Partitioning of autozygosity in different age-based classes in cattle populations with different demographic histories

T. Druet[1*], L. Flori[2] and M. Gautier[3]

[1]Unit of Animal Genomics, GIGA-R, University of Liège, 4000 Liège, Belgium; [2]SELMET, INRAE, CIRAD, L'Institut Agro, Université de Montpellier, 34988 Montpellier, France; [3]CBGP, INRAE, CIRAD, IRD, L'Institut Agro, Université de Montpellier, 34988 Montpellier, France; [*]tom.druet@uliege.be

Abstract

Classification of homozygous-by-descent (HBD) segments or ROH in different age-related groups based on their length is useful for several applications. Here we compare the partitioning obtained with our newly developed multiple-HBD classes hidden Markov model (HMM) with an alternative approach consisting of length-based clustering of ROH. When the HMM was applied to simulated data sets, autozygosity was concentrated in the HBD-class associated with the ancestors causing inbreeding. This was however not observed for the ROH-classes corresponding to the expected length of ROH transmitted from these ancestors. On real cattle data, autozygosity was maximized with the HMM in the HBD-class matching the inferred period of reduced N_e. In summary, while ROH-based classifications must be interpreted cautiously, our newly developed multiple HBD-classes HMM seems promising to provide a better picture of the age-based partitioning of individual genomic inbreeding.

Introduction

Inbreeding is common in cattle populations as a consequence of intense selection of elite sires or due to their recent demographic history often characterized by declining effective population sizes or recent bottlenecks. Homozygous-by-descent (HBD) segments, genomic segments inherited twice and through different paths from a common ancestor, are present at high frequency in such inbred populations. HBD segments result in long stretches of homozygous genotypes referred to as run-of-homozygosity (ROH), the length of which being a function of the number of generations to the common ancestor. The distribution of HBD segment lengths is thus informative about the past demographic history of a population and is often summarized, in empirical studies, by grouping observed ROH in several discrete length classes; long/short ROH-classes being considered as resulting from recent/ancient inbreeding. Such classifications are valuable to understand past demographic events from the population, to monitor evolution of inbreeding levels, to infer the mating structure when pedigrees are unknown, to determine how deleterious variants are purged across generations, to compare the effects of recent and ancient inbreeding on phenotypes or even to estimate mutation rates. As an alternative approach, we recently developed an HMM with multiple-HBD classes to partition autozygosity in different age-related HBD classes (Druet and Gautier, 2017). We herein assess the accuracy and robustness of such partitioning of autozygosity based on simulated data sets and genotyping data available for three cattle populations with contrasted recent demographic histories, and compare these classifications with those obtained with ROH-based approaches.

Materials & Methods

Classification approaches. The joint identification and classification of HBD segments was first realized with RZooRoH (Bertrand et al., 2019) and a model with 12 HBD classes. In this multiple HBD-classes model, several nested layers of ancestors are successively modelled (Druet and Gautier, 2021). In each layer, the genome is described as a mosaic of HBD and non-HBD segments with an HMM. The length of HBD and non-HBD segments is exponentially distributed with a rate R_c specific to the layer c and related to the number of generations to the common ancestors (approximately $0.5*R_c$ generations). The frequency of HBD segments within layer c is defined by a parameter ρ_c, that can be interpreted as the inbreeding coefficient accumulated over the generations included in the layer. The non-HBD classes are modelled as a mixture of HBD classes from earlier generations and shorter non-HBD segments. Genotype probabilities in different states (HBD / non-HBD) are obtained from the allele frequencies and the genotyping error rates.
A ROH-based approach was additionally used. The clustering of ROH was carried out with GARLIC (Szpiech et al., 2017) that automatically selects the optimal number of SNPs per window. GARLIC was also ran with 12 classes by selecting ROH boundaries matching HBD-classes. More precisely, the boundary between ROH class c-1 and c corresponded to the ROH-length for which the exponential distributions with rates R_{c-1} and R_c have equal probability.

Simulated data sets. To evaluate the classification approaches, we simulated 500 genomes as mosaics of HBD and non-HBD segments using different values of ρ, corresponding to the inbreeding coefficient *F*. Their length was exponentially distributed with a rate R_c. The genome consisted of 25 chromosomes 100 cM long, and individuals were genotyped for a total of 25,000 SNPs. We subsequently used Argon (Palamara, 2016) to simulate more realistic data under a Wright-Fisher model. In these simulations, bottlenecks occurred 16 or 64 generations ago and effective population size (N_e) dropped to 20 or 50.

Real cattle data sets. Identification and classification of HBD segments or ROH was applied to three cattle data sets with distinct recent demographic history (i.e., decline, severe bottleneck and expansion, respectively): i) 145 Dutch Holstein (HOL) individuals (Alemu et al., 2021); ii) 18 individuals from the feral cattle population of Amsterdam Island (TAF), and iii) 22 Zebus from Magadascar (ZMA) (Magnier et al., 2021). After quality filtering, the samples were respectively genotyped for 37,675, 23,679 and 531,967 SNPs. Past N_e was first estimated with GONE (Santiago et al., 2020).

Results

The expected length for a HBD segment associated with a common ancestor present G generations ago is L=100/(2G) cM, G is thus often estimated as 100/(2L) for a ROH of length L. However, we observed that classes with L > 100/(2L) had larger contributions to autozygosity, more so when simulated inbreeding levels were high (Figure 1). Conversely, with the HMM, the HBD-class with a rate R_c equal to 2G had the main contribution to autozygosity. Similar patterns were observed for simulations with different values of R_c.
The behaviour of the two approaches was confirmed when populations that experienced a bottleneck were simulated under a Wright-Fisher model. With the HMM based approach, the HBD-class with a rate R_c equal to 2G (where G is the timing in generations of the bottleneck), and its neighbours, captured the HBD-segments associated to the bottleneck. On the other hand, with the ROH-based approach, ROH longer than 100/(2G) had the highest contributions to autozygosity, more so when the bottleneck was stronger and inbreeding levels were higher.

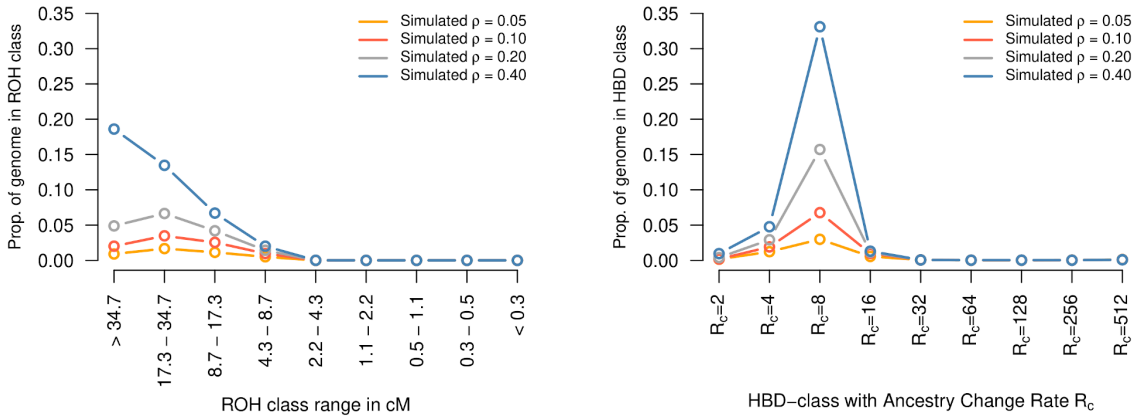

Figure 1. Classification of simulated HBD segments inherited from ancestors living 4 generations in the past (i.e., with an expected length L = 12.5 cM) with an ROH approach and with the multiple HBD-classes HMM (expected R_c = 8).

This trend was also confirmed when analysing real data on three cattle populations with very different recent histories (HOL, TAF and ZMA) inferred with GONE and which were found consistent with our prior expectations (Figure 2): i) HOL presented a recent decline with current N_e being equal to 70; ii) TAF experienced a recent and severe bottleneck (17 to 25 generations ago, N_e dropped to 5) in agreement with the historically reported introduction in the Amsterdam island of six founder individuals in 1861 followed by a rapid expansion of the population (up to Ne=1,300); and iii) ZMA is in expansion (100 generations ago, N_e = 100 up to a current Ne=3,000). The partitioning with the multiple HBD-classes model closely matches these demographic events (Figure 2). For instance, autozygosity is mainly associated with ancestors living around 16 generations in the past for TAF population, and around 100 generations in ZMA, just before the expansion. In HOL, inbreeding levels were associated to more recent ancestors. The inbreeding levels estimated with GARLIC were highly correlated with those from ZooRoH (0.963, 0.938 and 0.996 for the three populations) although values were lower, indicating that the smaller HBD segments were not captured. By decreasing the minimal ROH size to 20 SNPs with GARLIC, inbreeding levels were closer to those obtained with ZooRoH (although this is a rather small number of SNPs and the optimal values estimated by GARLIC were higher). As for the analyses on simulated data sets, the ROH-based approach associated most of autozygosity to classes with L > 100/R_c (R_c is the rate of HBD-classes with the main contribution to autozygosity). Interestingly, for lower inbreeding levels and more recent ancestors, classifications from both approaches were more concordant.

Discussion

In ROH-based approaches, the number of generations to the common ancestor of a given ROH is often estimated as G=100/(2L), based on the expected length of a ROH inherited from an ancestor living G generations ago. ROH are often classified according to their length and G is then estimated independently for each class. In simulations with ancestors present G generations ago, we observed that classes with L > 100/2G have larger contributions than classes with smaller ROH, leading to the incorrect interpretation that inbreeding might be associated with more recent ancestors. This result from the fact that ROH inherited from ancestors present G generations ago follow an exponential distribution and their length range spans generally over several classes, with only a subset of these ROH falling into the class

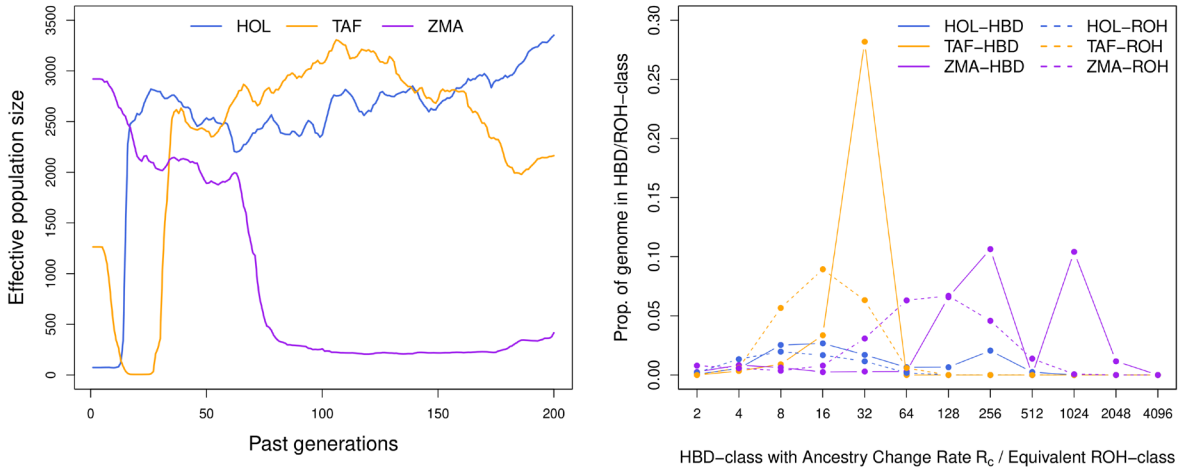

Figure 2. Past effective population size and partitioning in different HBD or ROH-classes for cattle populations with distinct demographic histories.

centered around L. Such an approach thus fails to account for the possibility that ROH from different classes might come from the same group of founders. A better estimate may be derived by combining information from multiple ROH (from their average length), providing a single ancestor (or multiple ancestors living in the same generation) contributed to the individual inbreeding. The classification varied also as a function of the inbreeding level ρ, possibly because for higher ρ, the probability that two consecutive short and independent ROH appear as one unique long ROH increases, shifting the distribution towards classes with longer segments. Overall, ROH-based classifications should be interpreted very cautiously. By contrast, the HMM allows joint estimation of the parameters of a mixture of exponential distributions that would result in the observed distribution of HBD segment lengths. We are currently evaluating how precisely the age (in terms of generation to the common ancestor) of each HBD segments / ROH can be estimated with different approaches in more realistic demographic histories. For instance, if we want to understand the relationship between HBD class contributions and Ne, we should determine which past generations are captured by each HBD classes. We must also account for the fact that contributions from the most ancient classes might be swept by autozygosity associated with more recent ancestors. However, we showed that the parameters ρ did not require such corrections (Druet and Gautier, 2021).

References

Alemu S.W., Kadri N.K., Harland C., Faux P., Charlier C. *et al.* (2021) Heredity 126:410–423. https://doi.org/10.1038/s41437-020-00383-9

Bertrand A.R., Kadri N.K., Flori L. Gautier M., and Druet T. (2019) Methods Ecol Evol 10(6):860–866. https://doi.org/10.1111/2041-210X.13167

Druet T., and Gautier M. (2017) Mol Ecol 26:5820–5841. https://doi.org/10.1111/mec.14324

Druet T., and Gautier M. (2021) bioRxiv https://doi.org/10.1101/2021.05.25.445246

Palamara P. (2016) Bioinformatics 32:3032-34.https://doi.org/10.1093/bioinformatics/btw355

Magnier J., Druet T., Naves J., Ouvrard M., Raoul S., et al. (2021) bioRxiv https://doi.org/10.1101/2021.10.08.463737

Santiago E., Novo I., Pardiñas A.F., Saura M., Wang J., *et al.* (2020) Mol Biol Evol. 37(12):3642-3653. https://doi.org/10.1093/molbev/msaa169

Szpiech Z.A., Blant A., and Pemberton T.J. (2017) Bioinformatics 33(13):2059–2062. https://doi.org/10.1093/bioinformatics/btx102