Genomic inbreeding coefficients using imputed genotypes: Assessing different estimators in Holstein-Friesian dairy cows

Christos Dadousis,1* Michela Ablondi,1 Claudio Cipolat-Gotet,1 Jan-Thijs van Kaam,2 Maurizio Marusi,2 Martino Cassandro,2,3 Alberto Sabbioni,1 and Andrea Summer1

1Department of Veterinary Science, University of Parma, 43126 Parma, Italy
2Associazione Nazionale Allevatori della Razza Frisona Bruna e Jersey Italiana, (ANAFIBJ), 26100 Cremona, Italy
3Department of Agronomy, Food, Natural Resources, Animals and Environment, University of Padova, 35020 Legnaro (PD), Italy

ABSTRACT

The objective of this study was to estimate inbreeding coefficients in Holstein dairy cattle using imputed SNPs data. A data set of 95,540 Italian Holstein dairy cows from the routine genomic evaluations of the Italian National Association of Holstein, Brown, and Jersey Breeders were analyzed, with 84,445 imputed SNP. Ten widely used genomic inbreeding estimators were tested, including 4 PLINK v1.9 estimators (F, FHAT1, FHAT2, FHAT3), 3 genomic relationship matrix (GRM)-based methods [VanRaden’s first method with observed allele frequencies (FGRM) or with fixed frequencies at 0.5 (FGRM05), VanRaden’s third method, allelic frequency free and pedigree regressed (FGRM2)], runs of homozygosity (ROH)-based estimators in a complete (FROH) and simplified version (FROH2), and proportion of homozygous SNP (FPH). Pairwise comparisons among them were made, including the comparison with traditional pedigree-based inbreeding coefficients (FPED). Our results showed variability among the genomic inbreeding estimators. Coefficients of FGRM and FHAT were >1, meaning that more variability has been lost than the variability that existed in the base population. Regarding the remaining ones, FGRM05, FROH, FROH2, and FPH provided coefficients within the [0,1] space and are considered comparable to FPED. Not comparable to FPED, yet with an interpretable value, can be considered the coefficients of F, FHAT2, and FGRM2. Estimators based on ROH had the highest correlation with pedigree-based coefficients (0.59–0.66), among all estimators tested. In this study, Spearman correlations were shown to possibly provide a clearer estimation of the strength of the relationship between estimators. We hypothesize that imputation might cause extreme genomic inbreeding values that deserves further investigation.

Key words: inbreeding, single nucleotide polymorphism, imputation, dairy cattle

INTRODUCTION

Modern dairy cattle industry has greatly advanced the last century using reproductive technologies, mainly due to the extensive use of AI. In the United States the set-up of AI centers started in the 1940s, whereas the global potential of this technology became clear in the 1970s after an immense cattle comparative trial organized at an international level by the Food and Agriculture Organization (Weigel et al., 2017). This stimulated the use of sire semen and distribution to dairy cattle farmers worldwide. Dairy cattle breeding programs, especially in Holstein-Friesian breed, used intense directional selection. This means that only few bulls from the entire population pool were selected to be sires for the next generations and were available in the AI market. As a consequence, at present, 99% of the Holstein bulls from the United States in AI stations can be traced back to only 2 bulls in the 1960s (Chief and Elevation), which result from 2 bulls born in the 1880s (Hulleman and Neptune H; Yue et al., 2015; Dechow et al., 2020). Similar situation has been recently reported for another widely used dairy cattle breed (the North American Jersey), with one bull (Champion Flying Fox) born in 1898 being the common ancestor of 5 bulls accounting for 100% of the North American sires born after 2010 (Dechow et al., 2020). Similar situation has been recently reported for another widely used dairy cattle breed (the North American Jersey), with one bull (Champion Flying Fox) born in 1898 being the common ancestor of 5 bulls accounting for 100% of the North American sires born after 2010 (Dechow et al., 2018). It must be clarified that these studies reported findings by tracing the Y chromosome. As reported in Adams et al. (2016), Chief, that is one of the top 25 influential sires in the US Holstein breed, estimated to share ~14.3% of the autosomal alleles to the entire breed (that was the maximum proportion found among the 4 sires studied). Unavoidably, this co-ancestry corroborates an increased inbreeding level in the population per year and generation. High levels of inbreeding, in turn, result in reduction of fertility (Falconer and Mackay, 1996) and
production characteristics in livestock species (Leroy, 2014; Baes et al., 2019; Doekes et al., 2019).

In livestock, inbreeding is practically observed when loops appear in pedigree and the parents of an individual are genetically related to a common ancestor. Inbreeding coefficient \( f \) is a measure of the level of inbreeding of an individual and is defined either as a correlation between homologous alleles of the two gametes that unite to form the individual (Wright, 1921, 1922) or as the probability of two homologous alleles at a given neutral locus (unlinked to any potential locus under selection) being identical-by-descent (IBD: Malécot, 1948). Historically, Wright justified this definition for allowing the prediction of heterozygote proportions in crosses and inbred lines, Malécot’s definition was easier to apply to single loci, whereas later on it was realized that the general view of Wright on relationship and inbreeding directly relates to the best linear unbiased prediction developed by Henderson (1973) and Hill (1989). In both cases, \( f \) is relative to a base population. Hence, although in many livestock populations a group of (unrelated) founders might establish the base population and IBD can be well defined, strictly speaking there is not a concrete and absolute IBD measure (Thompson, 2013). The concept of IBD in time depth has been reviewed by Thompson (2013) in human populations, but it is out of the scope of this work. In addition to the 2 previous definitions, \( f \) quantifies the reduction of heterozygosity as a result of inbreeding (Crow and Kimura, 1970). Traditionally, \( f \) is estimated via pedigree data (\( F_{\text{PED}} \)). Practically, \( F_{\text{PED}} \) is a direct function of the expected relatedness of the parents; it equals to the degree of relatedness of the parents, meaning that if parents are unrelated in pedigree \( f = 0 \). It should be noted that still an individual can have \( f > 0 \) even though both parents have \( f > 0 \), yet the parents are unrelated to each other. In livestock species, the inbreeding level is an important characteristic at an animal, herd, and population level. Moreover, apart from inbreeding depression, inbreeding is strongly related to the concepts of genetic progress, genetic diversity, mating schemes, number of selected parents and conservation in livestock.

Despite the wide use of pedigree in livestock populations, the main limitation of pedigree records is that they may contain errors and for an accurate estimate of inbreeding, pedigree quality and depth are key aspects. At the early stage of dairy cattle genomic evaluations, Wiggans et al., (2012) reported that ~14% of cows were found to have sire conflicts, whereas in an updated study a higher value of 22% was found (Nani et al., 2020). However, an incomplete pedigree might result in underestimating \( F_{\text{PED}} \) (Lutaaya et al., 1999; Cassell et al., 2003). A second criticism on pedigree-based inbreeding coefficients is that we calculate expectations based on theoretical background. This variation around the expected value is present mainly due to the result of the random process of meiotic recombination, known as Mendelian sampling (Franklin, 1977; Hill and Weir, 2011). Moreover, other evolutionary forces (e.g., mutation and drift) might contribute to further increase the variation around the expected value. Hence, inbreeding expectations differ from the realized inbreeding at genome level with a different proportion of a genome shared between individuals compared with what is estimated at pedigree level. The realized genomic similarity at a molecular level can be estimated with genomic data. However, although the wide use of genomic data, mainly in the form of SNP's nowadays, is expected to provide a more precise assessment, a series of new challenges appeared and have to be addressed. In brief, the new challenges could be summarized below:

- **Genotyping platforms**: A plethora of SNP chips is available varying on the quantity of SNP contained in the chip as well as the quality. For instance, some SNP panels have been developed by companies to fit the requirements of various breeds, whereas others have been developed to fit within a specific breed and the requirements of specific breeding associations and AI centers. Moreover, there are SNP panels that contain SNP targeting specific genes, that are associated with various economic important traits (e.g., milk yield and casein variants) and diseases (e.g., Bovine leukocyte adhesion deficiency). Even within a dairy cattle breeding scheme, it is very common that different cattle groups are genotyped with different SNP chips.

- **Genotyping errors and non-IBD homozygosity**: SNP data might also contain sampling errors, such as errors from the laboratory procedure. Moreover, homozygosity might exist but without necessarily reflecting IBD state.

- **SNP quality control (QC)**: It is very common to filter SNP data before any kind of genomic analysis. The most common filters in QC are the percentage of missing data per individual and SNP, parent-offspring SNP mismatch, Hardy-Weinberg deviation, minor allele frequency, and linkage disequilibrium thresholds. However, no golden standard exists; hence, different QC pipelines might differ in the number of filters selected as well as the thresholds applied in each filter. Although efforts have been recently made to quantify the effects of some of QC-filtering on genomic analysis (Meyermans et al., 2020) to a great extent, a
concrete background quantifying the effect of each of these filtering as well as their combination is lacking.

- **Imputation:** Nowadays, for routine genomic evaluations in dairy cattle breeding programs, imputed rather than the actual genotype data are used. Imputation might further include (1) a large number of SNP chip used to genotype various groups of animals in time, (2) various algorithms and pipelines (Whalen et al., 2018, 2019), and (3) both chip- and sequence-based SNP data. As a result, imputation accuracy is another parameter that should be considered in estimating inbreeding coefficients based on SNP data.

- **Genomic inbreeding methods:** In general, the existing methods to estimate genomic inbreeding coefficients could be grouped as measuring average over independent single-point estimates (i.e., summing homozygosity at individual SNP) and analyzing bigger consecutive parts of the DNA strand of a predefined minimum length, widely known as runs of homozygosity (ROH). Some methods, however, such as ROH-based inbreeding coefficients (F_{ROH}) require additional parameters to be defined before the analysis. Moreover, there are different methods to estimate F_{ROH}, for example, consecutive runs (Marras et al., 2015) versus sliding windows (Purcell et al., 2007; Bjelland et al., 2013). A more comprehensive view of ROH can be found in several studies (Curik et al., 2014; Nandolo et al., 2018; Meyermans et al., 2020). Hence, different algorithms and different set-up parameters within ROH estimators add another level of complexity that should be considered and its effect to be quantified.

- **Distinguishing autozygosity from homozygosity:** An IBD homozygosity at the DNA level is referred to autozygosity (homozygosity by descent; HBD). Methods proposed could be summarized in 2 main categories: (1) ROH-based methods, such as splitting ROH into classes of different genome length (Kirin et al., 2010; Pemberton et al., 2012; Marras et al., 2015) or age (Doeoes et al., 2019), and (2) HBD (Druet and Gautier, 2017) or age-based HBD (Solé et al., 2017). Various implementations of these methods have been proposed based on arbitrary thresholds (Doeoes et al., 2019), clustering approaches (Pemberton et al., 2012), hidden Markov models, and grid search algorithms (Sumreddie et al., 2021). The general objective though remains the same, that is to find a way to distinguish recent from old inbreeding or HBD from non-HBD DNA segments. In this concept, ROH-based inbreeding coefficient largely stands on the assumption that autozygosity perfectly reflects inbreeding, although this is not always the case, as some proportion of autozygosity may be due to selection.

- **Estimation of genomic inbreeding coefficients for ungenotyped animals:** Recently Legarra et al. (2020) proposed 3 methods for estimating genomic inbreeding coefficients for ungenotyped animals (linear projection of genomic inbreeding coefficients via pedigree inbreeding coefficients). These methods build on previous works on the single step genomic evaluations (Christensen and Lund, 2010; Legarra et al., 2020).

- **Given the above, in routine dairy cattle genomic breeding programs, genomic inbreeding might be derived from different methods using genomic relationship matrices (GRM) generated for genomic evaluations. Estimation of GRM-based genomic inbreeding and its effect in controlling the level of inbreeding in livestock has been discussed in recent works (Gebregiwerigis et al., 2020; Villanueva et al., 2021). However, the effect of this discrepancy (i.e., different genomic inbreeding estimators used in the genomic evaluations and controlling inbreeding) during the selection process of breeding candidates in breeding programs might be questionable.**

Although it has been more than a decade that SNP data have been routinely used for genomic evaluations in dairy cattle and other livestock species, knowledge around genomic inbreeding is still not fully explored and to a certain degree, it can be considered a challenging topic. From a practical point of view, a consistent, robust, and simple measure of genomic inbreeding should be available for AI centers, breeding associations and farmers for selection purposes and organizing matings. Hence, there are still compelling reasons to study genomic inbreeding. Indeed, the effect of different genomic inbreeding estimators in controlling co-ancestry and genetic diversity in livestock populations has been recently investigated (Gebregiwerigis et al., 2020; Maltecca et al., 2020; Meuwissen et al., 2020), as well the effect of genomic inbreeding on various economic traits (Howard et al., 2017; Baes et al., 2019). However, research on genomic inbreeding, either on theoretical or practical background, is mainly focused on actual SNP genotypes rather than imputed SNP data. For many genomic breeding programs, the latter is what is practically used. Moreover, despite the various genomic inbreeding estimators explored in scientific literature, which is difficult to review, there is still no consensus on which estimator is appropriate (Goudet et al., 2018; Alemu et al., 2021; Villanueva et al., 2021).
That said, the present work addresses genomic inbreeding from a practical perspective in dairy cattle breeding programs where imputed data are analyzed in the routine genomic evaluations; hence expanding the debate on genomic inbreeding estimators (Villanueva et al., 2021). In brief, 10 of the most well-known and practically applied genomic inbreeding estimators were tested; 4 estimates provided in the PLINK v.1.9 software (Purcell et al., 2007; http://pngu.mgh.harvard.edu/purcell/plink/), 3 estimates derived by constructing GRM, 2 F_{ROH}, and a simplified version of measuring homozygosity. All pairwise comparisons, including F_{PED} were made.

In literature, inbreeding coefficient is abbreviated either as $f$ or $F$. To avoid misperception with $F$ genomic inbreeding estimator of PLINK v.1.9 software, we opted herein as $f''$ to denote inbreeding coefficient. Moreover, the term genomic inbreeding hereafter describes estimates of genomic inbreeding coefficients based on SNP (imputed) data.

**MATERIALS AND METHODS**

No animals were used in this study, and ethical approval for the use of animals was thus deemed unnecessary.

**Animals and Genotypes**

In total, 95,540 Italian Holstein dairy cows from the Italian National Association of Holstein, Brown and Jersey Breeders (ANAFIBJ) were analyzed. Cows were born between 1998 and 2020 (Figure 1a) and genotyped with 30 different SNP chips of varying densities (from 3K to 777K). The ratio of sires to dams on the total population per year and generation followed a constant pattern as shown in Figure 1b. All genotypes were imputed to a set of 84,445 carefully preselected SNP that are used in the routine genomic evaluations of ANAFIBJ. The imputation was performed with a modified version of PedImpute (Nicolazzi et al., 2013). After imputation, genotypes were set to 0,1,2 for the SNP of $k$ individuals. In PLINK’s flag –ibc, 3 estimates derived by constructing GRM, 2 F_{ROH}, and a simplified version of measuring homozygosity. All pairwise comparisons, including F_{PED} were made.

Inbreeding Coefficients

Pedigree-Based Coefficients. To estimate $F_{PED}$, 393,607 cattle were used. Analysis was carried out in the _pedigree_ R package (Coster, 2013; R Core Team, 2013), which estimates traditional $f$ without considering genetic groups, assigning for missing ancestors $f = 0$. Moreover, to examine the quality of the pedigree, the widely used parameters of the equivalent, maximum, and complete generations (Maingel et al., 1996; Kadlečík et al., 2016), as well as the pedigree completeness index (MacCluer et al., 1983), were estimated in the R package _optiSel_ (Wellmann, 2021).

Genomic-Based Coefficients. In this work, we investigated genomic inbreeding coefficients based on 10 widely used methods or estimators. A detailed description is presented in Table 1. For simplicity and to improve readability, the estimators were grouped in 3 categories. Moreover, the original notation that appear in software was adopted. Starting with the 4 estimators available in PLINK v1.9 (Purcell et al., 2007), the 10 genomic inbreeding estimators were as follows.

PLINK Estimators and Simplified Homozygosity. The first genomic inbreeding estimator we considered ($F$) is implemented in PLINK v1.9 (Purcell et al., 2007). First proposed by Li and Horvitz (1953), it counts the proportion of SNP in homozygous state from the formula $O - E\over n - E$, where $O$ is the observed and $E$ the expected number of homozygous SNP under the Hardy-Weinberg principle. The estimator is based on SNP-by-SNP analysis and allele frequencies of a analyzed SNP. It can be obtained using the flag –het.

The estimator $F_{HAT1}$ (primarily developed in the GCTA software ($\hat{F}^L$; Yang et al., 2011) is the usual variance-standardized relationship minus 1, calculated as $1 - \frac{1\sum_{m=1}^{n}\left(X_{m} - 2p_{m}\right)^{2}}{2p_{m}q_{m}}$, where $X$ is the matrix ($n \times m$) of the genotypes coded based on the number of copies of the defined reference allele, $p$ and $q$ are the frequencies of the reference and alternative alleles, respectively, summarized across $m$ SNP of $k$ individuals. In PLINK v1.9 it is implemented via the flag –ibc.

The estimator $F_{HAT2}$ is analogous to the F estimator measuring the excess of homozygosity $\left\{1 - \frac{1\sum_{m=1}^{n}\left(X_{m} - 2X_{m}\right)}{2p_{m}q_{m}}\right\}$. The difference between $F$ and $F_{HAT2}$ is that $F$ is a ratio of sums, whereas $F_{HAT2}$ is a sum of ratios (Gazal et al., 2014). The estimator $F_{HAT2}$ was initially denoted as $\tilde{F}^{ll}$ (Yang et al., 2011), and was based on the second method of VanRaden (2008) to construct GRM.

The third genomic inbreeding estimator implemented in PLINK’s flag –ibc is $F_{HAT3}$ (first denoted by Yang et al., 2011, as $\tilde{F}^{ll}$) is the actual definition of Wright on inbreeding (correlation between uniting gametes; Wright, 1921, 1922; Yang et al., 2010), estimated as
The estimators $F_{\text{HAT1-3}}$ have been primarily developed in the GCTA software (Yang et al., 2011); they all represent a SNP-by-SNP analysis and can be obtained simultaneously with the flag --ibc in PLINK. The estimator $F_{\text{PH}}$ is a simplification of $F$, measuring percentage of SNP-based homozygosity.

Genomic Relationship Matrices-Based Estimators. The second group of methods include 3 GRM-based inbreeding coefficients. As a baseline, the first method of VanRaden (2008) was used (Leutenegger et al., 2003; Amin et al., 2007; VanRaden, 2008) following

$$ GRM = \frac{\mathbf{Z}^\prime \mathbf{Z}}{2 \sum p_i (1 - p_i)} $$

where $\mathbf{Z} = \mathbf{X} - 2p_i$ and $p$ is the frequency of the second allele (coded as 2 in $\mathbf{X}$). Follow-

Figure 1. (A) Number of genotyped cows per year of birth. (B) Number of dams and sires on the total population per generation (left) and year (right).
Table 1. Overview of the pedigree and genomic inbreeding estimators analyzed

<table>
<thead>
<tr>
<th>Method</th>
<th>Type</th>
<th>Software</th>
<th>Allele frequency depended</th>
<th>Original name(^1)</th>
<th>Source(^2)</th>
<th>Nomenclature(^3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pedigree</td>
<td>Pedigree</td>
<td>pedigree R</td>
<td>/</td>
<td>f</td>
<td>Wright, 1922</td>
<td></td>
</tr>
<tr>
<td>Genomic</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PLINK(^4)</td>
<td>SNP-by-SNP</td>
<td>PLINK v1.91 (–het)</td>
<td>NO</td>
<td>f</td>
<td>Li and Horvitz, 1953</td>
<td>F(_{ROM}) (Yengo et al., 2017; Alemu et al., 2021); F(_1) (Nietlisbach et al., 2019)</td>
</tr>
<tr>
<td>F(_{FHAT1})</td>
<td></td>
<td>PLINK v1.91 (–ibc)</td>
<td>YES</td>
<td>F(_I)</td>
<td>Yang et al., 2011</td>
<td>FHOM (Yengo et al., 2017; Alemu et al., 2021); FH (Nietlisbach et al., 2019)</td>
</tr>
<tr>
<td>F(_{FHAT2})</td>
<td></td>
<td></td>
<td>YES</td>
<td>F(_{II})</td>
<td></td>
<td>F(<em>{GCTA1}) (Gazal et al., 2014); F(</em>{GRM}) (Zhang et al., 2015; Mastrangelo et al., 2016); F(<em>{GRM2}) (Solé et al., 2017) VR2 (Morales-González et al., 2020) F(</em>{GRM3}) (Villanueva et al., 2021)</td>
</tr>
<tr>
<td>F(_{FHAT3})</td>
<td></td>
<td></td>
<td>YES</td>
<td>F(_{III})</td>
<td></td>
<td>F(<em>{GCTA3}) (Gazal et al., 2014); FGRM (Bérénos et al., 2016; Huisman et al., 2016); F(</em>{GRM2}) (Zhang et al., 2015; Solé et al., 2017; Yengo et al., 2017; Santiago et al., 2020; Alemu et al., 2021); SNP-Yang (Eynard et al., 2016); ΥAN (Morales-González et al., 2020); F(_{YAN}) (Villanueva et al., 2021)</td>
</tr>
<tr>
<td>GRM(^6)</td>
<td>SNP-by-SNP</td>
<td>In-house python script</td>
<td>YES</td>
<td>First method</td>
<td>VanRaden, 2008</td>
<td></td>
</tr>
<tr>
<td>F(_{GRM})</td>
<td></td>
<td>ANAFIBJ: editSNPimp.f90</td>
<td>YES</td>
<td>First method</td>
<td></td>
<td></td>
</tr>
<tr>
<td>F(_{GRM05})</td>
<td></td>
<td></td>
<td>NO</td>
<td>Third method</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ROH(^7)</td>
<td>Continuous homozygous SNP blocks</td>
<td>detectRUNS v. 0.9.5 R function: consecutiveRUNS.run</td>
<td>NO</td>
<td>F(_{ROH})</td>
<td>McQuillan et al., 2008; Marras et al., 2015; McQuillan et al., 2008; Purfield et al., 2012</td>
<td></td>
</tr>
<tr>
<td>F(_{ROH2})</td>
<td></td>
<td>ANAFIBJ: editSNPimp.f90</td>
<td>NO</td>
<td>F(_{ROH})</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\(^1\)Original name of inbreeding coefficient estimator as first appeared in literature.
\(^2\)Source of original name. In the case of 2 sources the first is related to the first appearance in literature and the second to the modification of the method.
\(^3\)Representative nomenclature of the estimator that can be found in literature.
\(^4\)Genomic inbreeding estimates based on PLINK v1.9 software.
\(^5\)Genomic inbreeding estimates based on commonly used genomic relationship matrices (GRM). F\(_{GRM}\) and F\(_{GRM05}\) are both referring to the first method proposed by VanRaden (2008), with the former using observed and the latter fixed at 0.5 allele frequencies.
\(^6\)ANAFIBJ: Italian National Association of Holstein, Brown and Jersey Breeders.
\(^7\)Genomic inbreeding estimates based on runs of homozygosity (ROH).
\(^8\)F\(_{PH}\) is a simplification of F, measuring percentage of SNP-based homozygosity, developed by ANAFIBJ.
ing from this, \( F_{GRM} = diag(GRM) - 1 \). We note that we have used observed allele frequencies, rather than allele frequencies calculated in the founders. To address this limitation, 2 modified versions of \( F_{GRM} \) were included, with the first setting all allele frequencies at 0.5 \( (F_{GRM05}) \), and the second avoiding estimations on allele frequencies by first regressing the diagonal of \( XX' \) on \( f \) to get the mean and the slope and then obtain \( F_{GRM2} = \frac{diag(XX') - mean}{slope} \) (third method, VanRaden, 2008), where \( X \) is the matrix of the genotypes, and mean and slope are the estimates of the previous regression. We note that for \( F_{GRM} \) and \( F_{GRM2} \) analysis was carried out without considering a base population or AI sire information to calibrate the diagonal elements of \( XX' \), as has been reported in Rolf et al. (2010), a scenario that is better described by \( F_{GRM05} \).

**Runs of Homozygosity-Based Estimators.** A ROH-based inbreeding coefficient refers to the proportion of the autosomal genome of an individual being in a ROH and is generally calculated as the sum of ROH identified in an individual divided by the total genome length. Two versions of ROH-based inbreeding coefficients were tested. The first version, \( F_{ROH} \), was estimated with the consecutive runs method in the R (v. 3.6.3) package detectRUNS v. 0.9.5 (https://r-project.org; Marras et al., 2015; Biscarini et al., 2019). Prior parameters to define a ROH included a minimum of 15 SNPs/ROH and 1 Mbp length of ROH. To account for possible genotyping errors one heterozygous SNP within an ROH was allowed. The second version, \( F_{ROH2} \), was based on the method presented by Purfield et al. (2012), which is a simplified version of detecting ROH of the method reported by McQuillan et al. (2008) of 27 homozygous SNPs with a maximum of 3 nonfilled SNPs and no heterozygous SNPs.

The estimators of \( F_{GRM05}, F_{ROH2}, \) and \( F_{PH} \) were obtained from the official genomic evaluations of ANAFIBJ. It should be mentioned that the described grouping of the methods is not strict, because similarities exist among methods, and was simply made to distinguish among estimators used. For example, \( F, F_{HAT2}, \) and \( F_{PH} \) measure excess of SNP homozygosity, and \( F_{HAT3} \) is also derived by the diagonal of a GRM proposed by Yang et al., 2011. Moreover, \( F_{HAT1} \) is the second method reported in VanRaden (2008), whereas ROH analysis can also be considered as a GRM approach (de Cara et al., 2013; Luan et al., 2014; Rodríguez-Ramilo et al., 2015). Unfortunately, different nomenclature exists in literature (a representative list of which is presented in Table 1), which makes comparison among studies difficult. We recognize that our effort to identify the various notations under which each estimator is presented in literature is limited and cannot be considered complete. A recent and more concrete overview of the various notations used for GRM, PLINK, and GCTA software genomic inbreeding estimators in the literature can be found in the work of Villanueva et al. (2021).

**Comparison Among Methods**

Summary statistics and inspection of distributions were made for all measures. All pairwise comparisons of the 11 measures of inbreeding coefficients were made. Pearson (\( \rho \)) and Spearman (\( r \)) correlation coefficients were used to evaluate the strength of relationship among estimators. The latter provides with a better description of the relationship between 2 variables in the case of nonlinear but monotonic functions. As a baseline, \( \rho \) coefficients were estimated and x-y plots were inspected. Moreover, a principal component analysis (PCA) was conducted on the 11 measures of inbreeding coefficients using the R function `prcomp` (R Core Team, 2013) to investigate relationships among estimators. The PCA transforms the original data into a set of mutual orthogonal vectors (principal components; PC), ordered by decreasing variability, each of which is a linear combination of all the original variables.

**RESULTS**

**Descriptive Statistics of the Inbreeding Coefficients**

Summary statistics of pedigree and genomic inbreeding coefficients are summarized in Table 2 and Figure 2. The \( F_{PED} \) varied between 0 and 0.335 with mean and median coefficients of ~0.07. Regarding PLINK estimators, \( F \) and \( F_{HAT2} \) had both a maximum of 0.963, but with a minimum at −0.382 and −5.743 (\( F \) and \( F_{HAT2} \), respectively). However, \( F_{HAT1} \) and \( F_{HAT3} \) showed extreme positive coefficients up to 117.906 and 58.379, respectively, whereas the minimum was similar and close to that of \( F \) (−0.304 and −0.265, respectively). The GRM-based coefficients were in the space of [−1.123, 1.065], [0, 0.968] and [−2.850, 0.459] for \( F_{GRM}, F_{GRM05}, \) and \( F_{GRM2} \). Results were similar for the 2 ROH-based genomic inbreeding estimators, with inbreeding coefficients ranging between 0.005 and 0.942 for \( F_{ROH} \) and 0.002 to 0.846 for \( F_{ROH2} \). The \( F_{PH} \) measure of genomic homozygosity had the lowest range of coefficients [0.487, 0.984].

The distributions of the inbreeding coefficients among the 11 inbreeding estimators greatly varied (Table 1, diagonal of Figure 3a). \( F_{HAT1} \) and \( F_{HAT3} \) were right skewed (skewness of 256 and 281, respectively), long tailed and with a sharp peak (kurtosis of 74.399
and 84052, respectively); FHAT2 distribution was similar but left skewed (skewness = −6.6, kurtosis = 155). Less sharp but right skewed were also the distributions of FPED, FGRM05, FROH, FROH and FPH. Interestingly, a submixture of distributions was found for FGRM and FGRM2, which was more profound for the latter. For FGRM the 2 subdistributions had a peak at −1 and 0, whereas the peaks of FGRM2 were roughly at −1.5, −1.0, −0.5, 0, and 0.5, respectively. Moreover, for 766 cows FPED = 0 but $f_g$ varied between −5.743 and 6.0 (for FHAT2 and FHAT1, respectively).

Correlations Among Different Measures of Inbreeding Coefficients

All pairwise comparisons (scatterplots, $\rho$ and $r$ estimates) and the distributions of FPED and the 10 genomic inbreeding estimators were summarized in Figure 3. To avoid drawing wrong conclusions regarding pairwise $\rho$ and $r$ estimates of FHAT1 with the rest of the 10 inbreeding estimators, we excluded the extreme and influential value of 117.906 found for one cow. To illustrate this, by excluding the extreme value, the $\rho_{F_{PED}, FHAT1}$ was 0.428, whereas when considering that value, the $\rho_{F_{PED}, FHAT1}$ = 0.946, implying that the latter strong relationship was an artifact of this highly influential value (Supplemental Figure S1, https://figshare.com/s/386c45ee3d2ed797bf0). Overall, FPED was correlated more strongly with ROH-based estimators (0.655 and 0.592 for FROH2 and FROH, respectively), with the lowest $\rho$ found for FHAT1 (−0.287) and lowest in absolute values for FGRM2 (0.045). Regarding $\rho$ of FPED with the rest of $f_g$ estimators, $\rho = 0.46$ was found between FPED and F, FPH, and FGRM05, and slightly lower (0.44) for FHAT2. Close to zero correlations with FPED were observed for FHAT1, FHAT3, FGRM, and FGRM2. Correlations between FPED and all genomic inbreeding estimators were also calculated for the top 5% of FPED (Supplemental Figure S2, https://figshare.com/s/386c45ee3d2ed797bf0). In most of the cases, $\rho$ increased up to ~0.60 to 0.66. Only for FHAT1 and FGRM2 $\rho$ was close to 0 (−0.09 and −0.13, respectively), whereas for FGRM $\rho$ was increased to 0.46.

Considering all pairwise correlations between estimators, high $\rho$ (>0.9) was found between F and FPH, FGRM5, and FROH; FPH and FGRM5; $F_{ROH}$ with FPH and FGRM5. Notably, a perfect correlation ($\rho = 1$) was found between FGRM5 and FPH. Intermediate correlations (0.5 ≤ $\rho$ ≤ 0.9) were found between F and FHAT2 and FROH2; FHAT2 with FGRM5; FROH, FROH2, and FPH; FPED with FROH and FROH2; FGRM5; FROH2; and FPH. All the rest pairwise $\rho$ (with the exception of FGRM-FGRM2) were in the range...
−0.29 ≤ ρ ≤ 0.5. The only high negative correlation was found between \( F_{GRM} \) and \( F_{GRM2} \) (−0.89).

Pairwise scatterplots (lower triangle of Figure 3a) depicted also nonlinear relationships, some of which with a monotonic pattern. Spearman pairwise correlations provide a better description of the strength of relationship of monotonic functions (Figure 3b). Compared with ρ, r of \( F_{PED} \) with all genomic inbreeding estimators was higher, except for \( F_{GRM} \) that slightly decreased but remaining close to zero. The highest r with \( F_{PED} \) was again observed for \( F_{ROH2} \) and \( F_{ROH} \) but also for \( F_{HAT2} \) (0.686, 0.673, 0.672, respectively), followed by \( F \) (0.641), \( F_{GRM05} \) and \( F_{PH} \) (−0.630), \( F_{HAT1} \) (−0.437), \( F_{HAT3} \) (0.285) and \( F_{GRM} \) (−0.213), and \( F_{GRM2} \) (−0.008). Regarding the rest pairwise r estimates of genomic inbreeding estimators, compared with ρ the general strength of relationship was mainly higher. However, there were cases were \( r \approx ρ \), either low (for e.g., \( F-F_{GRM} \), \( F-F_{GRM2} \)) or high (\( F_{PH} \), \( F_{PH}-F_{PH} \)). In few cases \( r < ρ \) was observed (for e.g., \( F_{HAT3} \) with \( F \), \( F_{HAT1} \), and \( F_{HAT2} \)).

### Similarities Among Inbreeding Estimators Based on Principal Components Analysis

Results of PCA regarding the similarities of pedigree and genomic inbreeding estimators were summarized in Figure 4 (scatterplot of the loadings of pedigree and genomic inbreeding estimators on the first 2 PC). The first 2 PC cumulatively captured ~70.5% of the total variability (55.3 and 20.1% for PC1 and PC2, respectively). Principal component 1 distinguished \( F_{HAT1} \), \( F_{GRM} \), and \( F_{GRM2} \) from the rest. Further, PC2 distinguished \( F_{HAT1} \) and \( F_{GRM} \) from \( F_{GRM2} \) and grouped more closely \( F_{PED}-F_{HAT2} \), \( F_{ROH}-F_{ROH2} \), and \( F \) with \( F_{GRM05} \) and \( F_{PH} \). Moreover, \( F_{HAT3} \) was positioned between the 2 groups of \( F_{HAT1}-F_{GRM} \) and \( F_{GRM05}-F_{PH} \). Adding the third PC, the total amount of the original variability was 90%, and 6 PC explained cumulatively 99% of the total variability.

### DISCUSSION

Inbreeding is a widely used concept to characterize evolution, diversity, and the general structure of a population. Moreover, in dairy cattle breeding and herds, individual and average f are considered for organizing future matings and management. Inbreeding can be studied on different levels, starting from the individual, herd, and up to population level (Ablondi et al., 2022; Howard et al., 2017) or even within consortium [e.g., AI or dairy plant (Ablondi et al., 2021)]. This knowledge helps for a better management of diversity and the genetic progress, controlling for inbreeding depression and, in small populations, for conservation purposes. Although traditionally inbreeding was calculated based on pedigree data, the progress made by the scientific and technological community in the last 2 decades advanced the estimation of inbreeding from SNP data.

With dense SNP data, the expected values of inbreeding are replaced with realized measurements of homozygosity, which are thought to be a more precise way to assess inbreeding and better reflect the homozygosity level. However, it is important to keep in mind that homozygosity might be caused by either common ancestors (autozygosity) or by the aforementioned evolutionary processes. In the latter case, homozygosity represents an identical-by-state situation termed C. An inbreeding coefficient should be able to quantify the degree of genome-wide autozygosity, as opposed to allozygosity. In practice, this is not straightforward. To distinguish this information, classes of \( F_{ROH} \) based on different genome length in human (McQuillan et al., 2008; Kirin et al., 2010; Pemberton et al., 2012) and livestock populations (Schiavo et al., 2020; Ablondi et al., 2021; Dadousis et al., 2021b) or based on age (Doekes et al., 2019) have been previously reported. Moreover, HBD has been proposed as a good candidate for studying genomic inbreeding in livestock (Druet and Gautier, 2017; Fabbri et al., 2021; Lozada-Soto et al., 2021). Although these approaches are useful to study inbreeding within population and in time (helping in understanding the effect of mating schemes in inbreeding, evolutionary history, and level of genetic diversity of the population under study), our work was motivated from a practical and simplistic perspective in individual inbreeding coefficients. In spite of this, the possibility of further identifying non-HBD individuals, as recently reported in Fabbri et al. (2021), shall be considered as an attractive and complementary approach to the inbreeding coefficients derived from the 11 estimators analyzed in this study (IBS-based estimators), which could be combined together to better organize assortative matings and control inbreeding at a population level.

Various ways on how to control inbreeding in livestock populations based on pedigree or genomic information are well summarized in the recent symposium review of Maltecca et al. (2020). In practice, there are several ways to estimate genomic inbreeding, which unfortunately provide with different results. For example, some estimators of genomic inbreeding provide with coefficients out of −1 to 1 [if inbreeding is interpreted as a correlation measure (Wright, 1921, 1922)] or the 0–1 [if inbreeding is seen as a probability (Malécot, 1948) and the ROH-based genomic inbreeding coefficients] space, making hard to interpret and compare with traditional
Even in the case of inbreeding interpreted as loss or gain of variability (i.e., loss of heterozygosity) based on a base population, values of genomic inbreeding >1 might be observed; however, they are unreasonable for a population under intense and directional (Villanueva et al., 2021). Moreover, negative correlations between pairs of estimators might exist, meaning that an individual might have a high inbreeding coefficient based on one estimator and a low coefficient using another estimator. Strangely enough, although genomic breeding programs are routinely applied in some of the most famous dairy cattle breeds worldwide and genomic breeding values are supplied to breeders and farmers, a concrete and absolute measure useful as genomic inbreeding coefficient is still lacking. Hence, together with the opportunities of using genomic data new challenges also appear that must be addressed.

It is of interest that we still lack criteria to choose the optimal genomic inbreeding estimator. To address this, the topic of genomic inbreeding has recently challenged the scientific community to investigate in depth the various ways that inbreeding can be estimated from SNP data (Meuwissen et al., 2020; Caballero et al., 2021; Villanueva et al., 2021) and the effect of this information in managing genetic diversity and genetic progress in livestock populations (Gebregiwerigs et al., 2020; Maltecca et al., 2020; Meuwissen et al., 2020). However, studies were mainly focused on the theoretical background, simulated data, and real SNP genotypes derived from a single SNP panel of data, using few samples.

In dairy cattle, studies that investigated genomic inbreeding from a practical perspective, meaning analyzing genotypes that are a product of a routine genomic evaluation process that includes several filtering steps and imputation of tens or hundreds of thousands of cattle and tens of SNP panels and comparing several genomic inbreeding estimators, are limited (Dadousis et al., 2021a; Nani and VanRaden, 2021). However, it is common that semen companies provide farmers with inbreeding estimates of sires based on genomic data, information that is expected to help farmers in taking decisions on planning their matings within the herd and to further control diversity and genetic progress in their herd. However, given the unresolved debate (Villanueva et al., 2021) of genomic inbreeding, the usefulness of these measures is questionable; hence, future expectations on decisions made by using genomic inbreeding coefficients should be seen with caution.

Given the above, our goal was to expand on the debate of SNP-derived genomic inbreeding by considering imputed SNP data and investigating the performance of 10 genomic inbreeding estimators widely used in livestock in a data set of 95,540 Italian Holstein dairy cows with 84,445 imputed SNP. Data were from the routine genomic evaluations of ANAFIBJ.

Categorizing and Interpreting Genomic Inbreeding Estimators Using Imputed SNP Data

For a fair comparison of pedigree versus SNP-based inbreeding estimates, a correct $F_{\text{PED}}$ is vital. Pedi-
Figure 3. (A) Pairwise Pearson correlations (above diagonal) between each pair of the pedigree and the 10 genomic inbreeding estimators analyzed, pairwise x-y plots (below diagonal) between each pair of the pedigree and the 10 genomic inbreeding estimators analyzed, and densities (diagonal) of the inbreeding estimators analyzed. (B) Pairwise Spearman correlations between each pair of the pedigree and the 10 genomic inbreeding estimators analyzed. See Table 1 and Materials and Methods section for a detailed description of the inbreeding coefficient estimators.
Quality was high with mean values of equivalent generations and pedigree completeness index of 10.91 and 0.99, respectively (Supplemental Table S1, https://figshare.com/s/386c45ee3d2ed797bf0). Moreover, based on the history and management of the Italian Holstein population, we compared our \( F_{\text{PED}} \) estimates with the US Holstein, because a large proportion of the Italian Holstein sires are US bulls. Our average \( F_{\text{PED}} \) per year of birth between 2015 and 2019 varied between 6.62, 6.92, 7.52, 8.29, and 8.83%, which agrees with the US Holstein reports of 6.59, 6.90, 7.29, 7.72, and 8.18% (https://queries.uscdcb.com/eval/summary/inbrd.cfm?R_Menu=HO#StartBody), respectively.

In general, genomic inbreeding estimators could be categorized (Table 1) into 2 main groups: those that summarize SNP by SNP or those that analyze interrupted sequences of SNP of a predefined length (i.e., ROH). A second grouping can be made by those estimators that depend (or not) on the allele frequencies. For instance, \( F_{\text{GRM2}} \) uses pedigree information to adjust for homozygosity avoiding the use of allele frequency in calculations. Further, there are methods that focus on homozygosity excess (\( F, F_{\text{HAT2}} \)), variance of additively recoded genotypes (\( F_{\text{HAT1}}, F_{\text{GRM}}, F_{\text{GRM05}} \)) or correlation between uniting gametes (\( F_{\text{HAT4}} \)).

Moreover, an important concept is the interpretation of inbreeding coefficient, which is vital not only for comparing different genomic inbreeding estimators but also for explaining diversity in the population that further helps control the inbreeding level. If individual

---

**Figure 4.** Scatterplot of the loadings of the first 2 principal components (Dim1 and Dim2, respectively; in parenthesis the proportion of variance explained by each principal component). Principal component analysis (PCA) was performed on the matrix of the inbreeding coefficients of the pedigree and the 10 genomic inbreeding estimators analyzed. See Table 1 and Materials and Methods section for a detailed description of the inbreeding coefficient estimators. Color scale indicates the contribution (in %) of each inbreeding estimator on the first principal component, based on the squared cosine (\( \cos^2 \)) quality of representation of the variables of the principal components applied in the R package factoextra (Kassambara and Mundt, 2020); see Abdi and Williams (2010) for more details. Angles between each pair of the 11 inbreeding estimators represent the correlations between each pair. Positive correlated variables point to the same side of the plot. Negative correlated variables point to opposite sides of the graph.
inbreeding coefficients reflect a proportional loss (a main effect of directional intense selection of breeding programs) or even gain (possible but very unlikely in intense dairy breeding programs) of variability given an unselected base population with unrelated individuals, then negative genomic inbreeding coefficients make sense with an accepted space of \([-\infty,1]\) (Villanueva et al., 2021). The general space that the genomic inbreeding estimators belonged to was summarized as follows: \([0,1]\) for \(F_{\text{PED}}, F_{\text{GRM05}}, F_{\text{ROH}}, F_{\text{ROH2}},\) and \(F_{\text{PH}}\); \([-1,1]\) for \(F_{i}\); \([-\infty,1]\) for \(F_{\text{HAT1}}\) and \(F_{\text{HAT3}}\); \([-\infty,1]\) for \(F_{\text{HAT2}}\) and \(F_{\text{GRM2}}\); and \([-\infty,\infty]\) for \(F_{\text{GRM}}\). As such, \(F_{\text{GRM}}\) and \(F_{\text{HAT1,3}}\) estimators are hard to interpret for a practical application in breeding programs that apply an intense directional selection, as in the case of Holstein dairy cattle. However, estimators are not impossible to obtain, because they provide with coefficients >1, meaning that more variability has been lost than the variability that existed in the base population (Villanueva et al., 2021). Regarding the rest of the genomic inbreeding estimators, \(F_{\text{GRM05}}, F_{\text{ROH}}, F_{\text{ROH2}},\) and \(F_{\text{PH}}\) that provided estimates within the \([0,1]\) space are considered comparable to \(F_{\text{PED}}\), reflecting the definition of Malécot (1948). Not comparable to \(F_{\text{PED}}\), yet with an interpreted value can be considered the \(F_{i}, F_{\text{HAT2}},\) and \(F_{\text{GRM2}}\) estimators; with \(F\) directly related to the definition of (Wright, 1921, 1922) and the rest within the general concept of heterozygosity reduction. For these estimators, negative coefficients have the meaning (although questionable and very unlikely for an intense and directional selection) that compared with the base population, variability has been gained. Worthy to note, that although Wright defined \(f\) as a correlation measure, also stated that “For a natural coefficient of inbreeding, we want a scale which runs from 0 to 1, while the percentage of homozygosis is running from 50 per cent to 100 per cent.” (Wright, 1922, pages 332–333). These findings agree with the expected values reported in Villanueva et al. (2021). However, our \(F_{\text{GRM}}\) estimates did not follow the pattern of \(F_{\text{HAT1}}\) and \(F_{\text{HAT3}}\) reported by Villanueva et al. (2021) and slightly exceeded the \([-1,\infty]\) space downward, with a range of \([-1.123, 1.065]\). This could be partly attributed to the effect of imputation. For more on the theoretical background of some of the genomic inbreeding estimators, the interested reader is directed to the recent studies of Caballero et al. (2021) and Villanueva et al. (2021) and the complementary works of Gebregiwergis et al. (2020) and Meuwissen et al. (2020).

Despite this, genomic inbreeding coefficients close to 1, at least for the ROH-based, \(F\) and \(F_{\text{PH}}\) estimators that measure percentage of homozygosity, are unrealistic, as they resemble cows with almost the whole genome in homozygous state. We hypothesize that this is an artifact as a result of the imputation procedure. Indeed, investigating a small subset of cows (\(n = 679\); maximum \(F_{\text{PED}} = 0.28\)) that were genotyped with the 777k Illumina Infinium BovineHD BeadChip the maximum coefficients for \(F_{\text{ROH}}, \) and the \(F_{\text{HAT1,2,3}}\) were 0.34, 0.25, and 0.26, respectively. The same was observed for a subset of 642 cows (maximum \(F_{\text{PED}} = 0.29\)) genotyped with the GeneSeek Genomic Profiler HD-150K SNP panel with maximum coefficients of \(F, F_{\text{ROH}},\) and the \(F_{\text{HAT1,2,3}}\) at 0.31, 0.41, and 0.50 (we note that for the cow with \(F_{\text{HAT1}} = 0.50, F_{\text{PED}}, F_{\text{HAT3}}, \) and \(F_{\text{ROH}}\) were between 0.03 and 0.08), respectively. Hence, in none of the high density (HD) genotyped cows genomic inbreeding coefficients exceeded 0.50. Given that cows genotyped in HD are expected to be elite dams, genomic inbreeding coefficients close to 1 might be considered as an artifact of imputation (that further depends both on the imputed and unimputed alleles), that justifies further investigation.

Correlation measures might partly describe the strength of association between 2 measures and might not provide the full picture of their relationship. For example, \(F_{\text{GRM05}}-F_{\text{PH}}\) had \(r = 1.\) This perfect correlation, however, might be misleading and does not imply equality of the measures (Table 2). For example, the median values of \(F_{\text{GRM05}}\) and \(F_{\text{PH}}\) were 0.273 and 0.637, and the coefficients of variation were 0.140 and 0.03, respectively. Although with similar maximum values in both estimators, the minimum drastically differed (0.000 vs. 0.487, respectively) between them. In general, \(F_{\text{PH}}\) was the estimator with the lowest range (with coefficient of variation at 3%) among all genomic inbreeding estimators (Tables 2 and 3). In Table 3, the unique values in the total set of 95,540 cows analyzed was summarized. Estimators \(F_{\text{GRM}}\) and \(F_{\text{GRM2}}\) where the only ones that provided 95,540 unique values (we note, however, that \(F_{\text{GRM05}}, F_{\text{ROH2}},\) and \(F_{\text{PH}}\) were provided by ANAIBJ with 4 decimals points), followed by \(F_{\text{ROH}}\) (95,521). Considering a scenario of 4 decimals, that addresses practical applications, unique values per estimator ranged between 1,542 (\(F_{\text{PH}}\)) to 16,586 (\(F_{\text{GRM2}}\)). Except for \(F_{\text{GRM2}}\) (high number of unique genomic inbreeding coefficients), unique values of the rest of genomic inbreeding estimators ranged between 2,572 and 6,735 for \(F_{\text{ROH}}\) and \(F_{\text{HAT2}}\); respectively. The number of unique values reflects the precision of the inbreeding estimates in the same way that genomic breeding values provide a more accurate ranking for example between full-sibs compared with traditional progeny tested and pedigree-based breeding values.
Cows with Zero \( F_{\text{PED}} \)

In total, 766 cows were not found inbred based on pedigree information (\( F_{\text{PED}} = 0 \)). However, considerable variation existed in the genomic inbreeding coefficients of those cows (Supplemental Table S2, https://figshare.com/s/386c45ec3d2ed797bf0). Interestingly, the minimum coefficients observed of each genomic inbreeding estimator for those cows were those minimum coefficients observed in the entire data set, with the only exception of \( F_{\text{HAT1}} \), \(-0.141 \) vs. \(-0.304 \) found in the complete data set (Table 2). The same was not observed for the maximum coefficients. However, in general the maximum genomic inbreeding coefficients for these cows were close to the maximum reported in Table 2, with the exceptions of \( F_{\text{HAT1}} \) and \( F_{\text{HAT3}} \).

Regarding the pedigree quality of those 766 cows, as expected, the average equivalent generations and pedigree index was much lower at 5.56 and 0.27, respectively; compared with 10.91 and 0.99 found in the complete data.

Comparisons of the Most Inbred Cows Based on Pedigree Data

Apart from the discrepancy on those cows that have \( F_{\text{PED}} = 0 \) but nonzero genomic inbreeding coefficients, there were also cows with high \( F_{\text{PED}} \). If the breeding objective is to detect and eliminate highly inbred animals in the population, then genomic inbreeding estimators are expected to provide useful and more precise information on the level of homozygosity of each animal. Moreover, genomic information can be precise down to the level of specific regions that might be interesting for selection, given that there is enough knowledge on the effect of these regions and potential genes associated with desirable phenotypic characteristics (Cole and VanRaden, 2010; Cole and Null, 2013; Howard et al., 2017).

In general, the correlations of the top 5% \( F_{\text{PED}} \) estimates (\( F_{\text{PED}} \geq 0.116; 4,777 \) cows) with each of the genomic inbreeding estimators were higher, compared with observed \( \rho \) in the entire data (Supplemental Figure S2). High variation was observed in the top 5% inbred cows in common between \( F_{\text{PED}} \) and genomic inbreeding estimators, with a minimum number between \( F_{\text{PED}}-F_{\text{HAT1}} \) (180) and maximum for \( F_{\text{PED}}-F_{\text{ROH2}} \) (2,353). Similar to \( F_{\text{ROH}} \) with more than 2,000 (out of 4,777) cows in common were also \( F \), \( F_{\text{HAT2}} \), \( F_{\text{GRM05}} \), \( F_{\text{ROH}} \), and \( F_{\text{PH}} \) (2,305, 2,205, 2,266, 2,304, and 2,277, respectively). Slightly lower was the number of high inbred cows found in common for \( F_{\text{PED}}-F_{\text{HAT3}} \) (1,687), whereas values below 1,000 were found, apart from \( F_{\text{HAT1}} \), for \( F_{\text{GRM}} \) and \( F_{\text{GRM2}} \) (539 and 251, respectively).

Effect of Imputation on Genomic Inbreeding Coefficients

The effect of imputation success on the inbreeding estimates deserves a further investigation and quantification. In a real case scenario, where groups of animals

<table>
<thead>
<tr>
<th>Type</th>
<th>Estimator/Genotype</th>
<th>Complete</th>
<th>6</th>
<th>5</th>
<th>4</th>
<th>3</th>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pedigree</td>
<td>( F_{\text{PED}} )</td>
<td>88,153</td>
<td>51,308</td>
<td>11,439</td>
<td>1,800</td>
<td>263</td>
<td>34</td>
</tr>
<tr>
<td>PLINK(^2)</td>
<td>( F )</td>
<td>28,896</td>
<td>28,270</td>
<td>17,966</td>
<td>3,229</td>
<td>677</td>
<td>114</td>
</tr>
<tr>
<td>( F_{\text{HAT1}} )</td>
<td>92,696</td>
<td>81,523</td>
<td>33,447</td>
<td>6,548</td>
<td>1,213</td>
<td>220</td>
<td></td>
</tr>
<tr>
<td>( F_{\text{HAT2}} )</td>
<td>92,834</td>
<td>81,588</td>
<td>33,886</td>
<td>6,735</td>
<td>1,379</td>
<td>281</td>
<td></td>
</tr>
<tr>
<td>( F_{\text{HAT3}} )</td>
<td>92,673</td>
<td>62,432</td>
<td>15,156</td>
<td>2,770</td>
<td>656</td>
<td>130</td>
<td></td>
</tr>
<tr>
<td>GRM(^3)</td>
<td>( F_{\text{GRM}} )</td>
<td>95,540</td>
<td>80,157</td>
<td>30,393</td>
<td>5,264</td>
<td>962</td>
<td>177</td>
</tr>
<tr>
<td>Genomic</td>
<td>( F_{\text{GRM05}} )</td>
<td>2,503</td>
<td>2,503</td>
<td>2,503</td>
<td>529</td>
<td>86</td>
<td></td>
</tr>
<tr>
<td>Genomic</td>
<td>( F_{\text{GRM2}} )</td>
<td>95,540</td>
<td>90,963</td>
<td>62,754</td>
<td>16,586</td>
<td>2,251</td>
<td>291</td>
</tr>
<tr>
<td>Genomic</td>
<td>( F_{\text{ROH}} )</td>
<td>95,521</td>
<td>66,927</td>
<td>16,958</td>
<td>2,664</td>
<td>478</td>
<td>77</td>
</tr>
<tr>
<td>Genomic</td>
<td>( F_{\text{ROH2}} )</td>
<td>2,572</td>
<td>2,572</td>
<td>2,572</td>
<td>406</td>
<td>67</td>
<td></td>
</tr>
<tr>
<td>Genomic</td>
<td>( F_{\text{PH4}} )</td>
<td>1,542</td>
<td>1,542</td>
<td>1,542</td>
<td>1,542</td>
<td>328</td>
<td>47</td>
</tr>
</tbody>
</table>

1See Table 1 and the Materials and Methods section for a detailed description of the inbreeding coefficient estimators.
2Genomic inbreeding estimates based on PLINK v1.9.
3Genomic inbreeding estimates based on commonly used genomic relationship matrices (GRM).
4Those estimators have been calculated with 4 decimals points by the Italian National Association of Holstein, Brown and Jersey Breeders (ANAFIBJ).
5Genomic inbreeding estimates based on runs of homozygosity (ROH).
6\( F_{\text{PH}} \) is a simplification of \( F \), measuring percentage of SNP-based homozygosity, developed by ANAFIBJ.
are genotyped with different SNP arrays, this work becomes more complex and difficult to be addressed. Moreover, different animals might have different relationship with the core animals (animals that have been genotyped with high-density SNP panels or sequenced) adding another confounding effect. The behavior of various imputation algorithms and strategies (i.e., which and how many core animals should be used; how often this core population should be updated) is also important but out of the scope of this work.

In our data, on average, the fraction of nonmissing SNPs after imputation (hereafter declared as imputation fill rate) was 0.996 with minimum and maximum values between 0.616 and 1, respectively (Supplemental Figure S3a, https://figshare.com/s/386c45ee3d2ed797bfd0). A high value implies that a low number of SNP were imputed, which is mainly the case of the HD SNP panels. By plotting inbreeding coefficients versus the imputation filling rate, we showed that low values of the filling rate of imputation resulted in lower genomic inbreeding coefficients, with increased variability and increased imputation fill rate (Figure 5). Moreover, cows that had <0.95 fill rate also appear with extreme inbreeding coefficients. We hypothesized an effect of (1) different SNP panels to genotype cows and (2) SNP selection and imputation applied in the routine genomic evaluation process. To test the effect of imputation on genomic inbreeding, 4 medium-density (MD) SNP chips (each with >10K genotyped cows; GeneSeek Genomic Profiler3,4,MD and LabogenaMD) and 2 HD SNP chip (~650 genotyped cows each; GeneSeek Genomic ProfilerHD-150K and Illumina BovineHD BeadChip777K) were selected comparing the genomic inbreeding coefficients derived from the 10 aforementioned estimators with the imputed SNP. Indeed, primary results indicated differences among MD SNP panels with more robust found for the LabogenaMD (Dadousis et al., 2021a). This can be partly attributed to the fact that ~98% of the LabogenaMD SNP are also included in the set of 84,445 final SNP set, whereas for the rest of MD SNP panels this is between 55 and 60%. Consistent results were found also for the HD panels (data not shown).

**Similarities Among Inbreeding Estimators Based on Principal Components Analysis**

A PCA was conducted in all inbreeding estimators. The first 2 PC captured ~75% of the original variability of all pedigree and genomic inbreeding estimators. Three main groups were identified (Figure 4), with FGRM2 standing alone and with minor contribution on PC1; FHAT1 was grouped together with FGRM, both with slightly higher contribution on PC1 compared with FGRM2 and all the remaining estimators, including Fped, were clustered together. From the last group, the highest contribution on PC1 was of F, FGRM05, and FPH, followed by FROH and FROH2. Moreover, Fped-FHAT2 were placed together, whereas FHAT3 slightly differentiated from the rest. Obviously, FHAT1, FGRM, and FGRM2 had the highest loadings on PC2, describing the second axis of highest variation in the inbreeding coefficient estimators.

Theoretically different estimators of inbreeding coefficients share some properties, yet a clear description of their connections remains complex (Alemu et al., 2021). Given that inbreeding coefficients, pedigree- or SNP-based, vary among estimators, each of which captures a different part of information on the inbreeding status of an individual, a reasonable question is if all the information could be summarized in one single measure.

**Assessing the Strength of Relationship Between Inbreeding Estimators**

The strength of the relationship between different inbreeding estimators in the literature is usually presented via the Pearson correlation coefficient and in some cases without supported by scatterplots. However, $\rho$ assesses linear relationships and can underestimate the strength of association between 2 variables in a nonlinear but monotonically increased pattern. The latter case is better described by the Spearman rank correlation. A clear monotonic pattern was found between FHAT2-FROH and FHAT2-FROH2 (Figure 3a, b). For example, $\rho$ values of 0.540, 0.594, 0.590, and 0.540 were found between FHAT2 and FGRM05, FROH, FROH2, and FPH, respectively. The corresponding $r$ were 0.726, 0.777, 0.779, and 0.726. The most profound difference (more than a 3-fold magnitude) of a monotonic relationship between $\rho$ and $r$ estimates was found between FROH2-FGRM2 ($-0.02$ vs. $-0.09$).

Apart from the linear and not linear but monotonic relationships, nonlinear and substructured correlations were also found. In this case, both measures, $\rho$ and $r$ fail to describe the strength a relationship between 2 variables when applied to the entire data, hence their usefulness and interpretability should be seen with caution. Such substructure in the pairwise comparisons between the estimators was clear in the FGRM scatterplots.

**Future Perspectives**

Our work was driven from a practical perspective, meaning to find a consistent, robust, and simple measure of genomic inbreeding available for farmers, AI centers, and breeding associations. We recognize that
Figure 5. Scatterplot of the imputation fill rate (fraction of nonmissing SNPs after imputation; x-axis) in each cow versus inbreeding coefficients (y-axis) in 95,540 Italian Holstein dairy cows. See Table 1 and the Materials and Methods section for a detailed description of the genomic inbreeding coefficient estimators.
our attempt to investigate genomic inbreeding coefficients is limited and the topic cannot be exhausted to those 10 estimators applied in our study. Other genomic inbreeding estimators exist (Nejati-Javaremi et al., 1997; Eding and Meeussen, 2001), and recently Nani and VanRaden (2021) investigated the adjustment for the X-chromosome to better scale genomic to pedigree inbreeding coefficients and to account for differences between females and males. Moreover, measures of autozygosity and ROH classes provide additional knowledge on the form of homozygosity (i.e., distinguish IBS from IBD) and may be considered complementary to other genomic inbreeding estimators for organizing matings and controlling inbreeding. We believe that the effect of various sources of variation, as mentioned in the introduction, on genomic inbreeding coefficients, needs to be further investigated and quantified.

Because the imputed data are the final product of many interrelated and confounded factors, future research should identify and quantify all possible factors that influence SNP-based inbreeding coefficients. These factors include the quality and quantity of SNP panels, filtering steps and quality parameters of the imputation (e.g., number of SNP imputed per animal). As a side effect of our analysis, cows with unreasonable homozygosity levels (close to 1) after imputation have been detected. We evidenced that this might be directly linked to imputation. The effect that this might have on the genomic evaluations should be further investigated. Moreover, the effect of using different inbreeding coefficients for inbreeding management and for genomic evaluations (diagonal of GRM used for genomic evaluations) should be further considered.

Observed allele frequencies were used in our analysis for GRM-based inbreeding estimators, rather than allele frequencies estimated in the base population consisting of unrelated and not inbred animals. The effect of replacing the observed with calculated allele frequencies in the base population should be further considered, although it was partly addressed here by setting allele frequencies to 0.5 in $F_{GRM05}$.

**CONCLUSIONS**

In this second decade of routine genomic evaluations in livestock, there are still compelling reasons to study genomic inbreeding. We have expanded the debate of genomic inbreeding by considering imputed SNP data, which has become a routine practice in most dairy cattle breeding programs worldwide. Results varied among 10 estimators of SNP-based inbreeding coefficients. Our results question the wide use of imputed SNP-based genomic inbreeding estimates by farmers, at least without an information on how these estimates have been derived and without being accompanied with a clear meaning. Further research should quantify the effect of imputation on genomic inbreeding estimates. Finally, we would like to draw the attention to the scientific community of the difficulties posed by the inconsistent nomenclature of genomic inbreeding methods, a “Babel tower” that slowly builds up with a risk to create lack of communication and irreproducible research in the nearest future.

**ACKNOWLEDGMENTS**

This study was supported by “Latteco2 project, sottomisura 10.2 of the PSRN Biodiversity 2020–2023” (MIPAAF. D.M. no. 465907 del 24/09/2021, project unique code J12C21004080005). This study benefits from the HPC (High Performance Computing) facility of the University of Parma, Italy. The authors thank the 2 anonymous reviewers for their constructive reviewing and helpful comments on the manuscript. We also thank Ricardo Pong-Wong [Genetics and Genomics, the Roslin Institute and the R(D)SVS, University of Edinburgh, Easter Bush, UK] for providing fruitful viewing and helpful comments on the manuscript. We have expanded the debate of genomic inbreeding coefficients. The authors have not stated any conflicts of interest.

**REFERENCES**


ORCIDs

Christos Dadousis @ https://orcid.org/0000-0002-7689-6443
Michela Ablondi @ https://orcid.org/0000-0003-3700-1042
Claudio Cipolat-Gotet @ https://orcid.org/0000-0002-2318-4231
Jan-Thijs van Kaam @ https://orcid.org/0000-0002-2592-2461
Maurizio Marusi @ https://orcid.org/0000-0002-2359-3633
Martino Cassandro @ https://orcid.org/0000-0002-8709-2870
Alberto Sabbioni @ https://orcid.org/0000-0002-2141-6668
Andrea Summer @ https://orcid.org/0000-0002-4833-657X