

## **A whole-genome-based approach for estimation and characterization of individual inbreeding**

T. Druet\* and M. Gautier<sup>§,†</sup>

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

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

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

Corresponding author:

Tom Druet

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

Tel: +3243669172; Fax:+3243664151

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

**Keywords:** Inbreeding; Runs of Homozygosity; ROH; hidden Markov models; identity-by-descent

**Running title:** Individual genomic inbreeding characterization

## 1 **Abstract**

2 Inbreeding results from the mating of related individuals and has negative consequence because it brings together  
3 deleterious variants in one individual. Inbreeding is associated with recessive diseases and reduced production or  
4 fitness. In general, inbreeding is estimated with respect to a base population that needs to be defined. Ancestors  
5 in generations anterior to the base population are considered unrelated. We herein propose a model that estimates  
6 inbreeding relative to multiple age-based classes. Each inbreeding distribution is associated to a different time in  
7 the past: recent inbreeding generating longer homozygous stretches than more ancient. Our model is a mixture of  
8 exponential distribution implemented in a hidden Markov model framework that uses marker allele frequencies,  
9 genetic distances, genotyping error rates and the sequences of observed genotypes. Based on simulations studies,  
10 we show that the inbreeding coefficients and the age of inbreeding are correctly estimated. Mean absolute errors  
11 of estimators are low, the efficiency depending on the available information. When several inbreeding classes  
12 are simulated, the model captures them if their ages are sufficiently different. Genotyping errors or low-fold  
13 sequencing data are easily accommodated in the hidden Markov model framework. Application to real data sets  
14 illustrate that the method can reveal recent different demographic histories among populations, some of them  
15 presenting very recent bottlenecks or founder effects. The method also clearly identifies individuals resulting from  
16 extreme consanguineous matings.

## 17 **Introduction**

18 With his pioneering work on self-fertilization, Darwin early noticed that mating relatives generally leads to off-  
19 spring with a reduced fitness (Darwin, 1876). This phenomenon now referred to as inbreeding depression may  
20 mostly result from an increased homozygosity for (recessive) deleterious variants although a lack of heterozy-  
21 gosity at loci displaying heterozygous advantage (overdominance) might also be involved (Charlesworth & Willis,  
22 2009). Accordingly, populations displaying high levels of individual inbreeding show a higher prevalence of mono-  
23 genic disorders (e.g., Charlier *et al*, 2008) or complex diseases (e.g., Rudan *et al*, 2003). Inbreeding depression can  
24 thus increase the risk of extinction by reducing the population growth rate (Hedrick & Kalinowski, 2000; Keller  
25 & Waller, 2002) although it may be conversely favorable in some conditions by purging deleterious variants from  
26 the population (Estoup *et al*, 2016). Assessing individual inbreeding is then of paramount interest to improve the  
27 management of populations under conservation or selection, and from a more general evolutionary perspective to  
28 better understand the genetic architecture of inbreeding depression.

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

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

46 type of methods ranges from simple estimates of individual heterozygosities (e.g., Szulkin *et al*, 2010) or homozy-  
47 gosities (e.g., Bjelland *et al*, 2013) to more advanced approaches based on the estimation of the realized genomic  
48 relationship matrix (VanRaden, 2008; Yang *et al*, 2010) or moment-based estimators to correct for population-  
49 structure in the estimation of population allele frequencies (e.g., Manichaikul *et al*, 2010). Their accuracy depends  
50 strongly on the number and informativeness of the genotyped markers (Kardos *et al*, 2015) but they always remain  
51 global in the sense that they can only capture the total amount of individual inbreeding. With genetic map infor-  
52 mation, one may alternatively rely on the identification of stretches of homozygous markers also referred to Runs  
53 of Homozygosity (RoH) (e.g., McQuillan *et al*, 2008) to estimate individual inbreeding at both a local genome  
54 scale and genome-wide (as the proportion of the genome contained in locally inbred regions). RoH are indeed  
55 most often interpreted as IBD chromosome segments that were inherited from a common ancestor without recom-  
56 bination (and mutation) in neither of them. Assessing the distribution of RoH within individual genomes has thus  
57 become popular to characterize inbreeding in a wide range of model species including humans (Kirin *et al*, 2010;  
58 McQuillan *et al*, 2008; Pemberton *et al*, 2012) or livestock (Bosse *et al*, 2012; Ferencakovic *et al*, 2013). RoH  
59 also allows to distinguish between recent and more ancient inbreeding (Kirin *et al*, 2010; Pemberton *et al*, 2012;  
60 Purfield *et al*, 2012) since pairs of IBD chromosomal segments tracing back to more remote ancestors are expected  
61 to be shorter because of a higher number of historical recombination events.

62 However, the main limitations of RoH-based approaches lie in their underlying rule-based procedure. For  
63 instance, the definition of the minimal number of homozygous markers (and segment length) and the maximum  
64 proportion of allowed heterozygous markers (to account for genotyping error) is mostly arbitrary. As a model-  
65 based alternative, Broman & Weber (1999) proposed a formal statistical approach to assess the IBD (or autozy-  
66 gous) status of the RoH they identified by accounting for population allele frequencies and genotyping error rates.  
67 Leutenegger *et al* (2003) further provided a full probabilistic modeling of the IBD process along the chromo-  
68 somes by developing a Hidden Markov Model (HMM). The HMM framework allows to make use efficiently of  
69 the available genetic information contained in the sequences of both homozygous and heterozygous markers and  
70 the linkage maps and can handle whole-genome sequence data (Narasimhan *et al*, 2016) including those obtained  
71 from low-fold sequencing experiments (Vieira *et al*, 2016). Although powerful, the aforementioned methods rely  
72 on a two-states HMM considering each marker either belongs to an IBD or a non-IBD chromosome segments.  
73 The transition probabilities between the (hidden) states of successive markers then depend on their given genetic  
74 distances, a parameter controlling the rate of changes per unit of genetic distance and the individual inbreeding  
75 coefficient. Considering only two states (IBD or non-IBD) thus amounts to assume that all the individual inbreed-

76 ing originates from one or several ancestors in a single generation in the past and that all the IBD segments have  
77 the same expected length. However, in both natural and domesticated populations, the sources of individual in-  
78 breeding are multiple, since they are all related to their usually complex past demography history, making such an  
79 hypothesis of a single inbreeding event highly unrealistic.

80 We herein propose to extend previous HMM by considering several IBD-classes, each associated with a differ-  
81 ent inbreeding age. This new model allows to provide a better fit to individual genetic data (either genotyping or  
82 sequencing data) and to refine the genomic partitioning of inbreeding into stretches of IBD regions from possibly  
83 different ancestral origins. To evaluate the accuracy of the methods, we carried out comprehensive simulation  
84 studies. In addition, three real data sets from human, dog and sheep populations were analyzed in more detail to  
85 illustrate the range of application of the methods. As a by-product of this study, a freely available program, named  
86 ZooRoH was developed to implement inferences under the newly developed model.

## 87 **The Models**

88 In the following we describe our HMM to model individual genomes as mixtures of IBD and non-IBD segments.  
89 We first consider a model with only two states (one IBD or autozygous class and one non-IBD class) and then  
90 describe the extension of the model to combine several IBD classes with varying time to the common ancestor (age  
91 measured in generations). To deal with the specificities of Next-Generation Sequencing (NGS) data (whole genome  
92 sequencing, low-fold sequencing, genotype-by-sequencing) that may provide less accurate genotype call than SNP  
93 chip arrays, we also propose alternative emission probabilities functions that integrate over the uncertainties of  
94 each possible genotype.

### 95 **The two-states model (1G model)**

96 The 1G model is similar to the HMM previously proposed by Leutenegger *et al* (2003) and assumes that the  
97 genome is partitioned in either IBD and non-IBD tracts that actually correspond to the two hidden states ( $K = 2$ ).  
98 The 1G model further relies on a one order Markov process to define the transition probabilities between successive  
99 hidden states, such a modeling representing a good approximation of the IBD process along the chromosome in the  
100 absence of interference (Lander & Green, 1987; Leutenegger *et al*, 2003; Thompson, 2008). Consider two adjacent  
101 loci  $M_{l-1}$  and  $M_l$  separated by  $r_l$  Morgans ( $l > 1$ ) and let  $G$  represent the size of the inbreeding loop i.e. twice the  
102 number of generations from a common ancestor and  $\rho$  the mixing coefficient corresponding to the proportion of

103 IBD segments in the genome. Under the 1G model,  $\rho$  can be interpreted as a measure of the individual inbreeding  
 104 coefficient  $F$  (Leutenegger *et al*, 2003). Let further  $S_l$  denote the (hidden) state of  $M_l$  with  $S_l = 1$  and  $S_l = K = 2$   
 105 for an IBD and non-IBD state respectively. The four transition probabilities between the hidden states of every  
 106 pairs of consecutive markers are then defined as:

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

108 This amounts to assume that co-ancestry changes (leaving an IBD or non-IBD segment) between two adjacent  
 109 markers  $M_{l-1}$  and  $M_l$  occur with a probability equal to  $1 - e^{-r_l G}$ . It should thus be noticed that the same rate of  
 110 co-ancestry changes ( $G$ ) is used for both IBD and non-IBD tracks since we model the inheritance of chromosomal  
 111 segments present in a single generation (that of the common ancestor). Under such assumptions, the length of  
 112 IBD segments (inherited from a single ancestor) is exponentially distributed with an expected mean equal to  $\frac{1}{G}$ .  
 113 Because consecutive segments in the genome might belong to the same class, the overall lengths of the IBD and  
 114 non-IBD segments have expected means equal to  $\frac{1}{G(1-\rho)}$  and  $\frac{1}{G\rho}$  respectively (Leutenegger *et al*, 2003). Vieira *et al*  
 115 (2016) also used a similar approach to model the transition probabilities whereas Narasimhan *et al* (2016) relied  
 116 on a unique parameter for the transition probabilities that integrates both  $G$  and  $\rho$ .

### 117 **Extension to multi-states models (KG models)**

118 With a unique IBD class, the 1G model described above considers that all the IBD segments have approximately  
 119 the same age either because they originate from a single ancestor (one strong inbreeding event) or from multiple  
 120 ancestors in the same generation (e.g., during a bottleneck). Population history might however lead to far more  
 121 complex patterns. For instance, common ancestors tracing back to different generations can be frequent in small  
 122 populations, in populations under strong selection or in endangered populations with declining size. We therefore  
 123 propose to extend the model to  $K_{\text{IBD}}$  different IBD classes, each characterized by their own mixing coefficient  $\rho_c$   
 124 and rate  $G_c$  ( $c \in (1, K_{\text{IBD}})$ ). Note that  $G_c$  might be interpreted as twice the age (in generations) of the inbreeding  
 125 class  $c$ . Common ancestors from IBD class  $c$  transmitted IBD segments whose lengths are exponentially distributed  
 126 with a mean equal to  $\frac{1}{G_c}$ . For the sake of generality, we may include several non-IBD classes but in the present

127 study we only used one non-IBD class labeled  $K$  (i.e., the total number of classes  $K = K_{\text{IBD}} + 1$ ) with a mixing  
 128 proportion  $\rho_K$  and a change rate  $G_K$ . The transition probabilities between the hidden states  $S_{l-1}$  and  $S_l$  of two  
 129 adjacent loci  $M_{l-1}$  and  $M_l$  read:

$$130 \quad \begin{cases} \mathbb{P}[S_l = a \mid S_{l-1} = a] = e^{-r_l G_a} + (1 - e^{-r_l G_a}) \rho_a \\ \mathbb{P}[S_l = a \mid S_{l-1} = b] = (1 - e^{-r_l G_b}) \rho_a \end{cases} \quad (2)$$

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

### 135 **Emission probabilities and extension to NGS data.**

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

### 146 **Emission probabilities for genotyping data.**

147 Let  $p_{li}$  be the population allele frequency of allele  $A_{li}$  which is assumed to be known. If the two chromosomes are  
 148 IBD in  $M_l$  ( $S_l \neq K$ ), we define the emission probabilities of the genotype  $A_{li}A_{lj}$  as follows:

$$149 \quad \mathbb{P}[A_{li}A_{lj} \mid S_l \neq K, p_{li}, \epsilon] = \begin{cases} (1 - \epsilon)p_{li} & \text{if } i = j \\ \frac{2\epsilon}{I_l(I_l-1)} & \text{if } i \neq j \end{cases} \quad (3)$$

150 where  $\epsilon$  is the probability (assumed to be known) to observe an heterozygous marker when the two underlying  
 151 chromosomes are IBD in  $M_l$  either resulting from a genotyping error or a recent mutation. In other words, we  
 152 assume that the vast majority of the polymorphic markers were segregating in the population before the common  
 153 ancestors of the IBD segments and thus interpret recent mutations as genotyping errors. For non-IBD segments  
 154 (tracing back to much more ancient ancestors), each genotype emission probabilities are derived assuming Hardy-  
 155 Weinberg equilibrium and disregarding genotyping error (or recent mutation):

$$156 \quad \mathbb{P}[A_{li}A_{lj} | S_l = K, p_{li}, p_{lj}] = \begin{cases} p_{li}^2 & \text{if } i = j \\ 2p_{li}p_{lj} & \text{if } i \neq j \end{cases} \quad (4)$$

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

### 158 Emission probabilities for genotype likelihood data.

159 To account for genotype uncertainty, emission probabilities are obtained by integrating over all the possible geno-  
 160 types:

$$161 \quad \begin{cases} \mathbb{P}[Y_l | S_l \neq K] = \sum_{J_l} \mathbb{P}[Y_l | A_{li}A_{lj}] \mathbb{P}[A_{li}A_{lj} | S_l \neq K] \\ \mathbb{P}[Y_l | S_l = K] = \sum_{J_l} \mathbb{P}[Y_l | A_{li}A_{lj}] \mathbb{P}[A_{li}A_{lj} | S_l = K] \end{cases} \quad (5)$$

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

## 164 Materials and Methods

### 165 Inference

#### 166 Estimation of model parameters.

167 Assuming the population allele frequencies ( $p_{li}$ ) of each marker  $M_l$  and the error term  $\epsilon$  are known, the set of  
 168 parameters  $\Theta$  that needs to be estimated consists of the mixing proportions  $\rho$  and the rates (interpreted as ages  
 169 for the inbreeding classes)  $G$  of the defined IBD and non-IBD classes. Therefore,  $\Theta$  consists of two parameters  
 170 ( $\rho$  and one rate  $G$ ) for the 1G model and  $2K$  parameters for a multi-classes KG model (with  $K_{\text{IBD}} = K - 1$   
 171 inbreeding classes). For multiple-IBD models, we alternatively consider reducing the parameter space by pre-

172 defining the ages  $G_k$  of the  $K$  classes leading to only estimate the  $K$  mixing proportions  $\rho_k$  (hereafter called  
173 mixKG model). For all the models, parameter estimation was achieved with the Expectation-Maximization (EM)  
174 algorithm known as the Baum-Welch algorithm that is very popular in the HMM literature (Rabiner, 1989). The  
175 program ZooRoH implementing the algorithm for the different models is freely available at <https://github.com/tdruet/ZooRoH>. Unless otherwise stated, model parameters were estimated with 1000 iterations of the EM  
176 algorithm and setting  $\epsilon$  to 0.001.  
177

### 178 **Estimation of the realized local (locus-specific) inbreeding ( $\phi_l$ ).**

179 The Baum-Welch algorithm allows to estimate the local state probabilities that correspond in our case to the  $K$   
180 probabilities  $\mathbb{P}(S_l = c \mid \widehat{\Theta}, \mathbf{Y})$  that the two chromosome segments belong to the IBD class  $c$  ( $c \in (1, K_{\text{IBD}})$ ) or to  
181 the non-IBD class ( $c = K$ ) at the marker  $M_l$  position given the estimated parameter set  $\widehat{\Theta}$  and the observed genetic  
182 data  $\mathbf{Y}$ . These probabilities can be used to estimate both the realized genome-wide (over all the markers) and local  
183 (for each and every marker) inbreeding. Indeed, genetic data allows to directly infer the realized IBD status of an  
184 individual for each locus in the genome as opposed to pedigree-based inbreeding estimates that infer the expected  
185 IBD status for all the loci. More precisely, the local estimate  $\widehat{\phi}_l$  of the realized inbreeding at marker  $M_l$  is defined  
186 as the probability that this marker lies in an IBD segment and may thus be computed by summing over all its local  
187 IBD state probabilities (i.e., excluding the non-IBD class):

$$188 \quad \widehat{\phi}_l = \sum_{c=1}^{K_{\text{IBD}}} \mathbb{P}(S_l = c \mid \widehat{\Theta}, \mathbf{Y}) \quad (6)$$

### 189 **Estimation of the realized inbreeding associated to each IBD age-based classes ( $F_G^{(c)}$ ) and the genome-wide 190 inbreeding ( $F_G$ ).**

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

$$194 \quad \widehat{F}_G^{(c)} = \frac{1}{L} \sum_{l=1}^L \mathbb{P}(S_l = c \mid \widehat{\Theta}, \mathbf{Y}) \quad (7)$$

195 Finally, the genome-wide estimate of the realized individual inbreeding  $\widehat{F}_G$  is simply the average over the

196 genome of the local estimates obtained for the  $L$  markers:

$$197 \quad \widehat{F}_G = \frac{1}{L} \sum_{l=1}^L \widehat{\phi}_l = \sum_{c=1}^{K_{\text{IBD}}} \widehat{F}_G^{(c)} \quad (8)$$

### 198 **Model assessment.**

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

$$203 \quad \text{BIC} = -2\ln(\mathbb{P}(\mathbf{Y} | \widehat{\Theta})) + n_p \ln(L) \quad (9)$$

204 where  $\mathbb{P}(\mathbf{Y} | \widehat{\Theta})$  is the maximum of the likelihood function obtained with the estimated parameters  $\widehat{\Theta}$  (computed  
205 with the forward algorithm (Rabiner, 1989)),  $L$  is the number of markers and  $n_p$  is the number of parameters, i.e.,  
206  $n_p = 2K - 1$  for a KG model (with  $K-1$  IBD classes) and  $n_p = K - 1$  for a mixKG model (see above).

### 207 **Simulated data sets**

#### 208 **Simulation under the inference model.**

209 The model was first tested by simulating data under the inference models. We simulated genotyping data at bi-  
210 allelic markers (SNPs) for 500 individuals considering a genome that consisted of 25 chromosomes of 100 cM  
211 length (i.e., 100 Mb length assuming a cM to Mb ratio of 1). The marker density was set to 10, 100 or 1,000  
212 evenly spaced SNPs per Mb (i.e., 25,000, 250,000 or 2,500,000 SNPs in total). When simulating data under  
213 the 1G inference model, the individual genome is a mosaic of either IBD or non-IBD segments whose length  
214 is exponentially distributed with the same rate equals to the simulated  $G$  (twice the age in generations of the  
215 inbreeding event). For each chromosome in turn, we successively generated consecutive segments by sampling  
216 their length in the corresponding exponential distribution and randomly declaring them as IBD or non-IBD with  
217 a probability  $\rho$  and  $1 - \rho$  (where  $\rho$  represents the simulated mixing coefficients). The process stops when the  
218 cumulative length of the simulated segments was greater than 100 cM (the last simulated segment being trimmed  
219 to obtain a chromosome length exactly equal to 100 cM). Under the multi-states model with several IBD classes,  
220 simulations were performed sequentially, with successive waves of inbreeding starting with the most ancient. We  
221 started by simulating the most ancient IBD class with the process described above. Then, each new IBD class was

222 simulated similarly (with its own  $G_i$  and  $\rho_i$ ) except that new inbreeding (IBD) masked previous classes whereas  
223 non-IBD segments did not change previously simulated states.

224 To simulate genotyping data, we first randomly sampled for each SNP the population frequency of an arbitrarily  
225 chosen reference allele either i) from an empirical distribution derived from real cattle genotyping SNP assay  
226 and WGS data (Figure S1), or ii) from a (U-Shaped) distribution  $\beta(0.2, 0.2)$  that mimics NGS data (Figure S1).  
227 Given the simulated IBD status of the segments on which each SNP lie (see above), we used these sampled allele  
228 frequencies to simulate SNP genotypes as described for the emission probabilities above (eqs. 3 and 4). We used  
229 the parameter  $\epsilon$  set to either 0.1% or 1% to introduce random genotyping errors (changing one genotype to one of  
230 the two other genotypes) and to evaluate the robustness of the models.

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

$$237 \begin{cases} \mathbb{P}[Y_l | A_{l1}A_{l1}] &= 1^{c_{l1}} 0^{t_{l2}} \\ \mathbb{P}[Y_l | A_{l1}A_{l2}] &= \left(\frac{1}{2}\right)^{t_{l1}+t_{l2}} \\ \mathbb{P}[Y_l | A_{l2}A_{l2}] &= 1^{c_{l2}} 0^{t_{l1}} \end{cases} \quad (10)$$

238 Finally, to assess the accuracies of the model estimation, we computed the Mean Absolute Error (MAE) for  
239 each parameter  $\alpha$  of interest as:

$$240 MAE(\alpha) = \frac{1}{N} \sum_{n=1}^N |\hat{\alpha}_n - \alpha_n| \quad (11)$$

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

### 243 Simulations under a discrete time Wright-Fischer process.

244 The inference model we used is based on hypotheses (exponential distribution for length of IBD segments, Hardy-  
245 Weinberg equilibrium in non-IBD states, etc.) commonly used and that have been proven to work well (e.g.,  
246 Leutenegger *et al*, 2003; Vieira *et al*, 2016). Still, we performed additional simulations relying on population

247 genetics models to obtain simulated data less dependent on these assumptions. To that end we used the program  
248 ARGON (Palamara, 2016) that simulates data under a discrete time Wright-Fischer process.

249 With constant and large effective population size  $N_e$ , inbreeding is expected to be low and to be spread over  
250 many generations. To concentrate inbreeding in specific age classes we simulated bottlenecks keeping large  $N_e$   
251 outside these events to reduce the noise due to inbreeding coming from other generations. In the first scenario *WF1*,  
252 we considered an ancestral population  $P_0$  with a constant haploid effective population size equal to  $N_{e0}=20,000$  that  
253 split in two populations  $P_1$  and  $P_2$  at generation time  $T_s$  in the past with respective population sizes  $N_{e1}=10,000$  or  
254 100,000 (according to the scenario) and  $N_{e2}=10,000$ . During four generations centered around generation  $T_b \ll T_s$   
255 in the past,  $P_1$  experienced a bottleneck with an (haploid) effective population size equal to  $N_{eb}$  and recovered its  
256 initial size. Population  $P_2$  that always maintains a constant size is actually used to select markers that were also  
257 segregating in the ancestral population  $P_0$  (only markers segregating at  $MAF \geq 0.05$  in both populations  $P_1$  and  
258  $P_2$  were kept for further analyses). The different simulation parameters are expected to have various impacts on  
259 the distribution of inbreeding. For instance for larger  $T_s$ , inbreeding tends to accumulate after the two populations  
260 split and selected markers will have an older origin. Similarly, the larger  $N_{e1}$ , the less inbreeding is accumulating  
261 outside the bottleneck while with smaller  $N_{eb}$ , more inbreeding is created during the bottleneck. In total, 50  
262 diploid individuals were simulated in both populations  $P_1$  and  $P_2$  considering a genome that consisted of a single  
263 chromosome of 250 cM length (i.e., 250 Mb assuming a cM to Mb ratio of 1). The mutation rate was set to  
264  $\mu = 10^{-8}$  and we use the functionalities of ARGON to identify all the IBD segments  $> 10$  kb and to obtain their  
265 ages (generation time of the most recent common ancestor).

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

## 271 **Human, dog and sheep real data sets**

272 For illustration purposes, we used publicly available genotyping data from *i*) the Human Genome Diversity Panel  
273 (HGDP) (Jakobsson *et al*, 2008) as downloaded from [ftp://ftp.cephb.fr/hgdp\\_supp10/Harvard\\_HGDP-CEPH](ftp://ftp.cephb.fr/hgdp_supp10/Harvard_HGDP-CEPH;);  
274 *ii*) the dog LUPA project (Vaysse *et al*, 2011) as downloaded from <http://dogs.genouest.org/SWEEP.dir/Supplemental.html>; and *iii*) the Sheep Diversity panel (Kijas *et al*, 2012) as downloaded from the WIDDE

276 database (Sempere *et al*, 2015). We then used the software PLINK (Purcell *et al*, 2007) to process and filter the  
277 genotyping data by removing individuals with a genotyping call rate below 90% and only keeping autosomal  
278 SNPs that had call rate > 95% and a MAF > 0.01 (in the original data set). As a result, the final data sets consisted  
279 of 304,406, 152,151 and 48,872 SNPs in human, dog and sheep respectively. For each species, we restricted our  
280 analysis to a subset of six populations corresponding to *i*) Karitiana (n=13), Pima (n=14), Melanesian (n=11),  
281 Papuan (n=17), French (n=28) and Yoruba (n=22) in humans; *ii*) Doberman Pinschers (n=25), Irish Wolfhounds  
282 (n=11), Jack Russell Terriers (n=12), English Bulldogs (n=13), Border Terriers (n=25) and Wolves (n=12) for  
283 the dog data set; and *iii*) Soay (n=110), Wiltshire (n=23), Dorset Horn (n=21), Milk Lacaune (n=103), Rasa  
284 Aragonesa (n=22) and Rambouillet (n=102) in sheep. Note that, within each population, markers with a MAF  
285 below 0.01 (within a population) were discarded from the analysis.

## 286 Results

### 287 Performance of the different models

#### 288 Analyzing data simulated under the 1G inference model.

289 We first analyzed individual genomes of 2,500 cM (with a marker density of 10 SNPs per cM) that were simulated  
290 under the 1G inference model, i.e., the simplest model. Depending on the two chosen simulation parameters (age  
291 of inbreeding  $G$  and mixing proportion  $\rho$ ), these individual genomes thus consisted of a mosaic of IBD and non-  
292 IBD segments (in proportions  $\rho$  and  $1 - \rho$  respectively) that both originated from the same ancestral generation  
293 ( $G/2$  generations ago). In total, we analyzed with the 1G, the 2G, the 3G and the 4G models, 500 individuals per  
294 simulated scenarios, considering in total 33 different scenarios representatives of a wide range of values for both  
295  $G$  (from  $G = 2$  to  $G = 256$ ) and  $\rho$  (from  $\rho = 0.0075$  to  $\rho = 0.5$ ). As mentioned in the Model section above, under  
296 the 1G model that was used for these simulations,  $\rho$  is highly similar to the individual inbreeding  $F_G$ . The results  
297 obtained from the analyses under the 1G model are detailed in Table 1 for 20 different scenarios. In addition, tables  
298 S1 and S2 give the results from the analyses under all the four models (1G, 2G, 3G and 4G) for all the 33 different  
299 scenarios.

300 Overall, estimates of both model parameters ( $\widehat{G}$  and  $\widehat{\rho}$ ) and individual inbreeding  $F_G$  obtained under the 1G  
301 model (Table 1 and Table S1) were found virtually unbiased and quite accurate (small MAE) irrespective of the  
302 considered scenarios. As expected, the 1G model performed even better when the number of IBD segments was

303 higher and these were longer (smaller  $G$ ) since more SNPs are available for their identification. For instance, for a  
304 given simulated  $\rho$  (e.g.,  $\rho \approx F_G = 0.100$ ), the MAE of  $\widehat{F}_G$  increased with larger simulated  $G$  (e.g., from  $1.1 \times 10^{-3}$   
305 when  $G = 16$  to  $4.6 \times 10^{-3}$  when  $G = 256$ ). The performance of the 1G model to estimate local inbreeding ( $\phi_l$ )  
306 was further evaluated by computing the corresponding MAE either for all the SNPs ( $\widehat{\phi}_l$ ) or for the SNPs lying  
307 within IBD segments only ( $\widehat{\phi}_{l\text{IBD}}$ ) (Table 1 and Table S1). Note that for every simulated SNP  $l$ , the actual  $\phi_l$  value  
308 is known (i.e.,  $\phi_l = 0$  or  $\phi_l = 1$  if the SNPs is within a non-IBD or a IBD segment respectively). Hence, if the  
309 model performs well and all the  $\phi_l$  are accurately estimated (i.e.,  $\widehat{\phi}_l$  close to 0 or 1 for SNPs within a non-IBD or  
310 a IBD segment respectively), the MAE of  $\widehat{\phi}_l$  should be close to 0. Conversely, departure of the  $\widehat{\phi}_l$  MAE from 0  
311 indicates that IBD (respectively non-IBD) positions have non-zero probability to be non-IBD (respectively IBD).  
312 Besides, inspecting the  $\widehat{\phi}_{l\text{IBD}}$  MAE allows to restrict attention to the prediction accuracy of truly IBD segments. As  
313 shown in Table 1, when inbreeding is recent ( $G < 32$ , i.e. less than 16 generations ago) MAE for both  $\widehat{\phi}_l$  and  $\widehat{\phi}_{l\text{IBD}}$   
314 are close to 0 indicating that both IBD and non-IBD positions are correctly identified with a high support. Also, at  
315 constant level of overall (simulated) inbreeding (e.g.,  $\rho \approx F_G = 0.125$ ) the accuracy decreases with higher value of  
316  $G$  (e.g., from  $1.0 \times 10^{-2}$  when  $G = 4$  to  $2.1 \times 10^{-2}$  when  $G = 8$  for the  $\widehat{\phi}_{l\text{IBD}}$  MAE). When considering more ancient  
317 (and/or) lower simulated inbreeding values, the  $\widehat{\phi}_{l\text{IBD}}$  MAE increased faster than the overall  $\widehat{\phi}_l$  MAE. This indicates  
318 that there is not enough information (number of SNPs per IBD segments) to confidently classify some positions,  
319 in particular those within the shortest IBD segments, the longest IBS segments or the segments boundaries. It is  
320 however important to notice that the local inbreeding estimates  $\widehat{\phi}_l$  always remained very well calibrated, i.e., for  
321 any  $p \in (0, 1)$ , the proportion of SNPs truly lying within IBD segments among the SNPs with  $\widehat{\phi}_l \approx p$  was close to  
322  $p$  (Figure S2). Accordingly, and as mentioned above, the global estimators of individual inbreeding ( $F_G$ ) and the  
323 model parameters ( $\rho$  and  $G$ ) remained accurate (Table 1).

324 [Table 1 about here.]

325 As shown in Table S1, the estimates of  $G$  for the IBD class under the 2G model started to be substantially  
326 biased for scenario with  $G \geq 128$ . More interestingly, the performances of the 2G model (Table S1) and both the  
327 3G and 4G models (Table S2) were highly similar to those of the 1G model for the estimation of both genome-wide  
328 ( $F_G$ ) and local ( $\phi_l$ ) individual inbreeding.

329 **Analyzing simulated data with several underlying IBD classes.**

330 We further evaluated the performances of the different models on simulated data sets with more than one class  
331 for the underlying IBD segments, i.e. for which inbreeding originated from several sources of different ages  $G_k$   
332 and contributions  $F_G^{(k)}$  to the overall inbreeding. We detail hereafter the analyses of individual genomes of 2,500  
333 cM (with a marker density of 10 SNPs per cM) that were simulated under the 3G inference model, i.e., assuming  
334 two different classes for IBD segments and one non-IBD class. Each simulation scenario was thus defined by the  
335 ages of inbreeding ( $G_1$  and  $G_2$ ) and the mixing proportions ( $\rho_1$  and  $\rho_2$ ) of the two classes of IBD segments. It  
336 should be noticed that the simulated mixing proportions ( $\rho_1$  and  $\rho_2$ ) directly controlled (and are generally close to)  
337 the amount of inbreeding originating from their corresponding IBD class. As shown in Table 2 for six different  
338 scenarios (and Tables S3 and S4 for a total of 23 different scenarios), estimates of the overall individual inbreeding  
339 ( $F_G$ ), of the ages ( $G_1$  and  $G_2$ ) and of the inbreeding contributions ( $F_G^{(1)}$  and  $F_G^{(2)}$ ) for the two IBD classes were close  
340 (but slightly biased) to the simulated values providing the differences between the ages of the two IBD classes  
341 was large enough (e.g.,  $G_1/G_2 \geq 16$ ), i.e., the overlap between the distributions of the IBD segments lengths is  
342 reduced. As the difference between the ratio of successive  $G_i$  became smaller, all inbreeding tended to concentrate  
343 in the first IBD class that had an overestimated age for small simulated  $G_1$  (Table 2 and Table S3). For instance,  
344 for the scenario with  $G_1 = 4$  ( $\rho_1 = 0.125$ ) and  $G_2 = 16$  ( $\rho_1 = 0.100$ ),  $med(\widehat{F_G^{(1)}}) = 0.195$  (*med* standing for median)  
345 and  $med(\widehat{F_G^{(2)}}) = 0.004$  while  $med(\widehat{G_1}) = 7.20$  and  $med(\widehat{G_2}) = 391$  across the 500 simulated individuals (Table 2).  
346 Strikingly however, the overall individual inbreeding  $F_G$  always remained very well estimated with  $MAE \leq 0.005$   
347 for all scenarios (Table 2 and Table S4). Finally, as for the simulations under the 1G model previously considered,  
348 accuracy in the estimation of local inbreeding was found to mostly depend on the ages  $G_1$  and  $G_2$  (Table 2 and  
349 Table S5), the MAE for both  $\widehat{\phi}_l$  and  $\widehat{\phi}_{IBD}$  lying in a similar range than the one observed previously on data simulated  
350 under the 1G model. More precisely, given the relatively sparse SNP density considered, MAE remained accurate  
351 (i.e.,  $\leq 0.05$ ) while  $G_1 < G_2 \leq 64$  but started to increase for higher values probably due to the inclusion of smaller  
352 IBD segments.

353 [Table 2 about here.]

354 To provide insights on the behavior of our model to a misspecification of the underlying number of IBD classes,  
355 we also analyzed these data simulated under the 3G model with the 1G, the 2G and 4G models. As expected, when  
356 considering the 1G and 2G models, the estimated age of the single assumed IBD class was intermediate between  
357 the two simulated  $G_1$  and  $G_2$  actual values (Table S3). In agreement with previous findings, the 1G and 2G lead to

358 highly similar estimates except for large  $G_1$  and  $G_2$  for which the estimated  $G$  tended to be higher with the 2G than  
359 the 1G model (e.g.,  $med(\widehat{G}) = 181$  and  $med(\widehat{G}) = 201$  respectively for the scenario with  $G_1 = 128$  and  $G_2 = 256$ ).  
360 More interestingly, using the 1G and 2G models (i.e., with a single IBD class) to analyze these data resulted in an  
361 underestimation of  $F_G$  for scenarios with a marked differences between  $G_1$  and  $G_2$  (Table S4). Conversely, using  
362 an over-parameterized model such as the 4G did not introduce any additional bias compare to the 3G model. For  
363 instance, for the scenario with  $G_1 = 4$  ( $\rho_1 = 0.125$ ) and  $G_2 = 256$  ( $\rho_1 = 0.100$ ) that lead to a median realized  
364 inbreeding equal to 0.211 across the 500 simulated individuals, the median estimated inbreeding was equal to 0.162  
365 with both the 1G and 2G models while it was equal to 0.208 and 0.209 with the 3G and 4G models respectively  
366 (Table S4). This suggested that the 1G and 2G model failed to capture some inbreeding. Accordingly, when  
367 focusing on the estimation of local inbreeding (Table S5), although the 1G and 2G models displayed a lower MAE  
368 for  $\widehat{\phi}_l$  (i.e., computed over all the SNPs), this was essentially driven by SNPs lying in non-IBD segments. Indeed,  
369 both the 3G and 4G resulted in a lower MAE for  $\widehat{\phi}_{IBD}$  (i.e., computed over SNPs lying within IBD segments)  
370 suggesting these model allowed to better capture IBD segments at the expense of a slightly higher misassignment  
371 of SNP lying in non-IBD segments.

372 Overall, similar conclusions about the performance of the models to estimate the simulated parameters could  
373 be drawn when considering data sets with more than two underlying IBD classes (see Table S6 for results on data  
374 sets simulated and analyzed under the 4G model). It should however be noticed that increasing the number of IBD  
375 classes in the model also increased misassignment of IBD segments to their actual IBD class (Figure S3). In other  
376 words, some IBD segments, although correctly identified as IBD, might display a non-zero probability to belong to  
377 an incorrect IBD class (most generally a neighboring one). As a result, when increasing the number of simulated  
378 IBD classes, higher deviations of the estimated inbreeding age ( $G_c$ ) and contribution ( $F_G^{(c)}$ ) of each classes from  
379 their actual values could be observed (e.g., Table S6). Nevertheless, for higher ratio between successive class ages,  
380 these estimates remained fairly good. Importantly and as shown in previous simulations, the overall individual  
381 inbreeding ( $F_G$ ) was accurately estimated in all scenarios and MAE for local inbreeding mostly depended on the  
382 age of the IBD segments.

### 383 **Using a set of $K$ predefined IBD-classes (the mixKG model).**

384 For a given model, instead of estimating the ages  $G_k$  of the different IBD classes, an alternative is to use a set of  
385 predefined age-classes and to only estimate the mixing proportions ( $\rho_k$ ). To illustrate and evaluate this strategy  
386 we hereby considered models consisting of 9, 11 or 13 IBD-classes depending on the simulated marker density

387 (see below) and one non-IBD class leading to the so-called mix10G, mix12G and mix14G models according to  
388 our nomenclature. For each model, the predefined ages of the  $K - 1$  IBD-classes always ranged from 2 to  $2^{K-1}$   
389 (with  $G_k = 2^k$  for each class  $k \in (1, K - 1)$ ) while the age of the unique non-IBD class was the same as the  
390 older IBD class (i.e.,  $G_K = G_{K-1} = 8192$ ). Application of these mixKG models to the various data sets previously  
391 generated under the 1G, the 3G and the 4G inference models proved highly efficient (Table S7 and S8). For instance  
392 and in agreement with above results, the mix10G model provided accurate estimation of the overall inbreeding  
393  $F_G$  (MAE always lower than 0.005 irrespective of the simulated scenarios) but also of the local inbreeding as  
394 indicated by MAE's that were always as good as the best alternative model (e.g., compare Table S7 and Table S5).  
395 Moreover, such models with pre-defined ages for the IBD classes allowed to provide indications on the actual ages  
396 of inbreeding  $G_k$ . We indeed observed that the estimated inbreeding contributions ( $F_G^{(k)}$ ) for the  $K - 1$  IBD classes  
397 were mainly concentrated in those IBD-classes with pre-defined ages close to the true simulated ones as shown in  
398 Figure 1 for a dense SNP data sets (1000 SNPs per Mb) analyzed under the mix14G models and in Figures S4 to  
399 S8 for additional simulated data sets with smaller SNP density (either 10 or 100 SNPs per Mb) that were analyzed  
400 under mix10G or mix12G models.

401 [Figure 1 about here.]

#### 402 **Model comparisons and selection.**

403 We finally evaluated the BIC criteria to compare the models. When comparing different KG models (from 1G to  
404 6G) applied to various simulation scenarios (ranging from 1 to 4 simulated IBD-distributions), we observed that  
405 the BIC criterion tended to support the correct underlying models and never provided support for models with a  
406 number of classes  $K$  higher than the simulated ones (Tables S9 and S10). Nevertheless, for simulations involving  
407 IBD segments from several classes (i.e., simulated under the 3G to 5G inference models), BIC may favor a model  
408 with a smaller number of IBD classes than the actual ones when the ages between successive classes are too close,  
409 although increasing SNP density improves the BIC resolution (Table S10). It should also be noticed that the BIC  
410 criterion never provided a stronger support in favor of the MixKG model (as defined above) when compared to the  
411 6 others models considered (from 1G to 6G), possibly due to its higher number of parameters (e.g.,  $n_p = 13$  for  
412 the Mix14G model against  $n_p = 11$  for the 6G model) (Tables S11 and S12). Yet, for simulations with several IBD  
413 classes (Table S12), the BIC support was generally higher than for the 1G and 2G models.

#### 414 **Sensitivity of the models to genotyping error and marker informativeness**

415 As only partially investigated above, when analyzing data with different SNP density, we expected that SNP in-  
416 formation content, both in terms of marker density and genotyping accuracy, might be a key determinant of the  
417 resolution of the models. As a matter of expedience, we investigated this further by focusing on the 1G model (for  
418 both simulation and analyses) and evaluated the effect on its overall performances of changing the marker density  
419 and the SNP informativeness as summarized by the SNP allele frequency spectrum (AFS). Results confirmed that  
420 both the estimation of  $G$  and the identification of IBD positions associated to older inbreeding events always im-  
421 proved when increasing marker density and informativeness (Table 3 ). For instance, when the simulated  $G = 256$ ,  
422 the MAE for  $\widehat{G}$  (respectively  $\widehat{\phi}_{IBD}$ ) dropped from 36.9 (respectively 0.7313) with a marker density of 10 SNPs per  
423 cM and a  $\beta(0.2, 0.2)$  AFS to 8.06 (respectively 0.1994) with a marker density of 100 SNPs per cM and to 5.79  
424 (respectively 0.0824) if, in addition, AFS was array-like. We also observe a better assignation of IBD segment to  
425 the correct IBD class with higher marker density (Figure S3). It is interesting to note that, at least for the range of  
426 parameters considered,  $F_G$  was accurately estimated irrespective of the marker densities and informativeness.

427 [Table 3 about here.]

428 We also investigated the sensitivity of the 1G model to the quality of genotyping or sequencing data. As shown  
429 in Table S13, when considering genotyping data (analyzed by setting  $\epsilon = 0$  for comparison purposes), we found  
430 that the presence of genotyping errors (either 1% or 0.1%) had little impact on the estimation of  $F_G$ , moderate  
431 effects on the estimation of local inbreeding  $\phi_l$  but estimates of  $G$  were strongly affected with an upward bias and  
432 an increased MAE. The magnitude of these effects was actually a function of the number of incorrect genotypes  
433 per IBD segment that increased the probability of observing heterozygotes and thus to cut the IBD segment into  
434 smaller ROH. As a result, the impact of genotyping errors was stronger for more recent inbreeding, at higher  
435 marker density and for higher error rate (Table S13). Interestingly, when analyzing the genotyping data with an  
436 appropriate error term i.e., setting  $\epsilon = 0.01$  (respectively  $\epsilon = 0.001$ ) for data simulated with a genotyping error  
437 of 1% (respectively 0.1%), the estimates of  $G$  became unbiased (Table S13). The accuracies with a 0.1% error  
438 were similar than without error but the MAE still remained larger with 1% genotyping errors and older inbreeding  
439 origins. Note that including a small error term in the model ( $\epsilon \neq 0$ ) had little influence in the absence of genotyping  
440 errors.

441 We finally evaluated the sensitivity of the 1G model to various confidence levels in genotype calling by sim-  
442 ulating data that mimic low-fold sequencing (or GBS) data for which several genotypes may have a non-zero

443 probability. In these cases, read count data were simulated with a higher SNP density than above (1,000 SNP per  
444 cM) and variable coverage (from 1 to 10X). For each simulated SNP, the likelihood of the three possible genotypes  
445 were derived from the read count data as described in the Material and Methods section. The analyzed data sets  
446 then either consisted of i) the actual SNP genotypes (ideal situation) or ii) vectors of genotype likelihoods. As  
447 detailed in Table S14, we found that the model performed well in estimating the global parameters  $G$  and  $F_G$  with  
448 sequencing data. As expected, the performances improved with higher coverages and were similar than those ob-  
449 tained with the corresponding genotyping data as coverages  $\geq 5X$ . Lowering sequencing coverages might indeed  
450 be viewed as decreasing SNP informativeness thereby leading to less accurate estimates for the different parame-  
451 ters (increased MAE), particularly for simulation in which inbreeding had an older origin (smaller IBD segments).  
452 For instance, for simulated  $G \geq 512$  and 1X coverage, both  $F_G$  and  $G$  were slightly underestimated (and to a lesser  
453 extent with 2X coverage) while for  $G \leq 256$ , both global and local ( $\phi_l$ ) estimates were accurate even with coverage  
454 as low as 1X (Table S14).

## 455 **Simulations under a discrete time Wright-Fischer process**

456 To evaluate the robustness of the model to departure from model assumptions, we analyzed data simulated under  
457 a discrete-time Wright-Fisher process using the recently developed program ARGON (Palamara, 2016). For our  
458 purposes, a decisive advantage of ARGON is that it allowed to identify all the IBD segments (here we only  
459 considered those  $\geq 10$  kb) and to obtain their age (i.e., time to most recent ancestor or TMRCA). Inbreeding was  
460 generated by assuming population histories with either i) a strong bottleneck in the recent past followed by a rapid  
461 expansion as might be observed in invasive populations (WF1 scenarios) or ii) a reduced effective population size  
462 in the last twenty generations as might be observed in some domestic populations (WF2 scenarios). In total we  
463 considered 12 different WF1 scenarios and two different WF2 scenarios (see Material and Methods) and simulated  
464 50 diploid individuals per scenario. As illustrated in Figure 2A for one WF1 scenario (see Figures S9 and S10  
465 for all the 12 WP1 and the 2 WP2 scenarios respectively), the simulated history lead as expected to an enrichment  
466 in IBD segments that trace back to the bottleneck period within the simulated individual genomes (about 20% on  
467 average in Figure 2A). Yet, in most scenarios, a substantial proportion of inbreeding was associated to more ancient  
468 classes that accumulate inbreeding over many more generations. Indeed, a segment was considered IBD if it traced  
469 back to an ancestor from a generation more recent than the split time ( $T_s = 10^3$  or  $T_s = 10^4$  generations depending  
470 on the scenarios) of two modeled populations (see Material and Methods). Accordingly, in WF1 scenarios, this  
471 proportion increased with lower effective population size ( $N_{e1}$ ), older split time ( $T_s$ ) and to a lesser extent higher

472 bottleneck population size ( $N_{eb}$ ) and timing ( $T_b$ ) (Figures S9 and S10).

473 We analyzed all these simulated data sets with a mix10G model that consisted of 9 IBD-classes with predefined  
474 ages ranging from 2 to 512 (with  $G_k = 2^k$  for each class  $k$ ) and one non-IBD class that had the same age as the  
475 older IBD class (i.e.,  $G_{10} = G_9 = 512$ ). The choice for a MixKG model was motivated by our previous findings  
476 that demonstrated it was informative to date the origin of inbreeding and performed as well as other models in  
477 estimating local and overall inbreeding. In addition, it allowed to compare all the simulated individuals according  
478 to the same age-based partitioning of inbreeding.

479 [Figure 2 about here.]

480 As shown in Figure 2B (see Figures S11 and S12 for all the 12 WP1 and the 2 WP2 scenarios respectively),  
481 our HMM always allowed to efficiently identify IBD segments tracing back to common ancestors with TMRCA  
482 smaller than 80 generations, since the underlying SNPs displayed an estimated local inbreeding probability ( $\phi_l$ )  
483 close to one. In agreement with results obtained on simulations performed under the inference model (see above),  
484 the power to identify IBD segments of older origin gradually decreased (towards values almost always lower than  
485 20% for TMRCA older than 5000 generations). Note that analyses of data sets simulated under the inference model  
486 showed that although the power was below one, overall inbreeding remained correctly estimated (see above).  
487 In addition, the model was found to perform well in assigning the identified IBD segments associated to the  
488 simulated bottlenecks since they were in their vast majority either assigned to their actual IBD class (i.e., with  
489 an age the closest to twice the age of the TMRCA) or to an immediately neighboring one. For instance, in the  
490 scenario considered in Figure 2, the estimated proportions of the individual genomes assigned to IBD segments  
491 were concentrated in the IBD class with predefined ages equal to 32 ( $G_{32}$ ), 16 ( $G_{16}$ ) and to a lesser (but less  
492 variable) extent in the oldest IBD-class ( $G_{512}$ ) (Figure 2C and Figures S13 and S14 for all the 12 WP1 and the  
493 2 WP2 scenarios respectively). This was in agreement with the actual characteristics of the simulated individuals  
494 since IBD segments with a  $TMRCA \approx 16$  that contributed on average to about 20% of their genome (Figure 2A)  
495 were mainly assigned (up to 70%) to the IBD classes  $G_{32}$  and  $G_{16}$  (Figure 2D). Note that the oldest IBD class  
496  $G_{512}$  also captured some of these IBD segments together with a small proportion of those with an older TMRCA  
497 probably because these older IBD classes then become more frequent and have higher mixing coefficients. This  
498 effect was stronger when the bottleneck contributed less to the overall inbreeding and when the bottleneck was  
499 older. The performances of the model to correctly assign IBD segments however declined as the timing of the  
500 bottleneck was older or more generally as the proportion of inbreeding resulting from the period of reduced  $N_e$

501 was lower (Figures S13 and S14). Note that misassignment of IBD segments might also result from simulated  
502 segments being smaller/larger than expectations for a given pre-defined age  $G_k$  of the IBD class due to the stochastic  
503 nature of the Wright-Fisher process. In all cases however, we observed a peak of inbreeding in the IBD-class(es)  
504 corresponding to the period of reduced  $N_e$  or its neighbors (Figures S13 and S14). Overall, this simulation study  
505 thus confirmed that our model correctly identifies IBD-segments and gives good indications of the inbreeding's  
506 age.

### 507 **Application to human, canine and ovine real data sets**

508 We applied our model to individuals from human, dog and sheep populations, i.e., species representative of a wide  
509 range of demographic histories. Individuals were genotyped, as part of previous experiments (see Material and  
510 Methods) with assays containing various number SNPs (ca. 300K, 150K and 50K for human, dog and sheep indi-  
511 viduals respectively) leading to different SNP density (ca., 1 SNP per 10kb, per 20 kb and per 60 kb respectively).  
512 As a result, and for the reasons mentioned above, the genotyping data were further analyzed with i) a mix14G  
513 model that consisted of 13 IBD-classes with predefined ages ranging from 2 to 8192 (with  $G_k = 2^k$  for each class  
514  $k$ ) and one non-IBD class that had the same age as the older IBD class (i.e.,  $G_{14} = G_{13} = 8192$ ) for humans and  
515 dogs; and ii) a mix9G model that consisted of 8 IBD-classes with predefined ages ranging from 2 to 256 (with  
516  $G_k = 2^k$  for each class  $k$ ) and one non-IBD class ( $G_{10} = G_9 = 512$ ) for sheep to account for the smaller SNP  
517 density. To interpret the results, it is useful to remind that the ages  $G_k$  of the predefined classes are approximately  
518 twice the TMRCA and that populations have variable ratio between genetic and physical distances when averaged  
519 between sexes: 1.16 cM/Mb for human (Kong *et al*, 2010), 1.26 cM/Mb for sheep (Johnston *et al*, 2016) and 0.88  
520 cM/Mb for dog (Campbell *et al*, 2016). Indeed, we used for the analyses the SNP position on the physical maps  
521 accompanying the respective data sets. The estimated contribution of each pre-defined IBD class (averaged over  
522 all the individuals) are detailed for each populations and each species in Figure 3.

523 [Figure 3 about here.]

524 Regarding humans, the six populations considered here (French, Yoruba, Melanesian, Papuan, Pima and Kari-  
525 tiana) have already been thoroughly analyzed in other studies (e.g., Jakobsson *et al*, 2008) including a study that  
526 provided a detailed assessment of the distribution of ROH of different lengths (Kirin *et al*, 2010). Our results  
527 showed that the amount of overall inbreeding increased from Africans, Europeans, Oceanians to Native Americans  
528 from Central and Southern America with a generally remote origin (Figure 3A,B and Figure S17). More precisely,

529 the ages of the main contributing IBD-classes that were generally consistent within population were clearly related  
530 to the  $N_e$  of the corresponding populations (the older the larger). Hence, the peak of inbreeding was i) in the class  
531 with  $G_k = 512$  for Pima and Karitianas; ii) in classes with  $G_k = 512$  and  $G_k = 1024$  for Papuans and Melanesians;  
532 iii) in the class with  $G_k = 1024$  for French; and iv) in the class with  $G_k = 2048$  for Yoruba. Interpretation of such  
533 old inbreeding, more related to population characteristics than individual differences, must be done with caution  
534 (see Discussion). Nevertheless it should be noticed that in French or Oceanian populations we observed some  
535 individuals with more recent inbreeding but this remained limited compared to Pima and Karitiana where there  
536 is strong evidence of recent inbreeding, some of the individuals having more than 10% inbreeding in very young  
537 classes from  $G_k = 2$  to  $G_k = 8$  (Figure 4A and Figure S17). These observations are consistent with previous  
538 findings by Kirin *et al* (2010) based on ROH that suggested the presence of both recent (long ROH) and ancient  
539 (short ROH) inbreeding in Native Americans. Conversely, individuals from Oceanian populations did not display  
540 long ROH (several Mb long) but had an excess of ROH of intermediate length (between 1 and 2 Mb) indicating  
541 a reduced  $N_e$  in the past. Finally, individuals from European and African populations mostly showed background  
542 inbreeding (short ROH) that correlated with the underlying  $N_e$ . One major difference of our results with the afore-  
543 mentioned study by Kirin *et al* (2010) is that they only considered ROH > 500 kb leading to a lower estimated  
544 value (most probably downwardly biased) for the overall individual inbreeding.

545 Modern dog breeds present large amounts of inbreeding and are known to have experienced strong bottlenecks  
546 associated with the recent breed creation from a small number of founders (e.g., Vaysse *et al*, 2011). In addition,  
547 strong artificial selection and matings in small closed populations further contributed to increase inbreeding in the  
548 last decades (Lewis *et al*, 2015). Accordingly, as shown in Figure 3C,D and Figure S18, we observed massive  
549 inbreeding (sometimes higher than 20%) in the IBD-class with  $G_k = 16$  (a common ancestor approximately  
550 8 generations ago) in all the five breeds we analyzed but the Jack Russell Terrier that has a larger  $N_e$  (Vaysse  
551 *et al*, 2011). As expected also, wolves that did not experienced domestication did not present such an excess of  
552 inbreeding in recent generations. In each population (including wolves), some individuals were found to be highly  
553 inbred with an  $F_G \approx 50\%$  and approximately 25% of this inbreeding associated to an estimated common ancestor  
554 living only one or two generations ago (Figure 4B and Figure S18).

555 Finally, among the six sheep populations we investigated, three (the Rasa Aragonesa, Milk Lacaune and Ram-  
556 bouillet) displayed a large  $N_e$  (> 700) as described in Kijas *et al* (2012). Hence, individuals from the Rasa  
557 Aragonesa displayed almost no trace of inbreeding (max = 1.3% when cumulated up to the IBD-class with  $G_k = 8$ )  
558 while the cumulative inbreeding remained lower than 5% on average for individuals from the Milk Lacaune and

559 Rambouillet breeds up to classes  $G_k = 32$  (Figure 3E,F and Figure S19). Yet, some Rambouillet individuals pre-  
560 sented high levels ( $> 20\%$ ) of recent inbreeding (Figure 4C and Figure S19). Conversely, the Wiltshire ( $N_e = 100$ )  
561 and Dorsethorn ( $N_e = 137$ ) populations that went through a strong reduction in size in the early 1900's (Dorsethorn  
562 to a lesser extent) were both found to have a high level of recent inbreeding (Figure 3 and Figure S19). The main  
563 contributing IBD-class was the one with age  $G_k = 16$  for Wiltshire and  $G_k = 4$  to  $G_k = 32$  for Dorsethorn.  
564 Interestingly, the Wiltshire individuals were sampled from a New-Zealand flock that experienced several strong  
565 and successive bottlenecks in its recent history. Indeed, its founders were imported in 1974 from Australia where  
566 the breed had previously been introduced in 1952 and survived as a remnant population of as few as 12 ewes  
567 (O'Connell *et al*, 2012). Assuming a generation time of approximately 4 years in sheep, the distribution of the  
568 contribution of the most recent classes to the overall inbreeding is thus consistent with this demographic history.  
569 The sixth sheep population we investigated was the well known Soay sheep that had an estimated  $N_e = 194$  (Kijas  
570 *et al*, 2012) and experienced a strong founder effect since the current population derives from a flock of 107 in-  
571 dividuals that were transferred on the Hirta island in 1932 and then lived in complete isolation (Clutton-Brock &  
572 Pemberton, 2004). We observed for this population a small amount of recent inbreeding (for IBD classes with age  
573  $G_k \leq 16$ ), even lower than in Milk Lacaune or Rambouillet, but rather high levels of inbreeding associated with  
574 IBD classes of ages between between 32 and 64 generations (Figure 3E,F and Figure S19). Integrating over all the  
575 generations, the Soay sheep thus appeared on average even more inbred than Dorsethorn, which explains the small  
576 estimated  $N_e$ . However, despite this strong founder effect and the high resulting inbreeding level, we observed  
577 almost no individual with an inbreeding  $F_G > 5\%$  in the most recent generations. The Soay breed represents an  
578 interesting example of a wild population resulting from a founder effect and in expansion. To summarize, our  
579 model allowed to provide deeper insights into the very different patterns of individual inbreeding observable in the  
580 sheep breeds. Indeed, these inbreeding patterns ranged from small as in the Rasa Aragonesa or limited level (with  
581 a few overly and recently inbred individuals) as in the Rambouillet breed, to moderate to high inbreeding level that  
582 either originated from strong bottleneck in the very recent (Wiltshire) or recent (Soay) past, or that resulted from  
583 the cumulative effect of a less pronounced population size reduction over more generations (Dorsethorn).

584 [Figure 4 about here.]

585 Importantly, besides providing a global estimator of inbreeding for each individual, the model also informs  
586 on the partitioning of this individual inbreeding which is highly valuable. For instance, individuals born from  
587 extremely consanguineous marriages might be easily identified. As an illustration, Figure 4B showed three dogs

588 (Doberman #1 and #7, Border Terrier #1) that displayed approximately 25% inbreeding associated with the  $G_k = 2$   
589 or  $G_k = 4$  IBD-class (ancestors living one or two generations ago) unlike other dogs from the same population  
590 (Doberman #12 and Border Terrier #13). These three individuals are likely resulting from matings between a  
591 sire and its daughter. This indicates that inbreeding is still present in these populations and is not only due to  
592 the breed creation event but to further management practices. High level of inbreeding associated to parents or  
593 grand-parents are also observed in sheep (19.2% for Rambouillet #92 in Figure 4C) and even in human (8.9% for  
594 Karitiana #13 in Figure 4A). For all these individuals, however, these recent events accounts only for a fraction of  
595 total inbreeding and a substantial proportion of inbreeding is due to more remote ancestors. More generally, by  
596 partitioning the total amount of inbreeding among ancestors from different generations, our model provides a better  
597 understanding of the origins of inbreeding in each individual. Hence, individuals with a similar overall inbreeding  
598 might display a quite different pattern of ancestral contributions captured by our model. For instance, for the  
599 three sheep individuals (Rambouillet #87, Wiltshire #4 and Soay #26) represented in Figure 4C that all displayed  
600 an overall inbreeding of approximately 20%, the inbreeding is mostly associated to the IBD-class  $G_k = 16$  for  
601 the Wiltshire #4, to the two IBD-classes  $G_k = 32$  and  $G_k = 64$  for the Soay #26 whereas for the Rambouillet  
602 #87 individual, ancestors contributing to inbreeding trace back to a wide spectrum of generations (from  $G_k = 4$   
603 to  $G_k = 256$ ). These observations are consistent with patterns at the population level. Interestingly, individuals  
604 with higher levels of inbreeding (Wiltshire #14 and Rambouillet #92) display comparable patterns with inbreeding  
605 concentrated in the IBD-class  $G_k = 16$  for Wiltshire #14 and associated to several IBD classes for Rambouillet #92  
606 (Figure 4C). In humans (Figure 4A), Native Americans from Central and Southern America were found to display  
607 different make-ups than Oceanians with similar levels of overall inbreeding (e.g., Karitiana #7 vs Melanesian #11  
608 or Pima #4 vs Papuan #16). As expected from previous results, Oceanians actually displayed little traces of very  
609 recent inbreeding but accumulated more inbreeding in distant generations.

## 610 Discussion

611 In this study, we developed and evaluated HMM models that use genomic data to estimate and to partition individ-  
612 ual inbreeding into classes of different ages. There actually exist a wide variety of methods to estimate individual  
613 inbreeding and these have different properties. Pedigree-based methods rely on a genealogy (the inbreeding can  
614 only result from individuals within the genealogy) and predict the expected IBD status at a locus whereas genomic  
615 measures estimate realized inbreeding (the observed level of inbreeding). Genomic estimates can either be global,

616 giving a unique measure per individual, or local. Obviously, these latter measures provide more information but  
617 require a higher marker density. Assessing the distribution of ROH within individual genome have recently be-  
618 come popular to characterize global and local inbreeding (Kirin *et al*, 2010; McQuillan *et al*, 2008; Pemberton  
619 *et al*, 2012). Most often, however, estimators relying on ROH are categorizing pairs of chromosome segments as  
620 IBD or non-IBD and do not provide intermediate values. They rely on the assumption that if stretches of homozy-  
621 gous markers are sufficiently long, they are IBD. Many parameters must be defined (including minimal number  
622 of homozygous markers, minimal length of an homozygous track, maximal spacing between successive markers,  
623 maximal number of heterozygous SNPs in a RoH) and these depend on the population under study and on the geno-  
624 typing technology used. HMM's as those developed in this study make a better use of all the information since they  
625 take into account the marker allele frequencies, the genotyping error rates, the genetic marker map (the genetic  
626 distance between successive markers) and the expected length of IBD tracks. Initially designed for genotyping  
627 arrays (Leutenegger *et al*, 2003), they can easily be extended to NGS data (Narasimhan *et al*, 2016) including  
628 low-fold sequencing data (Vieira *et al*, 2016) or genotype-by-sequencing data as done in our study, whereas simple  
629 ROH are inappropriate in such conditions. HMM's also allow to automatically estimate some parameter of interest  
630 such as the frequency of IBD segments (a measure similar to the expected inbreeding if only one IBD-class is  
631 modeled) and their expected length. Finally, when relying on the Forward-Backward algorithm (as in our study),  
632 these models integrate all the available information to estimate the IBD probabilities of each marker in opposition  
633 to a binary classification as obtained with ROH or with a Viterbi algorithm in HMM (Leutenegger *et al*, 2003;  
634 Narasimhan *et al*, 2016; Vieira *et al*, 2016). Using a probabilistic model is particularly valuable when information  
635 is sparser and classification is more uncertain (e.g., for smaller and older IBD tracts, at lower marker density or  
636 informativeness, with higher genotyping error rates or with low-fold sequencing).

637 The most simple HMM we considered consists of a single IBD state (1G model) and is similar to several previ-  
638 ously proposed ones (Leutenegger *et al*, 2003; Narasimhan *et al*, 2016; Vieira *et al*, 2016). This amounts to either  
639 assume that a single common ancestor is responsible for inbreeding or that the vast majority of IBD segments trace  
640 back to ancestors that lived in the same past generation. However, most populations have complex demographic  
641 histories, with varying  $N_e$  and common ancestors of IBD segments are thus expected to originate from many dif-  
642 ferent generations in the past. As shown by our application in real data sets, even in domestic populations for  
643 which inbreeding might be expected to result from a limited number of founder individuals, individual inbreeding  
644 generally result from ancestors in different generations back in time probably due to the subsequent intense use of  
645 some key (selected) breeders. Hence, extending the model to several IBD-classes is highly valuable and might be

646 viewed as defining multiple reference populations instead of a single one. Inbreeding is then captured as distantly  
647 in the past as made possible by the available marker density and informativeness. There is thus no need to arbitrar-  
648 ily define any base population with unrelated ancestors nor to select an arbitrary threshold below which stretches  
649 of homozygous markers are considered non-IBD. The first benefit of a multiple IBD-classes model is to better fit  
650 the data and to obtain more accurate estimators of inbreeding both locally and globally. Indeed, our simulations  
651 under the inference model with several IBD classes clearly showed that the 1G (and 2G) model underestimated  
652  $F_G$  as some IBD segments were missed while the power to detect IBD segments was decreased. In addition, in  
653 the presence of ancient inbreeding, 1G model will tend to interpret recent (and thus longer) IBD segments as  
654 consecutive smaller segments of older origins because the estimated age of the single IBD class would tend to be  
655 older. Of course, in the absence of genotyping errors, the entire segment would then be correctly declared IBD  
656 and would appear as a long tract. However, at higher genotyping error rates (as with NGS data) such segments  
657 would be cut into smaller pieces. This would not happen when analyzing data with a model with multiple classes  
658 since recent IBD segments would then be associated to a class with a smaller age and the penalty in the HMM  
659 to leave the IBD-class and start a new IBD segment would be too large. Under a single-IBD class model, the  
660 age of the longest ROH further tends to be overestimated which might introduce substantial biases in applications  
661 that rely on the age of the IBD tracts to estimate some parameters of interest (e.g., the mutation rate). With two  
662 states HMM (Leutenegger et al.,2003), LD pruning is sometimes used to get rid of background LD and to force  
663 the model to concentrate on recent inbreeding (and hence avoiding the aforementioned problem). With multiple  
664 IBD-classes model ( $> 2G$  models), ancient inbreeding associated with background population LD is automatically  
665 assigned to the eldest IBD classes making LD pruning unnecessary for that purpose. Also, HMM with multiple  
666 IBD classes allows to determine whether there is a single or multiple IBD distribution(s) and to obtain information  
667 on the relatively recent demographic history of the population, providing  $N_e$  was reduced at some recent time in the  
668 past such as in populations under conservation or invasive populations. Such a modeling actually explores more  
669 recent generations and can be considered as complementary to approaches that infer past  $N_e$  thousands generations  
670 ago and many more as proposed by Li & Durbin (2011). Application to real populations demonstrated that the  
671 model can capture very different patterns including presence or absence of consanguineous matings, large  $N_e$  and  
672 low inbreeding, bottlenecks at varying time in the past, founder effects and reduced  $N_e$  due to isolation in the past  
673 ( $G_i \geq 100$ ). Finally, with multiple classes, we can clearly identify individuals from extreme consanguineous mat-  
674 ings (sire x daughter, first cousins, etc) because the recent inbreeding due to this recent ancestor is distinguished  
675 from the background inbreeding. Such examples with 25% inbreeding in class  $G_i \leq 4$  were observed in dogs or

676 sheep populations.

677 Our modeling approach actually allows to explore inbreeding in several dimensions: the global ( $F_G$ ), the local ( $\phi_l$ ) and age-variable ( $F_G^{(k)}$ ). It has been stated that more ancient inbreeding should not be considered since  
678 deleterious variants are expected to be rapidly purged from populations. Yet, the number of generations for this  
679 purging to complete depends on the population history. For instance, strong bottlenecks tend to reduce the efficiency of purging deleterious variants (“The cost of domestication”) and artificial selection might favor some  
680 breeders carrying deleterious variants. Thanks to our model we could estimate the inbreeding depression associated with different age-classes. This requires appropriate data sets (individuals genotyped at high marker density  
681 to capture old inbreeding and with own fitness records) and sufficient variation in all IBD-classes. Alternatively, recent and old inbreeding can be compared by functional annotations of different segments. For instance, Szpiech  
682 *et al* (2013) showed that long ROH are enriched for deleterious variants in humans. We can also use our model to  
683 test for local inbreeding depression and identify regions or variants where homozygosity seems more deleterious  
684 (e.g., Leutenegger *et al*, 2006).

685 Several strategies can be used to infer inbreeding in populations with our model. First, when using only one  
686 IBD class as in Leutenegger *et al* (2003), we can either estimate a single age common to both IBD and non-IBD  
687 classes or a different value for both states. The first option results in a model similar to Leutenegger *et al* (2003) and  
688 Vieira *et al* (2016) (note that the model by Narasimhan *et al* (2016) does not estimate the age but a single transition  
689 parameters combining  $G$  and the mixing proportions) and results in better estimates of age. Next, we can select  
690 the best number of IBD-classes according to the BIC criterion to compare the different models. When evaluated  
691 under simulated data, the BIC appeared to be conservative since the selected values were smaller or equal to the  
692 simulated ones. Note that with this approach we select the number of classes that best fit the data (merging several  
693 close classes if necessary) and not the real number of classes. Finally, we can use a set of IBD (and non-IBD)  
694 classes with predefined ages (the so-called MixKG models). It is then recommended to well separate these ages  
695 (e.g., using a ratio of 2 between successive ages to limit the overlap between the exponential distributions assumed  
696 for the IBD segment lengths) and cover a range of generations compatible with the available marker density. That  
697 strategy proved particularly efficient in most cases since it provided accurate estimates of the overall and local  
698 inbreeding while providing insights into the partitioning of inbreeding in the different age-classes and more easily  
699 comparable results across individuals from the same population. Such a model was only sub-optimal when a single  
700 and rare IBD class was simulated (which might not be usual in real populations) but required larger computational  
701 resources since more classes are simultaneously fitted.

706 Some precautions must be taken regarding interpretation of results. In our model, the mixing proportion and  
707 the rate of the exponential distribution are both estimated contrary to the model by Narasimhan *et al* (2016) where  
708 a single parameter is used. The mixing proportion can be interpreted as an expected inbreeding (the proportion  
709 of IBD segments among all segments) only if we have one IBD class and a single estimator for the age  $G$ . When  
710 segments from different distributions have different lengths, that interpretation is no longer correct (see the model  
711 section). The estimation of  $G$  might further be influenced by approximations in the model since we assume that the  
712 map is known without error, the recombination rate is not variable, there is no mutation and the population allele  
713 frequencies are known and did not vary over time. The estimation of this parameter is based on the distribution  
714 of lengths of IBD segments but this is a random process, for the same true  $G$  we can obtain segments of different  
715 lengths. For estimation from few IBD segments, the relative variation is higher. The presence of multiple-IBD  
716 classes generates also noise and the estimated distributions are often combinations of true underlying distributions.  
717 Therefore, the estimated inbreeding distribution must not be considered as exact but rather indicative. This is par-  
718 ticularly true for ancient inbreeding classes for which there is less information and approximations are accumulated  
719 over many generations. Ancient inbreeding captures ancient demographic history (past  $N_e$  and resulting LD) and  
720 presents less variation among individuals (ancient inbreeding is the results of many lineages and variance decreases  
721 for large samples). Note that some methods do not consider as inbreeding such shorts ROH reflecting homozy-  
722 gosity for ancient haplotypes and contributing to local LD patterns although Broman & Weber (1999) declared  
723 that homozygosity resulting of "linkage disequilibrium is indeed the result of the mating of (very distantly) related  
724 individuals". With our model, such inbreeding is automatically associated to ancient IBD-classes and separated  
725 from more recent inbreeding. Hence, users are free to interpret this as true inbreeding or background LD and to  
726 carry out LD pruning prior to analysis. The purpose of our model is to estimate individual genomic inbreeding and  
727 we advice to interpret only classes presenting enough variation. For older events and population parameters, we  
728 recommend to rely on other complementary models. Globally, estimation of different parameters is less accurate  
729 when less information is available (fewer IBD segments and less informative marker per segment). The model  
730 relies on two important hypotheses. First, it is assumed that most of the variants trace back further in time than  
731 the ancestors: the mutation did not happen in the path between the individual and its ancestor. With standard  
732 mutation and recombination rates (e.g., as in human or cattle), few mutations per IBD segment are expected on  
733 these paths (the value is relatively constant regardless of the age since older segments are smaller but have more  
734 time for mutations). So, as long as enough SNPs are present per segment, the impact of mutations should be low  
735 and accounted for by the genotyping error rate parameter. In addition, markers from genotyping arrays are old due

736 ascertainment bias favoring polymorphism in several populations. Still the model should be used cautiously when  
737 this condition is not met e.g., in populations distantly related from all of those represented in the discovery panel  
738 of the genotyping array (monomorphic SNPs in all the individuals of the considered population should then be  
739 discarded). The second hypothesis is that the marker allele frequencies in the base populations are known but we  
740 have only estimates. A special attention must be taken when working with several very different populations and  
741 markers have been selected based on their frequencies in one of these. When many markers are not segregating in  
742 one population (due to ascertainment bias) but frequencies are estimated across populations, these markers will be  
743 considered variable. Their fixation in the breed might then be considered as inbreeding. It is therefore important  
744 either to estimate the frequencies within population or use markers segregating in all the populations.

745 We are working on several extensions of our model, for instance to better take into account the possibility of  
746 mutations or to estimate the allele frequencies. Another possible extension to capitalize on individual inbreeding  
747 for past demographic inference of the whole population would be to explicitly relate the contribution of each  
748 IBD-class to each and every individual inbreeding to the corresponding past effective population size and further  
749 consider all the individuals jointly to estimate these (hyper-)parameters. Such a development might be viewed as  
750 an extension of our individual-oriented model to the population level.

## 751 **Acknowledgements**

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

## 758 **References**

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

- 762 Bosse M, Megens HJ, Madsen O, Paudel Y, Frantz LAF, *et al* (2012) Regions of homozygosity in the porcine  
763 genome: consequence of demography and the recombination landscape. *PLoS genetics*, **8**, e1003100.
- 764 Broman KW, Weber JL (1999) Long homozygous chromosomal segments in reference families from the centre  
765 d'Etude du polymorphisme humain. *Am J Hum Genet*, **65**, 1493–500.
- 766 Campbell CL, Bharer C, Morrow BE, Boyko AR, Auton A (2016) A pedigree-based map of recombination in the  
767 domestic dog genome. *G3 (Bethesda)*.
- 768 Charlesworth D, Willis JH (2009) The genetics of inbreeding depression. *Nature Reviews Genetics*, **10**, 783–796.
- 769 Charlier C, Coppieters W, Rollin F, Desmecht D, Agerholm JS, *et al* (2008) Highly effective SNP-based association  
770 mapping and management of recessive defects in livestock. *Nat Genet*, **40**, 449–54.
- 771 Clutton-Brock TH, Pemberton JM (2004) *Soay Sheep: Dynamics and Selection in an Island Population*. Cam-  
772 bridge University Press.
- 773 Darwin C (1876) *The effects of cross and self fertilisation in the vegetable kingdom*. John Murray, London.
- 774 Estoup A, Ravigne V, Hufbauer R, Vitalis R, Gautier M, Facon B (2016) Is there a genetic paradox of biological  
775 invasion? *Annual Review of Ecology, Evolution, and Systematics*, **47**, 51–72.
- 776 Ferencakovic M, Hamzic E, Gredler B, Solberg TR, Klemetsdal G, *et al* (2013) Estimates of autozygosity derived  
777 from runs of homozygosity: empirical evidence from selected cattle populations. *Journal of Animal Breeding  
778 and Genetics*, **130**, 286–293.
- 779 Hedrick PW, Garcia-Dorado A (2016) Understanding inbreeding depression, purging, and genetic rescue. *Trends  
780 in Ecology and Evolution*.
- 781 Hedrick PW, Kalinowski ST (2000) Inbreeding depression in conservation biology. *Annual Review of Ecology and  
782 Systematics*, **31**, 139–162.
- 783 Hill WG, Weir BS (2011) Variation in actual relationship as a consequence of Mendelian sampling and linkage.  
784 *Genet Res (Camb)*, **93**, 47–64.
- 785 Jakobsson M, Scholz SW, Scheet P, Gibbs JR, VanLiere JM, *et al* (2008) Genotype, haplotype and copy-number  
786 variation in worldwide human populations. *Nature*, **451**, 998–1003.

- 787 Johnston SE, Berenos C, Slate J, Pemberton JM (2016) Conserved genetic architecture underlying individual re-  
788 combination rate variation in a wild population of soay sheep (*ovis aries*). *Genetics*, **203**, 583–598.
- 789 Kardos M, Luikart G, Allendorf FW (2015) Measuring individual inbreeding in the age of genomics: marker-based  
790 measures are better than pedigrees. *Heredity (Edinb)*, **115**, 63–72.
- 791 Kardos M, Taylor HR, Ellegren H, Luikart G, Allendorf FW (2016) Genomics advances the study of inbreeding  
792 depression in the wild. *Evolutionary Applications*, **9**, 1205–1218.
- 793 Keller LF, Waller DM (2002) Inbreeding effects in wild populations. *Trends in Ecology and Evolution*, **17**, 230–  
794 241.
- 795 Keller MC, Visscher PM, Goddard ME (2011) Quantification of inbreeding due to distant ancestors and its detec-  
796 tion using dense single nucleotide polymorphism data. *Genetics*, **189**, 237–49.
- 797 Kijas JW, Lenstra JA, Hayes B, Boitard S, Porto Neto LR, *et al* (2012) Genome-wide analysis of the world’s sheep  
798 breeds reveals high levels of historic mixture and strong recent selection. *PLoS Biology*, **10**, e1001258.
- 799 Kirin M, McQuillan R, Franklin CS, Campbell H, McKeigue PM, Wilson JF (2010) Genomic runs of homozygosity  
800 record population history and consanguinity. *PloS One*, **5**, e13996.
- 801 Kong A, Thorleifsson G, Gudbjartsson DF, Masson G, Sigurdsson A, *et al* (2010) Fine-scale recombination rate  
802 differences between sexes, populations and individuals. *Nature*, **467**, 1099–1103.
- 803 Lander ES, Green P (1987) Construction of multilocus genetic linkage maps in humans. *Proc Natl Acad Sci U S*  
804 *A*, **84**, 2363–7.
- 805 Leroy G, Danchin-Burge C, Palhiere I, *et al* (2012) An abc estimate of pedigree error rate: application in dog,  
806 sheep and cattle breeds. *Animal Genetics*, **43**, 309–314.
- 807 Leutenegger AL, Labalme A, Genin E, *et al* (2006) Using genomic inbreeding coefficient estimates for homozy-  
808 gosity mapping of rare recessive traits: application to Taybi-Linder syndrome. *Am J Hum Genet*, **79**, 62–6.
- 809 Leutenegger AL, Prum B, Genin E, *et al* (2003) Estimation of the inbreeding coefficient through use of genomic  
810 data. *American Journal of Human Genetics*, **73**, 516–23.
- 811 Lewis TW, Abhayaratne BM, Blott SC (2015) Trends in genetic diversity for all kennel club registered pedigree  
812 dog breeds. *Canine Genetics and Epidemiology*, **2**, 13.

- 813 Li H, Durbin R (2011) Inference of human population history from individual whole-genome sequences. *Nature*,  
814 **475**, 493–496.
- 815 Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, *et al* (2009) The Sequence Alignment/Map format and SAM-  
816 tools. *Bioinformatics*, **25**, 2078–9.
- 817 Malécot G (1948) *Les Mathématiques de l'hérédité*. Masson et Cie.
- 818 Manichaikul A, Mychaleckyj JC, Rich SS, Daly K, Sale M, Chen WM (2010) Robust relationship inference in  
819 genome-wide association studies. *Bioinformatics*, **26**, 2867–2873.
- 820 McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, *et al* (2010) The Genome Analysis Toolkit: a  
821 MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res*, **20**, 1297–303.
- 822 McQuillan R, Leutenegger AL, Abdel-Rahman R, Franklin CS, Pericic M, *et al* (2008) Runs of homozygosity in  
823 european populations. *American Journal of Human Genetics*, **83**, 359–372.
- 824 Narasimhan V, Danecek P, Scally A, Xue Y, Tyler-Smith C, Durbin R (2016) Bcftools/roh: a hidden markov model  
825 approach for detecting autozygosity from next-generation sequencing data. *Bioinformatics*, **32**, 1749–1751.
- 826 O'Connell D, Scobie D, Hickey S, Sumner R, Pearson A (2012) Selection for yearling fleece weight and its effect  
827 on fleece shedding in new zealand wiltshire sheep. *Animal Production Science*, **52**, 456–462.
- 828 Palamara PF (2016) ARGON: fast, whole-genome simulation of the discrete time Wright-fisher process. *Bioinfor-*  
829 *matics*, **32**, 3032–4.
- 830 Pemberton TJ, Absher D, Feldman MW, Myers RM, Rosenberg NA, Li JZ (2012) Genomic patterns of homozy-  
831 gosity in worldwide human populations. *American Journal of Human Genetics*, **91**, 275–292.
- 832 Purcell S, Neale B, Todd-Brown K, *et al* (2007) PLINK: a tool set for whole-genome association and population-  
833 based linkage analyses. *Am J Hum Genet*, **81**, 559–75.
- 834 Purfield DC, Berry DP, McParland S, Bradley DG (2012) Runs of homozygosity and population history in cattle.  
835 *BMC Genetics*, **13**, 70.
- 836 Rabiner LR (1989) A tutorial on hidden markov models and selected applications in speech recognition. In *PRO-*  
837 *CEEDINGS OF THE IEEE*, pp. 257–286.

- 838 Rudan I, Smolej-Narancic N, Campbell H, *et al* (2003) Inbreeding and the genetic complexity of human hyperten-  
839 sion. *Genetics*, **163**, 1011–21.
- 840 Sempere G, Moazami-Goudarzi K, Eggen A, Laloe D, Gautier M, Flori L (2015) WIDDE: a Web-Interfaced next  
841 generation database for genetic diversity exploration, with a first application in cattle. *BMC Genomics*, **16**, 940.
- 842 Szpiech ZA, Xu J, Pemberton TJ, *et al* (2013) Long runs of homozygosity are enriched for deleterious variation.  
843 *American Journal of Human Genetics*, **93**, 90–102.
- 844 Szulkin M, Bierne N, David P (2010) Heterozygosity-fitness correlations: a time for reappraisal. *Evolution*, **64**,  
845 1202–17.
- 846 Thompson EA (2008) The IBD process along four chromosomes. *Theor Popul Biol*, **73**, 369–73.
- 847 VanRaden PM (2008) Efficient methods to compute genomic predictions. *J Dairy Sci*, **91**, 4414–23.
- 848 Vaysse A, Ratnakumar A, Derrien T, Axelsson E, Rosengren Pielberg G, *et al* (2011) Identification of genomic  
849 regions associated with phenotypic variation between dog breeds using selection mapping. *PLoS Genetics*, **7**,  
850 e1002316.
- 851 Vieira FG, Albrechtsen A, Nielsen R (2016) Estimating ibd tracts from low coverage ngs data. *Bioinformatics*, **32**,  
852 2096–2102.
- 853 Wang J (2016) Pedigrees or markers: Which are better in estimating relatedness and inbreeding coefficient? *Theor*  
854 *Popul Biol*, **107**, 4–13.
- 855 Wright S (1922) Coefficients of inbreeding and relationship. *American Naturalist*, **56**, 330–338.
- 856 Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, *et al* (2010) Common SNPs explain a large proportion  
857 of the heritability for human height. *Nat Genet*, **42**, 565–9.

## 858 **Data Accessibility**

859 All data sets used in the present study are publicly available. the Human Genome Diversity Panel (HGDP)  
860 data was downloaded from [ftp://ftp.cephb.fr/hgdp\\_supp10/Harvard\\_HGDP-CEPH](ftp://ftp.cephb.fr/hgdp_supp10/Harvard_HGDP-CEPH), the dog LUPA project  
861 from <http://dogs.genouest.org/SWEEP.dir/Supplemental.html> and the Sheep Diversity panel from the

<sup>862</sup> WIDDE data base. The program ZooRoH implementing our model can be freely obtained at <https://github.com/tdruet/ZooRoH>.

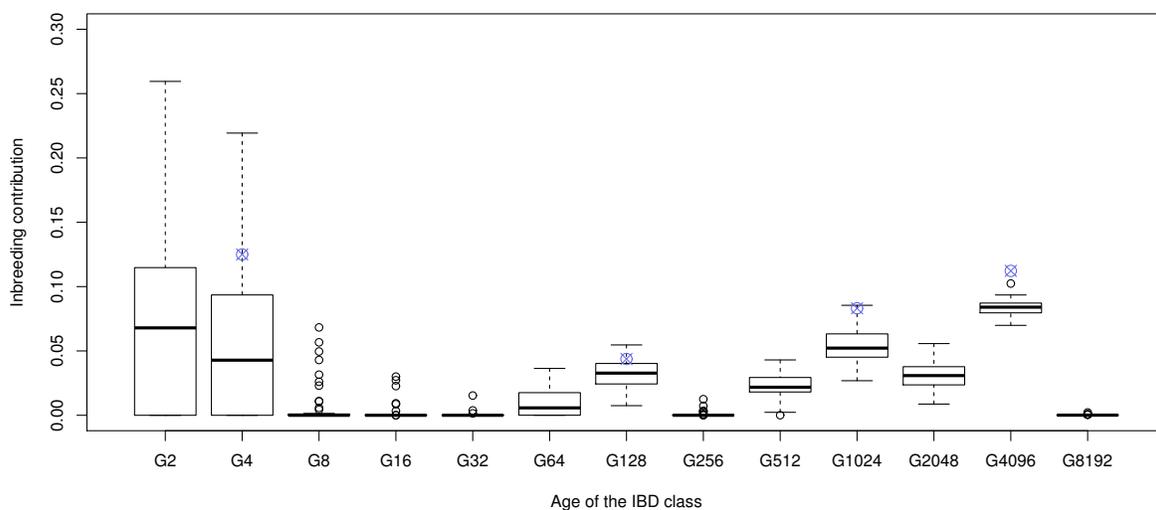
864 **List of Figures**

865 1 **Estimated inbreeding contributions  $F_G^{(k)}$  for 13 IBD classes with pre-defined ages (mix14G**  
866 **model) on data simulated under the 5G model (4 IBD classes).** The simulated genome consisted  
867 of 25 chromosomes of 100 cM with a marker density of 1000 SNPs per cM. Genotyping data  
868 for 50 individuals were simulated under the 5G inference model i.e., with 4 IBD-classes with  
869 the following realized ages (inbreeding contributions) as indicated by a star in the plot:  $G_1 = 4$   
870 ( $F_G^{(1)} = 0.125$ ),  $G_2 = 128$  ( $F_G^{(2)} = 0.08$ ),  $G_3 = 1024$  ( $F_G^{(3)} = 0.04$ ) and  $G_4 = 4$  ( $F_G^{(4)} = 0.11$ ). The  
871 data were analyzed with the mix14G that consisted of 13 IBD-classes with predefined ages ranging  
872 from 2 to 8192 (with  $G_k = 2^k$  for each class  $k$ ) and one non-IBD class that had the same age as the  
873 older IBD class (i.e.,  $G_K = G_{K-1} = 8192$ ). For each of these 13 IBD classes, the boxplots give the  
874 distribution of the estimated inbreeding contribution ( $\widehat{F}_G^{(k)}$ ) over the 50 simulated individuals. . . . 36

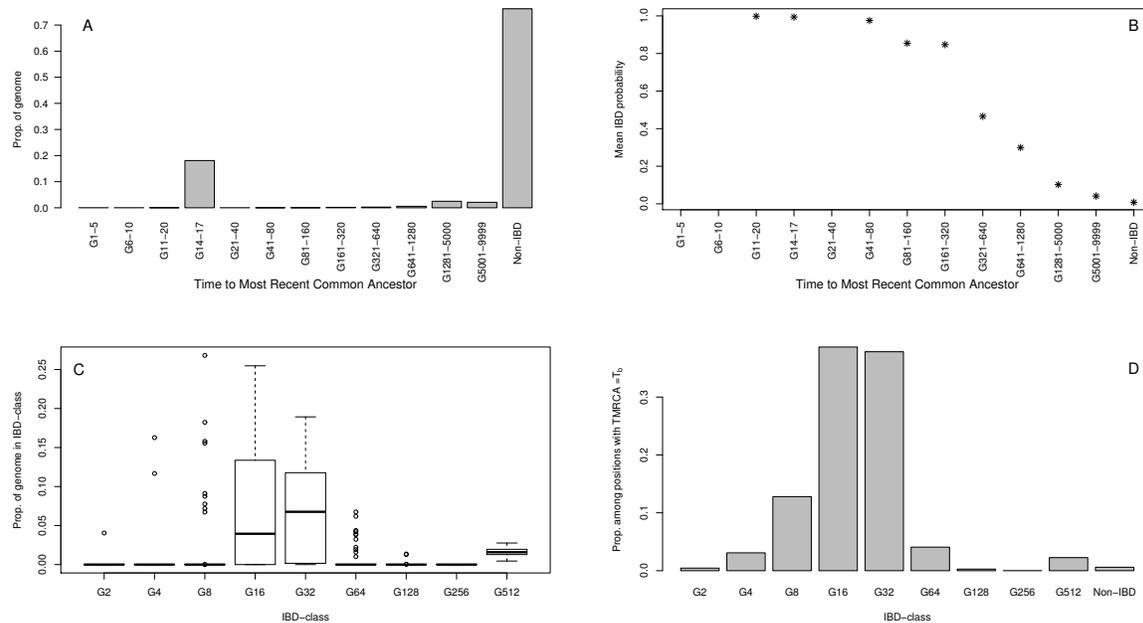
875 2 **Evaluation of the mix10G model on a data set consisting of 50 diploid individuals simulated**  
876 **under a Wright–Fisher demographic history with varying population sizes.** The population  
877 evolved under a WF1 scenario (see the Material and Methods section) with  $N_{e1} = 10^5$ ,  $T_s = 10^4$   
878 and a bottleneck lasting from generations 17 to 14 in the past and during which the population size  
879 was  $N_{eb} = 20$ . A) Realized distribution of the proportions of the simulated individual genomes  
880 lying within IBD segments as a function of their TMRCA (the interval G14-17 contains IBD  
881 segments tracing back to the bottleneck period, i.e., 14 to 17 generations backward in time) and  
882 within non-IBD segments (background). B) Estimated local inbreeding probabilities ( $\phi_l$ ) averaged  
883 over all the simulated individuals and markers as a function of the actual TMRCA of the underlying  
884 IBD segments. C) Distributions of the estimated proportion of the individual genomes assigned  
885 to each of the 9 predefined IBD classes (over the 50 simulated individuals). D) Proportion of the  
886 SNPs lying in IBD segments originating from the bottleneck period (i.e., 14 to 17 generations  
887 backward in time) that are assigned to the 9 different IBD classes of the mix10G model (summed  
888 over all the 50 individuals). . . . . 37

889 3 **Average estimated proportions of inbreeding contribution of a set of  $K$  predefined IBD classes**  
890 **for human (A,  $K = 13$ ), dog (C,  $K = 13$ ) and sheep (E,  $K = 8$ ) populations and corresponding**  
891 **average cumulative inbreeding (B, D and F for human, dog and sheep populations respec-**  
892 **tively).** . . . . . 38

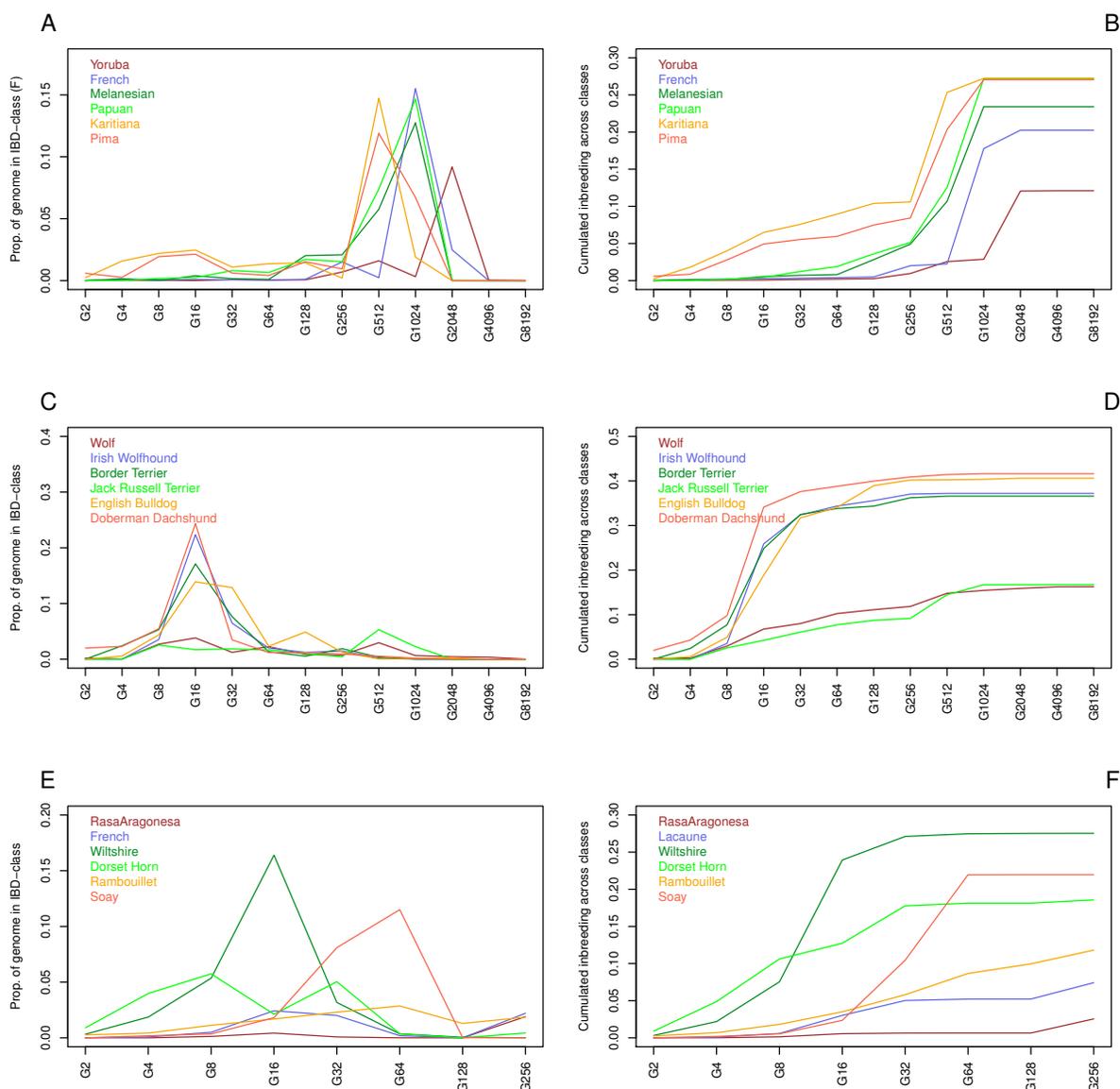
893 4 **Estimated partitioning of inbreeding in five humans (A), five dogs (B) and five sheeps (C).** . . . 39



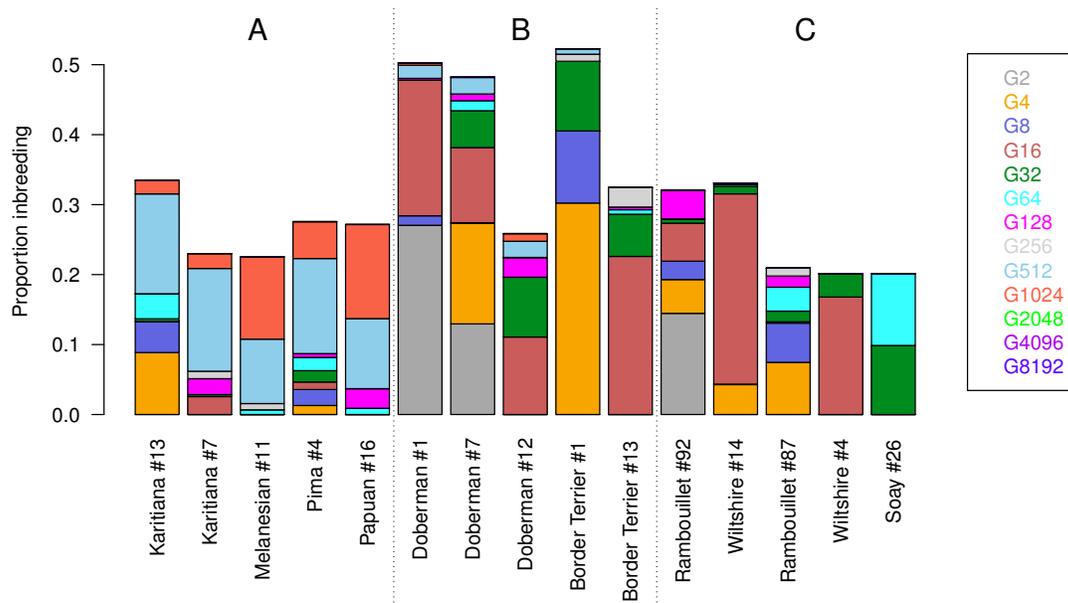
**Figure 1. Estimated inbreeding contributions  $F_G^{(k)}$  for 13 IBD classes with pre-defined ages (mix14G model) on data simulated under the 5G model (4 IBD classes).** The simulated genome consisted of 25 chromosomes of 100 cM with a marker density of 1000 SNPs per cM. Genotyping data for 50 individuals were simulated under the 5G inference model i.e., with 4 IBD-classes with the following realized ages (inbreeding contributions) as indicated by a star in the plot:  $G_1 = 4$  ( $F_G^{(1)} = 0.125$ ),  $G_2 = 128$  ( $F_G^{(2)} = 0.08$ ),  $G_3 = 1024$  ( $F_G^{(3)} = 0.04$ ) and  $G_4 = 4$  ( $F_G^{(4)} = 0.11$ ). The data were analyzed with the mix14G that consisted of 13 IBD-classes with predefined ages ranging from 2 to 8192 (with  $G_k = 2^k$  for each class  $k$ ) and one non-IBD class that had the same age as the older IBD class (i.e.,  $G_K = G_{K-1} = 8192$ ). For each of these 13 IBD classes, the boxplots give the distribution of the estimated inbreeding contribution ( $\widehat{F}_G^{(k)}$ ) over the 50 simulated individuals.



**Figure 2. Evaluation of the mix10G model on a data set consisting of 50 diploid individuals simulated under a Wright-Fisher demographic history with varying population sizes.** The population evolved under a WF1 scenario (see the Material and Methods section) with  $N_{e1} = 10^5$ ,  $T_s = 10^4$  and a bottleneck lasting from generations 17 to 14 in the past and during which the population size was  $N_{eb} = 20$ . A) Realized distribution of the proportions of the simulated individual genomes lying within IBD segments as a function of their TMRCA (the interval G14-17 contains IBD segments tracing back to the bottleneck period, i.e., 14 to 17 generations backward in time) and within non-IBD segments (background). B) Estimated local inbreeding probabilities ( $\phi_l$ ) averaged over all the simulated individuals and markers as a function of the actual TMRCA of the underlying IBD segments. C) Distributions of the estimated proportion of the individual genomes assigned to each of the 9 predefined IBD classes (over the 50 simulated individuals). D) Proportion of the SNPs lying in IBD segments originating from the bottleneck period (i.e., 14 to 17 generations backward in time) that are assigned to the 9 different IBD classes of the mix10G model (summed over all the 50 individuals).



**Figure 3. Average estimated proportions of inbreeding contribution of a set of  $K$  predefined IBD classes for human (A,  $K = 13$ ), dog (C,  $K = 13$ ) and sheep (E,  $K = 8$ ) populations and corresponding average cumulative inbreeding (B, D and F for human, dog and sheep populations respectively).**



**Figure 4. Estimated partitioning of inbreeding in five humans (A), five dogs (B) and five sheeps (C).**

## 894 List of Tables

895	1	<b>Performance of the 1G model on data simulated under the 1G inference model.</b> 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 1G inference model for each of 20 different scenarios defined by the simulated $G$ and $\rho$ values reported in the first two columns. The table reports the resulting median realized (true) values (across the 500 simulated individuals) for the age of inbreeding ( $G$ ), the mixing proportions ( $\rho$ ), the individual inbreeding ( $F_G$ ) and the number of IBD tracks ( $\#Tracks$ ). Similarly, the table gives the median estimated values and the Mean Absolute Errors (MAE) for the age of inbreeding ( $\widehat{G}$ ), the mixing proportions ( $\widehat{\rho}$ ) and the individual inbreeding ( $\widehat{F}_G$ ). Finally, the table gives the MAE for the estimated local inbreeding ( $\phi_l$ ) either for all the SNPs ( $\widehat{\phi}_l$ ) or for those actually lying within IBD segments ( $\widehat{\phi}_{IBD}$ ). . . . .	41
896	2	<b>Performance of the 3G model on data simulated under the 3G inference model (i.e., two IBD classes and one non-IBD class).</b> 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 3G inference model for each of 6 different scenarios defined by the simulated ages of inbreeding $G_1$ and $G_2$ (reported in the two first columns) and the corresponding mixing proportions $\rho_1$ and $\rho_2$ (reported in the third and fourth columns) of the two classes of IBD segments. The table reports the resulting median realized (true) values (across the 500 simulated individuals) for the ages of inbreeding ( $G_1$ and $G_2$ ), the amount of inbreeding originating from each IBD class ( $F_G^{(1)}$ and $F_G^{(2)}$ ) and the overall individual inbreeding ( $F_G$ ). The table further gives the median (and their associated MAE) of the estimated values ( $\widehat{G}_1$ , $\widehat{G}_2$ , $\widehat{F}_G^{(1)}$ , $\widehat{F}_G^{(2)}$ and $\widehat{F}_G$ ) obtained under the 3G model. The table also gives the MAE for the estimated local inbreeding ( $\phi_l$ ) either for all the SNPs ( $\widehat{\phi}_l$ ) or for those actually lying within IBD segments only ( $\widehat{\phi}_{IBD}$ ). . . . .	42
897	3	<b>Performance of the 1G model on simulated data sets with different SNP density and informativeness.</b> The simulated genome consisted of 25 chromosomes of 100 cM with a marker density of either 10 or 100 SNPs per cM. Allele frequency spectrum (AFS) of each SNP reference allele were either sampled from an empirical distribution (array-like) derived from a real (cattle) genotyping assay (i.e., close to uniform) or from a (U-shape) $\beta(0.2, 0.2)$ distribution that mimics NGS data. Genotyping data for 500 individuals were simulated under the 1G inference model for each of 3 different scenarios defined by the simulated $G$ and $\rho$ values reported in the first two columns. For each simulation, the table reports the resulting realized (true) median value (across the 500 simulated individuals) for the age of inbreeding ( $G$ ) and the individual inbreeding ( $F_G$ ) together with the median of their estimated values $\widehat{G}$ and $\widehat{F}_G$ and corresponding Mean Absolute Errors (MAE). Finally, the table gives the MAE for the estimated local inbreeding ( $\phi_l$ ) either for all the SNPs ( $\widehat{\phi}_l$ ) or for those actually lying within IBD segments only ( $\widehat{\phi}_{IBD}$ ). . . . .	43
898			
899			
900			
901			
902			
903			
904			
905			
906			
907			
908			
909			
910			
911			
912			
913			
914			
915			
916			
917			
918			
919			
920			
921			
922			
923			
924			
925			
926			
927			
928			

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

**Table 1. Performance of the 1G model on data simulated under the 1G inference model.** The simulated genome consisted of 25 chromosomes of 100 cM with a marker density of 10 SNPs per cM. Genotyping data for 500 individuals were simulated under the 1G inference model for each of 20 different scenarios defined by the simulated  $G$  and  $\rho$  values reported in the first two columns. The table reports the resulting median realized (true) values (across the 500 simulated individuals) for the age of inbreeding ( $G$ ), the mixing proportions ( $\rho$ ), the individual inbreeding ( $F_G$ ) and the number of IBD tracks ( $\#Tracts$ ). Similarly, the table gives the median estimated values and the Mean Absolute Errors (MAE) for the age of inbreeding ( $\widehat{G}$ ), the mixing proportions ( $\widehat{\rho}$ ) and the individual inbreeding ( $\widehat{F}_G$ ). Finally, the table gives the MAE for the estimated local inbreeding ( $\phi_l$ ) either for all the SNPs ( $\widehat{\phi}_l$ ) or for those actually lying within IBD segments ( $\widehat{\phi}_{IBD}$ ).

Scenario		Realized median values			Median estimated values (3G model)					
$G_1$ ( $\rho_1$ )	$G_2$ ( $\rho_2$ )	$G_1$ ( $F_G^{(1)}$ )	$G_2$ ( $F_G^{(2)}$ )	$F_G$	$\widehat{G}_1$ (MAE)	$\widehat{G}_2$ (MAE)	$\widehat{F}_G^{(1)}$ (MAE)	$\widehat{F}_G^{(2)}$ (MAE)	$\widehat{F}_G$ (MAE)	MAE for $\widehat{\phi}_l$ ( $\widehat{\phi}_{IBD}$ )
4 (0.125)	16 (0.100)	4.1 (0.12)	17 (0.09)	0.210	7.20 (3.06)	391 (288)	0.195 (0.075)	0.004 (0.074)	0.210 (0.002)	0.012 (0.025)
4 (0.125)	64 (0.100)	4.1 (0.12)	64 (0.09)	0.211	3.60 (1.01)	64.6 (9.53)	0.123 (0.007)	0.086 (0.007)	0.211 (0.002)	0.038 (0.089)
4 (0.125)	256 (0.100)	4.0 (0.12)	257 (0.09)	0.211	3.60 (0.65)	275 (35.9)	0.120 (0.001)	0.087 (0.004)	0.208 (0.004)	0.101 (0.238)
8 (0.100)	128 (0.100)	8.2 (0.10)	128 (0.09)	0.189	7.20 (1.48)	126 (14.8)	0.098 (0.004)	0.090 (0.005)	0.189 (0.003)	0.069 (0.182)
32 (0.100)	64 (0.100)	32 (0.10)	67 (0.09)	0.190	33.9 (7.08)	102 (140)	0.157 (0.058)	0.030 (0.057)	0.192 (0.003)	0.051 (0.132)
32 (0.100)	256 (0.100)	32 (0.10)	260 (0.09)	0.188	29.6 (4.31)	265 (38.0)	0.097 (0.007)	0.089 (0.007)	0.188 (0.004)	0.114 (0.302)

**Table 2. Performance of the 3G model on data simulated under the 3G inference model (i.e., two IBD classes and one non-IBD class).** The simulated genome consisted of 25 chromosomes of 100 cM with a marker density of 10 SNPs per cM. Genotyping data for 500 individuals were simulated under the 3G inference model for each of 6 different scenarios defined by the simulated ages of inbreeding  $G_1$  and  $G_2$  (reported in the two first columns) and the corresponding mixing proportions  $\rho_1$  and  $\rho_2$  (reported in the third and fourth columns) of the two classes of IBD segments. The table reports the resulting median realized (true) values (across the 500 simulated individuals) for the ages of inbreeding ( $G_1$  and  $G_2$ ), the amount of inbreeding originating from each IBD class ( $F_G^{(1)}$  and  $F_G^{(2)}$ ) and the overall individual inbreeding ( $F_G$ ). The table further gives the median (and their associated MAE) of the estimated values ( $\widehat{G}_1$ ,  $\widehat{G}_2$ ,  $\widehat{F}_G^{(1)}$ ,  $\widehat{F}_G^{(2)}$  and  $\widehat{F}_G$ ) obtained under the 3G model. The table also gives the MAE for the estimated local inbreeding ( $\phi_l$ ) either for all the SNPs ( $\widehat{\phi}_l$ ) or for those actually lying within IBD segments only ( $\widehat{\phi}_{IBD}$ ).

$G$	$\rho$	Simulation		Realized median value		Estimated median value		
		SNP per cM	AFS	$G$	$F_G$	$\widehat{G}$ (MAE)	$\widehat{F}_G$ (MAE)	MAE for $\widehat{\phi}_l$ ( $\widehat{\phi}_{IBD}$ )
4	0.125	10	Array-like	3.90	0.125	4.00 (0.57)	0.124 (0.001)	0.0026 (0.0101)
4	0.125	100	Array-like	4.00	0.123	4.00 (0.51)	0.123 (0.000)	0.0002 (0.0009)
4	0.125	10	$\beta(0.2, 0.2)$	4.10	0.119	4.00 (0.64)	0.120 (0.002)	0.0068 (0.0272)
4	0.125	100	$\beta(0.2, 0.2)$	4.10	0.120	4.00 (0.55)	0.120 (0.000)	0.0006 (0.0023)
64	0.100	10	Array-like	64.2	0.099	64.3 (4.06)	0.099 (0.002)	0.0410 (0.2056)
64	0.100	100	Array-like	64.6	0.099	64.4 (2.00)	0.099 (0.000)	0.0035 (0.0181)
64	0.100	10	$\beta(0.2, 0.2)$	64.2	0.100	64.1 (6.26)	0.100 (0.006)	0.0807 (0.4032)
64	0.100	100	$\beta(0.2, 0.2)$	64.1	0.099	64.2 (2.50)	0.099 (0.000)	0.0095 (0.0482)
256	0.100	10	Array-like	256	0.100	257 (16.7)	0.099 (0.005)	0.1134 (0.5689)
256	0.100	100	Array-like	255	0.100	256 (5.79)	0.100 (0.000)	0.0164 (0.0824)
256	0.100	10	$\beta(0.2, 0.2)$	257	0.100	252 (36.9)	0.100 (0.008)	0.1462 (0.7313)
256	0.100	100	$\beta(0.2, 0.2)$	256	0.100	255 (8.06)	0.100 (0.001)	0.0398 (0.1994)

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