Genomic characterization and conservation of genetic diversity in cattle

Harmen P. Doekes

Propositions

1. Not all inbreeding is depressing.
   (this thesis)

2. Every commercial livestock breed should have a gene bank collection.
   (this thesis)

3. For the development of science, it is essential that scientists are humble.

4. At least 25% of a PhD should consist of work outside the direct scientific field of the PhD-project.

5. The evolution of music, which is studied in musicology, has many parallels with animal breeding, including the need to maintain diversity.

6. We have an ethical obligation to ensure the wellbeing of domestic animals, but this does not imply that we have an ethical obligation to conserve domestic breeds.

7. One first has to get lost to find the right way.

Propositions belonging to the thesis, entitled

Genomic characterization and conservation of genetic diversity in cattle

Harmen P. Doekes
Wageningen, 18 September 2020

Genomic characterization and conservation of genetic diversity in cattle

Harmen P. Doekes

Thesis committee

Promotor
Prof. Dr. R.F. Veerkamp
Special professor of Numerical Genetics
Wageningen University & Research

Co-promotor
Dr. J.J. Windig
Researcher Animal Breeding and Genomics
Wageningen University & Research

Other members
Prof. Dr. B. Zwaan, Wageningen University & Research
Prof. Dr. J.E. Pryce, La Trobe University, Melbourne, Australia
Dr. G. Restoux, AgroParisTech, Jouy-en-Josas, France
Dr. M.J.M. Smulders, Wageningen University & Research

This research was conducted under the auspices of the Graduate School Wageningen Institute of Animal Sciences (WIAS).

Genomic characterization and conservation of genetic diversity in cattle

Harmen P. Doekes

Thesis

submitted in fulfilment of the requirements for the degree of doctor at
Wageningen University
by the authority of the Rector Magnificus
Prof. Dr. A.P.J. Mol,
in the presence of the
Thesis Committee appointed by the Academic Board
to be defended in public
on Friday 18 September 2020
at 4 p.m. in the Aula.

Doekes, H.P.
Genomic characterization and conservation of genetic diversity in cattle
216 pages

PhD thesis, Wageningen University, Wageningen, the Netherlands (2020)
With references, with summary in English

ISBN: 978-94-6395-423-5
DOI: https://doi.org/10.18174/523321

Abstract

Genetic diversity in livestock populations is important, because it forms the basis for these populations to adapt to changing environments and human demands. Traditionally, livestock genetic diversity has been characterized and conserved with pedigree-based measures of inbreeding and kinship. Thanks to the increasing availability of genomic information, in particular single nucleotide polymorphism (SNP) data, we now have additional opportunities to better manage genetic diversity. In this thesis, I utilized SNP data to characterize and conserve genetic diversity in Dutch cattle, both in *in situ* populations and *ex situ* gene bank collections. The Holstein Friesian (HF) breed was the main breed of interest, because of its importance in the Dutch and global dairy cattle sector. First, it was demonstrated how changes in breeding practices in the past have been accompanied by changes in genetic diversity trends in the Dutch-Flemish HF breeding program. Among others, it was shown that the introduction of genomic selection has been accompanied by an increase in pedigree-based and SNP-based rates of inbreeding and kinship. Second, the negative effects of inbreeding on performance ("inbreeding depression") were quantified for yield, fertility and udder health traits of Dutch HF cows. It was shown that recent inbreeding may be more harmful than ancient inbreeding, although results were mixed. It was also shown that, based on SNP data, the negative effects of inbreeding are quite equally distributed across the genome and well captured by genome-wide homozygosity. Third, the value of the Dutch cattle gene bank collection was demonstrated. It was shown that old HF gene bank bulls can be used in the current or future HF breeding program to increase (or recover) genetic diversity, or to improve genetic merit given a certain level of diversity. It was also shown that Dutch native breeds in the gene bank collection harbor genetic diversity, both within and across breeds, although some breeds showed substantial overlap. Last, it was discussed how genomic information (in particular SNP data) can be used to maintain genetic diversity in livestock populations. Based on our current knowledge and the availability of SNP data, I recommend to limit the increase in SNP-by-SNP similarity (and, thus, homozygosity) while performing selection. For gene bank collections, I envision a transition towards bio-digital resource centers, in which large amounts of genomic and phenotypic data are stored in addition to physical germplasm material. Overall, the findings of this thesis improve our understanding of (conservation of) genetic diversity in livestock and, thereby, contribute to sustainable livestock production.

Contents

1

General introduction

1.1 Introduction

Genetic diversity is the set of genetic differences between species, between populations within species, and between individuals within populations [1]. Genetic diversity is important for livestock (and other) populations, because it allows for adaptation to changing environments and for genetic improvement following human demands [1, 2]. In addition, livestock genetic resources have socio-economic, cultural and ecological value [3]. Hence, it is not surprising that there are various initiatives to monitor and conserve livestock genetic resources (Box 1.1).

Box 1.1 Examples of global, regional and national initiatives to monitor and conserve livestock genetic resources

The Convention on Biological Diversity, which entered into force in 1993, calls on countries to identify and monitor their biodiversity, including livestock genetic diversity [4]. In addition, one of the targets of the second Sustainable Development Goal of the United Nations is to "*maintain the genetic diversity of seeds, cultivated plants and farmed and domesticated animals and their related wild species ...*" [5].

At the global level, the Food and Agricultural Organization (FAO) of the United Nations coordinates the monitoring of livestock genetic resources. Global assessments have been performed in 2007 [6] and 2015 [2]. Based on the first of these assessments, a Global Plan of Action was developed [7]. FAO also maintains an online database for livestock genetic resources, which is called DAD-IS [8].

At the European level, the European Regional Focal Point for Animal Genetic Resources (ERFP) is a platform that, among others, facilitates the implementation of FAO's Global Plan of Action. Each member state has a National Focal Point and National Coordinator, assigned by the respective Ministry of the member state [9].

In the Netherlands, the Centre for Genetic Resources (CGN) is the National Focal Point. Among others, CGN monitors livestock genetic resources in the Netherlands and advises various stakeholders regarding conservation and sustainable use of these resources. Stakeholders include breeding organizations and the Dutch Ministry of Agriculture, Nature and Food Quality [10].

Within livestock species, animals are historically classified in groups called breeds. Although there is no universal definition of a breed, an operational definition is given by the Food and Agricultural Organization (FAO) of the United Nations: "*a breed is a subspecific group of domestic livestock with a common history whose members are treated in a common manner with respect to genetic management*" [11].

Over the last centuries, livestock breeding has undergone some major developments (Section 1.2). A recent example is the advancement in genomic technologies. The availability of genomic information (i.e. DNA information) has not only changed the way in which animals are being selected for breeding, but also

offers opportunities to better characterize and conserve genetic diversity compared to traditional methods.

In this thesis, I focus on the genomic characterization and conservation of genetic diversity in Dutch cattle, with an emphasis on the Holstein Friesian (HF) breed. HF is an important livestock breed, because it dominates the global and Dutch dairy cattle sector, with tens of millions of cows worldwide and approximately 1.3 million cows in the Netherlands [12, 13]. HF is also the first livestock breed in which genomic selection (Section 1.2) has been widely implemented. Therefore, it is an important breed to address questions related to genetic diversity and its conservation in the era of genomics. With this thesis, I aim to improve our understanding of genetic diversity and its conservation, and thereby contribute to sustainable livestock production.

In the rest of this introduction, I first give a short history of livestock breeding and describe the processes that influence genetic diversity in livestock breeds. I then introduce the concepts of inbreeding and kinship as measures of genetic diversity, describe the phenomenon of inbreeding depression and introduce conservation strategies. I finish the introduction with a section on genetic diversity in Dutch cattle, followed by the aim and outline of the thesis.

1.2 Short history of livestock breeding and genetic diversity

Livestock has been domesticated over the past 11,500 years [2, 14]. Ever since the first domestication events, livestock keepers have used selective breeding to change characteristics of their animals. Genetic improvement as we know it today, however, did not become routine until the industrial revolution, which started around 1760 in England [15]. Around this time, people began to systematically record phenotypes and pedigrees and the first breeders' associations were established [2]. In combination with geographical separation, this resulted in the development of more homogeneous breeds. Not all newly formed breeds were equally successful. Less successful breeds quickly disappeared, whereas more successful breeds were disseminated across the world, facilitated by developments in transportation in the 19th century and the rise of artificial insemination (AI) and cryopreservation in the decades following the Second World War [16-18]. As a result, a few successful transboundary breeds became very widespread, whereas many local breeds decreased in population size and became rare or extinct [2, 19, 20]. According to the Domestic Animal Diversity Information System (DAD-IS) of FAO, around 50 to 60% of breeds from the "big five" livestock species currently has an unknown risk status (Figure 1.1). Of the breeds that do have a known risk status, only a small percentage (<30%) is categorized as not at risk, whereas the majority (>60%) is categorized as endangered, critical or extinct.

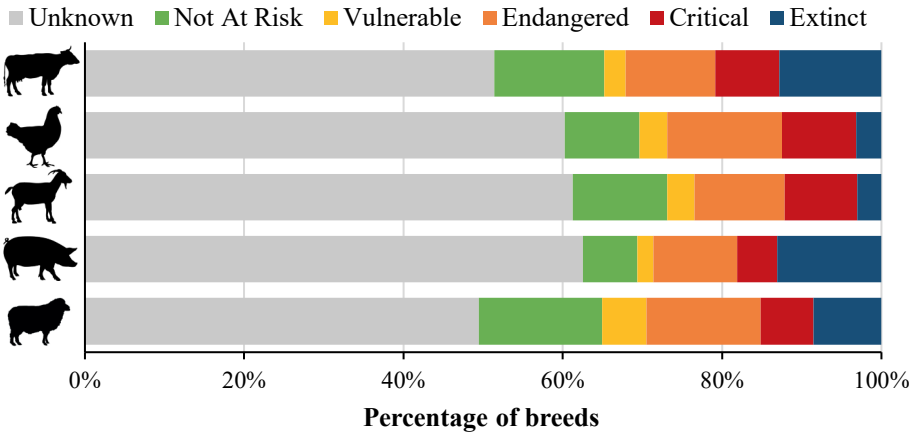


Figure 1.1 Percentage of breeds per risk category for cattle (n = 1433), chicken (n = 1653), goat (n = 691), pig (n = 712) and sheep (n = 712). Data were obtained from DAD-IS in October 2019 [8] and local and transboundary breeds were combined.

Within breeds, the rise of reproductive technologies such as AI has allowed for single males to produce many offspring. Examples are the popular dairy cattle bulls Sunny Boy (born 1985) and Toystory (born 2001), who produced more than 1.7 and 2.4 million straws of sperm, respectively [21]. Sunny boy was used for 25% of all inseminations in Dutch cows in 1990-1992 [22]. Although preferential use of specific individuals may reduce genetic diversity, the selection of genetically superior individuals also forms the basis of animal breeding (Box 1.2).

Box 1.2 The basics of animal breeding

Animal breeding is the process in which breeders select animals that are genetically superior to become parents of the next generation. Selection is repeated for many generations to improve the mean performance of the population over time. An example of successful selection is that of Dutch HF, where the mean protein yield in the first lactation of cows has increased from approximately 230 kg in 1990 to 290 kg in 2018, mainly as a result of genetic improvement [23].

Genetic improvement is only possible for traits that show heritable genetic variation. The heritable component of an animal's total genetic value is called the 'breeding value' and the expected performance of an animal's offspring is half of the animal's breeding value [24]. Although true breeding values are unknown, they can be estimated. Animals are selected based on estimated breeding values (EBVs).

In addition to the rise of reproductive technologies like AI, advancements in statistical methodology have had a major impact on animal breeding. In 1943, selection index theory was introduced as a means to simultaneously select for

multiple traits [25]. In 1949, Henderson introduced the mixed model procedure, which he further improved in following decades [26-28]. This procedure allows for best linear unbiased prediction (BLUP) of random effects, such as breeding values, while correcting for fixed factors such as herd-effects. In the 1980s, BLUP was implemented in breeding programs worldwide. In traditional BLUP, breeding values were estimated from performance records of animals themselves and/or of relatives, which were connected through a pedigree-based relationship matrix. With the advent of genomic technologies, this has changed. The development of economically attractive DNA arrays, with many single nucleotide polymorphisms or SNPs (i.e. substitutions of single nucleotides that occur at specific positions in the genome), has allowed genotyping of large numbers of animals at a relatively low cost. This has enabled the prediction of breeding values based on DNA profiles [29], a process called genomic prediction (or GBLUP). The use of genomic breeding values in selection is called genomic selection. Genomic selection was first implemented around 2009 in dairy cattle [30] and later also in other livestock species [31, 32]. One benefit of genomic selection, compared to traditional pedigree-based selection, is that estimated breeding values are more accurate at a very young age.

Breeding goals have also changed over time. Breeding programs initially focused on the appearance of animals, but with the introduction of BLUP and phenotyping (e.g. milk recording) the focus moved to improving production traits in the last century. Since then, breeding goals have become broader, including e.g. health and fertility traits, and they are expected to become even broader in future [33-37].

1.3 Forces that influence genetic diversity in populations

There are four major forces that influence genetic diversity in populations over time: mutation, migration, genetic drift and selection [24, 38, 39]. Mutation is the ultimate source of genetic variation. When mutations occur in germ cells, they are transmitted to the offspring and may be passed on to following generations. Although the mutation rate per base-pair is rather low, e.g. 1.21×10^{-8} in cattle [40], there are dozens of new mutations per genome per generation. The second force, migration, influences genetic diversity through exchange of genetic material between populations, such as breeds (e.g. by crossbreeding or introgression). Breeds typically have different allele frequencies, or even different alleles, and the level and direction of migration between them may strongly influence the diversity in the separate breeds as well as in the metapopulation of all breeds combined [38]. The third force, genetic drift, refers to random changes in allele frequency that occur across generations due to Mendelian sampling, i.e. due to offspring inheriting a random half of each parent's genetic material. The effect of genetic drift, and the probability that alleles are lost by chance, is larger for populations with a smaller

effective population size. The effective population size can be small due to a small census size, i.e. due to a small number of animals in the population, or due to an unequal contribution of animals to the reproduction process [24]. The latter can be seen as an indirect effect of selection, which is the last force shaping genetic diversity. In addition to its indirect effect through drift, selection reduces genetic diversity by acting directionally on allele frequencies. At any quantitative trait locus (QTL), selection is expected to increase the frequency of the favorable allele at the expense of the less favorable allele (unless there is heterozygous advantage). Consequently, the less favorable allele may be lost over time. Although this is essentially the aim of selection, there are two unfavorable side-effects. First, alleles at surrounding loci also change in frequency, due to linkage disequilibrium, and may ultimately be lost [41, 42]. Second, alleles that are currently deemed favorable may not be favorable in future, since environments and breeding goals change over time. Selection for traits that are currently of interest may thus result in a loss of alleles that are favorable for future traits of interest.

In addition to the forces mentioned above, recombination contributes to genetic diversity. During meiosis, crossovers occur between paternal and maternal chromosomes. Although this does not affect allele frequencies at population level, it does result in new genotypic combinations within individuals [43].

1.4 Genetic diversity measures: inbreeding and kinship

A wide range of measures can be used to characterize genetic diversity within and across populations [44, 45]. One of these measures is allelic richness, i.e. the mean number of alleles per locus. In addition to the number of alleles, the frequency at which these alleles occur determines genetic diversity. One of the most commonly used diversity measures, Nei's expected heterozygosity [46], reflects this principle. When allele frequencies at a locus are more equal, the expected heterozygosity (and, thus, diversity) is larger. In addition to the expected heterozygosity based on allele frequencies, the observed heterozygosity can be considered.

In practice, we typically do not have information on the entire genome. Therefore, other approaches are needed to estimate genetic diversity. In this thesis, I focus on three approaches: (1) pedigree-based inbreeding and kinship, (2) SNP-by-SNP homozygosity and similarity, and (3) segment-based inbreeding and kinship.

1.4.1 Pedigree-based inbreeding and kinship

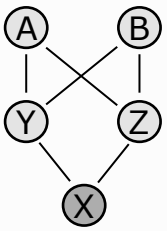
Traditionally, genetic diversity in livestock populations has been characterized and managed with pedigree-based coefficients of inbreeding and kinship. Inbreeding is the mating between relatives and the pedigree-based inbreeding of an individual is equal to the pedigree-based kinship between its parents [47]. Pedigree-based

inbreeding and kinship are probabilities that two alleles at a random locus in the genome, sampled within an individual (for inbreeding) or between individuals (for kinship) are identical-by-descent (IBD) with reference to a base population (Box 1.3). At population level, the proportional increases in mean inbreeding and mean kinship over time are related to losses in observed and expected heterozygosity, respectively [24]. These increases are known as the rates of inbreeding and kinship. As a general guideline, it is recommended to keep the rate of pedigree-based inbreeding in populations below 1% per generation, and preferably below 0.5% per generation [2].

Box 1.3 Definition and example of pedigree-based inbreeding and kinship

The pedigree-based inbreeding coefficient of animal X (F_{PED_X}) is the probability that two alleles at a neutral and selection-free locus in X are identical by descent (IBD) with reference to a base population (adapted from [24]). The pedigree-based kinship between animals Y and Z ($f_{PED_{YZ}}$) is the probability that two alleles at a locus, one randomly sampled from Y and one from Z, are IBD.

Example: in the pedigree on the right, animal X is inbred because its parents Y and Z are full-sibs. The probability that X is IBD for an allele from ancestor A is 0.5^3 and the probability that X is IBD for an allele from ancestor B is also 0.5^3. Therefore, the F_{PED_X} equals 0.25 (two times 0.5^3). The $f_{PED_{YZ}}$ equals the F_{PED_X} and, thus, also equals 0.25.



An advantage of the pedigree-based approach is that, once a herdbook is established and animals are systematically recorded, inbreeding and kinship can be easily calculated (Box 1.3). There are, however, also several limitations. First, pedigrees may be incomplete or incorrect, resulting in incorrect estimates of inbreeding and kinship [48]. Second, pedigree-based inbreeding and kinship are calculated with reference to a base population, which typically consists of the founders in the pedigree. These founders are assumed to be unrelated, while in practice they are likely to be related to some extent. Third, pedigree-based inbreeding and kinship are expectations of the proportion of the genome that is IBD within an individual or between two individuals, respectively. Realized proportions, however, vary from expectations due to Mendelian sampling [49]. Fourth, pedigree-based inbreeding and kinship are expectations for neutral and selection-free loci, i.e. loci that are not affected by selection and are not linked to loci affected by selection. Since complex traits are generally affected by thousands of loci, and the size of the genome is limited, it can be questioned whether such loci exist [50]. Last, pedigrees are generally recorded per breed, making it impossible to study diversity across breeds.

1.4.2 SNP-by-SNP homozygosity and similarity

With the increasing availability of genomic information, particularly SNP data, it has become possible to study diversity at the DNA level. One approach of measuring diversity based on SNP array data is to calculate SNP-by-SNP homozygosity and similarity. This approach relies on the concept of identical by state (IBS) and measures whether alleles at a SNP are identical or not (Box 1.4).

Box 1.4 Definition and example of SNP-by-SNP homozygosity and similarity

The SNP-by-SNP homozygosity of individual X (HOM_{SNP_X}) is the probability that two alleles at a SNP in X are identical by state (IBS). The SNP-by-SNP similarity between individuals Y and Z ($SIM_{SNP_{YZ}}$) is the probability that two alleles at a SNP, one randomly sampled from Y and one from Z, are IBS [51]. The mean HOM_{SNP_X} and mean $SIM_{SNP_{YZ}}$ equal one minus the observed and expected heterozygosity, respectively.

Example: in the SNP-data below, 26 of the 40 SNPs in individual X are homozygous (underlined). Therefore, the HOM_{SNP_X} equals 26/40 = 0.65.

Allele1: GTAGGAGTGCTTGTGCATTGCCACGAAACTGGATAGCTCG
Allele2: ATGGGAGTGCTTGTGCATCACTACAGAGCTAGTAAGCCGT

The advantage of genomic measures, compared to pedigree-based measures, is that they capture Mendelian sampling variation, account for founder relationships, allow for studying diversity at specific loci and allow for estimating relationships across breeds [52, 53]. Note that IBS-values are typically higher than IBD-values. This is due to the fact that alleles can be IBS because they are IBD, or because they are copies of two different alleles that were already IBS in the base population. To move from IBS to a SNP-based IBD estimate, the IBS-status can be scaled and centered by allele frequencies in the base population [54-57]. However, allele frequencies in the base population are often unknown and frequencies of the current population are used instead. Hence, the definition of the base population and the distinction between IBS and IBD have become less clear with the use of SNP data [58].

1.4.3 Segment-based inbreeding and kinship

In addition to measuring the IBS-status per SNP, it is also possible to measure the IBS-status of segments (Box 1.5). Long IBS-segments are expected to be IBD with reference to a relatively recent base population (Section 1.5). Thus, this approach can be used to estimate 'genomic' or 'realized' IBD. A drawback of this approach is that a set of criteria have to be predefined to identify segments [59-61].

Box 1.5 Definition and example of segment-based inbreeding and kinship

The segment-based inbreeding coefficient of individual X (F_{ROH_X}) is the proportion of the genome of X that is covered by regions of homozygosity (ROH), i.e. long stretches of homozygous SNPs [62]. The segment-based kinship coefficient between individuals Y and Z ($f_{SEG_{YZ}}$) is the amount of genome shared between Y and Z stored in haplotypes that are defined in the same way as ROH [63]. By considering long segments, this approach estimates genomic IBD.

Example: in the data below for individual X, there is one ROH that is 15 SNPs long (underlined). Assuming an equal distance between all 40 SNPs, this would give a F_{ROH_X} of 15/40 = 0.375.

Allele1: GTAGGAGTGCTTGTGCATTGCCACGAAACTGGATAGCTCG
Allele2: ATGGGAGTGCTTGTGCATCACTACAGAGCTAGTAAGCCGT

Pedigree-based and genomic approaches can also be combined when estimating kinships across individuals, e.g. when some individuals have both pedigree and genomic data, whereas others have only pedigree [64, 65]. It is challenging, however, to have the different approaches refer to the same base population [66, 67].

1.5 Inbreeding depression

Inbreeding is not only undesirable because it is associated with a loss in genetic diversity, but also because it reduces the mean performance of individuals [53, 68]. This phenomenon, now known as inbreeding depression, was already documented by Charles Darwin. In an experiment among various plant species, Darwin showed that offspring from self-fertilized plants were on average shorter, flowered later and produced fewer seeds than offspring produced by cross-fertilization of unrelated plants [69]. Darwin also had a personal interest in the topic (Box 1.6). Inbreeding depression has later also been documented for many livestock species and for, among others, production [70-72], reproduction [73, 74] and health traits [75, 76].

Traditionally, the degree of inbreeding depression has been assessed by regressing phenotypes on pedigree-based inbreeding coefficients. With genomic information, inbreeding depression can be studied in more detail for several reasons. First, genomic inbreeding coefficients are expected to be more accurate than pedigree-based coefficients (Section 1.4) and, therefore, could capture the negative effects on performance better. Second, the segment-based approach (Section 1.4.3) offers additional opportunities to infer the age of inbreeding. Long regions of homozygosity (ROH) are expected to represent more recent inbreeding than short

ROH, because recombination breaks up segments over time [77, 78]. Recent inbreeding is expected to be more harmful than ancient inbreeding, because the negative effects of ancient inbreeding may already have been purged from the population [79]. ROHs offer an additional way to test this hypothesis. Third, genomic information allows to study heterogeneity in inbreeding depression across the genome [80, 81]. Inbreeding may be more harmful in some regions of the genome than in others. If this is indeed the case, then it might be valuable to limit inbreeding in these regions more strictly than in other parts of the genome [82].

Box 1.6 Charles Darwin's personal interest in inbreeding depression

Charles Darwin, who is well known for his evolution theory, was married to his first cousin, Emma Wedgwood. Together they had ten children, who were often ill and of whom three died in childhood [83]. Darwin's work on negative consequences of self-fertilization in plant species [69] will likely have led him to suspect that the impaired health of his children may have been due to the marriage with his first cousin. His concern regarding inbreeding depression in humans is reflected by a letter he sent to his friend John Lubbock, member of Parliament, in which he asked Lubbock to request Parliament to include a question on consanguineous marriage in the 1871 Census of Great Britain and Ireland [84].

Charles Darwin (1809-1882)

In 2015, Álvarez et al. [85] showed that Darwin's concerns were legitimate. By studying 30 marriages in the Darwin-Wedgwood family, they found that higher inbreeding coefficients of fathers were associated with a reduced family size and a shorter reproductive period.

1.6 Conservation of genetic diversity: *in situ* and *ex situ*

Conservation efforts can be roughly divided into two classes: *in-situ* and *ex-situ* conservation. *In-situ* conservation refers to conservation of genetic diversity within breeding programs and production systems. *Ex-situ* conservation refers to conservation outside of such systems.

1.6.1 *In situ* conservation: optimal contribution selection

To conserve genetic diversity, it is important to restrict the loss of diversity within *in situ* populations. Therefore, it is important to limit the increase in kinship and

inbreeding in these populations. Inbreeding can be controlled at 2 levels: (1) at the level of mating, and (2) at the level of selection. At the level of mating, one could apply minimum-coancestry mating, i.e. mate sires and dams such that the kinship between them is minimized [86]. In the short term, this strategy may be very effective, because it minimizes inbreeding of the offspring. In the long term, however, it is not optimal, because it does not restrict the increase in mean kinship in the population [87]. Therefore, kinship and inbreeding should also be controlled at the level of selection. Examples are selection of individuals that have a low mean kinship with the rest of the population and the use of sire restrictions [87, 88].

At the level of selection, optimal contribution selection (OCS) is considered the golden standard to (1) maximise genetic gain while restricting the rate of inbreeding to a predefined value, or to (2) minimise the rate of inbreeding irrespective of genetic gain [89, 90]. In OCS, it is determined how much each selection candidate should contribute to the next generation to achieve the predefined objective (Box 1.7). Although originally based on pedigree information, OCS can also be used with genomic information [91]. Because OCS requires full control over the selection process, its implementation is not always feasible.

Box 1.7 Optimal contribution selection (OCS)

In optimal contribution selection (OCS) it is determined how much each of the n selection candidates from the current generation (t) should contribute to the next generation ($t+1$) to achieve a predefined objective. The traditional objective of pedigree-based OCS is to maximize the mean estimated breeding value in the next generation ($\overline{EBV}_{t+1}$), while restricting the mean pedigree-based kinship in the next generation ($\bar{f}_{PED_{t+1}}$) to a certain value. The $\overline{EBV}_{t+1}$ and $\bar{f}_{PED_{t+1}}$ are expressed as:

$$\overline{EBV}_{t+1} = \mathbf{c_t'EBV_t}$$

$$\bar{f}_{PED_{t+1}} = \mathbf{c_t'A_tc_t}/2$$

where $\mathbf{c_t}$ is a $(n \times 1)$ vector of contributions of the selection candidates to the next generation, $\mathbf{EBV_t}$ is a $(n \times 1)$ vector of estimated breeding values of the selection candidates, and $\mathbf{A_t}$ is a $(n \times n)$ matrix with pedigree-based relationships between the selection candidates (there is a division by two in the formula, because these relationships are two times the kinships). Contributions should be non-negative and are constrained to sum up to 0.5 per sex [89].

The concept of genetic contributions was first quantified in 1958 [92] and a relationship between long-term contributions and the rate of inbreeding was proven in 1990 [93]. The first algorithm for OCS was developed in the late 1990s [89]. This

algorithm uses Lagrangian multipliers to solve the optimisation problem. Over time, alternative algorithms have been proposed [94, 95].

1.6.2 *Ex situ* conservation: gene bank collections

For long-term conservation of genetic diversity, *in situ* conservation should be complemented by *ex situ* conservation. *Ex situ* conservation can be *in vivo*, by keeping live populations outside of production systems (e.g. in zoos or research farms), or *in vitro*, by cryoconserving genetic material (semen, ova, embryos or tissues) in gene banks. According to the latest report of FAO on global animal genetic resources, *in vitro* gene bank collections have been established by 64 countries and another 41 countries are planning to do so [2].

Within European *in vitro* gene banks, the most commonly stored type of material is semen and the most commonly stored species are cattle, sheep, horse, pig and goat [96, 97]. In the Netherlands, for example, a substantial semen collection has been established for most of the Dutch native breeds and for several transboundary breeds of various species (Table 1.1).

Table 1.1 Number (N) of breeds, donors and straws in the Dutch gene bank collection per species, as well as the birth years of the donors (status 2019).

Species	N breeds	N donors	N straws	Birth year of donors
Cattle	23	6,378	253,629	1966-2017
Chicken	31	270	18,662	1985-2009
Dog	7	19	612	1988-2012
Duck	3	67	1,591	2011-2013
Goat	6	82	6,476	2005-2015
Goose	1	11	102	2013-2014
Horse	13	253	4,538	1979-2017
Pig	33	767	21,946	1995-2017
Rabbit	8	62	1,957	2014-2015
Sheep	11	336	31,567	2001-2015

Despite the many gene bank collections worldwide, relatively little is known about these collections. To better understand the value of gene bank collections and to enhance their use, the project 'Innovative Management of Animal Genetic Resources' (IMAGE) was set up. The ultimate goal of this Horizon 2020 project was to demonstrate the benefits brought by gene banks for a more sustainable livestock production. This thesis is one of the scientific outputs of the IMAGE project.

1.7 Genetic diversity in Dutch cattle

In the Netherlands, as in most European countries, dual-purpose cattle were preferred over specialized dairy or beef types up to the 1970s [98]. Three native dual-

purpose breeds dominated the Dutch cattle population: Dutch Friesian (76%), Dutch Red and White or 'MRY' (22%), and Groningen White Headed (2%) [99]. Then, the 'Holsteinization' took place; HF bulls from the United States were used to upgrade cattle populations across the world. In the Netherlands, this resulted in the development of a HF dairy cattle population at the expense of the dual-purpose breeds. Nowadays, more than 98% of the Dutch milk-recorded population is HF [100]. The decrease in population sizes of native breeds has increased the risk of losing genetic diversity in these breeds.

Within HF, intense artificial selection has resulted in high realized genetic gains over time [23]. At the same time, the extensive use of a limited number of high-performing AI bulls has reduced the effective population size to less than 100 individuals [101, 102]. In recent decades, the Dutch-Flemish HF breeding program has undergone a few major changes that may have affected genetic diversity trends. One of these changes is the introduction of genomic selection (Section 1.2). Since 2009, animals are largely selected on genomic breeding values, rather than on breeding values estimated from progeny testing. This has accelerated genetic progress [103, 104] and has changed the selection process. Although the effect of genomic selection on diversity has been investigated in various theoretical and simulation studies [41, 91, 105], the effects in real life populations are largely unknown.

Since the early 1990s, genetic material from Dutch cattle breeds has been stored in the national gene bank collection. For the HF breed, 25 straws per AI bull have been stored. For native breeds, storage of material has been less systematic and has largely depended on availability of samples and financial resources. The current and future value of this stored material is largely unknown.

1.8 Aim and outline of thesis

The overall aim of this thesis was to utilize the wide availability of SNP data to obtain a better understanding of genetic diversity in Dutch cattle and how this diversity can be conserved. HF was used as main breed of interest, because of the national and global significance of this breed and the vast amount of genomic data available.

The first objective was to demonstrate how major changes in the Dutch-Flemish HF breeding program, such as the implementation of genomic selection, may have affected genetic diversity. To do so, we evaluated trends in genome-wide and region-specific genetic diversity from 1986 to 2015 (**Chapter 2**). The second objective was to better understand inbreeding depression for a variety of yield, fertility and health traits in HF cattle. We first compared the effects of ancient and recent inbreeding (**Chapters 3 & 4**), using pedigree-based and genomic approaches. We hypothesized that ancient inbreeding would be less harmful than recent inbreeding, because of

purging. We then evaluated heterogeneity in inbreeding depression across the genome (**Chapter 5**), with the hypothesis that inbreeding in some genomic regions might be more harmful than inbreeding in other regions. The third objective was to demonstrate the value of the Dutch cattle gene bank for conservation of genetic diversity. We first assessed the value of using old HF bulls from the gene bank in the current (or future) breeding program, both in terms of genetic diversity and genetic merit (**Chapter 6**). We then characterized genetic diversity conserved in the gene bank for native breeds and determined which genetic material could be stored in core sets in which expected heterozygosity was maximized (**Chapter 7**). Finally, in the general discussion (**Chapter 8**), I put the findings of this thesis in a wider context. I discuss the conservation of genetic diversity based on genomic information using approaches like OCS, describe the (potential) role of gene bank collections and stress the importance of genetic diversity for future livestock production. A schematic outline of the thesis, with the aim per chapter, is shown in Figure 1.2.

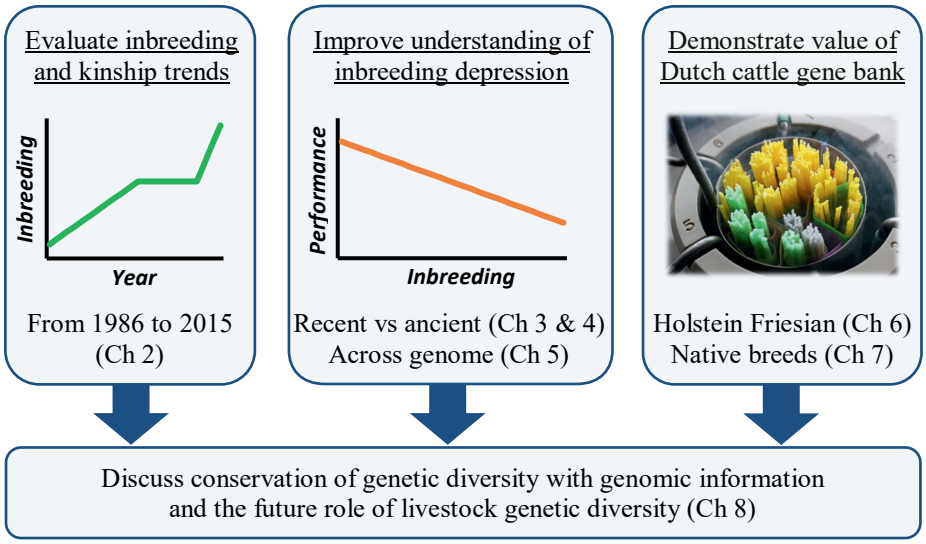


Figure 1.2 Outline of thesis with topic per chapter (Ch).

2

Trends in genome-wide and region-specific genetic diversity in the Dutch-Flemish Holstein Friesian breeding program from 1986 to 2015

Harmen P. Doekes[1,2], Roel F. Veerkamp[1], Piter Bijma[1],
Sipke J. Hiemstra[2], Jack J. Windig[1,2]

[1]Wageningen University & Research Animal Breeding and Genomics,
P.O. Box 338, 6700 AH, Wageningen, the Netherlands;
[2]Centre for Genetic Resources the Netherlands, Wageningen University &
Research, P.O. Box 16, 6700 AA, Wageningen, the Netherlands

Genetics Selection Evolution (2018) 50:15

Abstract

In recent decades, Holstein–Friesian (HF) selection schemes have undergone profound changes, including the introduction of optimal contribution selection (OCS; around 2000), a major shift in breeding goal composition (around 2000) and the implementation of genomic selection (GS; around 2010). These changes are expected to have influenced genetic diversity trends. Our aim was to evaluate genome-wide and region-specific diversity in HF artificial insemination (AI) bulls in the Dutch-Flemish breeding program from 1986 to 2015.

Pedigree and genotype data (~ 75.5 k) of 6280 AI-bulls were used to estimate rates of genome-wide inbreeding and kinship and corresponding effective population sizes. Region-specific inbreeding trends were evaluated using regions of homozygosity (ROH). Changes in observed allele frequencies were compared to those expected under pure drift to identify putative regions under selection. We also investigated the direction of changes in allele frequency over time.

Effective population size estimates for the 1986–2015 period ranged from 69 to 102. Two major breakpoints were observed in genome-wide inbreeding and kinship trends. Around 2000, inbreeding and kinship levels temporarily dropped. From 2010 onwards, they steeply increased, with pedigree-based, ROH-based and marker-based inbreeding rates as high as 1.8, 2.1 and 2.8% per generation, respectively. Accumulation of inbreeding varied substantially across the genome. A considerable fraction of markers showed changes in allele frequency that were greater than expected under pure drift. Putative selected regions harbored many quantitative trait loci (QTL) associated to a wide range of traits. In consecutive 5-year periods, allele frequencies changed more often in the same direction than in opposite directions, except when comparing the 1996–2000 and 2001–2005 periods.

In conclusion, genome-wide and region-specific diversity trends reflect major changes in the Dutch-Flemish HF breeding program. Introduction of OCS and the shift in breeding goal were followed by a drop in inbreeding and kinship and a shift in the direction of changes in allele frequency. After introduction of GS, rates of inbreeding and kinship increased substantially while allele frequencies continued to change in the same direction as before GS. These results provide insight in the effect of breeding practices on genomic diversity and emphasize the need for efficient management of genetic diversity in GS schemes.

2.1 Introduction

Genetic variation in (closed) livestock populations is largely driven by the fundamental processes of selection and genetic drift. While selection acts directionally on alleles that have a selective (dis)advantage and on alleles that are 'hitchhiking' [41, 42, 106], genetic drift acts across the whole genome, causing random changes in allele frequency from generation to generation as a result of sampling gametes in a finite population [24].

In Holstein Friesian dairy cattle (HF), intense artificial selection has been practiced over many years. The use of a limited number of elite sires has reduced the effective population to a size ranging from 49 to 115 [102, 107, 108]. This implies that, in spite of its census size of millions of individuals, the breed is subjected to the same rate of genetic drift and accumulation of inbreeding as an idealized population of 49 to 115 individuals [24]. To ensure adaptive capacity and limit inbreeding depression in the long term, it is important to monitor and manage genetic diversity in the HF population [53, 80].

Traditionally, genetic diversity has been characterized and managed with pedigree-based coefficients of inbreeding and kinship, which refer to the proportion of the genome that is expected to be identical by descent (IBD) within and between individuals, respectively. However, this genealogical approach has several limitations: (i) it strongly depends on pedigree completeness and quality [48]; (ii) it does not account for Mendelian sampling variation [49]; and (iii) it only provides a genome-wide expectation for loci that are selection-free, i.e. loci that are in complete linkage equilibrium with all loci under selection [109].

With the wide availability of dense single nucleotide polymorphism (SNP) data, it has become possible to obtain more accurate estimates of genome-wide inbreeding and kinship and to evaluate diversity for specific regions of the genome [52, 91, 110]. Two approaches have been widely used to characterize and manage diversity from SNP data: the marker-by-marker approach [51] and the segment-based approach [62, 63]. The former approach involves the calculation of the observed and expected fraction of SNPs for which alleles are identical by state (IBS). Thus, it captures relationships that are caused by common ancestors going back to a very distant theoretical base population in which all alleles were unique. The second approach considers IBS segments, rather than individual SNPs. Since the length of these segments follows an inverse exponential distribution with expectation $1/2G$ Morgan [111], where G is the number of ancestral generations to the common ancestor from which the segment was derived, this approach may be used to distinguish recent from distant relatedness and move from IBS to 'realized IBD' [62]. Both IBS and IBD are relevant for management. While IBS is the most direct diversity measure, (realized) IBD is more closely associated to inbreeding depression [63, 112, 113].

In recent decades, HF selection schemes have undergone profound changes with respect to inbreeding management, breeding goal composition and breeding value estimation. Around the year 2000, optimal contribution selection (OCS) was introduced to maximize genetic gain at a restricted rate of inbreeding [89]. Around the same time, national selection indices moved from production- and conformation-based only to more comprehensive indices that included traits related to production, conformation, longevity, health and reproduction [37]. More recently, genomic selection (GS) was introduced, which enabled the prediction of high-accuracy breeding values at a young age [30]. Since all these changes cause rearrangements in the ranking of artificial insemination (AI) bulls, they are expected to have influenced trends in genome-wide and region-specific genetic diversity. With the current availability of SNP-data, it is now possible to investigate this influence.

The aim of this study was to evaluate genome-wide and region-specific genetic diversity in HF AI bulls from 1986 to 2015, using genealogical, marker-by-marker and segment-based approaches. An important objective was to evaluate whether major changes in the Dutch-Flemish HF breeding program were accompanied by changes in inbreeding and kinship trends. A second objective was to investigate whether observed changes in allele frequency could be attributed to selection, and whether regions under selection could be linked to known quantitative trait loci (QTL). A last objective was to investigate how the direction of changes in allele frequency has evolved over time.

2.2 Material and methods

2.2.1 Animals and data

A total of 6,280 AI bulls (breed fraction ≥ 87.5% HF), born between 1986 and 2015 and genotyped by the Dutch-Flemish cattle improvement co-operative (CRV), were included in this study. Thus, the vast majority of AI bulls in the Dutch-Flemish breeding program were included. Figure 2.1 shows the number of bulls by birth year.

Pedigrees were extracted from the database of CRV and extended with publicly available data [114]. The total pedigree comprised 46,232 animals. Complete generation equivalents (CGE) were computed as the sum of $(1/2)^n$ over all known ancestors, with n being the generation number of a given ancestor. The average CGE increased from 9.6 in 1986 to 17.0 in 2015 and was equal to 13.3 when calculated across all years. The average number of completely known generations increased from 4.1 in 1986 to 8.1 in 2015. The generation interval (L), i.e. the average age of parents at birth of the bulls, was computed per year of birth for bull sires and bull dams separately, and for all parents combined (Figure 2.2). L decreased during the first decade and then increased slightly until it dropped steeply from 2009 onwards. The initial drop in L can be explained by an increased use of young unproven bull

sires, which, at the time, was expected to improve genetic gain. However, due to variable gains, the trend changed and, from 1998 onwards, almost exclusively proven bull sires were used. The drop in L from 2009 onwards was especially pronounced for bull sires and followed the implementation of GS. The average L across the whole 30-year period and for all parents combined was 5.0 years.

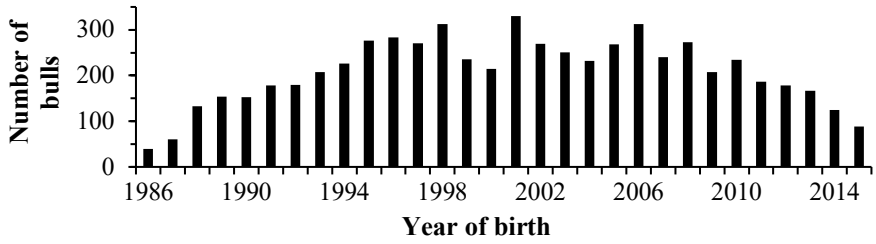


Figure 2.1 Number of genotyped bulls by year of birth.

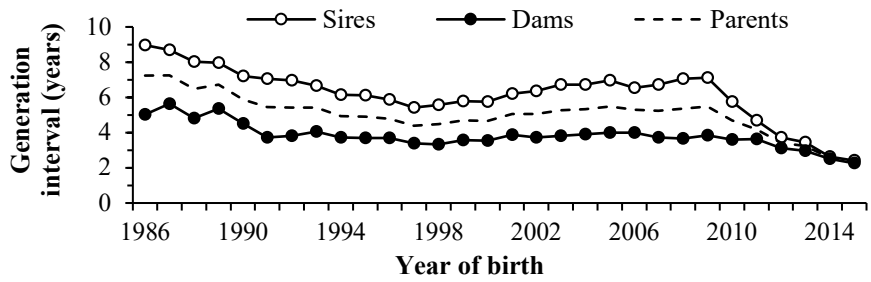


Figure 2.2 Generation interval for bull sires, bull dams and bull parents by year of birth.

Genotype data were provided by CRV and the final dataset comprised 75,538 autosomal SNPs. Bulls were genotyped with the Illumina BovineSNP50 BeadChip (versions v1 and v2) or CRV custom-made 60 k Illumina panel (versions v1 and v2). Genotypes were imputed to ~76 k from the different panels, following Druet et al. [115], and haplotypes were constructed with a combination of Beagle [116] and PHASEBOOK [117], by exploiting both familial and population information. Prior to imputation, SNPs with a call rate lower than 0.85, a MAF lower than 0.025 or a difference higher than 0.15 between observed and expected heterozygosity were discarded. SNP positions were obtained from the Btau4.0 genome assembly and those with unknown positions (N = 893) were discarded. The mean physical distance between two consecutive SNPs was 33.7 kb, with density varying substantially across the genome (see Figure S2.1 for the density). Black and white (N = 5021) and red and white (N = 1259) bulls were combined in all analyses, because a preliminary check on the mean SNP-based kinship within and between bulls of both groups indicated no major genetic differentiation across the 30-year period.

2.2.2 Genome-wide diversity

Genome-wide diversity was quantified with genealogical, marker-by-marker and segment-based approaches. Pearson correlations between genealogical, marker-by-marker and segment-based measures were calculated to compare the approaches.

Genealogical inbreeding and kinship

Genealogical coefficients of inbreeding (F_{PED_i}) and kinship ($f_{PED_{ij}}$) were defined as the pedigree-based probabilities that two alleles at an (imaginary) selection-free locus, sampled respectively within individual i or between individuals i and j, were IBD with reference to a base population [24]. Founders in the pedigree were considered as base population. Both F_{PED_i} and $f_{PED_{ij}}$ were calculated with *calc_grm* [118], according to the algorithms of Sargolzaei et al. [119] and Colleau [120].

Marker-by-marker homozygosity and similarity

Marker-by-marker homozygosity (HOM_{SNP_i}) and similarity ($SIM_{SNP_{ij}}$) were defined as the probabilities that two alleles at a random SNP, which were sampled respectively within individual i or between individuals i and j, were IBS. The HOM_{SNP_i} was obtained as the proportion of SNPs for individual i that were homozygous. The $SIM_{SNP_{ij}}$ was determined according to Malécot [51]:

$$SIM_{SNP_{ij}} = \frac{\sum_{k=1}^{n_{SNP}} (I_{11,k} + I_{12,k} + I_{21,k} + I_{22,k})}{4n_{SNP}}$$

where n_{SNP} is the total number of markers, $I_{xy,k}$ is an indicator variable that was set to 1 when allele x of individual i and allele y of individual j at marker k were IBS, and to 0 otherwise. Note that $SIM_{SNP_{ij}}$ is equivalent to VanRaden's genomic relationship G_{ij} [55] when allele frequencies of 0.5 are used in the computation of G_{ij} (except for the scale; see Additional file 1 of Eynard et al. [121] for derivation). Since self-similarities ($SIM_{SNP_{ii}} = \frac{1}{2}[1 + HOM_{SNP_i}]$) were included, the average similarity in a given cohort was also equivalent to the expected homozygosity in that cohort (i.e. the average sum of squared allele frequencies, $p^2 + q^2$, across all SNPs).

Segment-based inbreeding and kinship

Segment-based inbreeding (F_{ROH_i}) was defined as the proportion of the genome of individual i that was covered by long uninterrupted homozygous segments. Such regions of homozygosity (ROH) were detected by moving SNP by SNP across chromosomes and testing potential ROH against predefined criteria. The following criteria were used to define a ROH: (i) a minimum physical length of 3.75 Mb, (ii) a minimum of 38 consecutive homozygous SNPs (no heterozygous calls allowed), and

(iii) a maximum gap of 500 kb between two consecutive SNPs. The minimum length of 3.75 Mb was chosen to match the pedigree depth. Given the genetic distance of approximately 1 cM per Mb [122] and the average length of $1/2G$ Morgan for ROH derived from a common ancestor G generations ago [111], the F_{ROH_i} was expected to capture inbreeding over 13.3 ancestral generations (corresponding to the CGE of the pedigree). The latter two criteria were used to prevent calling of (potentially false positive) ROH in regions with low SNP density. The F_{ROH_i} was calculated as the fraction of the autosome in ROH [62]:

$$F_{ROH_i} = \frac{\sum_{m=1}^{n_{ROH_i}} l_{ROH_{i,m}}}{l_a}$$

where n_{ROH_i} is the total number of ROH in individual i, $l_{ROH_{i,m}}$ is the length of the m^{th} ROH and l_a is the length of the autosome covered by SNPs (i.e. the autosome length minus the summed length of gaps longer than 500 kb).

Segment-based kinship ($f_{SEG_{ij}}$) was defined as the expected F_{ROH} for an offspring of individuals i and j. Shared segments were identified by moving SNP by SNP across every possible pair of chromosomes, with one homolog of individual i and one of j, and testing potential segments against predefined criteria. The same criteria were used as for calling ROH. The $f_{SEG_{ij}}$ was computed following de Cara et al. [63]:

$$f_{SEG_{ij}} = \frac{\sum_{m=1}^{n_{SEG_{ij}}} \sum_{x_i}^{2} \sum_{y_j}^{2} l_{SEG_{ij,m}}}{4 l_a}$$

where $n_{SEG_{ij}}$ is the total number of shared segments between i and j, $l_{SEG_{ij,m}}$ is the length of the m^{th} shared segment measured over homolog x of individual i and homolog y of individual j and l_a is the length of the autosome covered by SNPs.

Rate of change and effective population size

For each genome-wide parameter, the annual rate of change (Δx_y) was obtained as the opposite of the slope of the regression of $LN(1-\bar{x})$ on year of birth, where $\bar{x}$ equaled the average of the parameter in a given year [123]. The annual rate was multiplied by L to obtain the rate per generation (Δx_{gen}) and, subsequently, the effective population size ($N_e = 1/(2\Delta x_{gen})$). To investigate trends over time, Δx_y and Δx_{gen} were calculated for five-year periods, taking changes in L into account.

2.2.3 Region-specific inbreeding

Accumulation of inbreeding across the genome over time was evaluated with ROH-based positional inbreeding coefficients. For every marker k in bull i, a positional

inbreeding coefficient ($F_{ROH_{i,k}}$) was set to 1 when k was encompassed by a ROH, and to 0 otherwise, following Kim et al. [124]. The F_{ROH_k} per five-year period was then calculated as the fraction of bulls born in that period for which k was encompassed by a ROH.

2.2.4 Changes in allele frequency and putative selected regions

Changes in allele frequency were computed as $\Delta p = p_t - p_0$, where p_t and p_0 were the frequency in the last (2011-2015) and first (1986-1990) five-year period, respectively. Since the average L was 5.0 years, the Δp-values were based on approximately five generations of drift and selection. To identify putative selected regions, the observed Δp-values were compared to those expected under pure genetic drift. The Δp-distribution under pure drift was obtained by gene dropping [125]. In each simulated gene drop, alleles for a single SNP were randomly assigned to founders and subsequently dropped through the pedigree following Mendelian principles (i.e. random sampling). To ensure a wide spectrum of p_0-values, founder minor allele frequencies (MAF) ranging from 0.5 to 50% were simulated. Realised p_0-values were classified into 100 MAF-classes, ranging from 0.0-0.5% to 49.5-50.0%, and the drift distribution per MAF-class was obtained based on 3000 replicates. Observed Δp-values above the 99.9% threshold ($P < 0.001$) of the empirical gene drop distribution were considered indicative of selection. To visualize systematic changes over the erratic pattern of individual SNPs, the moving average of 31 adjacent Δp-values was plotted against the physical position of the central SNP.

Genomic regions with an excess of putative selected SNPs were considered as putative selected regions. For the key regions of interest, we investigated which QTL were known in these regions, using AnimalQTLdb [126]. The complete CattleQTLdb, which contains 99,675 QTL, was first filtered; QTL mapped to chromosome X (N = 25,589), reported for non-HF breeds (N = 23,468) and/or with unknown start and end positions (N = 1737) were discarded. In addition, QTL associated to traits that were not clearly related to the Dutch-Flemish breeding bull-selection index, such as specific milk fatty acids or carcass traits, were removed (N = 21,195). This resulted in a final list of 27,662 QTL, associated to 61 traits classified in five trait categories: production (INET), conformation (CONF), longevity (LONG), reproduction (REPR) and udder health (UH). The final list of traits and number of QTL per trait and trait category is included as supplementary information (Table S2.1).

Changes in allele frequency were also computed within each five-year period as $\Delta p = p_t - p_0$, with p_t and p_0 being the frequencies in the last and first year of the period, respectively (e.g. $\Delta p = p_{1990} - p_{1986}$). Correlation coefficients between the Δp-values of the different five-year periods were calculated to investigate the direction of changes in allele frequency over time.

2.3 Results

2.3.1 Genome-wide diversity

Descriptive statistics for genome-wide parameters are shown in Table 2.1. The mean genealogical inbreeding and kinship were 5.2 and 6.5%, respectively. Segment-based coefficients were on average ~1.5% higher than genealogical coefficients. As expected, IBS coefficients showed a higher mean (64.4% for HOM_{SNP} and 64.8% for SIM_{SNP}), lower SD and lower CV than IBD coefficients. For all kinship parameters, the mean was higher than the median, which was indicative of the right-skewedness of the underlying distributions that was due to inclusion of self-kinships.

Table 2.1 Descriptive statistics for genome-wide inbreeding and kinship parameters in all years combined. Values are shown in percentages.

Parameter	N	Mean	SD	Median	Minimum	Maximum	CV
F_{PED}	6,280	5.21	2.25	5.10	0.00	17.88	0.432
F_{ROH}	6,280	6.75	2.89	6.43	0.67	25.38	0.429
HOM_{SNP}	6,280	64.36	1.18	64.22	58.43	71.84	0.018
f_{PED}	1,470,166	6.54	4.58	5.69	0.26	58.94	0.701
f_{SEG}	1,470,166	7.99	4.61	7.14	0.01	62.69	0.577
SIM_{SNP}	1,470,166	64.82	1.78	64.47	61.53	85.92	0.027

N: number of coefficients; SD: standard deviation; CV: coefficient of variation; F_{PED} and f_{PED}: genealogical inbreeding and kinship; F_{ROH} and f_{SEG}: segment-based inbreeding and kinship; HOM_{SNP} and SIM_{SNP}: marker-by-marker homozygosity and similarity.

Correlations between kinship parameters were considerably higher than those between inbreeding coefficients (Figure 2.3). Across all years, the highest correlations were found between the genomic parameters (on average 0.90 for HOM_{SNP} with F_{ROH} and 0.98 for SIM_{SNP} with f_{SEG}) and the lowest between the marker-by-marker and genealogical estimates (on average 0.60 for HOM_{SNP} with F_{PED} and 0.92 for SIM_{SNP} with f_{PED}). Correlations between genomic parameters remained relatively constant over years, whereas correlations between pedigree and genomic parameters decreased over time. For example, the correlation between f_{SEG} and f_{PED} decreased from 0.97 in 1986 to 0.88 in 2015. This divergence could be explained by the accumulation of Mendelian sampling variation over time, which is captured by genomic information, but not by pedigree data. When more generations are included in the calculation of f_{PED}, more sampling events are unaccounted for and f_{PED} is likely to deviate more from the realized genomic relationship. Correlations between pedigree and genomic inbreeding parameters seemed to increase slightly from 2009 onwards. However, this increase could also be due to random fluctuations, as the standard errors for inbreeding correlations were rather large (Figure 2.3).

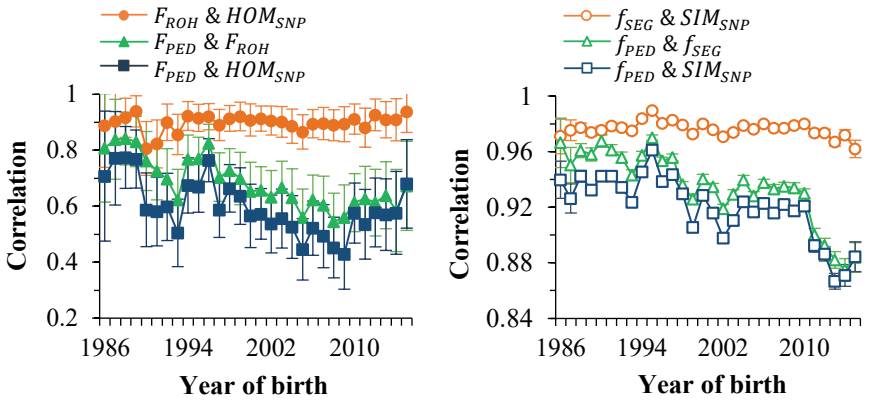


Figure 2.3 Correlations between different genome-wide estimates of inbreeding (left) and kinship (right) by year of birth. Note the different scales for the y-axes for inbreeding and kinship. Self-kinships were excluded from the computation to remove the influence of the number of bulls per year on the correlations. Error bars represent ± 2 standard errors. F_{PED} and f_{PED}: genealogical inbreeding and kinship; F_{ROH} and f_{SEG}: segment-based inbreeding and kinship; HOM_{SNP} and SIM_{SNP}: marker-by-marker homozygosity and similarity.

Roughly, genome-wide inbreeding increased from 1986 to 2000, remained rather constant for a decade and then steeply increased from 2011 onwards (Figure 2.4). Genome-wide kinship levels fluctuated more, but also increased from 1986 to 2000, temporarily dropped and then remained rather constant until a steep increase from 2009 onwards.

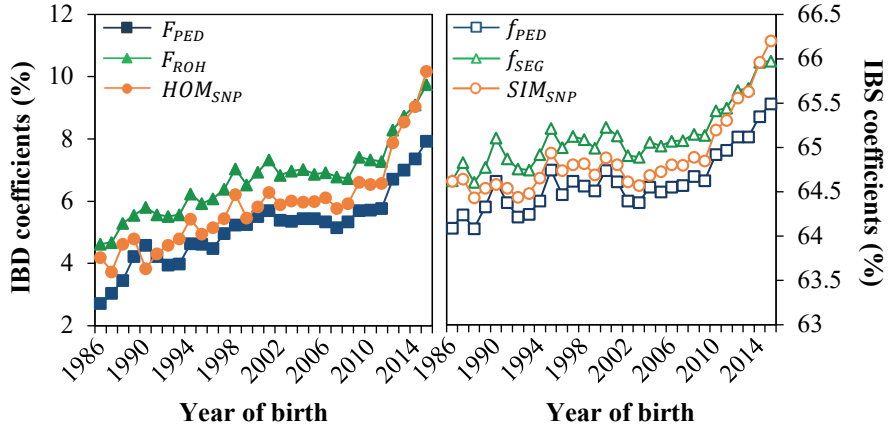


Figure 2.4 Average genome-wide inbreeding (left) and kinship (right) by year of birth. Coefficients of IBD (F_{PED}, F_{ROH}, f_{PED}, f_{SEG}) and IBS (HOM_{SNP}, SIM_{SNP}) are shown on the primary and secondary y-axis, respectively. F_{PED} and f_{PED}: genealogical inbreeding and kinship; F_{ROH} and f_{SEG}: segment-based inbreeding and kinship; HOM_{SNP} and SIM_{SNP}: marker-by-marker homozygosity and similarity.

Genome-wide rates of change per year and per generation for the 1986-2015 period are shown in Table 2.2. Estimates of N_e computed from ΔF_{PED}, ΔF_{ROH} and ΔHOM_{SNP} were equal to 79, 75 and 69, respectively. Rates of kinship were lower than rates of inbreeding, with a N_e estimated from Δf_{PED}, Δf_{ROH} and ΔSIM_{SNP} of 102, 100 and 91, respectively. The difference between inbreeding and kinship rates was largely due to the relatively high kinship levels in early years (Figure 2.4). In fact, the average kinship at the beginning of the period was more than two generations ahead of the average inbreeding, while a difference of a single generation is expected for a randomly mating population.

Table 2.2 Genome-wide rates of change and effective population size (N_e) for the period 1986-2015.

Parameter	Rate of change per year (%)	Rate of change per generation (%)	N_e
F_{PED}	0.1280	0.6354	78.67
F_{ROH}	0.1342	0.6663	75.04
HOM_{SNP}	0.1462	0.7261	68.86
f_{PED}	0.0984	0.4887	102.31
f_{SEG}	0.1001	0.4991	100.19
SIM_{SNP}	0.1108	0.5502	90.88

F_{PED} and f_{PED}: genealogical inbreeding and kinship; F_{ROH} and f_{SEG}: segment-based inbreeding and kinship; HOM_{SNP} and SIM_{SNP}: marker-by-marker homozygosity and similarity.

Rates of inbreeding and kinship were also computed for periods of five years, accounting for the change in L over time. Both rates per year and per generation decreased over the first four periods, were slightly negative between 2001 and 2005 and increased in the last two periods (Figure 2.5). In the 2011-2015 period, rates of ΔF_{PED}, ΔF_{ROH} and ΔHOM_{SNP} were as high as 1.8, 2.1 and 2.8% per generation, respectively. Rates of change were very similar across the three approaches, except in the first, third and last period. In the 1986-1990 period, the ΔHOM_{SNP} and ΔSIM_{SNP} were close to zero as a result of large fluctuations in IBS levels (Figure 2.4). In this period, ΔF_{PED} was also relatively high (i.e. 1% higher per generation than ΔF_{ROH}). In the 1996-2000 period, genealogical rates of inbreeding were slightly higher (0.1 to 0.2% higher per generation) than segment-based rates, which, in turn, were slightly higher (0.2 to 0.3%) than marker-based rates. In the last period, which showed almost no fluctuations, marker-based rates were considerably higher (0.7% per generation) than segment-based rates, which were in turn slightly higher (0.3% for ΔF and 0.1% forΔf) than genealogical rates of inbreeding.

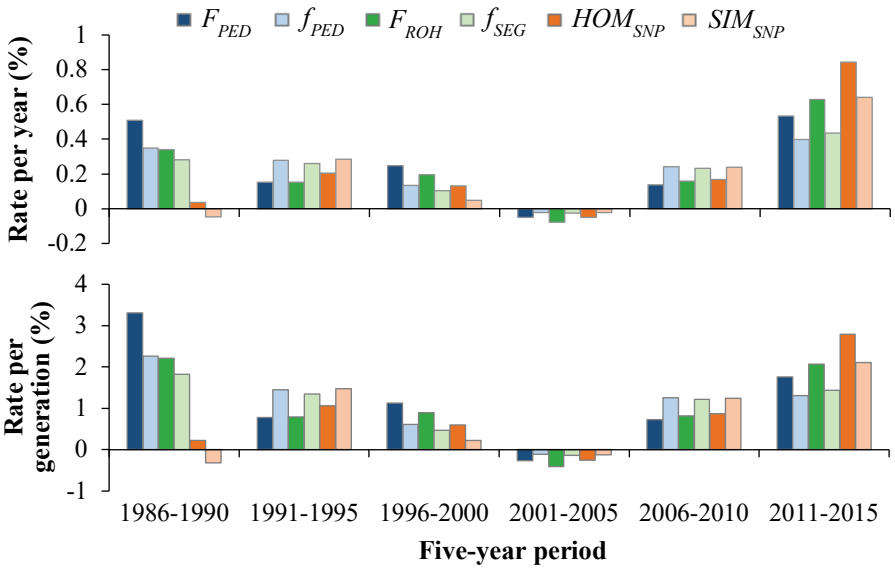


Figure 2.5 Rate of change per year (top) and generation (bottom) for genome-wide parameters within five-year periods. F_{PED} and f_{PED}: genealogical inbreeding and kinship; F_{ROH} and f_{SEG}: segment-based inbreeding and kinship; HOM_{SNP} and SIM_{SNP}: marker-by-marker homozygosity and similarity.

2.3.2 Region-specific inbreeding

Accumulation of inbreeding across the genome was evaluated with ROH-based positional inbreeding coefficients (F_{ROH_k}). Substantial heterogeneity was observed in the levels of F_{ROH_k} over time (Figure 2.6). There were, among others, regions with a continuous increase in inbreeding (e.g. the peaks on BTA10), regions with an increase followed by a decrease (e.g. around 40 Mb on BTA26) and regions with a constant inbreeding level over time (e.g. BTA18). Particularly striking was the strong increase in F_{ROH_k} in the last period for various regions (e.g. around 55 Mb on BTA4, around 40 Mb on BTA14 and around 25 Mb on BTA22). Overall, BTA10 showed the most prominent increase in F_{ROH_k}, from 5% in the 1986-1990 period to 20-30% in the 2011-2015 period at the peak regions. BTA20 also showed regions with a F_{ROH_k} of 20-30% in the 2011-2015 period, but these peaks had already a higher F_{ROH_k} at the start of the 30-year period (of 10 to 15%). Within the high peak on BTA10, there was a remarkable trough near 62.5 Mb, which could be due to incorrect SNP positions on the reference genome Btau4.0 (the 12 SNPs in this region were mapped near 71.5 Mb on UMD3.1). The trough within the peak on BTA4, near 55 Mb, might also be the result of incorrect SNP positions, although for this region there was no inconsistency between Btau4.0 and UMD3.1 positions.

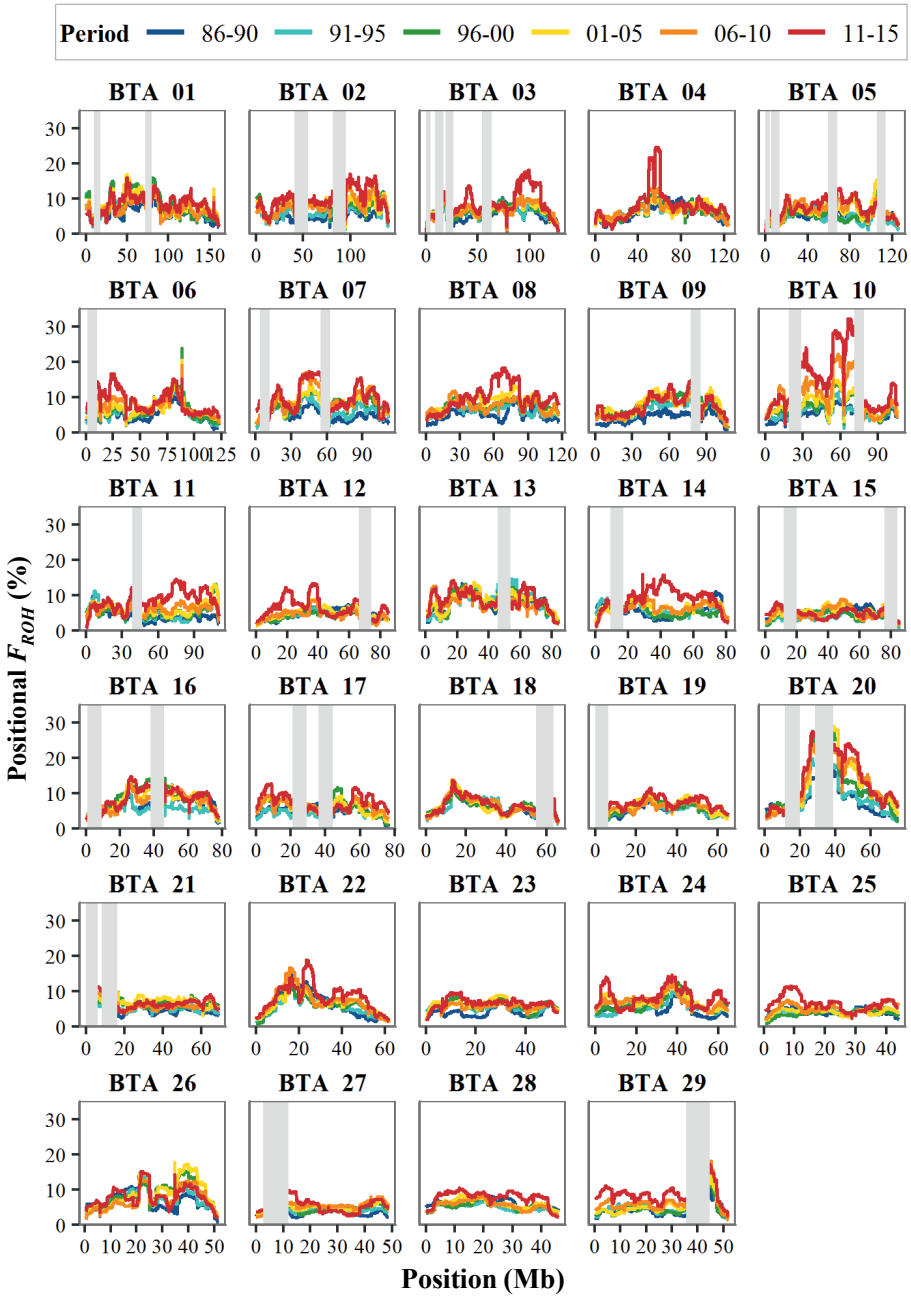


Figure 2.6 Positional inbreeding coefficients (F_{ROH}) per five-year period between 1986 and 2015. Grey bars cover gaps between consecutive markers of >500 kb (with an additional 3.75 Mb on both sides of the gap). BTA: Bos taurus autosome. Note that the scale of the x-axis differs between chromosomes.

2.3.3 Changes in allele frequency and putative selected regions

Absolute changes in allele frequencies from the 1986-1990 to the 2011-2015 period, $|\Delta p|$, were compared with those expected from gene dropping (Figure 2.7). Many SNPs showed higher $|\Delta p|$-values than would be expected under pure genetic drift. For example, there were 6,835 SNPs (9.05% of the total number) and 490 SNPs (0.65% of the total number) with a $|\Delta p|$ above the 95%- and 99.9%-thresholds of the gene drop distribution, respectively. The SNPs above the 99.9%-threshold were considered indicative of selection and, although they were spread across the whole genome, these SNPs were generally located in peaks of high $|\Delta p|$ (Figure 2.8). In line with the pattern observed for F_{ROH_k} (Figure 2.6), BTA10 showed the highest $|\Delta p|$ on average, with two wide peaks enriched with putative selected SNPs. However, on BTA20 no clear peak was observed and only three putative selected SNPs were detected. In contrast, BTA19 showed a narrow peak for $|\Delta p|$ that was not present in Figure 2.6. This could be explained by the extremely high SNP density in this region (Figure S2.1), which caused the moving average of 31 $|\Delta p|$-values to be based on a region of only 50 kb (while for ROH only regions longer than 3.75 Mb were considered). For 11 regions that were enriched with putative selected SNPs, we investigated whether QTL were known in these regions (Table 2.3). In general, the putative selected regions were large and overlapped with many QTL of different trait categories. Across all regions combined, there was a relatively large number of QTL for conformation traits and relatively few for production traits, when compared to QTL reported for the complete autosome. The relatively low fraction of QTL for production-traits could be explained by the fact that 39% of all production-QTL in the AnimalQTLdb are located on BTA14, whereas only a single short region on this chromosome was identified as a putative selected region.

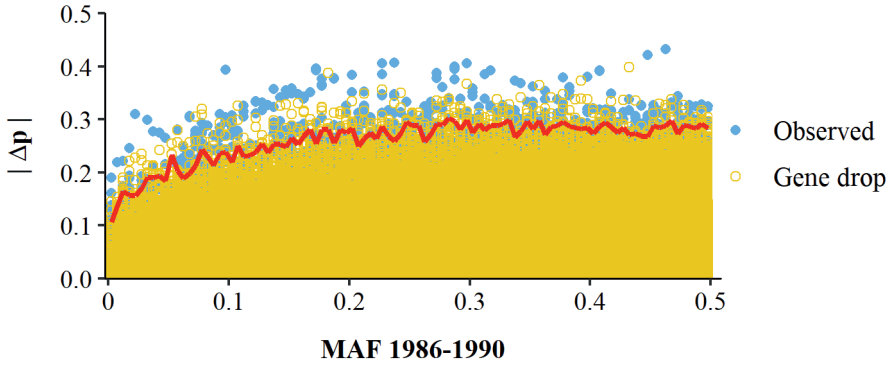


Figure 2.7 Absolute allele frequency changes from 1986-1990 to 2011-2015 ($|\Delta p|$) observed in data and gene drop. Changes are shown for different minor allele frequencies (MAF) in the 1986-1990 period, using MAF-classes of 0.005 (e.g. 0.0 to 0.005). The red line represents the 99.9%-threshold of the gene drop distribution per MAF-class.

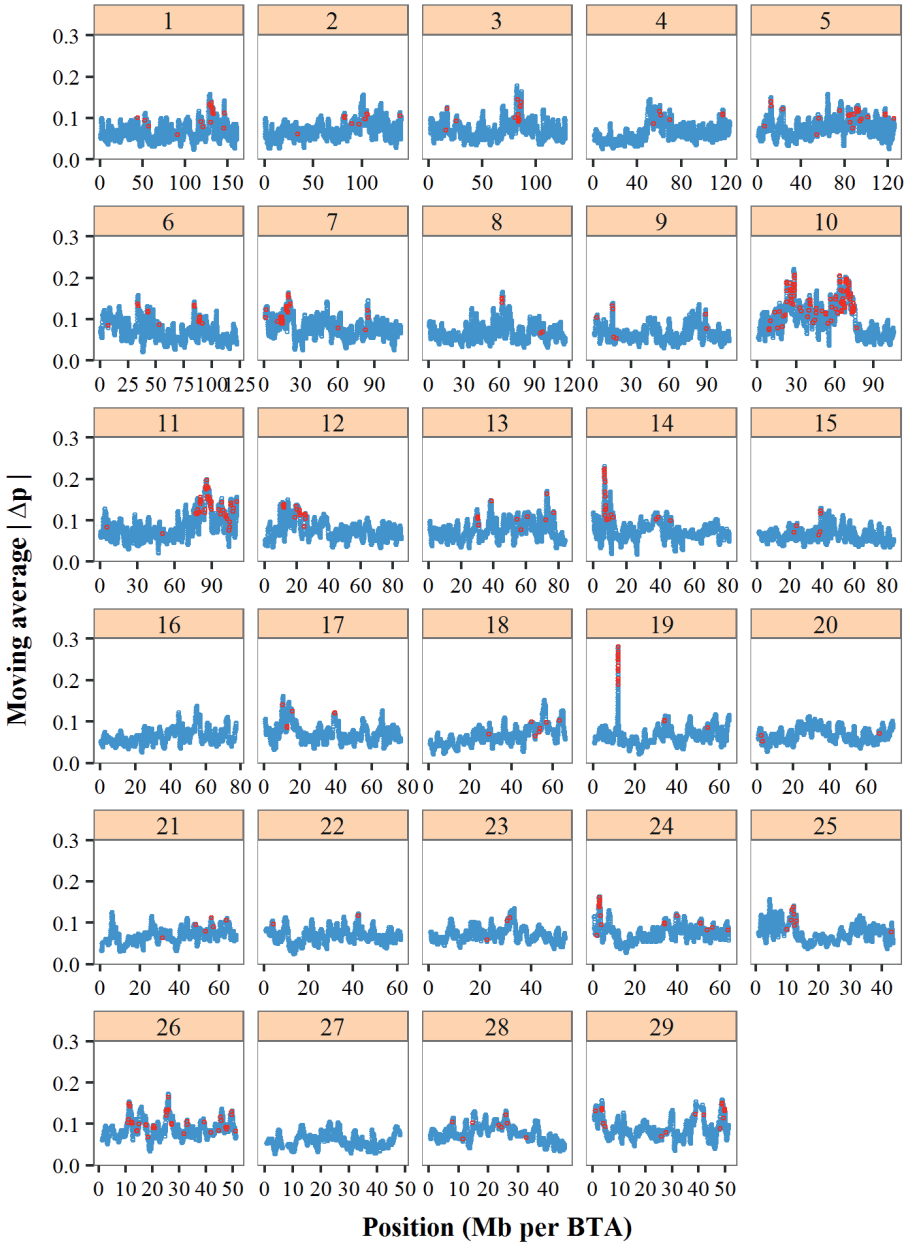


Figure 2.8 Moving average of changes in absolute allele frequency from the 1986-1990 to the 2011-2015 period ($|\Delta p|$). Moving average is based on 31 SNPs. SNPs in red (N = 490) have an allele frequency change above the 99.9%-threshold of the gene drop distribution (see Figure 2.7). Note that the scale of the x-axis differs between chromosomes.

Table 2.3 Putative selected regions based on changes in allele frequency from the 1986-1990 period to the 2011-2015 period and fraction of known QTL mapped to these regions per trait category.

			Fraction of QTL per trait category (%)				
BTA	Start – end (Mb)	n_{QTL}	INET	CONF	LONG	REPR	UH
1	128.0-133.0	83	4	34	8	46	8
3	80.0-86.0	68	6	47	7	31	9
7	10.5-22.0	169	7	36	19	31	8
10	19.0-29.0	43	37	30	5	23	5
10	60.0-75.0	111	9	76	2	9	5
11	76.0-89.5	342	15	70	1	14	1
12	19.0-26.0	42	29	21	7	38	5
14	6.0-8.0	44	91	2	2	2	2
19	11.5-12.0	1	0	0	0	0	100
24	1.5-4.0	26	35	27	8	27	4
26	25.0-27.5	30	17	50	3	23	7
Total putative selected regions		*959*	*17*	*51*	*6*	*22*	*4*
Total autosome		*27,662*	*38*	*25*	*8*	*26*	*3*

QTL were included when reported in AnimalQTLdb [126]. QTL were classified into five trait categories: INET (production index), CONF (conformation), LONG (longevity), REPR (reproduction) or UH (udder health). See Table S2.1 for classification of traits.

To evaluate the direction of allele frequencies over time, correlation coefficients between changes in allele frequency (Δp) within different five-year periods were calculated (Table 2.4). Except for the comparison between the 1996-2000 and 2011-2015 periods, all correlations were significantly different from 0 ($P < 0.0001$). Correlation coefficients for any two consecutive periods were positive (ranging from 0.08 to 0.26), except for the transition from the 1996-2000 period to the 2001-2005 period (-0.09).

Table 2.4 Correlations between allele frequency changes (e.g. $p_{1990} - p_{1986}$) within different five-year periods between 1986 and 2015.

Period	1986-1990	1991-1995	1996-2000	2001-2005	2006-2010
1991-1995	0.094				
1996-2000	0.089	0.082			
2001-2005	-0.062	-0.130	-0.094		
2006-2010	-0.028	-0.113	-0.087	0.092	
2011-2015	0.040	-0.041	0.001	-0.070	0.264

Standard errors of correlations ranged from 0.0004 (for 1996-2000 with 2011-2015) to 0.0035 (for 2006-2010 with 2011-2015).

2.4 Discussion

In this study, we evaluated genetic diversity across the genome of HF AI bulls from 1986 to 2015. An important objective was to investigate whether major changes in the Dutch-Flemish HF breeding program were accompanied by changes in diversity trends. We used genealogical, marker-by-marker and segment-based approaches to compare trends in expected IBD, IBS and realized IBD.

Genome-wide rates of inbreeding and kinship and corresponding estimates of N_e computed over the 1986-2015 period were similar to those previously reported for HF populations. Genealogical and genomic estimates of N_e for HF populations in Australia, Canada, Denmark, Spain, Ireland and the United States of America for (parts of) the 1975-2013 period range from 49 to 127 individuals [101, 102, 107, 108, 127]. A similar N_e across countries is expected, due to the extensive exchange of genetic material. Despite the global connectedness of the breed, there is some degree of genetic differentiation across countries [102, 128].

Genome-wide diversity trends showed two breakpoints. The first occurred around 2000, after which levels and rates of inbreeding and kinship temporarily dropped (Figures 2.4 and 2.5). The second occurred around 2010, after which inbreeding and kinship steeply increased. Both breakpoints coincided with major changes in the Dutch-Flemish breeding program.

The drop in inbreeding and kinship around 2000 followed a shift in breeding goal composition and the introduction of OCS. Although the Dutch-Flemish bull selection index has changed continuously over time, the major shift took place around 2000, when longevity, udder health and reproductive traits were added to the index within a few years' time (Table 2.5). The inclusion of a wide range of traits has resulted in a more diverse set of bulls with high estimated breeding values (EBV) and has thereby contributed to the (temporary) drop in inbreeding and kinship. From 2000 onwards, pedigree-based OCS has been applied to select bull-parents in the breeding program and restrict ΔF and Δf. However, the effect of OCS will have been limited due to practical difficulties. One such difficulty is that, in practice, not all candidates with allocated contributions are available for breeding. Another difficulty is that OCS considers all candidates at a single moment in time, while selection decisions in the breeding program are made on a daily basis. Despite these difficulties, the use of OCS will have restricted ΔF and Δf and its introduction will have contributed to the observed drop around 2000. A drop in ΔF and Δf around 2000 has also been observed in the Canadian and Danish HF populations [107, 127], although less pronounced than the drop in the current study. In these other HF populations, OCS was not (yet) introduced at that time. Stachowicz et al. [107] suggested that the drop in the Canadian population may be due to increased awareness and introduction of average relationship values (R-values) by the Canadian Dairy Network around 2000.

Table 2.5 Relative emphasis of trait categories in Dutch-Flemish bull selection index over time.

		Relative emphasis of trait category (%)					
Year	Index	INET	CONF	LONG	REPR	UH	Reference
1980	INET	100	-	-	-	-	[129]
1989	Stiersom	67	33	-	-	-	[129, 130]
1999	DPS	67	-	33	-	-	[131]
2003	DPS	58	-	26	12	4	[37]
2007	NVI	40	27	8	16	9	[132]
2012	NVI	26	30	11	19	14	[133]

Note that the relative emphasis of trait categories may not be calculated consistently across references. *INET*: production index combining milk, fat and protein yield; *CONF*: conformation traits, i.e. conformation of udder, legs, muscling and/or general stature; *LONG*: longevity or durability; *REPR*: reproductive traits including fertility and birth traits; *UH*: udder health or somatic cell count.

The steep increase in inbreeding and kinship rates around 2010 coincided with the implementation of GS. From the 2006-2010 period to the 2011-2015 period, there was a two- to four-fold increase in the annual rate of inbreeding. Rates per generation were also considerably higher since the implementation of GS, although the difference was less pronounced due to the decrease in L. Rates of ΔF_{PED}, ΔF_{ROH} and ΔHOM_{SNP} between 2011 and 2015 were as high as 1.8, 2.1 and 2.8% per generation, respectively (Figure 2.5). These rates correspond to an N_e of 18, 24 and 28, respectively. Rates of kinship were lower than rates of inbreeding, but were also well above the rates of 0.5 to 1% per generation recommended for livestock populations [2, 134]. The high rates per generation were rather unexpected, because, in theory, GS reduces ΔF_{gen} for a given genetic gain compared to traditional best linear unbiased prediction (BLUP) selection, by predicting Mendelian sampling terms and reducing the co-selection of sibs [91, 105].

Estimates of inbreeding and kinship rates in real life HF GS schemes are still scarce. Rodríguez-Ramilo et al. [108] recently evaluated genealogical and genomic inbreeding and kinship trends in the Spanish HF population. They reported N_e estimates that increased from 74-79 in the 1980-1999 period to 95-101 in the 2000-2013 period as a consequence of a reduction in L, but did not evaluate the years with GS separately [108]. For the global HF population, Miglior and Beavers [135] indicated that, although the number of AI bull sires has increased since GS, the number of sires that father 50% of the AI bulls has remained relatively constant. In North-American AI bulls, they also reported an increase of 1% in F_{PED} from 2011 to 2012 [135], which is in line with the 0.94% increase in the current study (Figure 2.4).

An important factor that contributes to the accumulation of kinship in GS schemes is the relationship of selection candidates with the reference population. In GS, genomic EBV (GEBV) are computed from the effects of SNPs, which are estimated

in a reference population of individuals with known genotypes and phenotypes [30]. The accuracy of an individual's GEBV is strongly affected by the genetic relationship between the individual and the reference population [136-138]. Pszczola et al. [136] indicated that the average squared relationship of a candidate with the reference population influences especially accuracy of GEBV. This means, for example, that having a single full sib in the reference population contributes more to a candidate's GEBV accuracy than having two half sibs. In general, candidates with a high average squared relationship with the reference population have a more accurate GEBV and are, therefore, more likely to be selected at a young age. This implies that, in a way, genetic variation in the reference population drives variation in selected individuals, which in turn drives variation at the population level. Thus, the composition of the reference population is an essential parameter that requires careful consideration for the management of diversity in the population.

Since the implementation of GS, rates of marker-by-marker homozygosity and similarity have been considerably higher (0.7%) than segment-based rates, which in turn have been slightly higher (0.1-0.3%) than genealogical rates. The higher rate for IBS suggests that relatedness due to distant common ancestors is increasing relatively fast compared to relatedness caused by common ancestors in more recent generations. This could be due to the discordance between the way breeding values are estimated and the way diversity is managed. In the current Dutch-Flemish breeding program, breeding values are predicted with genomic BLUP (GBLUP) and are, thus, based on marker-by-marker similarities weighted by allele frequencies [55]. However, diversity is managed on a genealogical basis by restricting f_{PED} with OCS. Although the relatively high correlations between f_{PED} and SIM_{SNP} and between f_{PED} and f_{SEG} (Figure 2.3) suggest that genomic IBD and IBS can be quite efficiently managed using f_{PED}, it is important to revisit this idea in view of OCS. In fact, when OCS is performed with GBLUP and a restriction on f_{PED}, the algorithm will search for selection candidates with a high GEBV and low f_{PED}, thereby putting emphasis on the Mendelian sampling terms that are not captured by the pedigree. As demonstrated by Sonesson et al. [91], the genomic inbreeding rate in such a scenario will substantially exceed the genealogical restriction. In addition, it will result in a IBD profile that is extremely variable across the genome [91]. Thus, controlling diversity at the genomic level should be a priority.

In this study, genomic diversity was characterized with marker-by-marker IBS and segment-based IBD. Both measures have clear advantages and drawbacks with regard to management. The main advantage of using marker-by-marker IBS in OCS is that it is the most effective in conserving diversity [121, 139]. However, a drawback is that it stimulates both alleles of biallelic loci to move to a frequency of 0.5, irrespective of their effects. Thereby, deleterious mutations continue to segregate in

the population. To expose and eliminate recessive deleterious mutations, it was suggested to combine OCS with inbred matings [140]. Alternatively, a segment-based IBD matrix can be used in OCS to restrict the increase in recent inbreeding. The rationale behind this approach is that recent inbreeding is more harmful than distant inbreeding, because the latter may have already been purged [141, 142]. In other words, the F_{ROH} is expected to be more closely associated with inbreeding depression than HOM_{SNP} [63, 112, 113]. Segment-based metrics can also be used to identify genomic regions that are prone to inbreeding depression, although the power of detection is limited by the fact that a single segment can contain multiple smaller haplotypes (or single SNPs) with different effects on the phenotype [53, 80]. Another drawback of the use of ROH and IBD-segments is their arbitrary definition. In this study, we defined the minimum length of IBD segments based on the average CGE of the pedigree, so that both genealogical and segment-based coefficients were expected to capture relatedness over 13.3 ancestral generations. However, the observed segment-based coefficients were on average ~1.5% higher than genealogical coefficients. Pedigree skewness, which is not completely accounted for by the CGE, will have contributed to this difference. For example, in an extreme scenario with 20 generations completely known on the sire's side, but with the dam unknown, the CGE of the offspring equals 10 while the F_{PED} equals 0 by definition. A second factor that strongly influenced the difference between genealogical and segment-based coefficients was the chosen maximum gap length between SNPs. For example, when the maximum gap size was set to 250 kb instead of 500 kb, the segment-based coefficients moved to the same scale as genealogical coefficients. Due to the large effect of such small changes, and the wide variety of criteria used in the literature [59, 124, 143], one should be extremely cautious when comparing segment-based coefficients across studies. A last drawback of the segment-based approach is that it is computationally rather intensive. In spite of these limitations, the use of segment-based metrics is considered a promising tool to determine the effect of inbreeding and, when applied in OCS, to maintain diversity and fitness simultaneously [53, 63, 112].

Selection has played an important role in shaping genetic variation across the HF genome over time. Although the identification of selection footprints was not the primary objective of this study, the regions in Table 2.3, enriched with 'significant' $|\Delta p|$ values, can be considered as putative signatures of selection. The most prominent peaks in $|\Delta p|$ were observed on BTA10 (Figure 2.8), which is in line with previously reported selection signatures for HF cattle [124, 144]. Using the extended haplotype homozygosity test (EHH) in German HF cattle, Qanbari et al. [144] detected 161 significant 'core regions' under selection, of which 17, 45, and 11 regions were located on BTA2, 10 and 20, respectively. We observed no clear peaks

on BTA2. For BTA20, a large region with high F_{ROH_k} (Figure 2.6) was observed, but it showed only small changes in allele frequency (Figure 2.8). This could be explained by the fact that F_{ROH_k} for this region was already high in 1986, which suggests that selection for this region occurred already before the Holsteinization (the large-scale introduction of HF into national dairy industries in the 1970s and early 1980s). The latter could also explain why this region is found in selection signature studies across various countries [124, 144, 145].

The important role of selection was also apparent from the fact that, in consecutive five-year periods, allele frequencies changed more often in the same direction than in opposite directions (Table 2.4). An exception was found when comparing allele frequency changes between the 1996-2000 and 2001-2005 periods, which suggests a change in the direction of selection around this time. Indeed, this change coincided with the implementation of OCS and the major shift in breeding goal composition. To further investigate the change in direction around 2000, a 'moving correlation' between Δp in the 1996-2000 period and Δp in the 2001-2005 period was computed for groups of 51 markers (Figure S2.2). There were several regions that showed a relatively strong negative correlation (Table S2.2) and which were rather large and harbored many known QTL associated with a wide range of traits. Although some of the identified regions showed a relatively large fraction of QTL related to traits such as reproduction (e.g. the region on BTA1), longevity (e.g. the region on BTA12) or udder health (e.g. the region on BTA13), these findings could not be specifically tied to the changes in breeding goal composition.

Substantial differences in $|\Delta p|$ (Figure 2.8) and in the accumulation of F_{ROH_k} (Figure 2.6) were observed across the genome. The emergence of such heterogeneity as a result of selection has been previously investigated in simulation and experimental studies [41, 91, 106]. These studies showed that GS acts more locally across the genome, with more pronounced hitchhiking effects compared to BLUP selection [41, 91, 106]. The striking increase in F_{ROH_k} from the 2006-2010 period to the 2011-2015 period for various genomic regions (Figure 2.6) could be the result of this local selection pressure. The peak regions showing high F_{ROH_k} remained fairly similar from the 2006-2010 period to the 2011-2015 period, which indicates that GS has not per se changed the regions that are under selection, but has especially increased the intensity of selection. This hypothesis is supported by the relatively strong positive correlation between Δp-values in the 2006-2010 period and those in the 2011-2015 period (Table 2.4).

An important question that should be raised is how heterogeneity in $|\Delta p|$ relates to maximizing genetic gain and maintaining genetic diversity. At some loci, it is desirable to increase the frequency of favorable alleles towards fixation. At other loci, a high level of genetic diversity is beneficial, for example to ensure a

population's capacity to combat a wide range of pathogens [146] or to limit inbreeding depression [80]. Thus, it is important to minimize the size of selection footprints [41, 53]. This can be achieved by slowly increasing the frequency of many favorable alleles with small effects, instead of strongly selecting for a few alleles with large effects [91, 147]. Although such an approach will not result in the highest gains in the short term, it will increase the long-term response [148, 149]. To maximize long-term gain further, it is desirable to select for rare favorable alleles, because this will increase the genetic variance [148]. Thus, to optimize long-term response while maintaining diversity, it is recommended to give less weight to SNPs that explain more variance and use a relatively uniform distribution of weights for the computation of GEBV [148, 149].

In general, genomic information offers many opportunities to manage genetic diversity and inbreeding more efficiently in the future (see [53] for a review). Among others, it can be used to control diversity at specific regions [82], select against multiple recessive disorders at the same time [150], estimate dominance effects for a better understanding of inbreeding depression [151], exploit variation in recombination rate across the genome [122] and characterize gene bank collections on the genomic level to optimize these collections and exploit stored material [102]. However, the practical benefit of such new insights and genomic tools in real-life selection schemes has yet to be explored.

2.5 Conclusions

There is substantial heterogeneity in diversity across the genome of HF AI -bulls over time as a result of selection and genetic drift. Trends in genome-wide and region-specific diversity reflect major changes in the Dutch-Flemish breeding program. The introduction of OCS and the shift in breeding goal, which both occurred around 2000, were followed by a temporary drop in inbreeding and kinship and were accompanied by a shift in the direction of changes in allele frequency. The recent introduction of GS around 2010 was accompanied by a substantial increase in the rates of inbreeding and kinship, both per year and per generation and especially at the IBS level. Allele frequencies continued to change in the same direction as before GS. These results provide insight in the effect of breeding practices on diversity across the genome and emphasize the need for efficient management of genetic diversity in HF GS schemes.

2.6 Acknowledgements

The research leading to these results has been conducted as part of the IMAGE project which received funding from the European Union's Horizon 2020 Research and Innovation Program under the grant agreement no 677353. The Dutch Ministry of Economic Affairs also contributed financially through the programs 'Kennisbasis

Dier' (code KB-12-005.03.001) and 'WOT' (code WOT-03-003-056). The authors gratefully acknowledge the Dutch-Flemish cattle improvement co-operative (CRV) for providing pedigree and genotype data. The authors would also like to thank the anonymous reviewers and the editors for their valuable comments and suggestions.

2.7 Supplementary information

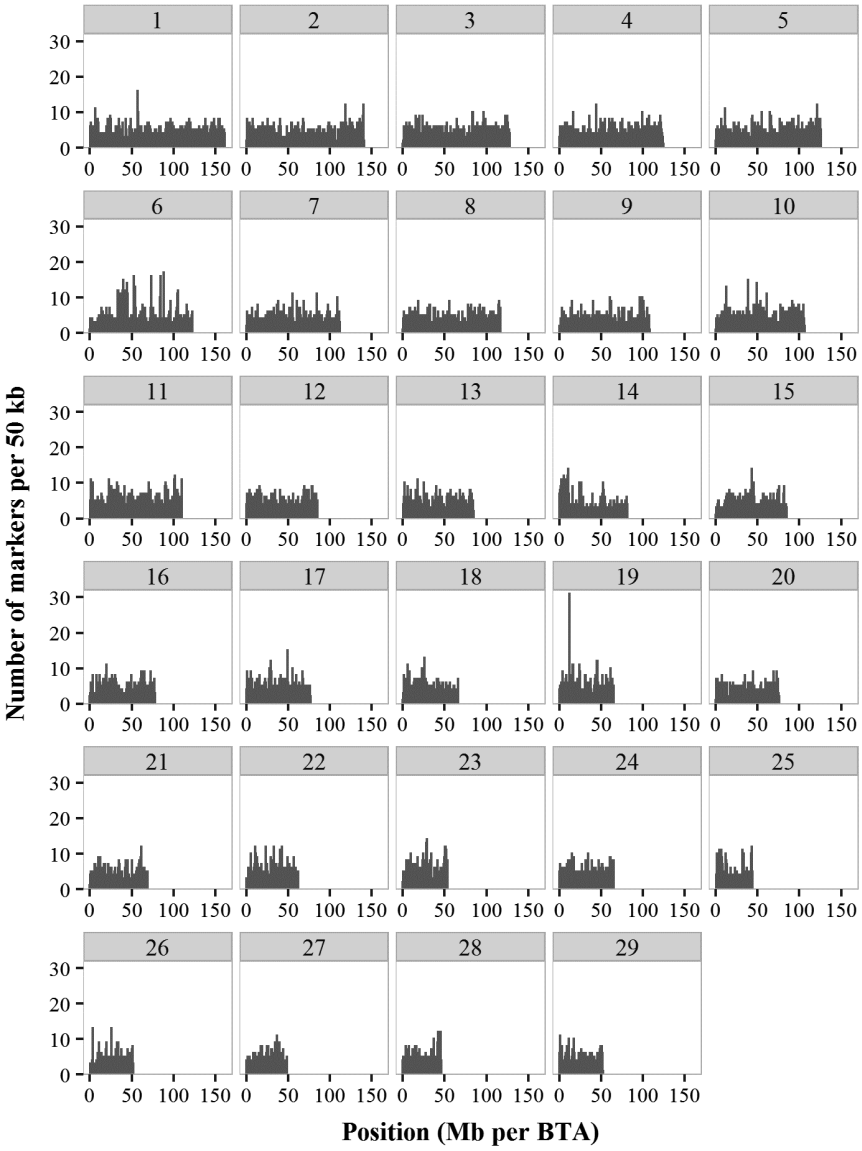


Figure S2.1 Number of SNPs per bin of 50 kb per Bos taurus autosome (BTA).

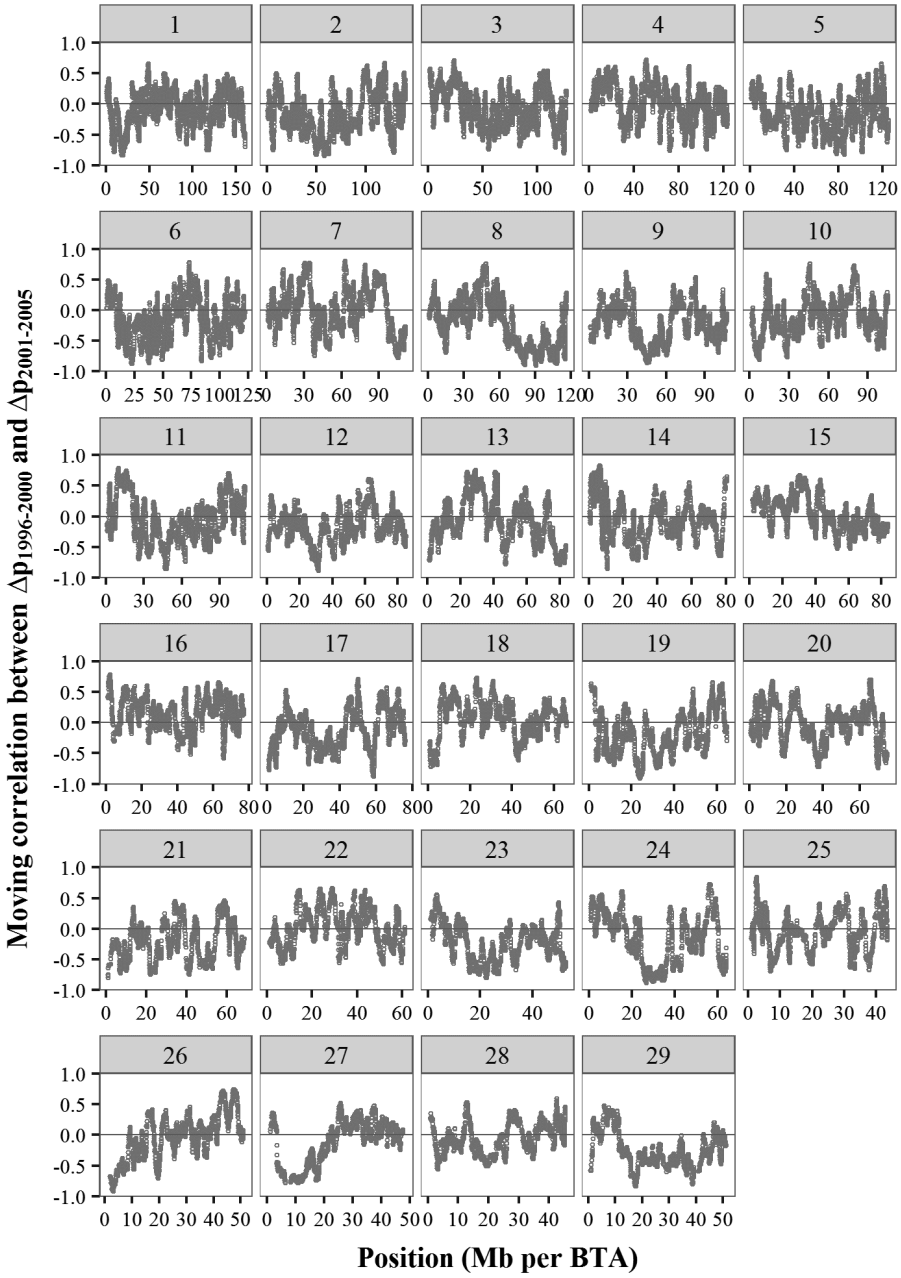


Figure S2.2 Moving correlation (of 51 markers) between changes in allele frequency in the 1996-2000 and 2001-2005 periods.

Table S2.1 Number of QTL extracted from AnimalQTLdb per trait and trait category (*continued on next page*).

Category	Trait	Number of QTL
INET	305-day milk yield	18
	Average daily milk yield	5
	Milk fat percentage	2,662
	Milk fat percentage (daughter deviation)	228
	Milk fat percentage (EBV)	128
	Milk fat yield	1,685
	Milk fat yield (daughter deviation)	384
	Milk lactose content	3
	Milk lactose yield	4
	Milk protein percentage	2,566
	Milk protein percentage (daughter deviation)	236
	Milk protein yield	925
	Milk protein yield (daughter deviation)	388
	Milk yield	876
	Milk yield (daughter deviation)	371
	Milk yield (EBV)	86
	Milk yield (ECM)	21
	Total INET	*10,586*
EXT	Body condition score	18
	Body depth	483
	Conformation score	12
	Dairy form	488
	Feet and leg conformation	592
	Foot angle	649
	Hind leg conformation	4
	Hoof and leg disorders	35
	Rear leg set	881
	Stature	579
	Teat length	281
	Teat number	5
	Teat placement	633
	Udder attachment	639
	Udder cleft	437
	Udder depth	655
	Udder height	476
	Udder structure	5
	Udder texture	10
	Udder width	1
	Total EXT	*6,883*
LONG	Length of productive life	2,092
	Lifetime profit index	53
	Total LONG	*2,145*
REPR	Birth index	17
	Calving ease	894
	Calving ease (maternal)	714
	Calving index	24

Table S2.1 (continued)

	Calving interval	49
	Calving to conception interval	46
	Conception rate	399
	Daughter pregnancy rate	781
	Fertility index	45
	Fertilization rate	20
	First service conception	10
	Inseminations per conception	1,611
	Interval from first to last insemination	1,463
	Interval from first to last insemination (EBV)	1
	Interval to first oestrus after calving	72
	Non-return rate	7
	Non-return rate (EBV)	13
	Stillbirth	682
	Stillbirth (maternal)	377
	Total	*1,744*
UH	Clinical mastitis	77
	Somatic cell count	33
	Somatic cell score	713
	Total	*823*

Table S2.2 Genomic regions of ≥ 7 Mb with strong negative correlation ($r \leq$ -0.6) between changes in allele frequency in the 1996-2000 and 2001-2005 periods, and fraction of QTL in these regions per trait category.

				Fraction of QTL per trait category (%)				
BTA	Start – end (Mb)	r	n_{QTL}	INET	CONF	LONG	REPR	UH
1	15.0 – 25.0	-0.60	34	24	26	6	38	6
2	46.0 – 63.0	-0.60	40	8	38	10	33	13
6	16.0 – 24.0	-0.63	51	20	45	8	25	2
8	75.0 – 110.0	-0.67	164	32	24	7	31	5
9	36.0 – 55.0	-0.60	115	47	25	5	18	4
10	3.0 – 10.0	-0.63	39	10	54	3	33	0
11	42.0 – 49.0	-0.68	36	22	53	6	19	0
12	25.0 – 32.0	-0.65	8	13	38	13	38	0
13	77.0 – 84.0	-0.66	19	26	0	0	21	53
24	25.0 – 37.0	-0.73	35	29	31	9	26	6
26	0.0 – 8.0	-0.68	70	21	47	10	20	1
27	4.0 – 13.0	-0.70	70	33	41	4	21	0
Total		*.*	*681*	*28*	*34*	*7*	*26*	*5*
Complete autosome		*-0.09*	*27,662*	*38*	*25*	*8*	*26*	*3*

QTL were included when reported in AnimalQTLdb [126]. QTL were classified into five trait categories: INET (production index), CONF (conformation), LONG (longevity), REPR (reproduction) or UH (udder health). See Table S2.1 for classification of traits.

3

Inbreeding depression due to recent and ancient inbreeding in Dutch Holstein Friesian dairy cattle

Harmen P. Doekes[1,2], Roel F. Veerkamp[1], Piter Bijma[1], Gerben de Jong[3], Sipke J. Hiemstra[2], Jack J. Windig[1,2]

[1]Wageningen University & Research Animal Breeding and Genomics, P.O. Box 338, 6700 AH, Wageningen, the Netherlands;
[2]Centre for Genetic Resources the Netherlands, Wageningen University & Research, P.O. Box 16, 6700 AA, Wageningen, the Netherlands
[3]Cooperation CRV, Wassenaarweg 20, 6843 NW Arnhem, the Netherlands

Genetics Selection Evolution (2019) 51:54

Abstract

Inbreeding decreases animal performance (inbreeding depression), but not all inbreeding is expected to be equally harmful. Inbreeding on recent ancestors is expected to be more harmful than inbreeding on more ancient ancestors, because selection decreases the frequency of deleterious alleles over time. The efficiency of selection is increased by inbreeding, a process called purging. Our objective was to investigate the effects of recent and ancient inbreeding on yield, fertility and udder health traits in Dutch Holstein Friesian cows, using various pedigree and genomic measures of inbreeding.

In total, 38,792 first-parity cows were included. Pedigree data were used to compute pedigree-based inbreeding (F_{PED}) and 76 k genotype data were used to compute genomic inbreeding measures, among others based on coverage of regions of homozygosity (ROHs) in the genome (F_{ROH}).

Inbreeding depression was observed, e.g. a 1% increase in F_{ROH} was associated with a decrease in 305-d milk yield of 36.3 kg (SE = 2.4), an increase in calving interval of 0.48 d (SE = 0.15) and an increase in mean somatic cell score for day 150 through to 400 of 0.86 units (SE = 0.28). These effects equaled -0.45, 0.12 and 0.05% of the corresponding trait means, respectively. Genomic inbreeding measures captured more inbreeding depression at the population level than pedigree-based inbreeding. When F_{PED} was split into generation-based components, inbreeding on recent generations was found to be more harmful for yield traits than inbreeding on more distant generations. In addition, there was evidence of purging based on pedigree. When F_{PED} was split into a new and an ancestral component, based on whether alleles were identical-by-descent for the first time or not, the new component was found to be more harmful than the ancestral component, especially for yield traits. For example, a 1% increase in the new component was associated with a decrease in 305-d fat yield of 2.42 kg (SE = 0.41), compared to an increase of 0.03 kg (SE = 0.71) for the ancestral component. There were no clear differences between effects of long ROHs (recent inbreeding) and short ROHs (ancient inbreeding).

Inbreeding depression was observed for yield, fertility and udder health traits. For yield traits and based on pedigree information, inbreeding on recent generations was more harmful than inbreeding on distant generations and there was evidence of purging. For all traits, both long and short ROHs contributed to inbreeding depression. In future work, inbreeding depression and purging should be assessed in more detail at the genomic level, using higher density information and genomic time series.

3.1 Introduction

Inbreeding depression is the decrease in mean performance due to mating between relatives. Many important traits in dairy cattle, such as yield and fertility traits, show inbreeding depression [80, 152-154]. The genetic basis of inbreeding depression is increased homozygosity with inbreeding, which increases the frequency of unfavorable genotypes [53, 68, 155]. Although overdominance and epistasis may contribute to inbreeding depression, partial dominance is expected to account for the major proportion of inbreeding depression [53, 79, 156].

A variety of methods can be used to assess inbreeding depression. Traditionally, inbreeding depression has been assessed by regression of phenotypes on pedigree-based inbreeding coefficients [157-159]. Nowadays, with the wide availability of genotype data, pedigree-based inbreeding coefficients can be replaced by genomic inbreeding coefficients [80, 152, 153]. Genomic inbreeding can be computed from a genomic relationship matrix (GRM) or from the proportion of the genome covered by regions (or runs) of homozygosity (ROHs) [54, 62]. Genomic inbreeding coefficients are expected to be more accurate than pedigree-based coefficients, because they account for Mendelian sampling variation (e.g. [49]) and do not depend on pedigree completeness and quality (e.g. [48]). Moreover, use of ROHs provides additional opportunities to distinguish recent from ancient inbreeding [80, 73, 113].

Not all inbreeding is expected to be equally harmful. Recent inbreeding (i.e. inbreeding arising from recent common ancestors) is expected to have a larger unfavorable effect than ancient inbreeding (i.e. inbreeding arising from more distant common ancestors). This hypothesis is based on the expected decrease in frequency of deleterious alleles over time, which is the result of (natural and/or artificial) selection. Since most deleterious alleles are (partially) recessive, inbreeding increases the efficiency of selection against these alleles by increasing homozygosity, which is called purging [79]. Purging is more likely to occur when there is strong selection pressure and when inbreeding accumulates slowly over time [79, 160].

With pedigree data, recent inbreeding may be distinguished from ancient inbreeding by including only a limited number of ancestral generations in the computation of inbreeding coefficients [72, 73]. Alternatively, one may use a purging-based approach to split the classical inbreeding coefficient into a new and an ancestral component, based on whether alleles are identical-by-descent (IBD) for the first time or have also been IBD in previous generations [161, 162]. The few studies that have applied the latter approach to commercial cattle populations found that the new inbreeding component was more harmful than the ancestral component, suggesting the presence of purging in these populations [154, 163].

With genomic data, age of inbreeding may be derived from the length of ROHs [77, 80, 113]. Longer ROHs reflect more recent inbreeding, because they have not

yet been broken up by recombination. More specifically, the length of ROHs derived from a common ancestor G generations ago roughly follows an exponential distribution with a mean of $1/2G$ Morgan [77, 78]. Only a few studies have investigated the effect of ROHs of different lengths on phenotypes in livestock, and the results of these studies vary [73, 74, 80].

The objective of this study was to evaluate the degree of inbreeding depression due to recent and ancient inbreeding in Dutch Holstein-Friesian dairy cattle. We expected to find stronger unfavorable effects for recent inbreeding compared to ancient inbreeding, because of selection against deleterious alleles over time (strengthened by purging). For a population of almost 40,000 genotyped cows, we determined the degree of inbreeding depression for yield, fertility and udder health traits. We used various pedigree-based and genomic inbreeding measures to compare these measures in terms of inbreeding depression. This study was performed in the context of artificial selection, meaning that all traits were under artificial selection and that natural selection will have had a relatively small contribution (or no contribution at all).

3.2 Material and methods

3.2.1 Animals and data

In total, 38,792 first-parity cows (fraction Holstein Friesian > 87.5%, either red or black) from 233 herds were included. These cows calved in the period 2012–2016 and were from herds with a data-agreement with the Dutch-Flemish cattle improvement cooperative (CRV; Arnhem, the Netherlands). Initially, 47,254 first-parity cows from 440 herds during the 2012–2016 period were considered. From this initial dataset, herds with less than 10 genotyped cows per year were discarded (n_{herds} = 207; n_{cows} = 8462) in order to exclude herds in which only a few cows were occasionally genotyped.

Pedigree, genotype and phenotype data were provided by CRV. The total pedigree comprised 167,924 individuals. To assess pedigree completeness, the number of complete generations (NCG) and the complete generation equivalent (CGE) were computed. The CGE was computed as the sum of $(1/2)^n$ of all known ancestors of an individual, with n being the number of generations between the individual and a given ancestor. To limit the effect of missing pedigree information on results, cows with a NCG lower than 3 and/or a CGE lower than 10 were excluded from pedigree-based analyses (n = 1,731). The mean NCG and CGE in the remaining cows equaled 6.5 generations and 12.5 generation-equivalents, respectively.

Cows were genotyped with the Illumina BovineSNP50 BeadChip (versions v1 and v2) or the CRV custom-made 60 k Illumina panel (versions v1 and v2). Genotypes were imputed to 76 k from the different panels, following Druet et al. [115]. Prior to

imputation, single nucleotide polymorphisms (SNPs) with a call rate lower than 0.85, a minor allele frequency (MAF) lower than 0.025 or a difference of more than 0.15 between observed and expected heterozygosity were discarded. In addition, SNPs with an unknown position on the Btau4.0 genome assembly were discarded. The final dataset contained 75,538 autosomal SNPs.

Yield, fertility and udder health traits were considered. For yield, the 305-day milk yield (MY; in kg), 305-day fat yield (FY; in kg), and 305-day protein yield (PY; in kg) were included. For fertility, the calving interval (CI; in days), interval calving to first insemination (ICF; in days), interval first to last insemination (IFL; in days), and conception rate (CR; in %) were included. For udder health, the mean somatic cell scores for day 5 through to 150 (SCS150; in units) and day 151 through to 400 (SCS400; in units) were included. Somatic cell scores were calculated as 1000 + 100*(log2 of cells/mL).

3.2.2 Inbreeding measures

Various inbreeding measures were used to assess inbreeding depression and distinguish recent from ancient inbreeding. These measures were divided into four groups: (1) pedigree generation-based measures, (2) pedigree purging-based measures, (3) ROH-based measures, and (4) GRM-based inbreeding.

Pedigree generation-based measures

The classical inbreeding coefficient based on all information in the pedigree (F_{PED}) was calculated with PEDIG [164]. The F_{PED} was defined as the pedigree-based probability that two alleles at a random locus in an individual were IBD [24]. In addition to F_{PED}, inbreeding coefficients based on the first n ancestral generations (F_{PEDn}), with n ranging from 4 to 8, were computed with the *vanrad.f* program in PEDIG [164, 165]. Inbreeding for specific age classes was computed as the difference between successive coefficients (e.g. inbreeding on ancestors from 5 generations ago was computed as $F_{PED5} - F_{PED4}$; abbreviated as F_{PED5-4}). The F_{PED8-7} was chosen as the most ancient category, because of the limited pedigree completeness for more ancient generations (e.g. only 78 cows had a NCG > 8; see Figure S3.1). The F_{PED4} was chosen as the most recent category, because very few individuals were inbred on ancestors in the first ancestral generations (Figure S3.2).

Pedigree purging-based measures

Based on the hypothesis of purging, a few additional pedigree-based measures were calculated. Following Kalinowski et al. [161], the F_{PED} was split into two components: an ancestral component (F_{ANC}) and a new component (F_{NEW}). The F_{ANC} was defined as the probability that alleles were IBD while they had already been IBD in at least

one ancestor, and F_{NEW} was the probability that alleles were IBD for the first time in the pedigree of the individual. The ancestral history coefficient (AHC) introduced by Baumung et al. [162] was also calculated. The AHC was defined as the number of times that a random allele had been IBD during pedigree segregation [162]. Kalinowski's inbreeding coefficients and the AHC were obtained by gene dropping, using 10^6 replications. The in-house script used for gene dropping is available upon request.

To illustrate the differences between all pedigree-based inbreeding measures, two example pedigrees are provided (Figure 3.1). In example (1), the F_{PED} of individual X equals 7.03%, since it is the sum of the inbreeding on ancestor A (0.5^7) and on ancestor D (0.5^4). Since ancestor A is in the 5th ancestral generation and D is in the first 4 generations, F_{PED5-4} equals the partial inbreeding on A (i.e. 0.5^7) and F_{PED4} equals the partial inbreeding on D (0.5^4). F_{ANC} is the probability that X is IBD for an allele that was already IBD in an ancestor, which in example (1) has to be ancestor E (since E is the only inbred ancestor). F_{ANC} can be manually calculated by multiplying the probability that E is IBD for an allele of A (0.5^4) with the probability that X inherits this allele from E given that E is IBD (1) and with the probability that X inherits this allele through D-F-G-X given that D is a carrier of the allele (0.5^3). Thus, it is equal to 0.78% (i.e. 0.5^7). In example (2), the F_{PED} of individual X is higher (31.25%) than in example (1), while F_{PED5-4} equals 0% based on the known information. The calculation of F_{ANC} in example (2) depends on both D and E, since both ancestors are inbred. F_{ANC} can be derived manually by tracing the possible genotype combinations. Individual A has two alleles, alleles 1 and 2. Consider the scenario in which individual B inherits allele 1 from A such that B has genotype 1/3, with 3 referring to a random allele inherited from the unknown parent of B. The possible genotypes of C are 1/4 and 2/4, where 4 is a random allele inherited from the unknown parent of C. If the genotype of C is 1/4, there are four possible genotypes for D and E (namely 1/1, 1/4, 3/1 and 3/4), resulting in 16 possible combinations of D and E and in 64 genotype possibilities for X. Among these 64 possibilities, there are 12 possibilities with X being 1/1 while D and/or E are 1/1 (four of which occur when D and E are both 1/1; the others occur when D or E is 1/1 while the other is 1/3 or 1/4). If C has genotype 2/4, while B is 1/3, there are also 64 genotype possibilities for X, but for none of these possibilities X will be IBD. Thus, if B is 1/3, there are 12 out of 128 possibilities for which X is IBD for allele 1 while D and/or E is also IBD for this allele. Similarly, if B is 2/3, there are 12 out of 128 possibilities for which X is IBD for allele 2 while D and/or E are also IBD for this allele. Therefore, the F_{ANC} equals 24 out of 256 (i.e. 9.38%).

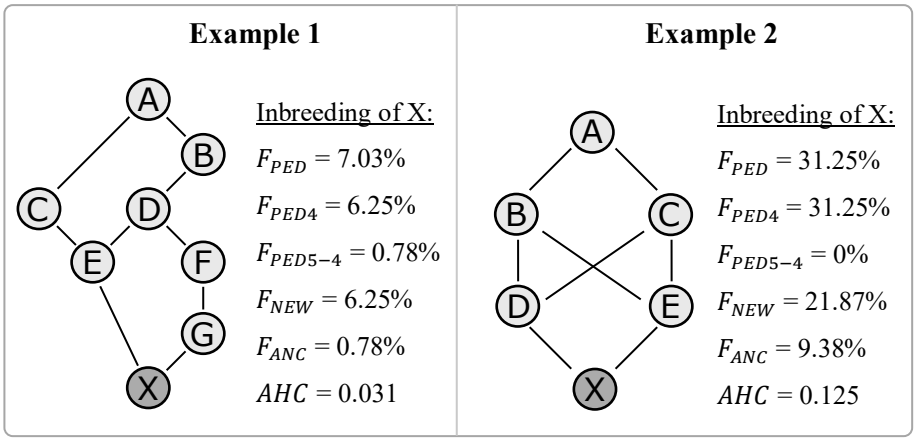


Figure 3.1 Example pedigrees illustrating differences between pedigree-based inbreeding measures for individual X. F_{PED}: pedigree inbreeding based on all available information; F_{PED4}: inbreeding based on first 4 generations; F_{PED5-4}: difference between inbreeding based on 5 and on 4 generations; F_{NEW}: Kalinowski's new inbreeding, i.e. probability that alleles in X are IBD for the first time; F_{ANC}: Kalinowski's ancestral inbreeding, i.e. probability that X is IBD for an allele that has already been IBD in an ancestor; AHC: ancestral history coefficient, i.e. the number of times that a random allele from X has been IBD during pedigree segregation.

ROH-based measures

The scanning window approach implemented in Plink 2.0 software [166] was used to identify ROHs. The following criteria were set to define a ROH: (i) a minimum physical length of 1 Mb, (ii) a minimum of 10 SNPs, (iii) a minimum density of one SNP per 100 kb, (iv) a maximum of one heterozygous call within a ROH, and (v) a maximum gap of 500 kb between consecutive SNPs. A scanning window of 10 SNPs, with a maximum of one heterozygote per window, was used.

After identification, ROHs were classified into five length classes: (i) > 16 Mb, (ii) 8 to 16 Mb, (iii) 4 to 8 Mb, (iv) 2 to 4 Mb, and (v) 1 to 2 Mb. The expected age of inbreeding increased from the first to the last class, since shorter ROHs reflect more ancient inbreeding. To illustrate this, the expected age was determined for each length category (Figure 3.2). The expected age of inbreeding was based on the concept that the length of ROHs derived from a common ancestor G generations ago follows an exponential distribution with mean $1/2G$ Morgan [77, 78]. For simplicity, a mean genetic distance of 1 Morgan per 100 Mb [122] was used and it was assumed that recombination rates were uniform across the genome and across sexes. Note that non-uniform recombination rates may result in deviations from the exponential distribution. For example, Speed and Balding [77] performed extensive simulations for the human genome and found that ROH length was best approximated with a gamma distribution with a shape parameter of 0.76. Since recombination rates may

differ across the bovine genome and across sexes [122], Figure 3.2 only provides a rough approximation of the expected length per ROH length class.

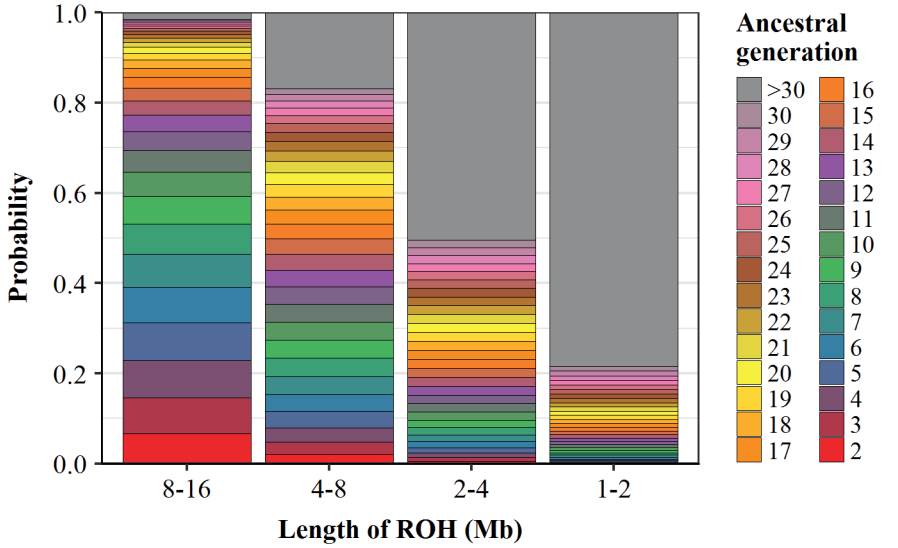


Figure 3.2 Expected age of inbreeding (in ancestral generations) for ROH classes, based on underlying exponential distributions. Note that this figure is an approximation, assuming a uniform distribution of inbreeding across ancestral generations, a uniform recombination rate across the genome and sexes, and a genetic distance of 1 Morgan per 100 Mb.

For each ROH length class, the inbreeding coefficient was calculated as the proportion of an individual's autosome that was covered by ROHs of that class (e.g. $F_{ROH>16}$). Autosome length was corrected for uncovered regions (i.e. ends of chromosomes and gaps of more than 500 kb without SNPs) and the corrected autosome length was 2469 Mb. A total inbreeding coefficient based on all ROHs (F_{ROH}) was also computed.

GRM-based inbreeding

Genomic inbreeding coefficients (F_{GRM}) were obtained as a measure of marker homozygosity. First, a genomic relationship matrix (GRM) was computed with *calc_grm* [118], according to the method of VanRaden [54]. Then, inbreeding coefficients were derived as the diagonal of the GRM minus 1 (since the relationship of an individual with itself equals 1 plus its inbreeding coefficient). When computing the GRM, allele frequencies were fixed to 0.5, such that F_{GRM} was equivalent to the proportion of homozygous SNPs, except for a difference in scale [167].

3.2.3 Statistical analysis

The degree of inbreeding depression was estimated by regressing phenotypes on inbreeding coefficients. For the total inbreeding measures (F_{PED}, F_{ROH} and F_{GRM}), the following linear mixed model was used:

$$(1)\quad y_{ijk} = \mu + HY_i + month_j + \alpha * age_k + \beta * F_k + cow_k + e_{ijk}$$

where HY_i is the i^{th} herd-year of calving (1,165 classes), $month_j$ is the j^{th} month of calving (12 classes), α is the regression coefficient for age_k, which was the age at calving for the k^{th} cow, β is the regression coefficient for F_k, which was the inbreeding coefficient for the k^{th} cow, cow_k is the random genetic effect for the k^{th} cow, and e_{ijk} is the random error term. The cow-effect was assumed to follow $\mathrm{N}(0, \mathbf{A}\sigma_a^2)$, where $\mathbf{A}$ was the numerator relationship matrix and σ_a^2 the additive genetic variance.

When F_{PED} or F_{ROH} was partitioned into classes based on the age of inbreeding, Model (1) was extended to fit these classes simultaneously (e.g. $F_{ROH>16}$, $F_{ROH8-16}$, F_{ROH4-8}, F_{ROH2-4} and F_{ROH1-2}):

$$(2)\quad y_{ijk} = \mu + HY_i + month_j + \alpha * age_k + \sum_{l=1}^{n} \beta_l * F_{kl} + cow_k + e_{ijk}$$

where β_l is the regression coefficient for F_{kl}, which was the inbreeding coefficient for the k^{th} cow and l^{th} inbreeding class, and n is the number of inbreeding classes.

All analyses were performed with ASReml 4.1 [168]. Regression coefficients and corresponding standard errors (SE) for inbreeding measures were obtained from output. In addition, P-values for the Wald test were obtained from output and were used to check for significance of the effects.

3.3 Results

3.3.1 Basic statistics for phenotypes and inbreeding measures

Descriptive statistics for the evaluated traits are in Table 3.1. Heritability estimates, obtained by running Model (1) without an inbreeding effect, were high for yield traits (0.36 to 0.47), moderate for somatic cell scores (0.11 and 0.14) and low for fertility traits (0.03 to 0.11).

Inbreeding based on ROH-coverage (F_{ROH}) was highly correlated with inbreeding based on marker homozygosity (F_{GRM}), with a Pearson correlation coefficient of 0.92 (Figure 3.3). Pedigree-based inbreeding (F_{PED}) was moderately correlated with F_{ROH} and F_{GRM}, with correlation coefficients of 0.66 and 0.61, respectively. The majority of cows (63%) were not inbred on ancestors in the first four ancestral generations, as illustrated by the distribution for F_{PED4} (Figure S3.2). For cows that were inbred

on the first four ancestral generations, clear peaks were visible at expected F_{PED4}-levels, for example at 0.78% (inbreeding on a single ancestor with an inbreeding loop of eight "steps") and at 1.56% (a single loop of seven steps, or two loops of eight steps). In line with pedigree-based results, only a few cows had very long ROHs (which indicate very recent inbreeding). About a fourth of the cows (26%) had no ROH > 16 Mb, 32% had a single ROH > 16 Mb, 21% had two ROHs > 16 Mb and the remaining 21% had three or more ROHs > 16 Mb. Pearson correlations suggest that the pedigree generation-based and the ROH-based measures partly captured the same age effects (Figure 3.3). For example, F_{PED4} showed a higher correlation with $F_{ROH>16}$ (r^2 = 0.50) than with $F_{ROH8-16}$ (0.34), F_{ROH4-8} (0.22), F_{ROH2-4} (0.10) and F_{ROH1-2} (−0.03). Similarly, F_{PED8-7} showed higher correlations with short ROHs than with long ROHs. Correlations among pedigree generation-based classes ranged from −0.23 to 0.27 and correlations among ROH-classes ranged from −0.10 to 0.26, suggesting rather independent inbreeding age classes. Notably, the F_{ROH1-2} showed a negative or very low correlation (ranging from −0.10 to 0.06) with all other calculated inbreeding measures, including overall homozygosity (F_{GRM}).

Table 3.1 Number of cows (N), mean, standard deviation (SD), corrected phenotypic standard deviation (σ_p), genetic standard deviation (σ_a) and heritability (h^2) for all evaluated traits

Trait	N	Mean	SD	σ_p	σ_a	h^2 (SE)
MY	38,778	8,091	1,375	1,199	825	0.47 (0.02)
FY	38,778	342	51.8	43.9	28.4	0.42 (0.02)
PY	38,778	283	44.7	36.6	22.0	0.36 (0.02)
CI	34,864	394	67.2	65.3	18.5	0.08 (0.01)
ICF	34,937	77.6	30.0	27.2	7.9	0.08 (0.01)
IFL	34,937	39.9	56.1	55.4	12.3	0.05 (0.01)
CR	34,774	63.8	36.1	35.7	6.1	0.03 (0.01)
SCS150	38,301	1,568	138	134	45.5	0.11 (0.01)
SCS400	37,068	1,581	133	129	48.9	0.14 (0.01)

MY: 305-day milk yield (kg); FY: 305-day fat yield (kg); PY: 305-day protein yield (kg); CI: calving interval (days); ICF: interval calving to first insemination (days); IFL: interval first to last insemination (days); CR: conception rate (%); SCS150 somatic cell score day 5 to 150 (units); SCS400: somatic cell score day 151 to 400 (units).

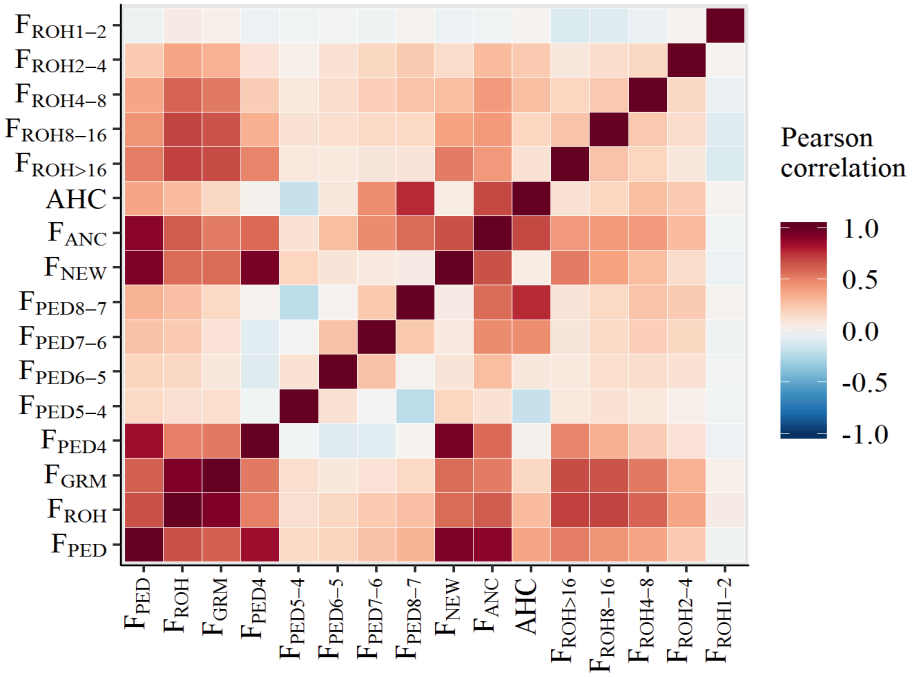


Figure 3.3 Heat map showing Pearson's correlations between different inbreeding measures. F_{PED}: pedigree inbreeding based on all generations; F_{ROH}: inbreeding based on all regions of homozygosity; F_{GRM}: inbreeding based on genomic relationship matrix computed with allele frequencies of 0.5. F_{PED4}: pedigree inbreeding based on first 4 generations; F_{PED5-4}: difference between pedigree inbreeding based on 5 and on 4 generations; F_{NEW}: Kalinowski's new inbreeding; F_{ANC}: Kalinowski's ancestral inbreeding; AHC: ancestral history coefficient; $F_{ROH>16}$: inbreeding based on regions of homozygosity longer than 16 Mb; $F_{ROH8-16}$: inbreeding based on regions of homozygosity of 8 to 16 Mb.

3.3.2 Depression for total inbreeding measures

Inbreeding depression was observed for each of the total inbreeding measures (F_{PED}, F_{ROH} and F_{GRM}) and the estimated effects were significant for most traits (Table 3.2). For example, a 1% increase in F_{ROH} was associated with a decrease in 305-day milk yield of 36.25 kg ($P < 0.01$), an increase in calving interval of 0.48 day ($P < 0.01$) and an increase in mean somatic cell score in day 151 to 400 of 0.80 units ($P < 0.01$). All estimated effects, including those that were not significant at the 0.05-level (e.g. for ICF), were unfavorable.

To further illustrate differences in performance associated with differences in inbreeding, the expected phenotypes of cows with low (5% percentile) and high (95% percentile) inbreeding coefficients were compared (Table 3.3). These differences were computed for traits that showed a significant depression effect for each of the total inbreeding measures. Differences between cows with low and high inbreeding coefficients were smaller for pedigree-based inbreeding than for genomic inbreeding

measures. For example, differences in 305-day milk yield between lowly and highly inbred cows were 198, 301 and 315 kg for F_{PED}, F_{ROH} and F_{GRM}, respectively.

Table 3.2 Estimates of inbreeding depression for all traits and total inbreeding measures, expressed as the change in expected phenotype per 1% increase in inbreeding. Significance for non-nullity is indicated by stars (*P < 0.05; **P < 0.01).

	F_{PED}		F_{ROH}		F_{GRM}	
Trait	Estimate	SE	Estimate	SE	Estimate	SE
MY	-37.95**	3.66	-36.25**	2.35	-48.07**	2.83
FY	-1.54**	0.14	-1.34**	0.09	-1.60**	0.11
PY	-1.27**	0.11	-1.20**	0.07	-1.55**	0.09
CI	0.46*	0.23	0.48**	0.15	0.62**	0.18
ICF	0.16	0.09	0.08	0.06	0.09	0.07
IFL	0.13	0.19	0.27*	0.12	0.42**	0.15
CR	-0.31*	0.12	-0.27**	0.08	-0.36**	0.09
SCS150	0.58	0.44	0.30	0.28	0.44	0.34
SCS400	0.86*	0.43	0.86**	0.28	1.15**	0.33

MY: 305-day milk yield (kg); FY: 305-day fat yield (kg); PY: 305-day protein yield (kg); CI: calving interval (days); ICF: interval calving to first insemination (days); IFL: interval first to last insemination (days); CR: conception rate (%); SCS150 somatic cell score day 5 to 150 (units); SCS400: somatic cell score day 151 to 400 (units).
F_{PED}: pedigree inbreeding based on all generations; F_{ROH}: inbreeding based on all regions of homozygosity; F_{GRM}: inbreeding based on genomic relationship matrix computed with allele frequencies of 0.5.

Table 3.3 Difference (Diff) between expected phenotypes of cows with low and high inbreeding, for significant traits and total inbreeding measures.

	F_{PED}			F_{ROH}			F_{GRM}		
Trait	Low	High	Diff	Low	High	Diff	Low	High	Diff
MY	8,175	7,977	198	8,227	7,926	301	8,232	7,917	315
FY	345.4	337.4	8.0	347.0	335.9	11.1	346.7	336.2	10.5
PY	285.8	279.2	6.6	287.5	277.5	10.0	287.5	277.4	10.1
CI	393.0	395.4	-2.4	392.2	396.2	-4.0	392.2	396.2	-4.0
IFL	39.6	40.3	-0.7	38.9	41.1	-2.2	38.7	41.4	-2.7
CR	64.5	62.9	1.6	64.8	62.6	2.2	64.9	62.5	2.4
SCS400	1,579	1,583	-4	1,578	1,585	-7	1,578	1,585	-7

MY: 305-day milk yield (kg); FY: 305-day fat yield (kg); PY: 305-day protein yield (kg); CI: calving interval (days); IFL: interval first to last insemination (days); CR: conception rate (%); SCS400: somatic cell score day 151 to 400 (units)
F_{PED}: pedigree inbreeding based on all generations; F_{ROH}: inbreeding based on all regions of homozygosity; F_{GRM}: inbreeding based on the genomic relationship matrix computed with allele frequencies of 0.5
Low and high inbreeding were defined as the 5% and 95% percentile, respectively. Low and high inbreeding equaled 2.8% and 8.0% for F_{PED}, 8.5% and 16.9% for F_{ROH} and 25.9% and 32.4% for F_{GRM}.

To compare depression effects across traits, the estimated regression coefficients from Table 3.2 were also expressed as percentages of the corresponding trait means, as well as in phenotypic and genetic standard deviations (Table S3.1). When expressed as percentages of trait means, yield traits showed a relatively large depression effect (of 0.39 to 0.47%) and somatic cell scores a relatively small effect (of 0.02 to 0.05%). The effect for fertility differed across traits and inbreeding measures. It was relatively high for CR and IFL (0.33 to 0.67%) and intermediate for CI and ICF (0.11 to 0.21%). When compared in phenotypic standard deviations, yield traits showed the highest degree of inbreeding depression. When compared in genetic standard deviations, yield traits also showed the highest degree of inbreeding depression, in spite of the lower heritability of fertility and udder health traits. Only conception rate, which had a heritability of 0.03, showed a depression effect similar to that of yield traits when compared in genetic standard deviations.

3.3.3 Depression for pedigree generation-based inbreeding classes

When F_{PED} was split into generation-based classes, recent inbreeding significantly reduced milk, fat and protein yield whereas more ancient inbreeding had a non-significant neutral or even favorable effect (Figure 3.4). For example, the estimated effects for 305-day protein yield from the most recent to the most ancient class were equal to −1.3 kg (for F_{PED4}), −1.4 kg (F_{PED5-4}), −0.6 kg (F_{PED6-5}), 0.3 kg (F_{PED7-6}) and 0.7 kg (F_{PED8-7}). For fertility and udder health traits, estimated effects were generally not significantly different from zero and no clear pattern was visible. For example, the interval between calving and first insemination seemed to be unfavorably affected by all classes, but none of the effects was significant. For all traits, standard errors increased with age of inbreeding. This may be explained by a lower degree of variation for more ancient inbreeding (Figure S3.2).

3.3.4 Depression for pedigree purging-based inbreeding components

When F_{PED} was split into Kalinowski's new (F_{NEW}) and ancestral (F_{ANC}) components, new inbreeding significantly reduced milk, fat and protein yield, whereas ancestral inbreeding did not (Figure 3.5). For example, a 1% increase in F_{NEW} was associated with a 2.42 kg (SE = 0.41) decrease in 305-day fat yield, while a 1% increase in F_{ANC} was associated with a 0.03 kg (SE = 0.71) increase in fat yield. For fertility and udder health traits, both new and ancestral inbreeding effects were not significantly different from zero. For most traits (MY, FY, PY, IFL, CR, SCS150, SCS400), the estimated effect of new inbreeding was more unfavorable than the effect of ancestral inbreeding. For some traits (e.g. IFL), the estimated effect of ancestral inbreeding was even slightly favorable, whereas the effect of new inbreeding was always unfavorable.

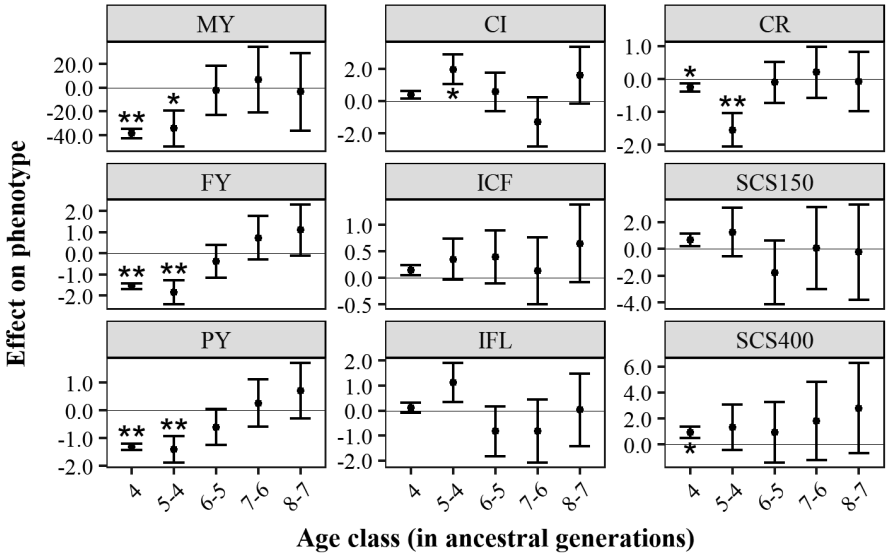


Figure 3.4 Effect of a 1% increase in pedigree inbreeding (F_{PED}) on phenotypes, for different age classes. Error bars represent one standard error and stars indicate significance for non-nullity (*P < 0.05; **P < 0.01). MY: 305-day milk yield (kg); FY: 305-day fat yield (kg); PY: 305-day protein yield (kg); CI: calving interval (days); ICF: interval calving to first insemination (days); IFL: interval first to last insemination (days); CR: conception rate (%); SCS150 somatic cell score day 5 to 150 (units); SCS400: somatic cell score day 151 to 400 (units).

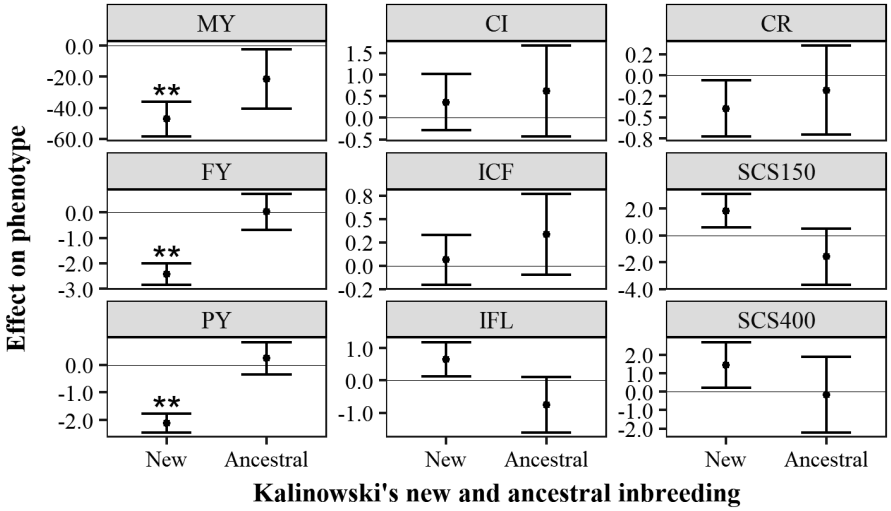


Figure 3.5 Effect of a 1% increase in Kalinowski's new and ancestral inbreeding on phenotypes. Error bars represent one standard error and stars indicate significance for non-nullity (*P < 0.05; **P < 0.01). MY: 305-day milk yield (kg); FY: 305-day fat yield (kg); PY: 305-day protein yield (kg); CI: calving interval (days); ICF: interval calving to first insemination (days); IFL: interval first to last insemination (days); CR: conception rate (%); SCS150 somatic cell score day 5 to 150 (units); SCS400: somatic cell score day 151 to 400 (units).

The AHC had no significant effect on traits, except for a favourable effect on 305-day protein yield (Table 3.4). When AHC was fitted simultaneously with F_{PED}, fat yield also tended to increase with an increase in AHC (P < 0.1). Interactions between AHC and F_{PED} were not significant.

Table 3.4 Effect of an increase in the ancestral history coefficient (AHC) on all traits, when a model with only the AHC or with the AHC and pedigree-based inbreeding (F_{PED}) was used. Significance for non-nullity is indicated by stars (*P < 0.05; **P < 0.01)

	Model with only AHC		Model with AHC and F_{PED}			
	AHC		AHC		F_{PED}	
Trait	Estimate	SE	Estimate	SE	Estimate	SE
MY	157.1	306.5	403.7	307.0	-38.3**	3.7
FY	9.4	11.1	20.2	11.1	-1.6**	0.14
PY	24.5**	9.2	34.1**	9.2	-1.31**	0.11
CI	11.5	14.7	7.0	14.9	0.44	0.23
ICF	6.8	6.1	5.3	6.2	0.15	0.09
IFL	-11.2	11.8	-12.9	12.0	0.17	0.19
CR	3.5	7.2	7.0	7.4	-0.34**	0.12
SCS150	-25.3	29.9	-31.4	30.2	0.64	0.44
SCS400	-3.2	30.0	-11.4	30.3	0.88*	0.43

MY: 305-day milk yield (kg); FY: 305-day fat yield (kg); PY: 305-day protein yield (kg); CI: calving interval (days); ICF: interval calving to first insemination (days); IFL: interval first to last insemination (days); CR: conception rate (%); SCS150 somatic cell score day 5 to 150 (units); SCS400: somatic cell score day 151 to 400 (units)

3.3.5 Depression for ROH length-based inbreeding components

When F_{ROH} was split into classes based on ROH length (> 16, 8–16, 4–8, 2–4 and 1–2 Mb), the effect of these classes differed across traits (Figure 3.6). For 305-day milk yield, for example, all five classes showed a significant decrease in yield per 1% increase in inbreeding, with a slightly stronger effect for ancient inbreeding (F_{ROH1-2}; effect of −60 kg) than for more recent inbreeding (longer ROH-classes; effects varying from −29 to −40 kg). For 305-day fat yield, an increase in $F_{ROH>16}$ and $F_{ROH8-16}$ was associated with a decrease in yield, while for shorter ROHs this decrease was less pronounced. For fertility and udder health traits, most effects were not significantly different from zero. However, some of these traits did show a trend. For calving interval and for the interval between calving and first insemination, inbreeding based on long ROHs seemed to increase these intervals, whereas that based on shorter ROHs seemed to decrease these intervals. In contrast, for somatic cell score for day 151 through to 400, there seemed to be a larger unfavorable effect of short ROHs compared to long ROHs. Across all traits, standard errors were larger for inbreeding based on short ROHs compared to long ROHs. This may be the result of less variation in inbreeding based on short ROHs (Figure S3.2).

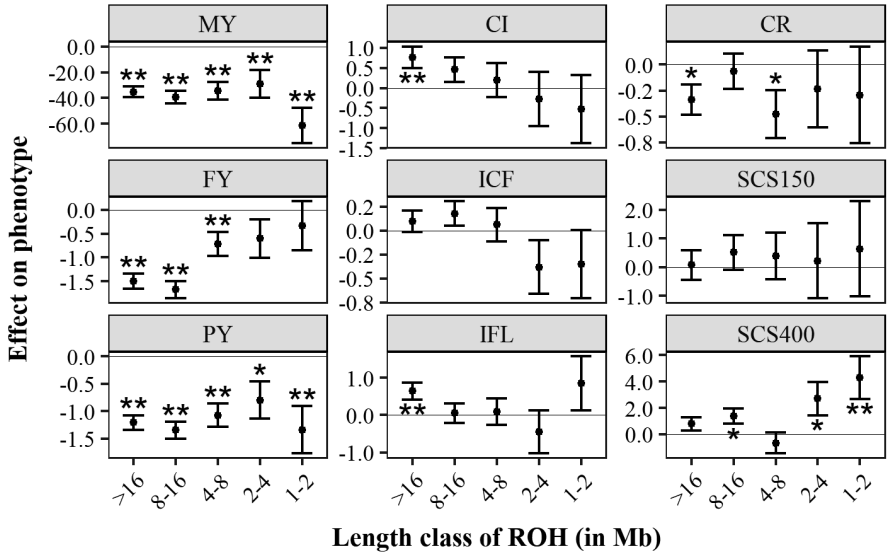


Figure 3.6 Effect of a 1% increase in ROH-based inbreeding (F_{ROH}) on phenotypes, for different ROH lengths. Error bars represent one standard error and stars indicate significance for non-nullity (*P < 0.05; **P < 0.01). MY: 305-day milk yield (kg); FY: 305-day fat yield (kg); PY: 305-day protein yield (kg); CI: calving interval (days); ICF: interval calving to first insemination (days); IFL: interval first to last insemination (days); CR: conception rate (%); SCS150 somatic cell score day 5 to 150 (units); SCS400: somatic cell score day 151 to 400 (units).

3.4 Discussion

3.4.1 Inbreeding depression and its costs

Estimates of pedigree-based inbreeding depression were comparable to those reported in previous studies. For example, a 1% increase in pedigree inbreeding has previously been associated with a reduction in 305-day milk yield of 20 to 30 kg [71, 159, 169] and with an increase in calving interval of 0.2 to 0.7 days [71, 157, 169]. Inbreeding depression for somatic cell score has also been observed before [76, 169, 170], but estimates were not directly comparable because of different scales and because of the use of separate measures for early (SCS150) and late (SCS400) lactation in the current study. In general, the accuracy of pedigree-based results depends largely on pedigree quality and completeness. Incomplete pedigrees may lead to downward bias of inbreeding coefficients and, therefore, to misleading estimates of inbreeding depression [171]. In an attempt to limit this bias, we decided to include only the individuals with a NCG of at least three generations and a CGE of at least 10 equivalents in this study.

Estimates of inbreeding depression based on genomic inbreeding measures were similar to those estimated for pedigree-based inbreeding and to those reported in other studies. In US Holstein Friesian cattle, Bjelland et al. [152] found a decrease in

205-day milk yield of 20 kg and 47 kg for a 1% increase in F_{ROH} and F_{GRM}, respectively. They also observed an increase in days open (a trait similar to calving interval) of 1.72 and 1.06 days for F_{ROH} and F_{GRM}, respectively. They did not observe an effect on SCS. In Australian Holstein Friesian cattle, Pryce et al. [80] estimated inbreeding depression based on a F_{GRM} measure that was corrected for allele frequencies of the contemporary population. They found that a 1% increase in their F_{GRM}-estimates was associated with a decrease in lactation yields for milk, fat and protein of 28 kg, 1.3 kg and 0.9 kg, respectively. In addition, they observed a slight increase in calving interval of 0.12 days, although this increase was not significant. As illustrated by the current and previous studies, genomic measures of inbreeding can be effectively used to estimate the effects of inbreeding on performance. In fact, we found that F_{ROH} and F_{GRM} captured more phenotypic differences between lowly and highly inbred cows than F_{PED} (Table 3.3), in spite of the larger estimated change in phenotype per 1% increase in F_{PED} compared to F_{ROH} (Table 3.2). This finding was in line with the results of Bjelland et al. [152] and is the result of a wider distribution for F_{ROH} compared to F_{PED} (Figure S3.2). The finding that F_{PED} captures less inbreeding depression than F_{ROH} and F_{GRM} may be explained by the random nature of recombination and segregation, which is captured with genomic measures but not with pedigree. Since there will be more measurement errors in pedigree inbreeding than in genomic inbreeding, there will be more attenuation or "flattening" of the slope towards zero for F_{PED} (a statistical phenomenon known as regression dilution). For the various inbreeding measures, which Keller et al. [113] investigated in their simulation study, ROH-based inbreeding showed the highest correlation with the homozygous mutation load. Our results suggest that F_{ROH} and F_{GRM} capture similar effects of inbreeding depression at the population level, which is not surprising because of the high correlation between these two measures (r^2= 0.93 in this study).

Costs of inbreeding should be considered in the framework of a breeding program. For example, for a trait such as 305-day milk yield, we estimated a reduction of around 38 kg per 1% increase in pedigree-based inbreeding. If we consider that the pedigree-based inbreeding level in Dutch Holstein-Friesian cattle has increased from around 0.5% in 1980 to around 4.5% in 2010 [172-174], this would roughly imply a mean loss of 150 kg due to inbreeding depression. Such a loss is small compared to the realized genetic progress in the same period, which was equal to approximately 2200 kg [175]. Although the rate of inbreeding has increased with the introduction of genomic selection [172], contrary to expectation [105], the increased genetic gains [104] are expected to still outweigh the losses caused by inbreeding depression. It should be noted that overall costs will be larger than the cost for single traits, especially since components of economic return may combine multiplicatively rather than additively [176]. In addition, it is important to realize that

inbreeding will also affect traits that were not included in the present study, such as stillbirths [163]. Previous economic analyses of inbreeding depression suggested lifetime losses per cow in the order of tens of US dollars per 1% increase in inbreeding [76, 157, 159]. These analyses confirm that, by affecting various traits, inbreeding depression reduces net income. Combined with the importance of conserving genetic diversity for future adaptability, the costs of inbreeding depression provide incentive to monitor and manage inbreeding in dairy cattle populations.

3.4.2 Recent inbreeding is more harmful than ancient inbreeding and evidence of purging

The main objective of this study was to evaluate the hypothesis that recent inbreeding is more harmful than ancient inbreeding. This hypothesis was based on the expected decrease in frequency of deleterious alleles over time as a result of selection, strengthened by purging. Computer simulations have shown that purging is more effective when selection pressure is strong and when inbreeding accumulates slowly over many generations [79, 160]. We expected that purging would have occurred in the Dutch Holstein-Friesian population, because the population has undergone decades of intense artificial selection and inbreeding has accumulated (at least until 2012) at approximately 0.13% per year [172-174].

Pedigree-based results support our hypothesis. For yield traits, inbreeding on recent generations was more harmful than inbreeding on more distant generations (Figure 3.4). In addition, there was evidence of purging for these traits (Figure 3.5). For most traits, Kalinowski's F_{NEW} was more harmful than Kalinowski's F_{ANC} (Figure 3.5). For some traits, the estimated effect of F_{ANC} was even favorable. In other words, to be IBD for alleles that were already IBD in the past had a neutral or favorable effect, whereas to be IBD for alleles for the first time was generally unfavorable. These findings are in line with the hypothesis of purging, which states that loci that have undergone inbreeding in the past have been exposed to an increased selection efficiency (against deleterious recessive alleles), compared to loci that have not undergone inbreeding before. Our results are largely in line with previous studies that have investigated purging in commercial cattle populations [154, 163]. In German Holstein-Friesian cattle, Hinrichs et al. [163] studied the effects of new and ancestral inbreeding on reproductive traits. They found that a 1% increase in F_{NEW} was associated with a decrease in birthweight of 11.9 kg, while a 1% increase in F_{ANC} was associated with an increase in birthweight of 41.6 kg. They also observed a significant increase in the rate of stillbirths for F_{NEW} , while F_{ANC} showed a slight reduction in stillbirths that was not significant. In Irish Holstein-Friesian cattle, McParland et al. [154] investigated the effects of new and ancestral

inbreeding on yield and fertility traits. They found that a 1% increase in F_{NEW} was associated with a decrease in 305-day milk, fat and protein yields of 32.4 kg, 2.4 kg and 1.1 kg, respectively. They also found unfavorable effects for F_{ANC}, but these effects were less strong, namely 8.9 kg, 0.5 kg and 0.3 kg, respectively. For calving interval, they estimated an increase of 4.1 and 0.6 days for F_{NEW} and F_{ANC}, respectively. Differences across studies may be partly explained by the way that F_{NEW} and F_{ANC} have been fitted. In this study and in the study of Hinrichs et al. [23], the F_{NEW} and F_{ANC} were fitted simultaneously in the model, thereby accounting for the correlation between the two measures (r^2 = 0.67 in this study). In the study of McParland et al. [154], however, F_{NEW} and F_{ANC} were fitted individually.

Differences between effects of recent and ancient inbreeding (Figure 3.4) and between effects of Kalinowski's F_{NEW} and F_{ANC} (Figure 3.5) were most apparent for yield traits, which is in accordance with McParland et al. [154]. This finding may be explained by the selection history of Dutch Holstein-Friesian cattle. Targeted selection for fertility and udder health has taken place only since these traits were included in the breeding goal around the year 2000, whereas selection for yield traits has taken place for many more decades [173]. Therefore, there has been less time for selection to act on alleles that affect fertility and udder health traits compared to alleles that affect yield traits.

In addition to Kalinowski's new and ancestral inbreeding, we also considered the ancestral history coefficient (AHC). AHC is defined as the number of times that a random allele in an individual has been IBD in the individual's pedigree [162]. The rationale behind this recently introduced measure is that purging is not fully efficient and that the probability of purging increases with the number of times the alleles have been IBD. In other words, an allele that has been IBD many times in an individual's pedigree is more likely to have a neutral or positive effect on traits under selection, compared to an allele that has been IBD only once or never before. An increase in AHC, therefore, is expected to be associated with a favorable effect on the phenotype. Indeed, we observed a few favorable effects, i.e. an increase in protein yield and a tendency for an increase in fat yield (Table 3.4). Most traits showed no significant effect, but the estimate was generally favorable. In Thoroughbred horses, Todd et al. [177] found a strong positive association between AHC and racing performance. Compared to their study, where the mean AHC was 1.97 (SD = 0.09), the mean AHC in the current study was rather low at 0.31 (SD = 0.05). This can be explained by the very comprehensive pedigree of the Thoroughbred population, which dates back to the late eighteenth century, with individuals from 2000 to 2010 having a mean CGE of 24.6 [177].

A purging-based measure that we did not include in this study is Ballou's [141] ancestral inbreeding coefficient (F_{ANC_BAL}). The F_{ANC_BAL} is defined as the probability

that any allele in an individual has been IBD in an ancestor at least once [141]. It can be calculated by using an iterative formula [141] or with gene dropping [178], where gene dropping provides more robust estimates by accounting for dependence between F_{ANC_BAL} and F_{PED} [179]. To assess the effect of purging, one has to include the product of F_{ANC_BAL} and F_{PED} in the model [141, 154, 179], because F_{ANC_BAL} does not consider the IBD-probability for an individual itself. The product of F_{ANC_BAL} and F_{PED} is the probability that an individual is IBD for an allele that was already IBD in at least one ancestor, which is in fact the definition of Kalinowski's F_{ANC} [161]. Similarly, the product of $(1 - F_{ANC\_BAL})$ and F_{PED} is equivalent to the F_{NEW} of Kalinowski. Because of this equivalence, we decided to include only Kalinowski's measures in this study.

More recently, an inbreeding-purging (IP) model was proposed to assess purging based on genealogical information [180]. This model, which was developed in a conservation biology context, predicts how fitness evolves in a population undergoing inbreeding by means of a purged inbreeding coefficient (g). g is the traditional inbreeding coefficient weighted by the reduction in frequency of deleterious alleles induced by purging. Using simulations, López-Cortegano et al. [181] showed that inbreeding depression estimates based on the IP model are similar to those obtained using Ballou's approach, with smaller standard errors for the IP model. We considered using the IP model for the current study. Since the model and associated software (PURGd) have been developed outside the context of artificially selected populations, various limitations exist for its application to livestock data. First, random effects cannot be fitted in the model, making it impossible to directly correct for additive genetic relationships. To overcome this limitation, one could first run an animal model in a different software environment (e.g. ASReml) and subsequently use the residuals as phenotypes for the IP model. This two-step process is not desirable, because it will affect the inbreeding depression estimates. Second, the model assumes that inbreeding load is due to deleterious alleles that have a low initial frequency in the (base) population. In the context of livestock breeding, where animals are selected based on a breeding goal composed of various traits [173], we do not expect that alleles that are deleterious for a single trait necessarily segregate at a low frequency. Given these limitations, we decided not to use the IP model in the current study. For future research, it would be valuable to explore the application of the IP model in (livestock) populations undergoing artificial selection.

3.4.3 Long and short ROHs contribute to inbreeding depression

We expected that inbreeding based on long ROHs (recent inbreeding) would be associated with stronger depression effects than inbreeding based on short ROHs

(ancient inbreeding). For some traits (e.g. fat yield and calving interval) our results were in line with this hypothesis, but for other traits there was no clear pattern across ROH-length classes or there was even a pattern in the opposite direction (Figure 3.6). Overall, both long and short ROHs seemed to contribute to inbreeding depression.

Only a few studies have investigated the effect of ROHs of different lengths on phenotypes in livestock populations, with various results [73, 74, 80]. In Austrian Fleckvieh, Ferenčaković et al. [73] found stronger inbreeding depression for number of spermatozoa when considering both long and short ROHs (e.g. > 2 Mb) than when considering only long ROHs (e.g. > 16 Mb). For autosome 3 in Iberian pigs, Saura et al. [74] observed that inbreeding based on long ROHs (> 5 Mb) significantly decreased the number of piglets born, whereas inbreeding based on short ROHs (0.5 to 5 Mb) had a non-significant favorable effect. In Australian Holstein Friesian cattle, Pryce et al. [80] observed a stronger depression effect for 305-day milk yield when only very long ROHs were included than when also shorter ROHs were included. To further investigate and compare our results to the findings of Pryce et al. [80], we also ran Model (1) for cumulative ROH-based inbreeding coefficients (i.e. $F_{ROH>16}$, $F_{ROH>8}$, $F_{ROH>4}$, $F_{ROH>2}$ and $F_{ROH>1}$). We obtained a similar trend (Figure S3.3) as Pryce et al. [80], with $F_{ROH>16}$ showing the strongest effect and the inclusion of shorter ROHs reducing the effect size. The difference between results for fitting multiple length classes simultaneously (Figure 3.6) and for fitting cumulative measures one by one (Figure S3.3) may be due to the correlations between classes. We believe that fitting length classes simultaneously provides the most accurate estimates, since this approach accounts for the correlations between classes.

Based on computer simulations, Keller et al. [113] concluded that long ROHs correlate better with the homozygous mutation load than short ROHs for a population with an effective population size of 100 (which is the approximate size of the Holstein Friesian population [172–174]). Functional predictions of deleterious variation have led to inconsistent conclusions as to whether short or long ROHs harbor more deleterious genetic variants [182, 183]. For the human genome, Szpiech et al. [182] predicted that long ROHs (of several Mb) are enriched with deleterious variants compared to short ROHs. In contrast, for four Danish cattle breeds Zhang et al. [183] predicted that short (< 0.1 Mb) and medium (0.1 to 3 Mb) ROHs are significantly enriched in deleterious variants compared to long (> 3 Mb) ROHs. For domestic dogs, Sams and Boyko [184] recently reported that the relative risk of a ROH carrying a known deleterious variant is similar across ROHs of different lengths, suggesting that ROHs of all lengths may contribute to inbreeding depression in dogs. This latter finding is more in line with our results, where both short and long ROHs seem to contribute to inbreeding depression.

There are various aspects that affect the accuracy of identification of ROHs and the inference of inbreeding age based on ROHs. First, the density of the SNP panel determines the size of ROHs that can be accurately identified. Previous studies have shown that the use of a 50k panel may result in false positive ROHs shorter than 5 Mb and especially in many false positives shorter than 2 Mb [59, 143]. For a more accurate estimation of ancient inbreeding, and to apply this approach to even more generations in the past, high-density SNP data or sequence data is required. Second, in this study we assumed a uniform recombination rate, while it actually varies across the genome [122]. A ROH of a given physical length in a region with high recombination will reflect more ancient inbreeding than a ROH of the same length in a region with low recombination. One may account for this effect by computing ROHs based on genetic distances. However, this is rarely done in practice, since it requires a high-quality linkage map [185]. Third, recent inbreeding may mask more ancient inbreeding [74]. If both chromosomes at a position in the genome trace back to a distant common ancestor, you expect to find a short ROH. If the same region also traces back to a recent common ancestor, then you would observe only the long ROH. As a result, one may expect a negative correlation between recent and ancient ROH-based inbreeding. In Iberian pigs, Saura et al. [74] report such a negative correlation of -0.641 between inbreeding based on short ROHs (0.5 to 5 Mb) and based on long ROHs (> 5 Mb). In this study, we found some negative correlations between the very short ROHs (F_{ROH1-2}) and the other classes (Figure 3.3). However, these negative correlations could also be an artefact of the unreliable estimation of short ROHs. To correct for the masking of ancient inbreeding by recent inbreeding, one could subtract the length of long ROHs of the total length of the genome when calculating F_{ROH} for short ROHs. The effect of this or other correction(s) should be investigated in future studies. Lastly, various approaches can be used to identify ROHs. In this study, we applied the sliding window approach implemented in Plink 2.0 [166], with a set of (rather arbitrary) rules to define a ROH. As an alternative to this rule-based approach, one may use a Hidden Markov model (HMM) to identify ROHs and infer age of inbreeding [185, 186]. In the future, it would be valuable to compare the different approaches and investigate the benefit of using linkage maps to infer inbreeding age based on ROHs.

As sequencing costs continue to decrease, genomic data (including that of cows) will become increasingly available. This offers opportunities to perform largescale analyses on genomic inbreeding depression based on high-density information, e.g. to identify regions associated with inbreeding depression [53, 73, 80]. In addition, genomic time series (consisting of genomic data of an individual and its ancestors) could be used to study purging in more detail at the genomic level.

3.5 Conclusions

Inbreeding depression was observed for yield, fertility and udder health traits in Dutch Holstein Friesian dairy cattle. Observed inbreeding depression was stronger for yield traits than for fertility and udder health traits, when compared in (phenotypic or genetic) standard deviations. Genomic inbreeding captured more inbreeding depression than pedigree-based inbreeding at the population level. For yield traits and based on pedigree information, inbreeding on recent generations was found to be more harmful than inbreeding on distant generations and there was evidence of purging. Based on ROHs, there was no clear difference between the effects of long ROHs (recent inbreeding) and short ROHs (ancient inbreeding). Future work should investigate inbreeding depression and purging in more detail at the genomic level, using higher density information and genomic time series.

3.6 Acknowledgements

The research leading to these results has been conducted as part of the IMAGE project, which received funding from the European Union's Horizon 2020 Research and Innovation Program under the grant agreement no. 677353. The study was co-funded by the Dutch Ministry of Agriculture, Nature and Food Quality (KB-34-013-002). The authors gratefully acknowledge the Dutch-Flemish cattle improvement cooperative (CRV) for providing pedigree and genotype data. The authors also thank the anonymous reviewers and editors for their valuable comments and suggestions.

3.7 Supplementary information

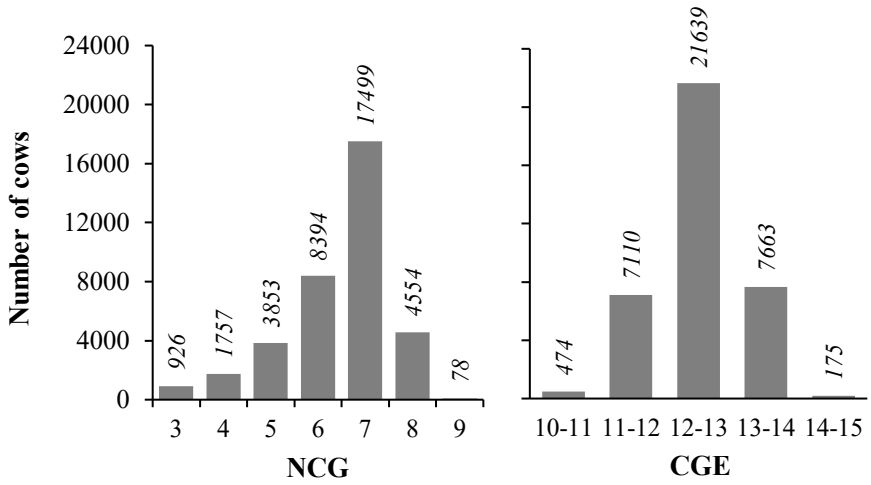


Figure S3.1 Distribution of the number of complete generations (NCG) and complete generation equivalent (CGE) for cows included in pedigree-based analyses (n = 37,061).

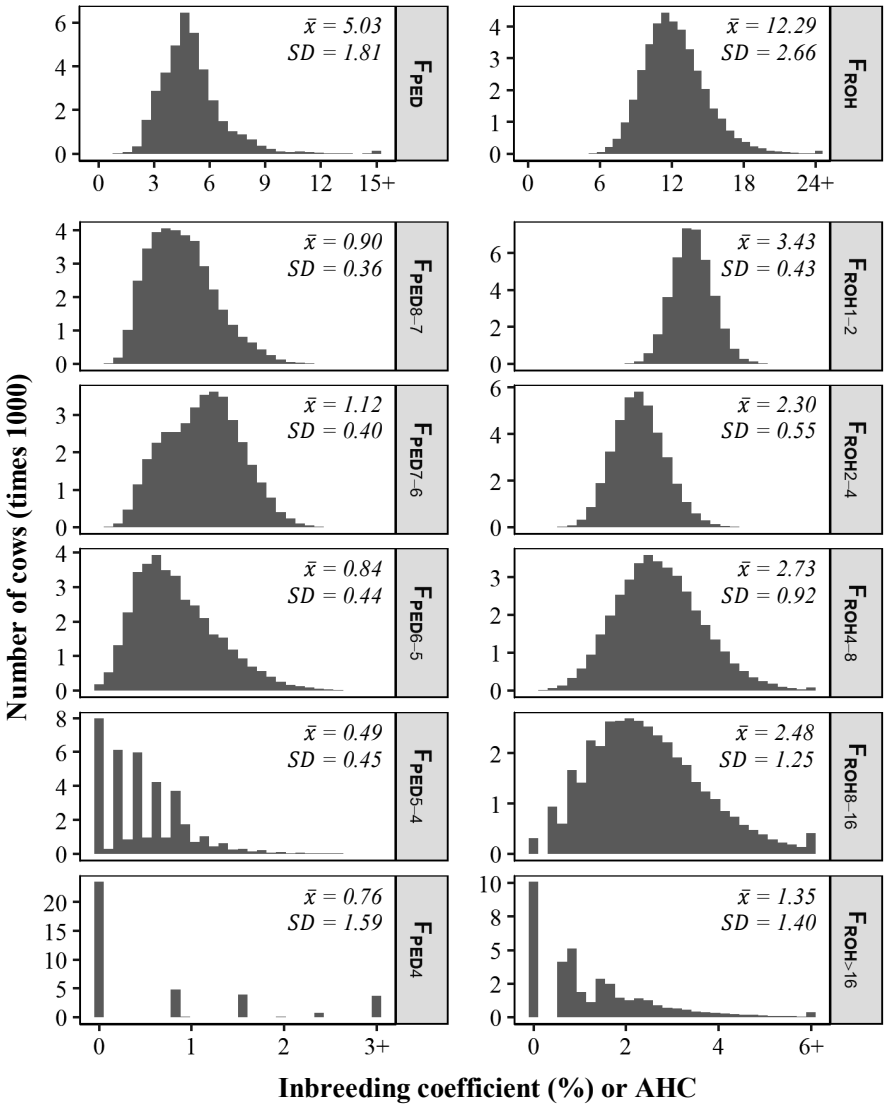


Figure S3.2 Distributions of inbreeding measures and the AHC (*continued on next page*). N = 37,061 for pedigree-based measures and n = 38,792 for genomic measures. The mean ($\bar{x}$) and standard deviation (SD) are also shown. F_{PED}: pedigree inbreeding based on all generations; F_{ROH}: inbreeding based on all regions of homozygosity; F_{GRM}: inbreeding based on genomic relationship matrix computed with allele frequencies of 0.5; F_{PED4}: pedigree inbreeding based on first 4 generations; F_{PED5-4}: difference between pedigree inbreeding based on 5 and on 4 generations; F_{NEW}: Kalinowski's new inbreeding; F_{ANC}: Kalinowski's ancestral inbreeding; AHC: ancestral history coefficient; $F_{ROH>16}$: inbreeding based on regions of homozygosity longer than 16 Mb; $F_{ROH8-16}$: inbreeding based on regions of homozygosity of 8 to 16 Mb.

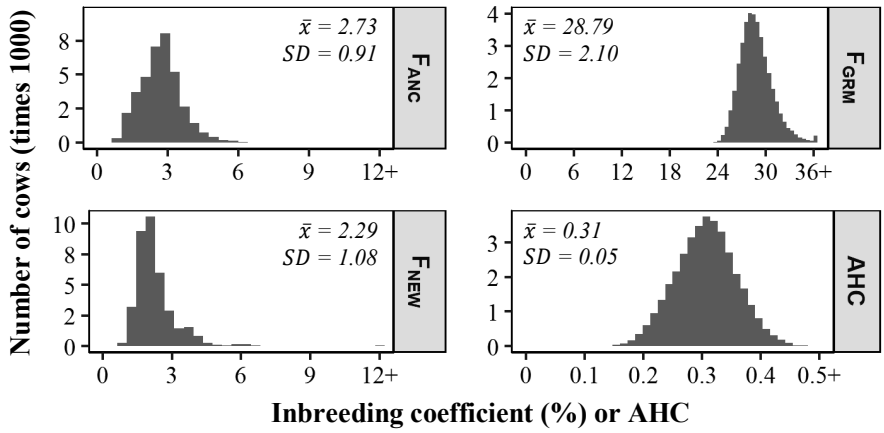


Figure S3.2 (continued)

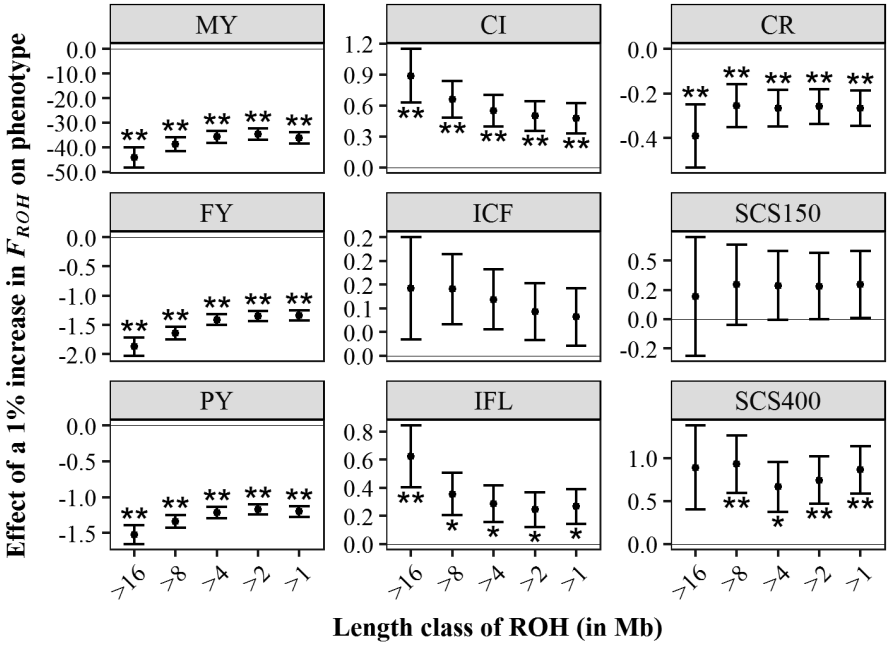


Figure S3.3 Effect of a 1% increase in ROH-based inbreeding (F_{ROH}) for cumulative measures. Error bars represent one standard error and stars indicate significance for non-nullity (*P < 0.05; **P < 0.01). MY: 305-day milk yield (kg); FY: 305-day fat yield (kg); PY: 305-day protein yield (kg); CI: calving interval (days); ICF: interval calving to first insemination (days); IFL: interval first to last insemination (days); CR: conception rate (%); SCS150 somatic cell score day 5 to 150 (units); SCS400: somatic cell score day 151 to 400 (units).

Table S3.1 Estimates of inbreeding depression for all traits and total inbreeding measures, expressed as percentage of trait means (% of $\bar{x}$), in corrected phenotypic standard deviations (σ_p) and in genetic standard deviations (σ_a). The results for σ_p and σ_a were multiplied by 100. Estimates correspond to the estimates in Table 3.2.

	F_{PED}			F_{ROH}			F_{GRM}		
Trait	% of $\bar{x}$	in σ_p	in σ_a	% of $\bar{x}$	in σ_p	in σ_a	% of $\bar{x}$	in σ_p	in σ_a
MY	-0.47	-3.16	-4.60	-0.45	-3.02	-4.39	-0.45	-4.01	-5.83
FY	-0.45	-3.51	-5.42	-0.39	-3.05	-4.72	-0.39	-3.64	-5.63
PY	-0.45	-3.46	-5.76	-0.42	-3.28	-5.46	-0.42	-4.24	-7.05
CI	0.12	0.70	2.47	0.12	0.73	2.57	0.12	0.95	3.35
ICF	0.21	0.60	2.07	0.11	0.30	1.04	0.11	0.34	1.17
IFL	0.33	0.24	1.08	0.67	0.48	2.16	0.67	0.76	3.42
CR	-0.49	-0.87	-5.08	-0.42	-0.74	-4.33	-0.42	-1.02	-5.93
SCS150	0.05	0.43	1.27	0.02	0.22	0.65	0.02	0.33	0.96
SCS400	0.05	0.67	1.76	0.05	0.67	1.77	0.05	0.89	2.35

MY: 305-day milk yield (kg); FY: 305-day fat yield (kg); PY: 305-day protein yield (kg); CI: calving interval (days); ICF: interval calving to first insemination (days); IFL: interval first to last insemination (days); CR: conception rate (%); SCS150 somatic cell score day 5 to 150 (units); SCS400: somatic cell score day 151 to 400 (units).
F_{PED}: pedigree inbreeding based on all generations; F_{ROH}: inbreeding based on all regions of homozygosity; F_{GRM}: inbreeding based on genomic relationship matrix computed with allele frequencies of 0.5.

4

Revised calculation of Kalinowski's ancestral and new inbreeding coefficients *(Communication)*

Harmen P. Doekes[1,2], Ino Curik[3], István Nagy[4],
János Farkas[5], György Kövér[5], Jack J. Windig[1,2]

[1]Wageningen University & Research Animal Breeding and Genomics,
P.O. Box 338, 6700 AH, Wageningen, the Netherlands;
[2]Centre for Genetic Resources the Netherlands, Wageningen University &
Research, P.O. Box 16, 6700 AA, Wageningen, the Netherlands;
[3]Department of Animal Science, Faculty of Agriculture, University of Zagreb,
Svetošimunska cesta 25, 10000 Zagreb, Croatia;
[4]Faculty of Agricultural and Environmental Sciences, Kaposvár University,
P.O. Box 16, H-7400 Kaposvár, Hungary;
[5]Faculty of Economic Science, Kaposvár University,
P.O. Box 16, H-7400 Kaposvár, Hungary;

Diversity (2020) 12(4):155

Abstract

To test for the presence of purging in populations, the classical pedigree-based inbreeding coefficient (F) can be decomposed into Kalinowski's ancestral (F_{ANC}) and new (F_{NEW}) inbreeding coefficients. The F_{ANC} and F_{NEW} can be calculated by a stochastic approach known as gene dropping. However, the only publicly available algorithm for the calculation of F_{ANC} and F_{NEW}, implemented in GRain v2.1 (and also incorporated in the PEDIG software package), has produced biased estimates. The F_{ANC} was systematically underestimated and consequently, F_{NEW} was overestimated. To illustrate this bias, we calculated F_{ANC} and F_{NEW} by hand for simple example pedigrees. We revised the GRain program so that it now provides unbiased estimates. Correlations between the biased and unbiased estimates of F_{ANC} and F_{NEW}, obtained for example data sets of Hungarian Pannon White rabbits (22,781 individuals) and Dutch Holstein Friesian cattle (37,061 individuals), were high, i.e., >0.96. Although the magnitude of bias appeared to be small, results from studies based on biased estimates should be interpreted with caution. The revised GRain program (v2.2) is now available online and can be used to calculate unbiased estimates of F_{ANC} and F_{NEW}.

4.1 Introduction

Inbreeding is the mating between (close) relatives and is unavoidable in genetically small populations. The degree of inbreeding is typically measured with pedigree-based inbreeding coefficients, as introduced by Wright [47]. Individuals with higher inbreeding coefficients show a lower phenotypic performance on average, a phenomenon known as inbreeding depression [53, 68, 79]. Inbreeding depression occurs because part of the genetic load in populations, known as inbreeding load, is only expressed in homozygotes [79]. Inbreeding depression is expected to be largely due to partial dominance, i.e. the existence of (partially) deleterious recessive alleles, although overdominance and epistasis may also play a role [53, 79, 155].

Inbreeding load in a population is not constant, but rather dynamic over time. New deleterious recessive alleles arise continuously by mutation and these alleles are eroded over time by (natural and/or artificial) selection and genetic drift [79]. Inbreeding increases the efficiency of selection against deleterious recessive alleles in a process called purging [79, 187].

To test for the existence of purging in populations, various pedigree-based methods have been proposed [141, 161, 180]. To test for purging in captive wildlife populations, Ballou [141] introduced the ancestral inbreeding coefficient, which is the probability that a random allele in an individual has been previously exposed to inbreeding, i.e., that this allele has been identical-by-descent (IBD) in at least one ancestor. While investigating purging in the captive breeding program of the Speke's gazelle (*Gazella spekei*), Kalinowski et al. [161] extended Ballou's concept by considering the IBD-status of the individual as well. In Kalinowski's approach the total pedigree-based inbreeding coefficient is decomposed into an ancestral (F_{ANC}) and a new (F_{NEW}) inbreeding coefficient. The F_{ANC} is the probability that alleles are IBD in the individual while they were already IBD in at least one ancestor, whereas F_{NEW} is the probability that alleles are IBD for the first time in the individual's pedigree [161].

To calculate F_{ANC} and F_{NEW} (and other inbreeding coefficients), a gene dropping based algorithm has been developed and implemented in GRain software [162]. The GRain algorithm has also been incorporated in the PEDIG package [164], in versions 2007 and later. Various studies have used the GRain algorithm, either in GRain itself [177, 188-191] or in PEDIG [154, 163, 192], to calculate F_{ANC} and F_{NEW}.

The objective of this study was to demonstrate that the previous version of GRain (v2.1) produced biased estimates of F_{ANC} and F_{NEW}. For several simple pedigrees, we show how F_{ANC} and F_{NEW} can be calculated by hand. We also investigate the magnitude of the bias for two example data sets of Hungarian Pannon White rabbits and Dutch Holstein Friesian dairy cattle. A revised version of GRain software (v2.2), which provides unbiased F_{ANC} and F_{NEW} estimates, is now available online.

4.2 Calculation of ancestral and new inbreeding by hand

For simple pedigrees, Kalinowski's ancestral inbreeding ($F_{ANC,X}$) and new inbreeding ($F_{NEW,X}$) coefficients of an individual X can be calculated by hand. To do so, Mendelian inheritance principles are followed, meaning that each allele has a probability of 0.5 to be passed on from parent to offspring. First, Wright's classical inbreeding coefficient (F_X) is determined. The F_X is defined as the probability that the two alleles at a random locus in individual X are IBD, and is calculated as [47]:

$$F_X = \sum_{i=1}^{n} (1 + F_i) \left(\frac{1}{2}\right)^{k_s + k_d + 1}$$

where n is the number of paths connecting the sire of X with the dam of X through the i^{th} common ancestor, F_i is the inbreeding coefficient of the i^{th} common ancestor, and k_s and k_d are the number of generations from, respectively, sire and dam (included) to the i^{th} common ancestor (excluded). Then, $F_{ANC,X}$ is calculated as the probability that X is IBD for an allele, given that this allele was also IBD in at least one of the ancestors of X. Finally, $F_{NEW,X}$ is obtained by subtracting $F_{ANC,X}$ from F_X, since the ancestral and new inbreeding sum up to the total inbreeding.

In Figure 4.1, four example pedigrees are shown. The corresponding inbreeding coefficients are provided in Table 4.1. In example (1), the F_X equals 0.0078 (0.5^7), because there is a single path that connects parents F and G through common ancestor A, which is of length 7 ($k_s + k_d + 1 = 7$), and ancestor A is non-inbred ($F_A = 0$). The $F_{ANC,X}$ for this example is 0, because none of the ancestors of X are inbred. Consequently, $F_{NEW,X}$ is equal to F_X (so 0.0078).

In example (2), the F_X equals 0.0703, because it is the inbreeding on ancestor D (0.5^4) multiplied with (1 + F_D), where F_D is the inbreeding coefficient of ancestor D (0.5^3). The $F_{ANC,X}$ is calculated as the probability that X is IBD for an allele that was IBD in D as well. Since D is the only inbred ancestor, we do not need to consider the IBD status of any other ancestors. The probability that D is IBD for an allele from its grandparent A, is the inbreeding coefficient of D on A and equals 0.125 (0.5^3). To obtain $F_{ANC,X}$, this probability has to be multiplied with the probabilities that the allele is passed on to X, through both the paths D-E-F-X and D-G-X. The probability that E inherits the allele from D is simply 1, because D is IBD. The probability that F inherits the allele from E is 0.5 and that X inherits it from F is also 0.5, so the total probability for the path D-E-F-X is 0.25 (0.5^2). Similarly, the probability for path D-G-X is 0.5. This gives a total probability of 0.125 × 0.25 × 0.5 = 0.0156 for $F_{ANC,X}$. Consequently, $F_{NEW,X}$ = F_X − $F_{ANC,X}$ = 0.0703 − 0.0156 = 0.0547. Note that, in this example, the $F_{ANC,X}$ can also be calculated as two times the inbreeding coefficient of X on D (0.5^4), multiplied with the inbreeding coefficient of D on A (0.5^3). However, it is important to realize that this reasoning only holds for scenarios in which one

inbreeding loop is "on top of the other", and not when there is an overlap in inbreeding loops, such as in examples (3) and (4).

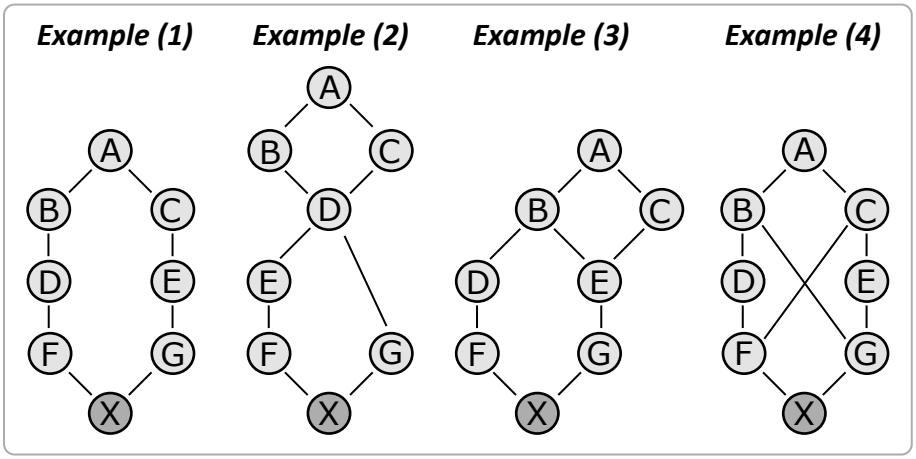


Figure 4.1 Example pedigrees for the calculation of classical and Kalinowski's inbreeding coefficients. X is the individual of interest and the other letters represent ancestors of X. Inbreeding coefficients for individual X, corresponding to the example pedigrees, are shown in Table 4.1.

Table 4.1 Inbreeding coefficients for four example pedigrees (Figure 4.1), estimated with revised and previous version of GRain.

		Revised Version (v2.2)		Previous Version (v2.1)		Difference
Example	F_X	$F_{ANC,X}$	$F_{NEW,X}$	$F_{ANC,X}$	$F_{NEW,X}$	in $F_{ANC,X}$
(1)	0.0078	0	0.0078	0	0.0078	0
(2)	0.0703	0.0156	0.0547	0.0156	0.0547	0
(3)	0.0390	0.0078	0.0312	0.0039	0.0351	0.0039
(4)	0.1641	0.0390	0.1250	0.0234	0.1406	0.0156

F_X: classical inbreeding coefficient of individual X; $F_{ANC,X}$: Kalinowski's ancestral inbreeding coefficient of individual X; $F_{NEW,X}$: Kalinowski's new inbreeding coefficient of individual X.

In example (3), the F_X equals 0.0390 and is the sum of inbreeding on ancestor A (0.5^7) and on ancestor B (0.5^5). The $F_{ANC,X}$ is calculated as the probability that X is IBD for an allele that was IBD in ancestor E as well. Since ancestor E is the only inbred ancestor, we do not need to consider the IBD status of any other ancestors. The probability that E is IBD for an allele from its grandparent A, is the inbreeding coefficient of E on A and equals 0.125 (0.5^3). This probability has to be multiplied by the probability that this allele is passed on to X through both the paths E-G-X and B-D-F-X. The probability that G inherits the allele from E is 1, because E is IBD. The probability that X inherits the allele from G is 0.5, so the total probability for the path E-G-X is 0.5. The probability that B carries the allele is 1, otherwise E could not have

been IBD. The probability that the allele is passed on from B to D to F and to X is 0.125 (0.5^3). This gives a total probability of 0.125 × 0.125 × 0.5 = 0.0078 for $F_{ANC,X}$. Consequently, $F_{NEW,X} = F_X - F_{ANC,X}$ = 0.0390 − 0.0078 = 0.0312.

In example (4), the F_X equals 0.1641 and is the sum of inbreeding on ancestor A ($0.5^7 + 0.5^5$), on ancestor B (0.5^4) and on ancestor C (0.5^4). The $F_{ANC,X}$ in this example is the probability that X is IBD for an allele that was also IBD in F and/or G (since F and G are inbred ancestors). The $F_{ANC,X}$ is the sum of the probabilities for three scenarios: (i) X is IBD for an allele that was IBD in both F and G, (ii) X is IBD for an allele that was IBD in F, but not in G, and (iii) X is IBD for an allele that was IBD in G, but not in F. The probability that F is IBD for an allele from A is the inbreeding coefficient of F on A and equals 0.0625 (0.5^4). If F is IBD for an allele from A, then both B and C must be carriers of that allele, and the probability that G is also IBD for that same allele is 0.125 (0.5^3), since this is the probability that G inherits that allele through B-G (0.5) multiplied with the probability that G inherits that allele through C-E-G (0.5^2). When F and G are IBD for the same allele, X has to be IBD for that allele as well. Therefore, the probability that scenario (i) happens is 0.0078 (i.e., 0.0625 × 0.125 × 1). If F is IBD for an allele from A, the probability that G carries two other "unknown" alleles is 0.375 (i.e., 0.5 × (1 − 0.5^2)), leaving 1 − 0.125 − 0.375 = 0.5 for the probability that G carries one copy of the allele and one copy of an unknown allele (scenario ii). In that case, the probability that the allele is inherited by X from G is 0.5. The total probability for scenario (ii) is therefore 0.0156 (i.e., 0.0625 × 0.5 × 0.5). Due to the symmetry in the pedigree, the probability for scenario (iii) is equal to that of scenario (ii), so 0.0156. Thus, the total probability that X is IBD for an allele that was also IBD in F and/or G, i.e., the $F_{ANC,X}$, equals 0.0078 + 0.0156 + 0.0156 = 0.0391. Consequently, $F_{NEW,X} = F_X - F_{ANC,X}$ = 0.1641 − 0.0391 = 0.1250.

4.3 Underestimation of ancestral inbreeding by previous version of GRain

In GRain, a stochastic approach known as gene dropping [125] is implemented to calculate inbreeding coefficients. In this approach, many independent simulations are run. In each simulation, alleles are dropped through the pedigree following Mendelian inheritance rules, and the IBD-status of individuals is stored. After all simulations are completed, the F_X is estimated as the fraction of simulations in which the alleles of individual X were IBD. Similarly, the $F_{ANC,X}$ is calculated as the fraction of simulations in which X was IBD for an allele that was already IBD in one of the ancestors of X. The accuracy of the estimated inbreeding coefficients is higher when more simulations are run. As shown by Baumung et al. [162], using 10^6 simulations provides estimates of inbreeding coefficients that show a correlation of >0.999 with inbreeding coefficients calculated using a deterministic approach (with only minor

differences at the fourth decimal). A more detailed explanation of the GRain program and its computational demands is given by Baumung et al. [162].

When $F_{ANC,X}$ was computed using the previous version of GRain (v2.1), the $F_{ANC,X}$ for examples (1), (2), (3) and (4) from Figure 4.1 equaled 0, 0.0156, 0.0039 and 0.0234, respectively (Table 4.1). Although the coefficients for examples (1) and (2) were correct, the $F_{ANC,X}$ for examples (3) and (4) was underestimated. Note that example (3) is equivalent to the example used by McParland et al. [154], in Figure 1A in their paper, for which they reported the incorrect $F_{ANC,X}$ estimate of 0.0039.

The underestimation of $F_{ANC,X}$ was occasionally caused by an incorrect tracking of IBD-status of ancestors throughout the pedigree. In the previous version of GRain (v2.1), every individual was given a flag that indicated whether one of their ancestors had been IBD (1 if true, 0 if false). This flag was calculated as the sum of the flags of the parents, divided by two. Thus, when both parents had a flag of 1, the flag of the offspring would also be 1, which is correct. However, when only one of the parents had a flag of 1 (and the other 0), the offspring would get a value of 0.5, which is incorrect (since it should be 1). In the revised version of GRain (v2.2), this issue was solved by obtaining the flag of an offspring as the maximum of the flags of its parents.

To clarify, in example (2) in Figure 4.1, whenever ancestor D was IBD, both parents F and G had a flag of 1 and X also got a flag of 1. Therefore, the $F_{ANC,X}$ was estimated correctly. In example (3), however, whenever ancestor E was IBD, parent G had a flag of 1 and parent F had a flag of 0 and, as a result, X got a flag of 0.5. Consequently, for simulations in which individual X was IBD for an allele that was also IBD in E, a value of 0.5 was stored (instead of 1) for the $F_{ANC,X}$ calculation. After simulations were completed, the stored values were summed across simulations and divided by the total number of simulations. Since stored values were underestimated by a factor two, the $F_{ANC,X}$ for example (3) was also underestimated by a factor two. In example (4), whenever both F and G were IBD, X got a flag of 1. This happened in 0.0078 of the simulations (see explanation in the previous section for calculation by hand, scenario (i)). When only parent F or parent G were IBD, while the other parent was not, X got a flag of 0.5. This happened in 0.0156 + 0.0156 = 0.0312 of the simulations (see explanation in the previous section for calculation by hand, scenarios (ii) and (iii)). Therefore, the $F_{ANC,X}$ for example (4) was underestimated by some factor between one and two. More specifically, the underestimated $F_{ANC,X}$ was equal to 0.0078 + (0.5 × 0.0312) = 0.0234.

4.4 Examples for Pannon White Rabbits and Holstein Friesian Cattle

To investigate the impact of the incorrect estimation, we computed F_{ANC} and F_{NEW} for two example data sets, using both the previous and revised version of GRain, and

10^6 simulations. The first data set was a pedigree of 22,781 rabbits of the Hungarian Pannon White (PW) breed. This pedigree included 6,760 rabbits (1,421 bucks and 5,339 does) with offspring and 16,021 rabbits without offspring. All rabbits were born between 1992 and 2016. To assess pedigree completeness, the number of complete generations (NCG) and the complete generation equivalent (CGE) were computed for each rabbit. The CGE was computed as the sum of $(1/2)^n$ of all known ancestors of an individual, with n being the number of generations between the individual and a given ancestor. The mean NCG in the PW pedigree was 4.0 (ranging from 0 to 10) and the mean CGE was 8.6 (ranging from 0 to 22.1). The second data set contained 37,061 Dutch Holstein Friesian (HF) cows, which were part of a larger pedigree of 167,924 individuals (19,363 bulls and 148,561 cows) and were used by Doekes et al. [193]. These HF cows were born between 2012 and 2016 and were filtered to have a high pedigree completeness (NCG ≥ 3 and CGE ≥ 10), and have phenotypic information on 305-day milk, fat and protein yields. The mean NCG in these HF cows was 6.5 generations (ranging from 3 to 9) and the mean CGE was 12.5 generation equivalents (ranging from 10.0 to 14.7). More details on the HF data set can be found in Doekes et al. [193].

For both the PW and HF data set, the total inbreeding coefficients (F) were identical across the previous and revised version of GRain. The F_{ANC} in the previous version however, was generally underestimated and the F_{NEW} was overestimated (Figure 4.2). For the PW data set and inbreeding coefficients above zero, the F_{ANC} from the previous version was on average 0.65 times the revised F_{ANC} (and the F_{NEW} was 1.27 times the revised F_{NEW}). For the HF data set and inbreeding coefficients above zero, the F_{ANC} from the previous version was on average 0.71 times the revised F_{ANC} (and the F_{NEW} was 1.36 times the revised F_{NEW}). Pearson correlation coefficients between coefficients estimated with the previous and revised version were high. For the PW data set, the correlations between the previous and revised version equaled 0.997 and 0.968 for F_{ANC} and F_{NEW}, respectively. For the HF data set, these correlations equaled 0.993 and 0.987, respectively. This indicates that the underestimation of F_{ANC} (and overestimation of F_{NEW}) did not strongly affect the ranking of animals.

For the HF data set, we also investigated the potential differences in inbreeding depression estimates for F_{ANC} and F_{NEW}, calculated with the previous and revised version of GRain. A linear mixed model was run in ASReml 4.1 [168], in which F_{ANC} and F_{NEW} were fitted as fixed effects and the regression coefficients on F_{ANC} and F_{NEW} were used as estimates of inbreeding depression (see Doekes et al. [193] for a detailed explanation). In general, differences between inbreeding depression estimates based on the previous and revised version of GRain were small (Figure 4.3). For example, the effect of a 1% increase in F_{NEW} on 305-day milk yield was

−46.4 kg (SE = 12.4 kg) for the previous version and −47.3 kg (SE = 11.2 kg) for the revised version. Standard errors for the inbreeding depression effects appeared smaller when the revised version was used to estimate F_{ANC} and F_{NEW}, compared to when the previous version was used. For example, the mean standard error of inbreeding depression estimates for fat and protein yields was 0.51 kg for the revised version, and 0.67 kg for the previous version. The overall conclusion, that F_{NEW} was associated with significant inbreeding depression, while F_{ANC} was not, was the same for both versions. Based on these findings, we expect that conclusions from other studies using F_{ANC} and F_{NEW} estimates from GRain v2.1 (e.g. [154, 163]) will also largely hold. However, they should be interpreted with caution.

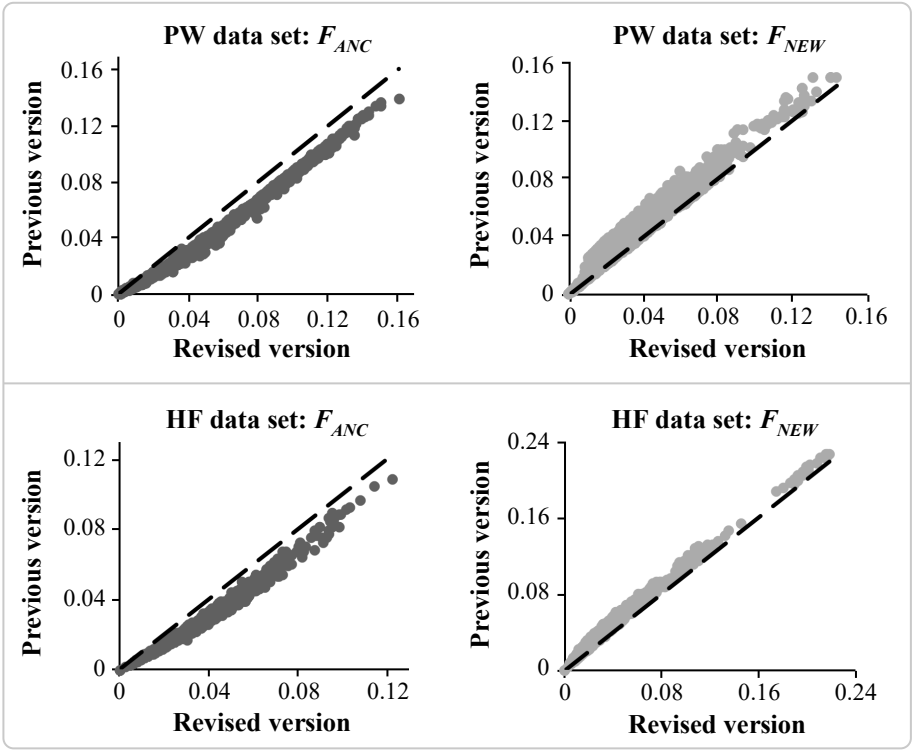


Figure 4.2 Relationship between Kalinowski's inbreeding coefficients calculated with previous (v2.1) and revised (v2.2) version of GRain, for example data sets of Pannon White rabbits (PW; n = 22,781) and Holstein Friesian cattle (HF; n = 37,061). The dashed line indicates *y = x*, i.e. a relationship in which there is no difference in estimation between the two GRain versions. F_{ANC}: Kalinowski's ancestral inbreeding. F_{NEW}: Kalinowski's new inbreeding.

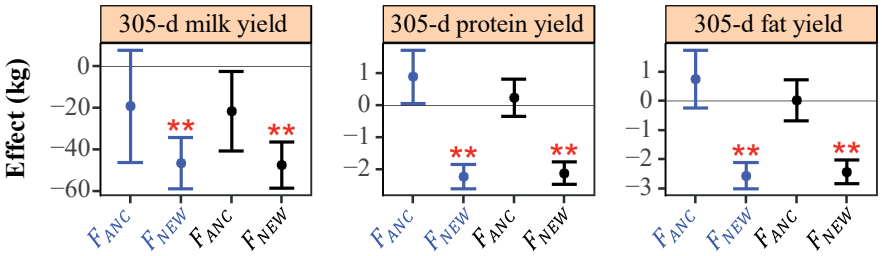


Figure 4.3 Effect of a 1% increase in Kalinowski's ancestral (F_{ANC}) and new (F_{NEW}) inbreeding on yield traits in Dutch Holstein Friesian cattle (n = 37,061), for F_{ANC} and F_{NEW} calculated with the previous (v2.1, in blue) and revised (v2.2, in black) version of GRain. Red stars indicate effects that significantly differed from zero (P < 0.001).

4.5 Conclusions

The previous version of GRain software (v2.1) systematically underestimated Kalinowski's ancestral inbreeding coefficients and, consequently, overestimated Kalinowski's new inbreeding coefficients. Although the magnitude of bias was rather small, results from studies based on biased estimates should be interpreted with caution. The GRain software has been revised, and the revised version (v2.2), which provides unbiased estimates of Kalinowski's coefficients, can be downloaded from [194] or [195].

4.6 Funding & acknowledgements

Calculations performed on the Dutch Holstein Friesian population were conducted as part of the IMAGE project, which received funding from the European Union's Horizon 2020 Research and Innovation Programme under the grant agreement no 677353. Calculations performed on the Pannon White rabbit population were supported by the Hungarian Scientific Research Fund (OTKA) K 128177 project. The study was co-funded by the Dutch Ministry of Agriculture, Nature and Food Quality (KB-34-013-002). The authors would like to thank the Dutch-Flemish cattle improvement cooperative (CRV) for providing the Holstein Friesian data.

5

Inbreeding depression across the genome of Dutch Holstein Friesian dairy cattle

Harmen P. Doekes[1,2], Piter Bijma[1], Roel F. Veerkamp[1],
Gerben de Jong[3], Yvonne C.J. Wientjes[2], Jack J. Windig[1,2]

[1]Wageningen University & Research Animal Breeding and Genomics,
P.O. Box 338, 6700 AH, Wageningen, the Netherlands;
[2]Centre for Genetic Resources the Netherlands, Wageningen University &
Research, P.O. Box 16, 6700 AA, Wageningen, the Netherlands
[3]Cooperation CRV, Wassenaarweg 20, 6843 NW Arnhem, the Netherlands

Submitted

Abstract

Inbreeding depression refers to the decrease in mean performance due to inbreeding. Inbreeding depression is caused by an increase in homozygosity and reduced expression of, on average favorable, dominance effects. Dominance effects and allele frequencies differ across loci, and consequently inbreeding depression is expected to differ across the genome. In this study, we investigated differences in inbreeding depression across the genome of Dutch Holstein Friesian cattle, by estimating dominance effects and effects of regions of homozygosity (ROH).

Genotype (75 k) and phenotype data of 38,792 first-parity cows were used. For nine yield, fertility and udder health traits, GREML models were run to estimate genome-wide inbreeding depression and estimate additive, dominance and ROH variance components. For this purpose, we introduced a ROH-based relationship matrix. Additive, dominance and ROH effects per SNP were obtained through backsolving. In addition, a single SNP GWAS was performed to identify significant additive, dominance or ROH associations.

Genome-wide inbreeding depression was observed for all yield, fertility and udder health traits. For example, a 1% increase in genome-wide homozygosity was associated with a decrease in 305-d milk yield of approximately 99 kg. For yield traits only, including dominance and ROH effects in the GREML model resulted in a better fit ($P < 0.05$), compared to a model with only additive effects. After correcting for genome-wide inbreeding depression, dominance and ROH variance explained less than 1% of phenotypic variance for all traits. Furthermore, dominance and ROH effects were evenly distributed across the genome. The most notable region with a favorable dominance effect for yield traits was found on chromosome 5, but overall few regions with large favorable dominance effects and significant dominance associations were found. No significant ROH-associations were found.

In conclusion, inbreeding depression was distributed quite equally across the genome and was well captured by genome-wide homozygosity. Based on these findings, there is little benefit of accounting for region-specific inbreeding depression in selection schemes.

5.1 Introduction

Inbreeding depression refers to the decrease in mean performance with increased levels of inbreeding [24]. Many important traits in dairy cattle show inbreeding depression [80, 152-154]. For example, a 1% increase in pedigree-based inbreeding has been associated with a decrease in 305-day milk yield of 20 to 38 kg and with an increase in calving interval of 0.2 to 0.7 days [71, 169, 193].

The reduction in mean performance is believed to be caused by the increase in homozygosity associated with inbreeding, reducing the expression of dominance effects [24, 53]. When dominance effects are on average favorable (i.e. when there is *directional dominance* in the favorable direction), their reduced expression results in a lower phenotypic performance.

Not all genomic loci are expected to contribute equally to inbreeding depression. The expected contribution of a locus depends on both its dominance effect (higher with larger dominance effect) and its allele frequency (higher at intermediate allele frequencies) [24, 53]. Interactions between loci, i.e. epistasis, may play a role in explaining inbreeding depression as well. Epistasis, however, is difficult to prove and difficult to account for in statistical models. Therefore, epistasis is typically ignored. When epistasis is ignored, the change in mean phenotypic performance due to inbreeding equals $-F\sum_i 2p_iq_id_i$, where F is the genome-wide inbreeding coefficient, d_i is the dominance effect at the i^{th} locus, and p_i and q_i are the allele frequencies [24].

The increasing availability of single nucleotide polymorphism (SNP) data enables to study differences in inbreeding depression across the genome. SNPs are expected to capture effects of quantitative trait loci (QTL) in linkage disequilibrium (LD) with the SNPs. Traditionally, single SNP genome-wide association studies (GWAS) have been conducted to identify significant dominance (and additive) associations [196-198]. In such studies, one SNP is fitted at a time and typically a pedigree-based or genomic relationship matrix is included to account for population structure and prevent inflation of type I errors (e.g. [198]). With more novel approaches, all SNP effects can be estimated simultaneously. For example, a dominance relationship matrix can be computed [199] and this matrix can be fitted in a genomic-relatedness-matrix residual maximum likelihood (GREML) model, after which dominance effects of single SNPs can be obtained through backsolving (e.g. [200]). A GREML model is similar to a genomic best linear unbiased prediction (GBLUP) model, but GREML estimates the polygenic SNP effects (e.g. breeding values) and variance components simultaneously, whereas GBLUP assumes known variances [201]. Benefits of GREML, as compared to a single SNP GWAS, are that all SNP effects are simultaneously estimated (i.e. accounting for other SNPs in LD) and effects are regressed towards

the mean depending on information in the data. In addition, GREML estimates the amount of phenotypic variance attributable to dominance variance.

In addition to the estimation of dominance effects, there is an increasing interest in the use of regions of homozygosity (ROH) to quantify inbreeding and inbreeding depression [53, 73, 80, 153]. The use of ROHs, as compared to homozygosity at single SNPs, has two potential advantages. First, ROHs may better capture homozygosity between SNPs and could therefore better capture dominance effects at loci between SNPs. Second, ROHs capture more recent inbreeding, which is expected to be more harmful than old inbreeding, although empirical results do not always support this hypothesis [80, 193]. In a simulation study, Keller et al. [113] found that, among the inbreeding measures they investigated, ROH-based inbreeding performed best in capturing the homozygous inbreeding load. Martikainen et al. [153] estimated the effect of ROH-based inbreeding on fertility traits in Finnish Ayrshire cattle, first per chromosome and then within chromosomes using a sliding window approach. Pryce et al. [80] performed a single SNP GWAS to study the effect of ROHs on yield traits and calving interval in Australian Holstein and Jersey cattle. In their approach, the ROH-status of a SNP was set to 1 when the SNP was in a ROH (irrespective of which ROH), and to 0 otherwise [80]. Ferenčaković et al. [73] performed a similar analysis for sperm quality traits in Austrian Fleckvieh bulls. Although these studies did report candidate regions associated with inbreeding depression, they did not consider how much of the total phenotypic variation was explained by ROH effects (in relation to additive and dominance effects).

The objective of this study was to investigate different measures of inbreeding depression (dominance, ROH) across the genome for Dutch Holstein Friesian dairy cattle and estimate their contribution to the phenotypic variance. For various yield, fertility and udder health traits, we first ran GREML models to estimate genome-wide inbreeding depression and estimate the amount of variance attributable to additive, dominance and ROH effects. We then obtained individual SNP effects through backsolving. We also performed a traditional single SNP GWAS to estimate additive, dominance and ROH effects per SNP and compared GWAS estimates with those obtained from the GREML approach.

5.2 Material and methods

5.2.1 Animals and data

A total of 38,792 first-parity cows (fraction Holstein Friesian > 87.5%, either red or black), which calved in the period 2012-2016 in 233 herds, were included. The same data set was used as in Doekes et al. [193]. Genotype and phenotype data were provided by the Dutch-Flemish cattle improvement co-operative (CRV; Arnhem, the Netherlands). Cows were genotyped with the Illumina BovineSNP50 BeadChip (v1

and v2) or CRV custom-made 60 k Illumina panel (v1 and v2). Genotypes were imputed to approximately 76 k, following Druet et al. [115]. The 75,538 SNPs used by Doekes et al. [193] were remapped to the ARS-UCD1.2 assembly, using the NAGRP Data Repository [202] and the NCBI Genome Remapping Service [203]. The final data set comprised 75,377 successfully remapped SNPs.

Phenotypic data included yield, fertility and udder health traits. For yield, the 305-day milk yield (MY; in kg), 305-day fat yield (FY; in kg) and 305-day protein yield (PY; in kg) were included. For fertility, the calving interval (CI; in days), interval calving to first insemination (ICF; in days), interval first to last insemination (IFL; in days) and conception rate (CR; in %) were included. For udder health, the mean somatic cell scores for day 5 through to 150 (SCS150; in units) and day 151 through to 400 (SCS400; in units) were included. Somatic cell scores were calculated as 1000+100*[log2 of cells/mL]. The number of cows with phenotypes ranged from 34,774 to 38,778. Descriptive statistics can be found in Doekes et al. [193].

5.2.2 Identification of ROH

Regions of homozygosity (ROH) were identified with Plink 2.0 [166]. The following criteria were used to define a ROH: (i) a minimum physical length of 1 Mb, (ii) a minimum of 15 SNPs, (iii) a minimum density of 1 SNP per 100 kb, (iv) a maximum of 1 heterozygous call within a ROH, and (v) a maximum gap of 500 kb between two consecutive SNPs. A scanning window of 15 SNPs was used, with a maximum of 1 heterozygote call per window. The Plink command was *"plink --cow --homozyg --homozyg-density 100 --homozyg-gap 500 --homozyg-het 1 --homozyg-kb 1000 --homozyg-snp 15 --homozyg-window-het 1 --homozyg-window-snp 15"*. The use of criteria like a maximum gap of 500 kb will have resulted in some SNPs having a lower probability to be in a ROH (e.g. there were 66 gaps of >500 kb), but will also have reduced the number of false positive ROHs.

5.2.3 Statistical models

Additive, dominance and ROH effects were estimated with two approaches: (i) a GREML model with backsolving, and (ii) a single SNP GWAS. For both approaches, the classical ("statistical") parametrization was used, which implies among others that additive effects were calculated as allele substitution effects (see [199]).

GREML with backsolving

GREML models were used to estimate all SNP effects simultaneously and to estimate variance components. For each trait, three models were run in mtg2 [204]: one with only additive effects (A), one with additive and dominance effects (AD), and one with additive, dominance and ROH effects (ADR). Model A was:

$$\text{(A)}\ \ \mathbf{y} = \mathbf{Xb} + \mathbf{Qc} + \mathbf{u} + \mathbf{e}$$

where $\mathbf{y}$ was a vector of phenotypes; $\mathbf{X}$ was an incidence matrix that related the observations to fixed effects; $\mathbf{b}$ was a vector of fixed effects that included herd of calving (233 levels), year of calving (5 levels), season of calving (4 levels, defined as the four quarters of a year), age at calving (as linear covariate) and genome-wide SNP homozygosity (as linear covariate, to account for genome-wide inbreeding depression); $\mathbf{Q}$ was an incidence matrix that related observations to random herd-year-season effects; $\mathbf{c}$ was a vector of random herd-year-season effects (4,596 levels), which were assumed to be distributed as $\mathbf{c}$ ~ N(0, $\mathbf{I}\sigma^2_{HYS}$), with $\mathbf{I}$ being an identity matrix and σ^2_{HYS} the herd-year-season variance; $\mathbf{u}$ was a vector of random polygenic additive effects (i.e. breeding values), which were assumed to be distributed as $\mathbf{u}$ ~ N(0, $\mathbf{G}\sigma^2_A$), with $\mathbf{G}$ being the genomic relationship matrix and σ^2_A the additive genetic variance; and $\mathbf{e}$ was a vector of random residuals, which were assumed to be distributed as $\mathbf{e}$ ~ N(0, $\mathbf{I}\sigma^2_E$), with $\mathbf{I}$ being an identity matrix and σ^2_E the residual variance.

Model A was extended to model AD by adding a dominance term:

$$\text{(AD)}\ \ \mathbf{y} = \mathbf{Xb} + \mathbf{Qc} + \mathbf{u} + \mathbf{v} + \mathbf{e}$$

where $\mathbf{v}$ was a vector of random polygenic dominance deviations, which were assumed to be distributed as $\mathbf{v}$ ~ N(0, $\mathbf{D}\sigma^2_D$), with $\mathbf{D}$ being the dominance relationship matrix and σ^2_D the dominance variance.

Model AD was further extended to model ADR by adding a ROH term:

$$\text{(ADR)}\ \ \mathbf{y} = \mathbf{Xb} + \mathbf{Qc} + \mathbf{u} + \mathbf{v} + \mathbf{w} + \mathbf{e}$$

where $\mathbf{w}$ was a vector of random polygenic ROH deviations, which were assumed to be distributed as $\mathbf{w}$ ~ N(0, $\mathbf{R}\sigma^2_{ROH}$), with $\mathbf{R}$ being a ROH-based relationship matrix and σ^2_{ROH} the ROH variance.

The additive genomic relationship matrix ($\mathbf{G}$) was computed with *calc_grm* [118], according to VanRaden [54]:

$$\mathbf{G} = \frac{\mathbf{ZZ}'}{\sum_i 2p_i q_i}$$

where p_i was the allele frequency of allele A at the *i*th SNP, q_i was the allele frequency of allele B at the *i*th SNP and $\mathbf{Z}$ was the additive marker covariate matrix with elements of $-2p_i$, $1 - 2p_i$, and $2 - 2p_i$ for genotypes BB, AB, and AA, respectively.

The dominance relationship matrix ($\mathbf{D}$) was computed with *calc_grm* [118], according to Vitezica et al. [199]:

$$\mathbf{D} = \frac{\mathbf{HH}'}{\sum_i (2p_i q_i)^2}$$

where $\mathbf{H}$ was the dominance marker covariate matrix with elements of $-2{p_i}^2$, $2p_i q_i$, $-2{q_i}^2$ for genotypes BB, AB, and AA, respectively.

The ROH-based relationship matrix ($\mathbf{R}$) was introduced here to quantify the effect of a SNP being in a ROH (irrespective of which ROH). The $\mathbf{R}$ was computed as:

$$\mathbf{R} = \frac{\mathbf{MM}'}{\sum_i p_i q_i}$$

where p_i was the frequency of SNP *i* being in a ROH, q_i was the frequency of SNP *i* not being in a ROH and $\mathbf{M}$ was the ROH marker covariate matrix with elements of $1 - p_i$ for being in a ROH and of $-p_i$ for not being in ROH. To obtain the $\mathbf{R}$-matrix, the 0/1 ROH-status was first converted to 0/2 values and then VanRaden's formula [54] was applied on these 0/2 values (such that scaled genotype counts were either $0 - 2p_i$ or $2 - 2p_i$) in *calc_grm* [118]. The relationships were divided by a factor 2 to adjust them to the right scale (see Supplementary file 1 for justification).

Goodness of fit of the A, AD and ADR models were compared using maximum likelihood (ML) ratio tests. Test statistics were defined as two times the difference between the maximum log likelihood of a reduced model (e.g. model A) and that of a full model (e.g. model AD). Approximate P-values were calculated as $0.5(1 - \mathrm{P}(\chi_1^2 \leq T))$, where T was the test statistic.

Variance components were directly obtained from mtg2 output. Relative variance components and corresponding standard errors were calculated using the "delta method" in mtg2 [204]. The relative dominance variance, for example, was calculated as σ_D^2/σ_P^2, where σ_P^2 was the phenotypic variance (which excluded σ_{HYS}^2).

To estimate additive effects ($\widehat{\boldsymbol{\alpha}}$), dominance effects ($\hat{\mathbf{d}}$) and ROH effects ($\hat{\mathbf{r}}$) per SNP, the polygenic additive effects ($\widehat{\mathbf{u}}$), dominance deviations ($\hat{\mathbf{v}}$) and ROH deviations ($\widehat{\mathbf{w}}$) were backsolved using the *"compute SNP-effects"* program of *calc_grm* [118], according to:

$$\widehat{\boldsymbol{\alpha}} = \frac{\mathbf{Z}'\mathbf{G}^{-1}\widehat{\mathbf{u}}}{\sum_i 2p_i q_i}$$

$$\hat{\mathbf{d}} = \frac{\mathbf{H}'\mathbf{D}^{-1}\hat{\mathbf{v}}}{\sum_i (2p_i q_i)^2}$$

$$\hat{\mathbf{r}} = \frac{\mathbf{M}'\mathbf{R}^{-1}\widehat{\mathbf{w}}}{\sum_i p_i q_i}$$

where all parameters were defined as before. Note that for additive and dominance effects the p_i and q_i were allele frequencies, whereas for ROH effects the p_i and q_i

were the frequencies of a SNP being in a ROH or not. The backsolving procedure was verified, by recalculating polygenic effects from the backsolved SNP effects.

Note that the dominance deviations ($\hat{\mathbf{v}}$) and dominance SNP effects ($\hat{\mathbf{d}}$) did not include directional dominance, because the mean dominance was already absorbed by the fixed regression on genome-wide homozygosity. The mean dominance effect across loci can be calculated as $-b/n_{SNP}$, where b is the regression coefficient for genome-wide homozygosity and n_{SNP} is the number of SNPs [205, 206]. In this study, we report the σ_D^2 and the dominance effects as obtained from GREML and backsolving output (thus, excluding mean dominance). However, we also computed the mean dominance effect (i.e. $-b/n_{SNP}$) and investigated the effect of correcting σ_D^2 for this mean dominance effect (see Discussion).

Single SNP GWAS

A single SNP GWAS was performed to estimate additive, dominance and ROH effects as fixed effects per SNP. For this purpose, GREML model A was extended by adding a fixed additive, dominance and ROH effect at a specific SNP. For each SNP, the following model was run with Snappy [207] in Wombat [208]:

$$\mathbf{y} = \mathbf{j}\alpha + \mathbf{k}d + \mathbf{l}r + \mathbf{Xb} + \mathbf{Qc} + \mathbf{u} + \mathbf{e}$$

where $\mathbf{j}$ was a vector with allele counts (coded as 0, 1, and 2 for genotypes BB, AB, and AA); α was the additive effect; $\mathbf{k}$ was a vector with heterozygosity status (coded as 0, 1, and 0 for genotypes BB, AB, and AA); d was the dominance effect; $\mathbf{l}$ was a vector with ROH status (coded as 1 when the SNP was in a ROH, or 0 otherwise); and r was the ROH-effect. The other parameters were defined as in GREML model A.

Solutions and t-statistics were obtained from the output, and corresponding P-values were computed. Genomic inflation was assessed using QQplots and genomic inflation factors. The latter were computed as the ratio of the observed median χ^2 statistic over the expected median of the χ^2 distribution under the null hypothesis [209]. To account for multiple testing, P-values were adjusted with the *p.adjust()* function in R, applying the approach of Benjamini & Hochberg [210]. A genome-wide false discovery rate (FDR) of 10% was used to declare associations as significant.

5.3 Results

5.3.1 Homozygosity and ROH-coverage across the genome

Genome-wide SNP homozygosity of cows approximately followed a normal distribution with a mean of 64.4% and a standard deviation of 1.0% (Figure 5.1A). A total of 3,910,969 ROHs were identified. As expected, these ROHs approximately followed an exponential distribution in terms of length (Figure 5.1B), with short ROH being more abundant than long ROH. The frequency of a SNP being in a ROH was on

average 11.5% and this frequency differed across the genome (Figure 5.1C). Chromosomes 10, 16 and 20 had the highest ROH-frequency. The highest local peak was observed on chromosome 1, with a ROH-frequency of up to 63.3%. There were also 62 SNPs that were never in a ROH. These SNPs were mostly located at the start or ends of chromosomes.

The homozygosity status and ROH status partly overlapped. Of all SNPs across all individuals, 11.3% was both homozygous and in a ROH, 53.1% was homozygous but not in a ROH, 0.2% was heterozygous and in a ROH, and the remaining 35.4% was heterozygous and not in ROH.

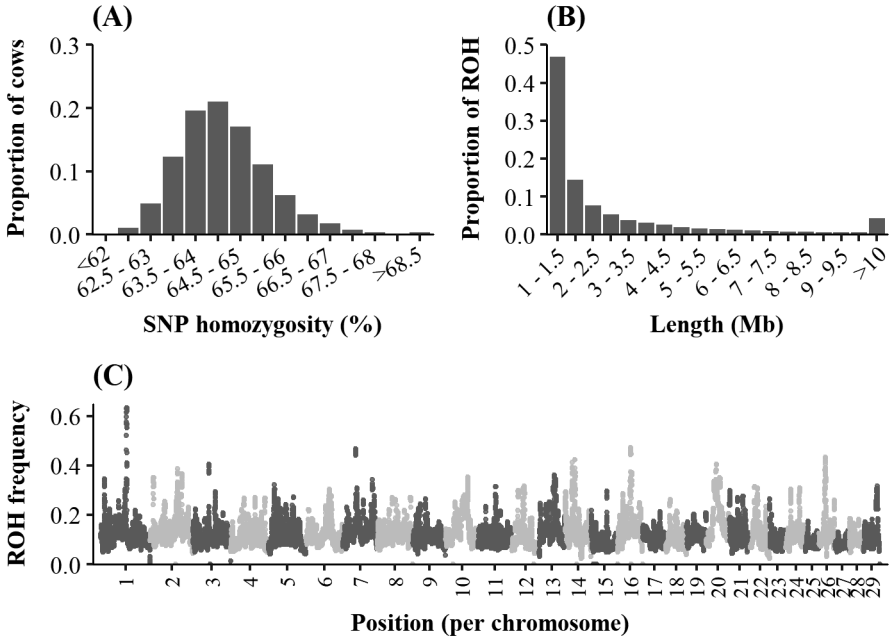


Figure 5.1 Summary statistics of SNP homozygosity and regions of homozygosity (ROH) across all cows. (A) distribution of genome-wide SNP homozygosity, (B) distribution of ROH length, (C) frequency of each SNP being in a ROH by genomic position.

5.3.2 Genome-wide inbreeding depression from GREML models

Genome-wide homozygosity had an unfavorable effect on all evaluated traits and across all GREML models (Table 5.1). For example, a 1% increase in homozygosity in model A was associated with a decrease in 305-d milk yield of 99.6 kg (SE = 5.2), an increase in calving interval of 1.1 days (SE = 0.4) and an increase in SCS400 of 2.3 units (SE = 0.7). These unfavorable effects of genome-wide homozygosity reflect the presence of (favorable) directional dominance. For example, the mean dominance effect of a SNP in model A was 0.13 kg for milk yield, -0.0015 days for calving interval

and -0.0030 units in SCS400 (Table 5.1). Estimated effects of genome-wide inbreeding depression were similar across the A, AD and ADR models.

Table 5.1 Effect of a 1% increase in genome-wide homozygosity (b), and mean dominance effect per SNP ($-b/n$; where n is the number of SNPs), for three models[1] and nine traits[2].

	Model A		Model AD		Model ADR	
Trait	b (SE)	$-b/n$	b (SE)	$-b/n$	b (SE)	$-b/n$
MY	-99.6 (5.2)	0.1322	-98.7 (6.1)	0.1310	-97.8 (6.7)	0.1298
FY	-4.10 (0.20)	0.0054	-4.04 (0.23)	0.0054	-4.01 (0.27)	0.0053
PY	-3.49 (0.17)	0.0046	-3.45 (0.20)	0.0046	-3.42 (0.23)	0.0045
CI	1.11 (0.35)	-0.0015	1.11 (0.35)	-0.0015	1.11 (0.38)	-0.0015
ICF	0.20 (0.15)	-0.0003	0.21 (0.16)	-0.0003	0.21 (0.17)	-0.0003
IFL	0.79 (0.30)	-0.0011	0.79 (0.30)	-0.0010	0.79 (0.30)	-0.0010
CR	-0.68 (0.19)	9.0E-06	-0.68 (0.19)	9.0E-06	-0.68 (0.19)	9.0E-06
SCS150	1.09 (0.69)	-0.0015	1.08 (0.70)	-0.0015	1.09 (0.71)	-0.0014
SCS400	2.28 (0.67)	-0.0030	2.26 (0.70)	-0.0030	2.26 (0.70)	-0.0030

[1]A: additive model; AD: additive + dominance model; ADR: additive + dominance + ROH model.
[2]MY: 305-day milk yield (kg); FY: 305-day fat yield (kg); PY: 305-day protein yield (kg); CI: calving interval (days); ICF: interval calving to first insemination (days); IFL: interval first to last insemination (days); CR: conception rate (%); SCS150 somatic cell score day 5 to 150 (units); SCS400: somatic cell score day 151 to 400 (units).

5.3.3 Variance components and goodness of fit of GREML models

Additive genetic variance was observed for all traits (Table 5.2). In model A, heritability estimates ranged from 2.36% (SE = 0.32%) for conception rate to 41.16% (SE = 0.81%) for milk yield. Heritability estimates were approximately identical across the A, AD and ADR models.

In model AD, 0.8 to 0.9% of phenotypic variance for yield traits and less than 0.4% of phenotypic variance for all other traits was attributable to dominance. When expressed as part of the total genetic variance, dominance variance explained on average 2.36% of genetic variance in the AD models (with a range of 0.07 to 5.24% across traits). The small contribution of dominance was also reflected by the goodness of fit of the different models. When moving from the A to the AD model, the maximum log likelihood increased significantly ($P < 0.05$) only for yield traits (Table 5.3). For these yield traits, the maximum log likelihood further increased ($P < 0.05$) when moving to the ADR model. In the ADR model, the relative ROH variance for yield traits was approximately 0.2% (Table 5.2), while the relative dominance variance was lower than that in the AD model (i.e. 0.5% instead of 0.8%).

The herd-year-season variance (data not shown) was similar across the A, AD and ADR models and was highest for yield traits (6.7% to 9.8% of total variance) and for the interval between calving and first insemination (5.8% of total variance). The latter trait is known to be strongly influenced by farmers' decision.

Table 5.2 Estimated variance components[1] for three GREML models[2] and nine traits[3], with standard errors in parentheses.

		Trait								
Model	Parameter	MY	FY	PY	CI	ICF	IFL	CR	SCS150	SCS400
A	σ_P^2	1356452	1813.04	1269.15	4215.19	717.658	3056.16	12.6991	17693.4	16336.5
	σ_A^2/σ_P^2 (%)	41.16 (0.81)	33.34 (0.82)	30.98 (0.81)	5.02 (0.43)	6.41 (0.49)	3.13 (0.35)	2.36 (0.32)	9.26 (0.54)	12.0 (0.60)
AD	σ_P^2	1356134	1812.49	1268.93	4215.18	717.690	3056.15	12.6991	17693.82	16336.80
	σ_A^2/σ_P^2 (%)	41.13 (0.81)	33.31 (0.82)	30.95 (0.81)	5.02 (0.43)	6.41 (0.49)	3.13 (0.35)	2.35 (0.32)	9.26 (0.54)	11.94 (0.60)
	σ_D^2/σ_P^2 (%)	0.77 (0.28)	0.90 (0.31)	0.87 (0.32)	0.00 (0.38)	0.35 (0.41)	0.09 (0.39)	0.04 (0.40)	0.16 (0.35)	0.34 (0.36)
	σ_A^2/σ_G^2 (%)	98.17 (0.66)	97.36 (0.89)	97.27 (0.99)	99.93 (7.63)	94.76 (5.78)	97.23 (11.89)	98.52 (16.45)	98.30 (3.66)	97.22 (2.87)
	σ_D^2/σ_G^2	1.83 (0.66)	2.64 (0.89)	2.73 (0.99)	0.07 (7.63)	5.24 (5.78)	2.77 (11.89)	1.48 (16.45)	1.70 (3.66)	2.78 (2.87)
ADR	σ_P^2	1355963	1812.48	1268.69	4215.08	717.624	3056.15	12.6991	17693.8	16336.8
	σ_A^2/σ_P^2 (%)	41.10 (0.81)	33.27 (0.82)	30.90 (0.81)	5.01 (0.43)	6.38 (0.49)	3.13 (0.35)	2.35 (0.32)	9.25 (0.54)	11.94 (0.60)
	σ_D^2/σ_P^2 (%)	0.51 (0.32)	0.54 (0.32)	0.52 (0.36)	0*	0.14 (0.47)	0.09 (0.39)	0.04 (0.40)	0.14 (0.40)	0.34 (0.36)
	$\sigma_{ROH}^2/\sigma_P^2$ (%)	0.17 (0.11)	0.25 (0.12)	0.24 (0.13)	0.07 (0.12)	0.13 (0.14)	0*	0*	0.01 (0.12)	0*
	σ_A^2/σ_G^2	98.38 (0.68)	97.68 (0.92)	97.62 (1.02)	98.53 (2.30)	95.94 (6.11)	97.23 (11.93)	98.52 (16.76)	98.39 (3.74)	97.22 (2.87)
	σ_D^2/σ_G^2	1.22 (0.75)	1.60 (1.01)	1.63 (1.12)	0*	2.09 (6.91)	2.77 (11.89)	1.48 (16.45)	1.46 (4.20)	2.78 (2.87)
	$\sigma_{ROH}^2/\sigma_G^2$	0.40 (0.26)	0.72 (0.37)	0.75 (0.40)	1.47 (2.25)	1.97 (2.21)	0*	0*	0.15 (1.29)	0*

*The corresponding variance component was fixed to 0, because its initial estimate was slightly negative.

[1]σ_P^2: phenotypic variance (excluding the herd-year-season variance); σ_A^2: additive genetic variance; σ_D^2: dominance variance; σ_{ROH}^2: ROH variance; σ_G^2: genetic variance ($\sigma_A^2 + \sigma_D^2$ for model AD, and $\sigma_A^2 + \sigma_D^2 + \sigma_{ROH}^2$ for model ADR).

[2]A: additive model; AD: additive + dominance model; ADR: additive + dominance + ROH model.

[3]MY: 305-day milk yield (kg); FY: 305-day fat yield (kg); PY: 305-day protein yield (kg); CI: calving interval (days); ICF: interval calving to first insemination (days); IFL: interval first to last insemination (days); CR: conception rate (%); SCS150 somatic cell score day 5 to 150 (units]); SCS400: somatic cell score day 151 to 400 (units).

Table 5.3 Comparison of goodness of fit of different GREML models[1] for nine traits[2].

	Difference in maximum log-likelihood		P-value	
Trait	AD - A	ADR - AD	AD vs A	ADR vs AD
MY	4.172	1.437	0.002	0.045
FY	5.014	2.355	0.001	0.015
PY	4.141	2.120	0.002	0.020
CI	0.000	0.236	0.500	0.246
ICF	0.382	0.436	0.191	0.175
IFL	0.026	0.000	0.410	0.500
CR	0.004	0.000	0.465	0.500
SCS150	0.107	0.007	0.322	0.453
SCS400	0.464	0.000	0.168	0.500

[1]A: additive model; AD: additive + dominance model; ADR: additive + dominance + ROH model.
[2]MY: 305-day milk yield; FY: 305-day fat yield; PY: 305-day protein yield; CI: calving interval; ICF: interval calving to first insemination; IFL: interval first to last insemination; CR: conception rate; SCS150 somatic cell score day 5 to 150; SCS400: somatic cell score day 151 to 400.

5.3.4 Comparison of GREML and GWAS effects

Estimated additive, dominance and ROH effects from backsolving in GREML and from single SNP GWAS models were approximately normally distributed with a mean of zero (Figure 5.2). The range and standard deviation of GWAS effects were substantially larger than those of GREML effects. For example, additive effects for milk yield estimated by GWAS ranged from -1069 kg to 1020 kg with a standard deviation of 36.6 kg, whereas those estimated by GREML ranged from -25.7 kg to 17.8 kg with a standard deviation of 1.1 kg. The difference between GWAS and GREML estimates was larger for dominance and ROH effects than for additive effects. For example, dominance effects for milk yield estimated by GWAS ranged from -1038 kg to 1120 kg with a standard deviation of 34.1 kg, whereas those estimated by GREML ranged from -0.25 kg to 0.25 kg with a standard deviation of 0.05 kg.

There was a moderate correlation between SNP effects estimated by GREML and GWAS. For milk yield, for example, this correlation was 0.50 for additive effects, 0.40 for dominance effects and 0.79 for ROH effects. For additive and dominance effects, many SNPs had large (absolute) GWAS effects but a GREML effect close to zero (Figure 5.3).

Estimated SNP effects were similar across the three GREML models. Correlations between additive effects estimated with the A, AD and ADR models, and between dominance effects of the AD and ADR models, were all above 0.998.

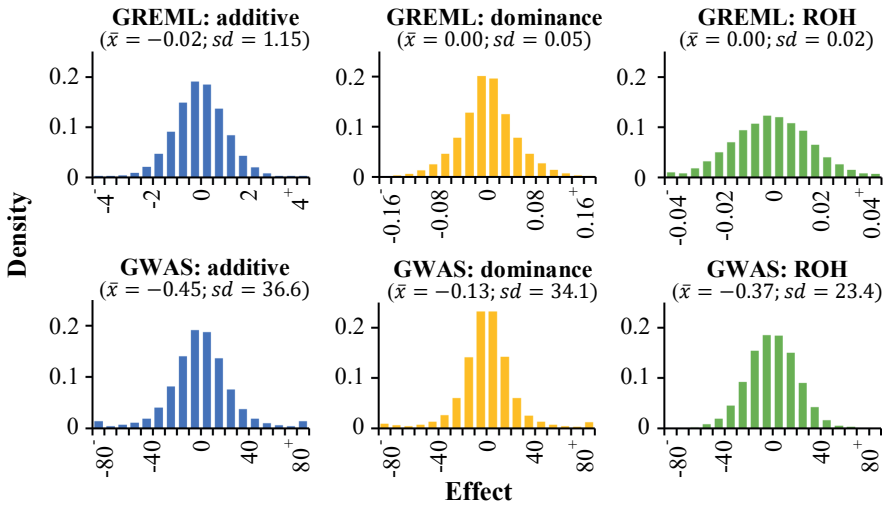


Figure 5.2 Distributions of SNP effects for 305-day milk yield (kg), estimated by GREML and single SNP GWAS. The mean ($\bar{x}$) and standard deviation (sd) of the effects are shown. Note that distributions were truncated such that the first and last bar represent "smaller than" and "bigger than" classes (i.e. the range was larger than shown here). Also note that the dominance effects shown here do not include the mean dominance effect that was absorbed by the fixed regression on genome-wide homozygosity.

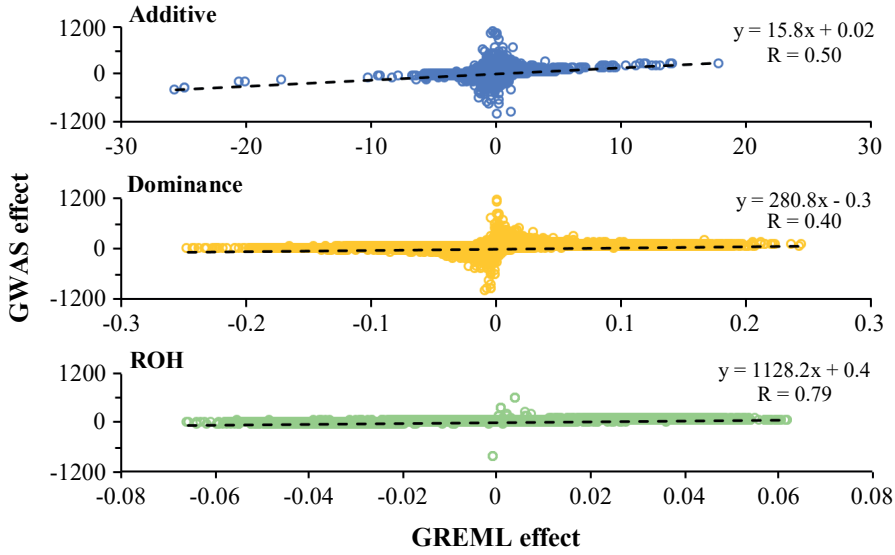


Figure 5.3 Scatterplot comparing SNP effects for 305-day milk yield (kg) estimated by GREML and single SNP GWAS. The dashed line is a linear trendline. The regression equation corresponding to this line and the Pearson correlation coefficient (R) are shown. Note that the dominance effects shown here do not include the mean dominance effect that was absorbed by the fixed regression on genome-wide homozygosity.

5.3.5 Additive effects across the genome

Estimated additive genetic effects followed the expectation from other GWAS studies. Manhattan plots of SNP effects for yield traits, obtained by backsolving from the GREML ADR models, are shown in Figure 5.4. As expected, SNPs with the largest additive effects for yield traits were located between 0 and 1 Mb on chromosome 14, surrounding the *DGAT1* gene [211]. The effects in this region were antagonistic, i.e. alleles that were favorable for milk and protein yields were unfavorable for fat yield. Two other regions had major additive effects on all yield traits, one on chromosome 5 (about 88.2 to 88.5 Mb), surrounding the *ABCC9* gene [211], and one on chromosome 6 (near 87 Mb), surrounding the *GC* gene [211]. For protein yield, there was an additional peak on chromosome 6 (about 85.4 to 85.7 Mb), which included the casein cluster, i.e. *CSN1*, *CSN2* and *CSN3* [211]. The abovementioned peaks also passed the 10% FDR threshold in the GWAS (Figure 5.5). Genomic inflation factors of the GWAS for additive and dominance effects were all <1.1, suggesting that there was no major inflation of P-values for these effects (Figure S5.1).

For fertility and udder health traits, there were less pronounced peaks of GREML additive effects than for yield traits (Figure S5.2). For fertility and udder traits, there were also fewer SNPs with a significant additive association in the GWAS (Figure S5.3). The most notable region with significant associations was a region on chromosome 19 for SCS400. This region consisted of two narrow subpeaks (one around 54.6 Mb and one around 55.3 Mb). For the interval between calving and first insemination (ICF), there were various significant additive associations in the GWAS. The most notable region, which was also identified by the GREML, was on chromosome 28 (near 35.8 Mb).

5.3.6 Dominance and ROH effects across the genome

GREML-based dominance effects showed less pronounced peaks than additive effects (Figures 5.4 and S5.2). In the GWAS, there were also fewer significant dominance associations than significant additive associations (Figures 5.5 and S5.3).

For yield traits, the most notable region with large favorable dominance effects in the GREML and with significant dominance associations in the GWAS was located on chromosome 5 (Figures 5.4 and 5.5). This region was rather wide, with significant associations between 13 and 40 Mb and the largest effects between 24 and 28 Mb (Figure S5.4). In addition to the region on chromosome 5, there were two other peaks that passed the 10% FDR in the GWAS, one near *DGAT1* for milk and fat yields and one on chromosome 23 (near 25.2 Mb) for milk yield.

For fertility and udder health traits, there were very few significant dominance associations in the GWAS (Figure S5.3). The only significant SNPs were found for the interval between calving and first insemination (ICF) and these SNPs did not cluster.

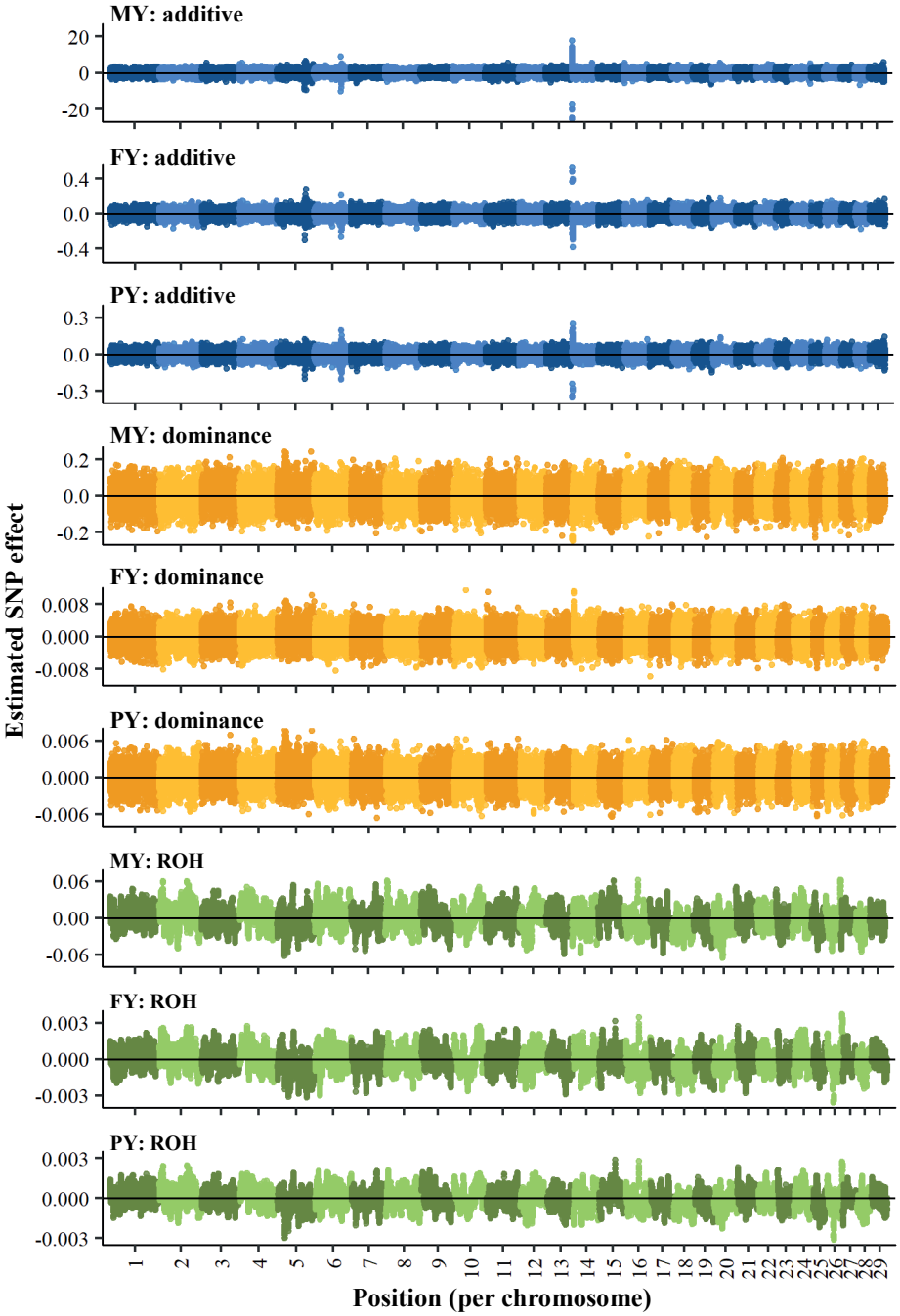


Figure 5.4 Additive, dominance and ROH effects for yield traits, estimated by GREML (model ADR) with backsolving. MY: 305-day milk yield (kg); FY: 305-day fat yield (kg); PY: 305-day protein yield (kg). Note that the dominance effects shown here do not include the mean dominance effect that was absorbed by the fixed regression on genome-wide homozygosity.

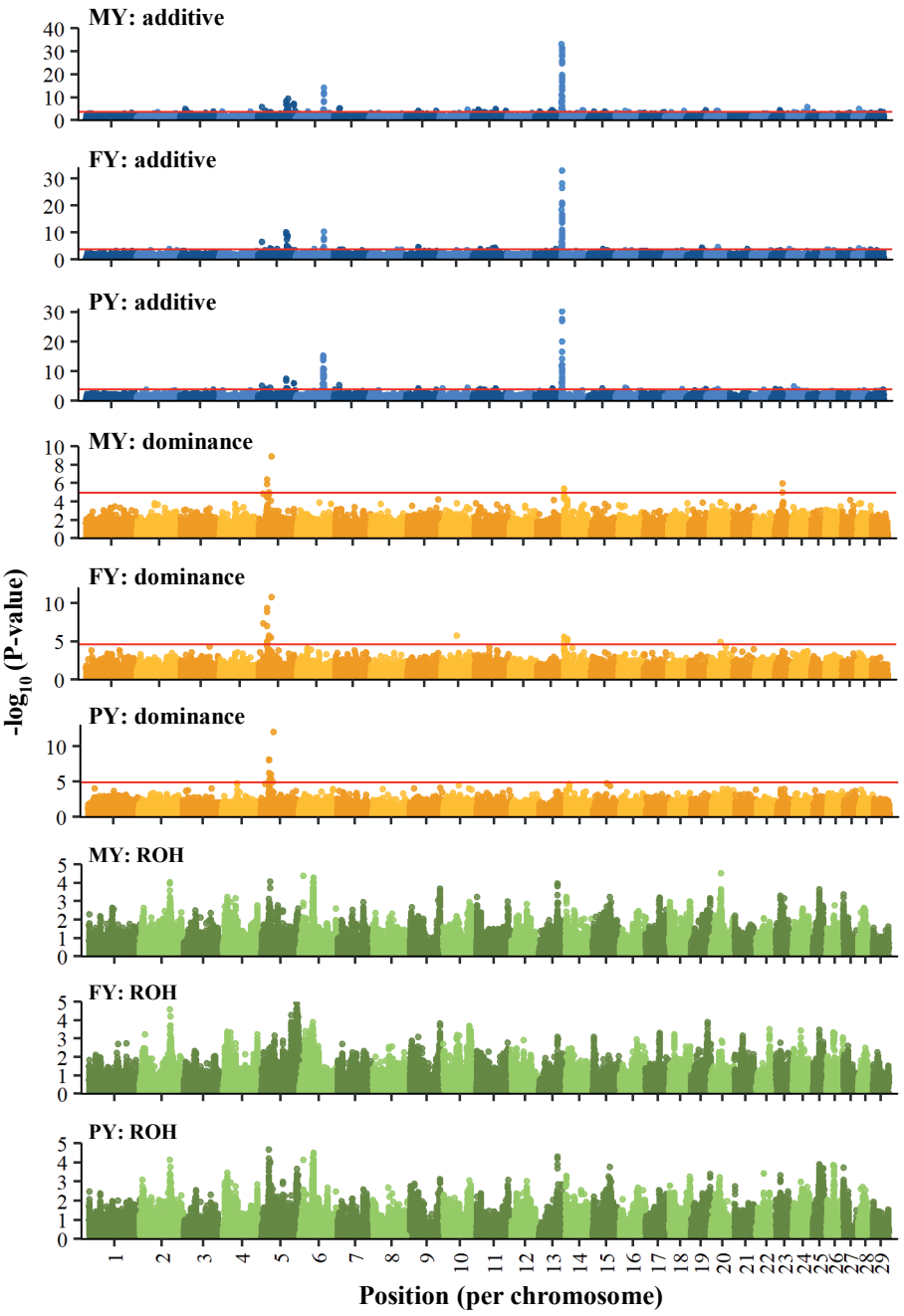


Figure 5.5 Significance of additive, dominance and ROH effects for yield traits from single SNP GWAS. MY: 305-day milk yield; FY: 305-day fat yield; PY: 305-day protein yield. The horizontal red line is a threshold based on a 10% false-discovery rate (absence of this line implies that all effects were below the threshold). The y-axis for MY additive effects was truncated at 40; in the peak on chromosome 14, there were 6 SNPs with a -$\log_{10}$(P-value) ranging from 40 to 94.

ROH effects showed many narrow peaks for all traits, both with GREML and GWAS (Figures 5.4, 5.5, S5.2 and S5.3). None of the ROH effects in the GWAS, however, passed the 10% FDR.

5.4 Discussion

The objective of this study was to obtain a better understanding of (differences in) inbreeding depression across the genome of Dutch Holstein Friesian dairy cattle. To fulfil this objective, we first estimated genome-wide inbreeding depression and estimated additive, dominance and ROH variance components with GREML models. We then investigated dominance and ROH effects across the genome for yield, fertility and udder health traits, using GREML (with backsolving) and a single SNP GWAS.

5.4.1 Genome-wide inbreeding depression

Genome-wide SNP homozygosity had an unfavorable effect on all traits, indicating the presence of directional dominance (Table 5.1). The estimated effects were comparable in size to those previously reported for similar traits in Holstein Friesian dairy cattle [80, 152, 193]. When comparing inbreeding depression estimates across studies, it is important to consider the variance of underlying inbreeding measures [212]. For example, the effect of a 1% increase in SNP homozygosity on milk yield in this study (of approximately -99 kg) may seem larger than the effect of a 1% increase in ROH-based inbreeding that we previously reported (of approximately -36kg [193]), but this difference can be largely explained by the different scale of the inbreeding measures. SNP homozygosity in this study had a mean of 64.4% and a standard deviation of 1.0%, whereas ROH-based inbreeding in our previous study had a mean of 12.3% and a standard deviation of 2.7% [193]. Therefore, a 1% increase in SNP homozygosity captures a larger effect at population level than a 1% increase in ROH-based inbreeding. To illustrate this effect of scale, we previously compared the phenotypes of highly inbred cows and lowly inbred cows, showing that different inbreeding measures may result in similar inbreeding depression at population level, despite the difference in estimated regression coefficients [193].

5.4.2 Dominance and ROH variance

Estimates of dominance and ROH variances were small and only differed significantly from zero for yield traits (Tables 5.2 and 5.3). Dominance and ROH variances explained less than 1% of phenotypic variance, and less than 5% of genetic variance.

Many other studies have estimated dominance variance in Holstein Friesian dairy cattle. In literature, the term “dominance heritability” is sometimes used for the ratio of dominance variance over phenotypic variance [200, 213, 214]. This term is

misleading, because dominance effects are, by definition, not heritable. Therefore, we used "relative dominance variance" instead. Estimates of relative dominance variance based on pedigree relationship matrices in Holstein cattle typically range from 1% to 5% [158, 170, 215, 216], although a few studies suggest a larger contribution of dominance effects [217, 218]. Estimates based on genomic relationship matrices are similar to, or slightly higher, than pedigree-based estimates (Table 5.4).

Table 5.4 Estimates of relative dominance variance from various studies that used genomic relationship matrices.

Study	Density	Accounted for GW ID[1]	Relative dominance variance
Aliloo et al. [200]	632 k (imputed)	Yes	≤ 1% for yield traits, 1% for calving interval
Aliloo et al. [200]	632 k (imputed)	No	3 to 4% for yield traits, 1% for calving interval
Sun et al. [213]	50 k (imputed)	Yes, in precorrection of phenotypes	3 to 4% for yield traits, 1% for SCS, 0% for daughter pregnancy rate
Jiang et al. [219][2]	50 k (imputed)	No	7 to 13% for yield traits, 0 to 15% for fertility traits, 9% for SCS
Alves et al. [220]	41k (imputed)	No	0 to 4% for fertility traits
Mao et al. [221]	36k	No	7% for interval first-last insemination, 4% for number of inseminations

[1]GW ID: genome-wide inbreeding depression

[2]In this particular study, an imprinting effect was also fitted. All other studies used AD models.

Our relative dominance variance estimates tend to be a bit lower than most estimates from literature. One explanation is that we corrected for genome-wide inbreeding depression in our GREML models (as discussed in the next section). Another reason for low dominance variance may be the limited SNP density we used, although most other studies have used similar densities (Table 5.4). It is well known that the additive variance captured by a SNP depends on r^2 (with r being the correlation between the SNP and a QTL), while the dominance variance captured by the SNP depends on r^4 [222]. In other words, detection of dominance effects relies more on high LD than detection of additive effects. Detection of dominance effects would, thus, benefit substantially from a higher SNP density. In addition to SNP density and the inclusion of a regression on genome-wide homozygosity, there are many other factors that may explain differences in relative dominance variance across studies. These factors include differences in trait definition, differences in the

way phenotypes are (pre)corrected for fixed effects, differences in how the dominance relationship matrix is calculated and population-specific differences [199, 223].

In this study, we introduced a ROH-based relationship matrix to estimate a ROH-based variance component. For yield traits, the ADR model showed a better fit than the AD model (Table 5.3), suggesting some benefit of including ROH effects. This benefit, however, could not be easily explained by the change in variance components. In fact, the proportion of variance explained by dominance in the AD model was higher than the combined proportion of variance explained by dominance and ROH effects in the ADR model (Table 5.2). However, the error variance in the ADR model was lower, resulting in the better model fit. When moving from the AD to the ADR model, the dominance variance from the AD model appeared to be split over dominance and ROH components, suggesting that dominance and ROH effects partly captured the same variation. This is not surprising, since both capture the effects of homozygosity (either based on single SNPs or regions).

5.4.3 Accounting for directional dominance when estimating dominance variance with GREML

In our GREML models, we corrected for genome-wide inbreeding depression by including a fixed regression on genome-wide SNP homozygosity. This correction is important to ensure that the model assumptions of a mean dominance effect of zero ($\mathrm{E}[\mathrm{v}] = 0$) and of no covariance between additive and dominance effects ($\mathrm{cov}[\mathrm{u}, \mathrm{v}] = 0$) hold, and to prevent the dominance variance from being inflated [196, 223]. Indeed, when we removed the fixed regression on genome-wide homozygosity from the AD model, the relative dominance variance for yield traits increased to approximately 3% (as compared to 0.8%), which are values similar to those reported by Aliloo et al. [223]. In addition, the mean backsolved dominance effect was no longer zero, but slightly favorable. The mean backsolved dominance effect for milk yield, for example, was 0.05 kg when not accounting for genome-wide homozygosity (as compared to 0.0005 kg when accounting for genome-wide homozygosity). Note that this 0.05 kg is smaller than the 0.13 kg ROH-from the fixed regression on genome-wide homozygosity (Table 5.1), which may be explained by shrinkage on the mean dominance when it was part of the random effect.

When a fixed regression on genome-wide SNP homozygosity is included in an AD model, the mean dominance effect across all loci is absorbed by this regression [205, 206]. Consequently, the σ_D^2 of such models is expected to be underestimated. Namely, the σ_D^2 of such models captures only the deviations of dominance effects (d_i) at individual loci from the mean dominance effect ($\bar{d}$) across all loci, $\sigma_D^2 = \sum_i (2p_i q_i (d_i - \bar{d}))^2$, while the full dominance variance equals $\sigma_{D_{FULL}}^2 =$

$\sum_i (2p_i q_i d_i)^2$. Thus, to obtain $\sigma^2_{D_{FULL}}$, a component related to the mean dominance effect across all loci should be added to the σ^2_D from GREML output. This additional component can be derived as:

$$\sigma^2_{D_{FULL}} = \sum_i (2p_i q_i d_i)^2$$
$$= \sum_i \left(2p_i q_i (d_i - \bar{d}) + 2p_i q_i \bar{d}\right)^2$$
$$= \sum_i (2p_i q_i (d_i - \bar{d}))^2 + \sum_i (2p_i q_i \bar{d})^2$$
$$= \sigma^2_D + n\overline{(2pq)^2}\bar{d}^2$$

where $\sigma^2_{D_{FULL}}$ is the full dominance variance, σ^2_D is the dominance variance obtained from the GREML output, n is the number of SNPs, $\overline{(2pq)^2}$ is the mean squared expected heterozygosity, and $\bar{d}^2$ is the squared mean dominance effect, where $\bar{d}$ can be obtained from the regression on genome-wide homozygosity. Note that, in the third line of the derivation above, a cross product has disappeared because $\sum(d_i - \bar{d}) = 0$.

To quantify the difference between σ^2_D and $\sigma^2_{D_{FULL}}$, we calculated $\sigma^2_{D_{FULL}}$ for the AD model, applying the reasoning above. The additional component, $n\overline{(2pq)^2}\bar{d}^2$, was found to be relatively small compared to σ^2_D. For milk yield, for example, $n\overline{(2pq)^2}\bar{d}^2$ equalled 189 kg^2, while σ^2_D equalled 10377 kg^2. Consequently, the relative dominance variance increased only marginally when accounting for the additional $n\overline{(2pq)^2}\bar{d}^2$ component (e.g. from 0.77% to 0.78% for milk yield).

In the ADR model, it was assumed that ROH effects were distributed as ~ N(0, $\mathbf{R}\sigma^2_{roh}$). This assumption may not hold, because of a potential average genome-wide ROH-effect being different from zero (similar to the genome-wide dominance effect). However, since genome-wide SNP homozygosity and genome-wide ROH coverage (the F_{ROH}) are highly correlated [193], we expected that the inclusion of genome-wide homozygosity would largely correct for a genome-wide ROH effect. In the ADR model, the means of the backsolved ROH effects were approximately zero (e.g. -0.0007 for milk yield), suggesting that the fixed effect for genome wide homozygosity indeed removed the mean ROH effect.

5.4.4 SNP effects estimated by GREML and GWAS

In this study, we estimated SNP effects with two approaches: by GREML with backsolving and a single SNP GWAS. Effect sizes were found to be much larger for the single SNP GWAS than for GREML and correlations between the effects of the two approaches were moderate (Figures 5.2 and 5.3).

The traditional single SNP GWAS, in which fixed effects are estimated for one SNP at a time, has some clear limitations. One of them is the large number of tests. To account for multiple testing, a stringent P-value threshold is typically used (in this study we used a 10% FDR), which may lead to many false negatives and overestimated effect sizes for significant associations [224]. Another limitation is that the effect of a QTL may be only partly captured by a single SNP due to imperfect LD. Since the effect of the QTL might be better captured by multiple SNPs surrounding the QTL, models in which all SNPs are fitted simultaneously, such as GREML, are increasingly used for association analyses [225-227].

In GREML, all SNPs are fitted simultaneously as a combined random polygenic effect. The polygenic effects and underlying SNP effects are shrunk towards zero. The magnitude of shrinkage depends on the standard error of the estimate of the effect, which in turn depends on the amount of data and the variance of the associated factor. There is more shrinkage for a factor with lower variance [228]. This partly explains why the effects of GREML are much smaller than those of the single SNP GWAS (Figures 5.2 and 5.3), especially for dominance (which explained a small amount of the variance) and ROH effects (which explained even less variance). Shrinkage can also explain why there were various SNPs with a large absolute additive and dominance effect for GWAS, but with a close to zero additive and dominance effect for GREML (Figure 5.3). When we further investigated these 'outliers', they were found to have a low minor allele frequency (MAF). As indicated by Gianola [228], shrinkage in GREML is independent on effect size but dependent on sample size and allele frequencies. For additive effects, the degree of shrinkage at a SNP (given a fixed sample size) depends on $2pq$ [228]. When we manually shrunk GWAS additive effects by multiplying them with $2pq$, the outliers indeed disappeared and the correlation between additive effects of GWAS and GREML increased from 0.50 to 0.86 (Figure 5.6).

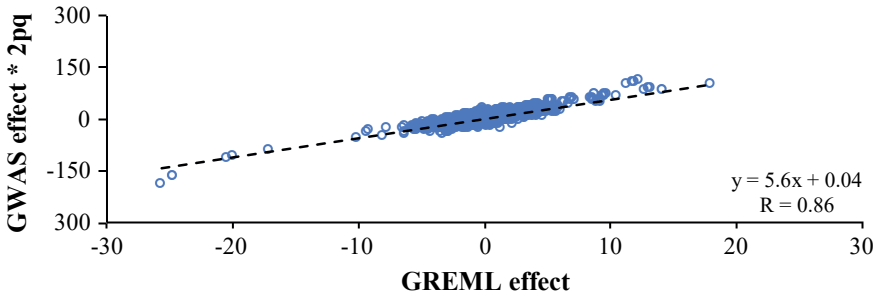


Figure 5.6 Scatterplot comparing additive effects for 305-day milk yield (kg) estimated by GREML and by GWAS with manual shrinkage. The GWAS effects were manually shrunk by multiplying them with 2pq. The dashed line is a linear trendline. The regression equation corresponding to this line and the Pearson correlation coefficient (R) are shown.

5.4.5 Region-specific inbreeding depression

We found limited evidence for region-specific inbreeding depression based on 75k SNP data for yield, fertility and udder health traits. For yield traits, we only found a few regions with large favorable dominance effects from GREML and significant dominance associations in the GWAS (Figures 5.4 and 5.5). These regions were similar to those identified in previous studies. For example, in a recent large scale GWAS with approximately 300k US Holsteins and 60k SNP data, Jiang et al. [197] also found the most significant dominance effects for yield traits on chromosome 5 between 24 and 28 Mb. The second most significant dominance peak for milk yield they found was located on chromosome 23 (near 18 Mb). The latter peak was also observed in an earlier GWAS with 43k Holsteins [219]. The peak that we observed on chromosome 23 (near 25 Mb) was not exactly at the same location, but close to the previously reported peak. For non-yield traits, we found very few significant dominance associations for fertility traits and no significant dominance associations for SCS, also similar to findings of Jiang et al. [197].

No significant GWAS associations for ROH effects were found. Pryce et al. [80], in contrast, reported various candidate regions associated with inbreeding depression for yield traits and calving interval based on a single SNP GWAS for ROH status in US Holsteins. This may be explained by two differences in the approach used. First, we estimated the additive, dominance and ROH effects simultaneously, whereas Pryce et al. [80] had no dominance effect in the model. Second, Pryce et al. [80] used a threshold of $-\log_{10}$(P-value) of 3 to identify candidate regions and mentioned that the FDR was high. When we applied a threshold $-\log_{10}$(P-value) of 3 in the GWAS, we indeed found various significant ROH peaks for all traits (Figures 5.5 and S5.3). Some of these peaks were favorable, potentially indicating selection signatures, whereas others were unfavorable, potentially indicating inbreeding depression. Also note that genomic inflation factors were approximately 1.3 for ROH effects (Supplementary file 2), suggesting substantial inflation. Because the ROH effects did not pass the 10% FDR, we decided not to correct for this inflation.

In this and other studies in which a single SNP GWAS for ROH-status was performed [73, 80], the ROH-status of a SNP was set to 1 when the SNP was in a ROH, irrespective of which ROH it concerned. As a result, the estimated ROH-effect of a SNP was a pooled effect of many distinct ROHs. This approach is in line with the reasoning behind inbreeding depression, namely that any homozygosity decreases performance (irrespective for which allele). It may, however, be of interest to know which specific ROH is unfavorable. Fine-mapping of individual ROH effects is not straightforward due to the large number of distinct ROHs (each of which has a low frequency). One possibility is to group ROHs based on common core regions and then try to associate ROH-groups with the phenotype, as is sometimes done in humans

(e.g. [229]). Alternatively, one could test each individual ROH, e.g. using the heuristic approach introduced by Howard et al. [230]. In the approach of Howard et al. [230], the mean phenotype of individuals with a specific ROH is compared to the mean phenotype of individuals without that ROH. A limitation of this approach is that it is computationally intensive, despite the various filtering criteria that can be used (such as minimum ROH-frequency). Consequently, the feasibility for data sets with many traits and large numbers of individuals is limited. The approach has recently been applied to smaller data sets (of < 10,000 individuals) of swine [230] and of Canadian Holstein cattle [81], and the estimated effects reported by these studies are rather large. For example, the average effect of unfavorable ROHs identified by Marras et al. [81] for 305-d milk yield in first parity cows was -295 kg with a standard deviation of 105 kg. These effects are likely to be overestimated, because of statistical biases similar to those in a single SNP GWAS (e.g. [224]). Also, there may be many false negatives due to initial filtering steps and the use of significance thresholds to account for multiple testing. Despite these limitations, the identified unfavorable ROH-haplotypes and their effects could be used to build an inbreeding load matrix (ILM), which provides some information on the expected inbreeding load of the offspring of a particular mating [230]. This could be valuable in mating programs but is of less importance for selection schemes.

An important observation in this study was that dominance (and ROH) effects shrunk substantially when the fixed regression on homozygosity was included in GREML models. This was also observed by Aliloo et al. [223]. As a result, ROH and dominance variances were small (< 1% of phenotypic variance) and only significant for yield traits. This suggests that, after correcting for genome-wide inbreeding depression, fitting dominance and ROH effects (based on 75k data) had little additional value. Overall, our findings suggest that the deleterious effect of homozygosity is quite evenly distributed across the genome and is well captured by genome-wide homozygosity in Holstein-Friesian dairy cattle. Based on these findings, there is little benefit of accounting for region-specific inbreeding depression in selection schemes.

5.5 Conclusion

Inbreeding depression was observed for yield, fertility and udder health traits in Dutch Holstein Friesian cattle. After correcting for genome-wide homozygosity, however, dominance and ROH effects explained very little variance in GREML models. A few regions with relatively large favorable dominance effects and significant dominance associations (based on 10% FDR) were identified for yield traits, based on both GREML and single SNP GWAS. Overall, however, inbreeding

depression appeared to be distributed quite equally across the genome and was well captured by genome-wide homozygosity.

5.6 Acknowledgements

The research leading to these results has been conducted as part of the IMAGE project, which received funding from the European Union's Horizon 2020 Research and Innovation Programme under the grant agreement n° 677353. The study was co-funded by the Dutch Ministry of Agriculture, Nature and Food Quality (KB-34-013-002). The authors gratefully acknowledge the Dutch-Flemish cattle improvement cooperative (CRV) for providing pedigree and genotype data. The authors would also like to thank Mario Calus (Animal Breeding and Genomics, Wageningen University & Research) for his valuable input.

5.7 Supplementary information

Box S5.1 Computation of ROH-based relationship matrix

The ROH-based relationship matrix (**R**) was built following the same reasoning as for computing the genomic relationship matrix (**G**) of VanRaden method 1 [54]. The ROH-based relationship (R_{jk}) between animals j and k was defined as:

$$R_{jk} = \frac{\sum_i (x_{ij} - p_i) * (x_{ik} - p_i)}{\sum_i p_i q_i}$$

where x_{ij} is the ROH-status (coded as 0/1) of animal j at the i^{th} SNP, x_{ik} is the ROH-status of animal k at the i^{th} SNP, p_i is the frequency of SNP i being in a ROH, and q_i is the frequency of SNP i not being in a ROH. In this notation, the numerator represents the covariance between the ROH-status of individuals, which is $(x_j - E[x]) * (x_k - E[x])$ for a single SNP, and the denominator represents the variance, which is $p(1-p)$ for a binary variable.

To compute **R**, we first converted the ROH-status from 0/1 values to 0/2 values and then applied VanRaden's formula for a genomic relationship matrix [25] on the 0/2 values with *calc_grm* [24]. According to VanRaden [25], the genomic relationship (G_{jk}) between animals j and k is:

$$G_{jk} = \frac{\sum_i (y_{ij} - 2p_i) * (y_{ik} - 2p_i)}{\sum_i 2p_i q_i}$$

where y and p are SNP allele counts and frequencies, respectively. In our case, the y-counts actually equalled $2x$ (where x was the 0/1 ROH). For a single SNP, replacing y by $2x$ in the formula results in a numerator of:

$$(2x_j - 2p) * (2x_k - 2p) = 4x_j x_k - 4px_j - 4px_k - 4p^2$$

instead of the intended numerator of:

$$(x_j - p) * (x_k - p) = x_j x_k - px_j - px_k - p^2.$$

Thus, the numerator was a factor 4 too big. At the same time, the denominator was a factor 2 too big, namely $2pq$ instead of pq. Consequently, the estimated ROH relationships were a factor 2 too big. To account for the difference in scale, we divided the obtained relationships by a factor 2. As expected, the average diagonal of the **R** after this correction was equal to 1.

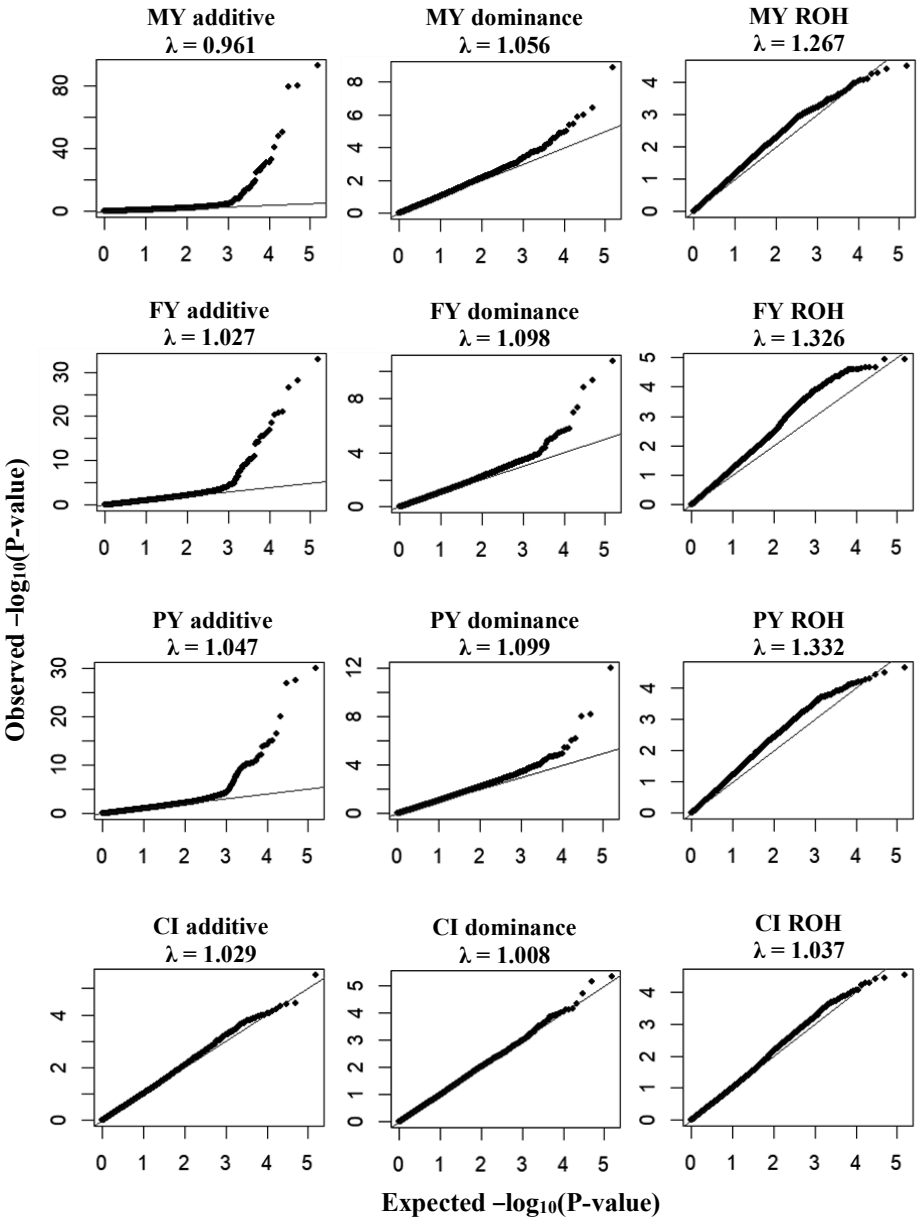


Figure S5.1 QQ-plots and genomic inflation factors (λ) for P-values corresponding to additive, dominance and ROH effects estimated by a single SNP GWAS for nine different traits (*continued on next page*). MY: 305-day milk yield; FY: 305-day fat yield; PY: 305-day protein yield; CI: calving interval; ICF: interval calving to first insemination; IFL: interval first to last insemination; CR: conception rate; SCS150 somatic cell score day 5 to 150; SCS400: somatic cell score day 151 to 400.

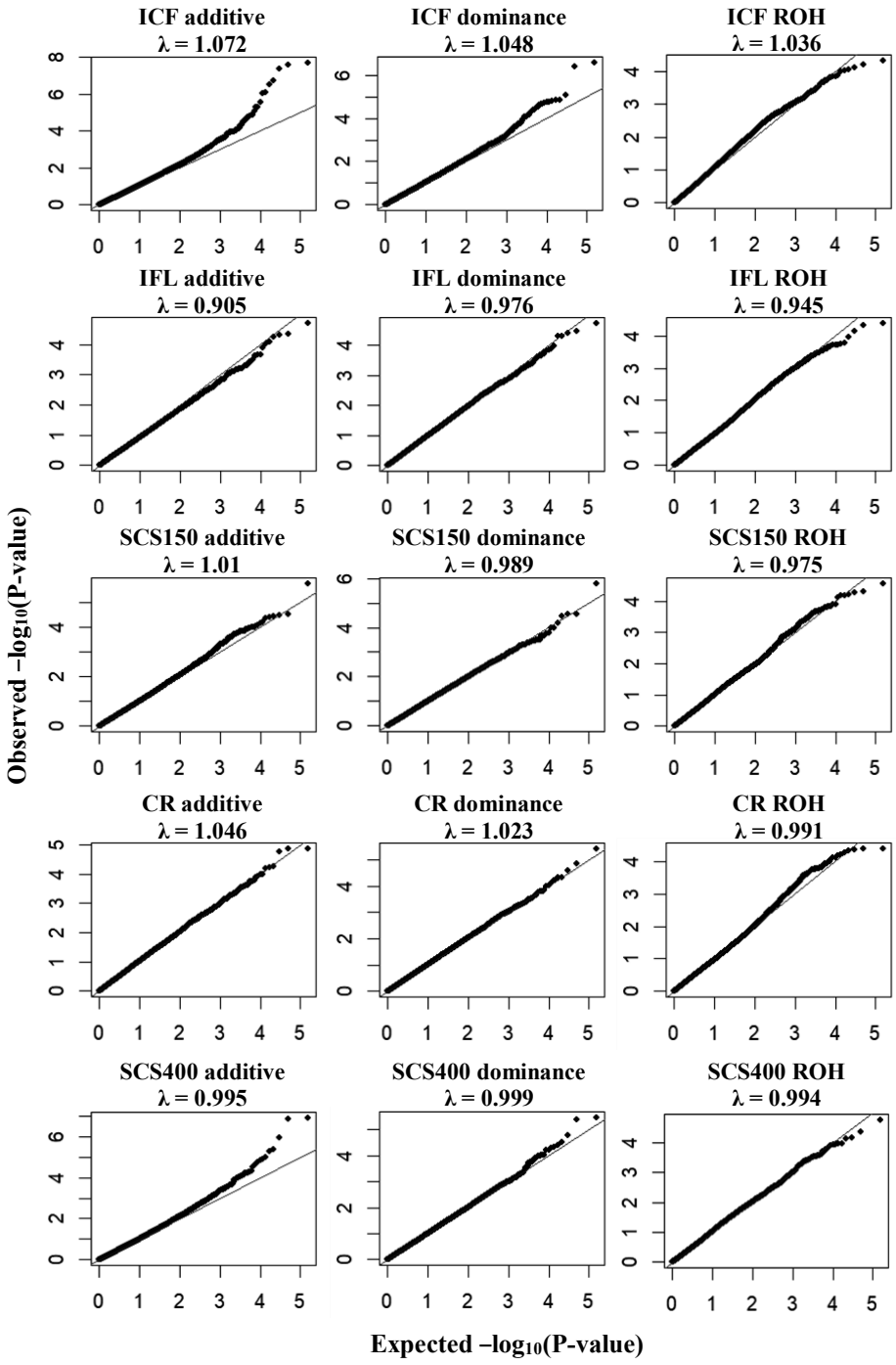


Figure S5.1 (continued)

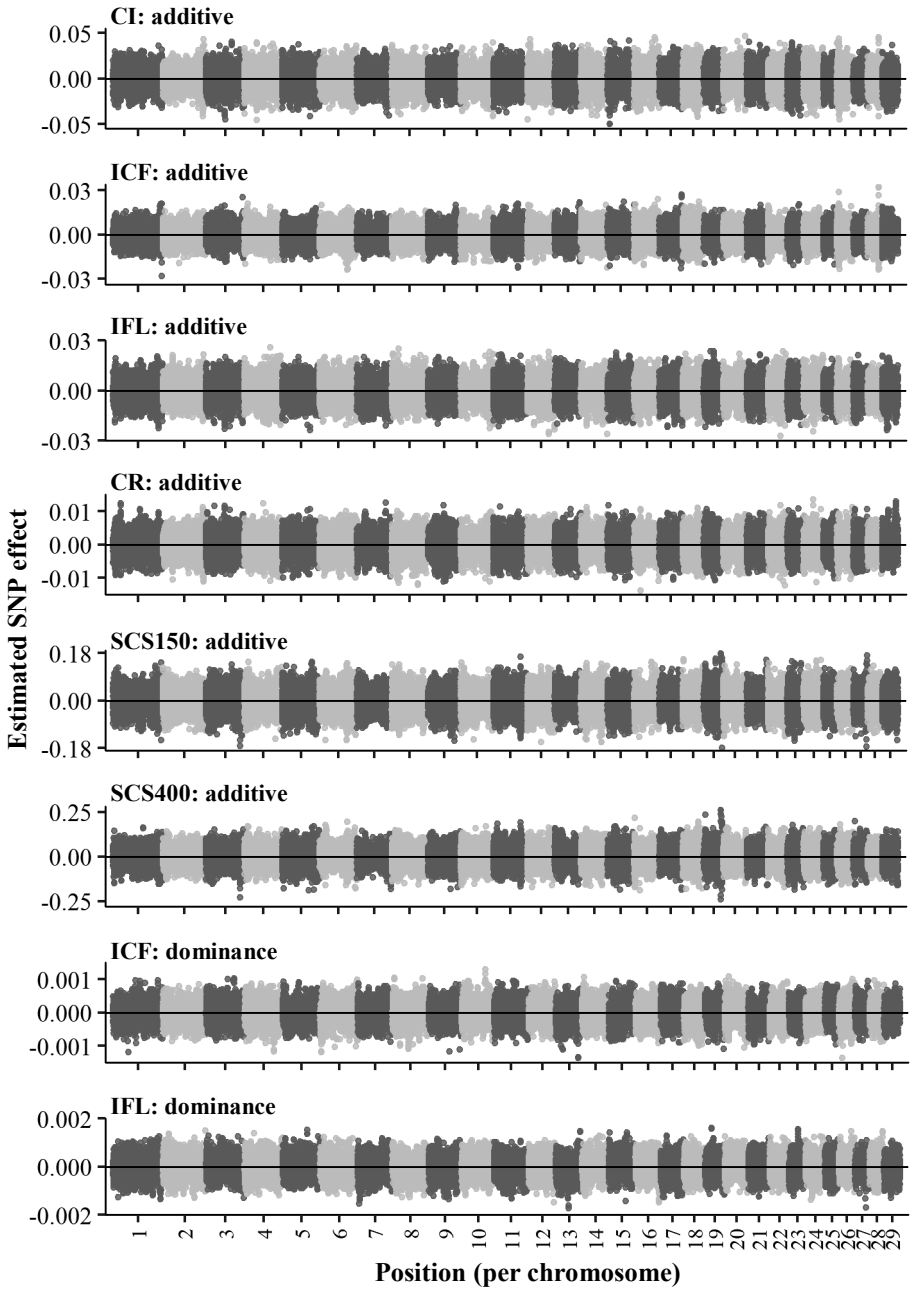


Figure S5.2 Additive, dominance and ROH effects for fertility and udder health traits, estimated by GREML model ADR with backsolving (*continued on next page*). CI: calving interval (d); ICF: interval calving to first insemination (d); IFL: interval first to last insemination (d); CR: conception rate (%); SCS150 somatic cell score day 5 to 150 (units); SCS400: somatic cell score day 151 to 400 (units). Note that dominance effects for CI and ROH effects for IFL, CR, and SCS400 are not shown, because the corresponding variances were fixed to zero (Table 5.2).

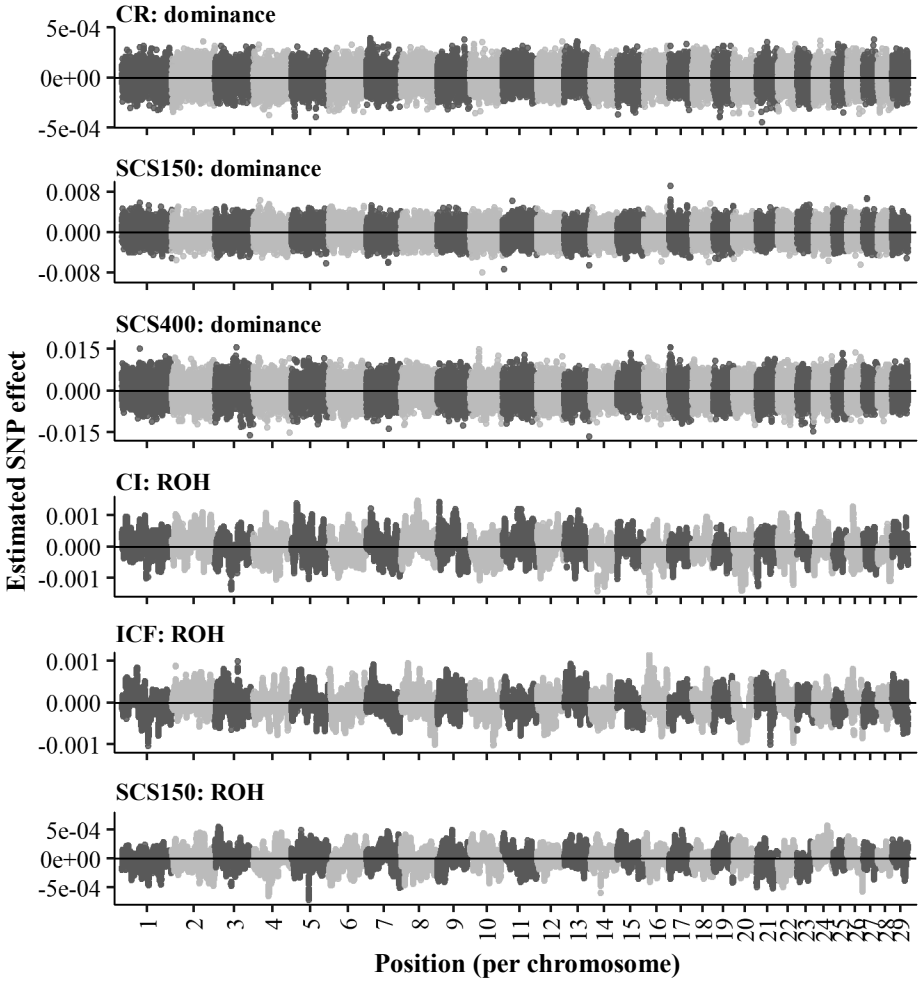


Figure S5.2 (continued)

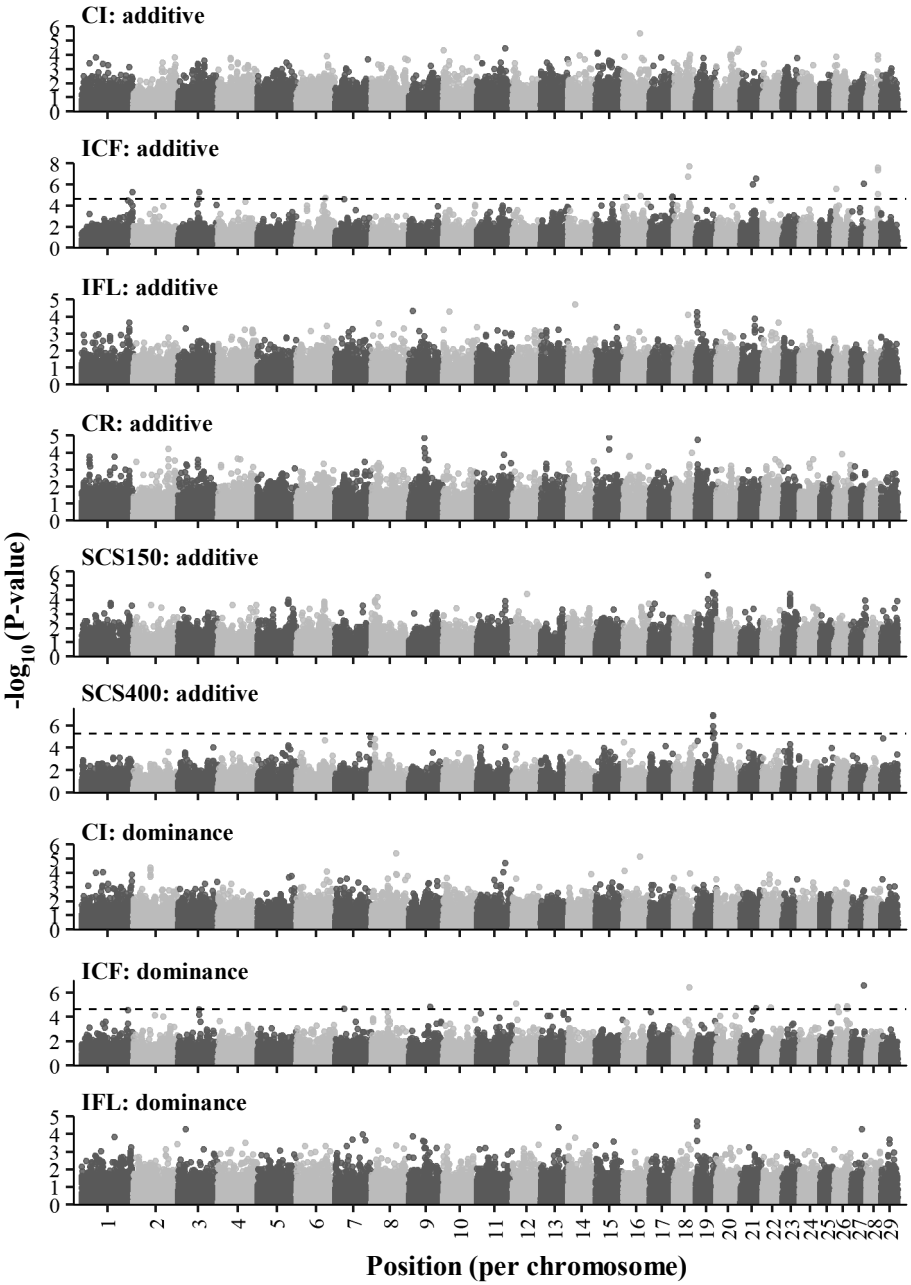


Figure S5.3 Statistical significance of additive, dominance and ROH effects for fertility and udder health traits based on single SNP GWAS (*continued on next page*). The horizontal red line is a threshold based on 10% false-discovery rate (absence of this line indicates that all effects were below the threshold). CI: calving interval (d); ICF: interval calving to first insemination (d); IFL: interval first to last insemination (d); CR: conception rate (%); SCS150 somatic cell score day 5 to 150 (units); SCS400: somatic cell score day 151 to 400 (units).

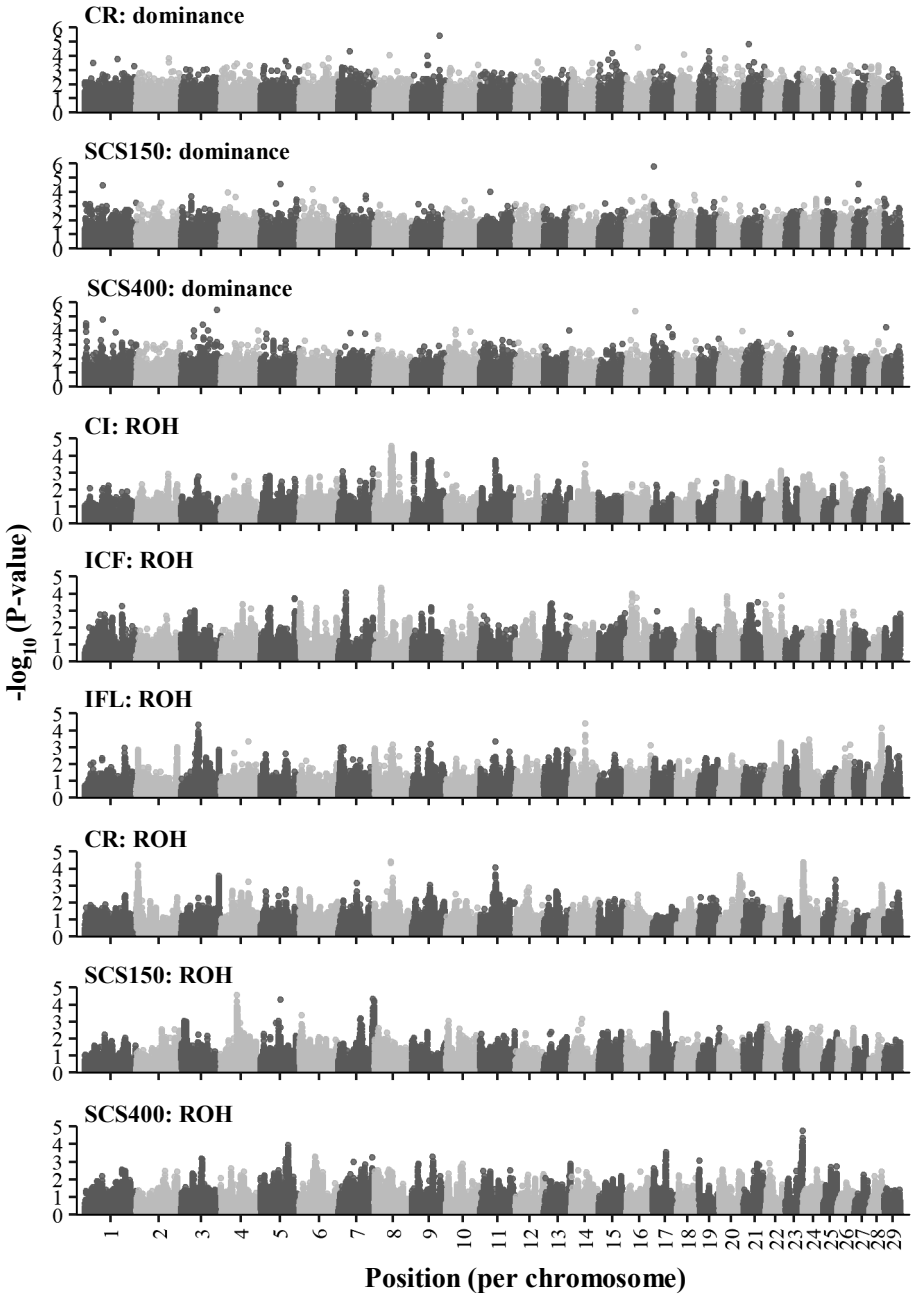


Figure S5.3 (continued)

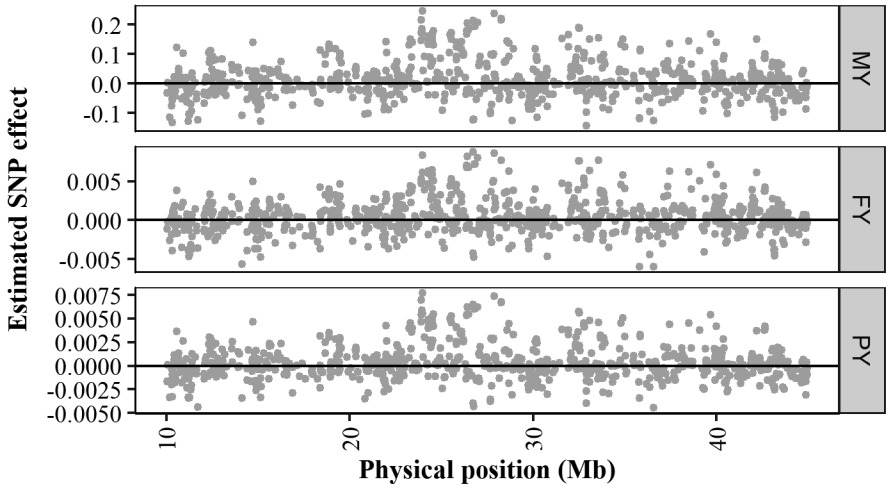


Figure S5.4 Dominance effects for yield traits, estimated by GREML (model ADR) with backsolving, for a region on chromosome 5 from 10 to 45 Mb. MY: 305-day milk yield (kg); FY: 305-day fat yield (kg); PY: 305-day protein yield (kg).

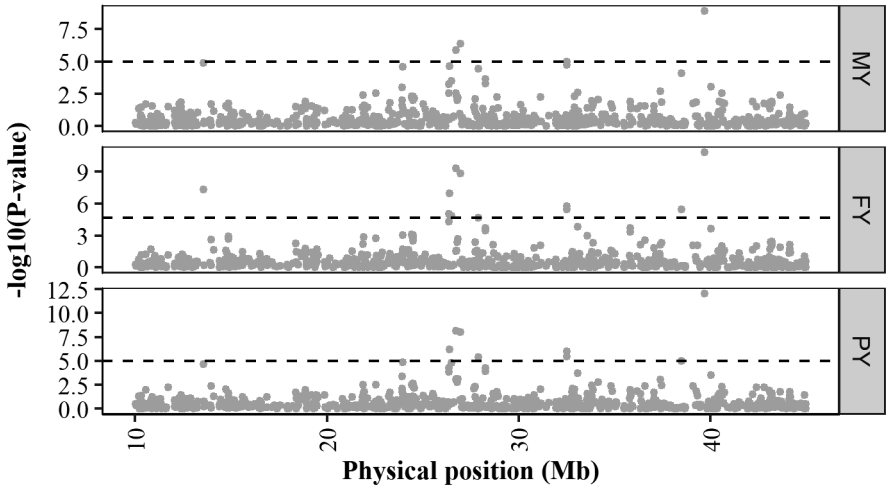


Figure S5.5 Statistical significance of dominance effects for yield traits, estimated by single SNP GWAS, for a region on chromosome 5 from 10 to 45 Mb. The horizontal dashed line is a threshold based on a 10% false-discovery rate. MY: 305-day milk yield; FY: 305-day fat yield; PY: 305-day protein yield.

6

Value of the Dutch Holstein Friesian germplasm collection to increase genetic variability and improve genetic merit

Harmen P. Doekes[1,2], Roel F. Veerkamp[1], Piter Bijma[1],
Sipke J. Hiemstra[2], Jack J. Windig[1,2]

[1]Wageningen University & Research Animal Breeding and Genomics, P.O. Box 338, 6700 AH, Wageningen, the Netherlands;
[2]Centre for Genetic Resources the Netherlands, Wageningen University & Research, P.O. Box 16, 6700 AA, Wageningen, the Netherlands

Journal of Dairy Science (2018) 101:10022-10033

Abstract

National gene bank collections for Holstein Friesian (HF) dairy cattle were set up in the 1990s. In this study, we assessed the value of bulls from the Dutch HF germplasm collection, also known as cryobank bulls, to increase genetic variability and improve genetic merit in the current bull population (bulls born in 2010–2015). Genetic variability was defined as 1 minus the mean genomic similarity (SIM_{SNP}) or as 1 minus the mean pedigree-based kinship (f_{PED}). Genetic merit was defined as the mean estimated breeding value for the total merit index or for 1 of 3 sub-indices (yield, fertility, and udder health). Using optimal contribution selection, we minimized relatedness (maximized variability) or maximized genetic merit at restricted levels of relatedness. We compared breeding schemes with only bulls from 2010 to 2015 with schemes in which cryobank bulls were also included. When we minimized relatedness, inclusion of genotyped cryobank bulls decreased mean SIM_{SNP} by 0.7% and inclusion of both genotyped and non-genotyped cryobank bulls decreased mean f_{PED} by 2.6% (in absolute terms). When we maximized merit at restricted levels of relatedness, inclusion of cryobank bulls provided additional merit at any level of SIM_{SNP} or f_{PED}, except for the total merit index at high levels of SIM_{SNP}. Additional merit from cryobank bulls depended on (1) the relative emphasis on genetic variability and (2) the selection criterion. Additional merit was higher when more emphasis was put on genetic variability. For fertility, for example, it was 1.74 SD at a SIM_{SNP} restriction of 64.5% and 0.37 SD at a SIM_{SNP} restriction of 67.5%. Additional merit was low to nonexistent for the total merit index and higher for the sub-indices, especially for fertility. At a SIM_{SNP} of 64.5%, for example, it was 0.60 SD for the total merit index and 1.74 SD for fertility. In conclusion, Dutch HF cryobank bulls can be used to increase genetic variability and improve genetic merit in the current population, although their value is very limited when selecting for the current total merit index. Anticipating changes in the breeding goal in the future, the germplasm collection is a valuable resource for commercial breeding populations.

6.1 Introduction

The Holstein Friesian (HF) breed is the dominating dairy cattle breed worldwide. Despite its census size of millions of individuals, the breed has an effective population size of 18-115 [102, 108, 173]. In the early 1990s, national HF gene bank collections were established to safeguard genetic variability [102]. Since then, *ex situ in vitro* conservation has been used as complementary strategy to *in situ in vivo* management of genetic variability in the breed [2].

In vitro conservation has several advantages and potential uses. One advantage is that the stored material harbors genetic variation of the population at the time of sampling, which may include variation that, since then, has been lost in vivo due to selection and drift. Material from cryobank individuals, therefore, could be used to restore or increase genetic variability in the current live population [90, 231]. In an extreme scenario in which the live population becomes highly endangered or extinct, e.g. due to a disease outbreak, the stored material could also be used to re-establish the population. Other potential uses of gene bank collections include the management of inbreeding in small populations [232, 233], the documentation of genetic trends [234, 235], and the introgression of specific genetic variants into live populations (e.g. introgression of the polled allele).

Recently, Doekes et al. [173] reported a decrease in genetic variability in the Dutch-Flemish HF breeding program from 1986 through to 2015, with a particularly fast decrease since the introduction of genomic selection. This recent decrease suggests that old bulls from the Dutch HF germplasm collection could be used to increase variability in the current population.

A disadvantage of old cryobank bulls is that their genetic level, measured by estimated breeding values (EBVs), is generally lower than that of recently born bulls. Consequently, the use of cryobank bulls to increase genetic variability in the current population is expected to reduce genetic merit. This hypothesis, however, does not have to hold for all traits, because not all traits currently of interest have been continuously selected for in the past. For example, while HF breeding goals consisted of mainly yield and conformation traits before 2000, they now also include many traits related to health, reproduction and longevity [34, 37, 173]. When the focus of the breeding goal would shift towards (one of) the latter traits, cryobank bulls might have value for the population in terms of both genetic variability and genetic merit.

The aim of this study was to assess the value of the Dutch HF germplasm collection to increase genetic variability and improve genetic merit in the current bull population. We considered three scenarios: (1) maximizing genetic variability, (2) maximizing genetic merit for the total merit index while maintaining variability, and (3) maximizing genetic merit for a sub-index (yield, fertility, or udder health) while maintaining variability. In addition to a SNP-based assessment, we performed a

pedigree-based evaluation to include bulls from the germplasm collection that had no genotype data.

6.2 Material and methods

6.2.1 Germplasm collection, groups and data

The Dutch HF germplasm collection was set up in 1993 and is now managed by the Centre for Genetic Resources the Netherlands (CGN) of Wageningen University & Research. The collection consists of 5,457 HF bulls (fraction HF > 87.5%, either red or black). The majority of these bulls is from progeny testing schemes of two commercial companies, the Dutch-Flemish cattle improvement co-operative (CRV; Arnhem, the Netherlands) and Alta Genetics (Feerwerd, the Netherlands).

In this study, we used 5,783 HF bulls (both cryobank and non-cryobank bulls). To assess the additional value of the germplasm collection to the current bull population, we defined four groups: very young bulls (VYB; n = 212), young bulls (YB; n = 762), cryobank bulls with genotype data (CBG; n = 2,888) and cryobank bulls with only pedigree data (CBP; n = 1,921). The VYB consisted of bulls born in 2014-2015, with EBVs based on only genomic and parental information. The YB consisted of bulls born in 2010-2013, with EBVs based on genomic and parental information (23% of bulls) or on genomic, parental and daughter information (77% of bulls). The mean reliability of EBVs was about 60% in the VYB and about 80% in the YB. Of the VYB and YB, respectively 68 bulls (= 32%) and 551 bulls (= 72%) were also stored in the germplasm collection. The CBG consisted of genotyped cryobank bulls born in 1985-2009. The CBP consisted of cryobank bulls born in 1978-2015 that had pedigree but no genotype data. Figure 6.1 shows the number of bulls by group and year of birth.

Pedigree and genotype data were provided by CRV. The total pedigree comprised 429,981 individuals. Pedigree completeness per bull was assessed with the number of ancestral generations completely known (NCG) and the complete generation equivalent (CGE). We computed the CGE for each bull as the sum of $(½)^n$ over all its' known ancestors, with n being the generation number of a given ancestor. Bulls with a NCG < 2 were excluded from the analyses (n = 29; these bulls had no genotype data and were excluded from the abovementioned group sizes). Genotyping was performed with the Illumina BovineSNP50 BeadChip (versions v1 and v2) or CRV custom-made 60 k Illumina panel (versions v1 and v2). Genotypes were imputed to 76 k from the different panels, following Druet et al. [115]. Prior to imputation, SNPs with a call rate lower than 0.85, a MAF lower than 0.025 or a difference of more than 0.15 between observed and expected heterozygosity were discarded. We also discarded SNPs with unknown position on the Btau4.0 genome assembly. After quality control and imputation, the dataset consisted of 75,538 autosomal SNPs per bull.

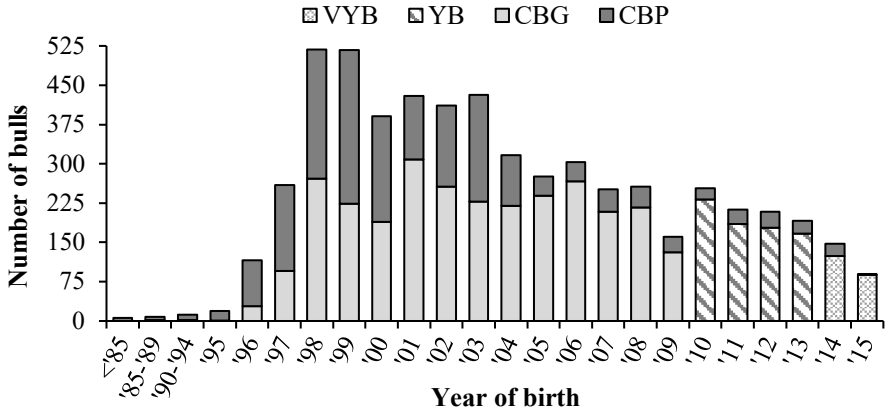


Figure 6.1 Number of bulls by year of birth and group. VYB = very young bulls (n = 212); YB = young bulls (n = 762); CBG = cryobank bulls with genotype data (n = 2,888); CBP = cryobank bulls with only pedigree data (n = 1,921).

6.2.2 Measures of genetic variability

Genetic variability was defined as one minus the mean relatedness in a population. We considered two measures of relatedness: the genomic similarity (SIM_{SNP}) and the pedigree-based kinship (f_{PED}).

The $SIM_{SNP_{ij}}$ was defined as the probability that two alleles at a random SNP, one sampled from bull i and one from bull j, were identical by state [51]. To calculate $SIM_{SNP_{ij}}$, we first computed a genomic relationship matrix for all genotyped individuals with allele frequencies fixed to 0.5, using *calc_grm* [118]. We then scaled the obtained relationships (G_{ij}) to genomic similarities, according to $SIM_{SNP_{ij}} = \frac{G_{ij}+2}{4}$ (see Additional file 1 of Eynard et al. [121] for derivation). Note that the mean SIM_{SNP} in a population is equal to 1 minus the expected heterozygosity.

The $f_{PED_{ij}}$ was defined as the pedigree-based probability that two alleles at a random (imaginary) selection-free locus, one sampled from bull i and one from bull j, were identical by descent with reference to a base population [24]. Founders in the pedigree were considered as base population. We first computed an additive genetic relationship matrix for all bulls with *calc_grm* [118], according to the algorithms of Sargolzaei et al. [119] and Colleau [120]. We then obtained the $f_{PED_{ij}}$ as half of the additive genetic relationship.

6.2.3 Measures of genetic merit

Genetic merit was defined as the mean EBV for a selection index. We considered four selection indices: a total merit index (NVI) and three sub-indices, namely yield (INET), daughter fertility (FERT) and udder health (UH). All EBVs were obtained from

the December 2017 publication of the organization for genetic evaluation of bulls in the Netherlands and Flanders [236]. The NVI is the Dutch-Flemish total merit index which includes INET, FERT, UH, longevity, conformation and birth traits with relative weights (based on the sum of genetic SDs) of 26%, 14%, 14%, 11%, 30% and 5%, respectively [237]. The breeding value for INET was composed of the EBVs for lactose yield (LACT), fat yield (FAT) and protein yield (PROT), and calculated as: $INET = 0.3 * LACT + 2.1 * FAT + 4.1 * PROT$ [238]. The breeding value for FERT was composed of the EBVs for the interval between first and last insemination (IFL) and the calving interval (CI), and calculated as: $FERT = 0.52 * (IEL - 100) + 0.52 * (CI - 100) + 100$ [239]. The breeding value for UH was composed of the EBVs for subclinical mastitis (SCM) and clinical mastitis (CM), and calculated as: $UH = 0.477 * (SCM - 100) + 0.641 * (CM - 100) + 100$ [240]. The EBVs for FERT and UH were rescaled such that the mean equaled 100 and the SD at population level was 4, whereas the NVI and INET were used on their original scales [237-240].

6.2.4 Optimal contribution selection

Since its introduction in the late 1990s [89, 241], OCS has become the golden standard to maximize the mean EBV in the next generation, while restricting the mean relatedness to a predefined value. The restriction on relatedness is generally based on the desired rate of inbreeding (e.g. 0.5% or 1%). In addition to balancing genetic merit and variability, OCS may also be used to maximize variability, by minimizing relatedness irrespective of genetic gain (e.g. [90]).

In this study, we used OCS to assess the value of cryobank bulls to the current bull population. We compared results of OCS-schemes with only (very) young bulls to those with both (very) young bulls and cryobank bulls. More specifically, we considered the following four populations: (1) VYB, (2) VYB + YB, (3) VYB + YB + CBG, and (4) VYB + YB + CBG + CBP. For each population, we first ran OCS to maximize variability (i.e. minimize mean SIM_{SNP} or mean f_{PED}). We then maximized the mean EBV for either NVI, INET, FERT or UH, while restricting the mean SIM_{SNP} or mean f_{PED} to predefined values. These predefined values ranged from a minimum (previously determined by minimizing SIM_{SNP} or f_{PED}) to a maximum of 68% for SIM_{SNP} and 12.5% for f_{PED}. The chosen maxima corresponded to a rate of inbreeding of about 5%, when considering the VYB as the current generation. Between the minimum and maximum, we ran scenarios at intervals of 0.04% for SIM_{SNP} and of 0.08% for f_{PED}. Analyses were performed with Gencont [242], which uses the Lagrangian multiplier approach to solve the optimization problem [89].

To compare results across traits, we will present the differences in merit between populations not only on the original index scales, but also in SDs (which equaled 128.3 for NVI, 136.0 for INET, 4.7 for FERT, and 5.1 for UH).

To further visualize the value of cryobank bulls, we evaluated the total contribution that was assigned to each subgroup when running OCS with all genotyped bulls (VYB + YB + CBG) for SIM_{SNP} or with all bulls with pedigree (VYB + YB + CBG + CBP) for f_{PED}.

6.3 Results

6.3.1 Genetic Trends

Mean NVI and INET have increased continuously over the last 30 years (Figure 6.2). Mean FERT and mean UH first decreased until they were included in the breeding goal around 2000. Since then, the genetic level for FERT and UH has steadily increased.

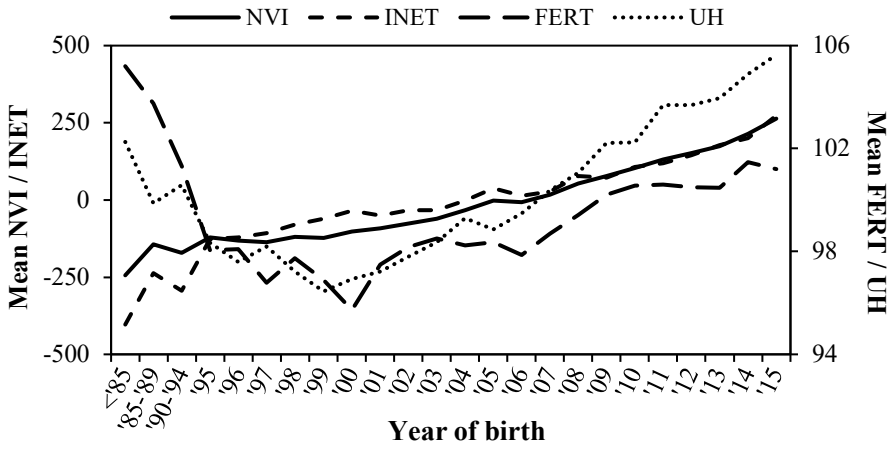


Figure 6.2 Mean estimated breeding value for four selection indices (NVI, INET, FERT and UH) for all bulls (n = 5,783) by year of birth. NVI = total merit index; INET = yield index; FERT = daughter fertility index; UH = udder health index. Note that NVI and INET are shown on the primary y-axis, and FERT and UH on the secondary y-axis.

6.3.2 Descriptive statistics across groups

Mean pedigree completeness (NCG and CGE), mean relatedness (SIM_{SNP} and f_{PED}) and mean EBV (for all traits) were greatest in the VYB, followed by the YB, CBG and CBP (Table 6.1). Means for the CBG and CBP were relatively similar, as compared to means in the VYB and YB. Cryobank bulls showed most variation in EBVs (i.e. greatest SD), followed by the YB and the VYB. Cryobank bulls also showed the lowest minimum EBV, followed by the YB and the VYB. The maximum of the total merit index NVI was smaller in the CBG (268) and CBP (287) than in the YB (374) and VYB (380). In fact, the maxima for NVI in the CBG and CBP were less than 1 SD above the mean EBV in the VYB. For the sub-indices (INET, FERT and UH), however, the maximum was similar across groups and the maxima for the CBG and CBP were in

the upper tail of the corresponding distributions in the VYB and YB (i.e., above the mean + 2 SD).

Table 6.1 Descriptive statistics for pedigree completeness (two measures), relatedness (two measures, in %) and estimated breeding values (four selection indices) by group of bulls.

Statistic & group[1]	Completeness[2]		Relatedness[3]		Estimated breeding values[4]			
	NCG	CGE	SIM_{SNP}	f_{PED}	NVI	INET	FERT	UH
Mean								
VYB	7.27	13.95	65.85	7.82	241.1	237.5	101.5	105.3
YB	6.61	12.67	65.25	6.91	141.8	139.5	100.7	103.4
CBG	5.92	10.86	64.48	5.20	-46.7	-7.38	98.0	98.5
CBP	5.71	10.48	NA[5]	5.14	-78.9	-51.1	97.8	98.2
SD								
VYB	0.89	0.44	0.97	2.11	46.7	73.7	2.2	2.9
YB	1.12	0.55	1.05	2.39	64.7	100.6	3.6	3.5
CBG	0.98	0.72	0.88	2.05	99.6	117.1	5.0	4.9
CBP	1.07	0.91	NA	2.19	101.5	118.0	4.3	4.8
Minimum								
VYB	4	12.79	63.13	4.00	124	36	96	95
YB	2	10.59	62.43	2.23	-96	-181	88	89
CBG	2	7.38	61.92	0.48	-413	-410	76	80
CBP	2	5.90	NA	0.19	-438	-487	84	75
Maximum								
VYB	9	14.89	75.23	31.43	380	448	107	112
YB	9	14.06	75.78	32.38	374	415	111	114
CBG	8	12.71	75.75	33.22	268	472	113	111
CBP	9	14.06	NA	31.47	287	382	114	113

[1]VYB = very young bulls (n = 212); YB = young bulls (n = 762); CBG = cryobank bulls with genotype data (n = 2,888); CBP = cryobank bulls with only pedigree data (n = 1,921). [2]NCG = number of completely known generations; CGE = complete generation equivalent. [3]SIM_{SNP} = genomic similarity (excluding self-similarities); f_{PED} = pedigree-based kinship (excluding self-kinships). [4]NVI = total merit index; INET = yield index; FERT = daughter fertility index; UH = udder health index. [5]NA = not applicable.

6.3.3 Maximizing genetic variability

Genetic variability was maximized by minimizing mean relatedness (either mean SIM_{SNP} or mean f_{PED}) with OCS. Minimization of SIM_{SNP} in the VYB decreased mean SIM_{SNP} from the current generation (65.9%) to the next generation (65.4%) by 0.5% (Table 6.2). This is equivalent to a 0.5% increase in mean heterozygosity. When the YB was added to the VYB, the mean SIM_{SNP} in the next generation further decreased by 0.9% (to 64.5%). Adding the CBG resulted in a further decrease of 0.7%. In other words, inclusion of genotyped cryobank bulls increased expected heterozygosity by

0.7% when compared to a scheme in which only very young and young bulls were used. Minimization of f_{PED} in the VYB decreased mean f_{PED} from the current generation (7.8%) to the next generation (7.0%) by 0.8%. Stepwise adding the YB, CBG and CBP resulted in further decreases of 1.8%, 1.8% and 0.8%, respectively. Thus, the inclusion of genotyped and non-genotyped cryobank bulls decreased mean f_{PED} by 2.6% when compared to the scenario in which only very young and young bulls were used. Note that although in absolute terms the realized decrease for f_{PED} was larger than that for SIM_{SNP}, in relative terms (i.e. scaled by the non-inbred part) they were quite similar. The difference between the VYB before OCS and the VYB + YB + CBG after OCS, for example, was in absolute terms 2.1% for SIM_{SNP} and 4.3% for f_{PED}, and in relative terms 6.1% for SIM_{SNP} and 4.7% for f_{PED}.

Table 6.2 Mean relatedness (in %) before and after minimizing relatedness with optimal contribution selection as well as number of selected bulls (nsel) for 4 populations and 2 relatedness measures.

		Minimized relatedness measure[2]					
		SIM_{SNP}			f_{PED}		
Population[1]	n	before	after	nsel	before	after	nsel
VYB	212	65.85	65.37	91	7.82	7.02	127
VYB + YB	974	65.33	64.51	151	7.02	5.24	232
VYB + YB + CBG	3,862	64.50	63.78	225	5.26	3.45	236
VYB + YB + CBG + CBP	5,783	NA[3]	NA	NA	5.16	2.63	247

[1]VYB = very young bulls (n = 212); YB = young bulls (n = 762); CBG = cryobank bulls with genotype data (n = 2,888); CBP = cryobank bulls with only pedigree data (n = 1,921). [2]SIM_{SNP}= genomic similarity; f_{PED} = pedigree-based kinship. [3]Not applicable.

The decrease in mean relatedness that was achieved by including additional groups of selection candidates was accompanied by an increase in the number of candidates that was selected (and by a decrease in the percentage of candidates that was selected). The increase in number of selected candidates was especially apparent when moving from VYB to VYB + YB for both SIM_{SNP} and f_{PED}, and when moving from VYB + YB to VYB + YB + CBG for SIM_{SNP} (Table 6.2). Moving from the VYB to larger populations was also accompanied by a redistribution of contributions among selected bulls.

Maximizing genetic variability, irrespective of genetic merit, decreased mean EBV for all indices (Figure 6.3). For example, minimizing mean SIM_{SNP} in the VYB decreased the mean NVI from the current generation (241.1) to the next generation (227.8) by 13.3 points. Stepwise adding the YB and CBG resulted in further decreases of 107.7 points (to 120.1) and 193.1 points (to -73.0), respectively. Thus, there was a clear cost in merit when selecting only for variability.

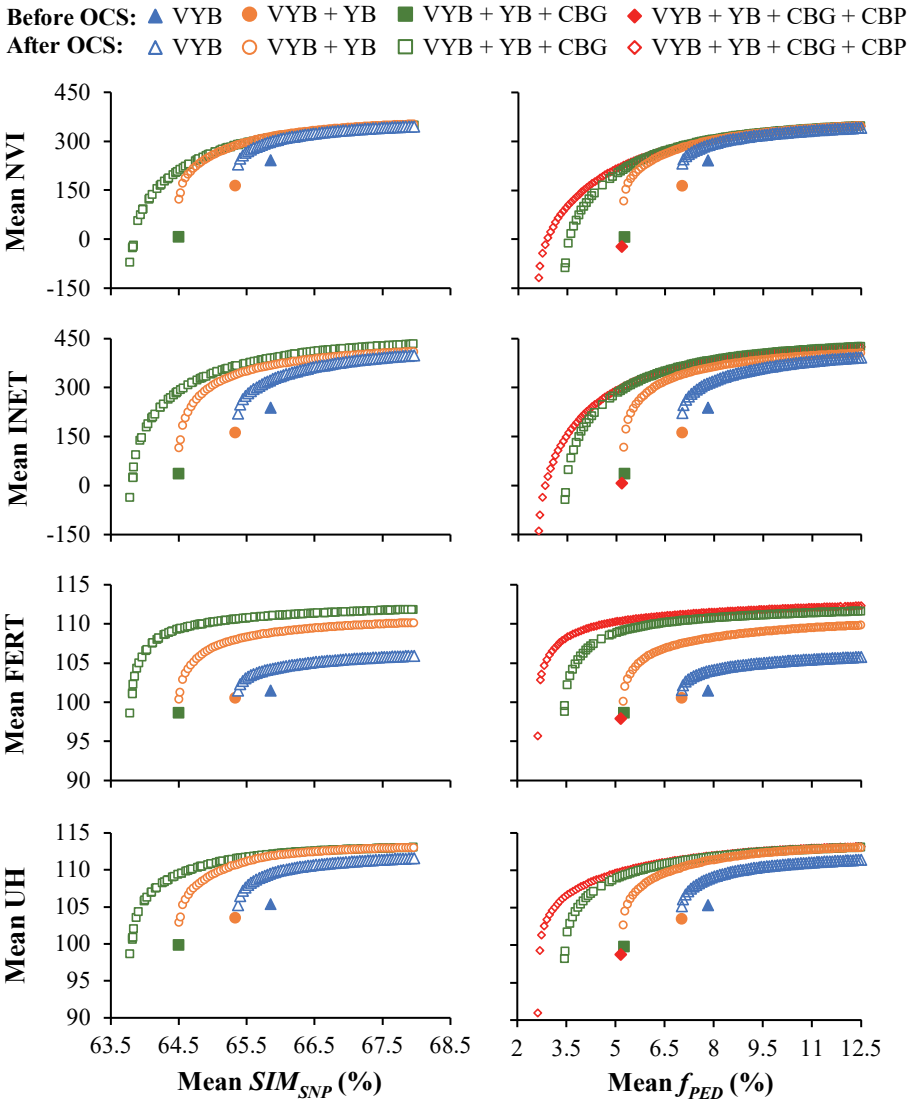


Figure 6.3 Mean genetic merit at restricted levels of relatedness (SIM_{SNP} left, f_{PED} right) before and after maximizing merit for one of four selection indices (NVI, INET, FERT and UH) with optimal contribution selection (OCS) in four populations (combinations of VYB, YB, CBG and CBP). Levels of relatedness ranged from a minimum, obtained by minimizing mean SIM_{SNP} or mean f_{PED}, to a maximum of 68% and 12.5% for SIM_{SNP} or f_{PED}, respectively. SIM_{SNP} = genomic similarity; f_{PED} = pedigree-based kinship; NVI = total merit index; INET = yield index; FERT = fertility index; UH = udder health index; VYB = very young bulls (n = 212); YB = young bulls (n = 762); CBG = cryobank bulls with genotype data (n = 2,888); CBP = cryobank bulls with only pedigree data (n = 1,921).

6.3.4 Maximizing genetic merit while maintaining genetic variability

The inclusion of additional groups of (old) selection candidates, such as cryobank bulls, resulted in more merit at the same level of variability (Figure 6.3). At a mean SIM_{SNP} equal to the mean SIM_{SNP} of the current VYB (65.58%), for example, maximization for INET resulted in a INET of 322.8 when using VYB, of 365.6 when using VYB + YB and of 389.3 when using VYB + YB + CBG. An exception was found for scenarios in which NVI was maximized at high levels of mean SIM_{SNP}. For these scenarios the VYB + YB + CBG provided slightly less merit than the VYB + YB (see Discussion).

The benefit of including additional groups of (old) selection candidates, such as genotyped cryobank bulls, was greater when more emphasis was put on genetic variability. In other words, the difference between the curves in Figure 6.3 was greater at lower levels of relatedness. For example, the additional merit for FERT obtained by adding the CBG to the VYB + YB at mean SIM_{SNP}-levels of 64.5%, 65.5%, 66.5% and 67.5% was 8.1 (= 1.74 SD), 2.4 (= 0.51 SD), 1.9 (=0.4 SD) and 1.7 (= 0.37 SD), respectively.

The benefit of including additional groups of (old) selection candidates, such as genotyped cryobank bulls, at specific levels of relatedness differed across selection indices (Figure 6.4). The additional merit obtained by adding the CBG to the VYB + YB at a mean SIM_{SNP} of 64.5%, for example, was 0.60 SD (= 71.2 points) for NVI, 1.06 SD (= 144.3 points) for INET, 1.74 SD (= 8.1 points) for FERT and 1.18 SD (= 6.0 points) for UH. For NVI, there was no additional merit of including the CBG at mean SIM_{SNP} above 65.24%. For UH, there was no additional merit (i.e. < 0.01 SD) of the CBG at mean SIM_{SNP} above 65.80%. For INET, additional merit of the CBG was relatively stable (at about 0.16 SD) for mean SIM_{SNP} above 65%. Of the four indices, FERT showed the greatest additional merit for CBG. Although the additional merit for FERT decreased when mean SIM_{SNP} increased, there was still benefit of including CBG at high levels of mean SIM_{SNP} (e.g. 0.36 SD more merit at a mean SIM_{SNP} of 67.5%).

Results for f_{PED} were similar to those for SIM_{SNP} (Figure 6.3). The VYB + YB + CBG + CBP resulted in more genetic merit compared to the VYB + YB + CBG, but this difference was only present at very low levels of mean f_{PED}. For the NVI, INET and FERT the additional merit of the CBP quickly decreased with increasing mean f_{PED} and for mean f_{PED} of ≥ 5% it was approximately zero. For FERT, there was still a bit of additional merit at higher levels of mean f_{PED} (about 0.13 SDs).

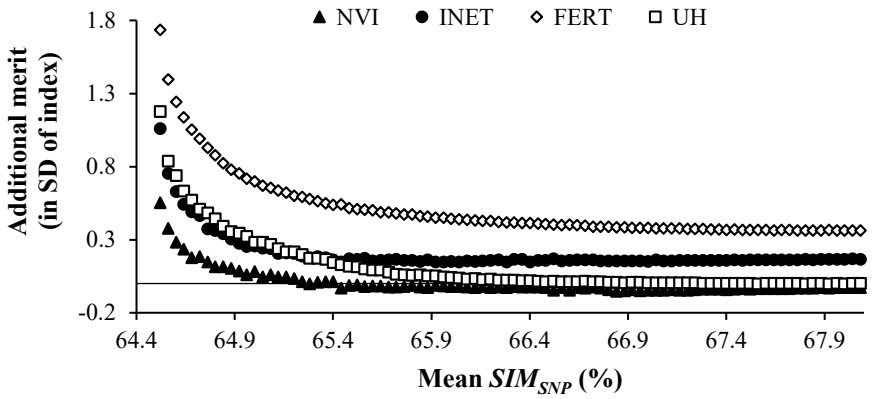


Figure 6.4 Additional merit (expressed in SDs of selection indices) achieved with VYB + YB + CBG compared to VYB + YB at various levels of genomic similarity (SIM_{SNP}). NVI = total merit index; INET = yield index; FERT = fertility index; UH = udder health index; VYB = very young bulls (n = 212); YB = young bulls (n = 762); CBG = cryobank bulls with genotype data (n = 2,888).

6.3.5 Contributions of groups

When SIM_{SNP} was minimized with all genotyped bulls (VYB + YB + CBG), 89% of contributions was assigned to the CBG, 10% to the YB and 1% to the VYB (Figure 6.5). When f_{PED} was minimized with all bulls (VYB + YB + CBG + CBP), 64% of contributions was assigned to the CBP, 34% to the CBG, 2% to the YB and nothing to the VYB.

As expected from Figure 6.3, the contribution of cryobank bulls (CBG and CBP) generally decreased when the restriction on relatedness became less stringent. The exact pattern differed across selection indices. For the total merit index NVI, the contribution of cryobank bulls continued to decrease with increasing levels of relatedness. At very high relatedness levels, only bulls from the VYB (about 70%) and YB (about 30%) were selected. For INET, the contribution of cryobank bulls also decreased with increasing relatedness, but remained stable (at about 40%) for mean SIM_{SNP} of ≥ 65.5% and for mean f_{PED} of ≥ 8%. For FERT, about 90% of the contributions was assigned to cryobank bulls at any level of relatedness. For UH, the contribution of cryobank bulls also decreased with an increase in relatedness. This index showed a relatively high contribution of the YB. For UH there was a single bull in the CBP (born in 2004) with a very high EBV (of 113), which was assigned a contribution of about 15% in the scenarios with high mean f_{PED}.

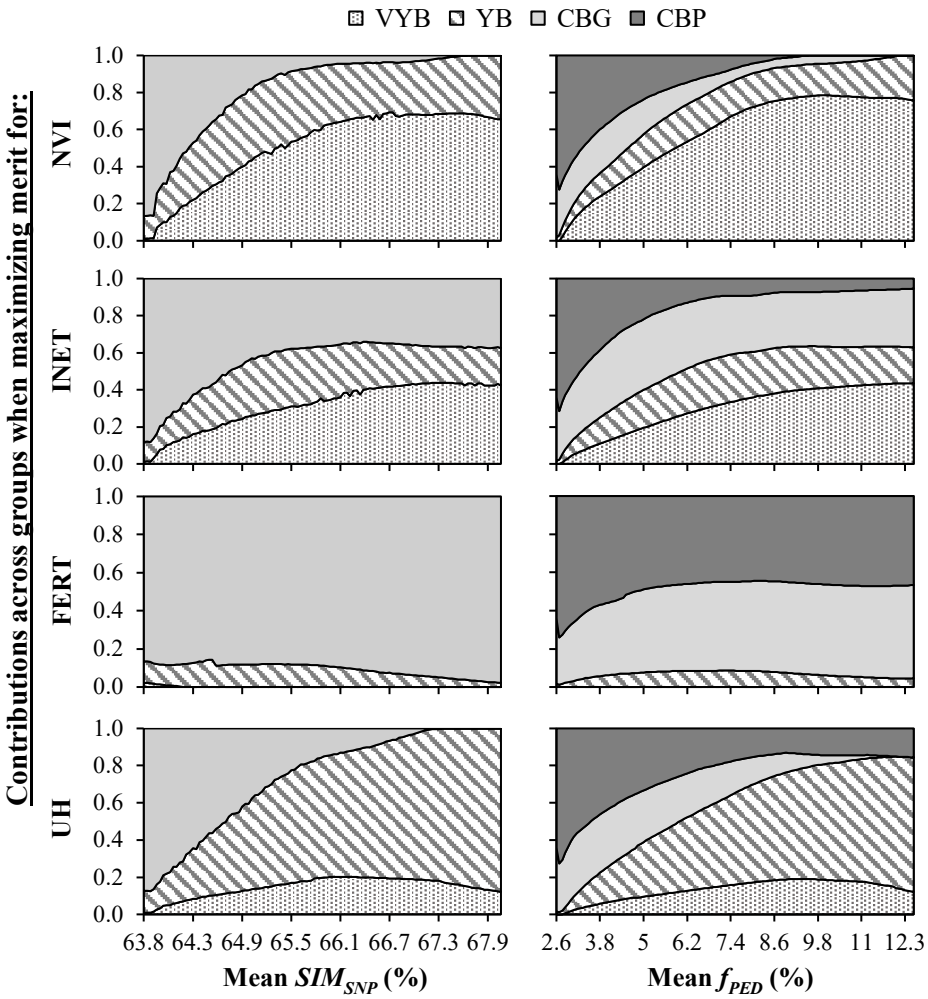


Figure 6.5 Contributions across groups when maximizing genetic merit for one of four selection indices (NVI, INET, FERT and UH) at restricted levels of mean SIM_{SNP} in the VYB + YB + CBG population (left), or at restricted levels of mean f_{PED} in the VYB + YB + CBG + CBP population (right). SIM_{SNP} = genomic similarity; f_{PED} = pedigree-based kinship; NVI = total merit index; INET = yield index; FERT = fertility index; UH = udder health index; VYB = very young bulls (n = 212); YB = young bulls (n = 762); CBG = cryobank bulls with genotype data (n = 2,888); CBP = cryobank bulls with only pedigree data (n = 1,921).

6.4 Discussion

6.4.1 Value of the germplasm collection

Our primary objective was to assess the value of the Dutch HF germplasm collection to increase genetic variability and improve genetic merit in the current bull population. The selection of almost exclusively cryobank bulls when maximizing

genetic variability (Figure 6.5) shows that diversity of the current bull population is well captured by the germplasm collection. This finding is in line with the results of Danchin-Burge et al. [102]. Results based on minimization of mean SIM_{SNP} and mean f_{PED} furthermore suggest that the germplasm collection can be used to increase genetic variability in the current population (Table 6.2). When genetic merit and genetic variability are balanced, the inclusion of cryobank bulls in addition to bulls from the current population may result in more merit at the same level of variability or, equivalently, in more variability at the same level of merit (Figure 6.3). We found that the benefit of using cryobank bulls depended on 2 factors: (1) the relative emphasis on variability and (2) the selection criterion (i.e., the index).

The additional merit of cryobank bulls, and the percentage of contributions assigned to cryobank bulls, was greater when more emphasis was put on genetic variability (Figures 6.3 and 6.4). This is not surprising because both relatedness and genetic merit have increased over time (Figure 6.2 and [173]).

Additional merit of cryobank bulls was relatively low for the current total merit index NVI and greater for the sub-indices INET, UH, and FERT (Figures 6.3 and 6.4). For the NVI, there was almost no value in including cryobank bulls as selection candidates (except when we put a lot of emphasis on genetic variability). The limited additional merit for the NVI can be explained by the fact that the NVI has been the main index under selection in recent decades and that bulls from the current population simply have the highest NVI (Figure 6.2 and Table 6.1). This finding also suggests that selection for NVI in the past has been effective, at least for the EBV. The observed difference between additional merit of cryobank bulls for the NVI on one hand and the sub-indices on the other reflects the principle that “single-trait selection often suffers from antagonistic correlations with traits not in the selection objective,” whereas “multiple-trait selection avoids those problems at the cost of less-than-maximal progress for individual traits” [34]. Past selection for NVI has resulted in less-than-maximal progress for the sub-indices. Some bulls that were assigned high contributions in this study when selecting for a sub-index (e.g., yield), may not have been used so much in practice because they scored relatively low for other traits. The relatively high additional merit for FERT can be explained by the availability of cryobank bulls with high FERT, which were born before the intense selection for yield traits in the last decades of the 20th century and, thus, before the associated decrease in FERT (Figure 6.2) that was due to the well-known antagonistic correlation between fertility and yield traits.

Cryobank bulls with only pedigree data (i.e., CBP) were assigned up to 70% of the contributions when maximizing merit at low levels of mean f_{PED} (Figure 6.5). This finding suggests that, based on pedigree, there are bulls in the CBP that are of interest with regard to genetic variability. To determine whether these bulls are also

less related at the genomic level or whether they were simply selected by OCS because they had limited pedigree completeness, they would have to be genotyped.

It is important to note that the actual breeding program is more dynamic than the schemes we evaluated. First, selection in practice occurs in both sexes, whereas we considered only bulls. We believe that the use of only bulls was sufficient for the purpose of this study because genetic variability in the population is largely driven by relatedness between bulls (due to substantial use of AI). Nevertheless, selection at the cow side may offer some possibilities for management of genetic variability, especially because cows show lower relatedness levels than AI bulls [172]. Second, the definition of the current bull population is not straightforward. We focused on bulls from a single breeding program, whereas in practice bulls from other breeding programs (from other countries) also may be used as selection candidates. Furthermore, we used 2 groups of bulls to represent the current bull population: 1 consisting of only VYB, born in 2013 to 2015, and 1 consisting of VYB and YB, born in 2010 to 2015. In practice, mostly top bulls from the VYB will be selected for breeding together with some top bulls from earlier years. To investigate the effect of focusing on top bulls, we performed OCS using the 50 top bulls from the VYB with the highest NVI. For any mean SIM_{SNP} above 66.3%, the NVI obtained with the top bulls was approximately identical to the NVI obtained with the entire VYB. When using only top bulls, a mean SIM_{SNP} below 66.3% could not be realized, whereas a mean SIM_{SNP} of 65.4% could be realized when using the entire VYB. This finding emphasizes that focusing only on top bulls further reduces genetic variability and highlights the importance of also storing lower ranked bulls in the gene bank collection.

In this study, we considered only a single generation of OCS. As shown by Leroy et al. [231], using cryobank bulls to increase genetic variability in a single generation will have no effect in the long term if their offspring are not selected subsequently. Genetic variability, therefore, should receive attention in subsequent generations as well. Continuous use of OCS may ensure that offspring of cryobank bulls are selected depending on the relative emphasis put on genetic variability. It would be interesting to investigate the decrease in long-term contributions of cryobank bulls considering various constraints on loss of genetic variability.

6.4.2 Limitation of the Lagrangian multiplier approach

An unexpected result was found when maximizing NVI at high levels of mean SIM_{SNP} in the VYB + YB and VYB + YB + CBG populations. For these scenarios, the obtained NVI was greater for the VYB + YB than for the VYB + YB + CBG (although the difference was small, i.e. < 0.05 SD; Figure 6.4). This result is theoretically impossible, because all bulls in the former population were also part of the latter population. If the CBG

would have provided no additional merit at all, the obtained NVI at a given mean SIM_{SNP} for the VYB + YB + CBG should have been at least equal to that for the VYB + YB. The observed pattern, therefore, has to be an artefact of the used algorithm. A difficulty of running OCS with the Lagrangian multiplier approach is that some contributions in the obtained optimal solution may be negative [89]. This problem is remedied in Gencont by eliminating selection candidates with negative contributions and repeating the optimization procedure until no negative contributions remain. The drawback of this approach, however, is that some of the candidates that were eliminated in early iterations, may have had a positive contribution in the true optimal solution [50, 95]. When we compared high-SIM_{SNP} scenarios between the VYB + YB and the VYB + YB + CBG, we observed that some bulls with moderate to high contributions in the former population were not selected at all in the latter population. This could be due to the elimination procedure. A possibility to get closer to the true optimal solution is to remove only a single candidate per iteration (the one with the most negative contribution), instead of eliminating all candidates with negative contributions. This remedy, however, would increase computation time. Alternatively, one may consider a completely different approach to solve the optimization problem, such as semidefinite-programming [95].

6.4.3 Future perspectives for HF germplasm collections

The Dutch HF germplasm collection is a rather unique collection containing material from many AI bulls over a period of approximately 40 yr. In addition to the national gene bank collection, AI companies and farmers have stored germplasm over time. These companies and individual farmers, however, cannot guarantee the availability of stored material in the long term. Systematically storing genetic material in national collections, therefore, is required to safeguard the material for future use.

Various national HF collections exist across the globe. For the collections of the Netherlands, France, and the United States, Danchin-Burge et al. [102] showed that there is substantial overlap in terms of the stored (pedigree-based) variability. An interesting question is whether national collections should be combined to reduce storage redundancy. We believe that although cooperation between gene banks is important to efficiently allocate resources and ensure that global HF variability is stored, there is substantial value in having separate national collections. The main advantage of having separate collections is that material from a national collection is more readily available for local breeding programs. In addition, having separate gene bank collections is an insurance against calamities. An important aspect of gene bank management is to determine which and how many individuals are stored in a collection. In the case of the Dutch HF, a vast majority of AI bulls have been stored in the collection over time, with generally 25 straws per bull. These preselected AI

bulls will not have covered all genetic variability and genetic potential for various traits present in the national population. To optimize collections, one could also store material from bulls that are not used for AI (and use OCS to determine exactly which bulls to store). The number of straws stored per individual is important with regard to the potential use, and therefore the value, of the collection. Today, fast genetic gain in HF is realized by producing many embryos through superovulation [243], which requires a lot of semen. One may question whether cryobank bulls, with a limited number of straws, can make a significant contribution in the current system. Producing 90% of all offspring in 1 generation with cryobank bulls (such as in the hypothetical FERT scenario in this study), for example, is not possible. Across multiple generations, however, gene bank material could be disseminated throughout the population.

In line with the simulation study of Leroy et al. [231], our results suggest that gene bank collections are mostly valuable when the aim is to increase genetic variability or when major changes in selection objectives or practices occur (and when the use of animals from other breeds is not preferred). Breeding goals for HF have changed in the past [34, 37, 173] and are expected to change further in the future. A complete shift from the total merit index to a sub-index such as FERT is very unlikely. Instead, we expect that relative weights for specific trait groups will gradually shift over time and that novel traits will be added to the breeding goal. Various factors may influence the shift in breeding goal composition, including production economics (e.g. milk quota), societal demands (e.g. animal welfare), environment (e.g. climate change), technology (e.g. midinfrared spectroscopy), and breeding value estimation (e.g. genomic prediction). By separating phenotype recording from the selection process, genomic prediction has removed the need for large-scale phenotyping and enabled selection for novel traits that are difficult to measure. An overview of such novel traits is provided by Boichard and Brochard [244], Egger-Danner et al. [245], and Cole and VanRaden [34]. Cryobank bulls may have relatively high EBV for novel traits because these traits have not been directly selected for in the past. Anticipating changes in the HF breeding goal in the future, the germplasm collection is a valuable resource in terms of both genetic variability and genetic merit.

6.5 Conclusions

Bulls from the Dutch HF germplasm collection can be used to increase genetic variability in the current breeding population. When genetic merit and genetic variability are to be balanced, the benefit of including cryobank bulls as selection candidates in addition to bulls from the current population depends on (1) the relative emphasis on genetic variability and (2) the selection criterion. Additional

merit from cryobank bulls is higher when more emphasis is put on variability. Additional merit from cryobank bulls is very low for the current total merit index but higher for the sub-indices INET, UH, and FERT (especially high for fertility). Anticipating changes in the HF breeding goal in the future, the germplasm collection is a valuable resource for commercial breeding populations in terms of both genetic variability and genetic merit.

6.6 Acknowledgements

The research leading to these results was conducted as part of the IMAGE project, which received funding from the European Union's Horizon 2020 Research and Innovation Program under grant agreement no. 677353. In addition, the Dutch Ministry of Agriculture, Nature and Food Quality (The Hague, the Netherlands) contributed financially (KB-21-004-003). The authors acknowledge CRV (Arnhem, the Netherlands) for providing pedigree and genotype data and thank Ina Hulsegge (Animal Breeding and Genomics, Wageningen University & Research, Wageningen, the Netherlands) for technical support.

7

Characterization of genetic diversity conserved in the gene bank for Dutch cattle breeds

Anouk E. van Breukelen[1], Harmen P. Doekes[1,2], Jack J. Windig[1,2], Kor Oldenbroek[1,2]

[1]Wageningen University & Research Animal Breeding and Genomics, P.O. Box 338, 6700 AH, Wageningen, the Netherlands;
[2]Centre for Genetic Resources the Netherlands, Wageningen University & Research, P.O. Box 16, 6700 AA, Wageningen, the Netherlands

Diversity (2019) 11(12):229

Abstract

In this study, we characterized genetic diversity in the gene bank for Dutch native cattle breeds. A total of 715 bulls from seven native breeds and a sample of 165 Holstein Friesian bulls were included. Genotype data were used to calculate genetic similarities. Based on these similarities, most breeds were clearly differentiated, except for two breeds (Deep Red and Improved Red and White) that have recently been derived from the MRY breed, and for the Dutch Friesian and Dutch Friesian Red, which have frequently exchanged bulls. Optimal contribution selection (OCS) was used to construct core sets of bulls with a minimized similarity. The composition of the gene bank appeared to be partly optimized in the semen collection process, i.e., the mean similarity within breeds based on the current number of straws per bull was 0.32% to 1.49% lower than when each bull would have contributed equally. Mean similarity could be further reduced within core sets by 0.34% to 2.79% using OCS. Material not needed for the core sets can be made available for supporting *in situ* populations and for research. Our findings provide insight in genetic diversity in Dutch cattle breeds and help to prioritize material in gene banking.

7.1 Introduction

Genetic diversity refers to all genetic differences between species, between breeds within species and between individuals within breeds measured as differences in DNA [1]. Genetic diversity is essential for sustainable livestock production, because it provides the base material for genetic improvement of livestock and their adaptation to changing socio-economic and environmental demands. In addition to the management of genetic diversity in *in situ* populations, genetic diversity can be conserved *ex situ* in gene banks. Gene bank collections are important for three main reasons [3, 7]: they (1) are an insurance against changes in market or environmental conditions; (2) are a safeguard against emerging diseases, political instability, and natural disasters; and (3) provide opportunities for research.

Prioritization of genetic material is an essential aspect of gene bank management, because financial and physical resources are generally limited [246]. Prioritization of material may occur at two levels. First, it has to be decided which animals from *in situ* populations are sampled to include in the gene bank. Second, genetic material within the gene bank can be divided into subsets based on their value for conservation. For the division of the total collection into subsets, different strategies may be used. One strategy is to set up a 'core collection' for each breed, consisting of cryo-conserved samples that would allow the reconstitution of that breed, with an effective population size of at least 50 [247]. This strategy focuses on the conservation of genotypes in single breeds and on having a backup in case of emergencies. It does not consider (overlap in) diversity across breeds. An alternative approach that focuses more on allelic diversity is to set up a 'core set', e.g. following the method by Eding et al. [248]. Such a core set comprises a subset of the total collection that is optimized in terms of diversity. It could be defined as the smallest set of individuals that still encompasses the genetic diversity within that breed (or within multiple breeds combined).

Over the last century, genetic diversity in domesticated European cattle populations has been affected by two major factors. First, starting in the late 1960s, the 'Holsteinization' took place. This upgrading process resulted in the development of a Dutch Holstein Friesian population at the expense of native cattle breeds, as cows from native cattle breeds were inseminated by Holstein Friesian bulls (Figure 7.1) [98]. Small population sizes in native breeds increase the risk on inbreeding and loss of genetic diversity through genetic drift [249]. Second, artificial selection, particularly in the Holstein Friesian breed, facilitated by techniques like artificial insemination and embryo transfer in combination with cryopreservation, has further reduced genetic diversity [102, 173].

In 1975, three native dual-purpose breeds dominated the Dutch cattle population (Figure 7.1): Dutch Friesian cattle (76%), Meuse-Rhine-Yssel cattle (22%),

and Groningen White Headed cattle (2%) [99]. Currently, more than 90% of the Dutch population consists of Holstein Friesian cattle [250]. After 1975, two additional native Dutch cattle breeds have been developed from the existing native MRY breed (Figure 7.1). These breeds were developed to conserve a breeding line with a characteristic color (Deep Red) and to develop a beef type breed (Improved Red and White). The herd books were established in 1988 and 2001, respectively.

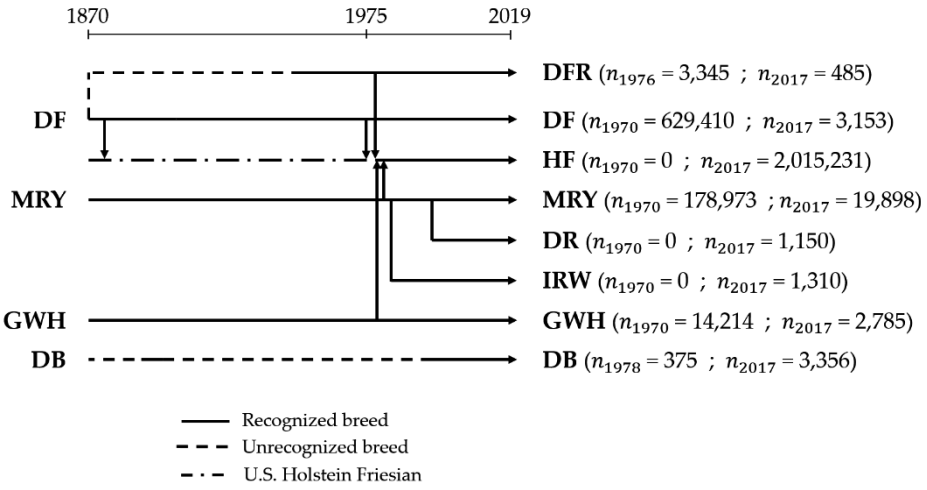


Figure 7.1 Development of Dutch cattle breeds from 1870 to 2019 and the number of cows per breed in 1970 (or in 1976 for DFR and 1978 for DB) and in 2017. Number of cows were based on [99, 251, 252]. DB: Dutch Belted; DF: Dutch Friesian; DFR: Dutch Friesian Red and White; DR: Deep Red; GWH: Groningen White Headed; IRW: Improved Red and White; MRY: Meuse-Rhine-Yssel.

Since the early 1990s, genetic material from the Holstein Friesian breed and from Dutch native cattle breeds has been stored in the Dutch gene bank. For the Holstein Friesian breed, 25 straws have been stored for each AI bull. For native breeds, storage has been less systematic and has largely depended on the availability of samples (e.g. surpluses of semen from bulls used for AI before 1990) and financial resources. In later years, genetic material for native breeds has partly been collected to increase genetic diversity in the gene bank collection, based on (limited) pedigree information. Recently, all gene bank bulls from native breeds were genotyped. This enables a genomic analysis of the stored genetic diversity, a comparison across breeds and an evaluation of the composition of the current gene bank collection.

The objectives of this study were to (1) characterize genetic diversity conserved in the Dutch gene bank for native cattle breeds and the Holstein Friesian breed, and (2) identify genetic material to set up core sets in which allelic diversity is optimized, either within or across breeds, and working sets with the remainder of material.

7.2 Material and methods

7.2.1 Data

The Dutch gene bank for livestock breeds is maintained by the Centre for Genetic Resources, the Netherlands (CGN) of Wageningen University & Research. From the CGN gene bank collections, a total of 715 bulls from seven Dutch native breeds, born between 1960 and 2015, were included in this study after data filtering (Figure 7.2). These bulls comprised all bulls from native breeds in the gene bank, except for a few bulls that were unsuccessfully genotyped. The seven breeds were Deep Red (DR), Dutch Belted (DB), Dutch Friesian (DF), Meuse-Rhine-Yssel (MRY), Friesian Red and White (DFR), Groningen White Headed (GWH), and Improved Red and White (IRW). Each bull was genotyped with the Illumina BovineSNP50 v2, Illumina BovineSNP50 v3 or Illumina BovineHD panel. After merging the different panels, 43,747 single nucleotide polymorphisms (SNPs) remained.

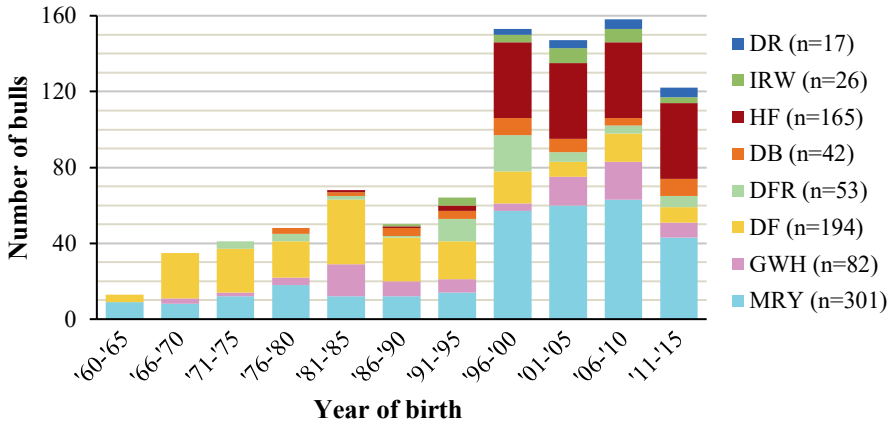


Figure 7.2 The number of bulls per breed and year of birth. DB: Dutch Belted; DF: Dutch Friesian; DFR: Dutch Friesian Red and White; DR: Deep Red; GWH: Groningen White Headed; IRW: Improved Red and White; MRY: Meuse-Rhine-Yssel.

In addition to the native breeds, a sample of 165 Holstein Friesian (HF) gene bank bulls was included to be able to determine relatedness between HF and the other breeds. This sample consisted of available genotyped HF gene bank bulls born before 2000 (n = 37), as well as 25 black and 25 red HF gene bank bulls that were randomly sampled per five-year period after 2000 (Figure 7.2). The HF bulls were genotyped with the Illumina BovineSNP50 panel (versions v1 and v2) or the CRV custom-made 60 k Illumina panel (versions v1 and v2) and imputed to 75 k (for details, see Doekes et al. [173]). After merging the HF genotypes with those of the native breeds, which was done based on the SNP identifiers, 36,779 SNPs remained. Note that the HF sample did not cover all HF bulls in the gene bank. For a more extensive analysis of the HF gene bank collection, see Doekes et al. [173].

The initial dataset, consisting of 880 bulls (735 bulls from native breeds and 165 HF bulls), was first filtered. Bulls with more than 10% missing SNPs (n = 3) were discarded. Bulls from native breeds with a HF fraction of 3/8 or more in the first three ancestral generations of their pedigree (n_{DF} = 4, n_{DFR} = 4, n_{GWH} = 6, n_{MRY}= 3) were also discarded, to prevent bulls with a high HF fraction being selected when constructing core sets for native breeds. In addition, SNPs with a call rate < 90% (n = 171) and SNPs that were non-polymorphic (n = 1,583) were removed. After filtering, 35,025 SNPs remained. The density of SNPs was rather uniform across the genome (Figure S7.1) with a mean distance between two successive SNPs of 69.08 Kb.

7.2.2 Characterization of genetic diversity within and between breeds

Genetic diversity was quantified with genetic similarities. The similarity between any two bulls was calculated as [51]:

$$SIM = \frac{\sum_{j=1}^{n_{SNP}}\left[I_{11,j} + I_{12,j} + I_{21,j} + I_{22,j}\right]}{4n_{SNP}}$$

where n_{SNP} is the number of SNPs, and I_{xy} is an indicator variable that, for the j^{th} SNP, equaled 1 when allele x of the first bull was identical to allele y of the second bull, and 0 otherwise.

Note that the mean similarity (including self-similarity) equals expected homozygosity. Consequently, one minus the mean similarity equals the expected heterozygosity, which is commonly used as diversity measure [46]. A 1% higher mean similarity thus means a 1% lower expected heterozygosity at SNP level. Genetic distances between bulls were obtained as one minus the similarity between those bulls. These distances were used to construct a neighbor-joining (NJ) tree with the package Ape in R v3.3.3 [253]. The NJ-tree was visualized in FigTree [254].

Population structure was further investigated using the variational Bayesian framework implemented in fastSTRUCTURE [255]. Advantages of this approach are that population structure is not defined *a priori* (i.e., there are no predefined entities such as breeds) and that the number of clusters can be increased one by one, such that the development of clusters can be investigated and the uppermost likely number of clusters can be identified. A simple prior approach was used. The predefined number of clusters (K) ranged from two to ten. For each K, 50 independent runs were executed. From these runs, the uppermost likely K was identified based on the change in likelihood between runs for each K, following the approach of Evanno et al. [256]. Results were visualized in R with the POPHELPER package [257].

7.2.3 Optimizing genetic diversity in collections

The mean similarity for a set of bulls was calculated following Berg and Windig [258]:

$$\overline{SIM} = \mathbf{c}'\mathbf{H}\mathbf{c}$$

where $\overline{SIM}$ is the mean similarity, $\mathbf{H}$ is the similarity matrix, and $\mathbf{c}$ is a vector of proportional contributions that sum up to one.

We applied this formula to the entire collection of native breeds (all breeds combined) as well as to individual native breeds. To determine a core set of bulls we considered three different scenarios for the contributions: (1) equal contributions, (2) current contributions, and (3) optimal contributions. In scenario (1), each bull contributed equally. This scenario was used to determine the mean similarity if from each bull the same number of straws would be set aside for the core set. In scenario (2), each bull contributed based on the current storage of straws. This scenario was used to determine the mean similarity if from each bull straws would be set aside for the core set according to the contributions in the current collection. This scenario also provided information on whether the current composition of the gene bank was already optimized, i.e., whether breeds and/or bulls that are important for the diversity have contributed more to the gene bank. In scenario (3), each bull was given an optimal contribution based on optimal contribution selection (OCS), such that the core set was constructed to minimize the mean similarity. Although OCS was originally introduced to maximize genetic progress while restricting loss of diversity [89, 241], it may also be used to optimize diversity irrespective of gain [90, 167]. In OCS, each selection candidate is assigned an optimal contribution, which can be interpreted as the percentage of offspring that would minimize the mean similarity in the next generation (in this study the contribution is the percentage of straws in the gene bank). In other words, it optimizes vector $\mathbf{c}$ to minimize the similarity. Optimal contributions were calculated with Gencont v2.0 [242].

7.3 Results

7.3.1 Characterization of genetic diversity within and between breeds

The mean genetic similarity within breeds ranged from 64.24% (HF) to 69.05% (GWH) (Table 7.1). Between breeds, the mean similarity ranged from 61.82% (between DF and HF) to 65.51% (between DF and DFR). The IRW and DR breeds, which both descend from MRY (Figure 7.1), showed a high mean similarity with MRY: 63.67% and 64.76%, respectively. IRW and DR also showed a high similarity with each other (63.64%). HF bulls were least similar to all other breeds, with a mean similarity ranging from 61.81% to 61.91%.

The NJ-tree based on genetic distances (i.e., one minus similarities) confirmed the abovementioned findings and provided insights for individual bulls (Figure 7.3).

For example, various DFR bulls clustered within DF. Similarly, DR bulls clustered within MRY. The IRW formed a distinct cluster from MRY, despite its relatively high mean similarity with MRY. The GWH and HF were most distant from other breeds.

Table 7.1 Mean genetic similarities (%) within breeds[1] (on diagonal, including self-similarities) and between breeds (below diagonal).

	DB	DF	DFR	DR	GWH	HF	IRW	MRY
DB	66.87							
DF	63.61	66.71						
DFR	63.58	65.51	66.47					
DR	63.04	62.89	63.01	65.80				
GWH	62.89	62.97	63.02	62.97	69.05			
HF	61.91	61.82	61.89	61.83	61.91	64.24		
IRW	63.04	63.06	63.11	63.64	63.02	61.75	64.69	
MRY	63.11	62.78	62.92	64.76	62.93	62.02	63.67	66.44

[1]DB: Dutch Belted; DF: Dutch Friesian; DFR: Dutch Friesian Red and White; DR: Deep Red; GWH: Groningen White Headed; HF: Holstein Friesian; IRW: Improved Red and White; MRY: Meuse-Rhine-Yssel.

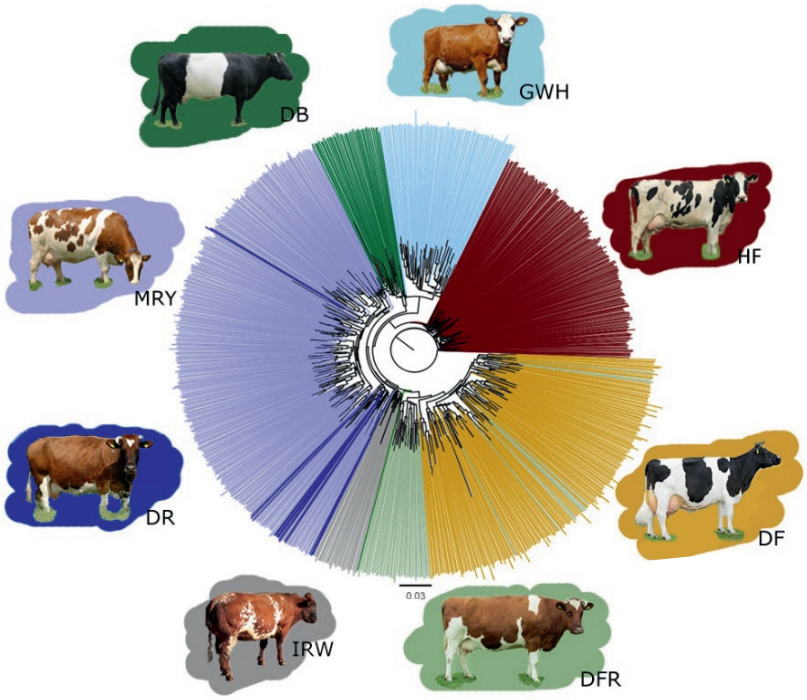


Figure 7.3 Neighbor-joining tree based on genetic distances (one minus genetic similarity) between bulls (each line is a bull). Each breed is shown with a distinct color. DB: Dutch Belted; DF: Dutch Friesian; DFR: Dutch Friesian Red and White; DR: Deep Red; GWH: Groningen White Headed; HF: Holstein Friesian; IRW: Improved Red and White; MRY: Meuse-Rhine-Yssel.

The fastSTRUCTURE results provided additional information on population admixture and breed divergence (Figure 7.4). The most likely number of clusters was K = 4, based on the change in likelihood (Figure S7.2). At K = 4, each of the three native breeds that were originally the most common breeds in the Netherlands (i.e., DF, GWH, and MRY) was part of a different cluster, and HF formed the fourth cluster. When adding additional clusters, at K = 5, DB bulls were assigned to the fifth cluster. At K = 8, however, the three remaining breeds (DR, MRY, and DFR) did not form clusters of their own. Instead, at K = 6 the cluster containing both DF and DFR bulls was split into two clusters, but these clusters did not separate the two breeds. At K = 7, a third cluster was formed in the DF–DFR cluster, and this cluster mainly contained DFR bulls, but also some DF bulls. At K = 8, MRY, IRW, and DR bulls were partly assigned to the eighth cluster, but still IRW and DR bulls did not cluster separately and shared most variation with the MRY bulls. In fact, all DR and IRW bulls were a mixture of 4 to 5 different clusters.

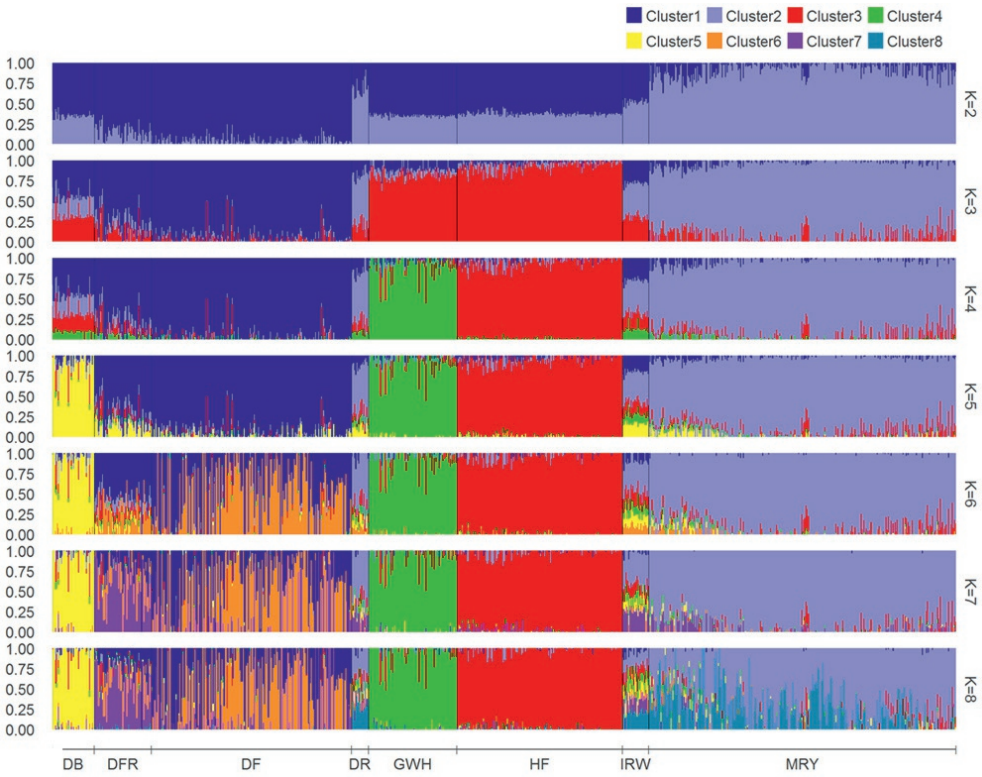


Figure 7.4 Population structure as estimated by fastSTRUCTURE divided into 2 to 8 clusters (K) for each breed (ordered alphabetically). DB: Dutch Belted; DF: Dutch Friesian; DFR: Dutch Friesian Red and White; DR: Deep Red; GWH: Groningen White Headed; HF: Holstein Friesian; IRW: Improved Red and White; MRY: Meuse-Rhine-Yssel.

7.3.2 Optimization of genetic diversity across breeds

The mean similarity of all bulls across the breeds based on equal contributions was 68.06% while similarity based on current contributions was 66.38%. When performing OCS, the mean genetic similarity further decreased to 64.78%. From the in total 715 bulls, 72 bulls received an optimal contribution higher than zero. These 72 bulls would be prioritized when the aim is to set up a core set collection that is optimized to conserve allelic diversity across native breeds.

The relative contributions of breeds differed across the three scenarios. For example, when considering equal contributions across bulls, the summed contribution was highest for MRY, namely, 42% (Figure 7.5). This is because MRY simply had the largest number of bulls used in this study (Figure 7.2). When considering the current storage of straws per bull, however, the summed contribution of MRY (25%) was lower than that of DF (of 28%). This is because the average number of straws per MRY bull was relatively low compared to the other breeds. When performing OCS, the summed contribution of MRY was even lower, namely, 16%. This suggests that, when the aim is to set up a core set to conserve allelic diversity across native breeds, the relative contribution of MRY should be lower than the current contributions in the entire collection, whereas the relative breed contributions should be increased for DB (from 2% to 10%), DF (28% to 37%), and IRW (2% to 17%), and decreased for DFR (19% to 3%), DR (6% to 4%), and GWH (18% to 13%).

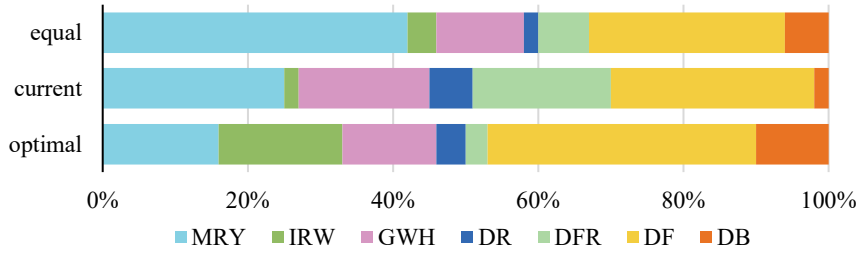


Figure 7.5 Summed contribution per native breed in the gene bank based on equal contributions across bulls, current contributions based on straws per bull, and optimal contributions based on minimizing the mean genetic similarity. DB: Dutch Belted; DF: Dutch Friesian; DFR: Dutch Friesian Red and White; DR: Deep Red; GWH: Groningen White Headed; IRW: Improved Red and White; MRY: Meuse-Rhine-Yssel.

7.3.3 Optimization of genetic diversity within breeds

Within each breed, the mean similarity based on current contributions was lower (ranging from 0.32% for MRY to 1.49% for DR) than the mean similarity for equal contributions (Table 7.2). When contributions were optimized, e.g. to set up a core set that is optimized to conserve allelic diversity within the breed, the mean similarity could be further reduced, ranging from 0.34% for DR, to 2.79% for IRW.

Table 7.2 Mean similarity (%) within breeds[1] based on equal contributions, current contributions and optimal contributions. The difference between the mean similarity for current and equal contributions, and for optimal and current contributions is also shown.

Breed	Equal	Current	Optimal	Current - Equal	Optimal - Current
DB	66.87	65.57	64.33	-1.3	-1.24
DF	66.71	65.92	63.36	-0.79	-2.56
DFR	66.47	65.24	64.77	-1.23	-0.47
DR	65.80	64.31	63.97	-1.49	-0.34
GWH	69.05	67.91	66.23	-1.14	-1.68
IRW	64.69	64.36	61.57	-0.33	-2.79
MRY	66.44	66.12	64.56	-0.32	-1.56

[1]DB: Dutch Belted; DF: Dutch Friesian; DFR: Dutch Friesian Red and White; GWH: Groningen White Headed: MRY: Meuse-Rhine-Yssel

Table 7.3 The number of selection candidates (n), the number of selected bulls (sel), and the summed contribution for selected bulls (c, in %) per birth period when performing optimal contribution selection within native breeds[1].

	DB			DF			DFR			DR			GWH			IRW			MRY		
Birth period	n	sel	c	n	sel	c	n	sel	c	n	sel	c	n	sel	c	n	sel	c	n	sel	c
1960–1969				24	4	11.9							2	1	0.6				16	5	13.9
1970–1979	3	1	0.8	44	7	10.8	6	4	9.9				5	1	2.6				29	14	28.4
1980–1989	6	6	36.9	47	9	29.5	2	1	0.6				23	6	22.8	1	1	1.1	20	6	3.5
1990–1999	10	9	26.1	45	4	16.1	28	22	63.6				10	2	9.6	6	5	57.8	64	7	11.2
2000–2009	14	11	16.2	20	4	31.6	8	7	11.1	11	11	62.7	32	14	41.6	13	13	27	123	14	22.5
2010–2015	9	6	20	10	0	0	9	8	14.8	6	6	37.3	10	7	22.7	6	6	14	51	14	20.5

[1]DB: Dutch Belted; DF: Dutch Friesian; DFR: Dutch Friesian Red and White; GWH: Groningen White Headed: MRY: Meuse-Rhine-Yssel

When optimizing contributions within native breeds, bulls from all ages were selected, with some differences across breeds (Table 7.3). For the majority of breeds, the summed contribution for bulls born before 2000 was higher than those for bulls born after 2000. More specifically, the summed contribution for bulls born before 2000 was 64% in DB, 68% in DF, 74% in DFR, 59% in IRW, and 57% in MRY. For DR, all bulls were born after 2000. For GWH, most of the contributions (64%) were assigned to bulls born after 2000. These findings suggest that old bulls harbor valuable genetic diversity not present in more recent bulls, with some differences across breeds.

7.4 Discussion

7.4.1. Characterization of cattle breeds in the Dutch gene bank

We observed that HF bulls were on average the least similar to bulls from native Dutch breeds (Table 7.1). This suggests that the native breeds harbor genetic variation that is not present in HF, and the other way around. In addition to HF, the three historically large native breeds, DF, MRY, and GWH, formed the most distinct clusters (Figures 7.3 and 7.4). The other breeds either clustered within one of the older breeds or showed clear admixture.

As expected from their breed development (Figure 7.1), IRW and DR bulls had a high mean similarity with MRY bulls (Table 7.1). The IRW bulls, however, clustered separately in the NJ tree (Figure 7.3). The distinctiveness of IRW may be explained by some influence of Belgian Blue (BBL) cattle, which have been used to improve meat quality traits in the IRW. Indeed, BBL ancestors were identified in the pedigree of a few of the IRW bulls, although the fraction of BBL was rather small with a maximum of 1/8 for the last three ancestral generations. The DR clustered only partially together with MRY, which could be explained by DR being developed more recently compared to the IRW (Figure 7.1) and DR being a dual-purpose breed like MRY.

We observed that DB had a relatively high mean similarity with DF and DFR bulls (Table 7.1), which was also observed by Eding et al. [248]. As registration for DB cattle started only in 1997, it is likely that before 1997 part of the DB cattle were upgraded by breeding DF females with DB bulls. The distinctiveness of DB is expected to be caused by random drift and the influence of American DB cattle, as various American DB ancestors were identified in the pedigree of DB gene bank bulls. The American DB is known to be partially founded by Dutch DB cattle, which were exported from the Netherlands in 1838, 1840, and 1858 [251].

Based on the multiple clusters that seemed to be present in DF cattle and MRY cattle in the fastSTRUCTURE results at K-levels above 6 (Figure 7.4), we decided to perform a principal component analysis (PCA) for DF and MRY. The PCA was based

on a genomic relationship matrix in PLINK v1.9 [259]. For DF, a distinct group of eighteen bulls was identified on the second principal component (Figure S7.3). Eleven out of the eighteen bulls in this cluster were from one of the so-called 'fundament breeders'. Since 1992, the DF breed society has applied fundament breeding, in which fundament breeders use their own bulls for breeding as much as possible [260]. The objective of this approach is to maintain genetic diversity by creating different groups of breeding animals, with each their own unique genetic diversity and a low kinship between groups. Ten breeders have been recognized as DF fundament breeder; three of which had more than one bull in the gene bank and five of which had a single bull in the gene bank. Based on the first two principal components, at least two out of three large fundament breeding groups appeared to offer unique genetic diversity within the DF breed (Figure S7.3). This finding gives incentive to make efforts to collect material from such fundament breeders for storage in the gene bank. For MRY, no visual subclusters were observed on the first two principal components (Figure S7.4).

7.4.2 Management and optimization of genetic diversity within breeds

The management and optimization of genetic diversity within breeds is relevant when conservation efforts are directed towards the conservation of individual breeds, to conserve their unique combination of alleles or genotypes. The conservation of breeds in gene banks facilitates the restoration of a breed in case of a disease outbreak or accumulation of recessive disorders due to inbreeding [261]. Within breeds, we first showed that the mean genetic similarity based on the current storage of straws was lower (0.32% to 1.49%, depending on the breed) than the mean similarity when each bull contributed equally (Table 7.2). The current composition of the gene bank appeared partly optimized in terms of genetic diversity based on pedigree information. A further reduction in the mean genetic similarity (of 0.34% to 2.79%, depending on the breed) could be achieved when using optimal contributions. Bulls with an optimal contribution larger than zero can be included in a core set for the specific breed. The number of straws stored in the core set across bulls should be in the same ratio as the optimal contributions to minimize the mean similarity in the core set. Note that this is sometimes not possible due to practical limitations (e.g. when a bull with a high optimal contribution has only few straws available). Material from bulls with an optimal contribution of zero can be relocated to a working set, where the material is accessible for, among others, the support of *in situ* populations, research, introgression of specific traits into breeds, and development of new breeds [261]. We furthermore recommend using OCS to determine which bulls from the *in situ* populations should be included to the gene

bank. This requires the analysis of genomic information (or well documented pedigree information) of both the *ex situ* and *in situ* populations.

7.4.3 Value of old and recent germplasm

We observed that both old bulls (i.e., bulls born before 2000) and recent bulls were selected when optimizing genetic diversity (Table 7.3). This suggests that also old gene bank bulls harbor valuable genetic diversity for *in situ* populations, although genomic information from *in situ* populations is required for further validation. A drawback of using old bulls is that, as a result of selection, these bulls are expected to have lower breeding values than bulls that were born more recently. Thus, by introducing their material, old bulls may increase genetic diversity at the cost of genetic merit. Previous studies have shown, however, that old gene bank bulls can be effectively used to maximize genetic merit for a given level of diversity [90, 167]. For the MRY breed, Eynard et al. [90] showed that the Dutch-Flemish total merit index (NVI) could be increased by a few points when using old gene bank bulls (born before 2000) in addition to current bulls (born after 2000). For the HF breed, Doekes et al. [167] found that the benefit of using gene bank bulls in addition to current AI-bulls depends on (1) the relative emphasis on genetic diversity and (2) the selection criterion. As expected, the relative benefit of using gene bank bulls was found to be larger when more emphasis was put on genetic diversity. Furthermore, the benefit was relatively small when selecting for the total merit index NVI, but higher when selecting for a specific index, such as fertility. Doekes et al. [173] concluded that, anticipating changes in breeding goal in the future (as result of changing environments and changing market demands), the gene bank collection is a valuable resource in terms of both genetic diversity and genetic merit.

7.4.4 Management and optimization of genetic diversity across breeds

The management and optimization of genetic diversity across breeds is relevant when conservation efforts are directed towards conservation of overall allelic diversity and not necessarily conservation of the different combinations of alleles (and thereby phenotypes). Across breeds, there seemed to be substantial overlap in the conserved genetic diversity. For example, when optimizing genetic diversity across all native breeds combined, only 72 out of 715 bulls received an optimal contribution above zero. This overlap in genetic diversity across breeds is not surprising, when considering breed history (Figure 7.1). To investigate the influence of HF on the optimal contributions of the native breeds, we also performed OCS including HF bulls. Inclusion of HF bulls lowered the contributions of native breeds,

because part of the contributions (37%) was assigned to HF bulls. The relative contributions of native breeds, however, remained very similar.

DFR bulls had a high genetic similarity to DF bulls (Table 7.1) and the optimal contribution of DFR when maximizing genetic diversity across breeds was only 1% (Figure 7.5). Based on these findings, one may question whether efforts should be made to conserve breeds like DF and DFR separately or whether they should be managed as a single breed. By managing them as a single breed, the population size increases, which could help to decrease inbreeding and drift effects.

In this study, we only considered the optimization of a single gene bank in a single country. However, there may also be overlap in the genetic diversity that is stored in gene banks worldwide. For the globally connected HF breed, for example, Danchin-Burge et al. [102] showed substantial overlap between US, French, and Dutch germplasm collections. However, they also indicated that there are various arguments in favor of this "redundancy". From a safety perspective it might be wise to have duplo-collections as are common in plant genetic resources. From a policy perspective, each country is supposed to manage its own genetic resources. From a practical point of view, germplasm stored in a national gene bank is more readily available, as exchange of animal germplasm over national borders must comply with international regulations, such as veterinary regulations and access and benefit sharing regulations under the Nagoya Protocol. However, gene bank collections are costly. Recently, de Oliveira Silva et al. [262] developed a mathematical model to optimize logistical decisions of conserving breeds in terms of economics. They evaluated alternative scenarios for reallocating genetic material currently stored in different European gene banks and showed that overall costs may be reduced by ~20% by selecting gene banks that have a relatively low combination of fixed and collection costs. Further work in this area would be valuable to economically and genetically optimize national and international gene bank collections.

Although the overlap in genetic diversity is expected to be less pronounced for native breeds compared to a transboundary breed like HF, there may still be some double-storage of genetic material across gene banks and/or countries. A major reason for this overlap might be the introgression of (e.g. transboundary) breeds into other (e.g. local) breeds. In this study, we partly corrected for the influence of HF by excluding bulls from native breeds with a HF fraction of 3/8 or more in the first three ancestral generations of their pedigree. Other breeds, however, may also have had an impact. For example, IRW bulls were found to have a fraction of BBL in their pedigree, which was likely part of the reason that IRW was assigned a relatively high contribution (17%) when performing OCS in Dutch native breeds (Figure 7.5). The fraction BBL that may be unique in the Dutch gene bank is likely to be well covered in the BBL population in Belgium. To account for influences of breeds like HF and BBL

in native cattle populations, OCS can be extended to minimize migrant contributions while maximizing genetic diversity [263]. In this extended OCS approach, a reference population (consisting of breeds that are likely to have contributed to the native breeds) is used to determine the migrant contributions. In future work, it would be valuable to consider these migrant contributions when optimizing gene bank collections.

7.4.5 From genotype to sequence data

As this study is based on 35 k SNP data, it is likely that rare variants are not considered (e.g. [264]). The effect of missing genetic variation may be stronger for local breeds than for a mainstream breed like HF, because of ascertainment bias (e.g. [265]). Sequencing costs are continuously decreasing and, as a result, increasing sequencing data will be available. This data will be extremely valuable for gene bank collections, since it will help, for example, to identify rare genetic variants that were lost over time or variants that are unique to specific breeds. The developed procedures in this study will be applicable to optimize gene banks based on sequence data.

7.5 Conclusions

Based on genotype data, bulls from native Dutch cattle breeds stored in the national gene bank are genetically distinct from the random sample of HF gene bank bulls. Old bulls (born before 2000) contribute considerably to the genetic diversity in the gene bank. Within breeds, the current collection is already partly optimized to maximize allelic diversity. Core sets could be set up using OCS based on genomic information. Across breeds, there is substantial overlap in the genetic diversity that is conserved in the gene bank. The increasing availability of genomic information and recent developments on economic modeling of gene bank collections and extension of OCS methodology may help to further optimize gene bank collections.

7.6 Acknowledgements

The research leading to these results has been conducted as part of the IMAGE project, which received funding from the European Union's Horizon 2020 Research and Innovation Program under the grant agreement n° 677353. The study was co-funded by the Dutch Ministry of Agriculture, Nature and Food Quality (KB-34-013-002). The authors gratefully acknowledge the Dutch-Flemish cattle improvement cooperative (CRV) for providing genotype data of Holstein Friesian bulls.

7.7 Supplementary information

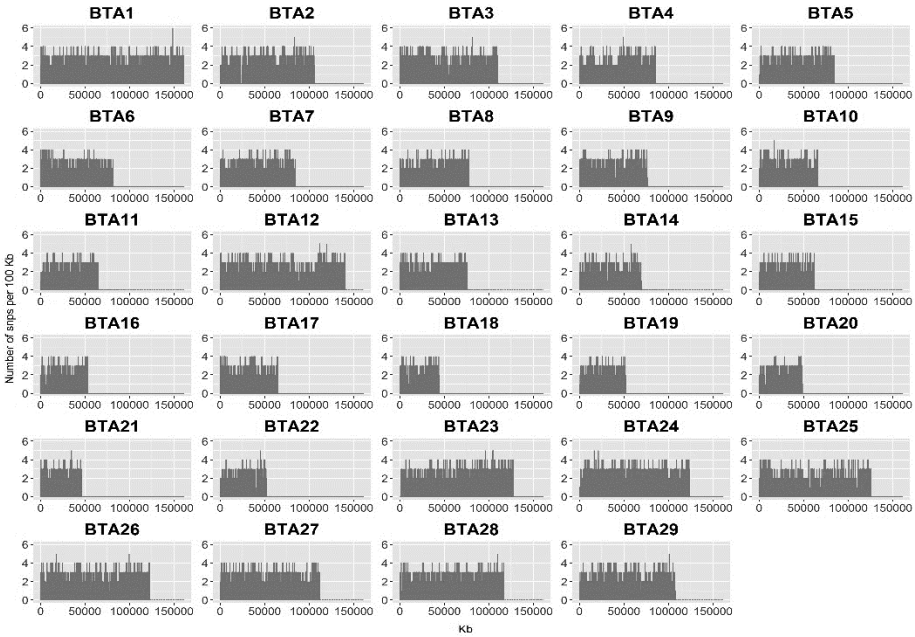


Figure S7.1 Number of SNPs per 100 Kb for each Bos Taurus Autosome (BTA)

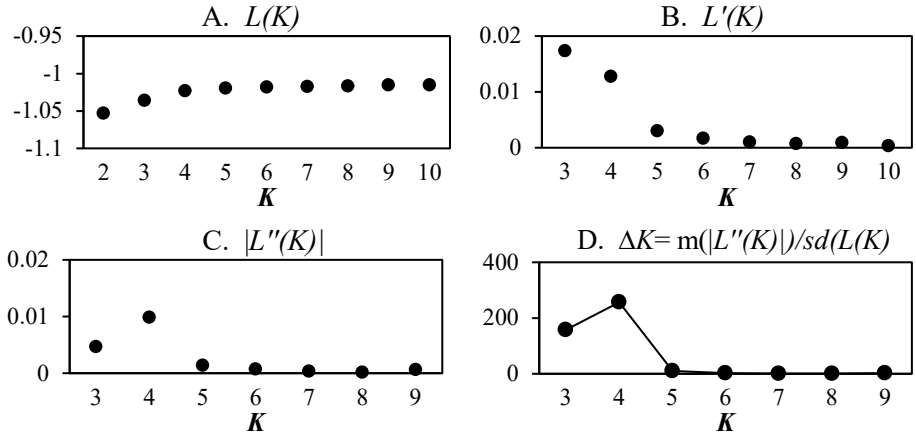


Figure S7.2 Estimation of uppermost number of clusters according to Evanno et al. [256]. A) Mean L(K) over 50 runs for each K-value. (B) Rate of change of likelihood distribution calculated as *L'(K) = L(K) – L(K– 1)*. (C) Absolute values of second order rate of change of the likelihood distribution calculated as *|L"(K)| = |L'(K + 1) – L'(K)|*. (D) ΔK calculated as *ΔK = m|L"(K)|/s[L(K)]*. The highest ΔK indicates the true K or the uppermost detected level of structure, which is here four clusters.

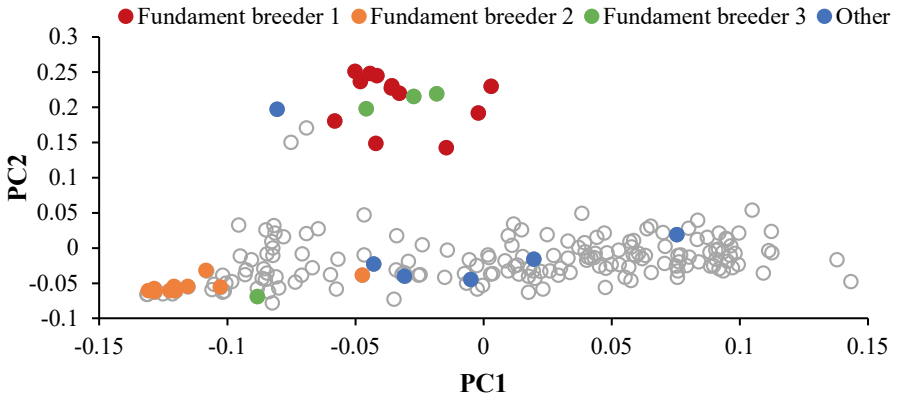


Figure S7.3 Principal component analysis within Dutch Friesian (DF) bulls showing groups of bulls from fundament breeders. The first two components, PC1 and PC2, accounted for 7.7% and 6.3% of the total variation, respectively.

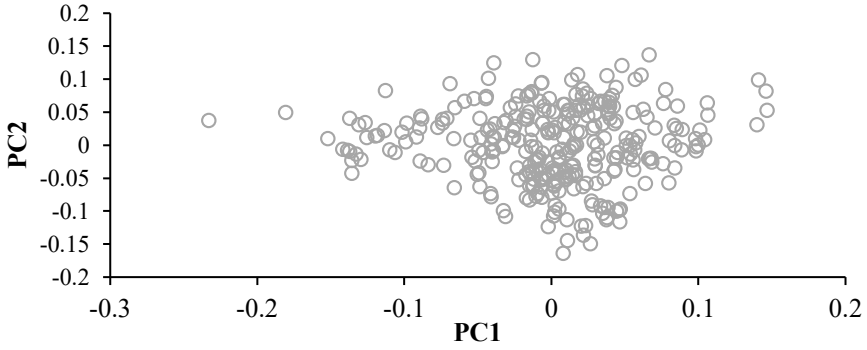


Figure S7.4 Principal component analysis within MRY bulls. The first two components, PC1 and PC2, accounted for respectively 6.5% and 6.0% of the total variation.

8

General discussion

8.1 Introduction

Recent advancements in genomic technologies have changed the way in which animals are selected and the speed at which genetic progress is realized in livestock breeding [18, 104]. At the same time, the availability of single nucleotide polymorphism (SNP) data has raised opportunities and questions regarding the characterization and conservation of genetic diversity [53, 266-268]. Before the era of genomics, there was general consensus on how to manage genetic diversity in livestock populations. Rates of pedigree-based inbreeding and kinship (ΔF and Δf) had to be limited to <1% per generation and pedigree-based optimal contribution selection (OCS) was the method of choice to do so. With the availability of genomic information, however, this consensus has somewhat disappeared.

In this general discussion, I address questions and opportunities related to conservation of genetic diversity based on genomic (in particular SNP) data. I discuss the unexpected increase in ΔF with genomic selection in the Holstein Friesian (HF) breed and address questions and opportunities related to SNP-based OCS. I also discuss the benefits of sequence data and provide perspectives on the (potential) role of gene bank collections. Last, I highlight the importance of genetic diversity for future livestock production by discussing expected changes in breeding goals.

8.2 Increase in inbreeding rate with genomic selection

Before its implementation, genomic selection (GS) was expected to reduce the ΔF (and Δf) per generation [105, 269, 270]. In Chapter 2, however, we found that the introduction of GS has been accompanied by an increase in ΔF in Dutch HF AI bulls, both per year and generation and both for pedigree-based and SNP-based measures. We later also observed this increase in a sample of genotyped Dutch HF cows, although the increase was less pronounced than for AI bulls and the ΔF in cows was still below 1% per generation with GS (Table 8.1).

Table 8.1 Inbreeding rates per year and generation (Gen) for Dutch HF AI bulls and Dutch HF cows before and after the introduction of genomic selection around 2009. The rates are shown in percentages (%). See also Doekes et al. [172]. F_{PED}: pedigree inbreeding; F_{ROH}: ROH-based inbreeding; HOM_{SNP}: SNP-by-SNP homozygosity.

		ΔF_{PED}		ΔF_{ROH}		ΔHOM_{SNP}	
Population	Period	Year	Gen	Year	Gen	Year	Gen
AI bulls	1986 - 2000	0.17	0.88	0.16	0.81	0.13	0.69
	2003 - 2009	0.02	0.11	0.03	0.15	0.05	0.27
	2009 - 2015	0.43	1.70	0.47	1.85	0.64	2.52
Cows	2003 - 2009	0.03	0.19	0.07	0.39	0.09	0.48
	2009 - 2017	0.16	0.79	0.15	0.71	0.19	0.93

8.2.1 Increase in inbreeding rate in HF populations worldwide

The faster increase in inbreeding since the implementation of GS has recently also been reported for other HF populations. For a combined data set of North-American bulls and cows, Forutan et al. [271] reported an annual increase in mean F_{PED} of 0.1% for animals born between 1990 and 2010 and of 0.3% for animals born after 2010. They also found a higher annual increase in mean F_{ROH} after implementation of GS than before GS (0.5% vs 0.1% per year). For a population of Polish HF bulls, Topolski and Jagusiak [272] estimated an annual increase in mean F_{PED} of less than 0.1% before GS (which was implemented in Poland in 2014), and of 0.6% after implementation of GS. It should be noted that the values reported by Forutan et al. [271] and Topolski et al. [272] were not calculated as inbreeding rates, because they were not expressed relative to the fraction of non-inbred loci [24]. For French HF bulls, Doublet et al. [273] found an increase in ΔF_{PED} and ΔF_{ROH} when comparing bulls from progeny-testing schemes between 2005 and 2010 with marketed bulls from GS schemes between 2012 to 2015. They estimated an increase in ΔF_{PED} from 0.5% to 1.4% per generation and an increase in ΔF_{ROH} from 0.4% to 1.4% per generation. For more than 200k North-American bulls and cows, Makanjuola et al. [274] reported similar trends as those found in Chapter 2 and by Doublet et al. [273]. They reported a ΔF_{PED} per generation of 1.3% for the 1990-1999 period, of 0.4% for the 2000-2009 period, and of 1.2% for the 2010-2018 period. For genomic measures (SNP-based and ROH-based), they reported a ΔF per generation of approximately 1.4% for the 1990-1999 period, 0.5% for the 2000-2009 period and 2.1% for the 2010-2018 period. Thus, the increase in ΔF with GS appears to be a global trend, which is not entirely surprising given the globalized market and connectedness of HF across countries [102, 275].

8.2.2 What is causing the unexpected increase in inbreeding rate?

The increase in ΔF with the implementation of genomic selection was unexpected. Although an increase in ΔF per year was anticipated, due to a decrease in generation interval [105], an increase in ΔF per generation was not [105, 269, 270]. Daetwyler et al. [105] argued that GS would reduce ΔF per generation by an earlier and better prediction of Mendelian sampling (MS) terms, and thereby a reduced co-selection of sibs as compared to traditional pedigree-based BLUP. The issue of co-selection may differ across species, because it depends on the size of full- and half-sib groups and on the information that is available to distinguish sibs. In dairy cattle, there are few full-sibs but many half-sibs. In pedigree-based BLUP, the estimation of MS terms of half-sib bulls depended on phenotypic records of the bulls' daughters. Since these records were not available at a young age, the EBVs of half-sib bulls were largely determined by the EBV of their sire and, consequently, remained similar for a long

time. With genomic information, MS terms can be predicted at a young age and half-sib bulls can be better distinguished. Therefore, it should theoretically be possible to reduce ΔF with GS while maintaining high genetic progress.

In the study of Doublet al. [273], the authors considered two national French breeds (Normande and Montbéliarde) in addition to HF. For these national breeds, no increase in ΔF per generation was found with GS, while genetic progress did increase. This finding suggests that it is not necessarily the methodology of GS that increased ΔF per generation, but rather a change in system for HF.

With GS, many more bulls are pre-screened than with traditional BLUP selection [266, 276]. In the end, however, fewer bulls are provided on the market [135, 276, 273]. The marketed AI bulls show the increase in ΔF and Δf that was observed in e.g. Chapter 2. Since these bulls are selected by farmers, the cow population is expected to follow the same trend (although less pronounced, because of additional mating programs at cow level). For North American young HF AI bulls, Miglior & Beavers [135] reported that, while the number of bulls that sires these young AI bulls has increased since GS, the number of bulls that sires 50% of young AI bulls has remained rather constant. In other words, the contributions of AI bulls to the next generation of AI bulls have remained skewed and may have become more skewed. Doublet et al. [273] compared distributions of number of offspring of HF, Normande and Montbéliarde bulls. They observed that, for all three breeds, the number of annually marketed GS bulls in 2012-2015 was lower than the number of annually progeny-tested bulls in 2005-2010 (N). For all breeds, they also found a drop in the effective number of bulls (N_e) based on variance in number of offspring. The relative drop in N_e compared to that in N, however, was larger for HF than for the national breeds. In other words, the number of progeny per bull was more skewed for HF than for the national breeds. This may partly explain the increase in ΔF for HF compared to the national breeds.

Another possible explanation for high observed ΔF with GS is the relatively slow transition to genomic control of inbreeding. While selection has been based on genomic information for over a decade now, inbreeding is still largely managed based on pedigree. From simulations it is known that pedigree-based measures will underestimate the genomic rate of inbreeding and that pedigree-based control may result in high local inbreeding peaks [91]. This could explain the higher ΔHOM_{SNP} than ΔF_{PED} with GS and the particularly high increase in F_{ROH} at specific genomic regions (Chapter 2), but it does not explain the increase in ΔF_{PED} with GS.

Another factor that may influence the increase in ΔF and Δf with GS is the composition of the reference population (see also Section 2.4). Recently, Eynard et al. [277] showed that the strategy used for updating the reference population may affect genetic gain and genetic diversity in the breeding population. Further research

is needed to better understand how the reference population should be optimized to maximize genetic gain and conserve genetic diversity in the breeding population.

Personally, I think that the increase in ΔF and Δf is also driven by the increase in expectations with GS. There are many traits of interest and there is a demand for bulls that perform well on national total merit indices, such as the Dutch-Flemish total index (NVI) or North American total index (TPI). Hence, AI companies are competing (both nationally and internationally) to provide the best bulls for their farmer clients and are limiting their products to a limited amount of diversity.

Overall, I would argue that the increase in ΔF and Δf with GS is not only a methodological issue, but also a system issue driven by expectations and demands. Hence, a solution should be sought at a system level, where commitment is needed from various stakeholders (AI companies, breeders and farmers). A key question that should be addressed is how much short-term gain the stakeholders are willing to give up to improve long-term gain and adaptability. Moreover, I recommend managing inbreeding and diversity at the genomic level, especially in genomic selection schemes (see also Section 8.3).

8.3 Optimal contribution selection based on SNP data: which relationship matrix to use?

The increasing availability of SNP data allows to move from pedigree-based OCS (POCS) to genomic OCS (GOCS). In GOCS, the pedigree-based relationship matrix ($\mathbf{A}$) in the OCS problem formulation (Box 1.7) is replaced with a genomic relationship matrix ($\mathbf{G}$). GOCS has been shown to outperform POCS for conservation of genetic diversity, if SNP density is sufficiently high [278-280]. Especially for genomic selection schemes, it is believed that genomic control of inbreeding is needed, because POCS underestimates genomic inbreeding rates [91, 138, 281, 282]. It is questioned, however, which SNP-based relationship matrix should be used. A related question is whether, with SNP data, the concept of identical-by-descent (IBD) is still relevant for conservation [58]. Below, I address these questions by discussing the use of SNP-based relationship matrices for two conservation objectives: (1) maintaining genetic variability to ensure long-term gain and adaptability, and (2) limiting inbreeding depression. I discuss these objectives separately, because it has been suggested that they may require the use of different relationship matrices [63].

Regarding objective (1), it should be noted that there is little empirical evidence of 'selection limits' being reached in selection schemes [284]. The absence of selection limits in practice can be due to many reasons, including short selection histories, changes in breeding goals over time, mutations, genotype-by-environment interactions (G x E), epistasis and epigenetics [284]. Regarding objective (2), it should be recognized that the economic losses due to inbreeding depression for single traits

are small compared to the genetic progress that has been realized [76, 157,159]. However, inbreeding has an unfavorable effect on many traits (Chapters 3 and 5), including traits that are not directly measured. Therefore, I believe that both objectives (1) and (2) are important for livestock populations.

8.3.1 Relationship matrices based on SNP array data

Different relationship matrices can be computed from SNP array data. A first matrix is the similarity matrix. This matrix consists of SNP-by-SNP similarities, which are probabilities of identical-by-state (IBS; see also Section 1.4.2). The SNP-by-SNP similarity ($SIM_{SNP_{jk}}$) between animals j and k is calculated as [51]:

$$SIM_{SNP_{jk}} = \frac{\sum_{i=1}^{n_{SNP}}(I_{11,i} + I_{12,i} + I_{21,i} + I_{22,i})}{4n_{SNP}}$$

where n_{SNP} is the total number of SNPs, $I_{xy,i}$ is an indicator variable that is 1 when allele x of animal j and allele y of animal k at the i^{th} SNP are IBS, and 0 otherwise.

A second group of relationship matrices is based on cross-products of allele counts at the SNPs. In the computation of these matrices, observed allele counts are centered by subtracting the expected allele counts at SNPs ($2p$), and scaled by dividing by the variance ($2pq$). As a result, the genomic relationship is an estimator of realized identical-by-descent (IBD) with reference to a base population with allele frequencies p and q. In VanRaden's method 1 [54], the genomic relationship (G_{jk}) between animals j and k is calculated as:

$$G_{jk} = \frac{\sum_{i=1}^{n_{SNP}}(x_{ij} - 2p_i) * (x_{ik} - 2p_i)}{\sum_i 2p_i q_i}$$

where, at the i^{th} SNP, x_{ij} is the count of allele A (coded as 0, 1 or 2) in animal j, x_{ik} is the count of allele A in animal k, p_i is the allele frequency of allele A and q_i is the allele frequency of allele B. In VanRaden's method 2 [55], G_{jk} is calculated as:

$$G_{jk} = \frac{1}{n_{SNP}} \sum_i \frac{(x_{ij} - 2p_i) * (x_{ik} - 2p_i)}{2p_i q_i}$$

which is the same as VanRaden's method 1, except that the scaling occurs per SNP (i.e. before summing across all SNPs). In Yang's method [56], G_{jk} is calculated as:

$$G_{jk} = \begin{cases} \frac{1}{n_{SNP}} \sum_i \frac{(x_{ij} - 2p_i) * (x_{ik} - 2p_i)}{2p_i q_i}, for\ j \neq k \\ 1 + \frac{1}{n_{SNP}} \sum_i \frac{x_{ij}^2 - (1 + 2p_i)x_{ij} + 2p_i^2}{2p_i q_i}, for\ j = k \end{cases}$$

which is the same as VanRaden's method 2, except that the diagonal of Yang's matrix is calculated in a different manner. Because diagonals and off-diagonals are calculated differently, Yang's matrix may be non semi-positive definite [56].

A last relationship matrix is based on IBS-segments (see also Section 1.4.3). The segment-based kinship ($f_{SEG_{jk}}$) between animals j and k is calculated as [63]:

$$f_{SEG_{jk}} = \frac{\sum_{m=1}^{n_{SEG_{ij}}} \sum_{x_i}^{2} \sum_{y_j}^{2} l_{SEG_{ij,m}}}{4 l_a}$$

where $n_{SEG_{ij}}$ is the total number of shared segments between j and k, $l_{SEG_{ij,m}}$ is the length of the m^{th} shared segment measured over homolog x of animal j and homolog y of animal k and l_a is the length of the genome covered by SNPs. The $f_{SEG_{jk}}$ is also interpreted as an estimator of realized IBD, where the base population depends on the length of the segments that is used.

8.3.2 Which matrix should be used to maintain genetic diversity and ensure long-term gain and adaptability?

To ensure long-term genetic gain and adaptability, it is important to maintain genetic variability underlying traits that are currently of interest or may become of interest in future. This implies that alleles should be conserved and genetic variance should be maintained. This can be realized by managing the expected heterozygosity across the genome. By maintaining expected heterozygosity, alleles are kept at relatively moderate allele frequencies, thereby limiting the probability that they are lost by drift. Moreover, in an additive model, expected heterozygosity at quantitative trait loci (QTL) determines the additive genetic variance of traits. This is reflected by the expression of additive genetic variance for a single biallelic QTL, which is $2pq\alpha^2$, where $2pq$ is the expected heterozygosity and α is the allele substitution effect [24].

There is a conflict between selection and maintaining expected heterozygosity. While for selection the frequency of favorable alleles should move towards fixation, for maintaining expected heterozygosity alleles should be kept at moderate frequencies. To partly resolve this conflict, it has been suggested to upweight SNPs with rare favorable alleles in genomic prediction to balance short-term and long-term gain [148, 281, 285, 286]. In addition, expected heterozygosity can be managed while performing selection, using an approach like GOCS.

It is questioned which SNP-based $\mathbf{G}$-matrix should be used in GOCS. To maintain heterozygosity at the SNPs, the similarity matrix is an obvious choice, since it directly measures the IBS status at the SNPs. To maintain heterozygosity at the rest of the genome (i.e. at unobserved loci), also other $\mathbf{G}$-matrices, which estimate realized IBD, can be considered. I agree with Powell et al. [58] that the aim of such IBD calculations

should be to capture IBS at unobserved loci. Only few studies have compared the efficiency of using different $\mathbf{G}$-matrices in GOCS for maintaining diversity at unobserved loci [63, 264]. Eynard et al. [264] applied one generation of GOCS with 277 HF bulls and compared the use of VanRaden's matrices, Yang's matrix and the similarity matrix. In their study, relationship matrices were computed from 50k SNP data and it was determined how many alleles were conserved at 16 million SNPs (obtained from sequence data). In addition, the expected heterozygosity at these 16 million SNPs was determined. It was found that the use of VanRaden's matrices conserved fewer alleles and maintained less heterozygosity compared to Yang's matrix and the similarity matrix. Yang's matrix appeared to perform best when there was no restriction on the number of selected animals, whereas the similarity matrix performed best when there was a restriction on the number of selected animals. The authors indicated that Yang's matrix might be suboptimal for conservation, because it favors animals that share common alleles. This can be seen from the formula in Section 8.3.1. For example, when for the i^{th} SNP animals j and k both carry one copy of an allele ($x_{ij} = x_{ik} = 1$) that has a p_i of 0.1 (or 0.9), the G_{jk} based on this SNP would be 3.56, whereas it would be 0.08 if p_i was 0.4 (or 0.6). Gómez-Romano [82] argued that use of the similarity matrix would drive allele frequencies to 0.5, whereas VanRaden's and Yang's matrices would aim to keep frequencies unchanged. Hence, when the objective is to maintain the *status quo* of the population, the latter matrices could be preferred, although rare alleles would be lost due to drift.

In a simulation study, De Beukelaer et al. [286] compared GOCS with VanRaden's method 1 to approaches in which they maximized a weighted index that consisted of the mean breeding value and a population diversity measure. They found that the use of IBS-status at the SNPs (their 'IND-HE' approach) maintained more heterozygosity at SNPs and unobserved loci than GOCS with VanRaden's method 1. When using GOCS with VanRaden's method 1, they observed that the constraints were not always met and showed that the increase in homozygosity at the SNPs is not simply related to the term $\mathbf{c}'\mathbf{G}\mathbf{c}$, but rather to the sum of $\mathbf{c}'\mathbf{G}\mathbf{c}$ and a second term involving cross-products of allele frequencies and their changes. This second term will likely disappear when using a similarity matrix. The authors, however, did not compare their IND-HE approach to GOCS with a similarity matrix.

In contrast to the finding of De Beukelaer et al. [286], constraints were met in the study of Sonesson et al. [91]. As argued by De Beukelaer et al. [286], this may be due to different numbers of unique IBD founder alleles that were simulated in the two studies. Further studies are needed to clarify this issue.

In another simulation study, De Cara et al. [63] compared the use of the similarity matrix and segment-based relationship matrix. They observed that the former resulted in higher heterozygosity at unobserved loci.

Overall, GOCS with a SNP-by-SNP similarity matrix appears to be an effective strategy to conserve alleles and maintain heterozygosity across the genome. In my opinion, the concept of IBD has largely lost it value for conservation with the availability of high density SNP data. However, more extensive studies (considering different SNP densities, direct comparisons of measures, multiple generations, etc.) are needed to justify the use of different **G**-matrices in GOCS for management of genetic diversity. In addition, the upweighting of rare favorable alleles in genomic prediction, potentially in combination with GOCS, should receive further attention.

8.3.3 Which matrix should be used to limit inbreeding depression?

Inbreeding depression occurs because of favorable dominance effects at QTL, which are expressed in heterozygotes (Chapters 3 and 5). Hence, to limit inbreeding depression, an intuitive objective is to limit the increase in homozygosity across the genome. As discussed in Section 8.3.1, the increase in homozygosity across the genome could be limited by using GOCS with a similarity matrix. However, not all homozygosity may be equally harmful. In Chapter 3, we found that, based on pedigree data, inbreeding on more recent ancestors may be more harmful than inbreeding on more distant ancestors. Especially when total inbreeding was split into new and ancestral components based on Kalinowski's approach [161], the new component was more strongly associated with inbreeding depression. A similar finding was also reported by other studies (e.g. [154]), even while these studies have used slightly biased estimates of Kalinowski's inbreeding coefficients (Chapter 4).

The premise that recent inbreeding is more harmful than ancient inbreeding may be an incentive to perform GOCS with a relationship matrix based on (long) IBS-segments. De Cara et al. [63] found in their simulation study that management of segment-based relationships may limit inbreeding depression better compared to management of SNP-by-SNP similarities, although the latter maintained more heterozygosity. Intuitively, this could be explained by the fact that managing SNP-by-SNP similarities tries to conserve alleles and move allele frequencies close to 0.5, also for deleterious alleles. In a sample of HF cows, Maltecca et al. [268] recently applied the approach of Druet and Gautier [185] to classify ROHs based on their expected age. They found a stronger unfavorable effect of inbreeding based on ROHs classified as 1 to 4 generations old than of inbreeding based on ROHs classified as 4 to 8 generations old [268]. Our results from Chapter 3 do not directly support the hypothesis that segment-based management would limit inbreeding depression better than SNP-by-SNP management. Namely, both ROH-based inbreeding and SNP-by-SNP homozygosity captured similar amounts of inbreeding depression at population level (Table 3.3) and no clear differences were found between effects of long and short ROHs (Figure 3.6).

The segment-based approach has several methodological drawbacks that may influence empirical results and thereby complicate the understanding of its potential benefits. First, the approach depends on haplotypes (for kinship calculations), which are not readily available from SNP data. Consequently, phasing is required [116], which may affect the results when the phasing accuracy is suboptimal. Second, it is assumed that when alleles at observed SNPs are identical, alleles at unobserved loci between the SNPs are identical as well. Especially with low SNP density, this may not be true [59, 287]. Third, to limit the number of false positive segments, many criteria can be used to define segments. In literature, a wide range of (rather arbitrary) settings is used and these settings are often poorly reported, making it almost impossible to compare results across studies [60]. Fourth, the segment-based approach is computationally intensive, since it requires scanning four combinations of homologous chromosomes for each pair of individuals. Last, segments are broken down over time and the speed with which they are broken down depends on the local recombination rate. ROH-hotspots may occur due to low local recombination rates [288, 289]. Although recombination rates in cattle [122] are relatively uniform across the genome as compared to e.g. chicken [290], it is expected that variation in recombination will influence ROH-based inference. To accurately identify IBD-segments, one should correct for differences in recombination rate, but this is almost never done in practice. It would be interesting to investigate the impact of using genetic maps, instead of physical maps, for identification of ROHs and estimation of ROH-based inbreeding and inbreeding depression.

Overall, it appears that GOCS with a similarity matrix is an effective approach to limit inbreeding depression in selection schemes. Although segment-based measures are expected to better capture inbreeding depression than SNP-by-SNP measures, empirical results are mixed. Due to the methodological drawbacks of the segment-based approach, and their seemingly poorer performance for conserving diversity (Section 8.3.1), I currently recommend the use of a SNP-by-SNP similarity matrix in GOCS to maintain genetic variability and limit inbreeding depression.

8.4 Conservation opportunities of extended genomic OCS

In addition to its value for maintaining genetic variability and limiting inbreeding depression (Section 8.3), genomic information offers opportunities for targeted conservation of genetic diversity. The traditional OCS problem, as described in Box 1.7, can be extended to include additional constraints. In this section, I first describe how constraints can be added to OCS and then discuss the use of additional constraints to (1) manage diversity at specific genomic regions, and (2) recover the original genetic background of a breed.

8.4.1 Including additional constraints in OCS

The original OCS algorithm, which is implemented in Gencont software and was used in Chapters 6 & 7, solves the OCS optimization problem with Lagrangian multipliers [89]. In this approach, an objective function $\mathrm{H_t}$ is maximized, given a vector of contributions $\mathbf{c_t}$ and Lagrangian multipliers λ_0 and $\boldsymbol{\lambda}$ [89]:

$$\mathrm{H_t} = \mathbf{c_t'EBV_t} - \left(\mathbf{c_t'A_tc_t} - 2\bar{f}_{PED,t+1}\right)\lambda_0 - (\mathbf{Q'c_t} - 1/2\ \mathbf{1})'\ \boldsymbol{\lambda}$$

where λ_0 and $\boldsymbol{\lambda}$ are Lagrangian multipliers ($\boldsymbol{\lambda}$ is a vector of two Lagrangian multipliers), $\mathbf{Q}$ is an $(n \times 2)$ incidence matrix indicating the sex of the candidates with 0's and 1's, $\mathbf{1}$ is a (2×1) vector of 1's, and the other parameters are as in Box 1.7. Drawbacks of this approach are that it does not guarantee the optimal solution (Section 6.4.2) and that it can be computationally intensive, although the latter issue can be overcome by circumventing the (repeated) inversion of the relationship matrix [291]. Another drawback is that the approach is not very flexible, i.e. adding constraints requires a complete reformulation of the optimization problem. Therefore, alternative algorithms are needed.

Pong-Wong and Woolliams [95] demonstrated how OCS can be reformulated as a semidefinite programming (SDP) problem, which can then be solved using general purpose software such as SDPA [292]. With SDP, a linear objective function is minimized, subject to a linear matrix inequality (LMI). In the context of OCS, the objective function is to maximize $\mathbf{c_t'EBV_t}$. Therefore, the minimization objective has to be multiplied by -1. The LMI is constructed from all the constraints, where quadratic constraints such as the constraint on kinship can be transformed to linear constraints with *Schur complements* [293]. The standard OCS formulation for SDP then becomes:

$$\begin{aligned}
&\text{Minimize:} && -\mathbf{c_t'EBV_t} && \text{(a)}\\
&\text{Subject to:} && \begin{bmatrix} \mathbf{A^{-1}} & \mathbf{c_t} \\ \mathbf{c_t'} & 2\bar{f}_{PED,t+1} \end{bmatrix} \geq 0 && \text{(b)}\\
& && \mathbf{c_t's} - 0.5 \geq 0 && \text{(c)}\\
& && -\mathbf{c_t's} + 0.5 \geq 0 && \text{(d)}\\
& && \mathbf{c_t'd} - 0.5 \geq 0 && \text{(e)}\\
& && -\mathbf{c_t'd} + 0.5 \geq 0 && \text{(f)}\\
& && \mathbf{c_t} \geq 0 && \text{(g)}
\end{aligned}$$

where $\mathbf{s}$ is a vector with 1s for sires and 0s for dams, $\mathbf{d}$ is a vector with 0s for sires and 1s for dams, and the rest of the parameters is the same as before. Constraint (b)

is the linear equivalent of the quadratic constraint on kinship, constraints (c), (d), (e) and (f) ensure that the contributions sum up to 0.5 per sex and constraint (g) ensures that contributions are non-negative. The LMI is constructed as a block diagonal matrix with all constraints on the diagonal, with $n + 1$ affine matrices, where n is the number of animals (for more details, see [89]). The advantages of the SDP approach are that, in contrast to the Langrangian multiplier approach, it computes the optimal contributions and allows for various constraints to be added, albeit at the cost of a more complex problem formulation. As an alternative to SDP, one may leave the strict optimization framework and maximize a weighted index containing genetic gain and inbreeding with differential evolution algorithms [138, 294].

8.4.2 Region-specific diversity management

Various studies have stressed the potential benefit of maintaining diversity in specific regions of the genome, such as the Major Histocompatibility Complex (MHC) or regions associated with inbreeding depression [82, 295, 296]. With an approach like SDP, it is technically possible to perform region-specific diversity management. For example, constraint (b) in the SDP formulation could be split into two different constraints, one with a genomic-relationship matrix for a specific region of interest, and one with a genomic-relationship matrix for the rest of the genome.

Gómez-Romano et al. [82] showed that SDP-based OCS can be used to limit the increase in SNP-by-SNP similarity at multiple target regions. Unsurprisingly, the optimisation was more successful when the targeted regions were on the same chromosome than when they were located on different chromosomes. They furthermore showed that restricting the increase in similarity at specific regions would result in an increase in similarity for the rest of the genome. This was also described by Engelsma et al. [296] and Roughsedge et al. [297]. By including a constraint on the increase in similarity for the rest of the genome, this increase could be limited, although it was still higher than when no region-specific management was applied [82].

Region-specific diversity management could also be useful when selecting for specific alleles underlying monogenic traits. An example is polledness in dairy cattle. In Europe, approximately 80% of dairy cattle is dehorned to prevent injuries among animals and increase safety for handlers [298, 299]. Dehorning, however, is an undesirable intervention [300, 301]. A promising alternative is to breed for naturally polled animals [302-304]. Polledness is a monogenic trait and the polled allele (P) is dominant over the horned allele [305]. The polled locus in HF is mapped to chromosome 1, between 1.6 and 2 Mb [305-307]. In HF, the frequency of the P-allele is low (Figure 8.1A) and the average EBV of polled bulls is lower compared to that of horned bulls (Figure 8.1B). Moreover, polled bulls are on average more related

amongst each other than horned bulls are, especially around the polled locus (Figure 8.1C). Selection for polledness without additional constraints would, therefore, result in a loss in genetic merit and an increase in similarity, especially around the polled locus. With the SDP-based OCS approach the EBV in the next generation could be maximized, while increasing the frequency of the P-allele to a target frequency, limiting the increase in similarity around the polled locus and limiting the increase in similarity across the rest of the genome. The target frequency of the P-allele could be included in the SDP formulation by adding the constraints $[\mathbf{c}_t'\mathbf{p} - \beta \geq 0]$ and $[-\mathbf{c}_t'\mathbf{p} + \beta \geq 0]$, where β is the target allele frequency in the next generation and $\mathbf{p}$ is a vector with half the number of copies of the P-allele per selection candidate. It would be interesting to investigate the effectiveness of this approach and compare it to other approaches, such as weighing polledness as trait in the total merit index (e.g. [308]) or the use of genome editing (Section 8.5.4).

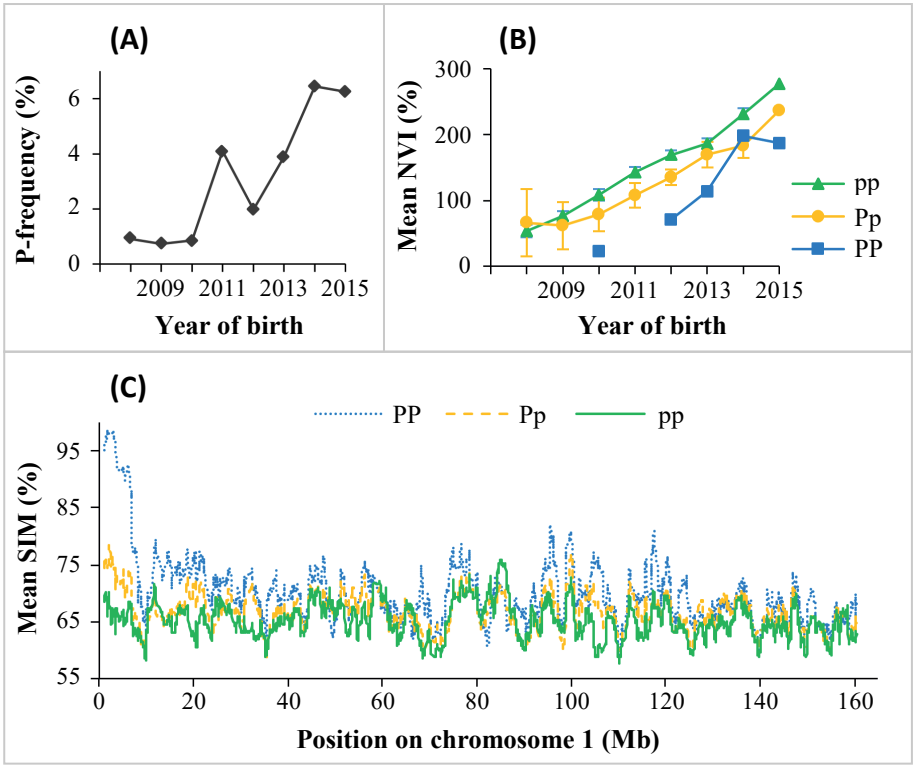


Figure 8.1 Statistics of 1443 Dutch HF AI bulls born in 2008-2015, of which 7 were homozygous for the polled allele (PP), 60 were heterozygous (Pp) and the other 1376 were horned (pp). Panel (A): frequency of the P-allele per year of birth. Panel (B): mean total merit index per year of birth. Panel (C): mean similarity across chromosome 1 as a moving average of 51 SNPs.

Region-specific diversity management with SDP has several technical drawbacks. First, constraints are not always met [82], which appears to be a more common issue with GOCS (Section 8.3.2) and should be further investigated. Second, the approach is computationally intensive, since it requires inverting a relationship matrix for each genomic region of interest (i.e. for each region for which a different constraint is set). For large data sets with multiple constraints it may, therefore, not be feasible. Last, the genomic relationship matrices need to be positive definite and invertible, which is not always the case in practice (especially not for small regions with few SNPs). This problem can be solved by adding a small constant to the diagonals of the matrices, but this may affect the outcome of the optimisation procedure.

Overall, it should be questioned whether the potential benefits of region-specific diversity management outweigh the risks of losing diversity in the rest of the genome. Results of Chapter 5 suggest that there is currently little benefit of region-specific inbreeding management in selection schemes to limit inbreeding depression. I agree with Maltecca et al. [309] that the identification of functional inbreeding is still a long-term objective and that management in the short term will mostly rely on genome-wide metrics. For a region like the MHC, I recognize that it is beneficial to have many different alleles in a population when the population is exposed to new pathogens (e.g. [146]). It can be questioned, however, whether MHC diversity is sufficiently captured by SNP data, or whether sequence data would be needed. For introgression of monogenic traits like polledness, region-specific management could also be beneficial, but this requires further investigation.

8.4.3 Recovering the genetic background of a population

An additional conservation objective could be to recover the original genetic background of a population. This may especially be relevant for local breeds that have been subject to introgression by high-productive breeds (i.e. 'upgrading') in the past. After introgression, genetic diversity in the local breed appears high due to the material that is coming from the donor breed. In Chapter 7, for example, we found substantial genetic diversity in the Dutch Improved Red and White (IRW) compared to other Dutch cattle breeds, which was likely due to the introgression of Belgian Blue in IRW. Another example is the upgrading of local cattle breeds with HF material, which happened globally in the 1980s (the 'Holsteinization'). If gene bank material is available from before the introgression occurred, then this material could be used for the recovery of the breed. If this material is not available, an alternative approach is needed. In recent years, OCS has been extended to allow for maximizing genetic gain, while conserving genetic diversity and decreasing migrant contributions [263, 310, 311]. In this approach, a constraint on migrant contributions is added to the traditional OCS formulation. Migrant contributions are determined

by investigating whether genomic segments may have been derived from donor breeds. For this purpose, a reference population of potential donor breeds is used. To accurately identify donor contributions, it is essential to know which breeds (and preferably which animals) may have been used as donors. Complete recovery is likely not possible, because of the random nature of Mendelian Sampling and recombination and because, for some genomic regions, none of the current animals will still carry the original material of the local breed.

8.5 From SNP arrays to whole genome sequence

With the on-going development of whole genome sequencing (WGS) and the decreasing costs of WGS techniques, it is anticipated that WGS data will replace SNP array data over time. The main benefit of WGS data is that it contains all variation in the genome. For example, in run 6 of the 1000 Bulls genome project, more than 40 million SNPs were identified in *Bos taurus* breeds [312].

Over the last five to ten years, many studies have investigated the benefit of WGS for genomic prediction in cattle. Results of these studies can be summarized in three main findings. First, the additional accuracy obtained by using WGS data is rather small for genomic prediction within breeds [313-316]. Second, the most promising results are obtained when preselecting informative SNPs from sequence data, e.g. based on GWAS results [314, 317]. Third, the largest gains in accuracy are found for genomic prediction across breeds [317, 318].

In addition to its potential benefits for genomic prediction, WGS offers opportunities to better characterize and conserve genetic diversity compared to SNP array data. Below, I discuss these opportunities.

8.5.1 Overcoming ascertainment bias: better capturing rare variants

SNP arrays are typically developed by sequencing a group of animals called the 'ascertainment group' [319, 320]. SNPs with a relatively high minor allele frequency (MAF) in this group are selected for the SNP array. Consequently, rare variants are not included in SNP arrays and are ignored in analyses based on these arrays. This is illustrated by the distribution of allele frequencies, which is rather uniform for SNP array data, but U-shaped for WGS data [121, 321]. Since mostly animals from large commercial breeds (such as HF) are used for array development, especially analyses of local breeds based on commercial SNP arrays are prone to ascertainment bias [322]. With WGS data, all variation is captured, allowing for a more accurate assessment of diversity across individuals and populations.

Within breeds, WGS can be used to better conserve rare variants across the genome. Eynard et al. [121] showed that the inclusion of rare variants of WGS data may result in slightly different relationship estimates, compared to the use of 50k

SNP data. They reported correlations between similarities estimated with 50k SNP data and WGS data ranging from 0.91 to 0.94. In another study, Eynard et al. [264] showed that the use of WGS in OCS can limit the loss of especially rare alleles, compared to the use of 50k SNP data. By better capturing rare variants, also *de novo* mutations are better captured. Mulder et al. [323] showed that genomic selection without using own performance data exploits mutational variance less effectively than traditional selection based on phenotypes (mass selection) or pedigree-based BLUP. With large amounts of sequenced individuals, it would be possible to screen for (un)favorable *de novo* mutations and exploit this information for breeding [323].

WGS data is also valuable to characterize rare alleles underlying phenotypic variation across breeds (see [324] for a review). It is known that various traits of interest have different genetic backgrounds across breeds (e.g. polledness [305]). Furthermore, WGS data may help to identify unique genetic variation in breeds. For example, while in Chapter 7 some breeds were hardly distinguishable based on SNP data, they may still harbor unique variation, which could be quantified with WGS data. This kind of information is valuable for prioritization purposes.

8.5.2 Identification of variation other than SNPs

WGS data allows for a more accurate detection and understanding of other types of genetic polymorphisms, such as structural variants [325, 326]. Copy number variants (CNVs) and other structural variants may play an important role in gene expression and phenotypic variation in livestock [325-328]. Bickhart et al. [329] estimated that around 3.1% of the cattle genome is copy number variable. The application of CNVs in livestock breeding and conservation, however, is still in its infancy.

8.5.3 Quantification of inbreeding and inbreeding load

With WGS data, inbreeding coefficients can be computed more accurately and the negative consequences of homozygosity across the genome can be studied in more detail. For example, short ROH can be identified more accurately [287] and the genome can be screened to identify lethal recessive alleles [230, 330]. In addition, time series of WGS could be used to study purging at a more detailed molecular level, compared to our analyses in Chapter 3. Hence, WGS can improve the understanding of functional inbreeding load.

8.5.4 Genome editing

The increasing availability of WGS data, combined with the increasing understanding of genetic variation underlying traits (which is facilitated by WGS data), may lead to the application of genome editing in livestock breeding. Since the development of

CRISPR/Cas9 technology in 2015, there has been a rapid increase in the number of publications on genome editing in livestock (for reviews, see [331-333]).

For monogenic traits, genome editing offers the opportunity to introduce or increase the frequency of the favorable allele, without the consequences of linkage drag associated with traditional introgression. Examples of edits that have been performed in cattle are those underlying double-muscling [334], polledness [335] and resistance to bovine tuberculosis [336]. Simulation studies have shown that with genome editing the frequency of e.g. the polled allele could be increased more rapidly than with conventional breeding, at reasonable increases in inbreeding [337, 338]. Bastiaansen et al. [338] concluded that the editing efficiency is an important factor that has to be considered, because it has a major impact on the required number of editing procedures and on the loss in selection response.

For complex (i.e. polygenic) traits, it has been suggested that genome editing can be used to promote alleles, i.e. PAGE [339], or to remove alleles, i.e. RAGE [340]. Personally, I do not expect that genome editing will lead to substantial benefits in genetic improvement for complex traits, at least not in the near future. This is because there is still limited knowledge of direct effects of loci, as well as of the interactions between loci and interactions across traits (e.g. antagonistic effects). The benefits of PAGE and RAGE depend on knowledge of the true causal variants. For example, Jenko et al. [339] reported an approximately 400% increase in genetic gains for PAGE compared to GBLUP, assuming that all causal variants were known and could be directly edited. Simianer et al. [341], however, first estimated SNP effects through ridge regression and then edited the markers with the largest effects, resulting in much lower additional genetic gains of approximately 12% more than GBLUP. Even if all true causal variants were known (or if the loci with largest estimated effects were used), still many edits per individual would be needed and it can be questioned whether this is feasible without side-effects (e.g. [342]). In the context of conservation, approaches such as PAGE and RAGE also pose a risk. Namely, these approaches make it attractive to quickly drive (presumably) favorable alleles to fixation, thereby losing the other allele (which may become of importance in the future). In addition to the mentioned limitations, there are various aspects including ethics, animal welfare, regulation and technical costs that should be carefully assessed before applying genome editing to livestock [331, 338, 343].

8.6 Gene banks: developments and future perspectives

As demonstrated in chapters 6 and 7, gene bank collections are valuable resources for, among others, conservation of genetic diversity. Many livestock breeds, however, have little or no material stored in gene bank collections [8]. In addition, gene bank collections are often perceived as static ‘museums’, due to their limited

exploitation and their long-term conservation objective [96]. Based on a recent survey among 51 European germplasm collections, the two most commonly mentioned objectives were to support *in situ* conservation of local/native breeds and to conserve genetic diversity as insurance in the long term, e.g. in case of breed extinction (Table 8.2).

Table 8.2 Conservation objectives of European germplasm collections. Data from a survey among European germplasm collections in 2018 [96].

Conservation objective	Percentage of collections
Support in situ conservation of local/native breeds	91%
Long-term conservation as insurance	80%
Research or genetic diversity studies	42%
Recreate breeds or breed lines lost	40%
Introduce diversity	22%
Reorient evolution/selection	11%
Develop new lines/breeds	4%

Developments in cryopreservation techniques, genomic technologies and infrastructure may help to further expand gene banks, optimize them and enhance their use. Below, I will discuss some recent insights in these areas, mostly obtained as part of the IMAGE project, and give perspectives on the future role of gene banks.

8.6.1 Expanding gene bank collections: more breeds, regular backups and different types of material

In a recent gap analysis among 15 European and 2 African countries, Leroy et al. [344] found that 15.9% of the 2,949 breeds registered in DAD-IS for these countries had material cryopreserved in gene banks, and 4.3% had sufficient material stored to reconstitute a breed (where 'sufficient' was defined as at least 25 male donors with at least *n* doses, with *n* depending on the species). The authors also observed that breeds not at risk were relatively well covered in gene bank collections compared to breeds at risk, and that transboundary breeds were better covered than local breeds. This is likely due to the relative ease of collection for transboundary breeds not at risk. It was concluded that there is a need for further expansion of gene bank collections, especially for local breeds at risk. Within the IMAGE project, it was concluded that there is a general need of expansion, also for breeds that are not (yet) at risk, and that regular backups are desired to capture changes over time [345].

Expansion of gene bank collections is restricted by limited financial resources and storage capacity. In addition, expansion depends on the availability of effective cryopreservation techniques. Here, differences across species may play an important role. For example, while cryopreservation of semen is very well developed in cattle

[18, 346], it is still more challenging in poultry due to relatively low and variable reproductive success rates following insemination with thawed semen [347]. Continuous new insights into which freezing protocol performs best, e.g. which cooling rate and which cryoprotective agent performs best, will help to optimize cryopreservation strategies per species [348-351]. In addition, while currently 99% of material in European germplasm collections consists of semen [96], also other types of materials can be cryopreserved. Developments in the vitrification of oocytes and embryo's, the cryopreservation and transplantation of gonadal tissues, and the storage of primordial germ cells offer opportunities for gene banking [351, 352]. An advantage of cryopreserving female reproductive material, in addition to semen, is that a breed could be re-established without many generations of backcrossing [353]. In my opinion, gene banks should only invest in novel cryopreservation techniques if these techniques help to conserve genetic diversity more effectively (this may differ across species). Moreover, before new techniques are implemented, ethical and regulatory aspects should be carefully considered (e.g. [354]).

8.6.2 Characterization, utilization and optimization of gene banks

Genomic characterization of gene bank collections is valuable. Among others, it helps to demonstrate the value of using gene bank material in current and future populations (e.g. Chapter 6) and to optimize gene bank collections through prioritization of donor animals (e.g. Chapter 7).

In addition to Chapters 6 & 7, various studies have demonstrated the potential benefits of genomic characterization and utilization of gene banks. Hulsegge et al. [355], for example, characterized the Dutch pig gene bank and showed that merging of commercial pig breeding lines has reduced the genetic diversity of the Landrace population in the Netherlands. The authors stressed the importance of conserving historical breeding lines in a gene bank [355]. Brekke et al. [356] evaluated genetic diversity within and between lines of the Norwegian live poultry gene bank and demonstrated how these lines contributed to genetic diversity in an international context. Dierks et al. [357] performed a case study in which they introgressed a specific trait from a gene bank collection into an *in situ* population. They introgressed the blue eggshell color from Araucana gene bank material into a White Leghorn laying line, through a marker-assisted backcrossing scheme. As a last example, Paris et al. [358] identified selection signatures from 25 years of gene bank data of the Spanish Asturiana de los Valles beef cattle breed. The authors used a method based on allele frequency trajectories [359], which could only be applied because of the availability of genomic time series data provided through the gene bank. These (and many other) case studies illustrate the value of gene bank collections for a wide range of objectives, including research and the (re)introduction of genetic diversity.

Conservation objectives and optimisation strategies may differ across gene bank collections, countries and stakeholders [96]. A common strategy of national gene banks is to conserve all national breeds, by establishing a core collection per breed that is sufficiently large to reconstitute that breed in case of an emergency. Such a strategy considers breeds as independent conservation units. From genomic analyses, however, it is known that breeds show admixture, both within and across countries (e.g. Chapter 7). It can be argued that losing a breed that was recently derived from another breed causes no major loss of genetic diversity. Therefore, it is important to balance the conservation of breeds as independent units with conservation of genetic diversity in the entire species. Tools for optimisation of genetic diversity within and across breeds are available [263, 360] and I expect that, with the increase in genomic information, these tools will become increasingly important for prioritization purposes.

Optimization of collections may also become feasible across different gene banks and countries. For this purpose, the economic optimization model of De Oliveira Silva et al. [262] could be used to minimize collection and storage costs across gene banks. This model could be extended to include a genomic component, aiming to limit overlap in genetic diversity stored across gene banks. Although I do not expect that exchange and storage of material across countries will become frequent in future, due to organizational and regulatory limitations (Section 7.4.4), I do believe that the transition towards bio-digital resource centers (Section 8.6.3) will allow for some genomic optimization across gene banks and countries.

It is expected that more and more (local) *in situ* populations and gene bank collections will be characterized in future. To facilitate genomic characterization, a cheap and globally available multi-species SNP array was recently developed [361]. The first version of this array contains 10k SNPs for cattle, pigs, chicken, horse, sheep and goats. Since this SNP array is focused on traditional breeds, analyses of genetic diversity in these breeds will be less prone to ascertainment bias compared to analyses based on conventional arrays (Section 8.5.1).

Overall, I expect that genomic characterization of gene bank collections will result in further optimization and utilization of these collections. The increasing amount of genomic (and other types of) data requires comprehensive information systems that link these data to physical gene bank collections.

8.6.3 From germplasm collections to bio-digital resource centers

The gene bank of the future is not only a physical germplasm collection, but rather a bio-digital resource center, as recently discussed by Mascher et al. [362] for plant gene banks. Bio-digital resource centers should provide detailed information on the

stored material, including genomic data, sample origin (e.g. geographical region), sample quality, freezing protocol and phenotypic information of the animal.

Currently, documentation of European germplasm collections is rather poor. In the survey of Passemard et al. [96], 95% of European germplasm collections indicated to use some sort of database, but 49% indicated that this was not more than an Excel file. Only 13% indicated to use CryoWEB [363], a dedicated web documentation system for animal gene banks. Moreover, the information that is documented is often limited. For example, sample identifier and collection date were documented for 85-90% of samples, sample quality for 68% of samples and freezing protocol for 51% of samples [96]. Hence, Passemard et al. [96] concluded that there is a great need for better documentation of the collections.

Ideally, bio-digital resource centers would be connected regionally or globally, thereby facilitating the optimization across centers and the comprehensive monitoring of stored diversity. One initiative for such a system is Animal-GRIN [364, 365], which is an information system that has been jointly developed by the United States, Brazil, and Canada. Although Animal-GRIN is not fully operational yet, users can explore the collections of all three countries, request samples, request genotypic information, view pedigrees, and compare phenotypic performance of animals in the collection. Another initiative is the European web portal that was developed as part of IMAGE [366]. This portal integrates gene bank collections with genomics data, geographical information system data, and other information generated by IMAGE. A challenge for such systems is the large amount of heterogeneous data that is distributed across gene banks, with different storage formats and different languages. In the IMAGE data portal, this challenge is addressed by (1) using an inject tool that supports gene bank managers to submit their data in a standardized way based on metadata rules, (2) storing data within the public BioSamples archive of EMBL-EBI [367], and (3) cross referencing to other gene bank and breeding databases such as DAD-IS [8]. Another European initiative is the portal of the European Genebank Network for Animal Genetic Resources (EUGENA) of ERFP [368]. The aim of this portal is to provide access to gene banks at a national level [368]. Although the information available through the IMAGE and EUGENA portals are limited so far [366, 368], I do think that the transition towards bio-digital resource centers are the future. This transition would help gene banks to become more accessible and enhance their use, thereby moving away from their static reputation.

8.7 Genetic diversity for future livestock production

Livestock production has increased substantially in the past and is expected to further increase in the future [33, 369, 370]. At the same time, production conditions, technology, market demands and societal demands change.

Consequently, breeding goals change. Genetic diversity allows to adapt to these changing breeding goals. In Chapter 6, for example, we showed how material of HF gene bank bulls is especially valuable when breeding goals change. In this last section, I discuss some expected changes in future breeding goals, focusing on dairy cattle. More comprehensive discussions are provided, among others, by Cole and VanRaden [34], Egger-Danner et al. [245], Hayes et al. [371] and Neeteson-van Nieuwenhoven et al. [372].

One trait that has recently received (renewed) interest in dairy cattle breeding is feed intake or feed efficiency [373, 374]. In the past, feed efficiency of dairy cattle has been improved through selection for milk production traits, resulting in a dilution of maintenance, i.e. in an increased portion of feed being partitioned towards milk instead of maintenance and body growth [373]. This dilution of maintenance effect, however, is expected to become less important in future [373]. Therefore, novel strategies are needed to improve digestive and metabolic efficiency, e.g. by selecting cows with low residual feed intake (RFI) [373, 374]. Traditionally, genetic evaluations for traits like RFI are costly, because they require large scale collection of feed intake and body weight data. Genomic prediction, however, has reduced the need to collect phenotypes at a large scale and allow for evaluation of many novel traits such as RFI [371, 375].

A second topic that is expected to receive more interest in future livestock breeding is climate change [34, 371, 376]. Livestock production affects climate change by emission of greenhouse gases (GHG) through various ways, including feed production and enteric fermentation by ruminants [376]. Many mitigation strategies have been proposed to reduce GHG emission from the livestock sector (see [377] for a review), including breeding for improved feed digestibility. While livestock production affects climate change, climate change also affects livestock. The International Panel on Climate Change has estimated that the global mean surface temperature in 2081-2100 will be 0.3 to 4.8 °C higher than in 1986-2005, where the range is due to different GHG emission scenario's [378]. Consequently, breeding for improved thermoregulation may become increasingly important. Australia, for example, has already launched genetic evaluations for heat tolerance in dairy cattle [379]. Genetic diversity across breeds may also play an important role here. For example, it has been suggested to introgress the SLICK haplotype, which underlies short sleek hair in tropical cattle breeds, into HF to improve thermoregulatory ability [380]. The increase in temperature also changes the geographical ranges of livestock pests such as ticks [381]. Consequently, breeding for traits like tick resistance may become important in areas where it is currently not. Emerging tools such as ecological modelling [382] and landscape genomics [383] may help to improve our understanding of the association between environments and genotypes.

A third group of traits that is anticipated to receive more emphasis in future breeding goals are animal health and welfare traits [384]. There is an increased pressure from society to increase the perceived welfare of livestock animals [385]. Examples are breeding for resilience [386], breeding against metabolic disorders in dairy cattle [387] and breeding against feather-pecking in laying hens [388].

Last, there are major developments at the food system level that may require a shift in livestock production and breeding. Currently, 77% of agricultural land is used for livestock production, including animal feed production, whereas the majority of the global calorie and protein supply is coming from plant-based foods [389]. There is an increasing awareness of the competition between the use of natural resources for, among others, food, feed, fuel, and nature conservation [372, 390-392]. In the Netherlands, the Ministry of Agriculture, Nature and Food Quality has recently adopted a vision of circular agriculture [393]. In this vision, plant biomass is the basis of the food system and should be primarily used to produce human food. Animals are primarily used to convert biomass inedible for humans into valuable food, manure and ecosystem services [393]. Hence, it may become increasingly important to breed for animals that can effectively convert by-products inedible for humans into valuable products. Here, genetic diversity across breeds (and, potentially, diversity stored in gene banks) may play an important role. For example, it has been suggested that Dutch Friesian cattle perform relatively well on grass-based systems. However, more studies are needed to determine which genetic background performs best for future livestock systems.

Overall, genetic diversity is essential to ensure that livestock systems can adapt to (un)expected changes in breeding goals, such as those described above. With this thesis, I have aimed to improve our understanding of how genomic information may be used to characterize and conserve genetic diversity in *in situ* populations and *ex situ* gene bank collections.

References

1. Woolliams JA, and Oldenbroek JK. 2017. Genetic diversity issues in animal populations in the genomic era. In: Genomic management of animal genetic diversity. Wageningen Academic Publishers, Wageningen.
2. FAO. 2015. The Second Report on the State of the World's Animal Genetic Resources for Food and Agriculture. FAO Commision on Genetic Resources for Food and Agriculture, Rome.
3. Oldenbroek JK. 2007. Introduction. In: Utilisation and conservation of animal genetic resources. Wageningen Academic Publishers, Wageningen.
4. UN. 1992. Convention on Biological Diversity. Visited October 21, 2019, *https://www.cbd.int/*
5. UN. 2017. Global indicator framework for the Sustainable Development Goals and targets of the 2030 Agenda for Sustainable Development. Visited October 21, 2019, *https://unstats.un.org/sdgs/indicators/indicators-list/*
6. FAO. 2007. The State of the World's Animal Genetic Resources for Food and Agriculture. FAO Commision on Genetic Resources for Food and Agriculture, Rome.
7. FAO. 2007. Global Plan of Action for Animal Genetic Resources and the Interlaken Declaration. FAO Commision on Genetic Resources for Food and Agriculture, Rome.
8. FAO. 2019. Domestic Animal Diversity Information System (DAD-IS): Data: Standard reports: Risk status of animal genetic resources. Visited October 22, 2019, *http://www.fao.org/dad-is/en/*
9. ERFP. 2018. About: ERFP: How do we operate: Assembly. Visited March 18, 2020, *https://www.animalgeneticresources.net*
10. CGN. 2020. Animal Genetic Resources. Visited March 18, 2020, *https://www.wur.nl/en/Research-Results/Statutory-research-tasks/Centre-for-Genetic-Resources-the-Netherlands-1/Animal-Genetic-Resources.htm*
11. FAO. 2013. In vivo conservation of animal genetic resources. FAO Commision on Genetic Resources for Food and Agriculture, Rome.
12. WHFF. 2019. Documentation: Statistics: 2018 Annual Statistics Report - World. Visited March 18, 2020, *https://www.whff.info/documentation/statistics.php*
13. CGN. 2019. Rassenlijst Nederlandse landbouwhuisdierrassen en hun risicostatus op basis van aantal volwassen vrouwelijke dieren in Nederland - september 2019. Visited March 18, 2020, *https://www.wur.nl/web/show/id=872562/langid=43*
14. Larson G, and Fuller DQ. 2014. The evolution of animal domestication. Annu Rev Ecol Evol Syst. 45:115-136.
15. Felius M, Beerling ML, Buchanan DS, Theunissen B, Koolmees PA, and Lenstra JA. 2014. On the history of cattle genetic resources. Diversity. 6(4):705-750.
16. Felius M, Theunissen B, and Lenstra J. 2015. Conservation of cattle genetic resources: the role of breeds. J Agric Sci. 153(1):152-162.
17. Foote R. 2010. The history of artificial insemination: Selected notes and notables. J Anim Sci. 80(E-suppl_2):1-10.
18. Fleming A, Abdalla EA, Maltecca C, and Baes CF. 2018. Invited review: Reproductive and genomic technologies to optimize breeding strategies for genetic progress in dairy cattle. Arch Anim Breed. 61(1):43.
19. Taberlet P, Valentini A, Rezaei H, Naderi S, Pompanon F, Negrini R, and Ajmone-Marsan P. 2008. Are cattle, sheep, and goats endangered species? Mol Ecol. 17(1):275-284.

20. Biscarini F, Nicolazzi EL, Stella A, Boettcher PJ, and Gandini G. 2015. Challenges and opportunities in genetic improvement of local livestock breeds. Front Genet. 6:33.
21. Peters P, and Brat I. 2015. A Breeder Apart: The Bull Who Sired 500,000 Is Gone - Fans Commemorate 'Toystory,' a Dairy Legend With a Ravenous Libido. Wall Street Journal, Eastern edition, New York.
22. van Arendonk JAM, Groen AF, and Brascamp EW. 1995. Invloed en risico van topstier : hoe groot moet de taart voor Sunny Boy zijn? Veeteelt. 12:648 - 649.
23. de Jong G (Head of Animal Evaluation Unit, CRV). 2020. Personal communication April 16, 2020.
24. Falconer DS, and Mackay TFC. 1996. Introduction to quantitative genetics, 4th edn. Longman Group Ltd, Harlow.
25. Hazel LN. 1943. The genetic basis for constructing selection indexes. Genetics. 28(6): 476-490.
26. Henderson CR. 1953. Estimation of variance and covariance components. Biometrics. 9(2):226-252.
27. Henderson CR. 1984. Applications of linear models in animal breeding. University of Guelph, Guelph.
28. Henderson CR. 1975. Best linear unbiased estimation and prediction under a selection model. Biometrics. 31:423-447.
29. Meuwissen THE, and Hayes BJ, Goddard ME. 2001. Prediction of total genetic value using genome-wide dense marker maps. Genetics. 157(4):1819-1829.
30. Hayes BJ, Bowman PJ, Chamberlain AJ, Goddard ME. 2009. Invited review: Genomic selection in dairy cattle: Progress and challenges. J Dairy Sci. 92(2):433-443.
31. Knol EF, Nielsen B, and Knap PW. 2016. Genomic selection in commercial pig breeding. Anim Front. 6(1):15-22.
32. Wolc A, Kranis A, Arango J, Settar P, Fulton J, O'Sullivan N, Avendano A, Watson K, Hickey J, and de los Campos G. 2016. Implementation of genomic selection in the poultry industry. Anim Front. 6(1):23-31.
33. Thornton PK. Livestock production: recent trends, future prospects. Philos Trans R Soc Lond B Biol Sci. 2010;365(1554):2853-2867.
34. Cole J, and VanRaden P. 2018. Symposium review: Possibilities in an age of genomics: The future of selection indices. J Dairy Sci. 101(4):3686-3701.
35. Dawkins M, and Layton R. 2012. Breeding for better welfare: genetic goals for broiler chickens and their parents. Anim Welf. 21(2):147.
36. Merks J, Mathur P, and Knol E. 2012. New phenotypes for new breeding goals in pigs. Animal. 6(4):535-543.
37. Miglior F, Muir BL, and van Doormaal BJ. 2005. Selection indices in Holstein cattle of various countries. J Dairy Sci. 88(3):1255-1263.
38. Charlesworth B, and Charlesworth D. 2010. Elements of Evolutionary Genetics. Roberts and Company Publishers, Greenwood Village.
39. Hill WG. 2000. Maintenance of quantitative genetic variation in animal breeding programmes. Livest Prod Sci. 63(2):99-109.
40. Harland C, Durkin K, Artesi M, Karim L, Cambisano N, Deckers M, Tamma N, Mullaart E, Coppieters W, and Charlier C. 2018. Rate of de novo mutation in dairy cattle and

potential impact of reproductive technologies. Proc 11th World Congr Genet Applied to Livest Prod. Auckland, New Zealand.
41. Liu H, Sørensen AC, Meuwissen THE, and Berg P. 2014. Allele frequency changes due to hitch-hiking in genomic selection programs. Genet Sel Evol. 46:8.
42. Barton NH. 2000. Genetic hitchhiking. Philos Trans R Soc Lond B Biol Sci. 355:1553-62.
43. Stapley J, Feulner PG, Johnston SE, Santure AW, and Smadja CM. 2017. Variation in recombination frequency and distribution across eukaryotes: patterns and processes. Philos Trans R Soc Lond B Biol Sci. 372(1736):20160455.
44. Lenstra J, Groeneveld L, Eding H, Kantanen J, Williams J, Taberlet P, Nicolazzi E, Sölkner J, Simianer H, and Ciani E. 2012. Molecular tools and analytical approaches for the characterization of farm animal genetic diversity. Anim Genet. 43(5):483-502.
45. Toro MA, Fernández J, and Caballero A. 2009. Molecular characterization of breeds and its use in conservation. Livest Sci. 120(3):174-195.
46. Nei M. 1973. Analysis of gene diversity in subdivided populations. Proc Natl Acad Sci USA. 70(12):3321-3323.
47. Wright S. 1922. Coefficients of inbreeding and relationship. Amer Nat. 56:330-338.
48. Oliehoek PA, and Bijma P. 2009. Effects of pedigree errors on the efficiency of conservation decisions. Genet Sel Evol. 41:9.
49. Hill WG, and Weir BS. 2011. Variation in actual relationship as a consequence of Mendelian sampling and linkage. Genet Res (Camb). 93(1):47-64.
50. Woolliams JA, Berg P, Dagnachew BS, and Meuwissen THE. 2015. Genetic contributions and their optimization. J Anim Breed Genet. 132(2):89-99.
51. Malécot G. 1948. Mathématiques de l'Hérédité. Masson & Cie, Paris.
52. Engelsma KA, Veerkamp RF, Calus MPL, Bijma P, and Windig JJ. 2012. Pedigree- and marker-based methods in the estimation of genetic diversity in small groups of Holstein cattle. J Anim Breed Genet. 129:195-205.
53. Howard JT, Pryce JE, Baes C, and Maltecca C. 2017. Invited review: Inbreeding in the genomics era: Inbreeding, inbreeding depression, and management of genomic variability. J Dairy Sci. 100(8):6009-6024.
54. VanRaden PM. 2008. Efficient methods to compute genomic predictions. J Dairy Sci. 91(11):4414-4423.
55. VanRaden PM, Olson KM, Wiggans GR, Cole JB, and Tooker ME. 2011. Genomic inbreeding and relationships among Holsteins, Jerseys, and Brown Swiss. J Dairy Sci. 94:5673-82.
56. Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, Nyholt DR, Madden PA, Heath AC, Martin NG, and Montgomery GW. 2010. Common SNPs explain a large proportion of the heritability for human height. Nat Genet. 42(7):565-569.
57. Toro MA, García-Cortés LA, and Legarra A. 2011. A note on the rationale for estimating genealogical coancestry from molecular markers. Genet Sel Evol. 43:27.
58. Powell JE, Visscher PM, and Goddard ME. 2010. Reconciling the analysis of IBD and IBS in complex trait studies. Nat Rev Genet. 11(11):800-805.
59. Ferenčaković M, Sölkner J, and Curik I. 2013. Estimating autozygosity from high-throughput information: effects of SNP density and genotyping errors. Genet Sel Evol. 45:42.

60. Meyermans R, Gorssen W, Buys N, and 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 Genom. 21:1-14.
61. Hillestad B. 2015. PhD-thesis: Inbreeding determined by the amount of homozygous regions in the genome. Norwegian University of Life Sciences, Ås.
62. McQuillan R, Leutenegger AL, Abdel-Rahman R, Franklin CS, Pericic M, Barac-Lauc L, Smolej-Narancic N, Janicijevic B, Polasek O, and Tenesa A. 2008. Runs of homozygosity in European populations. Am J Hum Genet. 83(3):359-372.
63. de Cara MÁR, Villanueva B, Toro MÁ, and Fernández J. 2013. Using genomic tools to maintain diversity and fitness in conservation programmes. Mol Ecol. 22(24):6091-6099.
64. Legarra A, Aguilar I, and Misztal I. 2009. A relationship matrix including full pedigree and genomic information. J Dairy Sci. 92:4656-4663.
65. Meuwissen THE, and Goddard ME. 2001. Prediction of identity by descent probabilities from marker-haplotypes. Genet Sel Evol. 33(6):605.
66. Legarra A, Christensen OF, Vitezica ZG, Aguilar I, and Misztal I. 2015. Ancestral relationships using metafounders: finite ancestral populations and across population relationships. Genetics. 200(2):455-468.
67. Christensen OF. 2012. Compatibility of pedigree-based and marker-based relationship matrices for single-step genetic evaluation. Genet Sel Evol. 44(37).
68. Leroy G. 2014. Inbreeding depression in livestock species: review and meta-analysis. Anim Genet. 45(5):618-628.
69. Darwin CR. 1876. The effects of cross and self fertilisation in the vegetable kingdom. John Murray, London.
70. Barczak E, Wolc A, Wójtowski J, Slósarz P, and Szwaczkowski T. 2009. Inbreeding and inbreeding depression on body weight in sheep. J Anim Feed Sci. 18(1):42-50.
71. McParland S, Kearney J, Rath M, and Berry D. 2007. Inbreeding effects on milk production, calving performance, fertility, and conformation in Irish Holstein-Friesians. J Dairy Sci. 90(9):4411-4419.
72. Silió L, Rodríguez M, Fernández A, Barragán C, Benítez R, Óvilo C, and Fernández A. 2013. Measuring inbreeding and inbreeding depression on pig growth from pedigree or SNP-derived metrics. J Anim Breed Genet. 130(5):349-360.
73. Ferenčaković M, Sölkner J, Kapš M, and Curik I. 2017. Genome-wide mapping and estimation of inbreeding depression of semen quality traits in a cattle population. J Dairy Sci. 100(6):4721-4730.
74. Saura M, Fernández A, Varona L, Fernández AI, de Cara MÁR, Barragán C, and Villanueva B. 2015. Detecting inbreeding depression for reproductive traits in Iberian pigs using genome-wide data. Genet Sel Evol. 47(1):1.
75. Sevinga M, Vrijenhoek T, Hesselink J, Barkema H, and Groen A. 2004. Effect of inbreeding on the incidence of retained placenta in Friesian horses. J Anim Sci. 82(4):982-986.
76. Sørensen AC, Madsen P, Sørensen MK, and Berg P. 2006. Udder health shows inbreeding depression in Danish Holsteins. J Dairy Sci. 89(10):4077-4082.
77. Speed D, and Balding DJ. 2015. Relatedness in the post-genomic era: is it still useful? Nat Rev Genet. 16(1):33-44.
78. Browning SR. 2008. Estimation of pairwise identity by descent from dense genetic marker data in a population sample of haplotypes. Genetics. 178(4):2123-2132.

79. Hedrick PW, and Garcia-Dorado A. 2016. Understanding inbreeding depression, purging, and genetic rescue. Trends Ecol. Evol. 31(12):940-952.
80. Pryce JE, Haile-Mariam M, Goddard ME, and Hayes BJ. 2014. Identification of genomic regions associated with inbreeding depression in Holstein and Jersey dairy cattle. Genet Sel Evol. 46(1):71.
81. Marras G, Howard J, Martin P, Fleming A, Alves K, Makanjuola B, Schenkel F, Miglior F, Maltecca C, and Baes CF. 2018. Identification of unfavourable homozygous haplotypes associated with with milk and fertility traits in Holsteins. Proc 11th World Congr Genet Applied to Livest Prod. Auckland, New Zealand.
82. Gómez-Romano F, Villanueva B, Fernández J, Woolliams JA, and Pong-Wong R. 2016. The use of genomic coancestry matrices in the optimisation of contributions to maintain genetic diversity at specific regions of the genome. Genet Sel Evol. 48(1):2.
83. Freeman R. 1982. The Darwin family. Biol J Linnean Soc. 17(1):9-21.
84. Berra TM, Alvarez G, and Ceballos FC. 2010. Was the Darwin/Wedgwood dynasty adversely affected by consanguinity? BioScience. 60(5):376-383.
85. Álvarez G, Ceballos FC, and Berra TM. 2015. Darwin was right: inbreeding depression on male fertility in the Darwin family. Biol J Linnean Soc. 114(2):474-483.
86. Sonesson AK, and Meuwissen THE. 2000. Mating schemes for optimum contribution selection with constrained rates of inbreeding. Genet Sel Evol. 32(3):1-18.
87. Windig J, and Oldenbroek K. 2015. Genetic management of Dutch golden retriever dogs with a simulation tool. J Anim Breed Genet. 132(6):428-440.
88. van Doormaal BJ, Miglior F, Kistemaker G, and Brand P. 2005. Genetic diversification of the Holstein breed in Canada and internationally. Interbull Bulletin. Uppsala, Sweden.
89. Meuwissen THE. 1997. Maximizing the response of selection with a predefined rate of inbreeding. J Anim Sci. 75(4):934-940.
90. Eynard SE, Windig JJ, Hulsegge I, Hiemstra SJ, and Calus MPL. 2018. The impact of using old germplasm on genetic merit and diversity—A cattle breed case study. J Anim Breed Genet. 135(4):1-12.
91. Sonesson AK, Woolliams JA, and Meuwissen THE. 2012. Genomic selection requires genomic control of inbreeding. Genet Sel Evol. 44(1):27.
92. James JW, and McBride G. 1958. The spread of genes by natural and artificial selection in closed poultry flock. J Genet. 56(1):55-62.
93. Wray NR, and Thompson R. 1990. Prediction of rates of inbreeding in selected populations. Genet Res. 55(01):41-54.
94. Berg P, Nielsen J, and Sørensen MK. 2006. EVA: Realized and predicted optimal genetic contributions. Proc 8th World Congr Genet Applied to Livest Prod. Belo Horizonte, Brazil.
95. Pong-Wong R, and Woolliams JA. 2007. Optimisation of contribution of candidate parents to maximise genetic gain and restricting inbreeding using semidefinite programming. Genet Sel Evol. 39(1):1.
96. Passemard A, Joly L, Duclos D, and Danchin-Burge C. 2018. Inventory and mapping of European animal genetic collections. IDELE, IMAGE (Innovative Management of Animal Genetic Resources) project report.
97. Hiemstra SJ, Martyniuk E, Duchev Z, and Begemann F. 2014. European Gene Bank Network for Animal Genetic Resources (EUGENA). Proc 10th World Congr Genet Applied to Livest Prod. Vancouver, Canada.

98. Theunissen B. 2008. Breeding without Mendelism: theory and practice of dairy cattle breeding in the Netherlands 1900-1950. J Hist Biol. 41(4):637-676.
99. Central Milkrecording Service the Netherlands. 1971. Jaarverslag. CMD, Arnhem.
100. Coöperatie CRV. 2019. Jaarstatistieken 2019. Visited March 25, 2020, *https://www.cooperatie-crv.nl/downloads/stamboek/bedrijven-en-koeien-in-cijfers/*
101. de Roos APW, Hayes BJ, Spelman RJ, and Goddard ME. 2008. Linkage disequilibrium and persistence of phase in Holstein-Friesian, Jersey and Angus cattle. Genetics. 179(3):1503-1512.
102. Danchin-Burge C, Hiemstra SJ, and Blackburn H. 2011. Ex situ conservation of Holstein-Friesian cattle: Comparing the Dutch, French, and US germplasm collections. J Dairy Sci. 94(8):4100-4108.
103. Meuwissen T, Hayes B, and Goddard M. 2013. Accelerating improvement of livestock with genomic selection. Annu Rev Anim Biosci. 1:221-237.
104. García-Ruiz A, Cole JB, VanRaden PM, Wiggans GR, Ruiz-López FJ, and van Tassell CP. 2016. Changes in genetic selection differentials and generation intervals in US Holstein dairy cattle as a result of genomic selection. Proc Natl Acad Sci USA. 113(28):E3995-E4004.
105. Daetwyler HD, Villanueva B, Bijma P, and Woolliams JA. 2007. Inbreeding in genome-wide selection. J Anim Breed Genet. 124(6):369-376.
106. Heidaritabar M, Vereijken A, Muir WM, Meuwissen THE, Cheng H, Megens HJ, Groenen MAM, and Bastiaansen JWM. 2014. Systematic differences in the response of genetic variation to pedigree and genome-based selection methods. Heredity (Edinb.). 113:503-13.
107. Stachowicz K, Sargolzaei M, Miglior F, and Schenkel FS. 2011. Rates of inbreeding and genetic diversity in Canadian Holstein and Jersey cattle. J Dairy Sci. 94:5160-75.
108. Rodríguez-Ramilo ST, Fernández J, Toro MA, Hernández D, and Villanueva B. 2015. Genome-wide estimates of coancestry, inbreeding and effective population size in the Spanish Holstein population. PLoS ONE. 10(4):e0124157.
109. Smith JM, and Haigh J. 1974. The hitch-hiking effect of a favourable gene. Genet Res. 23(1):23-35.
110. Kleinman-Ruiz D, Villanueva B, Fernández J, Toro MA, García-Cortés LA, and Rodríguez-Ramilo ST. 2016. Intra-chromosomal estimates of inbreeding and coancestry in the Spanish Holstein cattle population. Livest Sci. 185:34-42.
111. Fisher RA. 1954. A fuller theory of "junctions" in inbreeding. Heredity. 8:187-97.
112. Bosse M, Megens HJ, Madsen O, Crooijmans RPMA, Ryder OA, Austerlitz F, Groenen M, and de Cara MA. 2015. Using genome-wide measures of coancestry to maintain diversity and fitness in endangered and domestic pig populations. Genome Res. 25:970-81.
113. Keller MC, Visscher PM, and Goddard ME. 2011. Quantification of inbreeding due to distant ancestors and its detection using dense single nucleotide polymorphism data. Genetics. 189:237-49.
114. CDN. 2017. Animal Query. Visited March 17, 2017, *www.cdn.ca/query/individual.php*
115. Druet T, Schrooten C, and de Roos APW. 2010. Imputation of genotypes from different single nucleotide polymorphism panels in dairy cattle. J Dairy Sci. 93:5443-54.

116. Browning SR, and Browning BL. 2007. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. Am J Hum Genet. 81:1084-97.
117. Druet T, and Georges M. 2010. A hidden Markov model combining linkage and linkage disequilibrium information for haplotype reconstruction and quantitative trait locus fine mapping. Genetics. 184:789-98.
118. Calus MPL, and Vandenplas J. 2013. Calc_grm-a programme to compute pedigree, genomic, and combined relationship matrices. Wageningen University & Research Animal Breeding and Genomics, Wageningen.
119. Sargolzaei M, Iwaisaki H, and Colleau JJ. 2005. A fast algorithm for computing inbreeding coefficients in large populations. J Anim Breed Genet. 122:325-31.
120. Colleau JJ. 2002. An indirect approach to the extensive calculation of relationship coefficients. Genet Sel Evol. 34:409-21.
121. Eynard SE, Windig JJ, Leroy G, van Binsbergen R, and Calus MP. 2015. The effect of rare alleles on estimated genomic relationships from whole genome sequence data. BMC Genet. 16(1):24.
122. Ma L, O'Connell JR, VanRaden PM, Shen B, Padhi A, Sun C, Bickhart DM, Cole JB, Null DJ, Liu GE, Da Y, and Wiggans GR. 2015. Cattle sex-specific recombination and genetic control from a large pedigree analysis. PLoS Genet. 11:e1005387.
123. Pérez-Enciso M. 1995. Use of the uncertain relationship matrix to compute effective population size. J Anim Breed Genet. 112:327-32.
124. Kim ES, Cole JB, Huson H, Wiggans GR, van Tassell CP, Crooker BA, Liu G, Da Y, and Sonstegard TS. 2013. Effect of artificial selection on runs of homozygosity in US Holstein cattle. PLoS One. 8(11):e80813.
125. MacCluer JW, VandeBerg JL, Read B, and Ryder OA. 1986. Pedigree analysis by computer simulation. Zoo biology. 5:147-160.
126. Hu ZL, Park CA, and Reecy JM. 2016. Developmental progress and current status of the Animal QTLdb. Nucleic Acids Res. 44:D827-33.
127. Sørensen AC, Sørensen MK, and Berg P. 2005. Inbreeding in Danish dairy cattle breeds. J Dairy Sci. 88:1865-72.
128. Hanslik S, Harr B, Brem G, and Schlötterer C. 2000. Microsatellite analysis reveals substantial genetic differentiation between contemporary New World and Old World Holstein Friesian populations. Anim Genet. 31:31-8.
129. Leitch HW. 1994. Comparison of international selection indices for dairy cattle breeding. Interbull Bulletin. Ottawa, Canada.
130. de Graaf FM. 1989. Stiersom, kombinatie van produktie- en exterieurvererving. Veeteelt. 6:652-3.
131. de Jong G, Harbers A, Hamming I, Vollema AR, and van der Beek S. 1999. Fokkerijrevolutie: DPS lost Inet af: fokstieren vanaf augustus gerangschikt op unieke totaalwaarde: duurzame-prestatiesom. Veeteelt. 16:680-2.
132. van Drie I. 2007. Nvi vervangt dps: nieuwe totaalindex weegt behalve productie en gezondheid ook exterieur. Veeteelt. 24:36-8.
133. GES. 2015. Kengetallen NVI. In: Handboek Kwaliteit. Visited June 13, 2017, *www.gesfokwaarden.eu/nl/fokwaarden/pdf/E_20.pdf*

134. Meuwissen THE, and Oldenbroek JK. 2017. Genetic diversity in small in vivo populations. In: Genomic management of animal genetic diversity. Wageningen Academic Publishers, Wageningen.
135. Miglior F, and Beavers L. 2014. Genetic diversity and inbreeding: before and after genomics. Visited June 28, 2017, *www.progressivedairy.com/topics/a-i-breeding/genetic-diversity-and-inbreeding-before-and-after-genomics*
136. Pszczola M, Strabel T, Mulder HA, and Calus MPL. 2012. Reliability of direct genomic values for animals with different relationships within and to the reference population. J Dairy Sci. 95:389-400.
137. Habier D, Tetens J, Seefried FR, Lichtner P, and Thaller G. 2010. The impact of genetic relationship information on genomic breeding values in German Holstein cattle. Genet Sel Evol. 42:5.
138. Clark SA, Kinghorn BP, Hickey JM, and van der Werf JHJ. 2013. The effect of genomic information on optimal contribution selection in livestock breeding programs. Genet Sel Evol. 45:44.
139. Rodríguez-Ramilo ST, García-Cortés LA, and de Cara MÁR. 2015. Artificial selection with traditional or genomic relationships: consequences in coancestry and genetic diversity. Front Genet. 6:127.
140. de Cara MÁR, Villanueva B, Toro MÁ, and Fernández J. 2013. Purging deleterious mutations in conservation programmes: combining optimal contributions with inbred matings. Heredity (Edinb.). 110:530-7.
141. Ballou JD. 1997. Ancestral inbreeding only minimally affects inbreeding depression in mammalian populations. J Hered. 88:169-78.
142. Hinrichs D, Meuwissen THE, Ødegard J, Holt M, Vangen O, and Woolliams JA. 2007. Analysis of inbreeding depression in the first litter size of mice in a long-term selection experiment with respect to the age of the inbreeding. Heredity (Edinb.). 99:81-8.
143. Purfield DC, Berry DP, McParland S, and Bradley DG. 2012. Runs of homozygosity and population history in cattle. BMC Genet. 13:70.
144. Qanbari S, Pimentel EC, Tetens J, Thaller G, Lichtner P, Sharifi AR, and Simianer H. 2010. A genome-wide scan for signatures of recent selection in Holstein cattle. Anim Genet. 41:377-89.
145. Glick G, Shirak A, Uliel S, Zeron Y, Ezra E, Seroussi E, Ron M, and Weller JI. 2012. Signatures of contemporary selection in the Israeli Holstein dairy cattle. Anim Genet. 43:45-55.
146. Ellis SA, and Hammond JA. 2014. The functional significance of cattle major histocompatibility complex class I genetic diversity. Annu Rev Anim Biosci. 2:285-306.
147. Pedersen LD, Sørensen AC, and Berg P. 2010. Marker-assisted selection reduces expected inbreeding but can result in large effects of hitchhiking. J Anim Breed Genet. 127:189-98.
148. Bijma P. 2012. Long-term genomic improvement - new challenges for population genetics. J Anim Breed Genet. 129:1-2.
149. Goddard M. 2009. Genomic selection: prediction of accuracy and maximisation of long term response. Genetica. 136:245-57.
150. Cole JB. 2015. A simple strategy for managing many recessive disorders in a dairy cattle breeding program. Genet Sel Evol. 47:94.

151. Toro MÁ, and Varona L. 2010. A note on mate allocation for dominance handling in genomic selection. Genet Sel Evol. 42:33.
152. Bjelland DW, Weigel KA, Vukasinovic N, and Nkrumah JD. 2013. Evaluation of inbreeding depression in Holstein cattle using whole-genome SNP markers and alternative measures of genomic inbreeding. J Dairy Sci. 96:4697-706.
153. Martikainen K, Sironen A, and Uimari P. 2018. Estimation of intra-chromosomal inbreeding depression on female fertility using runs of homozygosity in Finnish Ayrshire cattle. J Dairy Sci. 101:11097-107.
154. McParland S, Kearney F, and Berry DP. 2009. Purging of inbreeding depression within the Irish Holstein-Friesian population. Genet Sel Evol. 41:16.
155. Charlesworth B, and Charlesworth D. 1999. The genetic basis of inbreeding depression. Genet Res. 74:329-40.
156. Charlesworth D, and Willis JH. 2009. The genetics of inbreeding depression. Nat Rev Genet. 10:783-96.
157. Smith LA, Cassell BG, and Pearson RE. 1998. The effects of inbreeding on the lifetime performance of dairy cattle. J Dairy Sci. 81:2729-37.
158. Miglior F, Burnside EB, and Kennedy BW. 1995. Production traits of Holstein cattle: estimation of non-additive genetic variance components and inbreeding depression. J Dairy Sci. 78:1174-80.
159. Croquet C, Mayeres P, Gillon A, Vanderick S, and Gengler N. 2006. Inbreeding depression for global and partial economic indexes, production, type, and functional traits. J Dairy Sci. 89:2257-67.
160. Boakes E, and Wang J. 2005. A simulation study on detecting purging of inbreeding depression in captive populations. Genet Res. 86:139-48.
161. Kalinowski ST, Hedrick PW, and Miller PS. 2000. Inbreeding depression in the Speke's gazelle captive breeding program. Conserv Biol. 14:1375-84.
162. Baumung R, Farkas J, Boichard D, Mészáros G, Sölkner J, and Curik I. 2015. GRAIN: a computer program to calculate ancestral and partial inbreeding coefficients using a gene dropping approach. J Anim Breed Genet. 132:100-8.
163. Hinrichs D, Bennewitz J, Wellmann R, and Thaller G. 2015. Estimation of ancestral inbreeding effects on stillbirth, calving ease and birthweight in German Holstein dairy cattle. J Anim Breed Genet. 132:59-67.
164. Boichard D. 2002. PEDIG: a fortran package for pedigree analysis suited for large populations. Proc 7th World Congr Genet Applied to Livest Prod. Montpellier, France.
165. VanRaden P. 1992. Accounting for inbreeding and crossbreeding in genetic evaluation of large populations. J Dairy Sci. 75:3136-44.
166. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, and Sham PC. 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 81:559-75.
167. Doekes HP, Veerkamp FR, Bijma P, Hiemstra SJ, and Windig JJ. 2018. Value of the Dutch Holstein-Friesian germplasm collection to increase genetic variability and improve genetic merit. J Dairy Sci. 101:10022-33.
168. Gilmour A, Gogel B, Cullis B, Welham S, and Thompson R. 2015. ASReml user guide release 4.1 structural specification. Hemel Hempstead: VSN international ltd.

169. Biffani S, Samoré A, and Canavesi F. 2002. Inbreeding depression for production, reproduction and functional traits in Italian Holstein cattle. Proc 7th World Congr Genet Applied to Livest Prod. Montpellier, France.
170. Miglior F, Burnside EB, and Dekkers JC. 1995. Non-additive genetic effects and inbreeding depression for somatic cell counts of Holstein cattle. J Dairy Sci. 78:1168-73.
171. Cassell B, Adamec V, and Pearson R. 2003. Effect of incomplete pedigrees on estimates of inbreeding and inbreeding depression for days to first service and summit milk yield in Holsteins and Jerseys. J Dairy Sci. 86:2967-76.
172. Doekes HP, Veerkamp RF, Hiemstra SJ, Bijma P, van der Beek S, and Windig JJ. 2018. Genomic selection and inbreeding and kinship in Dutch-Flemish Holstein-Friesian cattle. Proc 11th World Congr Genet Applied to Livest Prod. Auckland, New Zealand.
173. Doekes HP, Veerkamp RF, Bijma P, Hiemstra SJ, and Windig JJ. 2018. Trends in genome-wide and region-specific genetic diversity in the Dutch-Flemish Holstein-Friesian breeding program from 1986 to 2015. Genet Sel Evol. 50:15.
174. Koopman W. 2017. Inteelt blijft aandachtspunt, maar nadelige effecten tot nu toe minimaal. Veeteelt. Mei(2): 6-9.
175. Veerkamp RF, Kaal L, de Haas Y, and Oldham JD. 2013. Breeding for robust cows that produce healthier milk: ROBUSTMILK. Adv Anim Biosci. 4:594-9.
176. Kristensen TN, and Sørensen AC. 2005. Inbreeding-lessons from animal breeding, evolutionary biology and conservation genetics. Anim Sci. 80:121-33.
177. Todd ET, Ho SY, Thomson PC, Ang RA, Velie BD, and Hamilton NA. 2018. Founder-specific inbreeding depression affects racing performance in Thoroughbred horses. Sci Rep. 8:6167.
178. Suwanlee S, Baumung R, Sölkner J, and Curik I. 2007. Evaluation of ancestral inbreeding coefficients: Ballou's formula versus gene dropping. Conserv Genet. 8:489-95.
179. Boakes EH, Wang J, and Amos W. 2007. An investigation of inbreeding depression and purging in captive pedigreed populations. Heredity (Edinb.). 98:172-82.
180. García-Dorado A, Wang J, and López-Cortegano E. 2016. Predictive model and software for inbreeding-purging analysis of pedigreed populations. G3 (Bethesda). 6:3593-601.
181. López-Cortegano E, Bersabé D, Wang J, and García-Dorado A. 2018. Detection of genetic purging and predictive value of purging parameters estimated in pedigreed populations. Heredity (Edinb.). 121:38-51
182. Szpiech ZA, Xu J, Pemberton TJ, Peng W, Zöllner S, Rosenberg NA, and Li JZ. 2013. Long runs of homozygosity are enriched for deleterious variation. Am J Hum Genet. 93:90-102.
183. Zhang Q, Guldbrandtsen B, Bosse M, Lund MS, and Sahana G. 2015. Runs of homozygosity and distribution of functional variants in the cattle genome. BMC Genom. 16:542.
184. Sams AJ, and Boyko AR. 2019. Fine-scale resolution of runs of homozygosity reveal patterns of inbreeding and substantial overlap with recessive disease genotypes in domestic dogs. G3 (Bethesda). 9:117-23.
185. Druet T, and Gautier M. 2017. A model-based approach to characterize individual inbreeding at both global and local genomic scales. Mol Ecol. 26:5820-41.

186. Sole M, Gori AS, Faux P, Bertrand A, Farnir F, Gautier M, and Druet T. 2017. Age-based partitioning of individual genomic inbreeding levels in Belgian Blue cattle. Genet Sel Evol. 49:92.
187. García-Dorado A. 2012. Understanding and predicting the fitness decline of shrunk populations: inbreeding, purging, mutation, and standard selection. Genetics. 190:1461-1476.
188. Ács V, Bokor Á, and Nagy I. 2019. Population Structure Analysis of the Border Collie Dog Breed in Hungary. Animals. 9(5):e250.
189. Addo S, Schäler J, Hinrichs D, and Thaller G. 2017. Genetic Diversity and Ancestral History of the German Angler and the Red-and-White Dual-Purpose Cattle Breeds Assessed through Pedigree Analysis. Agr Sci. 8:1033.
190. Schäler J, Krüger B, Thaller G, and Hinrichs D. 2018. Comparison of ancestral, partial, and genomic inbreeding in a local pig breed to achieve genetic diversity. Conserv Genet Resour. 22:77-86.
191. Vostry L, Milerski M, Schmidova J, and Vostra-Vydrova H. 2018. Genetic diversity and effect of inbreeding on litter size of the Romanov sheep. Small Ruminant Res. 168:25-31.
192. Roos L, Hinrichs D, Nissen T, and Krieter J. 2015. Investigations into genetic variability in Holstein horse breed using pedigree data. Livest Sci. 177:25-32.
193. Doekes HP, Veerkamp RF, Bijma P, de Jong G, Hiemstra SJ, and Windig JJ. 2019. Inbreeding depression due to recent and ancient inbreeding in Dutch Holstein-Friesian dairy cattle. Genet Sel Evol. 51:54.
194. BOKU. 2020. Department für Nachhaltige Agrarsysteme: Institut für Nutztierwissenschaften: Software. Visited April 17, 2020, *https://boku.ac.at/nas/nuwi/software/*
195. ANGEN. 2020. Software: GRAIN 2-2. Visited April 17, 2020, *https://angen.agr.hr/hr/group/37/Grain+2-2*
196. Bolormaa S, Pryce JE, Zhang Y, Reverter A, Barendse W, and Hayes BJ. 2015. Non-additive genetic variation in growth, carcass and fertility traits of beef cattle. Genet Sel Evol. 47:26.
197. Jiang J, Ma L, Prakapenka D, VanRaden PM, Cole JB, and Da Y. 2019. A Large-Scale Genome-Wide Association Study in U.S. Holstein Cattle. Front Genet. 10:412.
198. MacLeod I, Hayes B, Savin K, Chamberlain A, McPartlan H, and Goddard M. 2010. Power of a genome scan to detect and locate quantitative trait loci in cattle using dense single nucleotide polymorphisms. J Anim Breed Genet. 127(2):133-142.
199. Vitezica ZG, Varona L, and Legarra A. 2013. On the additive and dominant variance and covariance of individuals within the genomic selection scope. Genetics. 195(4):1223-1230.
200. Aliloo H, Pryce JE, González-Recio O, Cocks BG, and Hayes BJ. 2016. Accounting for dominance to improve genomic evaluations of dairy cows for fertility and milk production traits. Genet Sel Evol. 48:8.
201. Calus M, Goddard M, Wientjes Y, Bowman P, and Hayes B. 2018. Multibreed genomic prediction using multitrait genomic residual maximum likelihood and multitask Bayesian variable selection. J Dairy Sci. 101(5):4279-4294.

202. NAGRP. 2019. ARS-UCD1.2 Cow Genome Assembly: Mapping of all existing variants. Visited September 15, 2019, *www.animalgenome.org/repository/cattle/*
203. NCBI. 2019. Genome Remapping Service. Visited September 15, 2019, *www.ncbi.nlm.nih.gov/genome/tools/remap*
204. Lee SH, and van der Werf JH. 2016. MTG2: an efficient algorithm for multivariate linear mixed model analysis based on genomic information. Bioinformatics. 32(9):1420-1422.
205. Xiang T, Christensen OF, Vitezica ZG, and Legarra A. 2016. Genomic evaluation by including dominance effects and inbreeding depression for purebred and crossbred performance with an application in pigs. Genet Sel Evol. 48:92.
206. Varona L, Legarra A, Toro MA, and Vitezica ZG. 2018. Non-additive Effects in Genomic Selection. Front Genet. 9:78.
207. Meyer K, and Tier B. 2012. "SNP Snappy": A strategy for fast genome-wide association studies fitting a full mixed model. Genetics. 190(1):275-277.
208. Meyer K. 2007. WOMBAT—A tool for mixed model analyses in quantitative genetics by restricted maximum likelihood (REML). J Zhejiang Univ Sci B. 8(11):815-821.
209. Price AL, Zaitlen NA, Reich D, and Patterson N. 2010. New approaches to population stratification in genome-wide association studies. Nat Rev Genet. 11(7):459.
210. Benjamini Y, and Hochberg Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Series B Stat Methodol. 57(1):289-300.
211. EMBL-EBI. 2020. *e!Ensemble:* Cow (ARS-UCD1.2). Visited March 9, 2020, *www.ensembl.org/Bos_taurus/*
212. Kardos M, Nietlisbach P, and Hedrick PW. 2018. How should we compare different genomic estimates of the strength of inbreeding depression? Proc Natl Acad Sci USA. 115(11):E2492-E2493.
213. Sun C, VanRaden PM, Cole JB, and O'Connell JR. 2014. Improvement of prediction ability for genomic selection of dairy cattle by including dominance effects. PLoS ONE. 9(8): e103934.
214. Da Y, Wang C, Wang S, and Hu G. 2014. Mixed model methods for genomic prediction and variance component estimation of additive and dominance effects using SNP markers. PLoS ONE. 9(1):e87666.
215. Kawahara T, Gotoh Y, Yamaguchi S, and Suzuki M. 2006. Variance component estimates with dominance models for milk production in Holsteins of Japan using method R. Asian-Australas J Anim Sci. 19(6):769-774.
216. van Tassell C, Misztal I, and Varona L. 2000. Method R estimates of additive genetic, dominance genetic, and permanent environmental fraction of variance for yield and health traits of Holsteins. J Dairy Sci. 83(8):1873-1877.
217. Tempelman RJ, and Burnside EB. 1990. Additive and nonadditive genetic variation for production traits in Canadian Holsteins. J Dairy Sci. 73(8):2206-2213.
218. Hoeschele I. 1991. Additive and nonadditive genetic variance in female fertility of Holsteins. J Dairy Sci. 74(5):1743-1752.
219. Jiang J, Shen B, O'Connell JR, VanRaden PM, Cole JB, and Ma L. 2017. Dissection of additive, dominance, and imprinting effects for production and reproduction traits in Holstein cattle. BMC Genom. 18:425.

220. Alves K, Brito LF, Baes CF, Sargolzaei M, Robinson JAB, and Schenkel FS. 2020. Estimation of additive and non-additive genetic effects for fertility and reproduction traits in North American Holstein cattle using genomic information. J Anim Breed Genet. 137(3):316-330
221. Mao X, Sahana G, Johansson AM, Liu A, Ismael A, Løvendahl P, de Koning D-J, and Guldbrandtsen B. 2020. Genome-wide association mapping for dominance effects in female fertility using real and simulated data from Danish Holstein cattle. Sci Rep. 10:2953.
222. Zhu Z, Bakshi A, Vinkhuyzen AA, Hemani G, Lee SH, Nolte IM, van Vliet-Ostaptchouk JV, Snieder H, Esko T, and Milani L. 2015. Dominance genetic variation contributes little to the missing heritability for human complex traits. Am J Hum Genet. 96(3):377-385.
223. Aliloo H, Pryce J, González-Recio O, Cocks B, Goddard M, and Hayes B. 2017. Including nonadditive genetic effects in mating programs to maximize dairy farm profitability. J Dairy Sci. 100(2):1203-1222.
224. Beavis W. 1998. QTL analyses: power, precision, and accuracy. In: Molecular dissection of complex traits. CRC Press, New York.
225. Schmid M, and Bennewitz J. 2017. Invited review: Genome-wide association analysis for quantitative traits in livestock-a selective review of statistical models and experimental designs. Arch Anim Breed. 60(3):335-346.
226. Legarra A, Ricard A, and Varona L. 2018. GWAS by GBLUP: single and multimarker EMMAX and Bayes factors, with an example in detection of a major gene for horse gait. G3 (Bethesda). 8(7):2301-2308.
227. Moser G, Lee SH, Hayes BJ, Goddard ME, Wray NR, and Visscher PM. 2015. Simultaneous discovery, estimation and prediction analysis of complex traits using a Bayesian mixture model. PLoS Genet. 11(4): e1004969.
228. Gianola D. 2013. Priors in whole-genome regression: the Bayesian alphabet returns. Genetics. 194(3):573-596.
229. Yang TL, Guo Y, Zhang JG, Xu C, Tian Q, and Deng HW. 2015. Genome-Wide Survey of Runs of Homozygosity Identifies Recessive Loci for Bone Mineral Density in Caucasian and Chinese Populations. J Bone Miner Res. 30(11):2119-2126.
230. Howard JT, Tiezzi F, Huang Y, Gray KA, and Maltecca C. 2017. A heuristic method to identify runs of homozygosity associated with reduced performance in livestock. J Anim Sci. 95(10):4318-4332.
231. Leroy G, Danchin-Burge C, and Verrier E. 2011. Impact of the use of cryobank samples in a selected cattle breed: a simulation study. Genet Sel Evol. 43:36.
232. Sonesson AK, Goddard ME, and Meuwissen TH. 2002. The use of frozen semen to minimize inbreeding in small populations. Genet Res. 80(1):27-30.
233. Shepherd RK, and Woolliams JA. 2004. Minimising inbreeding in small populations by rotational mating with frozen semen. Genet Res. 84(2):87-93.
234. Smith C. 1977. Use of stored frozen semen and embryos to measure genetic trends in farm livestock. J Anim Breed Genet. 94(1-4):119-130.
235. Garcia M, and Baselga M. 2002. Estimation of genetic response to selection in litter size of rabbits using a cryopreserved control population. Livest Prod Sci. 74(1):45-53.
236. GES. 2017. Fokwaarden: Perspublicaties met fokwaarden: Stieren. Visited January 29, 2018, *www.gesfokwaarden.eu/nl/fokwaarden/fokwaarden.php*

237. GES. 2015. Kengetallen NVI. In: Handboek Kwaliteit. Visited January 29, 2018, *www.gesfokwaarden.eu/nl/fokwaarden/pdf/E_20.pdf*
238. GES. 2015. Kengetallen INET. In: Handboek Kwaliteit. Visited January 29, 2018, *www.gesfokwaarden.eu/nl/fokwaarden/pdf/E_9.pdf*
239. GES. 2015. Fokwaarde vruchtbaarheid. In: Handboek Kwaliteit. Visited January 29, 2018, *www.gesfokwaarden.eu/nl/fokwaarden/pdf/E_17.pdf*
240. GES. 2017. Fokwaarde uiergezondheid. In: Handboek Kwaliteit. Visited January 29, 2018, *http://www.gesfokwaarden.eu/nl/fokwaarden/pdf/E_27.pdf*
241. Grundy B, Villanueva B, and Woolliams JA. 1998. Dynamic selection procedures for constrained inbreeding and their consequences for pedigree development. Genet Res. 72(02):159-168.
242. Meuwissen THE. 2002. GENCONT: an operational tool for controlling inbreeding in selection and conservation schemes. Proc 7th World Congr Genet Applied to Livest Prod. Montpellier, France.
243. Jaton C, Koeck A, Sargolzaei M, Malchiodi F, Price CA, Schenkel FS, and Miglior F. 2016. Genetic analysis of superovulatory response of Holstein cows in Canada. J Dairy Sci. 99(5): 3612-3623.
244. Boichard D, and Brochard M. 2012. New phenotypes for new breeding goals in dairy cattle. Animal. 6(4):544-550.
245. Egger-Danner C, Cole J, Pryce J, Gengler N, Heringstad B, Bradley A, and Stock KF. 2015. Invited review: overview of new traits and phenotyping strategies in dairy cattle with a focus on functional traits. Animal. 9(2):191-207.
246. Bennewitz J, Simianer H, and Meuwissen THE. 2008. Investigations on merging breeds in genetic conservation schemes. J Dairy Sci. 91:2512-2519.
247. Smith C. 1984. Genetic aspects of conservation in farm livestock. Livest Prod Sci. 11:37-48.
248. Eding H, and Meuwissen THE. 2003. Linear methods to estimate kinships from genetic marker data for the construction of core sets in genetic conservation schemes. J Anim Breed Genet. 120:289-302.
249. Fernández J, Meuwissen THE, Toro MA, and Mäki-Tanila A. 2011. Management of genetic diversity in small farm animal populations. Animal. 5:1684-1698.
250. Maurice - van Eijndhoven M. 2014. PhD thesis: Genetic variation of milk fatty acid composition between and within dairy cattle breeds. Wageningen University & Research, Wageningen.
251. van Helden W, and Minkema D. 1978. Inventarisatie van zeldzame huisdierrassen. SZH, Groningen.
252. CGN. 2017. Rassenlijst Nederlandse landbouwhuisdierrassen en hun risicostatus - udate 2017. CGN, Wageningen.
253. Paradis E, Claude J, and Strimmer K. APE: analyses of phylogenetics and evolution in R language. Bioinformatics. 2004;20:289-290.
254. Rambaut A, and Drummond AJ. 2012. Figtree v1.4. Visited 21 June, 2019, *http://tree.bio.ed.ac.uk/software/figtree/*
255. Raj A, Stephens M, and Pritchard JK. 2014. FastSTRUCTURE: variational inference of population structure in large SNP data sets. Genetics. 197:573-589.

256. Evanno G, Regnaut S, and Goudet J. 2005. Detecting the number of clusters of individuals using the software STRUCTURE : A simulation study. Mol Ecol. 14:2611-2620.
257. Francis RM. 2017. POPHELPER: An R package and web app to analyse and visualize population structure. Mol Ecol Resour. 17:27-32.
258. Berg P, and Windig JJ. 2017. Management of cryo-collecions with genomic tools. In: Genomic management of animal genetic diversity. 2017. Wageningen Academic Publishers, Wageningen.
259. Chang CC, Chow CC, Tellier LCAM, Vattikuti S, Purcell SM, and Lee JJ. 2015. Second-generation PLINK: Rising to the challenge of larger and richer datasets. Gigascience. 4:1-16.
260. van Mil R, and Nauta WJ. 2019. Fundamentfokkerij Fries Hollands vee. Louis Bolk Instituut, Driebergen.
261. FAO. 2012. Animal production and health guidelines: Cryoconservation of animal genetic resources. FAO Commision on Genetic Resources for Food and Agriculture, Rome.
262. de Oliveira Silva R, Ahmadi BV, Hiemstra SJ, and Moran D. 2019. Optimizing ex situ genetic resource collections for European livestock conservation. J Anim Breed Genet. 136:63-73.
263. Wang Y, Bennewitz J, and Wellmann R. 2017. Novel optimum contribution selection methods accounting for conflicting objectives in breeding programs for livestock breeds with historical migration. Genet Sel Evol. 49:45.
264. Eynard SE, Windig JJ, Hiemstra SJ, and Calus MPL. 2016. Whole-genome sequence data uncover loss of genetic diversity due to selection. Genet Sel Evol. 48:33.
265. Albrechtsen A, Nielsen FC, and Nielsen R. 2010. Ascertainment Biases in SNP Chips Affect Measures of Population Divergence. Mol Biol Evol. 27:2534-2547.
266. Baes CF, Makanjuola BO, Miglior F, Marras G, Howard JT, Fleming A, and Maltecca C. 2019. Symposium review: The genomic architecture of inbreeding: How homozygosity affects health and performance. J Dairy Sci. 102(3):2807-2817.
267. Eusebi PG, Martinez A, and Cortes O. 2020. Genomic Tools for Effective Conservation of Livestock Breed Diversity. Diversity. 12(1):8.
268. Maltecca C, Tiezzi F, Cole J, and Baes C. 2020. Symposium review: Exploiting homozygosity in the era of genomics—Selection, inbreeding, and mating programs. J Dairy Sci. 103(6): 5302-5313.
269. Dekkers J. 2007. Prediction of response to marker-assisted and genomic selection using selection index theory. J Anim Breed Genet. 124(6):331-341.
270. Habier D. 2010. More than a third of the WCGALP presentations on genomic selection. J Anim Breed Genet. 127(5):336-337.
271. Forutan M, Ansari Mahyari S, Baes C, Melzer N, Schenkel FS, and Sargolzaei M. 2018. Inbreeding and runs of homozygosity before and after genomic selection in North American Holstein cattle. BMC Genom. 19:98.
272. Topolski P, and Jagusiak W. 2019. Inbreeding in a population of Polish Holstein-Friesian young bulls before and after genomic selection. Ann Anim Sci. 20(1):71-83.
273. Doublet A-C, Croiseau P, Fritz S, Michenet A, Hozé C, Danchin-Burge C, Laloë D, and Restoux G. 2019. The impact of genomic selection on genetic diversity and genetic gain in three French dairy cattle breeds. Genet Sel Evol. 51:52.

274. Makanjuola BO, Miglior F, Abdalla EA, Maltecca C, Schenkel FS, and Baes CF. 2020. Effect of genomic selection on rate of inbreeding and coancestry and effective population size of Holstein and Jersey cattle populations. J Dairy Sci. 103(6):5183-5199.
275. Chavinskaia L, Ducrocq V, and Joly P. 2017. Interbull: Constructing international commensurability for dairy cattle selection. Interbull Bulletin. Tallinn, Estonia.
276. Pelikaan F. 2015. Huidig stiergebruik vraagt aanpassing fokprogramma. Veeteelt. Mei(2):26-27.
277. Eynard SE, Croiseau P, Laloë D, Fritz S, Calus MPL, and Restoux G. 2018. Which Individuals To Choose To Update the Reference Population? Minimizing the Loss of Genetic Diversity in Animal Genomic Selection Programs. G3 (Bethesda). 8(1):113-121.
278. Engelsma KA, Veerkamp RF, Calus MPL, and Windig JJ. 2011. Consequences for diversity when prioritizing animals for conservation with pedigree or genomic information. J Anim Breed Genet. 128(6):473-481.
279. de Cara MÁR, Fernández J, Toro MÁ, Villanueva B. 2011. Using genome-wide information to minimize the loss of diversity in conservation programmes. J Anim Breed Genet. 128(6):456-464.
280. Gómez-Romano F, Villanueva B, de Cara MÁR, and Fernández J. 2013. Maintaining genetic diversity using molecular coancestry: the effect of marker density and effective population size. Genet Sel Evol. 45:38.
281. Liu H, Meuwissen TH, Sørensen AC, and Berg P. 2015. Upweighting rare favourable alleles increases long-term genetic gain in genomic selection programs. Genet Sel Evol. 47:19.
282. Thomasen J, Liu H, and Sørensen A. 2020. Genotyping more cows increases genetic gain and reduces rate of true inbreeding in a dairy cattle breeding scheme using female reproductive technologies. J Dairy Sci. 103(1):597-606.
283. Meuwissen T, Sonesson AK, and Woolliams JA. 2018. Genomic management of inbreeding in breeding schemes. Proc 11th World Congr Genet Applied to Livest Prod. Auckland, New Zealand.
284. Berry DP. 2018 Symposium review: Breeding a better cow—Will she be adaptable? J Dairy Sci. 101(4):3665-3685.
285. Jannink JL. 2010. Dynamics of long-term genomic selection. Genet Sel Evol. 42:35.
286. de Beukelaer H, Badke Y, Fack V, and de Meyer G. 2017. Moving Beyond Managing Realized Genomic Relationship in Long-Term Genomic Selection. Genetics. 206(2):1127-1138.
287. Zhang Q, Calus MP, Guldbrandtsen B, Lund MS, and Sahana G. 2015. Estimation of inbreeding using pedigree, 50k SNP chip genotypes and full sequence data in three cattle breeds. BMC Genet. 16:88.
288. Bosse M, Megens HJ, Madsen O, Paudel Y, Frantz LAF, and Schook LB. 2012. Regions of homozygosity in the porcine genome: consequence of demography and the recombination landscape. PLoS Genet. 8(11):e1003100.
289. Purfield DC, McParland S, Wall E, Berry and DP. 2017. The distribution of runs of homozygosity and selection signatures in six commercial meat sheep breeds. PLoS ONE. 12(5):e0176780.

290. Elferink MG, van As P, Veenendaal T, Crooijmans RP, and Groenen MA. 2010. Regional differences in recombination hotspots between two chicken populations. BMC Genet. 11:11.
291. Dagnachew BS, and Meuwissen THE. 2016. A fast Newton-Raphson based iterative algorithm for large scale optimal contribution selection. Genet Sel Evol. 48:70.
292. Fujisawa K, Kojima M, Nakata K, and Yamashita M. 2002. SDPA (SemiDefinite Programming Algorithm) User's Manual—Version 6.2.0.
293. Vandenberghe L, and Boyd S. 1996. Semidefinite programming. SIAM Review. 38(1):49-95.
294. Carvalheiro R, Queiroz SAd, and Kinghorn B. 2010. Optimum contribution selection using differential evolution. Rev Bras Zootecn. 39(7):1429-1436.
295. Fernández J, Toro M, Gómez-Romano F, and Villanueva B. 2016. The use of genomic information can enhance the efficiency of conservation programs. Anim Front. 6(1):59-64.
296. Engelsma K, Veerkamp R, Calus M, and Windig J. 2014. Consequences for diversity when animals are prioritized for conservation of the whole genome or of one specific allele. J Anim Breed Genet. 131(1):61-70.
297. Roughsedge T, Pong-Wong R, Woolliams JA, and Villanueva B. 2008. Restricting coancestry and inbreeding at a specific position on the genome by using optimized selection. Genet Res. 90(2):199-208.
298. Gottardo F, Nalon E, Contiero B, Normando S, Dalvit P, and Cozzi G. 2011. The dehorning of dairy calves: Practices and opinions of 639 farmers. J Dairy Sci. 94(11):5724-5734.
299. Cozzi G, Gottardo F, Brscic M, Contiero B, Irrgang N, Knierim U, Pentelescu O, Windig J, Mirabito L, and Eveillard FK. 2015. Dehorning of cattle in the EU Member States: A quantitative survey of the current practices. Livest Sci. 179:4-11.
300. Caray D, Des Roches AdB, Frouja S, Andanson S, and Veissier I. 2015. Hot-iron disbudding: stress responses and behavior of 1-and 4-week-old calves receiving anti-inflammatory analgesia without or with sedation using xylazine. Livest Sci. 179:22-28.
301. Stafford KJ, and Mellor DJ. 2011. Addressing the pain associated with disbudding and dehorning in cattle. Appl Anim Behav Sci. 135(3):226-231.
302. Windig JJ, Hoving-Bolink RA, and Veerkamp RF. 2015. Breeding for polledness in Holstein cattle. Livest Sci. 179:96-101.
303. Gaspa G, Veerkamp RF, Calus MPL, and Windig JJ. 2015. Assessment of genomic selection for introgression of polledness into Holstein Friesian cattle by simulation. Livest Sci. 179:86-95.
304. Schafberg R, Swalve H. 2015. The history of breeding for polled cattle. Livest Sci. 179:54-70.
305. Medugorac I, Seichter D, Graf A, Russ I, Blum H, Göpel KH, Rothammer S, Förster M, and Krebs S. 2012. Bovine polledness-an autosomal dominant trait with allelic heterogeneity. PloS ONE. 7(6):e39477.
306. Allais-Bonnet A, Grohs C, Medugorac I, Krebs S, Djari A, Graf A, Fritz S, Seichter D, Baur A, and Russ I. 2013. Novel insights into the bovine polled phenotype and horn ontogenesis in Bovidae. PloS ONE. 8(5):e63512.

307. Rothammer S, Capitan A, Mullaart E, Seichter D, Russ I, and Medugorac I. 2014. The 80-kb DNA duplication on BTA1 is the only remaining candidate mutation for the polled phenotype of Friesian origin. Genet Sel Evol. 46:44.
308. Segelke D, Helge T, Jansen S, Pausch H, Reinhardt F, and Thaller G. 2014. Management of genetic characteristics. Interbull Bulletin. Berlin, Germany.
309. Maltecca C, Baes C, and Tiezzi F. 2019. The use of genomic information to improve selection response while controlling inbreeding in dairy cattle breeding programs. In: Advances in breeding of dairy cattle. Burleigh Dodds Science Publishing Limited, Cambridge.
310. Schäler J, Wellmann R, Bennewitz J, Thaller G, and Hinrichs D. 2018. Genetic diversity and historic introgression in German Angler and Red Dual Purpose cattle and possibilities to reverse introgression. Acta Agric Scand A Anim. 68(2):63-72.
311. Wellmann R. 2019. Optimum contribution selection for animal breeding and conservation: the R package optiSel. BMC Bioinform. 20(1):25.
312. Hayes BJ, and Daetwyler HD. 2019. 1000 bull genomes project to map simple and complex genetic traits in cattle: applications and outcomes. Annu Rev Anim Biosci. 7:89-102.
313. van Binsbergen R, Calus MP, Bink MC, van Eeuwijk FA, Schrooten C, and Veerkamp RF. 2015. Genomic prediction using imputed whole-genome sequence data in Holstein Friesian cattle. Genet Sel Evol. 47:71.
314. Veerkamp RF, Bouwman AC, Schrooten C, and Calus MP. 2016. Genomic prediction using preselected DNA variants from a GWAS with whole-genome sequence data in Holstein-Friesian cattle. Genet Sel Evol. 48:95.
315. VanRaden PM, Tooker ME, O'connell JR, Cole JB, and Bickhart DM. 2017. Selecting sequence variants to improve genomic predictions for dairy cattle. Genet Sel Evol. 49:32.
316. Brøndum RF, Su G, Janss L, Sahana G, Guldbrandtsen B, Boichard D, and Lund MS. 2015. Quantitative trait loci markers derived from whole genome sequence data increases the reliability of genomic prediction. J Dairy Sci. 98(6):4107-4116.
317. Raymond B, Bouwman AC, Schrooten C, Houwing-Duistermaat J, and Veerkamp RF. 2018. Utility of whole-genome sequence data for across-breed genomic prediction. Genet Sel Evol. 50:27.
318. MacLeod I, Bowman P, Vander Jagt C, Haile-Mariam M, Kemper K, Chamberlain A, Schrooten C, Hayes B, and Goddard M. 2016. Exploiting biological priors and sequence variants enhances QTL discovery and genomic prediction of complex traits. BMC Genom. 17:144.
319. Matukumalli LK, Lawley CT, Schnabel RD, Taylor JF, Allan MF, Heaton MP, O'Connell J, Moore SS, Smith TP, and Sonstegard TS. 2009. Development and characterization of a high density SNP genotyping assay for cattle. PLoS ONE. 4(4):e5350.
320. Illumina. 2016. Data Sheet: DNA Analysis: BovineSNP50 genotyping beadchip. Visited April 7, 2020, *https://www.illumina.com*
321. Pérez-Enciso M. 2014. Genomic relationships computed from either next-generation sequence or array SNP data. J Anim Breed Genet. 131(2):85-96.
322. Porto Neto LR, and Barendse W. 2010. Effect of SNP origin on analyses of genetic diversity in cattle. Anim Prod Sci. 50(8):792-800.

323. Mulder HA, Lee SH, Clark S, Hayes BJ, and van der Werf JH. 2019. The impact of genomic and traditional selection on the contribution of mutational variance to long-term selection response and genetic variance. Genetics. 213(2):361-378.
324. Leroy G, Besbes B, Boettcher P, Hoffmann I, Capitan A, and Baumung R. 2016. Rare phenotypes in domestic animals: unique resources for multiple applications. Anim Genet. 47(2):141-153.
325. Bickhart DM, and Liu GE. 2014. The challenges and importance of structural variation detection in livestock. Front Genet. 5:37.
326. Wellenreuther M, Mérot C, Berdan E, and Bernatchez L. 2019. Going beyond SNPs: the role of structural genomic variants in adaptive evolution and species diversification. Mol Ecol. 28(6):1203-1209.
327. Xu L, Cole JB, Bickhart DM, Hou Y, Song J, VanRaden PM, Sonstegard TS, van Tassell CP, and Liu GE. 2014. Genome wide CNV analysis reveals additional variants associated with milk production traits in Holsteins. BMC Genom. 15:683.
328. Yue X, Chang T, DeJarnette J, Marshall C, Lei C, and Liu W-S. 2013. Copy number variation of PRAMEY across breeds and its association with male fertility in Holstein sires. J Dairy Sci. 96(12):8024-8034.
329. Bickhart DM, Xu L, Hutchison JL, Cole JB, Null DJ, Schroeder SG, Song J, Garcia JF, Sonstegard TS, and van Tassell CP. 2016. Diversity and population-genetic properties of copy number variations and multicopy genes in cattle. DNA Res. 23(3):253-262.
330. Derks MF, Megens H-J, Bosse M, Lopes MS, Harlizius B, and Groenen MA. 2017. A systematic survey to identify lethal recessive variation in highly managed pig populations. BMC Genom. 18:858.
331. Eriksson S, Jonas E, Rydhmer L, and Röcklinsberg H. 2018. Invited review: Breeding and ethical perspectives on genetically modified and genome edited cattle. J Dairy Sci. 101(1):1-17.
332. Tait-Burkard C, Doeschl-Wilson A, McGrew MJ, Archibald AL, Sang HM, Houston RD, Whitelaw CB, and Watson M. 2018. Livestock 2.0-genome editing for fitter, healthier, and more productive farmed animals. Genome Biol. 19(1):204.
333. Petersen B. 2017. Basics of genome editing technology and its application in livestock species. Reprod Domest Anim. 52:4-13.
334. Proudfoot C, Carlson DF, Huddart R, Long CR, Pryor JH, King TJ, Lillico SG, Mileham AJ, McLaren DG, and Whitelaw CBA. 2015. Genome edited sheep and cattle. Transgenic Res. 24(1):147-153.
335. Carlson DF, Lancto CA, Zang B, Kim E-S, Walton M, Oldeschulte D, Seabury C, Sonstegard TS, and Fahrenkrug SC. 2016. Production of hornless dairy cattle from genome-edited cell lines. Nat Biotechnol. 34(5):479-481.
336. Gao Y, Wu H, Wang Y, Liu X, Chen L, Li Q, Cui C, Liu X, Zhang J, and Zhang Y. 2017. Single Cas9 nickase induced generation of NRAMP1 knockin cattle with reduced off-target effects. Genome Biol. 18(1):13.
337. Mueller M, Cole J, Sonstegard T, and van Eenennaam A. 2019. Comparison of gene editing versus conventional breeding to introgress the POLLED allele into the US dairy cattle population. J Dairy Sci. 102(5):4215-4226.

338. Bastiaansen JW, Bovenhuis H, Groenen MA, Megens H-J, and Mulder HA. 2018. The impact of genome editing on the introduction of monogenic traits in livestock. Genet Sel Evol. 50:18.
339. Jenko J, Gorjanc G, Cleveland MA, Varshney RK, Whitelaw CBA, Woolliams JA, and Hickey JM. 2015. Potential of promotion of alleles by genome editing to improve quantitative traits in livestock breeding programs. Genet Sel Evol. 47:55.
340. Johnsson M, Gaynor RC, Jenko J, Gorjanc G, de Koning D-J, and Hickey JM. 2019. Removal of alleles by genome editing (RAGE) against deleterious load. Genet Sel Evol. 51:14.
341. Simianer H, Pook T, and Schlather M. 2018. Turning the PAGE-the potential of genome editing in breeding for com-plex traits revisited. Proc 11th World Congr Genet Applied to Livest Prod. Auckland, New Zealand.
342. Wu X, Kriz AJ, and Sharp PA. 2014. Target specificity of the CRISPR-Cas9 system. Quant Biol. 2(2):59-70.
343. Bovenkerk B. 2020. Ethical perspectives on modifying animals: beyond welfare arguments. Anim Front. 10(1):45-50.
344. Leroy G, Boettcher P, Besbes B, Danchin-Burge C, Baumung R, and Hiemstra SJ. 2019. Cryoconservation of Animal Genetic Resources in Europe and Two African Countries: A Gap Analysis. Diversity. 11(12):240.
345. Hiemstra SJ (IMAGE Work Package 2 leader). Personal communication May 6, 2020.
346. Ugur MR, Saber Abdelrahman A, Evans HC, Gilmore AA, Hitit M, Arifiantini RI, Purwantara B, Kaya A, and Memili E. 2019. Advances in Cryopreservation of Bull Sperm. Front Vet Sci. 6:268.
347. Santiago-Moreno J, and Blesbois E. 2020. Strategy for biobanking avian resources: advantages and limits in the implementation of a sperm cryobank. Proc International Conference on Animal Genetic Resources. Madrid, Spain.
348. Woelders H, Matthijs A, Zuidberg C, and Chaveiro A. 2005. Cryopreservation of boar semen: equilibrium freezing in the cryomicroscope and in straws. Theriogenology. 63(2):383-395.
349. Oldenhof H, Bigalk J, Hettel C, de Oliveira Barros L, Sydykov B, Bajcsy AC, Sieme H, and Wolkers WF. 2017. Stallion sperm cryopreservation using various permeating agents: interplay between concentration and cooling rate. Biopreserv Biobank. 15(5):422-431.
350. Long JA, Purdy PH, Zuidberg K, Hiemstra S-J, Velleman SG, and Woelders H. 2014. Cryopreservation of turkey semen: effect of breeding line and freezing method on post-thaw sperm quality, fertilization, and hatching. Cryobiology. 68(3):371-378.
351. Woelders H. 2019. Novel gene banking approaches in poultry and pigs: PGCs, gonads, embryos, semen. In Symposium on new opportunities of animal reproductive and cryopreservation technologies for breeding and conservation. September 5, 2019. Wageningen, the Netherlands.
352. Woelders H, Windig J, and Hiemstra S. 2012. How developments in cryobiology, reproductive technologies and conservation genomics could shape gene banking strategies for (farm) animals. Reprod Domest Anim. 47:264-273.
353. Blackburn H. 2018. Biobanking genetic material for agricultural animal species. Annu Rev Anim Biosci. 6:69-82.

354. Hiemstra SJ. 2018. Ethical aspects of gene banks for livestock - stakeholder perspectives. Visited April 20, 2020, *https://www.wur.nl/en/newsarticle/Ethical-aspects-of-gene-banks-for-livestock-stakeholder-perspectives.htm*
355. Hulsegge I, Calus M, Hoving-Bolink R, Lopes M, Megens H-J, and Oldenbroek K. 2019. Impact of merging commercial breeding lines on the genetic diversity of Landrace pigs. Genet Sel Evol. 51:60.
356. Brekke C, Groeneveld LF, Meuwissen THE, Sæther N, Weigend S, and Berg P. 2020. Assessing the genetic diversity conserved in the Norwegian live poultry genebank. Acta Agr Scand A-An. 69(1-2):68-80.
357. Dierks C, Ha NT, Weigend A, Simianer H, Andersson B, Schmutz M, Cavero D, Preisinger R, and Weigend S. 2020. Painting eggs in blue. Proc International Conference on Animal Genetic Resources. Madrid, Spain.
358. Paris C, Boitard S, Servin B, Sevane N, and Dunner S. 2019. Annotation of selection signatures in the bovine breed Asturiana de Valles. Proc 70th Annual Meeting of EAAP. Ghent, Belgium.
359. Paris C, Servin B, and Boitard S. 2019. Inference of selection from genetic time series using various parametric approximations to the Wright-Fisher model. G3 (Bethesda). 9(12):4073-4086.
360. Eding H, Crooijmans RPMA, Groenen MAM, and Meuwissen THE. 2002. Assessing the contribution of breeds to genetic diversity in conservation schemes. Genet Sel Evol. 34(5):1.
361. Crooijmans R, Colli L, Stella A, Ajmone-Marsan P, Hiemstra S, and Tixier-Boichard M. 2019. The use of genomics in European livestock genebank collections. Proc 70th Annual Meeting of EAAP. Ghent, Belgium.
362. Mascher M, Schreiber M, Scholz U, Graner A, Reif JC, and Stein N. 2019. Genebank genomics bridges the gap between the conservation of crop diversity and plant breeding. Nat Genet. 51(7):1076-1081.
363. Duchev Z, Cong TVC, and Groeneveld E. 2010. CryoWEB: Web software for the documentation of the cryo-preserved material in animal gene banks. Bioinformation. 5(5):219-220.
364. Irwin G, Wessel L, and Blackburn H. 2012. The Animal Genetic Resource Information Network (AnimalGRIN) Database: A Database Design & Implementation Case. J Inf Syst Educ. 23(1).
365. Animal-GRIN. 2020. Animal Genetic Resources. Visited April 20, 2020, *https://agrin.ars.usda.gov/database_collaboration_page_dev#*
366. IMAGE. 2020. IMAGE Data Portal. Visited April 20, 2020, *https://www.image2020genebank.eu/*
367. EMBL-EBI. 2020. BioSamples. Visited April 20, 2020, *https://www.ebi.ac.uk/biosamples/*
368. EUGENA. 2020. Home. Visited May 6, 2020, *https://www.eugena-erfp.net/en/*
369. Alexandratos N, and Bruinsma J. 2012. World agriculture towards 2030/2050: the 2012 revision. ESA Working paper. FAO, Rome.
370. OECD & FAO. 2019. OECD-FAO Agricultural Outlook 2019-2028. OECD Publishing, Paris, and FAO, Rome

371. Hayes BJ, Lewin HA, and Goddard ME. 2013. The future of livestock breeding: genomic selection for efficiency, reduced emissions intensity, and adaptation. Trends Genet. 29(4):206-214.
372. Neeteson-van Nieuwenhoven A-M, Knap P, and Avendaño S. 2013. The role of sustainable commercial pig and poultry breeding for food security. Anim Front. 3(1):52-57.
373. VandeHaar M, Armentano L, Weigel K, Spurlock D, Tempelman R, and Veerkamp R. 2016. Harnessing the genetics of the modern dairy cow to continue improvements in feed efficiency. J Dairy Sci. 99(6):4941-4954.
374. Pryce J, Gonzalez-Recio O, Nieuwhof G, Wales W, Coffey M, Hayes B, and Goddard M. 2015. Hot topic: Definition and implementation of a breeding value for feed efficiency in dairy cows. J Dairy Sci. 98(10):7340-7350.
375. Calus M, de Haas Y, Pszczola M, and Veerkamp R. 2013. Predicted accuracy of and response to genomic selection for new traits in dairy cattle. Animal. 7(2):183-191.
376. Rojas-Downing MM, Nejadhashemi AP, Harrigan T, and Woznicki SA. 2017. Climate change and livestock: Impacts, adaptation, and mitigation. Clim Risk Manag. 16:145-163.
377. Herrero M, Henderson B, Havlík P, Thornton PK, Conant RT, Smith P, Wirsenius S, Hristov AN, Gerber P, and Gill M. 2016. Greenhouse gas mitigation potentials in the livestock sector. Nat Clim Chang. 6(5):452-461.
378. IPCC. 2014. Climate Change 2014: Synthesis Report. Contribution of Working Groups I, II and III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change. IPCC, Geneva.
379. Nguyen TT, Bowman PJ, Haile-Mariam M, Nieuwhof GJ, Hayes BJ, and Pryce JE. 2017. Implementation of a breeding value for heat tolerance in Australian dairy cattle. J Dairy Sci. 100(9):7362-7367.
380. Dikmen S, Khan F, Huson H, Sonstegard T, Moss J, Dahl G, and Hansen P. 2014. The SLICK hair locus derived from Senepol cattle confers thermotolerance to intensively managed lactating Holstein cows. J Dairy Sci. 97(9):5508-5520.
381. Bett B, Kiunga P, Gachohi J, Sindato C, Mbotha D, Robinson T, Lindahl J, and Grace D. 2017. Effects of climate change on the occurrence and distribution of livestock diseases. Prev Vet Med. 137:119-129.
382. Lozano-Jaramillo M. 2019. PhD-thesis: Predicting breed by environment interaction using ecological modelling. Wageningen University & Research, Wageningen.
383. Li Y, Zhang X-X, Mao R-L, Yang J, Miao C-Y, Li Z, and Qiu Y-X. 2017. Ten years of landscape genomics: challenges and opportunities. Front Plant Sci. 8:2136.
384. Rodenburg T, and Turner S. 2012. The role of breeding and genetics in the welfare of farm animals. Anim Front. 2(3):16-21.
385. Saitone TL, and Sexton RJ. 2017. Agri-food supply chain: evolution and performance with conflicting consumer and societal demands. Eur Rev Agric Econ. 44(4):634-657.
386. Berghof TV, Poppe M, and Mulder HA. 2019. Opportunities to improve resilience in animal breeding programs. Front Genet. 9:692.
387. Pryce J, Gaddis KP, Koeck A, Bastin C, Abdelsayed M, Gengler N, Miglior F, Heringstad B, Egger-Danner C, and Stock K. 2016. Invited review: Opportunities for genetic improvement of metabolic diseases. J Dairy Sci. 99(9):6855-6873.

388. Ellen ED, van der Sluis M, Siegford J, Guzhva O, Toscano MJ, Bennewitz J, van der Zande LE, van der Eijk JA, de Haas EN, and Norton T. 2019. Review of sensor technologies in animal breeding: Phenotyping behaviors of laying hens to select against feather pecking. Animals. 9(3):108.
389. Ritchie H. 2019. Half of the world's habitable land is used for agriculture (based on FAO data). Visited April 16, 2020, *https://ourworldindata.org/global-land-for-agriculture*
390. van Zanten HH, Herrero M, van Hal O, Röös E, Muller A, Garnett T, Gerber PJ, Schader C, and de Boer IJ. 2018. Defining a land boundary for sustainable livestock consumption. Glob Change Biol. 24(9):4185-4194.
391. Muscat A, de Olde E, de Boer IJ, and Ripoll-Bosch R. 2019. The battle for biomass: A systematic review of food-feed-fuel competition. Glob Food Sec. in press.
392. Tixier-Boichard M, Verrier E, Rognon X, and Zerjal T. 2015. Farm animal genetic and genomic resources from an agroecological perspective. 6:153.
393. de Boer IJ, and van Ittersum M. 2018. Circularity in agricultural production. Wageningen University & Research, Wageningen.

Summary

Genetic diversity is the basis for livestock populations to adapt to changing environments and human demands. Conservation of genetic diversity, therefore, is an important objective of sustainable livestock production. Genetic diversity can be conserved by managing the loss of diversity within breeding programs and production systems (*in situ*) and by establishing gene bank collections (*ex situ*).

Traditionally, livestock genetic diversity has been characterized and managed with pedigree-based measures of inbreeding and kinship. We now have additional opportunities to better characterize and conserve genetic diversity thanks to the increasing availability of genomic information, in particular single nucleotide polymorphism (SNP) data. These data allow, among others, to calculate more accurate coefficients of inbreeding and kinship, to study the negative effects of inbreeding on performance ('inbreeding depression') in more detail, and to estimate diversity across breeds. At the same time, the application of genomic information in selection schemes ('genomic selection') has raised questions on how to best conserve genetic diversity based on genomic information.

The overall aim of this thesis was to utilize the availability of SNP data to better characterize and conserve genetic diversity in Dutch cattle. The Holstein Friesian (HF) breed was used as the main breed of interest, because of its importance in the Dutch and global dairy cattle sector.

In **Chapter 2**, we used a time series of 6,280 genotyped bulls from 1986 to 2015 to investigate trends in genetic diversity in the Dutch-Flemish HF breeding program. We found major changes in diversity trends at two points in time. Around the year 2000, the introduction of optimal contribution selection (OCS) and shift in breeding goal were accompanied by a decrease in the rates of inbreeding (ΔF) and kinship (Δf). Around 2009, the implementation of genomic selection was accompanied by an increase in ΔF and Δf. For the period 2011-2015, the ΔF and Δf were estimated to be between 1.3% to 2.8% per generation. The observed increase in ΔF and Δf with implementation of genomic selection was remarkable and suggests a need for stricter management of inbreeding and kinship.

In **Chapters 3-5**, we studied the effects of inbreeding on performance of HF cows, using pedigree, genotype and phenotype data of 38,792 first-parity cows. As expected, we observed significant inbreeding depression for yield, fertility and udder health traits. At population level, genomic inbreeding measures were found to explain more inbreeding depression than pedigree-based measures. In **Chapter 3**, we furthermore investigated differences in the effects of recent and ancient inbreeding. Based on pedigree, recent inbreeding appeared to be more harmful than ancient inbreeding. Especially when the pedigree-based inbreeding coefficient was split into Kalinowski's new and ancestral components, based on whether alleles were identical-by-descent (IBD) for the first time or not, the new component was found to

be more harmful than the ancestral component. Based on SNP data, we did not find such clear differences in effects of recent and ancient inbreeding. When we considered the effect of regions of homozygosity (ROH), both long ROH (which reflect recent inbreeding) and short ROH (which reflect more ancient inbreeding) contributed to inbreeding depression. While computing inbreeding coefficients we also corrected the algorithm that is commonly used to calculate Kalinowski's coefficients (**Chapter 4**). In **Chapter 5**, we evaluated variation in inbreeding depression across the genome. We did so by estimating dominance and ROH-effects, either for one SNP at a time (through a single SNP GWAS) or for all SNPs simultaneously (through GREML with backsolving). Estimated dominance and ROH variances from GREML models were small, i.e. less than 1% of phenotypic variance. In both GWAS and GREML models, inbreeding depression appeared to be rather equally distributed across the genome and to be well captured by genome-wide homozygosity.

In **Chapters 6 and 7**, we demonstrated the value of the Dutch gene bank for HF and Dutch native cattle breeds. In **Chapter 6**, we showed that, although artificial selection has resulted in considerable genetic gains in the HF breed over time, (old) HF cryobank bulls can still be of value to today's breeding program. First, they can be used to increase genetic diversity in the current breeding program. When we minimized relatedness with OCS, the use of genotyped cryobank bulls decreased the mean SNP-by-SNP similarity by 0.7% and the use of both genotyped and non-genotyped cryobank bulls decreased mean pedigree-based kinship by 2.6% (in absolute terms). Second, cryobank bulls can be used to increase genetic merit for a given level of diversity. We showed that the additional value of cryobank bulls was higher when relatively more emphasis was put on genetic diversity. We also found that, although the additional value of cryobank bulls was limited for the current total merit index, it was substantial for sub-indices like fertility. Anticipating changes in breeding goals in future, we concluded that the gene bank collection is a valuable resource. In **Chapter 7**, we characterized genetic diversity in the gene bank across Dutch native cattle breeds. Besides the HF breed, 7 Dutch native cattle breeds are stored in the gene bank. We showed that, based on SNP data, Dutch native breeds genetically differed from the HF breed, suggesting they harbor unique genetic diversity. Among Dutch native breeds, there was admixture. Consequently, when we set up a 'core set' in which the expected heterozygosity was maximized through OCS, some breeds were assigned low contributions. Overall, our results show that gene bank collections are valuable resources, not only for small local breeds, but also for large commercial breeds.

In **Chapter 8**, the general discussion, I further addressed some major questions and opportunities for genomic management of genetic diversity. I showed that the

increase in ΔF and Δf with genomic selection in HF is a global trend and argued that the increase is not necessarily due to the methodology, but rather to a change in system. In addition, I discussed why the SNP-by-SNP similarity is an important measure for conservation of genetic diversity. Moreover, I discussed how OCS can be extended, providing additional conservation opportunities (e.g. for region-specific diversity management), and described the benefits of using sequence data. I also discussed how characterization of gene bank collections may enhance their utilization and how gene bank collections are expected to move towards bio-digital resource centers. Finally, I emphasized the importance of conservation of genetic diversity by discussing expected changes in future breeding goals.

Acknowledgements (Dankwoord)

Many people have contributed to this thesis. The cover, for example, was designed by 133 colleagues, friends and family members (this number would have been even higher if it wasn't for the intellectual lockdown). Hereby I would like to thank everyone who has supported me over the last years, in one way or another! A few (groups of) people I want to mention in particular, starting with my supervisors.

Jack, dat jij dagelijks begeleider zou worden was voor mij één van de redenen om voor deze PhD te gaan. Ik stel je vriendelijke en rustige persoonlijkheid erg op prijs, alsmede de manier waarop je studenten begeleidt. Gedurende het hele proces stond je deur altijd voor me open en bleef je vertrouwen houden in mij en mijn werk, wat me hielp me verder te ontwikkelen. Dank daarvoor! Roel, ook jou wil ik bedanken voor het vertrouwen dat je gaf. Wanneer ik zelf over mijn werk twijfelde, benadrukte jij dat mijn resultaten wél interessant waren. Daarnaast zorgden jij en Jack ervoor dat ik, als ik weer eens aan het zwemmen was in de methodes en details, me weer op de grote lijn ging richten (ookal bleef ik dan soms wat hangen). Sipke-Joost, dank voor al je input vanuit genenbank-perspectief en het actief betrekken bij CGN- en IMAGE-activiteiten. Ik waardeer je positieve houding en heb erg genoten van onze gesprekken (+ biertjes) tijdens o.a. IMAGE en EAAP meetings. Piter, je hulp met theorethische vraagstukken en scherpe opmerkingen hebben mijn werk veel goeds gedaan. Ik ben blij dat ik één letter promotie heb gemaakt en dat ik, nu IMAGE is afgerond, met je mag samenwerken binnen IMAGEN. Gerben, ik stel het zeer op prijs dat jij sinds het tweede jaar mijn begeleidingscommissie hebt versterkt. Bedankt voor je input vanuit de meer praktische kant van de rundveefokkerij.

In addition to my supervisors, I would like to acknowledge all other co-authors. Anouk, bedankt dat jij mijn eerste student wilde zijn en de uitdaging aan wilde gaan om samen met Kor (ook bedankt!), Jack en mijzelf een artikel te schrijven. Yvonne, dank voor de gezellige en nuttige brainstorm-sessies. Ino, István, János and György, thank you for your collaboration on the revision of GRain.

I am very grateful to all my colleagues from ABG and CGN for providing a great working atmosphere! I have enjoyed the discussions (QDG, TLM, CGN meetings), coffee breaks, days out, Sinterklaas events, end-of-year events, foosball games, drinks, dinners, Partycom events, conference trips and many other activities. A few people I want to mention by name, because they have particularly contributed to my work and personal development. Wilma en Lisette, bedankt voor al jullie administratieve ondersteuning. Ina, Mario en Jeremie, bedankt voor jullie hulp bij data- en software-gerelateerde vragen. Aniek, Henk, Martijn, Mirte, en Pascal, wat fijn dat jullie af en toe met me wilden sparren. John en Bart, bedankt dat jullie me hebben betrokken bij het onderwijs in Utrecht (was erg leuk om te doen!). Marieke, dankjewel voor de gezellige ADSA-week en voor het samen college geven op de Dairy campus. Malou, bedankt voor het proeflezen van mijn inleiding en discussie!

Daarnaast heb ik erg genoten van onze lunchwandelingen en bordspelletjes, wat later telefoon-wandelingen en online escape rooms werden i.v.m. thuiswerken. Thanks to all my office mates for the discussions, fun and daily hugs. Simion, Vinicius and Lisanne, it was great to get to know you. Meneer Onzima, you are such a friendly and warm person. I still smile every time I walk past your “geiten”-drawing in my apartment. Bedankt en tot ziens! Biaty, it was great to share experiences during the last stretches of our PhDs. Chiara, our paths have been so similar that some people even thought we were dating ;-)... Thanks for the fun and collaboration as fellow MSc-students, neighbours at work, neighbours at home, fellow PhDs in IMAGE and fellow WSD organizers. María, you are such a sweet, caring and energetic person. I loved discussing presentations and reviewers’ comments with you (and hope you realize now that some reviewers are more extreme “mierenneukers” than I am), as well as learning basic words in different languages, sharing feelings, and becoming friends. Finally, the Wageningen Sweater team, ai ni yo! Over the last years, you have become great friends and I am so grateful for our countless walks, hot pots, bīngjīlíng, movie nights (from Babe to Húli jīng), picnics, board games and day trips. You taught me some very useful sentences in Mandarin, including “wǒ bèi wén zi yǎo le yī ge bāo”. Shuwen, thank you for all your sweetness, being a dead animal, mastering Japanese garden puzzles like they were a PhD project and for laughing out loud while watching The Prince. Zhou (or should I say “boat” as you explained during our first WIAS course?), since your desk was opposite mine, you were generally the first who had to deal with my frustrations and questions. Thanks for being patient, and for showing that you cared by sharing food out of your endless supply. The topic of food brings me to Langqing, tengo hambre. You should really consider starting your own cooking show, since you obtained plenty of experience during the PhD ;-). Thanks for all the five o'clock foosball games and sharing beers and whisky. Zhou and Langqing, it was amazing to have you around for my entire PhD and I’m honoured that you accepted to be my paranymphs.

Besides my direct colleagues, I’ve met many nice and inspiring people during meetings, courses and conferences. A few of them I want to mention. John and Ricardo, I very much appreciate you hosting Jack and me in Edinburgh for a few days to discuss about my work and semi-definite programming. Anna-Charlotte, it was wonderful to get to know you and discuss our work during the IMAGE-course and EAAP (it’s a petty we couldn’t collaborate more in the end!).

Outside of the PhD, there are many friends that have supported me over the last years. Lieve Semprianen, wat ben ik blij dat ik in 2017 lid ben geworden! Het is heerlijk om na een dagje piekeren achter de computer even alles los te kunnen laten door samen te zingen, dansen en spelen. Ik heb onwijs genoten van alle repetities, voorstellingen, decordagen, kostuumdagen, bestuursvergaderingen, theater-uitjes,

after after after parties, muziekavonden (van barbershop tot Bach), kerstkoortjes, zeilweekenden, auto's shoppen, verjaardagen, BBQ's, Cicuto-bezoekjes, vakanties, oud en nieuw feestjes en andere activiteiten. En natuurlijk alle knuffels ☺. Jullie zijn stuk voor stuk toppers en velen van jullie zijn goede vrienden geworden. Bedankt voor de warmte en energie die jullie me de afgelopen jaren hebben gegeven! Teamgenootjes van HKC en KV Wageningen, korfbal is al leuk op zichzelf, maar nog veel leuker als je fijne mensen om je heen hebt! Dankjulliewel voor al het plezier op het veld, gedurende derde helften en tijdens teamuitjes. Whack Céol, ik geniet telkens weer van het samen muziek maken en kijk uit naar nog vele jamsessies en optredens. Vrienden van de Tafel, wat is het fijn dat we elkaar na tien jaar nog steeds niet uit het oog verloren zijn. Dank voor de gezelligheid tijdens verjaardagen en 5 mei feestjes. Dan zijn er nog vrienden die niet binnen één van bovengenoemde categorieën vallen; ook jullie heel erg bedankt voor de vele knuffels, wandelingen, museum-bezoekjes, filmavonden en spelletjes (zowel offline als online)!

Als laatste wil ik mijn familie, en met name mijn gezin, bedanken. Niemand van ons zal tien jaar geleden hebben verwacht dat ik een PhD zou gaan doen, maar blijkbaar valt de appel echt niet ver van de boom… Lieve papa en mama, betere ouders dan jullie kan ik me niet wensen. Jullie zijn zorgzaam, bieden altijd een luisterend oor en hebben me vele malen geholpen bij het maken van moeilijke keuzes voor mijn studie en loopbaan. Bedankt voor jullie onvoorwaardelijke steun. Papa, bedankt dat jij als kritische leek mijn inleiding en discussie wilde proeflezen. Lieve Hilje en Leo, wat was het fijn dat ik telkens weer van jullie B&B gebruik mocht maken wanneer ik 's ochtends in Utrecht moest zijn. Bedankt voor jullie gastvrijheid en voor al het plezier tijdens musiceren, vakanties en spelletjes (met Skype-pandemie als briljante uitvinding!). Grote zus, ik kan je niet genoeg bedanken voor al je hulp gedurende mijn school-, studie- en PhD-tijd. Ik ben onwijs trots dat we binnen één week allebei ons proefschrift mogen verdedigen en dat we inmiddels zelfs collega's zijn! Tot slot, lieve oma, bedankt voor alle mooie momenten samen. Deze zal ik nooit vergeten.

Curriculum vitae

About the author
Publications
Contributions to conferences
Training and education

About the author

Harmen Pieter Doekes was born on the 9th of August 1992 in Houten, the Netherlands. In 2014, he obtained his bachelor's degree in Animal Management from Van Hall Larenstein University of Applied Sciences. During his bachelor, Harmen specialized in societal and technical aspects of laboratory animal management, among others through an internship at the Netherlands Organization for Health Research and Development (ZonMw). In 2016, he obtained his MSc degree in Animal Sciences from Wageningen University & Research. During his MSc, Harmen specialized in animal breeding and genetics with two major research projects. In the first project, he analyzed the pedigree of two Dutch dog breeds at the Animal Breeding and Genomics group (ABG) of Wageningen University & Research. In the second project, he estimated genetic parameters for metabolic health indicators in dairy cattle at the Integrative Animal Sciences team of Scotland's Rural College (SRUC).

After graduation, Harmen started as a PhD-candidate at the ABG group. His project was part of the European consortium Innovative Management of Animal Genetic Resources (IMAGE) and results of this project are presented in this thesis. Since June 2020, Harmen works as a researcher/lecturer at the ABG group of Wageningen University & Research.

Publications

Doekes HP, Veerkamp RF, Bijma P, Hiemstra SJ, and Windig JJ. 2018. Trends in genome-wide and region-specific genetic diversity in the Dutch-Flemish Holstein-Friesian breeding program from 1986 to 2015. Genet Sel Evol. 50:15.

Doekes HP, Veerkamp FR, Bijma P, Hiemstra SJ, and Windig JJ. 2018. Value of the Dutch Holstein-Friesian germplasm collection to increase genetic variability and improve genetic merit. J Dairy Sci. 101:10022-33.

Windig JJ, and Doekes HP. 2018, Limits to genetic rescue by outcross in pedigree dogs. J Anim Breed Genet. 135(3):238-248.

Onzima RB, Upadhyay MR, Doekes HP, Brito LF, Bosse M, Kanis E, Groenen MA, and Crooijmans RP. 2018. Genome-wide characterization of selection signatures and runs of homozygosity in Ugandan goat breeds. Front Genet. 9:318.

Doekes HP, Veerkamp RF, Bijma P, de Jong G, Hiemstra SJ, and Windig JJ. 2019. Inbreeding depression due to recent and ancient inbreeding in Dutch Holstein-Friesian dairy cattle. Genet Sel Evol. 51:54.

van Breukelen AE, Doekes HP, Windig JJ, and Oldenbroek K. 2019. Characterization of genetic diversity conserved in the gene bank for Dutch cattle breeds. Diversity. 11(12):229.

Doekes HP, Curik I, Nagy I, Farkas J, Kövér G, and Windig JJ. 2020. Revised calculation of Kalinowski's ancestral and new inbreeding coefficients. Diversity. 12(4):155.

Doekes HP, Bijma P, Veerkamp RF, de Jong G, Wientjes YCJ, and Windig JJ. Inbreeding depression across the genome of Dutch Holstein Friesian dairy cattle. *Submitted*.

Contributions to conferences

Doekes HP, Veerkamp RF, Bijma P, Hiemstra SJ, and Windig JJ. 2017. Trends in genome wide and region specific diversity reflect Holstein Friesian breeding history. 68th Annual Meeting of the European Association of Animal Production (EAAP). Tallinn, Estonia.

Doekes HP, Veerkamp RF, Hiemstra SJ, Bijma P, van der Beek S, and Windig JJ. 2018. Genomic selection and inbreeding and kinship in Dutch-Flemish Holstein-Friesian cattle. 11th World Congress on Genetics Applied to Livestock Production (WCGALP). Auckland, New Zealand.

Doekes HP, Veerkamp RF, Bijma P, Hiemstra SJ, de Jong G, and Windig JJ. 2019. Inbreeding depression due to recent and ancient inbreeding in Dutch Holstein Friesian dairy cattle. WIAS Science Day. Lunteren, the Netherlands.

Doekes HP, Veerkamp RF, Bijma P, Hiemstra SJ, de Jong G, and Windig JJ. 2019. Not all inbreeding is depressing. Annual meeting of the American Dairy Science Association (ADSA). Cincinnati, USA.

Doekes HP, Veerkamp RF, Bijma P, Hiemstra SJ, de Jong G, and Windig JJ. 2019. Not all inbreeding is depressing. 70th Annual Meeting of the European Association of Animal Production (EAAP). Ghent, Belgium.

van Breukelen AE, Doekes HP, and Oldenbroek K. 2019. Genetic diversity and inbreeding in Dutch cattle breeds based on gene bank collections. 70th Annual Meeting of European Association of Animal Production (EAAP). Ghent, Belgium.

Windig JJ, Doekes HP, Veerkamp FR, Bijma P, and Hiemstra SJ. 2019. Value of gene bank material for the commercial breeding population of Dutch Holstein Friesian cattle. 70th Annual Meeting of European Association of Animal Production (EAAP). Ghent, Belgium.

Doekes HP, Veerkamp FR, Bijma P, Hiemstra SJ, and Windig JJ. 2020. Trade-offs between genetic diversity and genetic merit when using gene bank bulls. International Conference on Animal Genetic Resources. Madrid, Spain.

Training and education

The Basic Package (3.0 ECTS)	
WIAS Introduction day	2016
WIAS course Research Integrity & Ethics	2016
WIAS course on Essential skills	2017

Disciplinary Competences (17.0 ECTS)	
Writing WIAS research proposal	2016
Genetic improvement of livestock (Wageningen)	2016
Theory and application of inbreeding management (Ås, Norway)	2017
Discussion groups (quantitative genetics, Fortran)	2016-2020
Design of breeding programs with genomic selection (Wageningen)	2017
Interactive post-graduate course on characterization, management and exploitation of genomic diversity in animals (Wageningen)	2018

Professional competences (6.3 ECTS)	
WGS course Brain training	2017
WGS course Interpersonal Communication for PhD candidates	2017
Organization WIAS Science day 2018	2017-2018
WGS course Techniques for scientific writing and presenting	2018
WGS course Teaching and supervising thesis students	2018
WIAS course The Final Touch	2019

Presentation Skills (maximum of 4.0 ECTS)	
EAAP, Tallinn, Estonia (Oral)	2017
IMAGE meeting, Göttingen, Germany (Oral)	2017
WCGALP, Auckland, New Zealand (Oral)	2018
IMAGE meeting, Vienna, Austria (Oral)	2018
WIAS Science Day, Lunteren, the Netherlands (Oral)	2019
ADSA, Cincinnati, USA (Oral & Poster)	2019
IMAGE meeting, Bréscia, Italy (Oral)	2019
EAAP, Ghent, Belgium (Oral)	2019
Conference on Animal Genetic Resources, Madrid, Spain (Oral & Poster)	2020

Teaching competences (maximum of 6.0 ECTS)	
Assisting BSc-course Animal Breeding and Genetics	2017
Assisting MSc-course Genomics	2018
Teaching BSc-course on genetics at Utrecht University (3 times)	2018-2020
Supervising MSc thesis student (2 times)	2018-2019
Supervising BSc thesis student	2019

Total: 36.3 ECTS

Colophon

The research described in this thesis has been conducted as part of the IMAGE project, which received funding from the European Union's Horizon 2020 Research and Innovation Program under the grant agreement no. 677353. It was co-funded by the Dutch Ministry of Agriculture, Nature and Food Quality (KB-34-013-002). Data were provided by the Dutch-Flemish cattle improvement cooperative (CRV) and by the Centre for Genetic Resources, the Netherlands (CGN) of Wageningen University & Research.

The cover of this thesis was designed by a diverse group of 133 colleagues, friends and family members.

Printed by: DigiForce | Proefschriftmaken.nl