Evaluation of genomic tools to predict individual homozygosity-by-descent for the management of genetic diversity in small populations

Natalia Forneris[1] and Tom Druet[1] [1]Unit of Animal Genomics, GIGA-R & Faculty of Veterinary Medicine, University of Liège, Liège, Belgium

Motivation

In populations of small effective size (Ne), such as those in conservation programs, companion animals or livestock species, inbreeding management is key. In that context, homozygosity-by-descent (HBD) segments are valuable as they allow efficient estimation of the inbreeding coefficient, provide locus-specific information and their length is informative about the “age” of inbreeding.

Objective

Evaluate different methods to predict the HBD level in future offspring based on genotypes from their parents, a problem equivalent to identifying segments identical-by-descent (IBD) among the four parental chromosomes.

Methods

We evaluated 16 methods using trios (Box 1A) from both simulated and real data with small Ne, including a sequenced Dutch Holstein cattle pedigree and genotyped Mexican wolves, a population that faced extinction in the wild. Methods included model-based approaches, mostly hidden Markov models (HMM), that considered up to 15 IBD configurations among the four parental chromosomes, as well as more computationally efficient rule-based approaches such as those developed to analyze entire biobanks (Box 1B-D and Table 1).

Box 1. Approaches to predict future HBD levels

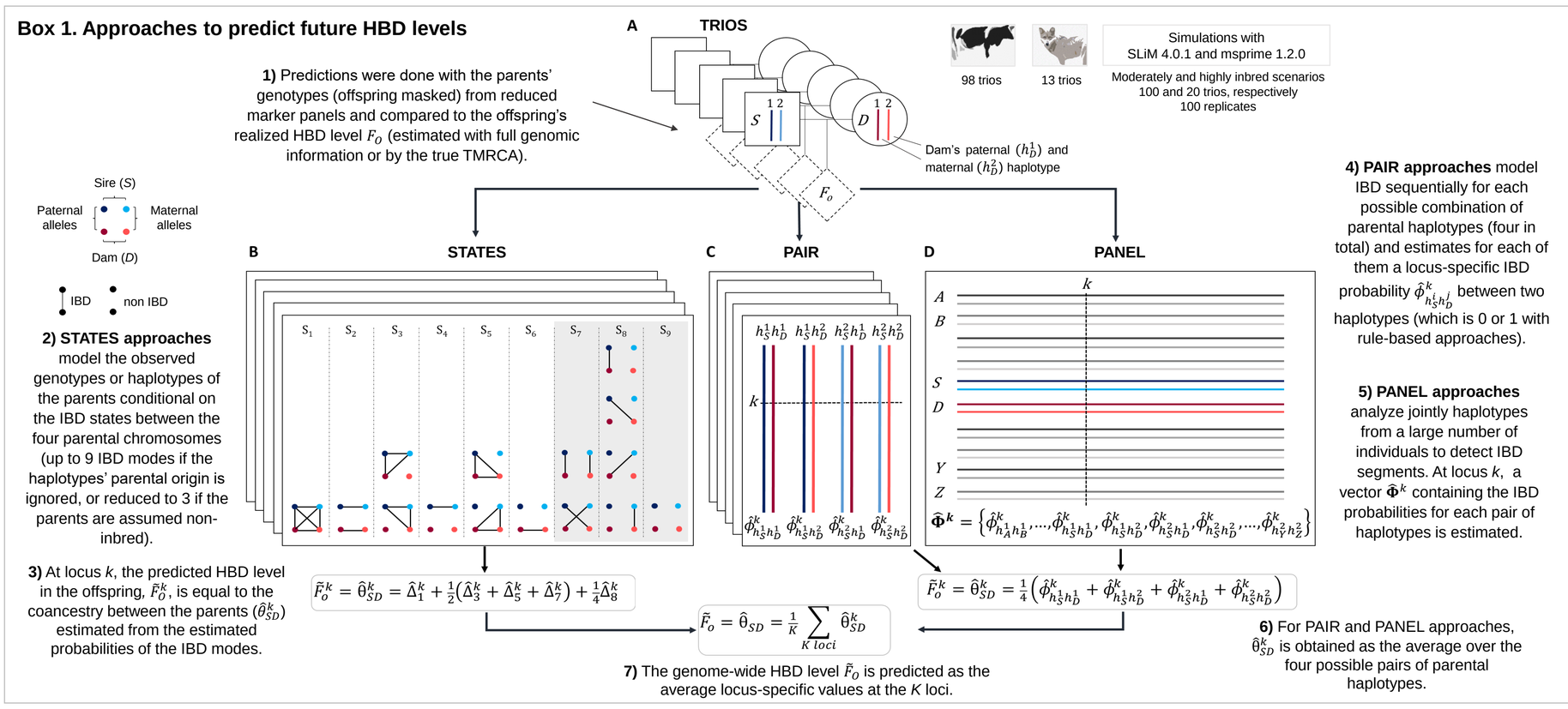

Table 1. Summary of the evaluated methods

Name	Data*	Approach		Software
IBD_Haplo15c	HAP	15-STATES	Model-based	IBD_Haplo
IBD_Haplo9c	GEN	9-STATES	Model-based	IBD_Haplo
GIBDLD	GEN	9-STATES	Model-based	IBDLD
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

*Indicates if the methods use genotypes (GEN) or haplotypes (HAP)
**Models with multiple HBD classes accounting for recent ZooRoH-125) or very recent inbreeding (ZooRoH-25) and a model with a single HBD class (ZooRoH-1R) were run.

Results

Relative performance of evaluated methods

Two HMM (IBD_Haplo15c and ZooRoH with multiple HBD classes) performed consistently well across different scenarios and were particularly efficient when probabilities were useful (i.e. with ROC curves) and information was reduced, at lower marker density and for locus-specific predictions (Figures 1 and 2). Two rule-based approaches (phasedibd and PLINK-ROH) were also efficient for genome-wide predictions. Locus-specific prediction accuracy of rule-based approaches decreased in some configurations (e.g., with LD and GBS-15K panels for PLINK ROH), while it improved at higher marker densities such as with the GBS-50K panel, despite being less efficient than the best model-based methods.

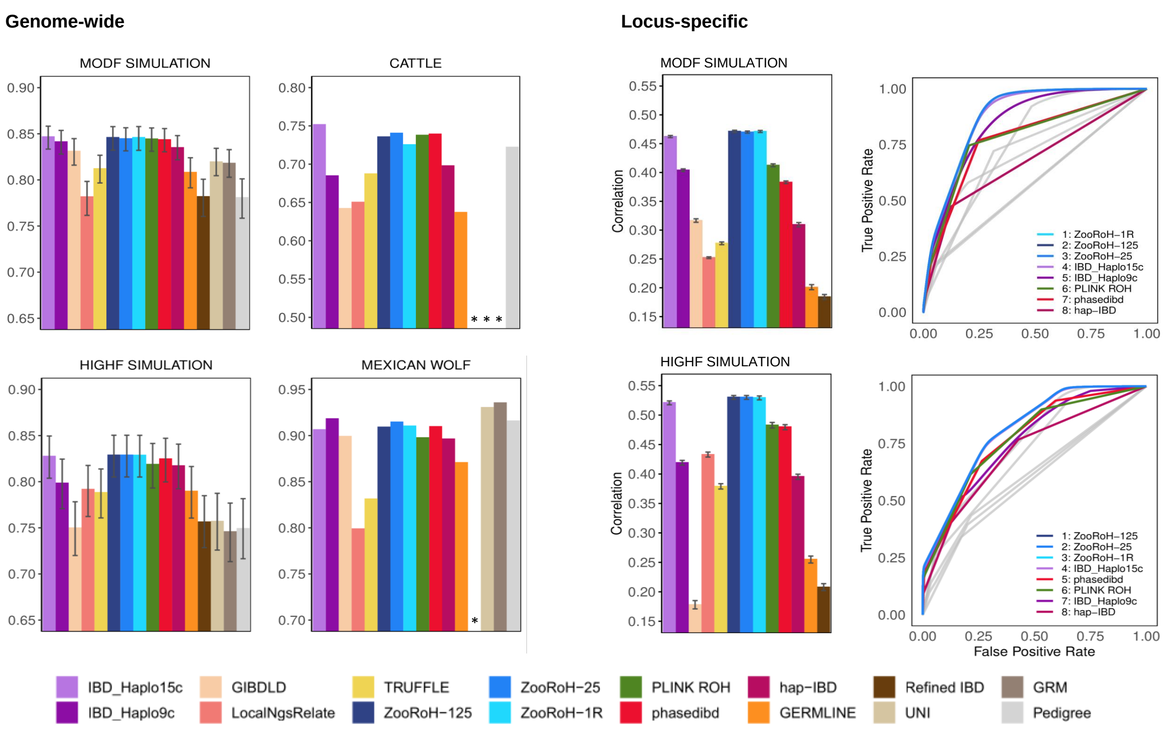

Figure 1. Correlations between predicted and reference genome-wide (left) or locus specific (right) HBD levels for the methods using a medium-density array on a moderately (MODF) or highly (HIGHF) inbred simulated population and on real data (*methods with values below 0.5). ROC curves are also shown (right) and methods with the best AUC values are highlighted.

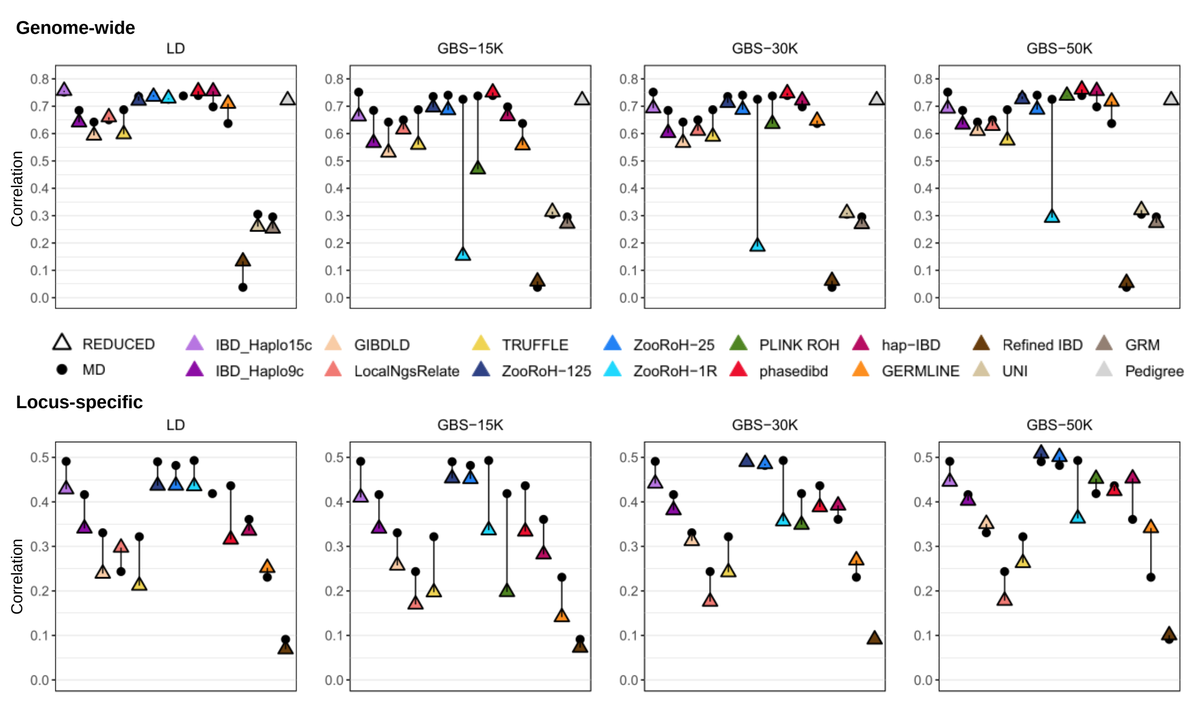

Figure 2. Correlations between predicted and reference HBD levels for the methods using reduced genotyping arrays (low-density (LD) or genotype-by-sequencing (GBS) panels with different number of markers) and compared to those achieved with the medium-density (MD) array (triangles versus dots) in the cattle data set.

Method features affecting predictions accuracy

Using phased data improved prediction accuracy, despite introducing errors. In addition, for some methods using allele frequencies (AFs) and genotypes, such us SNP-by-SNP approaches, performance could drop dramatically when sample instead of founder AF are used (Figure 3). In that case, approaches relying on the identification of long IBD segments, IBD_Haplo15c and ZooRoH (with multiple HBD classes) proved robust. The impact of used AF was marginal on locus-specific performance. Interestingly, when information is reduced, pedigree-based methods became more competitive for genome-wide predictions.

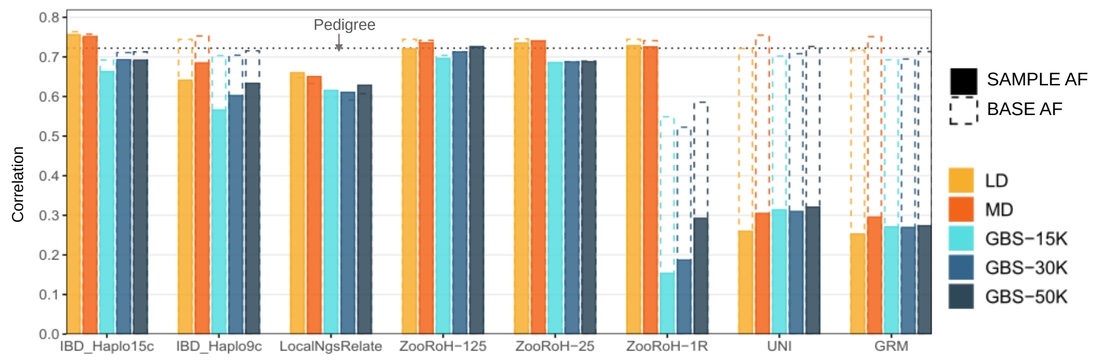

Figure 3. Impact of using founder versus sample AFs on correlations between predicted and reference genome-wide HBD levels, for different marker panels, in the dairy cattle data set. Results are shown for the methods that accept external AF as input.

Conclusions

- Large sequenced pedigree from livestock population allow to evaluate methods in realistic conditions, and are complementary to simulation approaches
- Our design allowed to highlight methods that perform well and identify sub-optimal approaches in populations with small Ne
- The study is also informative about the accuracy of the methods for estimating relatedness and identifying IBD segments between pairs of individuals
- Two model-based approaches relying on HMM proved efficient for both genome-wide and locus-specific prediction across scenarios and with reduced information
- Pedigree predictions were competitive for recent inbreeding when information is reduced

LIÈGE université GIGA institute
forneris@uliege.be