

## RESOURCE ARTICLE

# Effect of reduced genomic representation on using runs of homozygosity for inbreeding characterization

Eléonore Lavanchy<sup>1,2</sup>  | Jérôme Goudet<sup>1,2</sup> 

<sup>1</sup>Department of Ecology and Evolution, University of Lausanne, Lausanne, Switzerland

<sup>2</sup>Swiss Institute of Bioinformatics, University of Lausanne, Lausanne, Switzerland

**Correspondence**

Eléonore Lavanchy and Jérôme Goudet, Department of Ecology and Evolution, Biophore, University of Lausanne, CH-1050 Lausanne, Switzerland.  
Emails: [eleonore.lavanchy@unil.ch](mailto:eleonore.lavanchy@unil.ch); [jerome.goudet@unil.ch](mailto:jerome.goudet@unil.ch)

**Funding information**

Schweizerischer Nationalfonds zur Förderung der Wissenschaftlichen Forschung, Grant/Award Number: 31003A\_179358

**Handling Editor:** Kimberly J. Gilbert

**Abstract**

Genomic measures of inbreeding based on identical-by-descent (IBD) segments are increasingly used to measure inbreeding and mostly estimated on SNP arrays and whole-genome sequencing (WGS) data. However, some softwares recurrently used for their estimation assume that genomic positions which have not been genotyped are nonvariant. This might be true for WGS data, but not for reduced genomic representations and can lead to spurious IBD segments estimation. In this project, we simulated the outputs of WGS, two SNP arrays of different sizes and RAD-sequencing for three populations with different sizes and histories. We compare the results of IBD segments estimation with two softwares: runs of homozygosity (ROHs) estimated with PLINK and homozygous-by-descent (HBD) segments estimated with RZooRoH. We demonstrate that to obtain meaningful estimates of inbreeding, RZooRoH requires a SNPs density 11 times smaller compared to PLINK: ranks of inbreeding coefficients were conserved among individuals above 22 SNPs/Mb for PLINK and 2 SNPs/Mb for RZooRoH. We also show that in populations with simple demographic histories, distribution of ROHs and HBD segments are correctly estimated with both SNP arrays and WGS. PLINK correctly estimated distribution of ROHs with SNP densities above 22 SNPs/Mb, while RZooRoH correctly estimated distribution of HBD segments with SNPs densities above 11 SNPs/Mb. However, in a population with a more complex demographic history, RZooRoH resulted in better distribution of IBD segments estimation compared to PLINK even with WGS data. Consequently, we advise researchers to use either methods relying on excess homozygosity averaged across SNPs or model-based HBD segments calling methods for inbreeding estimations.

**KEYWORDS**

homozygous-by-descent, identical-by-descent, inbreeding, reduced genomic representations, runs of homozygosity

## 1 | INTRODUCTION

Inbreeding is defined as mating between relatives and has been observed across many taxa including humans (Bittles &

Black, 2010; Ceballos et al., 2018), livestock (Forutan, Ansari Mahyari, et al., 2018; Kim et al., 2013; Peripolli et al., 2017, 2018), wild animal populations (Åkesson et al., 2016; Huisman et al., 2016; Kardos et al., 2018; Keller & Waller, 2002) and

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial](https://creativecommons.org/licenses/by-nc/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.

© 2023 The Authors. *Molecular Ecology Resources* published by John Wiley & Sons Ltd.

plants (Kariyat & Stephenson, 2019; Keller & Waller, 2002; Zhang et al., 2019). Its quantification and the understanding of its deleterious consequences – called inbreeding depression – are central in many areas of biology, from human genetics to conservation biology (Keller & Waller, 2002). Indeed, increase in genome autozygosity has been associated with diseases, such as schizophrenia (Keller et al., 2012; Lencz et al., 2007) and Alzheimer's disease (Ghani et al., 2015; Nalls et al., 2009) as well as fitness costs in animals (Åkesson et al., 2016; Huisman et al., 2016) and plants (Menges, 1991; Zhang et al., 2019).

Individual levels of inbreeding are quantified with inbreeding coefficients ( $F$ ). Traditionally, inbreeding was measured by counting the size and number of loops in pedigrees ( $F_{PED}$ ) (Wright, 1922) a method with several downsides: (i) it estimates the expected individual coefficient which can differ from the realized individual coefficient due to recombination stochasticity and Mendelian segregation (Carothers et al., 2006; Franklin, 1977; Hill & Weir, 2011); (ii) it assumes founders of the pedigree are unrelated and noninbred; (iii) pedigrees must be correctly recorded which is extremely difficult in wild populations, although genetic data might be used to (re)construct links (Huisman, 2017; Jones & Wang, 2010). With the advancements in high throughput sequencing technologies it became possible to estimate with sufficient accuracy genomic-based  $F$ , and several studies have shown molecular estimates to be more accurate than pedigree-based estimates (Alemu et al., 2020; Kardos et al., 2015; Keller et al., 2011; Wang, 2016). Many different genomic-based  $F$  have been proposed, such as  $F_{HOM}$  (Chang et al., 2015),  $F_{AS}$  (Weir & Goudet, 2017),  $F_{UNI}$  and  $F_{GRM}$  (both described in Yang et al., 2011) but there is still no consensus on which is the most accurate (Alemu et al., 2020; Caballero et al., 2020; Goudet et al., 2018; Nietlisbach et al., 2019; Yengo et al., 2017). These estimates quantify average excess single nucleotide polymorphism (SNP) homozygosity or correlation between uniting gametes and treat all SNPs independently. However, parents transmit DNA to their offspring in large chromosomal segments rather than each base independently. Consequently, it has been suggested that measures of inbreeding should be based on identical-by-descent (IBD) segments rather than individual SNPs (McQuillan et al., 2008). Hence, a new  $F$  was proposed by McQuillan et al., (2008). This coefficient intends to quantify the proportion of IBD segments in the genome. From this point onward, we will call the true fraction of genome within IBD segments:  $F_{IBD}$  and its estimations (i)  $F_{ROH}$  when estimated from runs of homozygosity (ROHs) with observational-based approaches and (ii)  $F_{HBD}$  when estimated from homozygous-by-descent (HBD) segments from model-based approaches.

McQuillan et al. (2008) proposed to use ROHs, long consecutive homozygous segments, as a proxy for these IBD segments. ROHs were first described by (Broman & Weber, 1999) and shown to be ubiquitous in humans (Ceballos et al., 2018; Gibson et al., 2006; Pemberton et al., 2012) and across many different taxa (Kardos et al., 2018; Liu et al., 2020; Saremi et al., 2019).  $F_{ROH}$  is calculated as the proportion of the genome within ROHs and

several studies demonstrated that it was a reliable estimator of inbreeding (Alemu et al., 2020; Caballero et al., 2020; Nietlisbach et al., 2019). In addition to quantifying inbreeding, distribution of IBD segments (i.e., lengths and numbers) can inform about a population's past demography and history (Ceballos et al., 2018; Kirin et al., 2010; Pemberton et al., 2012): long segments reflect recent coalescence events while smaller segments indicate more distant coalescence and, if in high proportion, a history of small effective population size. Finally, IBD segments can be used for identifying rare deleterious recessive variants responsible for deleterious phenotypes by homozygosity mapping, which in short compares islands of IBD segments between affected and unaffected individuals (Alkuraya, 2013; Hildebrandt et al., 2009; Stoffel et al., 2021; Wang et al., 2009).

Two different methods for IBD segments detection are recurrently used in the literature: observation and model-based approaches (Ceballos et al., 2018). The most common method is a fast observation-based method (Ceballos et al., 2018) implemented in PLINK (Chang et al., 2015; Purcell et al., 2007). It makes use of a sliding window to identify continuous homozygous stretches, with a minimum size defined by the user, used as proxy for IBD segments. The other family of methods is model-based, and has been implemented in RZooRoH (Bertrand et al., 2019; Druet & Gautier, 2017, 2022), BEAGLE (Browning & Browning, 2010) and BCFTools (Narasimhan et al., 2016). It relies on hidden Markov models (HMM) and directly infers HBD segments from the genotypes by considering the distance between two markers, the mutation rate and even the recombination map if available. Consequently, these methods do not require a minimum threshold on segment length. HMM methods are computationally demanding (Ceballos et al., 2018) and a previous study suggested that PLINK outperformed HMM methods both in terms computation time and ROHs detection accuracy with simulated whole-genome-sequencing (WGS) data (Howrigan et al., 2011). However, few HMM methods were available at that time and the authors did not investigate the robustness of PLINK to genotyping errors. Observation-based approaches were designed for WGS data and assume that the region between two SNPs are entirely homozygous. However, many studies performing ROHs analyses with PLINK used reduced genomic representation techniques: often SNP arrays (Bjelland et al., 2013; Bosse et al., 2012; de Jong et al., 2020; Forutan, Mahyari, et al., 2018) where specific SNPs, chosen based on their position, effect on phenotype or minor allele frequency (MAF), are targeted and genotyped. ROHs have also been called with restriction-site associated DNA sequencing (RAD-sequencing) data, by cutting the genome near enzymes cutting sites and selecting and sequencing fragments based on their size. With both SNP arrays and RAD sequencing, only a fraction of the genome is sequenced resulting in a partial representation of the total polymorphism. Since PLINK assumes that genomic positions not included in the SNPs set are nonvariant, we expect that it will falsely consider nonsequenced heterozygous loci as homozygous which can lead to spurious ROHs detection. On the contrary,

the HMM approach from Leutenegger et al. (2003), which models the genome as a mosaic of HBD and non-HBD segments and from which most current model-based approaches follow, was initially developed for SNP arrays. These models do not treat nonsequenced genomic regions as homozygous but as missing data. However, model-based approaches are rarely used for detecting IBD segments with reduced genomic data (but see Alemu et al., 2020; Dutsch et al., 2021; Sole et al., 2017). In addition, no precise benchmarking with large sample size has been performed on comparing how the different methods behave with these reduced genomic data compared to WGS data and precise guidelines such as which method is suitable with which data are missing.

In addition to the fraction of genome captured, we hypothesize that the effective size and level of polymorphism in a population might also affect the capacity of the different methods to accurately detect IBD segments. Small and inbred populations tend to harbour higher numbers of long such segments easier to accurately detect with reduced representations as the missing positions are more likely to be homozygous. Larger populations will tend to harbour many small IBD segments (Ceballos et al., 2018; Kirin et al., 2010) harder to detect when only a fraction of the polymorphism is available since these small segments require lots of nearby SNPs to reach the minimum density threshold for an accurate detection (Kardos et al., 2015; Sole et al., 2017; Zhang et al., 2015). On the other hand, larger populations will harbour higher levels of polymorphism and thus higher numbers of SNPs resulting in an increased SNPs density for the same fraction of genome sequenced with RAD-sequencing.

Here, we use simulated data to compare the performance of PLINK and RZooRoH with two reduced genomic representations – SNP arrays and RAD-sequencing – and WGS. We compare both softwares output to the true IBD segments extracted from the simulated data. We hypothesize that the quality of detection depends on SNP density. In addition, since model-based approaches take into account the distance between each SNP, we predict that they will perform better when dealing with sparse data (Druet & Gautier, 2017). We show that both detection methods can be used to correctly estimate IBD segments with SNP arrays and RAD-sequencing providing that a sufficient proportion of the genome has been sequenced. This proportion varied between IBD segments detection methods and population sizes: the model-based method implemented in RZooRoH as well as the large population require a substantially smaller fraction of the genome to obtain correct inbreeding and distribution of IBD segments estimates.

## 2 | MATERIALS AND METHODS

All scripts used in this project are available on GitHub: <https://github.com/EleonoreLavanchy/ROHsReducedRep>. A general workflow of the study can be found in Figure 1 and additional details about the simulations and analyses performed can be found in the Supporting Information.

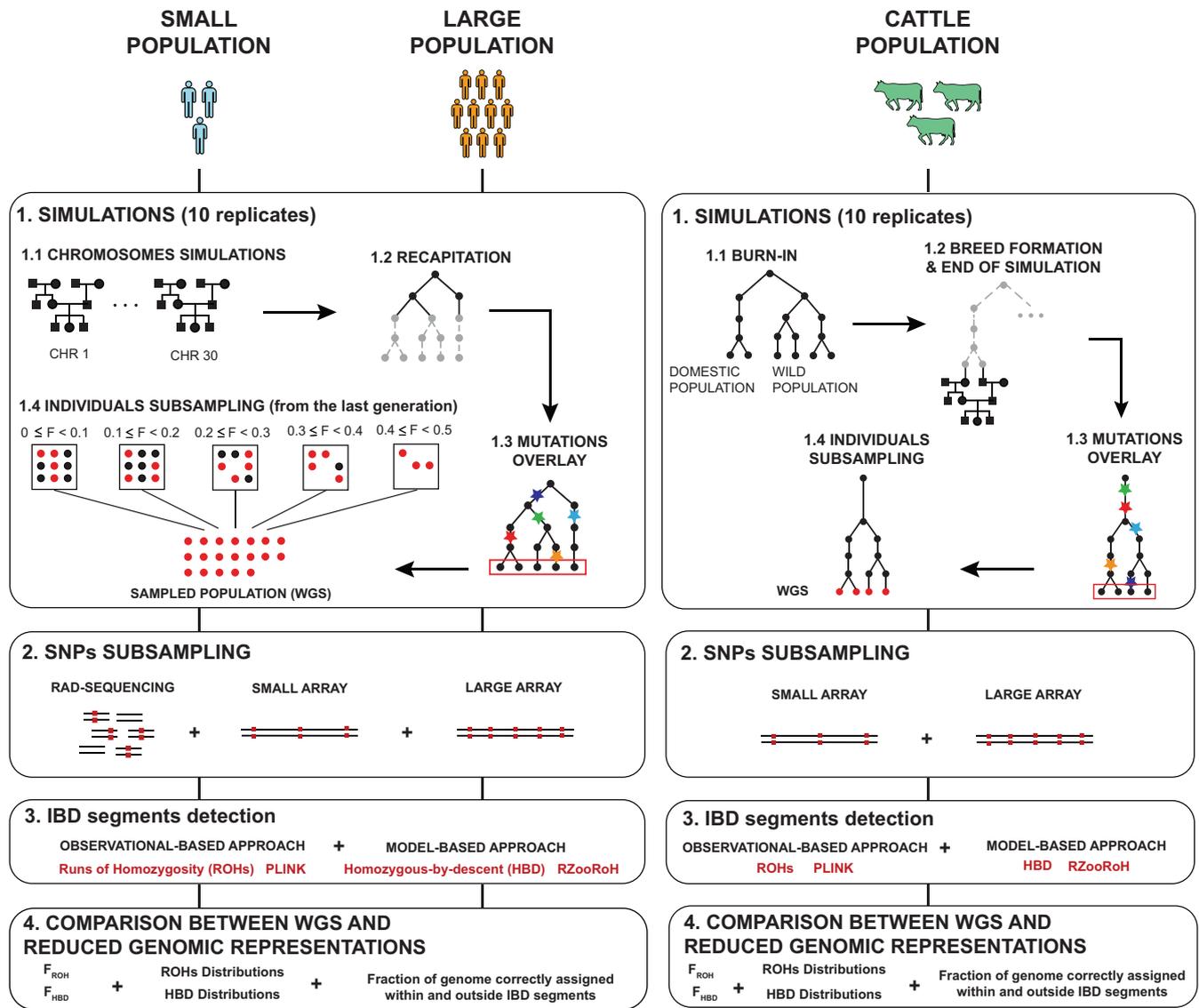
### 2.1 | Simulations

We simulated two hermaphroditic populations ( $N = 1000$  and  $N = 10,000$ ) using SLiM3, a forward-in-time individual-based simulation software (Haller & Messer, 2019). We used a “non-Wright-Fisher” model with nonfixed population sizes and overlapping generations (Haller & Messer, 2019). Population size was regulated via a patch carrying capacity where individuals were removed based on their overall fitness at the end of each simulation cycle. Individuals' fitness decreased with age which varied between 0 and 3: older individuals had higher probabilities to die. Individuals were able to reproduce from the age of 1 and selfing was not allowed. For each individual, its mate was chosen among the other individuals based on their age (with older individuals less likely to be chosen) and on their pedigree-based coancestry with the focal individual (related individuals had higher chances to be chosen). This resulted in a population mostly practicing random mating but ensured that some inbreeding would occur at each generation. We simulated 10 replicates for both population sizes and each simulation lasted for 1000 reproductive cycles. We used a human-like genetic map with a nonhomogenous recombination rate simulated with FREGENE as described in Chadeau-Hyam et al. (2008). This resulted in genomes of 3000 centimorgans (cM). Individuals from both populations carried 30 chromosomes each 100Mb long. The burnin were performed via recapitation in msprime (Kelleher et al., 2016). All mutations were added at the end of the simulation (after the burnin) based on a human-like mutation rate of  $2.5 \times 10^{-8}$  per site per generation (Nachman & Crowell, 2000).

At the end of the entire simulation process, we performed a random stratified sampling to ensure that the individuals used in subsequent analyses would cover the entire range of inbreeding. We are aware that this scheme is rare and hard to apply empirically but it allowed us to investigate whether the entire spectra of inbreeding was correctly estimated. Whenever possible, we subsampled 20 individuals with  $F_{PED}$  between 0 and 0.1, 20 individuals with  $F_{PED}$  between 0.1 and 0.2, 20 individuals with  $F_{PED}$  between 0.2 and 0.3, 20 individuals with  $F_{PED}$  between 0.3 and 0.4 and 20 individuals with  $F_{PED}$  between 0.4 and 0.5. The average ( $\pm$ SD) number of sampled individuals per replicate were  $87.3 \pm 5.03$  for the small population and  $67.40 \pm 4.09$  for the large population. The lower number of individuals subsampled in the large populations are because they contained fewer individuals with high  $F$ . At the very end, the mean ( $\pm$ SD) number of SNPs per simulated population was  $1.6 \times 10^6 \pm 2.0 \times 10^4$  SNPs for the small populations and  $1.6 \times 10^7 \pm 1.7 \times 10^5$  SNPs for the large populations.

### 2.2 | SNPs subsampling

In order to investigate the effect of reduced genomic representations on identical-by-descent (IBD) segments estimation, we mimicked different sequencing techniques by subsampling SNPs from whole-genome data. We simulated both RAD-sequencing and two SNP arrays of different sizes.



**FIGURE 1** General workflow of the study. Simulations were first performed in SLiM3. Within one population, each chromosome shared the same pedigree. Burnin, recapitation and mutation overlay were then performed in msprime. Single nucleotide polymorphism (SNP) subsampling was performed with bedtools for RAD-sequencing and homemade python script for both arrays. Runs of homozygosity (ROHs) were called with PLINK and homozygous-by-descent (HBD) segments with RZooRoH. For distribution of ROHs and HBD segments, segments were divided into six length classes. The true identical-by-descent (IBD) segments were extracted from simulated TreeSequences. Fractions of genomes correctly and incorrectly assigned within and outside IBD segments were then estimated as the overlap between ROHs or HBD segments and these true IBD segments

RAD-sequencing uses restriction enzymes to digest the genome in small fragments, which are then selected on size and sequenced (Andrews et al., 2016). Consequently, these fragments are not homogeneously distributed along the genome. For this purpose, we randomly selected 500 base-pair (bp) fragments (Andrews et al., 2016) using bedtools version 2.29 (Quinlan, 2014). Afterwards, SNPs within these windows were subsampled using --bed function from VCFTools (Danecek et al., 2011). Given that the proportion of the genome sequenced with RAD-sequencing varies greatly depending on the organism and on the restriction enzymes used, we varied this number of fragments so that they covered between 0.05% and 10% of the genome, and between 0.002% and 1%, for the small and the

large populations respectively. We did so because these percentages resulted in similar SNPs densities between the small and large populations and because SNPs density is an accurate indicator for the accuracy of runs of homozygosity (ROHs) or homozygous-by-descent (HBD) segments detection (see results). In addition, we subsampled different percentages of genome (resulting in different SNP densities) between PLINK and RZooRoH. This is because RZooRoH requires smaller SNP densities compared to PLINK to reach the same accuracy of IBD segments estimation. We performed 100 replicates for each RAD-sequencing subsampling percentage.

To simulate SNP array sequencing, we mimicked two arrays initially developed for cattle and widely used for ROHs analyses: the

Illumina BovineSNP 50 beadchip (~50,000 SNPs) hereafter “small array” - and the Illumina BovineHD BeadChip - (~777,000 SNPs) hereafter “large array”. Common features of both arrays are the homogenous distances between SNPs and the focus on common SNPs, hence, we first filtered our WGS data on MAF 5%. We then selected windows with size corresponding to the median distances between SNPs in the real arrays - 40kb for the small array and 3 kb for the large array - and selected the SNP with higher MAF within each window (if at least one SNP was present). We use the term “small array” for what is usually considered as a medium-density array in the literature.

### 2.3 | IBD segments estimation: ROHs And HBD segments

We compared two methods for IBD segments detection to investigate whether we observe a difference in their capacity to handle reduced genomic data. We chose one observational-based approach - the --homozyg method, implemented in PLINK (Chang et al., 2015) - and one model-based approach - the RZooRoH method, implemented as a R package (Bertrand et al., 2019; Druet & Gautier, 2017, 2022). PLINK makes use of a sliding window to identify homozygous segments, used as a proxy for IBD segments. Consequently, PLINK results will be called ROHs. On the contrary, RZooRoH models autozygous segments directly from the genotypes and its results will be referred to as HBD segments.

For ROHs detection with PLINK, we varied parameters according to the SNP density in the reduced data set as proposed by Kardos et al. (2015) and performed by Duntsch et al. (2021). In particular, we varied the window size (--homozyg-window-snp) as well as the minimum SNP density (--homozyg-density) and number of SNPs (--homozyg-snp) required for a homozygous segment to be called a ROH. We required lower numbers of SNPs for low SNPs densities data sets. We also varied the maximum number of heterozygous SNPs allowed in a ROH (--homozyg-het). On the contrary, we fixed some parameters: we authorized 1 heterozygous SNPs per window (--homozyg-window-het = 1), maximum 1 Mb in between two adjacent SNPs (--homozyg-gap = 1000) and a minimum ROH size of 100 kb (--homozyg-kb = 100). These parameters were consistent for every replicate per subsampling method. A more precise description and justification of how each parameter value was chosen according to the SNP density can be found in the Supporting Information (pages 3 to 5; Table S1 and Table S2; Figure S8 and Figure S9).

We called HBD segments with the RZooRoH package version 0.3.1 with a four HBD classes model with rates (R) equals 10, 100, 1000 and 10,000 for the HBD classes and 10,000 for the non-HBD class. These HBD classes correspond to different coalescence event ages: the rate corresponds to the expected number of generations since the coalescence event divided by 2 (i.e., 5, 50, 500 and 5000 generations ago, respectively). For each of these classes, the expected length of the HBD segments are defined as  $1/R$  [in M]: 10 cM, 1 cM, 0.1 cM and 0.01 cM. Even though we chose a model with few

HBD classes, we expect that these classes cover all IBD segments length as the variances associated to these average lengths are extremely large (Speed & Balding, 2015). We used a value of  $5 \times 10^{-5}$  for genotype uncertainty, which represents the probability that any allele mutated in one of the ancestors in the last 1000 reproductive events:  $\#meiosis \times mutation\ rate = 1000 \times 2 \times 2.5 \times 10^{-8}$ .

### 2.4 | SNPs-independent measures of inbreeding: $F_{HOM}$

To test the performance of a SNPs-independent based  $F$ , we estimated  $F_{HOM}$ , implemented in the --het method from PLINK for all SNPs densities presented for RAD-sequencing in this manuscript and for WGS.  $F_{HOM}$ ,  $F_{ROH}$  and  $F_{HBD}$  have different definitions and different assumptions about the ‘base-population’.  $F_{HOM}$  aims at identifying excess homozygosity relative to a random-mating population.  $F_{ROH}$  aims at identifying IBD segments relative to an ancient “base-population” and the minimum length threshold chosen will set how far this base-population is (in coalescence time) compared to our current population. Finally,  $F_{HBD}$  does not use a minimum size threshold but identifies IBD SNPs by including the allelic frequencies into the emission probabilities of the HMM.

### 2.5 | True IBD segments

We compared the estimations obtained from both softwares and all fractions of genome subsampled to the true fraction of genome within IBD segments. We choose to consider a segment IBD if both haplotypes coalesced less than 100 reproductive cycles ago (independently of their length). Since we have overlapping generations and four age classes in our model, these 100 time-steps correspond to 25 generations. In supplementary material, we also benchmark PLINK and RZooRoH results with IBD segments coalescing less than 1000 time-steps (i.e., 250 generations) ago to account for more ancient coalescence events. This was done with the tmrca method from the tskit module in python.

### 2.6 | Statistical analysis

For PLINK, we estimated individual ROHs-based inbreeding coefficient ( $F_{ROH}$ ) as the proportion of the genome within ROHs:

$$F_{ROH} = \frac{\sum Length_{ROH}}{genome\ length}$$
 (McQuillan et al., 2008). For RZooRoH, we estimated  $F_{HBD}$  as the average of posterior HBD probabilities across all markers in the data set (the @realized) as suggested in the RZooRoH documentation. We then compared these to the true fraction of genome within IBD segments ( $F_{IBD}$ ).

Since distribution of IBD segments can inform about the population history (Ceballos et al., 2018), we divided these segments into six length classes following Kirin et al. (2010): (i) between 100kb and

2 Mb (i.e., between 0.1 cM and 2 cM), (ii) between 2 Mb and 4 Mb (i.e., between 2 cM and 4 cM), (iii) between 4 Mb and 6 Mb (i.e., between 4 cM and 6 cM), (iv) between 6 Mb and 10 Mb (i.e., between 6 cM and 10 cM), (v) between 10 Mb and 16 Mb (i.e. between 10 cM and 16 cM) and (vi) larger than 16 Mb (i.e., larger than 16 cM). Distribution of IBD segments are represented as the mean total length per individual among simulation and subsampling replicates. For the sake of comparison and because we can benchmark these with the true distributions of IBD segments, we used these length classes with both softwares even if they are traditionally used with PLINK as RZooRoH models and partitions HBD segments according to the rates chosen when constructing the model. Consequently, with RZooRoH we used the “most probable” distribution of HBD segments detected with the Viterbi algorithm rather than the average SNP probabilities of belonging to any length class. We then compared these distributions to the distribution of true IBD segments extracted from the tree sequences.

We use four metrics to evaluate the accuracy of ROHs detection for each subsampling technique: (i) the fraction of genome correctly assigned within IBD segments (true-IBD), that is, ROHs or HBD segments which were detected and which were IBD; (ii) the fraction of genome correctly assigned outside IBD segments (true-non-IBD), that is, genomic regions which were not classified as ROHs nor HBD segments with neither softwares and were not IBD segments; (iii) the fraction of genome inappropriately assigned within IBD segments (false-IBD), that is, genomic regions which were classified as ROHs or HBD segments but were not IBD; (iv) the fraction of genome inappropriately assigned outside IBD segments (false-non-IBD), that is, IBD segments which were not assigned as ROHs nor HBD segments. We compared ROHs and HBD segment estimation to the true IBD segments for every individual in every replicate, subsampling method and simulation. We then averaged individual's fractions among simulation and subsampling replicates to obtain one measure per subsampling event.

## 2.7 | Additional simulations

We performed an additional batch of simulations based on a real, 57 years deep, cattle pedigree from Walloon beef cattle. We used a genetic map estimated from male Holstein cattle by Qanbari and Wittenburg (2020). In the simulation, a domestic population ( $N_e = 1500$ ) got separated from a large wild population ( $N_e = 50,000$ ) 10,000 generations ago with a migration rate of  $3 \times 10^{-5}$  (Frantz et al., 2020). To mimic the strong selective pressure which occurred during breed formation 200 generations ago and which resulted in high levels of inbreeding (Frantz et al., 2020), 200 individuals were randomly selected from the domestic population and used as founders for the rest of the simulation. The remaining 200 generations were then simulated in SLiM3 from these 200 founders. As the real pedigree was only covering the last 57 years, a first round of simulations was run to obtain 200 generations-deep simulated pedigree, which was then used to complete the real pedigree by assigning a genealogy from the simulated pedigree to each founder from the

real pedigree. At the end, only the individuals from the real pedigree were kept for the analyses.

Since we showed with the first batch of simulations that the accuracy of ROHs detection with RAD-sequencing depends on the proportion of genome subsampled, we only mimicked SNP arrays-like subsampling for these simulations. We did so as previously described. At the end of the sliding windows process, we obtained a lower number of SNPs than expected as some windows did not contain any SNPs. Additional SNPs were chosen randomly (but still with  $MAF > 0.05$ ) to account for empty windows and to reach the same number of SNPs as in real arrays.

Concerning IBD segments detection, we used the same parameters as we did for the small and large populations with PLINK and both SNP arrays. However, with WGS data we increased the maximum number of heterozygous SNPs authorized in a ROH (--homozyg-het) to 64 to optimize distribution of ROHs estimation. Concerning RZooRoH, we used the exact same model as before. For this cattle-like population, we compare the estimations obtained from both softwares to IBD segments coalescing less than 1000 (rather than 100) time-steps ago (independently of their length). This is because in this population, a large part of the inbreeding comes from old coalescence events: we show in Figure S10 that there is poor concordance between both softwares estimation and IBD segments coalescing less than 100 reproductive cycles ago.

## 2.8 | SNP density

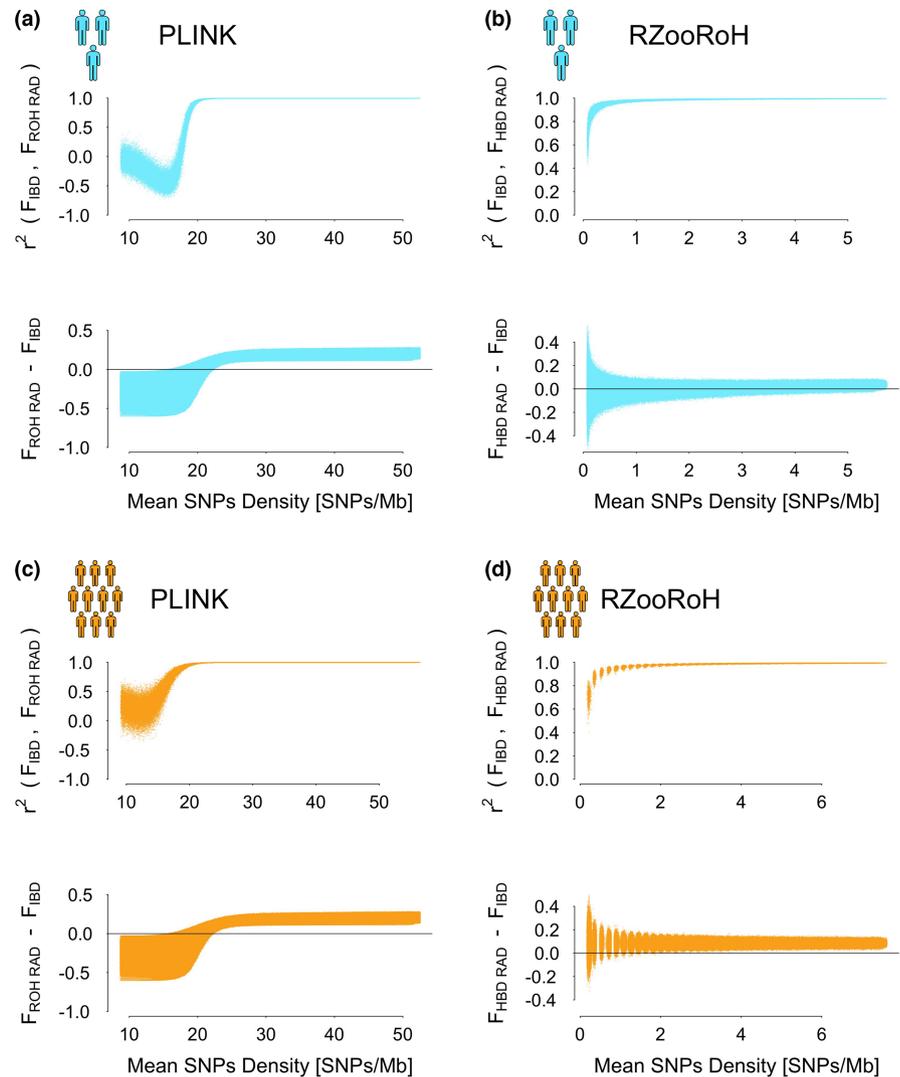
We calculated SNP densities with VCFtools (Danecek et al., 2011) method: --SNPdensity as the number of SNPs per each windows of 1 Mb. We then estimated the mean SNP density of each replicate as the mean density among the windows.

## 3 | RESULTS

### 3.1 | $F_{ROH}$ and $F_{HBD}$

We used simulated genomes to investigate the influence of different sequencing techniques, IBD calling methods and population size on IBD segments detection. Figure 2 shows the correlation ( $r^2$ ) between  $F_{IBD}$  (the true fraction of genome within IBD segments coalescing less than 100 reproductive cycles ago and calculated from simulated data) and its estimations:  $F_{ROH}$  (estimated with PLINK, Figure 2a,c) or  $F_{HBD}$  (estimated with RZooRoH, Figure 2b,d) according to SNP density as well as the absolute difference between the estimated  $F_{ROH \text{ or } HBD}$  and  $F_{IBD}$  according to the same SNP density. Figure 2 shows  $F_{ROH}$  and  $F_{HBD}$  can be correctly estimated with reduced genomic representations providing that a sufficient fraction of the genetic variation is captured. To retain conserved ranking among individual  $F$  (represented by a correlation of one between  $F_{ROH \text{ or } HBD}$  and  $F_{IBD}$ ), PLINK (Figure 2a,c; Table S3) required a SNP density of 22 SNPs/Mb, 11 times higher compared to

**FIGURE 2** Correlation ( $r^2$ ) and difference between  $F_{IBD}$  and  $F_{ROH}$  (panels a and c) or  $F_{HBD}$  (b and d) estimated with RAD-sequencing data according to single nucleotide polymorphism (SNP) density in the reduced data set for both the small (a and b) and large (c and d) populations. Pearson's correlations were estimated per simulation and subsampling replicate. The difference was calculated by subtracting  $F_{IBD}$  from  $F_{ROH}$  or  $F_{HBD}$  per simulation and subsampling replicate. (a) Small population; runs of homozygosity (ROH) have been called with PLINK. (b) Small population; homozygous-by-descent (HBD) segments have been called with RZooRoH. (c) Large population; ROHs have been called with PLINK. (d) Large population; HBD segments have been called with RZooRoH.



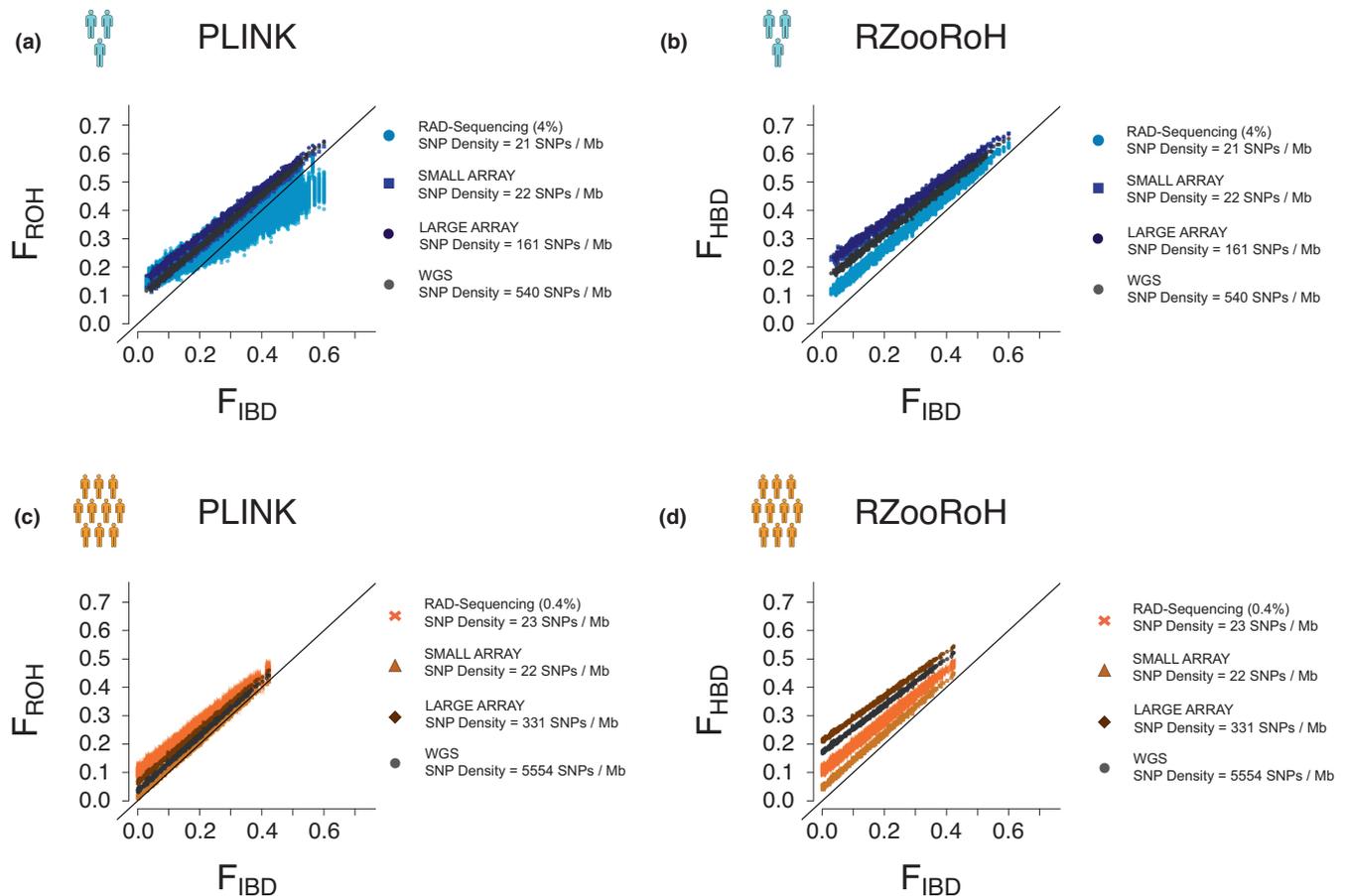
RZooRoH (Figure 2b,d; Table S3) which only required two SNPs/Mb. Interestingly, these minimum SNP densities were similar between the two population sizes suggesting that SNP density might be a key metric for assessing the accuracy of  $F_{IBD}$ . With a SNP density below 20 SNPs/Mb, PLINK resulted in negative correlations between  $F_{ROH}$  and  $F_{IBD}$  in the small population (Figure 2a). On the other hand, correlations were always higher than 0.5 with RZooRoH (Figure 2b,d). We should stress that the conservation of inbreeding ranks does not imply a correct estimation of the “absolute” value of the inbreeding coefficient. Indeed, with both softwares and populations,  $F_{ROH}$  and  $F_{HBD}$  are constantly slightly above  $F_{IBD}$  (estimated with segments coalescing less than 100 time-steps ago) even when the SNP density increases (Figure 2a–d). PLINK did not detect ROHs with SNP densities below 10 SNPs/Mb, resulting in  $F_{ROH}$  of zero for all individuals (Figure S1a). RZooRoH always detected HBD segments, independently of the SNP density we tested. The variance among subsampling replicates was large and the rank of individuals' inbreeding was poorly conserved for SNP densities below 1 SNPs/Mb (Figure 2b,d; Figure S1b). We show in the Supporting Information that  $F_{HOM}$ , an estimator of inbreeding coefficient relying on the difference between the

observed and expected heterozygosity under Hardy–Weinberg yielded similar results to RZooRoH for the same SNPs densities (Figure S2; Table S3).

Concerning medium and high SNP densities (both SNP arrays and WGS), we see little effect of the sequencing method or the software used on  $F_{ROH}$  or  $F_{HBD}$  estimation: inbreeding coefficients estimates were always consistent with  $F_{IBD}$  (Figure 3,  $r^2 > 0.97$ ). All sequencing methods resulted in slightly higher inbreeding coefficients, especially both arrays, but the rank of inbreeding was always conserved among individuals (Figure 3; Table S3). Interestingly, with PLINK and at similar densities a homogeneous spacing between SNPs (the SNP arrays) resulted in better correlations with  $F_{IBD}$  compared to RAD-sequencing in the small population (Figure 3a,c; Table S3).

### 3.2 | Distribution of ROHs and HBD segments

Figure 4 shows distribution of ROHs and HBD segments among the different length classes as the mean per individual (among simulation and subsampling replicates) total ROHs or HBD segments length falling within each length class. Horizontal black



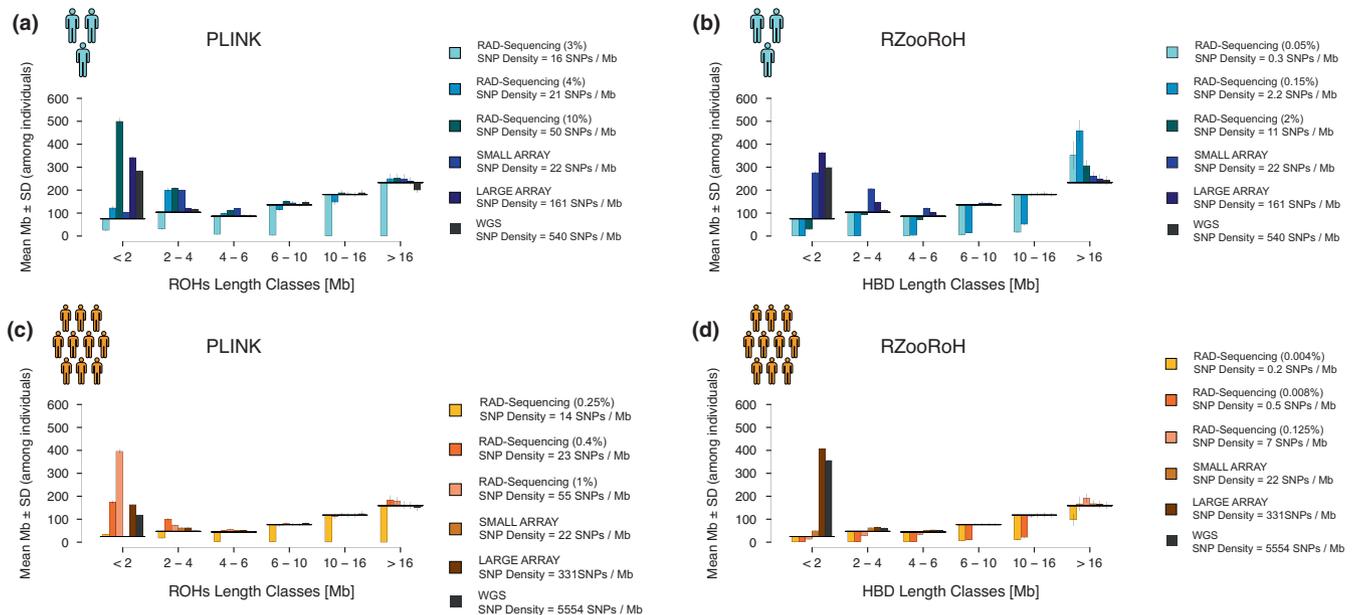
**FIGURE 3** Comparison between runs of homozygosity (ROH) or homozygous-by-descent (HBD) estimated with different sequencing methods on the y-axis and identical-by-descent (IBD) segments (the true fraction of genome within IBD segments coalescing less than 100 time-steps ago) on the x-axis. Each point represents one individual (for one subsampling replicate within one simulation replicate). The black line represents the equality line ( $x = y$ ). Blue points represent individuals from the small population and orange from the large population. Within these two colour categories, a change in shade represents an increase in single nucleotide polymorphism (SNP) density (fraction of genome subsampled indicated between the parentheses for RAD-sequencing). (a) Small population; ROHs were called with PLINK. Please note that points for the small array and whole-genome sequencing (WGS) perfectly overlap. (b) Small population; HBD segments were called with RZooRoH. Please note that points for both arrays perfectly overlap. (c) Large population; ROHs were called with PLINK. Please note that points for the large array and WGS (almost) perfectly overlap. (d) Large population; HBD segments were called with RZooRoH.

lines represent our gold standard: the true mean (among simulation replicate) individual total IBD segment lengths estimated from simulated data for each length class. Bar plots represent the mean (among simulation and subsampling replicate) difference between the estimated distributions (ROHs or HBD segments) and the truth (IBD segments). Thus, bar plots above the horizontal black segment indicate an overestimation while bar plots below the segment an underestimation. In addition, the y-axis starts at 0 indicating that no IBD segments of the particular length class has been detected if the bar reaches the bottom of this axis. Compared to the true distribution of IBD segments (relative to a reference population from 100 reproductive cycles ago), almost all sequencing methods and both softwares resulted in higher mean length of ROHs or HBD segments falling into the smaller length class (Figure 4). At similar SNPs densities (i.e. with both SNP arrays), this overestimation was stronger for RZooRoH compared to PLINK, especially in the small population. However, there is no such overestimation in the small

population when we compare these distributions to older true IBD segments which coalesced less than 1000 reproductive cycles ago suggesting that these segments are not wrongly identified, they simply come from older coalescence events (Figure S3).

With WGS, we can correctly estimate distribution of ROHs and HBD segments larger than 2 Mb (Figure 4). In addition, both SNP arrays, allowed correct estimation of total lengths of ROHs and HBD segments larger than 4 Mb in the small populations and larger than 2 Mb in the large population (Figure 4). These results suggest that medium and high SNPs density datasets can be confidently used for ROHs and HBD segments detection with both PLINK and RZooRoH.

Concerning RAD-sequencing, PLINK allowed the correct estimation of distribution of ROHs with a SNP density around 22 SNPs/Mb (Figure 4a,c). On the other hand, RZooRoH yielded accurate distribution of HBD segments with a SNP density of 11 SNPs/Mb and seven SNPs/Mb in the small and large populations, respectively



**FIGURE 4** Comparison of distribution of runs of homozygosity (ROHs) (a and c) and homozygous-by-descent (HBD) segments (b and d) between the different sequencing methods and the true distributions of identical-by-descent (IBD) segments (defined as segments which coalesced less than 100 reproductive cycles ago). Black horizontal lines correspond to the total length of IBD segments per individual (y-axis) falling into the different length classes (x-axis). Bar plots show the mean ( $\pm$  SD) difference between the mean total length of IBD segment and their estimation (ROHs and HBD segments) for each sequencing method. Bar plots below the horizontal black line indicate an underestimation while bar plots above the horizontal black line indicate an overestimation of the total length of segments. Mean ( $\pm$  SD) are among individuals, simulation and subsampling replicates. (a) Distribution of ROHs from the small population; ROHs were called with PLINK. (b) Distribution of HBD segments from the small population; HBD segments were called with RZooRoH. (c) Distribution of ROHs from the large population; ROHs were called with PLINK. (d) Distribution of HBD segments from the large population; HBD segments were called with RZooRoH

(Figure 4b,d). For lower SNP densities, RZooRoH tended to merge small adjacent HBD segments into larger ones (Figure 4d).

We also investigated ROHs calling accuracy with *PLINK* and the default parameters for all reduced genomic representations. We show in the Supporting Information that similarly to what was observed in Figure 2,  $F_{ROH}$  can be correctly estimated with SNP densities higher than 22 SNPs/Mb (Figure S4 and Figure S5). However, the distribution of ROHs are always biased with a large overestimation of small IBD segments and underestimation of large IBD segments even with both SNP arrays (Figure S6). These results emphasize the importance of fine-tuning *PLINK* parameters when working with reduced genomic representations.

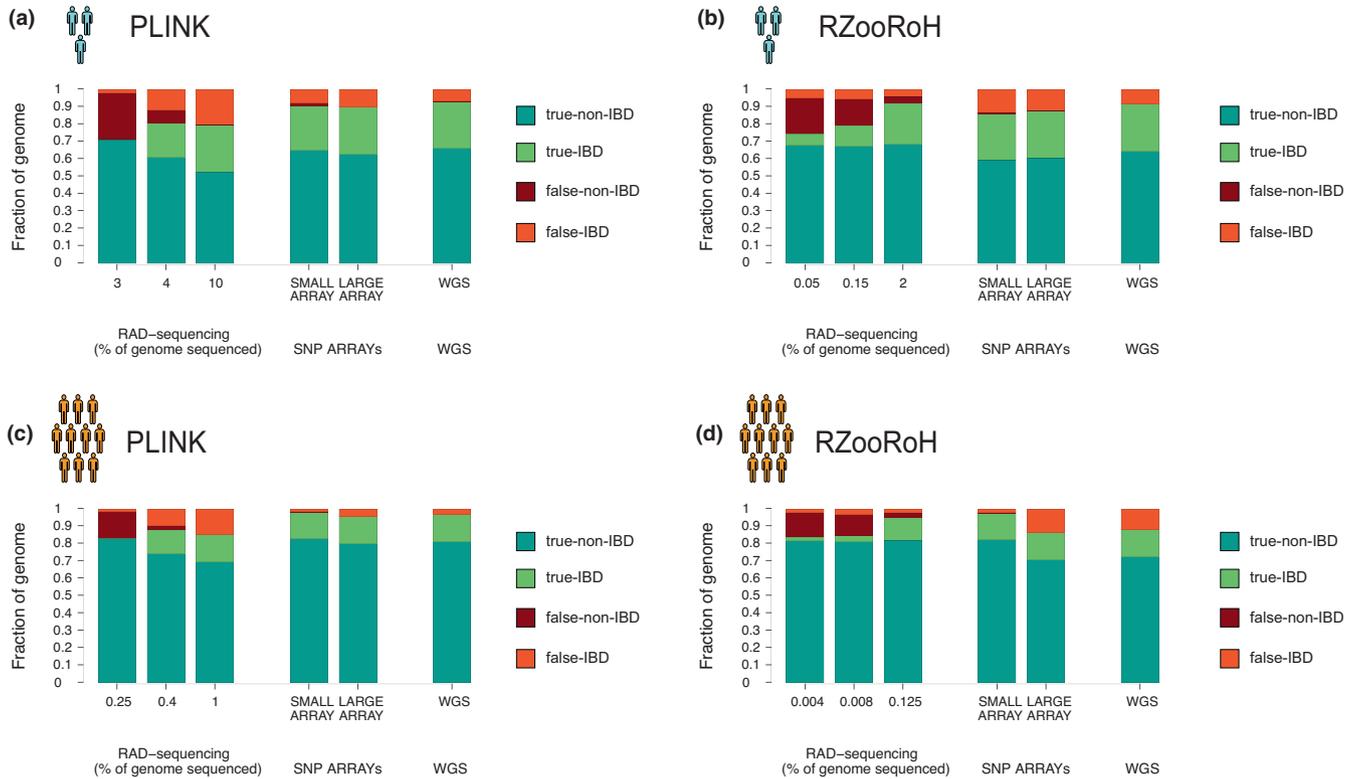
### 3.3 | Fraction of genome assigned within and outside IBD segments

Figure 5 shows the mean fraction of genome which has been correctly (true-non-IBD and true-IBD) and incorrectly (false-non-IBD and false-IBD) assigned to ROHs or HBD segments. Concerning RAD-sequencing, *PLINK* resulted in high fractions of genome incorrectly assigned to ROHs (false-IBD) (Figure 5a,c, left column). On the other hand, RZooRoH resulted in 90 and 95% of the genome correctly assigned within our outside IBD segments with 2% (SNP density = 11 SNPs/Mb) and 0.125% (SNP density = 7 SNPs/Mb) of

the genome sequenced in the small and large populations respectively (Figure 5b,d, right column). Concerning SNP arrays and WGS, the “incorrectly” assigned ROHs and HBD segments were mostly false positive (false-IBD). However, we show in Figure S7 that these false positive become true positive when compared to IBD segments which coalesced less than 1000 reproductive cycles ago. This indicates that these fragments come from coalescence events older than 100 reproductive cycles ago. In the large population, RZooRoH still resulted in a few false-positive when compared to distribution of IBD segments from less than 1000 reproductive cycles ago, with both the large array and WGS (Figure S7d).

### 3.4 | Cattle simulations

Figure 6 shows the comparison of  $F_{ROH}$  and  $F_{HBD}$  estimates, distribution of ROHs and HBD segments and fractions of genome correctly assigned within and outside IBD segments (coalescing less than 1000 reproductive cycles ago) for WGS and both SNP arrays in the cattle population. For this population, we consider IBD segments coalescing less than 1000 reproductive cycles ago because the major part of inbreeding comes from ancient coalescence events and both softwares resulted in poor IBD segments estimation when compared to segments coalescing less than 100 reproductive cycles ago (Figure S10). In this population,



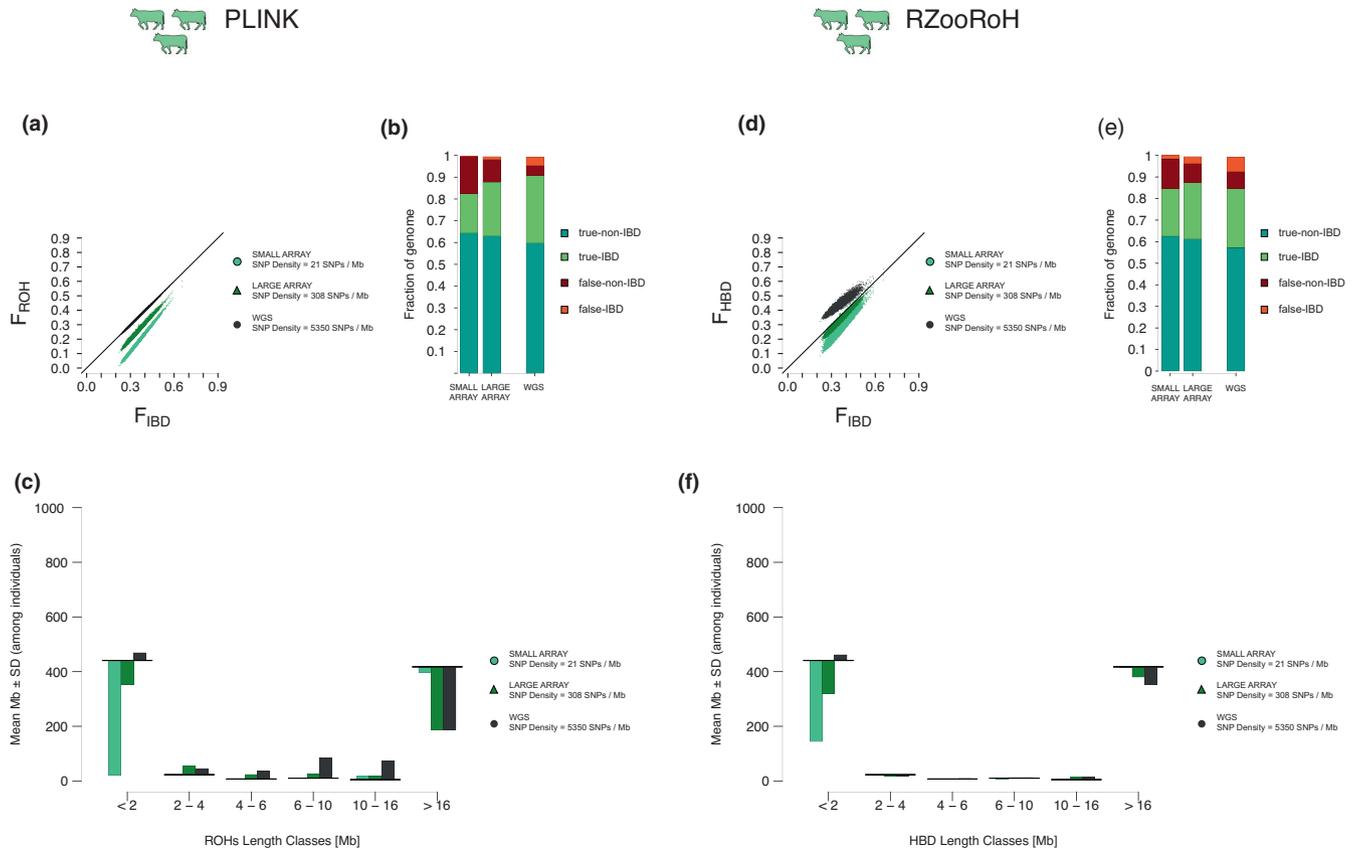
**FIGURE 5** For each sequencing method, fraction of genome correctly assigned outside identical-by-descent (IBD) segments (true-non-IBD), correctly assigned within IBD segments (true-IBD), incorrectly assigned outside IBD segments (false-non-IBD) and incorrectly assigned within IBD segments (false-IBD) are represented. Values are averaged among individuals as well as both simulation and subsampling replicates. (a) Small population; runs of homozygosity (ROHs) were called with PLINK. (b) Small population; HBD segments were called with RZooRoH. (c) Large population; ROHs were called with PLINK. (d) Large population; homozygous-by-descent (HBD) segments were called with RZooRoH

PLINK resulted in perfect estimation of the inbreeding coefficient (Figure 6a;  $r^2(F_{ROH\ WGS}, F_{IBD}) = 0.999$ ;  $r^2(F_{ROH\ LARGE\ ARRAY}, F_{IBD}) = 0.993$ ;  $r^2(F_{ROH\ SMALL\ ARRAY}, F_{IBD}) = 0.996$ ) but biased distribution of ROHs even with WGS (Figure 6c). Compared to the true distribution of IBD segments (coalescing less than 1000 reproductive cycles ago), PLINK identified higher numbers of ROHs smaller than 16 Mb (bar plots above the horizontal line) and fewer ROHs larger than 16 Mb (bar plots below the horizontal line) (Figure 6c). On the other hand with RZooRoH the correlation between  $F_{HBD}$  and  $F_{IBD}$  were slightly lower (Figure 6d:  $r^2(F_{HBD\ WGS}, F_{IBD}) = 0.937$ ,  $r^2(F_{HBD\ LARGE\ ARRAY}, F_{IBD}) = 0.983$ ;  $r^2(F_{HBD\ SMALL\ ARRAY}, F_{IBD}) = 0.955$ ) but distribution of HBD segments were closer to the true distribution of IBD segments (Figure 6f). Similar to previous observations, RZooRoH detected fewer small HBD segments (<2Mb) with the small and to a lesser extend the large SNP array. Finally, concerning the fraction of genome correctly and incorrectly assigned to IBD segments, 83, 88 and 91% of the genome were correctly assigned within or outside ROHs with PLINK using the small array, large array and WGS, respectively (Figure 6b). With RZooRoH, 84, 87 and 85% of the genome were correctly assigned within or outside HBD segments with the small array, the large array and WGS (Figure 6e). With both arrays and softwares, the wrongly assigned fraction of genome was mainly false-negative (false-non-IBD) but both softwares resulted in a few false-positive (false-IBD) with WGS data (Figure 6b,e).

## 4 | DISCUSSION

### 4.1 | Summary

We investigated the capacity of WGS, SNP arrays and RAD-seq to perform IBD segments analyses using either the observational-based runs of homozygosity (ROHs) calling approach implemented in PLINK (Chang et al., 2015) or the model-based homozygous-by-descent (HBD) segments calling approach implemented in RZooRoH (Bertrand et al., 2019; Druet & Gautier, 2017). We show that both methods can be used with medium to high SNPs density data sets in simulations with constant population sizes and proportion of inbreeding. However, for RAD-seq, PLINK required a SNP density above 22 SNPs/Mb to keep the ranking among individuals  $F_{ROH}$  and correct estimates of distribution of ROHs. On the other hand, RZooRoH only required a SNP density of 11 SNPs/Mb to obtain correct distribution of HBD segments and the rank of individual inbreeding coefficients  $F_{HBD}$  was conserved when the SNP density was above 2 SNPs/Mb. We also show that in the cattle population, PLINK did not estimate distribution of IBD segments as accurately as RZooRoH even with WGS data. Finally, we show in Supporting Information that  $F_{HOM}$ , a non-IBD segments-based estimate of individual  $F$ , is as accurate as  $F_{HBD}$  estimated with the model-based approach for the same SNP density.



**FIGURE 6** Comparison of runs of homozygosity (ROHs) and homozygous-by-descent (HBD) segment detection with whole-genome sequencing (WGS) and both single nucleotide polymorphism (SNP) arrays in the cattle population and for both identical-by-descent (IBD) segment detection methods. (a) Comparison of  $F_{ROH}$  estimated with PLINK on WGS data and both SNPs arrays with regard to the true fraction of genome within IBD segments (coalescing less than 1000 reproductive cycles ago):  $F_{IBD}$ . (b) Fraction of genome correctly assigned outside IBD segments (true-non-IBD), correctly assigned within IBD segments (true-IBD), incorrectly assigned outside IBD segments (false-non-IBD) and incorrectly assigned within IBD segments (false-IBD) for WGS and both SNP arrays. ROHs were called with PLINK. (c) Comparison of distribution of ROHs with WGS and both arrays with regard to the true distribution of IBD segments (coalescing less than 1000 generations ago). Horizontal black lines represent the true mean (among simulation replicates) individual total length of IBD segments estimated from simulated data for each length class. Bar plots represent the mean (among simulation and subsampling replicate) difference between the estimated distributions (ROHs) and the truth (IBD segments). ROHs were called with PLINK. (d) Comparison of  $F_{HBD}$  estimated with RZooRoH on WGS data and both SNP arrays with regard to the true fraction of genome within IBD segments (coalescing less than 1000 generations ago):  $F_{IBD}$ . (e) Fraction of genome correctly assigned outside IBD segments (true-non-IBD), correctly assigned within IBD segments (true-IBD), incorrectly assigned outside IBD segments (false-non-IBD) and incorrectly assigned within IBD segments (false-IBD) for WGS and both arrays. HBD segments were called with RZooRoH. (f) Comparison of distribution of HBD segments with WGS and both arrays with regard to the true distribution of IBD segments (coalescing less than 1000 generations ago). Horizontal black lines represent the true mean (among simulation replicate) individual total length of IBD segments estimated from simulated data for each length class. Bar plots represent the mean (among simulation and subsampling replicate) difference between the estimated distributions (HBD segments) and the truth (IBD segments). HBD segments were detected with RZooRoH

## 4.2 | PLINK vs. RZooRoH

With WGS and in the small and large populations, the two softwares yielded similar results concordant with the true IBD segments (extracted from simulated data as IBD segments which coalesced less than 100 reproductive cycles ago) for all segments larger than 2 Mb. We showed that both methods, but especially RZooRoH, identify a larger number of small IBD segments (<2 Mb) than the truth. However, we also show that these segments come from more ancient coalescence events. It makes sense that RZooRoH identifies a higher number of smaller HBD segments

compared to PLINK since it has no constraint on HBD minimum segments size. From the formula presented in Thompson (2013), the length of the IBD segment in centimorgans (cM)  $l$  is a simple function of the number of generations  $g$  since the coalescence event:  $l = 100/2g$ , our PLINK threshold corresponds to 0.1 cM (100 kb) in humans and thus coalescence events 500 reproduction events ago on average (with a very large variance [Speed & Balding, 2015]).

It is important to remember that even with WGS, IBD segment detection is still challenging. Beside the IBD segments detection method, many parameters can influence the result of IBD segments

calling. For instance, a minimum size threshold can be chosen. However, it has been shown with simulations that small segments can also result from recent coalescence events (Speed & Balding, 2015), thus neglecting smaller regions might lead to an underestimation of the inbreeding status of the individual or the population. Another important parameter is whether mutations (and sequencing errors) should be considered by allowing heterozygous SNPs in IBD segments and how many heterozygous markers are to be allowed can greatly differ among studies. To summarize, no consensus exist nowadays on which method and parameters are the best and further investigation is needed.

In this project, we used the true IBD values for parameter optimization with PLINK, not available with empirical data. We note that previous studies used WGS as the “gold-standard” and/or high coverage assembly data for parameters optimization (Duntsch et al., 2021; Meyermans et al., 2020; Mueller et al., 2022). Mueller et al., (2022) showed adapting parameters allow to get similar results between WGS and reduced genomic representations with smaller fractions of genomes: the authors used PLINK to call ROHs with RAD-sequencing, then tested ROHs calling with several different settings for three individuals for which they also had WGS data and extracted the settings which best conserved the rank of inbreeding found with WGS. We want to stress that WGS data might not be available or might not result in accurate IBD segments detection as observed with our cattle population. Finally, varying settings need to be done with caution: different settings can increase the number of ROHs detected but also the likelihood of noncorrect calls and maybe bias the individuals inbreeding ranking (Meyermans et al., 2020).

#### 4.3 | Estimation of $F_{ROH}$ and $F_{HBD}$

With WGS, both arrays and when a large portion of the genome is sequenced with RAD-sequencing, ranks of inbreeding were always conserved among individuals. Our results are consistent with Kardos et al. (2018) who showed (see Supporting Information) that  $F_{HBD}$  estimated with 10,000 loci, and a home-made script based on a likelihood ratio method (adapted from Pemberton et al., 2012), are similar to  $F_{HBD}$  estimated with WGS in an inbred wild wolf population.

With RAD-sequencing, accurate results of the fraction of genome within IBD segments depends on the fraction of genome sequenced, which drastically differ between both softwares. RZooRoH required a fraction of the genome 11 times smaller compared to PLINK for conserved ranking of inbreeding estimates among individuals. It is expected that model-based approaches perform better with low SNPs-densities as the distances between SNPs are taken into account in the model (Bertrand et al., 2019). Hence, we strongly recommend RZooRoH when working with reduced genomic data. Nevertheless, when the SNPs density is sufficient, the conservation of inbreeding ranks was observed with both softwares and is consistent with other studies: Duntsch et al. (2021) compared ROHs and HBD segments estimates from PLINK and

RZooRoH with RAD-sequencing, a custom-made array and WGS in few hihi (*Notiomystis cincta*; a nonmodel bird species) individuals. They found conserved individual inbreeding ranks with all reduced representations.

In the cattle population and with both SNP arrays,  $F_{ROH}$  and  $F_{HBD}$  estimates were lower than  $F_{IBD}$ . The missing portion of inbreeding was coming from small IBD segments with more ancient coalescence events. Indeed, recent inbreeding is easier to capture: with lower marker density you are not able to capture small segments (Druet & Gautier, 2017). This is supported by Sole et al. (2017) who compared SNP arrays of different sizes in cattle and showed that all arrays capture the same levels of recent inbreeding but higher densities allow older inbreeding to be captured.

Inbreeding coefficients are used for inbreeding depression studies, which are key for understanding the evolution of populations and for conservation managements in endangered species (Lynch & Walsh, 1998). If the rank of individuals'  $F$  is conserved, it ensures that the direction of the correlation between inbreeding estimates and phenotypes (the sign of the regression slope), and thus the general effect of inbreeding on the trait, is correctly estimated. However, underestimating the inbreeding coefficients can lead to an overestimation of the magnitude of inbreeding depression (the regression slope absolute value will be steeper). We showed in Supporting Information that  $F_{HOM}$  performed as well as model-based RZooRoH approach with similar SNPs densities. This last result is consistent with another study which showed that 5000 markers are sufficient to obtain genomic kinship values similar to pedigree estimates (Goudet et al., 2018). Hence, we strongly advise to use SNP independent measures (i.e., not based on IBD segments) or model-based HBD segments approaches for inbreeding studies when SNP density is low. However, it is important to keep in mind that  $F_{HOM}$  can only be used if the number of individuals is large enough to allow correct estimation of the allelic frequencies.

#### 4.4 | Distribution of ROHs and HBD segments

Distribution of IBD segments are used to describe the demographic history of the populations (Ceballos et al., 2018; Kirin et al., 2010) as it is possible to link the length of an IBD segment to the number of generations back to the common ancestor (Thompson, 2013). Our study suggests that with both WGS and SNP arrays, length of ROHs and HBD segments values are correctly estimated. On the contrary, with RAD-sequencing, length of ROHs and HBD segments should not be trusted at low SNP densities, meaning that age estimation for these segments is currently impossible. In the cattle population and with WGS data, RZooRoH resulted in more accurate distribution of HBD segments compared to PLINK. This was not the case in the constant sizes population simulations. This difference is due to the more complex demographic history of the cattle population which underwent both a strong bottleneck and intense selective pressures. We hypothesize that the difference observed between both softwares is due to the way they handle

heterozygous SNPs. RZooRoH uses the probability of observing a heterozygous SNP in an HBD segment per base-pair, automatically adapting the number of heterozygous SNPs allowed in an HBD segment to its size. This is not the case with PLINK. Hence, manually increasing the maximum number of heterozygous SNPs allowed in a ROH will merge small adjacent ROHs belonging to the same IBD segment into a larger one; but might also increase the number of false positives. Consequently, we advise researchers to use model-based approaches when working with populations with complex demographic history.

As we expect all the bias mentioned above to be similar among populations, relative comparisons between populations genotyped with the same sequencing method lead to reliable results as shown by Kirin et al. (2010): these authors used the HGDP data set genotypes (from 2009) with Illumina 650Y product and PLINK and compared distribution of ROHs among various human populations. The populations undergoing contemporaneous inbreeding such as West and South Asians and Oceanians harboured a higher fraction of genome within long ROHs while populations from central America harboured lower numbers of large ROHs but higher numbers of small ROHs, consistent with an history of ancient small effective population size. Similarly, Mastrangelo et al. (2016) compared distribution of ROHs estimated with PLINK from three dairy cattle breeds genotyped with the “small array” (Illumina BovineSNP 50 beadchip) to compare the different breeds and to assess their inbreeding status. The authors found that Italian Holstein individuals harboured a high number of short ROHs suggesting that inbreeding in this breed is mostly caused by ancient relatedness within the population rather than recent mating between relatives. Individuals from two local breeds *Modicana* and *Cinisara* harboured a high number of large ROHs, suggesting recent mating events between relatives. The authors concluded that the implementation of a monitored breeding programme aimed at reduced consanguinity was necessary in these local breeds.

With WGS, both SNP arrays and high SNPs density RAD-sequencing, we obtained correct assignment of the genome within and outside IBD segments above 90%. The remaining incorrectly assigned regions correspond mostly to IBD segments with older coalescence times or segments in between adjacent IBD segments. This trend indicates that IBD regions are correctly detected with medium to high densities and thus suggests that they can be confidently used for homozygosity mapping studies.

#### 4.5 | Effect of population size

Population size influenced the minimum proportion of genome required: the larger population required a smaller fraction of genome to obtain accurate IBD estimation. However, the larger population harboured higher genetic variation and thus a higher number of SNPs. Consequently, we were more likely to have SNPs in the subsampled regions with RAD-sequencing. Since the simulations scheme forced mating between related individuals and since we performed a random stratified individual sampling independently

of the effective size of the population, we do not expect distribution of IBD segments to reflect both populations “true” distributions. Consequently, in this study we cannot quantify the true effect of population size (other than the number of SNPs) on IBD segments detection. To disentangle these effects, one could vary the mutation rate so that both populations have similar SNP densities.

#### 4.6 | Limitations

We want to stress that the fractions of the genome sequenced used in the present study are indicative and do not correspond to “true” proportion of genome sequenced needed to obtain meaningful results. They come from simulated data and this fraction will be lower after quality filtering with empirical data, especially for RAD-sequencing and WGS where genotyping rates are lower. In addition, we did not include the effect of allele dropout when we simulated RAD-sequencing, which could influence the accuracy of IBD segments detection. Indeed, individuals heterozygous at restriction sites might appear homozygous which can bias further population genetic inferences, especially in large populations (Arnold et al., 2013; Gautier et al., 2013). Finally, we did not include in this project the method for IBD segments detection based on the odds ratio comparison of the likelihood of the genotype if autozygous or allozygous (Broman & Weber, 1999; Pemberton et al., 2012) but we expect that this method can handle reduced genomic representations since it takes linkage disequilibrium into account.

#### 4.7 | Conclusion

Using simulated data, we compared  $F_{ROH}$  and  $F_{HBD}$  estimates as well as distribution of ROHs and HBD segments to the true IBD (from simulated data and calculated as segments which coalesced less than 100 reproductive cycles ago).  $F_{ROH}$  and  $F_{HBD}$  can be correctly estimated with all sequencing methods when the SNP density is above 22 SNPs per Mb with PLINK and above two SNPs per Mb with RZooRoH. With lower SNPs densities,  $F_{HOM}$ , a genomic estimate of inbreeding coefficients not based on ROHs, is as accurate as the model-based estimate, for a fraction of the computing time. We would therefore recommend using independent SNPs-based genomic estimates such as  $F_{HOM}$  for inbreeding quantification with reduced genomic representation, unless the number of individuals analysed is too small to allow a correct estimation of the population alleles frequency or mean coancestry. Regarding distribution of ROHs and HBD segments, even though the majority of the genome is correctly assigned within and outside IBD segments, both softwares failed to capture IBD segments with older coalescence times at low densities. This still allows comparing populations analysed with the same methodology but prevents comparing distribution of ROHs or HBD segments from studies using different reduced genomic methods. In addition, in a population with a more complex demographic history, only RZooRoH resulted in accurate

distribution of HBD segments with WGS data. To conclude, we find little advantages in using IBD segments-based estimates of inbreeding at low SNPs densities and show that only model-based approaches can be used for distribution of HBD segments quantification at such low SNP densities and in populations with complex demographic histories.

#### AUTHOR CONTRIBUTIONS

Eléonore Lavanchy and Jérôme Goudet conceptualized the study. Eléonore Lavanchy conducted the study, performed all simulations and analyses and drafted the manuscript. Eléonore Lavanchy and Jérôme Goudet wrote and revised the manuscript.

#### ACKNOWLEDGEMENTS

We are grateful to the Association Wallonne de l'Elevage (awé) for accepting to share their cattle pedigree with us and to Tom Druet who kindly asked them on our behalf and provided feedbacks on this manuscript. We also want to thank Marty Kardos for his comments on this manuscript. Finally, we want to thank the reviewers: Anna Santure and an anonymous reviewer for their great comments and feedbacks. This study was funded by the Swiss National Science Foundation with grant 31003A\_179358 to JG. Open Access Funding provided by University of Lausanne.

#### CONFLICT OF INTEREST

The authors declare no conflict of interest.

#### DATA AVAILABILITY STATEMENT

All scripts necessary to generate genomic data and perform analyses used in this study have been made available on GitHub: <https://github.com/EléonoreLavanchy/ROHsReducedRep>.

#### ORCID

Eléonore Lavanchy  <https://orcid.org/0000-0003-4951-9332>

Jérôme Goudet  <https://orcid.org/0000-0002-5318-7601>

#### REFERENCES

- Åkesson, M., Liberg, O., Sand, H., Wabakken, P., Bensch, S., & Flagstad, Ø. (2016). Genetic rescue in a severely inbred wolf population. *Molecular Ecology*, 25(19), 4745–4756. <https://doi.org/10.1111/mec.13797>
- Alemu, S. W., Kadri, N. K., Harland, C., Faux, P., Charlier, C., Caballero, A., & Druet, T. (2020). An evaluation of inbreeding measures using a whole-genome sequenced cattle pedigree. *Heredity*, 126, 410–423. <https://doi.org/10.1038/s41437-020-00383-9>
- Alkuraya, F. S. (2013). The application of next-generation sequencing in the autozygosity mapping of human recessive diseases. *Human Genetics*, 132(11), 1197–1211. <https://doi.org/10.1007/s00439-013-1344-x>
- Andrews, K. R., Good, J. M., Miller, M. R., Luikart, G., & Hohenlohe, P. A. (2016). Harnessing the power of RADseq for ecological and evolutionary genomics. *Nature Reviews Genetics*, 17(2), 81–92. <https://doi.org/10.1038/nrg.2015.28>
- Arnold, B., Corbett-Detig, R. B., Hartl, D., & Bomblies, K. (2013). RADseq underestimates diversity and introduces genealogical biases due to nonrandom haplotype sampling. *Molecular Ecology*, 22(11), 3179–3190. <https://doi.org/10.1111/mec.12276>
- Bertrand, A. R., Kadri, N. K., Flori, L., Gautier, M., & Druet, T. (2019). RZooRoH: An R package to characterize individual genomic autozygosity and identify homozygous-by-descent segments. *Methods in Ecology and Evolution*, 10(6), 860–866. <https://doi.org/10.1111/2041-210X.13167>
- Bittles, A. H., & Black, M. L. (2010). Consanguinity, human evolution, and complex diseases. *Proceedings of the National Academy of Sciences*, 107(suppl 1), 1779–1786. <https://doi.org/10.1073/pnas.0906079106>
- Bjelland, D. W., Weigel, K. A., Vukasinovic, N., & Nkrumah, J. D. (2013). Evaluation of inbreeding depression in Holstein cattle using whole-genome SNP markers and alternative measures of genomic inbreeding. *Journal of Dairy Science*, 96(7), 4697–4706. <https://doi.org/10.3168/jds.2012-6435>
- Bosse, M., Megens, H.-J., Madsen, O., Paudel, Y., Frantz, L. A. F., Schook, L. B., Crooijmans, R. P. M. A., & Groenen, M. A. M. (2012). Regions of homozygosity in the porcine genome: Consequence of demography and the recombination landscape. *PLoS Genetics*, 8(11), e1003100. <https://doi.org/10.1371/journal.pgen.1003100>
- Broman, K. W., & Weber, J. L. (1999). Long homozygous chromosomal segments in reference families from the Centre d'Étude du Polymorphisme Humain. *The American Journal of Human Genetics*, 65(6), 1493–1500. <https://doi.org/10.1086/302661>
- Browning, S. R., & Browning, B. L. (2010). High-resolution detection of identity by descent in unrelated individuals. *The American Journal of Human Genetics*, 86(4), 526–539. <https://doi.org/10.1016/j.ajhg.2010.02.021>
- Caballero, A., Villanueva, B., & Druet, T. (2020). On the estimation of inbreeding depression using different measures of inbreeding from molecular markers. *Evolutionary Applications*, 14, 416–428. <https://doi.org/10.1111/eva.13126>
- Carothers, A. D., Rudan, I., Kolcic, I., Polasek, O., Hayward, C., Wright, A. F., Campbell, H., Teague, P., Hastie, N. D., & Weber, J. L. (2006). Estimating human inbreeding coefficients: Comparison of genealogical and marker heterozygosity approaches. *Annals of Human Genetics*, 70(5), 666–676. <https://doi.org/10.1111/j.1469-1809.2006.00263.x>
- Ceballos, F. C., Joshi, P. K., Clark, D. W., Ramsay, M., & Wilson, J. F. (2018). Runs of homozygosity: Windows into population history and trait architecture. *Nature Reviews Genetics*, 19(4), 220–234. <https://doi.org/10.1038/nrg.2017.109>
- Chadeau-Hyam, M., Hoggart, C. J., O'Reilly, P. F., Whittaker, J. C., De Iorio, M., & Balding, D. J. (2008). Fregene: Simulation of realistic sequence-level data in populations and ascertained samples. *BMC Bioinformatics*, 9(1), 364. <https://doi.org/10.1186/1471-2105-9-364>
- Chang, C. C., Chow, C. C., Tellier, L. C., Vattikuti, S., Purcell, S. M., & Lee, J. J. (2015). Second-generation PLINK: Rising to the challenge of larger and richer datasets. *GigaScience*, 4(1), 7. <https://doi.org/10.1186/s13742-015-0047-8>
- Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., Handsaker, R. E., Lunter, G., Marth, G. T., Sherry, S. T., McVean, G., Durbin, R., & 1000 Genomes Project Analysis Group. (2011). The variant call format and VCFtools. *Bioinformatics*, 27(15), 2156–2158. <https://doi.org/10.1093/bioinformatics/btr330>
- de Jong, J. F., van Hooft, P., Megens, H.-J., Crooijmans, R. P. M. A., de Groot, G. A., Pemberton, J. M., Huisman, J., Bartoš, L., Iacolina, L., van Wieren, S. E., Ydenberg, R. C., & Prins, H. H. T. (2020). Fragmentation and translocation distort the genetic landscape of ungulates: Red Deer in The Netherlands. *Frontiers in Ecology and Evolution*, 8, 365. <https://doi.org/10.3389/fevo.2020.535715>
- Druet, T., & Gautier, M. (2017). A model-based approach to characterize individual inbreeding at both global and local genomic scales. *Molecular Ecology*, 26(20), 5820–5841. <https://doi.org/10.1111/mec.14324>

- Druet, T., & Gautier, M. (2022). A hidden Markov model to estimate homozygous-by-descent probabilities associated with nested layers of ancestors. *Theoretical Population Biology*, 145, 38–51. <https://doi.org/10.1016/j.tpb.2022.03.001>
- Duntsch, L., Whibley, A., Brekke, P., Ewen, J. G., & Santure, A. W. (2021). Genomic data of different resolutions reveal consistent inbreeding estimates but contrasting homozygosity landscapes for the threatened Aotearoa New Zealand hihi. *Molecular Ecology*, 30(23), 6006–6020. <https://doi.org/10.1111/mec.16068>
- Forutan, M., Ansari Mahyari, S., Baes, C., Melzer, N., Schenkel, F. S., & Sargolzaei, M. (2018). Inbreeding and runs of homozygosity before and after genomic selection in north American Holstein cattle. *BMC Genomics*, 19(1), 98. <https://doi.org/10.1186/s12864-018-4453-z>
- Forutan, M., Mahyari, S. A., Baes, C., Melzer, N., Schenkel, F. S., & Sargolzaei, M. (2018). Inbreeding and runs of homozygosity before and after genomic selection in north American Holstein cattle. *BMC Genomics*, 19(1), 98. <https://doi.org/10.1186/s12864-018-4453-z>
- Franklin, I. R. (1977). The distribution of the proportion of the genome which is homozygous by descent in inbred individuals. *Theoretical Population Biology*, 11(1), 60–80. [https://doi.org/10.1016/0040-5809\(77\)90007-7](https://doi.org/10.1016/0040-5809(77)90007-7)
- Frantz, L. A. F., Bradley, D. G., Larson, G., & Orlando, L. (2020). Animal domestication in the era of ancient genomics. *Nature Reviews Genetics*, 21(8), 449–460. <https://doi.org/10.1038/s41576-020-0225-0>
- Gautier, M., Gharbi, K., Cezard, T., Foucaud, J., Kerdelhué, C., Pudlo, P., Cornuet, J.-M., & Estoup, A. (2013). The effect of RAD allele dropout on the estimation of genetic variation within and between populations. *Molecular Ecology*, 22(11), 3165–3178. <https://doi.org/10.1111/mec.12089>
- Ghani, M., Reitz, C., Cheng, R., Vardarajan, B. N., Jun, G., Sato, C., Naj, A., Rajbhandary, R., Wang, L.-S., Valladares, O., Lin, C.-F., Larson, E. B., Graff-Radford, N. R., Evans, D., De Jager, P. L., Crane, P. K., Buxbaum, J. D., Murrell, J. R., Raj, T., ... Genetics, A. D. (2015). Association of Long Runs of homozygosity with Alzheimer disease among African American individuals. *JAMA Neurology*, 72(11), 1313–1323. <https://doi.org/10.1001/jamaneurol.2015.1700>
- Gibson, J., Morton, N. E., & Collins, A. (2006). Extended tracts of homozygosity in outbred human populations. *Human Molecular Genetics*, 15(5), 789–795. <https://doi.org/10.1093/hmg/ddi493>
- Goudet, J., Kay, T., & Weir, B. S. (2018). How to estimate kinship. *Molecular Ecology*, 27(20), 4121–4135. <https://doi.org/10.1111/mec.14833>
- Haller, B. C., & Messer, P. W. (2019). SLiM 3: Forward genetic simulations beyond the Wright–fisher model. *Molecular Biology and Evolution*, 36(3), 632–637. <https://doi.org/10.1093/molbev/msy228>
- Hildebrandt, F., Heeringa, S. F., Rüschenhoff, F., Attanasio, M., Nürnberg, G., Becker, C., Seelow, D., Huebner, N., Chernin, G., Vlangos, C. N., Zhou, W., O'Toole, J. F., Hoskins, B. E., Wolf, M. T. F., Hinkes, B. G., Chaib, H., Ashraf, S., Schoeb, D. S., Ovunc, B., ... Otto, E. A. (2009). A systematic approach to mapping recessive disease genes in individuals from outbred populations. *PLoS Genetics*, 5(1), e1000353. <https://doi.org/10.1371/journal.pgen.1000353>
- Hill, W. G., & Weir, B. S. (2011). Variation in actual relationship as a consequence of Mendelian sampling and linkage. *Genetics Research*, 93(1), 47–64. <https://doi.org/10.1017/S0016672310000480>
- Howrigan, D. P., Simonson, M. A., & Keller, M. C. (2011). Detecting autozygosity through runs of homozygosity: A comparison of three autozygosity detection algorithms. *BMC Genomics*, 12(1), 460. <https://doi.org/10.1186/1471-2164-12-460>
- Huisman, J. (2017). Pedigree reconstruction from SNP data: Parentage assignment, sibship clustering and beyond. *Molecular Ecology Resources*, 17(5), 1009–1024. <https://doi.org/10.1111/1755-0998.12665>
- Huisman, J., Kruuk, L. E. B., Ellis, P. A., Clutton-Brock, T., & Pemberton, J. M. (2016). Inbreeding depression across the lifespan in a wild mammal population. *Proceedings of the National Academy of Sciences*, 113(13), 3585–3590. <https://doi.org/10.1073/pnas.1518046113>
- Jones, O. R., & Wang, J. (2010). COLONY: A program for parentage and sibship inference from multilocus genotype data. *Molecular Ecology Resources*, 10(3), 551–555. <https://doi.org/10.1111/j.1755-0998.2009.02787.x>
- Kardos, M., Akesson, M., Fountain, T., Flagstad, O., Liberg, O., Olason, P., Sand, H., Wabakken, P., Wikenros, C., & Ellegren, H. (2018). Genomic consequences of intensive inbreeding in an isolated wolf population. *Nature Ecology & Evolution*, 2(1), 124–131. <https://doi.org/10.1038/s41559-017-0375-4>
- Kardos, M., Luikart, G., & Allendorf, F. W. (2015). Measuring individual inbreeding in the age of genomics: Marker-based measures are better than pedigrees. *Heredity*, 115(1), 63–72. <https://doi.org/10.1038/hdy.2015.17>
- Kariyat, R. R., & Stephenson, A. G. (2019). Inbreeding depression: It's not just for population biologists. *American Journal of Botany*, 106(3), 331–333. <https://doi.org/10.1002/ajb2.1256>
- Kelleher, J., Etheridge, A. M., & McVean, G. (2016). Efficient coalescent simulation and genealogical analysis for large sample sizes. *PLoS Computational Biology*, 12(5), e1004842. <https://doi.org/10.1371/journal.pcbi.1004842>
- Keller, L. F., & Waller, D. M. (2002). Inbreeding effects in wild populations. *Trends in Ecology & Evolution*, 17(5), 230–241. [https://doi.org/10.1016/S0169-5347\(02\)02489-8](https://doi.org/10.1016/S0169-5347(02)02489-8)
- Keller, M. C., Simonson, M. A., Ripke, S., Neale, B. M., Gejman, P. V., Howrigan, D. P., Lee, S. H., Lencz, T., Levinson, D. F., Sullivan, P. F., & The Schizophrenia Psychiatric Genome-Wide Association Study Consortium. (2012). Runs of homozygosity implicate Autozygosity as a schizophrenia risk factor. *PLoS Genetics*, 8(4), e1002656. <https://doi.org/10.1371/journal.pgen.1002656>
- Keller, M. C., Visscher, P. M., & Goddard, M. E. (2011). Quantification of inbreeding due to distant ancestors and its detection using dense single nucleotide polymorphism data. *Genetics*, 189(1), 237–249. <https://doi.org/10.1534/genetics.111.130922>
- Kim, E.-S., Cole, J. B., Huson, H., Wiggans, G. R., Van Tassell, C. P., Crooker, B. A., Liu, G., Da, Y., & Sonstegard, T. S. (2013). Effect of artificial selection on runs of homozygosity in U.S. Holstein cattle. *PLoS One*, 8(11), e80813. <https://doi.org/10.1371/journal.pone.0080813>
- Kirin, M., McQuillan, R., Franklin, C. S., Campbell, H., McKeigue, P. M., & Wilson, J. F. (2010). Genomic runs of homozygosity record population history and consanguinity. *PLoS One*, 5(11), e13996. <https://doi.org/10.1371/journal.pone.0013996>
- Lencz, T., Lambert, C., DeRosse, P., Burdick, K. E., Morgan, T. V., Kane, J. M., Kucherlapati, R., & Malhotra, A. K. (2007). Runs of homozygosity reveal highly penetrant recessive loci in schizophrenia. *Proceedings of the National Academy of Sciences*, 104(50), 19942–19947. <https://doi.org/10.1073/pnas.0710021104>
- Leutenegger, A.-L., Prum, B., Génin, E., Verny, C., Lemaître, A., Clerget-Darpoux, F., & Thompson, E. A. (2003). Estimation of the inbreeding coefficient through use of genomic data. *The American Journal of Human Genetics*, 73(3), 516–523. <https://doi.org/10.1086/378207>
- Liu, L., Bosse, M., Megens, H.-J., de Visser, M., Groenen, M. A. M., & Madsen, O. (2020). Genetic consequences of long-term small effective population size in the critically endangered pygmy hog. *Evolutionary Applications*, 14(3), 710–720. <https://doi.org/10.1111/eva.13150>
- Lynch, M., & Walsh, B. (1998). *Genetics and analysis of quantitative traits* (Vol. 1). Sinauer Associates.
- Mastrangelo, S., Tolone, M., Di Gerlando, R., Fontanesi, L., Sardina, M. T., & Portolano, B. (2016). Genomic inbreeding estimation in small populations: Evaluation of runs of homozygosity in three local dairy cattle breeds. *Animal*, 10(5), 746–754. <https://doi.org/10.1017/S1751731115002943>
- McQuillan, R., Leutenegger, A.-L., Abdel-Rahman, R., Franklin, C. S., Pericic, M., Barac-Lauc, L., Smolej-Narancic, N., Janicijevic, B., Polasek, O., Tenesa, A., MacLeod, A. K., Farrington, S. M., Rudan, P.,

- Hayward, C., Vitart, V., Rudan, I., Wild, S. H., Dunlop, M. G., Wright, A. F., ... Wilson, J. F. (2008). Run of homozygosity in European populations. *The American Journal of Human Genetics*, 83(3), 359–372. <https://doi.org/10.1016/j.ajhg.2008.08.007>
- Menges, E. S. (1991). Seed germination percentage increases with population size in a fragmented prairie species. *Conservation Biology*, 5(2), 158–164. <https://doi.org/10.1111/j.1523-1739.1991.tb00120.x>
- Meyermans, R., Gorssen, W., Buys, N., & Janssens, S. (2020). How to study runs of homozygosity using PLINK? A guide for analyzing medium density SNP data in livestock and pet species. *BMC Genomics*, 21(1), 94. <https://doi.org/10.1186/s12864-020-6463-x>
- Mueller, S. A., Prost, S., Anders, O., Breitenmoser-Würsten, C., Kleven, O., Klinga, P., Konec, M., Kopatz, A., Krojerová-Prokešová, J., Middelhoff, T. L., Obexer-Ruff, G., Reiners, T. E., Schmidt, K., Sindičič, M., Skrbinšek, T., Tām, B., Saveljev, A. P., Naranbaatar, G., & Nowak, C. (2022). Genome-wide diversity loss in reintroduced Eurasian lynx populations urges immediate conservation management. *Biological Conservation*, 266, 109442. <https://doi.org/10.1016/j.biocon.2021.109442>
- Nachman, M. W., & Crowell, S. L. (2000). Estimate of the mutation rate per nucleotide in humans. *Genetics*, 156(1), 297–304.
- Nalls, M. A., Guerreiro, R. J., Simon-Sanchez, J., Bras, J. T., Traynor, B. J., Gibbs, J. R., Launer, L., Hardy, J., & Singleton, A. B. (2009). Extended tracts of homozygosity identify novel candidate genes associated with late-onset Alzheimer's disease. *Neurogenetics*, 10(3), 183–190. <https://doi.org/10.1007/s10048-009-0182-4>
- Narasimhan, V., Danecek, P., Scally, A., Xue, Y., Tyler-Smith, C., & Durbin, R. (2016). BCFtools/RoH: A hidden Markov model approach for detecting autozygosity from next-generation sequencing data. *Bioinformatics*, 32(11), 1749–1751. <https://doi.org/10.1093/bioinformatics/btw044>
- Nietlisbach, P., Muff, S., Reid, J. M., Whitlock, M. C., & Keller, L. F. (2019). Nonequivalent lethal equivalents: Models and inbreeding metrics for unbiased estimation of inbreeding load. *Evolutionary Applications*, 12(2), 266–279. <https://doi.org/10.1111/eva.12713>
- Pemberton, T. J., Absher, D., Feldman, M. W., Myers, R. M., Rosenberg, N. A., & Li, J. Z. (2012). Genomic patterns of homozygosity in world-wide human populations. *The American Journal of Human Genetics*, 91(2), 275–292. <https://doi.org/10.1016/j.ajhg.2012.06.014>
- Peripolli, E., Munari, D. P., Silva, M. V. G. B., Lima, A. L. F., Irgang, R., & Baldi, F. (2017). Runs of homozygosity: Current knowledge and applications in livestock. *Animal Genetics*, 48(3), 255–271. <https://doi.org/10.1111/age.12526>
- Peripolli, E., Stafuzza, N. B., Munari, D. P., Lima, A. L. F., Irgang, R., Machado, M. A., Panetto, J. C., do, C., Ventura, R. V., Baldi, F., & da Silva, M. V. G. B. (2018). Assessment of runs of homozygosity islands and estimates of genomic inbreeding in Gyr (Bos indicus) dairy cattle. *BMC Genomics*, 19(1), 34. <https://doi.org/10.1186/s12864-017-4365-3>
- Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., Maller, J., Sklar, P., de Bakker, P. I. W., Daly, M. J., & Sham, P. C. (2007). PLINK: A tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics*, 81(3), 559–575. <https://doi.org/10.1086/519795>
- Qanbari, S., & Wittenburg, D. (2020). Male recombination map of the autosomal genome in German Holstein. *Genetics Selection Evolution*, 52(1), 73. <https://doi.org/10.1186/s12711-020-00593-z>
- Quinlan, A. R. (2014). BEDTools: The Swiss-Army tool for genome feature analysis. *Current Protocols in Bioinformatics*, 47(1), 11.12.1–11.12.34. <https://doi.org/10.1002/0471250953.bi1112s47>
- Saremi, N. F., Supple, M. A., Byrne, A., Cahill, J. A., Coutinho, L. L., Dalén, L., Figueiró, H. V., Johnson, W. E., Milne, H. J., O'Brien, S. J., O'Connell, B., Onorato, D. P., Riley, S. P. D., Sikich, J. A., Stahler, D. R., Vilella, P. M. S., Vollmers, C., Wayne, R. K., Eizirik, E., ... Shapiro, B. (2019). Puma genomes from north and South America provide insights into the genomic consequences of inbreeding. *Nature Communications*, 10(1), 4769. <https://doi.org/10.1038/s41467-019-12741-1>
- Sole, M., Gori, A.-S., Faux, P., Bertrand, A., Farnir, F., Gautier, M., & Druet, T. (2017). Age-based partitioning of individual genomic inbreeding levels in Belgian blue cattle. *Genetics Selection Evolution*, 49, 92. <https://doi.org/10.1186/s12711-017-0370-x>
- Speed, D., & Balding, D. J. (2015). Relatedness in the post-genomic era: Is it still useful? *Nature Reviews Genetics*, 16(1), 33–44. <https://doi.org/10.1038/nrg3821>
- Stoffel, M. A., Johnstone, S. E., Pilkington, J. G., & Pemberton, J. M. (2021). Genetic architecture and lifetime dynamics of inbreeding depression in a Wild mammal. *Nature Communications*, 12(1), 2972. <https://doi.org/10.1038/s41467-021-23222-9>
- Thompson, E. A. (2013). Identity by descent: Variation in meiosis, across genomes, and in populations. *Genetics*, 194(2), 301–326. <https://doi.org/10.1534/genetics.112.148825>
- Wang, J. (2016). Pedigrees or markers: Which are better in estimating relatedness and inbreeding coefficient? *Theoretical Population Biology*, 107, 4–13. <https://doi.org/10.1016/j.tpb.2015.08.006>
- Wang, S., Haynes, C., Barany, F., & Ott, J. (2009). Genome-wide autozygosity mapping in human populations. *Genetic Epidemiology*, 33(2), 172–180. <https://doi.org/10.1002/gepi.20344>
- Weir, B. S., & Goudet, J. (2017). A unified characterization of population structure and relatedness. *Genetics*, 206(4), 2085–2103. <https://doi.org/10.1534/genetics.116.198424>
- Wright, S. (1922). Coefficients of inbreeding and relationship. *The American Naturalist*, 56(645), 330–338. <https://doi.org/10.1086/279872>
- Yang, J., Lee, S. H., Goddard, M. E., & Visscher, P. M. (2011). GCTA: A tool for genome-wide complex trait analysis. *The American Journal of Human Genetics*, 88(1), 76–82. <https://doi.org/10.1016/j.ajhg.2010.11.011>
- Yengo, L., Zhu, Z., Wray, N. R., Weir, B. S., Yang, J., Robinson, M. R., & Visscher, P. M. (2017). Detection and quantification of inbreeding depression for complex traits from SNP data. *Proceedings of the National Academy of Sciences*, 114(32), 8602–8607. <https://doi.org/10.1073/pnas.1621096114>
- Zhang, C., Wang, P., Tang, D., Yang, Z., Lu, F., Qi, J., Tawari, N. R., Shang, Y., Li, C., & Huang, S. (2019). The genetic basis of inbreeding depression in potato. *Nature Genetics*, 51, 374–378. <https://doi.org/10.1038/s41588-018-0319-1>
- Zhang, Q., Calus, M. P. L., Gulbrandsen, B., Lund, M. S., & Sahana, G. (2015). Estimation of inbreeding using pedigree, 50k SNP chip genotypes and full sequence data in three cattle breeds. *BMC Genomics*, 16, 88. <https://doi.org/10.1186/s12863-015-0227-7>

## SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

**How to cite this article:** Lavanchy, E., & Goudet, J. (2023). Effect of reduced genomic representation on using runs of homozygosity for inbreeding characterization. *Molecular Ecology Resources*, 00, 1–16. <https://doi.org/10.1111/1755-0998.13755>