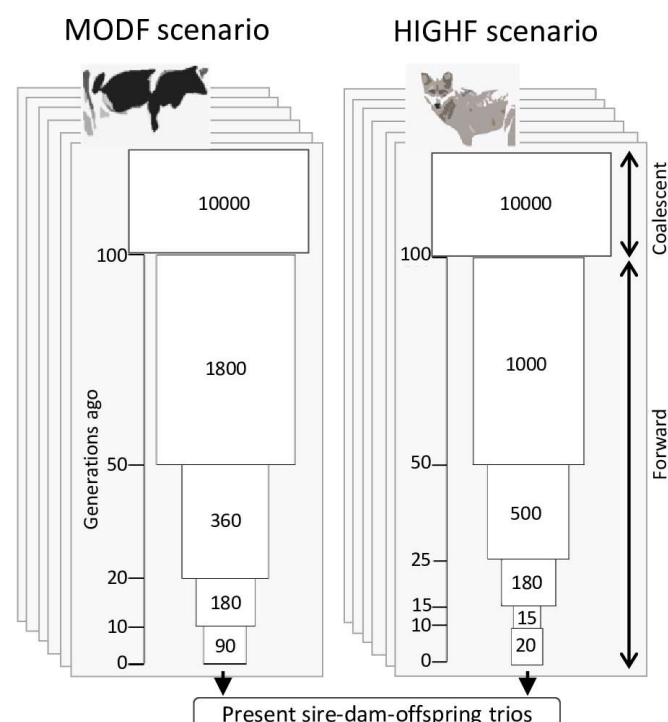


Figure 2. Simulated demographic scenarios. We simulated a moderately inbred (MODF) population corresponding to a typical livestock population (e.g. dairy cattle), and a highly inbred (HIGHF) population corresponding to an endangered wild population under conservation (similar to the Mexican Wolf). In both scenarios, we run a forward-in-time simulation with an initial population of  $N_e = 10,000$  and equal sex ratio. In the MODF and HIGHF scenario, the proportion of males for the subsequent generations was set to 0.1 and 0.5, respectively. The figure represents the evolution of  $N_e$  across generations.

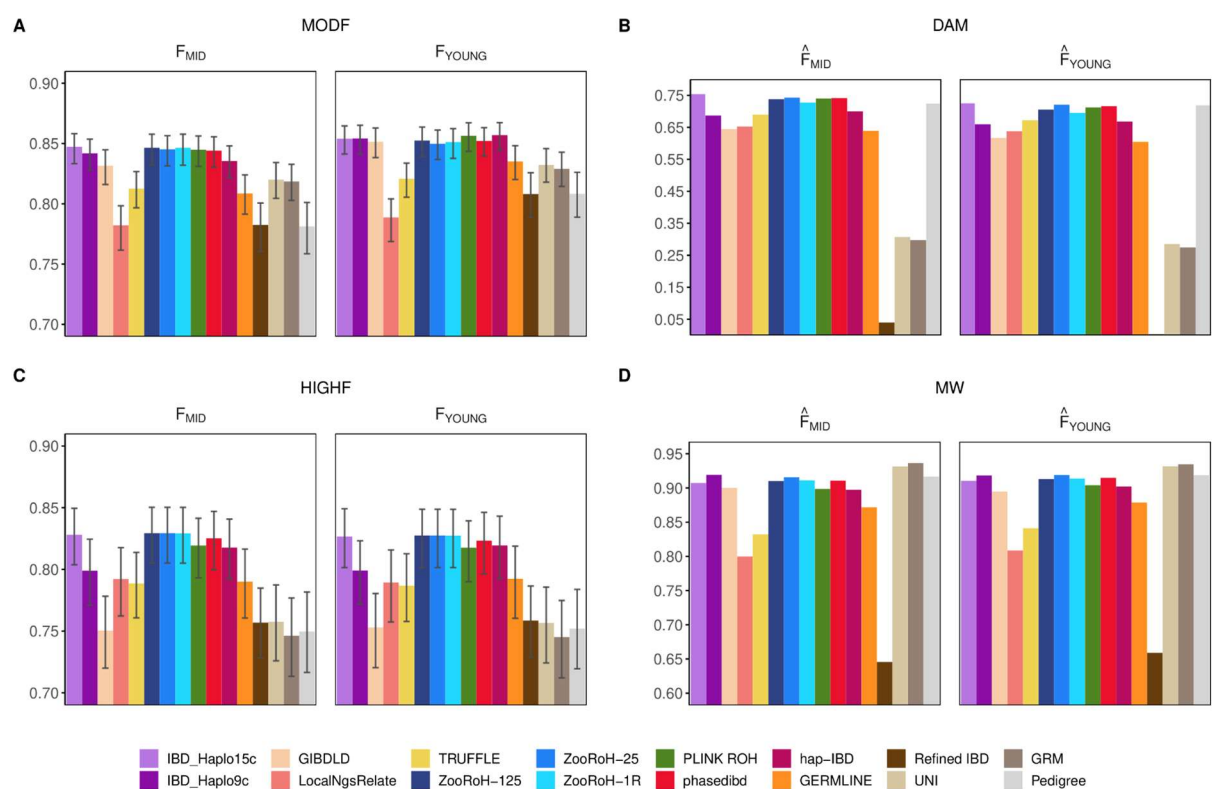


Figure 3. Correlations between predicted and reference genome-wide levels of HBD for the 16 methods compared and the four scenarios, using a medium density array. Methods and their abbreviation are described in Table 1. A) moderately inbred simulated population (MODF); B) DAMONA cattle data set (DAM); C) highly inbred simulated population (HIGHF); D) Mexican Wolf data set (MW). The reference levels were defined for either a recent or an intermediate base population (approximately 15 or 50 generations ago, respectively). Mean and 99% confidence intervals are shown for the simulated scenarios.

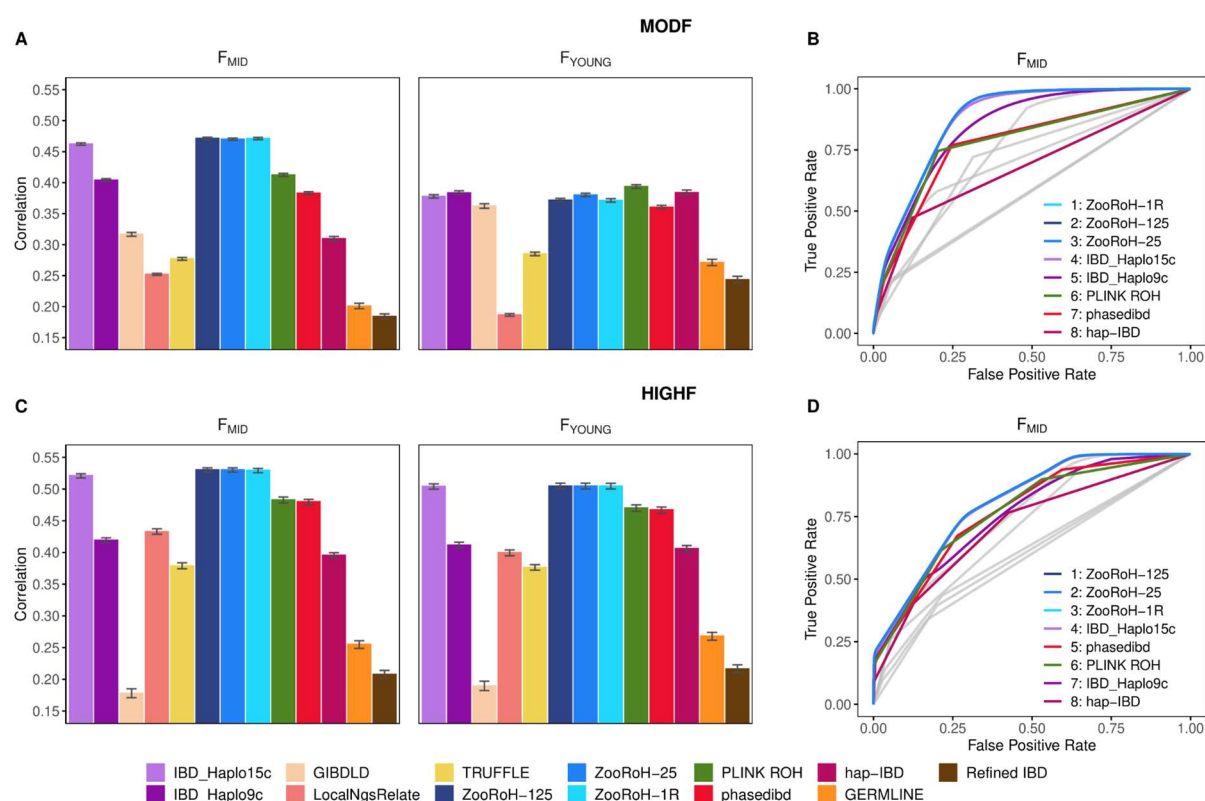


Figure 4. Locus-specific accuracy of 13 HBD prediction methods in the simulated data sets using a medium density array. Methods and their abbreviation are described in Table 1. Accuracy was assessed in the MODF (A-B) and HIGHF (C-D) scenario using correlations (mean and 99% confidence intervals) between predicted and reference locus-specific HBD levels (A & C). The reference levels are the true HBD status at every marker position and defined using either a recent (15 generations ago - F<sub>YOUNG</sub>) or an intermediate (50 generations - F<sub>MID</sub>) base population. The Receiver operating characteristic (ROC) curve for each method were also computed (B & D), with colored curves for the eight best methods. Their relative ranking in terms of AUC is also specified.

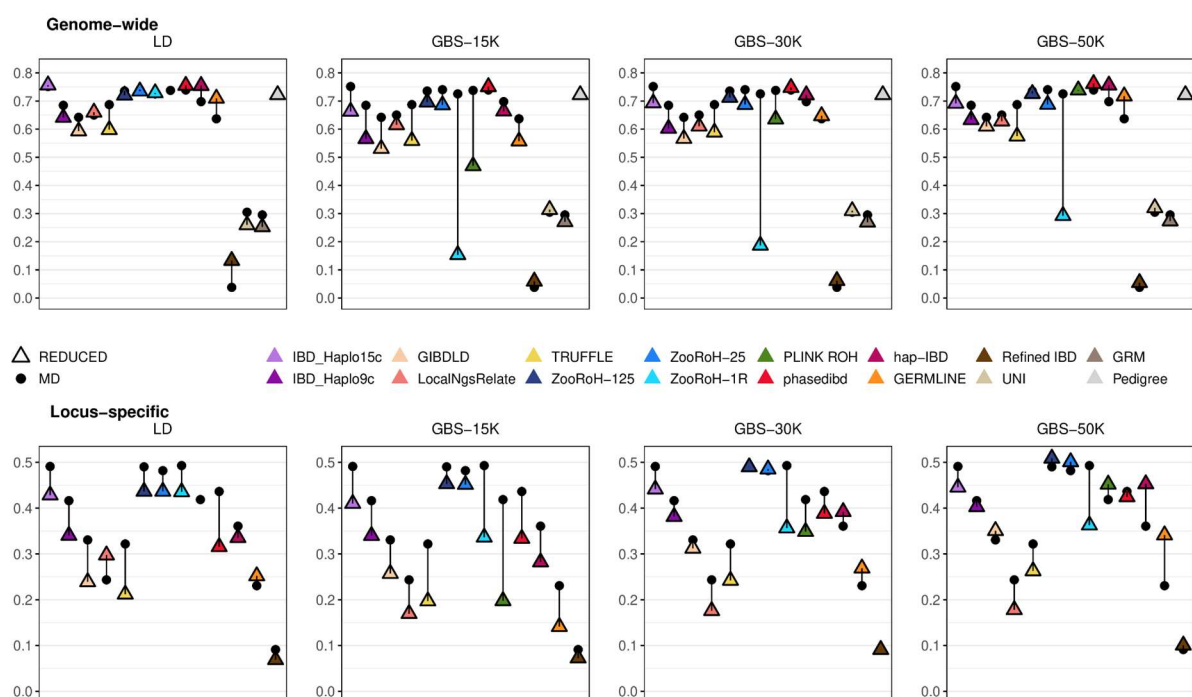


Figure 5. Correlations between predicted and reference genome-wide or locus-specific HBD levels for the 16 evaluated methods in the DAMONA cattle data set using reduced genotyping arrays. Methods and their abbreviation are described in Table 1. Correlations obtained with low-density (LD) or genotype-by-sequencing (GBS) data were compared to those achieved with the medium-density (MD) array (triangles versus dots). Reference inbreeding were estimated using the intermediate base population  $\hat{F}_{MID}$ .

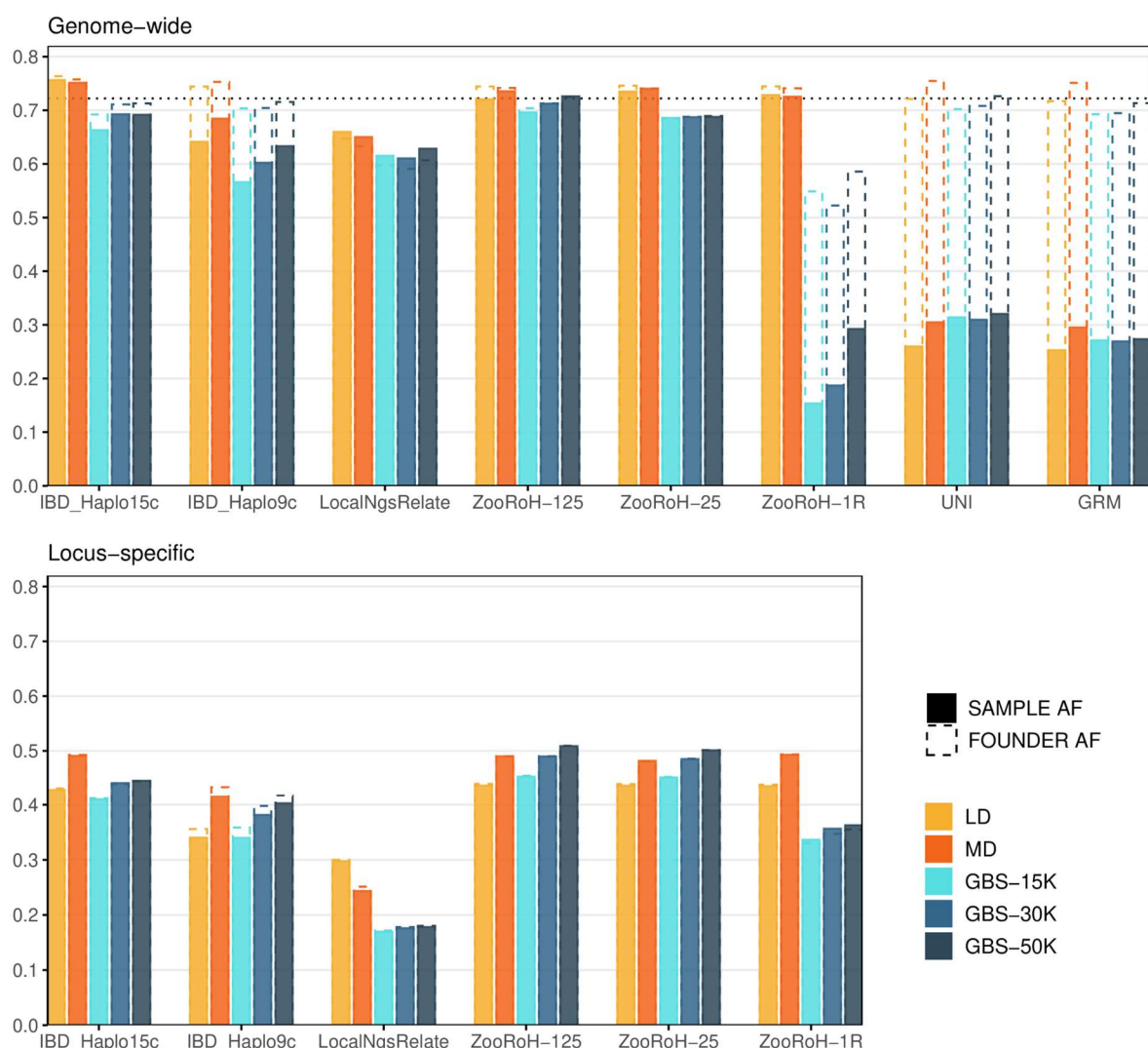


Figure 6. Impact of using founder versus sample allele frequencies on correlations between predicted and reference genome-wide and locus-specific HBD levels in the DAMONA cattle data set, for the methods described in Table 1 that accept external allele frequencies as input. Reference inbreeding were estimated using the intermediate base population  $\hat{F}_{MID}$ . Correlations are shown for different marker panels including low density (LD), medium density (MD), and GBS panels with different number of markers. Correlations are compared for predictions computed using founder versus sample allele frequencies. The horizontal dashed line is the correlation obtained with the pedigree.